DOCTORAL THESIS

The Molecular Characterisation of the Vernalisation Response in Safflower via the Development of Genomic and Transcriptomic Resources

Author:
Darren CULLERNE
BAppSci/BIT
MBiotech(*Research*)

Supervisors:
Dr. Andy EAMENS
Dr. Craig WOOD
Dr. Ben TREVASKIS
Prof. Christopher GROF

A thesis submitted in fulfilment of the requirements
for the degree of Doctor of Philosophy (Biological Sciences)

in the

Faculty of Science and Information Technology
School of Environmental and Life Sciences

May 3, 2017

Statements of Collaboration and Originality

I, Darren CULLERNE, declare that:

(a) This thesis, titled 'The Molecular Characterisation of the Vernalisation Response in Safflower via the Development of Genomic and Transcriptomic Resources' contains no material which has been accepted for the award of any other degree or diploma in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. I give consent to the final version of my thesis being made available worldwide when deposited in the University's Digital Repository**, subject to the provisions of the Copyright Act 1968.
**Unless an Embargo has been approved for a determined period.

(b) The work embodied in this thesis has been conducted in collaboration with and carried out at CSIRO Black Mountain in Canberra, in association with the University of Newcastle. As part of this collaboration, I have undertaken the research contained within this Thesis using CSIRO facilities under Dr Craig Wood in the Safflower Engineering laboratories.

Signed: _____

Date: _____

Acknowledgements

Firstly, I would like to thank my supervisors, Dr Craig Wood, Dr Ben Trevaskis, Dr Andy Eamens and Professor Chris Grof. Each of these people provided a unique depth of knowledge and expertise in their guiding contributions to this project.

I would like to acknowledge: The University of Newcastle, the Grains Research and Development Corporation and the CSIRO's Office of the Chief Executive for providing both operating funds and scholarship stipends. My involvement in this project would not have been possible without their financial support.

The bioinformatics community at CSIRO. In particular, my conversations with Dr Stuart Stephen, Mr Andrew Spriggs and Dr Jen Taylor, many of which were facilitated by the consumption of coffee. They proved to be invaluable for the correct planning and analysis of the countless samples and data files processed as part of this project.

The many postdoctoral scientists and technicians in the CSIRO Plant and Oil Engineering and the CSIRO Crop Genomics groups. Their suggestions of protocols, troubleshooting help, molecular biology techniques and assistance in getting my hands dirty on the bench and in the glasshouses is much appreciated.

The other PhD students, both within and external to CSIRO, some who have already graduated and others who yet to complete their projects. Their camaraderie and support helped me through some rough times during this project. I hope I have helped them in the same way. I would like to particularly thank Kyle Reynolds for his mental support, baked goods and coffee breaks throughout this project.

My family and friends for their ongoing support of my partner and I. Whether the support was financial, psychological, fermented or nutritional, I could not have accomplished this project without them.

Lastly, but most importantly, I'd like to thank my partner Michelle for support and encouragement throughout my studies and during this project in particular. Her endless patience and unwavering belief in me were crucial to getting me over the finishing line.

To my grandparents, Eileen Dotridge Comans, Thomas Joseph Bailey and Monica Ann Bailey and my great uncle Edmund Theopolis Cullerne. You have uniquely influenced, supported, and inspired me to excel.

Contents

List of Figures

List of Tables

Abbreviations

AGRF	**A**ustralian **G**enome **R**esearch **F**acility
Arabidopsis	*Arabidopsis thaliana*
AraTha	***Ara**bidopsis **tha**liana*
At	***A**rabidopsis **t**haliana*
BLAST	**B**asic **L**ocal **A**lgorithm **S**earch **T**ool (Software)
BLASTN	**BLAST N**ucleotide (Software)
BLASTP	**BLAST P**rotein (Software)
BUSCO	**B**enchmarking **U**niversal **S**ingle-**C**opy **O**rthologs (Software)
bp	**b**ase **p**air
Bv	***B**eta **v**ulgaris*
CarTin	***Car**thamus **tin**ctorius*
CEGMA	**C**ore **E**ukaryotic **G**enes **M**apping **A**pproach (Software)
ChrLav	***Chr**ysanthemum **lav**andulifolium*
ChrMor	***Chr**ysanthemum **mor**ifolium*
Ci	***C**hicory **i**ntybus*
cM	**C**enti**m**organs
CSIRO	**C**ommonwealth **S**cientific and **I**ndustrial **R**esearch **O**rganisation
Ct	***Car**thamus **tin**ctorius*
CTAB	**C**etyl **t**rimethyl **a**mmonium **b**romide
DArT	**D**iversity **Ar**rays **T**echnology
DEPC	**Die**thyl**p**yro**c**arbonate
EDTA	**E**thylene**d**iamine**t**etraacetic **a**cid
Eg	*Eustoma **g**randiflorum*
EGTA	**E**thylene **g**lycol-bis(β-aminoethyl ether)-N,N,N′,N′-**t**etraacetic **a**cid
E/O	**E**ocene / **O**ligocene boundary
EST	**E**xpressed **S**equence **T**ag
g	**g**ravitational force
Gbp	**G**iga**b**ase **p**air (1 Gbp = 1,000,000,000 bp)
GLA	**g**amma **l**inolenic **a**cid
GM	**G**enetic **M**odification
GRDC	**G**rains **R**esearch **D**evelopment **C**orporation
ha	**h**ect**a**re
HorVar	***Hor**deum **vul**gare*
Hv	***H**ordeum **v**ulgare*
MP	**M**ate **p**air (reads)
MSA	**M**ultiple **s**equence **a**lignment
Mt	***M**edicago **t**runcatula*
M. truncatula	*Medicago truncatula*

mya	**M**illions of years (**A**nnus)
NCBI	**N**ational **C**enter for **B**iotechnology **I**nformation
NEB	**N**uclear **E**xtraction **B**uffer
NGS	**N**ext **g**eneration **s**equencing
$OD_{x/y}$	**O**ptical **d**ensity (absorbance) at wavelength x and y
PacBio	**Pac**ific **Bio**sciences (sequencing technology)
PCR	**p**olymerase **c**hain **r**eaction
PE	**P**aired **e**nd (reads)
PHD-PRC2	**P**lant **H**omeo-**D**omain **P**olycomb **R**epression **C**omplex **2**
PIPES	piperazine-N,N′-bis(2-ethanesulfonic acid)
PVP	**p**oly**v**inyl**p**yrrolidinone
RIN	**R**NA **in**tegrity score
RPKM	**R**eads **p**er **k**ilobase of transcript per **m**illion mapped reads
RT-qPCR	**R**everse **t**transcriptase **q**uantitative **p**olymerase **c**hain **r**eaction
SAM	**S**hoot **a**pical **m**eristem
SCUBAT	**S**caffolding **C**ontigs **U**sing **B**LAST-like **A**lignment **T**ool
SDS	**S**odium **d**odecyl **s**ulfate
SHO	**S**uper **h**igh **o**leic
SNP	**s**ingle **n**ucleotide **p**olymorphism
TAIR	**T**he **A**rabidopsis *I*nformation **R**esource
Tris-HCl	**Tris**(hydroxymethyl)aminomethane hydrochloride
UTR	**Un**translated **r**egion

Gene and Protein Abbreviations

AP1	APETALA 1
CAL	CAULIFLOWER-A
BTC1	BOLTING TIME CONTROL 1
FAD2	FATTY ACID DESATURASE 2
FD	BASIC-LEUCINE ZIPPER (BZIP) TRANSCRIPTION FACTOR
FL	FLC-LIKE (*Beta vulgaris*)
FLC	FLOWERING LOCUS C
FLCL	FLC-LIKE (*Eustoma* spp.)
FL1	FLC-LIKE 1 (*Chicory intybus*)
FRI	FRIGIDA
FT	FLOWERING LOCUS T
FTL	FT-LIKE
LFY	LEAFY
MAF1-5	MADS AFFECTING FLOWERING 1-5
OS2	ODDSOC2
PEP1	PERPETUAL FLOWERING 1
SOC1	SUPPRESSOR OF OVEREXPRESSION OF CONSTANS 1
SOC1L	SOC1-LIKE
VIN3	VERNALISATION INSENSITIVE 3
VRN1-2	VERNALISATION 1-2

Nomenclature 1 Italicised represents the DNA locus or RNA transcript of the gene. Non-italicised represents the encoded protein.

Nomenclature 2 A lower case gene or encoded protein represents a mutant or recessive allele of that gene or protein.

Nomenclature 3 A two letter abbreviation before a gene represents the species where it is found, e.g. *At* = *Arabidopsis*.

Abstract

The Molecular Characterisation of the Vernalisation Response in Safflower via the Development of Genomic and Transcriptomic Resources

by Darren CULLERNE

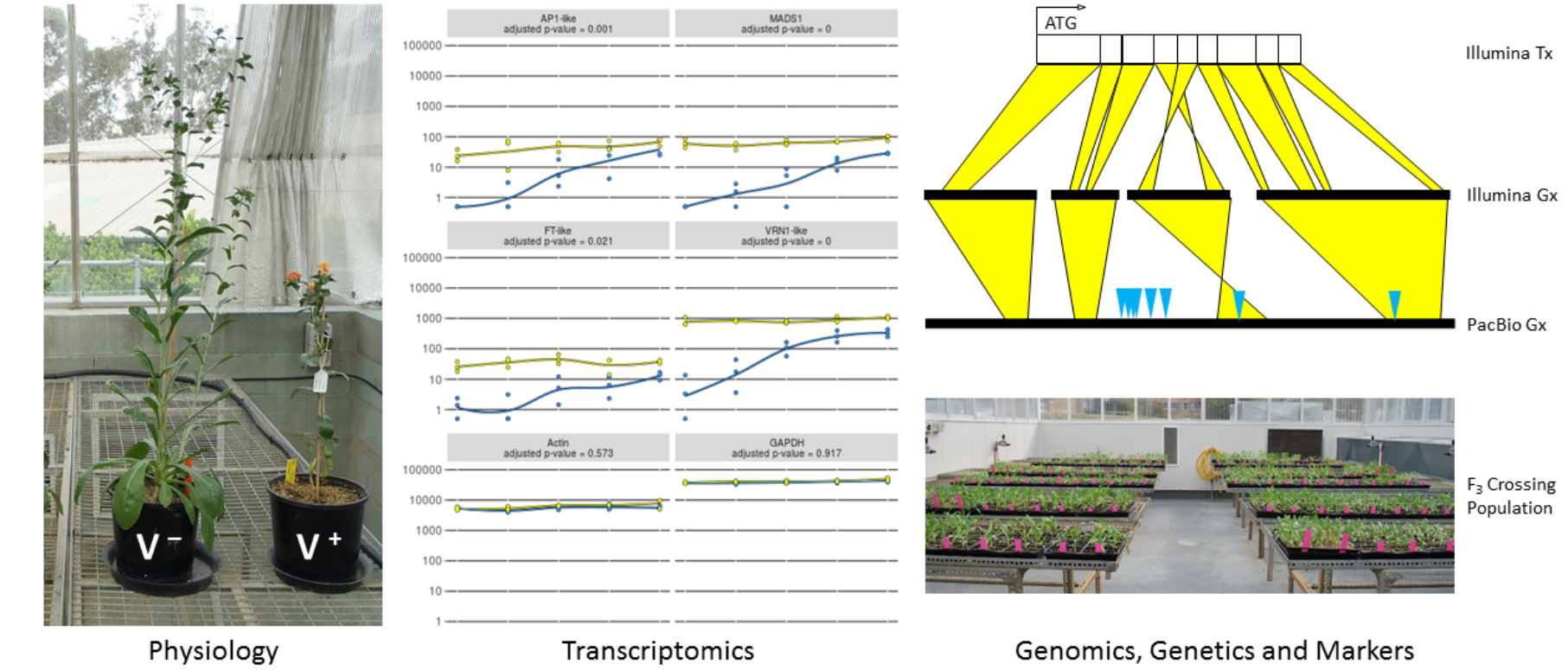
Safflower (*Carthamus tinctorius*) is an oilseed grown globally. There is an interest in modifying safflower to cope with climate change and to adapt to new agronomic trends. While almost all cropped safflower are spring varieties, a number of wild safflower varieties are noted as 'winter hardy'. These display characteristics found in plants that respond to vernalisation (an extended period of non-freezing cold). Because flowering traits, including the vernalisation response, are linked to yield and adaptability to climate, this PhD project sought to understand the vernalisation response in safflower, as a component of flowering time.

The vernalisation response in 'winter hardy' and spring safflower was investigated. It was confirmed that winter safflower does respond to vernalisation conditions, similar to other plant species. The vernalisation response in winter safflower is saturated after approximately 2 weeks exposure to vernalisation conditions. It is inheritable, epigenetic and appears to be dependent on a single recessive allele, as shown by segregation ratios in a crossing population created from winter and spring parents.

Two approaches were developed to characterise this vernalisation response. Firstly, RNA sequencing (RNA-Seq) was performed on RNA extracted from winter and spring safflower before, during and after exposure to vernalisation conditions. Differential expression analyses on the resulting RNA-Seq datasets tentatively identified four genes as directing functional roles in the vernalisation response of safflower: *APETALA 1-LIKE* (*CtAP1-LIKE*), *MADS-BOX DOMAIN CONTAINING 1* (*CtMADS1*), *FLOWERING LOCUS T-LIKE* (*CtFT-LIKE*) and *VERNALISATION 1-LIKE* (*CtVRN1-LIKE*). This analysis also identified 33 additional gene products (annotated transcripts or transcripts of unknown function) as candidates for further experimental investigation.

In addition to the transcriptomic data, genomic resources were developed to further characterise the molecular basis of the vernalisation response. A high quality *de novo* assembly was constructed using Illumina reads from spring safflower, covering approximately 80% of the estimated 1,400,000,000 base pair (1.4 Gbp) spring safflower genome. Using this draft genome in combination with F_3 crossing families and a genetic marker approach, 27 genetic markers for vernalisation were identified. Furthermore, these markers were mapped to a recent genetic map of safflower (Bowers 2016), clustering in close proximity to one another on chromosome 8. A single differentially expressed transcript, identified in the transcriptomic analyses, was located on the same chromosome. However, the transcript of interest was mapped to a chromosome 8 position some distance away from the identified markers.

These high quality transcriptomic and genomic resources were used to identify the molecular basis for vernalisation in safflower. The investigative approaches developed in this project can also be utilised to characterise the molecular mechanisms of other traits in safflower.



Graphical abstract of this PhD project. The first panel shows winter safflower, unvernalised and vernalised. The second panel shows four characterised transcripts that are differentially expressed after vernalisation of winter safflower, with two non-differentially expressed transcripts at the bottom. Blue are from winter safflower, yellow are from spring safflower The third panel shows a transcript mapping to multiple Illumina genomic contigs, which in turn map to a single PacBio contig. The bottom of the third panel shows the F_3 crossing population.

Chapter 1

Introduction

1.1 Safflower: An Ancient Crop with a Future Based on Biotechnology

Safflower (*Carthamus tinctorius* L.) is a member of the *Asteraceae* family of dicotyledonous flowering plants. This domesticated oilseed crop, native to the eastern and southern Mediterranean, the Middle East and India (Knowles 1960), is grown in over 60 countries. Historically, safflower was used in traditional medicines, and, using the anthocyanin that is rich in floral parts harvested prior to seed set, as a textile dye (Zohary and Hopf 1993). In Australia, the United States, Mexico, Kazakhstan and India, safflower is predominantly grown as an oilseed for the production of vegetable oil.

Safflower currently produces two distinct types of oil in the seed; high linoleic or high oleic, with the ancestral oil type rich in linoleic acid. The traditional oil contains approximately 80% linoleic acid, and has both nutritional and oleochemical uses. The arrangement of the double bonds in linoleic acid cannot be produced in humans and is, therefore, classified as an essential fatty acid. However, the double bond is oxidatively unstable. Therefore, it is used as an additive to many modern paints and polymers in the oleochemical industries (Knowles 1949; Işigigür et al. 1995; Gecgel et al. 2007). The second and more recently developed oil is called high oleic and is approximately 80% oleic acid, with low levels of linoleic acid. This oil is predominantly used in nutrition and is considered to be one of the lowest sources of polyunsaturated fatty acids (United States Department of Agriculture 2016).

Globally, safflower covers just over 1 million hectares (ha), with Mexico, India and Kazakhstan being the dominant producers, generating approximately 70% of the global safflower harvest. In 2014, Mexico produced approximately 530,000 tonnes (t) (approximately 30%), India produced approximately 255,000 t (18%) and Kazakhstan produced approximately 127,000 t (15%; Fig. 1.1). In Australia, the area sown to safflower reached a historical high in the late 1980s at about 80,000 ha, with most safflower grown in the wheat belt along the south-eastern seaboard. The emergence of more profitable crops such as high yield and high oleic oil canola, sunflower and cotton, has gradually decreased the acreage of safflower grown. With the reduced demand for safflower oil and the emergence of other more profitable crops, safflower research and breeding in Australia was phased out in the 1980s (Smith 1996). In 2014, just 5,030 t was harvested in Australia, making safflower a niche crop.

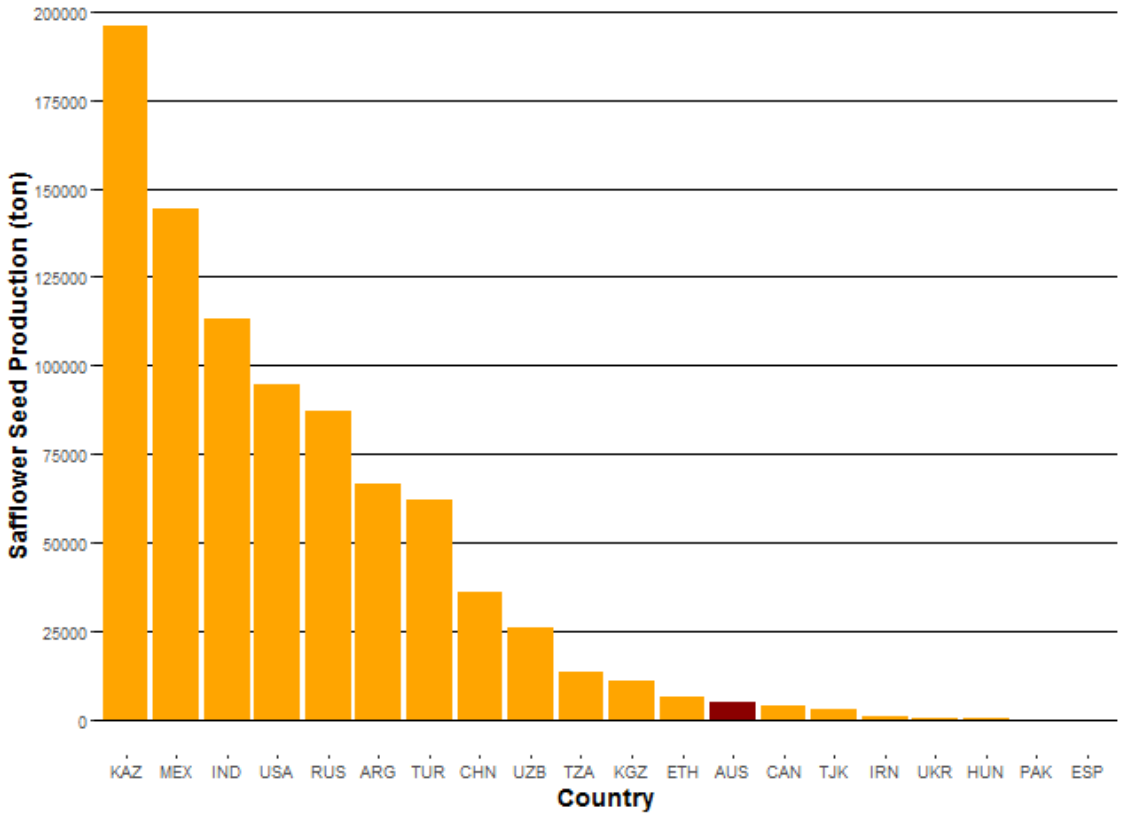


FIGURE 1.1: Global safflower production in 2014 ordered by quantity. Australia (AUS, highlighted) was ranked 13th in the world for safflower seed production, at approximately 5,030 t (United Nations 2016).

In 2010, the Commonwealth Scientific and Industrial Research Organisation (CSIRO) and the Grains Research Development Corporation (GRDC) started a research initiative to develop high-value oils in safflower seed. These high-value traits rely on advances in biotechnology and the genetic engineering of plants. This collaboration between major research bodies in Australia resulted in a robust genetic modification protocol for safflower (Belide et al. 2011) from which high-value traits can be introduced. Similar biotechnology has been deployed in North America for the development of genetically engineered safflower producing gamma-linolenic acid (GLA), the primary component of evening primrose oil. GLA has a unique arrangement of double bonds in the oil that imparts a range of nutraceutical benefits for human nutrition (Nykiforuk et al. 2012). CSIRO has recently developed a unique oil in safflower called super high oleic (SHO), an oil that possesses a unique resistance to oxidative degradation and therefore could potentially be widely used in oleochemical industries (Wood et al. 2016, in preparation). SHO has been licensed commercially and is poised to be released to growers in 2018.

Both GLA and SHO are genetically modified (GM) based traits unique to safflower and are unlikely to be transferred in other crops. As demand for these unique oils increases, so will the need to develop safflower for improved yields and to meet the range of new challenges in the agricultural growing cycle. It is also possible that cropping of

safflower will need to expand beyond the traditional growing regions into more marginal growing areas. Although the success of GM traits, such as SHO in safflower, remains unclear, it is apparent that future improvements of safflower will be hindered by a lack of breeding resources. The lack of diversified germplasm and modern genomic selection tools, such as marker assisted selection and genomic reference databases, could seriously hinder the development and expansion of safflower in a changing climate and into more challenging growing regions.

Across the world, all commercially cropped safflower cultivars are 'spring safflower' varieties, i.e. planted in late winter to early spring as temperatures and day lengths start to increase (Jochinke et al. 2008; Knights 2010; Knowles 2012). These varieties take approximately 100 to 120 days to flower, with high summer temperatures assisting the dessication of seed heads prior to harvest. In contrast, a number of wild 'winter hardy' safflower varieties, sourced from eastern China, performed best when planted in autumn and early winter. Spring safflower cultivars planted at the same time did not survive (Johnson and Li 2008). While these 'winter hardy' safflower cultivars express a phenotype that resembles the vernalisation response in other plant species, i.e. a large rosette and late elongation (Carapetian 2001), there is no information regarding whether or not they possess a true vernalisation response.

The remainder of this introductory chapter outlines the major biological theme of this thesis, the molecular basis of vernalisation in a range of plants. Chapter 2 examines phenotypes of 'winter hardy' safflower to confirm if it is indeed a vernalisation responsive safflower species. Chapters 3 and 4 are detailed accounts of the extensive use of next-generation sequencing (NGS) to develop the transcriptomics and genomic databases and marker assisted selection. The final chapter, Chapter 5, syntheses the advances detailed in each of the preceding chapters and outlines future work for the improvement of safflower. If the coordination of flowering time in safflower is, in fact modifiable, this is likely to have an enormous impact. It could extend the regions in Australia where safflower could be cultivated as a commercially viable crop, increase its utility as a winter break crop and improve safflower's ability to adapt to climate change.

1.2 The Vernalisation Response

The vernalisation (from the Latin *vernus*, meaning spring) response in plants is characterised by an accelerated transition from the vegetative growth stage to the reproductive stage of development following an extended period of exposure to non-freezing cold. In vernalisation responsive *Arabidopsis thaliana* (*Arabidopsis*) lines, without exposure to these non-freezing conditions, the plant will continue to grow Vegetatively for far longer than vernalised plants of the same variety (Fig. 1.2).

FIGURE 1.2: The effect of vernalisation on *Arabidopsis thaliana*. There is no response to vernalisation in summer (or spring) *Arabidopsis* annuals (a). Winter *Arabidopsis* annuals have a delayed elongation response when not exposed to vernalisation (b), but elongate and bolt in a similar fashion to spring annuals when exposed to vernalisation conditions (c). Sung and Amasino (2004a).

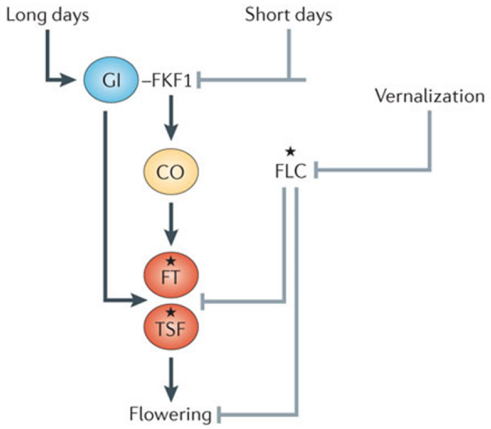
There are a number of common elements that define the vernalisation response. These include the:

- temperature during exposure is below approximately 10°C but non-freezing
- length of exposure is greater than seven days
- effects of the cold exposure is not seen during cold exposure itself
- effects of extended cold exposure are reset in the next generation

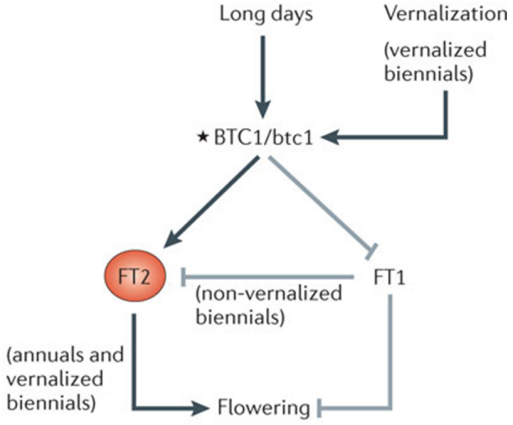
A number of features are not associated with the vernalisation response:

- cold triggered bud formation
- cold triggered bud dormancy breakage
- stratification requirement for germination
- lower ambient temperatures above 15°C reducing time to flowering

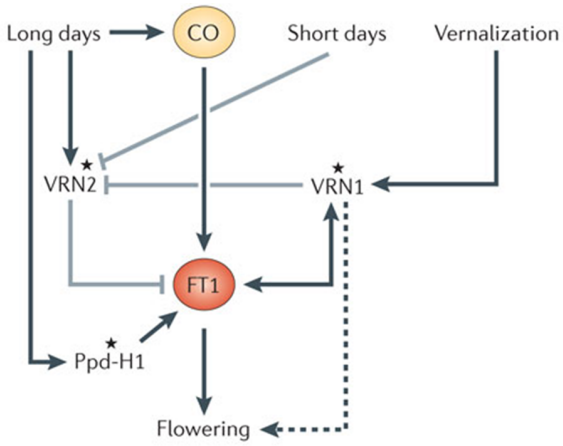
The implications of the vernalisation response on harvest time and crop yield has been an important focus of plant biology research for over 150 years (Klippart 1857). McKinney (1940), leveraging off the earlier work of Klippart and Gassner (1918), developed the concept of 'Growth Phases'. McKinney observed that while conditions such as vernalisation may result in similar phenotypes in different species, exposure of the same species to the same growth conditions, but at different growth phases, will not necessarily result in the same phenotypic response. While phenotypes of the vernalisation response appear to be consistent between both monocotyledonous and dicotyledonous angiosperms, there are a number of different molecular mechanisms that underpin the vernalisation response in a range of angiosperms (Fig. 1.3).



(a) The vernalisation gene model for *Arabidopsis*.



(b) The vernalisation gene model for sugar beet (*Beta vulgaris*).



(c) The vernalisation gene model for wheat (*Triticum aestivum*) and barley (*Hordeum vulgare*).

FIGURE 1.3: Gene interaction models from different angiosperms that respond to vernalisation environmental cues. (Andres and Coupland 2012).

1.3 Mechanisms of the Vernalisation Response

While a number of different plant species exhibit a similar phenotypical response to vernalisation conditions, the underlying mechanisms are very different.

1.3.1 *Arabidopsis thaliana* use FLC (and MAFs)

Arabidopsis is a genetic model organism for plant biology research and is the species where the genes involved in the vernalisation response were first characterised. In some *Arabidopsis* ecotypes, such as C24, vernalisation accelerates flowering (Simpson and Dean 2002). The vernalisation response is characterised in *Arabidopsis* by a reduced rosette leaf number and an early elongation of the inflorescence stem, both indicators of an early transition to reproductive development (Nordborg and Bergelson 1999). While vernalisation is not essential to trigger flowering in *Arabidopsis* (a facultative vernalisation response), vernalisation triggers flowering earlier than in an unvernalised plant, as evidenced by growing vernalised and non-vernalised *Arabidopsis* plants of the same ecotype grown alongside one another (Levy et al. 2002).

The key regulator of the vernalisation response in *Arabidopsis* is the MADS-box transcription factor FLC (Fig. 1.3a). High levels of FLC protein accumulate during vegetative growth in pre-cold periods (i.e. autumn) to repress flowering. During vernalisation, *FLC* expression is stably repressed, allowing the transition of *Arabidopsis* to flowering in spring, as temperatures rise and day length increases. During meiosis, DNA is recombined into male and female haploid gametes and the epigenetic regulation that silences the expression of *FLC* is released. After fertilisation and seed set, *Arabidopsis* progeny once again produce high levels of FLC and require vernalisation exposure to once again remove the molecular blocks that prevent downstream transcripts and proteins from triggering stem elongation and flowering (Sheldon et al. 2008). This is evidenced by *Arabidopsis* flc mutants harbouring non-functional FLC, mutant lines that transition to flowering earlier than wild-type *Arabidopsis* without the need to be exposed to vernalisation (Michaels and Amasino 2001).

Chromatin modifications regulate *FLC* expression via the Plant Homeo Domain Polycomb Repression Complex 2 (PHD-PRC2), a protein complex that is responsible for the epigenetic regulation of a number of critical plant systems, not just vernalisation (Gehring 2013). During vegetative growth, the PHD-PRC2 complex remains bound to the *FLC* locus (Köhler and Villar 2008). The presence of the PHD-PRC2 complex maintains *FLC* in an open conformation via histone H3 acetylation, loosening the interaction between nucleosomes and *FLC*, allowing the transcriptional machinery to access and promote *FLC* expression (De Lucia et al. 2008). During vernalisation in *Arabidopsis*, histone H3 is deacetylated and H3K9 and H3K27 are trimethylated across the *FLC* locus (Sung and Amasino 2004a). Further, the *FLC* promoter region is also

demethylated at H3K4 (Finnegan et al. 2005), blocking the transcriptomic machinery from accessing the *FLC* locus to repress *FLC* expression (Finnegan and Dennis 2007). This epigenetic repression of *FLC* is stable and irreversible, ensuring that the transition of vernalised *Arabidopsis* from vegetative to reproductive development is permanent (Levy et al. 2002). Mutations in the genes that encode the PHD-PRC2 complex proteins, such as VERNALISATION INSENSITIVE 3 (VIN3; Wood et al. 2006; Sung and Amasino 2004b), interfere with the action of this complex and negate its ability to regulate *FLC* expression by removing the ability to trigger the transition to flowering. The full extent of the effect of mutations on various proteins within the PHD-PRC2 complex has been the focus of other research (Levy et al. 2002; Mylne et al. 2006; Sung et al. 2006; Greb et al. 2007), but is beyond the scope of this introduction.

In a vernalisation sensitive cultivar of *Arabidopsis*, *CONSTANS* (*CO*) is expressed when exposed to long day conditions, which in turn expresses the two transcripts *FLOWERING LOCUS T* (*FT*) and *TWIN SISTER OF FT* (*TSF*). But when FLC is present, such as during vegetative growth pre-exposure to vernalisation, the expression of *FT*, *TSF* and *SUPPRESSOR OF OVEREXPRESSION OF CONSTANS 1* (*SOC1*) are blocked (Sheldon et al. 2000). After vernalisation, *FT* is expressed in leaf tissue, with FT travelling through the phloem to the shoot apical meristem (SAM; Corbesier et al. 2007; Jaeger and Wigge 2007). The SAM is composed of region-specific pluripotent stem cells which slowly divide and differentiate into the various progenitor cells necessary for vegetative growth (Fletcher 2002). After floral induction, i.e. conditions suitable for the plant to transition from vegetative to reproductive development, the SAM pluripotent stem cells differentiate into the progenitor cells necessary for floral tissue and structural development. Once FT arrives in the SAM, it interacts with FD (a basic-leucine zipper (bZIP) transcription factor; Abe et al. 2005) and triggers flowering via the expression of gene networks, including *LEAFY* (*LFY*) and *APETALA1* (*AP1*), two primary mediators of floral apical meristem development (Amasino 2004), as well as *SOC1* (Lee and Lee 2010), are all up regulated. Stout (1945) showed that in *Beta vulgaris* (Sugar Beet), grafting SAM containing vernalised scions onto non-vernalised root stock resulted in an early flowering phenotype, while grafting a non-vernalised scion onto vernalised root stock resulted in a late flowering phenotype. Wellensiek (1962) showed a similar effect in *Lunaria biennis* (Moon Wort). This indicates that while the entire plant may be exposed to vernalisation conditions, it is only pluripotent cells in the SAM that undergo a transition from the vegetative growth stage to the reproductive growth habit after exposure to vernalisation conditions. Presumably, a similar effect would be seen in *Arabidopsis*.

