Gene expression variation under dynamic growth conditions in *Arabidopsis thaliana*

Kevin Murray

Borevitz Lab, ANU

Thursday 24th October, 2013

Thesis submitted in partial fulfilment of the requirements of the degree of

Bachelor of Philosophy (Science) (Honours)

Aproximate Chapter Word Counts:

Introduction:	2010 words
Dynamic Growth Conditions:	1250 words
Improved RNAseq Analysis:	1545 words
Transcriptome Variation:	1870 words
Discussion:	3320 words

Abstract

The study of plant interaction with the environment, and in particular responses to stresses imposed by their environment, is of crucial importance to wide-ranging areas of botany, from agricultural crop development to global change biology. In this study, I develop methods to reproducibly grow plants in growth conditions which encompass combinations of climactic variables analogous to those that can appear naturally, termed dynamic growth conditions. Additionally, I develop bioinformatic protocols enabling the analysis of data from experiments aiming to determine environmental and genetic effects on genome-wide gene expression (the transcriptome). I present and analyse a preliminary dataset examining the transcriptomes of plants growth under dynamic growth conditions, aiming to test plant response to altered light intensity in the context of combinatorial abiotic stresses as recent abiotic stress research advocates. Analysing this dataset revealed a dataset characterised by high levels of biological variation in expression, and uncovered limited differential expression. Together with quantitative PCR examination of excess light-responsive genes, these data provide tentative evidence of field-like hardening to excess light stress by plants grown under dynamic growth conditions, a finding that warrants further investigation.

Contents

Acknowledgements

For their instrumental assistance in the creation of this thesis, I repay the following people with the meagre sum of my eternal gratitude:

- My supervisors, Prof. Justin Borevitz and Prof. Barry Pogson, for the opportunity to conduct this project, and for their advice, encouragement and tolerance throughout the year.
- Peter Crisp and Norman Warthmann, who deserve special mention for their evenings and weekends spend showing me how it's all done.
- The entire Borevitz and Pogson labs, who have been extraordinarily helpful, and a pleasure work with.
- My parents, for editorial advice and so much more.
- My friends, for putting up with my absence and keeping me within reach of sanity throughout the year.
- And finally, my examiners, Prof. Owen Atkin, Dr. Arun Yadav and A/Prof. Georg Weiller, for their encouragement, advice, and constructive criticism.

Thank you all,
Kevin

Chapter 1

Introduction

Nearly all terrestrial biomass, including human life, depends on the primary productivity of higher plants (photosynthesis) (Johnston et al. 2009). This primary productivity is reduced when plants are stressed by interaction with their environment (Mittler 2006; Mittler and Blumwald 2010). Abiotic stresses alone account for over 100 billion dollars of crop losses per annum in the United States alone (Mittler 2006). Therefore, the study of plant-environment interactions is crucial to improving agricultural yield and understanding functional plant ecology in a changing climate.

Central to increasing the understanding of plant-environment interactions is the development of laboratory study systems. Specifically, systems that allow classical mechanistic studies of abiotic stress response to be placed in the context of the conditions plants experience in natural or cultivated environments outside laboratories are of importance to studies of stress response. In this thesis, I discuss, develop and apply novel techniques that enable the global study of gene expression in laboratory growth conditions which mimic field-like combinations of light, temperature and humidity. I examine the effect of light intensity on global gene expression, within the framework of growth conditions whose temperature, humidity and light follow trends approximating those observed in regional climates.

1.1 Abiotic Stress: A limit to Plant Productivity

Abiotic stresses are the non-living stresses imposed upon plants by their growth environment. This includes deleterious extremes in environmental variables such as temperature, humidity, osmotic potential, water availability, light quality and quantity or combinations of these variables. Specifically, osmotic and drought stresses cause a decrease in the ability of plants to transpire and obtain inorganic carbon vital to photosynthesis (Seki et al. 2003; Mittler 2006). Excess light creates harmful reactive oxygen species (ROS) that damage the delicate photosynthetic apparatus (Niyogi 1999; Apel and Hirt 2004; Asada 2006; Li et al. 2009; Foyer and Noctor 2009; Mubarakshina et al. 2010). Extremes of temperature impair and damage many enzymes, particularly those involved in metabolism (Atkin and Tjoelker 2003). Together, abiotic stresses limit plant productivity through their detrimental effects upon the ability of plants to assimilate biomass, and secondarily via coping mechanisms which some plants have evolved (Mittler 2006). As a result, enormous effort has been directed towards the elucidation and improvement of these coping mechanisms. In particular, the mechanisms by which excess light impacts upon plants, and how plants respond to this stress, have been major foci of research.

Plants have developed a myriad of mechanisms to respond to abiotic stress. Therefore responses to abiotic stress must be studied via a variety of techniques (Demmig-Adams and Adams 1992). Studies of photosynthetic characteristics by chlorophyll fluorescence (reviewed in Baker 2008) have illustrated the negative effects of abiotic stress on photosynthesis (Külheim, Ågren, and Jansson 2002; Mishra et al. 2012; Alter et al. 2012; Tikkanen et al. 2012). Phenotypic analysis, particularly of plant morphology, has given insight into the effects of abiotic stress (Armstrong, Wardlaw, and Atkin 2007; Wituszyńska et al. 2013). Studies of the transcriptome, or global cellular pool of expressed genes, have provided insight into plant responses to many abiotic stresses and combinations thereof (Seki et al. 2001; Rossel, Wilson, and Pogson 2002; Kimura et al. 2003; Atkinson, Lilley, and Urwin 2013). Natural genetic variation in these responses exists, and allows for mechanistic study of such responses (Li et al. 2006).

1.1.1 Coping with Detrimental Effects of Excess Light: Photo-oxidative damage and Photoprotection

The quantity of light a plant receives is in excess when plants are unable to utilise energy obtained from absorbed photons to fix carbon (Li et al. 2009). Excess light is particularly damaging, causing reductions in photosynthetic ability, termed photoinhibition, cellular or tissue damage and death (Niyogi 1999; Asada 2006; Li et al. 2009; Suzuki et al. 2012). The indispensability of photosynthesis has caused the evolution of mechanisms by which the detrimental effects of excess light can be minimised. These mechanisms, collectively termed photoprotection, work to dissipate excess energy, reduce the amount of light absorbed, and prevent or repair any damage caused (Niyogi 1999; Takahashi and Badger 2011).

1.1.2 Physiological Responses to Excess Light

Upon exposure to excess light, plants immediately begin to mitigate its detrimental effects (Niyogi 1999; Demmig-Adams and Adams 1992). These include several mechanisms of response and damage mitigation within the chloroplast described in Figure 1.1 , as well as responses on a whole-cell scale. Chloroplast avoidance movement, the movement of chloroplasts parallel to the incident angle of light, decreases the amount of absorbed light (Kasahara et al. 2002). Transcriptional induction of heat shock proteins and proteins involved in antioxidant scavenging and photo-oxidative damage repair occurs following excess light (discussed in detail in subsection 1.1.3) (Niyogi 1999; Rossel, Wilson, and Pogson 2002; Jung et al. 2013). Production of anthocyanin, a protective class of pigments, is induced by the production of reactive oxygen species due to excess light (Vanderauwera et al. 2005). Together, these responses serve to reduce the photo-oxidative damage of cells and tissues due to excess light.

1.1.3 Transcriptional Responses to Light Stress

Transcriptomics, or the global study of gene expression (discussed further below), has been used to study the response of plants to excess light.The excess light induced expression of

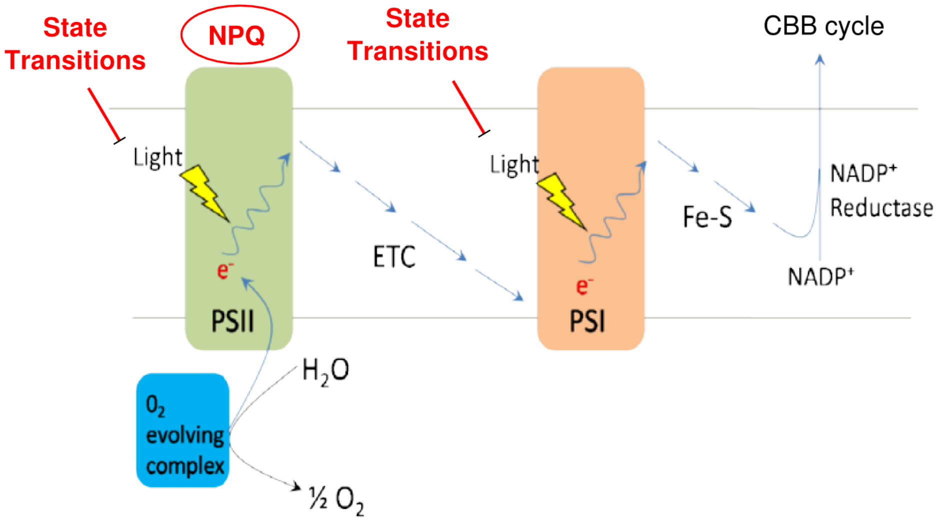

Figure 1.1: Chloroplastic mechanisms of excess light damage avoidance. Non-photochemical quenching (NPQ) dissipates excess energy from excited state chlorophyll molecules as heat (Müller, Li, and Niyogi 2001). It occurs in photosystem II (PS II), and is particularly important during rapid changes in light intensity (Külheim, Ågren, and Jansson 2002). Cyclic electron flow occurs in photosystem I (PS I) and acts by decreasing the pH of the thylakoid lumen, which in turn is thought to stabilise the oxygen evolving complex and aid NPQ (Takahashi et al. 2009). State transitions act via phosphorylation of PS II and light harvesting complex II (LHC II) proteins. State transitions reversibly alter the balance of excitation energy between PS I and PS II (Tikkanen et al. 2006). State transitions lower the light harvesting ability of PS II and prevent absorption of excess light by PS II, via reversible dissociation of the LHC II and PS II (Johnson et al. 2011). The *Arabidopsis* mutants *non-photochemical quenching 1* (*npq1*) and *non-photochemical quenching 4* (*npq4*), *proton gradient regulator 5* (*pgr5*), and *state transition 7* (*stn7*) and *state transition 8* (*stn8*) are defective in these three mechanisms respectively, and enable the study of the mechanisms underlying chloroplastic response to excess light and fluctuations in light intensity. Figure adapted from Nelson and Ben-Shem (2004)

a number of genes including *ASCORBATE PEROXIDASE 2* (*APX2*) (Rossel, Wilson, and Pogson 2002; Mühlenbock et al. 2008), *EARLY LIGHT INDUCED PROTEINS 1* and *2* (*ELIP1* and *ELIP2*) (Adamska 1997; Rossel, Wilson, and Pogson 2002), and various heat shock proteins (HSPs) (Rossel, Wilson, and Pogson 2002) is well established. These proteins are induced by discrete but overlapping mechanisms, allowing for differentiation between oxidative induction of retrograde chloroplastic stress signals (*APX2*) (Karpiński et al. 1999), ROS-mediated but photoreceptor-dependent induction of *ELIP1* (Kleine et al. 2007) and ROS-induced, heat shock transcription factor (HSF) mediated induction of HSPs (Miller and Mittler 2006). The light induced down-regulation of *LIGHT HARVESTING CHLOROPHYLL a/b BINDING* (*LHCB*) family genes is also well established (Rossel, Wilson, and Pogson 2002; Mishra et al. 2012). Therefore, these genes are used in this thesis as markers of transcriptional response to excess light. This enables comparison of transcriptional responses observed in this thesis to be compared to relevant previous works.

The rapid response of the *Arabidopsis* transcriptome to changes in light intensity has been demonstrated (Rossel, Wilson, and Pogson 2002). Acclimation to excess light conditions affects the steady-state level of expression of several classes of genes. In the cyanobacterium *Synechocysitis* sp. PCC 6803, genes involved in light capture are down-regulated and homologues of heat-shock proteins up-regulated after 15 hours of excess light treatment (Hihara et al. 2001). In rice, similar transcriptional down-regulation of light harvesting and up-regulation of photoprotection after 24 or 72 hours of excess light treatment has been observed (Murchie et al. 2005). Using quantitative real-time PCR, a method to assay relative expression of single genes, Gordon et al. (2013) have show that repeated excess light treatments lead to acclimation and reduced induction of excess light responsive transcripts. These transcriptomic responses to light overlap with response to drought, pathogen or oxidative stresses as well as hormone response, reiterating the requirement of stresses to be considered in combination.

In addition to light quantity, plants perceive alterations in light quality, or the spectral composition of light. Gordon et al. (2013) demonstrate that excess light induced expression was dependent on wavelength of excess light in *Arabidopsis*. Light quality is particularly important in studies of photo-oxidative damage, as the extent of photosystem II damage is not consistent across the visible spectrum (Takahashi et al. 2010), and photo-oxidative damage is relatively more severe under light of wavelengths between 580-620nm than in the remainder of the visible spectrum, which coincides with a density peak in the spectral power density of fluorescent lamps (see Figure 2.7). The integration of light quality into assays that test plant responses to altered light intensity is lacking in the vast majority of works on the topic. Plant growth methods developed in this thesis provide methods to examine such interactions.

1.1.4 Plant Growth Under Field Conditions

Plants grown in the field show markedly different phenotypes and responses to those grown in the laboratory. Under field conditions, plants respond to the combination of stresses they encounter in complex, natural environments. However, in the laboratory plants are exposed to pre-determined stresses (or combinations thereof) hypothesised to provoke a mechanism or response. This includes changes in metabolic profiles (Jänkänpää et al. 2012), transcript abundances (Mishra et al. 2012; Wituszyńska et al. 2013), photosynthetic differences (Mishra et al. 2012), and survival and fecundity (Külheim, Ågren, and Jansson 2002). Several authors report overlap between genes expressed upon treatment with excess light and those constitutively expressed in field conditions (Mishra et al. 2012; Wituszyńska et al. 2013).

1.2 Abiotic Stresses: Studied Individually, Encountered Combinatorially

Despite the large body of published works elucidating responses to abiotic stresses, the majority of works focus on an abiotic stress in isolation (Atkinson and Urwin 2012; Mittler 2006). While there is much merit to this reductionist approach, in natural or agricultural situations, plants rarely experience an abiotic stress in isolation. Unfortunately, this reductionist approach contributes to the difficulty in translating stress-tolerant lines of plants developed in labs to agricultural crops that are more field hardy (Mittler 2006;

Mittler and Blumwald 2010).

Recent laboratory studies have demonstrated that temperature interacts with light stress in a detrimental fashion. At cold temperatures, unacclimated plants are less able to dissipate excess energy to avoid photo-oxidative damage, leading to altered PS I/PS II redox poise (Armstrong, Wardlaw, and Atkin 2007). Many authors demonstrate temperature dependence on the transcriptional response to excess light (Rossel, Wilson, and Pogson 2002; Jung et al. 2013). Interactions with temperature have been observed in the transcriptional response to drought (Seki et al. 2001; Seki et al. 2003; Rizhsky, Liang, and Mittler 2002; Rizhsky et al. 2004). Additionally, interactions between biotic and abiotic stresses exist, but are beyond the scope of this thesis (Mittler 2006; Atkinson and Urwin 2012; Atkinson, Lilley, and Urwin 2013). The data these authors present warrants a paradigm shift towards the study of abiotic stresses in combination, as advocated by Mittler (2006).

Field-grown plants tend to demonstrate reduced survival, reproduction and altered transcriptional responses compared to those grown in supposedly comparable laboratory conditions. For example, in *Arabidopsis thaliana* the reduction in reproductive success caused by the *npq1* and *npq4* mutations was severe in the field but had no impact under static sufficient light in the laboratory (Külheim, Ågren, and Jansson 2002). Similarly, photoinhibition, physiological symptoms of abiotic stress and expression of excess light-induced transcripts were more severe in field-grown *Solidago altissima* than those grown in comparable laboratory conditions (Barua and Heckathorn 2006). Furthermore, Wituszyńska et al. (2013) observe increased steady-state expression of genes involved in response to excess light and photo-oxidative damage under field conditions compared to laboratory-grown plants, and concomitant physiological acclimation to variable light intensity. Similar findings are presented by Mishra et al. (2012), who found increased NPQ and expression of ELIP proteins and decreased LHC-associated protein abundances in field-grown *Arabidopsis*. These field studies reiterate interactions noted in laboratory studies, and provide impetus for the study of field-like combinations of abiotic stresses under controlled laboratory conditions. Natural variation in these interactive stress responses may facilitate the mapping of quantitative trait loci (QTLs) for combinatorial stress tolerance (Li et al. 2006; Li et al. 2010).

1.3 Transcriptomics: Assessing Global Expression

By studying how, when and to what extent each gene in the genome is expressed, we can gain insight into the response to any perturbation to a plant's environment. RNA sequencing (RNAseq) is a modern method of whole-transcriptome quantification. By creating cDNA from the cellular pool of mRNA, and sequencing this cDNA using high-throughput sequencing (HTS), quantification is possible (Wang, Gerstein, and Snyder 2009; Lister, Gregory, and Ecker 2009). RNAseq analysis takes raw sequence data from high-throughput sequencing of RNAseq libraries and generates data that can be interpreted in the context of the biological basis underlying the experiment. The process of RNAseq is presented graphically in Figure 1.2

1.4 Thesis Aims

The overall aim of this project is to examine the effect of light intensity on *Arabidopsis*, within the framework of realistic combinations of abiotic stresses. The remainder of this thesis is devoted to the examination of the following aims:

1. Design and implement "Dynamic Growth Conditions" which mimic regional climates
2. Select optimal software for the high-throughput study of transcriptome dynamics using High-throughput Sequencing, and implement a framework for generation of analysis pipelines to do so
3. Determine the transcriptional response of *Arabidopsis thaliana* to combinatorial application of abiotic stresses, using dynamic growth conditions

Due to technical obstacles and delays, the aims of this project have evolved over the year. As it became apparent that technical difficulties would prevent high-throughput RNAseq experiments required to map expression QTLs in the time available, the aims "Examine the extent of genetic variation in gene expression under these dynamic light conditions, and elucidate gene regulation networks controlling gene expression." and "

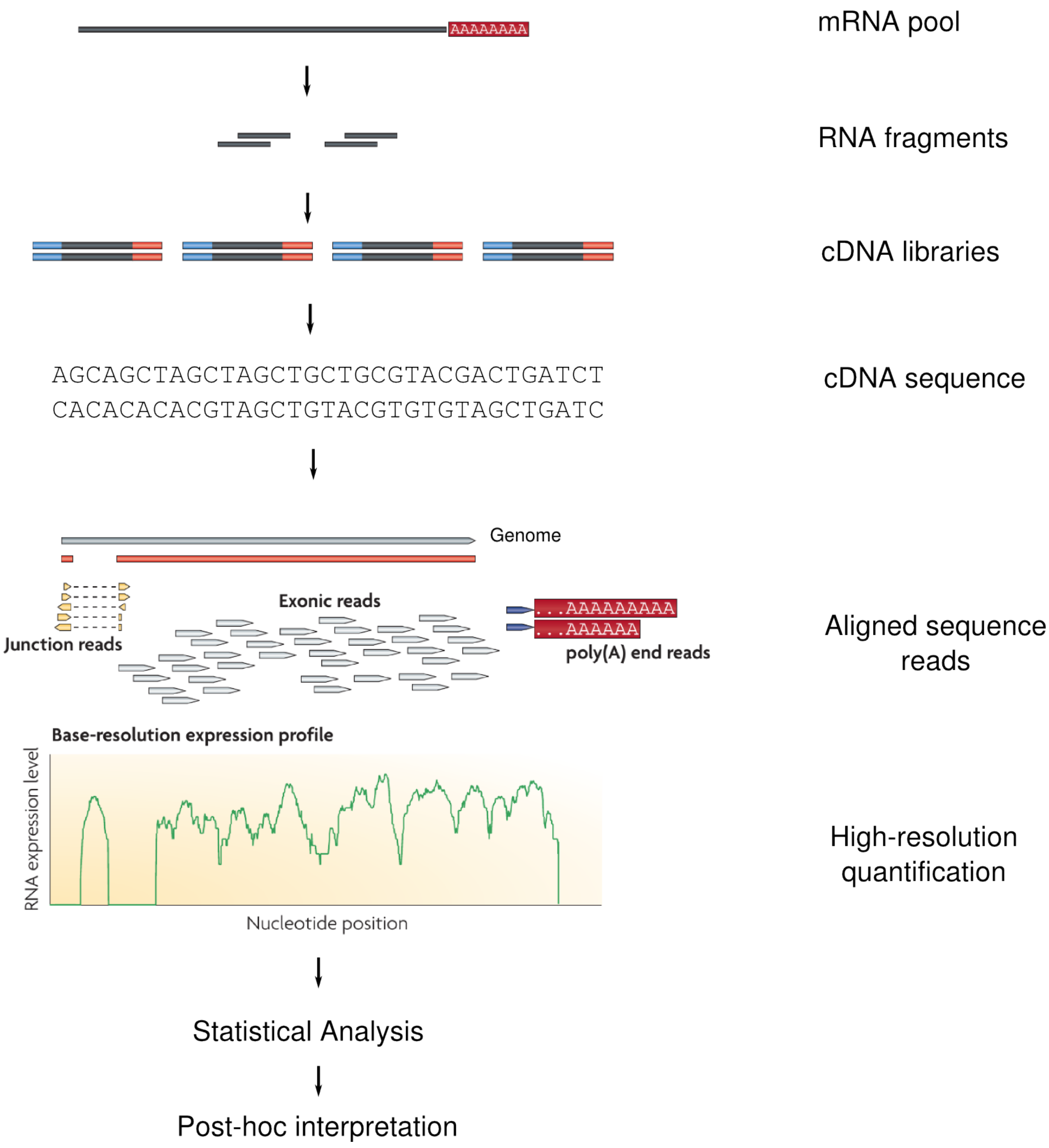

Figure 1.2: The molecular biology of RNAseq library preparation. The process of RNAseq library creation involves the extraction and purification of intact mRNAs, their conversion to complimentary DNA (cDNA), fragmentation, end-repair and A-tailing to allow sequencing adaptor ligation, ligation of sequencing adaptors and library amplification (Wang, Gerstein, and Snyder 2009; Kumar et al. 2012). When sequenced using HTS, raw sequence data is obtained with sequences derived from mRNA molecules proportional to their abundance in the sample mRNA pool. Computational analysis is required to provide quantification from this raw sequence data (Nookaew et al. 2012; Van Verk et al. 2013). This process involves raw sequence quality control, alignment of raw sequence reads to the genome, and quantification of gene-wise expression by aggregating the number of sequence that align to each gene. Quantitative data undergoes statistically rigorous normalisation before hypothesis testing occurs (Robinson and Oshlack 2010; Robinson, McCarthy, and Smyth 2010). Post-hoc analyses of expression data can be performed to glean summarised biological meaning from genome-wide differential expression patterns, including gene ontology (GO) term enrichment analysis (Berardini et al. 2004; Avraham et al. 2008) and gene set enrichment analysis (Subramanian et al. 2005; Kim 2012; Väremo, Nielsen, and Nookaew 2013; Yi, Du, and Su 2013).

Explore the effect of genotype-environment interactions on gene expression under dynamic light conditions.” were discontinued.

Chapter 2

Design and Implementation of Dynamic Growth Conditions

2.1 Background, Aims and Hypotheses

As biologists study organisms or mechanisms that have evolved in a given environment, it follows logically that the study environment should be as similar as possible to the environment in which the subject has evolved. In areas of molecular plant science however, our study organisms are often placed in environments highly dissimilar to those in which our subjects are hypothesised to have evolved (Mittler 2006; Mittler and Blumwald 2010). In this chapter I describe artificial growth conditions that I have created which vary on diurnal and circannual cycles in an analogous manner to the regional climates cultivated or naturally growing plants experience. This class of laboratory growth condition are termed dynamic growth conditions is in contrast with the highly static growth conditions typically used in the experimentation of plants in laboratory settings. I hypothesise that plants grown under dynamic growth conditions will exhibit phenotypes more similar to those grown outdoors under natural environments, as these dynamic growth conditions are more similar to natural environments when compared to static, benign static growth conditions. This hypothesis is tested in later chapters of this thesis. Additionally, I aim to create software to allow dynamic growth conditions to be implemented with existing hardware at the ANU. Successful completion of this aim has allowed the implementation of dynamic growth conditions, and their use in research into plant-environment interactions in *Arabidopsis*.

2.2 Materials and Methods

2.2.1 The SpectralPhenoClimatron and Implementation of `spcControl`

The SpectralPhenoClimatron is a new facility within the Research School of Biology, consisting of computer controllable plant growth cabinets featuring multi-spectral LED lamps, and real-time imaging hardware. Conviron PGC20 reach-in growth chambers (Conviron, Winnipeg, Canada) have been retro-fitted with four Heliospectra Model L4A Series 10 multi-spectral LED growth lamps (Heliospectra AB, Sweden) per chamber, and an image-based phenomics systems (Canon EOS DSLR cameras and other consumer hardware). The Conviron PGC20 cabinets have a capacity of 320 5cm by 5cm plant growth containers, or 16 250x200mm standard nursery seed trays (e.g. Garden City Plastics part TRSR00). The Heliospectra L4A Series 10 LED lamps contain 7 LED wavelength channels: 400nm (sub-blue), 420nm (blue), 450nm (blue), 530nm (green), 630nm (red), 660nm (red) and 735nm (far red).

Both the Heliospectra LED lamps and Conviron growth cabinets can be controlled via the Telnet protocol, and custom control software was created to utilise this feature. The `spcControl` program is invoked with a regional climate model in comma-separated value (CSV) format, describing the temperature, humidity, and intensity of each LED wavelength for climate models per SolarCalc calculations (Spokas and Forcella 2006). This software simultaneously sends telnet commands that control growth chamber temperature and humidity to each growth cabinet, and commands that control light quantity and quality to each of four LED arrays per cabinet. The success of each set of control commands is reported to an off-site database.

2.2.2 Design of dynamic growth conditions

SolarCalc was used to create climate models underlying dynamic growth conditions (Spokas and Forcella 2006). Model settings that I used in the creation of the dynamic conditions are described in Table 2.1 where they deviate from program defaults. As SolarCalc by default simulates the climate of a location without variable weather, post-processing work was required to create conditions that mimic cloudy and intermittently cloudy days. A SolarCalc model with a neutral density shade such that model sunlight intensity was 45% that of an unshaded model was created, and the result of both the shaded and unshaded models were spliced together to form a third condition whose light intensity fluctuated on a two hour sufficient light, one-hour excess light rotation, using the `spliceSolarCalc.py` script described in subsection 6.2.1. The temperature, humidity and light quality was preserved across these conditions.

2.2.3 Measurement of Spectral Power Density

Spectral power density, or distribution of light intensity across the visible spectrum, was quantified using a spectrophotometer to record spectral power density across the spectrum between light of wavelengths 400nm to 800nm, with 2nm wavelength resolution. Sun and shade spectra were obtained on July 18, 2013 at the Acton campus of the ANU, in a clear, open space and under heavy shade from mature trees of various species, in the courtyard between buildings 46 and 48. Spectra of laboratory growth conditions were obtained from a Conviron PGC20 by placing the spectrophotometer on the lowest shelf level while fluorescent lamps or Heliospectra L4A series 10 lamps were illuminated at their highest intensities. In the case of the Heliospectra L4A series 10 LED lamps, measurements from directly under a single unit were recorded. Intensity-normalised spectral power density was calculated by normalising the intensity recorded for each wavelength by total light source intensity.

