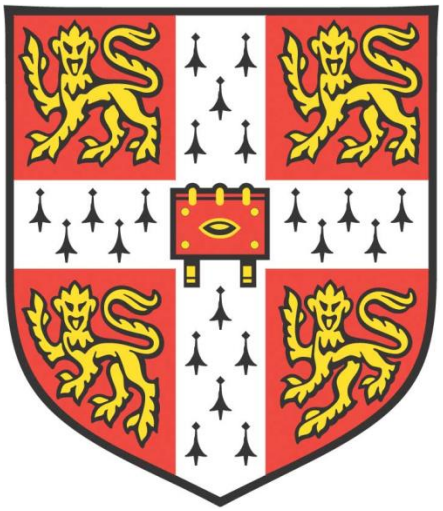
PLASTID TO NUCLEUS SIGNALLING AND THE EVOLUTION OF C_4 PHOTOSYNTHESIS



Clare Hall

Department of Plant Sciences
University of Cambridge

This thesis is submitted for the degree of Doctor of Philosophy
November 2020

DECLARATION

- This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as specified in the text.
  I further state that no substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as specified in the text.
- It does not exceed the prescribed word limit for the School of Biological Sciences Degree Committee.

Plastid to Nucleus Signalling and the Evolution of C_4 Photosynthesis

Robyn Lynch Phillips

ABSTRACT

Plastid to nucleus signals have been shown to regulate core photosynthetic genes, in both C_3 and C_4 species. However, the regulatory mechanisms of this signalling are not well understood. C_4 photosynthesis has evolved from the ancestral C_3 state in over sixty lineages of plants. Although commonly described as the most remarkable example of convergent evolution known to biology, more recent analysis indicates that the C_4 pathway is also underpinned by parallel evolution. For example, orthologous genes from separate C_4 lineages have repeatedly been recruited into the C_4 pathway such that they are co-regulated with existing genes of C_3 photosynthesis. For decades, the mechanisms allowing co-regulation of C_4 and C_3 photosynthesis genes have been unclear, but it was recently shown that in C_3 *Arabidopsis thaliana* C_4 orthologues can be controlled by plastid-to-nucleus signalling (Burgess et al. 2016). This strongly implies that evolution has re-enforced existing regulatory networks that operate in the C_3 state to control expression of C_4 genes. Currently, the extent to which this phenomenon is true in additional C_3 lineages is not known. The work presented here aimed to understand in more detail which genes are under plastid control in a eudicotyledon and monocotyledon model and identify the factors involved.

To investigate this, *Arabidopsis thaliana* and *Oryza sativa* were subjected to inhibitors of chloroplast development and differential gene expression analysis performed by mRNA sequencing. Comparisons between illuminated and dark-grown plants for each treatment enabled the complementary roles of plastid-to-nucleus signalling and light regulation to be identified. Over 20,000 genes were detected for each sample, with over 1000 genes differentially expressed in lincomycin-treated plants compared to controls, and over 3500 genes differentially expressed when plants were treated with norflurazon. Furthermore, over half of the C_4 orthologues in these two C_3 species were regulated by both light and the chloroplast. Genes whose transcript abundance was significantly affected by chloroplast inhibition were scanned for known motif binding sites. A number of putative transcription factors were identified that bind to motifs found in both *A. thaliana* and rice chloroplast responsive genes. An additional method for identifying potential *cis*-regulatory elements involved in plastid to nucleus signalling is ATAC-seq. A trial has allowed this method, including the analysis pipeline, to be optimised such that in the future it could be employed to create an atlas of transcription factor binding sites involved in plastid to nucleus signalling.

ACKNOWLEDGEMENTS

I would like to thank my supervisor Julian Hibberd for his guidance. He is the rare supervisor who can give immense insight, swift feedback all whilst showing a great level of patience and understanding. He made this work not only possible, but enjoyable. My thanks to the Hibberd lab for always being a friendly and welcoming environment and giving valuable feedback during lab meetings and discussions, particular thanks to Sean and Patrick for all their help with computational analyses, and Pallavi for her never-ending generosity.

I have a great deal of gratitude for my mother, Lynn Phillips, for instilling in me a passion for education and scientific discovery. She has always been my greatest support, and I'm thankful for all the many pages she has proofread. Thanks for the support of my grandmother, Jean Phillips, who despite never knowing what I was on about was always very encouraging. My fiancé Joe for never failing to cheer me up and for making lockdown much more enjoyable. Finally, our cavachon Dylan for his support via cuddles and Croutons the cat for many in-depth conversations.

CONTENTS

LIST OF TABLES

List of Figures

LIST OF ABBREVIATIONS AND ACRONYMS

ATAC-seq	Assay of Transposase Accessible Chromatin sequencing
ATP	Adenosine TriPhosphate
bp	base pairs
CA	Carbonic Anhydrase
ChIP-Seq	Chromatin ImmunoPrecipitation sequencing
DAP-seq	DNA Affinity Purification sequencing
DHS	DNAseI Hypersensitive Site
DNase-seq	DNase I sequencing
DTA	Differential Transcript Abundance
Ex1	Executor1
FAIRE-seq	Formaldehyde-Assisted Isolation of Regulatory Elements
FC1	FerroChelatase 1
FU	Fluorescence Units
F_v/F_m	Maximum quantum yield of PSII
GO	Gene Ontology
gun	genome uncoupled
LHCA	Light-Harvesting Chlorophyll A binding protein
LHCB	Light-Harvesting Chlorophyll B binding protein
MEcPP	2-C-Methyl-Erythritol-2,4-cycloPyroPhosphate
MEM1	Mesophyll-Enhancing Module
MEP	MethylErythritol Phosphate
MNase-seq	Micrococcal Nuclease digestion with deep sequencing
mRNA	Messenger RNA
NAD-ME	NAD-Malic Enzyme
NADP-ME	NADP-Malic Enzyme
NASC	Nottingham Arabidopsis Stock Centre
Ns	unassigned bases
OAA	OxaloAcetic Acid
PAP	3′-PhosphoAdenosine 5′-Phosphate
PCA	Principal Components Analysis
PCNA	Proliferating Cell Nuclear Antigen
PCR	Polymerase Chain Reactions
PEPC	PhosphoEnolPyruvate Carboxylase
PEPCK	PhosphoEnolPyruvate CarboxyKinase
PhANGs	Photosynthesis Associated Nuclear Genes
PPC	PhosphoenolPyruvate Carboxylase
PPDKA	Pyruvate OrthoPhosphate DiKinase A
PPR	PentatricoPeptide Repeat
PRANGs	Plastid Redox Associated Nuclear Genes
PSI	PhotoSystem I
PSII	PhotoSystem II
PWM	Position Weight Matrix.
QC	Quality Control
RAP-DB	The Rice Annotation Project DataBase
RIN	RNA Integrity Numbers
ROS	Reactive Oxygen Species
RS	Retrograde Signalling
RSRS	Rapid Stress Response Element

RuBisCO	Ribulose-1,5-Bisphosphate Carboxylase/Oxygenase
SORGs	Singlet Oxygen-Responsive Genes
TAIR	The Arabidopsis Information Resource
THS	Transposase Hypersensitive Site
TIC	Translocon of the Inner Chloroplast
TOC	Translocon of the Outer Chloroplast
TPM	Transcripts Per Million
TSS	Transcription Start Site

List of Appendices

1 INTRODUCTION

1.1 The Evolution of C_3 Photosynthesis

The evolution of oxygenic photosynthesis changed the atmosphere of our planet and facilitated the evolution of a diverse range of eukaryotes (Fischer et al., 2016). Oxygenic photosynthesis is believed to have originated in bacteria and required the evolution of two reaction centres to harvest light energy, referred to as Photosystem I and Photosystem II (Blankenship, 2014). Photosystem I evolved from a reaction centre that reduces ferredoxin to produce NADH, such a reaction centre can be found within the Chlorobi, Firmicutes and Acidobacteria. Photosystem II evolved from a reaction centre that reduces quinones to produce NADH, such a reaction centre can be found within the Proteobacteria, Chloroflexi and Gemmatimonadetes (Cardona, 2015). Bacteria which have only one reaction centre carry out anoxygenic photosynthesis. Cyanobacteria contain both PSI and PSII, possibly attained via horizontal gene transfer, and are believed to be the origin of oxygenic photosynthesis (Mulkidjanian et al., 2006). It is theorised that the chloroplasts of algae and land plants are derived from a cyanobacterium engulfed by a eukaryotic cell. There has been debate as to whether this endosymbiosis event occurred once or multiple times, and the shopping bag theory predicts that varied levels of gene transfer occurred during a series of endosymbioses (Howe et al., 2008). Around 18% of the nuclear genome in *A. thaliana* is predicted to derive from Cyanobacteria (Martin et al., 2002).

Photosystem I and Photosystem II are integral components of the photosynthetic electron transport chain, also known as the light-dependent reactions, which generate ATP and NADPH (Berg et al., 2002). Energy from photons is used by PSII to split water into hydrogen and oxygen. Oxygen is released and the electrons generated are passed to plastoquinone, to the Cyt b6f complex, and then to PSI where a photon further excites the electron. Ultimately the electron reduces NADP and H^+ to produce NADPH. The electron transport chain pumps protons into the thylakoid lumen to generate a proton gradient which is utilised by ATP synthase to generate ATP. The NADPH and ATP are then used in the light independent reactions to drive the Calvin Benson Bassham cycle. The Calvin Benson Bassham cycle is initiated when CO_2 is fixed by Ribulose-1,5-Bisphosphate Carboxylase/Oxygenase (RuBisCO) to produce 3-phosphoglycerate, a three-carbon compound (Calvin, 1956). The cycle regenerates RuBP from 3-phosphoglycerate through a series of reactions utilising ATP, however, every sixth molecule of 3-phosphoglycerate generated by the cycle is diverted for other functions such as sucrose or starch synthesis. Photorespiration occurs as a result of RuBisCO binding oxygen rather than CO_2 (Bauwe et al., 2010). This is considered detrimental as it produces 2-phosphoglycolate, a toxic compound which inhibits several enzymes in the Calvin Benson Bassham cycle. Moreover, to regenerate RuBisCO from 2-phosphoglycerate requires the dephosphorylation of ATP and the loss of CO_2.

1.2 C_4 Photosynthesis

C_4 photosynthesis is a carbon concentrating mechanism which typically decreases photorespiration by separating CO_2 assimilation into two cell types. Photorespiration occurs when RuBisCO, the primary carboxylase associated with photosynthesis, binds oxygen rather than CO_2 (Bauwe et al., 2010). To decrease photorespiration, many C_4 species have evolved Kranz anatomy where the bundle sheath cells closely surround the vein, and the mesophyll cells form an outer layer around the bundle sheath. Chloroplasts typically differ in both their morphology and protein complement between the two cell types, with a greater number of larger chloroplasts found in bundle sheath cells compared to mesophyll cells (Sedelnikova et al., 2018). In addition to changes in the cellular localisation and abundance of classical photosynthesis proteins in C_4 plants, an additional biochemical pathway operates alongside C_3 photosynthesis to deliver high concentrations of CO_2 to RuBisCO (Hatch, 1987).

In the most common form of C_4 photosynthesis, carbon capture is initiated in the mesophyll cells where carbonic anhydrase (CA) generates HCO_3 which can then be fixed by phospho*enol*pyruvate carboxylase (PEPC), to generate the C_4 acid oxaloacetate. Unlike RuBisCO, PEPC has no oxygenase activity. The C_4 metabolites produced are then shuttled to the bundle sheath and decarboxylated producing tenfold higher concentrations of CO_2 than are found in mesophyll cells and so oxygenation is reduced. C_4 photosynthesis has evolved independently over 60 times, and there is variation in the biochemical pathway across species. The C_4 metabolite moving between the mesophyll and bundle sheath can differ, being either malate or aspartate, and the enzyme mainly responsible for decarboxylating this C_4 metabolite in the bundle sheath can also vary between C_4 species. For example, subtypes use different combination of NADP-malic enzyme (NADP-ME), NAD-malic enzyme (NAD-ME), or phospho*enol*pyruvate carboxykinase (PEPCK) (Schlüter & Weber, 2020). The evolution of C_4 photosynthesis requires genes involved in this carbon concentrating pathway to become upregulated compared to their C_3 ancestors.

In addition to the evolution of C_4 photosynthesis requiring upregulation of C_4 gene expression, it also needs changes in spatial regulation of genes. The existence of so-called C_3-C_4 intermediates provides more insight into how these changes in regulation may have occurred. For example, species of *Flaveria* show increased expression of *PEPC* from C_3 to C_3-C_4 intermediates to fully C_4 species (Engelmann et al., 2003). In this case, expression is thought to be upregulated because of increased trimethylation of histone H3K4 in the *PEPC* promoter (Heimann et al., 2013). In addition to being upregulated, gene expression of *PEPC* becomes mesophyll specific in C_4 species. A region known as the mesophyll-enhancing module (MEM1), within the *PEPC* promoter was identified as being necessary for mesophyll expression, suggesting that spatial regulation was controlled, to some degree, by transcriptional regulation (Gowik et al., 2004). Changes in spatial regulation can also occur within the cell. CA which is also important for carbon fixation in mesophyll cells, becomes abundant in the cytosol in C_4 species through the loss of a plastid transit peptide (Ludwig, 2012). These sorts of changes also take place in the bundle sheath. For example, increased expression of enzymes of the Calvin Benson Bassham cycle has been reported in species including maize (Denton et al., 2017) and *G. gynandra* (Aubry et al., 2014). In some C_4 species, RuBisCO activity is limited to the bundle sheath largely by posttranscriptional regulation (Berry et al., 2016). In summary, whilst it has been known for some time that compared with the ancestral C_3 state, C_4 genes must become more strongly expressed and compartmented to specific cells in C_4 leaves, the mechanisms initiating these responses have been poorly characterised.

1.3 The Role of Light and Chloroplast Signalling in C_4 Photosynthesis

C_4 photosynthesis is hypothesized to have arisen over 60 times in flowering plants (Sage et al., 2011) making it one of the most remarkable examples of convergent evolution across the tree of life. The propensity for plants to evolve this adaptation could be because C_3 plants contain all the core C_4 genes, and they are simply rewired into existing gene regulatory networks. In the ancestral C_3 state these genes have functions in pathways unrelated to the chloroplast such as gluconeogenesis and ammonium assimilation (Eastmond et al., 2015; Masumoto et al., 2010). In contrast, in C_4 leaves, C_4 genes must become tightly integrated into photosynthesis gene regulatory networks and therefore regulated by light and chloroplast to nucleus signalling (Hibberd & Covshoff, 2010). It is not clear if these genes are loosely linked to these processes in C_3 ancestors. Thus the mechanisms by which these genes are recruited into the C_4 pathway remain unclear, and it has yet to be established how these genes become co-regulated with genes from the C_3 pathway.

The 60 independent origins of C_4 photosynthesis provide natural replicates for researchers to understand how C_4 photosynthesis arose. Phylogenetic studies have shown parallel evolution could also underpin the repeated recruitment of C_4 genes (Williams et al., 2012). Despite enzymatic activity being conserved within gene families, the same orthologues from multi-gene families have repeatedly been recruited during the evolution of the C_4 trait (Christin et al., 2013) suggesting that although C_4 photosynthesis appears to have arisen independently, constraints may exist that require it to follow the same evolutionary path. If function is not responsible for this, it is possible that a shared mechanism of genetic regulation could be.

Consistent with this latter notion, evidence has found facets of C_4 regulation can be found within the ancestral C_3 state (Reyna-Llorens & Hibberd, 2017). Regulatory elements are conserved between divergent C_4 species and are present in the ancestral C_3 state in both monocotyledon and eudicotyledons (Brown et al., 2011). Elements sufficient for mesophyll cell specificity in C_4 species are also present in both the 5′ and 3′ UTRs of *PPDKA* and *CA* in *A. thaliana* (Kajala et al., 2012). Recently, this has shown to be true for magnitude of gene expression as well as cell specificity. A number of C_4 orthologues are controlled by plastid-to-nucleus signalling in C_3 *Arabidopsis thaliana* (Burgess et al., 2016). This strongly implies that evolution has re-enforced existing regulatory networks that operate in the C_3 state to control expression of C_4 genes.

1.4 Light and retrograde signalling in the regulation of photosynthesis

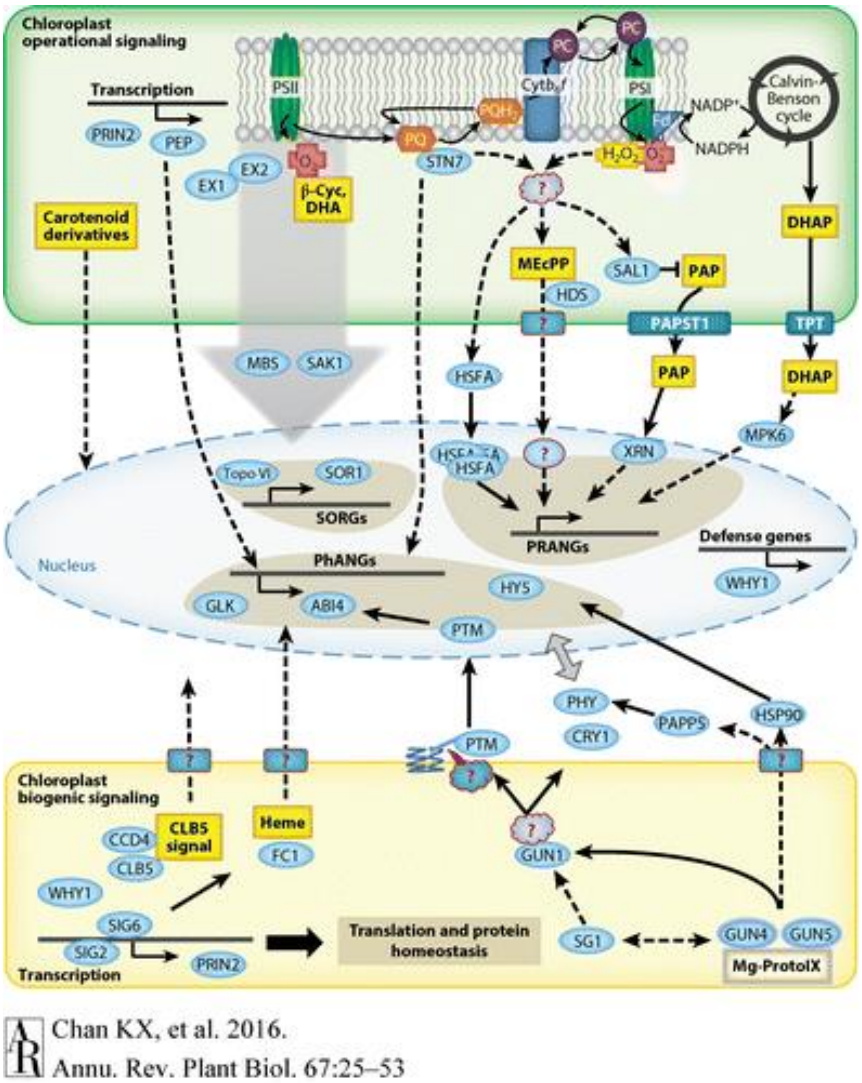
The vast majority of photosynthetic activity in land plants is associated with leaves. This is due to the high density of chloroplasts where this fundamental process takes place. The chloroplast is an organelle thought to have originated from a free-living cyanobacteria engulfed by a eukaryotic cell (Keeling, 2010). The chloroplast genome, also known as the “plastome”, no longer contains all the genes associated with a free-living organism and it is theorised that the majority of genes were either lost or transferred to the nucleus shortly after the symbiosis was established. Under strong selective pressures the transfer of genetic material from chloroplast to nucleus has been demonstrated (Stegemann et al., 2003) and in parasitic plants the chloroplast genome decreases even further (Wolfe et al., 1992). Despite the loss and transfer of genes a small number (~100) remain in the chloroplast. A number of hypotheses have been proposed to explain the retention of genetic material in the chloroplast. These include hydrophobicity of the proteins they encode, the need for colocation of gene and gene product for redox regulation of gene expression (the CoRR hypothesis) and the requirement for a chloroplast encoded glutamyl-tRNA essential for tetrapyrrole biosynthesis (Allen, 2015;

Barbrook et al., 2006; Selosse et al., 2001). Thus, photosynthesis is associated with both nuclear and chloroplast genomes.

Due to this symbiotic relationship of the chloroplast and nuclear genome, protein complexes in chloroplasts are composed of proteins encoded by both the nucleus and plastome (Abdallah et al., 2000). These protein complexes are involved directly in photosynthesis but also in processes indirectly associated with photosynthesis such as plastid maintenance. For example, Photosystem I (PSI), Photosystem II (PSII) and enzymes of the Calvin Benson Bassham cycle are all required for oxygenic photosynthesis, whereas chloroplast ribosomes allow assembly of proteins into the chloroplast (Pfannschmidt, et al., 2001). The movement of many genes to the nucleus, but the retention of some genes in the plastid, has created a requirement for the assembly of complexes containing both plastome-encoded and nuclear-encoded proteins. For protein complexes, or indeed proteins that simply collaborate in a metabolic pathway, communication between the nucleus and the chloroplast is required.

Communication between the nucleus and the chloroplast is bilateral. Anterograde signalling involves signals that emanate from the nucleus and are perceived by the chloroplast (Bräutigam et al., 2007). In contrast, retrograde signalling (RS) involves information flow from chloroplast to nucleus. Although RS was traditionally viewed as communication between nucleus and chloroplast and allowed correct assembly of chloroplast protein complexes, its role in regulating homeostasis in the mature plastid is now also recognised (Chan et al., 2016). In addition, retrograde signalling can have effects beyond the chloroplast, interacting with broader signalling networks and causing cell-wide stress responses and even cell death (Kim et al., 2012). RS has been further split into two categories (Figure 1.1) known as biogenic and operational signalling. Biogenic signalling allows chloroplast development whilst operational signalling allows the regulation of nuclear genes in response to an environmental change in mature chloroplasts.

There is a growing understanding of the complexity of retrograde signalling (RS) and the signalling pathways with which it interacts. Of the pathways elucidated so far, many are linked to light regulatory networks. Light is an essential requirement for photosynthesis, so its integral role in chloroplast development is unsurprising. For example, cryptochromes and phytochromes have long been known to be involved in the integration of plastid and light signals (Jiao et al., 2007). Many light responsive pathways act through G-box promoter elements, which are common to Photosynthesis Associated Nuclear Genes (PhANGs) (Larkin & Ruckle, 2008). Light can also act as an indirect inducer of retrograde signalling. For example, high light intensity increases production of ROS including singlet oxygen in the chloroplast. Singlet Oxygen-Responsive Genes (SORGs) have been characterised as a group that RS commonly targets during operational signalling (Gonzalez-Perez et al., 2011). Whether light directly impacts plastome expression or interacts downstream of initial retrograde signals is unclear (Finster et al., 2013; Liang et al., 2016).



Chan KX, et al. 2016.
Annu. Rev. Plant Biol. 67:25–53

Figure 1.1: Operational and Biogenic retrograde signalling pathways, and their nuclear targets grouped into PhANGs (Photosynthesis Associated Nuclear Genes, SORGs (Singlet Oxygen-Responsive Genes) and PRANGs (Plastid Redox Associated Nuclear Genes). Dashed lines show mechanisms not yet known and question marks denote unidentified genes. This diagram is not comprehensive as there are still features of chloroplast development and function that have not yet been elucidated. Image taken from Chan et al. (2016).

1.5 Operational Retrograde Signalling

To date it appears that operational signalling acts on Singlet Oxygen-Responsive Genes (SORGs) and Plastid Redox Associated Nuclear Genes (PRANGs). Reactive Oxygen Species (ROS) can affect both, with SORGs being directly upregulated by the presence of singlet oxygen and PRANGs affected by the change in ROS leading to altered redox poise within the plastid (Chan et al., 2016; Muthuramalingam et al., 2013). As increased ROS can be highly damaging and cause destruction of DNA, proteins and other important cellular components, rapid detection is desirable. Singlet oxygen can be detected by signalling components dependent on either β-cyclocitral or Executor proteins. The exact mechanism of action of β-cyclocitral, which is derived from oxygenation of carotenoids, is unknown. However, it has been shown to prevent a decrease in photosynthetic efficiency during cold or light stress (Ramel et al., 2012). Executor proteins are thought to repress defence responses under ambient conditions (Uberegui et al., 2015). Regulation by Executor proteins has impacts beyond

chloroplast function and can in some instances even lead to programmed cell death (Lee et al., 2007). Executor1 (Ex1) has been shown to 'sense' singlet oxygen through post translational modification, through the oxidation of the Ex1 DUF3506 domain by singlet oxygen (Dogra et al., 2019).

2-C-Methyl-Derythritol-2,4-cyclopyrophosphate (MEcPP) is a metabolite of the methylerythritol phosphate (MEP) pathway that allows synthesis of isoprenoids and is also proposed to act as an operational retrograde signal. MEcPP accumulates in response to stress such as wounding and acts as a global stress sensor (Xiao et al., 2012), for example, upregulating synthesis of the hormone salicylic acid and enhancing resistance to *Pseudomonas syringae*. MEcPP impacts nuclear gene expression via a *cis*-element termed the rapid stress response element (RSRS) (Benn et al., 2016). This suggests that a high throughput system for identifying nuclear *cis*-elements related to plastid affected genes could identify important regulators in RS pathways. Another metabolite, 3′-phosphoadenosine 5′-phosphate (PAP) acts in a retrograde signalling pathway that responds to high light and stress (Estavillo et al., 2011), PAP accumulates during stress and prevents downstream inhibition of stress related nuclear gene expression. PAP levels are regulated by the enzyme SAL1 which degrades PAP to AMP.

1.6 Biogenic Retrograde Signalling

Biogenic retrograde signalling (RS) occurs during development, and coordinates chloroplast biogenesis. Chloroplast biogenesis is typically initiated in the dark when proplastids in leaf primordia differentiate into etioplasts, which contain chlorophyll precursors and immature thylakoid membranes. Upon exposure to light etioplasts develop into mature fully functional chloroplasts (Pogson & Albrecht, 2011). This process has been shown to complete hours after illumination (Rudowska et al., 2012). Despite biogenic signalling often being defined exclusively as regulation during biogenesis (Chan et al., 2016), it is also often used to describe regulation that occurs in developing leaves. For example, many experiments sample plants that have been illuminated for days (Martín et al., 2016) where mature chloroplast division and leaf development, rather than biogenesis, are likely to be the predominant activities of photosynthetic cells.

Biogenic retrograde signalling mainly affects photosynthesis associated genes (PhANGs), and much of our current understanding is derived from work on *genome uncoupled (gun)* mutants. In wildtype plants, chloroplast inhibition causes a decrease in expressions of PhANGs, including genes encoding the light harvesting complexes *LHCB2.1* and *LHCB1.2* (Kacprzak et al., 2019a). The C_4 orthologue, *Carbonic Anhydrase (CA1)*, also shows a sharp decline in transcript abundance in response to chloroplast inhibition in C_3 and C_4 species. In contrast, in *gun* mutants PhANG expression is maintained, albeit at a slightly reduced level, after chloroplast inhibitors have been applied, implying that retrograde signalling pathways have been disrupted (Larkin et al., 2014; Susek et al., 1993). This approach therefore represents an important tool for phenotyping putative retrograde signals, as mutants in important transcription factors may show a *gun* phenotype in the chloroplast responsive genes they regulate.

Six *gun* mutants have been described to date. The functional *GUN2-6* alleles encode proteins involved in tetrapyrrole biosynthesis. Tetrapyrroles contain four rings each containing four carbon atoms and one nitrogen atom, and there are four classes within plants; chlorophyll, heme, siroheme, and phytochromobilin (Tanaka & Tanaka, 2007). GUN2, GUN3 and GUN6 affect heme synthesis which activates expression of PhANGs (Susek et al., 1993; Woodson et al., 2011). GUN4 and GUN5 affect Mg-proto IX, a chlorophyll precursor (Larkin, 2003; Mochizuki et al., 2001). The role for GUN1, a pentatricopeptide repeat (PPR) protein, has not been so clear and it has been referred to as a 'jack of

all trades' (Colombo et al., 2016). GUN1 was found to be necessary for a two stage corrective response observed in a slow greening mutant *cue8*, where the activity of nuclear encoded polymerases increases whilst activity of plastid encoded polymerases decreases compared to wildtype (Loudya et al., 2020). The mechanism of GUN1 has now been explicitly linked to tetrapyrrole synthesis, shown to directly bind heme and affect ferrochelatase 1 (FC1) activity (Shimizu et al., 2019). There has been further support for the importance of FC1 in retrograde signalling, with *FC1* overexpression within the chloroplasts of wildtype plants treated with chloroplast inhibitors rescuing *LHCB* expression, showing a *gun*-like phenotype (Page et al., 2020).

HY5, which is ubiquitinated and degraded in the dark, binds to G-box elements to promote gene expression of PhANGs in the light (Martín et al., 2016). However, *HY5* expression is antagonistic to that of *ABI4*, a transcription factor which inhibits PhANGs. *ABI4* has been proposed to be regulated through retrograde signalling via the proteolytic cleavage of a PHD-type transcription factor with transmembrane domains (PTM) located in the chloroplast envelope (Sun et al., 2011) but the role of both ABI4 and PTM as retrograde signals has now been refuted (Kacprzak et al., 2019a; Page et al., 2017). More promising retrograde signals have been proposed, including singlet oxygen. Treatment with far red light causes an accumulation of singlet oxygen and subsequently the inhibition of select nuclear genes (Page et al., 2017a). It is proposed that singlet oxygen acts as part of a distinct retrograde signalling pathway that represses nuclear gene expression, which in combination with the activating heme pathway can finetune chloroplast control. Whichever signals are acting during chloroplast biogenesis, mathematical analysis of nuclear gene expression suggests that retrograde signalling only impacts a distinct second phase of chloroplast biogenesis and not its initiation (Dubreuil et al., 2018).

1.7 Chloroplast Retrograde Signalling in Monocots

Whilst most recent research into chloroplast retrograde signalling has been carried out on *A. thaliana*, much of the initial experimental work investigated the system in monocotyledons. Genome uncoupled (*gun*) mutants, where photosynthesis associated nuclear gene (PhANG) expression is unchanged after chloroplast inhibition, were named in *A. thaliana* (Susek et al., 1993). This discovery was built upon previous work that established PhANG expression was disrupted in plants whose chloroplasts were perturbed. Bradbeer et al (1979) identified barley mutants that did not accumulate pigment and found two plastid enzymes synthesised in the cytoplast were no longer present in these plants. Harpster et al (1984) similarly identified maize mutants which did not green, and found that compared with wild-type, these mutants did not accumulate high levels of transcripts in light-harvesting chlorophyll a/b binding protein (*LHCB*) and other photosynthesis associated genes. The same was found by Hess et al (1994) in barley where mutants absent in chloroplast ribosomes failed to accumulate light-harvesting chlorophyll a/b binding protein, as well as nitrate reductase.

Work in barley expanded from chloroplast development mutants to the application of chloroplast inhibitors. Batschauer et al (1986) found treatment with norflurazon inhibited accumulation of transcripts encoding the light harvesting chlorophyll a/b protein. This work also identified mutants where *LHCB* accumulated to wildtype levels despite chloroplast perturbation, a *gun* mutant in everything but name. These chlorophyll-deficient *xantha* mutants are blocked in the tetrapyrrole biosynthesis pathway after protoporphyrin IX or Mg-protoporphyrin, similar to *gun4* and *gun5* in *A. thaliana*. Norflurazon's mode of action is to block carotenoid synthesis. Burgess & Taylor (1988) grew carotenoid deficient mutants in low light levels to limit chloroplast damage and found *LHCB* levels were unaffected, indicating that carotenoids do not specifically play a role in plastid to nucleus signals, and clarifying that it is norflurazon's perturbation of chloroplast function that affects PhANG

expression. *LHCB* expression was found to be inversely proportional to thylakoid development in maize (Rocca et al., 2000) and barley (La Rocca et al., 2000).

More recent studies have looked at the effect of chloroplast inhibition by the herbicide norflurazon on rice seedlings. Park & Jung (2017) showed 3 week old rice seedlings after 40 hours of a spray norflurazon treatment had decreased maximum quantum yield (Fv/Fm), a proxy for PSII activity, and the abundance of some LHCA (LHCA1, LHCA2, LHCA3 and LHCA4) and LHCB (LHCB1, LHCB2 and LHCB6) proteins were greatly decreased. Additionally, images taken with transmission electron microscopy showed norflurazon treated plants had fewer smaller plastids which had arrested development during thylakoid synthesis. Norflurazon treated rice seedlings were consequently found to show increased peroxidase activity (Park & Jung, 2018), which could have a photoprotective function in the chloroplast. Chloroplast inhibition by an antibiotic, lincomycin, has also been shown to cause bleaching and a decrease in *LHCB* expression in rice, although there were no other changes detected in leaf morphology (Duan et al., 2020). This evidence suggests that lincomycin and norflurazon cause similar effects within *A. thaliana* and rice, both experiencing decreased PSII activity and *LHCB* expression.

1.8 Evolution of Chloroplast Retrograde Signalling

To date few studies have focused on the evolution of chloroplast retrograde signalling across deep time. The close connection between the genomes and even the important role of anterograde and retrograde signalling for proper function indicate that this process likely originated shortly after endosymbiosis or around the time that the nuclear genome and plastome began exchanging genes. The discovery of *gun*-like mutants in barley (Batschauer et al., 1986), which block the tetrapyrrole biosynthesis pathway in a similar manner to *gun4* and *gun5* in *A. thaliana*, suggests that biogenic chloroplast retrograde signalling may have been present in the ancestral state. Chloroplast inhibition has also been shown to decrease *LHCB* expression within pine, a gymnosperm, suggesting that chloroplast retrograde signalling controlled photosynthesis associated gene expression within pine's last common ancestor with angiosperms, over 150 million years ago (Hills et al., 2015).

Moreover, the ABA response linked to retrograde signalling has been found in streptophyte algae (Hori et al., 2014). Many of the proteins associated with RS exhibit high degrees of sequence similarity across land plants (C. Zhao et al., 2019). A detailed analysis of the alignment and the molecular rate inferred for the gene phylogeny of *SAL1*, nucleotide phosphotase SAL1 protein, showed that the protein is highly conserved and recapitulates many of the well-known relationships in land plants (C. Zhao et al., 2019). The recapitulation of phylogeny indicates the conserved sites in the alignment co-vary and thus are likely ancestral states of the SAL1 protein and not a product of convergence. In concert with the conservation of 3′-phosphoadenosine-5′-phosphate (PAP) metabolism, this indicates that the SAL1-PAP retrograde signalling pathway, which helps with stomatal closure, likely existed in the ancestor of land plants and potentially facilitated the movement onto land.

1.9 Inhibitors of Chloroplast Function and Retrograde Signalling

There are multiple chemical compounds which cause plant death by interfering with the chloroplast and consequently the plant's photosynthetic capacity. Of particular relevance, lincomycin stops chloroplast translation by inhibiting chloroplast ribosomes (Ellis, 1975), whilst not affecting

cytoplasmic protein synthesis (Mulo et al., 2003). Moreover, norflurazon is another inhibitor of the chloroplast whose mode of action is to impede carotenoid synthesis by competitively binding phytoene desaturase and blocking binding of necessary co-factors (Breitenbach et al., 2001). Lincomycin and norflurazon treatment have both been shown to lead to bleaching of leaves (Zhao et al., 2018) and reduced maximum quantum yield of PSII in *A. thaliana* (Beisel et al., 2011). Bleached leaves and decreased maximum quantum yield were also observed in rice treated with lincomycin and norflurazon (Park & Jung, 2017; Zulfugarov et al., 2014).

Lincomycin and norflurazon have been employed in studies researching chloroplast retrograde signalling by analysing the effect of chloroplast inhibition on both gene expression and proteomics in wildtype and mutant plants. An advantage of using lincomycin is that it has a light independent mode of action so the effect of chloroplast inhibition on dark-grown grown plants can be investigated. Norflurazon's mode of action only affects the chloroplast in the light when the lack of protective carotenoids causes photooxidative damage. However, it does have other documented effects in dark-grown plants which include decreasing total lipid quantity and increasing dry weight (Di Baccio et al., 2002; Magnucka et al., 2007). Whilst norflurazon's more diverse phenotypic effect is disadvantageous, other choices of inhibitors come with their own shortcomings. Oxyfluorfen works in a similar manner to norflurazon by blocking protoporphyrin synthesis and causing oxidative damage to the chloroplast, however it is documented to cause larger amounts of oxidative damage compared with norflurazon and it also generates H_2O_2 (Park & Jung, 2018) which could interfere with retrograde signalling. Other inhibitors include DMCU, which only partially inhibits PSII function, and amitrole, which has a temperature dependent effect (Agnolucci et al., 1996).

1.10 Methods to Investigate Regulation and Transcription Factor Binding

Transcriptomic analysis of seedlings treated with chloroplast inhibitors will not only give insight into which genes are chloroplast responsive, but also how these genes are regulated. Regulation of gene expression at the transcriptional level often occurs via transcription factors, proteins which contain a DNA binding domain. This domain recognises short sequences, often referred to as motifs, in the genome which initiate binding (Gill, 2001). To investigate how a transcription factor regulates gene expression, a logical first step is to identify which transcription factor is binding and where. Several methods have been developed to investigate transcriptional regulation by identifying transcription factor binding sites, each requiring different prerequisites. Instead of looking at signals which could potentially emanate from the chloroplast, this body of work will be using an alternate approach by identifying chloroplast responsive genes and using a selection of the methods to identify transcription factors which bind them.

To identify which motifs a transcription factor will bind, chromatin immunoprecipitation sequencing (ChIP-Seq) cross-links bound proteins to DNA. The genome and crosslinked factors are extracted and fragmented, and the fragments bound to a transcription factor of interest are captured with antibodies. By sequencing these fragments, the binding motif is identified in the form of a position weight matrix which shows the likelihood of binding for each base (Barski et al., 2007; Johnson et al., 2007). To perform ChIP-Seq, an antibody against the transcription factor of interest is required. In contrast, the yeast-one hybrid assay requires previous knowledge of the motif of interest. The motif acts as a bait sequence to identify which transcription factors will bind (Ouwerkerk & Meijer, 2001). On a wider scale, DAP-seq involves fragmenting the genome and combining with beads which have transcription factors bound by affinity tags. This allows the genome fragments which bind to a specific transcription factor to be sequenced, revealing all the potential sequences it will bind (O'Malley et al., 2016).

ChIP-seq has identified the binding motifs of transcription factors from a wide range of plant species and DAP-Seq has been used to identify the binding motifs of 529 *A. thaliana* transcription factors. The position weight matrices of these binding motifs have been collated in the JASPAR database (Mathelier et al., 2014). Motifs in the JASPAR database can be searched for in sequences of genes that respond to a common treatment. One method of doing this involves the MEME software suite (Bailey et al., 2009) which includes FIMO (Grant et al., 2011), a program that identifies individual occurrences of motifs within an input sequence and provides a q-value to quantify the accuracy of the match. Another MEME program, AME (McLeay & Bailey, 2010) looks at all occurrences of a motif within the input sequences and provides a p-value which quantifies the likelihood that the motif is enriched above what would be expected by chance. The enriched motifs can highlight putative transcription factors which could act as a common signal.

Methods have now been published which can identify likely transcription factor binding sites within the genome without previous knowledge of the transcription factors or DNA motifs which are binding. These methods work at a genome wide scale and include DNase-I sequencing (DNaseI-seq), micrococcal nuclease digestion with deep sequencing (MNase-seq), Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE-seq) and Assay of Transposase Accessible Chromatin sequencing (ATAC-seq) (Klein & Hainer, 2020). All these techniques use deep sequencing to identify regions of open chromatin. FAIRE-seq crosslinks DNA and bound proteins, which are then fragmented and separated by weight using centrifugation. The protein bound fragments are heavier and can therefore be separated and purified. Sequencing the resulting reads indicates which regions are associated with protein binding. DNase-Seq, MNase-Seq and ATAC-seq all treat nuclei with an enzyme that preferentially cuts open chromatin and sequence the resulting fragments to identify the distribution of cuts. This distribution can be used to predict accessible regions as well as transcription factor footprints, short regions where chromatin is protected. Assay of Transposase Accessible Chromatin sequencing (ATAC-seq) identifies these accessible regions by digesting nuclei with a transposase and is faster and requires less tissue than the other methods (Buenrostro et al., 2013). All of the methods described will allow an investigation of transcription factor binding within genes whose transcript abundance is significantly changed after chloroplast perturbation.

1.11 Objectives

The overarching aim of the work presented in this thesis was to better understand the role of plastid-to-nucleus signalling in regulating genes of the C_4 cycle in the ancestral C_3 state. Three central objectives were designed to achieve this aim:

- Genes involved in plastid-to-nucleus signalling will be identified through RNA-seq of C_3 *Arabidopsis thaliana* and *Oryza sativa* after chloroplast function has been perturbed. The response of C_4 orthologues within the eudicotyledon and monocotyledon model will thus be compared.

- *Cis*-elements in the nuclear genome involved in the regulation of plastid-regulated genes will be predicted through analysis of enriched binding motifs in genes responsive to chloroplast perturbation.

- The methodology and data analysis pipeline of ATAC-seq will be optimised, to discover the distribution of predicted open chromatin regions across the rice genome and to allow future work which could generate an atlas of transcription factor binding sites with changed occupancy in response to chloroplast perturbation.

2 METHODS

2.1 Growth of plants

Growth of Arabidopsis thaliana

Seeds of *A. thaliana* cultivar Columbia were sterilised for 2 minutes in 70% (v/v) ethanol, followed by 5 minutes in 2.5% (v/v) sodium hypochlorite. After 5 washes in MiliQ water, seeds were submerged and stratified at 4˚C for 4 days. 150 µmol m^{-2} s^{-1} white light were then supplied for an hour to induce germination. 0.2g of seeds per plate were evenly distributed on 300 µm aperture nylon mesh (Normesh) placed on 30ml half-strength Murashige and Skoog salts and 0.8% (w/v) agar (pH 5.8 KOH) with or without 0.5mM lincomycin or 5µM norflurazon. An equivalent volume of ethanol was added to norflurazon controls. Plants were dark-grown at 22˚C for 5 days, prior to half being exposed to 150 µmol m^{-2} s^{-1} white light for 24 hours. Samples of hypocotyl and cotyledons were frozen in liquid nitrogen and stored at -80˚C.

For stocks, Columbia was grown for 8 weeks in a growth chamber with a 16 hour photoperiod at a light intensity of 200 µmol m^{-2} s^{-1} photon flux density with 65% relative humidity and a temperature cycle of 24°C (day) and 20°C (night). Plants were then dried for a further two weeks before seeds were harvested by shaking siliques.

Growth of Oryza sativa

Seeds of *Oryza sativa* cultivar Kitaake had seed coats removed and were sterilised in 2.5% (v/v) sodium hypochlorite with 0.5% (v/v) Tween 20 for 15 minutes, washed five times in MiliQ water, and then incubated for 24 hours in the dark at 30˚C. To allow germination, seeds were placed on filter paper for a further 24 hours. Subsequently, nine seedlings were placed in each pot (7.5cm diameter by 8cm depth) on 50ml half-strength Murashige and Skoog salts and 0.8% (w/v) agar (pH 5.8 KOH) with or without 10mM lincomycin or 100µM norflurazon (Sigma Aldrich). Norflurazon is soluble in ethanol and so norflurazon controls contained an equivalent volume of ethanol. Plants were dark-grown at 30˚C for 5 days, then half were exposed to 150 µmol m^{-2} s^{-1} white light for 24 hours. Leaves were sampled after a total of 6 days growth, frozen in liquid nitrogen and stored at -80˚C.

For stocks, Kitaake was grown for 10 weeks in a growth chamber with a 12 hour photoperiod at a light intensity of 300 µmol m^{-2} s^{-1} photon flux density with 60% relative humidity and a temperature cycle of 28°C (day) and 25°C (night). Plants were then dried for a further two weeks before seeds were harvested.

2.2 Chlorophyll Fluorescence Measurements

Fluorescence parameters were obtained using the Fluor*Imager* chlorophyll fluorescence imaging system (Technologica Ltd). The maximum quantum efficiency of PSII photochemistry (F_v/F_m) was

measured after 30 minutes dark-adaptation using a saturating pulse of 6172 µmol photons m^{-2} s^{-1}. Plants were then exposed to 150 µmol m^{-2} s^{-1} actinic light, and a time-course undertaken to determine quantum efficiency of PSII photochemistry (F_q'/F_m').

2.3 RNA Extraction and quantification

RNA was extracted from frozen samples using the RNeasy Plant Mini Kit (QIAGEN) with an additional DNase step performed using the RNase-Free DNase Set (QIAGEN). A preliminary check of RNA concentration and quality was performed with the Nanodrop 2000. RNA Integrity Numbers (RIN) and electropherograms were then generated after running diluted samples on the Agilent Bioanalyzer 2100 with a RNA 6000 Pico Chip (Agilent Technologies, Inc.). All samples taken forward for sequencing had an RNA Integrity Number (RIN) of > 7.

2.4 Library preparation for RNA-seq and Sequencing

Libraries were prepared from 0.5µg RNA with the QuantSeq 3′ mRNA-Seq Library Prep Kit FWD for Illumina (Lexogen). Libraries were quality controlled using the Agilent Bioanalyzer 2100 with a High Sensitivity DNA Chip (Agilent Technologies, Inc.). Libraries were sequenced on a NextSeq500 (Illumina) Mid Output 150 cycle run at the Sainsbury Laboratory University of Cambridge.

2.5 Transcriptome Analysis

2.5.1 Alignment and Differential Transcript Abundance Analysis

Concatenated fasta files were analysed using the QuantSeq 3' mRNA Seq Integrated Data Analysis Pipeline on the Bluebee platform (Lexogen). Quality control reports were generated both post and pre-trimming in FastQC. The pipeline trimmed with BBDuk from the BBMap suite, trimming to remove low quality tails, poly(A)read-through and adapter contamination. An alignment was generated using STAR aligner (Dobin et al., 2013) with slightly modified ENCODE settings (--outFilterMismatchNoverLmax 0.6) to allow more mismatches relative to read length than standard settings. Read counts were calculated with HTSeq-count (Anders et al., 2015), using mode intersection-nonempty which includes reads only partially mapped to a gene. Log2fold changes in transcript abundance were calculated using DESeq2 (Love et al., 2014), significantly different transcript abundance was defined as those having an adjusted p-value < 0.05. Transcripts Per Million (TPM) for expression figures were calculated using Salmon (Patro et al., 2017), employing the –noLengthCorrection option as with the Lexogen Quantseq kit the number of sequenced fragments deriving from a target are independent of that target's length. Principal Component Analysis was performed using PCAExplorer (Marini & Binder, 2019). Gene Ontology enrichment analysis was performed with AgriGO v2.0 (Tian et al., 2017) using the Fisher statistical test with Yukitieli multi-test adjustment methods. Figures illustrating expression in the photosynthetic pathway were generated with MapMan (Usadel et al., 2009). Rice specific orthologues of essential genes in the C_4 pathway were identified using OrthoFinder (Emms & Kelly, 2015).

2.5.2 Enriched Motif Analysis

Preliminary analysis looked for enriched 6-mers 1000bp upstream of the transcription start site using the TAIR web-based tool 'Statistical Motif Analysis in Promoter or Upstream Gene Sequences'. More in-depth analysis was then performed using two tools from the MEME suite, FIMO and AME. FIMO was run using standard settings, significant enrichment was defined as a q-value < 0.05 (Grant et al., 2011). AME was run in 'totalhits' mode which counts the number of times a motif matches with at least 0.25 of the maximum possible log-odds score for the position weight matrix (McLeay & Bailey, 2010). To inform a refinement of the input sequences, the distribution of DNAseI hypersensitive sites (DHS) from available DNaseI-seq data was plotted with ggplot2 (Arabidopsis data NCBI ID: GSE53322, Rice data NCBI ID: GSE75794). The distance between the midpoint of the DHS peaks from the closest transcription start site was calculated using the 'closest' function in the BEDTools suite (Quinlan & Hall, 2010).

2.6 Nuclei extraction and Library preparation for ATAC-Seq

The protocol closely followed that used in Maher et al (2018), and that is described in Bajic et al (2018). For nuclei extraction and quantification, nuclei were extracted from samples by sucrose sedimentation. Frozen ground tissue or fresh tissue, chopped for 5 minutes with a razor, was suspended in 10mL of freshly prepared ice-cold Nuclei Purification Buffer (20mM MOPS pH 7, 40mM NaCl, 90mM KCl, 2mM EDTA, 0.5mM EGTA, 0.5mM spermidine, 0.2mM spermine, and 1× Roche Complete protease inhibitors). The solution was then filtered through a 70µm nylon cell strainer and centrifuged at 1200g for 10 minutes at 4°C. The supernatant was discarded, and the pellet resuspended in 1mL of ice-cold Nuclei Extraction Buffer (0.25M Sucrose, 10mM Tris–HCl pH 8, 10mM MgCl2, 1% Triton X-100, and 1× Roche Complete Protease Inhibitors). After another centrifugation at 1200g for 10 minutes at 4°C, the supernatant was discarded and the pellet resuspended in 300µL of Nuclei Extraction Buffer 2 (1.7M Sucrose , 10mM Tris–HCl pH 8, 2mM MgCl2, and 0.15% Triton X-100, 1× Roche Complete Protease Inhibitors). This was layered on top of 300µL of Nuclei Extraction Buffer 2 in a fresh Eppendorf tube. After centrifugation at 16,000g for 10 minutes at 4°C, the supernatant was discarded and pellet resuspended in 1mL of ice-cold Nuclei Purification Buffer. 25µL of this nuclei suspension was combined with 1µL of DAPI stain and placed on a haemocytometer which enabled counting of nuclei and thus calculation of the nuclei concentration.

2.6.1 Transposase Treatment and Library Preparation

A volume of nuclei suspension containing 50,000 nuclei was centrifuged at 1500g for 7 minutes at 4°C. The supernatant was discarded and pellet resuspended in 50µL of ice-cold transposition reaction mix. The transposition reaction mix contained 2.5µL TDE1 Transposase (a hyperactive variant of Tn5 Transposase) and 25µL 2× TD buffer from the Illumina Nextera DNA Library Preparation Kit, made up to 50µL with water. This was incubated for 30 minutes at 37°C, mixing by hand every 5 minutes. The resulting solution contained fragmented DNA which was purified using the Qiagen MinElute PCR Purification Kit according to manufacturer's instructions. DNA was amplified using the NEBNext® High-Fidelity 2X PCR Master Mix with custom ATAC-seq primers (Maher et al., 2018). The resulting libraries were purified by adding 45µL of library to 30µL of Beckman Coulter AMPure XP beads and incubating for 5 minutes at room temperature. The beads were gathered and held at the bottom of the tube by magnets allowing the supernatant to be discarded. Two washes of

200µL of 80% (v/v) ethanol were undertaken without disturbing the beads, and the beads left to dry for 5 minutes. The beads were resuspended in 20µL 10mM Tris-HCl pH 8 for 2 minutes, before the supernatant containing the purified library was collected. Libraries were sequenced on a NextSeq500 (Illumina) High Output 150 cycle run at the Sainsbury Laboratory University of Cambridge.

2.7 Computational Analysis of ATAC-Seq

2.7.1 Quality Control and Alignment

Quality control reports were generated both post and pre-trimming in FastQC. Reads were trimmed with TrimGalore-04.4, with 17bp removed at the 5' end and 4bp at the 3' end. An alignment was generated using Bowtie2 (Langmead & Salzberg, 2012), with reads aligned to Release 40 of the IRGSP-1.0 *Oryza sativa* genome. Alignment quality was analysed by checking q scores with SAMtools (H. Li et al., 2009), and GC bias and insert size were calculated using the 'CollectGcBiasMetrics' and 'CollectInsertSizeMetrics' functions from Picard Tools respectively.

2.7.2 Peak Calling

Before peak calling, duplicate reads were removed using the 'MarkDuplicates' function from Picard Tools. The genome size of Release 40 of the IRGSP-1.0 *Oryza sativa* genome was calculated with KentUtils. Peaks were called with MACS2 (Zhang, 2008) with the cutoff p-value set to 1e-2. Spot scores were calculated by using the 'intersect' function of BEDtools to analyse overlap between all peaks and a randomly selected subset of reads. ATAC-seq footprints were predicted with Wellington (Piper et al., 2013) with a maximum allowed false discovery rate of -10.

2.8 Genotyping of T-DNA Insertion Mutants

T-DNA insertion mutants for promising transcription factors were obtained from the Nottingham Arabidopsis Stock Centre (NASC). Seed stocks were attained by growing t-DNA insertion mutants for 8 weeks in a growth chamber with a 16 hour photoperiod at a light intensity of 200 µmol m^{-2} s^{-1} photon flux density with 65% relative humidity and a temperature cycle of 24°C (day) and 20°C (night). Plants were then dried for a further two weeks before seeds were harvested by shaking siliques.

DNA was extracted from mutant lines with Edwards solution (0.2M Tris pH 8.0, 0.25M NaCl, 0.025M EDTA, 0.5% SDS). Frozen tissue was ground in the Qiagen TissueLyser II for 1 minute at 28 Hz. 500µL of Edwards solution was added before centrifuging for 5 minutes at 10,000g. The supernatant was transferred to a new Eppendorf tube and added to 500µL of ice-cold isopropanol. After another centrifugation for 5 minutes at 10,000g, the supernatant was discarded and the pelleted DNA washed with 500µL ice-cold ethanol before resuspension in 100µL water. DNA was amplified by PCR with each reaction including 2µL of the extracted DNA, 0.25µL Bioline BIOTAQ DNA Polymerase, 2.5µL 10x NH4 Reaction Buffer, 1.5µL 50 mM MgCl2 Solution, 2.5µL 10mM dNTPs and 1.5µL forward and reverse primer mix. Primers for genotyping were designed using the Signal SALK T-DNA Primer Design Tool. PCR product was run on a gel made up of 1% agar 0.5X TBE

buffer combined with 0.1µl/ml of SYBR safe dye (Thermo Fisher). Samples were ran in 0.5X TBE buffer (50mM Tris pH 8.0, 50mM Boric acid, 1mM EDTA) alongside 1kb HyperLadder I (BIOLINE).

3 CHLOROPLAST AND LIGHT DEPENDENT REGULATION OF C_4 ORTHOLOGUES IN *ARABIDOPSIS THALIANA*

3.1 Introduction

The majority of recent research on chloroplast retrograde signalling has been carried out in *Arabidopsis thaliana*. This includes the identification and analysis of genome uncoupled (*gun*) mutants, arguably the most definitive work on biogenic retrograde signalling to date. In wildtype plants, chloroplast inhibition results in a significantly decreased expression of PhANGs, such as *LHCB*. In *gun* mutants, however, PhANG expression does not drastically decrease after chloroplast inhibitors have been applied, implying that retrograde signalling pathways have been disrupted (Susek et al., 1993; Larkin, 2014). Six *gun* mutants have been identified. The *GUN2-6* alleles have now been well characterised and are involved in tetrapyrrole synthesis. GUN2, GUN3 and GUN6 encode proteins involved in the synthesis or turnover of heme, a positive regulator of PhANG expression; GUN4 and GUN5 are involved in chlorophyll synthesis, affecting the concentrations of Mg-proto IX (Martín et al., 2016).

Operational signalling has also been largely investigated in *A. thaliana*. Current understanding suggests these signals are controlled by the redox potential of the cell and ROS concentration. The redox potential of the chloroplast is thought to be transmitted through the malate valve, which involves the key enzyme NADP-MDH (Heyno et al., 2014). When the NADPH pool is over-reduced, NADP-MDH oxidises this pool by using the reduction power to convert OAA to malate which is then exported to the cytosol. Furthermore, ROS can act as a signal by affecting the redox potential of the cell, however, as increased ROS can be highly damaging and cause destruction of DNA, proteins and other important cellular components, the chloroplast has developed more than one detection pathway. Singlet oxygen can be detected by signalling components dependent on either β-cyclocitral or Executor proteins. The exact mechanism of action of β-cyclocitral, which is derived from oxygenation of carotenoids, is unknown (Ramel et al., 2012). Executor proteins are thought to repress defence responses under ambient conditions, and are responsible for taking the brake off the defence response under stress (Uberegui et al., 2015). An additional operational signal is 2-C-Methyl-Derythritol-2,4-cyclopyrophosphate (MEcPP). MEcPP is a metabolite of the methylerythritol phosphate (MEP) pathway that allows synthesis of isoprenoids. MEcPP accumulates in response to stress such as wounding and acts as a global stress sensor (Xiao et al., 2012) upregulating synthesis of the hormone salicylic acid. MEcPP impacts on nuclear gene expression via a *cis*-element termed the Rapid Stress Response Element (Benn et al., 2016).

Recent work in *A. thaliana* has proposed that chloroplast retrograde signalling could help explain facets of C_4 evolution. C_4 photosynthesis has evolved independently in over 60 lineages (Sage et al.,

2011) and phylogenetic studies have shown parallel evolution could underpin the repeated recruitment of C_4 genes (Williams et al., 2012). Despite enzymatic activity being conserved within gene families, the same orthologues from multi-gene families have repeatedly been recruited during the evolution of the C_4 trait (Christin et al., 2013). If function is not responsible for this, it is possible that a shared mechanism of genetic regulation could be. Consistent with this latter notion, recent evidence indicates that in C_3 species, C_4 genes may be already controlled by retrograde signalling pathways. Burgess et al (2016) found seven of sixteen C_4 orthologues tested in *Arabidopsis thaliana* are controlled by plastid-to-nucleus signalling.

Currently, it is not known how C_4 orthologues are regulated by plastid-to-nucleus signalling in *Arabidopsis thaliana* and whether this regulation occurs in other C_3 species, particularly within monocotyledons. The objective of the work presented in this chapter was to identify the genes that may be involved in this signalling through an RNA-seq experiment in C_3 *Arabidopsis thaliana* after chloroplast function has been perturbed. The results could provide further support for the hypothesis that retrograde control helps to explain the repeated recruitment of genes in the ~60 independent origins of C_4 photosynthesis. This work also aims to shed light on how this regulation occurs and identify components of the chloroplast to nucleus signalling pathway. By carrying out RNA-seq on plants with perturbed chloroplast function, an extensive study of C_4 genes can be carried out and the proportion which are regulated by retrograde signalling in *A. thaliana* can be elucidated. By looking more broadly at the genes affected, and by grouping these genes by function using Gene Ontology terms, general trends could be discerned in the type of genes whose expression is controlled through RS. Additionally, these genes can be scanned for enriched binding motifs allowing identification of potentially important *cis*-elements. The output of this analysis could lead to the identification and characterisation of a chloroplast derived signal modifying nuclear gene expression.

The wealth of chloroplast retrograde signalling research in *A. thaliana* simplifies the methodological setup for this work. Sublethal concentrations of the two chloroplast inhibitors selected for the study, lincomycin and norflurazon, are reported in the literature (McCormac & Terry, 2004). Additionally, there are many bioinformatic resources associated with the species. This could prove useful in the analysis of transcriptomic data and in future follow-up work. These resources include well annotated datasets, such as the Plant Cistrome Database which provides a list of known motifs identified through DAP-seq (O'Malley et al., 2016). This can be used to look for enriched motifs in genes which respond to chloroplast inhibition. Resources associated with *A. thaliana* have also been shown to allow fast functional testing, which will likely prove useful once characterisation targets are identified. Microarray data are available for *A. thaliana* wildtype seedlings and *gun* mutants treated with norflurazon, although the low number of replicates reduces the statistical power of any conclusions made with the dataset (Moulin et al., 2008).

3.2 Objectives

Genes affected by plastid-to-nucleus signalling will be identified through RNA-seq in C_3 *Arabidopsis thaliana* after chloroplast function has been perturbed.

Hypotheses

Photosynthesis related genes will be disproportionally overrepresented in the genes affected by inhibitor treatment as there is much evidence that the chloroplast retrograde signalling pathway is intertwined with the light signalling pathways known to regulate photosynthesis.

Multiple C_4 orthologues will be regulated by chloroplast to nucleus signalling, providing further support for the hypothesis that retrograde control explains the repeated recruitment of genes in the parallel evolution of C_4 photosynthesis.

Putative transcription factors involved in the regulation of plastid-regulated genes will be elucidated by analysing enriched binding motifs.

Hypotheses

There will be significant enrichment of binding motifs in genes responsive to chloroplast perturbation compared to background as these genes share a regulatory pathway.

The motifs significantly enriched in plastid-regulated C_4 orthologues will differ to those enriched in the other C_4 orthologues, as they do not wholly share a regulatory pathway.

3.3 Results

3.3.1 Impact of lincomycin and norflurazon on chlorophyll accumulation and chlorophyll fluorescence

Previously, sublethal concentrations of lincomycin and norflurazon that inhibit greening of *A. thaliana* cotyledons by impeding chloroplast development have been reported (McCormac & Terry, 2004). When this protocol was used here, chlorophyll accumulation was also compromised (Figure 3.1).

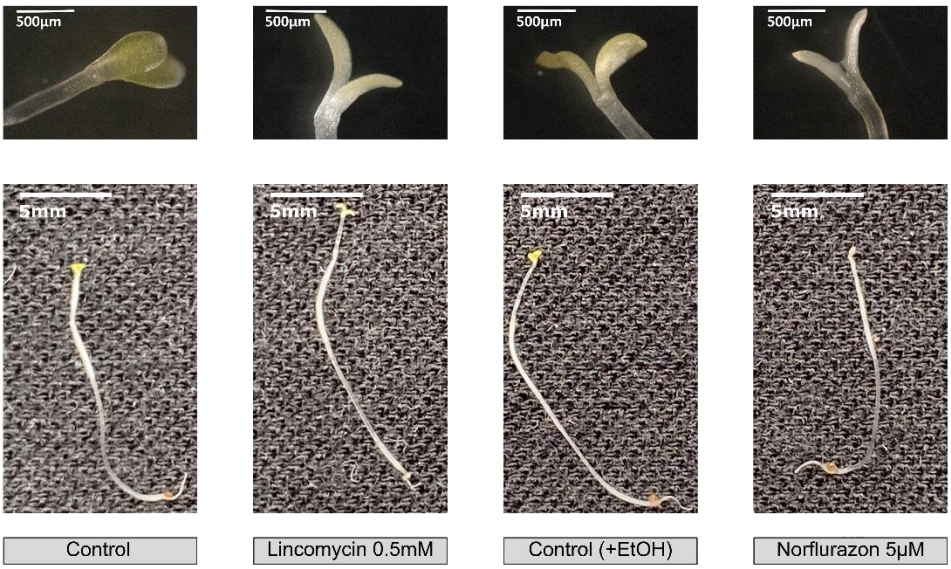


Figure 3.1: Accumulation of chlorophyll is impaired in seedlings exposed to either lincomycin or norflurazon. A. thaliana seedlings dark-grown for 5 days and grown in 150 µmol m^{-2} s^{-1} white light for 24 hours. EtOH = 0.001% (v/v) ethanol.

To quantify the impact of reduced greening on photosynthesis, pulse-modulated chlorophyll fluorescence was assessed. Maximum quantum yield of PSII (F_v/F_m) provides insight into the number of PSII reaction centres capable of absorbing and utilising light. This gives a reliable indication of photosynthetic performance (Maxwell & Johnson, 2000). Consistent with very low amounts of chlorophyll, seedlings of *Arabidopsis thaliana* maintained in the dark showed low values of F_v/F_m and addition of either lincomycin or norflurazon had no impact on F_v/F_m (Figure 3.2). However, when plants were transferred to the light values of F_v/F_m were significantly increased compared to the dark-grown controls. Treatment with norflurazon or lincomycin caused a greater than two-fold decrease in F_v/F_m in illuminated seedlings compared with controls. The effect on the photosynthetic apparatus is striking and shows a decrease in the ability of chloroplasts to harvest light. As these inhibitors reduced accumulation of chlorophyll, which is dependent on nuclear gene expression, these treatments were considered suitable for analysis of how genome-wide patterns of transcript abundance are perturbed during this process.

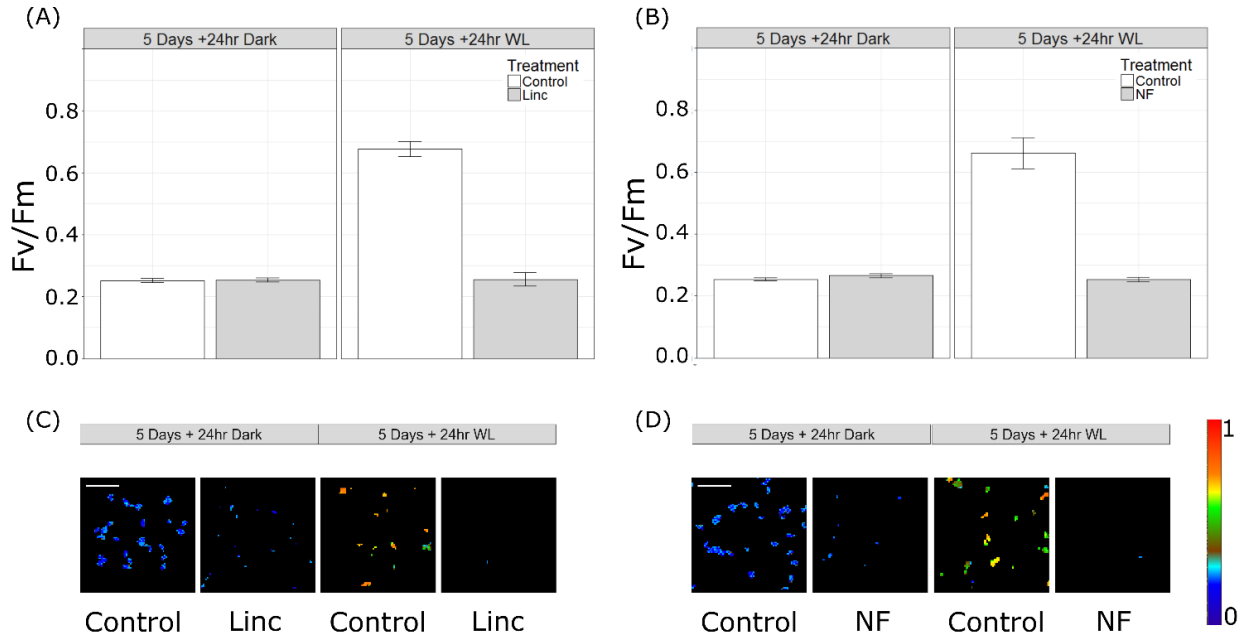


Figure 3.2: Effect of lincomycin and nuroflurazon on chlorophyll fluorescence in *A. thaliana*. Dark-grown seedlings showed low values of F_v/F_m. (A&B) Bar chart quantifying the impact of transferring seedlings from dark to light, and treatment with either lincomycin (A) or norflurazon (B) on F_v/F_m (n=3 plates of seedlings). Data are presented as means with one standard error. Consistent with the greening of cotyledons in response to light, illumination led to an increase in F_v/F_m. In both cases, application of lincomycin and norflurazon reduced F_v/F_m to levels similar to those detected in dark grown seedlings. Statistical significance was calculated using a student's t-test (illuminated control mean F_v/F_m = 0.64, dark-grown control mean F_v/F_m = 0.25): t(10) = -13.35, $p < 0.001$, (illuminated control mean F_v/F_m = 0.64, illuminated inhibitor-treated mean F_v/F_m = 0.25): t(10) = -12.631, $p < 0.001$. (C&D) Representative images captured by the FluorImager chlorophyll fluorescence imaging system showing a heatmap of F_v/F_m. The false colour scale represents an F_v/F_m of 0 (blue) to 1 (red). Scale bars represents 5mm. Linc = 0.5mM lincomycin, NF =5µM norflurazon, WL = 150 µmol m^{-2} s^{-1} white light.

3.3.2 Extraction and analysis of RNA prior to sequencing

To confirm that RNA extracted after each treatment was of sufficient quality for deep sequencing, its integrity was analysed using a microfluidic electrophoresis system. This generated electropherograms and an RNA Integrity Number (RIN), which gives an indication of how much degradation the RNA sample has undergone. The calculation of an RIN takes into account the ratio between the 18S and 25S peaks that derive from the small and large subunit of nuclear ribosomes respectively, as well as the overall distribution of the electropherogram trace. This trace also shows a 16S-related peak, which derives from the small subunit of the chloroplast ribosome (Delius & Koller, 1980). By comparing the height of this peak compared with the 18S and 25S peaks, the ribosome composition and subsequently the potential for chloroplast translation can be inferred (Figure 3.3).

This analysis corroborates findings from the physiological analysis above. Smaller 16S peaks can be seen in the dark-grown compared with the 24-hour illuminated sample (Figure 3.3). This is consistent with dark-grown cotyledons containing etioplasts which have not yet fully developed and proliferated into chloroplasts. The lincomycin treatment shows no 16S peak in either dark-grown or illuminated samples. This is likely partially due to a lack of plastid encoded polymerase as chloroplast translation

is blocked by lincomycin, and partially due to reduced expression of nuclear encoded polymerase subunits, such as rpoTp, via retrograde signalling. In contrast, a 16S peak was present in the dark-grown *A. thaliana* treated with norflurazon, but considerably reduced on exposure to light. By including both lincomycin and norflurazon in the study, the unintended effects of individual inhibitors can hopefully be disentangled from the results of interest. Combined with the data above (Figure 3.2), these results indicate that both inhibitors have perturbed chloroplast development, decreased the number of PSII reaction centres located in the thylakoid membranes and eliminated accumulation of chloroplast rRNA in light-exposed seedlings.

The RIN for each sample was assessed, samples with an RIN over 7 were considered intact and suitable for next generation sequencing. All libraries were prepared from samples with an RIN of 7 or greater.

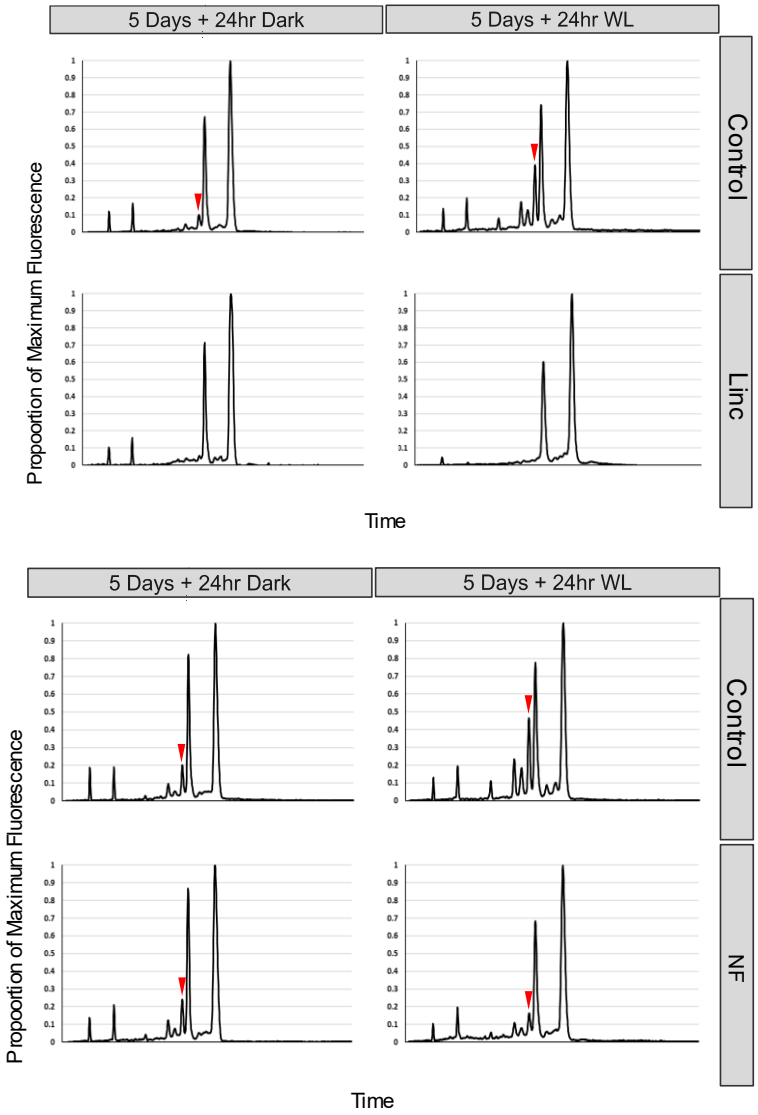


Figure 3.3: Electropherograms of *A. thaliana* RNA for each treatment obtained on the Agilent Bioanalyzer 2100. Red arrow indicates the 16S peak, which is indicative of the presence of chloroplast ribosomes. As maximum fluorescence varies according to sample concentration, all samples were normalised to their maximum value to simplify visual comparisons. Linc = 0.5mM lincomycin, NF =5µM norflurazon, WL = 150 µmol m^{-2} s^{-1} white light.

3.3.3 Quality control of transcriptomic libraries

Libraries were sequenced, and samples underwent quality control. FastQC is a tool that allows intuitive visual inspection of sequences and provides a pass or fail grading to QC components. Analysis of samples showed they were good quality, both at sequence and base level. There were no unassigned bases (Ns) present in the data, and no adapter contamination (Figure 3.4).

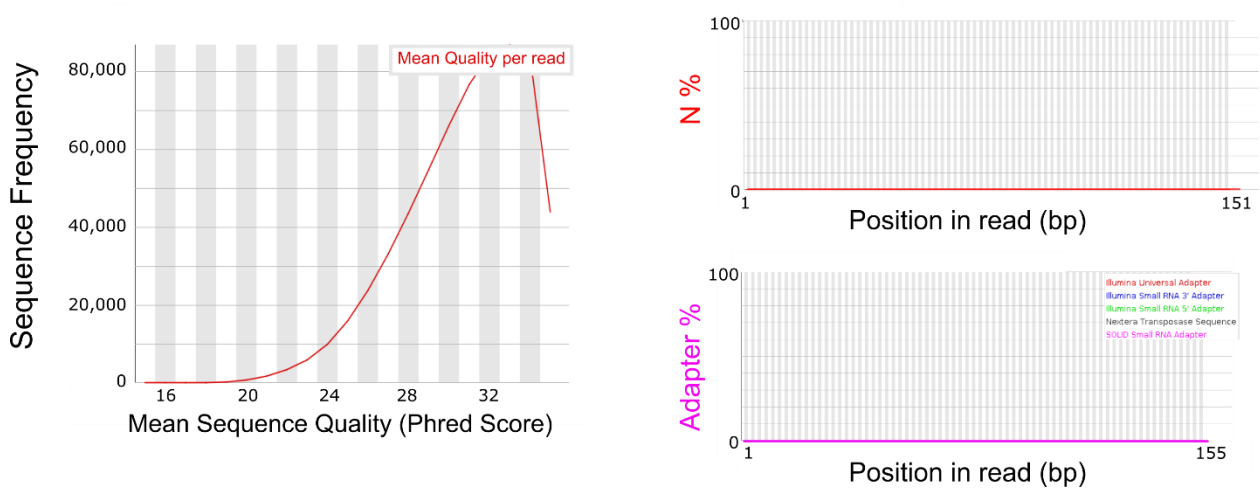


Figure 3.4: FastQC generated plots illustrating sequences were of a good quality and contained no unassigned bases or adapter contamination. (A) Representative plot showing frequency against mean sequence quality, with most sequences having a mean phred score of 33. (B) Representative plot showing Ns (unassigned bases) were not present at any position along the read. (C) Representative plot showing no adapters were present at any position in the read.

However, two FastQC categories failed for all samples; 'Per Sequence GC Content' and 'Kmer Content'. The assumptions FastQC makes in assessing quality control for these categories only holds true for kits which generate libraries with sequences from across the whole transcript. Libraries were prepared using the QuantSeq 3′ mRNA-Seq Library Prep Kit from Lexogen. This kit generates libraries of sequences close to the 3' end of polyadenylated RNA, making it ideal for gene expression analysis. In fact, Lexogen confirmed the results for 'Per Sequence GC Content' and 'Kmer Content' are a by-product of the QuantSeq kit's mode of action which means the assumptions of FastQC's grading method no longer hold true (Lexogen, personal communication, 2016). This is because the QuantSeq FWD reads containing the random primer sequence in the reads, and the binding energy efficiency of these primers is often G-biased. The validity of these results is further corroborated by comparison against QuantSeq generated *Zea mays* libraries from peer-reviewed data (Kremling et al., 2018) (Figure 3.5).

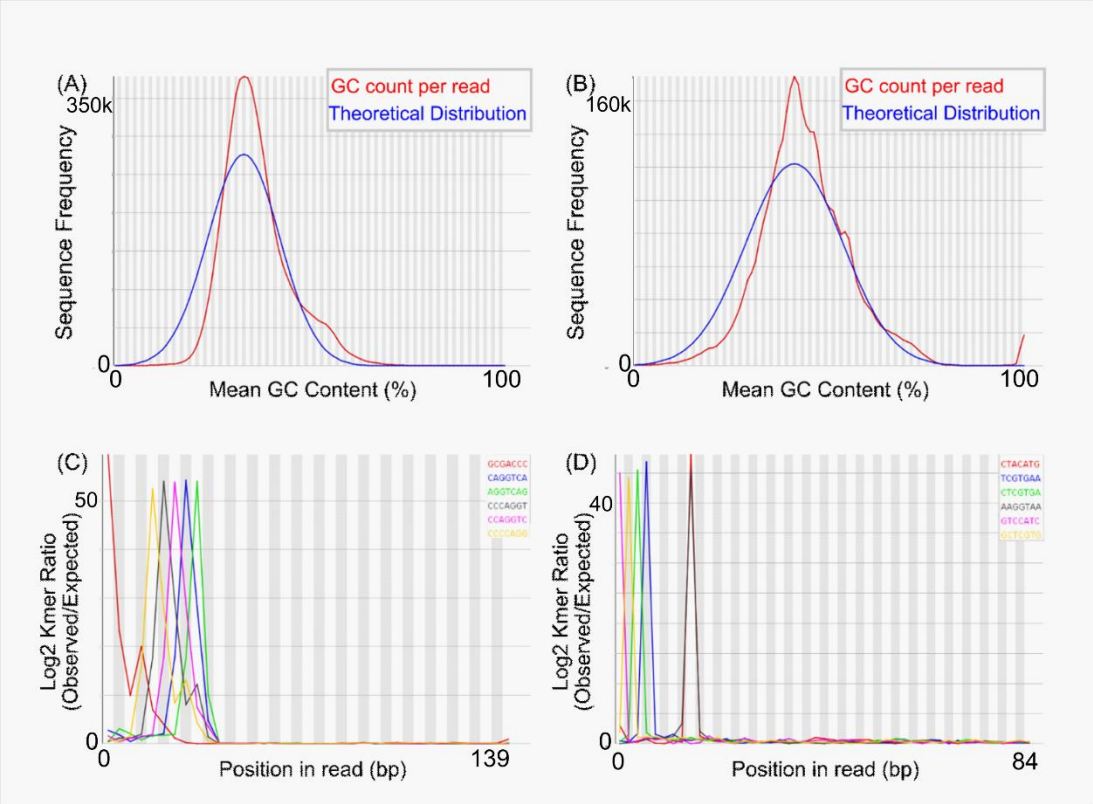


Figure 3.5: Representative FastQC plots for 'Per Sequence GC Content' and 'Kmer Content', showing the similarity between the libraries generated here (A and C), and from peer-reviewed work (Kremling et al., 2018; SRA ID SRR5911960) (B and D). (A) and (B) Representative plot showing frequency against mean GC content, theoretical distribution is based on assumptions which do not hold true for the Lexogen QuantSeq kit. (B) and (D) Representative plot showing Kmer enrichment across the read.