Natural allelic variants and targeted mutations in *Arabidopsis* have introduced variation in how individual ecotypes respond to cold and, therefore, variations in flowering time. In the *Arabidopsis* ecotype 'Cape Verde Islands 0' (*Cvi*-0), a natural variation that results in lower expression of the *FRIGIDA* (*FRI*) gene, which, in turn, promotes earlier

flowering (Werner et al. 2005). Conversely, mutants that inhibit the expression of *VIN3* lack the ability to detect cold. Therefore, these mutant lines no longer respond to vernalisation, resulting in delayed flowering regardless of the mutant's exposure to vernalisation conditions (Sung and Amasino 2004a).

Arabidopsis encodes five additional *FLC* homologues that are also regulated by vernalisation. Expression of *MADS AFFECTING FLOWERING 1* (*MAF1*) to *MAF4* are down regulated post vernalisation, whereas *MAF5* expression is up regulated (Ratcliffe et al. 2003). *MAF2*, like *FLC*, is expressed at high levels before exposure to vernalisation conditions and repressed after exposure. However, unlike *FLC*, *MAF2* requires a greater length of time exposed to vernalisation conditions to be repressed to the same degree as *FLC* (Airoldi et al. 2015). Further, *Arabidopsis* mutants without a functioning MAF2, but still able to encode a functional FLC, are unable to repress flowering at low but non-vernalising temperatures (16°C), but retain the ability to respond to vernalisation conditions by flowering early. Together, these data indicate that while FLC remains the primary regulator of flowering time in the vernalisation response, it is not the sole regulator of flowering time in *Arabidopsis*, as MAF2 also inhibits flowering at low non-vernalised ambient temperatures. Finally, it is unclear if the PHD-PRC2 complex also binds to and regulates the expression of these additional *MAF* loci during vernalisation, or alternatively, if there are additional and distinct molecular mechanisms that interact with these *FLC*-like genes.

In the perennial *Arabis alpina*, a close relative of *Arabidopsis*, the MADS-box gene *PERPETUAL FLOWERING 1* (*PEP1*) is functionally similar to *FLC*, in that PEP1 is an inhibitor of flowering. Because perennials flower every year, flowering needs to be repressed in winter and restored with warmer and longer days in a cyclical fashion. To accommodate this, the inhibitory mechanism of PEP1 is transient. During vernalisation conditions, like *FLC*, *PEP1* expression decreases to a level that is inversely proportional to the time of cold exposure. When warmer conditions return, *PEP1* expression is once again promoted (Wang et al. 2009). While PEP1 is a MADS-box transcription factor, this transient behaviour differentiates it from FLC.

Caution must be taken if the assumption is made that all FLC and FLC-like proteins are critical to the vernalisation response. Texas Bluebell (*Eustoma* spp., in the order *Gentianales*) is an ornamental flowering plant native to the southern United States, Central America and the northern regions of South America. *Eustoma* have a similar vernalisation response to *Arabidopsis*, where the time to flowering is shortened inversely proportional to the length of cold exposure (Pergola 1992). *Eustoma grandiflorum* encodes homologues to *Arabidopsis* FLC, FT and SOC1 (*Eg*FLC-like (FLCL), *Eg*FT-like (FTL) and *Eg*SOC1-like (SOC1L) respectively). *EgFTL* and *EgSOC1L* appear to be functional homologues of their *Arabidopsis* counterparts with similar expression profiles (lowly expressed until exposed to vernalisation conditions). Where as *EgFLCL* is lowly

expressed during vegetative growth and only increases with the onset of vernalisation, the opposite expression profile that was observed for *Arabidopsis FLC* (Nakano et al. 2011). Curiously, when *EgFLCL* is transformed into *Arabidopsis* with a non-functional FLC, it has a restorative effect, with an expression profile similar to *AtFLC* (Nakano et al. 2011). So while *At*FLC might be the primary transcriptional mediator responsible for the vernalisation response in *Arabidopsis*, in other species, this may not be the case (see Section 1.3.3).

The mechanisms that underpin the vernalisation response are epigenetic in nature and are reset in the next generation. But how similar is the vernalisation response in species outside *Arabidopsis* species? Physiologically, species that share a facultative vernalisation response behave in a similar way to *Arabidopsis*, but the underlying molecular mechanisms directing these physiological responses are quite distinct.

1.3.2 Legumes use a Family of FTs

Several species of the Fabaceae (legume) family, namely pea (*Pisum sativum* L.; Reid and Murfet 1975), sweet pea (*Lathyrus odoratus*; Ross and Murfet 1986) and Lupins (*Lupinus albus*, *L. angustifolius* and *L. luteus*; Gladstones and Hill 1969; Landers 1995), also respond to vernalisation by transitioning to flowering early. Research in both pea and sweet pea describes a 'transmissible element' that, during vernalisation, reduces the presence of an inhibiting factor that delays flowering time. Like *Arabidopsis*, when vernalised pea and sweet pea scions were grafted onto unvernalised root stock, the scions flowered significantly earlier than unvernalised scions grafted onto vernalised root stock (Reid and Murfet 1975; Ross and Murfet 1986).

These molecular mechanisms have been further explored in the model legume *Medicago truncatula*. *Arabidopsis* encodes a single *FT*, however, *M. truncatula* encodes five *FT* loci: *FTa1*; *FTa2*; *FTb1*; *FTb2*; and *FTc* (Laurie et al. 2011). When the *FT* genic sequences from angiosperms are aligned, three clades are distinctly apparent. Conservation of the *FTa*, *FTb* and *FTc* clade is unique to legumes, as it is not present in any other angiosperm (Hecht et al. 2011). Of the five *FT* genes in *M. truncatula*, *MtFTa1* expression is only observed after vernalisation exposure (Jaudal et al. 2013). Similarly, over expression of *MtFTa1* correlates with very early flowering (Putterill et al. 2013). *Medicago* species also appear to lack orthologues for *Arabidopsis* FLC and MAF (Hecht et al. 2005), but do encode a *Mt*FRI-LIKE protein. When *MtFRI-LIKE* was transformed into *Arabidopsis*, flowering time was delayed in the resulting transformants, indicating that in *Arabidopsis*, *Mt*FRI-LIKE directs a similar functional role to *At*FRI, i.e. the promotion of *FLC* expression (Chao et al. 2013). Even in the absence of an *FLC* homologue, the *Fabaceae* are still capable of responding to vernalisation by decreasing the time to flowering. While the *MtFT* gene family appear to play a central role in this response, similar to *Arabidopsis*, the exact molecular mediators that trigger the vernalisation response in legumes are yet to be determined.

1.3.3 Sugar Beet uses BTC1

Sugar beet (*Beta vulgaris*) is extensively cultivated worldwide for its large, sucrose rich root organ. Unlike *Arabidopsis*, sugar beet has an absolute vernalisation response, requiring both vernalisation and increased day length to flower (Dijk et al. 1997). Because of this, breeders have selected for ecotypes that display vegetative and root growth maintenance phenotypes in order to maximise root yield (Owen et al. 1940). Wild sugar beet (*Beta vulgaris* spp. *maritima*), like *Arabidopsis*, has naturally occurring variants originating from warmer climates where, in addition to higher average temperatures, there is less variation in day length between winter and summer. These variants have a far less pronounced response to vernalisation conditions or fail to respond at all (Dijk et al. 1997). Conversely, naturally occurring northern ecotypes express an absolute vernalisation requirement phenotype. The molecular mechanisms underpinning the vernalisation response in sugar beet have identified two paralogous *FT* genes, *BvFT1* and *BvFT2*, that are controlled by BOLTING TIME CONTROL 1 (*Bv*BTC1, Fig. 1.3b) and are central to the regulation of flowering time (Pin et al. 2012). *Bv*BTC1 regulates the expression of *BvFT1* and *Bv*FT1 in turn regulates the expression of *BvFT2* (Pin et al. 2010). This mechanism of an antagonistic pair of FT proteins to regulate flowering, rather than a single FT protein (as observed in *Arabidopsis*), is specific to *B. vulgaris*. While an *At*FLC homologue has been identified in sugar beet (FLC-LIKE (*Bv*FL); Reeves et al. 2007), it is not involved in the vernalisation response (Vogt et al. 2014). This indicates that while *Arabidopsis* and sugar beet both express similar vernalisation response phenotypes, the molecular machinery and/or pathways that mediate the vernalisation response in *B. vulgaris* are distinct to those in *Arabidopsis*.

1.3.4 Cereals use VRN1, VRN2 and FT

Klippart (1857) observed that vernalisation resulted in early flowering in some wheat cultivars. Monocots, such as bread wheat (*Triticum aestivum*), barley (*Hordeum vulgare*) and *Brachypodium distachyon* (a model monocot) respond to vernalisation much like *Arabidopsis*, in that seedlings exposed to extended non-freezing cold followed by increasing day length and temperature flowered earlier than non-vernalised seedlings (Trevaskis et al. 2003; Oliver et al. 2009; Woods et al. 2014). However, cereal species have a vernalisation pathway distinct to that of *Arabidopsis*, and other dicots in general. In many flowering dicots, the regulation of flowering time is controlled by the expression or repression of a MADS-box gene, such as *FLC* or *FLC-like*, which in turn, at the protein level, regulates the expression of downstream targets such as *FT*. In cereals, *OS2*, a MADS-box gene and an ancient *FLC* orthologue (Ruelens et al. 2013), appears to be a downstream target in the flowering pathway rather than a causative regulator that is itself responding to vernalisation (Deng et al. 2015). Further, it is the interplay between three genes; *VRN1* (another MADS-box transcription factor), *VRN2* (which in cereals is distinct to *Arabidopsis VRN2*; Yan et al. 2004) and *FT* (Trevaskis et al. 2007a)

that collectively regulate the vernalisation response in a number of cereal species (Fig 1.3c). It is believed that the interplay of these three genes is consistent across all the Pooideae (McKeown et al. 2016).

VRN1 serves two purposes; as a key meristem identity gene and as a regulator of *VRN2* expression (Trevaskis et al. 2007b). Winter wheats require vernalisation for induction of the early flowering phenotype, otherwise a late flowering phenotype is expressed. Spring wheat varieties are naturally early flowering and therefore do not express a phenotype in response to vernalisation (Trevaskis et al. 2003). During vernalisation of winter barley, H3K27 demethylation and H3K4 trimethylation occur at the *VRN1* locus. This modifies local chromatin confirmation to open the *VRN1* locus for expression (Oliver et al. 2009). Similar to the chromatin modification of the *Arabidopsis FLC* locus, this epigenetic modification is stable. However, in winter barley, and contrary to the consequence of chromatin modifications surrounding *FLC*, conformational changes at the *VRN1* locus promotes *VRN1* expression rather than repressing it.

Similar to the *Arabidopsis* flowering pathway where FLC represses *FT* expression, VRN2 represses the expression of *FT* (Ream et al. 2014). Prior to vernalisation of winter cereals, the floral repressor OS2 is abundant, and functions together with VRN2 to maintain the cereal in vegetative growth (Greenup et al. 2010). After vernalisation, increased VRN1 abundance represses *VRN2* expression, and low VRN2 levels enables the expression of *FT*. Increased FT ultimately triggers the transition to flowering in winter barley (Trevaskis et al. 2006). High VRN1 levels post exposure to vernalisation also stably inhibits *OS2* expression, promoting the transition to flowering. This mechanism of VRN1 repression of *VRN2* expression is readily observed in spring cereals, as *VRN1* is expressed in spring varieties regardless of their exposure, or lack thereof, to vernalisation. Similar to *Arabidopsis*, there are naturally occurring cereal variants containing mutations to the *VRN1*, *VRN2* or *FT* loci which change the requirement for vernalisation to transition to the reproductive state (Yan et al. 2006). Following VRN1-mediated repression of *VRN2*, decreased VRN2 abundance promotes *FT* expression. FT subsequently interacts with VRN1 which, in addition to acting as a repressor of *VRN2*, is also a promoter of meristem identity and the developmental transition to flowering (Deng et al. 2015). It is only once *VRN2* is repressed and *FT* is expressed that cereals can transition from vegetative to reproductive development. Taken together, the lack of an FLC homologue (assuming that OS2, while a MADS-box gene, is not a functional homologue of FLC) and the dual functionality of VRN1 as both a meristem identity factor and as a repressor of *VRN2* expression, demonstrates a major genetic divergence of the cereals from dicot species in regards to their response to vernalisation.

1.3.5 What Regulates the *Asteraceae*?

While the bulk of vernalisation research has historically focused on the genetic model species *Arabidopsis* and important crop species such as cereals and sugar beet, there is comparatively little research into the vernalisation response of one of the largest and most diverse flowering plant families, the *Asteraceae*. Early research reported a vernalisation response in lettuce (*Lactuca sativa*), with vernalised lettuce seedlings flowering up to four weeks earlier than unvernalised lettuce (Fig. 1.4; Gray 1942; Warne 1947; Rappapport et al. 1956). Subsequent studies confirmed that lettuce indeed responds to vernalisation (Waycott 1995). This research additionally demonstrated that flowering occurred even without exposure to vernalisation conditions (a facultative vernalisation response), similar to *Arabidopsis*.

Chicory (*Cichorium intybus*) is an *Asteraceae* that responds to vernalisation. Similar to other species, there appears to be significant natural variation in the vernalisation response of chicory, with ecotypes showing both absolute and facultative vernalisation responses (Schittenhelm 2001). Examining vernalisation-related molecular mechanisms in chicory, *FLC-LIKE 1* (*CiFL1*), a MADS-box transcription factor with significant sequence homology to *Arabidopsis FLC*, is expressed during vegetative growth. Similar to the expression profile of *FLC*, *CiFL1* expression is repressed following exposure to vernalisation (Périlleux et al. 2013). Furthermore, when *CiFL1* was over expressed in *Arabidopsis*, the resulting progeny displayed a significant delay in the onset of flowering, regardless of vernalisation exposure. This indicates that at the molecular level, *Ci*FL1 is functionally similar to FLC. However, there is little information regarding the intronic structure of *CtFL1*, which is known to regulate the expression of *AtFLC* (Michaels et al. 2003). In the chicory plants that have been vernalised, and subsequently returned to warmer growth conditions, *CiFL1* expression is again enhanced, revealing that vernalisation-mediated repression of *CiFL1* expression is transient, with an expression profile closer to *AaPEP1* rather than that of *AtFLC*.

While an FLC homologue may be present in many species (Reeves et al. 2007), based on what is seen in *Eustoma*, there may be no functional homologue of FLC in *Asteraceae*. Similar to the use of *M. truncatula* as the model species for legume research, a model for examining the vernalisation response, or for the general characterisation of flowering time within the *Asteraceae* would be highly beneficial. The identification of a 'flowering model' within the *Asteraceae* would allow for the determination of the molecular mechanisms and triggers that influence flowering time in the largest family of flowering plant species.

Unlike for *Arabidopsis* and other major crop species, such as cereals, legumes and sugar beet, the genomic and transcriptomic information available for the *Asteraceae* is comparatively limited. For safflower, the resources available at the time of authoring this Thesis included unpublished expressed sequence tags in global sequence databases,

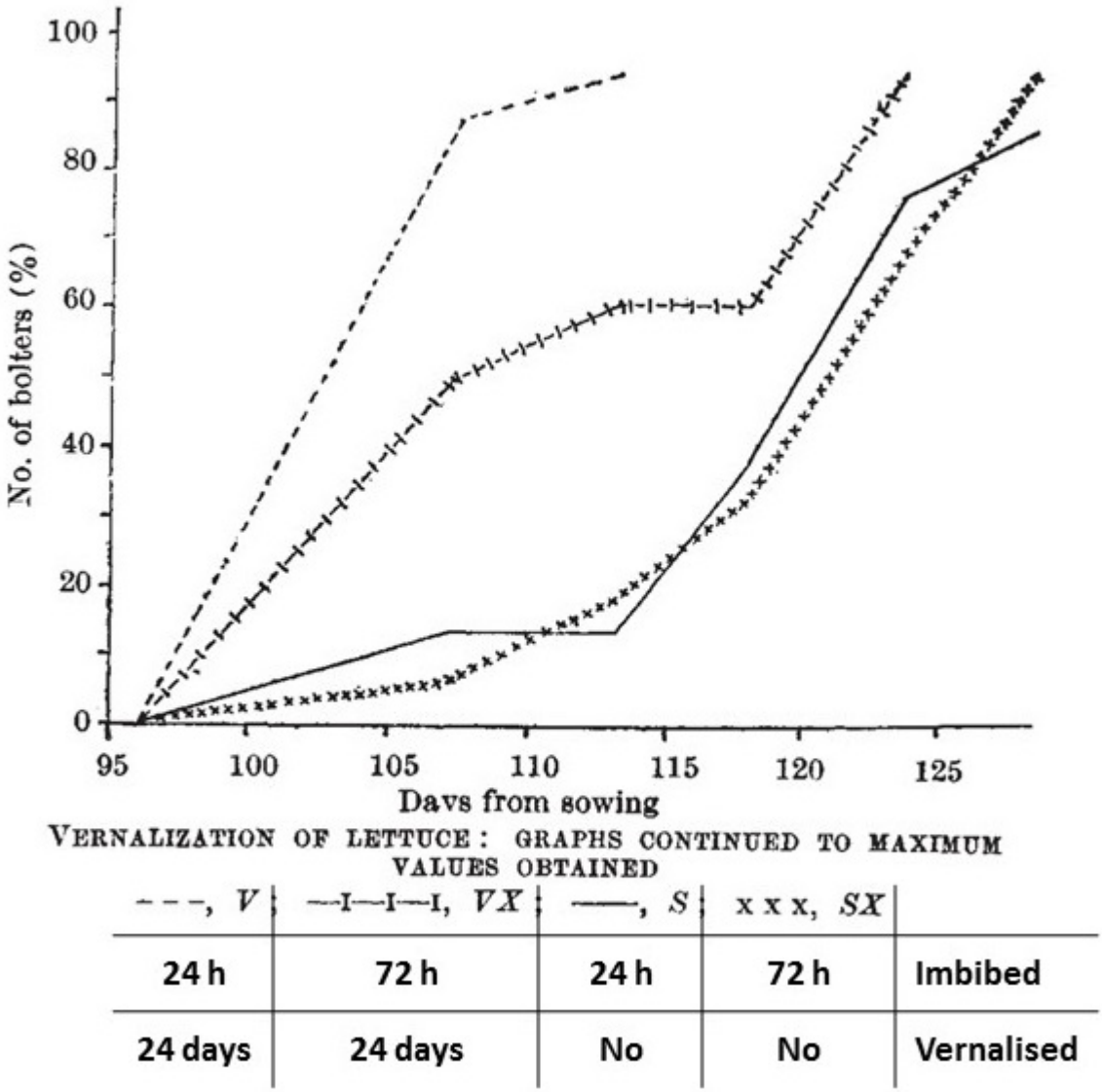


FIGURE 1.4: The effect of vernalisation on days to bolting in lettuce. Samples with a **V** have been vernalised for 24 days while those with an **S** have not. Samples with an **X** were imbibed for 72 hours, those without an **X** were imbibed for 24 hours. Modified from Warne (1947).

a number of analyses targeted at specific pathways (Li et al. 2012; Lulin et al. 2012; Cao et al. 2013; Liu et al. 2015), a safflower chloroplast (Lu et al. 2016) and a recently published SNP dataset covering just 15% of the genome (Bowers et al. 2016). There are few, if any, transcriptomic resources that exist which characterise the vernalisation response in safflower, or any aspect of the flowering response pathway in other species of *Asteraceae*.

1.4 Phylogenetic Analysis of Vernalisation Responsive Species

The divergence of monocots and dicots is estimated to have occurred approximately 180-220 million years ago (mya; Fig. 1.5; Wolfe et al. 1989). As the plant species diverged, mutation and speciation events would have occurred separately and independently of each other. This fits with the sequence homology but functional differences exist between some monocots and dicots. For example, the sequence homology between *ODDSOC2* (*OS2*) in wheat and *FLOWERING LOCUS C* (*FLC*) in *Arabidopsis thaliana* (*Arabidopsis*) imply that both are MADS-box transcriptional factors. But in terms of their functional role in the flowering pathway of these two species, *FLOWERING LOCUS T* (*FT*) is regulated by the presence of in *Arabidopsis* FLC (discussed below), whereas in wheat, the reverse is observed. *OS2* is the target of regulation by VERNALISATION 1 (VRN1) rather than being the regulator itself (Greenup et al. 2010; Deng et al. 2015).

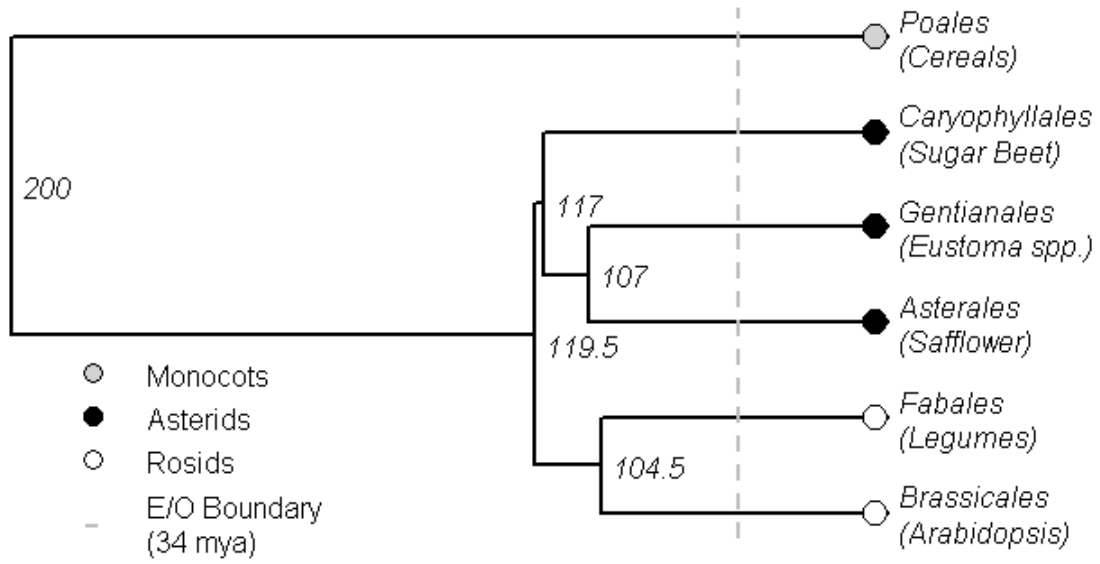


FIGURE 1.5: Phylogenetic tree of the six families of angiosperms investigated. Node labels indicate age of most recent common ancestor (mya). At the boundary marking the transition between the tropical Eocene and temperate Oligocene ages (the E/O boundary, estimated at 34 mya), a mass extinction event occurred where the global temperatures decreased and temperate climates emerged. Compiled from Wolfe et al. (1989), Wikström et al. (2001) and Smith et al. (2011).

Approximately 34 mya, there was a geological transition from the tropical Eocene age to the modern, temperate Oligocene age (Silva and Jenkins 1993; Speelman et al. 2009), known as the Eocene/Oligocene (E/O) boundary. Using marine temperatures as a proxy for land temperatures (Ivany et al. 2000; Wade et al. 2012), the drop in the average minimum winter temperatures beyond the E/O boundary resulted in a terrestrial mass extinction event (Prothero 1994). This, in part, could explain the distinct mechanisms of the vernalisation response. More specifically, novel but non-functional permutations of regulatory mechanisms could have been generated through natural mutation. As the

climate shifted to a lower mean temperature during the winter months, species that were unable to regulate flowering and adapt to the new, temperate climatic regions perished. Any remaining species that were able to delay the onset of flowering until warmer spring and summer conditions arrived would have thrived. This selection pressure would have continued throughout the newly temperate regions, with mild winters allowing for diversification of vernalisation mechanisms by allowing a greater quantity of plants to survive. The harsher winters would have tested the effectiveness of these newly diversified mechanisms, with only the fittest surviving in the Oligocene climate.

1.5 Further Questions on the Vernalisation Response in Safflower

In the vernalisation responsive plant varieties discussed above, *FT* and its variants are expressed in true leaves, with the downstream targets active in the SAM. While the downstream effects of extended cold exposure can be observed in vernalisation responsive phenotypes, the molecular mechanisms of how plants detect cold in the first instance is still poorly understood and difficult to elucidate. In all of the examples of vernalisation examined to date, a gene regulated by vernalisation has been repeatedly demonstrated. But the species specific molecular mechanism which triggers this effect remains to be identified. Helliwell et al. (2015) postulated that physical changes brought on by vernalisation conditions modulates the way that the DNA molecule mechanistically behaves in cells, essentially removing the natural elasticity of the DNA. Therefore, if the elastic nature of DNA allows an open conformation during the cold, it will remain open for a longer amount of time, allowing prolonged access to the site for the associated molecular machinery. Possibly, it is this lack of elasticity that allows access to the *AtFLC* locus by the PHD-PRC2 complex, or even genes that make up the PHD-PRC2 complex itself, such as *VIN3* and *VRN1*. This physical change to the elasticity of DNA and the resulting change to the conformational shape is another avenue of investigation to better characterise a common effect that vernalisation conditions have on all vernalisation responsive species.

Examining the phylogenetic tree for flowering plants that possess a vernalisation response reveals that the most recent common ancestor to the *Rosids* and the *Asterids* is estimated to have emerged at 120 mya (Wikström et al. 2001; Smith et al. 2011). The closest family to *Arabidopsis* is the *Fabaceae*, where an *AtFLC* homologue is absent. Perhaps the *Rosids*, containing the *Fabaceae* and the *Brassicales*, have diverged from other flowering plant clades in the way that vernalisation affects *FLC* expression, but the *Fabaceae* have diverged even further, shedding *FLC* altogether while still maintaining a functional vernalisation response.

The *Asterids* contain the ancient *Caryphyllales* family, the large and diverse *Asterales* and the *Gentianales*. The *Caryphyllales*, a primitive lineage of flowering plants (Wang 2010), diverging relatively shortly after the *Rosids*, approximately 117 mya, with the most recent common ancestor of the *Gentianales* and the *Asterales* estimated at 107 mya. Similar to what was seen in the *Fabaceae*, the *Caryphyllales*, which contain sugar beet, evolved a unique vernalisation response mechanism in the antagonistic *FT1* and *FT2* genes, controlled by the distinct BTC1. While the *Gentianales*, which include *Eustoma* species, have an *FLC-like* gene, the exact molecular mechanisms for the vernalisation response in the *Asterales* is currently unclear. Another area of future research could be to investigate whether members of the *Asterales* have genetic mechanisms resembling the vernalisation response of the *Gentianales*, with an FLC-like transcription factor regulating the expression of *FT-like*, or the *Caryphyllales*, which contains two antagonistic *FT* genes triggered by a non-FLC-like mechanism.

Many of these flowering plants display similar physiological responses to that of the *Arabidopsis* vernalisation response, but the underlying molecular mechanisms have diverged and evolved in different ways. During vernalisation in *Arabidopsis*, *FLC* expression is repressed via epigenetic modification of the *FLC* locus. Histone methylation and the associated condensing of the surrounding chromatin blocks the transcriptional machinery from accessing the *FLC* locus. The repression of *FLC* expression removes the molecular block that allows the expression of *FT* and other genes downstream of FLC in the *Arabidopsis* flowering pathway. In barley however, repression of *VRN2* expression by VRN1 promotes the expression of *FT*, allowing barley to transition from vegetative to reproductive development. Sugar beet utilises a different mechanism again, encoding two counteracting *FT* homologues, *BvFT1* and *BvFT2*, with opposing functional roles in its flowering pathway. It has also been shown that while both *Eustoma* and *Arabidopsis* encode FLC homologues, increased levels of *Eg*FLCL promotes *EgFTL* expression in *Eustoma*, whereas in *Arabidopsis*, decreasing levels of *At*FLC promote *AtFT* expression. Together, this data brings to light a further question: is the behaviour of FLC and its effect on *FT* expression unique to the *Brassicaceae*? Are the molecular mechanisms that regulate the vernalisation response in the *Brassicaceae*, in fact, the exception, rather than the rule?

1.6 Next Generation Sequencing in the Context of the Vernalisation Response in Safflower

In the last 15 years, Next Generation Sequencing (NGS) has greatly expanded the quantity and quality of genetic information available. Today, genomic, transcriptomic and proteomic information is available at levels never seen before. Furthermore, this data now requires significantly less time and a fraction of the resources to generate (Wetterstrand 2014). The hope was that this expansion of data generation capability

would quickly equate to an advanced understanding of both fundamental and complex genetic pathways. NGS technology has provided insights into well characterised mechanisms, such as regulation of the flowering pathway. In spite of this, investigations into these datasets has failed to shed light on more fundamental questions, such as the mechanisms plants use to detect exposure to cold in the first instance.

In this context, the research plan of this PhD Thesis was conceived to expand and explore the breeding resources and genomic tools available for accelerated breeding of safflower. This Thesis focuses almost entirely on vernalisation and flowering of safflower as a trait. While this phenotype has been reported in *Asteraceae* such as lettuce (*L. sativa*; Warne 1947), current literature only refers to 'winter hardy' safflower, with no mention of a vernalisation response (Johnson et al. 2006; Johnson and Li 2008), presumably due to nearly all commercially grown safflower being spring cultivars. With the vernalisation response as the focal point, this project developed a range of additional generic tools for advanced breeding of this ancient crop. Flowering time and vernalisation has been intricately linked to yield in other important crops (Trevaskis et al. 2007a; Deng et al. 2015). Numerous reviews (Amasino 2005; Dennis and Peacock 2009; Putterill et al. 2013) have shown for a range of crops, it is crucial to accurately match flowering time with the sowing date, the prevailing temperature and day length to improve yields. This focus on flowering time and vernalisation in safflower can be placed in the context of improving yield and profitability of safflower across a wider range of growing conditions, not only in Australia, but across the world.

1.7 Summary

While wheat was the first crop where the vernalisation response was observed, *Arabidopsis* was the first species where vernalisation was molecularly characterised, revealing *At*FLC as the primary regulator of flowering time. As research into vernalisation expanded to other angiosperm species, it became readily apparent that *Arabidopsis* and the *Brassicaceae* are unique in their repression of *FLC* in response to vernalisation. Each family has evolved its own molecular mechanisms to respond to vernalisation. While each mechanism is unique to each individual plant family, the resulting phenotype is the same, i.e. the longer the exposure to conditions of non-freezing cold, the less time the plant remains in vegetative development due to an early transition to reproductive growth. Because of this uniqueness, expanding research into the vernalisation response in other angiosperm families is needed to develop genetic markers specific to the species or family of interest. These markers could then be used to guide breeding, ultimately increasing the yield and providing more flexibility in the planting schedules for an ever widening array of cultivated crop species.

The initial focus of this PhD Thesis was the identification of a 'winter hardy' safflower variety that is demonstrated to be responsive to vernalisation conditions. By crossing this vernalisation responsive 'winter hardy' safflower with a vernalisation unresponsive spring safflower, a genetically segregated population was generated. This confirmed the heritable nature of the vernalisation responsive trait in safflower. Using this population, as well as several NGS techniques, a series of datasets were generated that allowed a thorough insight into the genetic and molecular basis of the vernalisation response in safflower. These insights are discussed relative to the knowledge of the vernalisation response in other flowering plants.

Chapter 2

Physiology of the Vernalisation Response in Safflower Varieties

2.1 Outline

It has been previously reported that there are a number of 'winter hardy' varieties of safflower (Johnson et al. 2006), but no attempt is made to characterise whether any of these varieties exhibit a true vernalisation response. The aim of this experimental chapter was to take one of the 'winter hardy' safflower cultivars and examine it to determine if there was a response to vernalisation conditions, similar to what is seen in other plant species. If this cultivar did respond, further experimentation would identify and characterise the parameters and temperature cues that triggered this response. These would then be compared against the list of vernalisation response characteristics, as described in Chapter 1, to determine if they could be classified as a vernalisation responsive safflower cultivar. By characterising the physiological response of safflower to vernalisation, a more detailed understanding of the potential agronomic benefits of the vernalisation responsive safflower cultivars could be obtained with the potential for this trait to be incorporated into future commercial breeding lines.

2.2 Materials and Methods

2.2.1 Cultivars

Two safflower cultivars were used for this experiment. S317, a commercially grown elite cultivar sourced from CSIRO (referred to hereafter as 'spring safflower') and C311, a wild variety imported from Eastern China (referred to hereafter as 'winter safflower'; CSIRO Plant Industry Accession: 154311; USDA-ARS accession number: WSRC03; described as 'winter hardy' by Johnson and Li (2008)). To ensure there was no variation in the genetics of the winter safflower line, seeds from the imported winter safflower, Q_0, were grown in a glasshouse, producing the Q_1 generation. Seeds harvested from a number of single Q_1 plants were propagated and grown, producing the Q_2 generation. Unless otherwise stated, experiments on winter safflower cultivars were conducted using seed sourced from these single seed descent Q_2 lines.