2.3 Results

2.3.1 Computer Control of the SpectralPhenoClimatron

To create dynamic growth conditions which change on diurnal and circannual cycles, high temporal resolution is required. Whilst SpectralPhenoClimatron hardware can produce static growth conditions, external software is required to enable the creation of such dynamic conditions. Thus, I have created software, `spcControl`, to do so. This software

Parameter	Temora Setting
Simulation End Date	31/12/12
Simulation Start Date	01/09/12
Chamber Max RH	85
Chamber Min Temp	5C
Chamber Date	01/03/13
Site Elevation	300m
Site Latitude	-34.446556
Site Longitude	147.53334
Timezone	Sydney
Update Frequency	1 min
Weather Year	2010
Lighting	LED
Set to Threshold	Yes
LED1 Power Threshold	5.00%
LED1 Wavelength	400nm
LED1 Weight Multiplier	5.2
LED2 Power Threshold	5.00%
LED2 Wavelength	420nm
LED2 WeightMultiplier	4.12
LED3 Power Threshold	5.00%
LED3 Wavelength	450nm
LED3 Weight Multiplier	4
LED4 Power Threshold	5.00%
LED4 Wavelength	530nm
LED4 Weight Multiplier	3.92
LED5 Power Threshold	5.00%
LED5 Wavelength	630nm
LED5 Weight Multiplier	4.88
LED6 Power Threshold	5.00%
LED6 Wavelength	660nm
LED6 Weight Multiplier	0.68
LED7 Power Threshold	5.00%
LED7 Wavelength	740nm
LED7 Weight Multiplier	3.64

Table 2.1: Parameter settings of SolarCalc used in the creation of dynamic conditions. LEDs 1-7 correspond to wavelengths 400nm (sub-blue), 420nm (blue), 450nm (blue), 530nm (green), 630nm (red), 660nm (red) and 735nm (far red).

can implement dynamic growth conditions given a climate model generated by SolarCalc (Spokas and Forcella 2006). `spcControl` will, at time intervals specified in the climate model, send control commands to both the LED arrays and plant growth cabinet, updating LED intensity, temperature and humidity. This process takes around 30-45 seconds, and thus can occur up to every minute, giving extreme temporal resolution in growth condition control. Additionally, as the commands are sent out synchronously, lighting, temperature and humidity will never go "out of sync" if a power outage or device failure occurs. Furthermore, to ensure reliable operation and detection of faults, at every timepoint specified in the SolarCalc model, the success or failure is communicated to an off site database, and any error message is emailed to an administrator. An additional program, `spcControl.monitor`, polls this database and guards against failure of hardware, control computers and software, informing an administrator upon any failure.

The `spcControl` python module runs with python version 3.2 or later. It is modular in design, with a main program loop which sends a control line to each sub-module per the schedule given by the SolarCalc model. Sub-modules then parse this line, and a configuration file, to formulate commands sent to the relevant device(s); sub-modules for Heliospectra L4S LED lamps and Conviron PCG20 chambers have been implemented. This modular design means that, given hardware specifications, creating new sub modules to control other hardware configurations would be relatively trivial. Status reports are sent to an external PostgreSQL database, and email error messages are generated from within python if an error occurs. In all, this consists of over 740 lines of code and configuration. This software is actively maintained; as software bugs, hardware limitations and feature requests are discovered solutions are provided. The codebase has expand from a simple script to a fully-fledged python module with 16 versions released thus far (Appendix subsection 6.3.1).

2.3.2 Sufficient, Excess and Fluctuating Dynamic Growth Conditions

To investigate the effect of altered light intensity on the *Arabidopsis* transcriptome, three novel growth conditions were specifically designed to mimic the dynamic nature of temperature, humidity and light intensity and quality that occurs outdoors. The sufficient light dynamic growth condition corresponds to approximately the same daily integral of light as “standard static growth conditions” of 120-150 μ mol photons $m^{-2}\,s^{-1}$with a 12-hour photoperiod. The excess light condition is approximately 250% brighter than the sufficient condition. The fluctuating light condition varies between sufficient and excess growth conditions on a 2 hour-1 hour basis, and is designed to simulate the pattern of light intensity variation caused by partial cloud. These conditions simulate the spring season, and display circannual or seasonal variation in temperature, light and humidity. As spring progresses, daily minimal and maximal temperature and peak light intensity increase, while minimal relative humidity decreases.

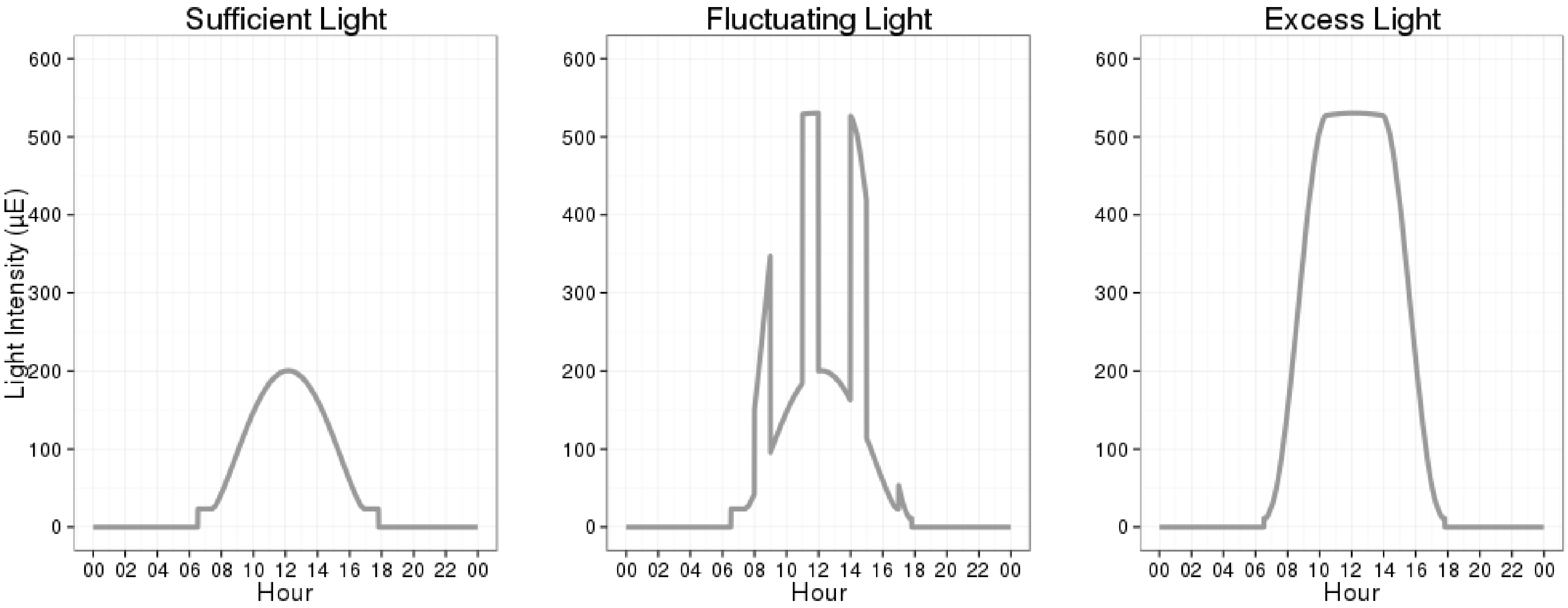

Figure 2.1: Diurnal variation in approximate light intensity of sufficient, fluctuating and excess light dynamic growth conditions (for model date 1 March). Note the altered light intensity between dynamic growth conditions, and identical photoperiod between conditions.

The light quality of all dynamic growth conditions are markedly different to other light sources. Compared to the fluorescent lamps typically used in laboratory growth chambers the spectral power density, or distribution of light intensity across the visible spectrum, of LED arrays is less variable across the visible and adjacent spectrum, with intensity-normalised spectral power density of fluorescent and LED array light sources of 1.00 ± 1.74 and 1.00 ± 0.92 μmol photons $m^{-2}s^{-1}nm^{-1}$ per total μmol photons $m^{-2}s^{-1}$respectively (means ± SD; see Figure 2.7). The intensity-normalised spectral power density of sunlight on a clear day is remarkably even (1.00 ± 0.19 μmol photons $m^{-2}s^{-1}nm^{-1}$ per total μmol photons $m^{-2}s^{-1}$; Figure 2.7). The intensity normalised spectral power density (spectral power density per unit total light intensity) of canopy shaded light is similar to sunlight at wavelengths lower than approx 700nm, above which sunlight is not filtered by vegetation and thus is over-represented. Overall, the spectral power density of LED lamps is more even than that of fluorescent lamps, however it still deviates notably from that of sunlight.

The overall light intensity of these natural and laboratory light sources varies drastically. Representative measurements of intensity reveal open sunlight to have an intensity of 2480 μ mol photons $m^{-2}s^{-1}$in the photosynthetically active spectrum on the day of measurement (July 18, 2013 in Canberra, ACT, Australia). The intensity of tree canopy shaded sunlight measured close by on the same day is much lower, at approximately 88 μ mol photons $m^{-2}s^{-1}$. This is in contrast with the light intensity under a single Heliospectra L4A series 10 LED lamp of 370 μ mol photons $m^{-2}s^{-1}$, and the intensity of light from fluorescent lamps was 140 μ mol photons $m^{-2}s^{-1}$(data shown graphically in Figure 2.7).

2.4 Summary and Technical Discussion

In this chapter, I present software enabling the implementation of dynamic growth conditions that mimic diurnal and circannual trends in temperature, humidity, photoperiod, and light quality and quantity observed in regional climates. This software enables the use of regional climate models to govern growth conditions in laboratory growth chambers, allowing reproducible and reliable implementation of dynamic growth conditions.

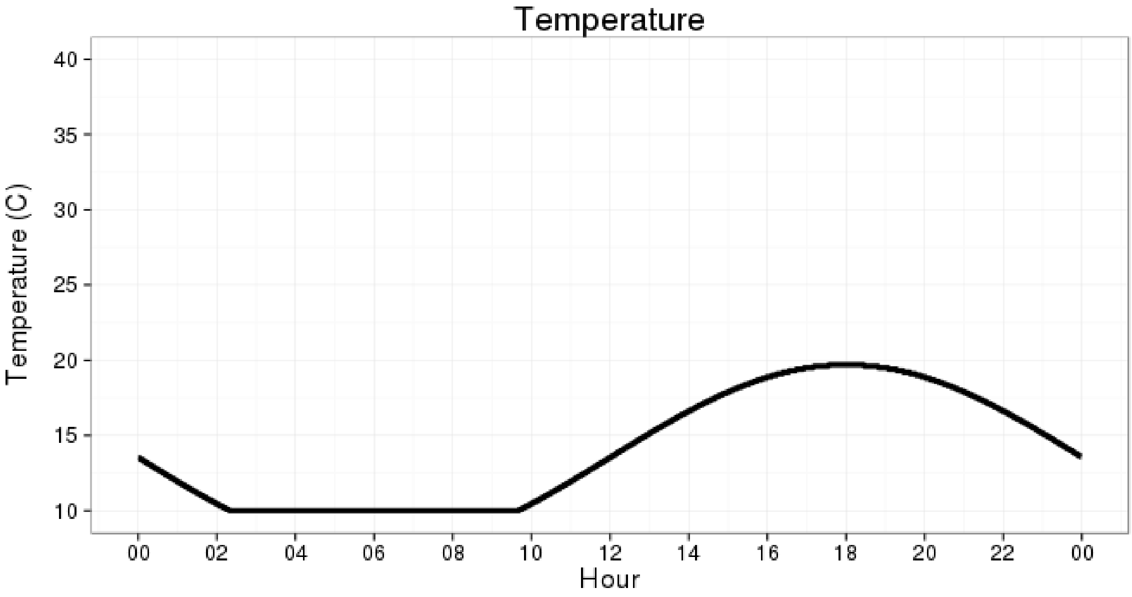

Figure 2.2: Diurnal variation in dynamic growth condition model temperature (for model date 1 March). Temperatures follow an approximation of those observed in temperate climates, reaching a minimum before sunrise (06:00), steadily increasing after sunrise to a peak immediately prior to sunset (18:00). Growth chamber hardware limitations prevent temperatures falling below 10 °C for extended periods, thus the model "bottoms out" where temperatures below 10 °C would have occurred (02:00 - 10:00).

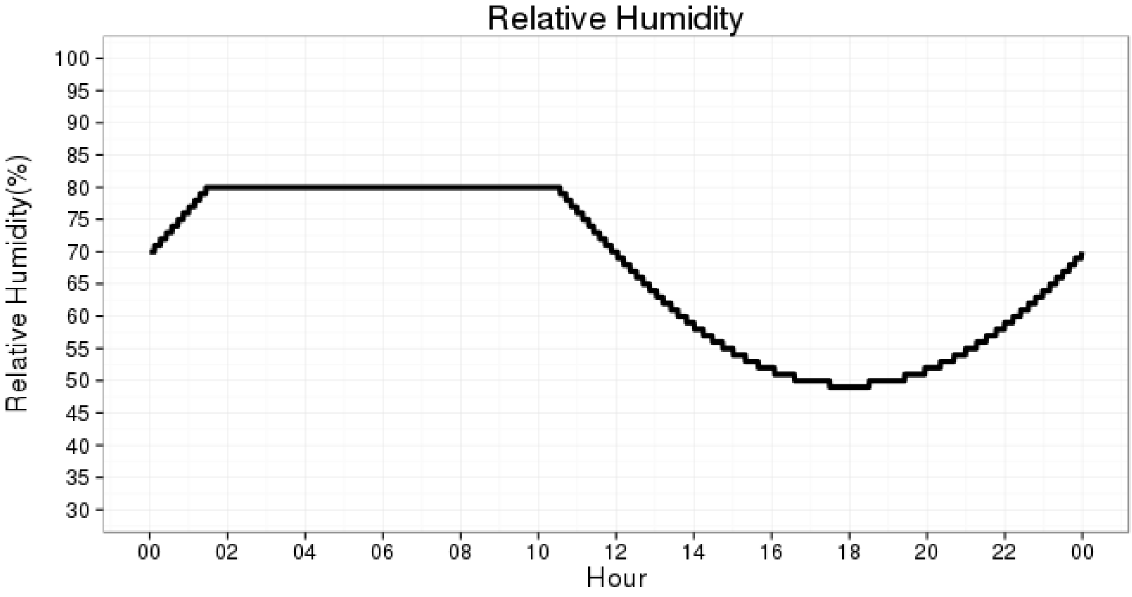

Figure 2.3: Diurnal variation in Dynamic Growth Condition model relative humidity (for model date 1 March). Humidity follows an inverse trend to temperature, peaking before sunrise (06:00) and reaching its minimum at approximately sunset (18:00). Similarly to temperature, growth chamber hardware limitations prevent relative humidities greater than 80% for long periods, and therefore humidity is capped at 80% between the hours of 02:00 and 10:00.

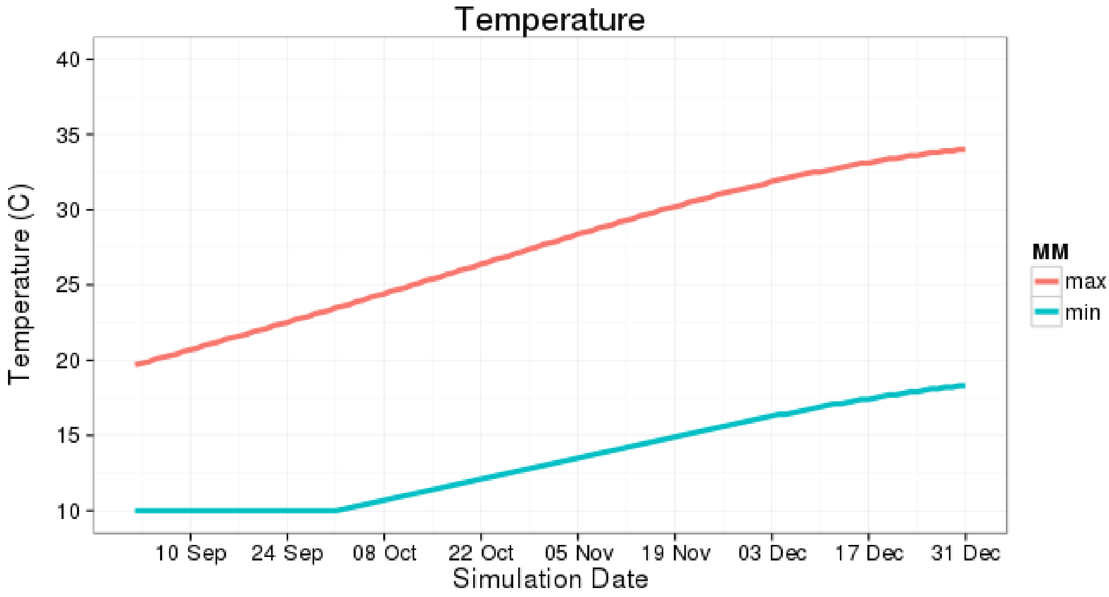

Figure 2.4: Circannual variation in daily minimal (blue) and maximal (red) Dynamic Growth Condition model temperature. Starting at model date September 1, both minimal and maximal daily temperature gradually increase, throughout spring and the first month of summer. Note also the hardware limitation of minimal daily temperature is alleviated after daily minimal temperature exceeds 10 °C (on approximately October 1st).

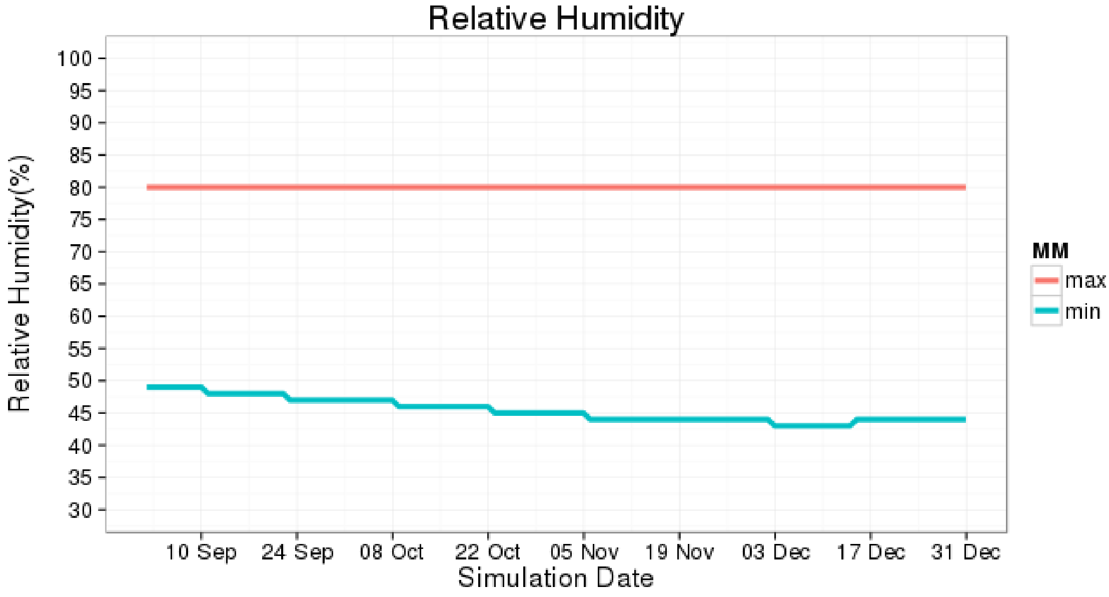

Figure 2.5: Circannual variation in daily minimal (blue) and maximal (red) Dynamic Growth Condition model relative humidity. Unlike the pattern observed in temperature, daily maxima do not increase across the modelled period, due to hardware limitations. However, daily minima in humidity does decrease over the modelled period, concomitant with observed daily maxima, partially preserving the inverse relationship between temperature and humidity.

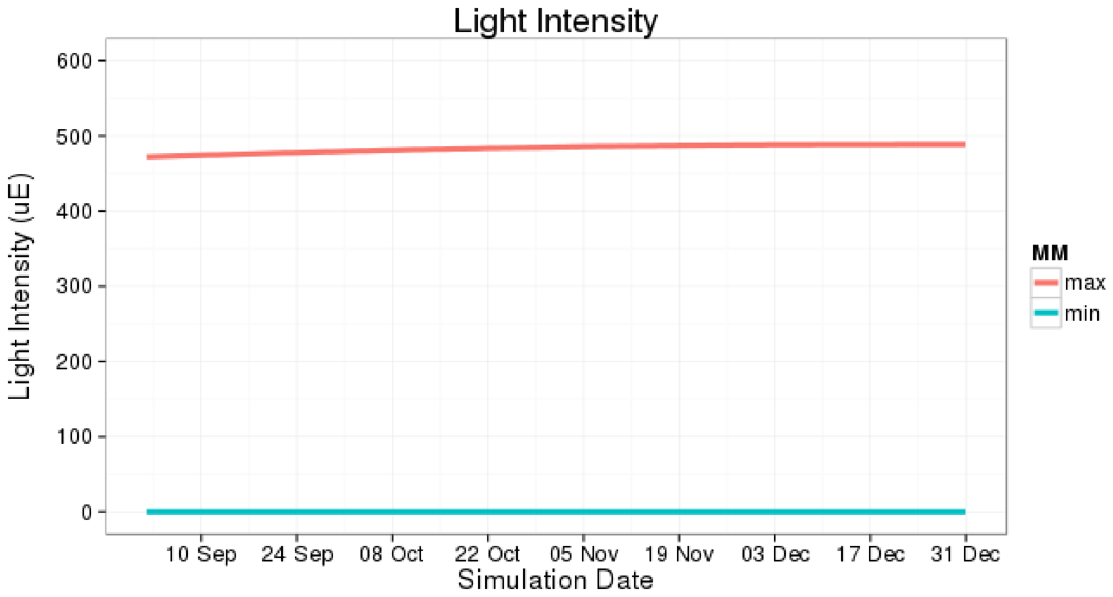

Figure 2.6: Circannual variation in daily minimal (blue) and maximal (red) Dynamic Growth Condition model light. Little change in daily maximal light intensity occurs, due to hardware limitations on the brightness of LED arrays.

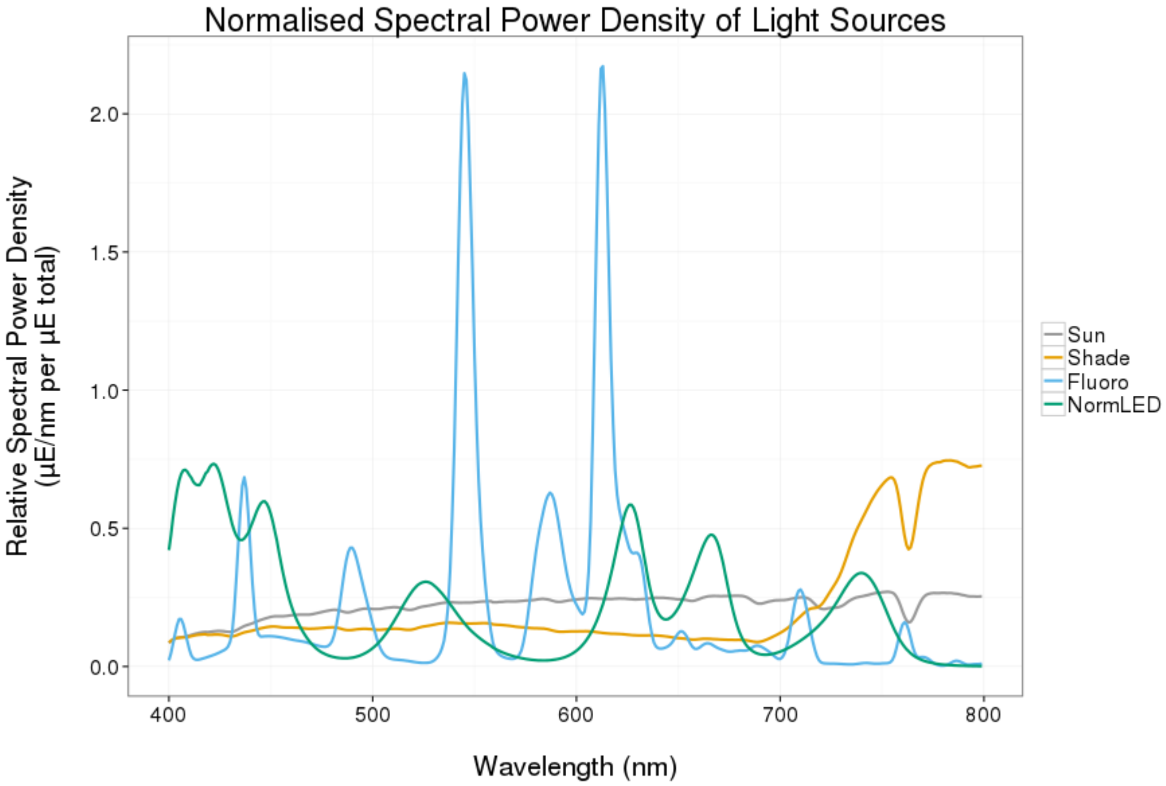

Figure 2.7: Intensity-normalised spectral power density of sunlight, shaded sunlight, fluorescent lamps and Heliospectra L4A series 10 LED lamps. Note the almost flat spectral density of sunlight, compare to the broad peaks of intensity of LED arrays, and large spikes of intensity (Mercury emission peaks) in fluorescent lamp spectra.

Practically, there are shortcomings in the SpectralPhenoClimatron. These are largely limitations inherent to the hardware from which it is constructed, and include the limited light intensity, temperature and humidity of the SpectralPhenoClimatron. The limited light intensity and the evenness of the spectral power density of Heliospectra L4A Series 10 lamps will be improved in an upcoming upgrade (pers. comm., Justin Borevitz). Despite these minor shortcomings, the SpectralPhenoClimatron is a phenomenal tool with which to study plant-environment interactions. It has been designed with large scale studies that elucidate underlying genetic mechanisms and examine genetic variation in these interactions in mind, and has been applied by colleagues to aims beyond the scope of this thesis (see subsection 6.4.2).

Three dynamic growth conditions to examine the effect of light intensity in field-like interaction with temperature, humidity and light quality were implemented using the SpectralPhenoClimatron. These conditions allow for light stress to be studied in the framework of combinatorial application of stresses, as recent literature has advocated (Mittler 2006; Wituszyńska et al. 2013). Examination of genetic variation in transcriptional, physiological or phenomic responses to altered light intensity may provide insight into mechanisms underlying response to light stress in field-like combinations with other abiotic stresses (Li et al. 2006; Li et al. 2010).

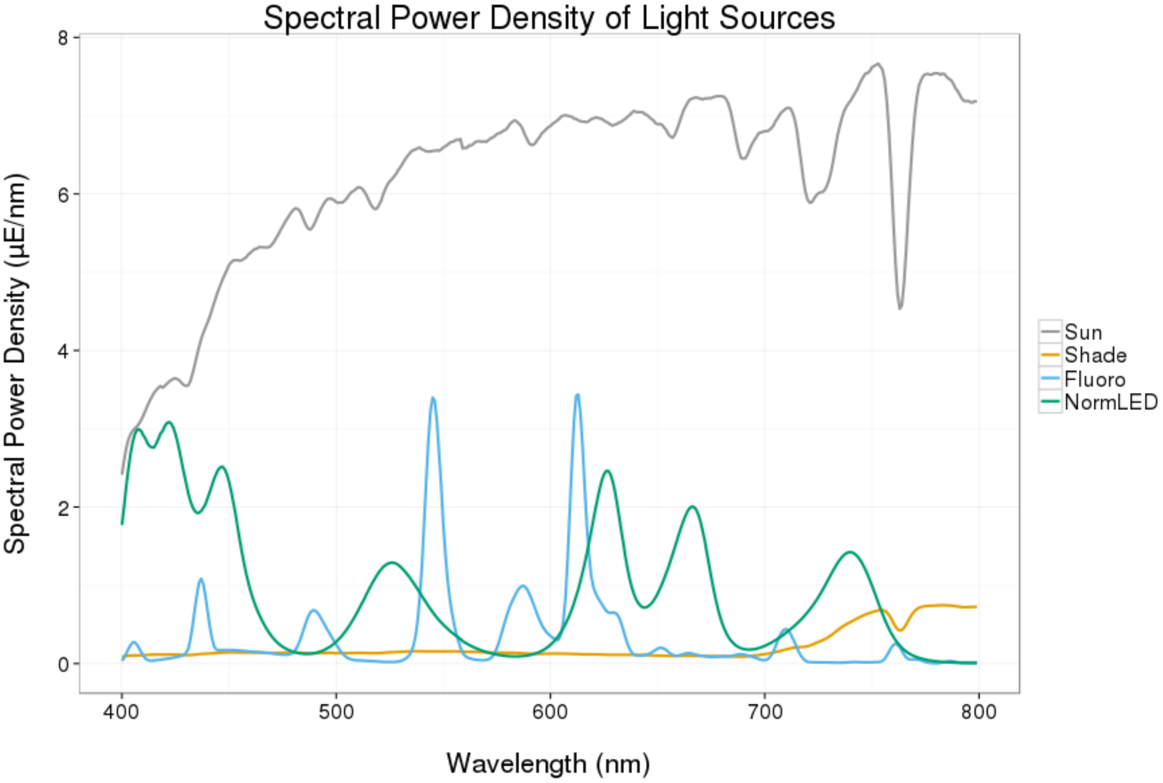

Figure 2.8: Spectral power density of sunlight, shaded sunlight, fluorescent lamps and Heliospectra L4A series 10 LED lamps. Note the higher intensity and less even spectral density of sunlight compared with other light sources.