Thus, the QC analysis indicated that all libraries generated good quality data, and so further analysis was initiated. After trimming and quality control, 165 million reads were obtained in total. A mean of 80.5% of reads mapped uniquely to the *A. thaliana* TAIR10 reference genome.

3.3.4 Hierarchical clustering and principal component analysis to determine sample similarity

Read counts estimated from the sample alignments were subjected to hierarchical clustering and principal component analysis to confirm biological replicates were behaving in a similar manner and identify which treatment caused the greatest alteration to transcript abundance. Hierarchical clustering was performed by calculating the distance between samples, providing a quantitative measure of sample similarity (Figure 3.6). Data underwent variance stabilising transformation which normalises with respect to library size. This also involves logarithmic transformation to ensure that a small number of highly expressed genes do not skew the results.

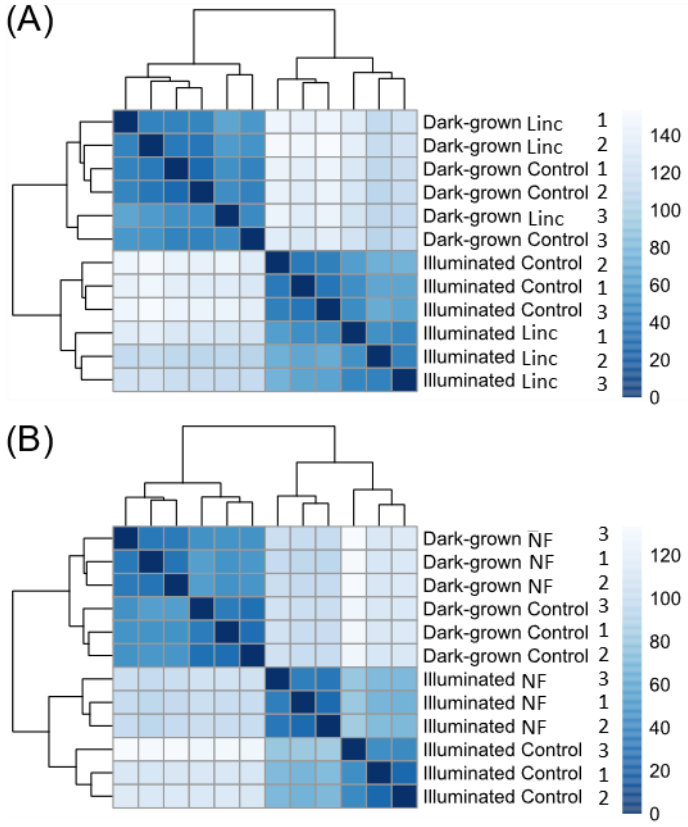


Figure 3.6: Heatmaps showing hierarchical clustering of sampling distance, indicating sample similarity and showing replicates cluster together. Images generated by pheatmap, with a false colour scale, units are variance stabilising transformed normalised counts. (A) Linc = 0.5mM Lincomycin treated (B) NF = 5µM Norflurazon treated.

Hierarchical clustering showed replicates generally clustered together, implying the majority of variance is explained by sample treatment rather than biological noise (Figure 3.6). One exception was the clustering of one dark-grown lincomycin treated replicate and one dark-grown control replicate. As dark-grown seedlings contain etioplasts rather than fully functioning chloroplasts, this is not completely unexpected as it indicates that the treatment in the dark has a less clear effect on transcript abundance

Principal components analysis (PCA) is a class free method for explaining variance between samples. Variance is attributed to unknown principal components, and plotting these components can give an indication of sample similarity or disparity. Over 95% of the variance between samples could be attributed to two principal components (Figure 3.7).

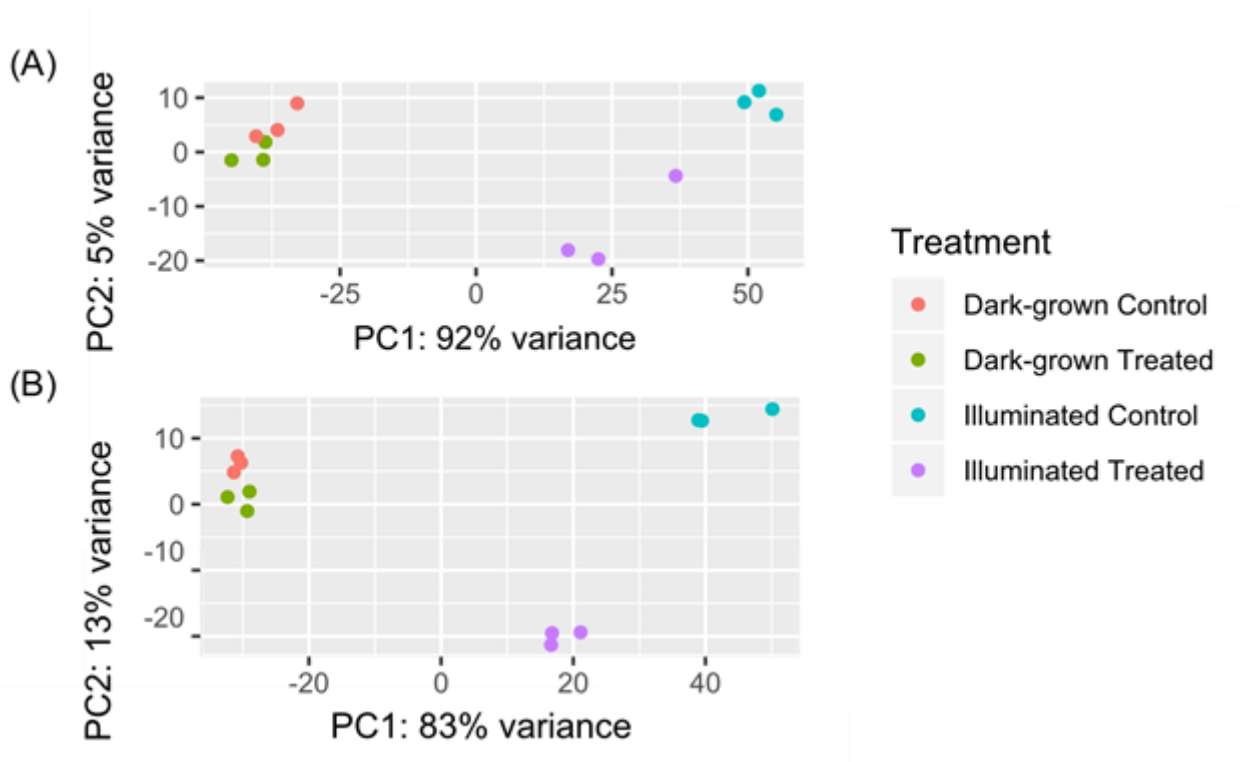


Figure 3.7: Principal components analysis of transcript abundance data obtained from dark-grown and illuminated inhibitor treated samples and corresponding controls. (A) Analysis of lincomycin treated samples and corresponding controls (B) Analysis of norflurazon treated samples and corresponding controls. PC1 = principal component 1, PC2 = Principal component 2.

The three replicates for each treatment generally clustered together, corroborating the previous findings that sample treatment is responsible for the majority of variation in the dataset (Figure 3.7). The spatial separation of samples across PC1 and PC2 was also consistent between the two inhibitors. PC1 accounted for 92% and 83% of the variation in samples, for lincomycin and norflurazon respectively, and most of this variation appears to be in response to light treatment. However, in illuminated samples, inhibitor treatment caused an approximate 20% shift in PC1. PC2 accounts solely for the variation between illuminated control and illuminated treated samples. One of the three replicates treated with lincomycin appeared to have been less affected than the other two, possibly due to reduced uptake of the inhibitor.

3.3.5 Differential Transcript Abundance in Response to Chloroplast Inhibition

Sample read counts consequently underwent a pairwise differential expression analysis. Pairwise comparisons were carried out between samples where only one factor changed, ie - either the light treatment or the inhibitor treatment (Table 3.1). Significance was defined as an adjusted p-value of less than 0.05.

(A)

	Number of Genes with Significant DTA (Total)			
	Lincomycin	Norflurazon	Shared	% Genes Shared DTA
Control: **Dark-grown vs Illuminated**	7352	7244	5616	76%
Inhibited: **Dark-grown vs Illuminated**	5628	7096	4114	73%
Dark-grown: **Control vs Inhibited**	2	1457	1	50%
Illuminated: **Control vs Inhibited**	1010	3607	713	71%

(B)

	Number of Genes Significantly Upregulated			
	Lincomycin	Norflurazon	Shared	% Genes Shared Upregulated
Control: **Dark-grown vs Illuminated**	3821	3758	3052	80%
Inhibited: **Dark-grown vs Illuminated**	2893	3463	2084	72%
Dark-grown: **Control vs Inhibited**	0	649	0	0%
Illuminated: **Control vs Inhibited**	181	1277	71	39%

(C)

	Number of Genes Significantly Downregulated			
	Lincomycin	Norflurazon	Shared	% Genes Shared Downregulated
Control: **Dark-grown vs Illuminated**	3531	3486	2558	72%
Inhibited: **Dark-grown vs Illuminated**	2735	3633	1979	72%
Dark-grown: **Control vs Inhibited**	2	808	1	50%
Illuminated: **Control vs Inhibited**	829	2330	623	75%

Table 3.1: Number of genes showing significant differential transcript abundance between treatments, split into (A) all genes with significant differential transcript abundance, (B) upregulated genes only and (C) downregulated genes only. Significance is defined as an adjusted p-value <0.05. The percentage of genes showing significant differential transcript abundance in response to both inhibitors as a percentage of the lincomycin related list. DTA = differential transcript abundance.

Comparisons between control dark-grown and illuminated samples yielded the most differentially expressed genes between any of the treatments (Table 3.1). This is consistent with the class-free analysis presented above, and expected as large transcriptomic changes in response to light are well documented (Jiao et al., 2005). Treatment with either lincomycin or norflurazon dampened this

response, however, even without competent chloroplasts present the majority of genes were still light responsive.

Treatment of illuminated samples with norflurazon disrupted transcript abundance for three times as many genes as lincomycin. However, over 70% of the genes responsive to lincomycin were also responsive to norflurazon. Of these 713 transcripts that responded to both inhibitors, 640 genes were downregulated in response to inhibitor treatment. Overall, the number of genes displaying alterations in transcript abundance between treatments was consistent with the PCA analysis, with increased variance relating to an increased number of genes showing differential transcript abundance.

3.3.6 Function of genes most responsive to chloroplast perturbation

From the results of the differential gene analysis, genes showing the largest log2fold change between control and inhibited samples in illuminated conditions were identified. The ten most downregulated and upregulated genes out of the 3607 and 1010 genes, for norflurazon and lincomycin respectively, were assessed (Table 3.2, Table 3.3).

Gene ID	TAIR Gene Symbol(s)	TAIR Description	Log2fold Change	Down or Up Regulated By RS	Adjusted P-value
AT2G34430	LHB1B1, LHCB1.4, LIGHT-HARVESTING CHLOROPHYLL-PROTEIN COMPLEX II SUBUNIT B1	Photosystem II type I chlorophyll a/b-binding protein The mRNA is cell-to-cell mobile.	-8.410021696	Upregulated	8.04E-205
AT1G29910	AB180, CAB3, CHLOROPHYLL A/B BINDING PROTEIN 3, LHCB1.2, LIGHT HARVESTING CHLOROPHYLL A/B BINDING PROTEIN 1.2	member of Chlorophyll a/b-binding protein family	-6.984378905		2.86E-221
AT2G05070	LHCB2, LHCB2.2, LIGHT-HARVESTING CHLOROPHYLL B-BINDING 2, PHOTOSYSTEM II LIGHT HARVESTING COMPLEX GENE 2.2	Encodes Lhcb2.2. Belongs to the Lhc super-gene family encodes the light-harvesting chlorophyll a/b-binding (LHC) proteins that constitute the antenna system of the photosynthetic apparatus.	-6.559600852		1.50E-145
AT4G05180	PHOTOSYSTEM II SUBUNIT Q, PHOTOSYSTEM II SUBUNIT Q-2, PSBQ, PSBQ-2, PSII-Q	Encodes the PsbQ subunit of the oxygen evolving complex of photosystem II.	-6.134962841		8.74E-191
AT1G37130	ARABIDOPSIS NITRATE REDUCTASE 2, ATNR2, B29, CHL3, CHLORATE RESISTANT 3, NIA2, NIA2-1, NITRATE REDUCTASE, NITRATE REDUCTASE 2, NR, NR2	Identified as a mutant resistant to chlorate. Encodes nitrate reductase structural gene. Involved in nitrate assimilation. Has nitrate reductase activity. Up-regulated by the fungus P. indica. Binds transcription factor At2g35940. The mRNA is cell-to-cell mobile.	-6.031463789		1.45E-161
AT5G54270	LHCB3, LHCB3*1, LIGHT-HARVESTING CHLOROPHYLL B-BINDING PROTEIN 3	Lhcb3 protein is a component of the main light harvesting chlorophyll a/b-protein complex of Photosystem II (LHC II).	-5.816159303		1.58e-172
AT1G44575	CP22, NONPHOTOCHEMICAL QUENCHING 4, NPQ4, PHOTOSYSTEM II SUBUNIT S, PSBS	Encoding PSII-S (CP22), a ubiquitous pigment-binding protein associated with photosystem II (PSII) of higher plants. Involved in nonphotochemical quenching rather than in photosynthesis. Mutant has a normal violaxanthin cycle but has a limited capacity of quenching singlet excited chlorophylls and is tolerant to lipid peroxidation.	-5.792088164		1.72e-94
AT5G38420	DEG24, RBCS2B, RUBISCO SMALL SUBUNIT 2B	Encodes a member of the Rubisco small subunit (RBCS) multigene family: RBCS1A (At1g67090), RBCS1B (At5g38430), RBCS2B (At5g38420), and RBCS3B (At5g38410). Activated by OXS2 under the treatment of salt.	-5.680516421		4.56e-113
AT2G34620	MTERF10	Mitochondrial transcription termination factor family member.	-5.569769833		1.36e-32
AT2G34420	LHB1B2, LHCB1.5, PHOTOSYSTEM II LIGHT HARVESTING COMPLEX GENE 1.5, PHOTOSYSTEM II LIGHT HARVESTING COMPLEX GENE B1B2	Photosystem II type I chlorophyll a/b-binding protein	-5.344515006		2.23e-138
AT3G45060	ARABIDOPSIS THALIANA HIGH AFFINITY NITRATE TRANSPORTER 2.6, ATNRT2.6, HIGH AFFINITY NITRATE TRANSPORTER 2.6, NRT2.6	member of High affinity nitrate transporter family	3.78585697	Downregulated	4.94E-08
AT5G11920	6-&1-FRUCTAN EXOHYDROLASE, ATCWINV6, CWINV6	Encodes a protein with fructan exohydrolase (FEH) activity acting on both inulin and levan-type fructans (1- and 6-FEH). The enzyme does not have invertase activity.	3.582803468		8.46E-41
AT5G51480	SKS2, SKU5 SIMILAR 2	GPI anchored protein, highly expressed in reproductive tissues.	3.022902405		2.50E-07
AT3G16150	ASPARAGINASE B1, ASPGB1	Encodes an asparaginase that catalyzes the degradation of L-asparagine to L-aspartic acid and ammonia. The mRNA is cell-to-cell mobile.	2.995738308		1.05E-27 2.54E-15
AT5G58750	PRISE, PROGESTERONE 5β-REDUCTASE	Putative PRISE (progesterone 5β-reductase and/or iridoid synthase-like 1,4-enone reductases).	2.885980368		
AT5G24820		Eukaryotic aspartyl protease family protein	2.791161672		2.24e-05
AT1G19510	ATRL5, RAD-LIKE 5, RADIALIS-LIKE SANT/MYB 4, RL5, RSM4	RAD-like 5	2.707243877		9.21e-13
AT5G20240	PI, PISTILLATA	Floral homeotic gene encoding a MADS domain transcription factor. Required for the specification of petal and stamen identities.	2.670138303		5.30e-37
AT2G33160	NIMNA (SANSKRIT FOR '"SUNKEN" OR "LOW"), NMA	-	2.646019197		2.85e-05
AT3G21720	ICL, ISOCITRATE LYASE	Encodes a glyoxylate cycle enzyme isocitrate lyase (ICL) involved in salt tolerance	2.615092084		9.09e-31

Table 3.2: The 10 genes with the greatest negative log2fold change in expression, and the 10 genes with the greatest positive log2fold change in expression when illuminated controls are compared to illuminated norflurazon treated samples. Log2fold changes and adjusted p-values were calculated using DESeq2. TAIR = *The Arabidopsis Information Resource*. RS = Retrograde signalling.

Gene ID	TAIR Gene Symbol(s)	TAIR Description	Log2fold Change	Down or Up Regulated By RS	Adjusted P-value
AT1G73870	B-BOX DOMAIN PROTEIN 16, BBX16, COL7, CONSTANS-LIKE 7	B-box type zinc finger protein with CCT domain-containing protein	-4.02848911	Upregulated	8.35E-22
AT2G27420	-	Cysteine proteinases superfamily protein	-3.525055019		8.22E-23
AT1G05560	UDP-GLUCOSE TRANSFERASE 1, UDP-GLUCOSYLTRANSFERASE 75B1, UGT1, UGT75B1	A UDP-glucose transferase localized in the phragmoplast. It has been co-purified with the callose synthase complex and may transfer UDP-glucose from sucrose synthase to the callose synthase and thus help form a substrate channel for the synthesis of callose at the forming cell plate. Induced by salicylic acid. Independent of NPR1 for their induction by salicylic acid.	-3.131875543		2.07E-24
AT2G15620	ARABIDOPSIS THALIANA NITRITE REDUCTASE, ATHNIR, NIR, NIR1, NITRITE REDUCTASE, NITRITE REDUCTASE 1	Involved in the second step of nitrate assimilation. Its expression is induced by nitrate. The mRNA is cell-to-cell mobile.	-2.626544363		9.95E-22
AT3G06145	-	RING zinc finger protein	-2.605411789		1.99E-11
AT1G72930	ATTN10, TIR, TIR-NUCLEOTIDE BINDING SITE FAMILY 10, TN10, TOLL/INTERLEUKIN-1 RECEPTOR-LIKE	Toll/interleukin-1 receptor-like protein (TIR) Mrna	-2.557771170		3.50e-08
AT4G09650	ATP SYNTHASE DELTA-SUBUNIT GENE, ATPD, PDE332, PIGMENT DEFECTIVE 332	Encodes the chloroplast ATPase delta-subunit. The mRNA is cell-to-cell mobile.	-2.502232464		4.08e-18
AT1G32900	GBSS1, GRANULE BOUND STARCH SYNTHASE 1	UDP-Glycosyltransferase superfamily protein	-2.448058750		4.61e-08
AT4G10120	ATSPS4F, SPS4F, SUCROSE PHOSPHATE SYNTHASE 4F	Encodes a protein with putative sucrose-phosphate synthase activity.	-2.445458532		3.79e-11
AT3G48310	"CYTOCHROME P450, FAMILY 71, SUBFAMILY A, POLYPEPTIDE 22", CYP71A22	putative cytochrome P450	-2.431108397		9.68e-07
AT1G36180	ACC2, ACETYL-COA CARBOXYLASE 2	acetyl-CoA carboxylase 2 (ACC2) The mRNA is cell-to-cell mobile.	4.128386599	Downregulated	3.92E-43
AT2G18193	-	P-loop containing nucleoside triphosphate hydrolases superfamily protein	3.780653173		5.37E-41
AT3G58270	-	phospholipase-like protein (PEARLI 4) with TRAF-like domain protein	3.032676713		1.77E-24
AT5G55270	-	hypothetical protein (DUF295)	3.01355143		1.78E-09
AT3G54520	-	hypothetical protein	2.968497217		1.23E-09
AT1G17960	-	Threonyl-tRNA synthetase	2.921047934		8.13e-21
AT1G30170	ATDOA4, DUF295 ORGANELLAR A 4	Hypothetical protein (DUF295) of unknown function	2.905921605		1.01e-08
AT5G24280	GAMMA-IRRADIATION AND MITOMYCIN C INDUCED 1, GMI1	Encodes GMI1, a structural-maintenance-of-chromosomes-hinge domain-containing protein. Involved in somatic homologous recombination.	2.777250726		1.21e-12
AT5G03495	-	RNA-binding (RRM/RBD/RNP motifs) family protein	2.750761110		1.01e-08
AT5G64060	ANAC103, NAC DOMAIN CONTAINING PROTEIN 103, NAC103	NAC domain containing protein 103	2.671560738		4.30e-08

Table 3.3: The 10 genes with the greatest negative log2fold change in expression, and the 10 genes with the greatest positive log2fold change in expression, when illuminated controls are compared to illuminated lincomycin treated samples. Log2fold changes and adjusted p-values were calculated using DESeq2. TAIR = *The Arabidopsis Information Resource.* RS = Retrograde signalling.

The genes most downregulated by norflurazon treatment were the photosynthesis-related genes. Lincomycin, however, does not elicit a large response in photosynthesis associated gene expression. In fact, for lincomycin, seven of the ten genes assessed had no associated TAIR gene symbol and vague gene descriptions, highlighting that this is an area that still requires a great deal of further research. Apart from the expression of reductases in the nitrogen pathway significantly decreasing in treated seedlings, there was not a large overlap in genes with the greatest change in expression between inhibitors (Figure 3.3 and 3.4). However, each inhibitor significantly altered the abundance

of over a thousand transcripts. As a broader picture of these genes would give a more accurate comparison of the effects of lincomycin and norflurazon, Gene Ontology (GO) terms that were enriched in each of these categories were assessed next.

3.3.7 Enrichment analysis of genes impacted by chloroplast perturbation in illuminated seedlings

Gene Ontology provides umbrella terms for gene functionality, and so the analysis can provide an overview of the impact of each treatment on global patterns of transcript abundance. To investigate genes that were affected most by the chloroplast and its signals, enrichment analysis was performed on data from illuminated inhibitor treated samples compared to controls. This showed that 79% of the significantly enriched terms from the lincomycin responsive gene set are also enriched in response to norflurazon (Figure 3.8).

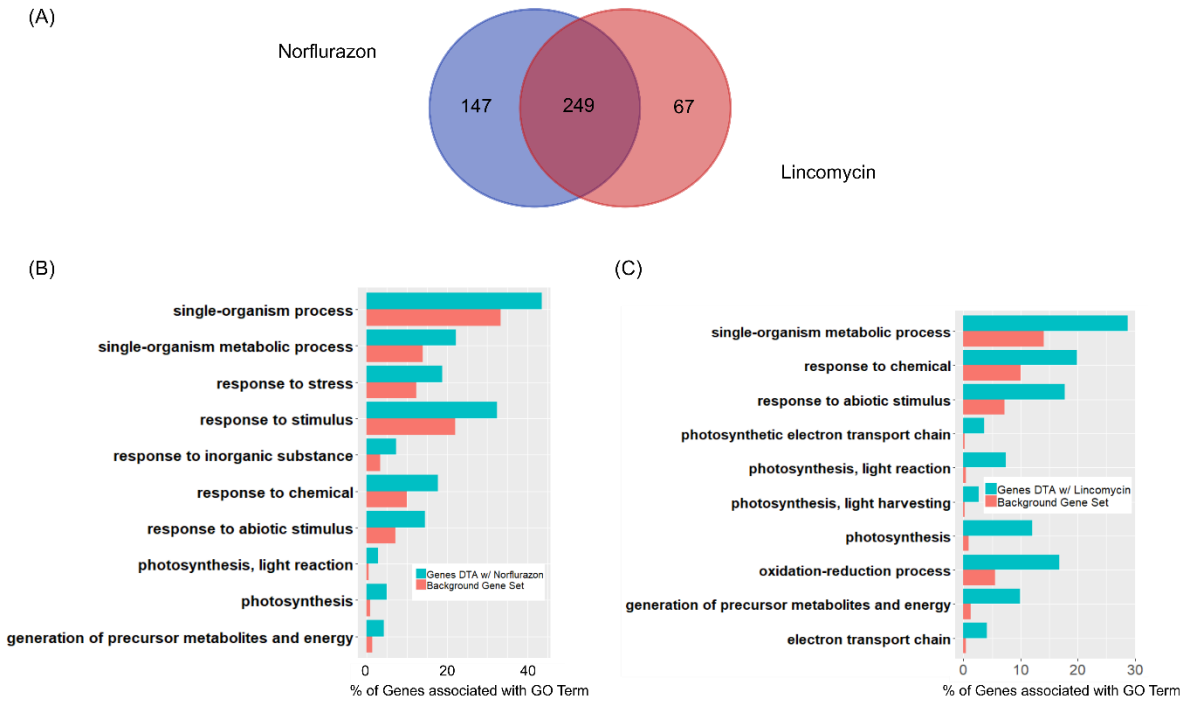


Figure 3.8: Gene Ontology analysis indicated that treatment with each inhibitor affected similar processes, with photosynthesis genes being some of the most disproportionally affected genes when seedlings are treated with either norflurazon or lincomycin. (A) Venn Diagram showing the overlap in significantly enriched GO terms after chloroplast perturbation by lincomycin and norflurazon. (B) Bar chart showing enrichment for each GO term in the inhibitor responsive genes compared to the rest of the genome. The Background Gene Set used contained all available *A. thaliana* annotated genes. GO = Gene Ontology. DTA = Significantly Different Transcript Abundance. Linc = 0.5mM lincomycin, NF =5µM norflurazon.

Photosynthesis and light harvesting are two of these mutually enriched terms for both lincomycin and norflurazon. This overlap in large-scale molecular changes suggests that despite their different modes

of action, norflurazon and lincomycin led to broadly similar impacts on gene expression. Further analysis employing MapMan (Usadel et al., 2009) demonstrated that photosynthetic genes were downregulated in the presence of compromised chloroplasts (Figure 3.9).

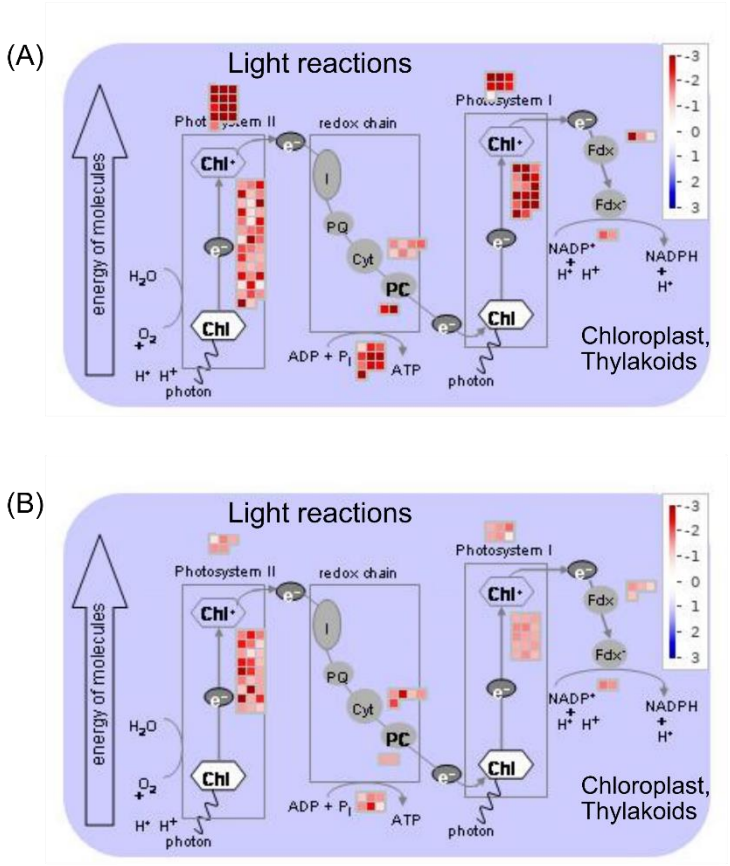


Figure 3.9: Chloroplast perturbation caused a decrease in expression of most photosynthetic genes. Image generated using MapMan software showing analysis of differential expression data for (A) lincomycin treated seedlings compared with illuminated controls or (B) norflurazon treated seedlings compared to illuminated controls. False colour scale represents log2fold change.

Whilst all genes involved in the light reactions of photosynthesis were downregulated by chloroplast perturbation, genes encoding Photosystem I and Photosystem II showed greater downregulation in the presence of lincomycin. These protein complexes are comprised of both nuclear and plastome encoded subunits (Pfannschmidt, et al., 2001), which could explain their high responsiveness.
Gene ontology analysis confirmed that both lincomycin and norflurazon have a similar effect on differential transcript abundance. This supports the assumption that looking at genes showing similar responses after seedlings were treated with each of these inhibitors could give robust insights into the effects of chloroplast inhibition, rather than off-target effects.

3.3.8 Expression Patterns of Retrograde Signals and Well-characterised Responders

Transcript abundance of genes with a well characterised response such as *LHCB* were analysed to check the consistency of these data with previously published findings. A number of genes have been proposed to have a role in retrograde signalling. Six *GUN* genes have been shown to be definitively involved, although the mechanism by which they effect gene expression is unclear. These genes were

identified in *gun* mutants; plants which showed no change in *LHCB* expression when treated with either lincomycin or norflurazon. Additionally, there are genes, such as those encoding Executor proteins and HY5, with supporting evidence for a prospective role in chloroplast retrograde signalling. The response of these proposed retrograde signals to chloroplast perturbation was analysed.

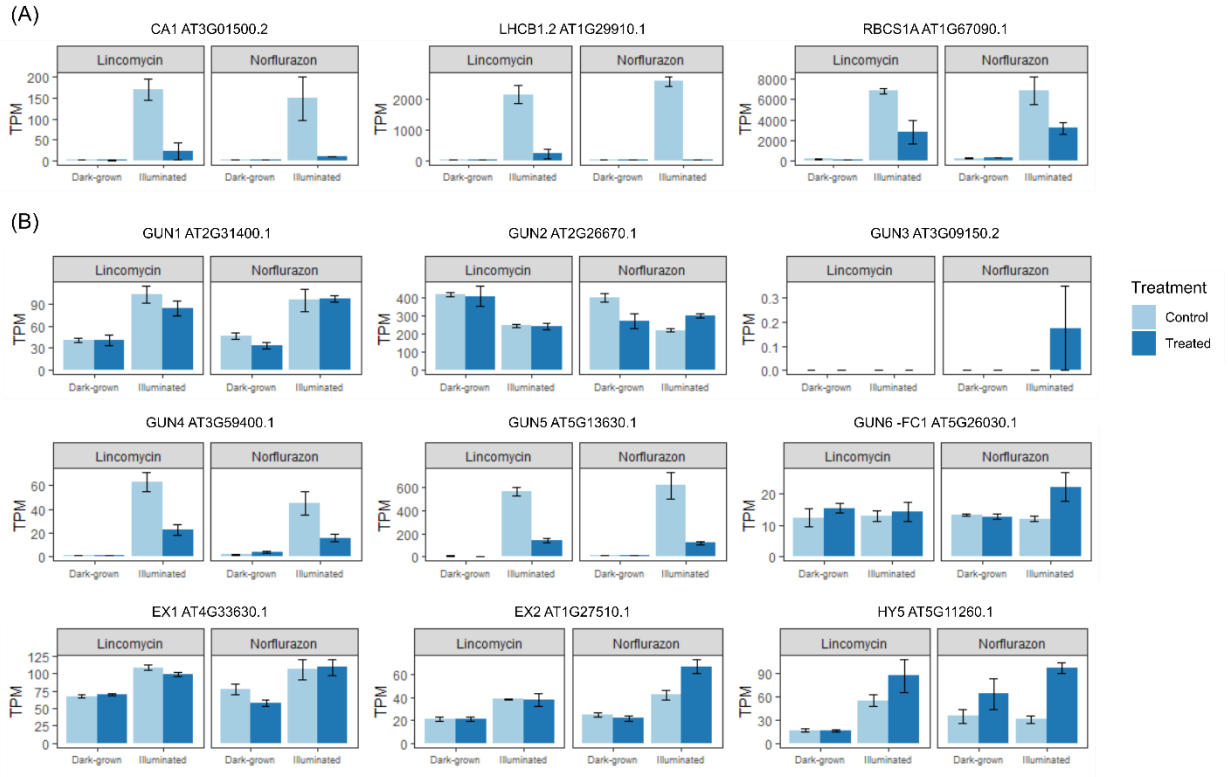


Figure 3.10: Histogram showing average Transcripts Per Million (TPM) values of well-characterised retrograde responders and genes associated with retrograde signals. Dark-grown = seedlings dark-grown for 6 days; Illuminated = seedlings dark-grown for 5 days and grown in 150 µmol m^{-2} s^{-1} white light for 24 hours. Error bars represent one standard error of the mean. *CA1 = Carbonic anhydrase 1, LHCB1.2 = Chlorophyll a-b binding protein 3, RBCS1A = Ribulose bisphosphate carboxylase small chain 1a, GUN 1-6 = Genomes uncoupled 1-6, EX1 = Executer 1, EX2 = Executer 2, HY5 = Long hypocotyl 5.*

Expression of *LHCB1.2*, *Carbonic Anhydrase* (*CA1*) and *Ribulose Bisphosphate Carboxylase Small Sub-unit* (*RBCS1A*) have been shown to decrease in response to chloroplast inhibition in wildtype seedlings (Kacprzak et al., 2019a). This response is clearly seen in the data presented here (Figure 3.10). *GUN4* and *GUN5* showed a similar expression pattern, implying that their mode of regulation could take place at the transcriptional level. However, most of the retrograde signalling genes appear to be affected by light only, implying their regulatory effect is not proportional to transcript abundance.

3.3.9 Expression Pattern of C_4 Orthologues in *A. thaliana*

An initial inspection of transcript abundance in chloroplast inhibited seedling implied that *CA1* was not the only C_4 orthologue subject to regulation by the chloroplast (Figure 3.11).

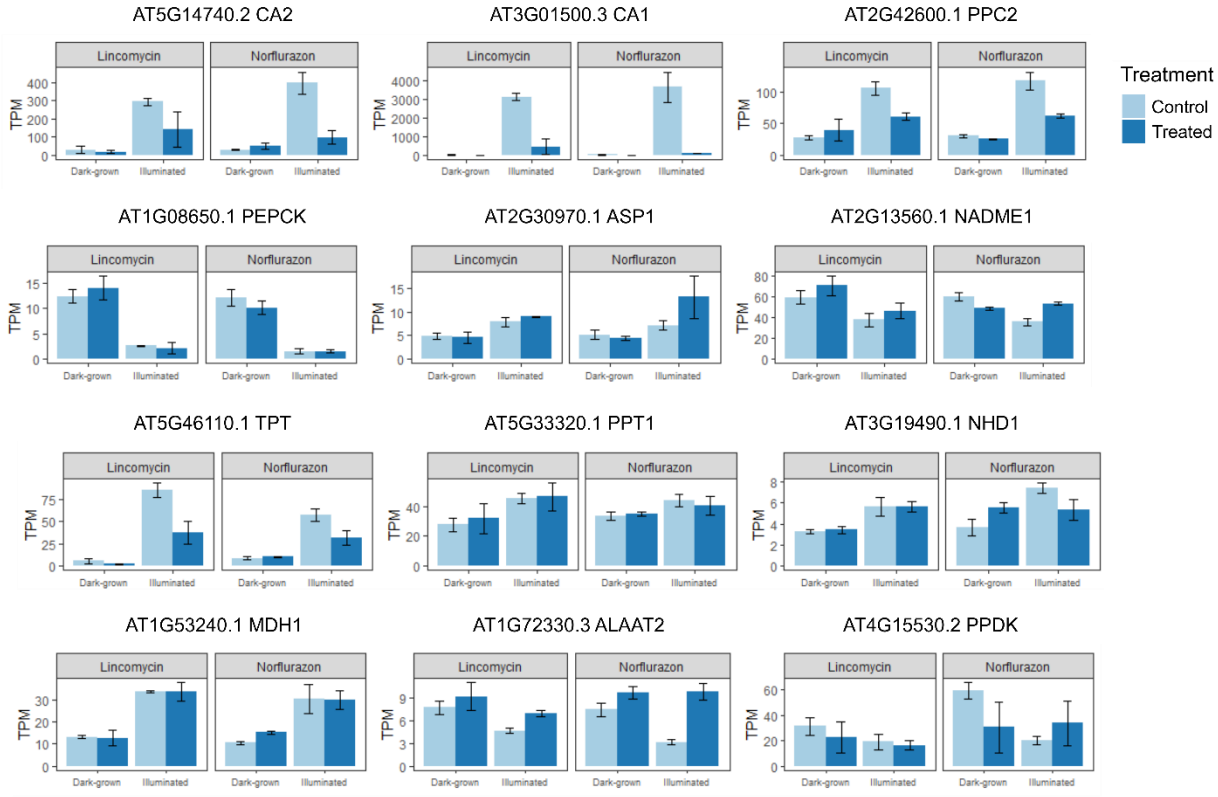


Figure 3.11: Histogram showing average TPM values of C4 orthologues in A. thaliana. Dark-grown = seedlings dark-grown for 6 days; Illuminated = seedlings dark-grown for 5 days and grown in 150 µmol m−2 s−1 white light for 24 hours. Error bars represent one standard error of the mean. CA2 = Carbonic anhydrase 2, CA1 = Carbonic anhydrase 1, PPC2 = Phosphoenolpyruvate carboxylase 2, PEPCK = Phosphoenolpyruvate carboxylase kinase 1, ASP1 = Aspartate aminotransferase, NADME1 = NAD-dependent malic enzyme 1, TPT = Triose phosphate/phosphate translocator, PPT1 = Phosphoenolpyruvate/phosphate translocator 1, NDH1 = Sodium/proton antiporter 1, MDH1 = Malate dehydrogenase 1, ALAAT2 = Alanine aminotransferase 2, PPDK = Pyruvate, phosphate dikinase 1.

Chloroplast inhibition caused a significant decrease in transcript abundance in *CA1* and *PPC2*, genes which function in the initial carbon dioxide fixation within mesophyll cell (Figure 3.12).

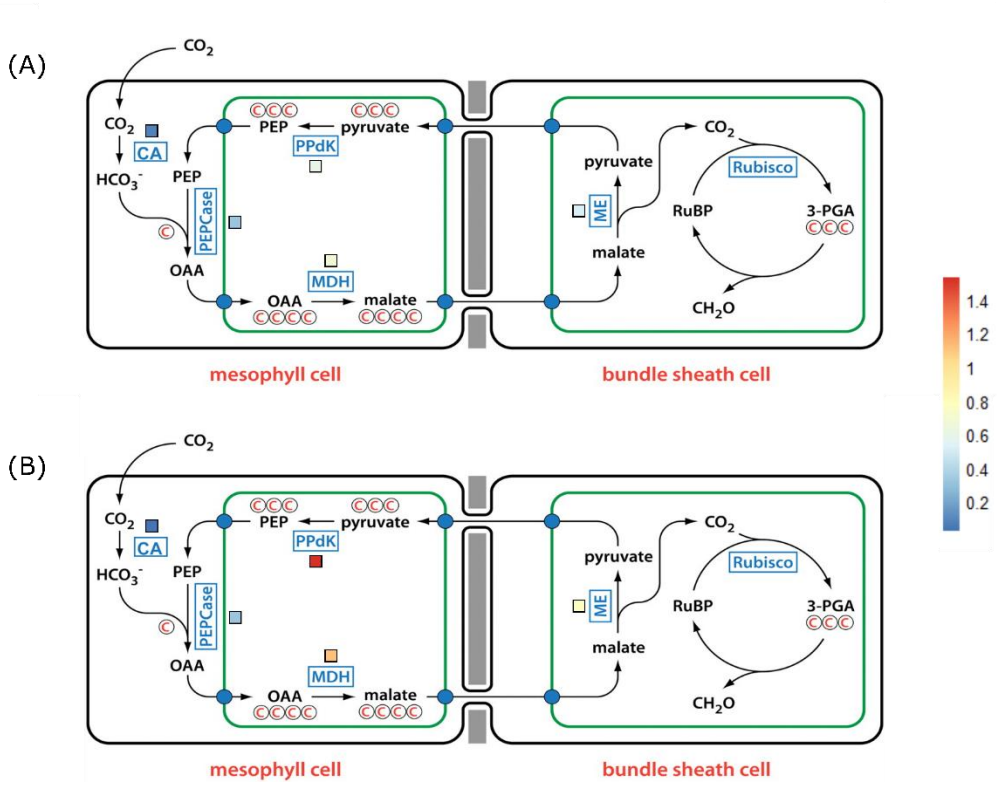


Figure 3.12: Chloroplast perturbation caused a decrease in expression of genes involved in CO_2 fixation in mesophyll cells. Image adapted from Langdale (2011) showing the ratio of transcript abundance in (A) lincomycin treated or (B) norflurazon treated seedlings compared to controls. False colour scale represents transcript abundance ratio.

To statistically analyse whether the expression of C_4 orthologues in *Arabidopsis thaliana* was significantly affected by chloroplast perturbation, pairwise comparisons were performed between all treatments using DESeq2. Comparison of inhibitor-treated seedlings versus controls was used to investigate if retrograde signalling was involved in regulation. To determine if retrograde signalling was interconnected with light regulation, this comparison was carried out in both dark-grown and illuminated samples. Very few genes were regulated by light only. Most differentially expressed C_4 orthologues were affected by both the chloroplast and light (Figure 3.13).

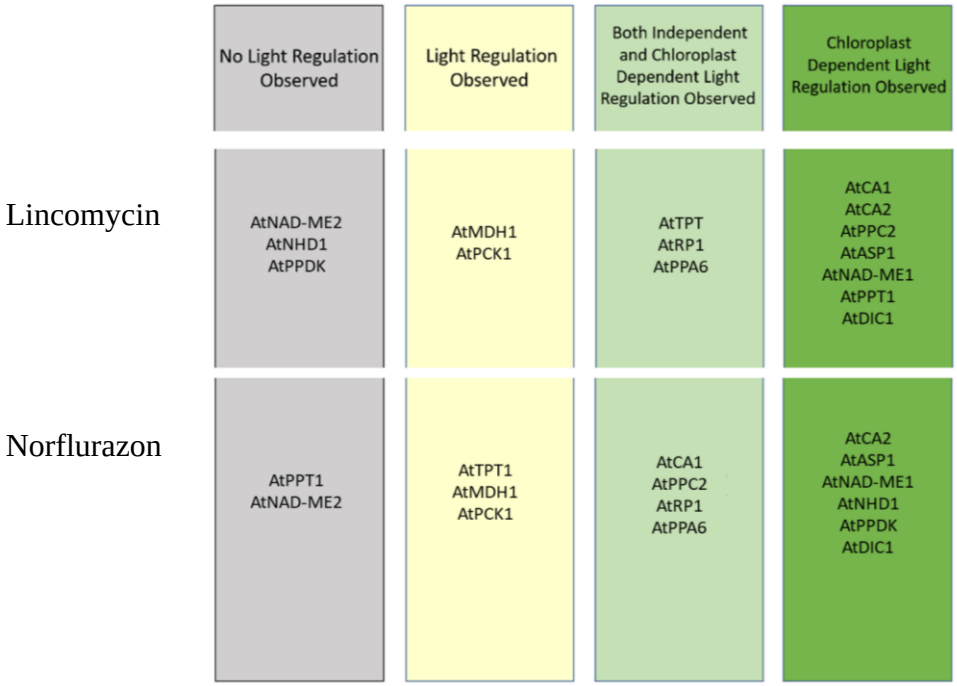


Figure 3.13: The majority of C_4 orthologues are subject to some extent of regulation by the chloroplast. C_4 orthologues (n = 15) separated by genetic regulation, based on differential expression between four different treatments. Chloroplast-dependent regulation defined as genes which are significantly differentially expressed in controls compared to seedlings which have undergone chloroplast perturbation by (A) Lincomycin or (B) Norflurazon.

The response of genes did differ slightly between the two inhibitors. However, the majority of C_4 orthologues are under chloroplast control. Chloroplast regulation was always found to be light dependent, with no C_4 orthologues showing consistent differences between controls and inhibited seedling in both dark-grown and illuminated conditions. 8 of the 15 analysed C_4 orthologues were differentially expressed both when lincomycin was applied, and when norflurazon was applied (Figure 3.13).

3.3.10 Patterns of Gene Expression in Response to Chloroplast Inhibition

To detect groups of genes which have a strong consistent response to chloroplast perturbation, a clustering analysis was carried out. Clust analysis finds broad patterns of gene expression occurring in transcriptomic data (Abu-Jamous & Kelly, 2018). Read counts are analysed by the programme, which then clusters together genes which share similar expression patterns in response to all experimental treatments. The Clust analysis found 12 gene clusters for each inhibitor experiment (Figure 3.14).

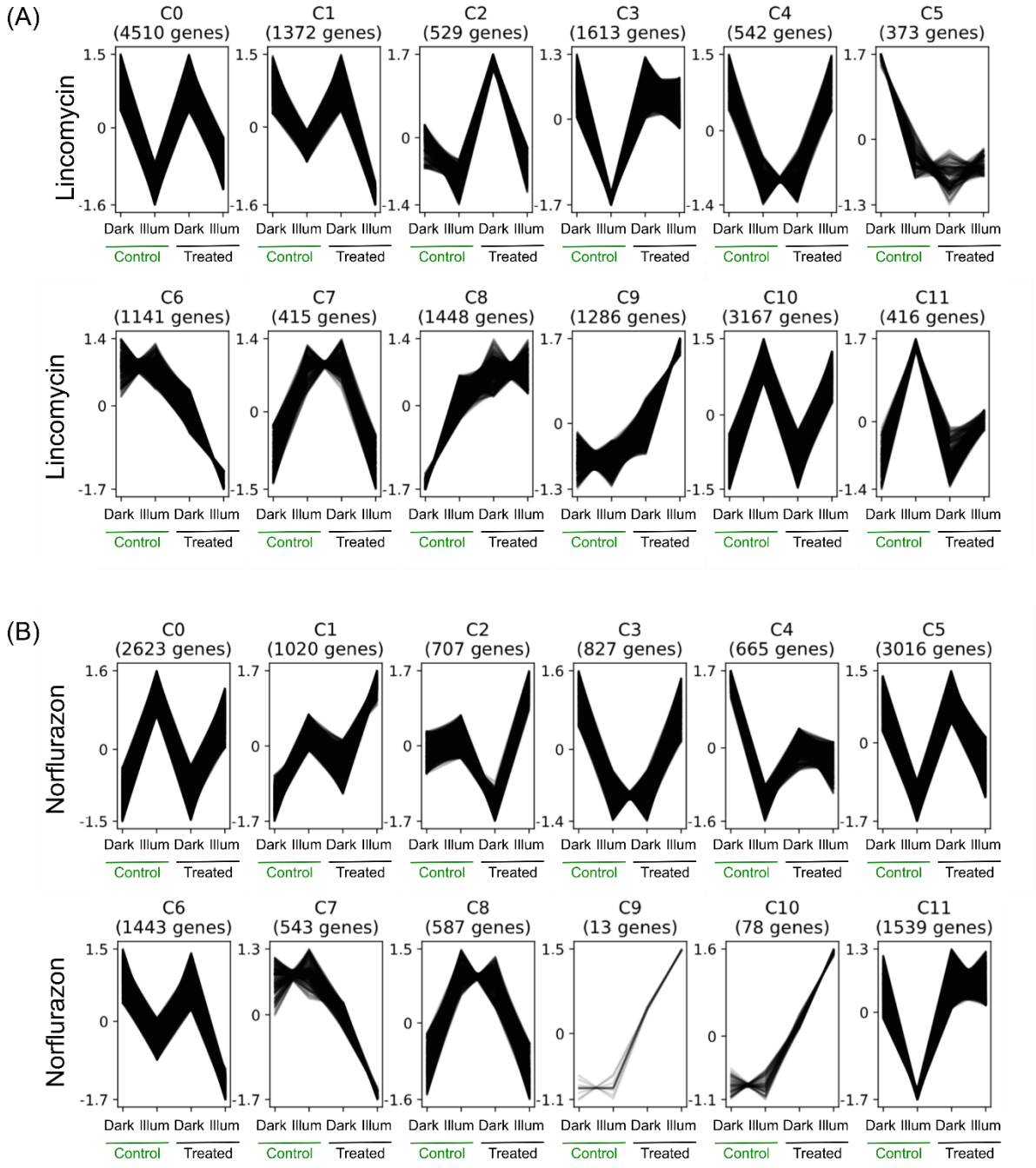


Figure 3.14: Cluster profiles generated by Clust showing expression patterns for (A) lincomycin and (B) norflurazon. Cluster number is arbitrary and for identification only. Lincomycin = 0.5mM lincomycin, Norflurazon = 5µM, Dark = 6 days dark-grown, Illum = 5 days dark-grown + 24 hours 150 µmol m^{-2} s^{-1} white light.

These clusters comprise 41% of the total input genes for lincomycin, and 32% for norflurazon. Clust analysis showed that the largest shared regulatory pattern contained light-responsive genes. Cluster C10 for lincomycin and C0 for norflurazon, containing 3167 and 2623 genes respectively, represented genes which were upregulated in the light. Conversely, Cluster C0 for lincomycin and C5 for norflurazon, containing 4510 and 3016 genes respectively, represented genes which were downregulated in the light. The analysis identified clusters that showed response to chloroplast inhibition, and some of the response patterns were shared by both inhibitors. For example, C3 for

lincomycin and C11 for norflurazon contained genes which in control samples were downregulated in response to light, however this response was lost when the chloroplast was perturbed. A basic GO analysis of each cluster was carried out to identify if any gene expression pattern correlated with function, however, the clusters did not show any specifically enriched terms.

3.3.11 Motif Enrichment Analysis

The analysis above identified 713 genes which have differential transcript abundance in both lincomycin and norflurazon treated seedlings compared to controls (Table 3.1). This list of genes enables a search for enriched motifs; short sequences of DNA recognised by transcription factors to facilitate binding. Enriched motifs can highlight putative transcription factors which could act as a chloroplast derived signal modifying nuclear gene expression.

3.3.12 Identifying Enriched Repetitive 6 base-pair Elements

To identify a shared regulatory mechanism which could potentially be a chloroplast to nucleus signal, the sequences of genes responsive to chloroplast inhibition were searched for known binding motifs. The simplest method computationally for this analysis is to scan for enriched 6-mer sequences using the web-based tool TAIR Motif Analysis. This creates an index of every possible 6 base sequence within the gene's promoter and compares the frequency of this motif in the genome to its frequency in the list of responsive genes to generate a p-value. The input to this analysis tool is the TAIR IDs of the genes of interest and a selection of how many base pairs upstream of the transcription start site (TSS) is to be analysed. A 1000bp upstream of the TSS of each chloroplast responsive gene was analysed, resulting in 3889 6-mers being significantly enriched (p-value < 0.05). Enriched motifs included the G-box, ABA responsive elements and motifs that have been reported to be involved in sugar metabolism (Table 3.4).

However, whilst this approach is rapid, it has limitations and is particularly constrained by the set motif length. Repeated sequences appeared in the enriched 6-mer list; the 'TATA' box occurred in both 'ATATAG' and 'CTATAT'. Additionally the enriched 'TGGATA' sequence contained the characterised 4-base motif GATA, known to be enriched in genes under light-dependent and nitrate-dependent regulation (Reyes et al., 2004). This method does not allow the enrichment of GATA or TATA to be analysed individually. The approach is also constrained by its lack of flexibility; in the cell transcription factors do not often have exclusive binding to one sequence. Variability in bases at certain positions may decrease the likelihood of binding, however, it will not stop binding from happening.

OligoMer	No of Plastid-responsive Genes containing OligoMer	No of Genes containing OligoMer in Genome	P-value	Description
CACGTG	142/713	4741/34292	1.73e-06	G-box
ACGTGG	185/713	6707/34292	6.28e-06	ABA-responsive element
CCACGT	185/713	6707/34292	6.28e-06	ABA-responsive element
GGACCA	200/713	7528/34292	2.27e-05	Class II TCP binding site
TGGTCC	200/713	7528/34292	2.27e-05	Class II TCP binding site
TATCCA	359/713	14947/34292	4.17e-05	Sugar-responsive element
TGGATA	359/713	14947/34292	4.17e-05	GATA binding domain
ATATAG	379/713	20511/34292	4.61e-05	TATA-like
CTATAT	379/713	20511/34292	4.61e-05	TATA-like
CGTGGC	121/713	4245/34292	7.24e-05	bHLH binding domain

Table 3.4: G-box and ABA responsive elements are enriched in genes responsive to chloroplast perturbation. 12 most enriched 6-mers within the 1000bp upstream of genes which are significantly downregulated by both inhibitors in light. Calculated using the TAIR tool ‘Statistical Motif Analysis in Promoter or Upstream Gene Sequences’.

3.3.13 Motif Enrichment in Chloroplast Responsive Genes

FIMO, part of the MEME software suite, is a programme which searches for known motifs by using experimentally verified motifs identified using methods such as DAP-seq and CHIP-seq. These motifs are stored as letter probability matrices, or position weight matrices, accounting for the inexact nature of transcription factor binding. The 5’UTR and 1000bp upstream of plastid-responsive genes was analysed using FIMO (Table 3.5).

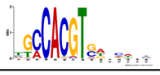
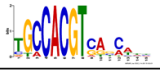
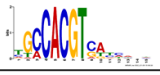
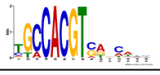
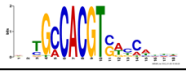
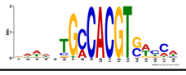
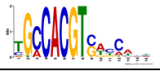
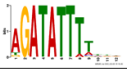
Motif	Total Motif Frequency	Fold change in Frequency compared to Average Background	Number of Genes in which the Motif Occurs	PWM
MYBrelated_tnt.AT3G10113_col_a_m1	26	4.64	21	
MYBrelated_tnt.At5g52660_colamp_a_m1	27	4.5	23	
bZIP_tnt.GBF3_colamp_m1	184	4.44	101	
bZIP_tnt.GBF3_col_m1	128	4.41	133	
bZIP_tnt.bZIP28_col_a_m1	130	4.36	101	
bZIP_tnt.bZIP68_col_a_m1	151	4.34	112	
bZIP_tnt.ABI5_col_v3h_m1	172	4.28	115	
bZIP_tnt.ABF2_col_v3a_m1	179	4.26	122	
bZIP_tnt.ABI5_colamp_v3b_m1	173	4.24	115	
MYBrelated_tnt.At3g09600_col_a_m1	22	4.23	18	

Table 3.5: The ten most enriched motifs within the promoter (1000bps upstream of TSS) and 5' UTR of chloroplast responsive genes with a total frequency over 20. Enrichment measured as fold change in the frequency of significant matches as calculated by FIMO compared to the average of 5 random background gene sets. PWM = position weight matrix.

This analysis showed that binding motifs that were enriched more than 2.5-fold compared to background in chloroplast responsive genes were from five major transcription factor families: MYB, TCP, bZIP, bHLH and MYB-related. A natural consequence of chloroplast inhibition is strong downregulation of photosynthesis related genes, meaning the two are confounded. To distinguish the response to chloroplast inhibition specifically, genes relating to the 'photosynthesis' GO term were also analysed. MYB-related binding motifs were enriched in this dataset but motifs for the other four transcription factor families were not. Binding sites for transcription factors from the same family can share similar motif sequences, as can be seen by the highly similar PWMs (Table 3.5). By plotting the location of similar motifs, such as ABI5 and ABF2, across the promoter of HEME1 we can see that these predicted sites are redundant, although they do vary in the confidence of the match represented in the p-value on the y-axis (Figure 3.15). This could be biological, with multiple transcription factors able to bind at a single site with varied affinities, or this could be unavoidable error in the predictive analysis.

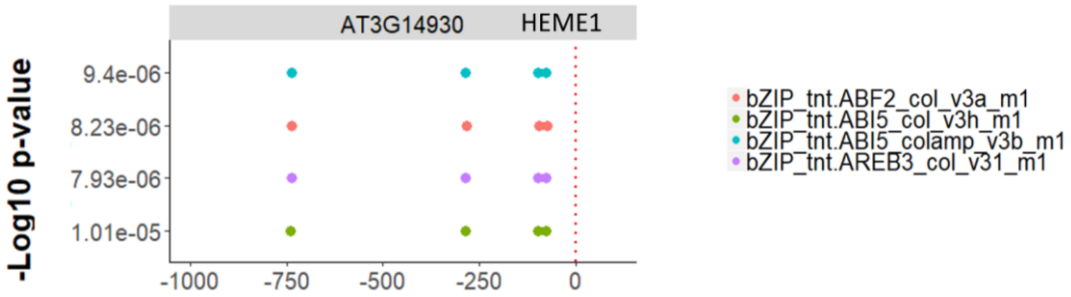


Figure 3.15: Transcription factor within a family are predicted to bind the same sites, although they do vary in the confidence of the match represented in the p-value on the y-axis. Schematic of the position of predicted enriched motifs along the gene promoter (1000bps upstream of TSS) and 5' UTR of *HEME1*. P-value decreases, and therefore likelihood of enrichment increases, along y-axis. 0 = TSS.

Additional transcription factor candidates were screened by analysing enriched motifs in the promoter of known responders, such as *LHCB2.1* and *CA1* whose expression is known to decrease in response to chloroplast perturbation. Two potential candidates were identified; AZF1 and AT3G46070 (Figure 3.16). However, AT3G46070 is not expressed in any samples making it likely AZF1 is binding to these sites.

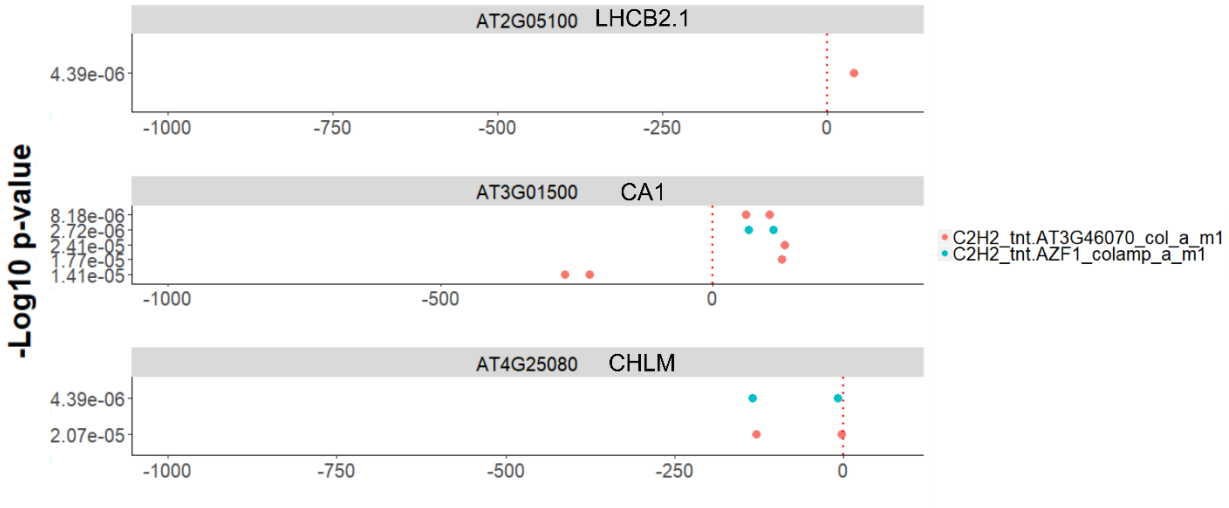


Figure 3.16: AZF1 is predicted to bind to motifs in the promoter of known chloroplast responders and is a candidate chloroplast to nucleus regulator. Schematic of the position of predicted enriched motifs in known chloroplast responders along the gene promoter (1000bps upstream of TSS) and 5' UTR. P-value decreases, and therefore likelihood of enrichment increases, along y-axis. 0 = TSS.

3.3.14 Motif Enrichment Analysis of Chloroplast Responsive Genes using AME

An additional tool available in the MEME software suite is AME (Analysis of Motif Enrichment). Whilst FIMO (Find Individual Motifs) scans a set of sequences for individual matches to each of the motifs you provide, AME identifies motifs that are relatively enriched in your sequences compared with control sequences. Analysis of chloroplast regulated genes with AME found five main transcription factor families were enriched: AP2/EREBP, bZIP, HMG, BZR and LOBAS2. The overall ranking of significantly enriched motifs was compared between the AME and FIMO approach (Figure 3.17).

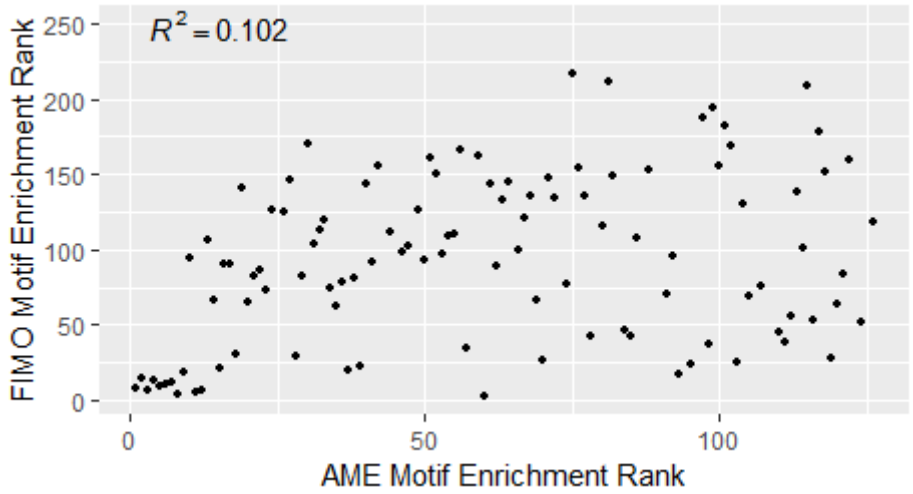


Figure 3.17: The resulting motif enrichment from an AME analysis compared to a FIMO analysis are disparate. Scatter plot of the rank for each motif based on its enrichment reported in AME plotted against the rank for each motif based on its enrichment above background in the FIMO analysis (filtered for motifs with a total frequency > 20).

Overall, we see there is little correlation between the two. This initially may seem surprising as both methods are looking for enrichment. However, their methods differ in that FIMO prioritises individual matches and looks for high affinity binding, whereas, in AME the frequency is prioritised. This means AME highlights numerous weak matches over FIMO's prioritisation of less frequent strong matches. It is worth noting however, that the left side of the plot shows a brief increase in correlation within motifs with the very highest enrichment (Figure 3.17). This means that FIMO and AME identify the same motifs within the highest enrichment category, likely due to these motifs occurring at a high frequency as well as having a high binding affinity. A more detailed look shows that the nine most enriched motifs within the AME analysis are also in the twenty most enriched motifs within the FIMO analysis (Table 3.6).

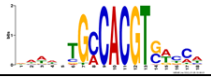
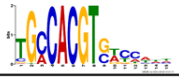
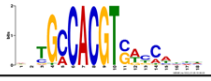
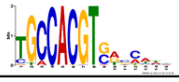
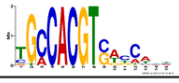
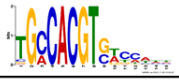
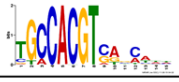
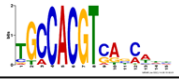
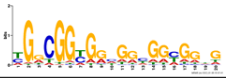
Motif	P-value for Enrichment	Rank of Enrichment in Fimo Analysis	PWM
bZIP_tnt.ABF2_col_v3a_m1	3.28E-13	8	
bZIP_tnt.AREB3_col_v31_m1	1.3E-12	14	
bZIP_tnt.ABI5_col_v3h_m1	1.68E-12	7	
bZIP_tnt.bZIP16_col_v3a_m1	1.79E-12	13	
bZIP_tnt.ABI5_colamp_v3b_m1	2.08E-12	9	
bZIP_tnt.AREB3_colamp_a_m1	2.17E-12	11	
bZIP_tnt.bZIP16_colamp_a_m1	1.18E-11	12	
bZIP_tnt.GBF3_col_m1	3.45E-10	4	
bZIP_tnt.GBF5_colamp_a_m1	9.55E-10	18	
AP2EREBP_tnt.AT1G71450_col_a_m1	2.84E-09	240	

Table 3.6: The ten most enriched motifs within the promoter (1000bps upstream of TSS) and 5' UTR of chloroplast responsive genes calculated by AME. Enrichment measured as the p-value of likelihood motif occurs more frequently in the genes set of interest over background. PWM = position weight matrix.

These motifs are likely most enriched in both the FIMO and AME analyses as they occur frequently within chloroplast-regulated genes and they also have a high affinity for binding. Therefore, these six motifs are promising candidates to analyse in future work.

3.3.15 Distribution of open chromatin in *A. thaliana*

It has been shown that transcription factors are more likely to bind areas of open chromatin. Looking at the distribution of open chromatin regions can indicate which parts of the genetic architecture are likely rich in binding motifs. The distribution of open chromatin was therefore analysed to inform a refined motif analysis, where only the most accessible elements of the genetic architecture are considered.

DNaseI-Seq treats intact isolated nuclei with DNaseI which preferentially cuts open chromatin. By sequencing the resulting DNA, the cut sites can be identified. In the literature, these open chromatin regions can be referred to as DNase hypersensitive sites (DHSs) or peaks. This is because there is a peak in the number of cuts at these sites. The above analysis looked only at the 5' UTR and 1000 base pairs upstream of the transcription start site (Section 3.3.13 and 3.3.14). This information could

inform the most relevant areas of a gene's sequence to input into the motif analysis. DNaseI-Seq data was used from one-week old *A. thaliana* seedlings grown in darkness for 7 days and then exposed to light from 0 minutes to 24 hours (Sullivan et al., 2014). The bed file containing the coordinates for these DHSs site was downloaded from NCBI (GSE53322) and input into PAVIS2 to produce pie charts showing how these DHSs distribute in the *A. thaliana* genome (Figure 3.18). PAVIS2 is a tool to facilitate visualisation of the distribution of peaks across genomic features (Huang et al., 2013).

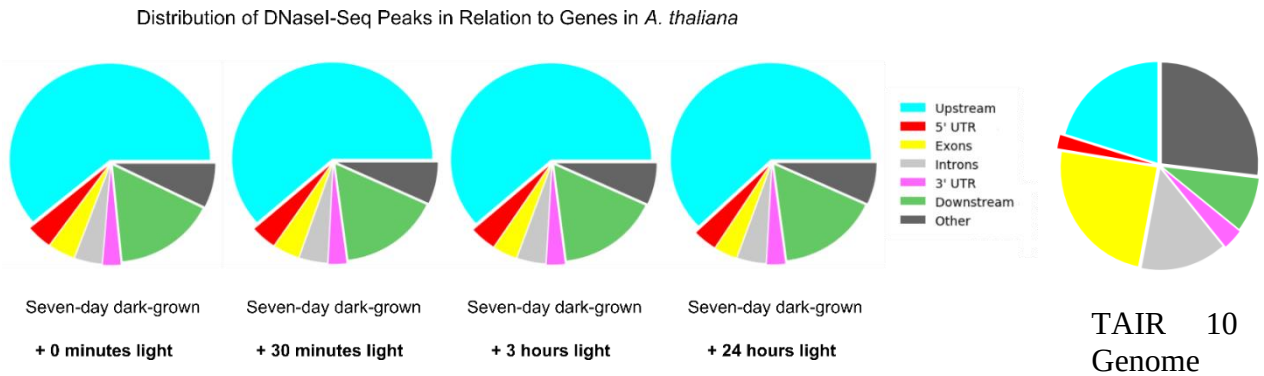


Figure 3.18: The distribution of accessible chromatin sites between gene features is in dark-grown and light exposed seedlings (Sullivan et al., 2014; NCBI (GSE53 rts generated using PAVIS2 software and the TAIR10 genome annotation. Upstream is defined as 5000bp upstream of the transcription start site. Downstream is defined as 1000bp downstream of the transcription termination site.

This analysis showed that peak distribution did not change with light treatment, and therefore this is not a further consideration in the analysis. The plots show that ~8.5% of the accessible peaks occur within the gene body. Whilst this could be considered a small proportion, these motifs are easily attributed to a specific gene and therefore are a valuable asset in the analysis. The analysis pipeline was therefore edited to incorporate the gene body in the input sequence. Over 60% of peaks were found within 5000bp upstream of the TSS in the promoter region. To investigate the distribution of these peaks, their distance from the TSS was plotted (Figure 3.19).

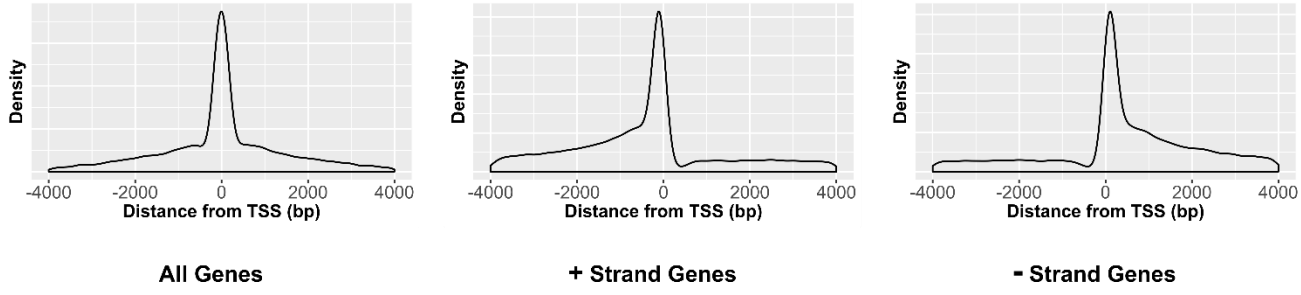


Figure 3.19: Distribution of DHS (DNaseI Hypersensitive sites) around the TSS for all annotated *A. thaliana* genes, showing that most DHSs are covered by 2000bp of the promoter region (Sullivan et al., 2014; NCBI (GSE53322). 0 = TSS. + Strand represents positive strand genes which read from 5' to 3', – Strand represents negative strand genes which read from 3' to 5'.

Over 50% of DNaseI hypersensitive sites occur less than 2000 base pairs upstream of the TSS, making this a suitable length of promoter to include in further motif analysis. However, 434 genes either overlap or have their TSS directly adjacent to another gene and 18486 genes have less than 2000bp between their TSS and the coding sequence of another gene. This means the sequence selected for the analysis would often partially include an adjacent gene to the gene of interest. To ensure that a 'one size fits all' approach analysing 2000bp upstream along with the gene body is appropriate, Sullivan DHS distribution plots were generated for genes with a low inter-gene distance between the 3' end and the next gene to ascertain whether they require a different analysis method (Figure 3.20).

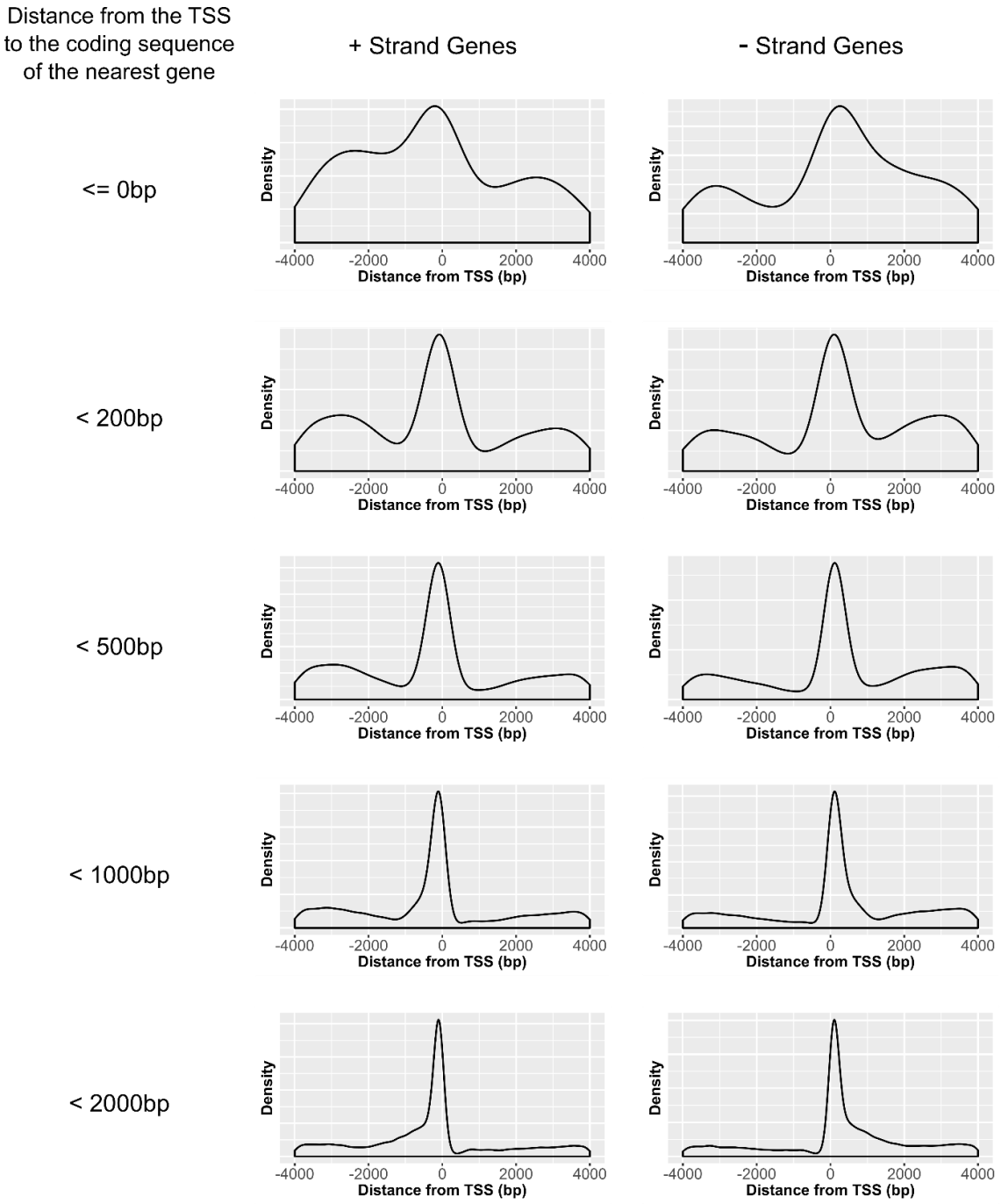


Figure 3.20: Distribution of DHS (DNaseI Hypersensitive sites) around the TSS for genes whose promoter region either overlaps or is less that 2000 base pairs from the next gene (Sullivan et al., 2014; NCBI (GSE53322). Number of genes for each category shown top left. Genes where the TSS

overlaps, or is directly adjacent to, the coding sequence of another gene. 0 = TSS. + Strand represents positive strand genes which read from 5' to 3', – Strand represents negative strand genes which read from 3' to 5'.

The DHS distributions for genes with a short distance between their TSS and the adjacent genes show a similar shaped distribution to those for all genes. Genes with shorter distances do show a shift of DHSs into the gene body, but the proposed analysis was to include this already. Additionally, shorter distances show a greater proportion of DHSs to the periphery. This is most likely caused by an increase in background; shorter distance plots have fewer input DHSs, and therefore there is a decrease in the signal to noise ratio.

3.3.16 Calculating the Background Rate of Motif Enrichment in the *A. thaliana* Genome

To identify which motifs are enriched in a gene group of interest, a background rate of enrichment must be calculated for normalisation. To ensure an accurate background enrichment the analysis was ran on 1000 sets of randomly chosen genes. To ensure any possible size effects of FIMO's algorithm were accounted for, the gene sets were size matched to the genes of interest. Therefore, for plastid-regulated genes 1000 random sets of 713 genes were analysed, and for the plastid-regulated C_4 orthologues, 1000 random sets of 8 genes. Once analysed, the rank of each motif for 713 genes averaged over 1000 runs was compared to 8 genes averaged over 1000 runs (Figure 3.21).

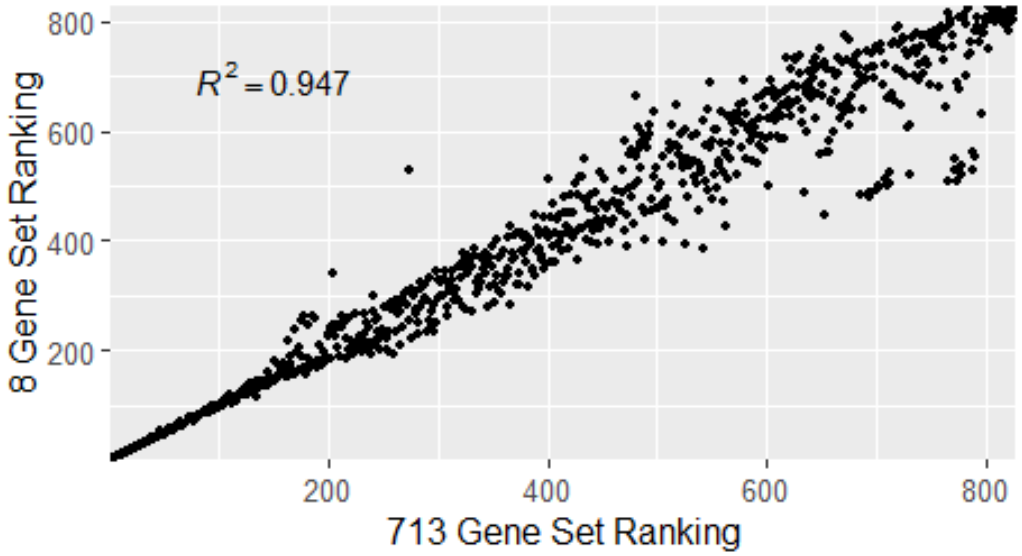


Figure 3.21: Motif enrichment is similar between gene sets of different sizes. Rank for average frequency of a motif in 1000 runs of 8 randomly selected genes plotted against a 1000 runs of 713 randomly selected genes.