2.2.2 Growth Conditions

2.2.2.1 Breaking Seed Dormancy

To break the seed dormancy, safflower seeds were placed into a volume of distilled H_2O that equated to twice the weight of the seeds, and incubated overnight at 4°C. Two different techniques were used for germination.

2.2.2.2 Petri Dish Germination and Vernalisation

Twenty seeds were placed in a 10 cm plastic Petri dish lined with a piece of sterile tissue paper moistened with $1.0\ \mathrm{mL}$ of distilled H_2O. The Petri dishes were sealed and placed into a 28°C incubator overnight. Germinated seeds were then transferred to seedling containers filled with soil, comprised of 30% river sand, 25% perlite, 25% vermiculite and 20% recycled and composted soil (referred to hereafter as 'Maria's Mix' soil). Ungerminated seeds were discarded. Unless the seeds were to be vernalised, they were placed into a growth room and exposed to 16 hours of fluorescent light at approximately 450 $\mu\mathrm{M/m^2s}$ at 28°C. After two weeks in the growth room, seedlings were transplanted into 20 cm pots containing 73% recycled and composted soil and 27% perlite (referred to hereafter as 'Cotton Mix' soil). Petri dishes containing seeds to be vernalised were wrapped in foil and refrigerated at 4°C for four weeks before being transferred to the glasshouse. After the 4 week treatment, the vernalised seeds were transferred to 20 cm pots of 'Maria's Mix' soil and grown in either the glasshouse or growth cabinets for the remainder of the growing period.

2.2.2.3 Measuring Cylinder Germination and Vernalisation

Seeds were placed in a test tube of distilled H_2O and aerated overnight at room temperature, changing the water once. After aeration, germinated seeds were then transferred to 'Cotton Mix' soil and any ungerminated seeds discarded. Seedlings to be vernalised were imbibed and germinated using this method were planted in seedling trays of 'Cotton Mix' soil, covered in foil and placed into a large growth room set between 4°C and 6°C for the vernalisation period. Seedlings were then transferred to the glasshouse or growth cabinets for the remainder of the growing period.

2.2.2.4 Glasshouse Growth Conditions

Temperatures in the glasshouse were set to a 26°C to simulate daylight temperature and 18°C to simulate nighttime temperature. Incandescent lights emitting between 200 $\mu\mathrm{M/m^2s}$ and 300 $\mu\mathrm{M/m^2s}$ were used to provide extended light exposure when necessary, exposing plants to approximately 16 hours of light and 8 hours of darkness to simulate long day conditions.

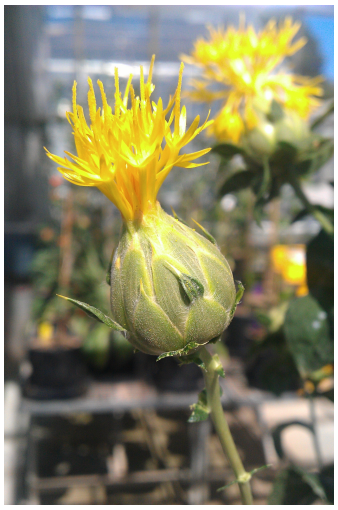
2.2.2.5 Growth Cabinet Conditions

To confirm the need for long day conditions to trigger elongation and flowering post vernalisation, both safflower plants were exposed to long and short day conditions. To simulate long day conditions, the plants in the growth cabinets were exposed to approximately 450 $\mu\mathrm{M}/\mathrm{m}^2\mathrm{s}$ of light for 16 hours and 8 hours of darkness. When simulating short day conditions, plants in the cabinets were exposed to approximately 450 $\mu\mathrm{M}/\mathrm{m}^2\mathrm{s}$ of light for 8 hours and 16 hours of darkness. Temperatures in cabinets were set at a 26°C to simulate daylight temperature and at 18°C to simulate nighttime temperature.

2.2.3 Generation of Crossing Population

A crossing population was created, as described in Mündel and Bergman (2010), to examine the vernalisation responsive phenotype, using late elongation as a proxy for vernalisation. Approximately two to three hours after sunrise (glasshouse plants) or post turning the light source on (growth cabinet plants), candidate floral heads from individual flowers that had begun to emerge were selected in both spring and winter safflower varieties (Fig. 2.1a). All leaves from the stem of the selected flowering head were removed (Fig. 2.1b). Then the outer calyx was removed (Fig. 2.1c) to expose the flowers within the selected flowering head. Any individual flowers that had opened, even partially, were removed at this stage (Fig. 2.1d). Using a sharp dissection needle, the corolla tube of the floret was cut approximately 2 mm above the junction between the style and ovary (Fig. 2.1e). The cut corolla tube was slid off the stigma, taking care not to damage the stigma in the process (Fig. 2.1f). This was repeated with as many florets as possible, with any florets harbouring stigmas suspected of being damaged during the corolla tube removal process discarded (Fig. 2.1g). Once the preparation had been completed, each flowering head was bagged to isolate it from any stray pollen flow from other plants. After 24 hours, if the stigmas had extended and were brightly coloured (Fig. 2.1h), the emasculated flower was assumed to be ready for pollen reception, and pollen from a donor plant was deposited on the stigma. The bag was replaced to ensure no stray pollen contaminated the crossed head. The first generation of these plants, F_1, were grown to maturity in the glasshouse as described in Section 2.2.2.4. The second generation, F_2, were planted in June and cultivated in an enclosed field at the CSIRO Black Mountain site in Acton, Australia (latitude:-35.26927938, longitude: 149.1112002, elevation: 603 m). Where at least 50 seeds were produced, 24 of the F_3 seeds from every surviving and seed producing F_2 plant were grown for four weeks in the glasshouse. Only F_3 families where at least 15 of the 24 planted seeds grew into seedlings were used to characterise this crossing population.

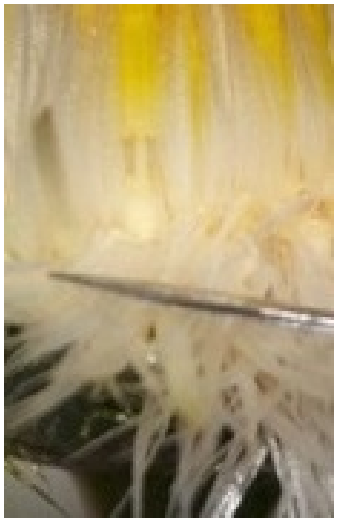
(a) A candidate safflower head



(b) Leaves are removed from the stem

(c) Calyx is removed, exposing flowers

(d) Any opened flowers are removed



(e) Incision made, cutting the corolla tube

(f) Corolla tube is slid off, exposing the stigma

(g) A candidate head prepared for crossing

(h) A candidate head receptive to pollen

FIGURE 2.1: Procedure for crossing safflower.

2.2.4 Characterisation of Vernalisation and Day Length Effect

Spring and winter safflower seeds were germinated using the Petri dish method (section 2.2.2.2). One plate of each cultivar was vernalised for four weeks with the unvernalised plate germinated at the end of the vernalisation period. Germinated seeds from both the vernalised and unvernalised cultivars were split into long and short day growth period groupings. The resulting plants were cultivated in temperature controlled growth cabinets until the first flower emerged on any plant. Independent t-tests ($\alpha = 0.05$) were used to determine if there was a significant difference between the vernalised and unvernalised plants as well as between spring and winter safflower cultivars.

2.2.5 Time to Vernalisation Saturation

Fifty seeds of both spring and winter safflower underwent 'cold break' dormancy (section 2.2.2.1) and were germinated in measuring cylinders (section 2.2.2.3). Seeds were then exposed to vernalisation conditions for periods of increasing length and cultivated (section 2.2.4). At two to three day intervals, two random seedlings were transferred to 20 cm pots containing 'Cotton Mix' soil and exposed to long day growth conditions (section 2.2.2.5). Plants were grown until the first flower emerged along the primary stem, at which point the plant was scored as 'flowered'.

2.2.6 Optimum Vernalisation Temperature

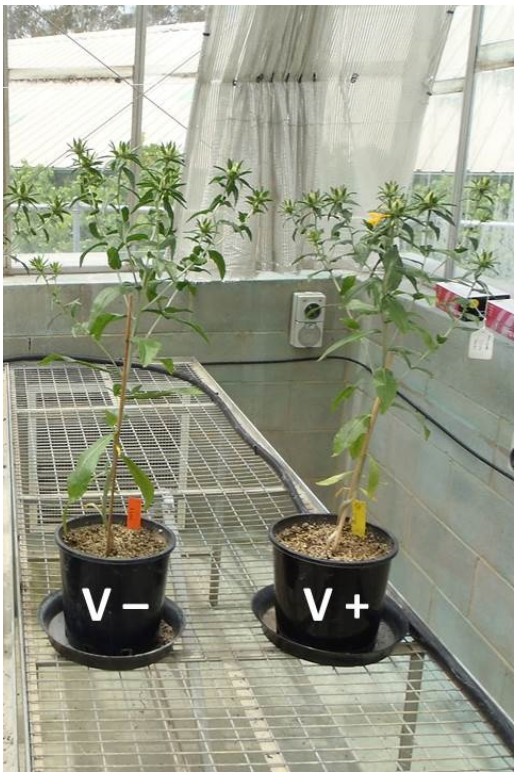
One hundred seeds of both spring and winter safflower underwent cold break dormancy (section 2.2.2.1) and were germinated in measuring cylinders (section 2.2.2.3). Approximately five cm of 'Cotton Mix' soil, 1 cm of cotton wool and 2 mL of Thymol [1.4 g/L] (as an antifungal treatment) were added to 15 mL blue capped tubes (in that order) prior to the addition of two seeds, either two spring or two winter seeds per tube. Next, six tubes of these spring and winter safflower seeds were evenly but randomly distributed into five different temperature blocks, set to 0°C, 4°C, 8°C, 12°C and 16°C, so each block had tubes which contained an equal number of spring and winter safflower seeds. Tubes were subjected to these five temperatures for four weeks before seeds were transferred to 20 cm pots containing 'Cotton Mix' soil. The seedlings were cultivated under long day conditions in temperature controlled growth cabinets until the first flower on the primary stem emerged, after which the plant was scored as 'flowered'. Plants from each temperature were analysed using a single factor ANOVA ($\alpha = 0.05$) with a TukeyHSD post-hoc test ($\alpha = 0.05$) used to compare the results of each temperature treatment.

2.3 Results

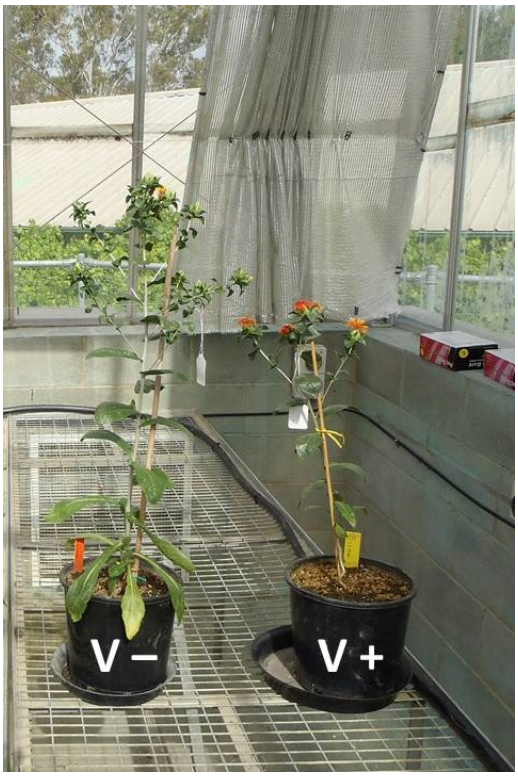
2.3.1 Initial Characterisation of the Vernalisation Response in Winter Hardy Safflower

The 'winter hardy' safflower cultivar was examined in parallel to a spring cultivar. After germination and 4 weeks of vernalisation at 4°C, the spring safflower seeds continued to grow and did not appear to respond to the vernalisation conditions. In the winter cultivar, however, almost no difference in growth rate was observed between plants from the vernalised and unvernalised seed (Fig. 2.2). As the plants continued to grow and mature, little phenotypical difference in growth behaviour between vernalised and unvernalised spring safflower was observed (Fig. 2.3a), but a substantial difference between the vernalised an unvernalised Q_0 single seed descent winter safflower (lines 3, 5 and 6; Fig. 2.3b, Fig. 2.3c and Fig. 2.3d respectively) was evident. In contrast, there was little difference observed between the three assessed Q_0 lines.

FIGURE 2.2: Seeds of both spring (left) and winter (right) safflower. Seeds above the horizontal line were imbibed and germinated using the measuring cylinder method, then immediately sown, seeds below the horizontal line were imbibed and germinated using the measuring cylinder method then vernalised for 4 weeks before sowing. While the metabolism was slowed in the vernalised spring safflower imbibed seeds they still grew while under vernalisation conditions, whereas the winter safflower seeds grew very little during the vernalisation conditions.



(a) Spring safflower.



(b) Winter safflower, single line 3.

(c) Winter safflower, single line 5.

(d) Winter safflower, single line 6.

FIGURE 2.3: Spring safflower (a) and winter safflwoer single seed descent lines (b, c and d) grown until first flowers emerged in either plant. V+ indicates the plant has been vernalised. V- indicates the plant has not been vernalised. Flowers in the vernalised and unvernalised spring safflower and vernalised winter safflower cultivars flowered after approximately seven weeks.

Two sample t-tests for population means compared flowering time between different safflower cultivars in the Q_0 population (Fig. 2.4). There was insufficient evidence of a significant difference in the time to flowering between vernalised and unvernalised spring safflower ($t = -1.25$; $df = 13.91$; $p - value = 0.23$) or the time to flowering between vernalised winter safflower and spring safflower ($t = -0.14$; $df = 15.51$; $p - value = 0.89$). There was, however, a significant difference in the time to flowering between vernalised and unvernalised winter safflower ($t = -3.97$; $df = 16.40$; $p - value = 0.00$) and between unvernalised winter and spring safflower ($t = -3.97$; $df = 11.72$; $p - value = 0.00$).

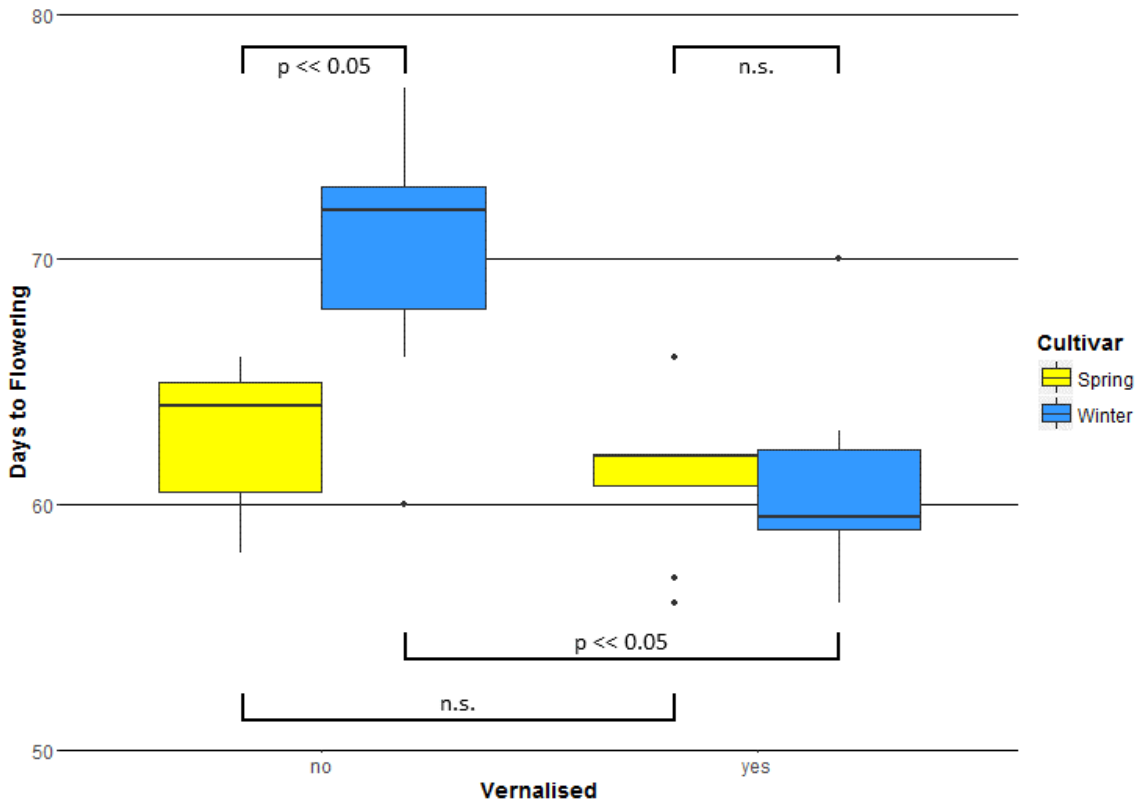


FIGURE 2.4: The (Q_0) generation of safflower plants (winter and spring safflower), grouped by vernalised and unvernalised. There was a significant difference in flowering time between unvernalised and vernalised winter safflower, and in the flowering time of unvernalised winter and spring safflower ($\alpha = 0.05$).

The progeny of the next generation of winter and spring safflower (Q_1) were examined to further characterise the vernalisation response in these cultivars (Fig. 2.5). Unfortunately, due to time and glasshouse space constraints and problems germinating this specific batch of safflower seeds, three biological replicates for every line, cultivar and vernalisation condition was not possible. For the three different winter safflower lines, this was not a problem, as the three different lines show comparable similar means and variances, which are all significantly different from the spring safflower line.

There was a significant difference in the number of leaves that developed on the primary stem ($t = -22.37$; $df = 34.30$; $p - value = 0.00$) and the final height of the plants ($t = -13.76$; $df = 14.34$; $p - value = 0.00$) when vernalised winter safflower was compared to unvernalised winter safflower. A significant difference in was again observed for the time to flowering ($t = -20.01$; $df = 33.36$; $p - value = 0.00$) between the vernalised and unvernalised winter safflower. Conversely, there was very little difference observed in these characteristics when comparing vernalised and unvernalised spring safflower (Table 2.1).

TABLE 2.1: Expressed phenotypes and growth attributes for vernalised and unvernalised Q_1 safflower, for both spring safflower and three single seed descent winter safflower lines. 'n' is the number of individual plants tested for each line, μ is the mean and σ is the variance of each measurement. Letters after each attribute group them together as being similar.

Winter, Line 3	Vernalised, n=2		Unvernalised, n=12	
	μ	σ	μ	σ
Days to Flowering	57^{a}	0	74^{b}	4.32
Leaf Number (Stem)	14^{q}	1	38^{r}	5.28
Height (mm)	493^{x}	27.5	742^{y}	60.29

Winter, Line 5	Vernalised, n=2		Unvernalised, n=9	
	μ	σ	μ	σ
Days to Flowering	58^{a}	1	75^{b}	1.87
Leaf Number (Stem)	15^{q}	0.5	38^{r}	3.82
Height (mm)	493^{x}	27.5	885^{y}	46.67

Winter, Line 6	Vernalised, n=2		Unvernalised, n=11	
	μ	σ	μ	σ
Days to Flowering	59^{a}	0.5	77^{b}	4.58
Leaf Number (Stem)	13^{q}	1	35^{r}	5.00
Height (mm)	475^{x}	15	858^{y}	94.83

Spring Safflower	Vernalised, n=3		Unvernalised, n=1	
	μ	σ	μ	σ
Days to Flowering	63^{a}	0	67^{c}	0
Leaf Number (Stem)	22^{q}	1.25	26^{q}	0
Height (mm)	637^{z}	44.97	685^{z}	0

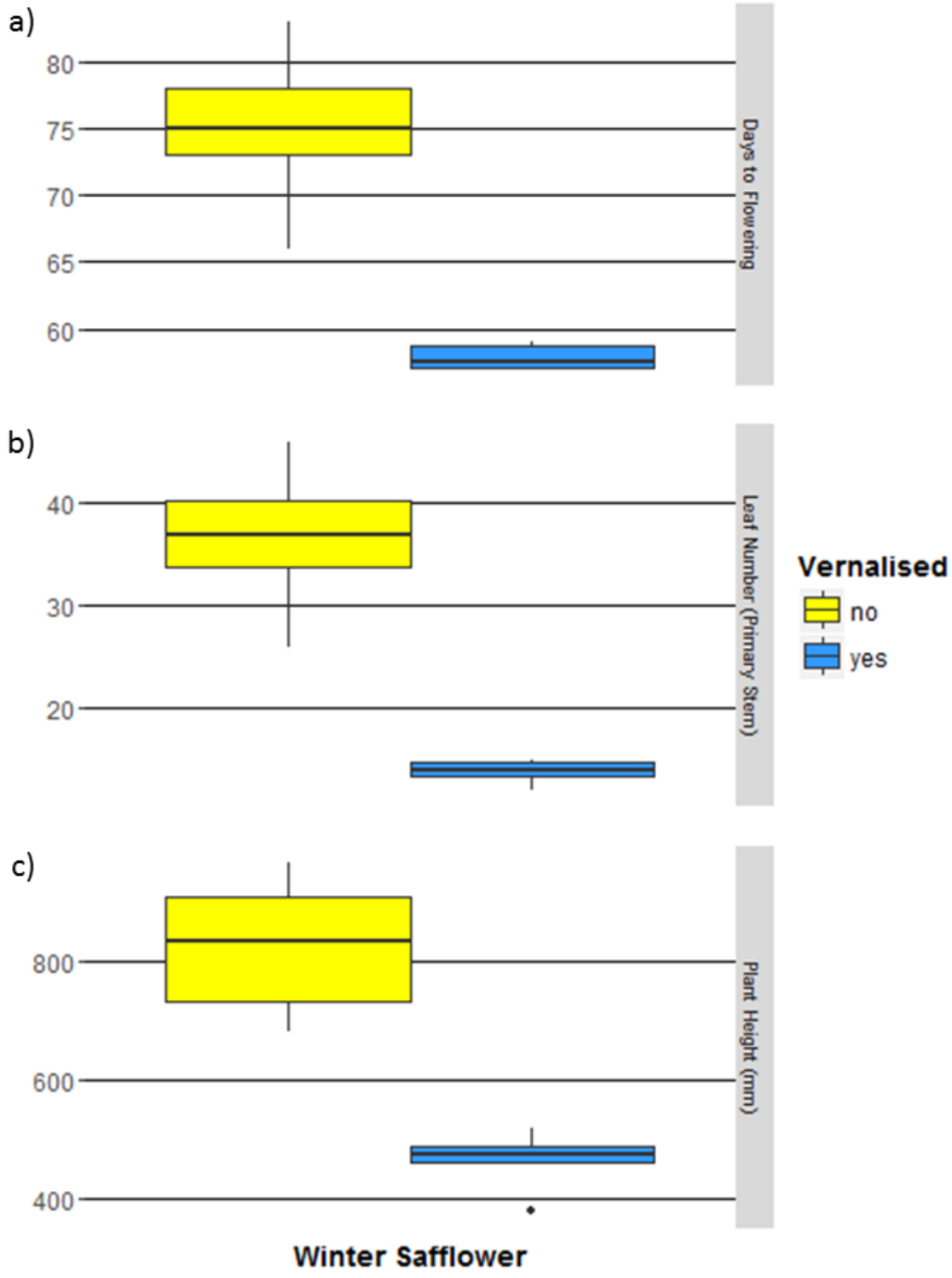


FIGURE 2.5: Characterisation of aspects of the vernalisation response in the Q_1 generation of winter safflower. Panel (a) is the number of days to flowering, panel (b) is the number of leaves counted on the primary stem and panel (c) is the height of the plant (soil line to highest point on the primary stem) in mm. For each aspect tested, there was a significant difference between the vernalised and unvernalised winter safflower ($p < 0.05$).

2.3.2 Resetting of the Vernalisation Response in Safflower

Interestingly, it was observed in each of the single seed descent lines that the exposure of parent plants to vernalisation conditions had very little, if any, observable influence on the growth behaviour of the progeny plants. When seeds from a Q_1 plant were sown (Q_2 seeds), a vernalisation treatment was required to express an early flowering phenotype in these Q_2 plants, regardless of whether the Q_1 plant had been vernalised or not. It was unknown whether vernalisation of a Q_1 spring safflower seedling affected the Q_2 generation, as no vernalisation response was observed. No statistical analysis of this observation was undertaken.

2.3.3 Vernalisation Exposure Timecourse to Determine the Saturation Point of the Vernalisation Response

When the vernalisation temperature remained fixed at 4°C while the length of time that the safflower varieties were exposed to 4°C was extended. The longer time exposed to vernalisation conditions caused no substantial change in the time to flowering in spring safflower ($\mu = 49$ days, $\sigma = 4.42$ days). But in winter safflower, there was a large overall reduction in the time to flowering ($\mu = 51$ days, $\sigma = 13.19$ days), which was significantly different ($t = -2.39$; $df = 25.07$; $p - value = 0.02$) when compared to spring safflower. In winter safflower, as the time exposed to the vernalisation treatment was extended, the time to flowering decreased (Fig. 2.6) in an asymptotic pattern. The effect of vernalisation on the time to flowering was most pronounced at the earlier time points, with the time to flowering reduced by 25 days after the seeds were exposed to just 10 days of vernalisation at 4°C. Extending exposure to this temperature for longer than approximately 2 weeks resulted in minimal observable decrease in the time to flowering in winter safflower.

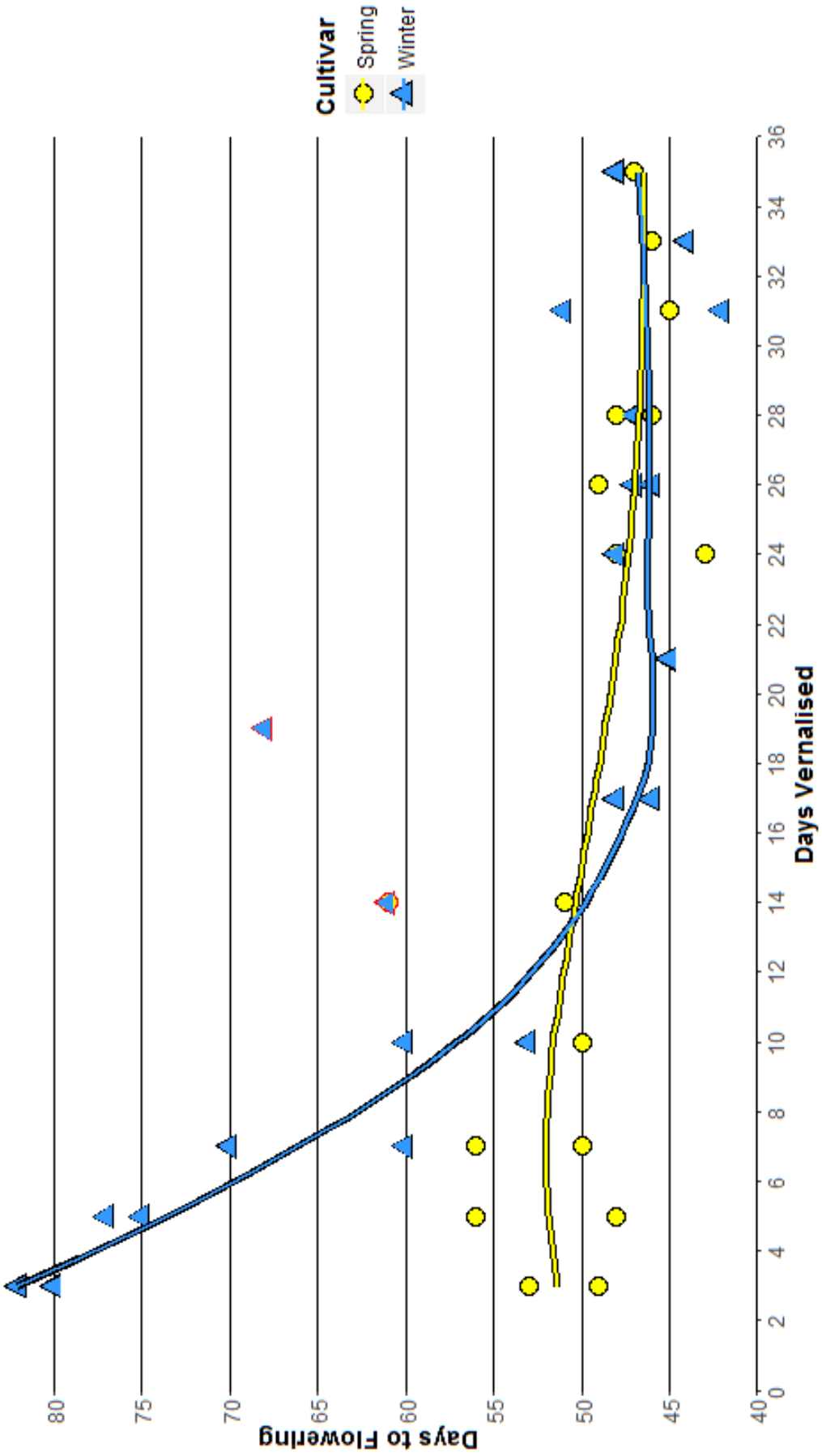


FIGURE 2.6: As the duration of vernalisation exposure is increased, there was a significant difference in the winter population vs the spring population with regard to time to flowering ($\alpha = 0.05$). Outlier data points have been outlined in red and do not contribute to the lines of best fit.

2.3.4 Temperature Timecourse to Determine the Vernalisation Temperature for Safflower

There was a greater variability in the time to flowering in winter safflower when compared to spring safflower following exposure to this temperature gradient (Fig. 2.7). Spring safflower showed little-to-no variation in time to flowering between any treatment (*single factor ANOVA;* $F = 0.38$; $df\,(4, 24)$; $p - value = 0.82$). Conversely, winter safflower varied significantly in the time to flowering across the time course (*single factor ANOVA;* $F = 6.56$; $df\,(4, 34)$; $p - value = 0.00$). The greatest decrease in the time to flowering was for the vernalisation temperature of 8°C which, in turn, differed significantly from the results for the 0°C and 16°C treatments (*TukeyHSD;* $p - value = 0.00$ and 0.00 respectively). Vernalisation treatment at 4°C and 12°C were not significantly different to each other or to the 8°C temperature treatment. There was no significant difference in flowering time results when comparing the 0°C to 16°C treatments (*TukeyHSD;* $p - value = 0.96$) and 4°C to 12°C (*TukeyHSD;* $p - value = 0.90$). Unexpectedly, for temperature gradient treatments of less than 8°C, the time to flowering actually increased. Below 8°C, the longest time to flowering ($\mu = 90$ days) resulted from the 0°C exposure, followed by the 4°C treatment ($\mu = 82$ days).

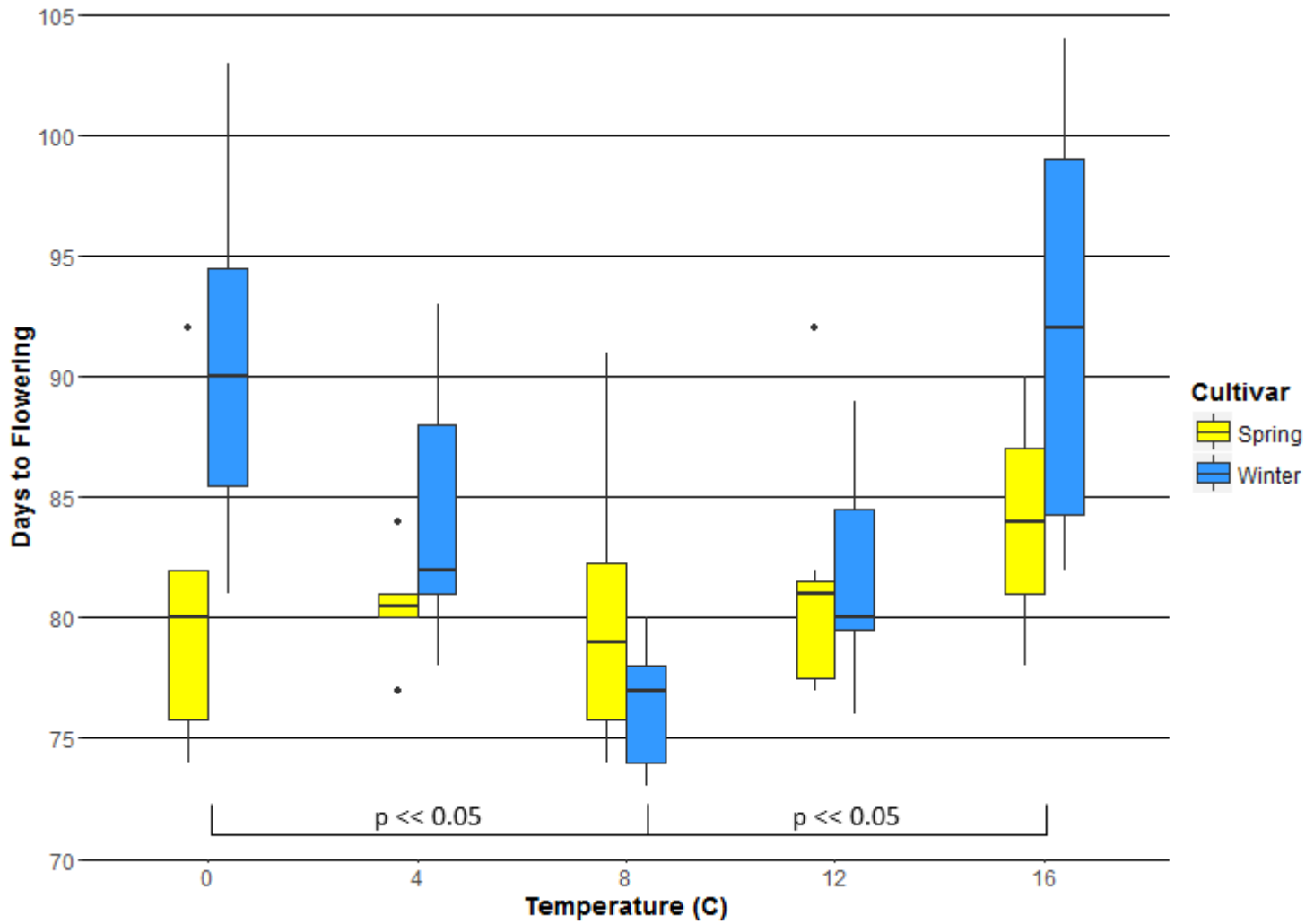


FIGURE 2.7: Vernalisation effect on winter and spring safflower at five different temperatures for 28 days. Significantly different groups indicated ($\alpha = 0.05$).