Chapter 3

Improved Methodology for High-throughput RNAseq Experiments

3.1 Background, Aims and Hypotheses

RNA sequencing (RNAseq) is a modern method for genome-wide expression (transcriptome) quantification. RNAseq works by sequencing the mRNA pool of a cell, tissue or organisms (Wang, Gerstein, and Snyder 2009; Martin et al. 2013). As high-throughput, large-scale experiments such as QTL mapping and Genome Wide Association Studies (GWAS) become more prevalent, increasingly subtle biological contrasts are being examined. RNAseq is an incredibly sensitive tool to assay subtle changes in, for example, plant growth environment or subtle genetic variation (Martin et al. 2013). Typically, RNAseq analytic methods have been developed to investigate stark biological comparisons, such as comparison of healthy and diseased or mutant and wild-type tissues or individuals (Wang, Gerstein, and Snyder 2009). In order to analyse these large scale datasets, updated molecular and analysis methodology must be used.

To gain biological meaning from raw RNAseq data, an analysis pipeline must be employed. In this context, a pipeline is a series of software components that, in succession, manipulate a dataset to obtain a biologically relevant result. Best practice pipelines for RNAseq analysis exist (e.g. Van Verk et al. 2013), but often must be manipulated to suit the idiosyncrasies of each experiment. Thus, an aim in this chapter of my thesis is to create a framework allowing easy creation of pipelines by non-expert bioinformaticians with limited programming experience, and to use this framework to create high-performance pipelines suited to analysis of high-throughput RNAseq experiments.

Additionally, I have conducted an *in silico* experiment to test the effect of sequencing depth on statistical power of RNAseq experiments. Despite the rapid and continuing reduction in the cost of high throughput sequencing, it is still a very large component of the overall cost of RNAseq experiments (Wang, Gerstein, and Snyder 2009). This is particularly evident with regards to high throughput experiments, and is often combated by increasing multiplexing, i.e. reducing the amount of raw sequence data each sample yields, at a cost to statistical power (Kumar et al. 2012). Therefore, I am to determine the optimal trade-off between sequencing cost and statistical power. Similar experiments suggest an optimal depth of 10 million reads per sample for Chicken tissue samples (Wang et al. 2011). However due to its smaller transcriptome size, I hypothesise that the optimal sequencing depth for *Arabidopsis* will be smaller, specifically between 2 and 5 million reads, or between 48 and 96 libraries per Illumina HiSeq 2500 sequencing lane (that yield 200 million reads apiece, (Glenn 2011)).

3.2 Methods

3.2.1 External RNAseq Datasets

RNAseq datasets created by Peter Crisp and Barry Pogson were used both as trial datasets and external references in this thesis. The Rapid Recovery Gene Silencing excess light time-course experiment (hereafter referred to as the RRGS time-course) consists of samples taken in triplicate from an eleven-point excess light stress and recovery time-course. *A. thaliana* reference accession Col-0 were grown for 3 weeks under standard laboratory growth conditions (≈ 150 μ mol photons $m^{-2} s^{-1}$light intensity, 12 hour photoperiod, 21 °C daytime temperature, 21 °C night-time temperature). Whole rosette samples were taken before any treatment, after 30, 60 and 120 minutes of 8x excess light (1000 μ mol photons $m^{-2} s^{-1}$, unfiltered light from a sodium vapour lamp, hereafter EL), after 60 minutes of EL followed by 7.5, 15, 30 and 60 minutes of recovery under standard growth conditions, after 60 minutes of EL, followed by 60minutes of recovery, followed by another 60 minutes of EL, and before and after 60 minutes of EL 24 hours after the original 60 minutes of EL. This complex time-course is illustrated in Figure 3.1. RNA extracted from five plants per replicate was pooled, before Illumina libraries were created using the TruSeq V2 library preparation kit (part number 15026495) per manufacturer's instructions. These libraries were sequenced across two Illumina HiSeq 2500 sequencing lanes, yielding the RRGS timecourse RNAseq dataset. This dataset studies a timecourse over a treatment highly similar to that conducted in the dynamic growth condition experiment, allowing development and validation of bioinformatic protocols for experiments of a similar nature.

3.2.2 Development of an Improved Analysis Pipeline

Bioinformatic experiments were used to validate pipelines against the "gold standard" RNAseq analysis pipeline. In these experiments, programs selected through both literature review and searches of pre-publication software releases (e.g. software on `github.com`) were tested against a published best-practice pipeline (Van Verk et al. 2013). Specifically, the computational speed and efficiency, and the results obtained with these newer programs were compared to the analysis pipeline of Van Verk et al. (2013). This enables the development of higher-performance analysis pipelines suitable to high throughput experiments, with no cost to the quality of results obtained. A summary of program selections is described in Table 3.1.

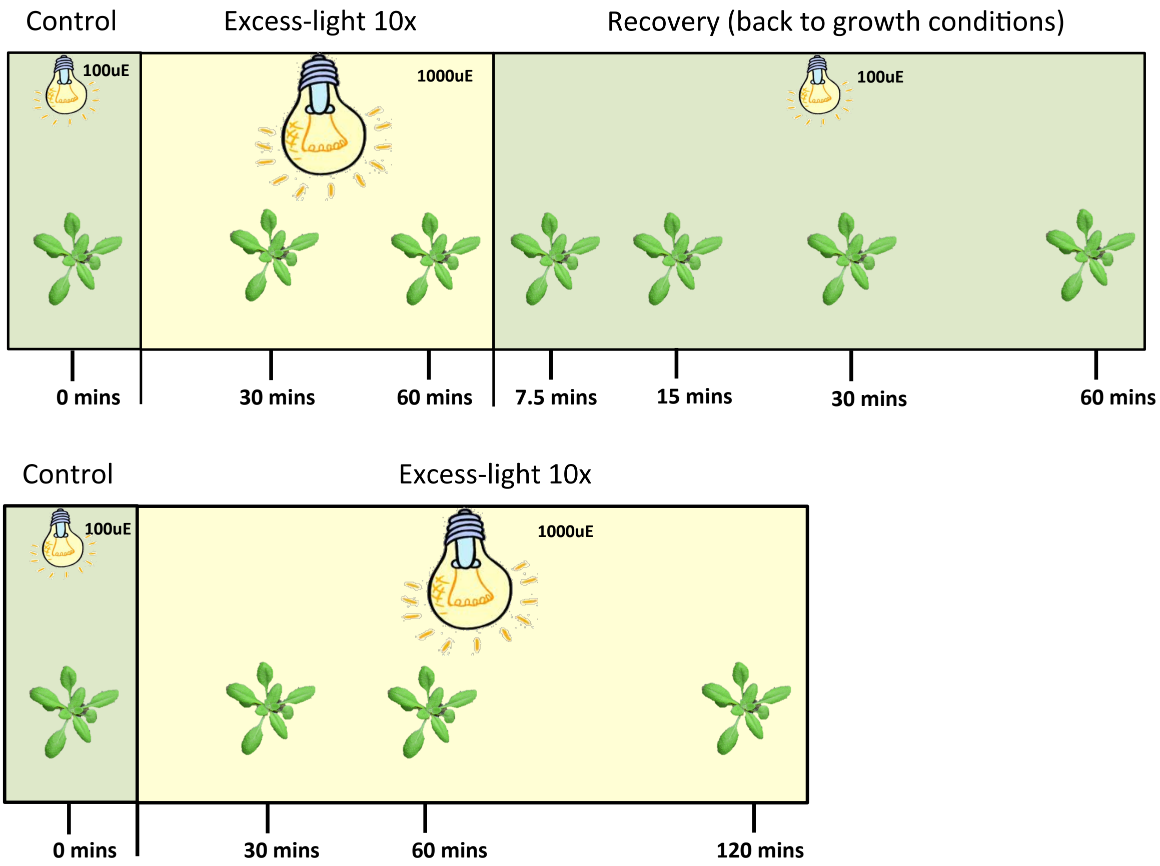

Figure 3.1: Illustration of the RRGS timecourse. Entire rosettes were harvested in triplicate at each indicated time-point along the excess light stress and recovery time-course. Two timepoints after a 24 hour recovery period are not show. This figure was created by Peter Crisp, and is reproduced with his permission

Program	Program's Role	Reasons for Selection	References
`fastqc`	Determine raw sequence quality of datasets	Easy of use and detailed reporting	(Andrews 2012)
`scythe`	Remove Illumina adaptor sequence from 3' ends of reads, allowing more accurate mapping	Author's claims of increased accuracy and speed	(Buffalo 2013)
`seqtk`	Remove sequences with low base-level quality from analysis	Mott trimming algorithm; fast	(Li 2013)
`subread`	Align short reads to genome	Fast RNAseq compatible	(Liao, Smyth, and Shi 2013b)
`tophat2`	Align short reads to genome while detecting mRNA splicing *de novo*	Capable of de-novo splicing detection	(Kim et al. 2013)
`subjunc`	Align short reads to genome while detecting mRNA splicing *de novo*		(Liao, Smyth, and Shi 2013b)
`featureCounts`	Aggregate gene-wise counts of aligned reads to quantify expression	Fast and well supported	(Liao, Smyth, and Shi 2013a)
`edgeR`	Perform statistical normalisation and hypothesis testing	Statistically rigorous; supports multi-factor experiments	(Robinson, McCarthy, and Smyth 2010; McCarthy, Chen, and Smyth 2012)
`goseq`	Perform gene ontology term enrichment	Improved RNAseq-compatible statistical basis	(Young et al. 2010)

Table 3.1: Selection of pipeline components and their roles in an RNAseq analysis pipeline. These programs were selected by review of relevant literature (References). Brief reasons for selection are given, along with references that describe the implementation and validation of these software.

Comparisons between the computational cost of four pipelines were conducted using a sub-sampled dataset. To demonstrate the improved performance of the `aln_subread` pipeline, it was compared to the `aln_tophat`, `aln_tophat_htseq` and `aln_subread_htseq` pipelines. The `time` UNIX command was used to summarise the computational cost of these four pipelines across five identical, independent, non-simultaneous runs. Four time-points of the RRGS timecourse dataset were sub-sampled to 500,000 reads by running `seqtk sample -s 10 500000` on both forward and reverse read files, which extracts 500000 random read pairs preserving read pairing. An ANOVA analysis was performed to find significant differences in runtime and CPU utilisation between analysis pipelines.

To ensure that the `subread` aligner and `featureCounts` produced comparable results to the analysis pipeline of Van Verk et al. (2013), several diagnostic measures were used. Firstly, the percentage of reads mapped to the genome, and to protein coding loci within the genome was computed and compared. Then, sample-wise correlations between gene-wise counts calculated by each pipeline were calculated. Finally, genes called differentially expressed by each pipeline were compared. These measures allow verification of pipeline performance at three major stages in an analysis pipeline: alignment of short reads to a genome, gene-wise count summarisation, and statiscial testing for differential expression.

3.2.3 Measuring the Effect of Sequencing Depth on Analysis of Differential Expression

Six samples from the RRGS-Timecourse experiment (see subsection 3.2.1) were sub-sampled to allow investigation of the effect of sequencing depth on statistical power. To do so, the command `seqtk sample -s 10` X was run on each pair of read files for these six samples, with X (number of reads to sample) set to 1000, 10000, 20000, 50000, 100000, 200000, 500000, 1000000, 2000000, 5000000 and 10000000. This sub-sampled dataset allows for titration of the optimal sequencing depth (or multiplexing level) for high throughput experiments, balancing sequencing cost with statistical power.

For each subsampled dataset, the `km_subread` pipeline followed by the `de_pairwise` pipeline were applied to find differential expression between the control and 30 minute excess light timepoints (these pipelines are described in subsection 3.3.2). Several metrics were then used to summarise the effect of sequencing depth on the statistical power of differential expression analysis. The number of genes called as differentially expressed at each sequencing depth was calculated, as was the common biological coefficient of variation. A third measure, the log-transformed mean expression level of the least-expressed differentially expressed gene and the overall least-expressed gene were calculated. These metrics were plotted against sequencing depth to give a graphical overview of the effect of reduction of sequencing depth Figure 3.3.

3.3 Results

3.3.1 A Framework for the Creation of RNAseq Analysis Pipelines

To allow creation of diverse analysis pipelines suited to the multitude of RNAseq experimental designs and methodologies, I have implemented a generic framework for the creation of RNAseq data analysis pipelines. This framework takes the form of “wrapper scripts”, which act as wrappers around programs which other authors have created, and “pipeline” scripts, which combine these wrapped programs to perform an analysis. Wrapper scripts are the workhorses of a pipeline created with this framework, accepting three arguments: an input directory, and output directory, and arguments to be passed to the underlying program. Given these, the wrapper script will run the underlying program, automatically detecting input files from the input folder and automatically accounting for experimental features such as single or paired-end sequence data. Pipeline scripts describe processes of analysis of RNAseq data. They combine wrapped programs together to perform an analysis specific to a dataset. Each pipeline is run in parallel to utilise multi-processor computers, and every command is comprehensively logged, ensuring reproducibility. By removing the complexity of command syntax and increasing readability and reproducibility of pipeline workflows, I have enabled their use by a larger community of biologists without detailed training in bioinformatics, and ensured the reproducibility of results obtained.

3.3.2 An Improved Analysis Pipeline for Large Plant RNAseq Datasets

A series of two-step pipelines to analyse RNAseq datasets have been developed. Step one (`aln_subread`, `aln_subjunc` and `aln_tophat`) taking raw sequence reads and produces summarised gene-wise counts. Step two (`de_glm` or `de_pairwise`) applies statistical normalisation techniques and tests for differential expression. When applied combinatorially, these pipelines allow for different experimental designs to be analysed. Table 3.1 describes each software element of these pipelines, reasons for their selection, and references to literature describing each component.

The `aln_subread` pipeline

In this pipeline, quality of sequencing data is checked using `fastqc` (Andrews 2012), sequencing adaptors are removed with `scythe` (Buffalo 2013), and `seqtk` removes low quality sequences, before the quality is again checked using `fastqc`. The `subread` aligner, a very fast RNAseq-compatible short read aligner, then aligns reads to the current TAIR10 *A. thaliana* reference genome. Gene expression is summarised gene-wise by counting the number of reads which align to genic loci with `featureCounts`, completing the pipeline.

The `aln_tophat` pipelines

For studies examining alternative splicing of mRNA transcripts, an aligner able to detect splicing *de novo* is required (Kim et al. 2013). `Tophat2`, one of the most popular RNAseq aligners, is able to align short reads while detecting slicing isoforms (Kim et al. 2013). `aln_tophat` is identical to the `aln_subread` pipeline, except it uses the `Tophat2` aligner, allowing study of alternative splicing. However, this *de novo* detection of splicing comes at a performance cost, and is not necessary for simple quantitation of gene expression.

`de_pairwise`

For paired experimental designs, pairwise statistical tests can be performed between samples. The `de_pairwise` pipeline implements these tests using the `edgeR` R package. This pipeline first reads count files, removes loci not detected above statistical noise and non-protein coding loci, normalises counts using the trimmed Mean of M values normalisation method of Robinson and Oshlack (2010). Transcriptome-wide and gene-wise dispersion are then calculated and tests are conducted between groups described in a keyfile. Tables of differential expression and diagnostic plots are also created.

`de_glm`

If the experimental design is multi-factorial, pairwise analysis is inadequate and Generalised Linear Model-based hypothesis testing functions of `edgeR` are required. This pipeline fits such models, and tests user-defined contrasts for differential expression. Processes analogous to those in the `de_pairwise` pipeline are performed, yielding similar tabular and graphical descriptions of differential expression.

3.3.3 Comparison of Differential Expression Pipelines

A large difference in computational cost between four pipelines (`aln_subread`, `aln_tophat`, `aln_tophat_htseq` and `aln_subread_htseq`). These differences were significant (ANOVA, 3 degrees of freedom (DF), F=15738 p<2e-16) As is shown graphically in Figure 3.2, the `aln_subread` is the fastest, followed by the `aln_subread_htseq` pipeline, the `aln_tophat` and the `aln_tophat_htseq`. The computational time can become a limitation in very large and heavily replicated data sets such as those needed to identify important effects of environment on common genetic variation.

Quantification of gene expression by the `aln_subread` pipeline is comparable to that obtained by the `aln_tophat` pipeline. As shown in Figure 3.3, there is a very tight relationship between counts produced by the Tophat2 and subread aligners. The slope of $log(n+1)$ transformed raw count data when the model $tophatcounts \sim subreadcounts$ is fitted is 0.994, with p < 2e-16 and R^2 of 0.993. This indicates that there is approximately 0.7% statistical variation between these aligners. Table 3.2 illustrates the increased percentage of reads the `aln_subread` pipeline is able to align to both the entire genome and to the protein-coding transcriptome, when compared to the `aln_tophat` pipeline.

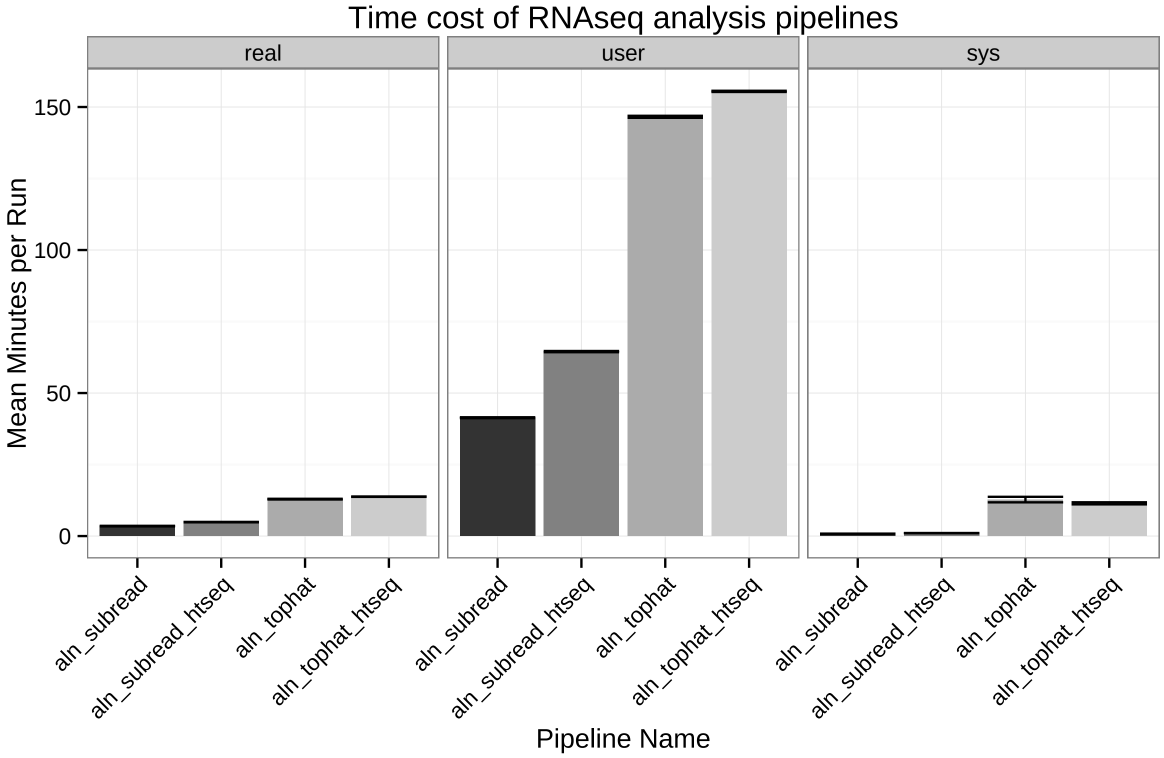

Figure 3.2: Computational cost of the `aln_subread`, `aln_tophat`, `aln_tophat_htseq` and `aln_subread_htseq` RNAseq analysis pipelines differs significantly. The "real" computational cost describes the number of minutes each pipeline took to complete. The "user" and "sys" metrics describe the number of CPU-minutes spent running user code (i.e. the pipeline components) and performing kernel operations on behalf of user code (e.g. input/output, memory (de)allocation and other system calls) for each pipeline execution.

Pipeline Name	Percentage Reads Mapped	
	Entire Genome	Protein-coding Genes
aln_subread	99.43	98.99
aln_tophat	97.02	96.30

Table 3.2: Percentage of mapped reads to genome and transcriptome. The `aln_subread` pipeline is able to map a slightly higher percentage of reads to both the genome and to the transcriptome when compared to the `aln_tophat` pipeline. This indicates it's increased sensitivity, and echoes findings by its authors (Liao, Smyth, and Shi 2013b)

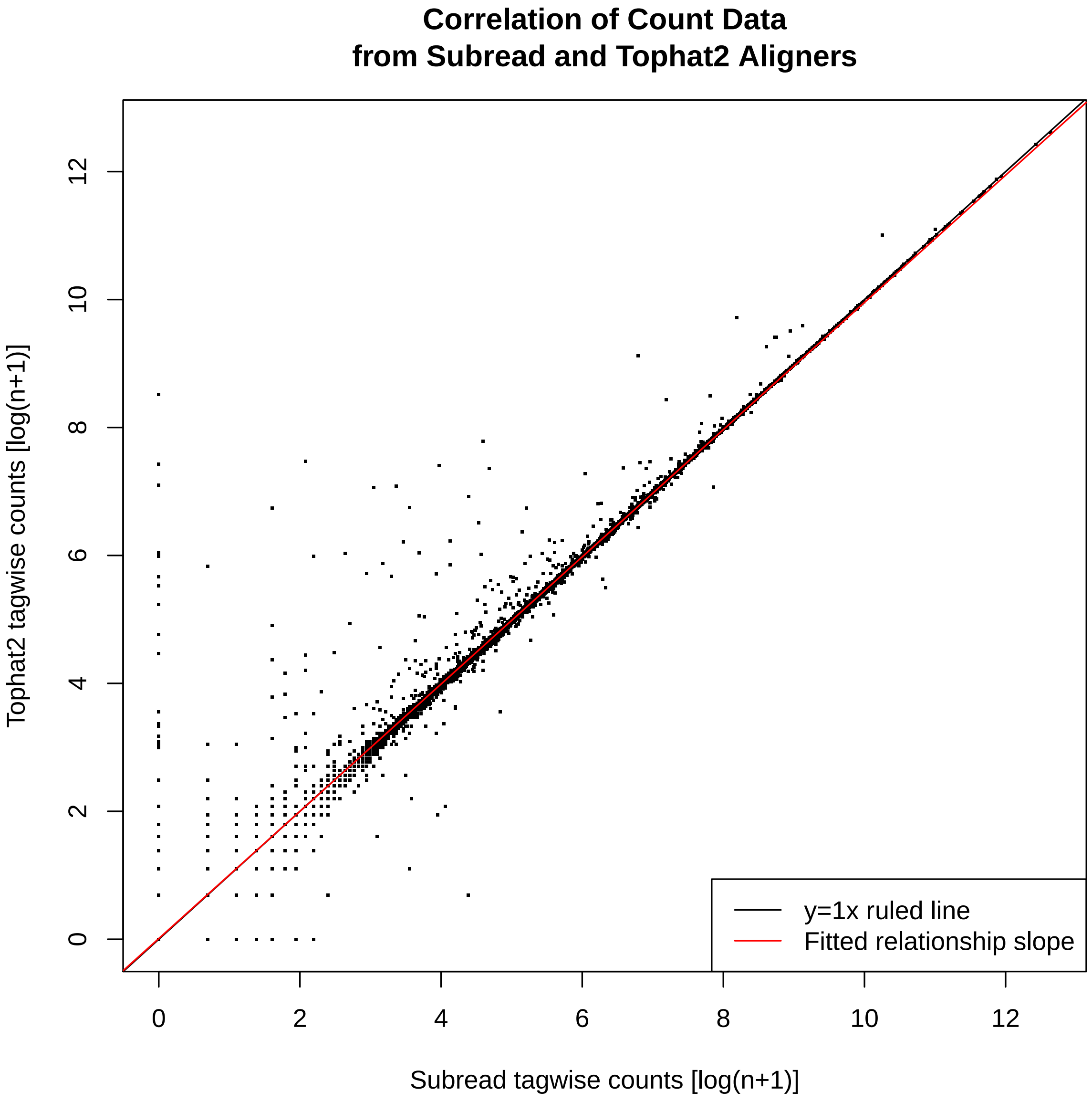

Figure 3.3: Comparison of gene-wise counts derived from the Tophat2 and subread short read aligners. A tight relationship is observed between count data from these aligners. If aligner had no effect, all points (genes) would fall exactly on the $y = x$ line. There are off-diagonal points, however these occur mostly at low expression levels and are likely due to the required log transformation of count data. Note that there are almost 18,000 genes plotted here, and off-diagonal scatter may, by eye, appear exaggerated.

3.3.4 Substantial reduction of RNAseq coverage is possible

While high-throughput sequencing has reduced sequencing costs dramatically, costs have stabilised in recent years (Glenn 2011). Multiplexing many samples per lane is important for high-throughput transcriptomics to increase replication to improve estimates of gene expression, and to increase sample breadth to better estimate transcriptome variation due to genetic or treatment effects. However, improvement to statistical power from increased sequencing depth diminishes as sequencing depth increases as the number of genes called as differentially expressed increases in a non-linear fashion with sequencing depth (Figure 3.4). Below a lower limit — 5 million reads — the number of genes called as differentially expressed reduces rapidly. This is concomitant with a decrease in number of genes whose expression is considered, and increases in biological coefficient of variation (data not shown). For the experimental system used in this experiment, I would recommend an optimal sequencing depth of approximately 5 million reads (or read pairs) to balance statistical power against additional sequencing cost. Budget permitting, more biological replicates or treatments should be included before sequencing is performed. These results have important implications for experimental design of future experiments.

3.4 Summary and Technical Discussion

In this chapter, I have implemented a framework allowing easier creation and optimisation of RNAseq analysis pipelines. Via literature review and brief experimental validation, I have selected software optimised for accuracy and speed in analysis of RNAseq data. Generic analysis pipelines that utilise these software have been created to allow investigation of common RNAseq experimental designs. Subtle transcriptional variation requires software to be optimised for accuracy and statistical rigour, while optimising software for performance allows end users to conduct their own analyses, ensuring and end-to-end understanding of their experiments. Emphasis has been placed on reproducible analysis of RNAseq, ensuring the validity and reproducibility of results obtained. These advances allow for the examination of subtle transcriptome variation, such as between different growth conditions or subtle genetic variation.

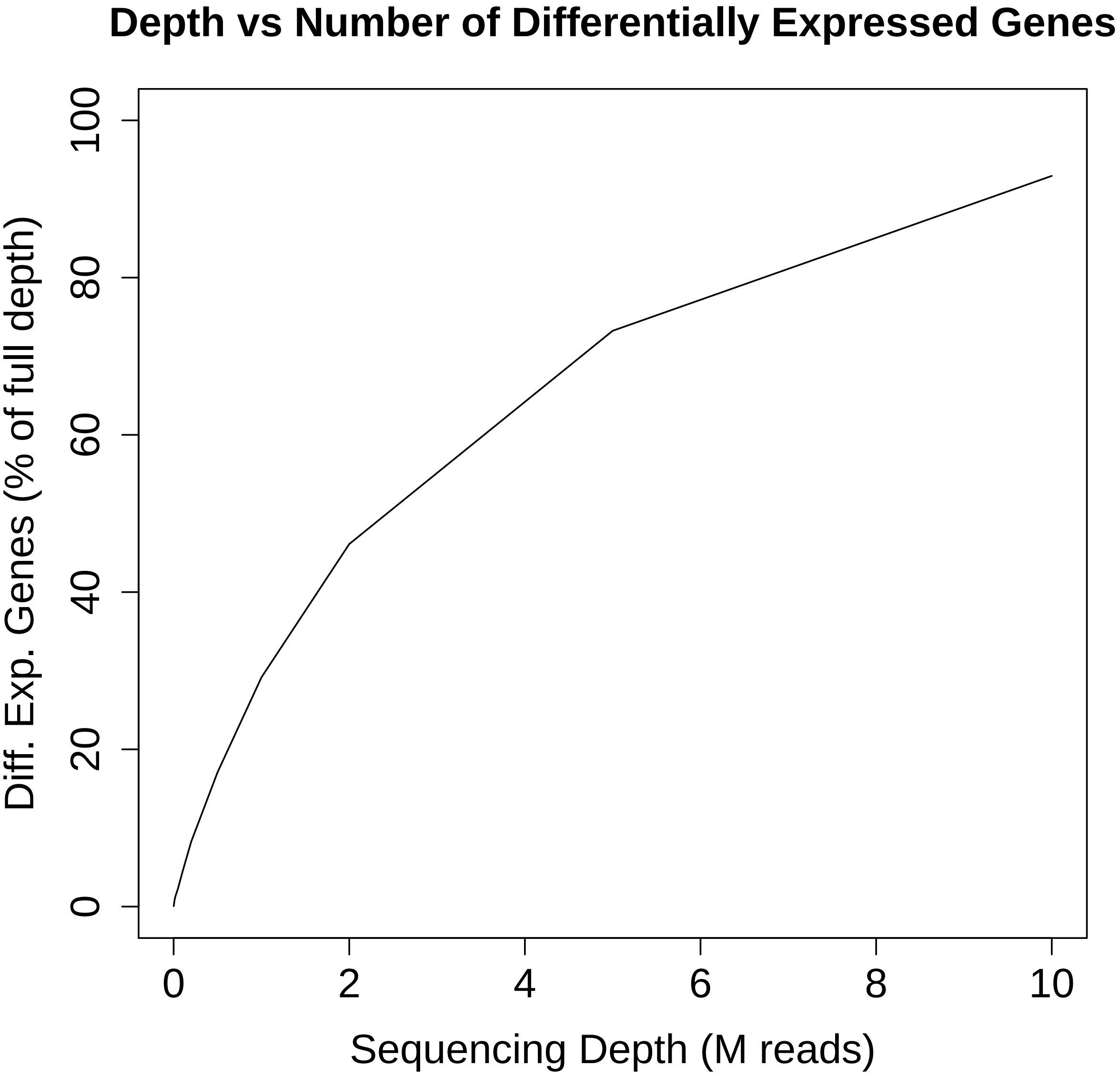

Figure 3.4: Decreasing sequencing depth per sample decreases the number of genes called as differentially expressed. This occurs because the total number of genes which can be examined for differential expression decreases in a similar fashion (data not shown). At 5 million reads per sample, approximately 80% of genes differentially expressed with full datasets (12-17 million reads) are found to be differentially expressed. At 2 million reads, this figure falls to approximately 50%, indicating at least 5 million reads are required to detect the majority of genes differentially expressed with three biological replicates.