There was a high correlation with an R^2 of 0.95, which means in any future time pressed motif enrichment studies one background set could be used for all analyses.

3.3.17 Motif Enrichment in the Gene Body and Most Accessible Regions of Chloroplast Responsive Genes

Based on the distribution of DNase hypersensitive sites in the genetic architecture, the gene body and 2000bp upstream of each plastid-regulated gene was analysed using FIMO (Table 3.7). Enrichment was normalised against an averaged background frequency of 1000 runs of same sized, randomly selected gene sets.

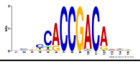
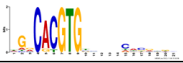
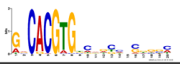
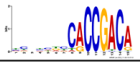
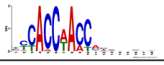
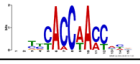
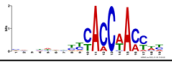
Motif	Total Motif Frequency	Fold change in Frequency compared to Average Background	Number of Genes in which the Motif Occurs	PWM
AP2EREBP_tnt.At2g44940_col_a_m1	21	3.56	18	
AP2EREBP_tnt.ERF1_colamp_a_d1	69	3.53	69	
bHLH_tnt.bHLH34_colamp_a_m1	57	2.87	39	
bHLH_tnt.bHLH104_col_b_m1	51	2.86	39	
AP2EREBP_tnt.AT1G01250_colamp_a_m1	23	2.84	21	
MYB_tnt.MYB13_colamp_a_m1	22	2.59	22	
bZIP_tnt.AREB3_colamp_a_m1	84	2.57	66	
MYB_tnt.MYB58_col_a_m1	28	2.48	26	
bZIP_tnt.bZIP42_colamp_a_m1	47	2.47	45	
C2H2_tnt.AT4G26030_col_a_m1	25	2.35	24	

Table 3.7: The ten most enriched motifs within the gene body and promoter (2000bps upstream of TSS) of chloroplast responsive genes. Enrichment measured as fold change in frequency compared to the average of 1000 random background gene sets of the same size. PWM = position weight matrix.

Motifs which bind the MYBS, bZIPs and bHLH transcription factor families are again enriched in the plastid-regulated gene set. However, expanding our analysis to the gene body and 2000bp upstream has highlighted the prevalence of AP2/EREBP binding motifs. AP2/EREBP transcription factors are thought to integrate metabolic, hormonal and environmental signals during the plant's response to stress (Dietz et al., 2010).

3.3.18 Motif Enrichment in Chloroplast Responsive and Non-Responsive C_4 Orthologues

The motif enrichment analysis was carried out on C_4 orthologues which were mutually affected by both inhibitors in illuminated seedlings (Figure 3.13), to identify which transcription factors may be involved in the regulation of C_4 photosynthesis (Table 3.8)

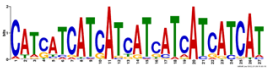
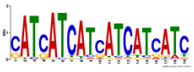
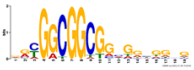
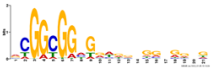
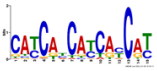
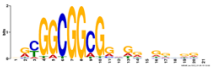
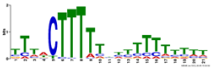
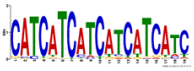
Motif	Total Motif Frequency	Fold change in Frequency compared to Average Background	Number of Genes in which the Motif Occurs	PWM
C2C2gata_tnt.GATA16_col_a_m1	41	4.62	5	
C2C2gata_tnt.GATA15_col_b_m1	38	4.01	6	
AP2EREBP_tnt.ERF104_col_a_m1	32	3.59	7	
HMG_tnt.3XHMGBOX1_col_a_m1	58	3.47	8	
C2C2gata_tnt.GATA1_col_a_m1	32	3.47	6	
AP2EREBP_tnt.ERF10_col_a_m1	32	3.41	7	
C2C2dof_tnt.OBP3_col_a_m1	85	3.35	8	
AP2EREBP_tnt.ERF105_colamp_a_m1	35	3.33	7	
C2C2gata_tnt.ZML2_col_a_m1	41	3.29	5	
BBRBPC_tnt.BPC6_col_a_m1	116	3.23	8	

Table 3.8: The ten most enriched motifs within the gene body and promoter (2000bps upstream of TSS) of chloroplast responsive C_4 orthologues. Enrichment measured as fold change in frequency compared to the average of 1000 random background gene sets of the same size. PWM = position weight matrix.

Two families of transcription factors had overrepresented binding motifs within the plastid-regulated C_4 orthologues; AP2/EREBP and C2C2GATA. GATA transcription factors are named for the 4bp sequence of their binding motif and have been found to be enriched in the promoters of both light and circadian regulated genes (Argüello-Astorga & Herrera-Estrella, 1998). GATA15 and GATA16, which were found to be the two most enriched binding motifs, work redundantly with GATA21 and GATA22 (also referred to as CNG and CNL) to control greening through the regulation of the tetrapyrrole biosynthesis pathway (Ranftl et al., 2016).

To elucidate if these enriched motifs are specific to the chloroplast to nucleus signalling pathway, and not simply to C_4 orthologues as a whole, the C_4 orthologues which were not responsive to chloroplast perturbation were also analysed (Table 3.9).

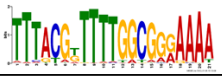
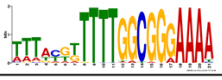
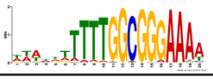
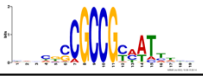
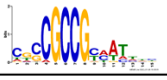
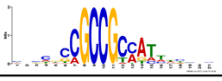
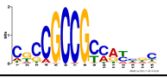
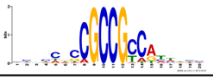
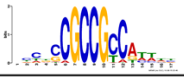
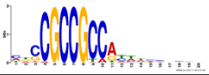
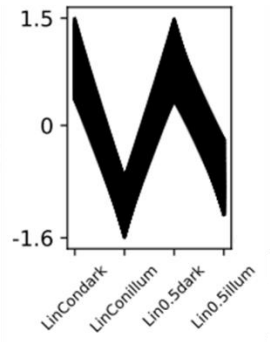
Motif	Total Motif Frequency	Fold change in Frequency compared to Average Background	Number of Genes in which the Motif Occurs	PWM
ARF_tnt.ARF16_col_b_m1	17	38.20	6	
ABI3VP1_tnt.REM16_col_a_m1	17	28.72	5	
E2FDP_tnt.DEL2_col_a_m1	11	26.57	1	
AP2EREBP_tnt.RRTF1_colamp_a_m1	17	18.09	8	
AP2EREBP_tnt.RRTF1_col_a_m1	18	17.48	6	
AP2EREBP_tnt.ABR1_colamp_a_m1	18	11.12	9	
AP2EREBP_tnt.ERF115_colamp_a_m1	24	10.02	11	
AP2EREBP_tnt.ERF11_colamp_a_m1	23	10.00	9	
AP2EREBP_tnt.ERF9_colamp_a_m1	13	9.77	6	
AP2EREBP_tnt.ERF10_colamp_a_m1	17	9.25	7	

Table 3.9: The ten most enriched motifs within the gene body and promoter (2000bps upstream of TSS) of chloroplast non-responsive C_4 orthologues. Enrichment measured as fold change in frequency compared to the average of 1000 random background gene sets of the same size. PWM = position weight matrix.

Whilst this analysis also found AP2/EREBP motifs to be highly enriched in C_4 orthologues who are not plastid-regulated, GATA binding motifs were not enriched at all with no single GATA binding motif occurring more than ten times in these genes. This supports the observation that certain C_4 orthologues are under chloroplast regulation and others are not.

3.3.19 Motif Enrichment in Chloroplast Responsive and Light Responsive Gene Expression Clusters

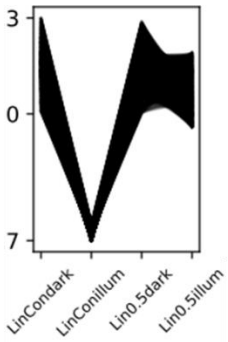
An analysis carried out using Clust identified 12 clusters which had a consistent transcriptomic response to chloroplast perturbation (Figure 3.14). Both inhibitor treatments led to two identifiable gene response clusters; a cluster where decreased transcript abundance was observed in illuminated samples compared to dark-grown samples regardless of inhibitor treatment, and a similar cluster where this response was seen in control plants, however treatment with a chloroplast inhibitor abolished this sharp decrease in expression. The motifs enriched between these two clusters were compared to see what transcription factors could be responsible for this change in expression between illuminated control and inhibitor treated samples.



C0

4510 Genes

Motif	Enrichment	PWM
AP2EREBP_tnt.ERF1_colamp_a_d1	65.89	
REM_tnt.REM19_col_a_m1	18.66	
HSF_tnt.HSFC1_col_a_m1	18.39	
ARF_tnt.ARF16_col_b_m1	17.62	
G2like_tnt.AT4G37180_colamp_a_m1	16.78	
bZIP_tnt.bZIP52_colamp_a_m1	16.22	
bZIP_tnt.bZIP52_col_v31_m1	15.96	
E2FDP_tnt.DEL2_col_a_m1	14.68	
E2FDP_tnt.DEL1_col_a_m1	14.60	
WRKY_tnt.WRKY18_col_a_m1	14.56	



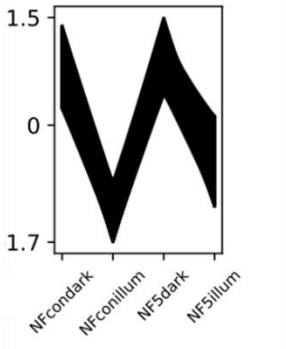
C3

1613 Genes

Motif	Enrichment	PWM
AP2EREBP_tnt.ERF1_colamp_a_d1	11.92	
ARF_tnt.ARF16_col_b_m1	7.05	
HSF_tnt.HSFC1_col_a_m1	6.73	
ABI3VP1_tnt.REM16_col_a_m1	6.52	
E2FDP_tnt.DEL1_colamp_a_m1	6.21	
E2FDP_tnt.DEL2_col_a_m1	6.13	
E2FDP_tnt.DEL2_colamp_a_m1	6.04	
E2FDP_tnt.DEL1_col_a_m1	6.01	
HB_tnt.ATHB21_colamp_a_m1	5.94	
REM_tnt.REM19_col_a_m1	5.59	

Figure 3.22: The ten most enriched motifs within the gene body and promoter (2000bps upstream of TSS) of genes within two gene expression clusters which have similar expression patterns in response to lincomycin treatment. Cluster number is arbitrary and for identification only (Figure 3.14). Enrichment measured as fold change in frequency compared to the average of 1000 random background gene sets of the same size. PWM = position weight matrix. Lin0.5 = 0.5mM lincomycin, dark = 6 days dark-grown, illum = 5 days dark-grown + 24 hours 150 µmol m^{-2} s^{-1} white light.

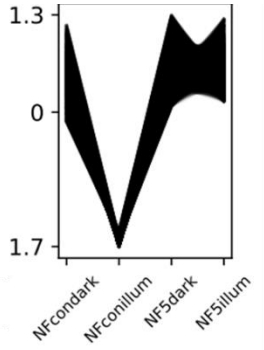
Looking at lincomycin treated samples and related controls, the motifs enriched within these two aforementioned clusters showed a large level of overlap, particularly within the ten most enriched motifs (Figure 3.22). However, a wider look at the one hundred most enriched motifs showed the two clusters only had 44 in common. The 66 motifs which were exclusive to the cluster where lincomycin abolished the decrease in transcript abundance to light observed in controls were bound mainly by C2C2, HSF and NAC transcription factors. Motifs binding the C2C2 transcription factor family were previously found to be enriched in chloroplast responsive C_4 orthologues (Table 3.8). The HSF and NAC transcription factor families are often involved with the perception of and response to stress (Guo et al., 2008; Jensen et al., 2010), however, the transcription factors predicted to bind these genes do not show any clear link to chloroplast function so whether this stress is directly related to chloroplast perturbation is unknown.



C5

3016 Genes

Motif	Enrichment	PWM
AP2EREBP_tnt.ERF1_colamp_a_d1	47.02	
MYB_tnt.MYB52_col_a_m1	13.59	
REM_tnt.REM19_col_a_m1	12.23	
C2H2_tnt.STOP1_col_a_m1	12.06	
GRF_tnt.GRF9_col_a_m1	11.71	
WRKY_tnt.WRKY18_col_a_m1	11.17	
ARF_tnt.ARF16_col_b_m1	10.77	
C2H2_tnt.STOP1_colamp_a_m1	10.12	
HSF_tnt.HSFC1_col_a_m1	9.87	
WRKY_tnt.WRKY30_col_a_m1	9.70	



C11

1539 Genes

Motif	Enrichment	PWM
WRKY_tnt.WRKY71_col_a_m1	34.57	
WRKY_tnt.WRKY29_col_a_m1	34.32	
WRKY_tnt.WRKY28_colamp_a_m1	32.98	
WRKY_tnt.WRKY15_col_b_m1	32.37	
WRKY_tnt.WRKY45_colamp_a_m1	31.05	
WRKY_tnt.WRKY21_col_m1	29.08	
WRKY_tnt.WRKY50_colamp_a_m1	22.70	
WRKY_tnt.WRKY46_colamp_a_m1	22.12	
WRKY_tnt.WRKY33_col_a_m1	18.97	
WRKY_tnt.WRKY3_col_a_m1	17.99	

Figure 3.23: The ten most enriched motifs within the gene body and promoter (2000bps upstream of TSS) of genes within two gene expression clusters which have similar expression patterns in response to norflurazon treatment. Cluster number is arbitrary and for identification only (Figure 3.14). Enrichment measured as fold change in frequency compared to the average of 1000 random background gene sets of the same size. PWM = position weight matrix. NF5 = Norflurazon 5µM, dark = 6 days dark-grown, illum = 5 days dark-grown + 24 hours 150 µmol m^{-2} s^{-1} white light.

A very different group of motifs were found to be enriched in those genes which did not show a large decrease in expression in illuminated samples when norflurazon is applied. Both in the top ten enriched (Figure 3.23) and a larger scale the top one hundred most enriched motifs were predicted to bind WRKY transcription factors. Several of these WRKYs have already been implicated in potential chloroplast to nucleus signalling pathways. WRKY33 for example is activated by chloroplast targeted proteins SIB1/2 (Sigma factor binding protein 1/2), allowing it to bind to the promoters of downstream nuclear genes (Phukan et al., 2016). However, it is unclear whether the difference between the cluster enriched motifs between the two inhibitors is due to their different methods of chloroplast perturbation or different unrelated side effects. Since neither Cluster 3 (Figure 3.22) or Cluster 11 (Figure 3.23) have any significant GO terms, a GO analysis cannot be used to further inform the analysis.

3.4 Discussion

3.4.1 Effectiveness of chloroplast perturbation

Communication from the chloroplast to the nucleus can be modified when chloroplast function is perturbed. Lincomycin and norflurazon, two chloroplast specific inhibitors were used to achieve this. These inhibitors both led to reduced greening and PSII activity, which mirrors results of other studies in the literature. Maximum quantum yield (F_v/F_m) measurements were taken in a study using norflurazon treated *A. thaliana* cultured cells to investigate the role of the chloroplast and ROS in programmed cell death (Doyle et al., 2010). F_v/F_m was found to be significantly lower in illuminated norflurazon treated samples compared to illuminated controls, but not significantly different to F_v/F_m in untreated dark-grown controls. Lincomycin has also been shown to decrease maximum quantum yield, with *A. thaliana* leaves infiltrated with 1mM lincomycin showing a 60% decrease in maximum quantum yield after 4 hours (Tarantino et al., 1999). Once it was established that lincomycin and norflurazon had the expected result on chloroplast formation and PSII activity, RNA-seq was performed to investigate its effect on gene expression.

3.4.2 Genes Responsive to Chloroplast Perturbation

Transcript abundance was significantly different between dark-grown and 24-hour light exposed seedlings in 26% of the detected transcripts, meaning over a quarter of the detectable transcriptome is light responsive. This is higher than the 20% predicted by Jiao et al (2005) based on microarray data comparing 6-day seedlings which were grown either entirely in the dark or entirely in the light. They also predicted that 20% of the rice transcriptome was light responsive, with qualitatively similar expression profiles found in both species. Both the qualitative and quantitative light response in *A. thaliana* and rice during deetiolation can be compared once a similar RNA-seq experiment is carried out in rice (Chapter 4). Of the 21319 genes whose transcripts were detected in the samples, 3607 genes showed significantly different transcript abundance when illuminated samples were treated with norflurazon as compared to untreated controls. However, only 1010 genes showed a significant response to lincomycin treatment. This fits with phenotypic data reported in the literature, where norflurazon application has been shown to cause effects other than bleaching which do not appear to directly link to its chloroplast perturbing properties. Changes caused by the inhibitor in dark-grown plants include decreasing total lipid quantity and increasing dry weight weight (Di Baccio et al., 2002; Magnucka et al., 2007).

Whilst norflurazon application had a larger impact on gene expression than lincomycin, there was overlap in the genes affected by each inhibitor with 70% of the genes affected by lincomycin treatment also affected by norflurazon treatment. It is therefore unsurprising that there was large overlap in the gene ontology annotations of the genes affected by each. The most enriched gene ontology terms include multiple photosynthesis associated terms for both inhibitors. Another shared term is the ‘Generation of precursor metabolites and energy’ which includes photosynthesis, oxidation and NADPH regeneration (Figure 3.8). Reactive oxygen species (ROS) have been implicated as a chloroplast retrograde signal in several potential pathways, including the redox potential of the chloroplast transmitted through the malate valve (Heyno et al., 2014), and the detection of singlet oxygen by executor proteins and β-cyclocitral (Ramel et al., 2012; Uberegui et al., 2015). The malate valve operates when the NADPH pool is over-reduced. NADP-MDH oxidises the pool using reduction power to convert OAA to malate which is then exported to the cytosol acting

as a signal. Genes associated with 'Single-organism metabolic processes' was also overrepresented in chloroplast responsive genes, this can include NADPH metabolism and oxidation-reduction processes. It can additionally include intermediary metabolites such as Mg-proto IX, an intermediary in the tetrapyrrole biosynthesis pathway directly regulated by GUN4 and GUN5 (Martín et al., 2016).

The gene *LHCB2.1* is known to have decreased transcript abundance in response to chloroplast perturbation in wildtype, and consequently a seedling lacking this response was the diagnostic for a *GUN* mutant. This classic response of *LHCB2.1* was observed alongside decreased transcript abundance of *CA1* and *GUN4* when seedlings were treated with norflurazon (Figure 3.10). These results mirror rt-qPCR expression data of 7-day old seedlings grown in 5µM norflurazon, with a minor difference in growth conditions as these seedlings were grown in 4 days dark followed by 3 days light (Page et al., 2017). Additionally, microarray data of norflurazon treated and control *A. thaliana* seedlings found *CA1* to be the most responsive alongside two light harvesting complex genes in the ten most upregulated in controls (Moulin et al, 2008). Whilst this does not exactly match the RNA-seq results, it is promising to see overlap and exact matches cannot be expected as the number of genes detected with the RNA-seq is much greater than that of the microarray analysis. Overall, the gene expression analysis shows consistency with previous studies as well as an overlap in the genes affected by each inhibitor. This suggests that they are comparable and focussing on genes affected by both increases the robustness of the findings.

3.4.3 The extent to which C_4 orthologues are subject to chloroplast retrograde signalling in *A. thaliana*

Of the fifteen C_4 orthologues analysed, eight were under some degree of plastid control; *AtCA1*, *AtCA2*, *AtPPC2*, *AtNAD-ME1*, *AtPPA6*, *AtRP1*, *AtASP1* and *AtDIC1*. Burgess et al (2016) had previously investigated this question by using rt-qPCR to detect differential transcript abundance in norflurazon and lincomycin treated seedlings. They found seven C_4 orthologues under chloroplast control, five of which were identified in this work; *AtCA1*, *AtPPC2*, *AtNAD-ME1*, *AtPPA6*, *AtRP1*, *AtNAD-ME2* and *AtTPT*. The increase in the number of responsive C_4 orthologues found further supports the hypothesis that parallel evolution repeatedly recruits genes into C_4 photosynthesis because these genes are already controlled by the same regulatory network as C_3 photosynthetic genes. Furthermore, regulation of these eight genes can have subsequent effects on other C_4 orthologues. AtNAD-ME1 forms a heterodimer in vivo with AtNAD-ME2 (Tronconi et al., 2008), therefore the expression of *AtNAD-ME1* has an impact on the activity of both the AtNAD-ME1 and AtNAD-ME2 protein.

The results of this work therefore support the hypothesis that evolution has re-enforced existing regulatory networks that operate in the C_3 state to control expression of C_4 genes. To further test this hypothesis, the next chapter reports analysis from an evolutionarily distant C_3 species, *Oryza sativa*, allowing comparison with a representative eudicotyledon and representative monocotyledon. If the findings are replicated, and C_4 orthologues are found to be under a large degree of chloroplast control in rice, this would support that the chloroplast to nucleus regulatory networks acting on these genes are ancestral. Whilst the findings regarding transcript abundance of C_4 orthologues during chloroplast perturbation give insight into the overall regulatory networks controlling these genes, the specific mechanisms involved in these networks remain elusive.

3.4.4 Regulatory elements enriched in chloroplast responsive genes

The list of chloroplast responsive genes generated in this analysis enabled a search for putative transcription factors likely involved, providing insight into how the change in gene regulation has occurred. Six families of transcription factors had enriched motif binding sites with chloroplast responsive genes: bZIP, bHLH, TCP, MYB, MYB-related, and AP2/EREBP. An additional family, the C2C2GATA transcription factors, had enriched binding sites in plastid regulated C_4 orthologues specifically. The prevalence of bZIPs could be expected after the 6-mer enrichment analysis showed the G-box (CACGTG) to be the most enriched 6bp sequence within the chloroplast regulated genes. Both the G-box and ABA-responsive elements found to be enriched in the 6-mer enrichment analysis contain the ACGT sequence. This sequence was found to be enriched in the promoters of 329 genes which were upregulated in both *gun1* and *gun5* seedlings, appearing 944 times (Koussevitzky et al., 2007).

The C2C2GATA transcription factors, whose binding motifs are enriched in chloroplast responsive C_4 orthologues, are involved in light and nitrate-dependent regulation. GATA15 and GATA16, which were found to be the two most enriched binding motifs, are thought to work redundantly with GATA21 and GATA22 (also referred to as GNC and GNL) to control greening (Rantfl et al., 2016). GATA21 and GATA22 have been shown to control greening through regulation of the tetrapyrrole biosynthesis pathway, and *gata21 gata22* mutants show decreased expression of chloroplast regulated genes such as *HEMA1* and *GUN4* (Hudson et al., 2011). GATA15 and GATA16 may work through a similar mechanism, regulating upstream of the tetrapyrrole retrograde signalling pathway. Of the four subfamilies which the C2C2GATA transcription factors are classified, three are shared between *A. thaliana* and rice which are theorised to have evolved before their most recent common ancestor (Reyes et al., 2004). The ancestral nature of this regulation means comparison between the enriched motifs in chloroplast responsive genes in *A. thaliana* and rice will provide further insight into which candidates warrant further attention, for both the C_4 orthologues and chloroplast regulated genes as a whole. Future work will focus on characterising the most promising transcription factor candidates.

4 CHLOROPLAST AND LIGHT DEPENDENT REGULATION OF C_4 ORTHOLOGUES IN *ORYZA SATIVA*

4.1 Introduction

Whilst most recent research into chloroplast retrograde signalling has been carried out on *A. thaliana*, much of the initial experimental work investigated the system in monocotyledons. Genome uncoupled (*gun*) mutants, where Photosynthesis Associated Nuclear Gene (PhANG) expression is unchanged after chloroplast inhibition, were so named after analysis of *A. thaliana* mutants (Susek et al. 1993). However, this discovery built on previous work that established PhANG expression was disrupted in other species whose chloroplasts were perturbed. For example, Harpster et al (1984) identified maize mutants that did not accumulate pigment by screening for bleached leaves. Moreover, compared with controls these mutants accumulated fewer transcripts encoding light-harvesting chlorophyll a/b binding protein (*LHCB*) and other photosynthesis associated genes such as Phosphoglycerate kinase (*PGK*) and Glyceraldehyde-3--phosphate Dehydrogenase (*GAPDH*). Similar results were reported in barley (Hess et al., 1994) where mutants lacking chloroplast ribosomes did not accumulate either light-harvesting chlorophyll a/b binding protein or nitrate reductase.

Work in barley expanded from chloroplast development mutants (Bradbeer et al., 1979) to the application of chloroplast inhibitors. Batschauer et al (1986) found that barley plants treated with norflurazon did not accumulate transcripts for the light harvesting chlorophyll a/b protein. This work also identified mutants where *LHCB* accumulated to wildtype levels despite chloroplast perturbation, ie. a *gun* mutant in everything but name. These chlorophyll-deficient *xantha* mutants blocked the tetrapyrrole biosynthesis pathway after protoporphyrin IX or Mg-protoporphyrin, similar to *gun4* and *gun5* in *A. thaliana.* Norflurazon's mode of action is to block carotenoid synthesis. Burgess and Taylor (1988) then grew carotenoid deficient mutants in low light levels to limit overall chloroplast damage and found *LHCB* levels were unaffected, indicating carotenoids do not specifically play a role in plastid to nucleus signalling and clarifying that it is norflurazon's perturbation of chloroplast function that affects PhANG expression. *LHCB* expression was found to be inversely proportional to thylakoid damage in maize (Rocca et al, 2000) and in barley (La Rocca et al, 2000).

More recent studies have looked at the effect of norflurazon on rice seedlings. Park and Jung (2017) showed 3 week old rice seedlings after 40 hours of norflurazon treatment had decreased maximum quantum yield (F_v/F_m), a proxy for PSII activity, and the abundance of some LHCA (LHCA1, LHCA2, LHCA3 and LHCA4) and LHCB (LHCB1, LHCB2 and LHCB6) proteins were decreased greatly. Additionally, images taken with transmission electron microscopy showed norflurazon treated plants had fewer and smaller plastids which had arrested development during thylakoid synthesis. Norflurazon treated rice seedlings were consequently found to show increased peroxidase activity (Park & Jung, 2018). This could potentially be involved in photoprotective pathways in the

chloroplast. Other photoprotective compounds, β-cyclocitral and Executor proteins, have been proposed as possible retrograde signals in *A. thaliana* (Ramel et al., 2012; Uberegui et al., 2015).

The work presented in this chapter aimed to continue the investigation of the role of plastid-to-nucleus signalling in regulating genes of the C_4 cycle in the ancestral C_3 state. Chloroplast derived signals have been shown to regulate C_4 photosynthesis genes in several species, including eudicotyledon *Gynandropsis gynandra* (Burgess et al., 2016) and monocotyledon *Zea Mays* (Tamada et al., 2003). C_4 orthologues within *A. thaliana* have been shown to respond to chloroplast perturbation. If this were also true in a distantly related species it would suggest an ancestral mechanism. To investigate this, an RNA-seq experiment was conducted on C_3 *Oryza sativa* after chloroplast function has been perturbed. The expectation was that analysis would confirm or refute if the previously observed response of C_4 orthologues in *A. thaliana* to chloroplast perturbation also take place in a C_3 monocotyledon. If this is the case, it would further support the hypothesis that retrograde control may help explain the repeated recruitment of genes in the ~60 independent origins of C_4 photosynthesis.

Rice and *A. thaliana* show similarities in their photosynthetic response to the chloroplast inhibitor norflurazon, with reduced chlorophyll fluorescence (F_v/F_m) and reduced thylakoid development (Park et al., 2017). However, only small subsets of the global transcriptomic response have been analysed. The data reported here allow a genome wide comparison of chloroplast responsive genes in *A. thaliana* and *O. sativa*. Whilst chloroplast control is hypothesised to be ancestral, many mechanisms regulating photosynthesis have diverged between these two species. They are thought to be separated by five genome duplications and a genome triplication (Lee et al., 2013), which means most *A. thaliana* transcription factors do not have one-to-one orthologues in rice (Wang et al., 2017). This work will therefore not only confirm the magnitude of the chloroplast response, but the function of genes under chloroplast control within a C_3 monocotyledon.

4.2 Objectives

Genes involved in plastid-to-nucleus signalling will be identified in C_3 monocotyledon *O. sativa* through RNA-seq after chloroplast function has been perturbed.

Hypotheses

- Photosynthesis related genes will be disproportionally overrepresented in genes affected by inhibitor treatment.
- Multiple C_4 orthologues in rice will be regulated by chloroplast to nucleus signalling, providing further support for the hypothesis that retrograde control of C_4 photosynthesis genes is ancestral.

Candidate transcription factors involved in the regulation of plastid-regulated genes will be identified by analysing enriched binding motifs.

Hypotheses

- There will be significant enrichment of binding motifs in genes responsive to chloroplast perturbation compared to background as these genes share a regulatory pathway.
- There will be common motifs enriched in *O. sativa* and *A. thaliana* as these regulatory pathways are ancestral.

4.3 Results

4.3.1 Impact of lincomycin and norflurazon on chlorophyll accumulation and chlorophyll fluorescence

Lincomycin and norflurazon inhibited greening of rice leaves, and this was particularly evident at concentrations of 10mM and 100µM respectively (Figure 4.1). This bleaching is indicative of chloroplast development being impeded.

Figure 4.1: Chlorophyll accumulation is impaired in rice seedlings exposed to either lincomycin (10mM) or norflurazon (100µM). Representative images are shown for 6-day old rice (5 days dark followed by 24 hours white light) grown in lincomycin or norflurazon. Scale bar represents 1cm.

To quantify the impact of reduced greening on photosynthesis, maximum quantum yield of PSII (F_v/F_m) was measured (Maxwell & Johnson, 2000). Seedlings of *O. sativa* maintained in the dark showed low values of F_v/F_m, consistent with the very low amount of chlorophyll expected in etiolated plants (Figure 4.2A). The addition of either lincomycin or norflurazon had no impact on photosynthetic performance of these seedlings. Compared with dark-grown controls, plants exposed to 24 hours illumination showed statistically significant increases in F_v/F_m. Treatment with norflurazon or lincomycin caused an approximate three-fold decrease in F_v/F_m in illuminated seedlings compared with controls. Measurements were consistent both between seedlings, and within individual seedlings (Figure 4.2B). Thus, although much higher concentrations of each inhibitor were required, lincomycin and norflurazon had similar effects on chlorophyll accumulation in rice and *A. thaliana* (see Chapter 3). In both cases, chloroplasts showed decreased ability to harvest light. To compare the basis of these responses between the two species, an RNA-seq experiment was undertaken.

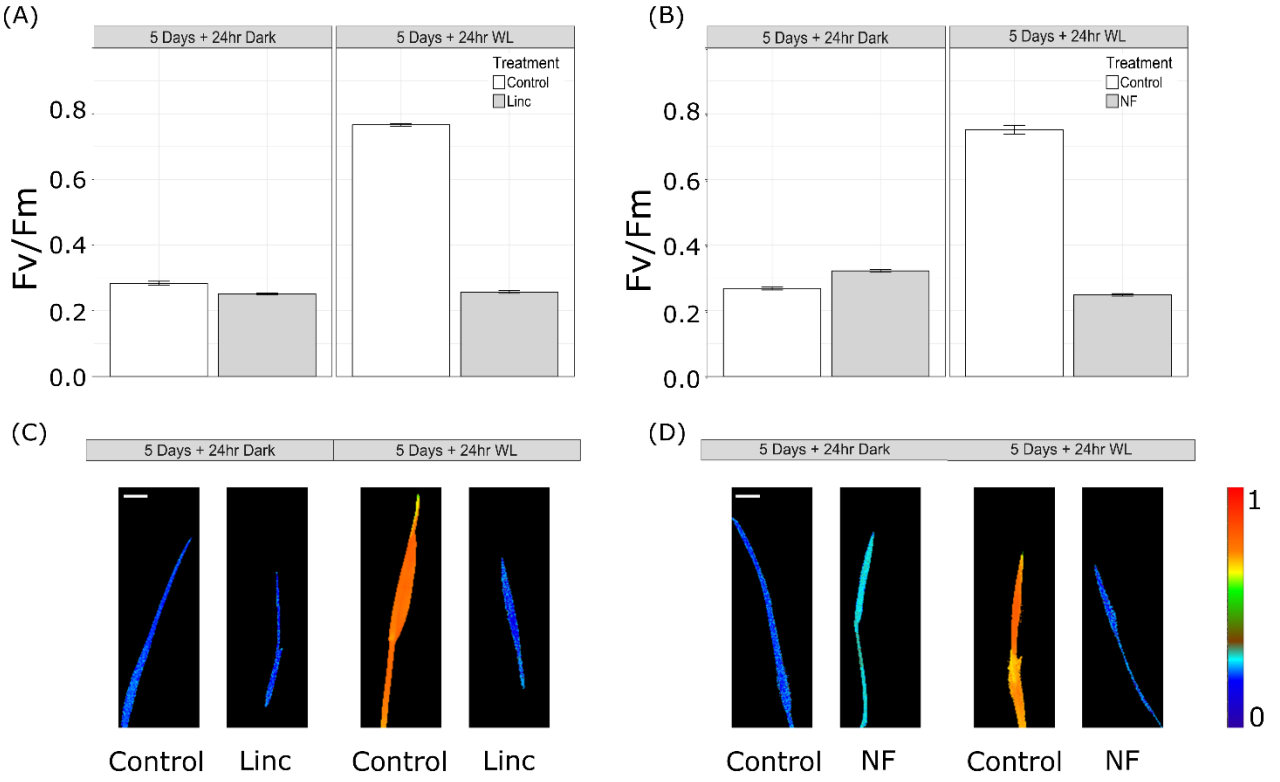


Figure 4.2: Effect of lincomycin and norflurazon on chlorophyll fluorescence in rice. (A&B) Bar charts quantifying the impact of transferring seedlings from dark to light, and treatment with either lincomycin (A) or norflurazon (B) on F_v/F_m (n=10 seedlings). Data are presented as means with one standard error. Consistent with the greening of leaves in response to light, illumination led to an increase in F_v/F_m. In both cases, application of lincomycin and norflurazon reduced F_v/F_m to levels similar to those detected in dark grown seedlings. Statistical significance was undertaken using a student's t-test (illuminated control mean F_v/F_m = 0.76, dark-grown control mean F_v/F_m = 0.28): t(38) = -58.32, $p < 0.001$, (illuminated control mean F_v/F_m = 0.76, illuminated inhibitor-treated mean F_v/F_m = 0.25): t(38) = -65.14, $p < 0.001$. (C&D) Representative images captured by the Fluor*Imager* chlorophyll fluorescence imaging system showing a heatmap of F_v/F_m. The false colour scale represents F_v/F_m values of 0 (blue) to 1 (red). Scale bars represent 1cm. Linc = 10mM lincomycin, NF =100µM norflurazon, WL = 150 µmol m^{-2} s^{-1} white light.

4.3.2 Extraction and analysis of RNA prior to sequencing

To confirm that RNA extracted after each treatment was of sufficient quality for deep sequencing, its integrity was analysed using a microfluidic electrophoresis system. This generated electropherograms and an RNA Integrity Number (RIN), which gives an indication of how much degradation the RNA sample has undergone. The calculation of an RIN takes into account the ratio between the 18S and 25S peaks that derive from the small and large subunit of nuclear ribosomes respectively, as well as the overall distribution of the electropherogram trace. This trace also shows a 16S-related peak, which derives from the small subunit of the chloroplast ribosome (Delius & Koller, 1980). By comparing the height of this peak compared to the 18S and 25S peaks, the ribosome composition and subsequently the potential for chloroplast translation can be inferred (Figure 4.3).

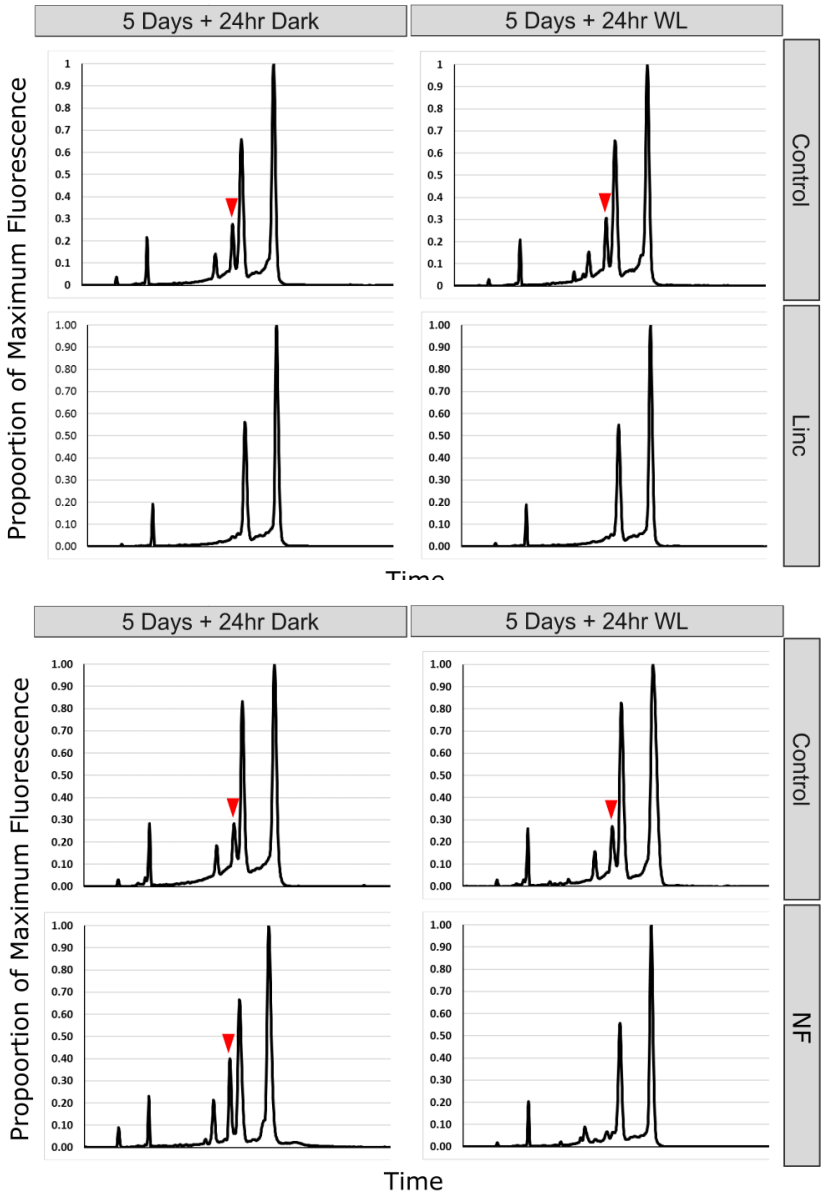


Figure 4.3: Electropherograms of *O. sativa* RNA obtained on the Agilent Bioanalyzer 2100. Red arrow indicates the 16S peak, which is indicative of the presence of chloroplast ribosomes. As maximum fluorescence varies according to sample concentration, to simplify comparison, all samples were normalised to their maximum value to simplify visual comparisons. Linc = 10mM lincomycin, NF =100µM norflurazon, WL = 150 µmol m−2 s−1 white light.

The lincomycin treatment shows no 16S peak in either dark-grown or illuminated samples. A 16S peak was present in dark-grown rice treated with norflurazon, but this was lost after exposure to light. Norflurazon treated dark-grown samples appeared to have higher proportions of chloroplast rRNA than controls. The exact reason for this is not known, but changes caused by the inhibitor in dark-grown plants have been documented including decreasing total lipid quantity and increasing dry weight (Di Baccio et al., 2002; Magnucka et al., 2007). By including both lincomycin and norflurazon in the study, the aim was to take into account unintended effects of individual inhibitors. Combined with the data above (Figure 4.2), these results indicate that both inhibitors have perturbed chloroplast development, decreased the number of PSII reaction centres located in the thylakoid membrane and eliminated the accumulation of chloroplast rRNA in light-exposed seedlings.

The RIN for each sample was assessed, and samples with an RIN over 7 were considered intact and suitable for next generation sequencing. Thus, all libraries were prepared from samples with an RIN of 7 or greater.

4.3.3 Quality control of transcriptomic libraries

Libraries were sequenced, and samples underwent quality control (QC). FastQC is a tool that allows intuitive visual inspection of sequences and provides a pass or fail grading to QC components. Analysis of samples showed they were good quality, both at sequence and base level. There were no unassigned bases (Ns) present in the data, and no adapter contamination (Figure 4.4).

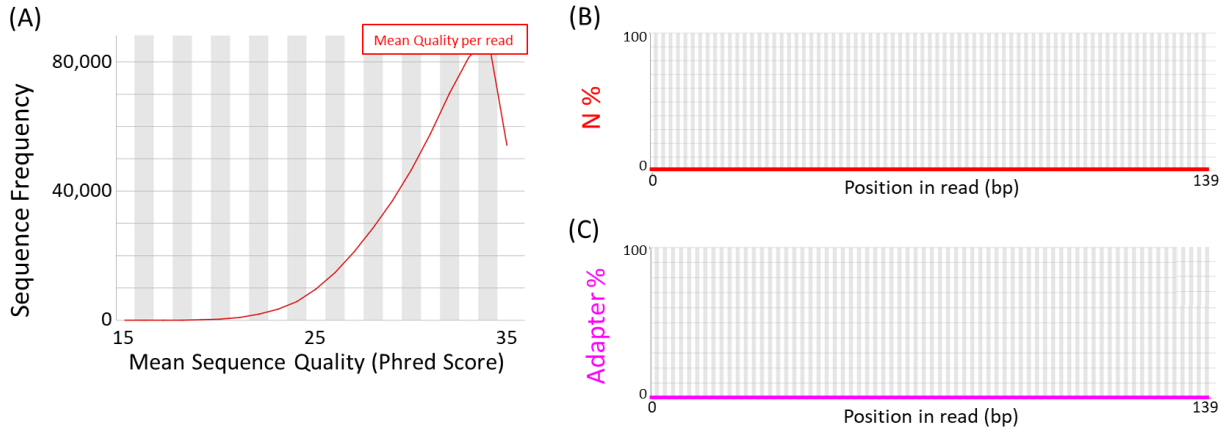


Figure 4.4: FastQC generated plots illustrating sequences were high quality and contained no unassigned bases or adapter contamination. (A) Representative plot showing frequency against mean sequence quality, with most sequences having a mean phred score of 33. (B) Representative plot showing Ns (unassigned bases) were not present at any position in the read. (C) Representative plot showing no adapters were present at any position in the read.

As was the case for the libraries prepared for *A. thaliana*, two FastQC categories failed for all samples; 'Per Sequence GC Content' and 'Kmer Content'. This is expected due to the mechanism of the library preparation kit, and does not affect the accuracy of the transcriptomic analysis as explained in the previous chapter (Figure 3.5). Thus, the QC analysis indicated that all libraries had generated good quality data, and so further analysis was initiated. After trimming and quality filtering, 149 million reads were obtained in total. A mean of 65.77% of reads mapped uniquely to the *Oryza_sativa*.IRGSP-1.0.37 reference genome.

4.3.4 Hierarchical clustering and principal component analysis to determine sample similarity

Hierarchical clustering was performed to confirm biological replicates were behaving in similar manner and to identify which treatment caused the greatest alteration to transcript abundance. Hierarchical clustering was performed by calculating the distance between samples, providing a quantitative measure of sample similarity. Data underwent variance stabilising transformation which

normalises with respect to library size. This also involves logarithmic transformation to ensure that a small number of highly expressed genes do not skew the results.

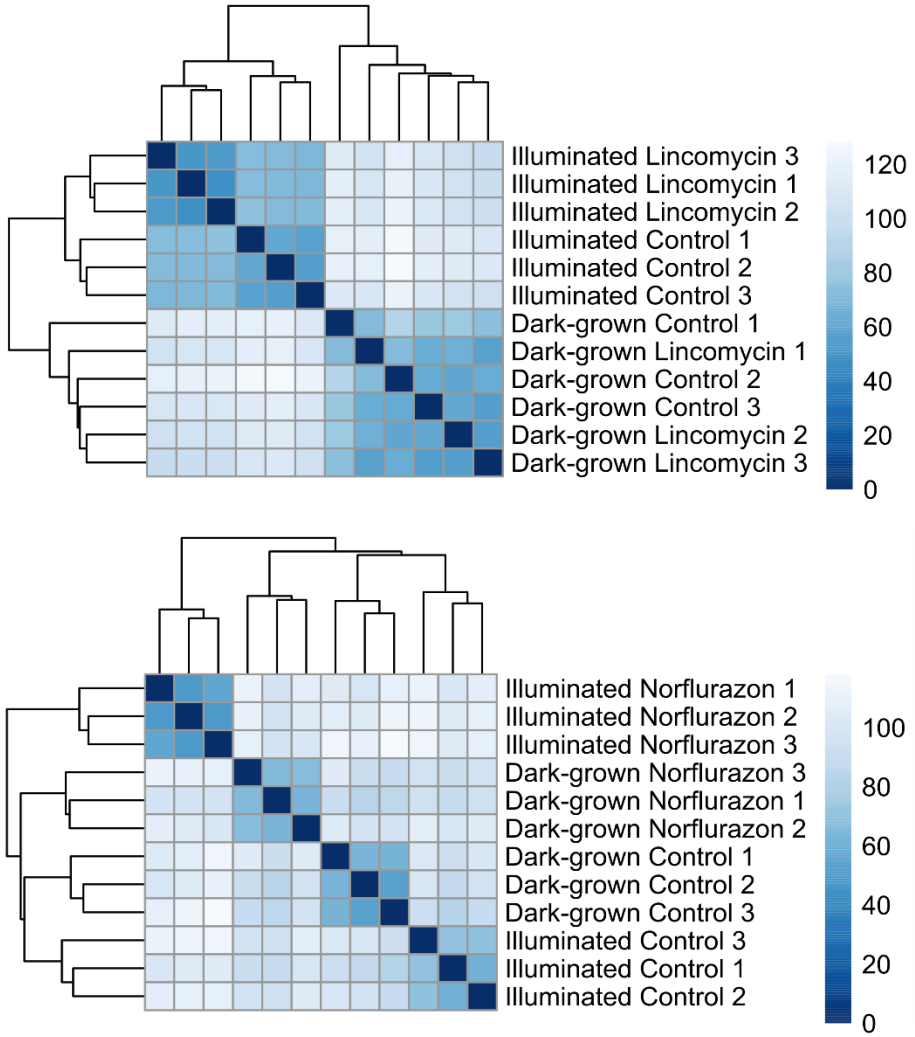


Figure 4.5: Heatmaps showing hierarchical clustering of sampling distance, indicating sample similarity and showing replicates cluster together. Images generated by pheatmap, with a false colour scale, units are variance stabilising transformed normalised counts.

Hierarchical clustering showed replicates generally clustered together, implying the majority of variance was explained by sample treatment rather than biological noise (Figure 4.5). An exception was the clustering of one dark-grown lincomycin treated replicate and one dark-grown control replicate. As dark-grown seedlings contain etioplasts rather than fully functioning chloroplasts, this is not completely unexpected as it indicates that the treatment in the dark has a less clear effect on transcript abundance.

Read counts estimated from the sample alignments were then subjected to principal components analysis (PCA), which is a class free method for explaining variance between samples. Variance is attributed to unknown principal components, and plotting these components can give an indication of sample similarity or disparity. This approach showed that over 80% of variance between samples could be attributed to two principal components (Figure 4.6).

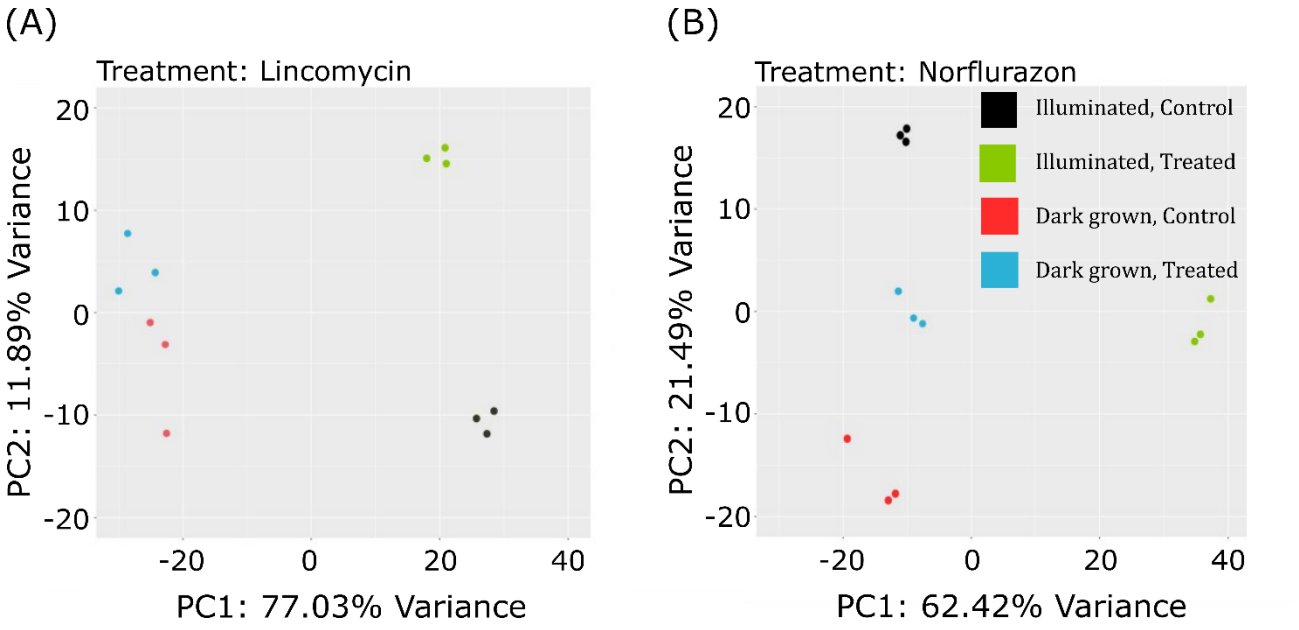


Figure 4.6: Principal components analysis of transcript abundance obtained from dark-grown and illuminated inhibitor treated samples and corresponding controls. (A) Analysis of lincomycin treated samples and corresponding controls (B) Analysis of norflurazon treated samples and corresponding controls. PC1 = principal component 1, PC2 = Principal component 2. Lincomycin = 10mM lincomycin, Norflurazon = 100µM norflurazon, Dark grown = 7 days dark grown. Illuminated = 6 day dark grown + 1 day 150 µmol m^{-2} s^{-1} white light.

The three replicates for each treatment generally clustered together, implying the majority of variance is explained by sample treatment rather than biological noise. However, it was notable that the dark-grown samples clustered less well for both lincomycin and norflurazon. However, most variation was associated with the dark and light treatments, visualised as a shift in PC1 (Figure 4.6A). Norflurazon exhibited a slightly different trend, whilst illumination of samples caused large variance it was not attributed to a single principal component. Dark-grown plants treated with norflurazon were more similar to control treatments than to illuminated norflurazon treated samples (Figure 4.6B). This could be caused by norflurazon's light-dependent mode of action, consistent with previous chlorophyll fluorescence and RNA electropherogram results. The differing PCA analysis between inhibitors in rice contrasts to *A. thaliana* where similar profiles were observed for lincomycin and norflurazon treated seedlings. This could cast doubt on whether a focus on genes which are affected by both inhibitors is robust, as it suggests a different mode of action between the two inhibitors. However, a comparison of the percentage of genes which are affected by both treatments and the GO overlap will confirm the similarity in the transcriptomic response (Table 4.1, Figure 4.7).

4.3.5 Differential Transcript Abundance between Treatments

Sample read counts consequently underwent a pairwise differential expression analysis (Table 4.1). PCA analysis is consistent with the results, with increased variance between treatments relating to an increased number of genes with significantly different transcript abundance (adjusted p-value < 0.05). Comparisons between chloroplast inhibited dark-grown and illuminated samples yielded more

differentially expressed genes than between dark-grown and illuminated control samples. This contrasts with the response in *A. thaliana*, where treatment with either lincomycin or norflurazon dampened the overall light response. Why more genes became light responsive in samples without a competent chloroplast present is unclear.

(A)

	Number of Genes with Significant DTA (Total)			
	Lincomycin	**Norflurazon**	**Shared**	**% Genes Shared DTA**
Control: **Darkgrown vs Illuminated**	3765	2612	1805	69%
Inhibited: **Dark-grown vs Illuminated**	4858	4204	1958	47%
Dark-grown: **Control vs Inhibited**	21	3336	14	67%
Illuminated: **Control vs Inhibited**	1370	4362	935	68%

(B)

	Number of Genes Significantly Upregulated			
	Lincomycin	**Norflurazon**	**Shared**	**% Genes Shared Upregulated**
Control: **Darkgrown vs Illuminated**	2167	1656	1270	77%
Inhibited: **Dark-grown vs Illuminated**	2695	2149	652	30%
Dark-grown: **Control vs Inhibited**	11	1333	6	55%
Illuminated: **Control vs Inhibited**	602	1985	310	51%

(C)

	Number of Genes Significantly Downregulated			
	Lincomycin	**Norflurazon**	**Shared**	**% Genes Shared Downregulated**
Control: **Darkgrown vs Illuminated**	1598	956	532	56%
Inhibited: **Dark-grown vs Illuminated**	2163	2055	590	29%
Dark-grown: **Control vs Inhibited**	10	2003	0	0%
Illuminated: **Control vs Inhibited**	768	2377	588	77%

Table 4.1: Number of genes showing statistically significant differential transcript abundance between treatments, split into (A) all genes with significant differential transcript abundance, (B) upregulated genes only and (C) downregulated genes only. Significance is defined as an adjusted p-value <0.05. The percentage of genes showing statistically significant differential transcript abundance in response to both inhibitors as a percentage of the lincomycin related list. DTA = differential transcript abundance.

After illumination, treatment with norflurazon disrupted transcript abundance for the expression of three times as many genes as lincomycin. However, there was overlap in the genes responding to the two inhibitors with 68% of the genes responsive to lincomycin treatment also responding to norflurazon. Of these 935 transcripts that responded to both inhibitors, 588 genes were downregulated in response to chloroplast inhibition. Overall, the number of significantly differentially expressed genes displaying alterations in transcript abundance between treatments was consistent with the PCA analysis, with increased variance relating to an increased number of differentially expressed genes.

4.3.6 Genes most responsive to chloroplast perturbation

Lincomycin and norflurazon are known to affect the expression of photosynthesis associated nuclear genes (PhANGs) (Ruckle et al., 2007). In contrast, in mutants of *GUN* genes which are involved in biogenic signalling chloroplast inhibitors no longer cause such a large change. It is therefore unsurprising that PhANGs feature on the list of most responsive genes when comparing controls to both lincomycin and norflurazon treated seedlings (most responsive in this context defined as the greatest log2 fold change in transcript abundance). Four of the top ten genes most responsive to lincomycin encoded Ribulose-1,5-bisphosphate carboxylase subunits which were downregulated by lincomycin treatment, and therefore upregulated by a competent chloroplast associated with chloroplast to nucleus signalling (Table 4.2). The most downregulated gene after lincomycin treatment was *prx29* encoding a peroxidase. This could potentially be involved in photoprotective pathways in the chloroplast alongside β-cyclocitral and Executor proteins. In a study investigating the effect of low level gamma radiation on the rice leaf transcriptome, expression of *Prx29* has been previously reported to correlate with expression of PhANGs (Hayashi et al., 2014). The genes most upregulated after lincomycin application, and therefore downregulated by retrograde signalling, are largely uncharacterised. Those with functional annotations seem to be related to defence and stress responses, such as *OsDR10* which is known to act in the defence response against pathogens.

The ten most downregulated genes after norflurazon treatment were also predominantly photosynthesis related (Table 4.3). *Prx29* features also, reinforcing the hypothesis that it could be involved in chloroplast specific photoprotection.

Gene ID	Oryzabase Gene Symbol(s)	RAP-DB Description	Log2fold Change	Down or Up Regulated By Lincomycin Treatment	Adjusted P-value
OS02G0240300	prx29	Similar to Class III peroxidase GvPx2b	-5.65	Downregulated	3.66E-46
OS01G0720500	Lhcb1.1 OsLhcb1.1, lhcB	Similar to Type I chlorophyll a/b-binding protein b	-5.14		5.00E-166
OS01G0791033	No Annotation Available	Similar to ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit.	-4.84		3.16E-47
OS01G0600900	cab2R, CAB-2, OsLhcp, Lhcb1	Chlorophyll a-b binding protein 2, chloroplast precursor (LHCII type I CAB-2)	-4.67		1.20E-203
OS12G0292400	RbcS, RbcS4, OsRBCS4, OsRbcS4	Similar to Petunia ribulose 1,5-bisphosphate carboxylase small subunit mRNA	-4.22		2.49E-27
OS12G0207600	No Annotation Available	Similar to ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit.	-4.03		2.42E-15
OS10G0356000	OsrbcL1, rbcL1	Similar to ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit.	-4.01		2.03E-33
OS01G0639900	OsCA, CA, OsbetaCA1, betaCA1	beta-carbonic anhydrase, Carbon assimilation, CO_2-mediated stomatal pore response	-3.72		8.35E-27
OS01G0238500	No Annotation Available	Similar to Branched-chain-amino-acid aminotransferase-like protein 3, chloroplast precursor.	-3.66		1.78E-61
OS02G0103800	OsLFNR1, LFNR1, LFNR2, Os-LFNR2, OsLFNR2, FNR	Leaf-type ferredoxin-NADP+-oxidoreductase, Regulation of electron partitioning in the chloroplast	-3.52		2.12E-21
OS08G0155900	OsDR10	Similar to Pathogen-induced defense-responsive protein	5.33	Upregulated	5.72E-71
OS05G0531100	Dfu584	Protein of unknown function DUF584 family protein	4.98		1.04E-103
OS01G0392600	No Annotation Available	Conserved hypothetical protein.	4.80		2.38E-27
OS08G0156100	No Annotation Available	Conserved hypothetical protein.	4.75		1.50E-35
OS08G0156000	No Annotation Available	Conserved hypothetical protein.	4.62		6.55E-54
OS04G0477300	BET1, ONAC081, ONAC81, NAC81	BORON EXCESS TOLERANT1, NAC domain-containing protein	4.49		3.93E-24
OS04G0600200	Ao1-1, AOX1a, OsAOX1a, AOX 1ª	Homodimeric diiron-carboxylate protein, Cyanide-resistant respiration in mitochondria, Reactive oxygen species (ROS) scavenging	4.43		1.91E-100
OS03G0151500	No Annotation Available	Conserved hypothetical protein.	4.39		9.57E-25
OS01G0695800	MDR4, OsABCB5, ABCB5, OsPGP5, OsMDR7, OsISC28, MRP4, OsABCB11, ABCB11	multidrug resistance protein 4, ABC transporter superfamily ABCB subgroup member 5, P-Glycoprotein 5, Iron-sulfur cluster protein 28	4.37		3.65E-19
OS02G0575000	No Annotation Available	Conserved hypothetical protein.	4.31		3.65E-84

Table 4.2: The 10 genes with the greatest negative log2fold change in expression, and the 10 genes with the greatest positive log2fold change in expression, when illuminated controls were compared to illuminated lincomycin treated samples. Log2fold changes and adjusted p-values were calculated using DESeq2. RAP-DB = The Rice Annotation Project Database.

Gene ID	Oryzabase Gene Symbol(s)	RAP-DB Description	Log2fold Change	Down or Up Regulated By Norflurazon Treatment	Adjusted P-value
OS01G0600900	cab2R, CAB-2, OsLhcp, Lhcb1	Chlorophyll a-b binding protein 2, chloroplast precursor (LHCII type I CAB-2)	-8.32	Downregulated	2.36E-280
OS01G0720500	No Annotation Available	Similar to Type I chlorophyll a/b-binding protein b	-8.04		1.77E-276
OS02G0240300	Prx29	Similar to Class III peroxidase GvPx2b	-7.77		2.64E-34
OS12G0274700	rbcS, RBCS-C, RbcS2, OsRBCS2, RbcS1, OsRbcS2, OsrbcS1	Similar to Petunia ribulose 1,5-bisphosphate carboxylase small subunit Mrna	-7.45		4.23E-206
OS12G0292400	RbcS, RbcS4, OsRBCS4, OsRbcS4	Similar to Petunia ribulose 1,5-bisphosphate carboxylase small subunit Mrna	-7.43		3.97E-228
OS08G0435900	cab, cab-n8	Similar to LHC I type IV chlorophyll binding protein	-7.36		7.04E-269
OS02G0197600	No Annotation Available	Chlorophyll a/b-binding protein type III	-7.36		2.98E-60
OS04G0414700	No Annotation Available	Photosystem I PsaO domain containing protein.	-7.20		3.81E-226
OS07G0148900	PsaK	Photosystem I protein-like protein	-7.16		5.31E-185
OS03G0251350	OsHAP5C, NF-YC, CBF-C, OsNF-YC2, Os-NF-YC2, NF-YC2, NFYC2, OsNF-YC4-1, NF-YC4-1	NUCLEAR FACTOR Y (NF-Y) transcription factor	-6.92		1.28E-17
OS12G0248600	No Annotation Available	Conserved hypothetical protein.	7.41	Upregulated	2.39E-21
OS01G0934800	Pir7b	Alpha/beta hydrolase fold-1 domain containing protein.	7.16		1.26E-98
OS06G0323100	No Annotation Available	indole-3-acetate O-methyl-transferase 1-like, jasmonate O-methyltransferase	7.03		2.43E-18
OS03G0245800	OsHSP26, Oshsp26, OsHSP26.7, HSP26.7	26kDa heat shock protein	6.90		1.94E-17
OS08G0155900	OsDR10	Similar to Pathogen-induced defense-responsive protein	6.79		1.34E-204
OS10G0527800	OsGSTU12, OsGSTU6.2, GSTU6.2	Glutathione S-transferase, Regulation of leaf senescence	6.64		1.54E-144
OS03G0841600	No Annotation Available	UDP-glucuronosyl/UDP-glucosyltransferase family protein.	6.38		6.51E-56
OS01G0373700	No Annotation Available	Conserved hypothetical protein.	6.31		5.49E-14
OS06G0146600	No Annotation Available	Hypothetical protein	6.30		5.98E-14
OS01G0949800	OsGSTU36	Similar to Glutathione S-transferase GST 28	6.25		1.99E-13

Table 4.3: The 10 genes with the greatest negative log2fold change in expression, and the 10 genes with the greatest positive log2fold change in expression, when illuminated controls were compared to illuminated norflurazon treated samples. Log2fold changes and adjusted p-values were calculated using DESeq2. RAP-DB = The Rice Annotation Project Database.

However, each inhibitor significantly altered the abundance of over a thousand transcripts. As a broader picture of these genes would give a more accurate comparison of the effects of lincomycin and norflurazon, Gene Ontology (GO) terms that were enriched in each of these categories were assessed.

4.3.7 Function of genes responsive to chloroplast perturbation

To provide an overview of the impact of each treatment on global patterns of transcript abundance, genes were classified into Gene Ontology (GO) terms - umbrella terms for gene functionality. To investigate genes that were most affected by the chloroplast and its signals, enrichment analysis was performed on data from illuminated inhibitor treated samples compared to controls. Of the 116 significantly enriched GO terms for lincomycin, ~70% were shared with norflurazon (Figure 4.7). However, this statistic only considers exact matches and so does not account for additional overlap associated with function; over 95% of the top 20 most enriched terms were related to metabolism for both inhibitors. This overlap in large-scale molecular changes suggests that despite their distinct modes of action, norflurazon and lincomycin led to broadly similar impacts on transcript abundance.

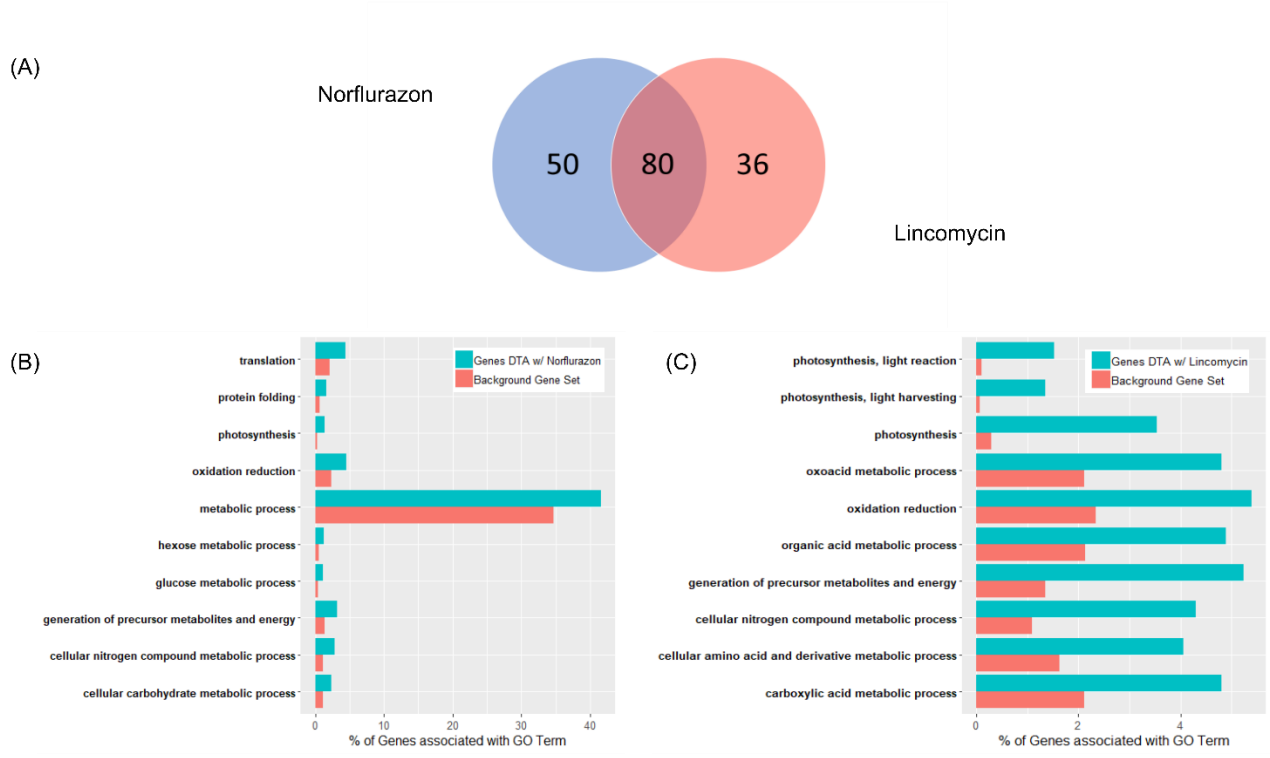


Figure 4.7: Gene Ontology analysis indicated that treatment with each inhibitor affected similar processes, with photosynthesis genes being some of the most disproportionally affected genes when seedlings were treated with either norflurazon or lincomycin. (A) Venn Diagram showing the overlap in significantly enriched GO terms after chloroplast perturbation by lincomycin and norflurazon. Bar chart showing enrichment for each GO term in the inhibitor responsive genes compared to the rest of the genome for norflurazon (B) and lincomycin (C). The Background Gene Set used contained all available *O. sativa* annotated genes. GO = Gene Ontology. DTA = Significantly Different Transcript Abundance. Linc = 10mM lincomycin, NF =100µM norflurazon.

GO terms encompassing photosynthesis, light harvesting and light reactions were all enriched after treatment with lincomycin and norflurazon. Further analysis employing MapMan (Usadel et al., 2009) demonstrated that photosynthetic genes were downregulated in the presence of an inhibited chloroplast (Figure 4.8). Notably, genes associated with both Photosystem I and Photosystem II were particularly downregulated by chloroplast perturbation. These protein complexes are comprised of both nuclear and plastome encoded subunits (Pfannschmidt et al., 2001), which could explain their high responsiveness.

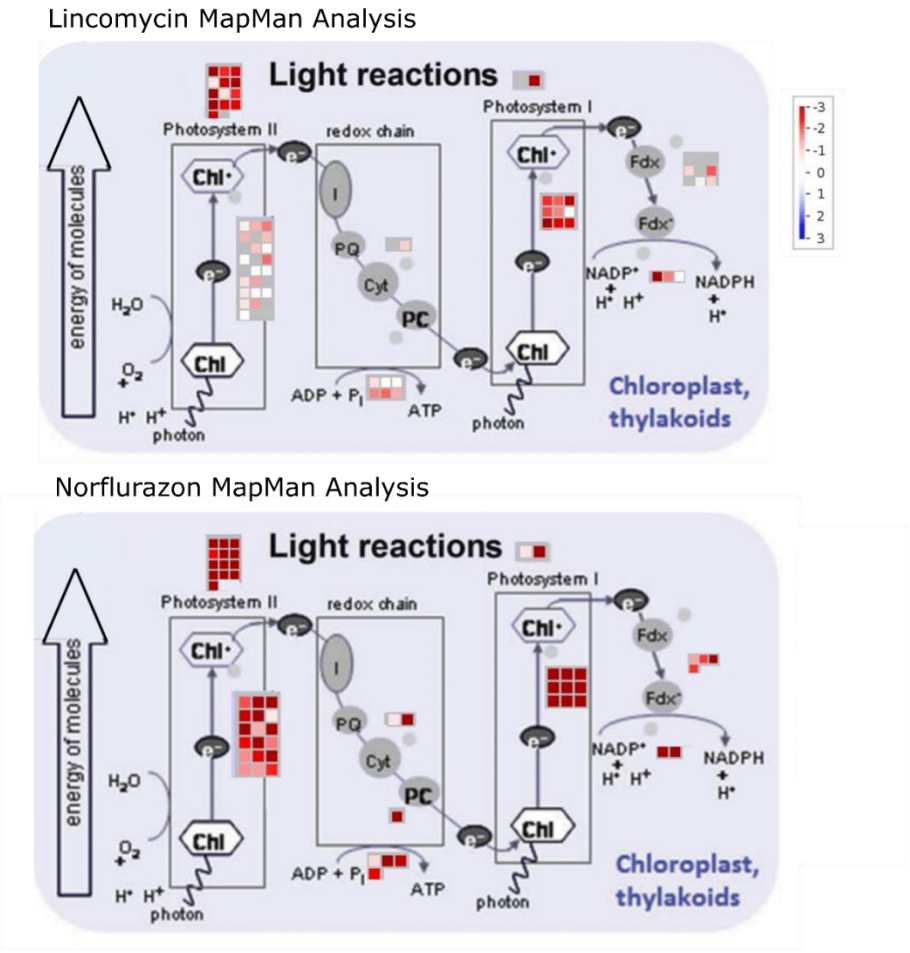


Figure 4.8: Chloroplast perturbation caused a decrease in expression of most photosynthetic genes. Transcripts derived from photosynthesis-related genes decreased in abundance after chloroplast inhibition from either lincomycin and norflurazon. Images generated using MapMan software showing analysis of differential mRNA levels for inhibitor treated samples compared with illuminated controls. False colour scale represents log2fold change.

4.3.8 Expression Patterns of Retrograde Signals and Well-characterised Responders

The transcript abundance of *LHCB1.2* decreased in illuminated samples after treatment with lincomycin or norflurazon (Figure 4.9).

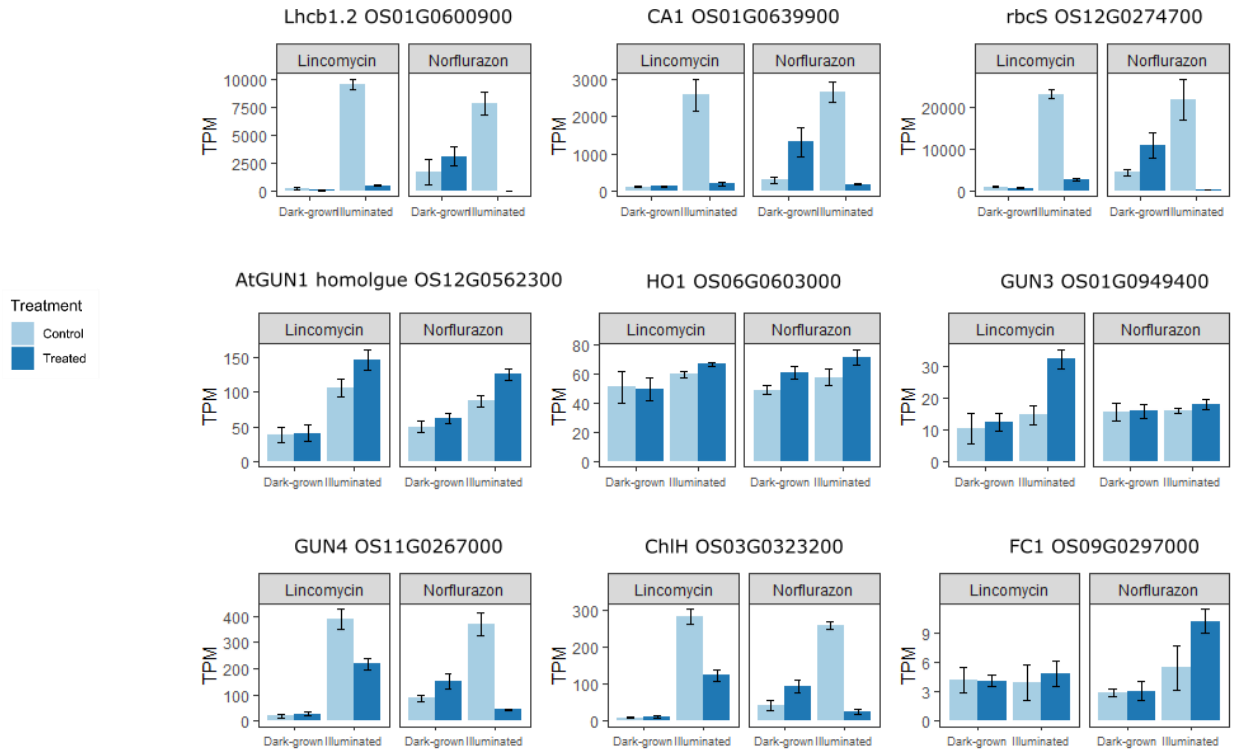


Figure 4.9: Histogram showing Transcripts Per Million (TPM) values of well-characterised retrograde responders and genes associated with retrograde signals. Dark-grown = seedlings dark-grown for 7 days; Illuminated = seedlings dark-grown for 6 days and grown in 150 µmol m^{-2} s^{-1} white light for 24 hours. Data are shown as mean and error bars represent one standard error of the mean. Lhcb1.2 = Chlorophyll a-b binding protein 2, *CA1 = Carbonic anhydrase 1, rbcS = Ribulose bisphosphate carboxylase small subunit, AtGUN1 = Arabidopsis thaliana Genome uncoupled 1, HO1 = Heme oxygenase 1, GUN3 = Genome uncoupled 3, GUN4 = Genome Uncoupled 4, ChlH = Magnesium chelatase H subunit, FC1 = Ferrochelatase 1.*

CA1 and *rbcS* also showed a decrease in transcript abundance in response to chloroplast inhibition. This is consistent with the literature, as is the similar pattern observed in transcript abundance of *CHLH*, an orthologue of *A. thaliana GUN5* (Park et al., 2017). The expression of the orthologues of the six *GUN* genes, which have been shown to have a definitive role in chloroplast to nucleus signalling in *A. thaliana*, were analysed. All six show consistent patterns of expression between rice and *A. thaliana* (Figure 3.10) with *GUN1* expression upregulated in the light, and *GUN4* and *CHLH* (*GUN5)* showing light activated expression that is repressed in sample without a competent chloroplast. *GUN2 (HO1), GUN3* and *GUN6 (FC1)* show no consistent differences in transcript abundance across treatments. *GUN* genes have been implicated to have a role in chloroplast retrograde signalling within monocotyledons; Batschauer et al (1986) found barley mutants in the tetrapyrrole pathway with a genome uncoupled phenotype when transcripts of *LHCB* did not accumulate when chloroplast inhibition occurred.

4.3.9 Chloroplast and light dependent regulation of C_4 orthologues

An initial inspection of transcript abundance in chloroplast inhibited seedling implied that *CA1* was not the only C_4 orthologue subject to regulation by the chloroplast (Figure 4.10).

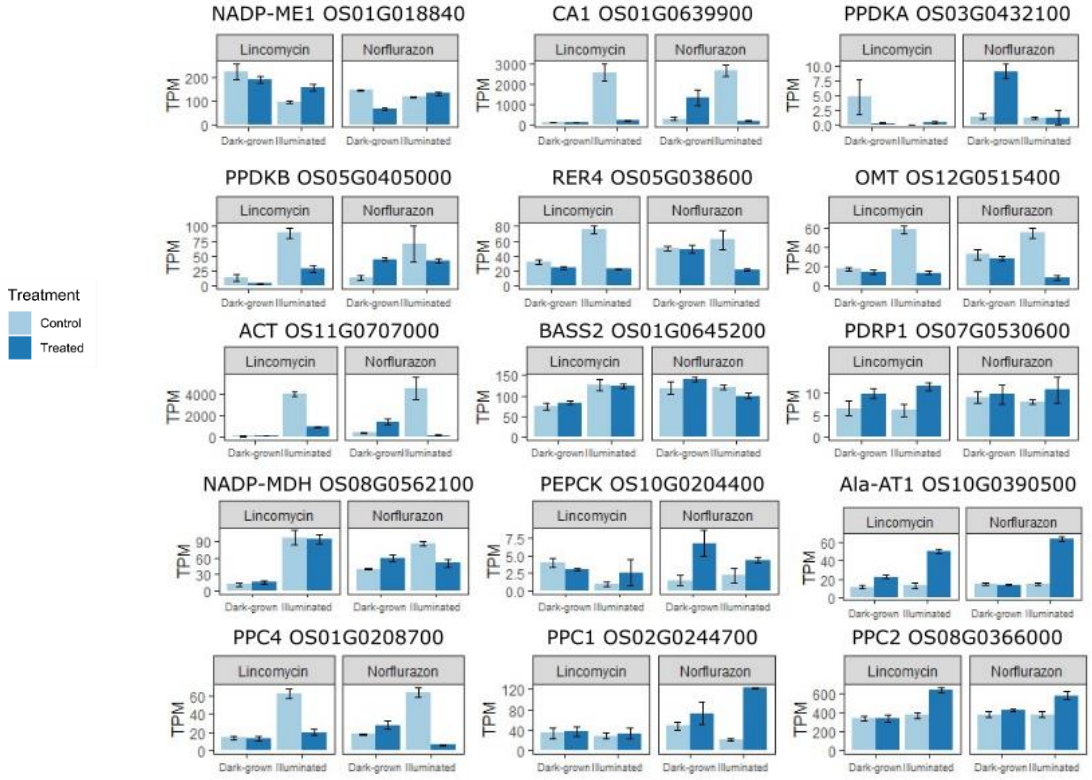


Figure 4.10: Histogram showing average Transcripts Per Million (TPM) values of well-characterised retrograde responders and genes associated with retrograde signals. Dark-grown = seedlings dark-grown for 7 days; Illuminated = seedlings dark-grown for 6 days and grown in 150 µmol m^{-2} s^{-1} white light for 24 hours. Error bars represent one standard error of the mean. NADP-ME1 = NADP-dependent malic enzyme 1, CA1 = Carbonic anhydrase 1, PPDKA/B = Pyruvate, phosphate dikinase A/B, RER4 = Reticula-related 4 , OMT = 2-oxoglutarate/malate transporter, ACT= Rubisco activase, BASS2 = Bile acid:sodium symporter family protein 2, PDRP1 = Phosphate dikinase regulatory protein 1, NADP-MDH = NADP-dependent malate dehydrogenase, PEPCK = Phosphoenolpyruvate carboxylase kinase 1, Ala-AT1= Alanine aminotransferase 1, PPC1/2/4 = Phosphoenolpyruvate carboxylase 1/2/4.