2.3.5 Inheritance of the Vernalisation Phenotype in Safflower

As there was such a distinct difference between unvernalised winter and spring safflower, a crossing population was created using winter and spring safflower parents. The progeny of this crossing population were examined to determine if the vernalisation trait was inherited in the following generations and to understand the genetic architecture of this newly discovered vernalisation response in safflower. Unfortunately, a complete reciprocal cross was not possible with this population. Emasculated spring safflower heads pollinated with winter safflower did not survive to seed maturity. Due to spring safflower elongating significantly earlier than unvernalised winter safflower and that the effect of vernalisation cannot be detected in early flowering phenotypes, the non-elongating winter rosette behaviour was used as a proxy for vernalisation, i.e. if the crossed progeny elongated early, this was taken as the plant being unresponsive to vernalisation.

In the F_1 crossing population, all plants expressed spring safflower elongation phenotypes, i.e. early elongation. F_2 seeds were grown in the field, but unfortunately, due to the time of planting, variations in growing time and exposure of the seedlings to wintering conditions at the field site used (Fig. 2.8a) made clear identification of F_2 late elongating phenotypes impossible (Fig. 2.8b). F_3 seeds grown under glasshouse conditions, were used instead to infer the segregation of the F_2 population (Appendix H, Table H.1). Out of the 408 F_2 plants grown in the field, only 147 plants survived and produced enough F_3 seed to plant in the glasshouse. These F_3 plants were used as a proxy to characterise the phenotypes expressed by the F_2 generation. Of the 147 F_2 seed populations assessed, only populations where at least 15 F_3 plants survived to phenotyping age, were analysed. This approach resulted in 52 lines of F_3 plants being phenotyped.

Four different gene models were examined to determine the nature of the genes involved in the late elongation response in families within the F_3 population that contained 15 or more plants that survived to phenoyping age (Table 2.2):

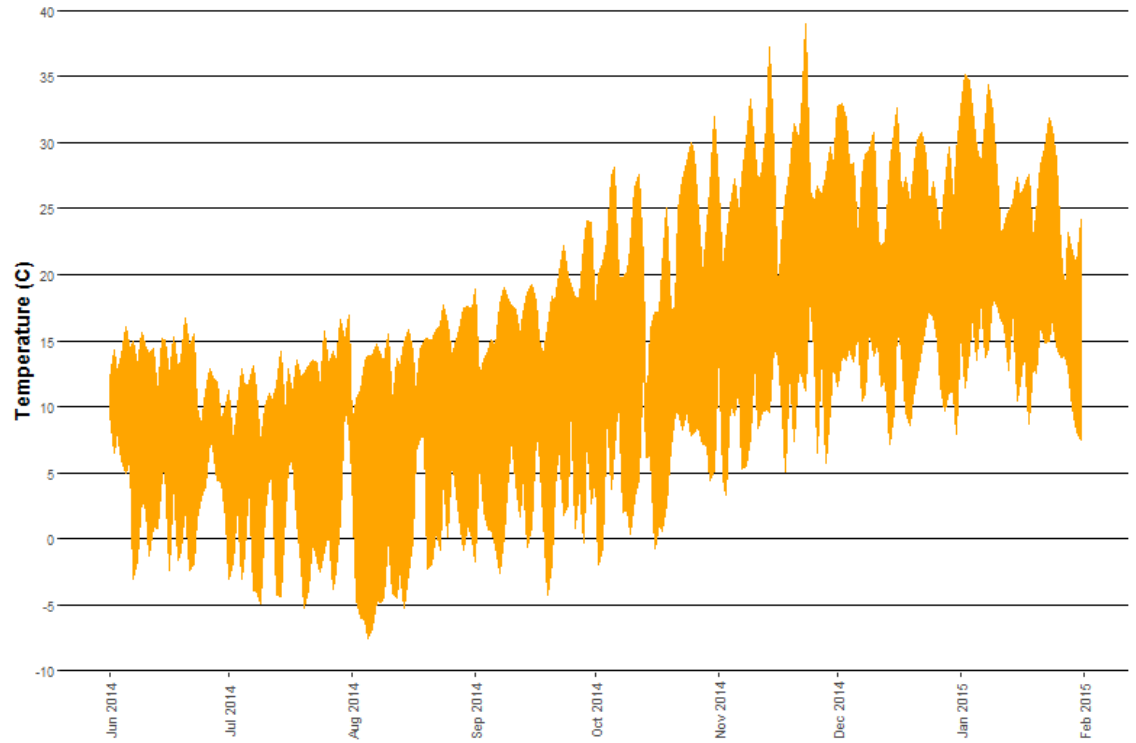
i) Model 1 - a single gene model

ii) Model 2 - a two gene model where both genes are recessive

iii) Model 3 - a two gene model where at least one gene is recessive

iv) Model 4 - a two gene model where one gene is dominant over a second, recessive gene

Of these, Models 1 and 4 best fitted the observed F_2 phenotypes inferred from the F_3 genetic screen ($\chi^2 = 2.62$; $p - value = 0.27$ and $\chi^2 = 0.39$; $p - value = 0.53$ respectively).



(a) Field plot temperatures at the Black Mountain site during the growth of the F_2 safflower crossing generation.

(b) Exposure of F_2 safflower crosses grown in the field to vernalisation conditions made phenotyping the vernalisation response difficult.

FIGURE 2.8: Effects of field conditions on phenotyping the vernalisation effect on safflower F_2 crosses.

TABLE 2.2: Different gene models examining the number of loci responsible for the vernalisation response in safflower in 52 families of the F_3 crossing population. These families contained 15 or more plants that survived to phenotyping age (4 weeks).

Model 1: Single recessive gene model

	Observed	Expected
Elongated (Homo)	14	13
Elongated (Hetero)	30	26
Non-elongated (Homo)	8	13
	χ^2	2.62
	p-value	0.27

Model 2: Two gene model, both recessive

	Observed	Expected
Elongated	44	48.75
Non-elongated	8	3.25
	χ^2	7.41
	p-value	0.01

Model 3: Two gene model, at least one recessive

	Observed	Expected
Elongated	44	29.25
Non-elongated	8	22.75
	χ^2	17.00
	p-value	0.00

Model 4: Two gene model, one gene dominant over another

	Observed	Expected
Elongated	44	42.25
Non-elongated	8	9.75
	χ^2	0.53
	p-value	0.39

2.4 Discussion

2.4.1 The Vernalisation Response is Observed in Winter Safflower Only

As discussed in Chapter 1, although the *Asteraceae* is a large family of flowering plants, a vernalisation response has been previously characterised in lettuce, but not safflower. Of the two safflower cultivars examined, only winter safflower significantly responded to vernalisation conditions. Vernalised winter safflower elongated earlier than unvernalised winter safflower. Further, this reduced elongation period observed was determined to be similar to that of spring safflower. This earlier elongation phenotype of vernalised winter safflower results in a reduced final number of leaves and earlier flowering. Germinated winter safflower seeds grew very little during the four weeks of exposure to vernalisation conditions. Spring safflower seeds continued to grow, regardless of whether they were exposed to vernalisation conditions, though these seeds grew at a slower rate than the unvernalised spring safflower seeds (Fig. 2.2). There was also very little difference in the growth rate in germinated unvernalised winter safflower seeds when compared to germinated unvernalised spring safflower. This indicates that there is an inhibiting mechanism affecting the growth of imbibed safflower seeds during vernalisation, but that this mechanism affects winter safflower to a greater degree than spring safflower. However, if vernalisation is indeed regulating gene expression in spring safflower, there is no way to score this phenotype visually, as spring safflower always elongates earlier than unvernalised winter safflower, regardless of its exposure to vernalisation conditions. Even so, the late elongation phenotype seen in winter safflower can be used as an indicator of a potential vernalisation responsive plant. Those plants that elongate early can be eliminated as vernalisation responsive, which substantially increases the speed of phenotyping vernalisation responsive safflower varieties.

In winter safflower, the greatest effect of vernalisation can be seen in the first two weeks of exposure of seedlings to 4°C where there is a marked decrease in the time to flowering. This decrease in time to flowering is proportional to the length of time the germinated seeds have been exposed to vernalisation conditions. After this two week period, the effect of vernalisation shows only a slight impact on the length of time to elongation and flowering. This asymptotic relationship between vernalisation exposure and flowering time is not seen in spring safflower. Rather, there appears to be a non-significant linear decrease in flowering time proportional to the length of time vernalised, which is less profound than in winter safflower both before and after the two week 'saturation' period. Based on this trend, and given enough time exposed to vernalisation conditions, the data presented here clearly indicates that winter safflower could flower significantly earlier than spring safflower, provided that they are planted in late autumn or winter and are exposed to vernalisation conditions.

In winter safflower (Fig. 2.7), the temperature that resulted in the greatest reduction in the time to flowering was 8°C, not the 4°C for *Arabidopsis* (Simpson and Dean 2002), the 5°C for *Brachypodium distachyon* (Ream et al. 2014), nor in the range of low temperatures (1°C to 7°C) as originally described by Chouard (1960). While there was no significant effect of different vernalisation temperatures on the flowering time in spring safflower, germinated seeds exposed to 16°C were slightly delayed in their time to flowering. Because safflower is native to hotter climates such as the Middle East, this effect could be attributed to a heightened metabolism in warmer temperatures, allowing the shoot apical meristem in spring safflower to produce more vegetative tissue in the early stages of development before elongation is triggered. In winter safflower, the non-significant difference when comparing 0°C to 16°C and comparing 4°C to 12°C could be attributed to passing a metabolic threshold where, rather than just altering gene expression, these lower temperatures actually start to slow cellular metabolism, inhibiting the vernalisation response. This fits with Chouard's observation that vernalisation temperatures below freezing have a proportionally lower effect on reduction in flowering time (Chouard 1960). The higher optimum vernalisation temperature for winter safflower of 8°C also fits with the model put forward by Angus et al. (1980), where different crops have different basal growth temperatures. Would other wild safflower lines, sourced from other locations across the world (for example, Turkey), express a similar response to vernalisation conditions?

2.4.2 The Vernalisation Response in Safflower is Recessive

Physiologically, the behaviour of winter and spring safflower is distinct. By creating a crossing population using winter and spring safflower varieties as parents, segregation in both the resulting F_1 and F_2 populations was used to putatively identify the nature of the vernalisation response in safflower. The vernalisation response can only be properly observed in a late elongating plant, so this was used as a proxy for vernalisation. Because all F_1 crosses resembled spring safflower (early elongation), the late elongation response is recessive. The F_2 population could not be accurately scored in the field due to unplanned exposure to conditions in the field that resembled vernalisation conditions. However, 52 of the 147 F_2 plants had segregations inferred from the F_3 families grown in the glasshouse.

While a single gene dominant:recessive model fits the segregated plants in the F_3 generation, a model with two factors, be they variations in promoter regions or alleles, where one is dominant and the other is recessive, is the best fit for the segregations observed. This also explains the slight, but non-significant, effect vernalisation conditions have on spring safflower (section 2.4.1). If a dominant vernalisation response factor is present in both spring and winter safflower, but a recessive factor for late elongation is only present in winter safflower, the vernalisation responsive phenotype cannot be expressed in spring safflower. Spring safflower, lacking the delayed

elongation phenotype, will always flower early. However, the 3:1 inheritance of the short rosette (spring): long rosette (winter) observed by Carapetian (2001) may, in fact, mean that the single genetic model fits better. Further investigations into the underlying genetic basis of the vernalisation response in safflower will be elucidated in Chapter 4.

There are a number of other traits found in winter safflower that were distinct from spring safflower, including seed oil composition and content, flower colour and 'spikiness'. Many of these traits also appeared to be recessive when crossed with spring safflower. There may be a great number of other traits that may be incorporated into commercial varieties of safflower, further increasing the overall value of safflower as a high value oilseed crop. Although readily observable, these additional phenotypic distinctions of winter safflower, not being related to the vernalisation response, were not investigated in any further detail.

2.4.3 The Vernalisation Response in Safflower is Epigenetic and Resets in the Next Generation

As described above, there is no evidence to show that the effect of vernalisation on parent safflower plants influences how the progeny respond to similar vernalisation conditions. This indicates that, for winter safflower, the regulation of the vernalisation response appears, as in *Arabidopsis* and many other species, to be epigenetic (Song et al. 2012). If this is indeed the case, the next avenue of investigation would be to identify homologues of genes found in the various flowering pathways, such as the approach adopted by Trevaskis et al. (2007a) to identify *FLC*-like and *FT*-like homologues, or a physiological mechanism resembling that directed by the PHD-PRC2 complex in *Arabidopsis* (De Lucia et al. 2008).

A minimum of two weeks of exposure to vernalisation conditions was required to reduce the time to flowering for winter safflower to that of spring safflower, regardless of vernalisation temperature). Beyond this two week period, little difference in time to flowering was observed. In *Arabidopsis*, the methylation of *FLC* by the PHD-PRC2 complex makes the transition to flowering permanent and stable. If a similar epigenetic mechanism is present in safflower, and the effect of this mechanism is stable and irreversible, once a target site of epigenetic regulation has been modified, it cannot be revert to its original state. In the case of *Arabidopsis*, epigenetic regulation represses *AtFLC* expression proportional to the time spent in vernalisation conditions. Eventually, the number of available sites for regulation crosses a 'critical mass' boundary and further exposure to vernalisation has little effect. In safflower, this appears to be after two weeks exposure and a vernalisation temperature of 8°C, but as yet, the underlying molecular mechanisms are unknown.

One limitation with this study was the lack of statistically robust data regarding the effect of vernalisation conditions on progeny plants. Anecdotally, it was observed that in winter safflower, whether or not a plant was vernalised, it had no effect on whether the resulting progeny responded to vernalisation environmental cues. Progeny from a vernalised winter safflower plant needed to be exposed to vernalisation conditions themselves, otherwise a late elongation phenotype resulted. This resetting mechanism in the next generation, with no impact of the parent plant's vernalisation exposure, is a characteristic of the vernalisation response. Safflower is not unique in this regard as this is observed in many other species, including *Arabidopsis* (Sheldon et al. 2000).

2.5 Conclusion

Together, these results strongly suggest that a vernalisation response is, indeed, present in winter safflower and that this is a true vernalisation response, as indicated by the traits listed in Chapter 1. Further, this response appears to be recessive and epigenetic in nature. The implications of this finding is that if this trait can be introduced and fixed into an elite safflower cultivar, it will permit late autumn or early winter planting of safflower seed in Australia, with seed remaining dormant or growing slowly until conditions allow it to germinate and to grow to maturity. This, in turn, could potentially 'free up' the spring and summer growth seasons for planting of other crops, such as wheat, barley or canola.

Chapter 3

Transcriptomic Analysis of the Vernalisation Response in Safflower

3.1 Outline

The experimental aim of this chapter was to identify transcripts that are differentially expressed in winter safflower as part of the vernalisation response. This involved the generation of a high quality *de novo* transcriptome for safflower for use as a reference. Subsequently, any transcripts of interest from winter safflower could be compared to the *de novo* spring safflower transcriptome. Further, the *de novo* spring safflower transcriptome could also be used to identify any single nucleotide polymorphisms (SNPs) or insertions/deletions in winter safflower transcripts. Finally, in silico analyses were confirmed via a reverse transcriptase (RT) quantitative polymerase chain reaction (qPCR) approach.

3.2 Materials and Methods

3.2.1 Selection of RNA Extraction Protocol

A number of RNA extraction techniques were assessed to determine the highest quality and most consistent protocol for extracting RNA material from safflower vegetative tissue. For this test, young leaf tissue extracted from four week old plants was ground to a powder in liquid nitrogen and used for the RNA extraction tests. The performance of each assessed RNA extraction technique was based on three criteria (Wilfinger et al., 1997):

i) how closely the Nanodrop sample measurement was to the optimum RNA absorbance ($OD_{260/280}$ = 2.0)

ii) how close the Nanodrop sample measurement was to the optimum absorbance for minimum contamination ($OD_{260/230}$ = 2.0 - 2.2)

iii) how consistent the replicate measurements were (by how closely data points cluster together)

For each of these methods, extracted RNA was tested on a 2% agarose at 110v for approximately 30 min to confirm the presence of RNA.

3.2.1.1 PureLink Based Method

The PureLink® reagent (cat#: 12322-012, Life Technologies™) RNA extraction protocol was followed as per the manufacturer's instructions.

3.2.1.2 Qiagen RNeasy Kit

The RNeasy® Plant Mini Kit (cat#: 74903, Qiagen™) protocol was followed as per the manufacturer's instructions.

3.2.1.3 TRIzol Based Method (Manufacturers Protocol)

The TRIzol® reagent (cat#: 15596018, Invitrogen™) protocol was followed as per the manufacturer's instructions.

3.2.1.4 Cetyl Trimethyl Ammonium Bromide (CTAB) Based Method

Fifteen mL of extraction buffer (2% CTAB, 2% polyvinylpyrrolidinone (PVP) K 30, 100 mM tris(hydroxymethyl)aminomethane hydrochloride (Tris-HCl; pH 8.0), 25 mM ethylenediaminetetraacetic acid (EDTA), 2.0 M NaCl, spermidine [0.5 $\mathrm{g/L}$], then autoclaved; 2% β-3-mercaptoethanol added just before use) was warmed to 65°C in a water bath before adding 2 to 3 g frozen and ground tissue. Tubes were mixed by inversion before adding an equal volume of chloroform:isoamyl alcohol (24:1 $\mathrm{v/v}$) to the solution each tube. Tubes were centrifuged at room temperature at 10,000 gravitational force (*g*) for 10 min. This step was repeated before adding one quarter volume of 10 M LiCl. Tubes were incubated overnight at 4°C before centrifugation at 4°C at 10,000 *g* for 20 min. Post centrifugation, the resulting supernatant was discarded and the pellet resuspended in 0.5% sodium dodecyl sulphate (SDS). An equal volume of chloroform:isoamyl alcohol (24:1 $\mathrm{v/v}$) was added to each tube before centrifugation at room temperature at 10,000 *g* for 10 min. Two volumes of 80% ethanol was added to the tubes before precipitating at -20°C for 2 hours. Tubes were centrifuged at room temperature at 10,000 *g* for 10 minutes, discarding the supernatant. The resulting pellets were air dried for approximately 30 min or until no ethanol was visible in the tubes before resuspending in 50 µL diethylprocarbonate (DEPC)-treated H_2O.

3.2.1.5 Hot Phenol Based Method

Acidified phenol (pH 4.7) was added to an equal part extraction buffer (100 mM Tris-HCl (pH 8.0), 100 mM LiCl, 10 mM EDTA (pH 8.0), 1% SDS) and heated to 80°C for at least 30 min. 1 g of ground frozen plant tissue was added to 1 mL of phenol:extraction buffer in a sterile RNase-free 1.5 mL microfuge tube and mixed using a sterile pipette tip. 500 µL chloroform was added to the sample and mixed using a rotation wheel. Samples were then centrifuged at room temperature at 14,000 *g* for 15 min and the resulting supernatant transferred to a new sterile RNase-free 1.5 mL microfuge tube. If there was a substantial quantity of material at the

supernatant/subnatant boundary, the chloroform purification step was repeated until a desired boundary was obtained. One third volume of 8 M LiCl was added and tubes mixed by inversion. Tubes were incubated at 4°C overnight before centrifuging at 4°C at 10,000 *g* for 30 min. The supernatant was discarded and the pellet washed with equal volume of 100% isopropanol. Tubes were incubated at room temperature for 10 min before centrifuging at room temperature at 10,000 *g* for 10 min. The supernatant was discarded and the tube allowed to air dry. The resulting pellet was washed with 2 M LiCl and the tube centrifuged at room temperature at 10,000 *g* for 5 min. The supernatant was discarded and the pellet washed with 500 µL 80% DEPC-treated ethanol. The samples were centrifuged at room temperature at 10,000 *g* for 5 min. The supernatant was discarded and the pellet air dried in a laminar flow hood for approximately 30 min or until the pellet was dry. The pellet was resuspended in 50 µL DEPC-treated H_2O.

3.2.1.6 TRIzol Based Method (Modified from Manufacturers Method)

One mL TRIzol® reagent (cat#: 15596018, Invitrogen™) was placed in a sterile RNase-free 1.5 mL microfuge tube. To this, 200 µg of frozen and ground tissue was added, briefly vortexed and incubated at room temperature for 15 min. Next, 200 µL chloroform was added to each tube and tubes inverted by hand before incubating at 4°C for 20 min. Tubes were then centrifuged at 4°C at 14,000 *g* for 15 min. The supernatant was transferred to a fresh sterile RNase-free 1.5 mL microfuge tube and 600 µL acidified phenol (pH 4.5) added, mixing by inversion. The samples were then incubated for 15 min at room temperature. Another 200 µL chloroform was added to each sample and tubes vortexed. The samples were incubated at 4°C for 20 min before centrifuging at 4°C at 14,000 *g* for 15 min. If substantial material was present at the supernatant/subnatant boundary, the two chloroform and phenol steps were repeated. Supernatant was transferred to a new RNase-free 1.5 mL microfuge tube and 0.6x volume RNase-free 100% isopropanol was added. The tubes were vortexed before incubating at -20°C for 30 min. The tubes were then incubated at room temperature for 30 min before being centrifuged at room temperature at 14,000 *g* for 10 min. The supernatant was then discarded and the pellet washed with 600 µL 80% ethanol and incubated at room temperature for 5 min. The samples were then centrifuged at room temperature at 14,000 *g* for 10 min before being left to air dry in a laminar flow hood. The pellet was resuspended in 50 µL DEPC-treated H_2O and incubated for 2 min at 65°C.

3.2.2 Primer Design and RT-qPCR Protocol

Primers for RT-qPCR were generated from transcripts extracted from the safflower *de novo* transcriptome using Oligo Explorer (`www.genelink.com/tools/gl-oe.asp`; v1.1.2) and confirmed with Netprimer (`www.premierbiosoft.com/netprimer/`; v3). Primers were designed to have a minimum of 18 nucleotides and an annealing

temperature above 62°C. Primers were designed to limit potential primer-dimer formation and, where possible, primers were designed towards the 3′ end of the transcript. Primers were resuspended in 10 mM Tris (pH 7.5, 0.1 $\mu\mathrm{M}$ EDTA) to produce a stock solution of 100 $\mu\mathrm{M}$ before diluting primers to a 10 $\mu\mathrm{M}$ working stock. The primers were tested by RT-qPCR against true leaf and shoot apical meristem, cotelydon and vegetative tissue. Where a generated primer did not produce a PCR product, the primer was redesigned using a different location on the transcript. All primers were manufactured by Sigma-Aldrich Inc. (Sydney, Australia; Appendix F).
A DNase treatment using the RQ1 DNase Treatment (cat#: M6101, Promega^{TM}) was conducted on all RNA samples that were used for RT-qPCR as per the manufacturer's protocol. First strand synthesis was performed on the DNase treated RNA samples using Maxima Reverse Transcriptase (Thermo Fisher Scientific^{TM}) as per the manufacturer's protocol.

For RT-qPCR runs in Experiment 1, three biological replicates were used for each gene tested with only a single technical replicate. For RT-qPCR runs in Experiment 2, a minimum of three technical replicates for the three biological replicates (only two with spring safflower, time point 10). For each RT-qPCR run, a master mix was prepared for each primer set (without cDNA template) using Fast SYBR Green Master Mix (Thermo Fisher Scientific^{TM}) and Platinum Taq Polymerase (Thermo Fisher Scientific^{TM}). Each master mix contained a final forward and reverse primer concentration of 0.25 $\mu\mathrm{M}$. Each 20 µL reaction contained 2 µL reaction buffer, 1.4 µL of 50 mM $MgCl_2$, 1 µL Fast SYBR Green, 0.5 µL of 10 $\mu\mathrm{M}$ forward and reverse primer, 0.8 µL of 10 mM dNTPs and 0.1 µL Platinum Taq Polymerase. One µL of first strand cDNA template solution was added before adding sufficient nuclease-free H_2O to bring the reaction volume to 20 µL.

PCR amplification was performed on a Rotor Gene Q (Qiagen^{TM}) beginning with denaturation by heating tubes to 95°C for 5 min. Fourty-five PCR amplification cycles were then completed, consisting of denaturation at 95°C for 20 sec, annealing at 59°C 20 sec with a single fluorescence measurement and extension at 72°C for 20 sec. The program finished with a melt curve analysis to detect the presence of an RT-qPCR product. The samples were cooled to 50°C before increasing the temperature to 99°C in 1°C increments, holding for 5 sec at each increment.

3.2.3 Assembly of *De Novo* Transcriptomic References for Safflower

3.2.3.1 RNA Extraction and Sequencing of Spring Safflower

Sixteen different tissues were isolated from spring safflower, including root, leaf, shoot apical meristem, pollen, dry seed and imbibed seed. RNA was extracted from these tissues using the Purelink method. Total RNA was prepared using the Illumina TruSeq Sample Prep Kit v2 according to the manufacturer's instructions. Libraries were individually barcoded and sequenced on an Illumina HiSeq2000 platform generating

100 base pair (bp) paired end (PE) reads for each of the 16 tissue types. The isolation of the tissues, extraction of RNA and preparation of the samples for sequencing was conducted by colleagues at CSIRO Black Mountain prior to the commencement of this project. Reads were archived in the CSIRO Data Access Portal (`https://data.csiro.au/dap/landingpage?pid=csiro:16250`).

3.2.3.2 Pre-processing of Spring Safflower Reads

Illumina reads were analysed and prepared prior to assembly. Reads from each paired end of the 16 spring safflower tissues were analysed using FastQC software (v0.10.1). Based on the results of FastQC, reads were trimmed by 30 bp on the 3′ end to remove the low quality region of the read and any residual adaptors. Any trimmed reads that were less than 70 bp were discarded.

3.2.3.3 *De Novo* Assembly of Spring Safflower Reads

The resulting trimmed reads were assembled with Trinity software (v2012-06-08; Grabherr et al. 2011) using the default settings, except that the path reinforcement distance was reduced to 15 (Appendix J.1). Spring safflower *de novo* contigs produced by the Trinity assembler have the following nomenclature:

<Species>_<type of assembly>_<variety or cultivar>_<*cluster*>_<*gene*>_<*isoform*>
e.g. '**CarTin_tx_s317_comp33397_c0_seq1**' is from the safflower (**CarTin**) transcriptome (**tx**) of spring safflower (**s317**), cluster (**comp**)**33397**, gene (**c**)**0**, isoform (**seq**) **1**.

3.2.3.4 Quality Assessment of the *De Novo* Spring Safflower Assembly

The quality of the resulting assembly was assessed using Biokanga 'Fasta2nxx' (v3.4.7, `https://github.com/csiro-crop-informatics/biokanga`), Core Eukaryotic Genes Mapping Approach (CEGMA; v2.4.010312; Parra et al. 2007) and Benchmarking Universal Single-Copy Orthologs (BUSCO; v1.1b1; Simão et al. 2015) software. Previously characterised genes in the *FATTY ACID DESATURASE 2* (*CtFAD2*) family (Cao et al. 2013) were also used to assess the quality of the *de novo* transcriptome. BLASTN (v2.2.28+; Altschul et al. 1997) was used to identify transcriptomic contigs that closely matched members of the *CtFAD2* family and a multiple sequence alignment was built using ClustalW (v2.1 via CLC Genomics Workbench v7.0.4). A phylogenetic tree of the *CtFAD2* gene family was also constructed using MEGA software (Build# 6140226), with a bootstrap value of 1000 to test the frequency of clades.

3.2.3.5 Back Alignment of Spring Safflower Reads to the *De Novo* Reference

Back alignment of raw reads to the *de novo* assembly serves two purposes. Firstly, as a quality control measure to identify if one or more libraries over or under contributed to the assembly. Secondly, for analysis of differential gene expression, with reads aligning against a transcript used to estimate the abundance of a transcript in a sample.

Raw reads from each of the 16 tissue libraries were back aligned against the *de novo* transcriptome assembly using Biokanga 'Align' software (v3.8.1; Appendix J.1). Only reads that uniquely aligned to gene bodies were accepted, as this avoided sequences of high similarity between different transcripts being artificially inflated. Reads that aligned in a chimeric fashion, i.e. at least half of the read aligned to a location but the remainder of the read did not, were trimmed back to the aligning segment. Paired end read libraries were combined and aligned as if they were single ended, with a substitution rate of 10% and a single ambiguous base pair permitted. This combination of loosely specific parameters and discarding reads aligning to multiple locations allowed a further reduction in the artificial inflation of counts across highly conserved sequences.

3.2.4 Assembly of a *De Novo* Winter Safflower Assembly

In addition to the spring safflower reference transcriptome, a *de novo* transcriptome was constructed from reads generated in Experiment 1 (Section 3.2.6.2). Due to an odd artefact found at 100 bp along the reads in every library, reads were trimmed to 100 bp before using Biokanga 'Assemb' and 'Scaffold' software (v3.5.3) with the default parameters (Appendix J.2.1 and J.2.2). The resulting winter safflower transcripts were used to examine allelic variation between winter and spring safflower and to identify any sequences present in winter safflower that may not have been expressed in spring safflower. Winter safflower *de novo* contigs produced by the Biokanga assembler have the following nomenclature:

<Species>_<type of assembly>_<variety or cultivar>_<scaffolded contig number>
e.g. '**CarTin_tx_WSRC03_Scaff65369**' is from the safflower (**CarTin**) transcriptome (**tx**) of winter safflower (**WSRC03**), scaffolded contig number **Scaff65369**.

3.2.5 Aligning the Spring and Winter *De Novo* Assemblies

To determine if there were any SNPs or indels present in the winter safflower transcripts that could be the source of the vernalisation response, the previously created spring and winter safflower transcriptomes were aligned against one another using Biokanga 'Blitz' software (v3.5.3) with a core length of 13, a minimum path score of 130, minimum extension threshold score of 12, a k-mer depth of 1,500, and a minimum sequence alignment of 25% (Appendix J.2.3). Alignments were then filtered to remove any alignments that were either less than 80% of the length of the winter safflower

transcript or contained more than 5% mismatches. Any winter safflower transcripts that were not found in the spring safflower transcriptome were translated and searched on the National Center for Biotechnology Information sequence database (referred to as NCBI) using BLASTP (v2.2.28+; Altschul et al. 1997).

3.2.6 Differential Expression (DE) Analysis

3.2.6.1 Growth Conditions for DE

Vernalised and unvernalised spring and winter safflower cultivars were used as described in Chapter 2.2.1. Cold break dormancy was performed as described in Chapter 2.2.2.1.

3.2.6.2 Experiment 1: Winter Safflower Before and After Vernalisation

Vernalised and unvernalised spring and winter safflower cultivars were used (Chapter 2.2.1). To test for differential expression, winter and spring safflower seeds were germinated and vernalised (Chapter 2.2.2.2). Vernalised and unvernalised germinated safflower seeds were grown in long day growth cabinets (Chapter 2.2.2.5), then the vegetative tissue harvested 3 mm below the hypocotyl junction. The harvested tissue was frozen in liquid nitrogen. RNA was extracted using the Purelink method. Three $\mu g/\mu L$ of total RNA from four biological replicates from the unvernalised winter safflower, four biological replicates from the vernalised winter safflower and four biological replicates from the unvernalised spring safflower (12 samples in total) were sent to a commercial sequencing supplier (the Australian Genome Research Facility; AGRF) for sequencing on a single lane of a HiSeq2500, producing 150 bp paired end (PE) reads as per the commercial supplier requirements. The CSIRO Data Access Portal (`https://data.csiro.au/dap/landingpage?pid=csiro:6416`) was used to archive reads. Read quality was determined using Biokanga 'Ngsqc' software (v3.8.1).