Chapter 4

Transcriptome Variation Under Dynamic Growth Conditions

4.1 Background, Aims and Hypotheses

Transcriptional responses to abiotic stress have been studied by many authors (Seki et al. 2001; Rossel, Wilson, and Pogson 2002). However, theses stresses are typically not studied in combination (Mittler 2006). This is despite evidence supporting the interactions between stresses (Mittler 2006; Atkinson and Urwin 2012). Additionally, many authors have observed responses to abiotic stresses encountered under field conditions that are not simple additions of the responses to individual stresses as elucidated under laboratory study (Seki et al. 2001; Mittler 2006; Atkinson and Urwin 2012; Atkinson, Lilley, and Urwin 2013; Wituszyńska et al. 2013). Together, these shortcomings warrant the investigation, under controlled laboratory conditions, of abiotic stresses in combinations which mimic those observed under field conditions.

To examine the effect of altered long-term light intensity in field-like combinations of climatic variables including temperature and humidity, dynamic growth conditions have been designed (see chapter 2). In this chapter, I present preliminary transcriptomic study of plants grown under these dynamic conditions. Specifically, I aim to test if plants exposed to modest excesses of light either continuously (excess light dynamic growth conditions) or intermittently (fluctuating light) exhibit increased hardening to excess light when compared with sufficient light dynamic growth conditions. I hypothesise that this increased hardening would manifest as increased steady-state transcription of stress-responsive genes (Wituszyńska et al. 2013; Gordon et al. 2013), and a reduced induction of stress responsive genes upon exposure of plants to an 8x hot excess light treatment. More specifically, I expect that the excess and fluctuating light conditions will exhibit increased steady-state expression of stress-responsive transcripts compared with the sufficient growth condition, and with steady-stated expression observed in previous studies of plants grown under static growth conditions. Furthermore, I hypothesise that the fluctuating light condition will exhibit higher steady-state expression of stress-responsive transcripts, as plants are generally less able to acclimate to fluctuations in light intensity than to constitutive excess light (Külheim, Ågren, and Jansson 2002; Alter et al. 2012; Gordon et al. 2013).

4.2 Methods

4.2.1 Growth and harvesting of *Arabidopsis*

Initially, I had aimed to conduct a experiment which mapped expression QTLs under dynamic growth conditions. Thus, 80 *A. thaliana* Recombinant Inbred Intercross (RIX) lines were planted (see Appendix Table 6.2, Zou et al. (2005)). Additionally, plants of reference accessions Cape Verde Islands (Cvi), Columbia (Col) and Landsberg erecta (Ler), the photoprotection mutants *stn8*, *pgr5*, *npq1* and *npq4* were planted. All lines were planted in triplicate for each dynamic growth condition. Plants were grown in a carefully controlled manner to minimise variation in germination time and the developmental state of plants at a given time. Seeds were sown directly onto Debco Seed Raising Mix (Debco Pty. Ltd.), mixed with 3 grams per litre Osmocote® slow release fertiliser. Following sowing, plants were lightly watered from above and vernalised in a 4 °C cool room for three days. Plants were germinated under static growth conditions, of approximately 120 μ mol photons $m^{-2} s^{-1}$, 21 °C , with a 12 hour photoperiod. Following thinning, the remaining plants were established under these conditions for two weeks, before being distributed between three SpectralPhenoClimatron growth cabinets, with temperature, humidity and lighting controlled according to the sufficient, excess and fluctuating light dynamic growth conditions described in subsection 2.3.2. Plants continued to grow under these conditions until and beyond harvesting. Throughout plant growth in the SpectralPhenoClimatron, high resolution images were captured at 20 minutes intervals to enable analysis of growth using high-throughput phenomics, a study beyond the scope of this thesis

To assay plant response to hot excess light, plants were exposed to one hour of hot excess light from sodium vapour and tungsten filament lamps in a Conviron growth cabinet. In this stress, hereafter referred to as hot excess light, temperature was maintained at approximately 30 °C and light intensity at approximately 1000 μ mol photons $m^{-2} s^{-1}$ .To enable study of transcriptomic responses to this assay, one fully expanded leaf was taken from each of the surviving plants between 0 and 10 minutes before and after this excess light treatment. To do so, petioles were cut with clean scissors, and leaves gently rolled without crushing to facilitate placement in 96 well 1.2 mL deep well plates. Plates were kept in dry ice while harvesting occurred, and were transferred to -80 °C freezers for storage.

4.2.2 RNAseq Library Preparation and Sequencing

Due to the failure of attempts to implement the high-throughput RNAseq protocol of Kumar et al. (2012) (briefly discussed in subsection 6.4.1), a subset of all samples were selected for RNAseq analysis. Specifically, the samples of all Col-0 reference accession plants were used, as described in Table 4.1. Tissue of selected samples was extricated from 96 well plates into pre liquid N_2 cooled 1.5ml micro-tubes, ensuring plates remained frozen. Samples were ground using a Qiagen TissueLyser II for two one minute pulses at 25Hz, cooling racks in liquid N_2 between each pulse.

Total nucleic acids (TNA) were extracted from samples using a commercial reagent (Trizol, Life Technologies) immediately after grinding. Total nucleic acids were extracted by adding 1mL of Trizol to each well-ground sample and shaking vigorously by hand, before adding 200 μLchloroform and shaking by hand again. Samples were incubated for 3 minutes at room temperature, before centrifugation at 14000 rcf for 10 min in a chilled centrifuge. The TNA contained in the aqueous phase was re-extracted with chloroform and precipitated with 500 μLice cold isopropanol before incubation at -20 °C overnight.

Total nucleic acids were precipitated by centrifugation for 20 minutes at 20000 rcf and 4 °C , washed with 1mL 75% Ethanol, and resuspended in 50 μLRNase free 10 mM Tris-HCl. The quality of RNA in extracted TNA was assayed using the Agilent Bioanalyser digital electrophoresis platform. Samples were loaded into a Plant Nano analysis chip, and analysis run according to manufacturer's protocol.

RNAseq libraries were then prepared using the Illumina TruSeq V2 RNAseq Sample Preparation kit (Part number RS-122-2002). As previous studies in the Pogson Lab indicated DNAse treatment of TNA samples before RNAseq library preparation was not necessary (pers. comm. Peter Crisp, 2013), TNA was diluted with 10mM Tris-HCl to 80 $ng\,\mu L^{-1}$ for use as input material. The manufacturer's protocol was then followed to produce RNAseq libraries, with modifications. RNA was fragmented by heating samples at the "Elution 2 - Frag - Prime" stage to 94 °C for 7 minutes, in place of the 8 minutes recommended by manufacturer guidelines to increase median insert size. To create cDNA, the SuperScript III reverse transcriptase (Life Technologies, part number 18080044) was used, and thus the incubation temperature was increased from 42 °C to 50 °C , per SuperScript III guidelines. During every SPRI clean-up step throughout the protocol, DNA bound to SPRI beads was washed with 180 μLethanol rather than the recommended 200 μL, allowing all liquid to be effectively removed with a P200 multi-channel pipettors.

A pilot enrichment PCR was conducted with a subset of samples, enabling estimation of optimal cycle number for final enrichment PCR. To do so, quarter-volume PCRs (12.5 μLmaster-mix, 2.5 μLsample) were run: samples 1 and 13 were run for 10 cycles, and 2 and 19 were run for 14 cycles, with the 60 °C annealing time extended to 45 seconds. Then, half-volume PCRs were used to amplify libraries for 12 cycles. Two libraries (samples 6 and 17) whose amplification failed with 12 PCR cycles were amplified using quarter-volume PCRs with 17 amplification cycles. These libraries were then cleaned up with SPRI beads, per TruSeq kit protocol. The success of these PCRs were assayed by digital electrophoresis, using the MultiNA instrument, as per manufacturer's protocols, both before and after SPRI cleanup.

Final sequencing libraries were created by diluting and pooling RNAseq libraries. Libraries were diluted to 10nM, as calculated from MultiNA quantification. They were then quantitated fluorometrically using the Qubit 2.0 instrument (Life Technologies) with the dsDNA BR assay kit per manufacturer's protocol, and diluted to 5nM accordingly. These 5nM libraries were re-quantitated fluorometrically as above. Samples 1-12 and 13-18 plus 4 additional samples from a colleague were pooled to equimolarity, forming two final sequencing libraries. Raw 100bp paired end sequence data was obtained by sequencing final sequencing libraries on two Illumina HiSeq 2500 sequencing lanes, performed at the Bio-molecular Resource Facility, John Curtain School of Medical Research, ANU.

4.2.3 Computational Analysis of RNAseq Data

Raw Illumina paired-end 100bp sequence data was obtained as gzipped FASTQ files from the BRF. To gauge the quality of the obtained sequence data, several analyses were conducted. Firstly, the number of reads obtained from each library were calculated using the code shown in listing 1 below. Then, PHRED scores (defined as $-10\,log_{10}(P)$, where P is the probability of error at a given position) for each sequence base in each library were analysed using analysis pipelines described in subsection 3.3.2.

Analysis pipelines described in subsection 3.3.2 were applied to this dataset, and the RRGS dataset described in subsection 3.2.1. Firstly, the `aln_subread` pipeline computed read counts gene-wise from raw short reads, using the keyfile below (listing 2). Then, the `de_glm` pipeline was used test for differential expression. A general linear model (GLM) of the form $\sim$ *Group* was fitted, according to the keyfile shown in listing 2. Statistical tests for differential expression between contrasts shown in Table 4.6.

Sample Number	Accession	Growth Condition	Light Treatment
1	Col-0	Sufficient	0h
2	Col-0	Sufficient	0h
3	Col-0	Sufficient	0h
4	Col-0	Sufficient	1hHL
5	Col-0	Sufficient	1hHL
6	Col-0	Sufficient	1hHL
7	Col-0	Fluctuating	0h
8	Col-0	Fluctuating	0h
9	Col-0	Fluctuating	0h
10	Col-0	Fluctuating	1hHL
11	Col-0	Fluctuating	1hHL
12	Col-0	Fluctuating	1hHL
13	Col-0	Excess	0h
14	Col-0	Excess	0h
15	Col-0	Excess	0h
16	Col-0	Excess	1hHL
17	Col-0	Excess	1hHL
18	Col-0	Excess	1hHL

Table 4.1: Samples analysed by RNAseq and qPCR. These samples are referred to by their sample numbers for the remainder of this thesis. Treatments 0h and 1hHL refer to samples taken before and after one hour treatment with hot excess light. Growth conditions of sufficient, fluctuating and excess refer to their respective dynamic growth conditions.

```
for fqfile in `find -name *.fastq.gz`
do
    echo "$fqfile $(($(zcat $fqfile | wc -l) / 4))"
done
```

Listing 1: Count the number of reads in raw sequence files

```
Ordinal     Sample     GrowthCondition     Treatment     Group
1     Sample_BJP_K1_1_index1     Sufficient     0h     Sufficient.0h
2     Sample_BJP_K1_2_index3     Sufficient     0h     Sufficient.0h
3     Sample_BJP_K1_3_index8     Sufficient     0h     Sufficient.0h
4     Sample_BJP_K1_4_index9     Sufficient     1hHL     Sufficient.1hHL
5     Sample_BJP_K1_5_index10     Sufficient     1hHL     Sufficient.1hHL
6     Sample_BJP_K1_6_index11     Sufficient     1hHL     Sufficient.1hHL
7     Sample_BJP_K1_7_index20     Fluctuating     0h     Fluctuating.0h
8     Sample_BJP_K1_8_index21     Fluctuating     0h     Fluctuating.0h
9     Sample_BJP_K1_9_index22     Fluctuating     0h     Fluctuating.0h
10     Sample_BJP_K1_10_index23     Fluctuating     1hHL     Fluctuating.1hHL
11     Sample_BJP_K1_11_index25     Fluctuating     1hHL     Fluctuating.1hHL
12     Sample_BJP_K1_12_index27     Fluctuating     1hHL     Fluctuating.1hHL
13     Sample_BJP_K2_13_index1     Excess     0h     Excess.0h
14     Sample_BJP_K2_14_index3     Excess     0h     Excess.0h
15     Sample_BJP_K2_15_index8     Excess     0h     Excess.0h
16     Sample_BJP_K2_16_index9     Excess     1hHL     Excess.1hHL
17     Sample_BJP_K2_17_index10     Excess     1hHL     Excess.1hHL
18     Sample_BJP_K2_18_index11     Excess     1hHL     Excess.1hHL
```

Listing 2: The `kevin-hons-glm.key` keyfile

4.2.4 Quantitative Real-time PCR Quantification of Gene Expression

To assay expression of genes shown to respond to hot excess light in previous studies, quantitative PCR (qPCR) was used. RNA extracted for RNAseq analysis (see subsection 4.2.2) was treated with DNAse, to remove genomic DNA contaminants, using Turbo DNAse (Life Technologies). Approximately 15-20 $\mu g$of RNA in 90 $\mu L$was combined with 10 μLTurbo DNAse buffer and 1 μLTurbo DNAse, before incubation at 37 °C for 30 minutes, at which point an additional 1 μLTurbo DNAse was added to sample. Samples were mixed, and incubated for a further 30 minutes at 37 °C . RNA was recovered by phenol-chloroform extraction, adding 100 μL1:1 phenol-chloroform mixture to the DNAse reaction solution, mixing, separating phases by centrifugation for 10 minutes at 14000 rcf and 4 °C , and precipitating RNA by adding 200 μLice cold isopropanol, incubating at -20 °C overnight, centrifuging to pellet RNA, washing RNA pellet with 70% ethanol and re-suspending RNA in 20 μLDEPC-treated MilliQ water. RNA quality was assayed by visualising denatured RNA on a 1% Agarose gel, prepared using buffers made with DEPC-treated water to prevent RNA degradation.

Complimentary DNA (cDNA) was synthesised using Invitrogen SuperScript III First-strand cDNA synthesis kit. RNA was diluted to 100 $ng\,\mu L^{-1}$, denatured at 65 °C for 5 minutes, and cDNA synthesis reactions consisting of 10 μLdenatured RNA sample, 1 μL50 $nmol\,L^{-1}$dT(18)VN primer, 1 μL10 $mmol\,L^{-1}$dNTPs, 2.5 μLnuclease free water, 4 μL5x reaction buffer, 1 μL100 $mmol\,L^{-1}$DTT, and 0.5 μLSuperScript III enzyme. This reaction solution was mixed, centrifuged briefly, and incubated at 50 °C for 60 minutes, before enzyme inactivation at 70 °C for 15 minutes. Samples were then stored at -20 °C until use.

Expression was quantified with qPCR using the Roche Sybr Gold master mix kit (part number 04-707-516-001) in a Roche LightCycler 480 thermocycler. Reactions of 10 μLconsisting of 1 μLcDNA, 3.6 μLnuclease free water, 0.2 $\mu L$of each 20 $\mu mol\,L^{-1}$primer, and 5 μLSybr Gold 2x master mix. Reactions were conducted in technical triplicate in 384 well plates sealed with qPCR-compatible plate seals. Expression of eight genes (APX2, ELIP1, ELIP2, LHCB1.4, as well as reference genes PP2AA3 and GAP) was quantified using primer sets described in Table 4.2 in 24 samples. These cDNA samples were samples 1-18 of the dynamic growth condition dataset described in Table 4.1, as well as cDNA samples prepared using identical methods as described above from plants grown under static conditions (100 μ mol photons $m^{-2}\,s^{-1}$light intensity, 12 hour photoperiod, 21 °C) for five weeks, before and after exposure to 10x hot excess light for one hour. In addition, template-less and reverse transcriptase free controls were conducted for some primer sets to ensure absence of genomic DNA contamination.

Raw quantification curves were obtained by thermocycling qPCR reactions in a LightCycler 480 thermocycler. Reactions were heated to 95 °C for 10 minutes, before cycling between 95 °C for 30 seconds, 60 °C for 45 seconds, and 72 °C for 60 seconds for 45 cycles. A final incubation at 72 °C for 5 minutes was followed by a slow ramping of temperature from 45 to 95 °C with continuous quantitative analysis, to obtain a melting curve. Raw quantification data was analysed with LinRegPCR (Ruijter et al. 2009) to create N0 values, a statistically rigorous arbitrary unit of quantification suitable for relative quantification. Quantification relative to PP2AA3 was then calculated samplewise for all primer sets analysed, using custom analysis code in R (see subsection 6.2.3). An ANOVA model was fitted to relative quantification, with Tukey's honest significant differences post-hoc testing to determine specific effects.

Name	AGI	NCBI Acc.	Sequence	Length	Tm (°C)	GC %	Amplicon Size
APX2_ej4_F	AT3G09640	NM_001035587.2	GCCGTTAGGCTTCTTGACCC	20	58.9	60.00	146
APX2_ej4_R			GGCTCAACTTTGTCCAGTCTACC	23	58.9	52.17	
GAPC2_5´F	AT1G13440	NM_101214.3	TCGGAAGAATCGGTCGTTTGG	21	58.5	52.38	105
GAPC2_5´R			TGTATGTCATGTACTCGGTGG	21	55.0	47.62	
ELIP2_F_UPL101	AT4G14690	NM_117551.2	CCACCACAAATGCCACAG	18	54.5	55.56	73
ELIP2_R_UPL101			GCAAATCTCCAAACTTCGTACTC	23	55.9	43.48	
PP2AA3_3´F	AT1G13320	NM_001035958.1	CGACCAAGCGGTTGTGGAGA	20	60.6	60.00	161
PP2AA3_3´R			CACAATTCGTTGCTGTCTTCTTT	23	56.2	39.13	
LHCB1.4_F	AT2G34430.1	NM_128995.2	TCCTCTCCTGCTTTGACCGG	20	59.4	60.00	87
LHCB1.4_R			TTTGCGCATGGTGATTCGGC	20	60.4	55.00	
KM_ELIP1_F	AT3G22840.1	NM_113183.3	AGATGCATGGCTGAGGGAGG	20	59.5	60.00	136
KM_ELIP1_R			AGTCGCTAAACTTTGTGCTCACC	23	59.4	47.83	

Table 4.2: QPCR primer sequences and characteristics.

4.3 Results

4.3.1 Quantification of Transcriptome-wide Responses to Altered Light Intensity Under Novel Growth Conditions

Successful Preparation of RNAseq Libraries

RNA of suitable yield and quality for RNAseq library preparation was obtained. Figure 4.1 and Table 4.3 illustrate the high quality and yield of RNA samples. Illumina RNAseq libraries of the expected size and concentration were successfully prepared from all 18 RNA samples. Figures 4.2 and 4.3 demonstrate the successful creation of RNAseq libraries. Two library amplifications (Samples 6 and 17) failed to reach a suitable yield after 12 PCR cycles, so the PCR was re-run with 17 amplification cycles, which created libraries of a suitable concentration (Figure 4.3).

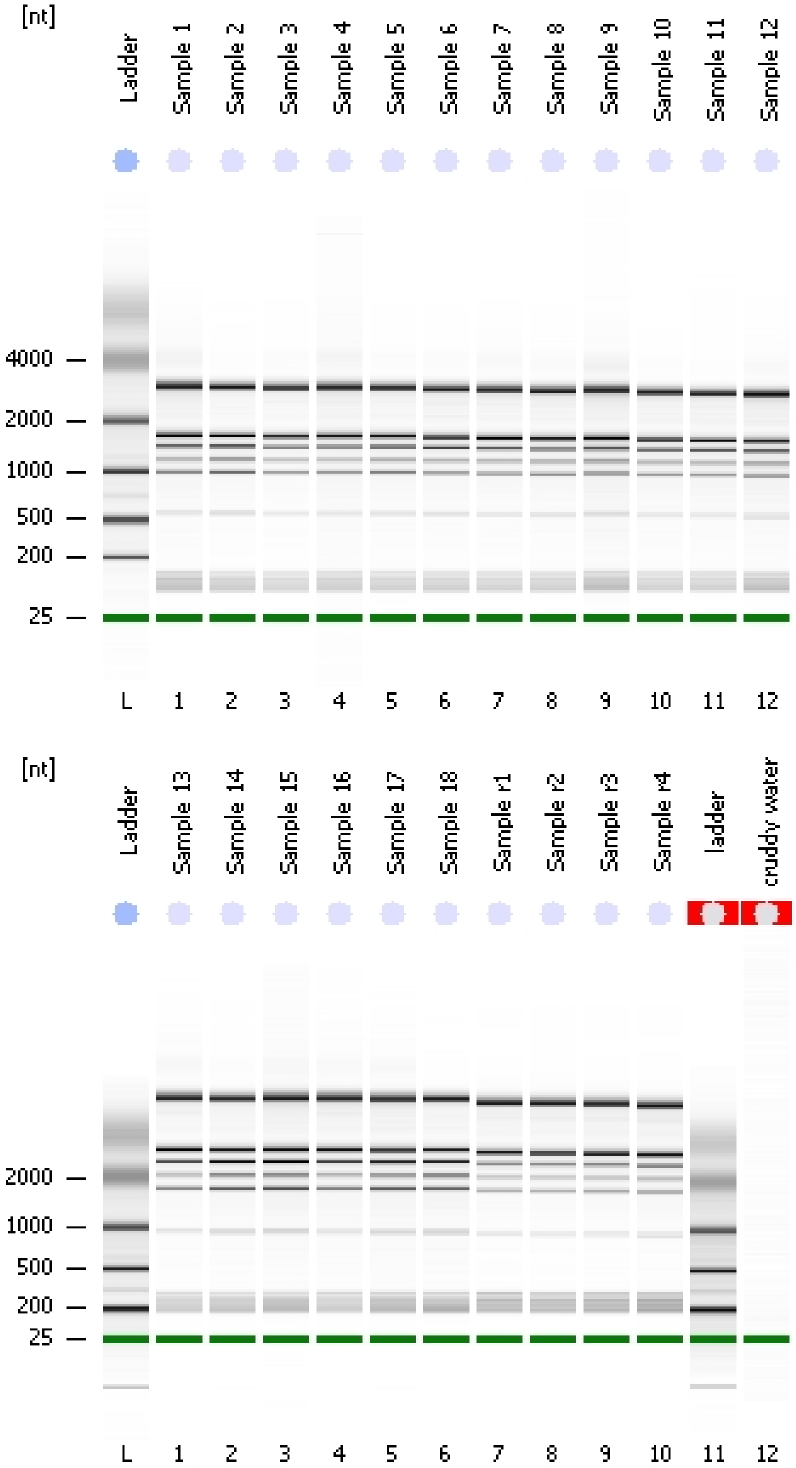

Figure 4.1: Bioanalyser 2100 digital electrophoretograms show expected rRNA peaks and mRNA smear, with no evidence of large-scale sample degradation. Samples 1-18 are numbered according to Table 4.1. The strong, clear bands at approximately 1900 and 3700nt are derived from the 18s and 25s nuclear rRNA species respectively (Babu and Gassmann 2011). The absence of a broad smear indicates minimal degradation of RNA has occurred (Babu and Gassmann 2011)

Sample	Conc. ($ng\,\mu L^{-1}$)	RIN
1	451	6.9
2	470	6.5
3	532	7
4	455	7.1
5	453	6.7
6	461	6.8
7	473	6.9
8	528	6.8
9	217	6.9
10	524	6.8
11	505	6.7
12	548	6.7
13	164	6.9
14	205	6.4
15	93	6.6
16	198	6.8
17	149	6.6
18	239	6.3

Table 4.3: RNA sample yield and RNA Integrity Number (RIN). RINs greater than 7 indicate high quality RNA, and RINs greater than 6 are acceptable (Babu and Gassmann 2011). Overall yield is sufficient, and all samples have a RIN of at least 6, indicating acceptable quality.

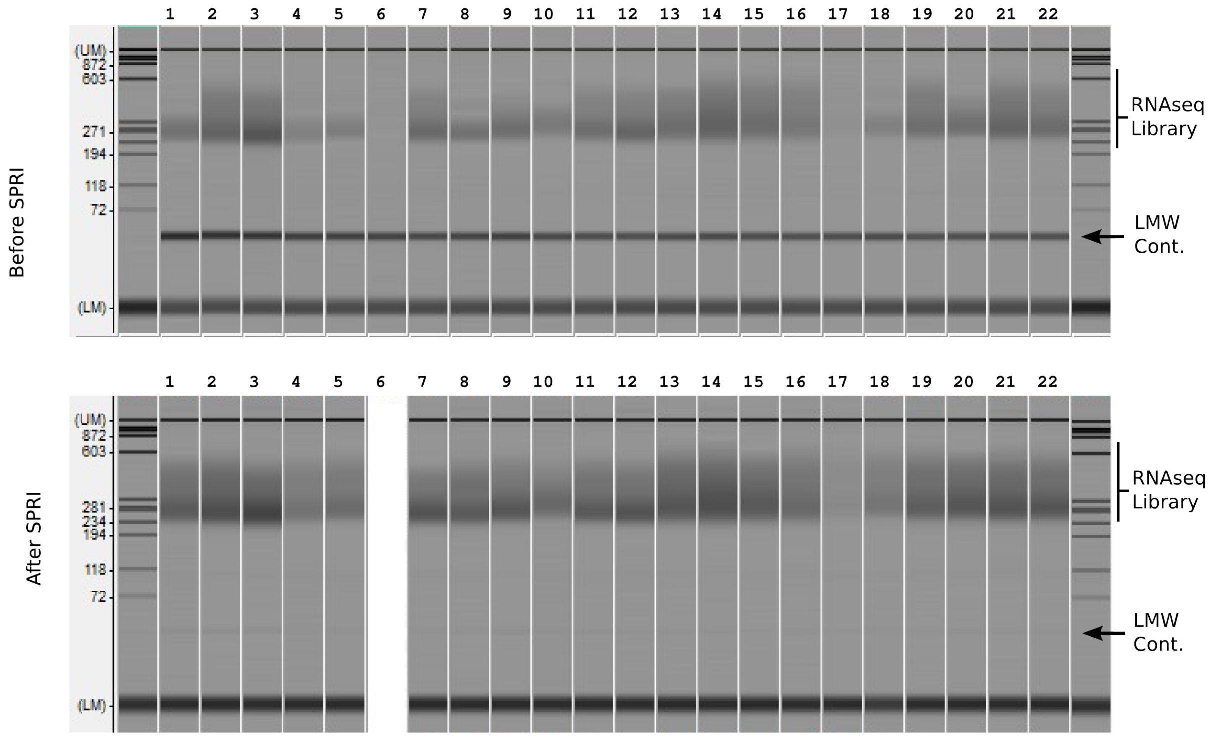

Figure 4.2: RNAseq libraries before (top) and after (bottom) final solid-phase reversible immobilisation (SPRI) cleanup. A broad smear of nucleic acid between approximately 200 and 600bp long is expected, ideally with a peak around 200-300b (Illumina 2012). Additionally, after PCR, a band of low molecular weight (approximately 60bp) contaminant is expected, and observed. Clean-up with SPRI is expected, and observed, to remove this contaminant efficiently.