As within *A. thaliana*, chloroplast inhibition caused a significant decrease in transcript abundance in *CA1* and *PPC4* (the orthologue of *AtPPC2*), genes which function in the initial carbon dioxide fixation within mesophyll cell (Figure 4.10). To further investigate whether the expression of C_4 orthologues in *Oryza sativa* was significantly affected by chloroplast perturbation, pairwise comparisons were performed between all treatments. Comparison of inhibitor-treated seedlings versus controls was used to investigate if retrograde signalling was involved in regulation. To determine if retrograde signalling was interconnected with light regulation, this comparison was carried out in both dark-grown and illuminated samples. Very few genes were only regulated by light. Most differentially expressed C_4 orthologues were affected by both the chloroplast and light (Figure 4.11). Ten of the nineteen analysed C_4 orthologues were differentially expressed both when lincomycin was applied, and when norflurazon was applied.

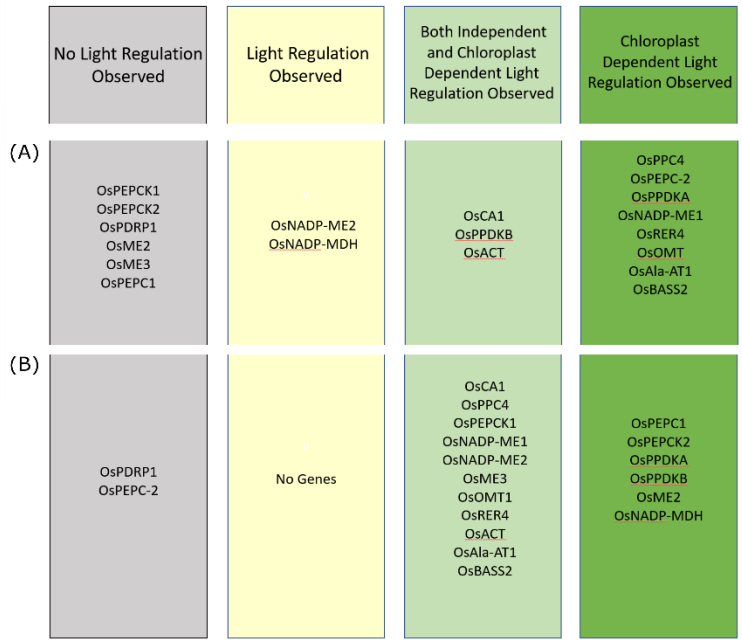


Figure 4.11: C_4 orthologues (n = 19) separated by genetic regulation, based on differential expression between four different treatments. Chloroplast-dependent regulation defined as genes which were significantly differentially expressed in controls compared to seedlings which have undergone chloroplast perturbation by (A) Lincomycin or (B) Norflurazon.

The prevalence of chloroplast signalling and light-dependent regulation within C_4 orthologues supports the hypothesis that parallel evolution has repeatedly facilitated gene recruitment into C_4 photosynthesis because these genes are already controlled by the same regulatory network as C_3 photosynthetic genes. Chloroplast perturbation caused the largest log2fold difference in the expression of *Carbonic Anhydrase 1 (OsCA1)* as compared to the other C_4 orthologues studied. Therefore, this has been highlighted as of gene of interest for future work. These data therefore show that retrograde signalling impacts the regulation of core C_4 cycle genes in both rice and *A. thaliana*.

4.3.10 Patterns of Gene Expression in Response to Chloroplast Inhibition

A Clust analysis was carried out to find broad patterns of gene expression occurring in the transcriptomic data; 11 clusters were acquired for lincomycin (Figure 4.12), and 15 clusters for norflurazon and their respective controls (Figure 4.13).

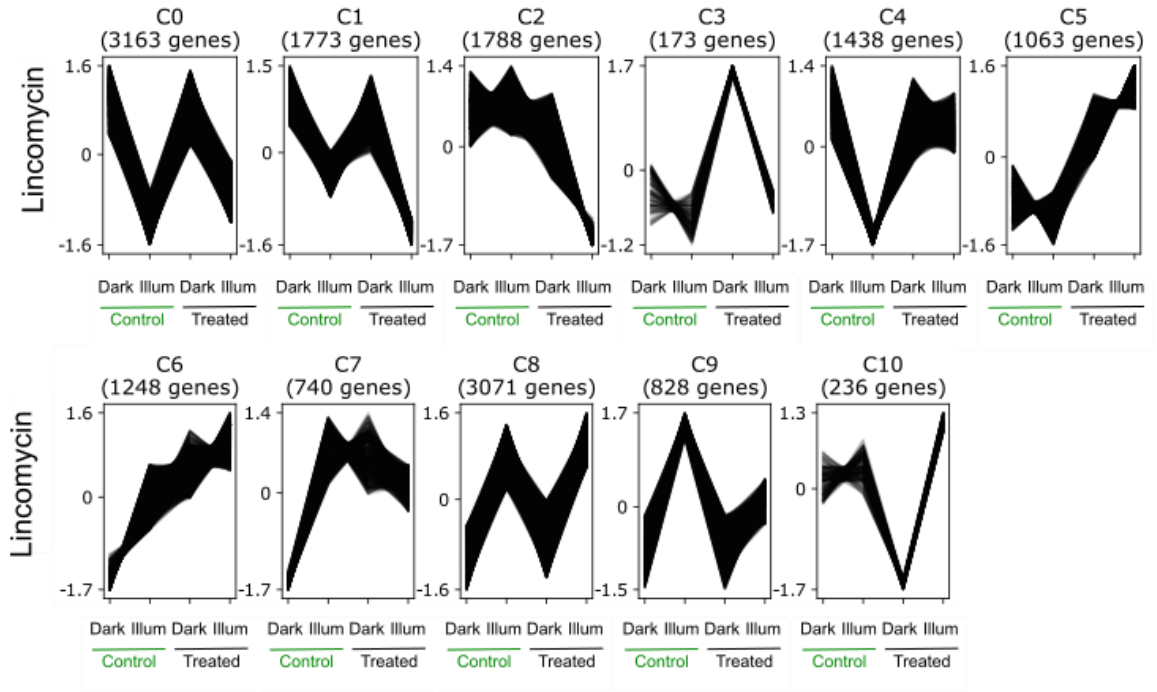


Figure 4.12: Cluster profiles generated by Clust showing expression patterns for lincomycin and related control samples. Cluster number is arbitrary and for identification only. Lincomycin = 10mM lincomycin, Dark = 7 days dark-grown, Illum = 6 days dark-grown + 24 hours 150 µmol m^{-2} s^{-1} white light.

For lincomycin, the two most common regulatory patterns were Cluster C0, containing 3163 genes, and Cluster C8, containing 3071 genes. Both clusters are representative of light-responsive genes, with Cluster C0 containing genes which were downregulated in the light, regardless of inhibitor treatment, and Cluster C8 the opposite, representing genes upregulated in illuminated samples. Additionally, there were clusters which showed a shared response to lincomycin treatment. For example, C9 represents genes whose light activation is repressed by lincomycin treatment similar to the expression pattern of *LHCB*. A similar cluster was present in data relating to norflurazon treatment (Figure 4.13).

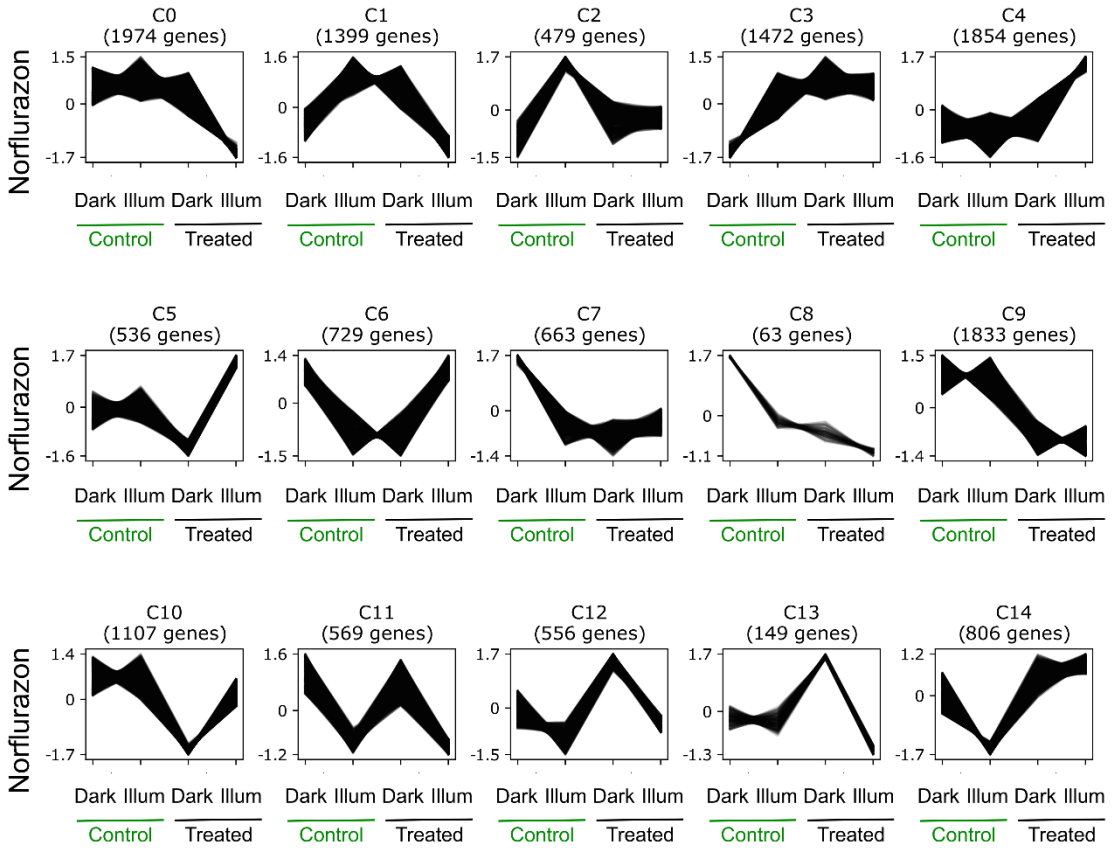


Figure 4.13: Cluster profiles generated by Clust showing expression patterns for norflurazon and related control samples. Cluster number is arbitrary and for identification only. Norflurazon = 100µM, Dark = 7 days dark-grown, Illum = 5 6days dark-grown + 24 hours 150 µmol m^{-2} s^{-1} white light.

For norflurazon, Cluster C2 shows that 479 genes had their light activation repressed by the herbicide treatment (Figure 4.13). Overall, 7 of the 11 clusters occurring in the lincomycin data appeared in the related norflurazon data. However, there were consequently 8 clusters in the norflurazon related data which do not appear in the lincomycin counterpart analysis. This is congruent with the earlier findings of the PCA analysis, that showed the inhibitors having different effects on the overall transcriptome. This analysis differs from the results found in *A. thaliana* where lincomycin and norflurazon related data garnered not only the same number of clusters, but also similar patterns of expression.

4.3.11 Motif Enrichment in Chloroplast Responsive Genes

FIMO, part of the MEME software suite, is a programme which searches for known motifs by using experimentally verified motifs identified using methods such as DAP-seq and CHIP-seq. These motifs are stored as letter probability matrices, or position weight matrices, accounting for the inexact nature of transcription factor binding. The 5'UTR and 1000bp upstream of plastid-responsive genes was analysed using FIMO. 935 genes were defined as plastid-responsive which have differential transcript abundance in both illuminated lincomycin and norflurazon treated seedlings compared to controls (Table 4.1). Enriched motifs can highlight putative transcription factors which could act downstream of a chloroplast derived signal modifying nuclear gene expression. This analysis showed that binding motifs that were enriched more than 2.5-fold compared to background in chloroplast responsive genes were from five major transcription factor families: MYBs, bZIPs, bHLH, MYB-

related and C2C2 GATA in addition to AZF1. Four of these transcription factor families associated with enriched motifs in chloroplast responsive genes in *A. thaliana*.

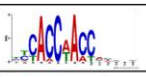
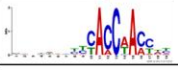
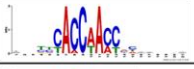
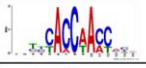
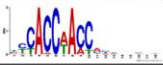
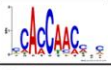
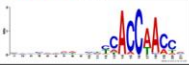
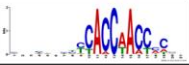
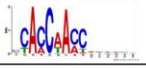
Motif	Total Motif Frequency	Fold change in Frequency compared to Average Background	Number of Genes in which the Motif Occurs	PWM
MYB_tnt.MYB13_col_a_m1	36	7.30	31	
C2H2_tnt.AZF1_colamp_a_m1	23	6.75	17	
C2H2_tnt.AT4G26030_col_a_m1	28	4.97	26	
MYB_tnt.MYB58_colamp_a_m1	40	4.94	36	
MYB_tnt.MYB58_col_a_m1	28	4.86	23	
MYB_tnt.MYB13_colamp_a_m1	34	4.67	32	
MYB_tnt.MYB83_col_a_m1	42	4.36	36	
MYB_tnt.ATY19_col_a_m1	55	4.30	43	
MYB_tnt.MYB63_col_a_m1	51	4.06	44	
MYB_tnt.MYB83_colamp_a_m1	36	3.98	31	

Table 4.4: The ten most enriched DAP-seq motifs within the promoter (1000bps upstream of TSS) and 5' UTR of chloroplast responsive genes with a total frequency over 20. Enrichment measured as fold change in the frequency of significant matches as calculated by FIMO compared to the average of 5 random background gene sets. PWM = position weight matrix.

The PWMs of the ten most enriched motif show a strong prevalence of the motif CACCAACC, which binds several members of the MYB transcription factor family (Table 4.4). One exception to this was AZF1 binding motif, which was also found to be present in the promoters of known chloroplast responders in *A. thaliana*; *LHCB, CA1* and *CHLM* (Figure 3.16). As the DAP-Seq motifs have all been experimentally verified in *A. thaliana*, another FIMO analysis was carried out searching for occurrences of known motifs collated from the JASPAR motif database which included motifs from an array of plant species.

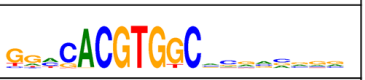
Motif ID	Motif Name	Species	Total Motif Frequency	Fold change in Frequency compared to Average Background	PWM
MA0968.1	bZIP68	*A. thaliana*	38	3.83	
MA0931.1	ABI5	*A. thaliana*	27	3.46	
MA1410.1	StBRC1/ TCP18	*S. tuberosum*	55	2.98	
MA1190.1	RVE5	*A. thaliana*	25	2.82	
MA0570.1	ABF1	*A. thaliana*	65	2.81	
MA1338.1	AREB3	*A. thaliana*	39	2.79	
MA1050.1	Osl_08196/ PCF2	*O. sativa*	26	2.70	
MA1066.1	TCP23	*A. thaliana*	26	2.70	
MA1062.1	TCP15	*A. thaliana*	28	2.70	
MA1349.1	bZIP16	*A. thaliana*	70	2.56	

Table 4.5: The ten most enriched JASPAR motifs within the promoter (1000bps upstream of TSS) and 5' UTR of chloroplast responsive genes with a total frequency over 20. Enrichment measured as fold change in the frequency of significant matches as calculated by FIMO compared to the average of 5 random background gene sets. PWM = position weight matrix.

Half of the ten most enriched motifs from the JASPAR database contained a, G-box (CACGTG), an element known to be frequent in light-responsive genes and to be an important cis-element controlling the regulation of *CHLH* (aka. *GUN5*) expression (Kobayashi et al., 2012). The G-box was also shown to be enriched in *A. thaliana* (Table 3.5) along with the GATATT motif, found here in RVE5, which binds MYB-like transcription factors. This analysis still predominately found motifs experimentally verified in *A. thaliana*, which is not surprising as the majority of plant research has been carried out in this model species. However, there were two exceptions; a TCP binding motif verified in potato and a TCP binding motif verified in rice which binds the transcription factor PCF2. PCF2 binds to the rice proliferating cell nuclear antigen (*PCNA*) gene which has a role in meristematic tissue-specific expression (Kosugi & Ohashi, 1997). However, due to many transcription factors from rice not yet having verified binding motifs this match should be taken as an indicator of which transcription factor family to focus on, rather than focusing on this specific transcription factor.

4.3.12 Motif Enrichment Analysis of Chloroplast Responsive Genes using AME

An additional tool available in the MEME software suite is AME (Analysis of Motif Enrichment). Whilst FIMO (Find Individual Motifs) scans a set of sequences for individual matches to each of the motifs you provide, AME identifies motifs that are relatively enriched in your sequences compared with control sequences.

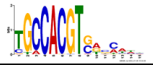
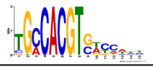
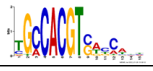
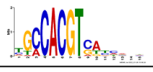
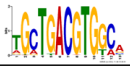
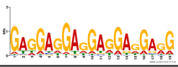
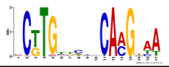
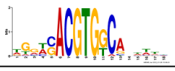
Motif	P-value for Enrichment	PWM
bZIP_tnt.bZIP16_col_v3a_m1	2.40E-10	
bZIP_tnt.AREB3_colamp_a_m1	1.04E-08	
bZIP_tnt.AREB3_col_v31_m1	2.43E-08	
bZIP_tnt.ABI5_colamp_v3b_m1	8.44E-08	
bZIP_tnt.ABI5_col_v3h_m1	9.82E-08	
bZIP_tnt.bZIP28_col_a_m1	5.85E-08	
bZIP_tnt.bZIP42_colamp_a_m1	2.94E-07	
bZIP_tnt.GBF3_col_m1	2.22E-07	
C2H2_tnt.TF3A_col_a_m1	8.88E-08	
bZIP_tnt.bZIP3_col_a_m1	2.85E-07	
NAC_tnt.ANAC017_colamp_v31_m1	8.56E-07	
bZIP_tnt.ABF2_col_v3a_m1	9.31E-07	

Table 4.6: The ten most enriched DAP-seq motifs within the promoter (1000bps upstream of TSS) and 5' UTR of chloroplast responsive genes. Enrichment measured as the p-value of likelihood motif occurs more frequently in the genes set of interest over background as calculated by AME. PWM = position weight matrix.

G-box features heavily in the enriched binding sites identified by AME. This suggests that the G-box is more frequent than the MYB-binding motifs found with FIMO, however, there is lower affinity for binding at these G-box sites. The affinity of binding for a motif can be seen in the position weight matrix (PWM), with base changes at certain positions increasing the likelihood of a transcription factor binding a motif, represented by a taller letter in the PWM (Table 4.6). Many of the bZIPs shown to be enriched in this analysis were also enriched in *A. thaliana.*

4.3.13 Distribution of open chromatin in *O. sativa*

Transcription factors are more likely to bind areas of open chromatin. The above analysis looked only at the 5' UTR and 1000 base pairs upstream of the transcription start site. By looking at the distribution of open chromatin regions, areas of the genetic architecture that are likely rich in binding motifs can be identified. The distribution of open chromatin was therefore analysed to inform a more refined motif analysis, in which only the most accessible elements of the genetic architecture are considered. This information could inform the most relevant areas of a gene's sequence to input into the motif analysis, and should provide a more stringent list of motifs that are relevant to the regulation of the genes being assessed.

ATAC-Seq is a method which treats intact isolated nuclei with a transposase enzyme which preferentially cuts open chromatin. To identify these regions of open chromatin, also referred to as transposase hypersensitive sites (THSs), ATAC-seq data was used from Wilkins et al (2016). The bed file containing the coordinates for these THSs site was downloaded from NCBI (GSE75794). To investigate their distribution, the distance of each THS from the transcription start site (TSS) was plotted (Figure 4.14).

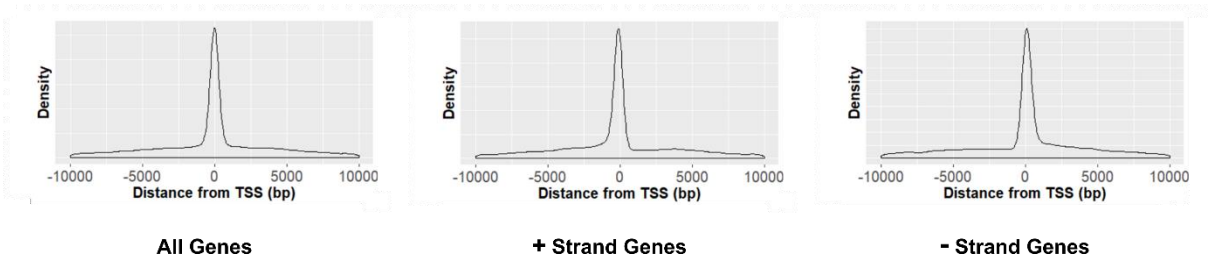


Figure 4.14: Distribution of THS (Transposase Hypersensitive sites) around the TSS for all annotated *O. sativa* genes, showing that most THSs were covered by 2000bp of the promoter region (Wilkins et al., 2016; NCBI GSE75794). 0 = TSS. + Strand represents positive strand genes which read from 5' to 3', – Strand represents negative strand genes which read from 3' to 5'.

Over half of THSs occur less than 2000 base pairs upstream of the TSS, making this a suitable length of promoter to include in further motif analysis. However, 6336 genes either overlap or have their TSS directly adjacent to another gene and 14706 genes have less than 2000bp between their TSS and the coding sequence of another gene. This means the sequence selected for the analysis would often partially include an adjacent gene to the gene of interest. To ensure that a 'one size fits all' approach analysing 2000bp upstream along with the gene body is appropriate, THS distribution plots were generated for genes with a low inter-gene distance between the 3' end and the next gene to ascertain whether they require a different analysis method (Figure 4.15).

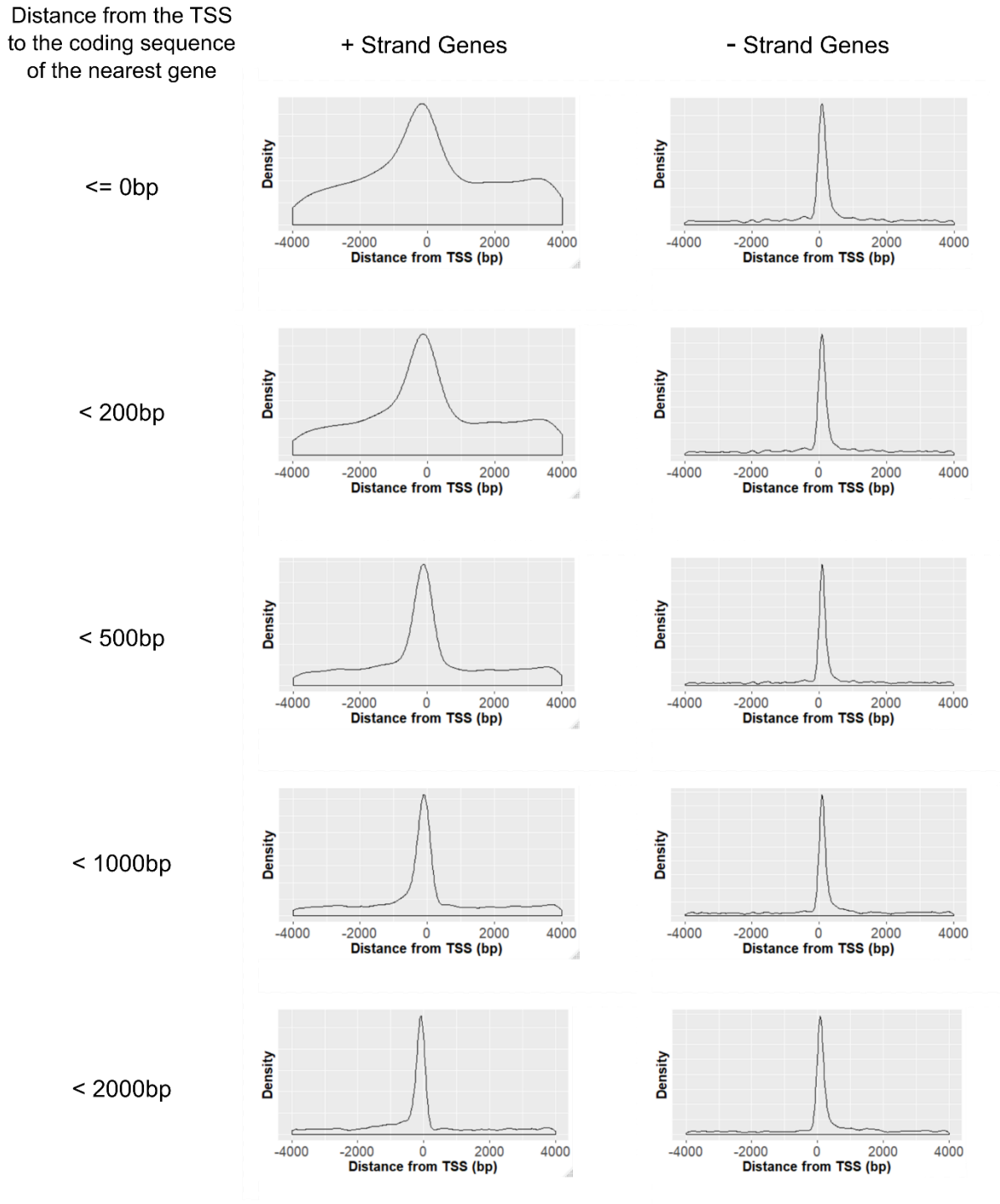


Figure 4.15: Distribution of THS (Transposase Hypersensitive sites) around the TSS for genes whose promoter region either overlaps or is >2000 base pairs from the next gene (Wilkins et al., 2016; NCBI GSE75794). 0 = TSS. + Strand represents positive strand genes which read from 5' to 3', – Strand represents negative strand genes which read from 3' to 5'.

The THS distributions for genes with a short distance between their TSS and the coding sequence of the next closest gene show a similar shaped distribution to those for all genes. Genes with shorter distances do show a shift of THSs into the gene body. Additionally, shorter distances show a greater proportion of THSs to the periphery. This is most likely caused by an increase in background; shorter

distance plots have fewer input THSs, and therefore there is a decrease in the signal to noise ratio. Additionally, THSs in negative strand genes are closer to the TSS than positive strand genes in cases of low inter-gene distance. The reason for this phenomenon is unknown, and curiously does not occur within *A. thaliana* (Figure 3.20).

4.3.14 Motif Enrichment in the Gene Body and Most Accessible Regions of Chloroplast Responsive Genes

Based on the distribution of Transposase hypersensitive sites in the genetic architecture, the gene body and 2000bp upstream of each plastid-regulated gene was analysed using FIMO (Table 4.7). Enrichment was normalised against an averaged background frequency of 1000 runs of randomly selected gene sets that were the same size as the gene set being tested.

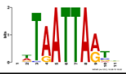
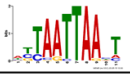
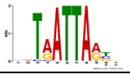
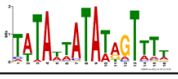
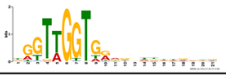
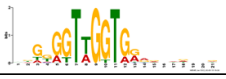
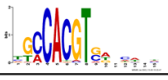
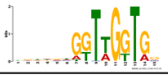
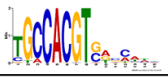
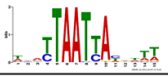
Motif	Total Motif Frequency	Fold change in Frequency compared to Average Background	PWM
ZFHD_tnt.ATHB23_colamp_a_m1	34	3.32	
HB_tnt.WOX11_col_a_m1	30	3.29	
ZFHD_tnt.ATHB25_col_a_m1	29	2.90	
ARID_tnt.AT1G04880_col_a_m1	33	2.88	
MYB_tnt.ATY19_col_a_m1	67	2.74	
MYB_tnt.MYB63_col_a_m1	62	2.60	
bZIP_tnt.GBF3_colamp_m1	45	2.56	
MYB_tnt.MYB83_colamp_a_m1	32	2.47	
bZIP_tnt.bZIP16_col_v3a_m1	52	2.43	
ZFHD_tnt.ATHB24_col_a_m1	39	2.35	

Table 4.7: The ten most enriched motifs within the gene body and promoter (2000bps upstream of TSS) of chloroplast responsive genes. Enrichment measured as fold change in frequency compared to the average of 1000 random background gene sets of the same size. PWM = position weight matrix.

Motifs which bind the MYB and bZIP transcription factor families were again enriched in the plastid-regulated gene set. However, expanding our analysis to the gene body and 2000bp upstream has highlighted the prevalence of ZF-HD binding motifs. ZF-HD transcription factors affect flower

development and have been shown to bind phosphoenolpyruvate carboxylase (*PPC*) gene in the C_4 species of the Flaveria genus (Windhövel et al., 2001). This transcription factor family is prevalent in both species; with 17 ZF-HD transcription factors in *A. thaliana* and 14 in rice (Hu et al., 2008).

4.3.15 Motif Enrichment in Chloroplast Responsive C_4 Orthologues

The motif enrichment analysis was carried out on C_4 orthologues which were mutually affected by both inhibitors in illuminated seedlings (Figure 4.10), to identify which transcription factors may be involved in the regulation of C_4 photosynthesis (Table 4.8).

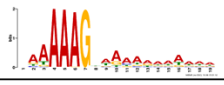
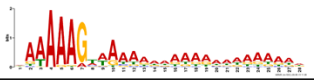
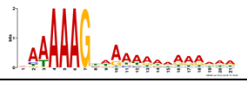
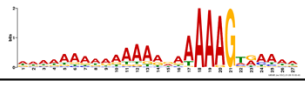
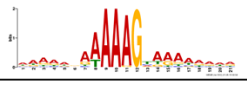
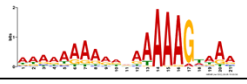
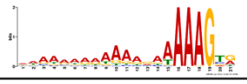
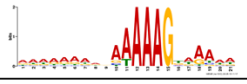
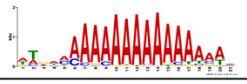
Motif	Total Motif Frequency	Fold change in Frequency compared to Average Background	PWM
C2C2dof_tnt.OBP3_colamp_a_m1	42	3.62	
C2C2dof_tnt.At4g38000_col_a_m1	318	3.15	
C2C2dof_tnt.AT2G28810_col_a_m1	33	3.14	
C2C2dof_tnt.AT1G69570_col_a_m1	68	2.73	
C2C2dof_tnt.Adof1_col_a_m1	36	2.73	
C2C2dof_tnt.OBP3_col_a_m1	143	2.64	
C2C2dof_tnt.AT1G69570_colamp_a_m1	24	2.49	
C2C2dof_tnt.At5g62940_col_a_m1	25	2.37	
BBRBPC_tnt.BPC1_colamp_a_m1	79	2.36	
REM_tnt.REM19_colamp_a_m1	350	2.33	

Table 4.8: The ten most enriched motifs within the gene body and promoter (2000bps upstream of TSS) of chloroplast responsive C_4 orthologues. Enrichment measured as fold change in frequency compared to the average of 1000 random background gene sets of the same size. PWM = position weight matrix.

The binding motifs of the C2C2dof transcription factor family were overrepresented within the plastid-regulated C_4 orthologues. One member of this family, OBP3, is a nuclear localised transcription factor which modulates phytochrome and cryptochrome signalling in *A. thaliana* (Ward et al., 2005).

4.4 Discussion

4.4.1 Effectiveness of chloroplast perturbation

Two inhibitors of the chloroplast, lincomycin and norflurazon, were used to perturb chloroplast function and in so doing investigate genes regulated by chloroplast derived signals. These inhibitors both led to reduced greening and PSII activity, leading to an approximate three-fold decrease in F_v/F_m in illuminated treated seedlings compared with controls. This is consistent with the effect of the inhibitors in *A. thaliana*, although higher concentrations of each inhibitor were required to attain the same result. These results also support previous work on rice; 2mM lincomycin treatment lead to pale leaves in one-week old rice seedlings (Duan et al., 2020). However, leaves were not entirely bleached, likely due to the lower concentration of lincomycin used compared with the work presented in this chapter (10mM). Moreover, three-week-old rice plants treated with 50μM norflurazon showed a 37% decrease in F_v/F_m compared to controls 40 hours after application (Park & Jung, 2017). This is a smaller response compared with the >60% decrease in F_v/F_m observed in this work, likely due to the difference in seedling age as here norflurazon was applied from germination onwards. Once it was established that lincomycin and norflurazon had the expected result on chloroplast formation and PSII activity in rice, RNA-seq was performed to investigate its effect on gene expression.

4.4.2 Genes Responsive to Chloroplast Perturbation

Transcript abundance was significantly different between dark-grown and 24-hour light exposed seedlings for 1805 genes which represents 5.5% of all detected transcripts. This is much less than the 20% predicted to be light responsive by microarray data (Jiao et al., 2005), and contrasts with *A. thaliana* where 26% of genes were found to be light responsive, despite the two species being predicted to have the same magnitude of response. Of the 33064 rice genes whose transcripts were detected in the analysis reported here, 4362 showed significantly different transcript abundance when illuminated samples were treated with norflurazon as compared to untreated controls. However, only 1370 genes showed a significant response to lincomycin treatment. This fits with the ratio observed in *A. thaliana*, and with phenotypic data reported in the literature where norflurazon application has been shown to cause seemingly unrelated side effects such as decreased total lipid quantity and increased dry weight weight (Di Baccio et al., 2002; Magnucka et al., 2007).

Whilst norflurazon application had a larger impact on gene expression than lincomycin, there was overlap in the genes affected by each inhibitor with 68% of the genes affected by lincomycin treatment also affected by norflurazon treatment. Despite PCA indicating differences in the response of gene expression to each inhibitor, there was large overlap not only in affected genes but in GO annotations which were enriched. This suggests that, overall, they are comparable and focussing on genes affected by both increases the robustness of the findings. For both inhibitors the most enriched gene ontology terms include multiple photosynthesis associated terms. Other shared GO terms included ‘Generation of precursor metabolites and energy’, which encompasses photosynthesis, oxidation and NADPH regeneration, and ‘Oxidation Reduction’ (Figure 4.7). Reactive oxygen species (ROS) have been implicated as a chloroplast retrograde signal in several potential pathways, including the redox potential of the chloroplast transmitted through the malate valve (Heyno et al., 2014), and the detection of singlet oxygen by executor proteins and β-cyclocitral (Ramel et al., 2012; Uberegui et al., 2015). The malate valve operates when the NADPH pool is over-reduced, NADP-

MDH oxidises the pool by using the reduction power to convert OAA to malate which is then exported to the cytosol acting as a signal.

The gene *LHCB2.1* is known to show decreased transcript abundance in response to chloroplast perturbation in wildtype, and consequently a seedling lacking this response was considered the diagnostic for defining a *gun* mutant. In rice, this classic response of *LHCB2.1* was observed, with strong induction in response to light, but a repressed induction in the presence of chloroplasts inhibitors. The same response was detected for rice genes *CA1* and *RbcS* (Figure 4.9). Decreased expression of *RbcS* and *CHLH* in rice after norflurazon treatment has previously been reported (Park & Jung, 2017). However, *HO1* (GUN2) expression was also reported to decrease expression (Park et al., 2017). This response was not observed in the work presented here. Overall, the data presented here indicate that the response of genes classified as *gun* mutants were consistent between rice and *A. thaliana*. Taking into account *gun*-like mutants have been found in barley (Batschauer et al., 1986), these findings therefore provide evidence to support the notion that *GUN* retrograde signalling pathway is derived from the last common ancestor of rice and *A. thaliana*.

4.4.3 The extent to which C_4 orthologues are subject to chloroplast retrograde signalling in *O. sativa*

Ten C_4 genes from rice were significantly affected by both inhibitors and are therefore likely subject to chloroplast-dependent regulation in this C_3 species. However, only three of these C_4 specific orthologues were found to be differentially expressed in *Arabidopsis thaliana* (Burgess et al., 2016). This low degree of overlap in response between the two species is likely associated with a number of factors relating to the genetic distance between *Arabidopsis thaliana* and *Oryza sativa.* For example, multiple genome duplications have occurred since their divergence ~140 million years ago (Lee et al., 2013). The C_4 genes analysed in this study are not thought to be directly involved in photosynthesis of rice, however, many are already known, or at least predicted, to be plastid localised. An example is Phosphoenolpyruvate Carboxylase 4 (*PPC4*), which shares 73% amino acid identity with the *PPC* gene recruited into C_4 photosynthesis in sorghum (Wang et al., 2009). This gene is plastid localised and has been shown to function in ammonium assimilation (Masumoto et al., 2010). Whilst its location means plastid regulation may be expected, the results showed light was also an important regulator of *PPC4*.

Carbonic anhydrase 1 (*CA1*) was the most responsive gene to chloroplast perturbation. Microarray data has previously been used to show that this gene is one of the most responsive to norflurazon treatment of *A. thaliana* (Mochizuki et al., 2008). *CA1* is predicted to be plastid-localised in rice. Loss of a chloroplast transit peptide is responsible for the cytosolic localisation in C_4 species (Clayton et al., 2017; Tanz et al., 2009). Although plastid-localised genes were prominent in the list of genes found to be subject to regulation through retrograde signalling, some genes such as *Pyruvate Orthophosphate Dikinase A* (*PPDKA*) localise to the cytosol (Wang et al., 2020). This shows that the repeated recruitment of genes in C_4 evolution which are subject to regulation by retrograde signalling cannot be explained by prior plastid-localisation of gene product or a known plastid-related function.

4.4.4 Regulatory elements enriched in chloroplast responsive genes

There were large differences in the transcription factor binding sites found to be most enriched in chloroplast responsive genes in rice and *A. thaliana*. This is unsurprising considering many transcription factors involved in photosynthetic regulation in *A. thaliana* have no direct orthologues

in rice (Wang et al., 2017). Moreover, the low commonality between DNA sequence motifs between the two species suggest that transcription factor binding could differ significantly (Cserhati, 2015). However, of the six families of transcription factors which had enriched motif binding sites within chloroplast responsive genes in *A. thaliana* (bZIP, bHLH, TCP, MYB, MYB-related, and AP2/EREBP), five were also enriched in rice (bZIP, bHLH, TCP, MYB, MYB-related). Additionally, when the motif search was carried out in AME to look for the most frequent binding motifs, both species were enriched for bZIP binding motifs containing the G-box (CACGTG). The G-box is an element known to be frequent in light-responsive genes and is an important *cis*-element controlling the regulation of *CHLH* (aka. *GUN5*) expression (Kobayashi et al., 2012) therefore already being implicated to have a role in a chloroplast retrograde signalling pathway.

As well as overlaps in the families of transcription factors involved, there were many instances of a binding motif for a specific transcription factor being enriched in both species. For example, the AZF1 binding motif, found to be enriched 1000bp upstream in chloroplast responsive rice genes (Table 4.4) was also found to be present in the promoters of known chloroplast responders in *A. thaliana*; *LHCB*, *CA1* and *CHLM* (Figure 3.16). AZF1 has been shown to be involved in the abiotic stress response by negatively regulating abscisic acid-repressive genes (Kodaira et al., 2011). The most notable overlap between the two species is that of the ten most enriched motifs output from the AME analysis in rice and *A. thaliana*, seven were shared. These seven motifs represented five transcription factors; ABF2, AREB3, ABI5, bZIP16 and GBF3. These are all promising candidates for future characterisation, hopefully being involved in an ancestral chloroplast to nucleus regulatory pathway. Less overlap was observed between enriched motifs in C_4 orthologues between the two species. However, one transcription factor was shared, OBP3, a nuclear localised transcription factor which modulates phytochrome and cryptochrome signalling in *A. thaliana* (Ward et al., 2005). OBP3's nuclear localisation means it could function as the final component in a theoretical chloroplast to nucleus signalling pathway which controls C_4 genes in a wide array of species.

5 Optimisation of ATAC-seq for identification of open chromatin regions and transcription factor binding footprints in Rice

5.1 Introduction

5.1.1 How ATAC-seq Identifies Routes of Gene Regulation

Gene expression can be regulated at multiple levels including transcriptional, translational or the processing, transport and degradation of mRNA. Several methods have been developed to investigate transcriptional regulation, including approaches to identify transcription factor binding sites. These include ChIP-Seq which can identify the DNA sequences a specific protein binds in vivo (Barski et al., 2007; Johnson et al., 2007) and the yeast-one hybrid assay that defines sequences bound by a protein in yeast (Ouwerkerk & Meijer, 2001). More recently, methods have been published that aim to define patterns of transcription factor binding at a genome wide scale. These methods can identify where transcription factors are likely to bind within the genome without previous knowledge of the transcription factors or DNA motifs subject to binding. Examples of such methods include DNase I hypersensitive sites sequencing (DNase-seq), micrococcal nuclease digestion with deep sequencing (MNase-seq), Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE-seq) and most recently Assay of Transposase Accessible Chromatin sequencing (ATAC-seq) (Klein & Hainer, 2020).

All these techniques use deep sequencing to identify regions of open chromatin. FAIRE-seq crosslinks DNA and bound proteins, and then separates fragments by weight using centrifugation. The heavier protein bound fragments are then purified and sequenced and the resulting reads indicate which regions are associated with protein binding. DNase-Seq, MNase-Seq and ATAC-seq all treat nuclei with an enzyme that preferentially cuts open chromatin and sequences the resulting fragments to identify the distribution of cuts. This distribution can be used to predict accessible chromatin. Assay of Transposase Accessible Chromatin sequencing (ATAC-seq) is a method which identifies regions of bound DNA by digesting nuclei with a transposase that preferentially cuts open chromatin (Buenrostro et al., 2013). After deep sequencing, aligning the resulting sequences to the genome shows where the transposase has cut. This can allow predictions of which regions of a chromosome are accessible to transcription factors, and at a deeper resolution the specific location of transcription factor binding sites. When done in concurrence with mRNA-seq, the predicted binding sites which

intersect with significantly differentially expressed genes form a list of putative *cis*-elements which can inform further characterisation work.

ATAC-seq has several advantages over similar methods. The Tn5 (transposase 5) enzyme cuts open chromatin and simultaneously attaches Illumina adapters. By performing digestion and adapter ligation concurrently, the time required for library preparation is reduced. Additionally, ATAC-seq only requires 50,000 nuclei per sample (Lu et al., 2017). Equivalent techniques require substantially more tissue, for example, accurate footprinting using DNaseI-seq requires 200 million nuclei per sample. This makes ATAC-seq a cost efficient and quick method for genome wide footprinting of transcription factors, with library preparation taking hours instead of days.

5.1.2 ATAC-Seq Results in Plant Tissue

ATAC-seq was initially developed to analyse human cells for medical research (Buenrostro et al., 2013) and has been predominately used in animal studies. However, it has also been employed in plants. It has previously been used to look at the stress response in rice seedlings exposed to different temperatures and water deprivation (Wilkins et al., 2016). Another study which used this technique analysed root tip cells in four plant species; *Arabidopsis thaliana, Medicago truncatula, Solanum lycopersicum*, and *Oryza sativa* (Maher et al., 2018). This revealed interesting inter-species variation in the location of open chromatin across the genome. For example, in *A. thaliana* 50% of open chromatin regions, also known as transposase hypersensitive sites (THS), were found to be in the promoter region, defined as 2000bp upstream of the TSS. This is considerably more than the other species studied, with ~40% in rice and only ~12.5% in *Solanum lycopersicum* (tomato).

5.1.3 Potential Applications of ATAC-seq in Understanding Plastid-to-nucleus Signalling

The aim of the work presented in this chapter was to optimise the method and pipeline analysis for ATAC-seq on rice tissue. Once this optimisation has taken place, the approach could be used to test if transcription factor binding differs between experimental treatments. For example, it could be carried out on chloroplast inhibited seedlings versus controls. This analysis would show if any C_4 orthologues have *cis*-elements whose binding is impacted by the chloroplast, informing further work towards the ultimate aim of identifying the regulatory mechanisms through which plastid signals impact gene expression. The dataset will also be of more general relevance as it can act as an atlas of transcription factor binding sites involved in chloroplast retrograde signalling.

5.1.4 Optimisation of Tissue Preparation and the ATAC-seq Data Analysis Pipeline

The recent creation and expansion of the technique means there is not yet a single established methodology or analysis pipeline. Two possible methods of tissue collection are first, flash freezing samples in frozen nitrogen and grinding, or second, sampling fresh tissue and finely chopping this tissue before nuclei extraction. It is unknown if the grinding process could affect bound transcription factors before nuclease treatment. Chopping however, whilst done on ice, is not rapidly preserving the cell's state and its possible that the cell's stress response could change the binding profile before the nuclease treatment can be applied. This work aimed to identify if these two methods led to significant difference in output, both in terms of the quality of the results and in location of the accessible regions identified.

ATAC-seq allows predictions of accessible chromosome regions, and when done at a high resolution, specific transcription factor binding sites. To use aligned reads to predict open regions, software specifically designed to find these peaks is employed. Several peak finding programmes exist and different software use different predictive algorithms. Due to the relative novelty of ATAC-seq there is no clear established analysis pipeline considered as the gold standard. Two different software were employed here and the resulting peaks compared to see how different the output were. Two such software which identify regions of the genome with more aligned reads than expected by random chance are HOMER (Heinz et al., 2010) and MACS2 (Zhang, 2008). This will determine if the software employed in the analysis pipeline will lead to changes in the final predicted open chromatin regions.

5.2 Objectives

The methodology and data analysis pipeline for ATAC-seq will be optimised, and open chromatin regions in the cotyledons of rice seedlings will be predicted.

Hypotheses

- There will be no major differences in predicted open chromatin regions between samples which have been prepared by chopping, or ground after flash freezing.
- There will be no major differences in predicted open chromatin regions between different peak calling software.

The distribution of predicted open chromatin regions in the genomes of rice seedling cotyledons will be elucidated.

Hypotheses

- The highest frequency of predicted open chromatin regions will occur immediately around the transcription start site.
- Most open chromatin regions will be predicted to be within 2000pb upstream of the nearest gene's transcription start site.

5.3 Results

5.3.1 Nuclei Extraction

The first step in performing ATAC-seq is to extract nuclei from samples. This can be done on flash frozen ground tissue or on fresh chopped tissue. No consensus has so far been reached on whether there was a difference in these methods, and if there is, which is preferable. This experiment aimed to identify which method is most appropriate. To confirm that nuclei were intact and not clumped they were dyed with DAPI stain and viewed under a fluorescence microscope. This is important as it ensures the transposase has equal access to nuclei and no bias is introduced.

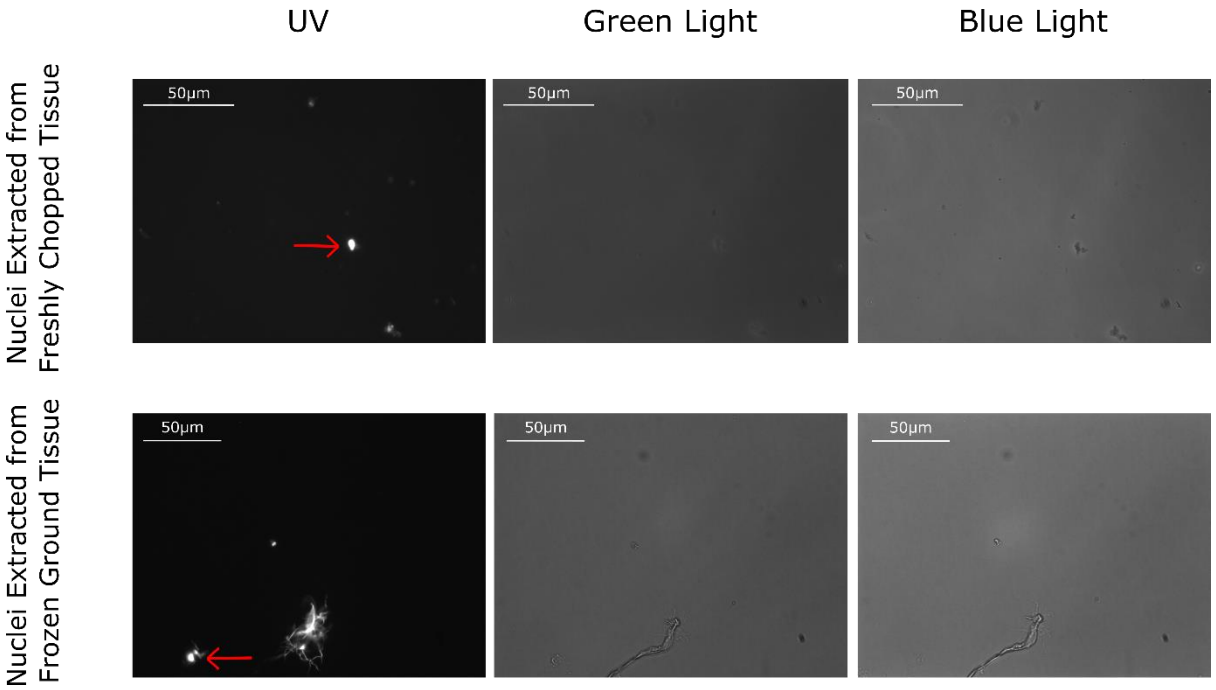


Figure 5.1: DAPI-stained nuclei extracted from freshly chopped and flash frozen ground tissue from rice seedlings. Red arrows point to nuclei. Scale bar represents 50µm.

Individual nuclei could be viewed in both preparations (Figure 5.1). Occasional debris with UV fluorescence material could be seen in ground samples, and also possible indications of nuclei clumping. Once nuclei had been extracted and a subset had been viewed under the microscope to check the nuclei shape and number, the transposase treatment was carried out and libraries were generated.

5.3.2 Analysis of ATAC-seq library prior to sequencing

To confirm that libraries were of a sufficient quality for deep sequencing, their integrity was analysed using a microfluidic electrophoresis system (Figure 5.2). This generated electropherograms, which provides an indication on the distribution of fragment sizes obtained.

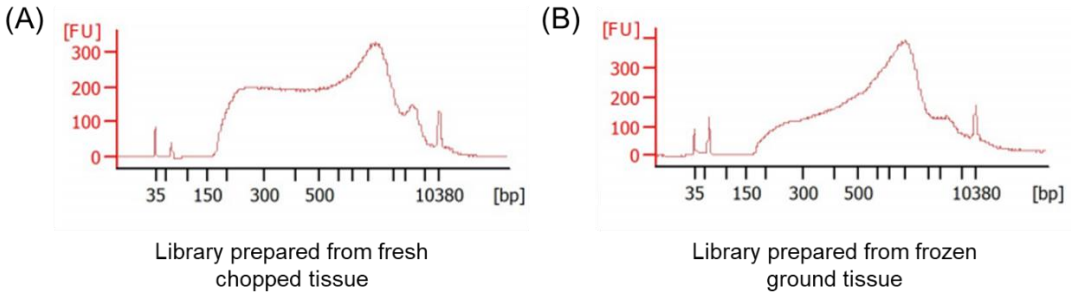


Figure 5.2: Electropherograms of rice ATAC-seq libraries from (A) fresh chopped tissue and (B) frozen ground tissue obtained on the Agilent Bioanalyzer 2100. X-axis represents fragment size and Y-axis represents fluorescence, which increases with the number of base pairs detected for a set fragment size. The higher the fluorescence at a set fragment size, the higher the concentration of fragments. FU = fluorescence units, bp = base pairs.

Reads for both libraries ranged from 200bp to 2000bp in length (Figure 5.2). Libraries prepared from freshly chopped tissue had a greater enrichment of shorter reads, with increased fluorescence detected from reads between 150 and 500bp long compared to ground tissue. Although fluorescence signal increased for longer reads in both libraries, it is worth noting that this is partially due to a greater number of bases and therefore is not a direct indicator of read number.

5.3.3 Quality control of transcriptomic libraries

Libraries were sequenced, and samples subjected to quality control. Trimming was carried out to remove low quality bases and led to a loss of 4% of bases. Visual inspection of sequences and analysis of QC components was carried out using FastQC. Analysis of samples showed they were good quality, both at sequence and base level. There were no unassigned bases (Ns) present in the data, and no adapter contamination (Figure 5.3).

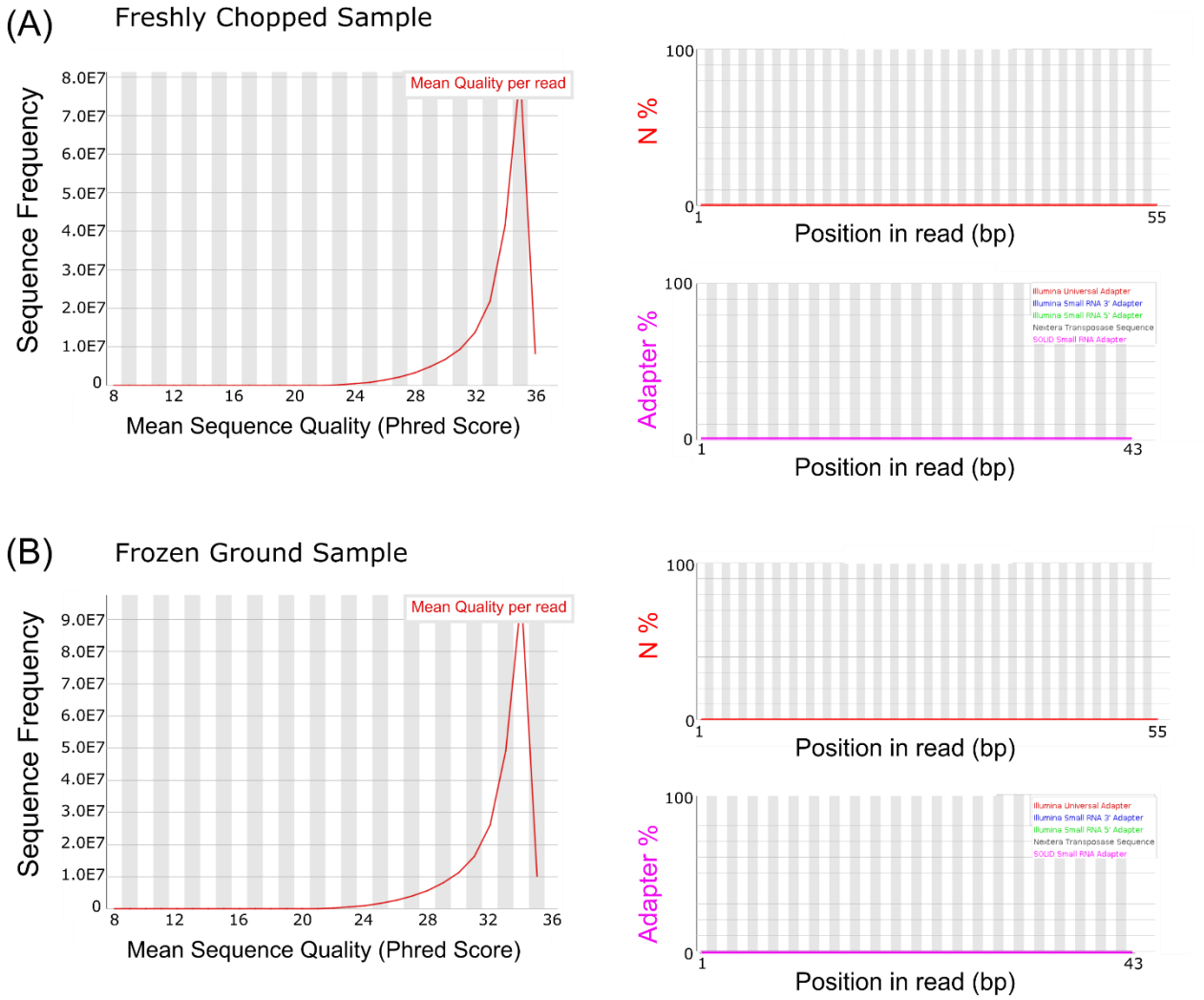


Figure 5.3: FastQC generated plots illustrating sequences were of a good quality and contained no unassigned bases or adapter contamination. Representative plots showing frequency against mean sequence quality, Ns (unassigned bases) and adapters across the read for libraries generated from (A) freshly chopped sample and (B) frozen ground samples.

However, two FastQC categories failed; Sequence Duplication Levels and Kmer content. This is due to assumptions FastQC makes in assessing quality for these categories, as it is not designed to analyse ATAC-seq generated libraries. To confirm these failed categories were associated with the methodology, plots were compared to data from Maher et al (2018) (Figure 5.4).

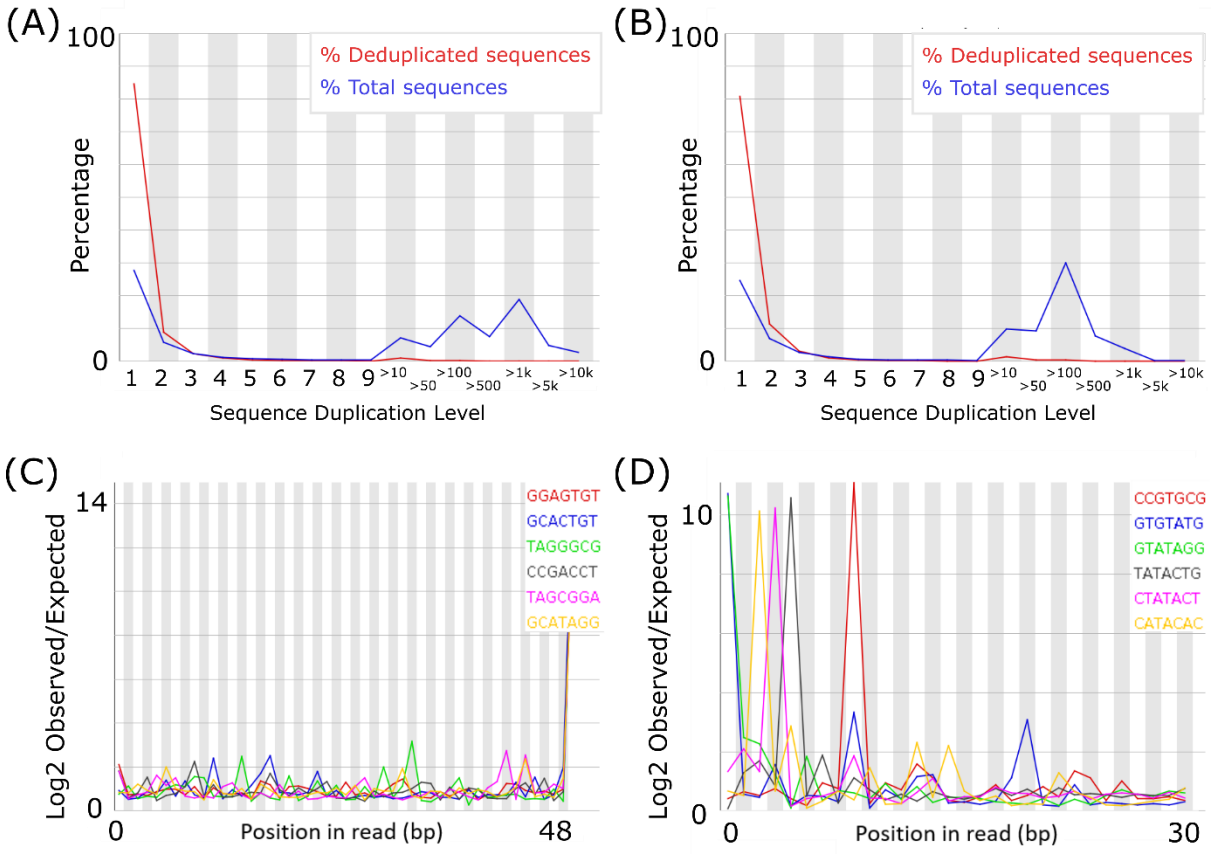


Figure 5.4: Representative FastQC plots for, showing the similarity between the libraries generated here (A and C), and from peer-reviewed work (Maher et al,. 2018; SRA ID SRP111984) (B and D). (A) and (B) Representative plot showing duplication levels, with many reads duplicated 100s or 1000s of times in both libraries (B) and (D) Representative plot showing Kmer enrichment across the read.

The data showed a large number of duplicate reads, with 67% of reads being duplicates in the chopped sample and 80% in the ground sample. This was also observed in FastQC reports produced from ATAC-seq reads generated by Maher et al (2018), where 70% of reads were duplicates. This implies that although this is not optimal, it is innate to the methodology and therefore does not impede analysis. This was similarly the case for enriched K-mers along the read (Figure 5.4).

Thus, the QC analysis indicated that all libraries had generated good quality data, and so further analysis was initiated.

5.3.4 Quality Control of Alignment

After trimming and quality control, 366 million reads were obtained for the chopped sample, of which 277 million (76%) mapped to the *Oryza sativa* IRGSP genome. For the ground sample, 453 million reads were obtained, and 365 million mapped. Ensuring an accurate read alignment is essential for ATAC-seq, as this is the input information on which open chromatin regions (or peaks) are predicted. Reads were filtered to discard reads with a MapQ, a value measuring the likelihood an alignment is real, less than 30 and all duplicate reads were removed. After this filter step, the quality of the alignment was checked by plotting the distribution of MapQ scores for each read's alignment (Figure 5.5).

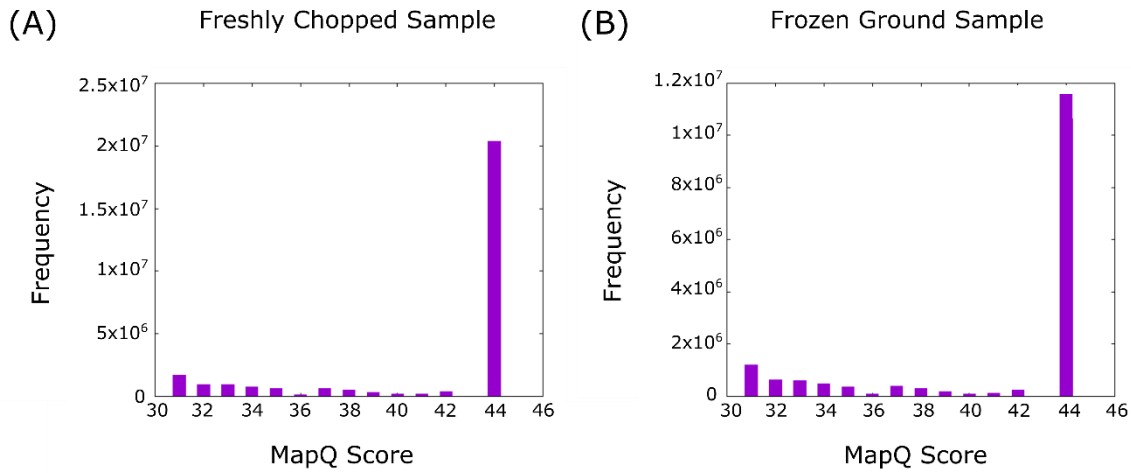


Figure 5.5: Alignments were estimated to be highly accurate for libraries generated by both sample preparation methods. Distribution of MapQ score, a measure of the confidence of each sequence's alignment, for (A) libraries generated from freshly chopped rice seedling cotyledons and (B) libraries generated from flash frozen and ground rice seedling cotyledons.

The range of available MapQ scores varies by aligner, but in this case 44 was the highest MapQ score available representing a perfect alignment of the read to the genome. Reads aligned well, with the majority having a MapQ score of 44. This mean that the alignment can be considered reliable and therefore used for peak calling to predict open regions of chromatin.

ATAC-seq uses pair-wise sequencing; this is important as the sequencing machine can only sequence a set number of bases (in this case 75 base pair). By sequencing both ends of the read, it can be determined more precisely where the transposase cut. Additionally, it allows the insert size to be calculated, which is the length of read between cuts.

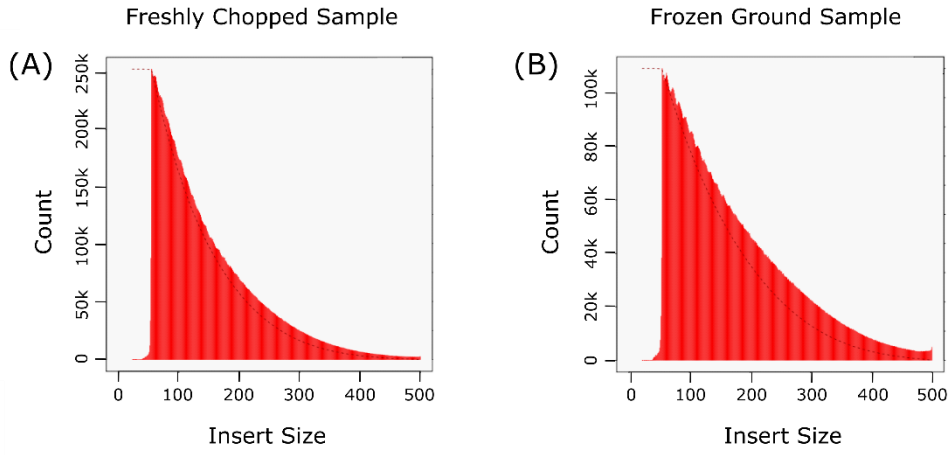


Figure 5.6: The distribution of insert sizes was consistent for libraries generated by both sample preparation methods. Distribution of insert size, the length of reads without adapters, for (A) libraries generated from freshly chopped rice seedling cotyledons and (B) libraries generated from flash frozen and ground rice seedling cotyledons.

The distribution shows that insert size peaked at 50bp and then declined in frequency to 500bp for both sample preparation methods (Figure 5.6). These results are replicated in rice ATAC-seq data (Wilkins et al., 2016, Supplementary Figure 4), confirming this is the expected distribution for libraries generated from rice.

The last step before peak calling, is to check that there is no GC bias within the reads. Several facets of GC content were plotted (Figure 5.7)

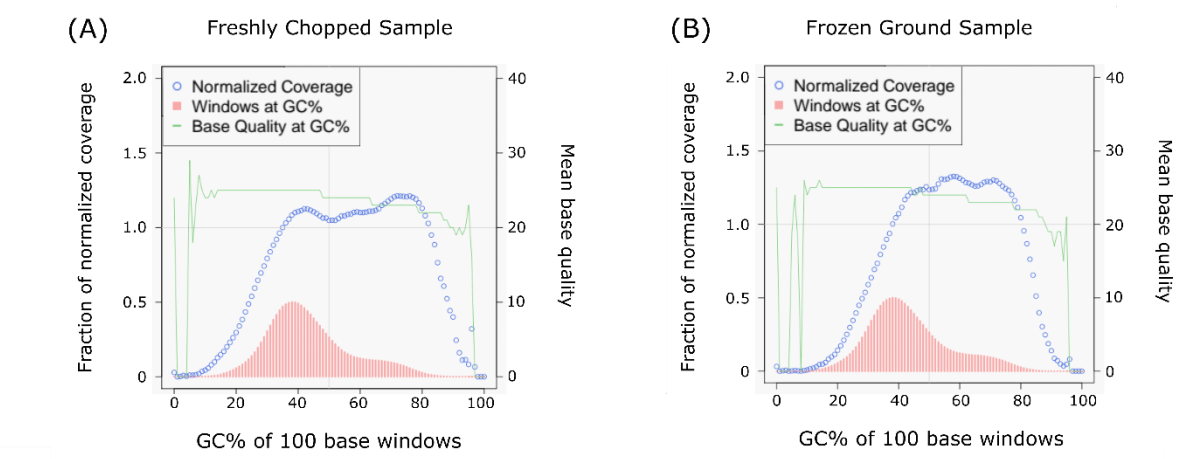


Figure 5.7: There was no GC bias for libraries generated by either sample preparation method. GC content quality measures calculated by Picard Tools for (A) libraries generated from freshly chopped rice seedling cotyledons and (B) libraries generated from flash frozen and ground rice seedling cotyledons. Normalized Coverage represents the frequency of 100bp windows of a set GC% in the reads over the frequency of 100bp windows of a set GC% in the genome.

Both chopped and ground samples showed consistency in GC content, with a skewed curve peaking at 40% trailing to 80% across 100bp windows. This replicates the GC content of the rice genome as a whole (Yu et al., 2002) and therefore there appears to have been no GC bias in the reads attained. To ensure that there was no difference in the quality of the reads across GC%, base quality was plotted (Figure 5.7). This score is generated by the sequencer and represents the confidence that a base is assigned appropriately. Mean base quality, averaged against the 100 bp in each window, was consistent across windows of different GC%. Additionally, the number of 100bp windows for each possible GC% in the reads was compared to the number of windows in the genome (Figure 5.7). This showed that extremities were depleted, with very low and very high GC% sequences occurring less in the ATAC-seq reads than in the genome. This was likely due to the size selection that occurs during the ATAC-seq methodology, and additionally, reads have been filtered out by quality controls during computational analysis. Overall, GC% content was not biased in either library, nor did GC% affect the quality of the sequences.

5.3.5 Prediction of open chromatin regions in the cotyledons of rice seedlings

Read alignment indicates where the transposase enzyme has cut genomic DNA. Accessible regions, or transposase hypersensitive sites (THSs), can be identified by looking at where the cuts are most frequent. As accessible regions occur at the peak of cut frequency, they are also often referred to as

peaks. Several different software exist for peak calling (Heinz et al., 2010; Zhang, 2008) of which Macs2 was the first software used here. 42733 peaks were called for chopped samples, and 69524 peaks were called in ground samples, 23415 of these peaks intersect.

One measure of the efficacy of peak calling is to generate a SPOT score. This score represents the fraction of reads that fall within predicted peaks, representing how many reads have been classified as signal instead of noise. Generally, a SPOT score of around 0.3 is desired. The ground sample had a SPOT score of around 0.3, falling within the expected signal to noise ratio. However, the chopped sample only generated a sample of 0.12 which means nearly 90% of reads did not fall within the predicted peaks. The length of the peaks for each sample were plotted, with peaks ranging from around 100bp to a maximum of around 1000bp (Figure 5.8).

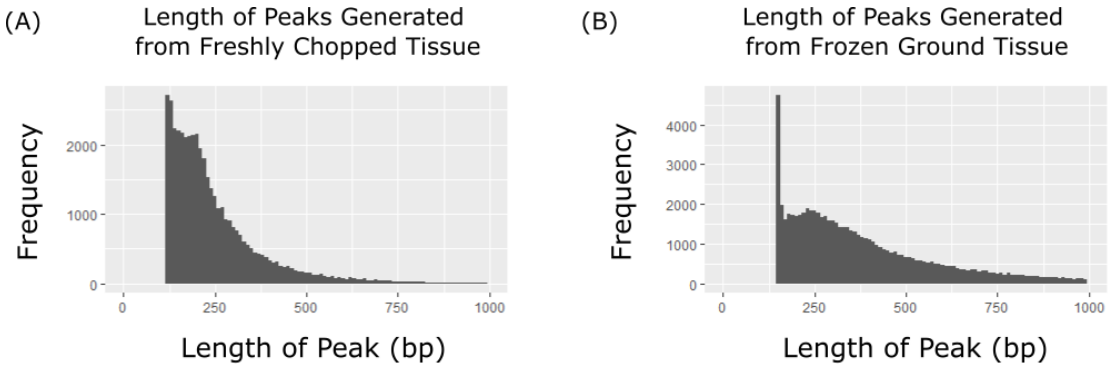


Figure 5.8: There was consistency in the range of peak lengths, but differences in the distribution of peak lengths for libraries generated by both sample preparation methods. Distribution of peak length for (A) libraries generated from freshly chopped rice seedling cotyledons and (B) libraries generated from flash frozen and ground rice seedling cotyledons. Bin width was 10bp.

Whilst the range of peak lengths predicted were the same for both sample preparation methods, the distributions differed. Peaks predicted from chopped samples were shifted left compared to those for ground samples (Figure 5.8), meaning there were a higher percentage of short peaks under 250bp in chopped samples.

The aim of this chapter was to optimise a protocol for ATAC-seq such that it could be used to investigate the regulatory effect of chloroplast inhibition in the future. Another output of these data however is a more general insight into which regions of the genome in rice cotyledons are most likely to bind transcription factors. This can be used to refine motif enrichment analyses such as that carried out in Chapter 4 (Figure 4.13).

The distance between the mid-point of each peak and the nearest transcription start site was calculated. Plotting these distances showed that open chromatin regions were most frequent immediately around the transcription start site (Figure 5.9).

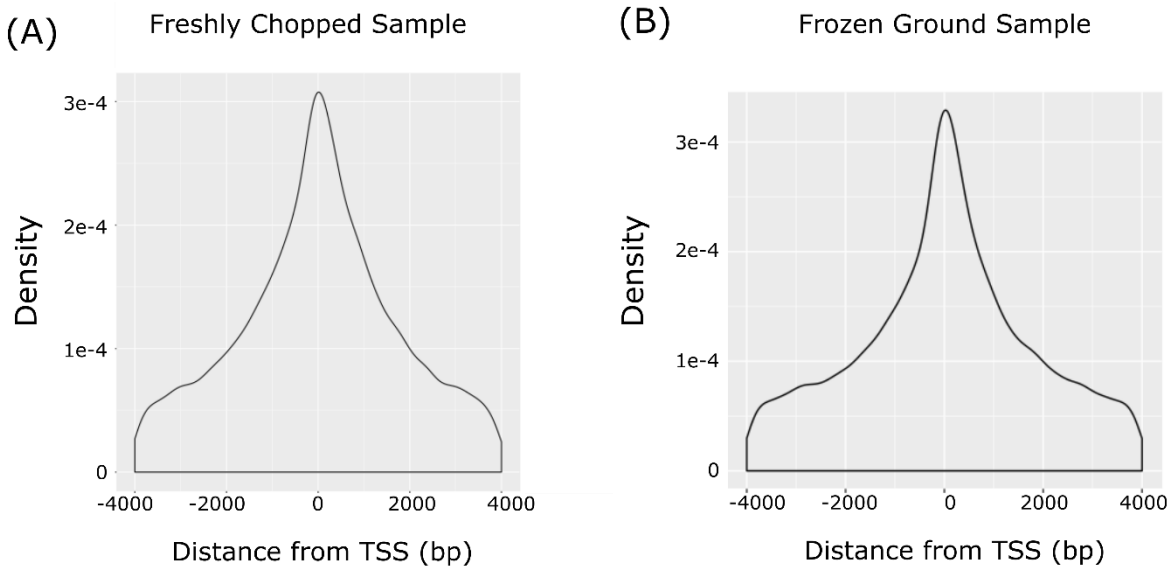


Figure 5.9: There was consistency in the distribution around the transcription start site (TSS) for libraries generated by both sample preparation methods, with most THSs occurring within 2000bp of the promoter region. Distribution of THS (Transposase Hypersensitive sites) around the TSS for all annotated *O. sativa* genes for (A) libraries generated from freshly chopped rice seedling cotyledons and (B) libraries generated from flash frozen and ground rice seedling cotyledons. 0 = TSS.

A high frequency of open regions around the transcription start site (TSS) is expected, as transcription factors are likely to bind there to facilitate initiation of transcription around the TSS. Consistent with this, 67% of peaks were found within 4000bp upstream for chopped, and 59% for ground.

The distribution of peaks was next analysed in more detail, by identifying where peaks occurred within the gene body (Figure 5.10). 29166 (68%) of loci from chopped samples and 41817 (60%) of loci from ground samples were associated with genes.

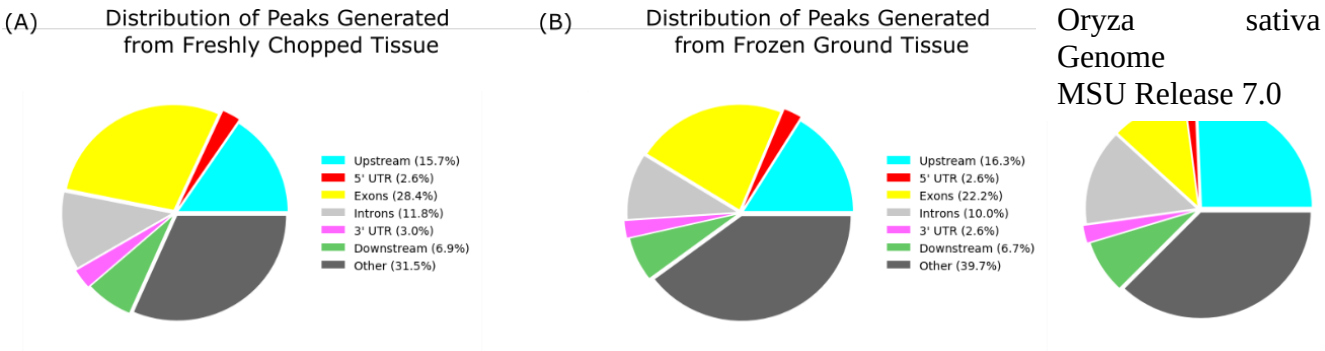


Figure 5.10: Most THSs were found within the intergenic region, the exons and 2000bp upstream of the transcription start site. Distribution of THSs in gene body generated by PAVIS for (A) libraries generated from freshly chopped rice seedling cotyledons and (B) libraries generated from flash frozen and ground rice seedling cotyledons. Upstream defined as 2000bp upstream of the transcription start site. Downstream defined as 1000bp downstream of the stop codon.