3.2.6.3 Experiment 2: Vernalisation of Safflower at Five Time Points

Experiment 1 identified and annotated a number of transcripts thought to be candidates in the vernalisation response pathway. To further examine these transcripts and to determine an initial sense of expression timing, the expression of these transcripts were examined as winter and spring safflower cultivars were exposed to longer periods of vernalisation conditions. For Experiment 2, germinated seeds from both winter and spring safflower were exposed to vernalisation conditions (Chapter 2.2.2.3) for 5, 10, 15 and 20 days, as based on the results from Chapter 2.3.3. Unvernalised seedlings for both cultivars were assessed in parallel. After exposure, vernalised and unvernalised seedlings were grown for one week in long day growth cabinets (Chapter 2.2.2.5). Vegetative tissue from three biological replicates of each cultivar and exposure time point were then harvested approximately 3 mm below the hypocotyl junction before

being frozen in liquid nitrogen. RNA was extracted using the Purelink method. Three $\mu g/\mu L$ of total RNA from three biological replicates from every time point and three biological replicates from the unvernalised winter and spring safflower (30 samples in total) were tested for quality on a Bioanalyzer 2100 (Agilent Technologies) using three Eukaryote Total RNA Nano chips (cat #: 5067-1511). An RNA integrity score (RIN) of 8 or more was an indication of good quality extracted RNA. Of the 30 samples assessed, only three returned a RIN indicating a sample quality that was questionable (Winter - 00 days, rep 1, RIN=5.9; Winter - 15 days, rep 3, RIN=5.5; Spring - 15 days, rep 3, RIN=4.2). Despite their low RIN score, these samples were included for sequencing, as any degraded or heavily fragmented RNA material would not align to the transcriptomic reference (due to the presence of ambiguous and incorrectly called bases) and subsequently be filtered out. The RNA samples were sent to a commercial supplier (AGRF) for sequencing on three lanes of a HiSeq2500, producing 100 bp PE reads, as per the commercial supplier's requirements. The CSIRO Data Access Portal (`https://data.csiro.au/dap/landingpage?pid=csiro:16251`) was used to archive reads. Read quality was determined using Biokanga 'Ngsqc' software (v3.8.1).

3.2.6.4 Analysis of Back Alignment Data

Reads generated from Experiment 1 (RNA extracted from vernalised and unvernalised winter safflower) and Experiment 2 (RNA from winter and spring safflower across five different time points) were back aligned to the spring safflower *de novo* transcriptome using Biokanga 'Align' (v3.8.1) as described above (section 3.2.3.5). The number of reads uniquely aligning to each contig in the spring safflower transcriptome (referred to hereafter as transcriptomic 'read counts') was recorded. Both Experiment 1 and Experiment 2 read counts were analysed using R software (v3.1.2), R Studio (v0.98.953) and the BiocParallel R package (v1.0.3).

For Experiment 1, transcriptomic read counts for each replicate were analysed with the DESeq2 R package (v1.6.3; Love et al. 2014). Any transcripts determined to be significantly differentially expressed, using an $\alpha = 0.01$ and minimum fold change of two, were translated into amino acid sequences in all six reading frames. While a significance threshold of $\alpha = 0.05$ is standard for most non-medical statistical analysis, decreasing the significance threshold to $\alpha = 0.01$ and increasing the minimum absolute fold change reduced the number of identified transcripts, which allowed hand curation of these differentially expressed transcrips.

For Experiment 2, transcriptomic read counts for each time point and replicate were also analysed with DESeq2 (v1.6.3). Transcripts that were identified as significantly differentially expressed, using an $\alpha = 0.05$. Those transcripts that were significantly differentially expressed between winter and spring safflower, were translated into amino acid sequences in all six reading frames.

3.2.6.5 Annotating Differentially Expressed Transcripts

Translated DE transcripts from the two differential expression experiments were aligned to the NCBI non-redundant amino acid database in all six reading frames using BLASTP (v2.2.28+) and filtered using the Entrez 'green plants' entry. After identification and annotation of transcripts via BLASTP and sequence homology alignment, multiple sequence alignments (MSAs) were generated using T-coffee (Notredame et al. 2000) with the safflower *de novo* transcripts sourced from both the spring and winter safflower transcriptomes, *Arabidopsis* transcripts sourced from The Arabidopsis Information Resource (TAIR10; `www.arabidopsis.org/`; Berardini et al. 2015) and sequences from other green plant species sourced from NCBI.

3.3 Results

3.3.1 Assessment of RNA Extraction Methods for Safflower Leaf Tissue

Six different RNA extraction methods were assessed to identify the protocol that consistently returned the highest quality RNA samples from safflower vegetative tissue. Of the six extraction techniques assessed, PureLink proved to be the optimum RNA extraction method (Fig. 3.1), being the only technique to consistently extract RNA from safflower with an absorbance (OD) $_{260\,nm/280\,nm}$ of 2.0 and $OD_{260\,nm/230\,nm}$ of between 2.0 and 2.2. The PureLink method was, therefore, used to extract RNA from safflower vegetative plant tissues collected for Experiment 1 and 2.

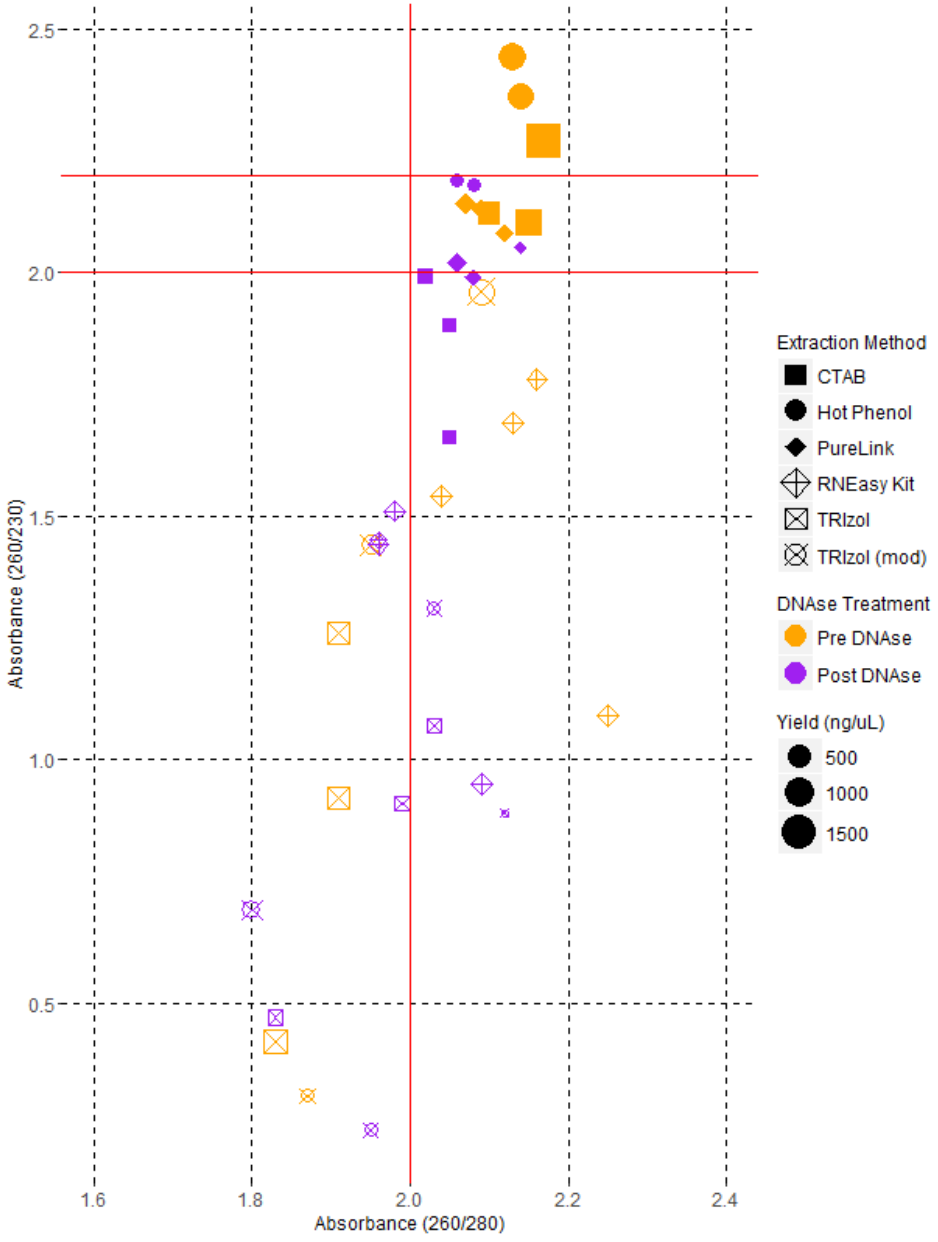


FIGURE 3.1: Results from testing different RNA extraction methods, thresholds indicated with a solid red line. A single pre DNase treated TRIzol (mod) data point with $OD_{260/280} < 0.5$ and $OD_{260/230} = 1.84$ is not shown.

3.3.2 Assembly of the *De Novo* Spring Safflower Transcriptomic Reference

3.3.2.1 Pre-processing of Reads from Safflower Tissues

For a number of the short read libraries created from the 16 isolated spring safflower tissues, the quality scores indicated that some libraries required trimming of 10 to 30 bp on the 3′ end. Two libraries, Stem, Pair 2, and Imbibed Seed, Pair 2, contained a substantial number of reads that were of questionable or poor quality, with over 50% of the read libraries receiving a Phred score of below 22 and 14 respectively. Despite the presence of the substantial number of low quality reads in some of the libraries, these two datasets were included in the assembly and every library was trimmed by 30 bp at the 3′ end, giving a final trimmed read length of 70 bp. Coverage of transcriptomic reads for each tissue type has been noted in Table 3.1.

TABLE 3.1: Coverage of each pair of read libraries from different safflower tissues. The number of reads from each library is multiplied by the read length (100 bp) to produce the number of total base pairs in the library, then divided by the estimated length of the genome (1.4 Gbp) to get the coverage of the genome.

Tissue	Reads	Genome Coverage (%)
Stage 1 Embryo	50,089,350	3.58
Stage 2 Embryo	32,103,570	2.29
Stage 3 Embryo	45,507,708	3.25
Embryo 10 days	51,465,842	3.68
Embryo 15 days	32,218,366	2.30
Embryo 20 days	37,566,892	2.68
Embryo 25 days	31,371,762	2.24
Embryo 50 days	24,863,610	1.78
Cotelydon	32,746,324	2.34
Root	28,962,326	2.07
Shoot Apical Meristem	36,907,338	2.64
Leaf	37,208,548	2.66
Stem	16,404,390	1.17
Pollen	24,062,494	1.72
Dry Seed	28,401,436	2.03
Imbibed Seed	39,127,700	2.79

3.3.2.2 Assembly of the *De Novo* Transcriptomic Reference

The final transcriptome size was approximately 145 Mbp. This was just over 10% of the estimated genome size of 1.4 Gbp (Garnatje et al. 2006) with a mean transcript length of just under 1 kbp and an n50 of just over 1.5 kbp (Table 3.2).

TABLE 3.2: Attributes of the *de novo* spring safflower transcriptome.

Total Size	**144,650,204 bp**
Contigs	146,780
Min Length	201 bp
n50	1,669 bp
Mean Length	985 bp
Max Length	16,781 bp

3.3.2.3 Quality Assessment with CEGMA and BUSCO

Using CEGMA, 247 of the 248 conserved eukaryotic sequences were found in the spring safflower transcriptomic assembly (Table 3.3). When assessed using BUSCO, 92% of the reference sequences were found at least once. Of these 461 sequences were identified once, an additional 427 BUSCO sequences were identified multiple times (Table 3.4).

TABLE 3.3: CEGMA analysis on the *de novo* spring safflower transcriptome, using 248 highly conserved protein sequences from Eukaryotic organisms aligned against spring safflower *de novo* nucleotide sequences.

	Proteins	Completeness	Total	Average	Orthologous
Complete	247	99.60	650	2.63	70.45
Group 1	66	100.00	186	2.82	77.27
Group 2	56	100.00	150	2.68	71.43
Group 3	60	98.36	159	2.65	71.67
Group 4	65	100.00	155	2.38	61.54
Partial	248	100.00	737	2.97	77.42
Group 1	66	100.00	205	3.11	83.33
Group 2	56	100.00	169	3.02	78.57
Group 3	61	100.00	184	3.02	77.05
Group 4	65	100.00	179	2.75	70.77

TABLE 3.4: BUSCO analysis on the *de novo* spring safflower transcriptome using protein sequences that are highly conserved amongst Eukaryotic organisms.

BUSCOs Searched	956	%
Complete Single-copy	461	48%
Complete Duplicated	427	44%
Fragmented	21	2.1%
Missing	47	4.9 %

3.3.2.4 Quality Assessment using the *CtFAD2* Gene Family

Arguably, the best characterised genes currently available for safflower are those in the *FATTY ACID DESATURASE 2* (*CtFAD2*) family described by Cao et al. (2013). These 11 *CtFAD2* transcripts were used to assess the quality of the assembled spring safflower transcriptome. A phylogram created from a ClustalW multiple sequence alignment

(Fig. 3.2) showed that 8 of the 11 *CtFAD2* transcripts aligned with one or more spring safflower *CtFAD2* transcripts in individual clades, each with a very high bootstrap score (95% or greater). *CtFAD2.9* clustered with two spring safflower *de novo* transcripts, CarTin_tx_s317_comp28476_c0_seq1 and CarTin_tx_s317_comp32267_c0_seq1, having a bootstrap score of 79%. Both *CtFAD2.4* and *CtFAD2.3* clustered in a clade against the spring safflower *de novo* transcript CarTin_tx_s317_comp32843_c0_seq1, with *CtFAD2.4* branching from the parent node of *CtFAD2.3*. This indicates that *CtFAD2.3* and *CtFAD2.4* could be isoforms of the same transcript. The *de novo* transcripts clustering with *CtFAD2.6* indicated three isoforms of the same transcript, CarTin_tx_s317_comp33397_c0_seq1, CarTin_tx_s317_comp33397_c0_seq2 and CarTin_tx_s317_comp33397_c0_seq3. While BLASTN returned a reasonably high homology between the *de novo* transcripts CarTin_tx_s317_comp8168_c0_seq2, CarTin_tx_s317_comp49814_c0_seq_1 and CarTin_tx_s317_comp8370_c0_seq1, and the *CtFAD2*s, they did not cluster closely with any other *CtFAD2* genes, indicating that these may be new and novel members of the *CtFAD2* family. The presence of homologues for all of the currently characterised *CtFAD2* transcripts, as well as potentially having identified three novel members of the *CtFAD2* family provides additional support for the *de novo* transcriptomic spring safflower assembly being accurate and of high quality. Because the *CtFAD2* family was only used to assess the quality of the constructed reference transcriptome, no further analysis was conducted.

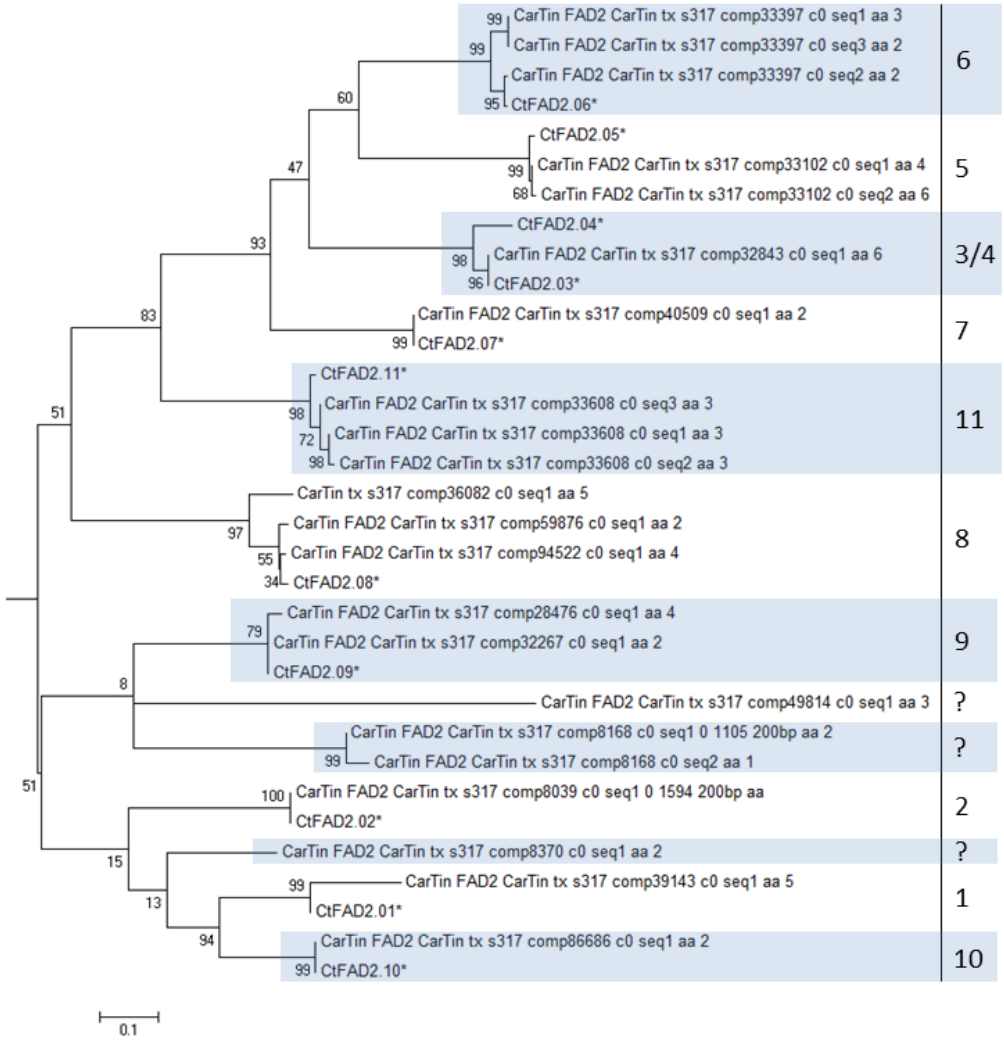


FIGURE 3.2: Phylogenetic tree showing similarity between of *CtFAD2* family genes and highly similar sequences from the spring safflower *de novo* transcriptome. Multiple Sequence Alignment was created using ClustalW. Those sequences marked with an '*' have been extracted from Cao et al. (2013). Numbers at each node represent the bootstrap score for that node i.e. the percentage of times the same alignment is seen after bootstrapping the sequence and recalculating the alignment. The scale bar at the bottom represents the number of substitutions per nucleotide position being, in this case, 10%.

3.3.2.5 Assessment of the Winter Safflower *De Novo* Assembly

Consistent with the spring safflower *de novo* transcriptomic assembly, the winter transcriptomic assembly (Table 3.5) was assessed using CEGMA and BUSCO. CEGMA reported that 227 of the 248 conserved sequences (approximately 92%) were found as complete sequences within the winter transcriptome, and 247 out of 248 were found as partial sequences (Table 3.6). Assessment with BUSCO showed that 47% of conserved BUSCO sequences were found once in the transcriptome, with a further 31% of these conserved sequences being found multiple times (Table 3.7).

TABLE 3.5: Attributes of the winter safflower *de novo* transcriptome, using RNA extracted from vernalised and unvernalised vegetative tissue.

Total Size	**65,932,538 bp**
Contigs	94,032
Min Length	300 bp
n50	786 bp
Mean Length	701 bp
Max Length	9,011 bp

TABLE 3.6: CEGMA analysis on the winter safflower *de novo* transcriptome, using 248 highly conserved protein sequences from Eukaryotic organisms aligned against winter safflower *de novo* nucleotide sequences.

	Proteins	Completeness	Total	Average	Orthologous
Complete	227	91.53	615	2.71	82.82
Group 1	56	84.85	149	2.66	82.14
Group 2	53	94.64	143	2.70	81.13
Group 3	57	93.44	151	2.65	85.96
Group 4	61	93.85	172	2.82	81.97
Partial	245	98.79	844	3.44	88.98
Group 1	64	96.97	202	3.16	85.94
Group 2	56	100.00	197	3.52	89.29
Group 3	60	98.36	216	3.60	88.33
Group 4	65	100.00	229	3.52	92.31

TABLE 3.7: BUSCO analysis on the winter safflower *de novo* transcriptome using protein sequences highly conserved amongst Eukaryotic organisms.

BUSCOs Searched	956	%
Complete Single-copy	456	47%
Complete Duplicated	298	31%
Fragmented	97	10%
Missing	105	10%

3.3.2.6 Aligning the Spring and Winter Safflower Transcriptomic Assemblies

Using Biokanga 'Blitz', when the *de novo* winter safflower transcriptome and the spring safflower transcriptome were compared, 93,346 of the 94,032 sequences in the winter transcriptome (99.3%) mapped to 52,392 transcripts in the spring safflower *de novo* transcriptome, in just under 330,000 alignments. After filtering out any alignments that were less than 80% of the winter transcriptome length or contained more than 5% mismatches across the length of the alignment, 66,266 (approximately 71%) of the winter transcripts aligned across 30,351 (approximately 58%) of the spring safflower alignments.

3.3.3 Experiment 1: Differentially Expressed Transcripts Before and After Vernalisation

Differential expression analysis was conducted between vernalised and unvernalised winter safflower, using both the spring and winter safflower *de novo* transcriptomic assemblies. Following the quality control assessment and reads being been trimmed to a final length of 100 bp, the number of reads uniquely aligned, aligned to multiple locations and unaligned was similar across all replicates for both the vernalised and non-vernalised winter safflower.

Using DESeq2 to analyse the read counts from vernalised and unvernalised winter safflower that were aligned against the spring safflower reference, there were 1000 significantly differentially expressed transcripts ($\alpha = 0.05$). Restricting the parameters further, DESeq2 identified that 273 transcripts were significantly differentially expressed ($\alpha = 0.05$) that had an increased or decreased log2 fold change of at least one. Restricting the differential expression criteria even further by limiting the criteria to contigs that were very significantly differentially expressed ($\alpha = 0.01$) that had increased or decreased log2 count difference of at least two, 30 transcriptomic contigs remained (Table 3.8; Appendix B, Figs. B.1 and B.3).

TABLE 3.8: Differentially Expressed Spring Safflower Transcriptomic Contigs. **Bold** faced rows indicate sequences annotated as involved in the vernalisation response in other organisms.

Contig	Mean Counts	log2 Fold Change	Adjusted p-value	Annotation
CarTin_tx_s317_comp33519_c0_seq70	**295**	**5.9**	**0.0**	**Vernalisation 1 (VRN1)-like**
CarTin_tx_s317_comp33367_c7_seq4	**61**	**4.2**	**0.0**	**MADS1 (MADS box containing)**
CarTin_tx_s317_comp26769_c0_seq1	**124**	**4.1**	**0.0**	**Apetala 1 (AP1)-like**
CarTin_tx_s317_comp29294_c0_seq1	155	3.0	0.0	Glucose and Ribitol Dehydrogenase
CarTin_tx_s317_comp26483_c0_seq1	287	2.4	0.0	Uncharacterised/Hypothetical Protein
CarTin_tx_s317_comp541778_c0_seq1	251	-2.1	0.0	No Hits
CarTin_tx_s317_comp67497_c0_seq1	115	2.5	0.0	(Multiple Oxidases/Hydroylases)
CarTin_tx_s317_comp561354_c0_seq1	48	-2.9	0.0	Ethylene Response Factor
CarTin_tx_s317_comp145452_c0_seq1	53	-2.9	0.0	Zeatin O-glucosyltransferase-like
CarTin_tx_s317_comp26765_c0_seq1	59	-2.6	0.0	Ethylene Response Factor
CarTin_tx_s317_comp26483_c0_seq2	153	2.4	0.0	Uncharacterised/Hypothetical Protein
CarTin_tx_s317_comp28573_c0_seq1	61	-2.5	0.0	Uncharacterised/Hypothetical Protein
CarTin_tx_s317_comp487373_c0_seq1	32	2.2	0.0	Proteinase Inhibitor
CarTin_tx_s317_comp176356_c0_seq1	22	-2.5	0.0	Non-specific Lipid-Transfer Protein-like
CarTin_tx_s317_comp147113_c0_seq1	23	2.4	0.0	Sulfur Deficiency-Induced 1
CarTin_tx_s317_comp32761_c0_seq1	**18**	**2.5**	**0.0**	**Flowering Locus T (FT)-Like**
CarTin_tx_s317_comp7986_c0_seq1	202	-2.4	0.0	Cysteine Proteinase COT44-like
CarTin_tx_s317_comp34117_c0_seq1	14	-2.3	0.0	Oleosin-like
CarTin_tx_s317_comp366899_c0_seq1	28	-2.2	0.0	Stellacyanin/Mavicyanin-like
CarTin_tx_s317_comp46857_c0_seq1	119	-2.3	0.0	Defensin-like
CarTin_tx_s317_comp26440_c0_seq1	36	-2.2	0.0	GDSL Esterase/Lipase
CarTin_tx_s317_comp34793_c0_seq1	17	-2.2	0.0	Seed Maturation/Late Embryogenesis
CarTin_tx_s317_comp14924_c0_seq1	23	2.1	0.0	Cytochrome B5-like
CarTin_tx_s317_comp32578_c0_seq1	27	-2.1	0.0	Nuclear Pore Complex Protein/RNA Helicase
CarTin_tx_s317_comp188549_c0_seq1	32	-2.1	0.0	Periaxin-like
CarTin_tx_s317_comp44200_c0_seq1	63	-2.1	0.0	Kirola-like
CarTin_tx_s317_comp4179_c0_seq1	18	2.1	0.0	Elongation Factor 1-alpha
CarTin_tx_s317_comp31932_c0_seq1	55	-2.0	0.0	Non-specific Lipid-Transfer Protein-like
CarTin_tx_s317_comp5504_c0_seq1	34	-2.0	0.0	Glucan Endo-1,3-beta-glucosidase 13-like
CarTin_tx_s317_comp4818_c0_seq1	103	-2.0	0.0	Non-specific Lipid-Transfer Protein-like

After performing a BLASTP search against the NCBI non-redundant amino acid database, four contigs contained substantial sequence homology to genes previously associated with flowering time regulation in other plants. These were *APETALA 1-LIKE* (*CtAP1-LIKE*), *MADS-BOX CONTAINING 1* (*CtMADS1*), *FLOWERING LOCUS T-LIKE* (*CtFT-LIKE*) and *VERNALISATION 1-LIKE* (*CtVRN1-LIKE*; Table 3.9). Based on annotations found in other plant species, the remaining 26 transcripts did not appear to contain any annotation or sequence homology that associated them with the vernalisation response.

TABLE 3.9: Key differentially expressed transcripts in winter safflower aligned to the spring safflower transcriptomic reference, after exposure to four weeks of vernalisation conditions. 'n' indicates the number of complete biological replicates in each sample. Sequence homology was identified by BLASTP (v2.2.28+), dashes in the adjusted p-value column could not be calculated.

	Experiment 1 (n=4)			Experiment 2 (n=3)			
Contig	Mean Counts	log2 Fold Change	Adjusted p-value	Mean Counts	log2 Fold Change	Adjusted p-value	Homology
CarTin_tx_s317_comp33519_c0_seq70	295	5.92	0.00	179	0.02	0.18	*CtVRN1-LIKE*
CarTin_tx_s317_comp26769_c0_seq1	124	4.05	0.00	23	0.07	-	*CtAP1-LIKE*
CarTin_tx_s317_comp33367_c7_seq4	61	4.17	0.00	15	0.07	-	*CtMADS1*
CarTin_tx_s317_comp32761_c0_seq1	18	2.46	0.00	7	0.03	-	*CtFT-LIKE*

As there may have been transcripts expressed in winter safflower that were not present in spring safflower, read counts aligned to the winter safflower *de novo* transcriptome were analysed for any differential expression. DESeq2 identified 506 winter safflower transcripts as differentially expressed ($\alpha = 0.05$). Restricting the analysis to those transcripts that were significantly differentially expressed ($\alpha = 0.05$) and had at least a one-fold change in expression, 285 transcripts were identified as differentially expressed. Further restricting the analysis to very differentially expressed transcripts ($\alpha = 0.01$) with at least a two-fold change in expression resulted in twenty transcripts (Table 3.10; Appendix B, Figs. B.2 and B.4).

TABLE 3.10: Differentially expressed winter safflower transcriptomic contigs. The transcripts in *italics* were not found in the differentially expressed transcripts from spring safflower. **Bold** faced rows indicate sequences annotated as involved in the vernalisation response in other organisms.

Contig	Mean Counts	log2 Fold Change	Adjusted p-value	Annotation
CarTin_tx_WSRC03_Scaff20021	**139**	**3.9**	**0.0**	**MADS1 (MADS box containing)**
CarTin_tx_WSRC03_Scaff43593	**95**	**3.4**	**0.0**	**Vernalisation 1 (VRN1)-like**
CarTin_tx_WSRC03_Scaff23886	**29**	**3.3**	**0.0**	**MADS1 (MADS box containing)**
CarTin_tx_WSRC03_Scaff32547	**90**	**3.1**	**0.0**	**Apetala 1 (AP1)-like**
CarTin_tx_WSRC03_Scaff23955	250	2.2	0.0	Uncharacterised/Hypothetical Protein
CarTin_tx_WSRC03_Scaff28193	109	2.3	0.0	(Multiple Oxidases/Hydroylases)
CarTin_tx_WSRC03_Scaff32883	80	2.5	0.0	Glucose and Ribitol Dehydrogenase
CarTin_tx_WSRC03_Scaff62842	24	-2.6	0.0	Zeatin O-glucosyltransferase-like
CarTin_tx_WSRC03_Scaff65648	**40**	**2.7**	**0.0**	**Apetala 1 (AP1)-like**
CarTin_tx_WSRC03_Scaff23396	105	-2.4	0.0	Ethylene Response Factor
CarTin_tx_WSRC03_Scaff57705	**34**	**2.6**	**0.0**	**Flowering Locus T (FT)-Like**
CarTin_tx_WSRC03_Scaff23972	26	2.5	0.0	Uncharacterised/Hypothetical Protein
CarTin_tx_WSRC03_Scaff27766	205	-2.6	0.0	Cysteine Proteinase COT44-like
CarTin_tx_WSRC03_Scaff35647	60	-2.5	0.0	Kirola-like
CarTin_tx_WSRC03_Scaff65369	*26*	*-2.2*	*0.0*	*Proteoglycan 4/Periaxin-like*
CarTin_tx_WSRC03_Scaff61146	*19*	*2.1*	*0.0*	*Xyloglucan Endotransglycosylase 1*
CarTin_tx_WSRC03_Scaff64947	20	-2.2	0.0	Non-specific Lipid-Transfer Protein-like
CarTin_tx_WSRC03_Scaff5488	81	-2.1	0.0	Non-specific Lipid-Transfer Protein-like
CarTin_tx_WSRC03_Scaff62404	*60*	*-2.1*	*0.0*	*SLE2/Late Embryogenesis Abundance Protein*
CarTin_tx_WSRC03_Scaff66630	26	-2.0	0.0	Uncharacterised/Hypothetical Protein

We compared the differentially expressed transcripts of both spring and winter safflower to determine if there were any transcripts expressed in winter safflower that were not present in spring safflower. When the 30 transcripts that were very differentially expressed in spring safflower ($\alpha = 0.01$) were aligned to the 20 transcripts that were differentially expressed in winter safflower ($\alpha = 0.05$), 17 of the 20 winter safflower transcripts aligned to very differentially expressed transcripts in the spring safflower, leaving three transcripts significantly differentially expressed in winter safflower but not found in the list of very significantly differentially expressed spring safflower transcripts (Table 3.10). When these three winter safflower transcripts were aligned against the complete spring safflower transcriptome using Biokanga 'Blitz', all alignments reported a high level of sequence homology against a spring safflower transcript (Table 3.11). After searching through the NCBI using BLASTX, each winter safflower transcript returned alignments to existing annotations of other proteins, none of the annotations were obvious candidates for the vernalisation response pathway in other species.

TABLE 3.11: Results of the three very significantly differentially expressed winter safflower transcripts not found in the very significantly differentially expressed spring safflower transcripts when aligned against the complete spring safflower transcriptome.