Obtaining a High Quality RNAseq Dataset

To quantify global gene expression accurately, it is important to ensure the quality of raw sequencing data is acceptable. As detailed in Table 4.4, library depth variation was within an order of magnitude across both lanes sequenced. Sequence quality was high; 25th percentile PHRED quality score exceeded 28 at every base in all sequence libraries before any quality control (see Figure 4.4). Following sequence quality control, the minimal 25th percentile PHRED score increased to 30. These basic statistics show the sequencing of RNAseq libraries was successful and the resulting data suitable for further analysis. The very high proportion of sequence reads which aligned to the genome, and moreover to protein coding loci, is a further indicator of dataset integrity (Table 4.5).

4.3.2 RNAseq Reveals a Noisy Transcriptome

Once reads were aligned to the genome and gene-wise counts obtained, statistical assessment of differential expression was conducted. After dataset filtering to remove lowly expressed or non-protein coding loci, 17948 loci remained. Figure 4.5 illustrates the high biological coefficient of variation across the majority of genes. This is evidenced by the spread of points upwards in samples from the dynamic growth condition dataset when compared to those from the static growth condition (RRGS) dataset described in subsection 3.2.1. The overall biological coefficient of variation was also much higher in the dynamic growth condition dataset (common BCV = 0.493), when compared to the RRGS dataset (common BCV = 0.128). It is crucial to note that the RRGS experiment is not a control for the dynamic growth condition experiment, but is still useful an external comparison. The high variation between replicates causes reduced statistical power to detect differential expression (Robinson et al. 2013).

The high biological variance in expression is also demonstrated via multiple-dimensional scaling, an unsupervised clustering algorithm that describes the transcriptome-wide similarity of samples. Samples grown under dynamic growth conditions have a higher scatter about both axes of the multiple-dimensional scaling plot when compared to the RRGS dataset (Figure 4.6). Replicates often cluster less closely than treatments, however, upon treatment with one hour of hot excess light, plants grown under both static and dynamic

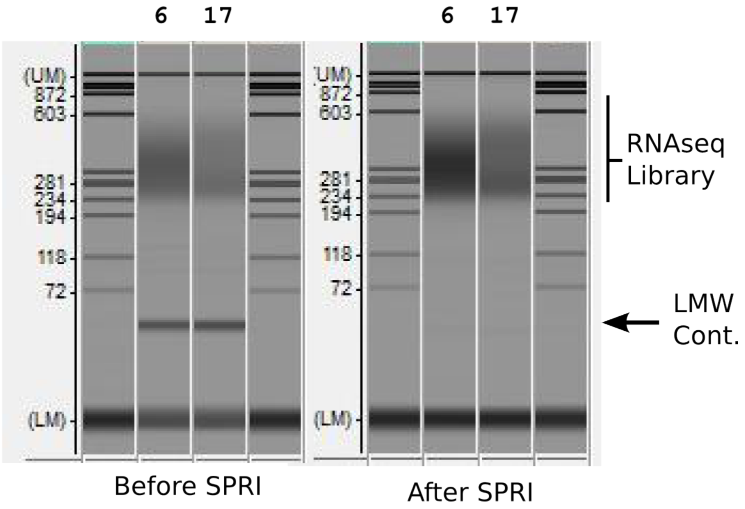

Figure 4.3: PCR amplification of samples that were not amplified after 12 PCR cycles. Note the presence of a band of between 200 and 600bp is expected, with a peak around 200-300b as expected, and as observed in Figure 4.2. Note also the efficient removal of low molecular weight contaminant after SPRI clean-up (right).

Library	Reads per Library (Millions)	
	Pre QC	Post QC
1	11.43	10.78
2	10.66	10.12
3	10.82	10.37
4	10.79	10.41
5	12.39	11.86
6	14.08	13.44
7	12.78	12.28
8	11.58	11.20
9	12.83	12.31
10	12.47	12.10
11	11.25	10.89
12	11.92	11.51
13	17.98	17.31
14	19.35	18.49
15	17.94	17.16
16	14.31	13.84
17	14.11	13.56
18	11.58	11.03
19	12.72	NA
20	16.86	NA
21	15.87	NA
22	12.36	NA

Table 4.4: RNAseq library sequencing depth. Reads per library before and after quality control refer to the length of raw sequence data, and to the size of the libraries immediately before statistical analysis of differential expression, after sparse tags and non-protein coding loci were removed. Libraries 19-22 were sequenced on behalf of a colleague, and do not form part of this thesis, thus have not been analysed for differential expression.

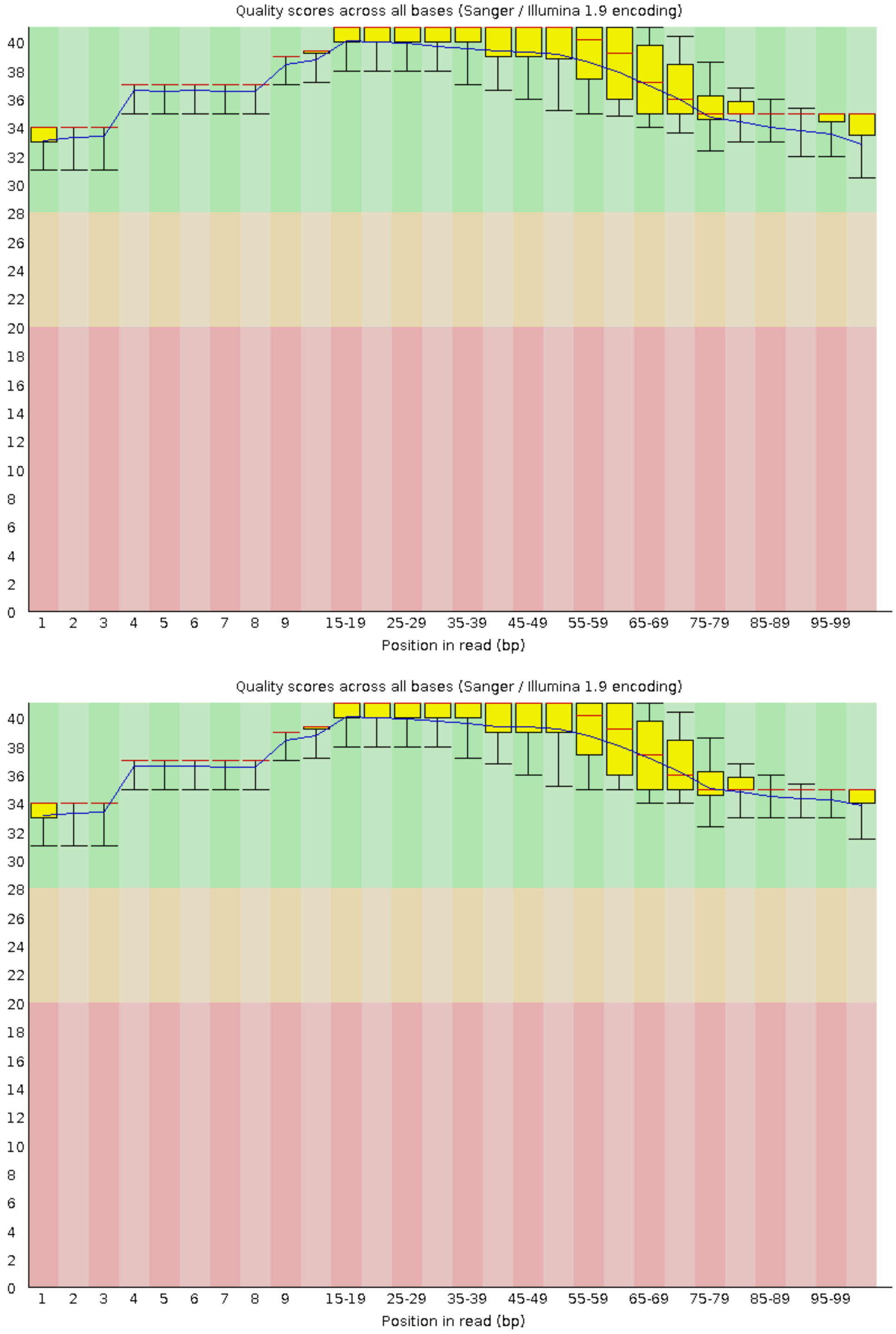

Figure 4.4: Per-base quality before (top) and after (bottom) quality control. These box-plots describe, per-position in each read (x-axis), quartiles and medians of PHRED quality score, which is related to probability of sequencing error at each position. PHRED scores of greater than 30 are considered good, and scores are expected to be lower towards the 3' end of read sequences, due to sequencing technology (Andrews 2012)

growth conditions exhibit similar patterns (in Figure 4.6, pre and post-treatment samples separate along a axis 1). It is also important to note that replicates within the RRGS dataset cluster together tightly compared with replicates from the dynamic growth condition dataset, a hallmark of the RRGS dataset's lower biological variance. Whist the variability observed in the dynamic growth condition dataset is concerning, promising qualitative trends transcriptome-wide patterns of differential expression warrant further — albeit cautious — investigation of this dataset.

Sample Number	Percentage Reads Mapped to:	
	Entire Genome	Protein-coding Genes
1	99.38%	97.59%
2	99.15%	98.10%
3	99.64%	98.60%
4	99.59%	98.92%
5	99.10%	98.48%
6	99.16%	98.26%
7	99.61%	98.82%
8	99.53%	99.07%
9	99.62%	98.84%
10	99.59%	99.24%
11	99.62%	99.14%
12	99.60%	98.94%
13	99.28%	98.88%
14	99.16%	98.88%
15	99.29%	98.70%
16	99.31%	99.09%
17	99.10%	98.86%
18	99.23%	98.59%

Table 4.5: Percentage of mapped reads to genome and transcriptome. These figures reiterate the successful creation of the RNAseq dataset, with very low (<4%) rates of contamination by non-protein coding RNAs or genomic DNA. This imparts confidence that any short read counted towards quantification of the expression of a protein coding gene is derived from mRNA transcribed from that gene.

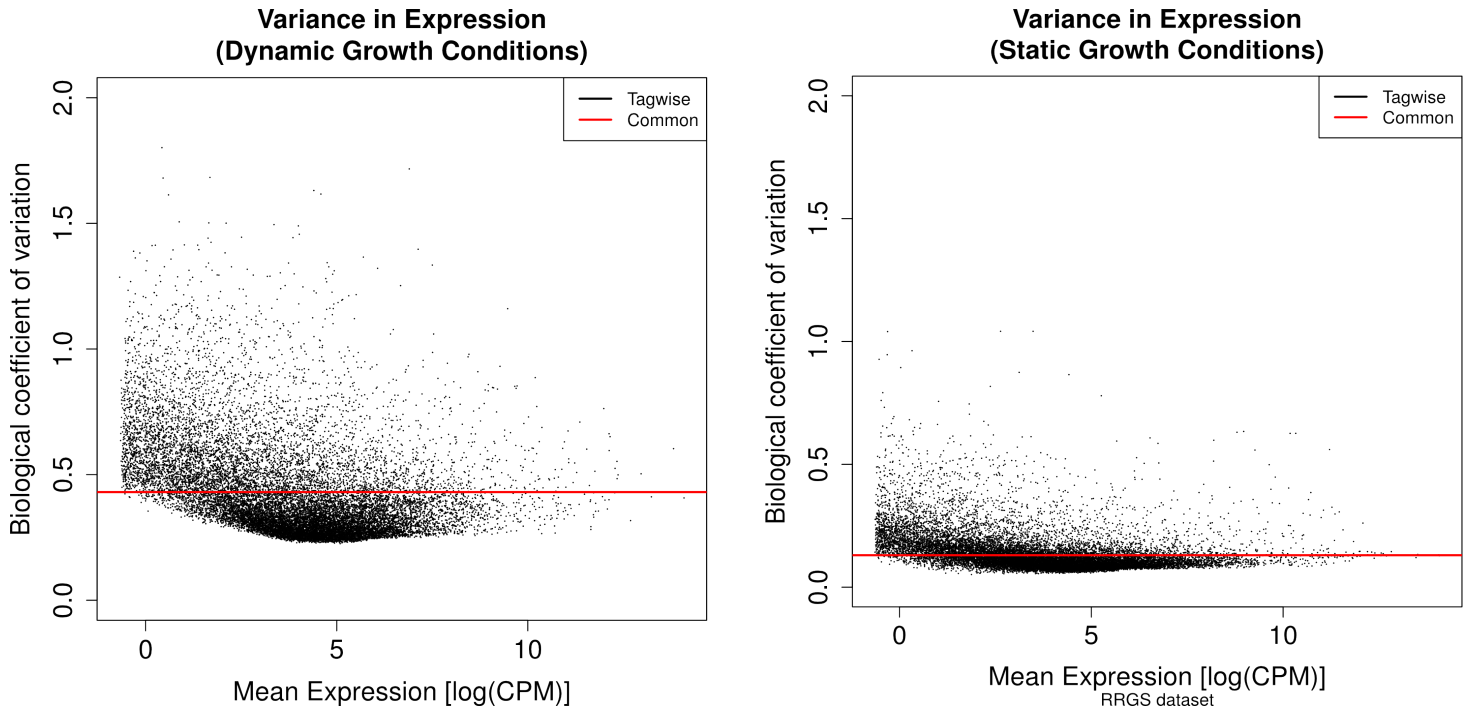

Figure 4.5: Biological coefficient of variation in dynamic (left) and static(right) growth conditions. Here, each point represents a gene, and the relationship between expression level (x-axis) and gene-wise biological coefficients of variation is plotted. Additionally, a common (analysis-wide) measure of the biological coefficient of variation is plotted (red line), for samples from the dynamic and static growth condition datasets, this equates to 0.493 and 0.128 respectively. The higher biological variance of the dynamic growth condition dataset when compared to the static growth condition (RRGS) dataset is evidenced by a more positive spread of gene-wise biological coefficients of variation, and by a higher common biological coefficient of variation.

Differential expression between dynamic growth conditions

Eight group contrasts were tested for differential expression. These contrasts test the effect of light intensity within the framework of dynamic growth conditions on both the steady-state transcriptome, and on the transcriptional response to exposure to hot excess light for one hour. A summary of statistically significant differential expression is shown in Table 4.6 and Figure 4.7. Small numbers of differentially expressed genes were observed between steady-state transcription in the sufficient growth condition and the excess and fluctuating growth conditions, but no differential expression was detected between the fluctuating and excess growth conditions. A transcriptional response to one hour of hot excess light was observed in plants grown under all growth conditions, however plants grown under excess light dynamic growth conditions showed the greatest number of differentially expressed genes in response to this treatment, followed by fluctuating and sufficient light conditions, in direct conflict with hypotheses. Tests for interaction between growth condition light intensity and treatment effect showed little or no significant differential expression Table 4.6. The differential expression observed between dynamic growth conditions is in contrast to the 3195 up-regulated and 3146 down-regulated genes differentially expressed after one hour of hot excess light treatment in plants grown under static growth conditions (the RRGS dataset described in subsection 3.2.1).

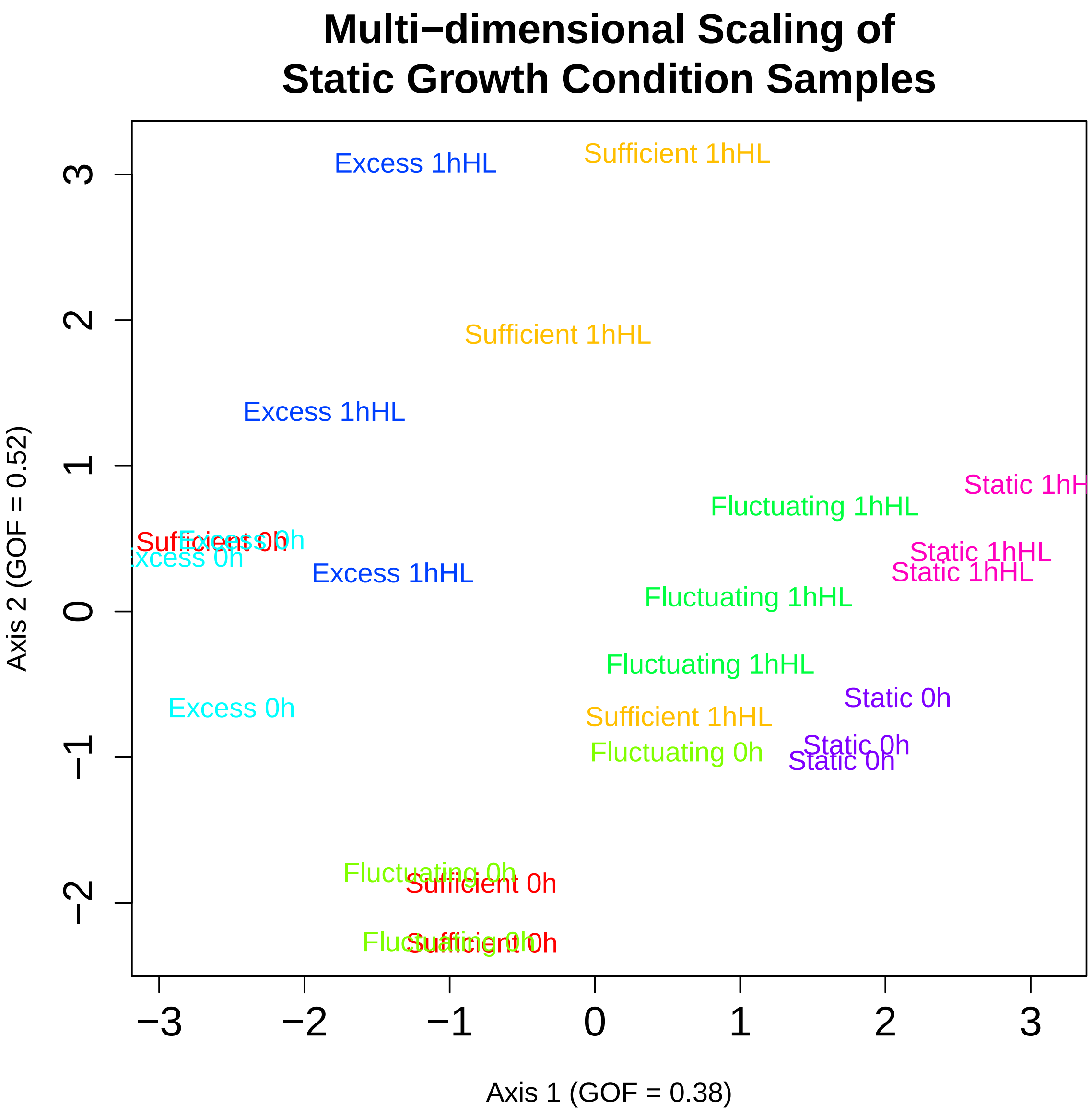

Figure 4.6: Multi-dimensional scaling of dynamic and static growth conditions. Replicate samples are represented in the same colour. Pre-hot excess light samples (0h) tend towards negative values on axis 2, while post-hot excess light samples (1hHL) tend towards positive values on axis 2. This trend is preserved across samples from plants grown under both dynamic and static growth conditions. Replicates within the rapid recovery gene silencing dataset (Static 0h and Static 1hHL) cluster together tightly compared with replicates from the dynamic growth condition dataset (Excess, Fluctuating and Sufficient 0h and 1hHL). Additionally, the is a trend towards higher growth condition light intensities towards negative values of axis 1. Note the meaning of axes is arbitrary; they are pseudo-variables that selected to separate samples based upon log-fold-changes between samples. Goodness of Fit (GOF) indicates that together, these two axes account for 90% of variation between samples, indicating that the majority of difference across entire transcriptomes between samples is described within this plot.

Contrast Name	Contrast Description	Genes Downregulated	Genes Up-regulated
Exc0-Suf0	Excess Light vs Sufficient Light	4	80
Flu0-Suf0	Fluctuating Light vs Sufficient Light	154	217
Exc0-Flu0	Excess Light vs Fluctuating Light	0	0
Suf1-Suf0	Sufficient Light pre- vs post-hot excess light	11	140
Exc1-Exc0	Excess Light pre- vs post-hot excess light	109	426
Flu1-Flu0	Fluctuating Light pre- vs post-hot excess light	94	201
Exc01h-Suf01h	Interaction between Fluctuating light and hot excess light treatment	0	0
Flu01h-Suf01h	Interaction between Sufficient light and hot excess light treatment	1	1

Table 4.6: Summary of differential expression between contrasts. The number of genes induced (up-regulated) or repressed (down-regulated) in each contrast with a false discovery rate of below 0.05 is described.

To gain biological insight from patterns of differential expression, gene ontology (GO) term enrichment analysis was used. Statistically significant enrichment of GO terms in genes up- and down-regulated in comparisons of steady-state expression and in transcriptional response to hot excess light. The GO terms enriched in genes up-regulated on exposure to hot excess light were highly conserved across all dynamic growth conditions and the RRGS dataset grown under static growth conditions (Figure 4.8). Specifically, terms including 'response to heat', 'response to high light intensity', 'response to hydrogen peroxide' and 'response to jasmonic acid stimulus' are among the most statistically over-represented genes induced by one hour of hot excess light in plants from all dynamic conditions. Moreover, these terms are also amongst the most statistically over-represented genes induced by one hour of hot excess light in plants grown under static growth conditions. Table Table 4.7 describes GO terms that are over-represented in genes induced by hot excess light treatment across all conditions; terms involved in biotic and abiotic stress represent the majority of the 37 such terms. Full details of the 30 most significantly enriched GO terms in all differential expression tests are described in appendix **??**.

4.3.3 Expression Patterns of Excess Light Marker Genes

Quantitative RT-PCR (qPCR) was used to examine expression patterns of excess light marker genes. Comparisons of both steady-state expression, and induction of expression upon treatment with hot excess light have been examined. *APX2*, a gene induced by oxidative stress and excess light, was up-regulated in sufficient, excess and fluctuating light dynamic light conditions (ANOVA, F=5.63, p=0.0072, 3 degrees of freedom (DF), with Tukey's honest significant differences demonstrating significant pairwise differences between expression in sufficient, excess and fluctuating light dynamic light conditions and in static growth conditions with $p<0.05$). Steady-state expression of LHCB1.4, a photosynthetic gene known to be down-regulated by excess light (Ruckle, DeMarco, and Larkin 2007), and was down-regulated in plants grown under dynamic growth conditions, with statistically significant down-regulation observed between excess and fluctuating dynamic light conditions and static growth conditions (ANOVA, F=4.45, 3 DF, p=0.019). Expression of ELIP1 is significantly up-regulated in excess and fluctuating dynamic light

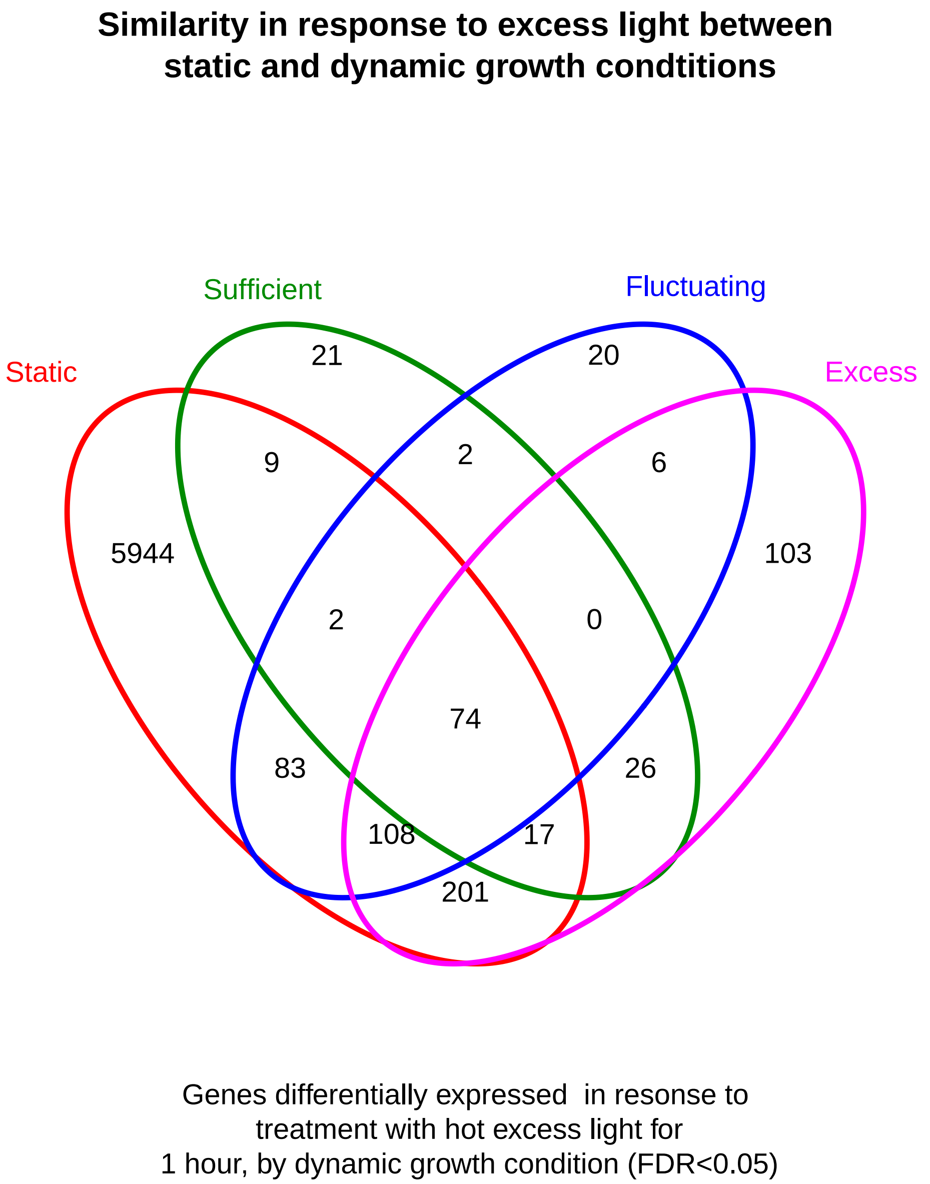

Figure 4.7: Extent of transcriptional response to a one hour treatment with hot excess light. Similarity between between sufficient, excess and fluctuating light dynamic growth conditions, and a similarly treated dataset grown under static growth conditions (from the RRGS dataset discussed in subsection 3.2.1).

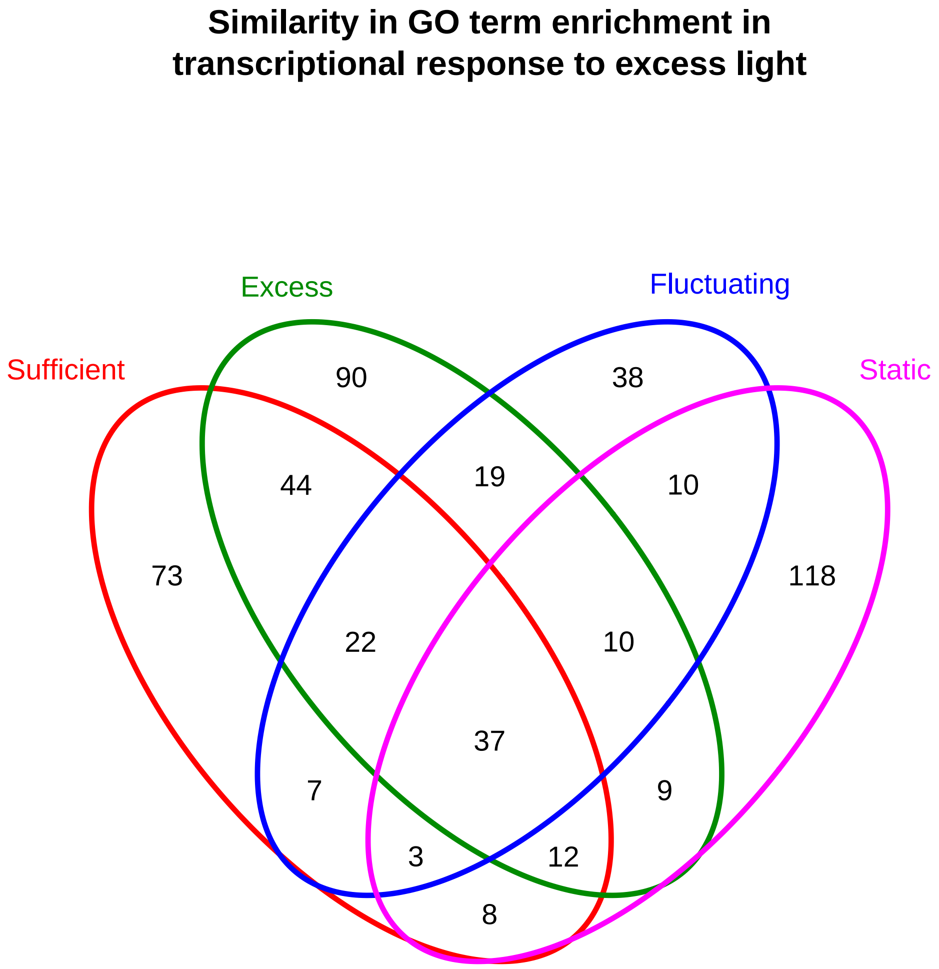

Figure 4.8: Similarity in GO terms between genes whose expression is induced by one hour of hot excess light in plants growth in Sufficient, Excess or Fluctuating dynamic growth conditions, as well as plants grown under static growth conditions (from the Rapid Recovery Gene Silencing dataset. A core set of GO terms is commonly over-represented in genes differentially expressed in all conditions (see Table 4.7). Numbers indicate the number of enriched gene ontology terms in common between each plant growth condition.