The distribution of peaks within the gene body was generally consistent between both sample preparation methods (Figure 5.10). More peaks were found to be intergenic in ground samples (40%) than in chopped samples (32%). Over 20% of the open chromatin regions were associated with exons in both samples.

5.3.6 Investigating Differences between Two Peak Calling Software

The open chromatin regions predicted with Macs2 (Zhang, 2008) were compared to those predicted with another peak calling software, HOMER (Heinz et al., 2010). Both pipelines predict open chromatin regions in a generally similar manner, by identifying where more cuts occur than would be expected by chance. However, there are subtle differences between the two. For example, under standard settings, HOMER tends to find peaks of a fixed width whilst Macs2 does not. 42733 and 81031 peaks were called by Masc2 and HOMER respectively for chopped samples. 23536 (55%) of these peaks overlapped between the two software. 69524 peaks were called by Masc2, and 27937 peaks were called by HOMER for ground samples. 23668 (85%) of these peaks overlapped between the two software. The overlap between the two sample preparation methods was around 50% for both peak calling outputs, with 23415 (55%) overlapping between chop and ground samples for Macs 2, and 14033 (50%) overlapping between the two for HOMER.

The SPOT score, an indicator of the signal to noise ratio, for the chopped sample increased from 0.11 with Macs2 to 0.19 with HOMER. However, the opposite was true for the peaks called from ground sample, with a decrease from 0.29 with Macs2 to 0.25 with HOMER. The length of the peaks for each sample were plotted (Figure 5.11)

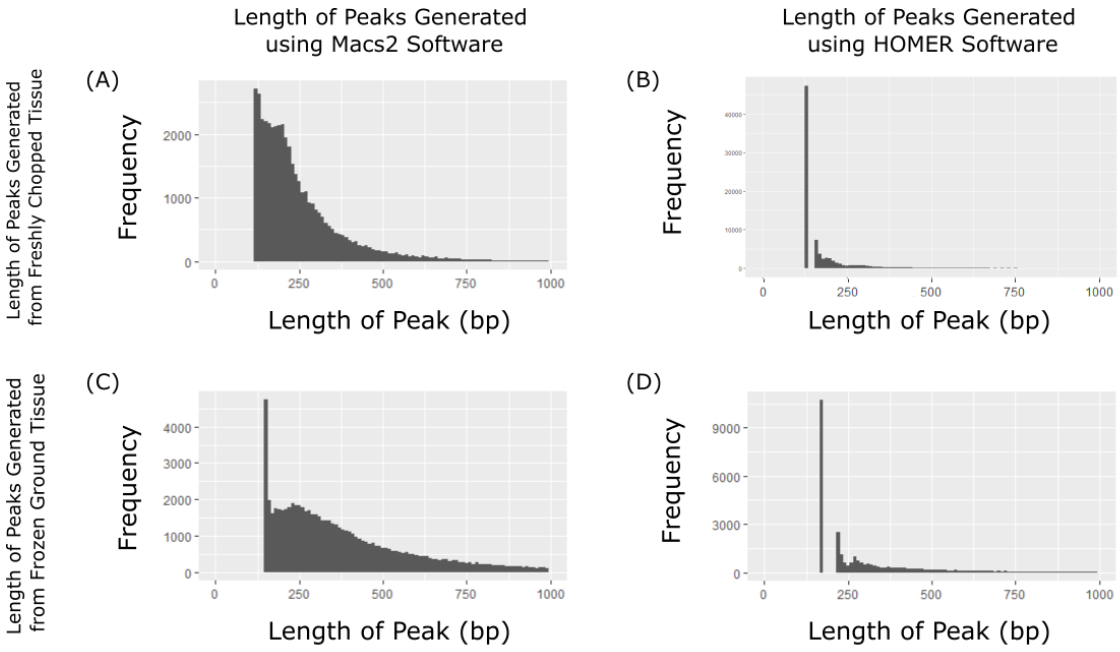


Figure 5.11: There was a large difference in both the range and distribution of peak lengths predicted by Macs2 (A and C) and HOMER (B and D). Distribution of peak length for (A and B) libraries generated from freshly chopped rice seedling cotyledons and (C and D) libraries generated from flash frozen and ground rice seedling cotyledons. Bin width was 10bp.

Fitting with its described mechanism of selecting a fixed width for peaks, HOMER had a high frequency of peaks that were less than 200bp long. For this reason, HOMER would be best used to focus on a region of a promoter for example, rather than detecting larger accessible regions, where it would be best to employ Macs2.

The distribution of the peaks generated by both HOMER and Macs2 around the transcription start site was compared (Figure 5.12)

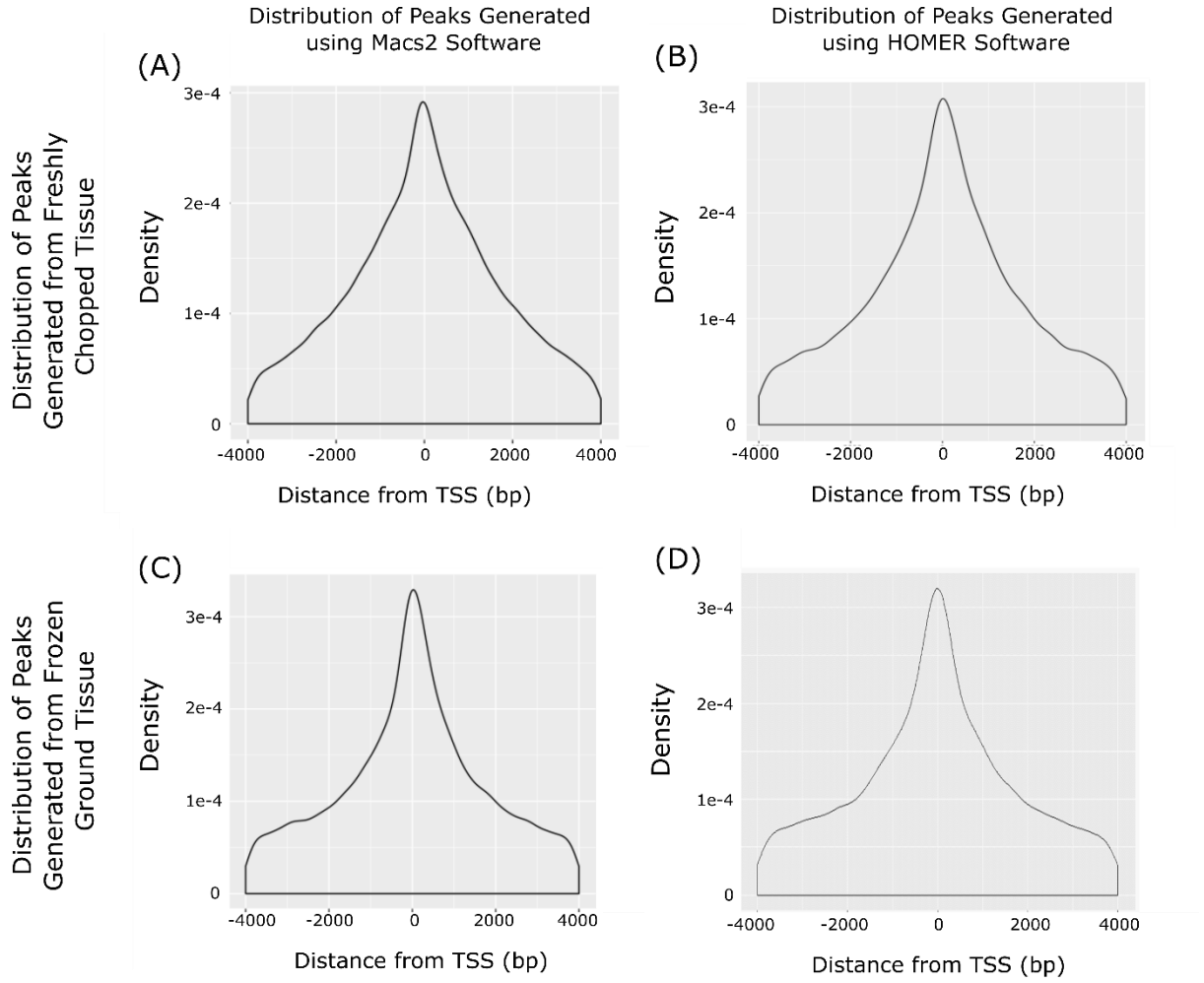


Figure 5.12: There was consistency in the distribution around the transcription start site (TSS) for peaks predicted by both Macs2 (A and C) and HOMER (B and D), with most THSs occurring within 2000bp of the promoter region. Distribution of THS (Transposase Hypersensitive sites) around the TSS for all annotated *O. sativa* genes for (A and B) libraries generated from freshly chopped rice seedling cotyledons and (C and D) libraries generated from flash frozen and ground rice seedling cotyledons. 0 = TSS.

The distribution of peaks around the TSS was very similar between the two methods. This similarity in distribution was also apparent at wider scales. For example, peaks predicted by Macs2 and

HOMER had similar distributions throughout the gene body. However, for the freshly chopped sample, Macs2 had ~10% more intergenic peaks, whereas, these ~10% peaks had shifted to the introns in the predicted distribution for HOMER (Figure 5.13).

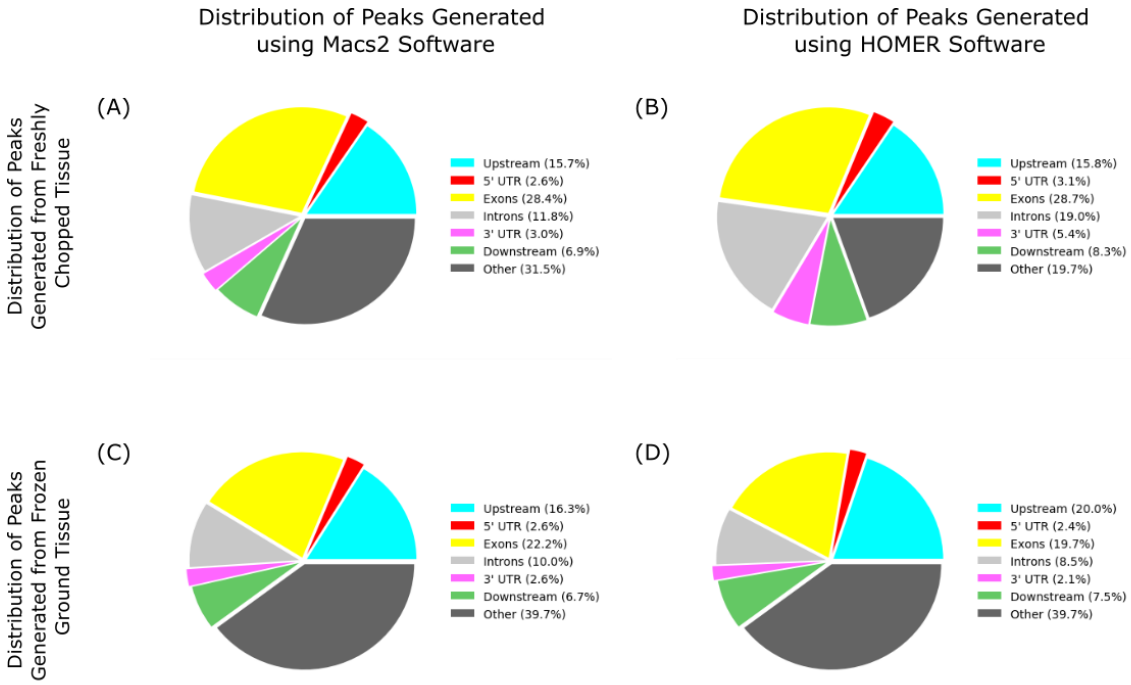


Figure 5.13: Most THSs predicted by Macs2 (A and C) and HOMER (B and D) were found within the intergenic region, the exons and 2000bp upstream of the transcription start site. Distribution of THSs in gene body generated by PAVIS for (A and B) libraries generated from freshly chopped rice seedling cotyledons and (C and D) libraries generated from flash frozen and ground rice seedling cotyledons. Upstream defined as 2000bp upstream of the transcription start site. Downstream defined as 1000bp downstream of the stop codon.

The overall distributions of called peaks were similar for the two software. Next, a more granular analysis was undertaken, by plotting predicted peaks within the chloroplast responsive gene *Carbonic Anhydrase 1 (CA1)* were plotted (Figure 5.14).

A histogram of cut sites across the gene was plotted to allow visualisation of peaks (Figure 5.14A). Consistent with HOMER's mechanism of choosing a short fixed width for peaks, there are more short peaks predicted for both sample preparation methods as compared with Macs2. Enriched motifs in chloroplast responsive genes in rice were identified in Chapter 4. One method for motif enrichment analysis is to use FIMO which identifies individual motif occurrences that can then be totalled and normalised against background. The loci for overlapping peaks between Macs2 and HOMER were searched for enriched motifs using the FIMO output from Chapter 4 (Table 4.7). There was one region where peaks overlapped for all analyses, from 25695931bp to 25697431bp that was particularly enriched in AP2EREBP transcription factors, with five occurrences of the binding motif for transcription ERF1. Another region from 25701928bp to 25702429bp, had peaks overlapping from both ground sample analyses and peaks predicted by HOMER for the chopped sample. This region was particularly enriched for BBRBP motifs with two occurrences of BPC1 and BPC6 transcription factor binding sites.

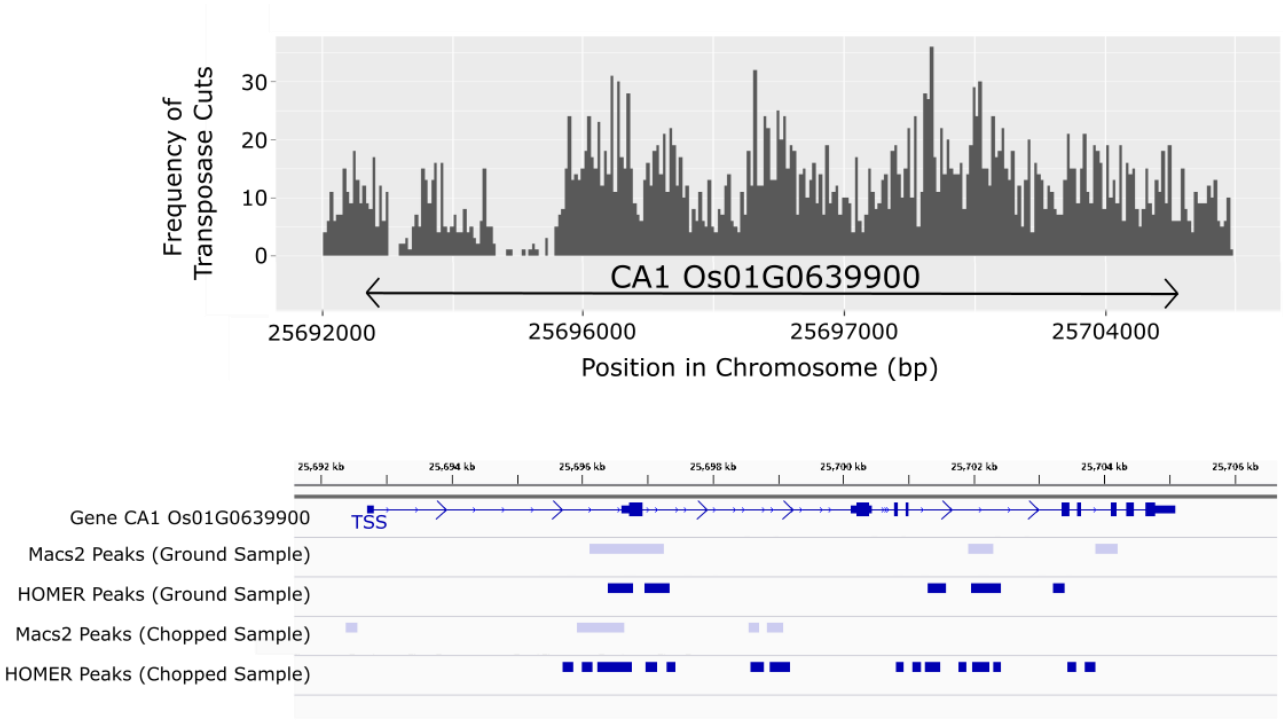


Figure 5.14: Several peaks have been predicted within the gene body of *CA1*. (A) A histogram showing the frequency of cuts along the gene body of *CA1*, with bins of 50bp. (B) A visualisation of peaks predicted by Macs2 and HOMER in the *CA*1 gene body in IGV genome viewer, along with the gene body of *CA1* from the *Oryza sativa* IRGSP 7.0 Genome release. Genome shown as UTRs (short blue boxes), exons (tall blue boxes), and introns (blue lines) for all possible splice variants.

5.3.7 Analysis following the Pipeline outlined in Maher et al (2018)

To this point, with the exception of the SPOT score, all quality checks were passed. As earlier comparisons to data generated by Maher et al (2018) showed similarities, the data generated in this chapter were analysed using their pipeline. The same software, bowtie2, was used to generate the initial alignment and then HOMER was used to call peaks. The main difference between this pipeline and the analysis detailed above is the filtering steps before analysis. After alignment, reads were discarded with a MapQ score less than 2, contrasting to the much more conservative 30 chosen above. Additionally, duplicate reads were not removed from the analysis.

71222 peaks were called for the chopped samples, and 45343 were called for ground, with 25866 peaks overlapping. The SPOT score for the ground sample increased to 0.49 using this pipeline, and more importantly, the chopped sample increased to 0.28. This suggests that the low SPOT scores attained earlier were due to the differences in analysis pipeline and not data quality. This data and the data from Maher et al (2018) had many duplicates reads (Figure 5.4). Their removal appears to cause a decrease in SPOT score. Whilst we cannot be completely certain that these duplicates are not genuine biological signal, it is more likely that they are a result of PCR amplification. This means

that the duplicates retained in Maher et al (2018) are causing the 'signal', in the signal to noise ratio, to be artificially amplified.

Whilst there was a large effect on SPOT score, the distributions of the peaks predicted in this pipeline compared to earlier analysis is unchanged (Figure 5.15). The distribution around the TSS is consistent with the Macs2 and HOMER analyses using more conservative filters (Figure 5.12).

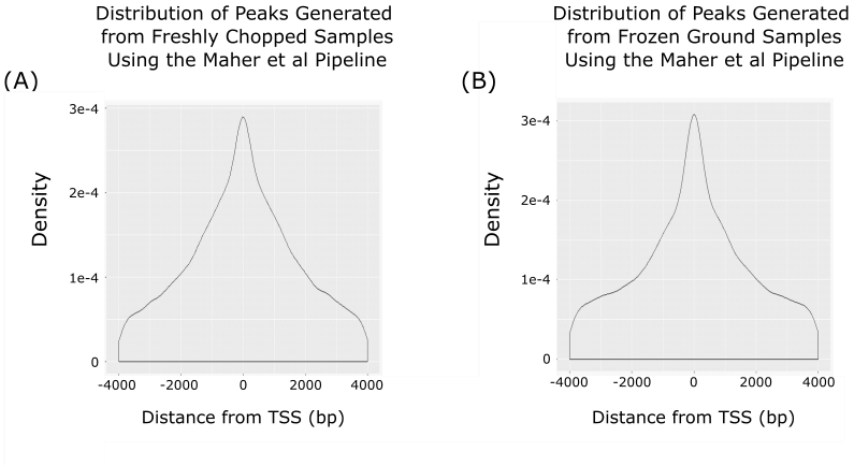


Figure 5.15: There was consistency in the distribution of THSs predicted using the Maher et al pipeline around the transcription start site (TSS) for libraries generated by both sample preparation methods, with most THSs occurring within 2000bp of the promoter region. Distribution of THS (Transposase Hypersensitive sites) around the TSS for all annotated *O. sativa* genes for (A) libraries generated from freshly chopped rice seedling cotyledons and (B) libraries generated from flash frozen and ground rice seedling cotyledons. 0 = TSS.

The general distribution of peaks across the gene body was also not changed, with over 18% of the open chromatin regions associated with exons in both samples (Figure 5.16). This is consistent with earlier results (Figure 5.13), and contrasts to the results in Maher et al (2018), where only 6% of peaks from rice roots were associated with exons. This suggests this distribution represents biological reality and is not an artifact of the data analysis pipeline.

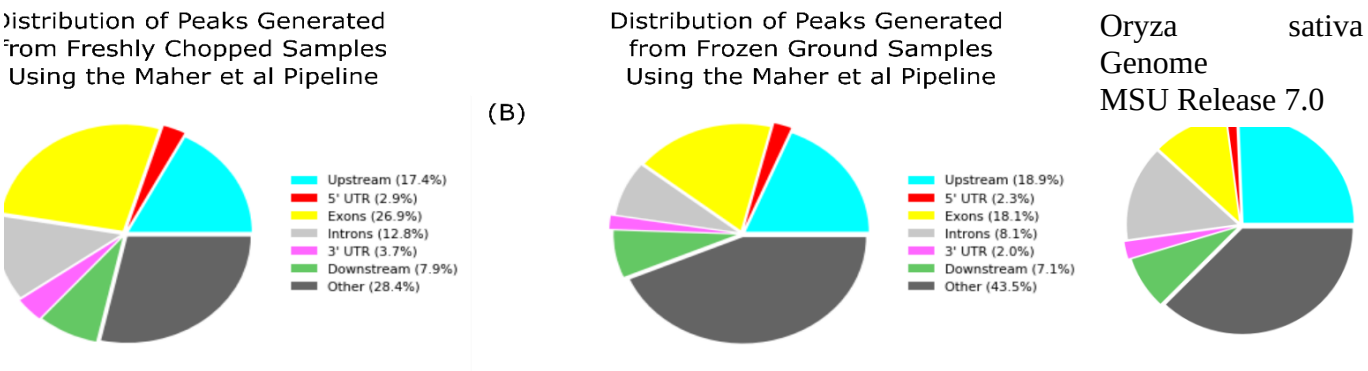


Figure 5.16: Most THSs predicted using the Maher et al pipeline were found within the intergenic region, the exons and 2000bp upstream of the transcription start site. Distribution of THSs in gene

body generated by PAVIS for (A) libraries generated from freshly chopped rice seedling cotyledons and (B) libraries generated from flash frozen and ground rice seedling cotyledons. Upstream defined as 2000bp upstream of the transcription start site. Downstream defined as 1000bp downstream of the stop codon.

The distribution of peaks has not changed at a wider scale, but a zoomed in view shows subtle differences within the chloroplast responsive gene *CA1* (Figure 5.17).

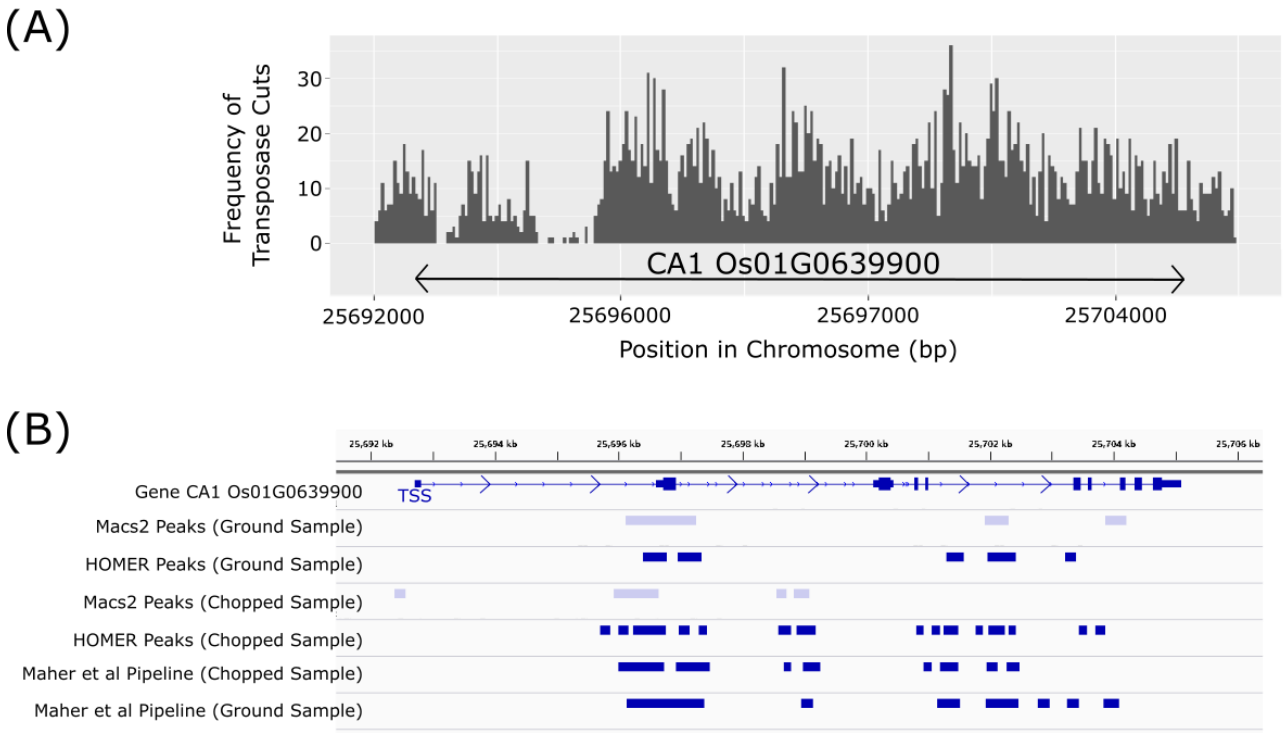


Figure 5.17: Several peaks were predicted within the gene body of *CA1*. (A) A histogram showing the frequency of cuts along the gene body of *CA1*, with bins of 50bp. (B) Visualisation of peaks predicted by Macs2 and HOMER in the *CA1* gene body in IGV genome viewer, along with the gene body of *CA1* from the *Oryza sativa* IRGSP 7.0 Genome release. Genome shown as UTRs (short blue boxes), exons (tall blue boxes), and introns (blue lines) for all possible splice variants.

A histogram of cut sites across the gene was plotted to allow visualisation of peaks in their literal form (Figure 5.17A). Peaks predicted by the Maher et al pipeline spanned more of the gene body, but often overlapped with those predicted from the more conservatively filtered data in earlier analyses (Figure 5.17B).

These results suggest that both sets of data gathered in the chapter are comparable to peer reviewed data, with both chopped and ground samples having attained passable SPOT scores using this pipeline.

5.3.8 Predicted transcription factor footprints in the genome of rice seedlings

Peak calling looks for accessibility at a larger scale, while footprinting searches for smaller transcription factor binding sites. Footprints were predicted used Wellington (Piper et al., 2013), based on peaks predicted by Macs2. 578 and 621 footprints were found in the chopped sample and ground sample respectively, of which 359 overlapped. The average footprint length was similar between the two, at 22bp long for the chopped sample and 23bp long for the ground sample.

Of the 359 footprints which overlapped in chopped and ground samples, 12 footprints were found either in the coding region or within 4000bp upstream of genes encoding retrotransposon Ty3/Gypsy proteins and 8 footprints were found either in the coding region or within 4000bp upstream of genes encoding ATP synthase. Additionally, a few of the footprints found in both chopped and ground samples occurred in genes whose function is associated with photosynthesis.

For example, *Ycf3* encodes a putative Photosystem I assembly protein that facilitates assembly of reaction centre subunits (Nellaepalli et al., 2018). A transcription factor binding footprint was predicted just before the transcription start site of this gene in both chopped and ground samples (Figure 5.18A). *CYB6* is a gene which putatively encodes Cytochrome b6, a subunit of the Cytochrome b6-f complex. This complex facilitates transfer of electrons from plastoquinol to plastocyanin at the thylakoid membrane. A transcription factor binding footprint was predicted in the gene body of *CYB6* in both chopped and ground samples (Figure 5.18B). Finally, *rps16* encodes a putative component of the 30S chloroplast ribosome. The chloroplast ribosome is responsible for the translation of important photosynthetic complexes, including subunits of PSI and PSII. A transcription factor binding footprint was predicted in the gene body of *rps16* in both chopped and ground samples (Figure 5.18C).

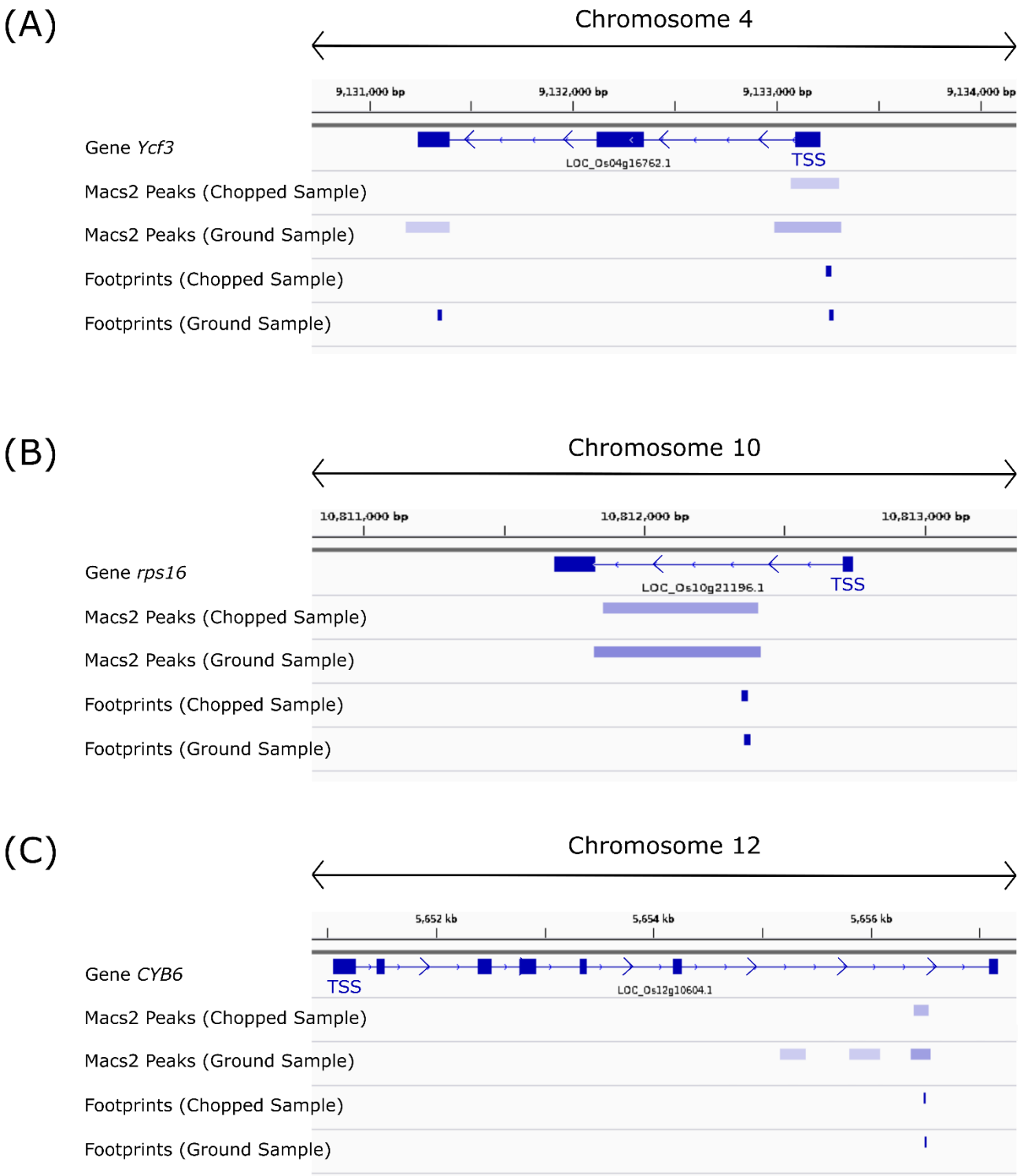


Figure 5.18: Visualisation of peaks predicted by Macs2 and footprints predicted by Wellington in or before photosynthesis associated genes. Visualised in IGV genome viewer, along with the gene body of (A) *Ycf3*, (B) *CYB6* and (C) *rps16* from the *Oryza sativa* IRGSP 7.0 Genome release.

5.4 Discussion

5.4.1 Chopping versus grinding prior to ATAC-seq

ATAC-seq is yet to have one established gold-standard methodology or analysis pipeline. The first step of the ATAC-seq protocol is to extract intact nuclei from samples, which are then treated with the transposase enzyme (Buenrostro et al., 2013). Nuclei can be extracted from tissue prepared by either chopping (Lu et al., 2017), or flash freezing in frozen nitrogen and grinding (Maher et al., 2018; Wilkins et al., 2016). The work reported in this chapter aimed to identify if these two sample preparation methods caused any differences in output. The quality of reads from each sample were equivalent, with high confidence base assignments and read alignments attained for both with no GC bias observed. The distribution of insert size differed from that of many animal ATAC-seq studies where clear peaks can be observed at 200bp and 400bp representing cuts around nucleosomes (Adams et al., 2019; Davie et al., 2015). However, these results are replicated in rice ATAC-seq data from Wilkins et al (2016), which suggests this is the expected distribution for libraries generated from rice, and possibly plants in general.

Assessable regions were called with two different peak calling programs. Peaks called with Macs2, which does not set a fixed width for peaks, differed slightly between sample preparation methods. Peaks predicted from reads from the chopped sample were shorter than those predicted with ground samples. This implies that DNA from nuclei extracted from chopped tissue is more accessible than ground tissue. It is possible that an increased level of nuclei clumping within ground tissue could cause such a phenomenon. Another difference between the two methods was the SPOT score, a measure of how much the peaks represent biological signal above the background rate. Ground samples had a higher SPOT score. When this is taken into account along with the other advantages, flash freezing and grinding appears to be the best method for sample preparation (Table 5.1). Advantages of grinding include speed, which allows for many sample to be processed at once, and more importantly ensures the transcription factor binding profile is not degraded during sampling.

	Freshly Chopped Tissue	Frozen Ground Tissue
Sample Preparation Time	5-10 minutes a sample. Impacts the ability to scale up the experiment and transcription factor binding could be affected	<1 minute a sample. Allows multiple treatment and biological replicates without risking a change in transcription factor binding
Quality of Nuclei Preparation	Great, with many isolated intact nuclei	Good, some debris and clumping observed but generally isolated intact nuclei
Quality of Sequencing Library	Good, comparable to peer reviewed data although a large number of duplicated reads	Good, comparable to peer reviewed data although a large number of duplicated reads
Quality of Alignment	Great, high quality alignment and no GC bias	Great, high quality alignment and no GC bias
Distribution of Peaks	As expected, most peaks predicted within 2000bp of the transcription start site	As expected, most peaks predicted within 2000bp of the transcription start site
SPOT score of peaks	0.28, acceptable signal to noise ratio	0.49, good signal to noise ratio

Table 5.1: Summary of the differences between ATAC-seq output for chopped and ground rice cotyledons.

5.4.2 Optimisation of Data Analysis Pipeline

ATAC-seq involves treating nuclei with a transposase enzyme that preferentially cuts open chromatin. By sequencing the libraries produced from these treated nuclei, the position of cuts can be determined by aligning the read to a reference genome. Two software packages, HOMER (Heinz et al., 2010) and MACS2 (Zhang, 2008), identify regions of the genome where there are more cut sites than expected by chance. Both output predictions of accessible chromosome regions, also referred to as peaks. HOMER predicted a greater number of shorter peaks than Macs2, because it employs a fixed width algorithm. This is because HOMER concentrates on identifying the precise location of DNA-protein contact. Both software predicted peaks of a similar distribution throughout the genome. Low SPOT scores were attained for both methods, and therefore data were analysed using the data analysis pipeline described in Maher et al (2018). This pipeline attained much higher SPOT scores,

likely due to reads not being deduplicated before alignment and peak calling. A potential future methodological improvement would be to reduce PCR cycles during library amplification, in the hope of reducing duplicate reads, as it is a waste of money and resources to sequence so many duplicates.

Overall, Macs2 and HOMER are both useful tools, and a future pipeline would likely employ both and focus initially on the overlap between the two. There has been an increased understanding of the potential bias which can be introduced during ATAC-seq since this work was carried out and updated computational methods have become available to mitigate such bias (Reske et al., 2020), which would be employed in a future analysis.

5.4.3 Distribution of open chromatin regions in cotyledons of rice seedlings

The distribution of open chromatin regions was consistent between sample preparation methods, and between peaks called using different software. The highest frequency of open chromatin regions was found immediately surrounding the transcription start site (TSS), with frequency decreasing further upstream. The distribution at a wider scale showed most peaks were found within the intergenic regions, but of those peaks that associated with a gene most were found within the exons and 2000bp upstream of the TSS. Across all analyses, 18.1% to 28.7% of the open chromatin regions were associated with exons. Previous ATAC-seq data from rice roots found only 6% associated with exons (Maher et al., 2018). This does fit however with DNase-Seq results from several grass species, where 20% of DNase hypersensitive sites were predicted to be in exons in foxtail millet, sorghum, maize and purple false brome (Burgess et al., 2019). The promoter region is often the focus in studies attempting to identify transcription factors which bind *cis*-elements. These results suggest that the gene body could be an important inclusion in such an analysis.

5.4.4 Transcription factor binding sites predicted in the genome of rice seedling cotyledons

By analysing cut sites at a finer resolution, short transcription factor binding sites around 30bp long can be identified. Only a small number of footprints were predicted in this analysis, which suggests that a greater depth of sequencing would be needed to obtain a fuller picture of transcription factor binding across the rice genome. Of the footprints identified, over 10 were found either within or before genes encoding retrotransposon Ty3/Gypsy proteins which are the most prolific repeat retrotransposons in the genomes of the Oryza genus (Zuccolo et al., 2008). This could potentially be an artefact due to cutting bias, it is also possible that retrotransposons are so highly represented as they insert themselves into open chromatin regions. Despite the low numbers predicted, several footprints were found within photosynthesis associated genes including a putative Photosystem I assembly protein, Cytochrome b6 and a component of the chloroplast ribosome. These footprints could be used to inform studies looking to identify *cis*-regulatory elements within these genes and give further insight into photosynthesis associated genetic regulation in rice. Although only a small number of footprints were called, the THS identified earlier allow predictions of where transcription factors are likely to bind. This is particularly beneficial at a genome wide scale, where the data can be used to refine motif enrichment analyses such as that carried out in Chapter 4 (Figure 4.13).

5.4.5 Future Application of ATAC-seq

A trial of ATAC-seq has been completed and both library preparation and the data analysis pipeline optimised. Time and resource limitations prevented this work from being utilised on treated samples to investigate the primary hypotheses being reported in the thesis. However, this analysis is now primed to take place in the future and the work informed other lab members' work. A future ATAC-seq analysis will investigate any differences in open chromatin regions and transcription factor footprints on chloroplast inhibited seedlings versus controls. This will show if any C_4 orthologues have *cis*-elements whose binding is impacted by the chloroplast, informing further work towards the ultimate aim of identifying the regulatory mechanisms through which plastid signals impact gene expression. The dataset will also be of more general relevance as it can act as an atlas of transcription factor binding sites involved in chloroplast retrograde signalling.

6 DISCUSSION

Plastid to nucleus signals have been shown to regulate core photosynthetic genes, in both C_3 and C_4 species. However, the regulatory mechanisms of this signalling are not well understood. A genome wide analysis was carried out to investigate changes in transcript abundance genes in response to chloroplasts inhibition in *A. thaliana*, a model eudicotyledon, and rice, a model monocotyledon. In this chapter the effect of chloroplast inhibitors on PSII activity in each species, and the quantitative and qualitative effect of chloroplast inhibition on the transcriptome will be discussed. Chloroplast responsive genes were searched for enriched motifs, to identify putative transcription factors which could act as a chloroplast derived signal modifying nuclear gene expression. A literature search was undertaken on the most promising candidate transcription factors, to search for a biological link between what is currently known about their function and chloroplast to nucleus signalling. Finally, the future directions are outlined, detailing the methods which could be used to investigate if these transcription factors are linked to a chloroplast retrograde signalling pathway.

6.1 Shared Response to Chloroplast inhibition in *A. thaliana* and Rice

Both *A. thaliana* and rice were treated with chloroplast inhibitors to investigate which genes responded. This could be in response to a signal which transmits the chloroplast's damaged status, or a response caused by the absence of a signal which can no longer be produced or conveyed due to chloroplast inhibition. Lincomycin, an inhibitor or plastid translation, and norflurazon, an inhibitor of carotenoid synthesis that causes photooxidative damage to the chloroplast in the light, were incorporated into the media on which seedlings were grown. Both inhibitors caused bleaching in illuminated seedlings and decreased maximum quantum yield of PSII. These results are corroborated by the literature. *A. thaliana* cultured cells treated with norflurazon had a significantly reduced maximum quantum yield when cells were exposed to light compared to controls, however, no significance difference was observed between treated and control cells in the dark (Doyle et al., 2010). Lincomycin treatment also causes a significant decrease in maximum quantum yield; *A. thaliana* leaves infiltrated with 1mM lincomycin showed a 60% decrease in maximum quantum yield (Tarantino et al., 1999) and three-week-old rice plants which had been treated with 50μM norflurazon showed a 37% decrease in maximum quantum yield compared to controls (Park & Jung, 2017).

However, differences were observed in the overall transcriptional response to the two chloroplast inhibitors. Norflurazon affected the transcript abundance of an additional 2500 genes compared to lincomycin in *A. thaliana* and rice. Although lincomycin affected the transcript abundance of fewer genes, ~70% of those genes were also affected by norflurazon in both species. This observation was consistent with the observed overlap in gene function; ~70% of gene ontology terms enriched in genes responsive to lincomycin treatment in rice, were also enriched in response to norflurazon treatment. This increased to 80% in *A. thaliana*. The similarity in which genes were affected and by how much between *A. thaliana* and rice suggests that inter-species comparisons of chloroplast inhibition are reasonable and warranted. One difference in the response to inhibitors between the two species was

emphasised in the principal component analysis, a measure of sample similarity. Dark-grown samples grouped together, whether treated with lincomycin of norflurazon, with one exception; norflurazon treatment in dark-grown rice seedlings caused 10% of the variance observed between norflurazon treated seedlings and controls. Norflurazon has been shown to affect another dark-grown monocotyledon, decreasing lipid content in barley (Di Baccio et al., 2002). The overlap in the effect on PSII, the genes affected, and gene function suggest both inhibitors are comparable.

6.2 Genes Responsive to Chloroplast Perturbation in *A. thaliana* and rice

Of the 21319 genes whose transcripts were detected in the *A. thaliana* samples, 743 genes showed significantly different transcript abundance when illuminated samples were treated with norflurazon or lincomycin. Of the 33064 genes whose transcripts were detected in the rice samples, 935 genes showed significantly different transcript abundance when illuminated samples were treated with either inhibitor. A similar magnitude in the transcriptomic response to chloroplast perturbation was observed, with 3% of detected transcripts affected in both *A. thaliana* and rice. Transcript abundance could have changed in response to a signal transmitting the chloroplast's damaged status or in response to the absence of a signal which previously conveyed chloroplast development was proceeding successfully. Whilst microarray data are available for *A. thaliana* treated with norflurazon (Moulin et al., 2008), the work presented here provides the first genome wide transcriptome analysis of chloroplast inhibition in rice, and to my knowledge, the first in a monocotyledon. Whilst the response to chloroplast inhibition was consistent between the two species, changes in transcript abundance in response to light varied dramatically. In *A. thaliana*, 26% of the detected transcripts had significantly different abundance between dark-grown and 24-hour light exposed seedlings, whilst in rice this was only 5.5% of the detected transcripts. This could be due to *A. thaliana* having few direct orthologues in rice (Wang et al., 2017) and the low commonality in binding motifs between the two species (Cserhati, 2015).

For both *A. thaliana* and rice, the most enriched gene ontology terms for genes responsive to the chloroplast were photosynthesis associated terms. Both species showed a strong response from genes associated with the 'Generation of precursor metabolites and energy', which includes photosynthesis, oxidation and NADPH regeneration. Reactive oxygen species (ROS) have been implicated as a chloroplast retrograde signal in several potential pathways, including the redox potential of the chloroplast transmitted through the malate valve (Heyno et al., 2014). The malate valve operates when the NADPH pool is over-reduced, NADP-MDH oxidises the pool by using the reduction power to convert OAA to malate which is then exported to the cytosol acting as a signal. Genes associated with 'Single-organism metabolic processes' were also overrepresented in chloroplast responsive genes, this can include NADPH metabolism and oxidation-reduction processes. It can additionally include intermediary metabolites such as Mg-proto IX, an intermediary in the tetrapyrrole biosynthesis pathway directly regulated by GUN4 and GUN5 (Martín et al., 2016). *A. thaliana* did show more of a stress response than was observed in rice, with genes associated with 'response to chemical' and 'response to abiotic stimulus' being enriched. However, overall, the inhibitors affected similar cellular processes.

Several genes have been reported to show decreased expression after chloroplast inhibition. A rt-qPCR study on 7-day old *A. thaliana* seedlings grown in 5µM norflurazon showed decreased expression in *LHCB2.1, CA1, GUN4* and *CHLH/GUN5* (Page et al., 2017). Decreased expression of *RbcS* and *CHLH* has been seen in rice after treatment with 50mM (Park & Jung, 2017). These results were mirrored in this work, with the five aforementioned genes showing decreased expression in *A.*

thaliana and rice. However, *HO1* (GUN2) expression was also reported to decrease in response to norflurazon treatment in rice (Park et al., 2017). This response was not observed in either norflurazon or lincomycin treated samples. The *GUN* gene expression patterns were consistent between rice and *A. thaliana*. Taking into account *gun*-like mutants have been found in barley (Batschauer et al., 1986), there is evidence to support that the *GUN* retrograde signalling pathway could be ancestral. Overall, the gene expression analysis shows consistency with previous studies as well as an overlap in the genes affected by each inhibitor.

6.3 The extent to which C_4 orthologues are subject to chloroplast retrograde signalling

Chloroplast derived signals have been shown to regulate C_4 photosynthesis associated genes in the C_4 eudicotyledon *Gynandropsis gynandra* (Burgess et al., 2016) and in the C_4 monocotyledon *Zea Mays* (Tamada et al., 2003), two species that are thought to have last shared a common ancestor over 100 million years ago. Chloroplast perturbations have also been shown to generate a response in C_4 orthologues within C_3 *A. thaliana*. Burgess et al (2016) used rt-qPCR to detect differential transcript abundance in norflurazon and lincomycin treated seedlings. They found seven C_4 orthologues under chloroplast control; *AtCA1, AtPPC2, AtNAD-ME1, AtPPA6, AtRP1, AtNAD-ME2* and *AtTPT*. Five of these were also found to be under chloroplast control in the work completed here; *AtCA1, AtPPC2, AtNAD-ME1, AtPPA6, AtRP1,* on top of an additional three; *AtCA2, AtASP1* and *AtDIC1*. Regulation of one gene can have subsequent effects on other C_4 orthologues, for example, AtNAD-ME1 forms a heterodimer *in vivo* with AtNAD-ME2 (Tronconi et al., 2008), therefore the expression of *AtNAD-ME1* has an impact on the activity of both the AtNAD-ME1 and AtNAD-ME2 protein. This small increase in the number of responsive C_4 orthologues compared to the previous study supports the conclusion that a number of C_4 genes are regulated by chloroplast to nucleus signalling in C_3 species.

An aim of this project was to test if these results could be repeated in another C_3 species, particularly a divergent species such as monocotyledon *O. sativa* (rice). Ten C_4 orthologues were significantly affected by both inhibitors in rice and are therefore likely subject to chloroplast-dependent regulation. However, only three of these C_4 specific orthologues were found to be differentially expressed in *A. thaliana*. This low degree of overlap in response between the two species is likely associated with a number of factors relating to the genetic distance between *A. thaliana* and rice. For example, multiple genome duplications have occurred since their divergence ~140 million years ago (Lee et al., 2013). The C_4 genes analysed in this study are not thought to be directly involved in photosynthesis of rice, however, many are already known, or at least predicted, to be plastid localised. For example, *Phosphoenolpyruvate Carboxylase 4* (*PPC4*) shares 73% amino acid identity with the PPC gene recruited into C_4 photosynthesis in sorghum (Wang et al., 2009). This gene is plastid localised and has been shown to function in ammonium assimilation (Masumoto et al., 2010). Whilst its location means plastid regulation may be expected, the results showed light was also an important regulator of *PPC4*. *Carbonic anhydrase 1* (*CA1*) was the most responsive gene to chloroplast perturbation. *CA1* is predicted to be plastid-localised in rice and loss of a chloroplast transit peptide is responsible for the cytosolic localisation in C_4 species (Tanz et al., 2009; Clayton et al., 2017).

Although many of the genes found to be subject to regulation through chloroplast retrograde signalling were plastid-localised, some genes such as *Pyruvate Orthophosphate Dikinase A* (*PPDKA*) encode proteins that localise to the cytosol (Wang et al., 2020). This shows that the repeated recruitment of genes in C_4 evolution which are subject to chloroplast regulation cannot be simply explained by prior plastid-localisation of gene product. The regulation of C_4 orthologues by the

chloroplast in two divergent C_3 species supports the hypothesis that parallel evolution repeatedly recruits genes into C_4 photosynthesis because these genes are already controlled by the same regulatory network as C_3 photosynthetic genes. Thus it appears that chloroplast retrograde control is an essential component of understanding parallel evolution in the ~60 independent origins of C_4 photosynthesis.

6.4 Regulatory elements enriched in chloroplast responsive genes

The list of chloroplast responsive genes generated in this analysis enabled a search for putative transcription factors likely involved in chloroplast to nucleus signalling. An analysis looking for transcription factors with strong affinity binding found six main families of transcription factors had enriched motif binding sites with chloroplast responsive genes in *A. thaliana*: bZIP, bHLH, TCP, MYB, MYB-related, and AP2/EREBP. Five of these transcription factor families were also predicted to bind to enriched motifs in rice; bZIP, bHLH, TCP, MYB, MYB-related. An analysis which prioritised the frequency of a binding motif over the likelihood of binding found both species were enriched for bZIP binding motifs containing the G-box (CACGTG). The G-box is an important *cis*-element which controls the regulation of *CHLH* (aka. *GUN5*) expression (Kobayashi et al., 2012) as well as many light-responsive genes. It has been previously linked to chloroplast retrograde signalling and was found to be enriched in the promoters of 329 genes which were upregulated in both *gun1* and *gun5* seedlings, appearing 944 times (Koussevitzky et al., 2007).

There was not only overlap at the level of families, but also multiple instances of a binding motif for a specific transcription factor being enriched in both species. The most notable overlap between the two species was that seven of the ten most enriched motifs identified in the analysis focused on overall motif frequency were shared between A. *thaliana* and rice. These seven enriched binding motifs represented five transcription factors; ABF2, AREB3, ABI5, bZIP16 and GBF3. These are all promising candidates for future characterisation, hopefully being involved in an ancestral chloroplast to nucleus regulatory pathway. Less overlap was observed between enriched motifs in C_4 orthologues between the two species. This is unsurprising considering many transcription factors involved in photosynthetic regulation in *A. thaliana* have no direct orthologues in rice (Wang et al., 2017). However, the large overlap in the transcription factors families predicted to bind chloroplast responsive genes between the two species could mean the regulation is ancestral.

In *A. thaliana*, the C2C2GATA transcription factors, thought to be involved in light and nitrate-dependent regulation, had enriched binding sites in plastid regulated C_4 orthologues. The two most enriched binding motifs GATA15 and GATA16, are thought to work redundantly with GATA21 and GATA22 (also referred to as GNC and GNL) to control greening (Ranftl et al., 2016). GATA21 and GATA22 have been shown to control greening through regulation of the tetrapyrrole biosynthesis pathway, and *gata21 gata22* mutants show decreased expression of chloroplast regulated genes such as *HEMA1* and *GUN4* (Hudson et al., 2011). GATA15 and GATA16 may work through a similar mechanism, regulating upstream of the tetrapyrrole retrograde signalling pathway. Of the four subfamilies of which the C2C2GATA transcription factors are classified, three are shared between *A. thaliana* and rice which are theorised to have evolved before their most recent common ancestor (Reyes et al., 2004). The binding motif of one transcription factor was explicitly found to be enriched in both species, OBP3, a nuclear localised transcription factor which modulates phytochrome and cryptochrome signalling in *A. thaliana* (Ward et al., 2005). OBP3's nuclear localisation means it could function as a late component in a theoretical chloroplast to nucleus signalling pathway which controls C_4 genes in a wide array of species.

6.5 Putative Transcription Factors Selected for Further Investigation

Factors to consider when judging the potential of a candidate transcription factor are level of motif enrichment, expression levels across treatments and current knowledge of function, particularly how this relates to chloroplast retrograde signalling. Four such transcription factors were highlighted based on these categories in the first motif enrichment analysis, and mutants of these transcription factors were ordered (Figure 6.1). GBF3 binds to motifs which were enriched 4.8 times above background in *A. thaliana*, and to a lesser extent enriched in rice, occurring 2.8 times more than in background. Its expression responded to light, with illuminated seedlings showing a decrease in expression compared to dark-grown seedlings. GBF3 has been shown to interact with GLK1 (Tamai et al., 2002), an interaction which could be part of a chloroplast to nucleus signalling pathway. Transcription factor GLK1 has been proposed to have a role in retrograde signalling due to its expression being correlated with that of known chloroplast-responsive photosynthesis associated genes, and GUN1 appears to downregulate *GLK1* in response to plastid damage (Kakizaki et al., 2009) through a ubiquitin-proteasome system (Tokumaru et al., 2017).

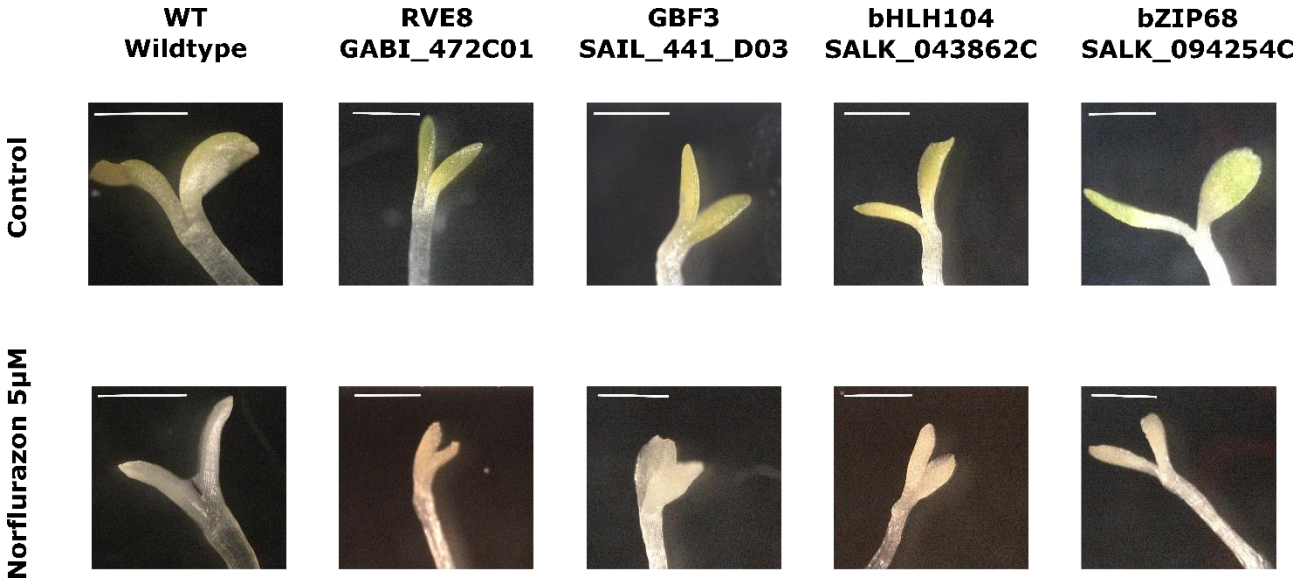


Figure 6.1: tDNA insertion mutants of four candidate transcription factors implicated in chloroplast retrograde signalling. Seedlings grown for 6-days dark and 24 hours 150 μmol m^{-2} s^{-1} white light on ½ MS media containing either 5μM norflurazon or 0.001% (v/v) ethanol. Scale bars represent 500μm.

Another candidate transcription factor, RVE8, binds to motifs enriched 4.2 times above background in *A. thaliana* and 2.6 times in rice. RVE8 is light responsive and shows increased transcript abundance in illuminated samples. This MYB-related transcription factor functions in the circadian clock by modulating histone 3 acetylation (Nguyen & Lee, 2016). One facet of this is diurnal regulation of the anthocyanin metabolic pathway (Pérez-García et al., 2015). GUN1 and GUN5 have been shown to regulate the accumulation of anthocyanins in the cell (Richter et al., 2020), and therefore the link between RVE8 and the anthocyanin biosynthesis pathway make it a promising

candidate. Binding motifs of bHLH104, another candidate whose function may relate to chloroplast retrograde signalling, was similarly enriched in *A. thaliana* and rice, occurring 2.9 times and 3.2 times above background respectively. Knockouts of bHLH104 have reduced iron content, whereas, overexpressers accumulate iron (Li et al., 2016). This may suggest a link to heme production, and therefore tetrapyrrole biosynthesis and GUN genes. Finally, bZIP68 binds motifs enriched 4.3 times above background in *A. thaliana*, and 2.6 times in rice. The transcription factor contains a redox-sensitive Cys320 residue and has a proposed role in balancing growth and stress tolerance (Li et al., 2019). Its ability to sense the redox state of the cell make it a promising candidate, as this is thought to be an important component of chloroplast retrograde signalling.

6.6 Future Directions

Genes regulated by chloroplast retrograde signalling were scanned for enriched binding motifs to identify potentially important *cis*-elements and the transcription factors that bind them. Future work could focus on how to experimentally validate these predictions, and how to refine the list through genome wide studies of transcription factor binding in chloroplast inhibited plants.

6.6.1 Looking for a Genome Uncoupled Phenotype in Transcription Factor Knockout Mutants

In wildtype plants, chloroplast inhibition results in a significantly decreased expression of photosynthesis associated genes, such as *LHCB*. However, six *gun* mutants have been identified in *A. thaliana* where *LHCB* expression does not decrease to wildtype levels after chloroplast inhibitors have been applied, implying that retrograde signalling pathways have been disrupted. If a *gun*-like phenotype is found in transcription factor knockouts treated with chloroplast inhibitors, it would confirm that the transcription factor has a role in chloroplast retrograde signalling pathways. A focus could therefore be on those transcription factors whose predicted binding motifs occur frequently in both *A. thaliana* and rice, as the chloroplast to nucleus regulatory networks acting on these genes have an increased likelihood of being ancestral. Experimentally, this work could involve growing transcription factor knockouts for seed collection and genotyping, then growing the collected seeds on either control MS or MS containing the chloroplast inhibitors lincomycin or norflurazon. An rt-qPCR analysis of the cotyledons will show if the genes predicted to bind these transcription factors have become unresponsive to chloroplast inhibition.

6.6.2 Mutating Binding Motifs in the Promoter of Chloroplast Responsive Genes

Another method to investigate if the aforementioned enriched motifs are binding transcription factors involved in chloroplast retrograde signalling pathways is to introduce mutations within the binding motifs. By mutating the motifs found within a chloroplast responsive gene, transcription factor binding would be perturbed. If this transcription factor is responsible for the decrease of gene transcript abundance in response to chloroplast inhibition, this phenotype would be lost in genes with mutated binding motifs and a *gun*-like phenotype would be observed. This method is particularly useful to study bZIP transcription factor binding, as many enriched motifs in both *A. thaliana* and rice are predicted to bind multiple transcription factors. ABI5, AREB3 and ABF2 are an example of transcription factors often predicted to bind at the same site, and it is unclear which one is truly

relevant, or indeed, if only one is relevant. By mutating the 2-3 motif occurrences in genes including *HEME1*, *PORB*, *CA1* and *PPC2* (Figure 6.2), the importance of these binding sites could be studied.

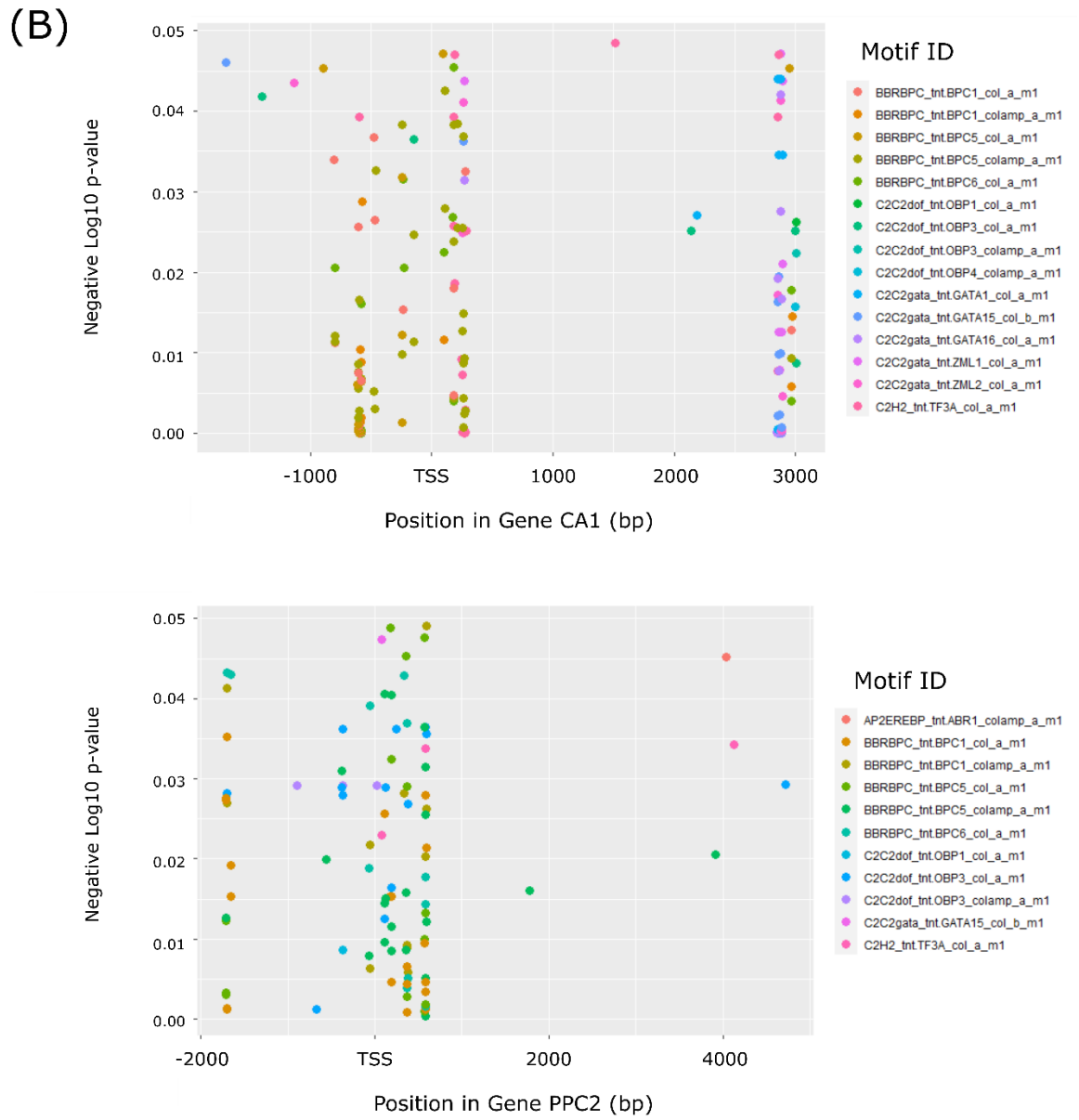


Figure 6.2: The loci of select enriched binding motifs across the promoter and gene body of (A) photosynthesis related genes and (B) C_4 orthologues.

Experimentally, this work will involve generating Golden Gate constructs comprised of the promoter or gene body with motifs deleted and a GUS reporter. *A. thaliana* will be transformed via dipping, and seed selected using the FAST marker (Shimada et al., 2010). Some preliminary results could be attained by analysing the T1 generation, and more definitive results attained from analysis of the T2 generation through a MUG assay to determine expression.

6.6.3 Identifying Transcription Factor Binding in Rice Seedlings with ATAC-seq

A trial of ATAC-seq was completed as part of this work and both library preparation and the data analysis pipeline were optimised. Time and resource limitations prevented this work from being utilised to investigate plastid to nucleus signalling. A future ATAC-seq analysis would investigate

any differences in open chromatin regions and transcription factor footprints on chloroplast inhibited seedlings versus controls. This will show if any C_4 orthologues have *cis*-elements whose binding is impacted by the chloroplast, informing further work towards the ultimate aim of identifying the regulatory mechanisms through which plastid signals impact gene expression. The dataset will also be of more general relevance as it can act as an atlas of transcription factor binding sites involved in chloroplast retrograde signalling. This data set could be combined with the RNA-seq data set to filter for genes with significantly differential transcript abundance, and further filter the list of promising transcription factor candidates.

6.7 An Alternative Approach to Investigate Alternative Splicing in Plastid to Nucleus Regulation

Gene expression can be regulated at multiple levels including transcriptional, translational or the processing, transport and degradation of mRNA. This body of work has focussed on elucidating transcriptional regulation via transcription factors. However, other elements of RNA metabolism are becoming increasingly implicated in chloroplast retrograde studies, including alternative splicing and RNA editing (Zhao et al., 2020). When the plastoquinone pool in *A. thaliana* chloroplasts is reduced, chloroplast retrograde signalling regulates nuclear genes through alternative splicing in response to the changed redox state (Petrillo et al., 2014). An alternative route of investigation could have been to look at genome wide change in splicing in response to chloroplast inhibition. Alternative splicing can be analysed from RNA-seq data (Gulledge et al., 2014), but only those generated from kits that produce libraries of reads from across the whole genome. Libraries in this work were prepared using the QuantSeq 3′ mRNA-Seq Library Prep Kit from Lexogen, which generates libraries of sequences close to the 3’ end of polyadenylated RNA. This makes the kit ideal for gene expression analysis, however, not genome assembly or alternative splicing studies. However, the QuantSeq kit provided a large advantage in reducing costs and library preparation time.

6.8 Concluding Remarks

This project aimed to understand the role of plastid-to-nucleus signalling in regulating genes of the C_4 cycle in the ancestral C_3 state. Genes involved in plastid-to-nucleus signalling were identified through RNA-seq in C_3 *Arabidopsis thaliana* and *Oryza sativa* after chloroplast function was perturbed. This, to my knowledge, constitutes the first genome wide look at transcriptional effects of chloroplast inhibition in a monocotyledon. This work has shown that over half of the C_4 orthologues in these two C_3 species were regulated by both light and the chloroplast. This strongly implies that evolution has re-enforced existing regulatory networks that operate in the C_3 state to control expression of C_4 genes. Genes whose transcript abundance was significantly affected by chloroplast inhibition were scanned for known motif binding sites. *Cis*-elements in the nuclear genome involved in the regulation of plastid-regulated genes were elucidated by analysing enriched binding motifs within genes responsive to chloroplast perturbation. A methodology and data analysis pipeline of ATAC-seq was optimised and the distribution of open chromatin was predicted across the rice genome. A surprising number of open chromatin regions were found within exons. This optimised protocol can be used in future work to generate an atlas of transcription factor binding sites with changed occupancy in response to chloroplast perturbation.