Winter Transcript	Length (bp)	Spring Transcript	Length (bp)	Alignment Length (bp)	Annotation (winter)
CarTin_tx_WSRC03_Scaff61146	380	CarTin_tx_s317_comp19211_c0_seq1	1088	380	Xyloglucan Endotransglycosylase 1
CarTin_tx_WSRC03_Scaff62404	616	CarTin_tx_s317_comp34877_c0_seq1	722	616	SLE2/Late Embryogenesis Abundance Protein
CarTin_tx_WSRC03_Scaff65369	431	CarTin_tx_s317_comp274893_c0_seq1	222	173	Proteoglycan 4/Periaxin-like

Of all of the transcripts that were found to be differentially expressed, there were four that were annotated to a degree that allowed for their identification as candidates for the vernalisation response in safflower. The next stage of the transcriptomic analysis was to see how these transcripts changed in their expression during the vernalisation response, and whether the transcripts characterised in the Experiment 1 would be identified as differentially expressed.

3.3.4 Experiment 2: Vernalisation at Five Time Points

Experiment 1 revealed a number of candidate transcripts involved in the vernalisation response. Experiment 2 examined the vernalisation response in spring and winter safflower in higher resolution using five time points, 0, 5, 10, 15 and 20 days exposed to vernalisation conditions instead of two, 0 and 28 days. Using the DESeq2 R package, transcripts that were significantly differentially expressed ($\alpha = 0.05$ across the five time points) were identified, where the expression profile of the transcript was significantly different between spring and winter safflower. There were 73 transcripts that were identified as differentially expressed (Table A.1), which included the four transcripts that were identified as differentially expressed and annotated from Experiment 1, *CtAP1-LIKE*, *CtMADS1*, *CtFT-LIKE* and *CtVRN1-LIKE*. These 73 transcripts were used to search the NCBI protein database using BLASTX and annotated based on the reported alignments. Based on these alignments, the 73 differentially expressed transcripts were broadly classified into four different categories (Appendix A, Table A.1):

i) annotated transcripts that are believed to be involved in the vernalisation response

ii) annotated transcripts that are differentially expressed but their role in the vernalisation response is unclear

iii) differentially expressed transcripts that have no annotation e.g. 'No hit found' or 'hypothetical protein'

iv) transcripts where the winter safflower counts did not change (up or down) as the time exposed to vernalisation conditions increased

Of the 73 differentially expressed transcripts, four were annotated and thought to be members of the vernalisation pathway, 22 were annotated but their role, if any, in the vernalisation pathway was unclear, 11 were not annotated, but were differentially expressed and 36 transcripts did not show differential expression in winter safflower cultivars, irrespective of the time exposed to vernalisation conditions (Fig. 3.3).

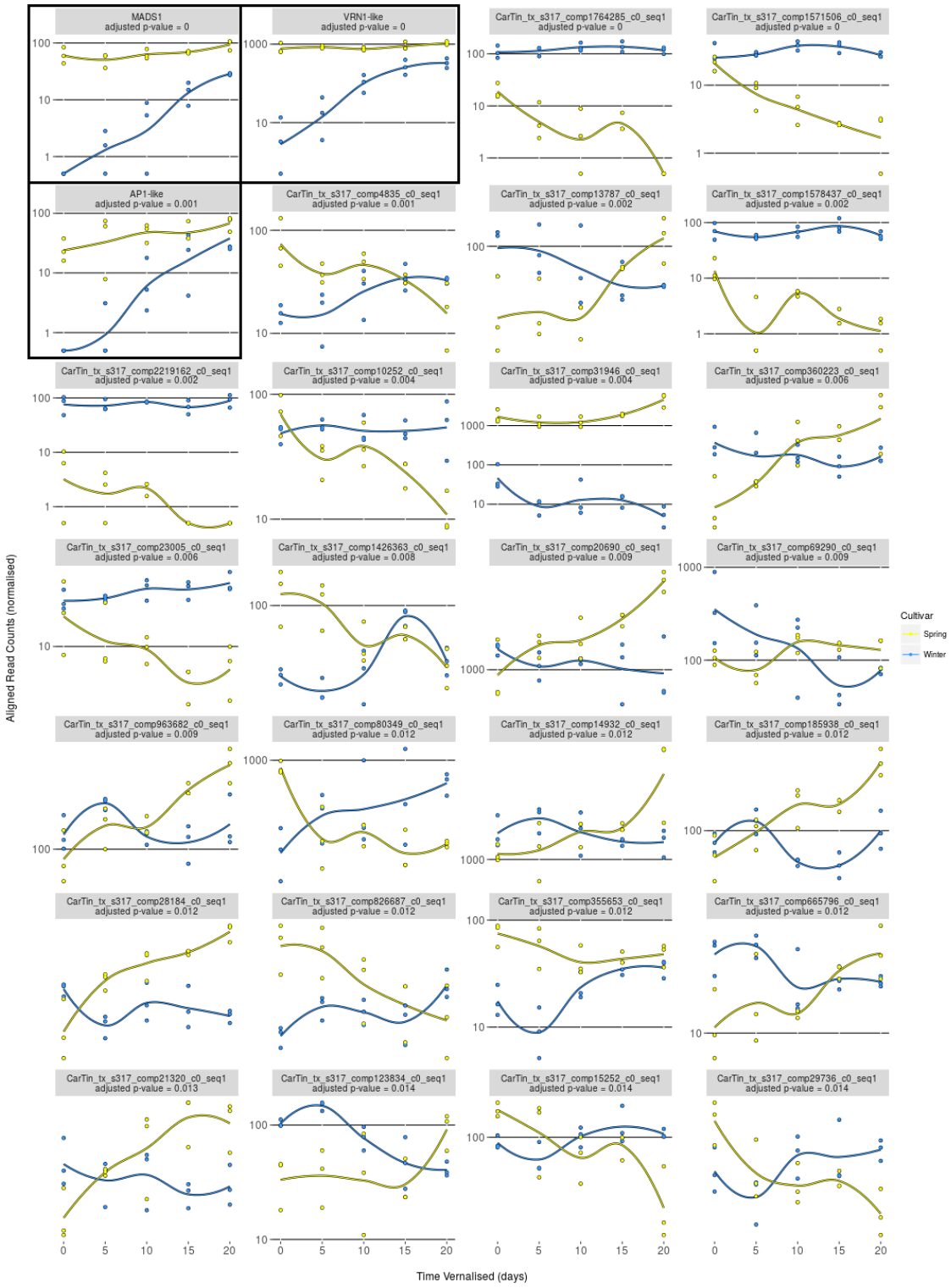


FIGURE 3.3: All significantly differentially expressed (using RNASeq data and DESeq2) transcripts ($\alpha = 0.05$) in spring and winter safflower from Experiment 2, where plants were exposed to vernalisation conditions from 0 days to 20 days. Annotated transcripts have been outlined. Part 1 of 3.

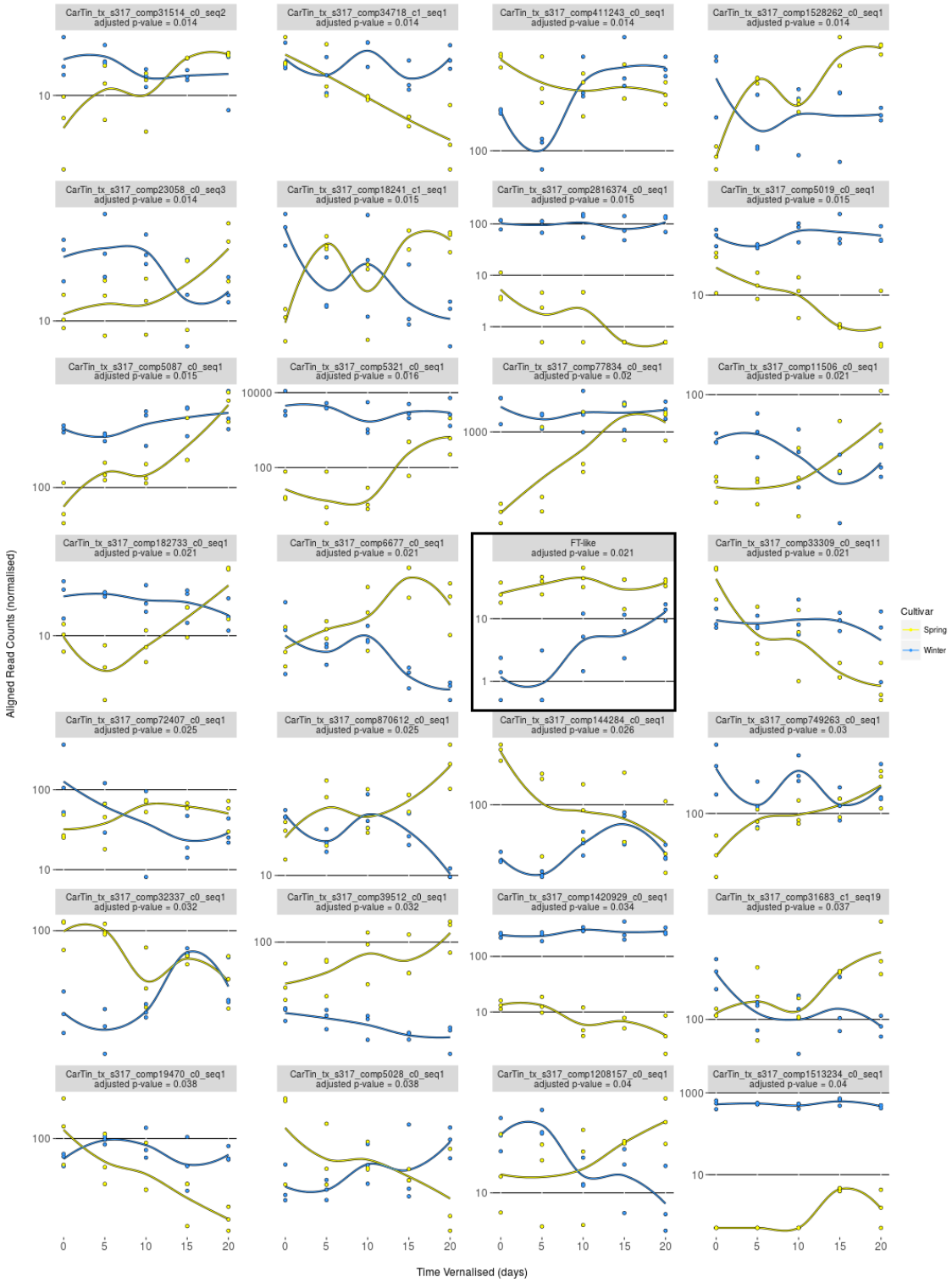


FIGURE 3.3: All significantly differentially expressed (using RNASeq data and DESeq2) transcripts ($\alpha = 0.05$) in spring and winter safflower from Experiment 2, where plants were exposed to vernalisation conditions from 0 days to 20 days. Annotated transcripts have been outlined. Part 2 of 3.

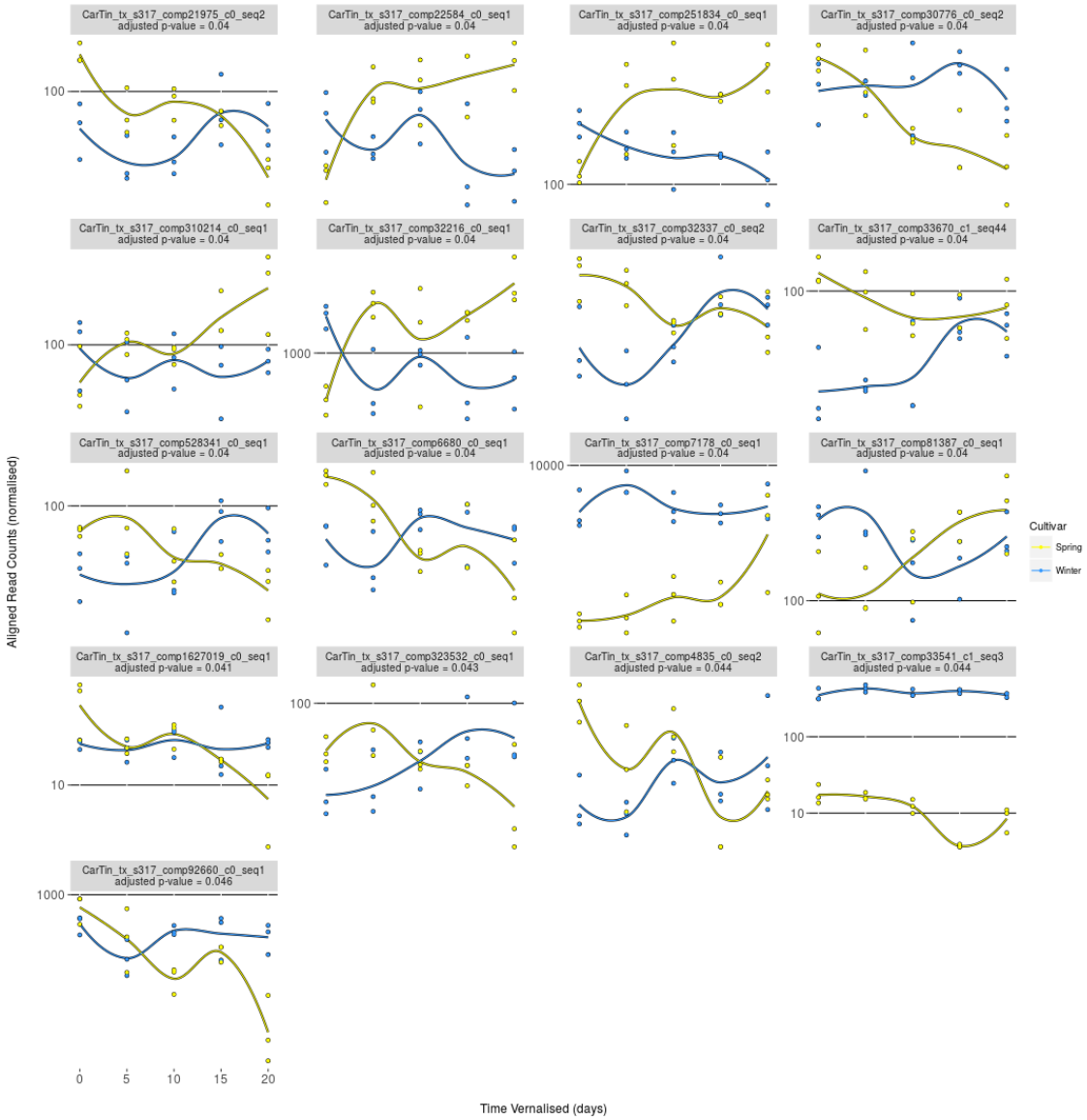


FIGURE 3.3: All significantly differentially expressed (using RNASeq data and DESeq2) transcripts ($\alpha = 0.05$) in spring and winter safflower from Experiment 2, where plants were exposed to vernalisation conditions from 0 days to 20 days. Part 3 of 3.

The four differentially expressed sequences characterised in Experiment 1 (Section 3.3.3), *CtAP1-LIKE*, *CtMADS1*, *CtFT-LIKE* and *CtVRN1-LIKE*, were examined in Experiment 2. Also examined were two housekeeping transcripts identified via sequence homology as *CtACTIN1-LIKE* (CarTin_tx_s317_comp36134_c0_seq1) and *GLYCERALDEHYDE 3-PHOSPHATE DEHYDROGENASE-LIKE* (*CtGAPDH-LIKE*; CarTin_tx_s317_comp34418_c0_seq1). Timecourse data for both winter and spring safflower showed that *CtAP1-LIKE*, *CtMADS1*, *CtFT-LIKE* and *CtVRN1-LIKE* were significantly differentially expressed (adjusted p-value < 0.05). As expected, *CtACTIN1-LIKE* and *CtGAPDH-LIKE* were not significantly differentially expressed for either cultivar, regardless of the time spent in vernalisation conditions (Fig. 3.4). The vernalisation timecourse also showed similar differential expression profiles for

CtAP1-LIKE, *CtMADS1*, *CtFT-LIKE* and *CtVRN1-LIKE* that increased in expression as the time exposed to vernalisation conditions increased. In spring safflower, these transcripts were expressed regardless of the time exposed to vernalisation conditions and did not vary in their expression levels.

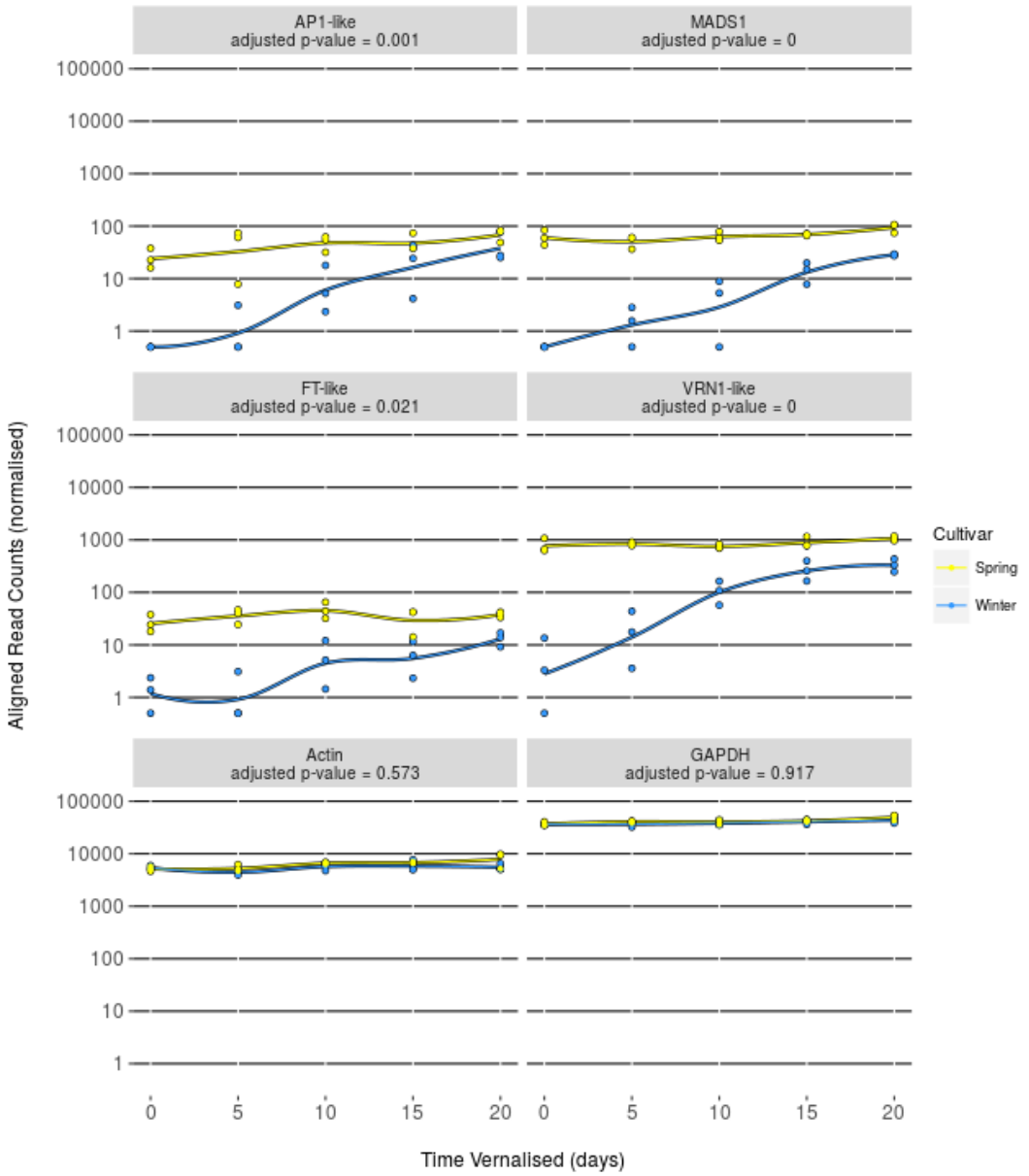


FIGURE 3.4: Differential expression (using RNASeq data and DESeq2) of transcripts in spring and winter safflower from Experiment 2 where the cultivars were exposed to vernalisation conditions from 0 days to 20 days.

An additional four significantly differentially expressed transcripts from Experiment 2 were identified as potentially involved in the vernalisation response in safflower. Two of these, CarTin_tx_s317_comp20690_c0_seq1 and CarTin_tx_s317_comp528341_c0_seq1, were characterised using sequence homology from BLASTP alignments from the NCBI as putative orthologues of Homeodomain-like/MYB transcription factors. Two,

CarTin_tx_s317_comp870612_c0_seq1 and CarTin_tx_s317_comp32216_c0_seq1 were identified as Zinc-Finger proteins RING/FYVE/PHD-type. The BLASTP alignments for the other differentially expressed transcripts did not reveal any other homology to proteins previously associated with the vernalisation response.

3.3.5 Alignment of Annotated Sequences from Spring and Winter Safflower

3.3.5.1 *APETALA 1-LIKE* (*CtAP1-LIKE*)

When the two differentially expressed *de novo* transcripts CarTin_tx_s317_comp26769_c0_seq1 and CarTin_tx_WSRC03_Scaff32547 were translated and aligned to one another (Appendix E, Fig. E.2), two regions of dissimilarity were observed. The first was two adjacent residues at position 63 (Y) and 64 (R) in winter *Ct*AP1-LIKE that were not present in spring *Ct*AP1-LIKE. The second was a region of consecutive residues (STSQA) at positions 233 to 237 that were present in spring *Ct*AP1-LIKE but absent in winter *Ct*AP1-LIKE.

Comparing these translated safflower *Ct*AP1-LIKE polypeptides against sequences identified in the BLASTP alignment to the NCBI, even with the missing residues at the start of the CtAP1-LIKE safflower transcripts, there was a high sequence homology in the N-terminus between the translated safflower transcripts *Arabidopsis* AP1 and the three sequences, AP1-LIKE, LFY-LIKE and the third MADS-box sequence sourced from the *Chrysanthemum* species. This similarity remains until approximately position 80, where the *Arabidopsis* AP1 sequence diverges. The *Chrysanthemum* and safflower polypeptides begin to diverge at position 106 and differing widely after residue 120. *Arabidopsis* LFY was another transcript that was extracted and compared to the two translated safflower transcripts due to the *FLY-LIKE* annotated AP1-LIKE from *Chrysanthemum lavandulifolium*. However, when the *Arabidopsis* LFY was compared against the two translated safflower *CtAP1-LIKE* sequences, almost no substantial sequence similarity was seen (data not shown).

3.3.5.2 *FLOWERING LOCUS T-LIKE* (*CtFT-LIKE*)

There were three translated safflower sequences that returned substantial sequence homology to FT in other *Asteraceae* (*Helianthus annuus*, *Lactuca sativa* and *Chrysanthemum morifolium*). CarTin_tx_s317_comp32761_c0_seq1 was extracted from the spring safflower transcriptome and CarTin_tx_WSRC03_Scaff57705 and CarTin_tx_WSRC03_Scaff93957 from the winter safflower transcriptome (Appendix E, Fig. E.1). Three differences were observed upon comparison of the translated sequences. The N-terminal residues for CarTin_tx_WSRC03_Scaff93957 were truncated to position 78. In CarTin_tx_s317_comp32761_c0_seq1 there was both an extended C-terminus consisting of an additional 28 amino acid residues and a substitution (M) at position 115, compared to an (I) residue at this position for the two winter safflower polypeptides. When these three safflower sequences were compared to *At*FT, a high

level of sequence homology was observed between position 80 through to the end of the *At*FT and the winter safflower translation product at position 175.

3.3.5.3 *MADS-BOX DOMAIN CONTAINING 1* (*CtMADS1*)

Four safflower sequences were determined to share substantial sequence homology to a number of *FLC* and *FLC-LIKE* sequences (Appendix E, Fig. E.3), CarTin_tx_s317_comp33367_c7_seq4 and CarTin_tx_s317_comp33367_c7_seq1 from spring safflower and CarTin_tx_WSRC03_Scaff20021 and CarTin_tx_WSRC03_Scaff23886 from winter safflower. In spring safflower, only one of the two *FLC-LIKE* transcripts were significantly differentially expressed, CarTin_tx_s317_comp33367_c7_seq4 but not CarTin_tx_s317_comp33367_c7_seq1. However, both winter safflower *FLC-LIKE* transcripts CarTin_tx_WSRC03_Scaff20021 and CarTin_tx_WSRC03_Scaff23886 were differentially expressed. When translated, there were a number of differences between these *Ct*MADS1-LIKE polypeptides. Between the two spring *Ct*MADS1 sequences, there is a substantial change in amino acid residues at positions 169 to 176 (VIRYNKVI) in CarTin_tx_s317_comp33367_c7_seq1. In the two winter *Ct*MADS1 sequences, there were missing N-terminus residues before position 11 and a single missing amino acid at position 87.

While there was not a large amount of sequence homology between *Arabidopsis* FLC and the safflower *Ct*MADS1 sequences, BLASTP identified the first 75 residues as part of the MADS-box family using sequences from the NCBI. There were also a number of BLASTP sequences in non-*Arabidopsis* angiosperms that identify this transcript as FLC-LIKE, including protein sequences determined to be FLC in *Carya cathayensis*, *Juglans regia* and *Theobroma cacao*. While the MADS-box motif was present in *At*FLC and other annotated FLC-LIKE proteins, the presence of alignments to a number of other MADS-box containing proteins e.g. CAULIFLOWER-A (CAL) and LEAFY (LFY) and other amino acid sequences annotated as 'MADS-box (domain) proteins' made it difficult to conclusively annotate these sequences as FLC-like.

3.3.5.4 *VERNALISATION 1-LIKE* (*CtVRN1-LIKE*)

Two safflower sequences were extracted from spring safflower with homology to *VRN1*, CarTin_tx_s317_comp33519_c0_seq70 and CarTin_tx_WSRC03_Scaff43593 (Appendix E, Fig. E.4). When translated, no differences were observed when comparing the winter and spring safflower *Ct*VRN1-LIKE. When *Ct*VRN1-LIKE was compared to VRN1 from *Arabidopsis* and barley, there was a greater sequence homology to barley VRN1.

3.3.6 RT-qPCR Validation of RNA-seq Data

In both Experiment 1 and 2, four RNA-seq transcripts were identified as differentially expressed using *in silico* techniques. Transcripts from Experiment 1 and 2 that were

identified as very significantly differentially expressed and annotated as candidates of the vernalisation response pathway in safflower i.e. *CtAP1-LIKE*, *CtFT-LIKE*, *CtMADS1-LIKE* and *CtVRN1-LIKE*, were validated using RT-qPCR. Primers (Appendix F) based on the RNA-seq data and a previously designed pair of *Ct*ACTIN1-specific primers were used as the normalising gene. When the RNA-seq expression profiles of *CtAP1-LIKE*, *CtFT-LIKE*, *CtMADS1* and *CtVRN1-LIKE* (Fig. 3.5) were compared against the RT-qPCR expression profiles (Fig. 3.6), both analyses showed that in all four, their level of expression increased significantly after exposure to vernalisation conditions.

In Experiment 2, the expression of *CtMADS1* and *CtFT-LIKE* was examined as the time exposed to vernalisation conditions increased. In spring safflower, both *CtMADS1* and *CtFT-LIKE* were expressed at varying levels across all time points in spring safflower. In winter safflower, there was no detectable expression of either transcript in the unvernalised samples. Both the RNA-seq and RT-qPCR approaches (Figs 3.7 and 3.8) clearly demonstrated that, in winter safflower, the expression of *CtMADS1* and *CtFT-LIKE* increases as the time exposed to vernalisation conditions increases, starting at approximately ten days of exposure.

First, all four transcripts, *CtAP1-LIKE*, *CtFT-LIKE*, *CtMADS1-LIKE* and *CtVRN1-LIKE*, were confirmed as differentially expressed. Second, a similar expression profile was observed for *CtFT-LIKE* and *CtMADS1-LIKE* when the results of the differential expression and RT-qPCR were compared. Finally, the results from the RT-qPCR validation experiment demonstrated the accuracy of the RNA-seq differential expression analysis for both Experiments 1 and 2.

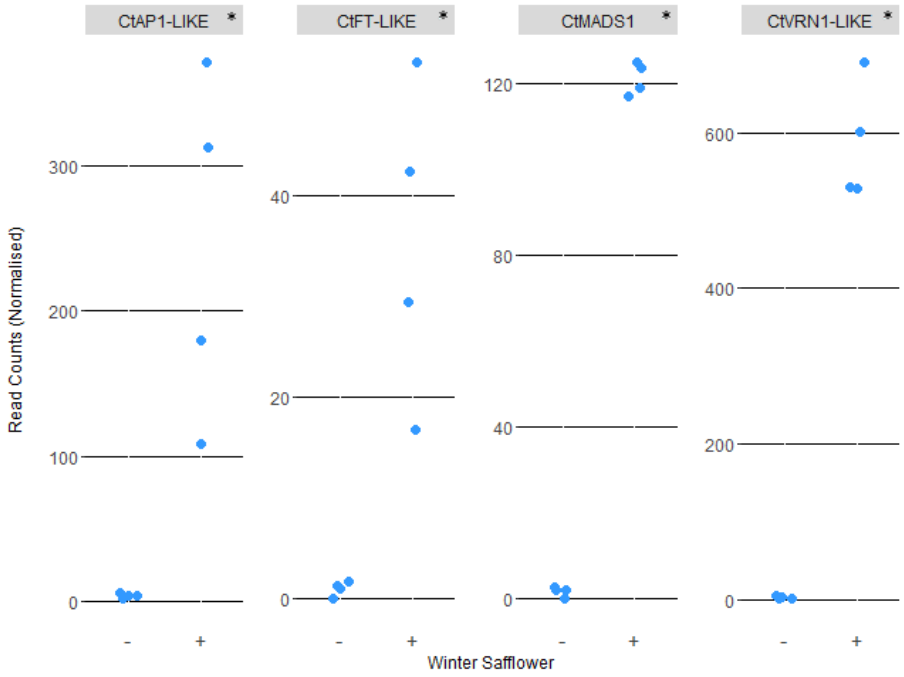


FIGURE 3.5: Expression of the four transcripts from winter safflower before vernalisation (-) and after vernalisation (+) using four biological replicates for each. Read counts were generated from back alignment data, aligning the winter safflower reads against the spring safflower reference and normalised using DESeq2. Using Walsh's t-test, significant differences ($\alpha = 0.05$) are indicated with an asterisk in the title bar.

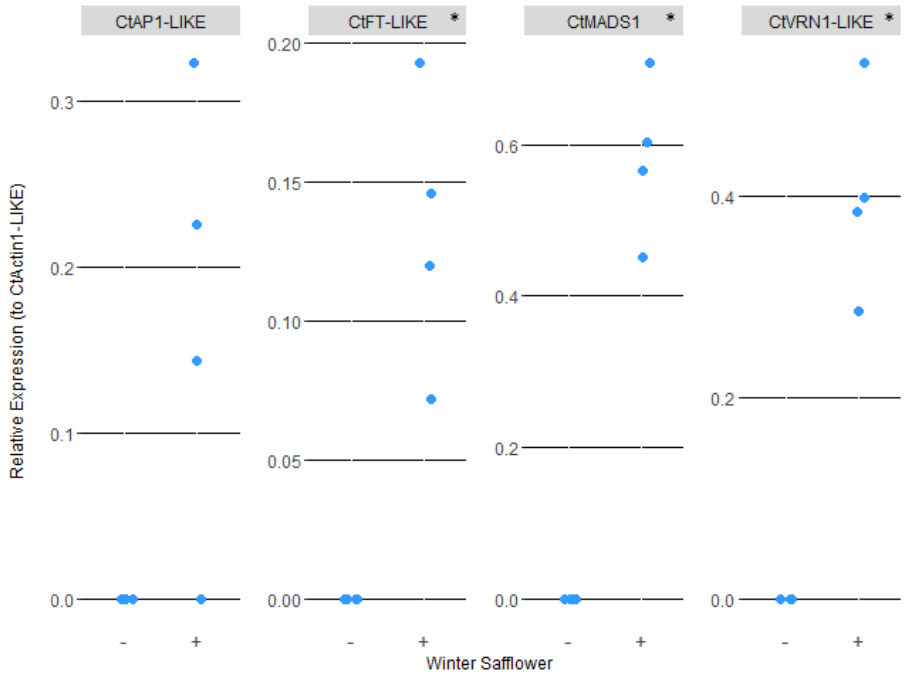


FIGURE 3.6: Expression of four transcripts from winter safflower before vernalisation (-) and after vernalisation (+), using four biological replicates for each. Using RT-qPCR, transcripts were normalised with *CtACTIN1-LIKE*. Using Walsh's t-test, significant differences ($\alpha = 0.05$) are indicated with an asterisk in the title bar.