Gene Ontology Term
abscisic acid mediated signaling pathway
anthocyanin-containing compound biosynthetic process
cellular response to heat
endoplasmic reticulum
endoplasmic reticulum lumen
gibberellic acid mediated signaling pathway
heat acclimation
hyperosmotic salinity response
jasmonic acid biosynthetic process
jasmonic acid mediated signaling pathway
jasmonic acid metabolic process
membrane
oxygen binding
protein disulfide isomerase activity
protein folding
response to abscisic acid stimulus
response to auxin stimulus
response to bacterium
response to cold
response to desiccation
response to endoplasmic reticulum stress
response to ethylene stimulus
response to fungus
response to gibberellin stimulus
response to heat
response to high light intensity
response to hydrogen peroxide
response to jasmonic acid stimulus
response to karrikin
response to osmotic stress
response to salt stress
response to symbiotic fungus
response to water deprivation
response to wounding
sequence-specific DNA binding transcription factor activity
signal transduction
transport

Table 4.7: Gene Ontology terms significantly enriched in genes differentially expressed after treatment with one hour of hot excess light in all dynamic growth conditions and static growth conditions. These terms are those preserved across all conditions, and describe a commonality of response involving genes known to be involved in abiotic stress response

conditions compared to sufficient and static growth conditions (ANOVA, F=18.51, 3 DF, $p<0.001$) and after hot excess light treatment (ANOVA, F=7.93, 1 DF, p=0.012). Similar patterns can be qualitatively observed in ELIP2, however high variance prevents statistical significance ($p>0.05$). These patterns of differential expression are summarised in Figure 4.9.

4.4 Summary of Findings

In this chapter I present findings obtained from a preliminary dataset examining the transcriptional response to altered light intensity within the framework of dynamic growth conditions. An RNAseq dataset characterised by high levels of biological noise was created; within this dataset differential expression was observed. Upon treatment of plants grown under sufficient, excess or fluctuating light dynamic growth conditions with one hour of hot excess light, differential regulation of gene classes involved in abiotic stress response was observed. Comparisons between sufficient, excess and fluctuating light dynamic growth conditions elucidated limited differential expression. Gene classes including genes involved in translation, biotic and abiotic stress response and metabolism were significantly enriched in genes differentially expressed between these dynamic growth conditions.

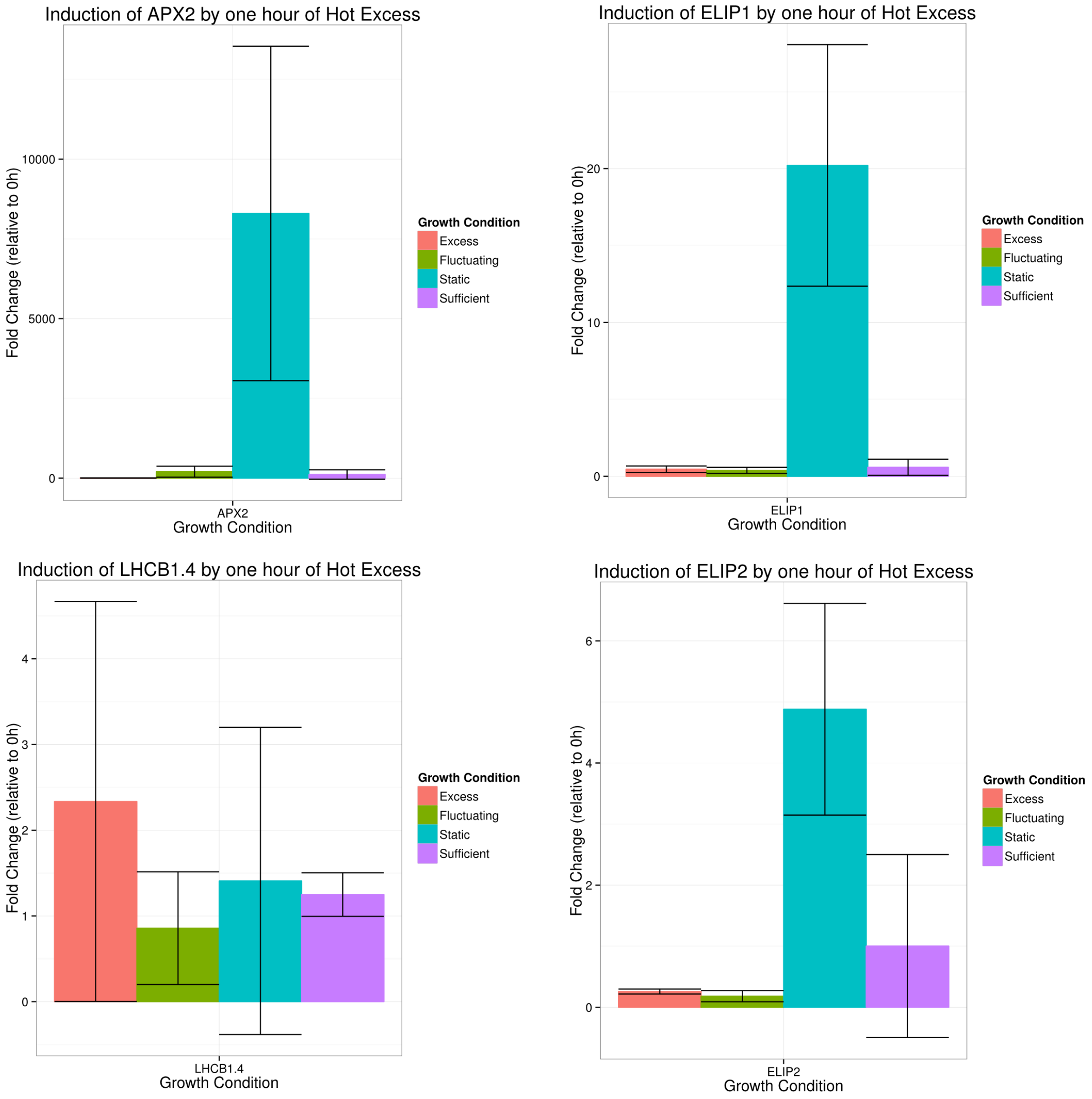

Figure 4.9: Differential induction of excess light marker genes by one hour of hot excess light. Statistically significant differential induction between static and dynamic growth conditions is observed for *APX2* and *ELIP1*; similar non-significant patterns are observed for *ELIP2*. No significant or qualitative difference in LHCB repression by hot excess light is observed.

Chapter 5

Discussion

The study of plant-environment interactions is crucial to improving agricultural yield and understanding functional plant ecology in a changing climate. The study of mechanisms by which plants tolerate or respond to abiotic stresses is of particular importance, as abiotic stresses cause many hundreds of billions of dollars of crop losses and incalculable environmental damage (Mittler 2006). Several recent reviews of abiotic stress research have suggested a changed focus for research, away from studying stresses individually towards more holistic combined stress studies (Mittler 2006; Mittler and Blumwald 2010). Plants in ecological or agricultural settings commonly experience environments unfavourable to optimal growth, composed of multiple abiotic (and biotic) stresses (Boyer 1982; Mittler 2006). Moreover, natural environments are dynamic, and challenge plants to rapidly adapt to heterogeneous environmental conditions on a daily basis. Laboratory study of these stresses should therefore account for recent findings suggesting that interactions between abiotic stresses in field conditions are not simply additive, rather are influenced by the dynamic combinations of stresses their environments impose upon them (Mittler 2006; Atkinson and Urwin 2012).

This thesis has developed methods to reproducibly induce abiotic stresses in laboratories in combinations analogous to those in which they can appear naturally and assay plant growth under such conditions, analyse plant response to these conditions, and presents a preliminary proof-of-concept transcriptomics dataset analysing light response in the context of combinatorial abiotic stresses.

5.1 Novel Dynamic Growth Conditions

Plant biologists often conduct laboratory experiments in controlled growth facilities. These facilities may lack the ability to mimic characteristics of their natural environment, as they are rarely built from hardware enabling an investigator to do so. While static growth conditions used in laboratories are sufficient to uncover core mechanisms of plant responses to stress, this is not always the case. In fact, evidence from field studies has uncovered phenotypes that are cryptic under static laboratory conditions, and that may be observed and studied closely under dynamic laboratory conditions (Külheim, Ågren, and Jansson 2002; Mittler 2006; Mishra et al. 2012; Wituszyńska et al. 2013). By creating software which connects existing but disparate technologies, I have enabled the implementation of growth conditions which can mimic regional climates or weather. The combination of computer controllable growth cabinets fitted with multi-spectral LED arrays allow for experimentation to consider important parameters that vary in natural regional climates, including temperature, humidity, and light quality and quantity.

In creating dynamic growth conditions, I aimed to combine some elements of the climates experienced by plants in nature with the reliability, reproducibility and convenience of laboratory growth conditions. Diurnal trends in light intensity, temperature and humidity follow those observed in the recent past at the model location (Bureau of Meteorology 2013). However, daily minima in temperature and maxima in humidity are tempered by hardware limitations. Similarly, daily integrated light intensity is lower due to limitations in the brightness of LED arrays used in the SpectralPhenoClimatron. The light produced by LED arrays does not cover the photosynthetically active portion of the visible and adjacent spectrum with an even intensity of light per wavelength as sunlight does, although LED arrays do so with broader peaks of increased spectral intensity than fluorescent or incandescent lamps. These are important limitations to the SpectralPhenoClimatron, and addressing them in future work is a priority. Circannual variation in temperature and humidity follow similar climactic trends as the historical mean observations of weather at the trial location, with temperature slowly increasing, humidity decreasing and photoperiod lengthening, however light intensity does not gradually increase in the same manner as the observed climate.

While the dynamic growth conditions implemented in the SpectralPhenoClimatron mimic elements of natural environments, they are not designed to be accurate representations of conditions encountered in nature. In nature, weather provides a layer of stochasticity upon the broad trends in climate. However, the developments I have described in this thesis facilitate the emulation of stressful growth environments which approximate weather-induced abiotic stresses plants experience in the natural environments. The system I have developed is not only applicable to studies discussed in this thesis, and has been used in experiments beyond the scope of this thesis including virtual reciprocal transplants and gene-by-environment interaction QTL mapping. The study of standing genetic variation within regional climates allows links between genotype and reaction to some environmental parameter, and can shed light on mechanistic links via methods including GWAS (Li et al. 2006; Li et al. 2010; Brachi, Morris, and Borevitz 2011). However, when modelling plant growth in, or reaction to, environmental conditions, the system I have developed allows for non-stochastic weather to be imposed over the climatic trends (e.g. the fluctuations in light modelling intermittent cloud cover used in this thesis).

5.2 Improved analytic methods for RNAseq

RNAseq is a precise method to quantify genome-wide expression (Wang, Gerstein, and Snyder 2009). RNAseq can reveal hidden phenotypes and subtle environmental effects of a plant's growth environment (Martin et al. 2013), giving insights into development, regulatory mechanisms, signalling pathways, acclimation and many other aspects of plant biology. It is sufficiently sensitive to measure the subtle differences in expression between both closely genetically related organisms (à la expression QTL mapping, Sun and Hu (2013)), and subtle environmental effects (e.g. differential response to dynamic growth conditions).

Reproducibility, accuracy, and performance of the computational analysis of any dataset is crucial. Inaccurate and poorly reproducible analyses have lead to embarrassing errors and retractions in many fields (Peng 2009; Herndon, Ash, and Pollin 2013). Computational performance of analysis software is central to interactive analysis of datasets;

enabling fast analysis of datasets allows researchers to explore their data without the requirement for expensive clusters or supercomputers. The design and implementation of the RNAseq analysis framework I have created specifically address reproducibility, accuracy, and performance, by selecting the latest advances in high-performance, accurate analysis software, and providing a simple structure to an analysis that allows exact reproduction of the entire analysis.

The specific algorithms used in analysis of RNAseq datasets is a field of active technical research. Validation of pipelines designed around the Subread aligner (Liao, Smyth, and Shi 2013b) reiterate the software authors' claims of increased speed and accuracy. A recent publication suggests superior performance of the trimmed mean of M values normalisation method proposed by Robinson and Oshlack (2010) and implemented by Robinson, McCarthy, and Smyth (2010) compared to its competitors (Rapaport et al. 2013). However, this review did not consider recently published statistical techniques such as QLspline (Lund et al. 2012). Further investigation of the state of the art in RNAseq statistical and computational analysis software is of great importance to studies of differential expression, and the framework presented in this thesis is specifically designed for modularity, enabling substitution of components for improved versions or alternatives with minimal effort.

When transcriptome variation within a study system is subtle, sensitive and accurate methods are required to glean information. This includes accurate analysis software (see above) and optimal experimental design. I examined the effect of sequencing depth on power to detect differential expression, and found that below 5 million reads per sample, power to detect differential expression with three replicates diminishes rapidly (see Figure 3.4). Several recent studies and reviews suggest that the current informal standard of RNAseq experimental design, which emphasises sample sequencing coverage over replication, is unwise. Rapaport et al. (2013) find increasing the level of replication over sequencing depth yields more differentially expressed genes, while Kliebenstein (2012) demonstrate that, even when sequence coverage was very low, nearly all expression QTLs could be mapped. In light of these data, my analysis of sequencing depth is conservative, as it does not consider the increases in replication made possible by increasing the number of samples sequenced per unit cost by a factor of two or more.

5.3 Elucidating Response to Light Intensity Under Dynamic Growth Conditions

The transcriptomes of *Arabidopsis* grown under dynamic growth conditions have been observed. A preliminary RNAseq dataset characterised by high levels of biological noise indicated differential expression of hundreds of transcripts between growth conditions and in response to treatment with hot excess light. QPCR analysis showed increased steady-state expression of know stress responsive genes under dynamic growth conditions, and reduced fold-change induction of stress genes upon application of hot excess light. Together, these data indicate patterns of differential expression observed under dynamic growth conditions, and provide limited and tentative support for the hypothesised hardening of plants to excess light.

The large amount of biological noise and resulting lack of power to detect differential expression is likely caused by a combination of factors. A rapid (within minutes) reduction in transcript abundance of hot excess light induced genes upon removal from hot excess light treatment has been observed (pers. comm. Peter Crisp). In light of these data, the high temporal error in sampling which exists in the dataset that I created may explain some of the biological variability. Additionally, as plants were of an advanced age (5 weeks) when samples were taken rapid estimation of the particular leaf that was sampled was infeasible, thus the largest expanded leaf was taken. This variation in leaf number may be an additional reason for the high variability in this dataset, as the transcriptome is dependent on leaf developmental stage (Gordon et al. 2013; Carmody 2013). A final possible explanation for the high biological noise is variation in harvesting time. While plants were harvested as quickly as possible, each replicate took 3 hours to harvest. Previous studies have found circadian effects to both the general transcriptome (Covington et al. 2008; Ptitsyn 2008) and in transcriptional response to biotic stress (Wilkins, Bräutigam, and Campbell 2010). The harvesting techniques and experimental design utilised in this study were suited to experiments that mapped eQTLs for stress-responsive genes. For simple detection of differential expression however, they were not optimal and likely contributed to the high level of biological noise in the obtained dataset. Ideally, true internal controls grown under static growth conditions, and higher levels of biological replication would be more appropriate. These shortcomings result in reduced statistical power, and are a possible cause for the lower than expected number of genes whose expression was induced or repressed upon treatment with hot excess light.

Given this high level of biological noise, all further analyses of differential expression should be approached with caution, even if statistical techniques are sufficiently advanced to allow detection of differential expression. Despite this, patterns of expression similar to those observed by other authors have been noted in this study. Hundreds of genes were differentially expressed in response to one hour of excess light in each dynamic light condition, a similar magnitude to previous examinations of similar stresses (Rossel, Wilson, and Pogson 2002; Rossel et al. 2007; Gordon et al. 2013; Kimura et al. 2003). Specifically, the induction of heat shock family proteins observed by Rossel, Wilson, and Pogson (2002) was also observed here. Additionally, gene ontology (GO) term analysis indicates commonality of the transcriptional response to hot excess light with heat and oxidative stresses, as observed in previous studies (Rossel, Wilson, and Pogson 2002). Together, these data suggest that the response of plants grown under dynamic light conditions to hot excess light is similar to that of plants growth under static light conditions, albeit with somewhat tempered induction of hot excess light-responsive marker genes. An understanding of the statistical power needed and variation present in an experiment examining dynamic growth conditions has been identified.

Limited transcriptomic evidence of the effect of light intensity within the framework of dynamic light conditions was observed. Gene ontology (GO) term enrichment provides evidence that plant response to hot excess light is preserved in plants grown under dynamic growth conditions hypothesised to induce acclimation to excess light or fluctuations in light intensity (Figure 4.8). Additionally, GO term enrichment analysis details the overlap and interaction with other abiotic stresses. GO terms enriched in genes with differential steady-state expression in plants grown under sufficient light dynamic growth conditions and excess or fluctuating light dynamic growth conditions exhibit limited overlap with those observed in previous studies of acute changes in light intensity, as hypothesised. However, this evidence of hardening is contrasted by the number of genes differentially expressed in response to hot excess light under dynamic growth conditions. If acclimation to modest excesses in light intensity result in reduced fold-change induction of stress-inducible genes, as qPCR quantification of APX2 expression suggest, then either the relative order of hardening hypothesised is incorrect, or the level of biological noise in this dataset prevented observation of such phenomena.

A more detailed analysis of this dataset may be warranted in light of these findings. In particular, examination of possible harvesting or treatment block effects may, if any effect exists, reduce the biological coefficient of variation (Robinson et al. 2013). A detailed analysis of all genes called as differentially expressed may be of use, as high biological noise may not affect the estimates of differential expression in all genes. I hypothesise that upon doing so, genes previously found to be differentially expressed between lab and field growth conditions would have differential steady-state expression between sufficient, fluctuating and excess light conditions, with expression being highest in the excess and lowest in sufficient light conditions.

Study of the transcriptome is an established method to assess the responses of plants to their environment, but it is far from the only method of doing so. Analysis of chlorophyll fluorescence of plants acclimating to fluctuating light demonstrated up-regulation of NPQ (Alter et al. 2012; Gordon et al. 2013), and study of such characteristics under dynamic light conditions is ongoing, but beyond the scope of this thesis. Analysis of accumulation of stress metabolites has elucidated metabolites involved in acclimation to combined stresses in field conditions (Jänkänpää et al. 2012), warranting similar analysis of metabolites in plants grown under dynamic growth conditions. However, transcriptomics remains a sensitive, accurate and useful tool to measure plants' responses to their environment, particularly in a systems biology approach that integrates metabolomic, phenomic, transcriptomic and physiological responses to applied stresses.

5.4 Interpreting Combined Abiotic Stress Responses

While the effect of altered light intensity on plants and their transcriptomes has been studied in depth, few authors have studied under laboratory conditions analogous to those to which plants are adapted. Mutants in pathways important in survival or fecundity of plants in ecological or agricultural settings may show little or no detrimental phenotype under benign laboratory conditions (Külheim, Ågren, and Jansson 2002). This further underlies the importance of examining the physiological, metabolomic and transcriptional responses of plants to field-like combinations of stresses (Jänkänpää et al. 2012; Mishra et al. 2012; Wituszyńska et al. 2013). By examining the responses of plants to their environment in conditions similar to those which they have evolved, and studying stress response to field-like combinations of stresses, it may be possible to obtain a more direct picture of the role of parts of the genome with little function in the artificially benign conditions under which many laboratory experiments on plants are conducted.

Studying the effect of hardening on the transcriptional response of plants to stress may reveal mechanistic insights of response genes. As plants acclimate to altered growth conditions, alteration of steady-state transcription often occurs (Hihara et al. 2001; Alter et al. 2012; Heinrich et al. 2012). Transcriptome profiling during acclimation to cold stress, (Fowler and Thomashow 2002; Chawade et al. 2007), excess light (Gordon et al. 2013; Page et al. 2012), and drought (Ding, Fromm, and Avramova 2012) reveals a transcriptional response to long-term stress exposure. The elucidation of genes underlying such acclimatory responses would yield insights into mechanisms by which plant tolerance of chronic abiotic stress could be improved.

Scientists have repeatedly developed crop lines tolerant to stress assays “in the lab” that, when trialled under agricultural conditions, either do not show stress tolerance or have increased susceptibility to other stresses or combinations of stresses (Mittler 2006; Mittler and Blumwald 2010; Atkinson and Urwin 2012; Wituszyńska et al. 2013). Some of the difficulty in translation of stress tolerance from lab to field can be explained by detrimental interaction of stresses (Mittler and Blumwald 2010). By performing genetic screens or genome wide association studies for stress tolerance under dynamic growth conditions, which mimic field-like combinations of stresses, it may be possible to elucidate mechanisms of tolerance to specific stresses which do not conflict with mechanisms of tolerance to other abiotic stresses, as tolerance is assayed in a mildly stressful environment that mimics that encountered under agricultural conditions.

5.5 Future Directions

The emergence of transcriptional patterns despite a dataset with high noise warrants further investigation. Utilising recent research on the underlying biology of transcriptional response (Peter Crisp, unpublished data) and RNAseq molecular methodology (Kumar et al. 2012) and experimental design (Rapaport et al. 2013), a repeated experiment which incorporates faster sampling and more accurate application of hot excess light stress, along with many more biological replicates (5-8 replicates) per condition and the addition of static growth condition controls may further elucidate the patterns of differential expression under dynamic growth conditions, and determine the effect of light intensity on the transcriptome under dynamic light conditions.

Expression QTL mapping presents an opportunity to simultaneously discover novel transcriptional patterns, and their underlying regulatory mechanisms (Keurentjes et al. 2007). Samples to perform a limited eQTL mapping experiment were collected but not processed due to constraints on time and statistical power. Once genetic material suitable for such analyses has been obtained, mapping of eQTLs controlling differential expression of would elucidate not only response mechanisms, but their underlying regulators. This information could be used by crop improvement programs to select for lines with improved ability to respond to or tolerate field-like combinations of stresses.

Application of real-time phenomic analysis to plants grown under dynamic growth conditions, combined with QTL mapping or GWAS analyses, may provide novel insights into physiological responses to combinations of abiotic stresses. Phenomic traits that correlate with transcriptional or physiological response to or tolerance of combinations of abiotic stresses experienced under dynamic growth conditions should be selected. Such traits may include chlorophyll fluorescence measurements of redox state or NPQ (Alter et al. 2012; Külheim, Ågren, and Jansson 2002), or visual estimation of pigmentation caused by anthocyanin accumulation in response to photooxidative stress. GWAS and QTL mapping for these non-destructive measures could be conducted concomitantly with mapping of eQTLs discussed above. Such experiments would present integrative results which give holistic insight to mechanisms of tolerance of or response to combinations of abiotic stresses.

5.6 Conclusions

This thesis has developed several methods to study abiotic stresses in the framework of dynamic conditions, and presents a preliminary dataset which examines plant response to such conditions. Aiming to design and implement dynamic growth conditions that mimic regional climates, within this thesis I have presented both software and protocols to implement such conditions, and have created conditions to test the effect of light intensity in combination with abiotic stresses. In aiming to select optimal software for the high-throughput study of transcriptome dynamics using High-throughput Sequencing, and implement a framework for generation of analysis pipelines to do so, I have presented a comprehensive framework for the creation of pipelines applicable to experiments beyond those discussed in this thesis. Additionally, I have, using this framework, implemented pipelines to analyse generic RNAseq datasets. Finally, aiming to determine the transcriptional response of *Arabidopsis thaliana* to the combinatorial application of abiotic stresses using dynamic growth conditions, I created a preliminary RNAseq dataset, and samples for preliminary eQTL and phenomic QTL mapping datasets. This RNAseq dataset exhibited high biological noise, preventing reliable detection of differential expression on similar scales to analogous published experiments, however yielded insight into patterns of differential expression that warrant further investigation. This work lays the foundation for further work elucidating the response to and tolerance of field-like combinations of abiotic stresses in a manner compatible with that suggested by recent reviews (Mittler 2006; Atkinson and Urwin 2012), and for the further discovery of the genetic architecture underlying any tolerance or response discovered.

Bibliography

Adamska, I (1997). ELIPs – Light-induced stress proteins. *Physiologia Plantarum* 100.4, pp. 794–805. DOI: 10.1111/j.1399-3054.1997.tb00006.x (cit. on p. 9).

Alter, P, Dreissen, A, Luo, FL, and Matsubara, S (2012). Acclimatory responses of Arabidopsis to fluctuating light environment: comparison of different sunfleck regimes and accessions. *Photosynthesis Research* 113.1-3. PMID: 22729524 PMCID: PMC3430843, pp. 221–237. DOI: 10.1007/s11120-012-9757-2 (cit. on pp. 6, 42, 76–78).

Andrews, S (2012). *FastQC A Quality Control tool for High Throughput Sequence Data* (cit. on pp. 32, 35, 57).

Apel, K and Hirt, H (2004). Reactive oxygen species: Metabolism, oxidative stress, and signal transduction. *Annual Review of Plant Biology* 55, pp. 373–399 (cit. on p. 6).

Armstrong, AF, Wardlaw, KD, and Atkin, OK (2007). Assessing the relationship between respiratory acclimation to the cold and photosystem II redox poise in Arabidopsis thaliana. *Plant, Cell & Environment* 30.12, pp. 1513–1522. DOI: 10.1111/j.1365-3040.2007.01738.x (cit. on pp. 6, 11).

Asada, K (2006). Production and Scavenging of Reactive Oxygen Species in Chloroplasts and Their Functions. *Plant Physiology* 141.2, pp. 391–396. DOI: 10.1104/pp.106.082040 (cit. on pp. 6, 7).

Atkin, OK and Tjoelker, MG (2003). Thermal acclimation and the dynamic response of plant respiration to temperature. *Trends in Plant Science* 8.7, pp. 343–351. DOI: 10.1016/S1360-1385(03)00136-5 (cit. on p. 6).

Atkinson, NJ, Lilley, CJ, and Urwin, PE (2013). Identification of Genes Involved in the Response of Arabidopsis to Simultaneous Biotic and Abiotic Stresses. *Plant Physiology* 162.4. PMID: 23800991, pp. 2028–2041. DOI: 10.1104/pp.113.222372 (cit. on pp. 6, 11, 41).

Atkinson, NJ and Urwin, PE (2012). The interaction of plant biotic and abiotic stresses: from genes to the field. *Journal of Experimental Botany* 63.10. PMID: 22467407, pp. 3523–3543. DOI: 10.1093/jxb/ers100 (cit. on pp. 10, 11, 41, 70, 77, 79).

Avraham, S, Tung, CW, Ilic, K, Jaiswal, P, Kellogg, EA, McCouch, S, Pujar, A, Reiser, L, Rhee, SY, Sachs, MM, Schaeffer, M, Stein, L, Stevens, P, Vincent, L, Zapata, F, and Ware, D (2008). The Plant Ontology Database: a community resource for plant structure and developmental stages controlled vocabulary and annotations. *Nucleic Acids Research* 36.suppl 1. PMID: 18194960, pp. D449–D454. DOI: 10.1093/nar/gkm908 (cit. on p. 13).

Babu, S and Gassmann, M (2011). *Assessing integrity of plant RNA with the Agilent 2100 Bioanalyzer* (cit. on pp. 52, 53).

Baker, NR (2008). Chlorophyll Fluorescence: A Probe of Photosynthesis In Vivo. *Annual Review of Plant Biology* 59.1. PMID: 18444897, pp. 89–113. DOI: 10.1146/annurev.arplant.59.032607.092759 (cit. on p. 6).

Barua, D and Heckathorn, SA (2006). The interactive effects of light and temperature on heat-shock protein accumulation in Solidago altissima (Asteraceae) in the field and laboratory. *American Journal of Botany* 93.1, pp. 102–109. DOI: 10.3732/ajb.93.1.102 (cit. on p. 11).