7 BIBLIOGRAPHY

Abdallah, F., Salamini, F., & Leister, D. (2000). A prediction of the size and evolutionary origin of the proteome of chloroplasts of Arabidopsis. *Trends in Plant Science*, *5*(4), 141–142. https://doi.org/10.1016/S1360-1385(00)01574-0

Abu-Jamous, B., & Kelly, S. (2018). Clust: automatic extraction of optimal co-expressed gene clusters from gene expression data. *Genome Biology*, *19*(1), 172. https://doi.org/10.1186/s13059-018-1536-8

Adams, E. J., Karthaus, W. R., Hoover, E., Liu, D., Gruet, A., Zhang, Z., Cho, H., DiLoreto, R., Chhangawala, S., Liu, Y., Watson, P. A., Davicioni, E., Sboner, A., Barbieri, C. E., Bose, R., Leslie, C. S., & Sawyers, C. L. (2019). FOXA1 mutations alter pioneering activity, differentiation and prostate cancer phenotypes. *Nature*, *571*(7765), 408–412. https://doi.org/10.1038/s41586-019-1318-9

Agnolucci, L., Vecchia, F. D., Barbato, R., Tassani, V., Casadoro, G., & Rascio, N. (1996). Amitrole Effects on Chloroplasts of Barley Plants Grown at Different Temperatures. *Journal of Plant Physiology*, *147*(5), 493–502. https://doi.org/10.1016/S0176-1617(96)80037-X

Allen, J. F. (2015). Why chloroplasts and mitochondria retain their own genomes and genetic systems: Colocation for redox regulation of gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(33), 10231–10238. https://doi.org/10.1073/pnas.1500012112

Anders, S., Pyl, P. T., & Huber, W. (2015). HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics*, *31*(2), 166–169. https://doi.org/10.1093/bioinformatics/btu638

Argüello-Astorga, G., & Herrera-Estrella, L. (1998). EVOLUTION OF LIGHT-REGULATED PLANT PROMOTERS. *Annual Review of Plant Physiology and Plant Molecular Biology*, *49*(1), 525–555. https://doi.org/10.1146/annurev.arplant.49.1.525

Aubry, S., Kelly, S., Kümpers, B. M. C., Smith-Unna, R. D., & Hibberd, J. M. (2014). Deep Evolutionary Comparison of Gene Expression Identifies Parallel Recruitment of Trans-Factors in Two Independent Origins of C4 Photosynthesis. *PLoS Genetics*, *10*(6), e1004365. https://doi.org/10.1371/journal.pgen.1004365

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J., Li, W. W., & Noble, W. S. (2009). MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Research*, *37*(Web Server), W202–W208. https://doi.org/10.1093/nar/gkp335

Bajic, M., Maher, K. A., & Deal, R. B. (2018). *Identification of Open Chromatin Regions in Plant Genomes Using ATAC-Seq* (pp. 183–201). https://doi.org/10.1007/978-1-4939-7318-7_12

Barbrook, A. C., Howe, C. J., & Purton, S. (2006). Why are plastid genomes retained in non-photosynthetic organisms? *Trends in Plant Science*, *11*(2), 101–108. https://doi.org/10.1016/j.tplants.2005.12.004

Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I., & Zhao, K. (2007). High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell*, *129*(4), 823–837. https://doi.org/10.1016/j.cell.2007.05.009

Batschauer, A., MÖSINGER, E., KREUZ, K., DÖRR, I., & APEL, K. (1986). The implication of a plastid-derived factor in the transcriptional control of nuclear genes encoding the light-harvesting

chlorophyll a/b protein. *European Journal of Biochemistry*, *154*(3), 625–634. https://doi.org/10.1111/j.1432-1033.1986.tb09444.x
Bauwe, H., Hagemann, M., & Fernie, A. R. (2010). Photorespiration: players, partners and origin. *Trends in Plant Science*, *15*(6), 330–336. https://doi.org/10.1016/j.tplants.2010.03.006
Beisel, K. G., Schurr, U., & Matsubara, S. (2011). Altered Turnover of -Carotene and Chl a in Arabidopsis Leaves Treated with Lincomycin or Norflurazon. *Plant and Cell Physiology*, *52*(7), 1193–1203. https://doi.org/10.1093/pcp/pcr069
Benn, G., Bjornson, M., Ke, H., De Souza, A., Balmond, E. I., Shaw, J. T., & Dehesh, K. (2016). Plastidial metabolite MEcPP induces a transcriptionally centered stress-response hub via the transcription factor CAMTA3 - supplement. *Proceedings of the National Academy of Sciences*, *16*(2003), 201602582. https://doi.org/10.1073/pnas.1602582113
Berg, J. M., Tymoczko, J. L., & Stryer, L. (2002). *Biochemistry, ; W. H*. Freeman: New York.
Berry, J. O., Mure, C. M., & Yerramsetty, P. (2016). Regulation of Rubisco gene expression in C4 plants. *Current Opinion in Plant Biology*, *31*, 23–28. https://doi.org/10.1016/j.pbi.2016.03.004
Blankenship, R. E. (2014). *Molecular Mechanisms of Photosynthesis*. Wiley. https://books.google.co.uk/books?id=jQztAgAAQBAJ
Bradbeer, J. W., Atkinson, Y. E., Borner, T., & Hagemann, R. (1979). Cytoplasmic synthesis of plastid polypeptides may be controlled by plastid-synthesised RNA. *Nature*, *279*(5716), 816–817. https://doi.org/10.1038/279816a0
Bräutigam, K., Dietzel, L., & Pfannschmidt, T. (2007). Plastid-nucleus communication: anterograde and retrograde signalling in the development and function of plastids (pp. 409–455). https://doi.org/10.1007/4735_2007_0243
Breitenbach, J., Zhu, C., & Sandmann, G. (2001). Bleaching Herbicide Norflurazon Inhibits Phytoene Desaturase by Competition with the Cofactors. *Journal of Agricultural and Food Chemistry*, *49*(11), 5270–5272. https://doi.org/10.1021/jf0106751
Brown, N. J., Newell, C. A., Stanley, S., Chen, J. E., Perrin, A. J., Kajala, K., & Hibberd, J. M. (2011). Independent and Parallel Recruitment of Preexisting Mechanisms Underlying C₄ Photosynthesis. *Science*, *331*(6023), 1436 LP – 1439. https://doi.org/10.1126/science.1201248
Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., & Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, *10*, 1213. http://dx.doi.org/10.1038/nmeth.2688
Burgess, D. G., & Taylor, W. C. (1988). The chloroplast affects the transcription of a nuclear gene family. *MGG Molecular & General Genetics*, *214*(1), 89–96. https://doi.org/10.1007/BF00340185
Burgess, S. J., Granero-moya, I., Grangé-guermente, M. J., Boursnell, C., Terry, M. J., & Hibberd, J. M. (2016). Ancestral light and chloroplast regulation form the foundations for C 4 gene expression. *Nature Plants*, *October*. https://doi.org/10.1038/nplants.2016.161
Burgess, S. J., Reyna-Llorens, I., Stevenson, S. R., Singh, P., Jaeger, K., & Hibberd, J. M. (2019). Genome-Wide Transcription Factor Binding in Leaves from C 3 and C 4 Grasses. *The Plant Cell*, *31*(10), 2297–2314. https://doi.org/10.1105/tpc.19.00078
Calvin, M. (1956). The photosynthetic carbon cycle. *Journal of the Chemical Society (Resumed)*, 1895. https://doi.org/10.1039/jr9560001895
Cardona, T. (2015). A fresh look at the evolution and diversification of photochemical reaction centers. *Photosynthesis Research*, *126*(1), 111–134. https://doi.org/10.1007/s11120-014-0065-x
Chan, K. X., Phua, S. Y., Crisp, P., Mcquinn, R., & Pogson, B. J. (2016). Learning the Languages of the Chloroplast : Retrograde Signaling and Beyond. *Annual Review of Plant Biology*, *67*(November 2015), 25–53. https://doi.org/10.1146/annurev-arplant-043015-111854
Christin, P. A., Boxall, S. F., Gregory, R., Edwards, E. J., Hartwell, J., & Osborne, C. P. (2013). Parallel recruitment of multiple genes into C4 photosynthesis. *Genome Biology and Evolution*, *5*(11), 2174–2187. https://doi.org/10.1093/gbe/evt168

Clayton, H., Saladié, M., Rolland, V., Sharwood, R., Macfarlane, T., & Ludwig, M. (2017). Loss of the Chloroplast Transit Peptide from an Ancestral C(3) Carbonic Anhydrase Is Associated with C(4) Evolution in the Grass Genus Neurachne. *Plant Physiology*, *173*(3), 1648–1658. https://doi.org/10.1104/pp.16.01893

Colombo, M., Tadini, L., Peracchio, C., Ferrari, R., & Pesaresi, P. (2016). GUN1, a Jack-Of-All-Trades in Chloroplast Protein Homeostasis and Signaling. *Frontiers in Plant Science*, *7*(September). https://doi.org/10.3389/fpls.2016.01427

Cserhati, M. (2015). Motif content comparison between monocot and dicot species. *Genomics Data*, *3*, 128–136. https://doi.org/10.1016/j.gdata.2014.12.006

Davie, K., Jacobs, J., Atkins, M., Potier, D., Christiaens, V., Halder, G., & Aerts, S. (2015). Discovery of Transcription Factors and Regulatory Regions Driving In Vivo Tumor Development by ATAC-seq and FAIRE-seq Open Chromatin Profiling. *PLOS Genetics*, *11*(2), e1004994. https://doi.org/10.1371/journal.pgen.1004994

Delius, H., & Koller, B. (1980). Sequence homologies between Escherichia coli and chloroplast ribosomal DNA as seen by heteroduplex analysis. *Journal of Molecular Biology*, *142*(2), 247–261. https://doi.org/https://doi.org/10.1016/0022-2836(80)90048-0

Denton, A. K., Maß, J., Külahoglu, C., Lercher, M. J., Bräutigam, A., & Weber, A. P. M. (2017). Freeze-quenched maize mesophyll and bundle sheath separation uncovers bias in previous tissue-specific RNA-Seq data. *Journal of Experimental Botany*, *68*(2), 147–160. https://doi.org/10.1093/jxb/erw463

Di Baccio, D., Quartacci, M., Vecchia, F., La Rocca, N., Rascio, N., & Navari-Izzo, F. (2002). Bleaching herbicide effects on plastids of dark-grown plants: lipid composition of etioplasts in amitrole and norflurazon-treated barley leaves. *Journal of Experimental Botany*, *53*(376), 1857–1865. https://doi.org/10.1093/jxb/erf035

Dietz, K.-J., Vogel, M. O., & Viehhauser, A. (2010). AP2/EREBP transcription factors are part of gene regulatory networks and integrate metabolic, hormonal and environmental signals in stress acclimation and retrograde signalling. *Protoplasma*, *245*(1–4), 3–14. https://doi.org/10.1007/s00709-010-0142-8

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Dogra, V., Li, M., Singh, S., Li, M., & Kim, C. (2019). Oxidative post-translational modification of EXECUTER1 is required for singlet oxygen sensing in plastids. *Nature Communications*, *10*(1), 2834. https://doi.org/10.1038/s41467-019-10760-6

Doyle, S. M., Diamond, M., & McCabe, P. F. (2010). Chloroplast and reactive oxygen species involvement in apoptotic-like programmed cell death in Arabidopsis suspension cultures. *Journal of Experimental Botany*, *61*(2), 473–482. https://doi.org/10.1093/jxb/erp320

Duan, L., Ruiz-Sola, M. Á., Couso, A., Veciana, N., & Monte, E. (2020). Red and blue light differentially impact retrograde signalling and photoprotection in rice. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1801), 20190402. https://doi.org/10.1098/rstb.2019.0402

Dubreuil, C., Jin, X., Barajas-López, J. de D., Hewitt, T. C., Tanz, S. K., Dobrenel, T., Schröder, W. P., Hanson, J., Pesquet, E., Grönlund, A., Small, I., & Strand, Å. (2018). Establishment of Photosynthesis through Chloroplast Development Is Controlled by Two Distinct Regulatory Phases. *Plant Physiology*, *176*(2), 1199–1214. https://doi.org/10.1104/pp.17.00435

Eastmond, P. J., Astley, H. M., Parsley, K., Aubry, S., Williams, B. P., Menard, G. N., Craddock, C. P., Nunes-Nesi, A., Fernie, A. R., & Hibberd, J. M. (2015). Arabidopsis uses two gluconeogenic gateways for organic acids to fuel seedling establishment. *Nature Communications*, *6*. https://doi.org/10.1038/ncomms7659

Ellis, R. J. (1975). Inhibition of chloroplast protein synthesis by lincomycin and 2- (4-methyl-2 , 6-dinitroanilino)-N-methylpropionamide. *Phytochemistry*, *14*, 89–93.
Emms, D. M., & Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology*, *16*(1). https://doi.org/10.1186/s13059-015-0721-2
Engelmann, S., Bläsing, O. E., Gowik, U., Svensson, P., & Westhoff, P. (2003). Molecular evolution of C4 phosphoenolpyruvate carboxylase in the genus Flaveria--a gradual increase from C3 to C4 characteristics. *Planta*, *217*(5), 717–725. https://doi.org/10.1007/s00425-003-1045-0
Estavillo, G. M., Crisp, P. A., Pornsiriwong, W., Wirtz, M., Collinge, D., Carrie, C., Giraud, E., Whelan, J., David, P., Javot, H., Brearley, C., Hell, R., Marin, E., & Pogson, B. J. (2011). Evidence for a SAL1-PAP Chloroplast Retrograde Pathway That Functions in Drought and High Light Signaling in Arabidopsis. *The Plant Cell*, *23*(11), 3992–4012. https://doi.org/10.1105/tpc.111.091033
Finster, S., Eggert, E., Zoschke, R., Weihe, A., & Schmitz-Linneweber, C. (2013). Light-dependent, plastome-wide association of the plastid-encoded RNA polymerase with chloroplast DNA. *The Plant Journal*, *76*(5), 849–860. https://doi.org/https://doi.org/10.1111/tpj.12339
Fischer, W. W., Hemp, J., & Johnson, J. E. (2016). Evolution of Oxygenic Photosynthesis. *Annual Review of Earth and Planetary Sciences*, *44*(1), 647–683. https://doi.org/10.1146/annurev-earth-060313-054810
Gill, G. (2001). Regulation of the initiation of eukaryotic transcription. *Essays in Biochemistry*, *37*, 33–43. https://doi.org/10.1042/bse0370033
Gonzalez-Perez, S., Gutierrez, J., Garcia-Garcia, F., Osuna, D., Dopazo, J., Lorenzo, O., Revuelta, J. L., & Arellano, J. B. (2011). Early Transcriptional Defense Responses in Arabidopsis Cell Suspension Culture under High-Light Conditions. *PLANT PHYSIOLOGY*, *156*(3), 1439–1456. https://doi.org/10.1104/pp.111.177766
Gowik, U., Burscheidt, J., Akyildiz, M., Schlue, U., Koczor, M., Streubel, M., & Westhoff, P. (2004). cis-Regulatory elements for mesophyll-specific gene expression in the C4 plant Flaveria trinervia, the promoter of the C4 phosphoenolpyruvate carboxylase gene. *The Plant Cell*, *16*(5), 1077–1090. https://doi.org/10.1105/tpc.019729
Grant, C. E., Bailey, T. L., & Noble, W. S. (2011). FIMO: scanning for occurrences of a given motif. *Bioinformatics*, *27*(7), 1017–1018. https://doi.org/10.1093/bioinformatics/btr064
Gulledge, A. A., Vora, H., Patel, K., & Loraine, A. E. (2014). *A Protocol for Visual Analysis of Alternative Splicing in RNA-Seq Data Using Integrated Genome Browser* (pp. 123–137). https://doi.org/10.1007/978-1-4939-0700-7_8
Guo, J., Wu, J., Ji, Q., Wang, C., Luo, L., Yuan, Y., Wang, Y., & Wang, J. (2008). Genome-wide analysis of heat shock transcription factor families in rice and Arabidopsis. *Journal of Genetics and Genomics*, *35*(2), 105–118. https://doi.org/https://doi.org/10.1016/S1673-8527(08)60016-8
Harpster, M. H., Mayfield, S. P., & Taylor, W. C. (1984). Effects of pigment-deficient mutants on the accumulation of photosynthetic proteins in maize. *Plant Molecular Biology*, *3*(2), 59–71. https://doi.org/10.1007/BF00040030
Hatch, M. D. (1987). C4 photosynthesis: a unique elend of modified biochemistry, anatomy and ultrastructure. *Biochimica et Biophysica Acta (BBA) - Reviews on Bioenergetics*, *895*(2), 81–106. https://doi.org/10.1016/S0304-4173(87)80009-5
Hayashi, G., Shibato, J., Imanaka, T., Cho, K., Kubo, A., Kikuchi, S., Satoh, K., Kimura, S., Ozawa, S., Fukutani, S., Endo, S., Ichikawa, K., Agrawal, G. K., Shioda, S., Fukumoto, M., & Rakwal, R. (2014). Unraveling low-level gamma radiation-responsive changes in expression of early and late genes in leaves of rice seedlings at litate Village, Fukushima. *Journal of Heredity*, *105*(5), 723–738. https://doi.org/10.1093/jhered/esu025
Heimann, L., Horst, I., Perduns, R., Dreesen, B., Offermann, S., & Peterhansel, C. (2013). A Common Histone Modification Code on C4 Genes in Maize and Its Conservation in Sorghum and Setaria italica. *Plant Physiology*, *162*(1), 456–469. https://doi.org/10.1104/pp.113.216721

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H., & Glass, C. K. (2010). Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Molecular Cell*, *38*(4), 576–589. https://doi.org/10.1016/j.molcel.2010.05.004

Hess, W. R., Müller, A., Nagy, F., & Börner, T. (1994). Ribosome-deficient plastids affect transcription of light-induced nuclear genes: genetic evidence for a plastid-derived signal. *MGG Molecular & General Genetics*, *242*(3), 305–312. https://doi.org/10.1007/BF00280420

Heyno, E., Innocenti, G., Lemaire, S. D., Issakidis-Bourguet, E., & Krieger-Liszkay, A. (2014). Putative role of the malate valve enzyme NADP–malate dehydrogenase in H 2 O 2 signalling in Arabidopsis. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1640), 20130228. https://doi.org/10.1098/rstb.2013.0228

Hibberd, J. M., & Covshoff, S. (2010). The regulation of gene expression required for C4 photosynthesis. *Annual Review of Plant Biology*, *61*, 181–207. https://doi.org/doi:10.1146/annurev-arplant-042809-112238

Hills, A. C., Khan, S., & López-Juez, E. (2015). Chloroplast Biogenesis-Associated Nuclear Genes: Control by Plastid Signals Evolved Prior to Their Regulation as Part of Photomorphogenesis. *Frontiers in Plant Science*, *6*. https://doi.org/10.3389/fpls.2015.01078

Hori, K., Maruyama, F., Fujisawa, T., Togashi, T., Yamamoto, N., Seo, M., Sato, S., Yamada, T., Mori, H., Tajima, N., Moriyama, T., Ikeuchi, M., Watanabe, M., Wada, H., Kobayashi, K., Saito, M., Masuda, T., Sasaki-Sekimoto, Y., Mashiguchi, K., … Ohta, H. (2014). Klebsormidium flaccidum genome reveals primary factors for plant terrestrial adaptation. *Nature Communications*, *5*(1), 3978. https://doi.org/10.1038/ncomms4978

Howe, C. ., Barbrook, A. ., Nisbet, R. E. ., Lockhart, P. ., & Larkum, A. W. . (2008). The origin of plastids. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1504), 2675–2685. https://doi.org/10.1098/rstb.2008.0050

Hu, W., DePamphilis, C. W., & Ma, H. (2008). Phylogenetic Analysis of the Plant-specific Zinc Finger-Homeobox and Mini Zinc Finger Gene Families. *Journal of Integrative Plant Biology*, *50*(8), 1031–1045. https://doi.org/10.1111/j.1744-7909.2008.00681.x

Huang, W., Loganantharaj, R., Schroeder, B., Fargo, D., & Li, L. (2013). PAVIS: a tool for Peak Annotation and Visualization. *Bioinformatics*, *29*(23), 3097–3099. https://doi.org/10.1093/bioinformatics/btt520

Hudson, D., Guevara, D., Yaish, M. W., Hannam, C., Long, N., Clarke, J. D., Bi, Y.-M., & Rothstein, S. J. (2011). GNC and CGA1 Modulate Chlorophyll Biosynthesis and Glutamate Synthase (GLU1/Fd-GOGAT) Expression in Arabidopsis. *PLoS ONE*, *6*(11), e26765. https://doi.org/10.1371/journal.pone.0026765

Jensen, M. K., Kjaersgaard, T., Nielsen, M. M., Galberg, P., Petersen, K., O'Shea, C., & Skriver, K. (2010). The Arabidopsis thaliana NAC transcription factor family: structure-function relationships and determinants of ANAC019 stress signalling. *The Biochemical Journal*, *426*(2), 183–196. https://doi.org/10.1042/BJ20091234

Jiao, Y., Lau, O. S., & Deng, X. W. (2007). Light-regulated transcriptional networks in higher plants. *Nat Rev Genet*, *8*(3), 217–230. http://dx.doi.org/10.1038/nrg2049

Jiao, Y., Ma, L., Strickland, E., & Deng, X. W. (2005). Conservation and divergence of light-regulated genome expression patterns during seedling development in rice and Arabidopsis. *The Plant Cell*, *17*(12), 3239–3256. https://doi.org/10.1105/tpc.105.035840

Johnson, D. S., Mortazavi, A., Myers, R. M., & Wold, B. (2007). Genome-Wide Mapping of in Vivo Protein-DNA Interactions. *Science*, *316*(5830), 1497–1502. https://doi.org/10.1126/science.1141319

Kacprzak, S. M., Mochizuki, N., Naranjo, B., Xu, D., Leister, D., Kleine, T., Okamoto, H., & Terry, M. J. (2019a). Plastid-to-Nucleus Retrograde Signalling during Chloroplast Biogenesis Does Not Require ABI4. *Plant Physiology*, *179*(1), 18 LP – 23. https://doi.org/10.1104/pp.18.01047

Kacprzak, S. M., Mochizuki, N., Naranjo, B., Xu, D., Leister, D., Kleine, T., Okamoto, H., & Terry, M. J. (2019b). Plastid-to-Nucleus Retrograde Signalling during Chloroplast Biogenesis Does Not Require ABI4. *Plant Physiology*, *179*(1), 18–23. https://doi.org/10.1104/pp.18.01047

Kajala, K., Brown, N. J., Williams, B. P., Borrill, P., Taylor, L. E., & Hibberd, J. M. (2012). Multiple Arabidopsis genes primed for recruitment into C4 photosynthesis. *The Plant Journal*, *69*(1), 47–56. https://doi.org/https://doi.org/10.1111/j.1365-313X.2011.04769.x

Kakizaki, T., Matsumura, H., Nakayama, K., Che, F.-S., Terauchi, R., & Inaba, T. (2009). Coordination of Plastid Protein Import and Nuclear Gene Expression by Plastid-to-Nucleus Retrograde Signaling. *Plant Physiology*, *151*(3), 1339–1353. https://doi.org/10.1104/pp.109.145987

Keeling, P. J. (2010). The endosymbiotic origin, diversification and fate of plastids. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1541), 729 LP – 748. http://rstb.royalsocietypublishing.org/content/365/1541/729.abstract

Kim, C., Meskauskiene, R., Zhang, S., Lee, K. P., Lakshmanan Ashok, M., Blajecka, K., Herrfurth, C., Feussner, I., & Apel, K. (2012). Chloroplasts of Arabidopsis Are the Source and a Primary Target of a Plant-Specific Programmed Cell Death Signaling Pathway. *The Plant Cell*, *24*(7), 3026–3039. https://doi.org/10.1105/tpc.112.100479

Klein, D. C., & Hainer, S. J. (2020). Genomic methods in profiling DNA accessibility and factor localization. Chromosome Research : An International Journal on the Molecular, Supramolecular and Evolutionary Aspects of Chromosome Biology, 28(1), 69–85. https://doi.org/10.1007/s10577-019-09619-9

Kobayashi, K., Obayashi, T., & Masuda, T. (2012). Role of the G-box element in regulation of chlorophyll biosynthesis in Arabidopsis roots. *Plant Signaling & Behavior*, *7*(8), 922–926. https://doi.org/10.4161/psb.20760

Kodaira, K.-S., Qin, F., Tran, L.-S. P., Maruyama, K., Kidokoro, S., Fujita, Y., Shinozaki, K., & Yamaguchi-Shinozaki, K. (2011). Arabidopsis Cys2/His2 Zinc-Finger Proteins AZF1 and AZF2 Negatively Regulate Abscisic Acid-Repressive and Auxin-Inducible Genes under Abiotic Stress Conditions. *Plant Physiology*, *157*(2), 742–756. https://doi.org/10.1104/pp.111.182683

Kosugi, S., & Ohashi, Y. (1997). PCF1 and PCF2 specifically bind to cis elements in the rice proliferating cell nuclear antigen gene. *The Plant Cell*, *9*(9), 1607–1619. https://doi.org/10.1105/tpc.9.9.1607

Koussevitzky, S., Nott, A., Mockler, T. C., Hong, F., Sachetto-Martins, G., Surpin, M., Lim, J., Mittler, R., & Chory, J. (2007). Signals from Chloroplasts Converge to Regulate Nuclear Gene Expression. *Science*, *316*(5825), 715 LP – 719. http://science.sciencemag.org/content/316/5825/715.abstract

Kremling, K. A. G., Chen, S.-Y., Su, M.-H., Lepak, N. K., Romay, M. C., Swarts, K. L., Lu, F., Lorant, A., Bradbury, P. J., & Buckler, E. S. (2018). Dysregulation of expression correlates with rare-allele burden and fitness loss in maize. *Nature*, *555*(7697), 520–523. https://doi.org/10.1038/nature25966

La Rocca, N., Dalla Vecchia, F., Barbato, R., Bonora, A., Bergantino, E., & Rascio, N. (2000). Plastid photodamage and Cab gene expression in barley leaves. *Physiologia Plantarum*, *109*(1), 51–57. https://doi.org/10.1034/j.1399-3054.2000.100108.x

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Larkin, R. M. (2003). GUN4, a Regulator of Chlorophyll Synthesis and Intracellular Signaling. *Science*, *299*(5608), 902–906. https://doi.org/10.1126/science.1079978

Larkin, Robert M, Larkin, R., Ruckle, M., Mullineaux, P., Baker, N., Chi, W., Sun, X., Zhang, L., Mochizuki, N., Susek, R., Chory, J., Ruckle, M., DeMarco, S., Larkin, R., Foudree, A., Aluru, M., Rodermel, S., Hassidim, M., Yakir, E., … Wilson, P. (2014). Influence of plastids on light signalling and development. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *369*(1640), 20130232. https://doi.org/10.1098/rstb.2013.0232

Larkin, Robert M, & Ruckle, M. E. (2008). Integration of light and plastid signals. *Current Opinion in Plant Biology*, *11*(6), 593–599. https://doi.org/10.1016/j.pbi.2008.10.004
Lee, K. P., Kim, C., Landgraf, F., & Apel, K. (2007). EXECUTER1- and EXECUTER2-dependent transfer of stress-related signals from the plastid to the nucleus of Arabidopsis thaliana. *Proceedings of the National Academy of Sciences of the United States of America*, *104*, 10270–10275. https://doi.org/10.1073/pnas.0702061104
Lee, T. H., Tang, H., Wang, X., & Paterson, A. H. (2013). PGDD: A database of gene and genome duplication in plants. *Nucleic Acids Research*, *41*(D1), 1152–1158. https://doi.org/10.1093/nar/gks1104
Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352
Li, X., Zhang, H., Ai, Q., Liang, G., & Yu, D. (2016). Two bHLH Transcription Factors, bHLH34 and bHLH104, Regulate Iron Homeostasis in Arabidopsis thaliana. *Plant Physiology*, *170*(4), 2478–2493. https://doi.org/10.1104/pp.15.01827
Li, Y., Liu, W., Zhong, H., Zhang, H., & Xia, Y. (2019). Redox-sensitive <scp>bZIP</scp> 68 plays a role in balancing stress tolerance with growth in Arabidopsis. *The Plant Journal*, *100*(4), 768–783. https://doi.org/10.1111/tpj.14476
Liang, C., Cheng, S., Zhang, Y., Sun, Y., Fernie, A. R., Kang, K., Panagiotou, G., Lo, C., & Lim, B. L. (2016). Transcriptomic, proteomic and metabolic changes in Arabidopsis thaliana leaves after the onset of illumination. *BMC Plant Biology*, *16*(1), 43. https://doi.org/10.1186/s12870-016-0726-3
Loudya, N., Okunola, T., He, J., Jarvis, P., & López-Juez, E. (2020). Retrograde signalling in a virescent mutant triggers an anterograde delay of chloroplast biogenesis that requires GUN1 and is essential for survival. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1801), 20190400. https://doi.org/10.1098/rstb.2019.0400
Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550. https://doi.org/10.1186/s13059-014-0550-8
Lu, Z., Hofmeister, B. T., Vollmers, C., DuBois, R. M., & Schmitz, R. J. (2017). Combining ATAC-seq with nuclei sorting for discovery of cis-regulatory regions in plant genomes. *Nucleic Acids Research*, *45*(6), e41–e41. https://doi.org/10.1093/nar/gkw1179
Ludwig, M. (2012). Carbonic anhydrase and the molecular evolution of C4 photosynthesis. *Plant, Cell & Environment*, *35*(1), 22–37. https://doi.org/10.1111/j.1365-3040.2011.02364.x
Magnucka, E. G., Suzuki, Y., Pietr, S. J., Kozubek, A., & Zarnowski, R. (2007). Effect of norflurazon on resorcinolic lipid metabolism in rye seedlings. *Zeitschrift Fur Naturforschung - Section C Journal of Biosciences*, *62*(3–4), 239–245.
Maher, K. A., Bajic, M., Kajala, K., Reynoso, M., Pauluzzi, G., West, D. A., Zumstein, K., Woodhouse, M., Bubb, K., Dorrity, M. W., Queitsch, C., Bailey-Serres, J., Sinha, N., Brady, S. M., & Deal, R. B. (2018). Profiling of Accessible Chromatin Regions across Multiple Plant Species and Cell Types Reveals Common Gene Regulatory Principles and New Control Modules. *The Plant Cell*, *30*(1), 15–36. https://doi.org/10.1105/tpc.17.00581
Marini, F., & Binder, H. (2019). pcaExplorer: an R/Bioconductor package for interacting with RNA-seq principal components. *BMC Bioinformatics*, *20*(1), 331. https://doi.org/10.1186/s12859-019-2879-1
Martín, G., Leivar, P., Ludevid, D., Tepperman, J. M., Quail, P. H., & Monte, E. (2016). Phytochrome and retrograde signalling pathways converge to antagonistically regulate a light-induced transcriptional network. *Nature Communications*, *7*(May), 11431. https://doi.org/10.1038/ncomms11431
Martin, W., Rujan, T., Richly, E., Hansen, A., Cornelsen, S., Lins, T., Leister, D., Stoebe, B., Hasegawa, M., & Penny, D. (2002). Evolutionary analysis of Arabidopsis, cyanobacterial, and

chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proceedings of the National Academy of Sciences*, *99*(19), 12246–12251. https://doi.org/10.1073/pnas.182432999

Masumoto, C., Miyazawa, S.-I., Ohkawa, H., Fukuda, T., Taniguchi, Y., Murayama, S., Kusano, M., Saito, K., Fukayama, H., & Miyao, M. (2010). Phosphoenolpyruvate carboxylase intrinsically located in the chloroplast of rice plays a crucial role in ammonium assimilation. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(11), 5226–5231. https://doi.org/10.1073/pnas.0913127107

Mathelier, A., Zhao, X., Zhang, A. W., Parcy, F., Worsley-Hunt, R., Arenillas, D. J., Buchman, S., Chen, C., Chou, A., Ienasescu, H., Lim, J., Shyr, C., Tan, G., Zhou, M., Lenhard, B., Sandelin, A., & Wasserman, W. W. (2014). JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Research*, *42*(D1), D142–D147. https://doi.org/10.1093/nar/gkt997

Maxwell, K., & Johnson, G. N. (2000). Chlorophyll fluorescence--a practical guide. *Journal of Experimental Botany*, *51*(345), 659–668. https://doi.org/10.1093/jexbot/51.345.659

McCormac, A. C., & Terry, M. J. (2004). The nuclear genes Lhcb and HEMA1 are differentially sensitive to plastid signals and suggest distinct roles for the GUN1 and GUN5 plastid-signalling pathways during de-etiolation. *The Plant Journal*, *40*(5), 672–685. https://doi.org/10.1111/j.1365-313X.2004.02243.x

McLeay, R. C., & Bailey, T. L. (2010). Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data. *BMC Bioinformatics*, *11*(1), 165. https://doi.org/10.1186/1471-2105-11-165

Mochizuki, N., Brusslan, J. A., Larkin, R., Nagatani, A., & Chory, J. (2001). Arabidopsis genomes uncoupled 5 (GUN5) mutant reveals the involvement of Mg-chelatase H subunit in plastid-to-nucleus signal transduction. *Proceedings of the National Academy of Sciences*, *98*(4), 2053–2058. https://doi.org/10.1073/pnas.98.4.2053

Mochizuki, Nobuyoshi, Tanaka, R., Tanaka, A., Masuda, T., & Nagatani, A. (2008). The steady-state level of Mg-protoporphyrin IX is not a determinant of plastid-to-nucleus signaling in Arabidopsis. *Proceedings of the National Academy of Sciences* , *105*(39), 15184–15189. https://doi.org/10.1073/pnas.0803245105

Moulin, M., McCormac, A. C., Terry, M. J., & Smith, A. G. (2008). Tetrapyrrole profiling in Arabidopsis seedlings reveals that retrograde plastid nuclear signaling is not due to Mg-protoporphyrin IX accumulation. *Proceedings of the National Academy of Sciences*, *105*(39), 15178–15183. https://doi.org/10.1073/pnas.0803054105

Mulkidjanian, A. Y., Koonin, E. V, Makarova, K. S., Mekhedov, S. L., Sorokin, A., Wolf, Y. I., Dufresne, A., Partensky, F., Burd, H., Kaznadzey, D., Haselkorn, R., & Galperin, M. Y. (2006). The cyanobacterial genome core and the origin of photosynthesis. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(35), 13126–13131. https://doi.org/10.1073/pnas.0605709103

Mulo, P., Pursiheimo, S., Hou, C.-X., Tyystjärvi, T., & Aro, E.-M. (2003). Multiple effects of antibiotics on chloroplast and nuclear gene expression. *Functional Plant Biology*, *30*, 1097. https://doi.org/10.1071/FP03149

Muthuramalingam, M., Matros, A., Scheibe, R., Mock, H.-P., & Dietz, K.-J. (2013). The hydrogen peroxide-sensitive proteome of the chloroplast in vitro and in vivo. *Frontiers in Plant Science*, *4*. https://doi.org/10.3389/fpls.2013.00054

Nellaepalli, S., Ozawa, S.-I., Kuroda, H., & Takahashi, Y. (2018). The photosystem I assembly apparatus consisting of Ycf3–Y3IP1 and Ycf4 modules. *Nature Communications*, *9*(1), 2439. https://doi.org/10.1038/s41467-018-04823-3

Nguyen, N. H., & Lee, H. (2016). MYB-related transcription factors function as regulators of the circadian clock and anthocyanin biosynthesis in Arabidopsis. *Plant Signaling & Behavior*, *11*(3), e1139278. https://doi.org/10.1080/15592324.2016.1139278

O'Malley, R. C., Huang, S. C., Song, L., Lewsey, M. G., Bartlett, A., Nery, J. R., Galli, M., Gallavotti, A., & Ecker, J. R. (2016). Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. *Cell*, *165*(5), 1280–1292. https://doi.org/10.1016/j.cell.2016.04.038

Ouwerkerk, P. B. F., & Meijer, A. H. (2001). Yeast One-Hybrid Screening for <scp>DNA</scp> - Protein Interactions. *Current Protocols in Molecular Biology*, *55*(1). https://doi.org/10.1002/0471142727.mb1212s55

Page, M. T., Garcia-Becerra, T., Smith, A. G., & Terry, M. J. (2020). Overexpression of chloroplast-targeted ferrochelatase 1 results in a genomes uncoupled chloroplast-to-nucleus retrograde signalling phenotype. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1801), 20190401. https://doi.org/10.1098/rstb.2019.0401

Page, M. T., Kacprzak, S. M., Mochizuki, N., Okamoto, H., Smith, A. G., & Terry, M. J. (2017). Seedlings Lacking the PTM Protein Do Not Show a *genomes uncoupled (gun)* Mutant Phenotype. *Plant Physiology*, *174*(1), 21–26. https://doi.org/10.1104/pp.16.01930

Page, M. T., McCormac, A. C., Smith, A. G., & Terry, M. J. (2017). Singlet oxygen initiates a plastid signal controlling photosynthetic gene expression. *New Phytologist*, *213*(3), 1168–1180. https://doi.org/10.1111/nph.14223

Park, J. H., & Jung, S. (2017). Perturbations of carotenoid and tetrapyrrole biosynthetic pathways result in differential alterations in chloroplast function and plastid signaling. *Biochemical and Biophysical Research Communications*, *482*(4), 672–677. https://doi.org/10.1016/j.bbrc.2016.11.092

Park, J. H., & Jung, S. (2018). Perturbations in carotenoid and porphyrin status result in differential photooxidative stress signaling and antioxidant responses. *Biochemical and Biophysical Research Communications*, *496*(3), 840–845. https://doi.org/10.1016/j.bbrc.2018.01.142

Park, J. H., Tran, L. H., & Jung, S. (2017). Perturbations in the photosynthetic pigment status result in photooxidation-induced crosstalk between carotenoid and porphyrin biosynthetic pathways. *Frontiers in Plant Science*, *8*(November), 1–11. https://doi.org/10.3389/fpls.2017.01992

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, *14*(4), 417–419. https://doi.org/10.1038/nmeth.4197

Pérez-García, P., Ma, Y., Yanovsky, M. J., & Mas, P. (2015). Time-dependent sequestration of RVE8 by LNK proteins shapes the diurnal oscillation of anthocyanin biosynthesis. *Proceedings of the National Academy of Sciences*, *112*(16), 5249–5253. https://doi.org/10.1073/pnas.1420792112

Petrillo, E., Godoy Herz, M. A., Fuchs, A., Reifer, D., Fuller, J., Yanovsky, M. J., Simpson, C., Brown, J. W. S., Barta, A., Kalyna, M., & Kornblihtt, A. R. (2014). A Chloroplast Retrograde Signal Regulates Nuclear Alternative Splicing. *Science*, *344*(6182), 427–430. https://doi.org/10.1126/science.1250322

Pfannschmidt, T., Allen, J. F., & Oelmüller, R. (2001). Principles of redox control in photosynthesis gene expression. *Physiologia Plantarum*, *112*(1), 1–9. https://doi.org/10.1034/j.1399-3054.2001.1120101.x

Phukan, U. J., Jeena, G. S., & Shukla, R. K. (2016). WRKY Transcription Factors: Molecular Regulation and Stress Responses in Plants . In *Frontiers in Plant Science* (Vol. 7, p. 760). https://www.frontiersin.org/article/10.3389/fpls.2016.00760

Piper, J., Elze, M. C., Cauchy, P., Cockerill, P. N., Bonifer, C., & Ott, S. (2013). Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data. *Nucleic Acids Research*, *41*(21), e201–e201. https://doi.org/10.1093/nar/gkt850

Pogson, B. J., & Albrecht, V. (2011). Genetic dissection of chloroplast biogenesis and development: an overview. *Plant Physiology*, *155*(4), 1545–1551. https://doi.org/10.1104/pp.110.170365

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

Ramel, F., Birtic, S., Ginies, C., Soubigou-Taconnat, L., Triantaphylides, C., & Havaux, M. (2012). Carotenoid oxidation products are stress signals that mediate gene responses to singlet oxygen in

plants. *Proceedings of the National Academy of Sciences*, *109*(14), 5535–5540. https://doi.org/10.1073/pnas.1115982109
Ranftl, Q. L., Bastakis, E., Klermund, C., & Schwechheimer, C. (2016). LLM-Domain Containing B-GATA Factors Control Different Aspects of Cytokinin-Regulated Development in Arabidopsis thaliana. *Plant Physiology*, *170*(4), 2295–2311. https://doi.org/10.1104/pp.15.01556
Reske, J. J., Wilson, M. R., & Chandler, R. L. (2020). ATAC-seq normalization method can significantly affect differential accessibility analysis and interpretation. *Epigenetics & Chromatin*, *13*(1), 22. https://doi.org/10.1186/s13072-020-00342-y
Reyes, J. C., Muro-Pastor, M. I., & Florencio, F. J. (2004). The GATA family of transcription factors in Arabidopsis and rice. *Plant Physiology*, *134*(4), 1718–1732. https://doi.org/10.1104/pp.103.037788
Reyna-Llorens, I., & Hibberd, J. M. (2017). Recruitment of pre-existing networks during the evolution of C4 photosynthesis. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1730), 20160386. https://doi.org/10.1098/rstb.2016.0386
Richter, A. S., Tohge, T., Fernie, A. R., & Grimm, B. (2020). The genomes uncoupled -dependent signalling pathway coordinates plastid biogenesis with the synthesis of anthocyanins. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1801), 20190403. https://doi.org/10.1098/rstb.2019.0403
Rocca, N. La, Barbato, R., Vecchia, F. dalla, & Rascio, N. (2000). Cab gene expression in bleached leaves of carotenoid-deficient maize. *Photosynthesis Research*, *64*(2–3), 119–126. https://doi.org/10.1023/A:1006477215572
Ruckle, M. E., DeMarco, S. M., & Larkin, R. M. (2007). Plastid signals remodel light signaling networks and are essential for efficient chloroplast biogenesis in Arabidopsis. *The Plant Cell*, *19*(12), 3944–3960. https://doi.org/10.1105/tpc.107.054312
Rudowska, Ł., Gieczewska, K., Mazur, R., Garstka, M., & Mostowska, A. (2012). Chloroplast biogenesis — Correlation between structure and function. *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, *1817*(8), 1380–1387. https://doi.org/10.1016/j.bbabio.2012.03.013
Sage, R. F., Christin, P. A., & Edwards, E. J. (2011). The C 4 plant lineages of planet Earth. In *Journal of Experimental Botany* (Vol. 62, Issue 9, pp. 3155–3169). https://doi.org/10.1093/jxb/err048
Schlüter, U., & Weber, A. P. M. (2020). Regulation and Evolution of C4 Photosynthesis. *Annual Review of Plant Biology*, *71*(1), 183–215. https://doi.org/10.1146/annurev-arplant-042916-040915
Sedelnikova, O. V, Hughes, T. E., & Langdale, J. A. (2018). Understanding the Genetic Basis of C4 Kranz Anatomy with a View to Engineering C3 Crops. *Annual Review of Genetics*, *52*(1), 249–270. https://doi.org/10.1146/annurev-genet-120417-031217
Selosse, M.-A., Albert, B., & Godelle, B. (2001). Reducing the genome size of organelles favours gene transfer to the nucleus. *Trends in Ecology & Evolution*, *16*(3), 135–141. https://doi.org/10.1016/S0169-5347(00)02084-X
Shimada, T. L., Shimada, T., & Hara-Nishimura, I. (2010). A rapid and non-destructive screenable marker, FAST, for identifying transformed seeds of Arabidopsis thaliana. *The Plant Journal*, *61*(3), 519–528. https://doi.org/10.1111/j.1365-313X.2009.04060.x
Shimizu, T., Kacprzak, S. M., Mochizuki, N., Nagatani, A., Watanabe, S., Shimada, T., Tanaka, K., Hayashi, Y., Arai, M., Leister, D., Okamoto, H., Terry, M. J., & Masuda, T. (2019). The retrograde signaling protein GUN1 regulates tetrapyrrole biosynthesis. *Proceedings of the National Academy of Sciences*, *116*(49), 24900–24906. https://doi.org/10.1073/pnas.1911251116
Stegemann, S., Hartmann, S., Ruf, S., & Bock, R. (2003). High-frequency gene transfer from the chloroplast genome to the nucleus. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(15), 8828–8833. https://doi.org/10.1073/pnas.1430924100
Sullivan, A. M., Arsovski, A. A., Lempe, J., Bubb, K. L., Weirauch, M. T., Sabo, P. J., Sandstrom, R., Thurman, R. E., Neph, S., Reynolds, A. P., Stergachis, A. B., Vernot, B., Johnson, A. K., Haugen, E., Sullivan, S. T., Thompson, A., Neri, F. V., Weaver, M., Diegel, M., … Stamatoyannopoulos, J.

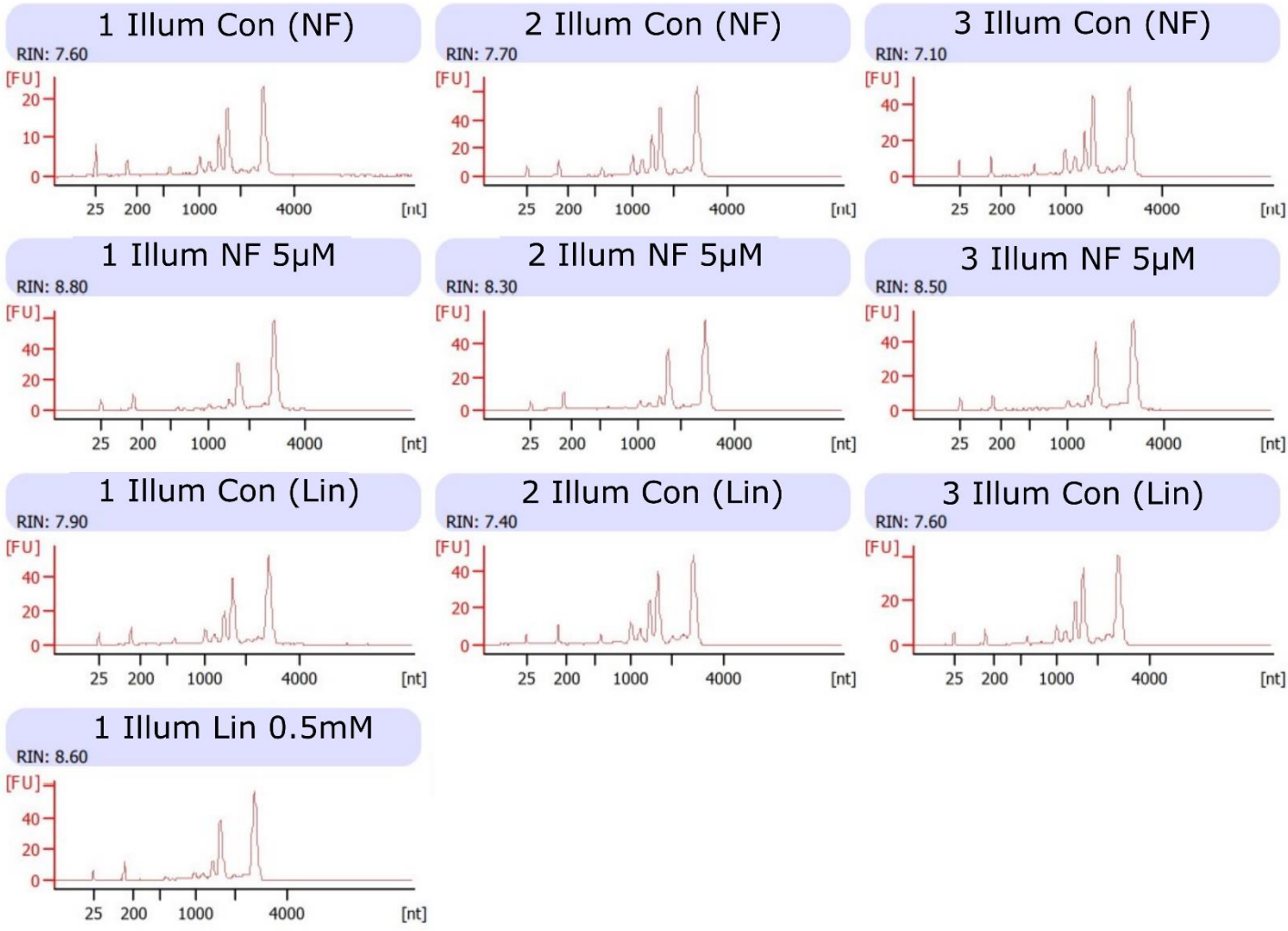
A. (2014). Mapping and Dynamics of Regulatory DNA and Transcription Factor Networks in A. thaliana. *Cell Reports*, *8*(6), 2015–2030. https://doi.org/10.1016/j.celrep.2014.08.019
Sun, X., Feng, P., Xu, X., Guo, H., Ma, J., Chi, W., Lin, R., Lu, C., & Zhang, L. (2011). A chloroplast envelope-bound PHD transcription factor mediates chloroplast signals to the nucleus. *Nature Communications*, *2*(1), 410–477. https://doi.org/10.1038/ncomms1486
Susek, R. E., Ausubel, F. M., & Chory, J. (1993). Signal transduction mutants of arabidopsis uncouple nuclear CAB and RBCS gene expression from chloroplast development. *Cell*, *74*(5), 787–799. https://doi.org/10.1016/0092-8674(93)90459-4
Tamada, Y., Imanari, E., Kurotani, K., Nakai, M., Andreo, C. S., & Izui, K. (2003). Effect of photooxidative destruction of chloroplasts on the expression of nuclear genes for C4 photosynthesis and for chloroplast biogenesis in maize. *Journal of Plant Physiology*, *160*(1), 3–8. https://doi.org/Doi 10.1078/0176-1617-00918
Tamai, H., Iwabuchi, M., & Meshi, T. (2002). Arabidopsis GARP Transcriptional Activators Interact with the Pro-Rich Activation Domain Shared by G-Box-Binding bZIP Factors. *Plant and Cell Physiology*, *43*(1), 99–107. https://doi.org/10.1093/pcp/pcf011
Tanaka, R., & Tanaka, A. (2007). Tetrapyrrole Biosynthesis in Higher Plants. *Annual Review of Plant Biology*, *58*(1), 321–346. https://doi.org/10.1146/annurev.arplant.57.032905.105448
Tanz, S. K., Tetu, S. G., Vella, N. G. F., & Ludwig, M. (2009). Loss of the Transit Peptide and an Increase in Gene Expression of an Ancestral Chloroplastic Carbonic Anhydrase Were Instrumental in the Evolution of the Cytosolic C₄ Carbonic Anhydrase in Flaveria *Plant Physiology*, *150*(3), 1515 LP – 1529. http://www.plantphysiol.org/content/150/3/1515.abstract
Tarantino, D., Vianelli, A., Carraro, L., & Soave, C. (1999). A nuclear mutant of Arabidopsis thaliana selected for enhanced sensitivity to light-chill stress is altered in PSII electron transport activity. *Physiologia Plantarum*, *107*(3), 361–371. https://doi.org/10.1034/j.1399-3054.1999.100314.x
Tian, T., Liu, Y., Yan, H., You, Q., Yi, X., Du, Z., Xu, W., & Su, Z. (2017). agriGO v2.0: a GO analysis toolkit for the agricultural community, 2017 update. *Nucleic Acids Research*, *45*(W1), W122–W129. http://dx.doi.org/10.1093/nar/gkx382
Tokumaru, M., Adachi, F., Toda, M., Ito-Inaba, Y., Yazu, F., Hirosawa, Y., Sakakibara, Y., Suiko, M., Kakizaki, T., & Inaba, T. (2017). Ubiquitin-Proteasome Dependent Regulation of the GOLDEN2-LIKE 1 Transcription Factor in Response to Plastid Signals. *Plant Physiology*, *173*(1), 524–535. https://doi.org/10.1104/pp.16.01546
Tronconi, M. A., Fahnenstich, H., Gerrard Weehler, M. C., Andreo, C. S., Flügge, U.-I., Drincovich, M. F., & Maurino, V. G. (2008). Arabidopsis NAD-Malic Enzyme Functions As a Homodimer and Heterodimer and Has a Major Impact on Nocturnal Metabolism. *Plant Physiology*, *146*(4), 1540–1552. https://doi.org/10.1104/pp.107.114975
Uberegui, E., Hall, M., Lorenzo, ??scar, Schr??der, W. P., & Balsera, M. (2015). An Arabidopsis soluble chloroplast proteomic analysis reveals the participation of the Executer pathway in response to increased light conditions. *Journal of Experimental Botany*, *66*(7), 2067–2077. https://doi.org/10.1093/jxb/erv018
Usadel, B., Poree, F., Nagel, A., Lohse, M., Czedik-Eysenberg, A., & Stitt, M. (2009). A guide to using MapMan to visualize and compare Omics data in plants: A case study in the crop species, Maize. *Plant, Cell and Environment*, *32*(9), 1211–1229. https://doi.org/10.1111/j.1365-3040.2009.01978.x
Wang, H., Ham, T.-H., Im, D.-E., Lar, S. M., Jang, S.-G., Lee, J., Mo, Y., Jeung, J.-U., Kim, S. T., & Kwon, S.-W. (2020). A New SNP in Rice Gene Encoding Pyruvate Phosphate Dikinase (PPDK) Associated with Floury Endosperm. *Genes*, *11*(4), 465. https://doi.org/10.3390/genes11040465
Wang, P., Hendron, R.-W., & Kelly, S. (2017). Transcriptional control of photosynthetic capacity: conservation and divergence from *Arabidopsis* to rice. *New Phytologist*, *216*(1), 32–45. https://doi.org/10.1111/nph.14682

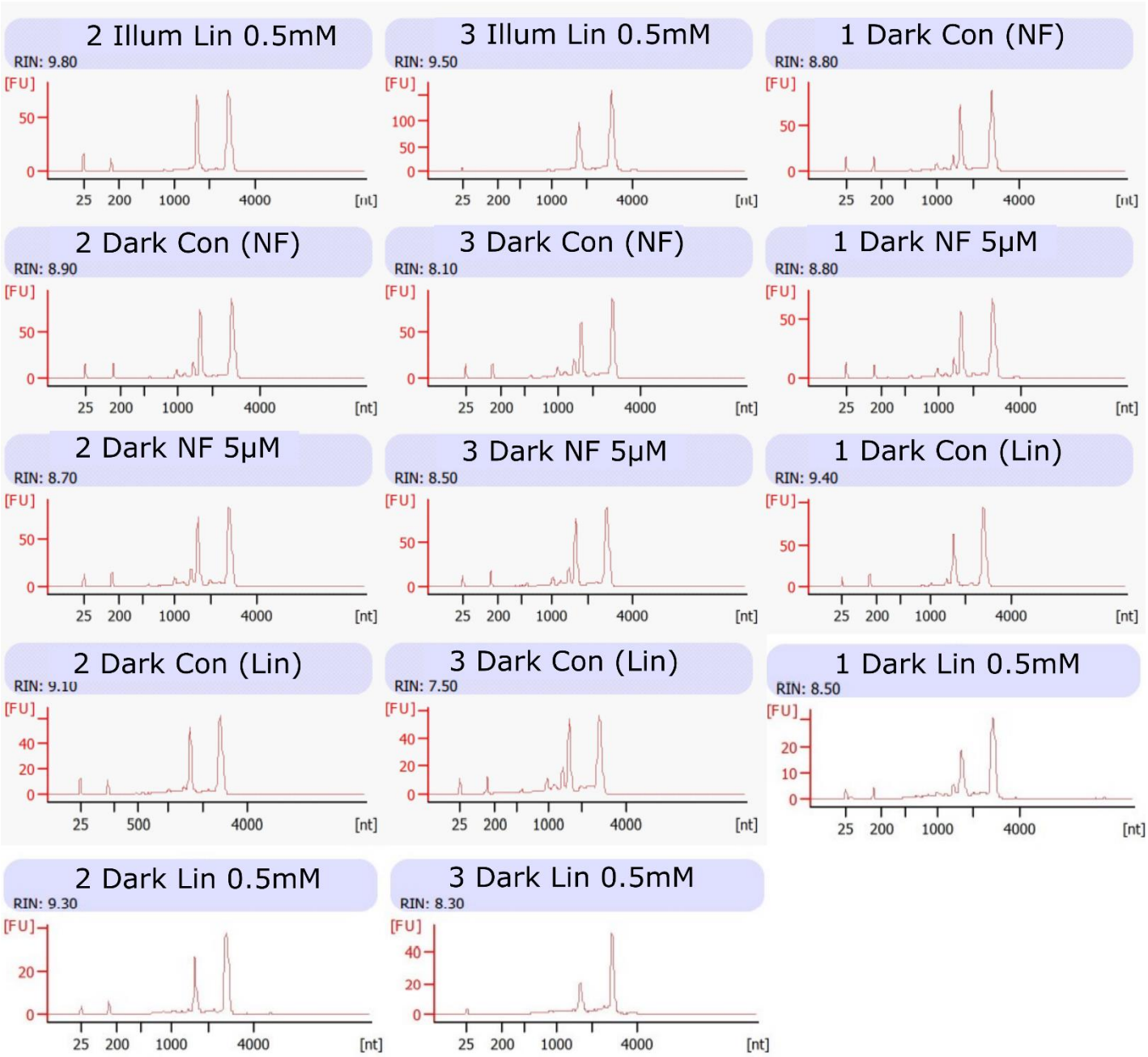
Wang, X., Gowik, U., Tang, H., Bowers, J. E., Westhoff, P., & Paterson, A. H. (2009). Comparative genomic analysis of C4 photosynthetic pathway evolution in grasses. *Genome Biology*, *10*(6), R68–R68. https://doi.org/10.1186/gb-2009-10-6-r68
Ward, J. M., Cufr, C. A., Denzel, M. A., & Neff, M. M. (2005). The Dof Transcription Factor OBP3 Modulates Phytochrome and Cryptochrome Signaling in Arabidopsis. *The Plant Cell*, *17*(2), 475–485. https://doi.org/10.1105/tpc.104.027722
Wilkins, O., Hafemeister, C., Plessis, A., Holloway-Phillips, M.-M., Pham, G. M., Nicotra, A. B., Gregorio, G. B., Jagadish, S. V. K., Septiningsih, E. M., Bonneau, R., & Purugganan, M. (2016). EGRINs (Environmental Gene Regulatory Influence Networks) in Rice That Function in the Response to Water Deficit, High Temperature, and Agricultural Environments. *The Plant Cell*, *28*(10), 2365–2384. https://doi.org/10.1105/tpc.16.00158
Williams, B. P., Aubry, S., & Hibberd, J. M. (2012). Molecular evolution of genes recruited into C 4 photosynthesis. In *Trends in Plant Science* (Vol. 17, Issue 4, pp. 213–220). Elsevier Ltd. https://doi.org/10.1016/j.tplants.2012.01.008
Windhövel, A., Hein, I., Dabrowa, R., & Stockhaus, J. (2001). Characterization of a novel class of plant homeodomain proteins that bind to the C4 phosphoenolpyruvate carboxylase gene of Flaveria trinervia. *Plant Molecular Biology*, *45*(2), 201–214. https://doi.org/10.1023/a:1006450005648
Wolfe, K. H., Morden, C. W., & Palmer, J. D. (1992). Function and evolution of a minimal plastid genome from a nonphotosynthetic parasitic plant. *Proceedings of the National Academy of Sciences*, *89*(22), 10648–10652. https://doi.org/10.1073/pnas.89.22.10648
Woodson, J. D., Perez-Ruiz, J. M., & Chory, J. (2011). Heme Synthesis by Plastid Ferrochelatase I Regulates Nuclear Gene Expression in Plants. *Current Biology*, *21*(10), 897–903. https://doi.org/10.1016/j.cub.2011.04.004
Xiao, Y., Savchenko, T., Baidoo, E. E. K., Chehab, W. E., Hayden, D. M., Tolstikov, V., Corwin, J. A., Kliebenstein, D. J., Keasling, J. D., & Dehesh, K. (2012). Retrograde Signaling by the Plastidial Metabolite MEcPP Regulates Expression of Nuclear Stress-Response Genes. *Cell*, *149*(7), 1525–1535. https://doi.org/10.1016/j.cell.2012.04.038
Yakandawala, N., Lupi, C., & Bilang, R. (2003). Lincomycin Treatment : A Simple Method to Differentiate Primary and Processed Transcripts in Rice (Oryza sativa L .) Chloroplasts. *Plant Molecular Biology Reporter*, *September*, 241–247.
Yu, J., Hu, S., Wang, J., Wong, G. K.-S., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y., Zhang, X., Cao, M., Liu, J., Sun, J., Tang, J., Chen, Y., Huang, X., Lin, W., Ye, C., Tong, W., … Yang, H. (2002). A Draft Sequence of the Rice Genome (Oryza sativa L. ssp. indica). *Science*, *296*(5565), 79 LP – 92. https://doi.org/10.1126/science.1068037
Zhang, Y. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biol*, *9*, R137.
Zhao, C., Wang, Y., Chan, K. X., Marchant, D. B., Franks, P. J., Randall, D., Tee, E. E., Chen, G., Ramesh, S., Phua, S. Y., Zhang, B., Hills, A., Dai, F., Xue, D., Gilliham, M., Tyerman, S., Nevo, E., Wu, F., Zhang, G., … Chen, Z.-H. (2019). Evolution of chloroplast retrograde signaling facilitates green plant adaptation to land. *Proceedings of the National Academy of Sciences*, *116*(11), 5015–5020. https://doi.org/10.1073/pnas.1812092116
Zhao, X., Huang, J., & Chory, J. (2018). genome uncoupled1 Mutants Are Hypersensitive to Norflurazon and Lincomycin. *Plant Physiology*, *178*(3), 960–964. https://doi.org/10.1104/pp.18.00772
Zhao, X., Huang, J., & Chory, J. (2020). Unraveling the Linkage between Retrograde Signaling and RNA Metabolism in Plants. *Trends in Plant Science*, *25*(2), 141–147. https://doi.org/10.1016/j.tplants.2019.10.009
Zuccolo, A., Ammiraju, J. S. S., Kim, H., Sanyal, A., Jackson, S., & Wing, R. A. (2008). Rapid and Differential Proliferation of the Ty3-Gypsy LTR Retrotransposon Atlantys in the Genus Oryza. *Rice*, *1*(1), 85–99. https://doi.org/10.1007/s12284-008-9002-y

Zulfugarov, I. S., Tovuu, A., Eu, Y.-J., Dogsom, B., Poudyal, R. S., Nath, K., Hall, M., Banerjee, M., Yoon, U. C., Moon, Y.-H., An, G., Jansson, S., & Lee, C.-H. (2014). Production of superoxide from Photosystem II in a rice (Oryza sativaL.) mutant lacking PsbS. *BMC Plant Biology*, *14*(1), 242. https://doi.org/10.1186/s12870-014-0242-2

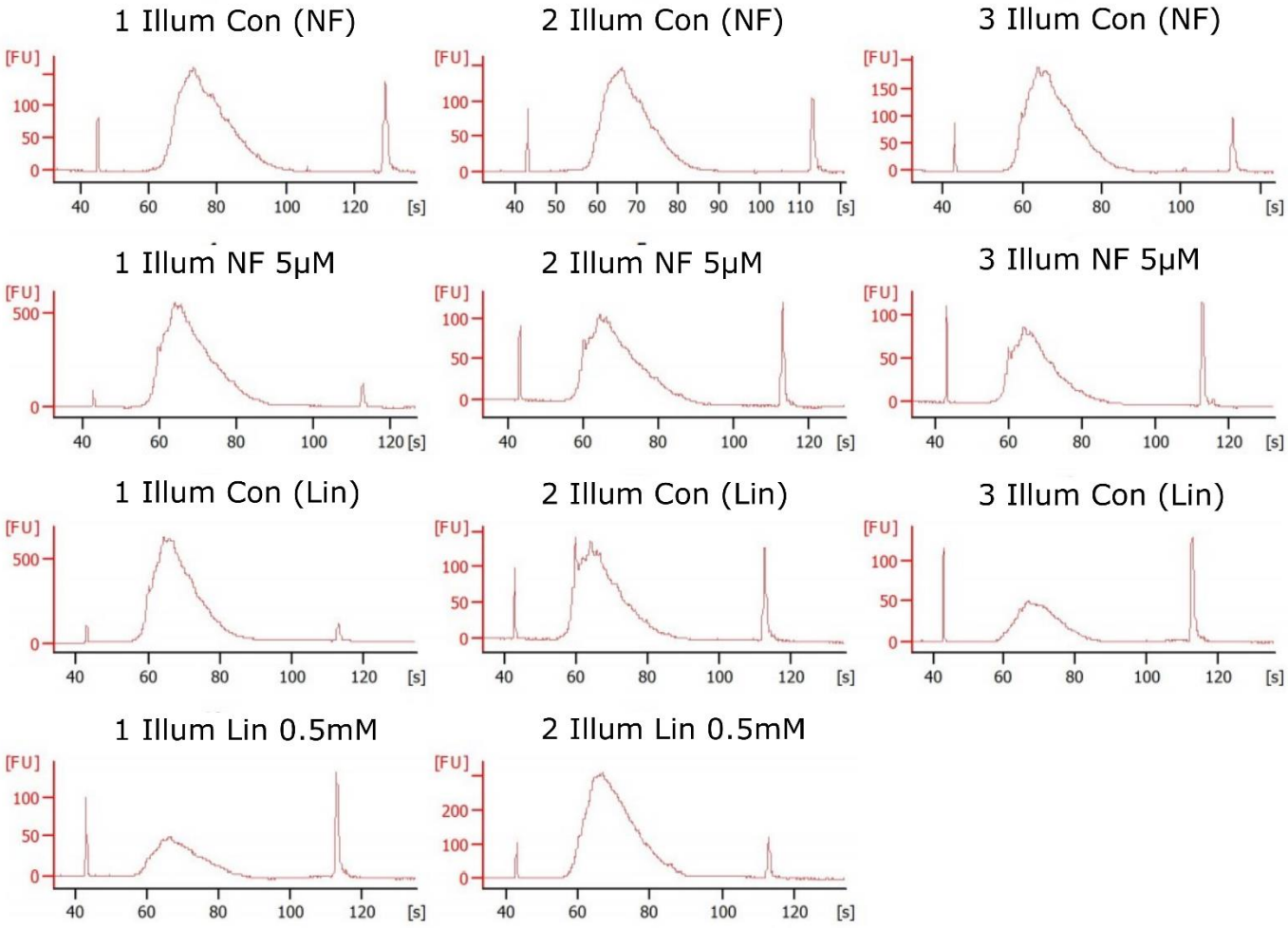
8 APPENDICES

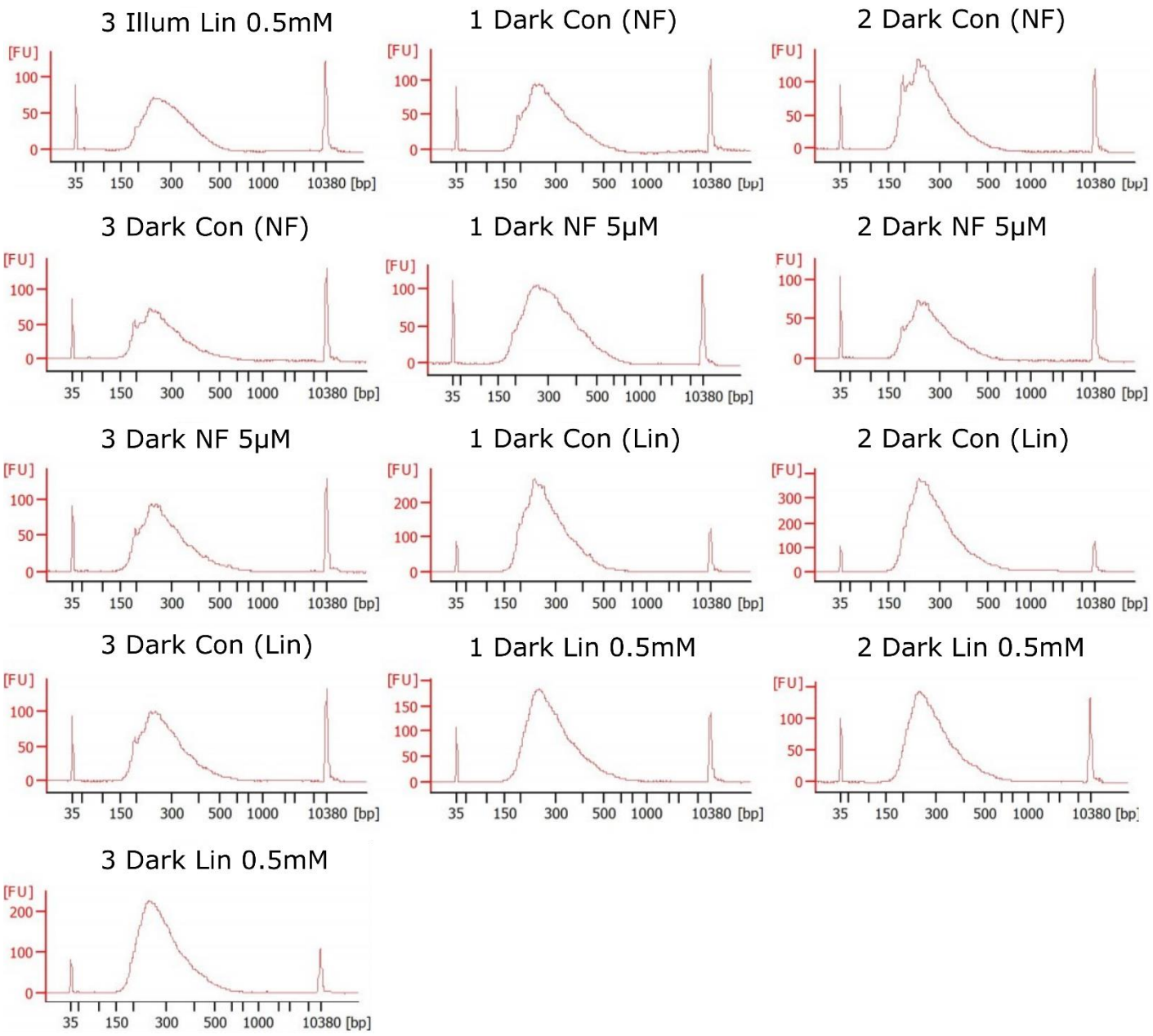
APPENDIX 1: ELECTROPHEROGRAMS OF *A. THALIANA* RNA




2 Illum Lin 0.5mM
RIN: 9.80
3 Illum Lin 0.5mM
RIN: 9.50
1 Dark Con (NF)
RIN: 8.80
2 Dark Con (NF)
RIN: 8.90
3 Dark Con (NF)
RIN: 8.10
1 Dark NF 5µM
RIN: 8.80
2 Dark NF 5µM
RIN: 8.70
3 Dark NF 5µM
RIN: 8.50
1 Dark Con (Lin)
RIN: 9.40
2 Dark Con (Lin)
RIN: 9.10
3 Dark Con (Lin)
RIN: 7.50
1 Dark Lin 0.5mM
RIN: 8.50
2 Dark Lin 0.5mM
RIN: 9.30
3 Dark Lin 0.5mM
RIN: 8.30
[FU]
[nt]

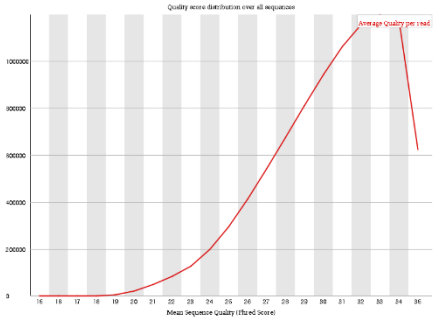
APPENDIX 2: ELECTROPHEROGRAMS OF *A. THALIANA* RNA-SEQ LIBRARIES



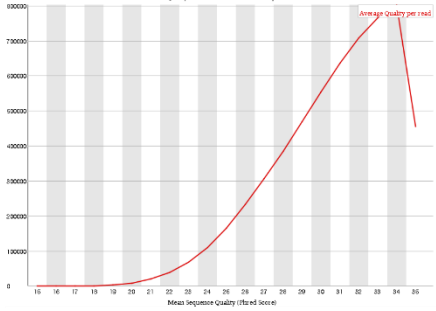

3 Illum Lin 0.5mM
1 Dark Con (NF)
2 Dark Con (NF)
3 Dark Con (NF)
1 Dark NF 5µM
2 Dark NF 5µM
3 Dark NF 5µM
1 Dark Con (Lin)
2 Dark Con (Lin)
3 Dark Con (Lin)
1 Dark Lin 0.5mM
2 Dark Lin 0.5mM
3 Dark Lin 0.5mM
[FU]
35 150 300 500 1000 10380 [bp]

APPENDIX 3: FASTQC REPORTS OF MEAN SEQUENCE QUALITY FOR *A. THALIANA* RNA-SEQ READS

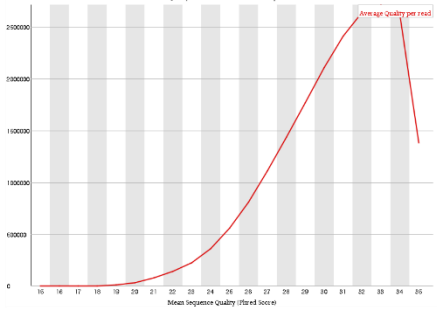
1 Illum Con (Lin)



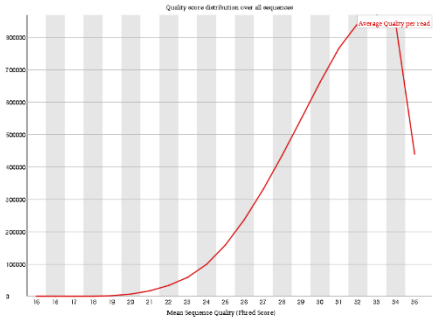
2 Illum Con (Lin)



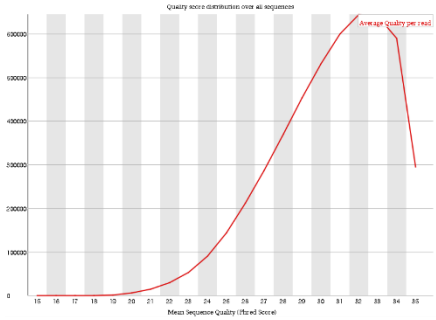
3 Illum Con (Lin)



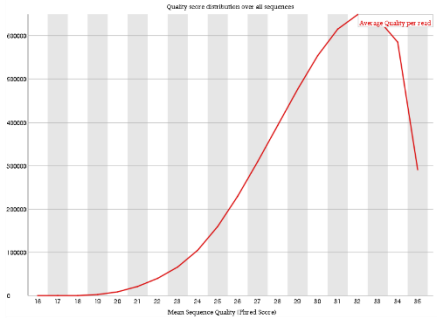
1 Illum Lin 0.5mM



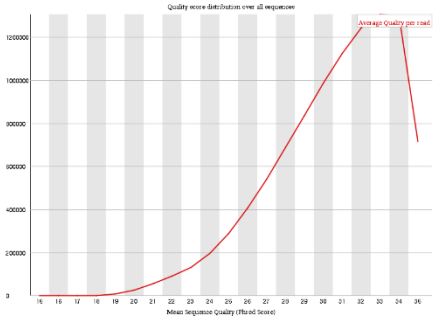
2 Illum Lin 0.5mM



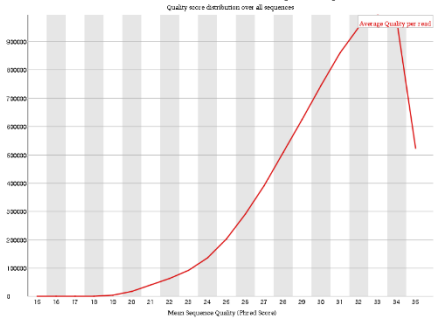
3 Illum Lin 0.5mM



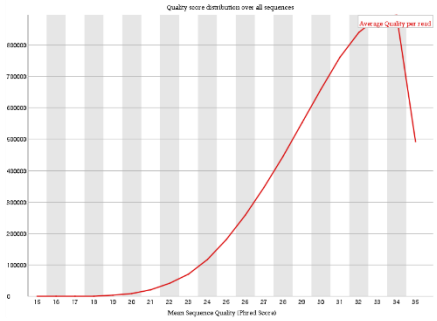
1 Dark Con (Lin)



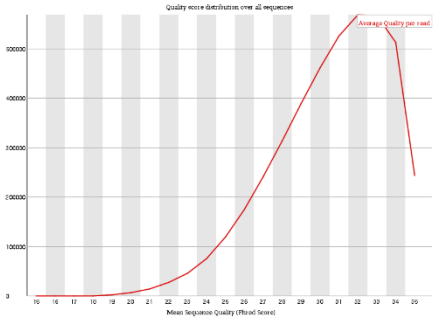
2 Dark Con (Lin)



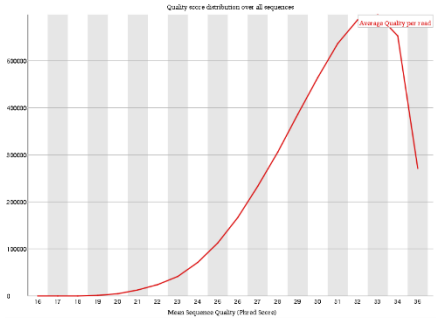
3 Dark Con (Lin)



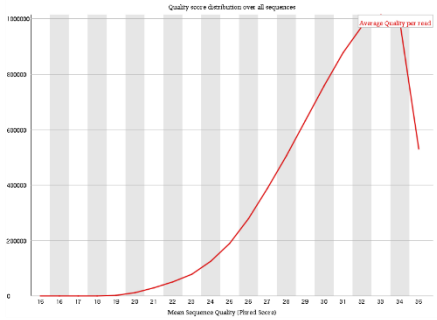
1 Dark Lin 0.5mM



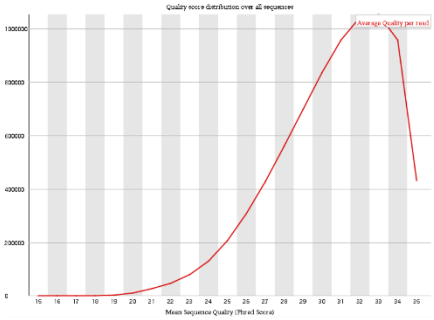
2 Dark Lin 0.5mM



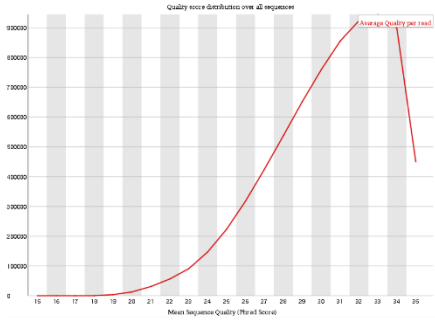
3 Dark Lin 0.5mM



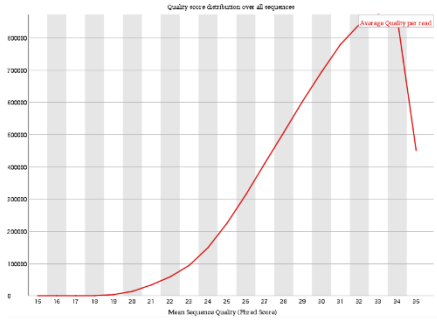
1 Illum Con (NF)



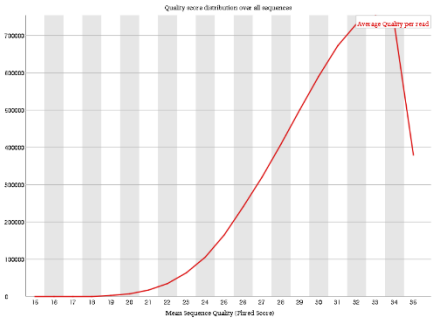
2 Illum Con (NF)



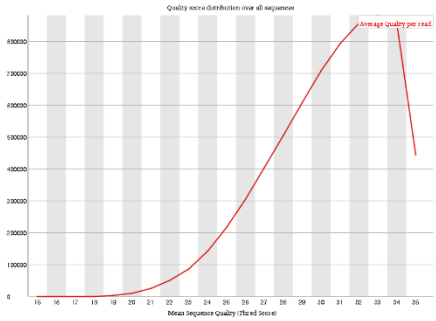
3 Illum Con (NF)



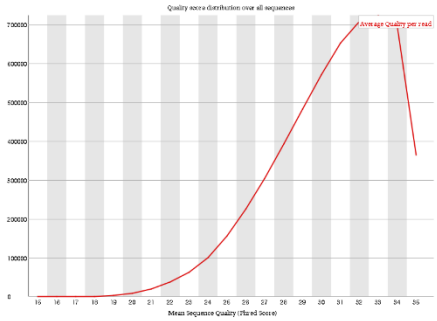
1 Illum NF 5µM



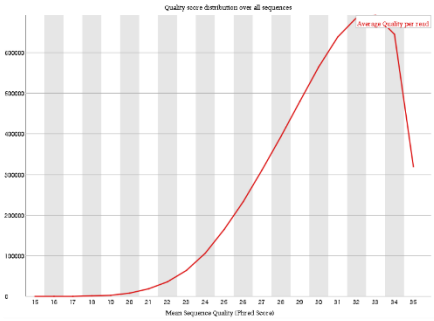
2 Illum NF 5µM



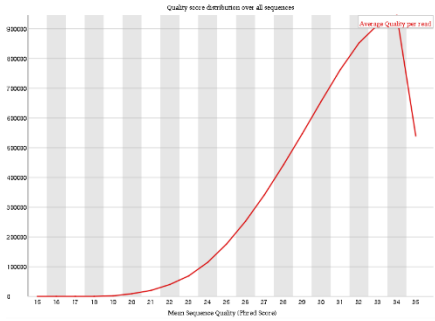
3 Illum NF 5µM



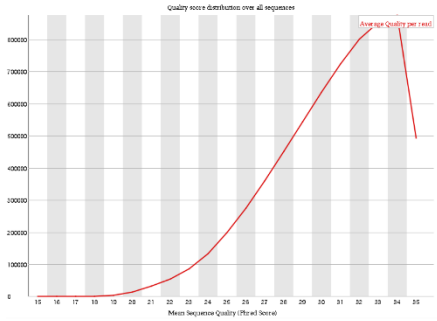
1 Dark Con (NF)



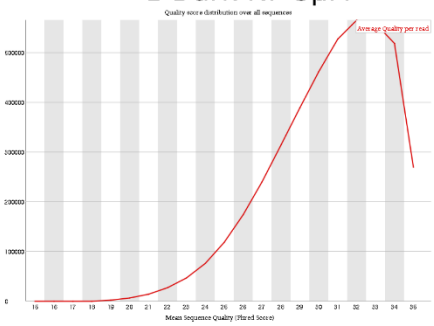
2 Dark Con (NF)



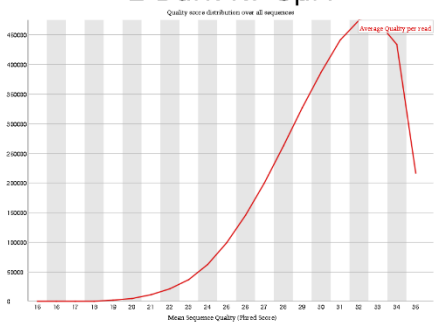
3 Dark Con (NF)



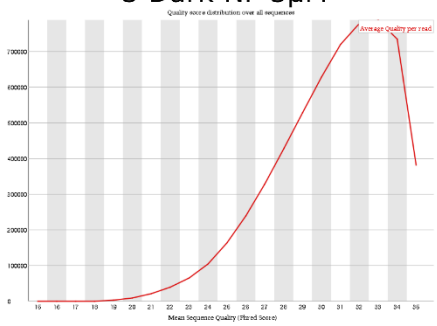
1 Dark NF 5µM



2 Dark NF 5µM



3 Dark NF 5µM



APPENDIX 4: THE 100 GENES WITH THE HIGHEST ABSOLUTE LOG2FOLD CHANGE IN TRANSCRIPT ABUNDANCE BETWEEN ILLUMINATED CONTROLS AND LINCOMYCIN TREATED *A. THALIANA* SEEDLINGS

Rank	TAIR Gene ID	Log2Fold Change	Adjusted p-value
1	AT1G36180	4.13	3.92E-43
2	AT1G73870	-4.03	8.35E-22
3	AT2G18193	3.78	5.37E-41
4	AT2G27420	-3.53	8.22E-23
5	ATCG00090	-3.16	2.62E-32
6	AT1G05560	-3.13	2.07E-24
7	ATCG00080	-3.10	3.42E-24
8	AT3G58270	3.03	1.77E-24
9	ATCG00690	-3.03	3.10E-12
10	AT5G55270	3.01	1.78E-09
11	AT3G54520	2.97	1.23E-09
12	AT1G17960	2.92	8.13E-21
13	AT1G30170	2.91	1.01E-08
14	ATCG00300	-2.90	1.77E-24
15	AT5G24280	2.78	1.21E-12
16	AT5G03495	2.75	1.01E-08
17	ATCG00730	-2.72	6.10E-23
18	AT5G64060	2.67	4.30E-08
19	AT2G04050	2.65	1.96E-17
20	AT2G18190	2.63	3.96E-07
21	AT2G15620	-2.63	9.95E-22
22	AT3G06145	-2.61	1.99E-11
23	AT1G72930	-2.56	3.50E-08
24	AT5G60250	2.55	4.59E-07
25	ATCG01010	-2.52	1.25E-12
26	AT3G16660	2.51	6.29E-17
27	AT4G09650	-2.50	4.08E-18
28	ATCG00310	-2.48	3.31E-13
29	AT1G32900	-2.45	4.61E-08
30	AT4G10120	-2.45	3.79E-11
31	AT3G48310	-2.43	9.68E-07
32	AT2G02240	2.43	6.83E-07
33	ATCG00330	-2.42	6.03E-20

Rank	TAIR Gene ID	Log2Fold Change	Adjusted p-value
34	ATCG00420	-2.41	2.06E-18
35	AT5G58610	2.41	7.07E-06
36	AT5G58770	-2.39	1.28E-11
37	ATCG00560	-2.39	6.42E-06
38	ATCG00070	-2.39	3.25E-11
39	AT1G72920	-2.37	4.34E-06
40	AT1G05490	2.37	1.85E-07
41	ATCG00930	-2.36	1.94E-06
42	ATCG00680	-2.36	4.76E-06
43	ATCG00490	-2.36	9.63E-06
44	AT3G01550	-2.35	4.18E-06
45	AT5G52570	-2.31	1.66E-08
46	AT1G52315	2.29	2.92E-08
47	ATCG00270	-2.28	1.46E-05
48	AT2G25510	-2.26	6.10E-06
49	AT1G65190	-2.25	4.02E-16
50	AT2G34620	-2.23	3.18E-05
51	ATCG00440	-2.23	2.07E-24
52	AT2G21640	2.23	1.58E-13
53	ATCG00220	-2.22	2.82E-16
54	AT3G05790	2.20	2.75E-05
55	AT3G27630	2.20	1.73E-08
56	AT4G13680	2.19	6.40E-05
57	AT3G46650	-2.18	3.44E-05
58	AT1G18730	-2.16	2.83E-13
59	AT5G06760	-2.14	4.44E-11
60	ATCG00020	-2.13	9.99E-05
61	ATCG00720	-2.13	4.17E-05
62	AT1G70985	-2.08	1.29E-06
63	AT1G78995	-2.08	7.12E-10
64	AT1G26761	-2.06	1.55E-07
65	ATCG00410	-2.06	8.36E-06
66	AT1G25440	-2.06	0.000115
67	AT5G12860	-2.06	2.15E-13

Rank	TAIR Gene ID	Log2Fold Change	Adjusted p-value
68	AT1G07050	-2.05	2.75E-05
69	AT2G27402	2.05	3.24E-07
70	ATCG00430	-2.04	3.76E-10
71	AT1G65870	-2.03	0.000289
72	AT1G32080	-2.02	8.90E-11
73	AT1G77760	-2.02	6.43E-05
74	AT2G03750	-2.02	1.97E-07
75	AT1G15870	2.01	0.000365
76	AT1G68238	-2.00	0.000156
77	AT1G02820	-1.98	1.25E-08
78	AT4G09010	-1.97	3.23E-10
79	AT1G10657	-1.97	0.000429
80	AT3G59400	-1.96	2.38E-10
81	AT5G17230	-1.95	1.67E-09
82	AT3G54050	-1.95	1.48E-09
83	AT3G63140	-1.94	3.30E-09
84	AT5G14490	1.94	0.000681
85	AT3G10060	-1.94	4.40E-13
86	AT1G15980	-1.93	2.59E-10
87	AT2G36750	-1.93	1.60E-06
88	AT3G46890	1.90	8.95E-09
89	AT5G48490	-1.90	5.49E-07
90	AT1G64780	-1.89	5.90E-07
91	AT3G24460	-1.89	2.48E-05
92	AT5G22390	-1.89	2.92E-08
93	AT4G28080	-1.88	4.56E-11
94	AT3G29320	-1.87	0.000194
95	AT1G30160	1.86	0.00028
96	AT1G19150	-1.85	0.000429
97	AT1G33170	-1.85	1.96E-05
98	AT5G65010	-1.85	3.45E-10
99	AT4G37560	-1.84	8.19E-05
100	AT4G18650	-1.84	0.000296

APPENDIX 5: THE 100 GENES WITH THE HIGHEST ABSOLUTE LOG2FOLD CHANGE IN TRANSCRIPT ABUNDANCE BETWEEN ILLUMINATED CONTROLS AND NORFLURAZON TREATED *A. THALIANA* SEEDLINGS