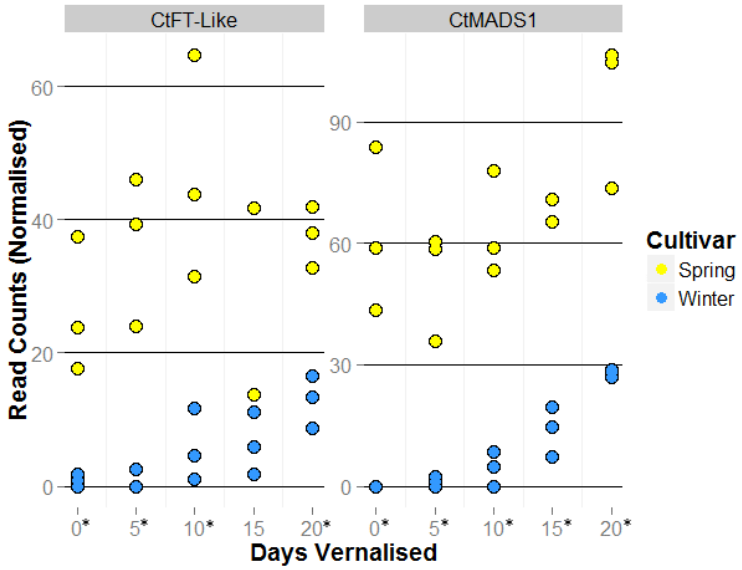


FIGURE 3.7: The expression of *CtMADS1* and *CtFT-LIKE* transcripts in Experiment 2. The RNA-Seq data shows spring safflower (yellow) expressing *CtFT-LIKE* and *CtMADS1* throughout the entire vernalisation time course. Whereas in winter safflower (blue), *CtFT-LIKE* and *CtMADS1* are not expressed in unvernalised winter safflower but gradually increases as the time spend in vernalisation conditions increase. There are three biological replicates for every time point and cultivar. Using Walsh's t-test, significant differences ($\alpha = 0.05$) between winter and spring safflower at each time point are indicated with an asterisk at the time point label.

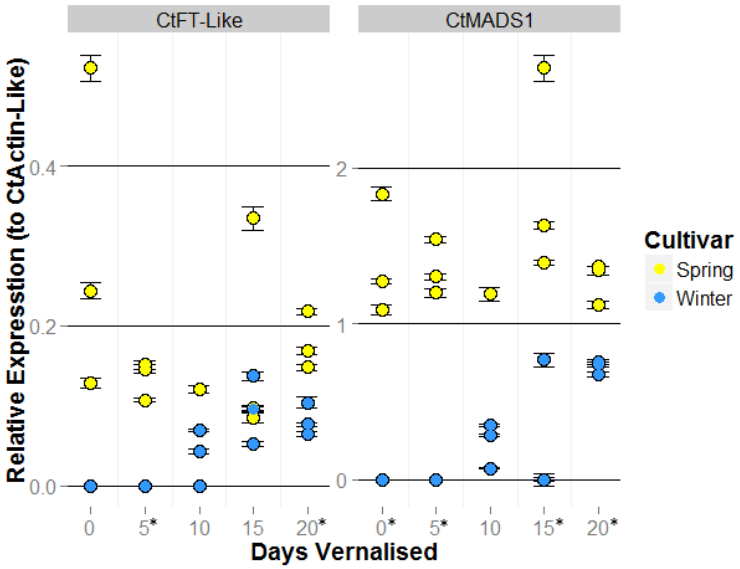


FIGURE 3.8: The expression of *CtMADS1* and *CtFT-LIKE* was identified as differentially expressed in Experiment 2 and validated using RT-qPCR. In spring safflower (yellow), both *CtFT-LIKE* and *CtMADS1* were expressed in some way throughout the entire vernalisation timecourse. In winter safflower, *CtFT-LIKE* and *CtMADS1* were not expressed until after ten days exposure to vernalisation conditions. There are three biological replicates for each time point, except at the ten day time point for spring safflower. At this time point in spring safflower for both *CtFT-LIKE* and *CtMADS1*, only two biological replicates were tested, and one result from *CtFT-LIKE* and *CtMADS1* was removed as an outlier (data not shown). Error bars indicate standard error of RT-qPCR technical replicates. Using Walsh's t-test, significant differences ($\alpha = 0.05$) between winter and spring safflower at each time point are indicated with an asterisk at the time point label.

3.4 Discussion

3.4.1 The Creation of a Reference Transcriptomic Assembly for Safflower

At the time of authoring this Thesis, there was no publicly available multi-tissue transcriptomic reference for safflower, which is essential for conducting any sort of differential expression analysis. Therefore, a *de novo* transcriptomic assembly for safflower was created and assessed.

One of the major outcomes of these experiments was the creation of a multi-tissue reference transcriptome for spring safflower. This transcriptome was comprised of approximately 147,000 contigs, with a total length of approximately 145 Mbp (145,000,000 bp), just under 10% of the estimated genome size (Garnatje et al. 2006). Analysis using CEGMA and BUSCO showed that this reference transcriptome for spring safflower contained almost all of the conserved transcripts found among all Eukaryotic organisms, indicating a high quality reference transcriptome.

TABLE 3.12: A comparison of the CSIRO safflower transciptome compared against previously published safflower trasncriptomes.

Publisher	**Transcriptome**	**Raw Read Size**	**Technology**	**Contigs**	**Average Length**
Li et al. (2012)	Seed Leaf Petal	4.5 Gbp 4.3 Gbp 4.3 Gbp	Solexa Solexa Solexa	68,889 51,702 100,650	499 bp 653 bp 528 bp
Liu et al. (2015)	Early and Full Flowering (combined)	(Unknown)	454	51,591	679 bp
CSIRO	Multiple Tissues (combined)	54.9 Gbp	Illumina	146,780	985 bp

The decision to include the tissue libraries that contained reads of questionable quality in the *de novo* assembly of the spring safflower transcriptome was made under two assumptions. Reads containing low confidence but correct base calls should be supported by other reads with higher confidence base calls in the same locations and, therefore, be assembled into longer, higher confidence contigs. Likewise, where reads contained incorrect base calls and low confidence scores, they would either be rejected with minimal support from other reads, or assembled into separate contigs with other incorrect reads, but not have RNA-seq reads aligned against them in later *in silico* experiments. What has been observed in the back alignment data for the safflower transcriptome constructed using the 16 different tissues is that, in fact, the two libraries that had the least number of reads align to the transcriptome was the two libraries with the highest number of low confidence reads; the stem and imbibed seed libraries.

The *de novo* transcriptome was supported by the high scoring alignments to members of the *CtFAD2* gene family, characterised previously (Cao et al. 2013). Most of the *CtFAD2* genes primarily aligned to a single *de novo* contig. However, there were a number of cases where multiple transcriptomic transcripts aligned to a single *CtFAD2* locus, *CtFAD2.5* (CarTin_tx_s317_comp33102_c0_seq1 and CarTin_tx_s317_comp33102_c0_seq2), *CtFAD2.6* (CarTin_tx_s317_comp33397_c0_seq1, CarTin_tx_s317_comp33397_c0_seq2 and CarTin_tx_s317_comp33397_c0_seq3), and *CtFAD2.11* (CarTin_tx_s317_comp33608_c0_seq1, CarTin_tx_s317_comp33608_c0_seq2 and CarTin_tx_s317_comp33608_c0_seq3). Alignment of multiple transcripts to a single *CtFAD2* locus strongly implied that splice variants were being transcribed from these loci. Further, the sequence homology for the transcripts aligned with *CtFAD2.8* and *CtFAD2.9* was not as high as was determined for the other members of the *CtFAD2* clades. There was also evidence, in the phylogenetic tree, that because *CtFAD2.3* and *CtFAD2.4* only align to a single *de novo* transcript, CarTin_tx_s317_comp32843_c0_seq1, they may, in fact, be isoforms or variants of the same *CtFAD2* transcript. On the branch of the phylogentic tree containing CarTin_tx_s317_comp8168_c0_seq1, CarTin_tx_s317_comp8168_c0_seq2 and CarTin_tx_s317_comp49814_c0_seq1, these three *de novo* transcripts appear to be uncharacterised members of the *CtFAD2* family. However, to confirm this, investigation into the *CtFAD2* is required, which is beyond the scope of validating the transcriptome using the *CtFAD2* family of genes.

All but three of the 20 very significantly ($\alpha = 0.01$) differentially expressed transcripts identified in the winter transcriptome were also detected in the 30 very significantly ($\alpha = 0.01$) differentially expressed transcripts from the spring transcriptome. This gives further confidence in the accuracy of the *de novo* transcriptomic assemblies for safflower, as they were constructed using different algorithms.

3.4.2 Differentially Expressed Transcripts During the Vernalisation Response

In Experiment 1, four of the 30 transcripts that were very significantly differentially expressed (*CtAP1*, *CtMADS1*, *CtFT-LIKE* and *CtVRN1-LIKE*) were initially characterised (via sequence homology) as part of the vernalisation response in spring safflower. When these transcripts were examined in Experiment 2, their expression was not only shown to increase with the extension of the vernalisation period, but that the expression profile of these transcripts was different between winter and spring safflower.

The expression profiles reported for these four transcripts using RNA-seq data was confirmed by the RT-qPCR data. Further, the expression profiles generated by using RT-qPCR returned the same expression trends as the RNA-seq approach, demonstrating that the *in silico* expression data was accurate. In addition, when the expression profiles for *CtFT* and *CtMADS1* were examined, similarity between the expression trends was again obtained using the RNA-seq and RT-qPCR methods. Together, these two distinct

approaches indicated the same functionality, i.e. these four genes are critical in the vernalisation response. But these may not, in fact, be the actual triggers themselves. Similar to the PHD-PRC2 complex in *Arabidopsis*, there may be a regulatory mechanism involved that has the role of a 'master regulator' of the vernalisation response. An additional four differentially expressed transcripts were identified via BLASTP sequence homology in Experiment 2, two Homeodomain-like/MYB transcription factors and two Zinc-finger RING/FYVE/PHD-type proteins (Appendix A, Table A.1). These four candidates may potentially be upstream triggers of differential gene expression for several of the transcripts characterised above (Amasino, 2004; Yan et al., 2014; Deng et al., 2015) and warrant further investigation as to whether they play a critical role in regulating transcripts in the vernalisation response in winter safflower.

The repeated detection of transcripts in both the spring and winter transcriptomes indicates that the cause of the differential expression of these transcripts between these two cultivars is most likely the result of epigenetic repression of genes via non-coding regions. However, identification of non-protein-coding regulatory mechanism is non-trivial without having a reference genome to compliment the transcriptomic data. This problem will be further explored in Chapter 4.

3.4.3 Characterisation of Genes in the Vernalisation Response

For the *CtMADS1* aligned transcripts, CarTin_tx_s317_comp33367_c7_seq4 and CarTin_tx_WSRC03_Scaff20021, the coding regions were determined to be identical. Because the winter transcript is only expressed after exposure to vernalisation conditions, the identical coding regions in both the winter and spring safflower indicates another source of this differential expression. This could be due to differences in a non-protein-coding region of the winter safflower genome, e.g. an additional intronic sequence or another regulator of *CtMADS1* being differentially expressed. The early truncation of CarTin_tx_s317_comp33367_c7_seq1 may be responsible for the lack of its expression in spring safflower despite having an identical amino acid sequence to CarTin_tx_s317_comp33367_c7_seq4 until position 168. In winter safflower, the single amino acid gap at position 87 in CarTin_tx_WSRC03_Scaff23886 does not appear to affect the expression of this gene when exposed to vernalisation conditions, nor does the truncation of the N-terminus of this transcript. Aligning CarTin_tx_WSRC03_Scaff23886 to a safflower genomic reference may provide enough information to characterise the 5′ end of this transcript and reveal any related homology to CarTin_tx_WSRC03_Scaff20021, CarTin_tx_s317_comp33367_c7_seq4 and CarTin_tx_s317_comp33367_c7_seq1.

Despite the BLASTP identifying a number of FLC and FLC-LIKE transcripts based on sequence homology to the above four transcripts, characterising them as 'FLC-LIKE' may be presumptuous. In *Arabidopsis*, many transcripts, such as *AtAP1*, *AtFLC* and the *AtMAF* group of genes, all contain MADS-box domains, and yet, only one is

characterised as *FLC*. In the BLASTP search of the NCBI amino acid database, using the translated sequences of the two spring and two winter safflower *CtMADS1* transcripts, a large number of alignments from other organisms annotated as *FLC-LIKE* were returned. However, this search also returned a large number of homologous sequences previously annotated as 'MADS-box domain proteins'. Based on this result, a more accurate descriptor than 'FLC-LIKE' is 'MADS-box domain containing' (or similar), at least until *Arabidopsis* transformant lines, or, better yet, knockout mutant lines in safflower can be developed to better elucidate the function of the gene encoding these transcripts. Even with *Arabidopsis* transformants, there could be a similar system to that seen in *Eustoma* species. *EgFLCL* shows a similar expression profile to that seen in winter safflower, where expression is increased after exposure to vernalisation conditions. But when transformed into an *FLC* knockout *Arabidopsis* mutant which lacks a vernalisation response, a restorative effect is not only seen, but an expression profile similar to that of *AtFLC* rather than *EgFLCL* was reported (Nakano et al. 2011). The generation of transformant lines for winter safflower where the loci encoding the transcripts CarTin_tx_WSRC03_Scaff20021 and CarTin_tx_WSRC03_Scaff23886 are knocked out would provide a clearer understanding of whether these gene products are critical in the vernalisation response pathway in safflower.

Two identical safflower *CtVRN1-LIKE* transcripts, CarTin_tx_s317_comp33519_c0_s70 in spring safflower and CarTin_tx_WSRC03_Scaff43593 in winter safflower, unexpectedly returned a higher sequence similarity to barley *VRN1* rather than to *Arabidopsis VRN1*. This could be the remnant of a gene in a common ancestor between the *Poales* and the *Asteraceae* that has remained functional in safflower for the last 200 million years. Like the *CtMADS1* transcripts, only the winter *CtVRN1-LIKE* transcript was differentially expressed, despite both winter and spring *CtVRN1-LIKE* genes having identical nucleotide sequences within their respective coding regions. Similar to *CtMADS1* above, aligning these *CtVRN1-LIKE* transcripts against a reference genome may reveal a non-protein-coding region or sequence responsible for the differential expression in winter safflower. Determining any synteny that exists between the coding and non-coding regions of safflower. *CtVRN1-LIKE* and *VRN1* in barley or *B. distachyon*, along with their respective upstream and downstream genomic regions, may also shed light on the regulatory mechanisms that exist for *CtVRN1-LIKE*.

3.5 Further Investigations

It is clear that there are a number of transcripts that are differentially expressed in winter safflower in response to vernalisation. But what genes these transcripts represent has only been investigated through sequence homology. From what has been seen in *Eustoma*, *EgFLC* expression increases as the time in vernalisation conditions was extended. This was the opposite to what was observed at the *AtFLC* locus, where expression is repressed in response to lengthening exposure to vernalisation conditions.

It was also previously found that when *EgFLC* is transformed into an FLC defective *Arabidopsis* line, *EgFLCL* expression takes on the expression profile of *AtFLC*. It could be expected that, similar to *Eustoma*, the transformation of *CtFLC-LIKE* from safflower into *Arabidopsis* will produce a similar result, which will demonstrate how *CtFLC-LIKE* operates in *Arabidopsis* but not in safflower.

The presence of a barley *VRN1* homologue is another interesting aspect of the safflower transcriptome. There is a substantial difference between translated *At*VRN1 and *Hv*VRN1 (Appendix E, Fig. E.4). While the sequence homology of the other three characterised transcripts, *CtAP1-LIKE*, *CtMADS1* and *CtFT-LIKE*, share homology with sequences from other dicotyledonous species, *CtVRN1-LIKE* shares homology with *VRN1* from a monocot species. Transforming a barley line with a non-functional *VRN1* using either a knockout or an existing cultivar with *CtVRN1-LIKE* would determine whether it is a homologue of *HvVRN1*. Further investigation is required to understand why this transcript, a major member of the vernalisation response, is present in both a monocot and a dicot when their vernalisation response is so markedly different.

Experiment 2 revealed that, while 73 transcripts were significantly differentially expressed, the number of counts for these transcripts was far lower than for Experiment 1, as evidenced by the differences in count and fold data for *CtAP1-LIKE*, *CtMADS1*, *CtFT-LIKE* and *CtVRN1-LIKE*. The RT-qPCR expression data of *CtMADS1* and *CtFT-LIKE* also reflected this when the expression data was compared between Experiment 1 and Experiment 2. In *Arabidopsis*, FT is produced in the leaves and travels through the phloem, regulating the genes in the shoot apical meristem and allowing the transition of the plant from vegetative to reproductive growth. In future experiments, targeting specific tissues and creating an expression profile for them, e.g. the leaves and the SAM, may provide a clearer picture of how the vernalisation response is regulated in safflower.

The creation of a high quality genomic reference for safflower will also allow further exploration of the vernalisation response as this will reveal details about non-protein-coding sequences or regions that cannot be analysed via a conventional RNA-Seq-based approach.

3.6 Conclusion

It is clear that the vernalisation response in winter safflower is underpinned by the differential expression of a number of key genes, identified by sequence homology rather than function. At least with regards to *MADS1*, the differences between winter and spring safflower may not be related to the coding region, rather, they potentially lie in the regulatory sequences flanking the encoding loci itself. Candidate sequences, identified as part of the vernalisation response, require further investigation via a

knock-out mutagenesis approach in either *Arabidopsis*, but preferably in safflower, to confirm their *in planta* function and to more accurately characterise their functional role in the vernalisation pathway of safflower.

Chapter 4

Genomic Basis of the Vernalisation Response in Safflower

4.1 Outline

High quality reference genomes are a fundamental resource for understanding the genetic mechanisms in any organism, not just safflower. Two different technologies were used in the assembly of the safflower genome, with each technology having its own distinct advantages and disadvantages. While Illumina read libraries contain low error rates (approximately 0.1%; Loman et al. 2012; Ferrarini et al. 2013), there is a limitation on the length of reads that can be produced, typically, around 100 bp PE reads to a maximum of 250 bp. Algorithms designed to assemble Illumina reads have difficulty correctly resolving highly repetitive regions of the genome, as the reads generated are not of a sufficient length to span these regions, leaving them unresolved or incorrectly assigned as truncated regions of high repeats, while reporting the location as being deeply covered. Conversely, reads generated using Pacific Biosciences (PacBio) technology are very long in comparison to Illumina reads, being no less than 500 bp to over 50,000 bp in some instances, but with an errors rate over one hundred times higher than Illumina read libraries, at approximately 18% (Ferrarini et al. 2013). As PacBio reads must be error corrected before assembly, they present challenges when assembling that are distinct to the assembly of Illumina reads.

The aim of this experiment was to identify a number of DNA-based markers that can be used to identify if a safflower plant or cultivar is responsive to vernalisation. To achieve this, a high quality *de novo* genome for safflower was created, using spring safflower as the reference cultivar and using reads generated with both Illumina and PacBio seuqencing technology. This combination of sequencing technology allows a far more accurate *de novo* assembly. The *de novo* genome was analysed using both parents and F_3 crosses sequences using DArTSeq technology to identify DNA fragments that could be used as markers to identify safflower varieties, cultivars and crosses that are responsive to vernalisation. The DNA fragments were compared against a previously generated genetic map (Bowers et al. 2016) to determine their location in the safflower genome.

While the original intention of this chapter was to use both the Illumina and PacBio reads for the safflower *de novo* genome assembly, due to time constraints, it was not possible to complete this assembly using the PacBio reads at the writing of this thesis. It

was also not possible to analyse the resulting digested DNA fragments of the DArTSeq analysis to confirm their application as markers for vernalisation in safflower.

I would like to acknowledge Dr Stuart Stephen from CSIRO for the development and ongoing support of the Biokanga and PacBiokanga software packages. His assistance and guidance with the creation and assessment of the draft genomic assemblies were invaluable.

4.2 Materials and Methods

4.2.1 Cultivars and Growth Conditions

Spring safflower cultivars (Chapter 2.2.1) were grown until 20 g of leaf tissue could be harvested (Chapter 2.2.2).

4.2.2 Extraction of Nuclear Genomic DNA

Every effort was made to isolate nuclear genomic DNA from safflower leaf tissue and reduce 'contamination' from mitochondrial and chloroplastic genomic material. This method was modified from Naim et al. (2012) for use with safflower.

4.2.2.1 Preparation of the Nuclear Extraction Buffer (NEB)

Nuclear genomic DNA was isolated and extracted from spring safflower leaf tissue using Nuclear Extraction Buffer (NEB). The NEB was prepared by gradually adding PVP-K30 to deionised H_2O, allowing it to fully dissolve, until a 2% final concentration was achieved. Mannitol was added to a final concentration of 0.5 M and allowed to dissolve completely before adding PIPES-KOH (10 mM final concentration, $MgCl_2$ (10 mM final concentration), L-lysine monohydrochloride (200 mM final concentration) and ethylene glycol-bis(β-aminoethyl ether)-N,N,N′,N′-tetraacetic acid (EGTA; 6 mM final concentration). After these reagents were fully dissolved, the pH was adjusted to 6.0 using 10 M NaOH. The solution was split into 500 mL aliquots and autoclaved. Just before use, 0.9 g sodium metabisulfate was added to two NEB batches and 0.2 mL β-mercaptoethanol added to a single batch (referred to as NEB-complete; NEB without β-mercaptoethanol is referred to as NEB-incomplete). Prior to use, all solutions were stored at 4°C and, wherever possible, extraction steps were performed at 4°C or on ice.

4.2.2.2 Isolation of the Nuclei

First, 20 grams of spring safflower leaf tissue, fresh or snap frozen, was added to 300 mL NEB-complete and processed in a food grade blender for three to five bursts of 10 seconds per burst. The homogenate was filtered into a measuring cylinder through four to six layers of cheesecloth, then through two to four layers of sterile miracloth. After adjusting the volume to 294 mL with NEB-complete, 6 mL of 25% Triton X-100,

prepared using NEB-complete, was gently added to the cylinder dropwise down the side of the cylinder, sealed with parafilm, gently mixed by inversion 10-20 times and incubated at room temperature for 30 min, inverting gently 10-20 times every 10 min. By gradually mixing the Triton X-100 into the slurry of safflower leaf tissue, it gently lyses the cells outer membrane, the chloroplast membrane and the mitochondrial membrane while keeping the nuclear membrane intact. The intact nuclear organelles were separated from the other fractions via centrifugation. The solution was then aliquoted evenly into six 50 mL tubes and centrifuged for at 4°C at 1,800 g for 10 min. The supernatant was discarded and the pellets gently resuspended in 50 mL NEB-incomplete. Tubes were centrifuged at 4°C at 1,800 g for 15 min and the supernatant discarded. Pellets were gently resuspended with another 5 mL of NEB-incomplete before transferring all samples to a single 50 mL tube. The volume was adjusted to 50 mL using NEB-incomplete before centrifuging at 4°C at 1,800 g for 15 min. The supernatant was then discarded. At this point, the colour of the pellet is critical and indicates if the isolation of nuclei from the plant cells has been successful. A white or slightly red pellet indicates the extracted material is particularly pure nuclear genomic material. A green coloured pellet indicated contamination with chloroplastic genetic material.

4.2.2.3 Extraction of Nuclear Genomic DNA

The pellet was gently suspended in 14 mL lysis buffer (0.5% SDS, 150 mM Tris-borate (pH 7.4), 5 mM EDTA) and incubated for 15 min at 37°C. Then 1.4 mL 5 M potassium acetate (pH 7) was added and the tube mixed by inversion. Next, 3.5 mL 100% ethanol was added before vortexing the tube for 30 s, then an equal volume chloroform:isoamyl alcohol (24:1, $\mathrm{v/v}$) was added. The tubes were mixed by gentle inversion 10 times before placing them onto an orbital shaker for 10 min at 20 rpm then centrifuging the tube for 10 min at 1,800 g at room temperature, using a slow stop to avoid mixing of the pellet and the supernatant. The supernatant was transferred to a new 50 mL tube and an equal volume of ice cold 100% isopropanol was added. The tube was mixed by inversion before incubating for at least overnight, preferably two to three days at -20°C. After incubation, tubes were centrifuged for 10 min at 1,800 g at room temperature with a slow stop. The supernatant was discarded and the pellet washed with 10 mL 70% ethanol. The tube was left to dry, either in a fume hood or gently in a non-heated vacuum dryer, for 30 min or until there was no visible ethanol in the tube. The pellet was then resuspended in 100 µL of 10 mM Tris Buffer (pH 8.0). This DNA was then centrifuged 4°C at 14,000 g at for 30 min to precipitate any starch granules from the sample. The supernatant was transferred to a new sterile 1.5 mL microfuge tube. The DNA quality and contamination was checked on a Nanodrop spectrophotometer (Thermo Fisher Scientific$^{\mathrm{TM}}$) and by running the samples on a 1% agarose gel to assess the presence of high molecular weight DNA. Other tests were conducted at sequencing facilities to assess the quality and quantity of DNA. Some of the DNA handling steps outlined above are specified by the DNA sequencing facility to ensure high molecular

weight DNA and compatibility with subsequent DNA handling steps, e.g. library preparation. A sample was considered suitable for sequencing if the $OD_{260/280}$ was close to 1.8 and the $OD_{260/230}$ was between 2 and 2.2.

4.2.3 *De Novo* Assembly using Illumina Reads

4.2.3.1 Illumina Sequencing

Extracted DNA was sent to the AGRF for Illumina based sequencing. Seven libraries of 100 bp paired end (PE) reads with a fragment length of 180 bp, and a single library of 36 bp mate pair (MP) reads with a fragment length range of 5,000 to 15,000 bp, were generated on a HiSeq2000 as per the manufacturers instructions. Reads were archived in the CSIRO Data Access Portal (`https://data.csiro.au/dap/landingpage?pid=csiro:8449`).

4.2.3.2 Pre-Processing of Illumina Reads

The quality of the Illumina reads in the PE and MP read libraries was analysed using FastQC (v0.10.1) before combining the PE libraries into a single large library. Several filtering steps were undertaken on the read libraries. Reads from each Illumina PE library were combined before filtering with Biokanga 'Filter' software (v3.1.1), discarding any duplicate reads and reads containing more than a single ambiguous nucleotide (N; Appendix J.3). Reads were also compared to each another, discarding any read that did not overlap with another read by at least 50%. PE reads were also treated as 'dependent', meaning if a single read from a pair of reads was discarded, the other read of that pair was also discarded.

4.2.3.3 Assembly of Illumina Reads

The assembly of the safflower genome was performed as a multi step process. Filtered PE reads were assembled using Biokanga 'Assemb' software (v3.1.1), requiring a 70 bp minimum overlap to merge reads into a contig and only allowing a maximum of one ambiguous N for every 100 bp in each contig (Appendix J.3.1). Once complete, the assembly process was repeated using the 36 bp MP reads and adding these MP reads to the PE assembly (Appendix J.3.2).

4.2.3.4 Scaffolding Using Library Information

Scaffolding of an assembly utilised the insert information of a read library to connect contigs together and decrease the total fragmentation in an assembly. Two scaffolding steps were performed using Biokanga 'Scaffold' software (v3.1.1). The PE reads were first scaffolded using a total fragment size of 180 bp and a directionality of Sense (5′-3′) for Pair 1 and Antisense (3′-5′) for Pair 2 (Appendix J.3.3). The second scaffolding step used the MP reads with a total fragment size of between 5,000 and 15,000 bp and a directionality of Antisense for Pair 1 and Sense for Pair 2 (Appendix J.3.4).

4.2.3.5 Scaffolding Using the Spring Safflower *De Novo* Transcriptome

Due to the high quality of the spring safflower *de novo* transcriptome (Chapter 3) and that RNA transcripts often contain large intronic regions of non-coding DNA, this transcriptome was used to further scaffold the draft safflower genome. After the MP scaffolding step was complete, the spring safflower *de novo* transcriptome was aligned to the Illumina *de novo* genome using the Biokanga 'Blitz' software (v3.9.8) with default parameters, except increasing the maximum k-mer seed depth to 15,000, reducing the minimum length of the alignment to 5% and increasing the core extension threshold to 16 (Appendix J.3.5). The resulting transcriptomic alignment further scaffolded the *de novo* genomic assembly by using the Scaffolding Contigs Using BLAST-like Alignment Tool ('SCUBAT') software package (at time of writing, `www.nematodes.org/bioinformatics/SCUBAT/index.shtml` or `github.com/elswob/SCUBAT`). The Illumina genomic assembly was considered 'frozen' at this point (hereafter referred to as the CSIRO draft safflower genome). The CSIRO draft safflower genome was analysed using Biokanga 'Fasta2nxx' software (v3.4.7), with the quality of the assembly assessed using CEGMA (v2.4.010312; Parra et al. 2007) and BUSCO software (v1.1b1; Simão et al. 2015).

4.2.3.6 Back Alignment of Illumina Reads

To further assess the assembly quality, each of the seven read libraries was back aligned to the CSIRO draft safflower genome using the Biokanga 'Align' software (v3.9.8), allowing a single ambiguous nucleotide and up to three nucleotide substitutions across the alignment. Two different fragment length parameters were used for the back alignments. The first was a fixed fragment size of 180 bp (as reported by the commercial supplier, AGRF) and the second, a varying fragment length of 100 bp to 500 bp (Appendix J.3.6 and J.3.7 respectively).

4.2.4 *De Novo* Assembly using Pacific Biosciences (PacBio) Reads

4.2.4.1 PacBio Sequencing

DNA was extracted from spring safflower using the above protocol on two separate occasions and sent for PacBio sequencing by the Queensland University of Technology's Diamantia Institute. These DNA samples were tested for quality, as per the Diamantia Institutes requirements, and prepared using the '20 Kb Blue Pippin' library method. These samples were sequenced on a Pacific Biosciences RSII sequencer, using version 6 chemistry, as per the manufacturer's instructions. Because of the high error rate and the large variation in the distribution of reads in this sequencing technology, the process for assembling PacBio read libraries differs from assembly of Illumina reads (Fig. 4.1). Both read libraries were archived in the CSIRO Data Access Portal (`https://data.csiro.au/dap/landingpage?pid=csiro:22653`).

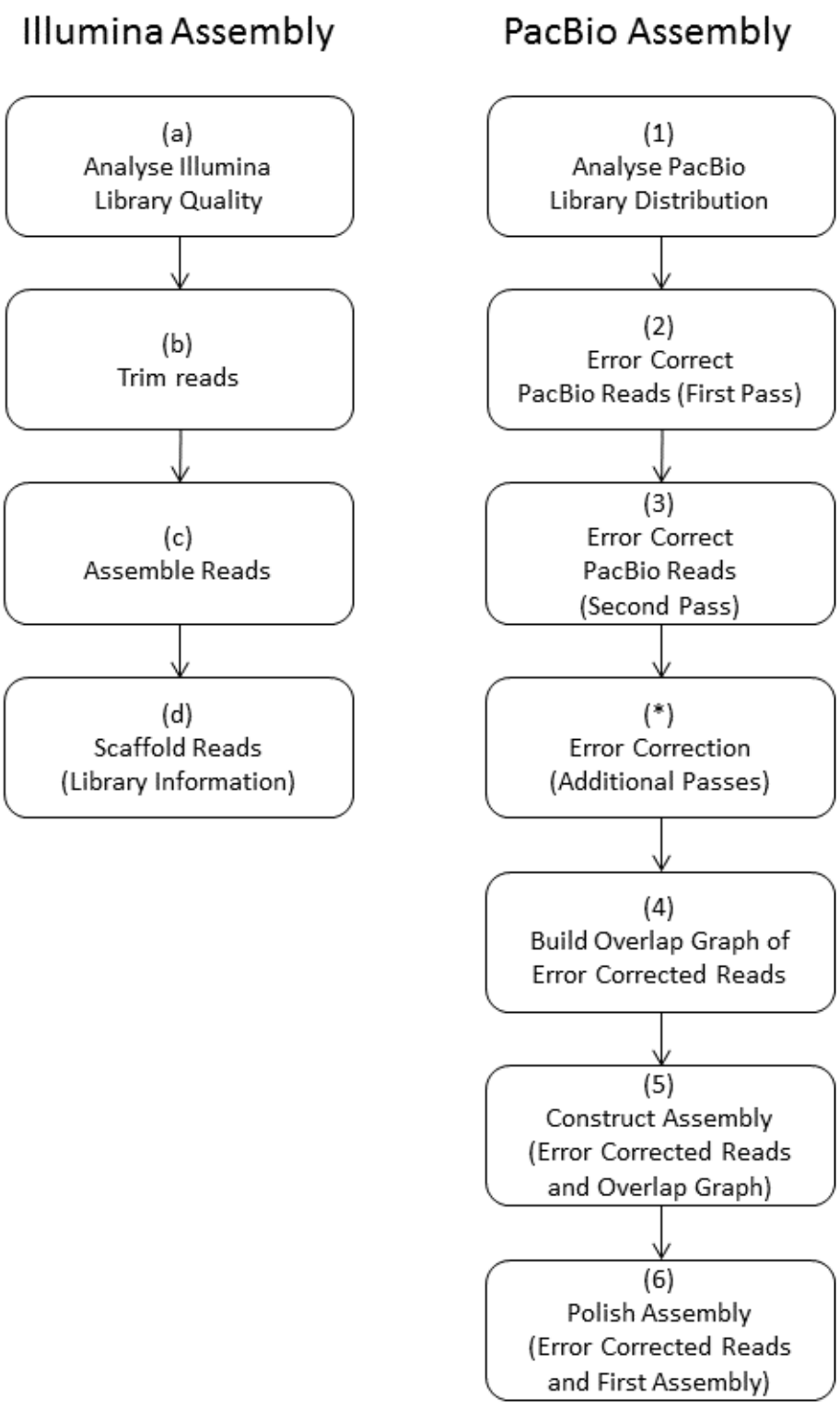


FIGURE 4.1: Steps needed to assemble Illumina sequencing libraries (left, letters) and PacBio sequencing libraries (right, numbers). The additional analysis of read distributions, error correction steps and polishing of the final assembly are required due to the high rates of errors seen in PacBio libraries (approx 18%) when compared to Illumina Sequencing (approx 0.1%).