Berardini, TZ, Mundodi, S, Reiser, L, Huala, E, Garcia-Hernandez, M, Zhang, P, Mueller, LA, Yoon, J, Doyle, A, Lander, G, Moseyko, N, Yoo, D, Xu, I, Zoeckler, B, Montoya, M, Miller, N, Weems, D, and Rhee, SY (2004). Functional Annotation of the Arabidopsis Genome Using Controlled Vocabularies. *Plant Physiology* 135.2. PMID: 15173566, pp. 745–755. DOI: 10.1104/pp.104.040071 (cit. on p. 13).

Boyer, JS (1982). Plant Productivity and Environment. *Science* 218.4571. PMID: 17808529, pp. 443–448. DOI: 10.1126/science.218.4571.443 (cit. on p. 70).

Brachi, B, Morris, GP, and Borevitz, JO (2011). Genome-wide association studies in plants: the missing heritability is in the field. *Genome biology* 12.10. PMID: 22035733, p. 232. DOI: 10.1186/gb-2011-12-10-232 (cit. on p. 72).

Buffalo, V (2013). *Scythe - A Bayesian adapter trimmer* (cit. on pp. 32, 35).

Bureau of Meteorology (2013). *Climate statistics for Australian locations* (cit. on p. 71).

Carmody, ME (2013). Rapid leaf-to-leaf communication of high light stress in Arabidiosis. PhD thesis (cit. on p. 74).

Chawade, A, Bräutigam, M, Lindlöf, A, Olsson, O, and Olsson, B (2007). Putative cold acclimation pathways in Arabidopsis thaliana identified by a combined analysis of mRNA co-expression patterns, promoter motifs and transcription factors. *BMC Genomics* 8.1. PMID: 17764576, p. 304. DOI: 10.1186/1471-2164-8-304 (cit. on p. 77).

Covington, MF, Maloof, JN, Straume, M, Kay, SA, and Harmer, SL (2008). Global transcriptome analysis reveals circadian regulation of key pathways in plant growth and development. *Genome Biology* 9.8. PMID: 18710561, R130. DOI: 10.1186/gb-2008-9-8-r130 (cit. on p. 74).

Demmig-Adams, B and Adams, WW (1992). Photoprotection and Other Responses of Plants to High Light Stress. *Annual Review of Plant Physiology and Plant Molecular Biology* 43.1, pp. 599–626. DOI: 10.1146/annurev.pp.43.060192.003123 (cit. on pp. 6, 7).

Ding, Y, Fromm, M, and Avramova, Z (2012). Multiple exposures to drought 'train' transcriptional responses in Arabidopsis. *Nature Communications* 3, p. 740. DOI: 10.1038/ncomms1732 (cit. on p. 77).

Fowler, S and Thomashow, MF (2002). Arabidopsis Transcriptome Profiling Indicates That Multiple Regulatory Pathways Are Activated during Cold Acclimation in Addition to the CBF Cold Response Pathway. *The Plant Cell Online* 14.8. PMID: 12172015, pp. 1675–1690. DOI: 10.1105/tpc.003483 (cit. on p. 77).

Foyer, CH and Noctor, G (2009). Redox Regulation in Photosynthetic Organisms: Signaling, Acclimation, and Practical Implications. *Antioxidants & Redox Signaling* 11.4, pp. 861–905 (cit. on p. 6).

Glenn, TC (2011). Field guide to next-generation DNA sequencers. *Molecular Ecology Resources* 11.5, pp. 759–769. DOI: 10.1111/j.1755-0998.2011.03024.x (cit. on pp. 29, 39).

Gordon, MJ, Carmody, M, Albrecht, V, and Pogson, B (2013). Systemic and local responses to repeated HL stress-induced retrograde signaling in Arabidopsis. *Frontiers* *in Plant Physiology* 3, p. 303. DOI: 10.3389/fpls.2012.00303 (cit. on pp. 9, 42, 74–77).

Heinrich, S, Valentin, K, Frickenhaus, S, John, U, and Wiencke, C (2012). Transcriptomic Analysis of Acclimation to Temperature and Light Stress in Saccharina latissima (Phaeophyceae). *PLoS ONE* 7.8, e44342. DOI: 10.1371/journal.pone.0044342 (cit. on p. 77).

Herndon, T, Ash, M, and Pollin, R (2013). *Does High Public Debt Consistently Stifle Economic Growth? A Critique of Reinhart and Rogoff* (cit. on p. 72).

Hihara, Y, Kamei, A, Kanehisa, M, Kaplan, A, and Ikeuchi, M (2001). DNA Microarray Analysis of Cyanobacterial Gene Expression during Acclimation to High Light. *The Plant Cell Online* 13.4, pp. 793–806. DOI: 10.1105/tpc.13.4.793 (cit. on pp. 9, 77).

Illumina (2012). *TruSeq RNA Sample Preparation V2 Kit Guide'*. Illumina, Inc (cit. on p. 54).

Jänkänpää, HJ, Mishra, Y, Schröder, WP, and Jansson, S (2012). Metabolic profiling reveals metabolic shifts in Arabidopsis plants grown under different light conditions. *Plant, Cell & Environment* 35.10, pp. 1824–1836. DOI: 10.1111/j.1365-3040.2012.02519.x (cit. on pp. 10, 76, 77).

Johnson, MP, Goral, TK, Duffy, CDP, Brain, APR, Mullineaux, CW, and Ruban, AV (2011). Photoprotective Energy Dissipation Involves the Reorganization of Photosystem II Light-Harvesting Complexes in the Grana Membranes of Spinach Chloroplasts. *The Plant Cell Online* 23.4, pp. 1468–1479. DOI: 10.1105/tpc.110.081646 (cit. on p. 8).

Johnston, DT, Wolfe-Simon, F, Pearson, A, and Knoll, AH (2009). Anoxygenic photosynthesis modulated Proterozoic oxygen and sustained Earth's middle age. *Proceedings of the National Academy of Sciences* 106.40. PMID: 19805080, pp. 16925–16929. DOI: 10.1073/pnas.0909248106 (cit. on p. 5).

Jung, HS, Crisp, PA, Estavillo, GM, Cole, B, Hong, F, Mockler, TC, Pogson, BJ, and Chory, J (2013). Subset of heat-shock transcription factors required for the early response of Arabidopsis to excess light. *Proceedings of the National Academy of Sciences*

110.35. PMID: 23918368, pp. 14474–14479. DOI: 10.1073/pnas.1311632110 (cit. on pp. 7, 11).

Karpiński, S, Reynolds, H, Karpińska, B, Wingsle, G, Creissen, G, and Mullineaux, P (1999). Systemic Signaling and Acclimation in Response to Excess Excitation Energy in Arabidopsis. *Science* 284.5414. PMID: 10213690, pp. 654–657. DOI: 10.1126/science.284.5414.654 (cit. on p. 9).

Kasahara, M, Kagawa, T, Oikawa, K, Suetsugu, N, Miyao, M, and Wada, M (2002). Chloroplast avoidance movement reduces photodamage in plants. *Nature* 420.6917, pp. 829–832. DOI: 10.1038/nature01213 (cit. on p. 7).

Keurentjes, JJB, Fu, J, Terpstra, IR, Garcia, JM, Ackerveken, Gvd, Snoek, LB, Peeters, AJM, Vreugdenhil, D, Koornneef, M, and Jansen, RC (2007). Regulatory network construction in Arabidopsis by using genome-wide gene expression quantitative trait loci. *Proceedings of the National Academy of Sciences* 104.5, pp. 1708–1713. DOI: 10.1073/pnas.0610429104 (cit. on p. 78).

Kim, D, Pertea, G, Trapnell, C, Pimentel, H, Kelley, R, and Salzberg, SL (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology* 14.4. PMID: 23618408, R36. DOI: 10.1186/gb-2013-14-4-r36 (cit. on pp. 32, 35).

Kim, JH (2012). Biological Knowledge Assembly and Interpretation. *PLoS Comput Biol* 8.12, e1002858. DOI: 10.1371/journal.pcbi.1002858 (cit. on p. 13).

Kimura, M, Yamamoto, YY, Seki, M, Sakurai, T, Sato, M, Abe, T, Yoshida, S, Manabe, K, Shinozaki, K, and Matsui, M (2003). Identification of Arabidopsis Genes Regulated by High Light–Stress Using cDNA Microarray. *Photochemistry and Photobiology* 77.2, pp. 226–233. DOI: 10.1562/0031-8655(2003)0770226IOAGRB2.0.CO2 (cit. on pp. 6, 75).

Kleine, T, Kindgren, P, Benedict, C, Hendrickson, L, and Strand, Å (2007). Genome-Wide Gene Expression Analysis Reveals a Critical Role for CRYPTOCHROME1 in the Response of Arabidopsis to High Irradiance. *Plant Physiology* 144.3, pp. 1391–1406. DOI: 10.1104/pp.107.098293 (cit. on p. 9).

Kliebenstein, DJ (2012). Exploring the Shallow End; Estimating Information Content in Transcriptomics Studies. *Frontiers in Plant Science* 3. PMID: 22973290 PMCID: PMC3437520. DOI: 10.3389/fpls.2012.00213 (cit. on p. 73).

Külheim, C, Ågren, J, and Jansson, S (2002). Rapid Regulation of Light Harvesting and Plant Fitness in the Field. *Science* 297.5578, pp. 91–93. DOI: 10.1126/science.1072359 (cit. on pp. 6, 8, 10, 11, 42, 71, 77, 78).

Kumar, R, Ichihashi, Y, Kimura, S, Chitwood, DH, Headland, LR, Peng, J, Maloof, JN, and Sinha, NR (2012). A high-throughput method for Illumina RNA-Seq library preparation. *Frontiers in Plant Genetics and Genomics* 3, p. 202. DOI: 10.3389/fpls.2012.00202 (cit. on pp. 13, 29, 43, 78, 112, 113).

Li, H (2013). *seqtk - Toolkit for processing sequences in FASTA/Q formats* (cit. on p. 32).

Li, Y, Huang, Y, Bergelson, J, Nordborg, M, and Borevitz, JO (2010). Association mapping of local climate-sensitive quantitative trait loci in Arabidopsis thaliana. *Proceedings of the National Academy of Sciences* 107.49. PMID: 21078970, pp. 21199–21204. DOI: 10.1073/pnas.1007431107 (cit. on pp. 12, 26, 72).

Li, Y, Roycewicz, P, Smith, E, and Borevitz, JO (2006). Genetics of Local Adaptation in the Laboratory: Flowering Time Quantitative Trait Loci under Geographic and Seasonal Conditions in Arabidopsis. *PLoS ONE* 1.1, e105. DOI: 10.1371/journal.pone.0000105 (cit. on pp. 6, 12, 26, 72).

Li, Z, Wakao, S, Fischer, BB, and Niyogi, KK (2009). Sensing and Responding to Excess Light. *Annual Review of Plant Biology* 60.1. PMID: 19575582, pp. 239–260. DOI: 10.1146/annurev.arplant.58.032806.103844 (cit. on pp. 6, 7).

Liao, Y, Smyth, GK, and Shi, W (2013a). *featureCounts: an efficient general-purpose read summarization program.* arXiv e-print 1305.3347 (cit. on p. 32).

— (2013b). The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Research* 41.10. PMID: 23558742, e108–e108. DOI: 10.1093/nar/gkt214 (cit. on pp. 32, 37, 73).

Lister, R, Gregory, BD, and Ecker, JR (2009). Next is now: new technologies for sequencing of genomes, transcriptomes, and beyond. *Current Opinion in Plant Biology* 12.2, pp. 107–118. DOI: 10.1016/j.pbi.2008.11.004 (cit. on p. 12).

Lund, SP, Dan, N, McCarthy, DJ, and Smyth, GK (2012). Detecting Differential Expression in RNA-sequence Data Using Quasi-likelihood with Shrunken Dispersion Estimates. *Statistical Applications in Genetics and Molecular Biology* 11.5, pp. 1–44 (cit. on p. 73).

Martin, LBB, Fei, Z, Giovannoni, JJ, and Rose, JKC (2013). Catalyzing plant science research with RNA-seq. *Frontiers in Plant Systems Biology* 4, p. 66. DOI: 10.3389/fpls.2013.00066. (cit. on pp. 28, 72).

McCarthy, DJ, Chen, Y, and Smyth, GK (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research* 40.10, pp. 4288–4297. DOI: 10.1093/nar/gks042 (cit. on p. 32).

Miller, G and Mittler, R (2006). Could Heat Shock Transcription Factors Function as Hydrogen Peroxide Sensors in Plants? *Annals of Botany* 98.2. PMID: 16740587 PMCID: PMC2803459, pp. 279–288. DOI: 10.1093/aob/mcl107 (cit. on p. 9).

Mishra, Y, Jänkänpää, HJ, Kiss, AZ, Funk, C, Schröder, WP, and Jansson, S (2012). Arabidopsis plants grown in the field and climate chambers significantly differ in leaf morphology and photosystem components. *BMC Plant Biology* 12.1. PMID: 22236032, p. 6. DOI: 10.1186/1471-2229-12-6 (cit. on pp. 6, 9–11, 71, 77).

Mittler, R (2006). Abiotic stress, the field environment and stress combination. *Trends in Plant Science* 11.1, pp. 15–19. DOI: 10.1016/j.tplants.2005.11.002 (cit. on pp. 5, 6, 10, 11, 15, 26, 41, 70, 71, 77, 79).

Mittler, R and Blumwald, E (2010). Genetic Engineering for Modern Agriculture: Challenges and Perspectives. *Annual Review of Plant Biology* 61.1. PMID: 20192746, pp. 443–462. DOI: 10.1146/annurev-arplant-042809-112116 (cit. on pp. 5, 11, 15, 70, 77).

Mubarakshina, MM, Ivanov, BN, Naydov, IA, Hillier, W, Badger, MR, and Krieger-Liszkay, A (2010). Production and diffusion of chloroplastic H2O2 and its implication to signalling. *Journal of Experimental Botany* 61.13. PMID: 20595239, pp. 3577–3587. DOI: 10.1093/jxb/erq171 (cit. on p. 6).

Mühlenbock, P, Szechyńska-Hebda, M, Płaszczyca, M, Baudo, M, Mateo, A, Mullineaux, PM, Parker, JE, Karpińska, B, and Karpiński, S (2008). Chloroplast Signaling and

LESION SIMULATING DISEASE1 Regulate Crosstalk between Light Acclimation and Immunity in Arabidopsis. *The Plant Cell Online* 20.9. PMID: 18790826, pp. 2339–2356. DOI: 10.1105/tpc.108.059618 (cit. on p. 9).

Müller, P, Li, XP, and Niyogi, KK (2001). Non-Photochemical Quenching. A Response to Excess Light Energy. *Plant Physiology* 125.4, pp. 1558–1566. DOI: 10.1104/pp.125.4.1558 (cit. on p. 8).

Murchie, EH, Hubbart, S, Peng, S, and Horton, P (2005). Acclimation of photosynthesis to high irradiance in rice: gene expression and interactions with leaf development. *Journal of Experimental Botany* 56.411, pp. 449–460. DOI: 10.1093/jxb/eri100 (cit. on p. 9).

Nelson, N and Ben-Shem, A (2004). The complex architecture of oxygenic photosynthesis. *Nature Reviews Molecular Cell Biology* 5.12, pp. 971–982. DOI: 10.1038/nrm1525 (cit. on p. 8).

Niyogi, KK (1999). PHOTOPROTECTION REVISITED: Genetic and Molecular Approaches. *Annual Review of Plant Physiology and Plant Molecular Biology* 50.1. PMID: 15012213, pp. 333–359. DOI: 10.1146/annurev.arplant.50.1.333 (cit. on pp. 6, 7).

Nookaew, I, Papini, M, Pornputtapong, N, Scalcinati, G, Fagerberg, L, Uhlén, M, and Nielsen, J (2012). A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in Saccharomyces cerevisiae. *Nucleic Acids Research* 40.20, pp. 10084–10097. DOI: 10.1093/nar/gks804 (cit. on p. 13).

Page, M, Sultana, N, Paszkiewicz, K, Florance, H, and Smirnoff, N (2012). The influence of ascorbate on anthocyanin accumulation during high light acclimation in Arabidopsis thaliana: further evidence for redox control of anthocyanin synthesis. *Plant, Cell & Environment* 35.2, pp. 388–404. DOI: 10.1111/j.1365-3040.2011.02369.x (cit. on p. 77).

Peng, RD (2009). Reproducible research and Biostatistics. *Biostatistics* 10.3. PMID: 19535325, pp. 405–408. DOI: 10.1093/biostatistics/kxp014 (cit. on p. 72).

Ptitsyn, A (2008). Comprehensive analysis of circadian periodic pattern in plant transcriptome. *BMC Bioinformatics* 9.Suppl 9. PMID: 18793463, S18. DOI: 10.1186/1471-2105-9-S9-S18 (cit. on p. 74).

Rapaport, F, Khanin, R, Liang, Y, Pirun, M, Krek, A, Zumbo, P, Mason, CE, Socci, ND, and Betel, D (2013). Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biology* 14.9. PMID: 24020486, R95. DOI: 10.1186/gb-2013-14-9-r95 (cit. on pp. 73, 78).

Rizhsky, L, Liang, H, and Mittler, R (2002). The Combined Effect of Drought Stress and Heat Shock on Gene Expression in Tobacco. *Plant Physiology* 130.3. PMID: 12427981, pp. 1143–1151. DOI: 10.1104/pp.006858 (cit. on p. 11).

Rizhsky, L, Liang, H, Shuman, J, Shulaev, V, Davletova, S, and Mittler, R (2004). When Defense Pathways Collide. The Response of Arabidopsis to a Combination of Drought and Heat Stress. *Plant Physiology* 134.4. PMID: 15047901, pp. 1683–1696. DOI: 10.1104/pp.103.033431 (cit. on p. 11).

Robinson, MD, McCarthy, DJ, and Smyth, GK (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26.1. PMID: 19910308 PMCID: PMC2796818, pp. 139–140. DOI: 10.1093/bioinformatics/btp616 (cit. on pp. 13, 32, 73).

Robinson, MD and Oshlack, A (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology* 11.3. PMID: 20196867, R25. DOI: 10.1186/gb-2010-11-3-r25 (cit. on pp. 13, 36, 73).

Robinson, M, McCarthy, D, Chen, Y, and Smyth, GK (2013). *edgeR: differential expression analysisof digital gene expression data - User's Guide* (cit. on pp. 55, 76).

Rossel, JB, Wilson, IW, and Pogson, BJ (2002). Global Changes in Gene Expression in Response to High Light in Arabidopsis. *Plant Physiology* 130.3, pp. 1109–1120. DOI: 10.1104/pp.005595 (cit. on pp. 6, 7, 9, 11, 41, 75).

Rossel, JB, Wilson, PB, Hussain, D, Woo, NS, Gordon, MJ, Mewett, OP, Howell, KA, Whelan, J, Kazan, K, and Pogson, BJ (2007). Systemic and Intracellular Responses to Photooxidative Stress in Arabidopsis. *The Plant Cell Online* 19.12, pp. 4091–4110. DOI: 10.1105/tpc.106.045898 (cit. on p. 75).

Ruckle, ME, DeMarco, SM, and Larkin, RM (2007). Plastid Signals Remodel Light Signaling Networks and Are Essential for Efficient Chloroplast Biogenesis in Arabidopsis. *The Plant Cell Online* 19.12. PMID: 18065688, pp. 3944–3960. DOI: 10.1105/tpc.107.054312 (cit. on p. 64).

Ruijter, JM, Ramakers, C, Hoogaars, WMH, Karlen, Y, Bakker, O, Hoff, MJB van den, and Moorman, AFM (2009). Amplification efficiency: linking baseline and bias in the analysis of quantitative PCR data. *Nucleic acids research* 37.6. PMID: 19237396, e45. DOI: 10.1093/nar/gkp045 (cit. on p. 49).

Seki, M, Kamei, A, Yamaguchi-Shinozaki, K, and Shinozaki, K (2003). Molecular responses to drought, salinity and frost: common and different paths for plant protection. *Current Opinion in Biotechnology* 14.2, pp. 194–199. DOI: 10.1016/S0958-1669(03)00030-2 (cit. on pp. 6, 11).

Seki, M, Narusaka, M, Abe, H, Kasuga, M, Yamaguchi-Shinozaki, K, Carninci, P, Hayashizaki, Y, and Shinozaki, K (2001). Monitoring the Expression Pattern of 1300 Arabidopsis Genes under Drought and Cold Stresses by Using a Full-Length cDNA Microarray. *The Plant Cell Online* 13.1. PMID: 11158529, pp. 61–72. DOI: 10.1105/tpc.13.1.61 (cit. on pp. 6, 11, 41).

Spokas, K and Forcella, F (2006). Estimating hourly incoming solar radiation from limited meteorological data. *Weed Science* 54.1, pp. 182–189. DOI: 10.1614/WS-05-098R.1 (cit. on pp. 16, 19).

Subramanian, A, Tamayo, P, Mootha, VK, Mukherjee, S, Ebert, BL, Gillette, MA, Paulovich, A, Pomeroy, SL, Golub, TR, Lander, ES, and Mesirov, JP (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* 102.43. PMID: 16199517, pp. 15545–15550. DOI: 10.1073/pnas.0506580102 (cit. on p. 13).

Sun, W and Hu, Y (2013). eQTL Mapping Using RNA-seq Data. *Statistics in Biosciences* 5.1, pp. 198–219. DOI: 10.1007/s12561-012-9068-3 (cit. on p. 72).

Suzuki, N, Koussevitzky, S, Mittler, R, and Miller, G (2012). ROS and redox signalling in the response of plants to abiotic stress. *Plant, Cell & Environment* 35.2, pp. 259–270. DOI: 10.1111/j.1365-3040.2011.02336.x (cit. on p. 7).

Takahashi, S and Badger, MR (2011). Photoprotection in plants: a new light on photosystem II damage. *Trends in Plant Science* 16.1, pp. 53–60. DOI: 10.1016/j.tplants.2010.10.001 (cit. on p. 7).

Takahashi, S, Milward, SE, Fan, DY, Chow, WS, and Badger, MR (2009). How Does Cyclic Electron Flow Alleviate Photoinhibition in Arabidopsis? *Plant Physiology* 149.3, pp. 1560–1567. DOI: 10.1104/pp.108.134122 (cit. on p. 8).

Takahashi, S, Milward, SE, Yamori, W, Evans, JR, Hillier, W, and Badger, MR (2010). The Solar Action Spectrum of Photosystem II Damage. *Plant Physiology* 153.3, pp. 988–993. DOI: 10.1104/pp.110.155747 (cit. on p. 10).

Tikkanen, M, Grieco, M, Nurmi, M, Rantala, M, Suorsa, M, and Aro, EM (2012). Regulation of the photosynthetic apparatus under fluctuating growth light. *Philosophical Transactions of the Royal Society B: Biological Sciences* 367.1608, pp. 3486–3493. DOI: 10.1098/rstb.2012.0067 (cit. on p. 6).

Tikkanen, M, Piippo, M, Suorsa, M, Sirpiö, S, Mulo, P, Vainonen, J, Vener, AV, Allahverdiyeva, Y, and Aro, EM (2006). State transitions revisited—a buffering system for dynamic low light acclimation of Arabidopsis. *Plant Molecular Biology* 62.4-5, pp. 779–793. DOI: 10.1007/s11103-006-9044-8 (cit. on p. 8).

Van Verk, MC, Hickman, R, Pieterse, CM, and Van Wees, SC (2013). RNA-Seq: revelation of the messengers. *Trends in Plant Science* 18.4, pp. 175–179. DOI: 10.1016/j.tplants.2013.02.001 (cit. on pp. 13, 28, 30, 33).

Vanderauwera, S, Zimmermann, P, Rombauts, S, Vandenabeele, S, Langebartels, C, Gruissem, W, Inzé, D, and Van Breusegem, F (2005). Genome-Wide Analysis of Hydrogen Peroxide-Regulated Gene Expression in Arabidopsis Reveals a High Light-Induced Transcriptional Cluster Involved in Anthocyanin Biosynthesis. *Plant Physiology* 139.2, pp. 806–821. DOI: 10.1104/pp.105.065896 (cit. on p. 7).

Väremo, L, Nielsen, J, and Nookaew, I (2013). Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods. *Nucleic Acids Research*. DOI: 10.1093/nar/gkt111 (cit. on p. 13).

Wang, Y, Ghaffari, N, Johnson, CD, Braga-Neto, UM, Wang, H, Chen, R, and Zhou, H (2011). Evaluation of the coverage and depth of transcriptome by RNA-Seq in chickens. *BMC Bioinformatics* 12.Suppl 10. PMID: 22165852, S5. DOI: 10.1186/1471-2105-12-S10-S5 (cit. on p. 29).

Wang, Z, Gerstein, M, and Snyder, M (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* 10.1, pp. 57–63. DOI: 10.1038/nrg2484 (cit. on pp. 12, 13, 28, 29, 72).

Wilkins, O, Bräutigam, K, and Campbell, MM (2010). Time of day shapes Arabidopsis drought transcriptomes. *The Plant Journal* 63.5, pp. 715–727. DOI: 10.1111/j.1365-313X.2010.04274.x (cit. on p. 74).

Wituszyńska, W, Gałązka, K, Rusaczonek, A, Vanderauwera, S, Van Breusegem, F, and Karpiński, S (2013). Multivariable environmental conditions promote photosynthetic adaptation potential in Arabidopsis thaliana. *Journal of Plant Physiology* 170.6, pp. 548–559. DOI: 10.1016/j.jplph.2012.11.016 (cit. on pp. 6, 10, 11, 26, 41, 42, 71, 77).

Yi, X, Du, Z, and Su, Z (2013). PlantGSEA: a gene set enrichment analysis toolkit for plant community. *Nucleic Acids Research* 41.W1, W98–W103. DOI: 10.1093/nar/gkt281 (cit. on p. 13).

Young, MD, Wakefield, MJ, Smyth, GK, and Oshlack, A (2010). Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biology* 11.2. PMID: 20132535, R14. DOI: 10.1186/gb-2010-11-2-r14 (cit. on p. 32).

Zou, F, Gelfond, JAL, Airey, DC, Lu, L, Manly, KF, Williams, RW, and Threadgill, DW (2005). Quantitative Trait Locus Analysis Using Recombinant Inbred Intercrosses. *Genetics* 170.3. PMID: 15879512 PMCID: PMC1451174, pp. 1299–1311. DOI: 10.1534/genetics.104.035709 (cit. on p. 42).

Chapter 6

Appendix

Notes:

- Code listings, where included, are illustrative. Full source code of all software developed is large (over 5000 lines of code), and will be distributed as a gzipped tar archive. The latest code for all pipelines, scripts, is available online. See Appendix section 6.1 and Table 6.1

6.1 Source Code Repositories

The following source code repositories have been created as part of this thesis.

Repository	URL
`spcControl` module	`http://github.com/borevitzlab/spcControl`
RNAseq Pipeline	`https://github.com/kdmurray91/RNAseqPipeline`

Table 6.1: Source code repositories

6.2 Miscellaneous Software

The following pieces of software are not part of any software package, however are used within this thesis.

Most scripts are available within the `bioscripts` repository on my github site, available at `https://github.com/kdmurray91/bioscripts`.

6.2.1 `spliceSolarCalc.py`

Create SolarCalc model files which fluctuate between two model files on a regular cycle. Available at `https://github.com/kdmurray91/bioscripts/blob/master/solarcalc/spliceSolarCalc.py`.

6.2.2 Analysis of RNAseq analysis pipeline computation cost

Statistical analysis of RNAseq pipeline computation cost.

`https://github.com/kdmurray91/hons-thesis-stats/blob/master/pltimes/pltimes.Rmd`

6.2.3 qPCR analysis Code

Statistical calculation of relative quantification and differential expression in qPCR datasets.

Available at `https://github.com/kdmurray91/hons-thesis-stats/blob/master/qpcr/qpcr.Rnw`

6.3 spcControl Module Implementation Details

6.3.1 Evolution of the spcControl codebase

As bugs were discovered, the initial code to relay control commands to the SpectralPhenoClimatron was developed into a fully fledged python module, of which the main program loop is shown below.