Rank	TAIR Gene ID	Log2Fold Change	Adjusted p-value
1	AT2G34430	-8.41002	8E-205
2	AT1G29920	-8.40594	0.0E+00
3	AT1G29910	-6.98438	2.9E-221
4	AT2G05070	-6.5596	1.5E-145
5	AT4G05180	-6.13496	8.7E-191
6	AT1G37130	-6.03146	1.4E-161
7	AT5G54270	-5.81616	1.6E-172
8	AT1G44575	-5.79209	1.72E-94
9	AT5G38420	-5.68052	4.6E-113
10	AT2G34620	-5.56977	1.36E-32
11	AT2G34420	-5.34452	2.2E-138
12	AT3G25770	-5.17355	7.81E-20
13	AT1G77760	-5.08034	7.2E-67
14	AT3G01500	-5.04872	3.4E-82
15	AT5G24420	-5.03994	3.41E-21
16	AT1G29930	-5.03255	1E-133
17	AT5G25610	-4.99004	6.87E-59
18	AT1G73870	-4.76261	8.56E-13
19	AT1G32900	-4.69676	6.64E-15
20	AT3G27690	-4.6928	1.65E-61
21	AT4G33550	-4.61511	7.55E-22
22	AT3G08940	-4.55661	1.4E-158
23	AT1G23130	-4.46111	2.77E-75
24	AT2G42220	-4.46014	2.8E-53
25	AT1G15820	-4.45553	6.7E-123
26	AT4G26530	-4.3455	1.57E-78
27	AT1G25440	-4.21134	2.25E-41
28	AT1G52230	-4.12939	1.46E-90
29	AT5G35630	-4.0748	6.29E-65
30	AT1G21500	-4.06509	6.04E-50
31	AT2G05100	-4.02807	8.2E-63
32	AT4G35160	-4.02222	2.11E-10
33	AT3G52720	-4.00631	4.14E-12

Rank	TAIR Gene ID	Log2Fold Change	Adjusted p-value
34	AT2G15620	-4.00596	1.68E-63
35	AT4G28706	-3.97255	4.07E-15
36	AT1G08380	-3.94209	1.3E-126
37	AT1G68238	-3.93194	NA
38	AT5G53870	-3.89069	5.85E-89
39	AT1G61300	-3.81899	5.1E-29
40	AT3G45060	3.785857	4.94E-08
41	AT1G70820	-3.75207	6.15E-21
42	AT4G10340	-3.71743	2.18E-67
43	AT1G06080	-3.70006	3.99E-15
44	AT1G03130	-3.69459	5.49E-62
45	AT2G15090	-3.6774	3.43E-11
46	AT2G32550	-3.66486	8.69E-22
47	AT5G20740	-3.64376	1.73E-10
48	AT3G54890	-3.63359	1.12E-88
49	ATCG00030	-3.63306	2.07E-57
50	AT3G21055	-3.6168	1.23E-68
51	AT3G48310	-3.61586	7.52E-08
52	AT3G53980	-3.58543	2.15E-08
53	AT5G11920	3.582803	8.46E-41
54	AT3G19480	-3.56701	1.82E-18
55	AT5G65010	-3.54837	4.56E-30
56	AT5G66400	-3.54555	1.05E-83
57	AT2G08950	-3.53476	5.07E-12
58	AT5G20630	-3.53104	6.81E-58
59	AT4G27520	-3.51658	3.68E-44
60	AT2G21210	-3.49444	NA
61	AT1G02205	-3.48201	2.89E-10
62	AT5G12860	-3.47144	8.43E-61
63	AT1G76100	-3.46548	1.23E-68
64	AT3G01550	-3.44718	3.72E-07
65	AT1G52690	-3.43991	9.13E-29
66	AT5G03760	-3.43211	2.32E-29
67	AT5G59320	-3.4157	7.5E-07

Rank	TAIR Gene ID	Log2Fold Change	Adjusted p-value
68	AT3G11090	-3.3864	1.66E-30
69	AT4G28250	-3.33459	1.89E-15
70	AT3G47470	-3.32971	3.39E-86
71	AT5G62720	-3.30752	9.45E-22
72	AT1G56190	-3.30196	6.73E-20
73	AT1G53520	-3.29841	1.38E-15
74	AT3G06145	-3.29382	3.74E-10
75	AT5G06790	-3.27309	3.21E-13
76	AT3G62570	-3.25613	3.78E-10
77	AT3G16140	-3.25023	3.07E-32
78	AT1G31330	-3.24487	6.56E-50
79	AT2G32100	-3.19359	4.64E-09
80	AT4G28750	-3.18565	4.33E-51
81	AT4G18650	-3.17752	1.81E-05
82	AT2G05790	-3.15706	9.86E-23
83	AT3G48280	-3.15533	1.39E-06
84	AT1G55670	-3.14899	2.13E-37
85	AT5G16030	-3.14222	7.53E-27
86	AT5G52300	-3.1319	2.04E-50
87	AT2G10940	-3.12976	1.18E-30
88	AT2G35960	-3.12962	2.23E-05
89	AT3G46780	-3.12832	7.89E-42
90	AT4G02770	-3.12595	1.35E-44
91	AT4G37800	-3.12291	1.16E-50
92	AT5G04370	-3.11933	1.03E-05
93	AT1G67870	-3.09617	1.34E-27
94	ATCG00350	-3.09476	6.71E-26
95	AT2G34060	-3.09332	4.04E-23
96	AT3G18890	-3.08392	9.68E-26
97	AT5G22390	-3.05762	7.17E-12
98	AT3G29320	-3.03966	2.45E-09
99	AT1G52770	-3.03909	2.7E-06
100	AT2G35260	-3.0377	1.14E-53

APPENDIX 6: LIST OF ENRICHED GENE ONTOLOGY (GO) TERMS IN *A. THALIANA* GENES THAT SHOW SIGNIFICANTLY DIFFERENT TRANSCRIPT ABUNDANCE IN RESPONSE TO LINCOMYCIN TREATMENT

GO term	Ontology	Description	Number in input list	Number in BG/Ref	p-value	FDR
GO:0015979	P	photosynthesis	121	253	8.3E-84	3.7E-80
GO:0019684	P	photosynthesis, light reaction	74	134	1.1E-54	2.4E-51
GO:0006091	P	generation of precursor metabolites and energy	99	364	6.1E-50	9E-47
GO:0055114	P	oxidation-reduction process	168	1566	2.3E-35	2.6E-32
GO:0044710	P	single-organism metabolic process	288	3970	2.6E-32	2.3E-29
GO:0009767	P	photosynthetic electron transport chain	36	56	6.6E-29	4.9E-26
GO:0009628	P	response to abiotic stimulus	178	2022	9.3E-28	5.9E-25
GO:0022900	P	electron transport chain	41	112	2.6E-25	1.5E-22
GO:0009765	P	photosynthesis, light harvesting	27	44	1.3E-21	6.2E-19
GO:0042221	P	response to chemical	199	2853	1.1E-19	5E-17
GO:0050896	P	response to stimulus	344	6250	2E-18	8E-16
GO:0006778	P	porphyrin-containing compound metabolic process	31	97	4.5E-18	1.7E-15
GO:0033013	P	tetrapyrrole metabolic process	31	98	5.8E-18	2E-15
GO:0015994	P	chlorophyll metabolic process	29	85	1.3E-17	4E-15
GO:0009416	P	response to light stimulus	81	744	1.6E-17	4.7E-15
GO:0009409	P	response to cold	56	384	2.6E-17	7.1E-15
GO:0009314	P	response to radiation	81	772	1.1E-16	2.9E-14
GO:0009266	P	response to temperature stimulus	67	559	1.6E-16	3.8E-14
GO:0009735	P	response to cytokinin	43	244	3.7E-16	8.7E-14
GO:0051186	P	cofactor metabolic process	54	389	6.3E-16	1.4E-13
GO:0044281	P	small molecule metabolic process	124	1597	2.3E-15	4.8E-13
GO:0046148	P	pigment biosynthetic process	33	148	3.2E-15	6.3E-13
GO:0042440	P	pigment metabolic process	36	181	3.8E-15	7E-13
GO:0018298	P	protein-chromophore linkage	21	46	3.8E-15	7E-13
GO:0006779	P	porphyrin-containing compound biosynthetic process	24	68	4.3E-15	7.6E-13
GO:0015995	P	chlorophyll biosynthetic process	23	61	5.1E-15	8.6E-13
GO:0044711	P	single-organism biosynthetic process	130	1735	5.5E-15	9E-13
GO:0010035	P	response to inorganic substance	86	926	8.6E-15	1.4E-12
GO:0009642	P	response to light intensity	32	144	9E-15	1.4E-12
GO:0033014	P	tetrapyrrole biosynthetic process	24	71	9.5E-15	1.4E-12
GO:0008152	P	metabolic process	547	12035	2.5E-14	3.6E-12
GO:0006950	P	response to stress	209	3506	1E-13	1.4E-11
GO:0009657	P	plastid organization	36	207	1.4E-13	1.8E-11
GO:0009773	P	photosynthetic electron transport in photosystem I	14	16	2.4E-13	3.1E-11
GO:0006979	P	response to oxidative stress	53	453	5.8E-13	7.3E-11
GO:0044699	P	single-organism process	444	9448	9.5E-13	1.2E-10
GO:0051188	P	cofactor biosynthetic process	36	229	1.9E-12	2.3E-10
GO:0009987	P	cellular process	563	12771	2.8E-12	3.2E-10
GO:0044237	P	cellular metabolic process	457	9879	3.9E-12	4.4E-10
GO:0044763	P	single-organism cellular process	324	6591	8.4E-11	9.3E-09
GO:0009658	P	chloroplast organization	26	150	3.5E-10	3.8E-08
GO:0009644	P	response to high light intensity	19	76	4.5E-10	4.5E-08
GO:1901700	P	response to oxygen-containing compound	106	1557	4.4E-10	4.5E-08
GO:0019685	P	photosynthesis, dark reaction	12	21	4.3E-10	4.5E-08
GO:0010207	P	photosystem II assembly	12	22	6.5E-10	6.4E-08
GO:0009768	P	photosynthesis, light harvesting in photosystem I	12	23	9.5E-10	9.2E-08
GO:0015977	P	carbon fixation	12	24	1.4E-09	1.3E-07
GO:0009414	P	response to water deprivation	39	347	1.9E-09	1.8E-07
GO:0019253	P	reductive pentose-phosphate cycle	11	20	3.2E-09	2.9E-07
GO:0043436	P	oxoacid metabolic process	74	973	3.3E-09	2.9E-07

GO term	Ontology	Description	Number in input list	Number in BG/Ref	p-value	FDR
GO:0009415	P	response to water	39	355	3.4E-09	3E-07
GO:1901564	P	organonitrogen compound metabolic process	148	2569	7.5E-09	6.4E-07
GO:0010109	P	regulation of photosynthesis	13	39	1.5E-08	1.2E-06
GO:0043623	P	cellular protein complex assembly	27	197	1.5E-08	1.2E-06
GO:0005982	P	starch metabolic process	16	66	1.5E-08	1.2E-06
GO:0010206	P	photosystem II repair	9	13	2E-08	1.6E-06
GO:0006082	P	organic acid metabolic process	78	1099	2E-08	1.6E-06
GO:0019752	P	carboxylic acid metabolic process	66	864	2.1E-08	1.6E-06
GO:0042548	P	regulation of photosynthesis, light reaction	11	26	2.6E-08	0.000002
GO:0010033	P	response to organic substance	121	2023	3.2E-08	2.3E-06
GO:0009725	P	response to hormone	103	1631	3.1E-08	2.3E-06
GO:0043467	P	regulation of generation of precursor metabolites and energy	11	28	4.9E-08	3.5E-06
GO:0045454	P	cell redox homeostasis	22	147	8.2E-08	5.8E-06
GO:0009719	P	response to endogenous stimulus	106	1732	8.6E-08	5.9E-06
GO:0001101	P	response to acid chemical	80	1181	8.5E-08	5.9E-06
GO:0046686	P	response to cadmium ion	35	343	1.1E-07	7.6E-06
GO:0044283	P	small molecule biosynthetic process	52	653	2E-07	0.000014
GO:0019252	P	starch biosynthetic process	11	35	3.1E-07	0.00002
GO:0034599	P	cellular response to oxidative stress	17	97	3.4E-07	0.000022
GO:0009772	P	photosynthetic electron transport in photosystem II	7	10	7.7E-07	0.000048
GO:0010038	P	response to metal ion	41	481	8.2E-07	0.000051
GO:0070271	P	protein complex biogenesis	30	294	9E-07	0.000055
GO:0030091	P	protein repair	9	24	1.1E-06	0.000068
GO:0006112	P	energy reserve metabolic process	8	18	1.6E-06	0.000096
GO:0005977	P	glycogen metabolic process	8	18	1.6E-06	0.000096
GO:0019682	P	glyceraldehyde-3-phosphate metabolic process	11	44	2.1E-06	0.00012
GO:0043155	P	negative regulation of photosynthesis, light reaction	7	14	4.1E-06	0.00023
GO:0010205	P	photoinhibition	7	14	4.1E-06	0.00023
GO:0006461	P	protein complex assembly	28	287	4.5E-06	0.00025
GO:0016108	P	tetraterpenoid metabolic process	10	42	8.9E-06	0.00049
GO:0016116	P	carotenoid metabolic process	10	42	8.9E-06	0.00049
GO:0006970	P	response to osmotic stress	46	636	0.000011	0.00059
GO:1905156	P	negative regulation of photosynthesis	7	17	0.000011	0.00059
GO:0006081	P	cellular aldehyde metabolic process	12	65	0.000011	0.00059
GO:0016109	P	tetraterpenoid biosynthetic process	9	34	0.000012	0.00062
GO:0016117	P	carotenoid biosynthetic process	9	34	0.000012	0.00062
GO:1901566	P	organonitrogen compound biosynthetic process	111	2076	0.000018	0.0009
GO:0000302	P	response to reactive oxygen species	19	166	0.000021	0.001
GO:0032787	P	monocarboxylic acid metabolic process	37	481	0.000023	0.0012
GO:0018198	P	peptidyl-cysteine modification	7	20	0.000026	0.0013
GO:0016123	P	xanthophyll biosynthetic process	5	7	0.00003	0.0015
GO:0006006	P	glucose metabolic process	9	40	0.000037	0.0018
GO:0016122	P	xanthophyll metabolic process	6	14	0.00004	0.0019
GO:1901605	P	alpha-amino acid metabolic process	23	242	0.000045	0.0021
GO:0042549	P	photosystem II stabilization	5	8	0.000047	0.0022
GO:0010114	P	response to red light	11	65	0.000053	0.0024
GO:0046501	P	protoporphyrinogen IX metabolic process	6	15	0.000055	0.0025
GO:0006782	P	protoporphyrinogen IX biosynthetic process	6	15	0.000055	0.0025
GO:0019637	P	organophosphate metabolic process	38	524	0.000059	0.0026
GO:0006783	P	heme biosynthetic process	7	24	0.000068	0.003
GO:0006520	P	cellular amino acid metabolic process	27	320	0.00007	0.003
GO:0071822	P	protein complex subunit organization	28	338	0.000069	0.003
GO:0017014	P	protein nitrosylation	6	16	0.000073	0.0031
GO:0018119	P	peptidyl-cysteine S-nitrosylation	6	16	0.000073	0.0031
GO:0009853	P	photorespiration	10	56	0.000077	0.0032
GO:0072593	P	reactive oxygen species metabolic process	18	169	0.00008	0.0033
GO:0006073	P	cellular glucan metabolic process	21	218	0.00008	0.0033
GO:0044042	P	glucan metabolic process	21	218	0.00008	0.0033
GO:0016051	P	carbohydrate biosynthetic process	28	342	0.000083	0.0034
GO:0009737	P	response to abscisic acid	40	578	0.000098	0.0039
GO:0034622	P	cellular macromolecular complex assembly	29	365	0.0001	0.0041
GO:0009769	P	photosynthesis, light harvesting in photosystem II	5	10	0.0001	0.0041
GO:0009064	P	glutamine family amino acid metabolic process	11	71	0.00011	0.0042
GO:0008610	P	lipid biosynthetic process	37	522	0.00011	0.0044
GO:0097305	P	response to alcohol	40	583	0.00012	0.0045
GO:1901135	P	carbohydrate derivative metabolic process	43	645	0.00012	0.0045
GO:0005996	P	monosaccharide metabolic process	13	99	0.00012	0.0045
GO:0019725	P	cellular homeostasis	26	314	0.00012	0.0047
GO:0015980	P	energy derivation by oxidation of organic compounds	16	146	0.00014	0.0054
GO:0010027	P	thylakoid membrane organization	8	38	0.00015	0.0055
GO:0006732	P	coenzyme metabolic process	23	264	0.00015	0.0055

GO term	Ontology	Description	Number in input list	Number in BG/Ref	p-value	FDR
GO:0042168	P	heme metabolic process	7	28	0.00015	0.0056
GO:0009636	P	response to toxic substance	13	102	0.00016	0.0056
GO:0009117	P	nucleotide metabolic process	25	302	0.00017	0.006
GO:0006739	P	NADP metabolic process	8	39	0.00017	0.0061
GO:0009668	P	plastid membrane organization	8	39	0.00017	0.0061
GO:0006753	P	nucleoside phosphate metabolic process	25	305	0.00019	0.0067
GO:0006094	P	gluconeogenesis	6	20	0.0002	0.0069
GO:0005978	P	glycogen biosynthetic process	5	12	0.0002	0.0069
GO:0016114	P	terpenoid biosynthetic process	14	120	0.00021	0.007
GO:0006721	P	terpenoid metabolic process	15	137	0.00023	0.0079
GO:0008299	P	isoprenoid biosynthetic process	16	154	0.00025	0.0085
GO:0044255	P	cellular lipid metabolic process	42	649	0.00026	0.0086
GO:0006662	P	glycerol ether metabolic process	7	31	0.00027	0.0087
GO:0018904	P	ether metabolic process	7	31	0.00027	0.0087
GO:0006720	P	isoprenoid metabolic process	17	172	0.00028	0.0091
GO:0055086	P	nucleobase-containing small molecule metabolic process	28	370	0.00028	0.0091
GO:0009637	P	response to blue light	11	81	0.0003	0.0096
GO:0072524	P	pyridine-containing compound metabolic process	14	125	0.0003	0.0096
GO:0044723	P	single-organism carbohydrate metabolic process	40	613	0.00031	0.0097
GO:0009167	P	purine ribonucleoside monophosphate metabolic process	17	174	0.00032	0.0099
GO:0009126	P	purine nucleoside monophosphate metabolic process	17	174	0.00032	0.0099
GO:0009161	P	ribonucleoside monophosphate metabolic process	18	191	0.00032	0.01
GO:0019693	P	ribose phosphate metabolic process	21	245	0.00035	0.011
GO:0042744	P	hydrogen peroxide catabolic process	11	83	0.00036	0.011
GO:0006541	P	glutamine metabolic process	7	33	0.00037	0.011
GO:0080167	P	response to karrikin	14	128	0.00038	0.011
GO:0009123	P	nucleoside monophosphate metabolic process	18	196	0.00043	0.013
GO:0009269	P	response to desiccation	6	24	0.00046	0.014
GO:0019319	P	hexose biosynthetic process	6	24	0.00046	0.014
GO:0019318	P	hexose metabolic process	10	72	0.00048	0.014
GO:0006631	P	fatty acid metabolic process	21	252	0.0005	0.014
GO:0065003	P	macromolecular complex assembly	30	424	0.0005	0.014
GO:0019362	P	pyridine nucleotide metabolic process	13	117	0.00052	0.015
GO:0010218	P	response to far red light	9	60	0.00055	0.016
GO:0071214	P	cellular response to abiotic stimulus	18	202	0.00059	0.017
GO:0009150	P	purine ribonucleotide metabolic process	18	202	0.00059	0.017
GO:0042743	P	hydrogen peroxide metabolic process	12	104	0.00062	0.017
GO:0009651	P	response to salt stress	37	574	0.00062	0.017
GO:0009116	P	nucleoside metabolic process	21	258	0.00066	0.018
GO:0046364	P	monosaccharide biosynthetic process	6	26	0.00066	0.018
GO:0006163	P	purine nucleotide metabolic process	18	204	0.00066	0.018
GO:0009639	P	response to red or far red light	19	222	0.00067	0.018
GO:0042742	P	defense response to bacterium	26	354	0.00069	0.019
GO:0033993	P	response to lipid	46	772	0.00072	0.019
GO:0006090	P	pyruvate metabolic process	11	91	0.00073	0.02
GO:0005983	P	starch catabolic process	5	17	0.00075	0.02
GO:0046394	P	carboxylic acid biosynthetic process	30	438	0.00081	0.021
GO:0009168	P	purine ribonucleoside monophosphate biosynthetic process	9	64	0.00084	0.022
GO:0009127	P	purine nucleoside monophosphate biosynthetic process	9	64	0.00084	0.022
GO:0006633	P	fatty acid biosynthetic process	15	158	0.00092	0.024
GO:0009058	P	biosynthetic process	264	6255	0.00096	0.025
GO:0009156	P	ribonucleoside monophosphate biosynthetic process	10	80	0.001	0.026
GO:0044724	P	single-organism carbohydrate catabolic process	11	95	0.001	0.026
GO:0043094	P	cellular metabolic compound salvage	11	95	0.001	0.026
GO:1901657	P	glycosyl compound metabolic process	24	326	0.0011	0.027
GO:0000413	P	protein peptidyl-prolyl isomerization	6	29	0.0011	0.027
GO:0016053	P	organic acid biosynthetic process	34	531	0.0011	0.028
GO:0006536	P	glutamate metabolic process	5	19	0.0011	0.028
GO:0042128	P	nitrate assimilation	7	41	0.0012	0.028
GO:0042126	P	nitrate metabolic process	7	41	0.0012	0.028
GO:0044264	P	cellular polysaccharide metabolic process	21	273	0.0013	0.031
GO:0071482	P	cellular response to light stimulus	13	130	0.0013	0.031
GO:0006733	P	oxidoreduction coenzyme metabolic process	13	130	0.0013	0.031
GO:0009611	P	response to wounding	18	217	0.0013	0.031
GO:0046034	P	ATP metabolic process	14	148	0.0014	0.033
GO:0046496	P	nicotinamide nucleotide metabolic process	12	115	0.0014	0.033
GO:0009124	P	nucleoside monophosphate biosynthetic process	10	85	0.0015	0.036
GO:0009259	P	ribonucleotide metabolic process	18	221	0.0015	0.036
GO:0009743	P	response to carbohydrate	13	133	0.0015	0.036
GO:0044712	P	single-organism catabolic process	36	585	0.0016	0.036
GO:0042214	P	terpene metabolic process	5	21	0.0017	0.038

GO term	Ontology	Description	Number in input list	Number in BG/Ref	p-value	FDR
GO:0009251	P	glucan catabolic process	7	44	0.0017	0.038
GO:0006098	P	pentose-phosphate shunt	6	32	0.0017	0.039
GO:0051156	P	glucose 6-phosphate metabolic process	6	32	0.0017	0.039
GO:0071478	P	cellular response to radiation	13	135	0.0017	0.039
GO:0009617	P	response to bacterium	29	441	0.0017	0.039
GO:0071941	P	nitrogen cycle metabolic process	7	46	0.0021	0.046
GO:0009631	P	cold acclimation	7	46	0.0021	0.046
GO:2001057	P	reactive nitrogen species metabolic process	7	46	0.0021	0.046
GO:0009205	P	purine ribonucleoside triphosphate metabolic process	14	155	0.0021	0.046
GO:0046395	P	carboxylic acid catabolic process	11	105	0.0021	0.046
GO:0044249	P	cellular biosynthetic process	246	5870	0.0022	0.048
GO:0044262	P	cellular carbohydrate metabolic process	27	407	0.0022	0.048
GO:0009144	P	purine nucleoside triphosphate metabolic process	14	157	0.0023	0.05
GO:0016491	F	oxidoreductase activity	134	1547	3.1E-20	4E-17
GO:0016168	F	chlorophyll binding	24	40	3.3E-19	2.1E-16
GO:0046906	F	tetrapyrrole binding	54	402	2.2E-15	9.3E-13
GO:0051536	F	iron-sulfur cluster binding	27	131	5.1E-12	1.3E-09
GO:0051540	F	metal cluster binding	27	131	5.1E-12	1.3E-09
GO:0009055	F	electron carrier activity	29	196	9.4E-10	2E-07
GO:0003824	F	catalytic activity	412	9101	2.7E-09	5E-07
GO:0016209	F	antioxidant activity	25	181	4.5E-08	7.3E-06
GO:0031409	F	pigment binding	10	22	6.7E-08	9.7E-06
GO:0008266	F	poly(U) RNA binding	9	19	2.3E-07	0.00003
GO:0043167	F	ion binding	193	3821	3.7E-07	0.000044
GO:0008187	F	poly-pyrimidine tract binding	9	21	4.6E-07	0.000046
GO:0045156	F	electron transporter, transferring electrons within the cyclic electron transport pathway of photosynthesis activity	7	9	4.6E-07	0.000046
GO:0016651	F	oxidoreductase activity, acting on NAD(P)H	19	127	6.2E-07	0.000057
GO:0051537	F	2 iron, 2 sulfur cluster binding	13	58	7.4E-07	0.000064
GO:0043169	F	cation binding	183	3678	0.000002	0.00017
GO:0048038	F	quinone binding	10	35	2.2E-06	0.00017
GO:0048037	F	cofactor binding	38	451	2.6E-06	0.00019
GO:0031072	F	heat shock protein binding	9	28	3.2E-06	0.00022
GO:0046872	F	metal ion binding	181	3673	4.2E-06	0.00027
GO:0016667	F	oxidoreductase activity, acting on a sulfur group of donors	17	124	6.8E-06	0.00042
GO:0016671	F	oxidoreductase activity, acting on a sulfur group of donors, disulfide as acceptor	11	51	7.2E-06	0.00042
GO:0016730	F	oxidoreductase activity, acting on iron-sulfur proteins as donors	6	11	0.000014	0.00079
GO:0003755	F	peptidyl-prolyl cis-trans isomerase activity	11	58	0.000021	0.0011
GO:0016668	F	oxidoreductase activity, acting on a sulfur group of donors, NAD(P) as acceptor	9	38	0.000026	0.0013
GO:0016859	F	cis-trans isomerase activity	11	61	0.000031	0.0016
GO:0004791	F	thioredoxin-disulfide reductase activity	8	31	0.000043	0.0021
GO:0051082	F	unfolded protein binding	12	81	0.000077	0.0036
GO:0020037	F	heme binding	29	360	0.000082	0.0036
GO:0016684	F	oxidoreductase activity, acting on peroxide as acceptor	16	139	0.000086	0.0037
GO:0051539	F	4 iron, 4 sulfur cluster binding	10	58	0.0001	0.0042
GO:0016830	F	carbon-carbon lyase activity	14	115	0.00014	0.0055
GO:0003727	F	single-stranded RNA binding	10	61	0.00014	0.0055
GO:0015036	F	disulfide oxidoreductase activity	12	87	0.00014	0.0055
GO:0047134	F	protein-disulfide reductase activity	7	28	0.00015	0.0057
GO:0004601	F	peroxidase activity	15	133	0.00017	0.0062
GO:0015035	F	protein disulfide oxidoreductase activity	11	79	0.00025	0.0087
GO:0016627	F	oxidoreductase activity, acting on the CH-CH group of donors	11	81	0.0003	0.01
GO:0016628	F	oxidoreductase activity, acting on the CH-CH group of donors, NAD or NADP as acceptor	7	32	0.00031	0.01
GO:0004176	F	ATP-dependent peptidase activity	6	24	0.00046	0.015
GO:0016829	F	lyase activity	29	406	0.00053	0.017
GO:0016853	F	isomerase activity	21	264	0.00086	0.027
GO:0046527	F	glucosyltransferase activity	15	158	0.00092	0.028
GO:0005488	F	binding	476	12081	0.0013	0.038
GO:0044435	C	plastid part	389	1455	6.1E-211	4.6E-208
GO:0044434	C	chloroplast part	385	1419	2.1E-210	7.9E-208
GO:0009507	C	chloroplast	533	4148	8.3E-170	2.1E-167
GO:0009536	C	plastid	536	4213	2.8E-169	5.4E-167
GO:0009579	C	thylakoid	239	584	3.6E-157	5.4E-155
GO:0009534	C	chloroplast thylakoid	217	508	5.6E-145	7.1E-143
GO:0031976	C	plastid thylakoid	217	509	7.8E-145	8.5E-143
GO:0044436	C	thylakoid part	195	463	1.2E-128	1.1E-126
GO:0009535	C	chloroplast thylakoid membrane	185	400	1.6E-127	1.4E-125
GO:0055035	C	plastid thylakoid membrane	185	401	2.3E-127	1.8E-125
GO:0042651	C	thylakoid membrane	187	421	1.9E-126	1.3E-124
GO:0034357	C	photosynthetic membrane	187	422	2.7E-126	1.7E-124
GO:0031984	C	organelle subcompartment	222	817	1.3E-114	7.9E-113
GO:0009532	C	plastid stroma	213	774	1.4E-110	7.7E-109

GO term	Ontology	Description	Number in input list	Number in BG/Ref	p-value	FDR
GO:0009570	C	chloroplast stroma	208	740	2.6E-109	1.3E-107
GO:0009526	C	plastid envelope	184	707	2.7E-91	1.3E-89
GO:0009941	C	chloroplast envelope	179	688	1.1E-88	4.9E-87
GO:0044422	C	organelle part	447	4894	8.1E-86	3.4E-84
GO:0044446	C	intracellular organelle part	446	4882	1.4E-85	5.4E-84
GO:0044444	C	cytoplasmic part	693	10923	1.6E-82	6.3E-81
GO:0031975	C	envelope	191	1170	2.6E-65	9E-64
GO:0031967	C	organelle envelope	191	1170	2.6E-65	9E-64
GO:0005737	C	cytoplasm	734	13406	1.9E-60	6.4E-59
GO:0009521	C	photosystem	58	96	1E-44	3.3E-43
GO:0010287	C	plastoglobule	42	80	8E-31	2.4E-29
GO:0031977	C	thylakoid lumen	43	96	2.5E-29	7.4E-28
GO:0009523	C	photosystem II	39	71	2.8E-29	7.9E-28
GO:0043227	C	membrane-bounded organelle	806	18217	3E-28	8E-27
GO:0043231	C	intracellular membrane-bounded organelle	805	18205	4.9E-28	1.3E-26
GO:0043226	C	organelle	809	18527	4.4E-26	1.1E-24
GO:0043229	C	intracellular organelle	808	18517	7.5E-26	1.8E-24
GO:0031978	C	plastid thylakoid lumen	35	81	1.1E-23	2.4E-22
GO:0009543	C	chloroplast thylakoid lumen	35	81	1.1E-23	2.4E-22
GO:0098796	C	membrane protein complex	80	597	3.4E-22	7.6E-21
GO:0009522	C	photosystem I	26	44	1.4E-20	3.1E-19
GO:0044424	C	intracellular part	853	20693	1.5E-19	3.2E-18
GO:0031969	C	chloroplast membrane	49	259	2.4E-19	5E-18
GO:0005622	C	intracellular	853	20721	2.8E-19	5.6E-18
GO:0042170	C	plastid membrane	49	265	5.6E-19	1.1E-17
GO:0048046	C	apoplast	63	486	4.5E-17	8.5E-16
GO:0010319	C	stromule	19	37	1.5E-14	2.8E-13
GO:0016020	C	membrane	416	8532	3.9E-14	7.1E-13
GO:0044464	C	cell part	891	22662	7.2E-14	1.3E-12
GO:0005623	C	cell	891	22664	7.5E-14	1.3E-12
GO:0098807	C	chloroplast thylakoid membrane protein complex	17	34	5.1E-13	8.6E-12
GO:0030095	C	chloroplast photosystem II	13	19	1.5E-11	2.5E-10
GO:0010598	C	NAD(P)H dehydrogenase complex (plastoquinone)	10	13	1.6E-09	2.6E-08
GO:0009539	C	photosystem II reaction center	9	11	7.3E-09	1.1E-07
GO:0009538	C	photosystem I reaction center	9	11	7.3E-09	1.1E-07
GO:0032991	C	macromolecular complex	152	2658	7.2E-09	1.1E-07
GO:0009654	C	photosystem II oxygen evolving complex	11	24	1.4E-08	2E-07
GO:0043234	C	protein complex	109	1834	2.2E-07	3.2E-06
GO:0009528	C	plastid inner membrane	15	82	0.000001	0.000015
GO:0030076	C	light-harvesting complex	9	25	1.5E-06	0.000021
GO:0009706	C	chloroplast inner membrane	14	79	3.2E-06	0.000045
GO:1990204	C	oxidoreductase complex	18	146	0.000014	0.00019
GO:0005840	C	ribosome	37	499	0.000048	0.00064
GO:0009533	C	chloroplast stromal thylakoid	5	10	0.0001	0.0014
GO:0009295	C	nucleoid	9	48	0.00013	0.0016
GO:0044425	C	membrane part	260	5962	0.00015	0.0019
GO:0009501	C	amyloplast	5	13	0.00027	0.0034
GO:0009547	C	plastid ribosome	8	43	0.00031	0.0038
GO:0009527	C	plastid outer membrane	8	51	0.00085	0.01
GO:0009707	C	chloroplast outer membrane	8	51	0.00085	0.01
GO:0000312	C	plastid small ribosomal subunit	5	20	0.0014	0.016
GO:0042646	C	plastid nucleoid	6	31	0.0015	0.017
GO:0000313	C	organellar ribosome	8	62	0.0026	0.03

APPENDIX 7: LIST OF ENRICHED GENE ONTOLOGY (GO) TERMS IN *A. THALIANA* GENES THAT SHOW SIGNIFICANTLY DIFFERENT TRANSCRIPT ABUNDANCE IN RESPONSE TO NORFLURAZON TREATMENT

GO term	Ontology	Description	Number in input list	Number in BG/Ref	p-value	FDR
GO:0015979	P	photosynthesis	174	253	3.3E-56	3.6E-52
GO:0009628	P	response to abiotic stimulus	516	2022	2.3E-44	1.3E-40
GO:0050896	P	response to stimulus	1163	6250	1.7E-41	6.2E-38
GO:0042221	P	response to chemical	634	2853	1.6E-38	4.3E-35
GO:0044710	P	single-organism metabolic process	797	3970	1.2E-35	2.6E-32
GO:0019684	P	photosynthesis, light reaction	99	134	2.8E-34	5E-31
GO:0044699	P	single-organism process	1560	9448	3.7E-33	5.6E-30
GO:0006091	P	generation of precursor metabolites and energy	152	364	5.5E-30	7.4E-27
GO:0010035	P	response to inorganic substance	262	926	8.6E-28	1E-24
GO:0006950	P	response to stress	673	3506	6.8E-25	7.3E-22
GO:0009266	P	response to temperature stimulus	180	559	2.2E-24	2.2E-21
GO:0055114	P	oxidation-reduction process	361	1566	8.1E-24	7.3E-21
GO:0009987	P	cellular process	1924	12771	7.2E-23	5.9E-20
GO:0010033	P	response to organic substance	430	2023	2.1E-22	1.6E-19
GO:0044281	P	small molecule metabolic process	360	1597	2.5E-22	1.8E-19
GO:0008152	P	metabolic process	1812	12035	1.7E-20	1.1E-17
GO:0009416	P	response to light stimulus	204	744	1.6E-20	1.1E-17
GO:0044763	P	single-organism cellular process	1081	6591	2.1E-19	1.3E-16
GO:0009719	P	response to endogenous stimulus	368	1732	3.7E-19	2.1E-16
GO:0009314	P	response to radiation	204	772	5.3E-19	2.8E-16
GO:1901700	P	response to oxygen-containing compound	338	1557	7.9E-19	4.1E-16
GO:0009725	P	response to hormone	349	1631	1.4E-18	6.6E-16
GO:0044711	P	single-organism biosynthetic process	364	1735	3.3E-18	1.5E-15
GO:0043436	P	oxoacid metabolic process	237	973	3.2E-18	1.5E-15
GO:0019752	P	carboxylic acid metabolic process	214	864	3.2E-17	1.4E-14
GO:0009409	P	response to cold	123	384	6.5E-17	2.7E-14
GO:0006082	P	organic acid metabolic process	253	1099	8.1E-17	3.2E-14
GO:0009767	P	photosynthetic electron transport chain	43	56	3.9E-16	1.5E-13
GO:0009735	P	response to cytokinin	90	244	1.2E-15	4.5E-13
GO:0010038	P	response to metal ion	136	481	6.1E-15	2.2E-12
GO:0001101	P	response to acid chemical	256	1181	2.2E-14	7.5E-12
GO:0046686	P	response to cadmium ion	107	343	2.8E-14	9.5E-12
GO:0009765	P	photosynthesis, light harvesting	35	44	1E-13	3.3E-11
GO:0022900	P	electron transport chain	54	112	1.8E-13	5.5E-11
GO:0044237	P	cellular metabolic process	1470	9879	2.6E-13	8E-11
GO:0032787	P	monocarboxylic acid metabolic process	129	481	7.3E-13	2.2E-10
GO:0009414	P	response to water deprivation	103	347	1.1E-12	3.1E-10
GO:0044283	P	small molecule biosynthetic process	159	653	1.3E-12	3.7E-10
GO:0009415	P	response to water	104	355	1.6E-12	4.4E-10
GO:0046394	P	carboxylic acid biosynthetic process	116	438	2E-11	5.4E-09
GO:0009642	P	response to light intensity	56	144	5.5E-11	1.4E-08
GO:0018298	P	protein-chromophore linkage	31	46	5.3E-11	1.4E-08
GO:0009408	P	response to heat	68	199	5.6E-11	1.4E-08
GO:0006979	P	response to oxidative stress	116	453	1.2E-10	2.8E-08
GO:0009617	P	response to bacterium	113	441	1.9E-10	4.6E-08
GO:0051186	P	cofactor metabolic process	102	389	5.1E-10	1.2E-07
GO:0016053	P	organic acid biosynthetic process	127	531	6.2E-10	1.4E-07
GO:0006970	P	response to osmotic stress	145	636	6.4E-10	1.4E-07
GO:0009768	P	photosynthesis, light harvesting in photosystem I	21	23	1.6E-09	3.5E-07
GO:0006631	P	fatty acid metabolic process	74	252	2.3E-09	5E-07

GO term	Ontology	Description	Number in input list	Number in BG/Ref	p-value	FDR
GO:0006633	P	fatty acid biosynthetic process	54	158	5E-09	1.1E-06
GO:0042742	P	defense response to bacterium	92	354	5.1E-09	1.1E-06
GO:0009651	P	response to salt stress	130	574	6.8E-09	1.4E-06
GO:0072330	P	monocarboxylic acid biosynthetic process	72	251	8.5E-09	1.7E-06
GO:0071704	P	organic substance metabolic process	1521	10672	1.4E-08	2.8E-06
GO:0009644	P	response to high light intensity	34	76	2.1E-08	4.1E-06
GO:0033993	P	response to lipid	159	772	4E-08	7.5E-06
GO:0019685	P	photosynthesis, dark reaction	18	21	4.6E-08	8.5E-06
GO:0006778	P	porphyrin-containing compound metabolic process	38	97	5.4E-08	9.9E-06
GO:0006779	P	porphyrin-containing compound biosynthetic process	31	68	6.3E-08	0.000011
GO:0033013	P	tetrapyrrole metabolic process	38	98	6.7E-08	0.000012
GO:0010114	P	response to red light	30	65	8.3E-08	0.000014
GO:0019253	P	reductive pentose-phosphate cycle	17	20	1.2E-07	0.00002
GO:0070271	P	protein complex biogenesis	76	294	1.3E-07	0.000021
GO:0033014	P	tetrapyrrole biosynthetic process	31	71	1.3E-07	0.000022
GO:0015994	P	chlorophyll metabolic process	34	85	1.8E-07	0.000029
GO:0015977	P	carbon fixation	18	24	1.9E-07	0.00003
GO:0009639	P	response to red or far red light	62	222	2E-07	0.000031
GO:0044712	P	single-organism catabolic process	125	585	2.1E-07	0.000033
GO:0015995	P	chlorophyll biosynthetic process	28	61	2.4E-07	0.000037
GO:0080167	P	response to karrikin	43	128	2.6E-07	0.000039
GO:0005982	P	starch metabolic process	29	66	3E-07	0.000045
GO:0006457	P	protein folding	66	247	3.1E-07	0.000046
GO:1901605	P	alpha-amino acid metabolic process	65	242	3.2E-07	0.000047
GO:0009607	P	response to biotic stimulus	228	1253	3.6E-07	0.000051
GO:0042440	P	pigment metabolic process	53	181	4.4E-07	0.000061
GO:0005975	P	carbohydrate metabolic process	211	1144	4.3E-07	0.000061
GO:0006073	P	cellular glucan metabolic process	60	218	4.6E-07	0.000062
GO:0044042	P	glucan metabolic process	60	218	4.6E-07	0.000062
GO:0016051	P	carbohydrate biosynthetic process	82	342	5.9E-07	0.000079
GO:0051707	P	response to other organism	220	1210	6.1E-07	0.000081
GO:0046148	P	pigment biosynthetic process	46	148	6.3E-07	0.000082
GO:0006520	P	cellular amino acid metabolic process	78	320	6.4E-07	0.000083
GO:0006629	P	lipid metabolic process	187	994	6.5E-07	0.000083
GO:0043207	P	response to external biotic stimulus	220	1212	6.7E-07	0.000085
GO:0009737	P	response to abscisic acid	121	578	8.1E-07	0.0001
GO:0044262	P	cellular carbohydrate metabolic process	92	407	1.1E-06	0.00014
GO:0097305	P	response to alcohol	121	583	1.2E-06	0.00014
GO:0009605	P	response to external stimulus	273	1582	1.2E-06	0.00015
GO:0006461	P	protein complex assembly	71	287	1.3E-06	0.00015
GO:0009123	P	nucleoside monophosphate metabolic process	54	196	1.6E-06	0.00019
GO:0009161	P	ribonucleoside monophosphate metabolic process	53	191	1.7E-06	0.0002
GO:0044264	P	cellular polysaccharide metabolic process	67	273	3.1E-06	0.00036
GO:0009637	P	response to blue light	30	81	3.3E-06	0.00038
GO:0051188	P	cofactor biosynthetic process	59	229	3.4E-06	0.00038
GO:0009056	P	catabolic process	247	1430	3.8E-06	0.00042
GO:0006753	P	nucleoside phosphate metabolic process	72	305	4.4E-06	0.00049
GO:0070887	P	cellular response to chemical stimulus	209	1176	4.4E-06	0.00049
GO:0071840	P	cellular component organization or biogenesis	443	2812	4.7E-06	0.00051
GO:0009773	P	photosynthetic electron transport in photosystem I	13	16	5.2E-06	0.00056
GO:0005976	P	polysaccharide metabolic process	93	432	5.6E-06	0.00059
GO:0009117	P	nucleotide metabolic process	71	302	5.8E-06	0.00061
GO:0006810	P	transport	383	2391	0.000006	0.00063
GO:0071822	P	protein complex subunit organization	77	338	6.4E-06	0.00066
GO:0006952	P	defense response	265	1566	6.5E-06	0.00066
GO:0044765	P	single-organism transport	182	1004	6.6E-06	0.00067
GO:0051179	P	localization	400	2517	7.1E-06	0.00071
GO:0018198	P	peptidyl-cysteine modification	14	20	7.6E-06	0.00076
GO:0051234	P	establishment of localization	384	2406	7.9E-06	0.00078
GO:0044723	P	single-organism carbohydrate metabolic process	121	613	9.3E-06	0.00091
GO:0019693	P	ribose phosphate metabolic process	60	245	0.00001	0.001
GO:0010109	P	regulation of photosynthesis	19	39	0.00001	0.001
GO:1902578	P	single-organism localization	186	1040	0.000011	0.001
GO:0045454	P	cell redox homeostasis	42	147	0.000011	0.001
GO:0044085	P	cellular component biogenesis	194	1095	0.000011	0.0011
GO:0009167	P	purine ribonucleoside monophosphate metabolic process	47	174	0.000011	0.0011
GO:0009126	P	purine nucleoside monophosphate metabolic process	47	174	0.000011	0.0011
GO:0043623	P	cellular protein complex assembly	51	197	0.000013	0.0012
GO:0019637	P	organophosphate metabolic process	106	524	0.000013	0.0012
GO:0006006	P	glucose metabolic process	19	40	0.000014	0.0012
GO:0044238	P	primary metabolic process	1377	9890	0.000016	0.0014

GO term	Ontology	Description	Number in input list	Number in BG/Ref	p-value	FDR
GO:1901657	P	glycosyl compound metabolic process	73	326	0.000018	0.0016
GO:0046395	P	carboxylic acid catabolic process	33	105	0.000019	0.0016
GO:0009769	P	photosynthesis, light harvesting in photosystem II	10	10	0.00002	0.0017
GO:0016043	P	cellular component organization	387	2456	0.00002	0.0017
GO:0034622	P	cellular macromolecular complex assembly	79	365	0.000023	0.002
GO:0071214	P	cellular response to abiotic stimulus	51	202	0.000024	0.002
GO:0044255	P	cellular lipid metabolic process	124	649	0.000027	0.0022
GO:0051704	P	multi-organism process	269	1631	0.000028	0.0023
GO:0032502	P	developmental process	492	3233	0.000033	0.0027
GO:0009269	P	response to desiccation	14	24	0.000034	0.0028
GO:0055086	P	nucleobase-containing small molecule metabolic process	79	370	0.000035	0.0028
GO:1901575	P	organic substance catabolic process	213	1249	0.000035	0.0028
GO:0005996	P	monosaccharide metabolic process	31	99	0.000035	0.0028
GO:0044282	P	small molecule catabolic process	43	162	0.000037	0.003
GO:0009156	P	ribonucleoside monophosphate biosynthetic process	27	80	0.000038	0.003
GO:0009124	P	nucleoside monophosphate biosynthetic process	28	85	0.00004	0.0031
GO:0044767	P	single-organism developmental process	477	3131	0.00004	0.0031
GO:0009657	P	plastid organization	51	207	0.000041	0.0031
GO:0043094	P	cellular metabolic compound salvage	30	95	0.000041	0.0032
GO:0009150	P	purine ribonucleotide metabolic process	50	202	0.000043	0.0033
GO:0016054	P	organic acid catabolic process	35	121	0.000045	0.0034
GO:0065003	P	macromolecular complex assembly	87	424	0.00005	0.0037
GO:0009259	P	ribonucleotide metabolic process	53	221	0.000053	0.004
GO:0006163	P	purine nucleotide metabolic process	50	204	0.000054	0.004
GO:0010207	P	photosystem II assembly	13	22	0.000059	0.0043
GO:0016052	P	carbohydrate catabolic process	59	257	0.00006	0.0044
GO:1901293	P	nucleoside phosphate biosynthetic process	38	139	0.000061	0.0044
GO:0008610	P	lipid biosynthetic process	102	522	0.000063	0.0046
GO:0042548	P	regulation of photosynthesis, light reaction	14	26	0.000066	0.0047
GO:0017014	P	protein nitrosylation	11	16	0.000081	0.0057
GO:0018119	P	peptidyl-cysteine S-nitrosylation	11	16	0.000081	0.0057
GO:0009853	P	photorespiration	21	56	0.000081	0.0057
GO:0098542	P	defense response to other organism	164	936	0.000086	0.006
GO:0044724	P	single-organism carbohydrate catabolic process	29	95	0.00009	0.0062
GO:0009772	P	photosynthetic electron transport in photosystem II	9	10	0.000089	0.0062
GO:0009165	P	nucleotide biosynthetic process	37	137	0.000093	0.0063
GO:0019252	P	starch biosynthetic process	16	35	0.000092	0.0063
GO:0042335	P	cuticle development	15	31	0.000093	0.0063
GO:0009141	P	nucleoside triphosphate metabolic process	43	170	0.000095	0.0064
GO:0072521	P	purine-containing compound metabolic process	54	233	0.000098	0.0065
GO:1901606	P	alpha-amino acid catabolic process	17	40	0.00011	0.0076
GO:0042278	P	purine nucleoside metabolic process	48	200	0.00011	0.0076
GO:0019318	P	hexose metabolic process	24	72	0.00012	0.0077
GO:0043467	P	regulation of generation of precursor metabolites and energy	14	28	0.00012	0.0079
GO:0002376	P	immune system process	76	369	0.00013	0.0084
GO:0006955	P	immune response	70	333	0.00014	0.009
GO:0006869	P	lipid transport	43	174	0.00015	0.0094
GO:0046128	P	purine ribonucleoside metabolic process	47	197	0.00015	0.0096
GO:0010118	P	stomatal movement	31	109	0.00015	0.0097
GO:0009168	P	purine ribonucleoside monophosphate biosynthetic process	22	64	0.00016	0.0097
GO:0009127	P	purine nucleoside monophosphate biosynthetic process	22	64	0.00016	0.0097
GO:0044248	P	cellular catabolic process	175	1024	0.00016	0.0097
GO:0045087	P	innate immune response	68	323	0.00017	0.01
GO:0006112	P	energy reserve metabolic process	11	18	0.00017	0.011
GO:0005977	P	glycogen metabolic process	11	18	0.00017	0.011
GO:0014070	P	response to organic cyclic compound	81	405	0.00018	0.011
GO:0010876	P	lipid localization	47	199	0.00018	0.011
GO:0006081	P	cellular aldehyde metabolic process	22	65	0.00019	0.011
GO:0009116	P	nucleoside metabolic process	57	258	0.00019	0.011
GO:0006732	P	coenzyme metabolic process	58	264	0.00019	0.011
GO:1901607	P	alpha-amino acid biosynthetic process	39	154	0.00019	0.011
GO:0006164	P	purine nucleotide biosynthetic process	26	85	0.0002	0.012
GO:0071310	P	cellular response to organic substance	174	1023	0.0002	0.012
GO:0000302	P	response to reactive oxygen species	41	166	0.00021	0.012
GO:0009743	P	response to carbohydrate	35	133	0.00022	0.012
GO:0048856	P	anatomical structure development	470	3146	0.00022	0.013
GO:0009064	P	glutamine family amino acid metabolic process	23	71	0.00023	0.013
GO:0043648	P	dicarboxylic acid metabolic process	23	71	0.00023	0.013
GO:0009144	P	purine nucleoside triphosphate metabolic process	39	157	0.00027	0.015
GO:0022607	P	cellular component assembly	108	586	0.00028	0.015
GO:0009063	P	cellular amino acid catabolic process	19	53	0.00028	0.015

GO term	Ontology	Description	Number in input list	Number in BG/Ref	p-value	FDR
GO:0000038	P	very long-chain fatty acid metabolic process	14	31	0.00028	0.015
GO:0007275	P	multicellular organism development	423	2811	0.00028	0.016
GO:0019682	P	glyceraldehyde-3-phosphate metabolic process	17	44	0.00028	0.016
GO:0009152	P	purine ribonucleotide biosynthetic process	25	83	0.00031	0.017
GO:0009814	P	defense response, incompatible interaction	41	170	0.00032	0.017
GO:0009611	P	response to wounding	49	217	0.00033	0.018
GO:0009199	P	ribonucleoside triphosphate metabolic process	40	165	0.00034	0.018
GO:0006094	P	gluconeogenesis	11	20	0.00035	0.019
GO:0061077	P	chaperone-mediated protein folding	13	28	0.00037	0.019
GO:0019319	P	hexose biosynthetic process	12	24	0.00037	0.019
GO:0008652	P	cellular amino acid biosynthetic process	34	132	0.00037	0.019
GO:1901135	P	carbohydrate derivative metabolic process	116	645	0.00038	0.02
GO:0009205	P	purine ribonucleoside triphosphate metabolic process	38	155	0.00039	0.021
GO:0010218	P	response to far red light	20	60	0.00042	0.022
GO:0009066	P	aspartate family amino acid metabolic process	20	60	0.00042	0.022
GO:0009260	P	ribonucleotide biosynthetic process	28	101	0.00044	0.023
GO:0046390	P	ribose phosphate biosynthetic process	28	101	0.00044	0.023
GO:0090407	P	organophosphate biosynthetic process	61	293	0.00044	0.023
GO:0032870	P	cellular response to hormone stimulus	150	879	0.00047	0.024
GO:0009119	P	ribonucleoside metabolic process	50	227	0.00048	0.024
GO:0006090	P	pyruvate metabolic process	26	91	0.00048	0.024
GO:0051716	P	cellular response to stimulus	426	2856	0.00049	0.025
GO:0009132	P	nucleoside diphosphate metabolic process	24	81	0.0005	0.025
GO:0071478	P	cellular response to radiation	34	135	0.00052	0.026
GO:0044707	P	single-multicellular organism process	428	2875	0.00055	0.027
GO:0009755	P	hormone-mediated signaling pathway	141	821	0.00054	0.027
GO:0071482	P	cellular response to light stimulus	33	130	0.00055	0.027
GO:0055085	P	transmembrane transport	118	667	0.00059	0.029
GO:0044743	P	intracellular protein transmembrane import	21	67	0.0006	0.029
GO:0046034	P	ATP metabolic process	36	148	0.00062	0.03
GO:0072522	P	purine-containing compound biosynthetic process	26	93	0.00062	0.03
GO:0045962	P	positive regulation of development, heterochronic	7	8	0.00063	0.03
GO:0046364	P	monosaccharide biosynthetic process	12	26	0.00064	0.031
GO:0009411	P	response to UV	30	115	0.00066	0.031
GO:0048366	P	leaf development	70	356	0.00071	0.034
GO:0006811	P	ion transport	114	645	0.00074	0.035
GO:0071495	P	cellular response to endogenous stimulus	150	890	0.00074	0.035
GO:0009645	P	response to low light intensity stimulus	10	19	0.00083	0.039
GO:0007568	P	aging	40	174	0.00083	0.039
GO:0019362	P	pyridine nucleotide metabolic process	30	117	0.00084	0.039
GO:0005978	P	glycogen biosynthetic process	8	12	0.0009	0.042
GO:0034285	P	response to disaccharide	20	65	0.00097	0.044
GO:0071555	P	cell wall organization	94	518	0.001	0.046
GO:0006733	P	oxidoreduction coenzyme metabolic process	32	130	0.001	0.047
GO:0006796	P	phosphate-containing compound metabolic process	268	1735	0.0011	0.049
GO:0072524	P	pyridine-containing compound metabolic process	31	125	0.0011	0.05
GO:0003824	F	catalytic activity	1432	9101	6.6E-21	2E-17
GO:0005488	F	binding	1802	12081	1.6E-18	2.4E-15
GO:0016491	F	oxidoreductase activity	323	1547	5.5E-16	5.6E-13
GO:0005515	F	protein binding	638	3669	4.8E-15	3.6E-12
GO:0016168	F	chlorophyll binding	34	40	6.2E-14	3.8E-11
GO:0043167	F	ion binding	632	3821	2.5E-11	1.3E-08
GO:0031409	F	pigment binding	20	22	4.1E-09	1.8E-06
GO:0043169	F	cation binding	591	3678	7E-09	2.7E-06
GO:0048037	F	cofactor binding	108	451	1.1E-08	3.8E-06
GO:0046872	F	metal ion binding	587	3673	1.6E-08	4.9E-06
GO:0005507	F	copper ion binding	68	238	2.5E-08	6.9E-06
GO:0009055	F	electron carrier activity	59	196	4.5E-08	0.000012
GO:0043168	F	anion binding	55	195	7.2E-07	0.00016
GO:0046906	F	tetrapyrrole binding	92	402	7.2E-07	0.00016
GO:0016651	F	oxidoreductase activity, acting on NAD(P)H	40	127	2.5E-06	0.00051
GO:0048038	F	quinone binding	19	35	3.1E-06	0.0006
GO:0051540	F	metal cluster binding	40	131	4.7E-06	0.00079
GO:0051536	F	iron-sulfur cluster binding	40	131	4.7E-06	0.00079
GO:0016853	F	isomerase activity	64	264	7.2E-06	0.0011
GO:0016209	F	antioxidant activity	49	181	7.2E-06	0.0011
GO:0019843	F	rRNA binding	32	97	0.000011	0.0016
GO:0008289	F	lipid binding	72	317	0.000014	0.0019
GO:0036094	F	small molecule binding	522	3434	0.00002	0.0027
GO:0016774	F	phosphotransferase activity, carboxyl group as acceptor	10	11	0.000034	0.0043
GO:0016859	F	cis-trans isomerase activity	23	61	0.000035	0.0043

GO term	Ontology	Description	Number in input list	Number in BG/Ref	p-value	FDR
GO:0005524	F	ATP binding	354	2248	0.000048	0.0053
GO:0051082	F	unfolded protein binding	27	81	0.000046	0.0053
GO:0003755	F	peptidyl-prolyl cis-trans isomerase activity	22	58	0.000048	0.0053
GO:0045156	F	electron transporter, transferring electrons within the cyclic electron transport pathway of photosynthesis activity	9	9	0.000052	0.0055
GO:0000166	F	nucleotide binding	506	3370	0.00008	0.0079
GO:1901265	F	nucleoside phosphate binding	506	3370	0.00008	0.0079
GO:0030554	F	adenyl nucleotide binding	367	2366	0.0001	0.0096
GO:0097367	F	carbohydrate derivative binding	412	2696	0.00011	0.01
GO:0032559	F	adenyl ribonucleotide binding	366	2364	0.00012	0.011
GO:0051537	F	2 iron, 2 sulfur cluster binding	21	58	0.00012	0.011
GO:0035639	F	purine ribonucleoside triphosphate binding	386	2515	0.00014	0.011
GO:0102337	F	3-oxo-cerotoyl-CoA synthase activity	12	21	0.00014	0.011
GO:0102336	F	3-oxo-arachidoyl-CoA synthase activity	12	21	0.00014	0.011
GO:0102338	F	3-oxo-lignoceronyl-CoA synthase activity	12	21	0.00014	0.011
GO:0001882	F	nucleoside binding	400	2618	0.00015	0.011
GO:0032549	F	ribonucleoside binding	400	2617	0.00014	0.011
GO:0016667	F	oxidoreductase activity, acting on a sulfur group of donors	34	124	0.00014	0.011
GO:0016830	F	carbon-carbon lyase activity	32	115	0.00017	0.012
GO:0005198	F	structural molecule activity	102	538	0.00017	0.012
GO:0032553	F	ribonucleotide binding	404	2664	0.00023	0.015
GO:0008266	F	poly(U) RNA binding	11	19	0.00025	0.016
GO:0001883	F	purine nucleoside binding	396	2612	0.00027	0.017
GO:0017076	F	purine nucleotide binding	399	2633	0.00026	0.017
GO:0032550	F	purine ribonucleoside binding	396	2612	0.00027	0.017
GO:0050662	F	coenzyme binding	60	281	0.00028	0.017
GO:0016829	F	lyase activity	80	406	0.00029	0.018
GO:0032555	F	purine ribonucleotide binding	398	2631	0.0003	0.018
GO:0003959	F	NADPH dehydrogenase activity	7	7	0.00037	0.021
GO:0008187	F	poly-pyrimidine tract binding	11	21	0.00047	0.026
GO:0004312	F	fatty acid synthase activity	11	21	0.00047	0.026
GO:0016740	F	transferase activity	552	3791	0.00051	0.028
GO:0016783	F	sulfurtransferase activity	9	14	0.00052	0.028
GO:0016668	F	oxidoreductase activity, acting on a sulfur group of donors, NAD(P) as acceptor	15	38	0.00053	0.028
GO:0015036	F	disulfide oxidoreductase activity	25	87	0.00056	0.029
GO:0005528	F	FK506 binding	11	22	0.00064	0.032
GO:0005527	F	macrolide binding	11	22	0.00064	0.032
GO:0015035	F	protein disulfide oxidoreductase activity	23	79	0.00079	0.038
GO:0004791	F	thioredoxin-disulfide reductase activity	13	31	0.00078	0.038
GO:0031406	F	carboxylic acid binding	25	91	0.00097	0.046
GO:0042802	F	identical protein binding	50	236	0.001	0.047
GO:0044434	C	chloroplast part	627	1419	9.9E-136	1.6E-132
GO:0044435	C	plastid part	631	1455	1.2E-133	1E-130
GO:0009507	C	chloroplast	1055	4148	1.1E-98	5.9E-96
GO:0009579	C	thylakoid	345	584	3.7E-98	1.5E-95
GO:0009536	C	plastid	1059	4213	1.9E-96	6.3E-94
GO:0009534	C	chloroplast thylakoid	305	508	7.6E-88	2.1E-85
GO:0031976	C	plastid thylakoid	305	509	1.1E-87	2.6E-85
GO:0009570	C	chloroplast stroma	361	740	5.7E-85	1.2E-82
GO:0009532	C	plastid stroma	368	774	3E-84	5.5E-82
GO:0044436	C	thylakoid part	277	463	2E-79	3.4E-77
GO:0034357	C	photosynthetic membrane	263	422	3.7E-78	5.6E-76
GO:0042651	C	thylakoid membrane	262	421	9.3E-78	1.3E-75
GO:0009535	C	chloroplast thylakoid membrane	254	400	1.1E-76	1.4E-74
GO:0055035	C	plastid thylakoid membrane	254	401	1.6E-76	1.9E-74
GO:0031984	C	organelle subcompartment	337	817	5.6E-65	6.2E-63
GO:0009526	C	plastid envelope	308	707	1.8E-63	1.9E-61
GO:0044422	C	organelle part	1053	4894	1.1E-62	1.1E-60
GO:0044446	C	intracellular organelle part	1050	4882	2.2E-62	2E-60
GO:0009941	C	chloroplast envelope	301	688	2.5E-62	2.2E-60
GO:0044444	C	cytoplasmic part	1858	10923	1.9E-52	1.6E-50
GO:0005737	C	cytoplasm	2162	13406	1.2E-49	9.4E-48
GO:0031975	C	envelope	363	1170	1E-45	7.3E-44
GO:0031967	C	organelle envelope	363	1170	1E-45	7.3E-44
GO:0005623	C	cell	3151	22664	8.4E-34	5.8E-32
GO:0044464	C	cell part	3150	22662	1.3E-33	8.7E-32
GO:0009521	C	photosystem	84	96	6.3E-33	4E-31
GO:0016020	C	membrane	1423	8532	6.7E-31	4.1E-29
GO:0009523	C	photosystem II	60	71	2.1E-23	1.3E-21
GO:0048046	C	apoplast	156	486	3.6E-21	2.1E-19
GO:0071944	C	cell periphery	764	4309	4.1E-20	2.3E-18
GO:0005829	C	cytosol	404	1993	5.3E-18	2.8E-16

GO term	Ontology	Description	Number in input list	Number in BG/Ref	p-value	FDR
GO:0031977	C	thylakoid lumen	58	96	1.1E-17	5.7E-16
GO:0010287	C	plastoglobule	53	80	1.5E-17	7.7E-16
GO:0030312	C	external encapsulating structure	186	706	2.6E-17	1.2E-15
GO:0005618	C	cell wall	186	706	2.6E-17	1.2E-15
GO:0005622	C	intracellular	2845	20721	5.4E-17	2.5E-15
GO:0044424	C	intracellular part	2841	20693	6.8E-17	3.1E-15
GO:0009522	C	photosystem I	39	44	3.5E-16	1.5E-14
GO:0043231	C	intracellular membrane-bounded organelle	2536	18205	1.8E-15	7.6E-14
GO:0043227	C	membrane-bounded organelle	2537	18217	2E-15	8.4E-14
GO:0031978	C	plastid thylakoid lumen	49	81	3.5E-15	1.3E-13
GO:0009543	C	chloroplast thylakoid lumen	49	81	3.5E-15	1.3E-13
GO:0043229	C	intracellular organelle	2571	18517	3.5E-15	1.3E-13
GO:0043226	C	organelle	2572	18527	3.7E-15	1.4E-13
GO:0098796	C	membrane protein complex	158	597	4.9E-15	1.8E-13
GO:0098807	C	chloroplast thylakoid membrane protein complex	29	34	4E-12	1.5E-10
GO:0005886	C	plasma membrane	625	3735	5.6E-12	2E-10
GO:0009506	C	plasmodesma	197	899	1.1E-11	3.7E-10
GO:0055044	C	symplast	197	899	1.1E-11	3.7E-10
GO:0005911	C	cell-cell junction	197	901	1.3E-11	4.2E-10
GO:0030054	C	cell junction	197	901	1.3E-11	4.2E-10
GO:0031969	C	chloroplast membrane	82	259	1.5E-11	4.9E-10
GO:0042170	C	plastid membrane	82	265	3.9E-11	1.2E-09
GO:0032991	C	macromolecular complex	456	2658	4.3E-10	1.3E-08
GO:0009505	C	plant-type cell wall	87	323	3.2E-09	9.8E-08
GO:0030076	C	light-harvesting complex	21	25	4.4E-09	1.3E-07
GO:0009654	C	photosystem II oxygen evolving complex	20	24	1.1E-08	3.3E-07
GO:0010319	C	stromule	24	37	1.3E-08	3.8E-07
GO:0043234	C	protein complex	323	1834	2.2E-08	6.1E-07
GO:0043232	C	intracellular non-membrane-bounded organelle	276	1543	7.6E-08	2.1E-06
GO:0043228	C	non-membrane-bounded organelle	276	1543	7.6E-08	2.1E-06
GO:0044425	C	membrane part	892	5962	8.9E-08	2.4E-06
GO:0030095	C	chloroplast photosystem II	16	19	3E-07	0.000008
GO:0031225	C	anchored component of membrane	70	280	1.1E-06	0.000028
GO:0009547	C	plastid ribosome	22	43	1.1E-06	0.000029
GO:0046658	C	anchored component of plasma membrane	37	109	1.4E-06	0.000035
GO:0009295	C	nucleoid	23	48	1.6E-06	0.000039
GO:0031224	C	intrinsic component of membrane	818	5540	2.4E-06	0.00006
GO:0005773	C	vacuole	237	1359	3.3E-06	0.00008
GO:0005840	C	ribosome	105	499	3.4E-06	0.000082
GO:0005576	C	extracellular region	463	2954	4.5E-06	0.0001
GO:0010598	C	NAD(P)H dehydrogenase complex (plastoquinone)	12	13	0.000005	0.00011
GO:0009539	C	photosystem II reaction center	11	11	7.5E-06	0.00017
GO:0031226	C	intrinsic component of plasma membrane	66	292	0.000034	0.00076
GO:0009538	C	photosystem I reaction center	10	11	0.000034	0.00076
GO:0000313	C	organellar ribosome	23	62	0.000043	0.00095
GO:0009528	C	plastid inner membrane	27	82	0.000054	0.0012
GO:0009706	C	chloroplast inner membrane	26	79	0.000074	0.0016
GO:1990204	C	oxidoreductase complex	37	146	0.00027	0.0057
GO:0042646	C	plastid nucleoid	14	31	0.00028	0.0057
GO:0030529	C	intracellular ribonucleoprotein complex	147	852	0.00035	0.0071
GO:0000312	C	plastid small ribosomal subunit	11	20	0.00035	0.0071
GO:1990904	C	ribonucleoprotein complex	147	852	0.00035	0.0071
GO:0009533	C	chloroplast stromal thylakoid	8	10	0.00038	0.0076
GO:0031090	C	organelle membrane	276	1760	0.00042	0.0081
GO:0016021	C	integral component of membrane	753	5324	0.00076	0.015
GO:0044459	C	plasma membrane part	81	428	0.00078	0.015
GO:0000311	C	plastid large ribosomal subunit	11	23	0.00085	0.016
GO:0005730	C	nucleolus	74	390	0.0012	0.022
GO:0045298	C	tubulin complex	8	13	0.0013	0.024
GO:0005759	C	mitochondrial matrix	29	116	0.0014	0.026
GO:0009986	C	cell surface	8	14	0.0019	0.034
GO:0000314	C	organellar small ribosomal subunit	12	31	0.0021	0.038

APPENDIX 8: GENE IDS FOR C_4 ORTHOLOGUES IN *A. THALIANA*

C_4 Orthologue	TAIR Gene ID
ASP1	AT2G30970
CA1	AT3G01500
CA2	AT5G14740
DIC1	AT2G22500
MDH1	AT1G53240
NADP-ME1	AT2G13560
NADP-ME2	AT4G00570
NHD1	AT3G19490
PCK1	AT4G37870
PPA6	AT5G09650
PPC2	AT2G42600
PPDK	AT4G15530
PPT1	AT5G33320
RP1	AT4G21210
TPT	AT5G46110

APPENDIX 9: LIST OF MOTIFS ENRICHED >2.5 ABOVE BACKGROUND IN THE FIMO ANALYSIS OF 1000BP UPSTREAM + 5' UTR OF *A. THALIANA* GENES THAT SHOW SIGNIFICANTLY DIFFERENT TRANSCRIPT ABUNDANCE IN RESPONSE TO CHLOROPLAST INHIBITION

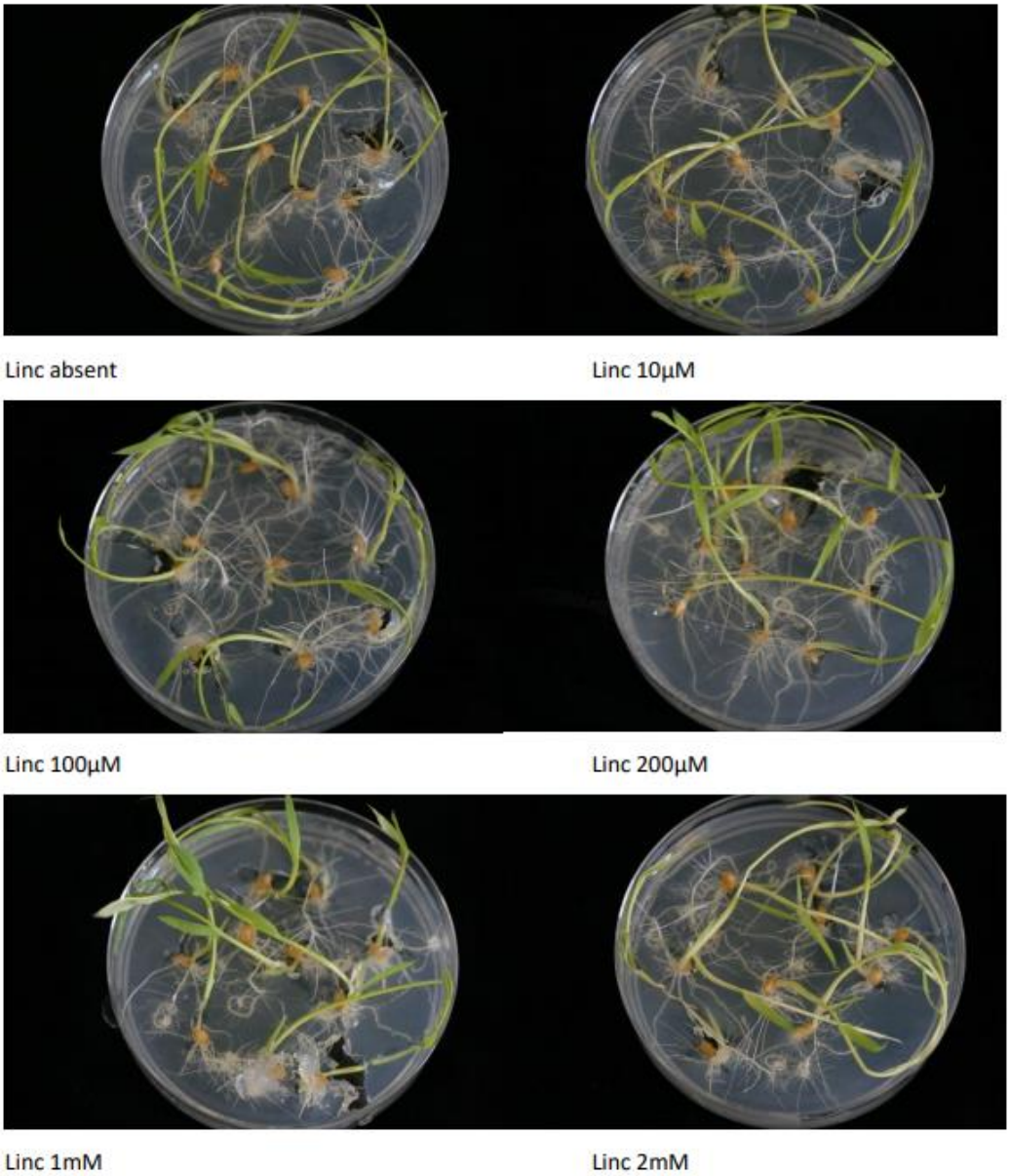
Motif ID	Total Motif Frequency	Ratio compared to Average Background Frequency
MYBrelated_tnt.AT3G10113_col_a_m1	26	4.64
MYBrelated_tnt.At5g52660_colamp_a_m1	27	4.50
bZIP_tnt.GBF3_colamp_m1	184	4.44
bZIP_tnt.GBF3_col_m1	128	4.41
bZIP_tnt.bZIP28_col_a_m1	130	4.36
bZIP_tnt.bZIP68_col_a_m1	151	4.34
bZIP_tnt.ABI5_col_v3h_m1	172	4.28
bZIP_tnt.ABF2_col_v3a_m1	179	4.26
bZIP_tnt.ABI5_colamp_v3b_m1	173	4.24
MYBrelated_tnt.At3g09600_col_a_m1	22	4.23
bZIP_tnt.AREB3_colamp_a_m1	176	4.21
bZIP_tnt.bZIP16_colamp_a_m1	140	4.09
bZIP_tnt.bZIP16_col_v3a_m1	171	4.07
bZIP_tnt.AREB3_col_v31_m1	185	3.99
MYBrelated_tnt.RVE1_col_a_m1	21	3.62
bZIP_tnt.bZIP43_col_a_m1	44	3.28
bHLH_tnt.PIF7_col_a_m1	96	3.22
bZIP_tnt.GBF5_colamp_a_m1	70	3.13
MYB_tnt.MYB67_col_a_m1	20	3.13
bZIP_tnt.GBF6_colamp_a_m1	69	3.05
bZIP_tnt.GBF5_col_v3a_m1	65	3.04
bZIP_tnt.GBF6_col_m1	54	3.03
TCP_tnt.TCP20_colamp_a_m1	25	2.98
bZIP_tnt.bZIP48_colamp_a_m1	54	2.97
bHLH_tnt.bHLH104_col_b_m1	105	2.92
bZIP_tnt.bZIP53_col_m1	53	2.91
MYB_tnt.MYB13_col_a_m1	30	2.88
TCP_tnt.At5g08330_colamp_a_m1	24	2.86
bZIP_tnt.bZIP48_col_a_m1	52	2.83
bZIP_tnt.bZIP3_col_a_m1	63	2.79
TCP_tnt.TCP20_col_a_m1	30	2.78
bHLH_tnt.bHLH31_col_m1	82	2.70
MYB_tnt.MYB58_colamp_a_m1	35	2.69
bHLH_tnt.bHLH34_col_m1	117	2.67
MYB_tnt.MYB13_colamp_a_m1	32	2.67
TCP_tnt.At2g45680_colamp_a_m1	24	2.67
bHLH_tnt.bHLH34_colamp_a_m1	124	2.65
MYB_tnt.MYB58_col_a_m1	30	2.63
bZIP_tnt.bZIP53_colamp_a_m1	47	2.61
MYB_tnt.MYB10_colamp_a_m1	24	2.61
TCP_tnt.At2g45680_col_b_m1	25	2.60
bZIP_tnt.bZIP44_colamp_a_m1	50	2.58
bZIP_tnt.bZIP44_col_m1	52	2.57
TCP_tnt.At1g72010_col_m1	24	2.55
bZIP_tnt.bZIP42_colamp_a_m1	43	2.53
bZIP_tnt.bZIP42_col_a_m1	42	2.50

APPENDIX 10: LIST OF MOTIFS ENRICHED WITH A P-VALUE >0.01 IN THE AME ANALYSIS OF 1000BP UPSTREAM + 5' UTR OF *A. THALIANA* GENES THAT SHOW SIGNIFICANTLY DIFFERENT TRANSCRIPT ABUNDANCE IN RESPONSE TO CHLOROPLAST INHIBITION

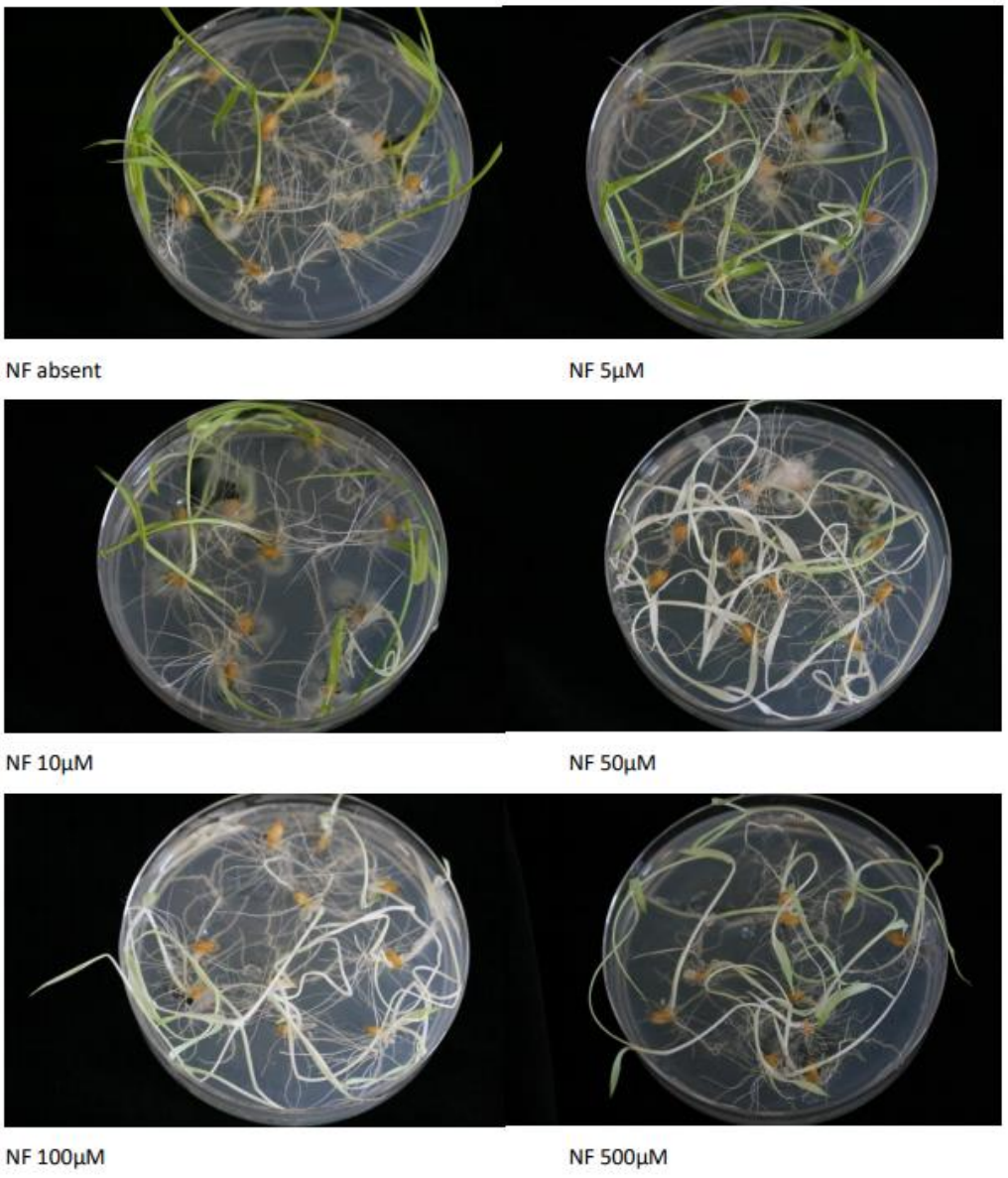
Rank	Motif ID	Adjusted p-value
1	bZIP_tnt.ABF2_col_v3a_m1	3.28E-13
2	bZIP_tnt.AREB3_col_v31_m1	1.30E-12
3	bZIP_tnt.ABI5_col_v3h_m1	1.68E-12
4	bZIP_tnt.bZIP16_col_v3a_m1	1.79E-12
5	bZIP_tnt.ABI5_colamp_v3b_m1	2.08E-12
6	bZIP_tnt.AREB3_colamp_a_m1	2.17E-12
7	bZIP_tnt.bZIP16_colamp_a_m1	1.18E-11
8	bZIP_tnt.GBF3_col_m1	3.45E-10
9	bZIP_tnt.GBF5_colamp_a_m1	9.55E-10
10	AP2EREBP_tnt.AT1G71450_col_a_m1	2.84E-09
11	bZIP_tnt.bZIP28_col_a_m1	5.86E-08
12	bZIP_tnt.bZIP68_col_a_m1	7.68E-08
13	HMG_tnt.3XHMGBOX1_col_a_m1	1.79E-07
14	AP2EREBP_tnt.At5g18450_col_a_m1	1.90E-07
15	bZIP_tnt.GBF5_col_v3a_m1	3.55E-07
16	AP2EREBP_tnt.ERF48_col_a_m1	7.86E-07
17	AP2EREBP_tnt.CEJ1_col_a_m1	8.14E-07
18	bZIP_tnt.bZIP3_col_a_m1	9.92E-07
19	LOBAS2_tnt.AS2_col_a_m1	1.60E-06
20	AP2EREBP_tnt.At1g75490_col_a_m1	1.74E-06
21	HMG_tnt.3XHMGBOX1_colamp_a_m1	1.90E-06
22	AP2EREBP_tnt.At5g65130_col_a_m1	2.12E-06
23	BZR_tnt.At4g18890_col_a_m1	3.26E-06
24	AP2EREBP_tnt.ERF10_col_a_m1	3.76E-06
25	BZR_tnt.At4g18890_colamp_a_m1	3.78E-06
26	AP2EREBP_tnt.CRF4_col_a_m1	5.20E-06
27	LOBAS2_tnt.LBD13_col_a_m1	5.66E-06
28	bZIP_tnt.bZIP48_col_a_m1	6.44E-06
29	AP2EREBP_tnt.At5g67000_col_a_m1	8.43E-06
30	AP2EREBP_tnt.ABR1_col_a_m1	9.74E-06
31	AP2EREBP_tnt.ERF7_col_a_m1	1.05E-05
32	AP2EREBP_tnt.ERF2_col_a_m1	1.40E-05
33	AP2EREBP_tnt.ERF115_col_a_m1	1.56E-05
34	AP2EREBP_tnt.At1g22810_col_m1	1.85E-05
35	AP2EREBP_tnt.ERF105_col_m1	2.09E-05
36	AP2EREBP_tnt.At4g16750_col_a_m1	2.36E-05
37	bZIP_tnt.GBF6_colamp_a_m1	2.74E-05
38	AP2EREBP_tnt.ESE3_col_a_m1	3.05E-05
39	bZIP_tnt.GBF6_col_m1	3.33E-05
40	LOBAS2_tnt.LOB_col_a_m1	5.17E-05
41	AP2EREBP_tnt.ERF104_col_a_m1	7.18E-05
42	AP2EREBP_tnt.AT1G71450_colamp_a_m1	8.06E-05
43	NAC_tnt.ANAC094_col_a_m1	9.98E-05
44	AP2EREBP_tnt.AT1G44830_col_a_m1	1.05E-04
45	AP2EREBP_tnt.At4g31060_colamp_a_m1	1.09E-04
46	BES1_tnt.BAM8_col_a_m1	1.10E-04
47	AP2EREBP_tnt.CRF10_col100_a_m1	1.42E-04
48	BZR_tnt.At1g78700_colamp_a_m1	1.44E-04
49	AP2EREBP_tnt.At4g28140_colamp_a_m1	1.46E-04
50	AP2EREBP_tnt.AT3G57600_col_a_m1	1.51E-04
51	AP2EREBP_tnt.ERF9_colamp_a_m1	1.62E-04

Rank	Motif ID	Adjusted p-value
52	AP2EREBP_tnt.ERF2_colamp_a_m1	1.78E-04
53	AP2EREBP_tnt.SHN3_col_a_m1	1.84E-04
54	AP2EREBP_tnt.ERF15_colamp_a_m1	1.87E-04
55	AP2EREBP_tnt.CRF4_colamp_a_m1	1.98E-04
56	bHLH_tnt.BIM3_col_a_m1	2.55E-04
57	bHLH_tnt.bHLH34_col_m1	3.04E-04
58	MYB_tnt.MYB30_col_a_m1	3.51E-04
59	AP2EREBP_tnt.ERF4_col_a_m1	3.76E-04
60	bZIP_tnt.GBF3_colamp_m1	3.86E-04
61	AP2EREBP_tnt.RAP26_col_a_m1	4.45E-04
62	AP2EREBP_tnt.DEAR2_col_a_m1	4.48E-04
63	AP2EREBP_tnt.AT1G12630_col_a_m1	4.85E-04
64	AP2EREBP_tnt.ERF13_col_b_m1	4.98E-04
65	bHLH_tnt.BIM1_colamp_a_m1	5.64E-04
66	mTERF_tnt.AT5G23930_col_a_m1	5.79E-04
67	AP2EREBP_tnt.ERF8_col_a_m1	6.50E-04
68	AP2EREBP_tnt.ERF105_colamp_a_m1	6.77E-04
69	AP2EREBP_tnt.ERF15_col_m1	6.94E-04
70	bZIP_tnt.bZIP53_col_m1	7.88E-04
71	AP2EREBP_tnt.ERF13_colamp_a_m1	8.14E-04
72	AP2EREBP_tnt.AT4G18450_col_a_m1	8.33E-04
73	bZIP_tnt.TGA2_colamp_v31_m1	9.01E-04
74	AP2EREBP_tnt.ESE1_col_a_m1	9.22E-04
75	AP2EREBP_tnt.RAP211_col_a_m1	9.45E-04
76	AP2EREBP_tnt.ESE3_colamp_a_m1	1.11E-03
77	AP2EREBP_tnt.LEP_col_a_m1	1.24E-03
78	bZIP_tnt.bZIP44_col_m1	1.40E-03
79	BZR_tnt.BZR1_col_a_m1	1.56E-03
80	AP2EREBP_tnt.ERF5_col_a_m1	1.57E-03
81	LOBAS2_tnt.LBD13_colamp_a_m1	1.73E-03
82	AP2EREBP_tnt.RAP212_col_a_m1	1.75E-03
83	MYBrelated_tnt.At1g74840_colamp_a_m1	1.78E-03
84	bZIP_tnt.bZIP42_col_a_m1	1.79E-03
85	bZIP_tnt.bZIP44_colamp_a_m1	1.86E-03
86	bHLH_tnt.bHLH74_col_a_m1	1.86E-03
87	BZR_tnt.At1g78700_col_a_m1	2.05E-03
88	AP2EREBP_tnt.ERF11_col_b_m1	2.19E-03
89	AP2EREBP_tnt.At2g44940_col_a_m1	2.26E-03
90	AP2EREBP_tnt.TINY_col_a_m1	2.28E-03
91	AP2EREBP_tnt.At1g77640_col_a_m1	2.69E-03
92	AP2EREBP_tnt.At4g28140_col_a_m1	2.82E-03
93	bHLH_tnt.PIF7_col_a_m1	3.09E-03
94	AP2EREBP_tnt.AT1G01250_colamp_a_m1	3.16E-03
95	bZIP_tnt.bZIP48_colamp_a_m1	3.27E-03
96	bHLH_tnt.bHLH10_col_a_m1	3.29E-03
97	AP2EREBP_tnt.At2g33710_colamp_a_m1	3.38E-03
98	bHLH_tnt.bHLH34_colamp_a_m1	3.50E-03
99	AP2EREBP_tnt.ERF10_colamp_a_m1	3.92E-03
100	AP2EREBP_tnt.ERF73_col_a_m1	4.04E-03
101	AP2EREBP_tnt.PUCHI_col_a_m1	4.07E-03
102	AP2EREBP_tnt.AT1G28160_col_a_m1	4.60E-03
103	bHLH_tnt.bHLH104_col_b_m1	4.89E-03
104	AP2EREBP_tnt.ERF9_col_a_m1	5.59E-03
105	BES1_tnt.BAM8_colamp_a_m1	5.64E-03
106	C3H_tnt.At5g08750_col_a_m1	5.75E-03
107	bHLH_tnt.BIM2_colamp_m1	6.45E-03
108	bHLH_tnt.bHLH74_colamp_a_m1	6.47E-03
109	AP2EREBP_tnt.AT1G01250_col_a_m1	6.86E-03
110	bZIP_tnt.bZIP42_colamp_a_m1	7.02E-03
111	bZIP_tnt.bZIP53_colamp_a_m1	7.05E-03
112	AP2EREBP_tnt.At1g36060_colamp_a_m1	7.15E-03
113	AP2EREBP_tnt.At2g33710_col_m1	7.17E-03
114	AP2EREBP_tnt.ERF38_col_a_m1	7.37E-03
115	AP2EREBP_tnt.ERF11_colamp_a_m1	7.57E-03
116	MYB_tnt.ATY19_col_a_m1	8.47E-03
117	AP2EREBP_tnt.PUCHI_colamp_a_m1	8.51E-03
118	C2H2_tnt.TF3A_col_a_m1	8.64E-03
119	TCP_tnt.At5g08330_colamp_a_m1	9.27E-03
120	AP2EREBP_tnt.DREB26_col_a_m1	9.39E-03
121	C2C2COlike_tnt.AT5G59990_colamp_a_m1	9.40E-03
122	AP2EREBP_tnt.ERF104_colamp_a_m1	9.53E-03
123	C2H2_tnt.At5g04390_col200_a_m1	1.04E-02
124	C2C2COlike_tnt.AT5G59990_col_a_m1	1.08E-02
125	GeBP_tnt.AT4G00250_col_a_m1	1.10E-02
126	bHLH_tnt.BIM2_col_v3b_m1	1.13E-02

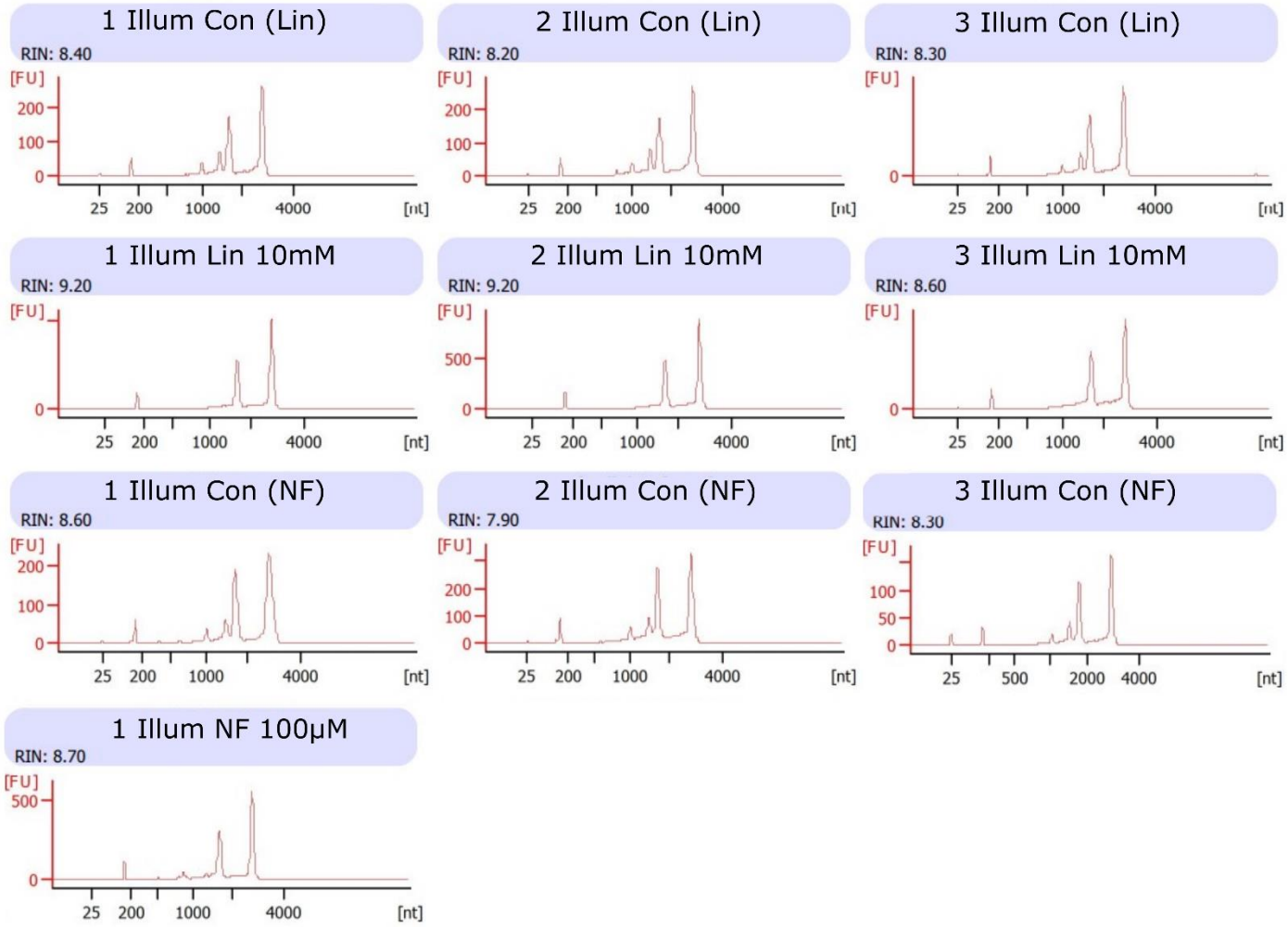
APPENDIX 11: LINCOMYCIN DOSE RESPONSE ASSAY IN *O. SATIVA*

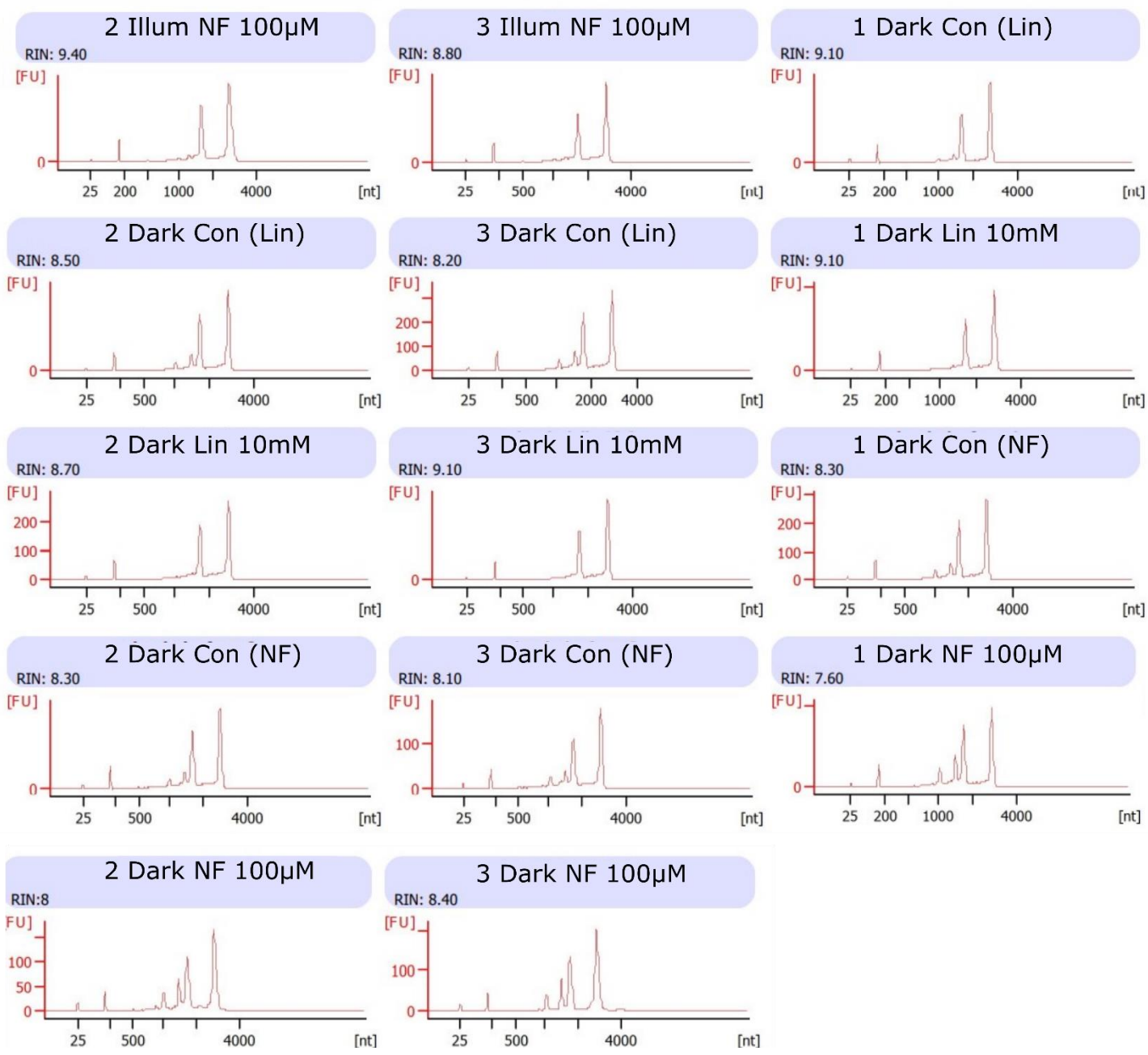


APPENDIX 12: NORFLURAZON DOSE RESPONSE ASSAY IN *O. SATIVA*

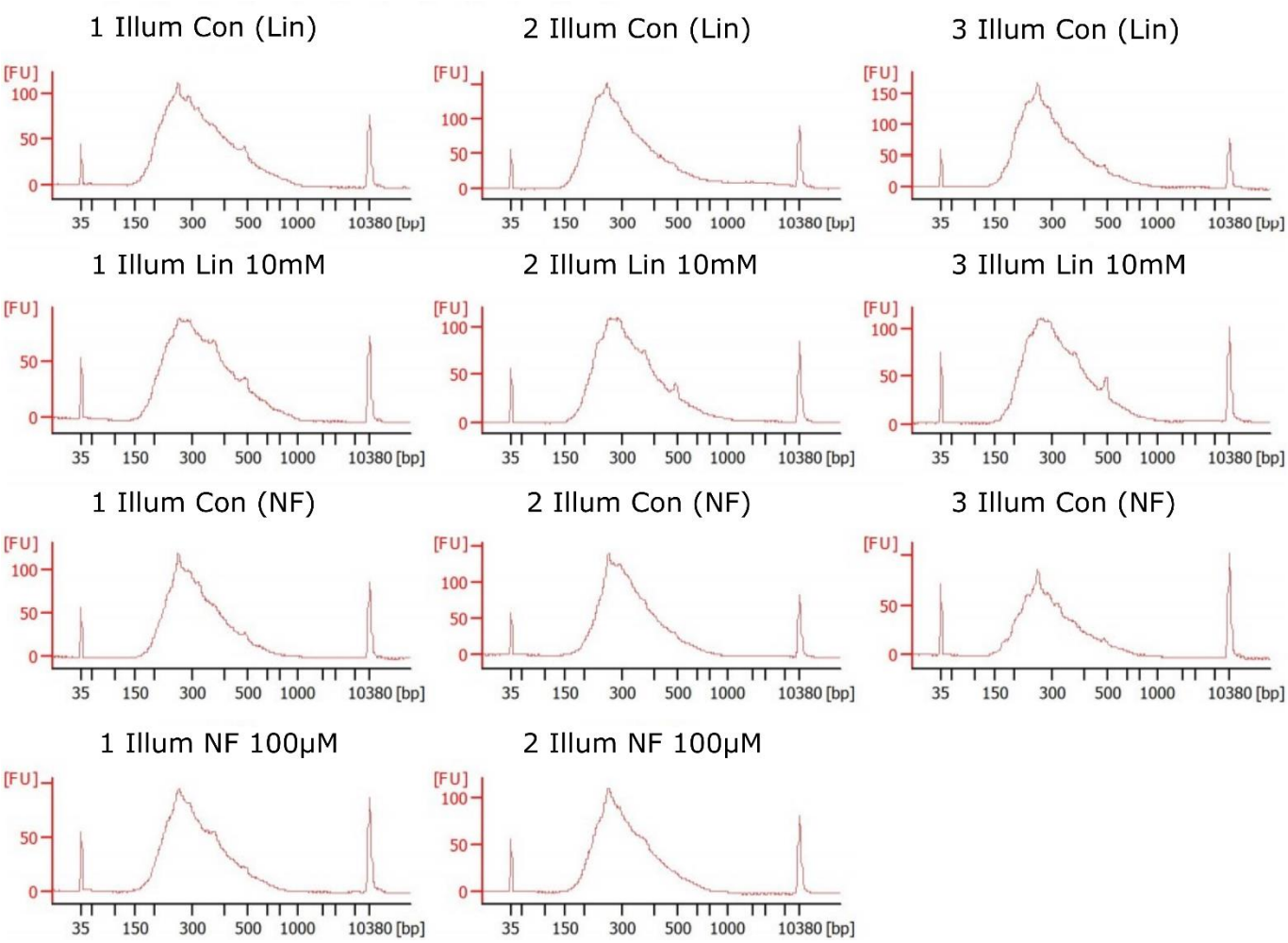


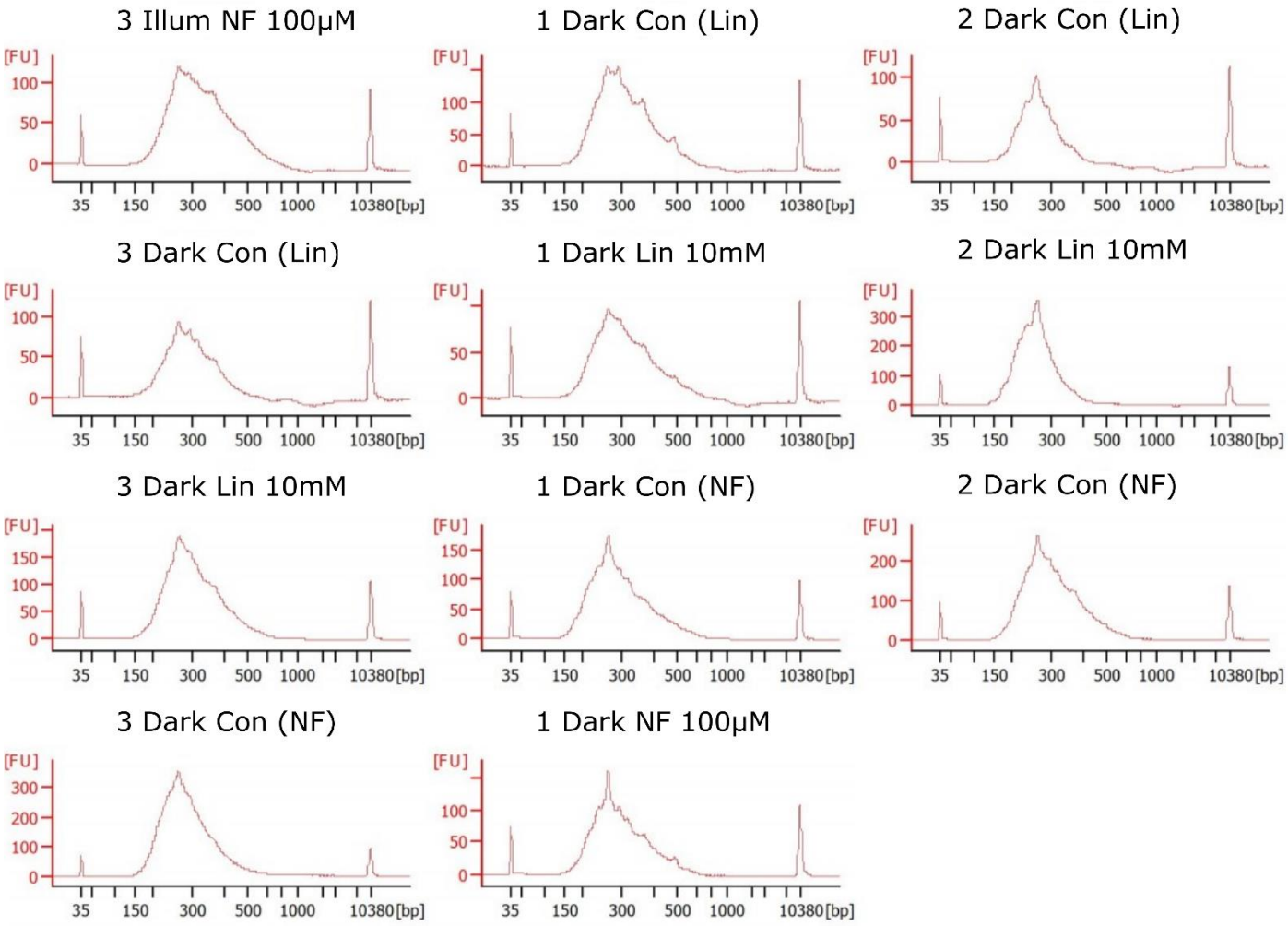
APPENDIX 13: ELECTROPHEROGRAMS OF *O. SATIVA* RNA




2 Illum NF 100µM
RIN: 9.40
3 Illum NF 100µM
RIN: 8.80
1 Dark Con (Lin)
RIN: 9.10
2 Dark Con (Lin)
RIN: 8.50
3 Dark Con (Lin)
RIN: 8.20
1 Dark Lin 10mM
RIN: 9.10
2 Dark Lin 10mM
RIN: 8.70
3 Dark Lin 10mM
RIN: 9.10
1 Dark Con (NF)
RIN: 8.30
2 Dark Con (NF)
RIN: 8.30
3 Dark Con (NF)
RIN: 8.10
1 Dark NF 100µM
RIN: 7.60
2 Dark NF 100µM
RIN:8
3 Dark NF 100µM
RIN: 8.40
[FU]
[nt]

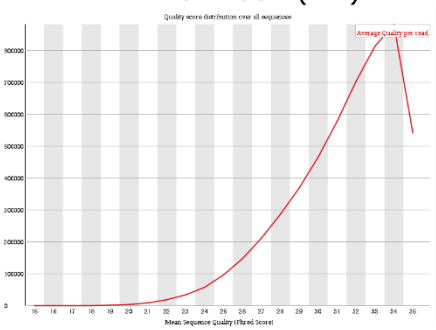
APPENDIX 14: ELECTROPHEROGRAMS OF *O. SATIVA* RNA-SEQ LIBRARIES



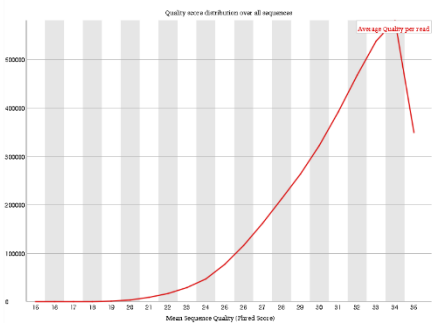

3 Illum NF 100µM
1 Dark Con (Lin)
2 Dark Con (Lin)
3 Dark Con (Lin)
1 Dark Lin 10mM
2 Dark Lin 10mM
3 Dark Lin 10mM
1 Dark Con (NF)
2 Dark Con (NF)
3 Dark Con (NF)
1 Dark NF 100µM
[FU]
35 150 300 500 1000 10380 [bp]

APPENDIX 15: FASTQC REPORTS OF MEAN SEQUENCE QUALITY FOR *O. SATIVA* RNA-SEQ READS

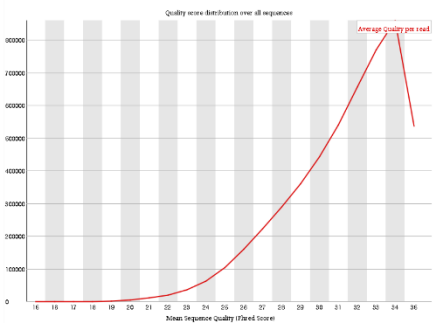
1 Illum Con (Lin)



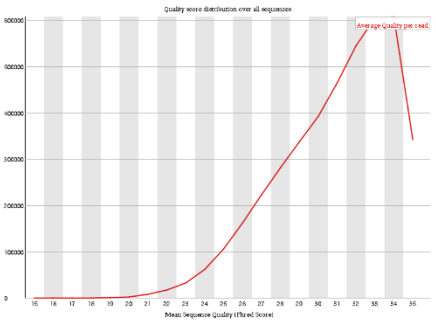
2 Illum Con (Lin)



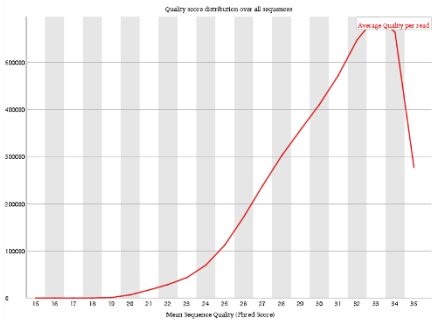
3 Illum Con (Lin)



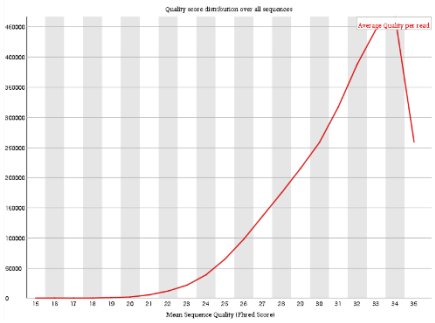
1 Illum Lin 10mM



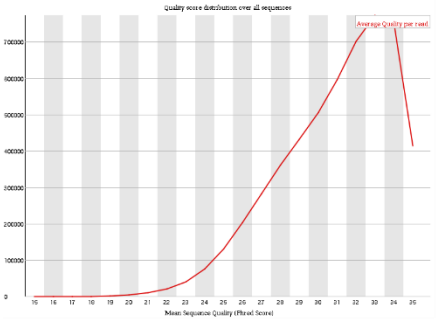
2 Illum Lin 10mM



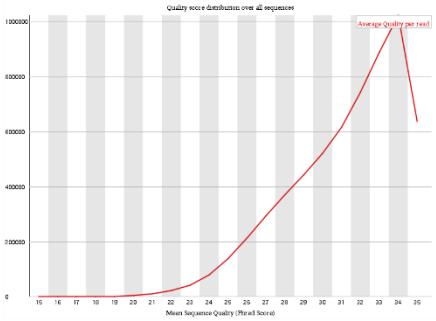
3 Illum Lin 10mM



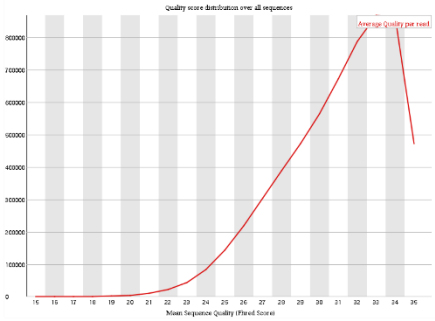
1 Dark Con (Lin)



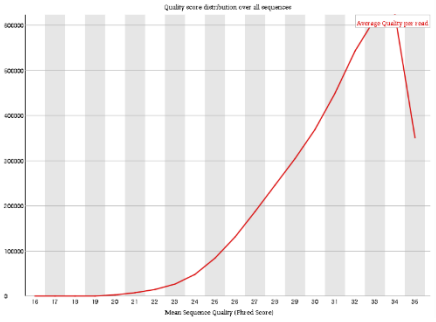
2 Dark Con (Lin)



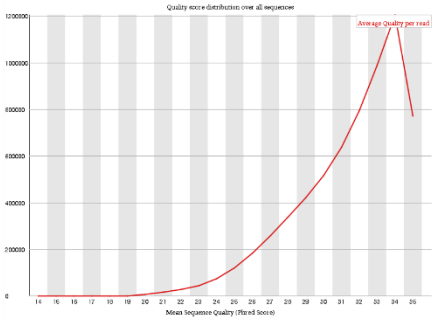
3 Dark Con (Lin)



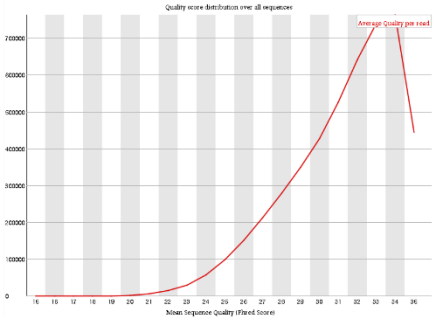
1 Dark Lin 10mM



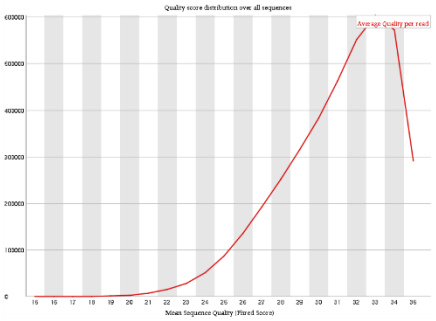
2 Dark Lin 10mM



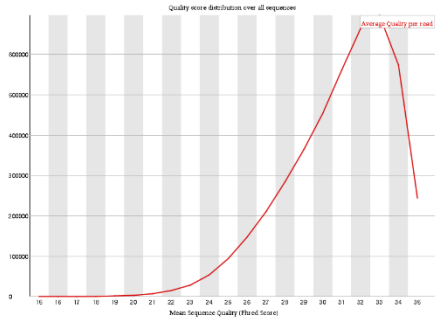
3 Dark Lin 10mM



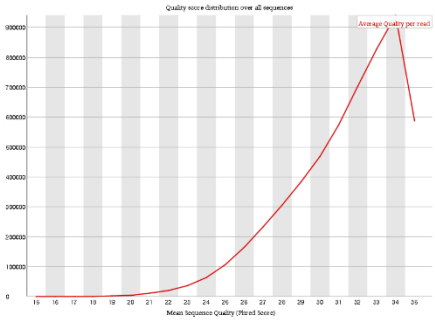
1 Illum Con (NF)



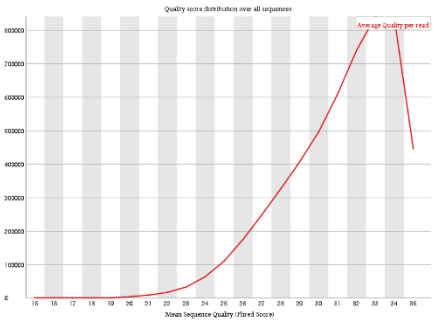
2 Illum Con (NF)



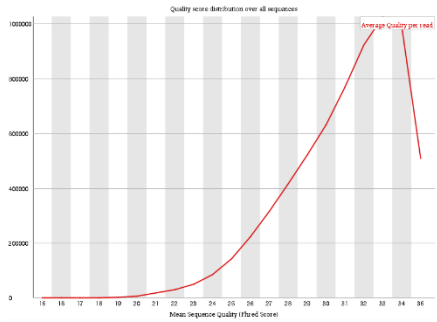
3 Illum Con (NF)



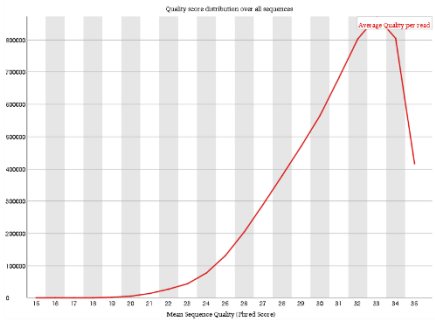
1 Illum NF 100µM



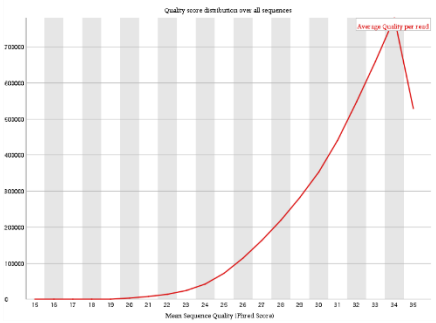
2 Illum NF 100µM



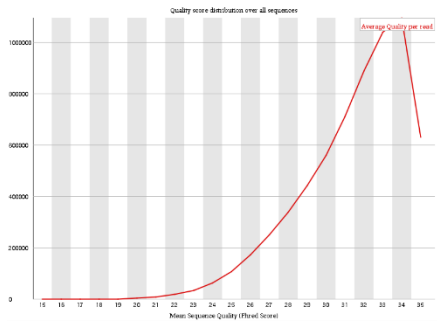
3 Illum NF 100µM



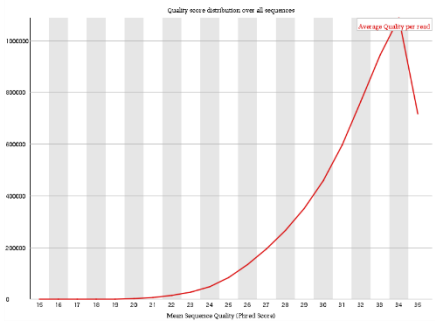
1 Dark Con (NF)



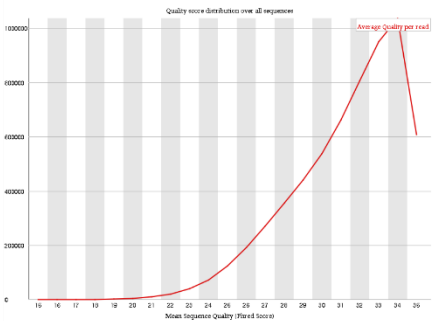
2 Dark Con (NF)



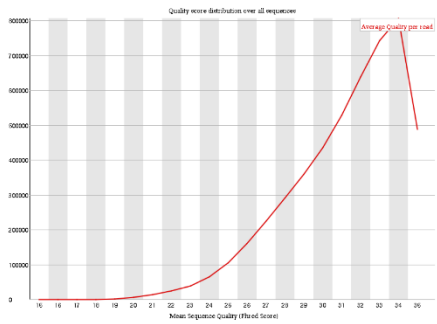
3 Dark Con (NF)



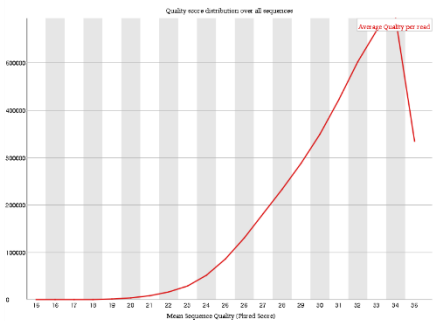
1 Dark NF 100µM



2 Dark NF 100µM



3 Dark NF 100µM



APPENDIX 16: THE 100 GENES WITH THE HIGHEST ABSOLUTE LOG2FOLD CHANGE IN TRANSCRIPT ABUNDANCE BETWEEN ILLUMINATED CONTROLS AND LINCOMYCIN TREATED *O. SATIVA* SEEDLINGS.

Rank	Ensembl Gene ID	Log2fold Change	Adjusted p-value
1	OS02G0240300	-5.645124968	3.66E-46
2	OS08G0155900	5.325211254	5.72E-71
3	OS01G0720500	-5.135457305	5.00E-166
4	OS05G0531100	4.980346384	1.04E-103
5	OS01G0791033	-4.835766977	3.16E-47
6	OS01G0392600	4.796043893	2.38E-27
7	OS08G0156100	4.753931735	1.50E-35
8	OS01G0600900	-4.672865608	1.20E-203
9	OS08G0156000	4.620688335	6.55E-54
10	OS04G0477300	4.487130764	3.93E-24
11	OS04G0600200	4.428647517	1.91E-100
12	OS03G0151500	4.392062037	9.57E-25
13	OS01G0695800	4.374573719	3.65E-19
14	OS02G0575000	4.307309694	3.65E-84
15	OS02G0533800	4.249006233	8.06E-127
16	OS12G0292400	-4.221099616	2.49E-27
17	OS12G0207600	-4.027322055	2.42E-15
18	OS10G0356000	-4.010560586	2.03E-33
19	OS01G0934800	3.960564242	6.59E-47
20	OS10G0527400	3.736690955	8.00E-19
21	OS01G0639900	-3.718281807	8.35E-27
22	OS03G0724500	3.712458589	1.35E-32
23	OS01G0238500	-3.660569675	1.78E-61
24	ENSRNA049472556	-3.57044448	3.49E-17
25	OS06G0227500	3.538310588	1.30E-10
26	OS02G0103800	-3.523913849	2.12E-21
27	OS03G0776900	3.485499501	8.97E-25
28	OS04G0414700	-3.459922069	2.36E-75
29	OS06G0146800	3.437406874	2.55E-16
30	OS12G0274700	-3.414913012	5.78E-131
31	OS02G0580500	3.383220179	8.62E-26
32	OS01G0866400	-3.369136318	7.66E-82
33	OS03G0802500	3.354346465	1.39E-59

Rank	Ensembl Gene ID	Log2fold Change	Adjusted p-value
34	OS07G0558300	-3.35084	4.57E-12
35	OS12G0186600	3.350293	5.63E-16
36	OS02G0758000	3.340068	6.02E-81
37	OS04G0650700	-3.32606	2.05E-52
38	OS10G0127900	3.278038	1.42E-29
39	OS12G0291100	-3.24067	4.34E-89
40	OS07G0142900	-3.22705	3.09E-21
41	OS10G0492300	-3.18358	3.11E-44
42	OS03G0592500	-3.17524	1.05E-85
43	OS03G0724600	3.160998	1.41E-19
44	OS01G0692000	3.133404	5.15E-12
45	OS09G0541000	-3.10278	6.84E-10
46	OS03G0626100	3.088324	2.57E-08
47	OS03G0841700	3.018432	4.46E-28
48	OS08G0371200	-2.99792	4.00E-07
49	OS03G0251350	-2.97223	1.24E-12
50	OS08G0529100	2.964766	4.97E-09
51	OS02G0197600	-2.96083	3.64E-13
52	OS08G0435900	-2.95345	5.82E-53
53	OS05G0542500	2.889188	1.29E-24
54	OS01G0692100	2.874667	7.16E-10
55	OS07G0148900	-2.84165	9.49E-67
56	OS05G0567600	-2.81785	7.24E-08
57	OS09G0460500	-2.77732	5.98E-12
58	OS06G0320500	-2.75484	1.01E-68
59	OS09G0346500	-2.7538	4.35E-58
60	OS01G0734800	2.753694	3.73E-13
61	OS05G0194500	2.731871	3.24E-08
62	OS01G0511600	-2.71897	3.74E-27
63	OS07G0147500	-2.68514	2.58E-08
64	OS01G0123900	-2.67649	4.52E-07
65	OS08G0485900	-2.67063	1.89E-23
66	OS11G0242800	-2.66711	1.39E-59
67	EPlOSAG00000035288	-2.6358	2.01E-05

Rank	Ensembl Gene ID	Log2fold Change	Adjusted p-value
68	OS01G0734600	-2.63163	2.89E-11
69	OS04G0352400	2.62473	1.21E-11
70	OS05G0427800	-2.61652	9.24E-06
71	OS12G0189400	-2.60274	6.63E-77
72	OS07G0558400	-2.59703	5.50E-72
73	OS07G0635500	2.583764	5.83E-07
74	OS10G0527800	2.571288	1.12E-15
75	OS02G0815400	-2.56796	2.19E-08
76	EPlOSAG00000023814	2.561107	2.27E-11
77	OS10G0503800	-2.55422	4.52E-29
78	OS03G0283200	-2.54683	4.03E-05
79	OS02G0181900	2.544817	1.44E-11
80	OS12G0291400	-2.53056	3.56E-52
81	OS10G0552700	-2.52693	5.99E-34
82	OS07G0562700	-2.52226	2.00E-53
83	OS02G0194700	-2.49605	2.30E-06
84	OS12G0420400	-2.48964	1.34E-43
85	OS02G0731700	-2.48316	3.80E-13
86	OS07G0162100	2.480703	2.17E-18
87	EPlOSAG00000019385	-2.47656	7.92E-05
88	OS07G0443500	-2.47429	4.64E-05
89	OS01G0975900	-2.46242	2.94E-05
90	OS12G0113600	2.454957	4.48E-11
91	OS11G0707000	-2.45309	2.04E-57
92	OS01G0760000	-2.42174	3.35E-09
93	OS06G0239200	-2.41652	3.97E-08
94	OS12G0204550	-2.40939	9.53E-05
95	OS12G0515400	-2.38653	8.42E-19
96	OS06G0608700	-2.38537	3.40E-64
97	OS06G0136600	2.370897	5.73E-32
98	OS12G0512100	2.36673	2.27E-18
99	OS07G0409500	2.351043	0.000169
100	OS04G0282400	-2.34616	0.000269

APPENDIX 17: THE 100 GENES WITH THE HIGHEST ABSOLUTE LOG2FOLD CHANGE IN TRANSCRIPT ABUNDANCE BETWEEN ILLUMINATED CONTROLS AND NORFLURAZON TREATED *O. SATIVA* SEEDLINGS.

Rank	Ensembl Gene ID	Log2fold Change	Adjusted p-value
1	OS01G0600900	-8.31573	2.36E-280
2	OS01G0720500	-8.04254	1.77E-276
3	OS02G0240300	-7.7749	2.64E-34
4	OS12G0274700	-7.47763	4.23E-206
5	OS12G0292400	-7.42715	3.97E-228
6	OS12G0248600	7.409774	2.39E-21
7	OS12G0291100	-7.38404	0
8	OS08G0435900	-7.3644	7.04E-269
9	OS02G0197600	-7.35685	2.98E-60
10	OS04G0414700	-7.19645	3.81E-226
11	OS07G0148900	-7.164	1.00E-194
12	OS01G0934800	7.155612	1.26E-98
13	OS06G0323100	7.027481	2.43E-18
14	OS03G0251350	-6.94212	1.28E-17
15	OS03G0592500	-6.91551	1.26E-251
16	OS03G0245800	6.904205	1.94E-17
17	OS09G0460500	-6.85761	3.51E-20
18	OS06G0320500	-6.81488	1.51E-188
19	OS08G0155900	6.790677	1.34E-204
20	OS09G0541000	-6.76548	6.31E-81
21	OS10G0527800	6.637388	1.54E-144
22	OS12G0291400	-6.6141	1.62E-272
23	OS01G0191200	-6.52783	4.56E-47
24	OS03G0841600	6.37945	6.51E-56
25	OS09G0346500	-6.36563	3.37E-288
26	OS01G0373700	6.30957	5.49E-14
27	OS06G0146600	6.30409	5.98E-14
28	OS01G0949800	6.247712	1.99E-13
29	OS03G0757600	6.245518	1.69E-13
30	OS10G0527400	6.217282	1.05E-78
31	OS01G0511600	-6.20913	3.58E-53
32	OS07G0638400	6.166406	4.40E-80
33	OS01G0915900	6.128946	6.72E-13

Rank	Ensembl Gene ID	Log2fold Change	Adjusted p-value
34	OS09G0493000	6.095726	1.24E-12
35	OS12G0226900	-6.0704	2.70E-35
36	OS11G0242800	-6.05926	6.03E-166
37	OS01G0695800	6.046811	3.12E-38
38	OS10G0492300	-6.04595	7.54E-99
39	OS04G0107900	6.036034	3.92E-15
40	OS01G0866400	-5.94879	6.54E-105
41	OS04G0339400	5.938559	9.40E-15
42	OS04G0457000	-5.93841	5.78E-144
43	OS06G0608700	-5.90668	1.12E-132
44	OS05G0567600	-5.90004	4.22E-11
45	OS07G0635200	5.857865	1.68E-11
46	OS01G0795200	5.817045	5.52E-13
47	OS11G0138300	5.811775	3.97E-11
48	OS01G0238500	-5.76356	2.09E-92
49	OS07G0558400	-5.76204	1.52E-165
50	OS08G0139700	-5.73729	3.31E-10
51	OS11G0303800	5.712718	1.07E-10
52	OS12G0134900	5.688442	6.79E-13
53	OS08G0156100	5.657931	1.06E-69
54	OS10G0378100	-5.6549	6.89E-10
55	OS08G0371200	-5.58933	9.18E-10
56	OS03G0711100	-5.56958	1.16E-33
57	OS07G0577600	-5.54965	4.98E-187
58	OS01G0373800	5.514177	8.81E-10
59	OS08G0379400	-5.50583	5.29E-59
60	OS02G0103800	-5.4991	8.61E-141
61	OS01G0585200	5.488546	1.15E-09
62	OS07G0562700	-5.48113	2.04E-223
63	OS01G0615100	5.476311	1.69E-09
64	OS05G0508900	-5.45924	1.55E-27
65	OS01G0303300	5.459136	1.44E-09
66	OS02G0698000	-5.45619	2.94E-22
67	OS04G0206600	5.444068	1.83E-09

Rank	Ensembl Gene ID	Log2fold Change	Adjusted p-value
68	OS01G0692100	5.410854	3.63E-87
69	OS02G0110200	-5.40001	1.28E-30
70	OS12G0420400	-5.39025	2.18E-159
71	OS07G0625400	5.369912	1.61E-23
72	OS02G0815300	-5.36838	1.22E-46
73	OS01G0123900	-5.34494	6.36E-12
74	OS07G0558300	-5.31362	1.45E-13
75	OS03G0724600	5.306229	2.84E-45
76	OS12G0420200	-5.30309	4.28E-102
77	OS06G0227500	5.301048	4.72E-13
78	OS12G0189400	-5.22896	2.52E-115
79	OS06G0107100	-5.21303	3.35E-35
80	OS09G0279400	-5.20533	6.85E-67
81	OS04G0127300	-5.16078	1.51E-16
82	OS10G0344000	5.136732	4.52E-10
83	OS07G0443500	-5.13165	1.73E-12
84	OS10G0528300	5.103205	4.37E-130
85	OS01G0791033	-5.09773	1.12E-105
86	OS12G0204550	-5.07962	1.47E-07
87	OS04G0659100	-5.06295	4.77E-114
88	OS12G0207600	-5.05789	1.38E-19
89	OS10G0205700	-5.05763	1.89E-11
90	OS05G0147800	5.051104	7.49E-08
91	OS11G0506800	5.049829	7.49E-08
92	OS06G0697500	5.046287	1.21E-09
93	OS11G0533400	5.042556	1.28E-16
94	OS03G0782200	-5.03355	1.35E-08
95	OS11G0707000	-5.01105	4.83E-32
96	OS11G0147150	4.986328	2.96E-21
97	OS12G0124000	-4.97016	2.20E-25
98	OS01G0647200	-4.93681	2.07E-10
99	OS01G0155000	4.919474	5.33E-27
100	OS01G0949900	4.919433	2.60E-07

APPENDIX 18: LIST OF ENRICHED GENE ONTOLOGY (GO) TERMS IN *O. SATIVA* GENES THAT SHOW SIGNIFICANTLY DIFFERENT TRANSCRIPT ABUNDANCE IN RESPONSE TO LINCOMYN TREATMENT

GO term	Ontology	Description	Number in input list	Number in BG/Ref	p-value	FDR
GO:0015979	P	photosynthesis	42	70	1.6E-26	2.5E-23
GO:0006091	P	generation of precursor metabolites and energy	62	313	2.3E-17	1.8E-14
GO:0034641	P	cellular nitrogen compound metabolic process	51	252	7.7E-15	4E-12
GO:0019684	P	photosynthesis, light reaction	18	24	2.4E-13	9.4E-11
GO:0009765	P	photosynthesis, light harvesting	16	17	4.8E-13	1.5E-10
GO:0055114	P	oxidation reduction	64	543	5.7E-09	0.0000015
GO:0006082	P	organic acid metabolic process	58	493	0.000000031	0.0000069
GO:0006519	P	cellular amino acid and derivative metabolic process	48	376	0.000000054	0.00001
GO:0043436	P	oxoacid metabolic process	57	492	0.000000066	0.00001
GO:0019752	P	carboxylic acid metabolic process	57	492	0.000000066	0.00001
GO:0042180	P	cellular ketone metabolic process	57	499	0.0000001	0.000014
GO:0006520	P	cellular amino acid metabolic process	40	290	0.00000011	0.000014
GO:0044106	P	cellular amine metabolic process	40	291	0.00000012	0.000014
GO:0044271	P	cellular nitrogen compound biosynthetic process	31	201	0.00000035	0.000039
GO:0015977	P	carbon fixation	9	13	0.00000042	0.000043
GO:0016053	P	organic acid biosynthetic process	33	235	0.00000099	0.00009
GO:0046394	P	carboxylic acid biosynthetic process	33	235	0.00000099	0.00009
GO:0008652	P	cellular amino acid biosynthetic process	23	128	0.0000012	0.000099
GO:0006006	P	glucose metabolic process	20	100	0.0000014	0.00011
GO:0019318	P	hexose metabolic process	22	123	0.0000021	0.00016
GO:0008152	P	metabolic process	489	8041	0.0000031	0.00023
GO:0065003	P	macromolecular complex assembly	22	131	0.0000052	0.00036
GO:0034728	P	nucleosome organization	14	56	0.0000058	0.00038
GO:0006334	P	nucleosome assembly	14	56	0.0000058	0.00038
GO:0009308	P	amine metabolic process	41	359	0.0000062	0.00038
GO:0009309	P	amine biosynthetic process	23	144	0.0000067	0.0004
GO:0031497	P	chromatin assembly	14	57	0.0000069	0.0004
GO:0006323	P	DNA packaging	14	58	0.0000082	0.00045
GO:0022607	P	cellular component assembly	22	137	0.0000097	0.00052
GO:0034622	P	cellular macromolecular complex assembly	21	127	0.00001	0.00053
GO:0065004	P	protein-DNA complex assembly	14	60	0.000011	0.00057
GO:0022900	P	electron transport chain	24	162	0.000013	0.00063
GO:0044262	P	cellular carbohydrate metabolic process	33	274	0.000019	0.00087
GO:0043933	P	macromolecular complex subunit organization	22	144	0.000019	0.00088
GO:0034621	P	cellular macromolecular complex subunit organization	21	139	0.000034	0.0015
GO:0005996	P	monosaccharide metabolic process	23	162	0.000036	0.0015
GO:0045454	P	cell redox homeostasis	18	112	0.00006	0.0025
GO:0006544	P	glycine metabolic process	7	16	0.000078	0.0032
GO:0009069	P	serine family amino acid metabolic process	10	40	0.00012	0.0049
GO:0009081	P	branched chain family amino acid metabolic process	8	25	0.00014	0.0056
GO:0044275	P	cellular carbohydrate catabolic process	15	90	0.00017	0.0064
GO:0019725	P	cellular homeostasis	18	125	0.00021	0.0075
GO:0009082	P	branched chain family amino acid biosynthetic process	6	13	0.00021	0.0075
GO:0019320	P	hexose catabolic process	14	84	0.00028	0.0093
GO:0006007	P	glucose catabolic process	14	84	0.00028	0.0093
GO:0044085	P	cellular component biogenesis	24	201	0.00027	0.0093
GO:0042592	P	homeostatic process	18	132	0.00038	0.012
GO:0006457	P	protein folding	20	159	0.00047	0.015
GO:0006066	P	alcohol metabolic process	24	210	0.00049	0.015
GO:0015995	P	chlorophyll biosynthetic process	5	10	0.00054	0.017

GO term	Ontology	Description	Number in input list	Number in BG/Ref	p-value	FDR
GO:0006096	P	glycolysis	11	67	0.0014	0.041
GO:0018130	P	heterocycle biosynthetic process	10	57	0.0014	0.042
GO:0046164	P	alcohol catabolic process	15	113	0.0014	0.042
GO:0016491	F	oxidoreductase activity	131	1387	8.8E-11	0.000000075
GO:0048037	F	cofactor binding	55	419	2.4E-09	0.000001
GO:0016984	F	ribulose-bisphosphate carboxylase activity	8	12	0.0000023	0.00065
GO:0005198	F	structural molecule activity	41	367	0.00001	0.0021
GO:0019842	F	vitamin binding	22	140	0.000013	0.0022
GO:0030170	F	pyridoxal phosphate binding	19	113	0.000022	0.0023
GO:0070279	F	vitamin B6 binding	19	113	0.000022	0.0023
GO:0050662	F	coenzyme binding	34	288	0.00002	0.0023
GO:0016829	F	lyase activity	32	277	0.000051	0.0049
GO:0016830	F	carbon-carbon lyase activity	16	96	0.0001	0.0089
GO:0016831	F	carboxy-lyase activity	13	67	0.00012	0.0094
GO:0016616	F	oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor	18	122	0.00016	0.011
GO:0051287	F	NAD or NADH binding	11	53	0.00024	0.015
GO:0003824	F	catalytic activity	464	7902	0.00024	0.015
GO:0003735	F	structural constituent of ribosome	30	286	0.0004	0.02
GO:0004365	F	glyceraldehyde-3-phosphate dehydrogenase (phosphorylating) activity	5	9	0.00038	0.02
GO:0008943	F	glyceraldehyde-3-phosphate dehydrogenase activity	5	9	0.00038	0.02
GO:0016765	F	transferase activity, transferring alkyl or ary (other than methyl) groups	13	82	0.00069	0.033
GO:0016614	F	oxidoreductase activity, acting on CH-OH group of donors	18	143	0.00088	0.04
GO:0004364	F	glutathione transferase activity	9	44	0.00095	0.041
GO:0009536	C	plastid	391	4949	1.5E-19	5.7E-17
GO:0009507	C	chloroplast	58	257	1.4E-18	2.7E-16
GO:0009579	C	thylakoid	33	72	2.2E-18	2.8E-16
GO:0034357	C	photosynthetic membrane	30	66	9.1E-17	8.7E-15
GO:0044436	C	thylakoid part	24	54	1.6E-13	1.2E-11
GO:0009521	C	photosystem	18	35	2.7E-11	1.7E-09
GO:0044435	C	plastid part	27	99	6.1E-11	3.4E-09
GO:0044434	C	chloroplast part	26	95	1.3E-10	6.2E-09
GO:0032991	C	macromolecular complex	119	1227	1.6E-10	6.9E-09
GO:0009534	C	chloroplast thylakoid	16	32	4.7E-10	0.000000016
GO:0031976	C	plastid thylakoid	16	32	4.7E-10	0.000000016
GO:0042651	C	thylakoid membrane	17	38	5.1E-10	0.000000016
GO:0031984	C	organelle subcompartment	16	35	1.3E-09	0.000000038
GO:0009535	C	chloroplast thylakoid membrane	14	30	0.000000011	0.00000029
GO:0055035	C	plastid thylakoid membrane	14	30	0.000000011	0.00000029
GO:0044422	C	organelle part	89	902	0.000000018	0.00000041
GO:0044446	C	intracellular organelle part	89	902	0.000000018	0.00000041
GO:0009523	C	photosystem II	11	21	0.00000017	0.0000037
GO:0005737	C	cytoplasm	826	14583	0.0000012	0.000025
GO:0009522	C	photosystem I	8	12	0.0000023	0.000042
GO:0000786	C	nucleosome	14	51	0.0000023	0.000042
GO:0043232	C	intracellular non-membrane-bounded organelle	66	684	0.0000026	0.000044
GO:0043228	C	non-membrane-bounded organelle	66	684	0.0000026	0.000044
GO:0009532	C	plastid stroma	12	37	0.0000029	0.000047
GO:0032993	C	protein-DNA complex	14	54	0.0000041	0.000063
GO:0044424	C	intracellular part	890	16030	0.0000076	0.00011
GO:0009570	C	chloroplast stroma	11	35	0.0000096	0.00014
GO:0009654	C	oxygen evolving complex	8	17	0.000016	0.00022
GO:0044444	C	cytoplasmic part	803	14318	0.00002	0.00026
GO:0048492	C	ribulose bisphosphate carboxylase complex	6	9	0.000045	0.00054
GO:0009573	C	chloroplast ribulose bisphosphate carboxylase complex	6	9	0.000045	0.00054
GO:0005622	C	intracellular	899	16347	0.000044	0.00054
GO:0009538	C	photosystem I reaction center	5	5	0.000056	0.00065
GO:0043234	C	protein complex	68	792	0.00006	0.00068
GO:0043229	C	intracellular organelle	858	15631	0.00021	0.0022
GO:0043226	C	organelle	858	15631	0.00021	0.0022
GO:0005840	C	ribosome	33	316	0.00022	0.0023
GO:0030529	C	ribonucleoprotein complex	37	384	0.00041	0.0042
GO:0045259	C	proton-transporting ATP synthase complex	6	22	0.002	0.019
GO:0031090	C	organelle membrane	31	335	0.0021	0.02
GO:0019898	C	extrinsic to membrane	5	15	0.0023	0.021
GO:0000785	C	chromatin	15	120	0.0024	0.022
GO:0005829	C	cytosol	14	112	0.0033	0.03

APPENDIX 19: LIST OF ENRICHED GENE ONTOLOGY (GO) TERMS IN *O. SATIVA* GENES THAT SHOW SIGNIFICANTLY DIFFERENT TRANSCRIPT ABUNDANCE IN RESPONSE TO NORFLURAZON TREATMENT

GO term	Ontology	Description	Number in input list	Number in BG/Ref	p-value	FDR
GO:0006412	P	translation	167	480	5.8E-16	7.3E-13
GO:0008152	P	metabolic process	1563	8041	4.7E-16	7.3E-13
GO:0034641	P	cellular nitrogen compound metabolic process	107	252	5E-15	4.2E-12
GO:0006091	P	generation of precursor metabolites and energy	120	313	3.8E-14	1.9E-11
GO:0015979	P	photosynthesis	51	70	3.2E-14	1.9E-11
GO:0055114	P	oxidation reduction	169	543	1.1E-12	4.6E-10
GO:0044262	P	cellular carbohydrate metabolic process	90	274	2.8E-08	0.00001
GO:0006457	P	protein folding	60	159	1.3E-07	0.000042
GO:0006006	P	glucose metabolic process	44	100	2.3E-07	0.000064
GO:0019318	P	hexose metabolic process	49	123	5.2E-07	0.000097
GO:0043436	P	oxoacid metabolic process	132	492	4.9E-07	0.000097
GO:0042180	P	cellular ketone metabolic process	134	499	3.8E-07	0.000097
GO:0006082	P	organic acid metabolic process	132	493	5.3E-07	0.000097
GO:0019752	P	carboxylic acid metabolic process	132	492	4.9E-07	0.000097
GO:0044249	P	cellular biosynthetic process	547	2735	1.8E-06	0.0003
GO:0009765	P	photosynthesis, light harvesting	16	17	2.2E-06	0.00035
GO:0009058	P	biosynthetic process	570	2873	2.4E-06	0.00036
GO:0005975	P	carbohydrate metabolic process	172	713	3.1E-06	0.00044
GO:0051186	P	cofactor metabolic process	52	146	3.5E-06	0.00046
GO:0019684	P	photosynthesis, light reaction	18	24	4.8E-06	0.00061
GO:0044237	P	cellular metabolic process	1074	5837	8.2E-06	0.00099
GO:0044275	P	cellular carbohydrate catabolic process	36	90	0.000015	0.0017
GO:0019320	P	hexose catabolic process	34	84	0.000021	0.0022
GO:0006007	P	glucose catabolic process	34	84	0.000021	0.0022
GO:0022900	P	electron transport chain	52	162	0.000036	0.0035
GO:0005996	P	monosaccharide metabolic process	52	162	0.000036	0.0035
GO:0044271	P	cellular nitrogen compound biosynthetic process	60	201	0.000055	0.0052
GO:0033013	P	tetrapyrrole metabolic process	18	31	0.000059	0.0054
GO:0033014	P	tetrapyrrole biosynthetic process	16	25	0.000065	0.0057
GO:0006778	P	porphyrin metabolic process	17	29	0.000086	0.0073
GO:0006779	P	porphyrin biosynthetic process	15	23	0.000093	0.0076
GO:0042440	P	pigment metabolic process	17	30	0.00012	0.0093
GO:0044248	P	cellular catabolic process	93	366	0.00012	0.0093
GO:0046148	P	pigment biosynthetic process	15	24	0.00013	0.0096
GO:0008299	P	isoprenoid biosynthetic process	21	44	0.00013	0.0096
GO:0006096	P	glycolysis	27	67	0.00015	0.01
GO:0033036	P	macromolecule localization	78	296	0.00016	0.011
GO:0046394	P	carboxylic acid biosynthetic process	65	235	0.00018	0.012
GO:0016053	P	organic acid biosynthetic process	65	235	0.00018	0.012
GO:0018130	P	heterocycle biosynthetic process	24	57	0.0002	0.013
GO:0065003	P	macromolecular complex assembly	41	131	0.00035	0.022
GO:0015994	P	chlorophyll metabolic process	11	15	0.00039	0.024
GO:0046483	P	heterocycle metabolic process	71	273	0.00042	0.025
GO:0009060	P	aerobic respiration	15	28	0.00046	0.027
GO:0016052	P	carbohydrate catabolic process	46	156	0.00048	0.027
GO:0015995	P	chlorophyll biosynthetic process	9	10	0.00049	0.027
GO:0044106	P	cellular amine metabolic process	74	291	0.00055	0.03
GO:0043933	P	macromolecular complex subunit organization	43	144	0.00058	0.03
GO:0046164	P	alcohol catabolic process	36	113	0.00059	0.031
GO:0006066	P	alcohol metabolic process	57	210	0.00063	0.032

GO term	Ontology	Description	Number in input list	Number in BG/Ref	p-value	FDR
GO:0044265	P	cellular macromolecule catabolic process	71	279	0.00069	0.034
GO:0022607	P	cellular component assembly	41	137	0.00073	0.036
GO:0006520	P	cellular amino acid metabolic process	73	290	0.00076	0.036
GO:0032787	P	monocarboxylic acid metabolic process	49	175	0.00083	0.039
GO:0006732	P	coenzyme metabolic process	33	103	0.00091	0.041
GO:0051187	P	cofactor catabolic process	14	27	0.0009	0.041
GO:0046365	P	monosaccharide catabolic process	34	108	0.00098	0.044
GO:0008610	P	lipid biosynthetic process	52	191	0.001	0.044
GO:0015980	P	energy derivation by oxidation of organic compounds	19	46	0.0011	0.044
GO:0006099	P	tricarboxylic acid cycle	12	21	0.0011	0.044
GO:0009109	P	coenzyme catabolic process	12	21	0.0011	0.044
GO:0019748	P	secondary metabolic process	40	136	0.0011	0.044
GO:0046356	P	acetyl-CoA catabolic process	12	21	0.0011	0.044
GO:0006084	P	acetyl-CoA metabolic process	12	21	0.0011	0.044
GO:0006720	P	isoprenoid metabolic process	27	79	0.0012	0.046
GO:0005198	F	structural molecule activity	158	367	5.1E-22	8.6E-19
GO:0003735	F	structural constituent of ribosome	132	286	1.3E-20	1.1E-17
GO:0016491	F	oxidoreductase activity	372	1387	1.2E-17	7E-15
GO:0003824	F	catalytic activity	1458	7902	0.000000019	0.0000081
GO:0048037	F	cofactor binding	121	419	0.000000059	0.00002
GO:0050662	F	coenzyme binding	84	288	0.0000043	0.0012
GO:0016765	F	transferase activity, transferring alkyl or aryl (other than methyl) groups	35	82	0.0000064	0.0015
GO:0004364	F	glutathione transferase activity	24	44	0.0000082	0.0017
GO:0016853	F	isomerase activity	60	190	0.000014	0.0027
GO:0009055	F	electron carrier activity	158	667	0.000018	0.003
GO:0005506	F	iron ion binding	143	616	0.000097	0.015
GO:0016859	F	cis-trans isomerase activity	22	47	0.00011	0.015
GO:0003755	F	peptidyl-prolyl cis-trans isomerase activity	22	47	0.00011	0.015
GO:0032561	F	guanyl ribonucleotide binding	51	171	0.00019	0.02
GO:0005525	F	GTP binding	51	171	0.00019	0.02
GO:0016830	F	carbon-carbon lyase activity	34	96	0.00017	0.02
GO:0070003	F	threonine-type peptidase activity	14	23	0.00027	0.025
GO:0004298	F	threonine-type endopeptidase activity	14	23	0.00027	0.025
GO:0019001	F	guanyl nucleotide binding	51	175	0.00031	0.027
GO:0016209	F	antioxidant activity	50	177	0.00063	0.048
GO:0019842	F	vitamin binding	42	140	0.00061	0.048
GO:0016829	F	lyase activity	71	277	0.00059	0.048
GO:0005840	C	ribosome	141	316	7.2E-21	4.8E-18
GO:0030529	C	ribonucleoprotein complex	154	384	3.5E-19	1.2E-16
GO:0009536	C	plastid	1040	4949	2.4E-17	5.3E-15
GO:0005737	C	cytoplasm	2623	14583	2.7E-16	4.6E-14
GO:0032991	C	macromolecular complex	330	1227	7.5E-16	1E-13
GO:0043232	C	intracellular non-membrane-bounded organelle	204	684	1.4E-13	1.4E-11
GO:0043228	C	non-membrane-bounded organelle	204	684	1.4E-13	1.4E-11
GO:0044444	C	cytoplasmic part	2556	14318	1.8E-13	1.5E-11
GO:0009507	C	chloroplast	96	257	4.2E-11	3.2E-09
GO:0044424	C	intracellular part	2797	16030	5.2E-11	3.5E-09
GO:0009579	C	thylakoid	43	72	3.3E-10	0.00000002
GO:0005622	C	intracellular	2827	16347	0.000000003	0.00000017
GO:0034357	C	photosynthetic membrane	38	66	7.3E-09	0.00000038
GO:0043229	C	intracellular organelle	2707	15631	0.000000018	0.0000008
GO:0043226	C	organelle	2707	15631	0.000000018	0.0000008
GO:0044436	C	thylakoid part	32	54	0.000000067	0.0000028
GO:0009521	C	photosystem	25	35	0.00000014	0.0000056
GO:0005829	C	cytosol	44	112	0.0000026	0.000096
GO:0009523	C	photosystem II	17	21	0.0000043	0.00014
GO:0044464	C	cell part	3115	18466	0.0000041	0.00014
GO:0005623	C	cell	3115	18467	0.0000042	0.00014
GO:0044434	C	chloroplast part	38	95	0.0000086	0.00027
GO:0044435	C	plastid part	38	99	0.000018	0.00053
GO:0009654	C	oxygen evolving complex	14	17	0.000027	0.00075
GO:0044422	C	organelle part	199	902	0.000062	0.0016
GO:0044446	C	intracellular organelle part	199	902	0.000062	0.0016
GO:0043231	C	intracellular membrane-bounded organelle	2594	15273	0.00013	0.0033
GO:0043227	C	membrane-bounded organelle	2594	15274	0.00013	0.0033
GO:0042651	C	thylakoid membrane	19	38	0.00017	0.0039
GO:0031984	C	organelle subcompartment	18	35	0.00019	0.0043
GO:0009534	C	chloroplast thylakoid	17	32	0.00021	0.0045
GO:0031976	C	plastid thylakoid	17	32	0.00021	0.0045
GO:0005839	C	proteasome core complex	14	23	0.00027	0.0055
GO:0019898	C	extrinsic to membrane	11	15	0.00039	0.0079
GO:0009570	C	chloroplast stroma	17	35	0.00047	0.0091
GO:0055035	C	plastid thylakoid membrane	15	30	0.0008	0.014
GO:0009535	C	chloroplast thylakoid membrane	15	30	0.0008	0.014
GO:0009532	C	plastid stroma	17	37	0.00077	0.014
GO:0033279	C	ribosomal subunit	18	42	0.001	0.018
GO:0009522	C	photosystem I	9	12	0.0012	0.02
GO:0000502	C	proteasome complex	14	28	0.0012	0.02
GO:0000786	C	nucleosome	20	51	0.0013	0.021
GO:0032993	C	protein-DNA complex	20	54	0.0023	0.036

APPENDIX 20: GENE IDS FOR C_4 ORTHOLOGUES IN *O. SATIVA*

C_4 Orthologue	Ensembl Gene ID
ACT	Os11g0707000
Ala-AT1	Os10g0390500
BASS2	Os01g0645200
CA1	Os01g0639900
ME2	Os05g0186300
ME3	Os01g0743500
NADP-MDH	Os08g0562100
NADP-ME1	Os01g0188400
NADP-ME2	Os01g0723400
OMT1	Os12g0515400
PDRP1	Os07g0530600
PEPC1	Os02g0244700
PEPC2	Os08g0366000
PEPCK1	Os10g0204400
PEPCK2	Os03g0255500
PPC4	Os01g0208700
PPDKA	Os03g0432100
PPDKB	Os05g0405000
RER4	Os05g0388600

APPENDIX 21: LIST OF MOTIFS ENRICHED >2.5 ABOVE BACKGROUND IN THE FIMO ANALYSIS OF 1000BP UPSTREAM + 5' UTR OF *O. SATIVA* GENES THAT SHOW SIGNIFICANTLY DIFFERENT TRANSCRIPT ABUNDANCE IN RESPONSE TO CHLOROPLAST INHIBITION

Motif ID	Total Frequency	Ratio compared to Average Background Frequency
MYB_tnt.MYB13_col_a_m1	36	7.297137
C2H2_tnt.AZF1_colamp_a_m1	23	6.751948
C2H2_tnt.AT4G26030_col_a_m1	28	4.966107
MYB_tnt.MYB58_colamp_a_m1	40	4.935262
MYB_tnt.MYB58_col_a_m1	28	4.864758
MYB_tnt.MYB13_colamp_a_m1	34	4.668598
MYB_tnt.MYB83_col_a_m1	42	4.360484
MYB_tnt.ATY19_col_a_m1	55	4.295715
MYB_tnt.MYB63_col_a_m1	51	4.057754
MYB_tnt.MYB83_colamp_a_m1	36	3.980256
bHLH_tnt.bHLH34_colamp_a_m1	28	3.667279
bHLH_tnt.bHLH34_col_m1	33	3.648568
Orphan_tnt.AT1G24250_col_a_m1	21	3.575597
bZIP_tnt.AREB3_col_v31_m1	39	3.494944
MYBrelated_tnt.At3g09600_colamp_a_m1	26	3.458539
MYBrelated_tnt.At4g01280_colamp_a_m1	26	3.458539
MYBrelated_tnt.At4g01280_col_a_m1	25	3.325518
bZIP_tnt.AREB3_colamp_a_m1	38	3.301086
bHLH_tnt.bHLH104_col_b_m1	23	3.209943
MYBrelated_tnt.EPR1_colamp_a_m1	23	3.209943
bZIP_tnt.bZIP16_col_v3a_m1	70	3.136489
bZIP_tnt.GBF3_colamp_m1	59	2.989799
MYB_tnt.MYB96_col_b_m1	29	2.974536
MYBrelated_tnt.At5g52660_col_a_m1	22	2.926456
MYBrelated_tnt.LCL1_colamp_a_m1	21	2.837775
MYB_tnt.MYB61_col_a_m1	44	2.837775
bZIP_tnt.GBF5_colamp_a_m1	51	2.819348
bZIP_tnt.ABI5_col_v3h_m1	46	2.81736
bZIP_tnt.GBF3_col_m1	63	2.793435
bZIP_tnt.GBF5_col_v3a_m1	53	2.71811
bZIP_tnt.ABF2_col_v3a_m1	30	2.717019
C2C2gata_tnt.ZIM_col_a_m1	38	2.695887
bZIP_tnt.bZIP3_col_a_m1	60	2.688419
C2C2gata_tnt.ZML1_colamp_a_m1	45	2.660414
bZIP_tnt.bZIP16_colamp_a_m1	58	2.640497
bZIP_tnt.ABI5_colamp_v3b_m1	47	2.63241
MYBrelated_tnt.At3g09600_col_a_m1	21	2.629115
bZIP_tnt.bZIP68_col_a_m1	60	2.553998

APPENDIX 22: LIST OF MOTIFS ENRICHED WITH A P-VALUE >0.01 IN THE AME ANALYSIS OF 1000BP UPSTREAM + 5' UTR OF *O. SATIVA* GENES THAT SHOW SIGNIFICANTLY DIFFERENT TRANSCRIPT ABUNDANCE IN RESPONSE TO CHLOROPLAST INHIBITION

Rank	Motif ID	Adjusted p-value
1	bZIP_tnt.bZIP16_col_v3a_m1	2.94E-06
2	bZIP_tnt.AREB3_colamp_a_m1	7.51E-05
3	bZIP_tnt.AREB3_col_v31_m1	1.84E-04
4	bZIP_tnt.ABI5_colamp_v3b_m1	7.13E-04
5	bZIP_tnt.ABI5_col_v3h_m1	7.33E-04
7	bZIP_tnt.bZIP28_col_a_m1	1.21E-03
8	bZIP_tnt.bZIP42_colamp_a_m1	1.53E-03
9	bZIP_tnt.GBF3_col_m1	2.74E-03
10	C2H2_tnt.TF3A_col_a_m1	2.76E-03
11	bZIP_tnt.bZIP3_col_a_m1	4.51E-03
12	NAC_tnt.ANAC017_colamp_v31_m1	6.90E-03
13	bZIP_tnt.ABF2_col_v3a_m1	7.45E-03