Listing 3: Initial code to relay control commands to the SpectralPhenoClimatron

```
from __future__ import print_function
from telnetlib import Telnet
from time import strptime, sleep, mktime
import datetime
import csv
import argparse


CSV_FIELDS = {
      "Date": 0,
      "Time": 1,
      "Temperature": 2,
      "Humidity": 3,
      "Light 1": 4,
      "Light 2": 5,
      "Light 3": 6,
      "Light 4": 7,
      }
SET_COMMAND = "pcoset"
DEVICE_ID = "0"  # as string
DATATYPES = {
      "Temperature": "I",
      "Humidity": "I",
      # Need info from Conviron for this
```

```
    "Light 1": "I",
    "Light 2": "I",
    "Light 3": "I",
    "Light 4": "I",
    }
INDICIES = {
    "Temperature": 105,
    "Humidity": 106,
    # Need info from Conviron for this
    "Light 1": "107",
    "Light 2": "",
    "Light 3": "",
    "Light 4": "",
    }

STRP_FORMAT = "%m/%d/%Y %I:%M %p"

def get_args():
  parser = argparse.ArgumentParser(description="Daemon to update"
      " environmental conditions of Conviron cabinets in real time")
  parser.add_argument("-H", "--host", action="store", required=True,
      help="Host of the Conviron, to Telnet into", dest="host")
  parser.add_argument("-c", "--csv-file", action="store", required=True,
      help="The CSV file describing the environmental conditions",
      dest="csvfile")
  parser.add_argument("-u", "--user", action="store", default="root",
      help="The login username for the conviron", dest="user")
  parser.add_argument("-v", "--verbose", action="count", dest="verbosity",
      help="Verbosity level. The more -v's, the more verbose")
  parser.add_argument("-p", "--password", action="store", default="froot",
      help="The password of the user for the conviron", dest="passwd")
  return parser.parse_args()


def communicate_line(args, line):
  print("Communicating:", line)
  cmd_str = SET_COMMAND + " " + DEVICE_ID + " I "


  # Establish connection
  telnet = Telnet(args.host)
  response = telnet.read_until(b"login: ")
  if args.verbosity > 0:
    print(response)

  # Username
  payload = bytes(args.user + "\n", encoding="UTF8")
  telnet.write(payload)
  response = telnet.read_until(b"Password: ")
  if args.verbosity > 0:
    print(payload)
    print(response)

  # Password
  payload = bytes(args.passwd + "\n", encoding="UTF8")
  telnet.write(payload)
  response = telnet.read_until(b"#")
  if args.verbosity > 0:
    print(payload)
    print(response)
```

```
# PCOSET initialisation header
payload = bytes(cmd_str + " 100 26\n", encoding="UTF8")
telnet.write(payload)
response = telnet.read_until(b"#")
if args.verbosity > 0:
   print(payload)
   print(response)

payload = bytes(cmd_str + " 101 1\n", encoding="UTF8")
telnet.write(payload)
response = telnet.read_until(b"#")
if args.verbosity > 0:
   print(payload)
   print(response)

payload = bytes(cmd_str + " 102 1\n", encoding="UTF8")
telnet.write(payload)
response = telnet.read_until(b"#")
if args.verbosity > 0:
   print(payload)
   print(response)

# PCOSET send temperature
payload = bytes(cmd_str + " 105 243\n", encoding="UTF8")
telnet.write(payload)
response = telnet.read_until(b"#")
if args.verbosity > 0:
   print(payload)
   print(response)
```

```
# PCOSET send humidity
payload = bytes(cmd_str + " 106 66\n", encoding="UTF8")
telnet.write(payload)
response = telnet.read_until(b"#")
if args.verbosity > 0:
   print(payload)
   print(response)

# PCOSET send light 1
payload = bytes(cmd_str + " 107 0\n", encoding="UTF8")
telnet.write(payload)
response = telnet.read_until(b"#")
if args.verbosity > 0:
   print(payload)
   print(response)

# PCOSET footer
payload = bytes(cmd_str + " 123 1\n", encoding="UTF8")
telnet.write(payload)
response = telnet.read_until(b"#")
if args.verbosity > 0:
   print(payload)
   print(response)

payload = bytes(cmd_str + " 121 1\n", encoding="UTF8")
telnet.write(payload)
response = telnet.read_until(b"#")
if args.verbosity > 0:
   print(payload)
```

```
    print(response)

  # Wait 3 seconds and clear write flag
  sleep(3)
  payload = bytes(cmd_str + " 120 0\n", encoding="UTF8")
  telnet.write(payload)
  response = telnet.read_until(b"#")
  if args.verbosity > 0:
    print(payload)
    print(response)

  # Wait 3 seconds and force program reload
  sleep(3)
  payload = bytes(cmd_str + " 100 7\n", encoding="UTF8")
  telnet.write(payload)
  response = telnet.read_until(b"#")
  if args.verbosity > 0:
    print(payload)
    print(response)

  payload = bytes(cmd_str + " 101 1\n", encoding="UTF8")
  telnet.write(payload)
  response = telnet.read_until(b"#")
  if args.verbosity > 0:
    print(payload)
    print(response)
  payload = bytes(cmd_str + " 102 1\n", encoding="UTF8")

  telnet.write(payload)
  response = telnet.read_until(b"#")
```

```
if args.verbosity > 0:
    print(payload)
    print(response)

payload = bytes(cmd_str + " 123 1\n", encoding="UTF8")
telnet.write(payload)
response = telnet.read_until(b"#")
if args.verbosity > 0:
    print(payload)
    print(response)

payload = bytes(cmd_str + " 121 1\n", encoding="UTF8")
telnet.write(payload)
response = telnet.read_until(b"#")
if args.verbosity > 0:
    print(payload)
    print(response)

# Wait 3 seconds and clear write flag
sleep(3)
payload = bytes(cmd_str + " 120 0\n", encoding="UTF8")
telnet.write(payload)
response = telnet.read_until(b"#")
if args.verbosity > 0:
    print(payload)
    print(response)

# Wait 3 seconds, and clear busy flag
sleep(3)
payload = bytes(cmd_str + " 123 0\n", encoding="UTF8")
```

```
    telnet.write(payload)
    response = telnet.read_until(b"#")
    if args.verbosity > 0:
        print(payload)
        print(response)

    # Close telnet session
    telnet.close()



def main():
    args = get_args()

    csv_fh = open(args.csvfile, "rt")
    csv_reader = csv.reader(csv_fh, delimiter=',',
            quoting=csv.QUOTE_NONE)

    line = next(csv_reader)
    date_time = line[CSV_FIELDS["Date"]] + " " + line[CSV_FIELDS["Time"]]
    try:
        last_time = datetime.datetime.fromtimestamp(mktime(strptime(date_time, STRP_FOR
    except ValueError:
        line = next(csv_reader)
        date_time = line[CSV_FIELDS["Date"]] + " " + line[CSV_FIELDS["Time"]]
        last_time = datetime.datetime.fromtimestamp(mktime(strptime(date_time, STRP_FOR
    print(last_time)

# Ensure that the current date
    reached_now = False
    for line in csv_reader:
        date_time = line[CSV_FIELDS["Date"]] + " " + line[CSV_FIELDS["Time"]]
        time = datetime.datetime.fromtimestamp(mktime(strptime(date_time, STRP_FORMAT))
        now = datetime.datetime.now()
        # use a window of 10 minutes to find current time in file
        timedelta = datetime.timedelta(minutes=10)
        #print(time, now)
        if time < now < time + timedelta:
            reached_now = True
            break
        elif time > now + timedelta:
            raise ValueError("The file starts too far into the future.")
    if not reached_now:
        raise ValueError("No date in the CSV file matches the current time.")

    line = next(csv_reader)
    date_time = line[CSV_FIELDS["Date"]] + " " + line[CSV_FIELDS["Time"]]
    last_time = datetime.datetime.fromtimestamp(mktime(strptime(date_time, STRP_FORMAT
    print(last_time)



    for line in csv_reader:
        date_time = line[CSV_FIELDS["Date"]] + " " + line[CSV_FIELDS["Time"]]
        time = datetime.datetime.fromtimestamp(mktime(strptime(date_time, STRP_FORMAT))
        timediff = time - last_time
        last_time = time
        wait_sec = timediff.days * 24 * 60 * 60 + timediff.seconds
        print("Waiting %i secs." % wait_sec)
        sleep(0.0001)
        #sleep(wait_sec)
        communicate_line(args, line)
```

```
if __name__ == "__main__":
    main()
```

Listing 4: Main program loop in the `spcControl` module.

```
from time import strptime, sleep, mktime, time
import datetime
import csv
import socket
import sys
import time
import traceback
from spcControl import (
    get_config_file,
    get_config,
    chamber,
    heliospectra,
    email_error,
    )


timepoint_count = 0
config = get_config(get_config_file())


def _email_traceback(traceback):
  message_text = "Error on chamber %i\n" % \
      config.getint("Global", "Chamber")
  message_text += traceback
  subject = "Conviron Error (Chamber %i)" % \
      config.getint("Global", "Chamber")
  email_error(subject, message_text)
```

```
def _log_to_postgres(log_tuple):
    try:
        import psycopg2
    except ImportError:
        return
    try:
        con = psycopg2.connect(
                host=config.get("Postgres", "Host"),
                port=config.getint("Postgres", "Port"),
                user=config.get("Postgres", "User"),
                password=config.get("Postgres", "Pass"),
                )
        cur = con.cursor()
        statement = config.get("Postgres", "InsertStatement")
        cur.execute(statement, log_tuple)
        con.commit()
        cur.close()
        con.close()
    except Exception as e:
        traceback_text = traceback.format_exc()
        _email_traceback(traceback_text)


def communicate_line(line):
    """This processes each line, and handles any errors which they create elegantly.
    """
    global timepoint_count
    timepoint_count += 1
    if config.getboolean("Global", "Debug"):
        print("Csv line is:", line)
    now = datetime.datetime.now()
    log_str = "Running timepoint %i at %s" % (timepoint_count, now)
    print(log_str, end='... ')
    chamber_num = config.get("Global", "Chamber")
    sys.stdout.flush() # flush to force buffering, so above is printed
    try:
        if config.getboolean("Conviron", "Use"):
            chamber.communicate(line)
        if config.getboolean("Heliospectra", "Use"):
            heliospectra.communicate(line)
        print("Success")
        log_tuple = (chamber_num, "FALSE", log_str)
    except Exception as e:
        print("FAIL")
        if config.getboolean("Global", "Debug"):
            traceback.print_exception(*sys.exc_info())
        traceback_text = traceback.format_exc()
        _email_traceback(traceback_text)
        log_tuple = (chamber_num, "TRUE", "%s\n%s" % (log_str, traceback_text))
    _log_to_postgres(log_tuple)


def main():
    """Main event loop. This just handles the files, and passes lines to be
      processed to communicate_line()
      """

    # open the CSV file, and make the csv reader
```

```
try:
    csv_file = sys.argv[1]
    csv_fh = open(csv_file)
except (KeyError, IOError):
    print("ERROR: csv file must exist\n"
            "Usage:\n"
            "\tpython3 -m spcControl <csv_file> [<ini.file>]"
            )
    exit(-1)

csv_reader = csv.reader(csv_fh, delimiter=',',
        quoting=csv.QUOTE_NONE)
# Define these for short/easy reference later
datefield = config.getint("GlobalCsvFields", "Date")
timefield = config.getint("GlobalCsvFields", "Time")

# Detect if the file has a header, by trying to get a date and time from
# the first two field of the first row.
try:
    first_line = next(csv_reader)
    date_time = first_line[datefield] + " " + first_line[timefield]
    # If this fails to strip the datetime from date_time, it raises a
    # ValueError, and means the first line isn't valid (i.e. probably a
    # header)
    first_time = datetime.datetime.fromtimestamp(
            mktime(strptime(
                date_time, config.get("Global", "CsvDateFormat")
                ))
            )
except ValueError:
    # If an error occurs in this block, there's something wrong with the
    # csv file, so we don't want to catch it.
    first_line = next(csv_reader)
    date_time = first_line[datefield] + " " + first_line[timefield]
    first_time = datetime.datetime.fromtimestamp(
            mktime(strptime(
                date_time,
                config.get("Global", "CsvDateFormat")
                ))
            )
if config.getboolean("Global", "Debug"):
    print("First time in file is:", first_time)

## Find current time in CSV file ##
now = datetime.datetime.now()
# use a window of 10 in
timedelta = datetime.timedelta(minutes=10)

# Check if file starts too far into the future
if first_time > now + timedelta:
    raise ValueError("The file starts too far into the future.")
line = []  # Make the line variable local to the main() function

# Read through the file to find the current date and time
while not first_time < now < first_time + timedelta:
    try:
        line = next(csv_reader)
    except StopIteration:
        raise ValueError(
                "No date in the CSV file matches the current time."
                )
    date_time = line[datefield] + " " + line[timefield]
    first_time = datetime.datetime.fromtimestamp(
            mktime(strptime(
                date_time, config.get("Global", "CsvDateFormat")
                ))
            )

# Run the first line
prev_run_length = 0
start = time.time()
communicate_line(line)
prev_run_length = time.time() - start
previous_time = first_time

# Loop through the rest of the lines, running each one
for line in csv_reader:
    date_time = line[datefield] + " " + line[timefield]
    csv_time = datetime.datetime.fromtimestamp(
            mktime(strptime(
                date_time,
                config.get("Global", "CsvDateFormat")
                ))
            )
    diff = csv_time - previous_time
    wait_sec = max(
            diff.days * 24 * 60 * 60 + diff.seconds - prev_run_length,
            0  # We don't want to wait a negative number of seconds
            )
    if config.getboolean("Global", "Debug"):
            print("Waiting %i secs." % wait_sec)
        sleep(wait_sec)
        start = time.time()
        communicate_line(line)
        prev_run_length = time.time() - start
        previous_time = csv_time

main()
```

6.3.2 Invocation of `spcControl`

Some important notes about the invocation of `spcControl`:

- `SolarCalc` is a java program. The implementation of `strftime` function, which formats date objects into character strings (text) in either `SolarCalc` or java does not comply with the standards which the `datetime` module of python does. Thus, each time in the hour after midnight is recorded as, for example, `07/01/12 00:15:00AM`, for 15 minutes after midnight, July 1, 2012. As a result, we cannot *directly* use SolarCalc output. There is a very simple fix however, finding all occurrences of `00:` in the SolarCalc output file and replacing them with `12:` solves this issue. As these files are very large, this is best accomplished with `sed`, the Stream EDitor, as show below in listing 5.

```
sed -ibak -e 's/00:/12:/g' <file>
```

Listing 5: Code to ensure `SolarCalc` makes dates which are compatible with python

6.4 Supplementary Information

6.4.1 High Throughput RNAseq Library Preparation Protocol

To economically generate RNAseq libraries from the hundreds of samples I have collected, non kit-based protocols must be used. Here, I briefly summarise the failed implementation of the RNAseq library preparation protocol of Kumar et al. (2012), a published protocol enabling the preparation of RNAseq libraries in high throughput using 96 well plates. Specifically, I used a slightly modified version of the High-Throughput RNAseq protocol described in Supplementary Methods 1 of Kumar et al. (2012) (hereafter referenced as the HTR protocol). To test and optimise this protocol, leaf tissue collected from surplus 5-week old *A. thaliana* Col-0 from a colleague's experiment was used. This tissue had been collected into Qiagen 1.2mL collection tubes containing a single steel ball bearing, and snap frozen in liquid N_2 before grinding in a TissueLyser for two one minute pulses at 25Hz at a later date. In collection tubes, 750 μLDynabeads Lysis/Binding Buffer was added, before sample was ground for a further 30s in a TissueLyser as before. Then, lysates were prepared per Steps 1.1.6-1.1.8 of the HTR protocol described in Supplementary Methods 1 of Kumar et al. (2012). Isolated mRNA was obtained and cDNA synthesised and fragmented according to steps 1.2 to 3 of the HTR protocol of Kumar et al. (2012). Working with Dr. Norman Warthmann, the remaining steps of the HTR protocol were validated using four cDNA samples, following the HTR protocol with modifications as described below.

To test all sequencing adaptors, sonicated genomic DNA was used as input material, due to it's similar size and fragmentation properties to fragmented cDNA, and due to the scarcity of large quantities of cDNA of little value. This DNA was obtained from *Oryza sativa* seedlings, diluted to a concentration of approximately $7 ng \mu L^{-1}$ (500 ng in 70 μL) before sonication in a Diagenode Bioruptor DNA sonicator. This sonicated DNA was cleaned up per steps 3.5-3.8 of the HTR protocol, with the modification that 30 μL bead binding buffer and 40 μLAmpure XP SPRI beads were used. Then, a modified step 4 of the HTR protocol was used to create unamplified sequencing libraries. In step 4.1 and 4.2, double reactions were performed, however the same quantity of SPRI cleanup reagents were used. Before adaptor ligation, the A-tailed libraries were eluted using a mixture of 5 μLdiluted adaptor oligonucleotide, 1 μL10x ligation buffer and 2 μLwater. The DNA ligase was diluted in the remaining 1.5 μL water and added to each reaction, before proceeding with protocol steps 4.3.3 onwards. The adaptors used were not those specified by Kumar et al. (2012), instead custom sequencing adaptors were used. These adaptors were designed by Dr Norman Warthmann, and are compatible with the T/A overhang ligation method Kumar et al. (2012) utilise. The protocol described in step 5 of the HTR protocol was used, to amplify the libraries.

RNA quality and quantity was assayed using the Agilent BioAnalyser digital electrophoresis system. The RNA samples were loaded into a Plant RNA Pico analysis chip and an analysis run per manufacturer's protocol. The effectiveness of various steps in this protocol was assayed by digital electrophoresis with the Shimadzu MultiNA instrument, using the DNA1000 kit. The pre-mix protocol was used: 2 μLsample was added to 4 μLDNA1000 marker solution, the solution mixed, and loaded into the instrument, which was run according to manufacturer's DNA1000-PreMix protocol. Quantitative PCR was performed to test the ligation and PCR efficiency of each adaptor. To do so, 2 $\mu L$of each pre-amplification library was combined with 5 μLSybr Green qPCR master mix, 1 μLeach of the forward and index 1 reverse primers (see **??**), and 1 μLUracil-Specific Excision Reagent (USER) enzyme mix.

Sections of this protocol was successfully implemented. Messenger RNA was extracted in a 96 well plate format using oligo-d(T) magnetic beads, albeit with low yield that expected (Kumar et al. 2012) (Figure 6.1, Figure 6.2). Complimentary DNA (cDNA) was prepared from this mRNA. Enzymatic fragmentation of cDNA, end repair and A-tailing were performed. However, a cumulation of possible ligation and PCR biases caused order-of-magnitude variation in library abundance (data not shown). This variation in library abundance was confirmed with diagnostic qPCR (data not shown). For this reason, the use of this protocol in my experiment was abandoned, as the optimisation of these difficulties may have been very time consuming and was not feasible in the time remaining in my Honours year.

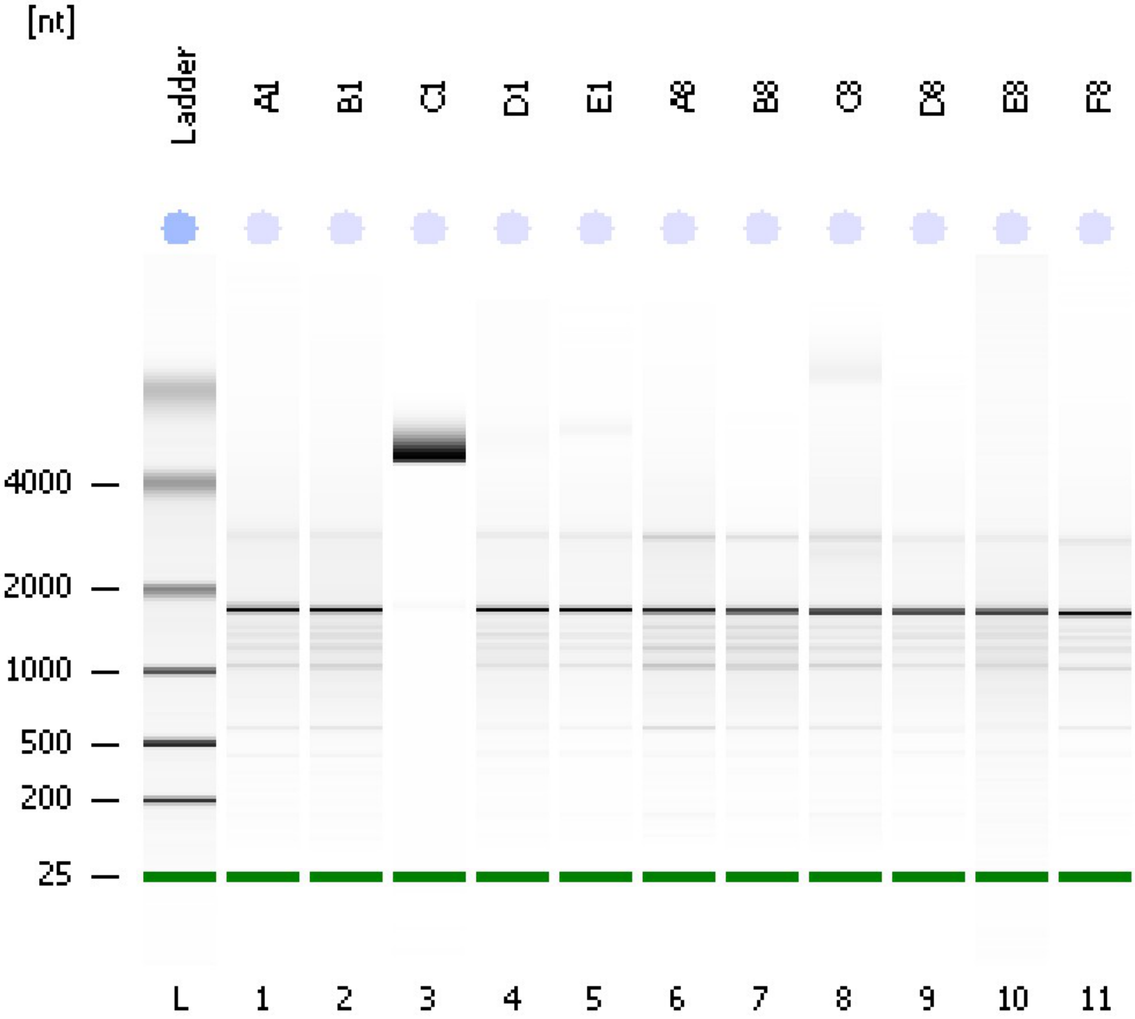

Figure 6.1: BioAnalyser digital gels of 10 RNA samples extracted. The extraction or quantification of C1 failed. Note the feint mRNA smear, and residual rRNA bands. This indicates successful direct extraction of mRNA molecules from plant tissue lysates, and residual carry-over of rRNA.

6.4.2 Additional Dynamic Growth Conditions

The `spcControl` software I have developed can be applied to many experimental designs. For example, it was further used to create conditions which examine hypotheses beyond the scope this thesis. This includes two conditions to test the overall effect of environments with higher light and more diurnal variation in temperature, such as may be experienced in inland regional climates, compared to conditions with lower light intensity and lower diurnal variation in temperature, such as those often encountered in coastal climates. These conditions, named “NSW inland” and “NSW coastal” respectively, generally are more harsh than the sufficient, fluctuating and excess light dynamic growth conditions created for my experiments. Additionally, conditions required to conduct a virtual reciprocal transplant of *Pelargonium* species from Australia and South Africa were created, by simulating regional climates in two locations, in coastal New South Wales, Australia and coastal Western Cape, South Africa. The specifics of these dynamic growth conditions are beyond the scope of this thesis.

6.4.3 RIX lines

Line Name	
CvL-1 x CvL-146	CvL-171 x CvL-143
CvL-10 x CvL-26	CvL-174 x CvL-34
CvL-101 x CvL-176	CvL-180 x CvL-157
CvL-102 x CvL-28	CvL-183 x CvL-118
CvL-145 x CvL-105	CvL-186 x CvL-27
CvL-107 x CvL-124	CvL-187 x CvL-190
CvL-109 x CvL-185	CvL-187 x CvL-69
CvL-109 x CvL-47	CvL-189 x CvL-133
CvL-110 x CvL-32	CvL-19 x CvL-173
CvL-112 x CvL-30	CvL-19 x CvL-67
CvL-113 x CvL-141	CvL-190 x CvL-176
CvL-114 x CvL-3	CvL-191 x CvL-31
CvL-114 x CvL-60	CvL-192 x CvL-189
CvL-115 x CvL-126	CvL-20 x CvL-138
CvL-117 x CvL-73	CvL-21 x CvL-22
CvL-118 x CvL-108	CvL-24 x CvL-171
CvL-118 x CvL-164	CvL-25 x CvL-9
CvL-119 x CvL-177	CvL-26 x CvL-74
CvL-12 x CvL-142	CvL-33 x CvL-58
CvL-122 x CvL-42	CvL-35 x CvL-120
CvL-125 x CvL-117	CvL-38 x CvL-35
CvL-128 x CvL-6	CvL-39 x CvL-27
CvL-129 x CvL-132	CvL-40 x CvL-74
CvL-133 x CvL-35	CvL-41 x CvL-70
CvL-134 x CvL-29	CvL-43 x CvL-131
CvL-135 x CvL-10	CvL-44 x CvL-50
CvL-135 x CvL-140	CvL-45 x CvL-23
CvL-136 x CvL-102	CvL-46 x CvL-29
CvL-165 x CvL-137	CvL-48 x CvL-160
CvL-139 x CvL-162	CvL-49 x CvL-158
CvL-139 x CvL-36	CvL-5 x CvL-172
CvL-14 x CvL-4	CvL-5 x CvL-188
CvL-146 x CvL-64	CvL-51 x CvL-111
CvL-147 x CvL-50	CvL-51 x CvL-18
CvL-147 x CvL-69	CvL-54 x CvL-183
CvL-149 x CvL-165	CvL-55 x CvL-18
CvL-150 x CvL-37	CvL-59 x CvL-116
CvL-152 x CvL-42	CvL-6 x CvL-131
CvL-153 x CvL-108	CvL-61 x CvL-162
CvL-153 x CvL-20	CvL-63 x CvL-151
CvL-154 x CvL-144	CvL-7 x CvL-46
CvL-156 x CvL-166	CvL-8 x CvL-61
CvL-16 x CvL-4	Ler x Ler
CvL-16 x CvL-66	Ler self
CvL-164 x CvL-7	Cvi x Ler
CvL-166 x CvL-25	Ler x Cvi
CvL-168 x CvL-22	Cvi-0
CvL-169 x CvL-175	
CvL-17 x CvL-21	
CvL-170 x CvL-24	

Table 6.2: Cvi Ler RIX lines planted and harvested as part of this thesis.

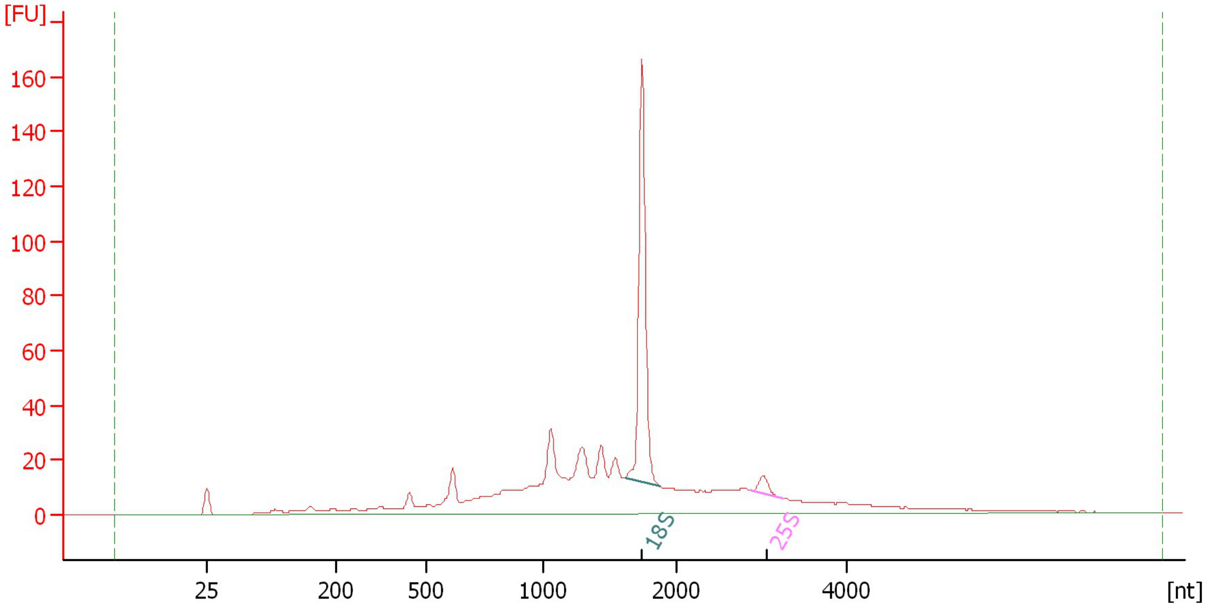

Figure 6.2: BioAnalyser digital electrophoretogram of a representative mRNA sample (Sample B1 in Figure 6.1). Note the smear-like quality of the mRNA sample, and the reduced or absent ribosomal RNA peaks, when compared with a total RNA sample such as in Figure 4.1