4.2.4.2 Error Correction of PacBio Reads

Due to the large variation in the length distribution of reads generated by the PacBio sequencing technology, the reads were assessed using Biokanga 'Fasta2nxx' software (v3.9.8) to determine the parameters for error correction. PacBiokanga software was used for the error correction and assembly of PacBio reads (`https://github.com/csiro-crop-informatics/biokanga`). The major parameters for the first error correction stage used the n50 of the read distribution as the minimum read length, rounded to the nearest 500 bp with a minimum overlap of approximately half this (rounded to the nearest 500 bp). The first error correction step of the first library, using PacBiokanga 'Ecreads' software (v1.8.1; Appendix J.4.1) accepted reads between 7,500 and 35,000 bp and required a minimum overlap of 5,000 bp. Error corrected consensus reads of at least 3,000 bp were accepted and used in subsequent steps of the assembly process. While these were less stringent than the guidelines initially estimated, the aim was to capture the greatest number of the raw reads possible for use in the error correction stage. The second error correction step, also using PacBiokanga 'Ecreads' software (v1.8.1; Appendix J.4.2) for Library 1 accepted reads of between 3,000 and 35,000 bp in length and required a minimum consensus sequence of 3,000 bp to be accepted.

Error correction of reads from PacBio Library 2 were performed using PacBiokanga 'Ecreads' software (v1.9.2; Appendix J.5.1) and accepted reads of between 9,000 and 35,000 bp with a minimum overlap of at least 5,000 bp to accept the error correction. At the time of authoring this Thesis, the first phase of error correction on the second PacBio read library remained incomplete. A sample of nine of the first stage error corrected reads, with lengths of between 8,739 and 17,452 bp, were aligned to the Illumina *de novo* genome using BLASTN (v2.2.28+). These alignments were visualised using 'Kablammo' (`kablammo.wasmuthlab.org/`).

4.2.4.3 Assembly of Error Corrected PacBio Reads

The first assembly stage of the first safflower PacBio library, using PacBiokanga 'Contig' software (v1.2.1; Appendix J.4.4), required a minimum sequence length of 5,000 bp and a minimum overlap length of 5,000 bp to merge the sequences into a single contig. The final contig error correction step, using PacBiokanga 'Eccontig' software (v1.2.1; Appendix J.4.5) filtered out any contigs of low confidence i.e. contigs with a depth of less than 30 error corrected reads and any contigs with a length of less than 10,000 bp.

4.2.4.4 Analysis of the Assembly of Library 1

There were a number of steps in the analysis of the assembly produced using reads from Library 1. First, the largest contig from the PacBio *de novo* genomic assembly generated from PacBio Library 1 (referred to herein as the draft safflower chloroplast

sequence) was aligned to nucleotide sequences in NCBI using BLASTN (2.2.28+), filtered by the Entrez query 'green plants'.

Next, the draft safflower chloroplast was analysed to determine how unique this contig was using Hamming Distances. A Hamming Distance is the number of changes required for one sequence of nucleotides or amino acids to match another and is used as a method of determining the uniqueness of an assembly (Pilcher et al. 2008). The Hamming Distances were calculated across the length of the draft safflower chloroplast using a 100 bp sliding window, up to a Hamming Distance of 10. Using the same back alignment parameters as described in Section 4.2.3, reads from each of the Illumina genomic read libraries were back aligned to the draft safflower chloroplast.

4.2.5 Generation of SNP-based Markers for the Vernalisation Response

4.2.5.1 Scoring of F_3 Phenotypes

A crossing population was developed from winter and spring safflower parents (Chapter 2.2.3). Due to time and cultivation space constraints, as well as complications with phenotyping plants from the F_2 generation, twenty-four seeds from the F_3 generation of this crossing population were cultivated under controlled glasshouse conditions (Chapter 2.2.2.4). It had been previously established that an early elongation phenotype could be used as a proxy to a vernalisation, which could be scored after four weeks of growth in long day conditions (Chapter 2.2.3). Vernalisation responsive plants have many leaves and are late to elongate, whereas spring safflower plants that do not respond to vernalisation have fewer leaves and elongate ('bolt') much earlier. The plants were segregated based on this winter and spring elongation behaviour as a proxy for the vernalisation response. Four families were used to identify DNA-based SNP markers. One hundred seeds were germinated from each of the F_3 crossing families X017, X030, X100 and X395, segregated in a 3:1 ratio of spring:winter, along with twenty spring and twenty winter safflower controls. The seeds were germinated and grown as described in Chapter 2.2.3. After four weeks, the plants in each of the four families were phenotyped based on whether they had elongated (early or late; Fig. 4.2). A young leaf was sampled from every plant and sent to Diversity Arrays Technology (DArT) to run a DArTSeq analysis as per their internal protocols.

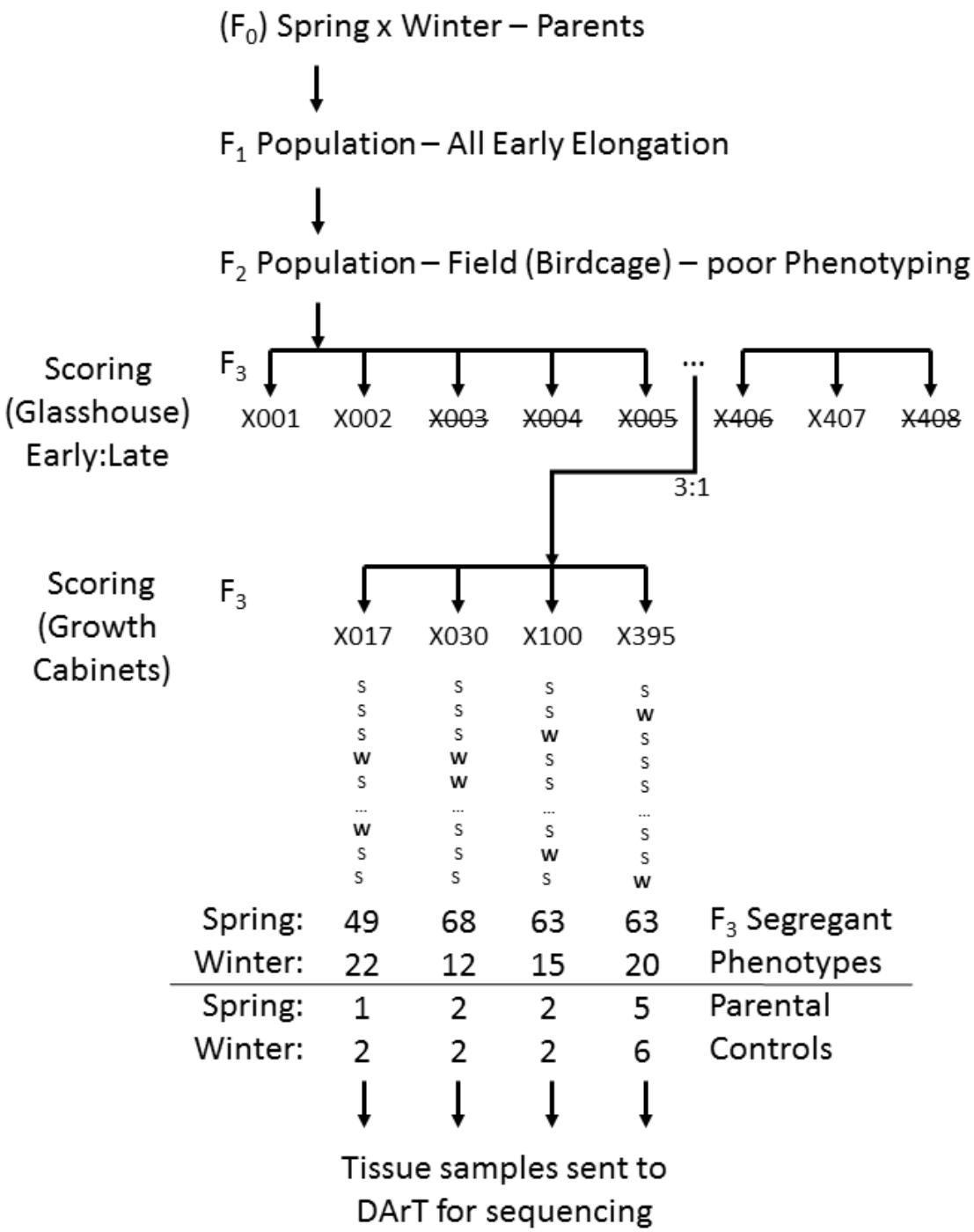


FIGURE 4.2: The preparation of the F_3 crossing population, assessment of time to elongation segregation and scoring of the samples sent to DArT.

4.2.5.2 Generation of Markers by DArT

The F_3 family leaf samples were sent to DArT in two batches. The first batch comprised of plants grown from F_3 cross X395. The second batch consisted of plants from the F_3 crosses X017, X030 and X100. The DArTSeq process extracted the DNA from each individual leaf sample then digested the DNA with two restriction endonucleases, *Pst*I, with a cut site of CTGCA'G and *Mse*I, with a cut site of T'TAA. The larger digested DNA fragments were filtered out based on size, with the remaining DNA fragments amplified. These fragments were then sequenced by DArT, producing sequence fragments of approximately 70 bp. DArT provided a report containing the presence and absence of fragments resulting from the restriction endonuclease digest (referred to herein as 'digest fragments') and the presence and absence of SNPs for each plant sample from every member of each cross and the spring and winter safflower control samples.

4.2.5.3 Comparison of Markers Across Families

For each digest fragment and SNP reported by DArT, the average score of the presence and absence in each expressed phenotype for every sample, excluding the controls, was recorded. The square of the difference between the early and late phenotype score average of the crosses, but not the controls, was calculated. The digest fragments and the SNPs were then ranked based on this score, highest to lowest. The first group of the digest fragments and SNPs from both batches of F_3 families were then extracted and compared against one another. DNA fragments from the digest markers and SNPs that were common between the two batches were aligned to the Illumina *de novo* safflower genome using BLASTN (v2.2.28+).

4.2.5.4 Mapping of Markers

The highest scoring contig from the CSIRO draft safflower genome that aligned to each digest fragment or SNP was then extracted and aligned to a genetic map of safflower constructed by another laboratory (referred to herein as the 'Bowers genetic map', with individual contigs originating from the Bowers genetic map referred to as 'Bowers contigs'; Bowers et al. 2016). The 73 differentially expressed transcripts previously identified in Chapter 3.3.4 were compared against the Bowers genetic map in a similar fashion. The highest scoring Bowers contigs that aligned to the CSIRO draft safflower genome were recorded and searched for using the Bowers genetic map. When a Bowers contig was found in the Bowers genetic map, the chromosome and location of that Bowers contig was recorded (Fig. 4.3).

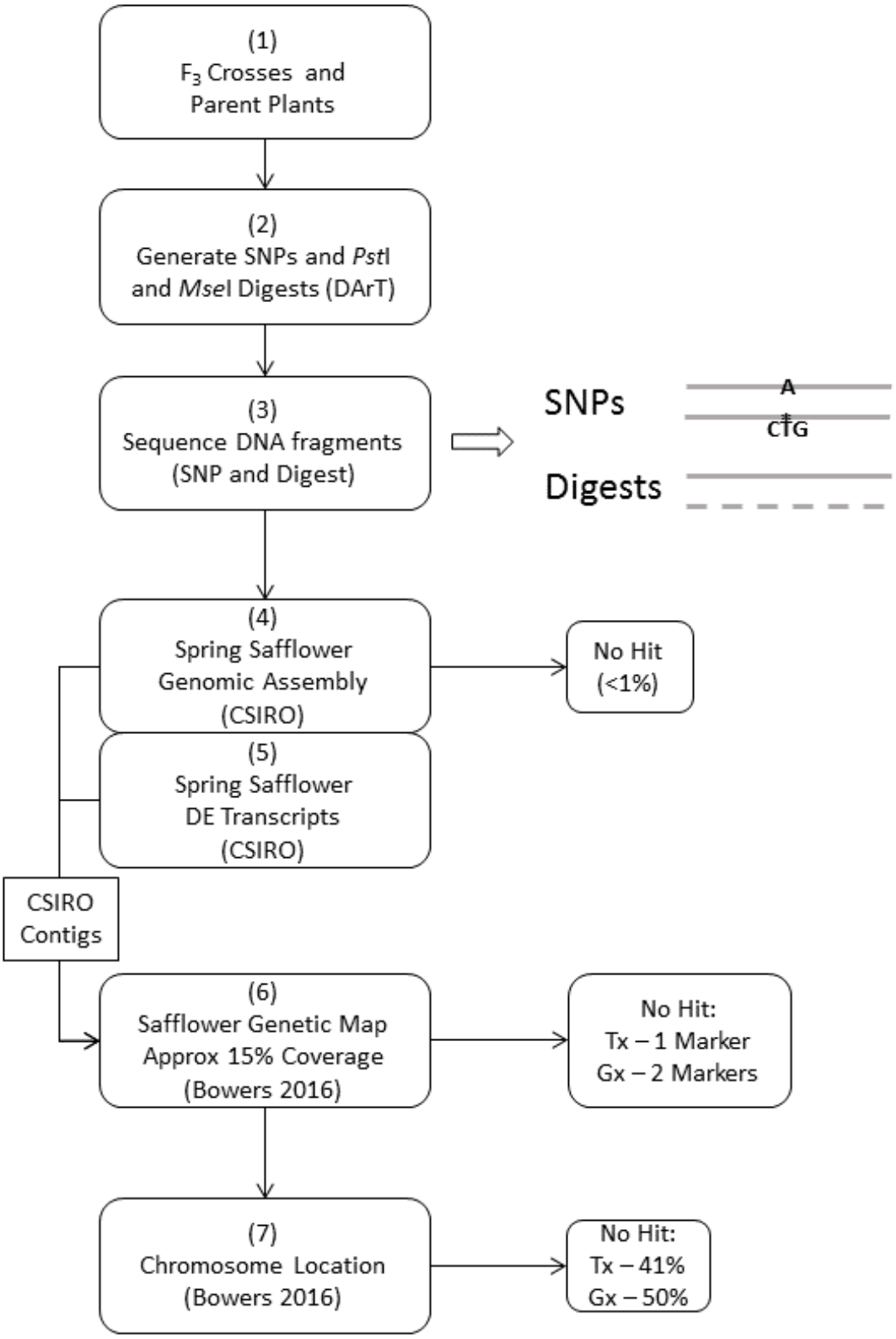


FIGURE 4.3: The process for creation of digest and SNP markers and determining if a marker is part of the vernalisation response in safflower. Tx are transcriptomic and Gx are genomic reads.

4.3 Results

4.3.1 A High Quality Draft Assembly of the Safflower Genome

4.3.1.1 Pre-processing of Illumina Reads

Seven 100 bp PE libraries and one 36 bp MP library were sequenced by the AGRF using DNA extracted from spring safflower. Using FastQC to assess the quality of the read libraries, all were found to be of good quality, with no indication of residual sequencing adaptors. All seven PE libraries were combined into a single PE library consisting of just over 1 billion pairs of reads, totalling approximately 125x coverage (200,256,499,200 bp) of the estimated safflower genome size. After filtering and deduplication, there were just under 250 million PE reads (approximately 125 million pairs of reads), totalling nearly 22 Gbp (approximately 15x coverage), used in the assembly, with an average read length of 89 bp. Before deduplication and filtering of the MP library, there were 130,770,079 pairs of 36 bp MP reads totalling 4,707,722,844 bp, approximately 3.3x coverage of the safflower genome. After filtering, there were 3,866,454 pairs of 36 bp reads, totalling 278,384,688 bp, less than 0.2x coverage. Once the quality assessment, filtering and deduplication steps were completed, the reads were ready for the assembly stage.

4.3.1.2 Assembly and Scaffolding

The Illumina *de novo* genomic spring safflower assembly was constructed using Biokanga 'Assemb' and 'Scaffold' (Table 4.1). This process resulted in an assembly of approximately 1.3 Gbp across over 2 million contigs. After the MP assembly stage, the total size decreased slightly, as did the number of contigs. After both of the scaffolding stages using the PE and MP reads, the resulting scaffolded assembly was approximately 1.15 Gbp across just over 900,000 contigs. The total number of contigs and total assembly size was slightly decreased again after processing the scaffolded assembly with SCUBAT (Table 4.1). At this point, the assembly was frozen.

TABLE 4.1: Attributes of the CSIRO draft safflower genome, constructed using Illumina Paired End and Mate Pair reads.

Sizes (bp)	PE Assembly	MP Assembly	PE Scaffolded	MP Scaffolded	**Frozen Assembly**
Total Size	**1,346,692,908**	**1,346,192,194**	**1,163,709,069**	**1,163,746,919**	**1,163,499,791**
Contigs	2,151,981	2,147,379	916,630	912,845	**904,199**
Min Length	100	100	300	300	**300**
n50	1,396	1,398	1,900	1,914	**1,940**
Mean Length	625	626	1,269	1,274	**1,286**
Max Length	21,658	21,658	25,812	25,812	**32,974**

4.3.1.3 Quality Assessment of the Assembly

The quality of the CSIRO draft safflower genome was assessed using two different software packages. CEGMA showed that 162 of the 248 conserved Eukaryotic sequences (just over 60%) were found as complete sequences, while 225 of the 248 were found as partial sequences (Table 4.2). BUSCO showed 613 of the 956 conserved sequences (approximately 64%) were found in the Illumina *de novo* genomic assembly, with 106 of the BUSCO sequences duplicated in the genome and 186 found as fragments (Table 4.3).

TABLE 4.2: CEGMA analysis on the draft safflower genome using 248 highly conserves protein sequences across Eukaryotes. The safflower *de novo* genome was constructed using Illumina Paired End and Mate Pair reads, scaffolded and refined using SCUBAT.

	Proteins	Completeness	Total	Average	Orthologous
Complete	162	65.32	333	2.06	57.41
Group 1	42	63.64	87	2.07	52.38
Group 2	31	55.36	52	1.68	41.94
Group 3	42	68.85	91	2.17	66.67
Group 4	47	72.31	103	2.19	63.83
Partial	225	90.73	598	2.66	77.78
Group 1	59	89.39	139	2.36	71.19
Group 2	49	87.50	110	2.24	65.31
Group 3	56	91.80	165	2.95	85.71
Group 4	61	93.85	184	3.02	86.89

TABLE 4.3: BUSCO analysis on the draft safflower genomeusing highly conserved protein sequences across Eukaryotic organisms. The safflower *de novo* genome was constructed using Illumina Paired End and Mate Pair reads, scaffolded and refined using SCUBAT.

BUSCOs Searched	956	%
Complete Single-copy	613	64%
Complete Duplicated	106	11%
Fragmented	186	19%
Missing	157	16 %

4.3.1.4 Back Alignment as a Method of Quality Assessment

The unfiltered reads from the seven PE libraries were aligned to the CSIRO draft safflower genome using two different fragment length parameters, a fixed length of 180 bp and a varying length ranging from 100 to 500 bp (Appendix G, Fig. G.4). In each library, regardless of the total size, when the fragment length of the read pair was fixed at 180 bp, as reported for each genomic library by the sequence provider, only a small fraction of each pair of reads back aligned to the the CSIRO draft safflower genome

(approximately 1%). Yet when the fragment length parameter was relaxed to allow read pairs to align to a varied length of between 100 and 500 bp, the number of reads that aligned increased to around 45%. This increased the read coverage from approximately 1.4x coverage (19,783,986 reads) when using the fixed fragment length to 65x coverage (918,922,580 reads) with the varying fragment length. This difference was only observed between the unique and unalignable reads. There was no change in the number of read pairs that align to multiple loci on the CSIRO draft safflower genome from back alignment using the two different fragment size parameters.

4.3.2 Determining Intron/Exon Boundaries for Identified Vernalisation Genes

By aligning the spring safflower genome and transcriptome, the intron/exon structure for transcripts could be inferred. Using Biokanga 'Blitz', the alignment of the *de novo* transcriptomic contigs (Chapter 3) against the *de novo* genomic contigs showed that 144,931 out of 146,780 (98.7%) of transcriptomic contigs aligned somewhere on the genome. Many of these transcriptomic contigs aligned across two or more genomic contigs, indicating an intron/exon structure. Vernalisation transcripts that were annotated in Chapter 3, *CtAP1-LIKE*, *CtFT-LIKE*, *CtMADS1* and *CtVRN1-LIKE* were examined in detail.

4.3.2.1 *APETALA 1-LIKE* (*CtAP1-LIKE*)

The transcriptomic contig CarTin_tx_s317_comp26769_c0_seq1, annotated as *CtAP1-LIKE*, aligned to two genomic contigs, CarTin_gx_s317_Scaff075165 and CarTin_gx_s317_Scaff298315, splitting the transcript into nine exons, with the 3′ untranslated region (UTR) adjacent to the last exon (Fig. 4.4). The first exon was aligned to CarTin_gx_s317_Scaff075165 from nucleotide position 1,343 to 1,499 in the same orientation. The remaining eight exons aligned along CarTin_gx_s317_Scaff298315 in the reverse orientation, between nucleotide positions 1,754 and 91. While no start codon was identified at the 5′ end of the *CtAP1-LIKE* transcript, 26 bp upstream on the genomic contig CarTin_gx_s317_Scaff075165 from the first aligned nucleotide from the transcriptomic contig, an ATG site was identified and is in the same reading frame as the remainder of the transcript. When translated, the N-terminus matched the amino acids present at the N-terminus of three Chrysanthemum amino acid sequences (Appendix E, Fig. E.2). There was no information available to accurately identify the length of the 5′ UTR of *CtAP1-LIKE*. A noteworthy feature for the *CtAP1-LIKE* locus is that the first intron was estimated to be at least 4,686 bp in length.

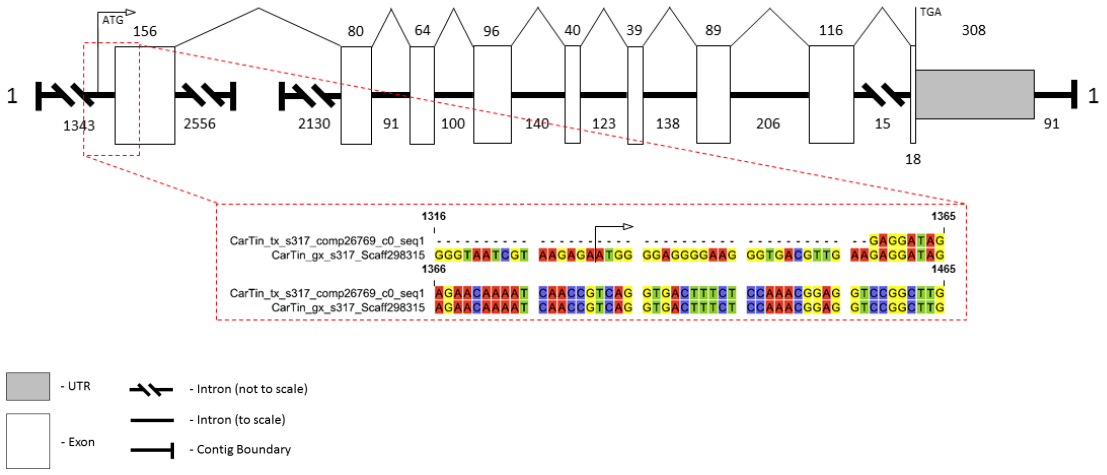


FIGURE 4.4: The gene model for *CtAP1-LIKE* created by the alignment of transcriptomic contig CarTin_tx_s317_comp26769_c0_seq1 to the draft the genomic contigs CarTin_gx_s317_Scaff075165 and CarTin_gx_s317_Scaff298315 (reverse orientation). The gene structure shows 9 exons and a 3′UTR. The inset shows a start codon that is present 26 nucleotides upstream of the first exon on the transcript. These 26 nucleotides on the genomic contig and the first nucleotide of the transcript translate to MGRGRVTLK which is present in the *Chrysanthemum* AP1, LFY and MADS_CDM8 amino acid sequences (Fig. E.2).

4.3.2.2 *FLOWERING LOCUS T-LIKE* (*CtFT-LIKE*)

CarTin_tx_s317_comp32761_c0_seq1, annotated as *CtFT-LIKE*, aligned to a single genomic contig (CarTin_gx_s317_Scaff108510), in the reverse orientation, between contig positions 4,648 bp and 908 bp (Fig. 4.5). The alignment split the transcript into four exons, with a single large intron of 2,116 bp between exon 1 and exon 2. Neither the 5′ or 3′ UTRs were separated by any genomic regions along *CtFT-LIKE* transcriptomic sequence.

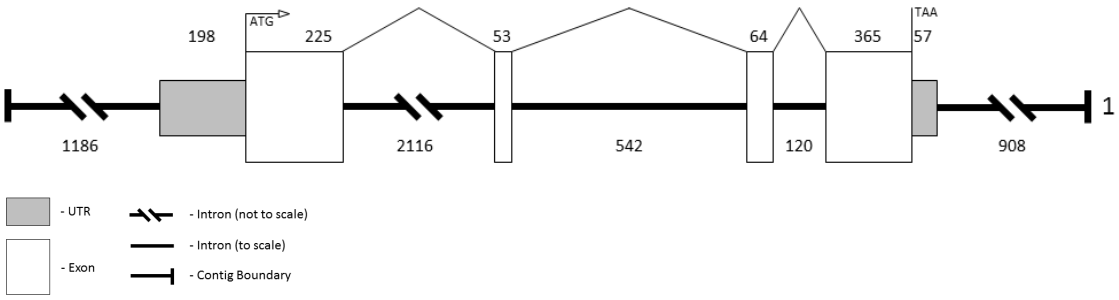


FIGURE 4.5: The gene model for *CtFT-LIKE*, which was produced by the alignment of transcriptomic contig CarTin_tx_s317_comp32761_c0_seq1 aligned to the draft genomic contig CarTin_gx_s317_Scaff108510 (reverse orientation). The gene structure contains 4 exons and a single 5′ and 3′UTR.

4.3.2.3 *MADS BOX DOMAIN CONTAINING 1* (*CtMADS1*)

The transcriptomic contig CarTin_tx_s317_comp33367_c7_seq4, annotated as *CtMADS1* in spring safflower, aligned to three genomic contigs, separating the *CtMADS1* encoding sequence in the transcript into seven exons, two 5′ UTRs and a single 3′ UTR (Fig. 4.6). The 5′ UTR, and the first exon, align onto genomic contig Scaffold_m10540 in the same orientation, from nucleotide position 2,395 and 2,902, with a 15 bp gap separating the 5′ UTR. The second exon aligned onto genomic contig CarTin_gx_s317_Scaff004438 in the same orientation, from nucleotide position 4,305 to 4,377. The remaining five exons and the 3′ UTR align to genomic contig CarTin_gx_s317_Scaff554696 in the reverse orientation, from nucleotide position 1 to 1,747. At the 5′ end of this genomic transcript, the last 36 bp of the 3′ UTR of contig CarTin_tx_s317_comp33367_c7_seq4 did not align to any other genomic contigs with enough confidence to incorporate into the *CtMADS1* gene model. The estimated lengths of the first and second introns are 5,519 bp and 1,627 bp respectively.

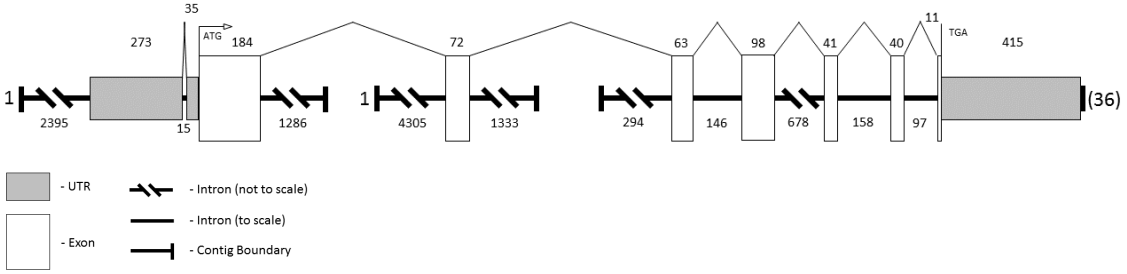


FIGURE 4.6: The gene model for *CtMADS1* was produced by the alignment of the transcriptomic contig CarTin_tx_s317_comp33367_c7_seq4 to the draft genomic contigs scaffold_m10540, CarTin_gx_s317_Scaff004438 and CarTin_gx_s317_Scaff554696. The gene structure contains seven exons, two 5′ UTRs and a 3′UTR. The third genomic contig ends before the end of the transcript, leaving the last 36 bp of the transcriptomic contig unaligned.

4.3.2.4 *VERNALISATION 1-LIKE* (*CtVRN1-LIKE*)

CarTin_tx_s317_comp33519_c0_seq70, annotated as *CtVRN1-LIKE*, aligned to three genomic contigs and split the transcript into seven exons and divided the 3′ UTR into two sequences (Fig. 4.7). The 5′ UTR and first exon align to the genomic contig CarTin_gx_s317_Scaff220104, in the reverse orientation, between nucleotide positions 2,075 and 1,667. The second, third, fourth and fifth exons align against genomic contig CarTin_gx_s317_Scaff206517, in the reverse orientation, between nucleotide positions 1,893 and 272. The last two exons and 3′ UTR align against the genomic contig CarTin_gx_s317_Scaff096311, in the reverse orientation, between nucleotide positions 1,367 and 724, with a 15 bp gap dividing the 3′ UTR. The estimated lengths of the first and fifth introns are at least 3,788 bp and 1,263 bp respectively.

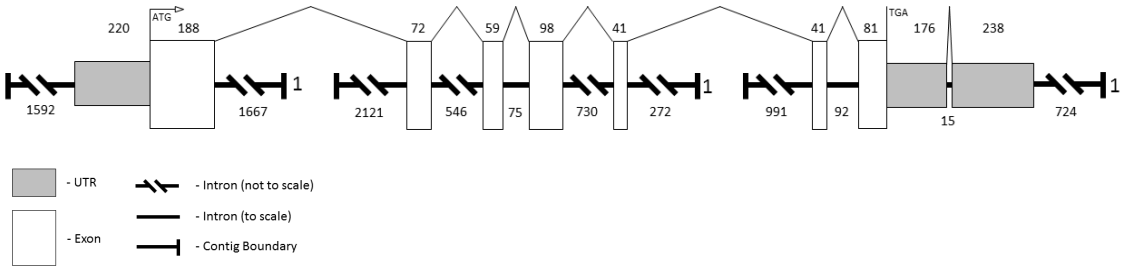


FIGURE 4.7: The gene model for *CtVRN1*, which was produced by the alignment of transcriptomic contig CarTin_tx_s317_comp33519_c0_seq70 aligned to the draft genomic contigs CarTin_gx_s317_Scaff220104, CarTin_gx_s317_Scaff206517 and CarTin_gx_s317_Scaff096311, all in the reverse orientation. The gene structure consists of seven exons, a 5′ UTR and two 3′UTRs.

4.3.3 The PacBio *De Novo* Assemblies

Both PacBio libraries had similar total sequence sizes and dimensions (Table 4.4), at approximately 71 Gbp and 60 Gbp coverage respectively. The read distributions for Library 1 and Library 2 (Fig. 4.8) were also a similar shape, although there was a larger peak of reads at approximately 10,000 in PacBio Library 1. Because PacBio Library 1 was believed to be contaminated with chloroplastic reads, a random PacBio read from Library 2 was aligned to the CSIRO draft safflower genome (Fig. 4.9). Based on this, the error rate was calculated to be approximately 13% in the PacBio libraries. While this is much lower than the estimated 18% error rate for PacBio read libraries (Ferrarini et al. 2013), error correction was still a crucial part of the PacBio assembly process.

TABLE 4.4: Dimensions of the PacBio Genomic Libraries.

	Library 1 *	Library 2 **
Total Size (bp)	**70,883,299,339**	**60,105,680,196**
Reads	8,338,235	7,132,753
Min Length (bp)	50	50
n50 (bp)	11,010	11,171
Mean Length (bp)	8,500	8,426
Max Length (bp)	54,622	53,479

* Highly contaminated with chloroplastic genomic material
** Primarily nuclear genomic DNA

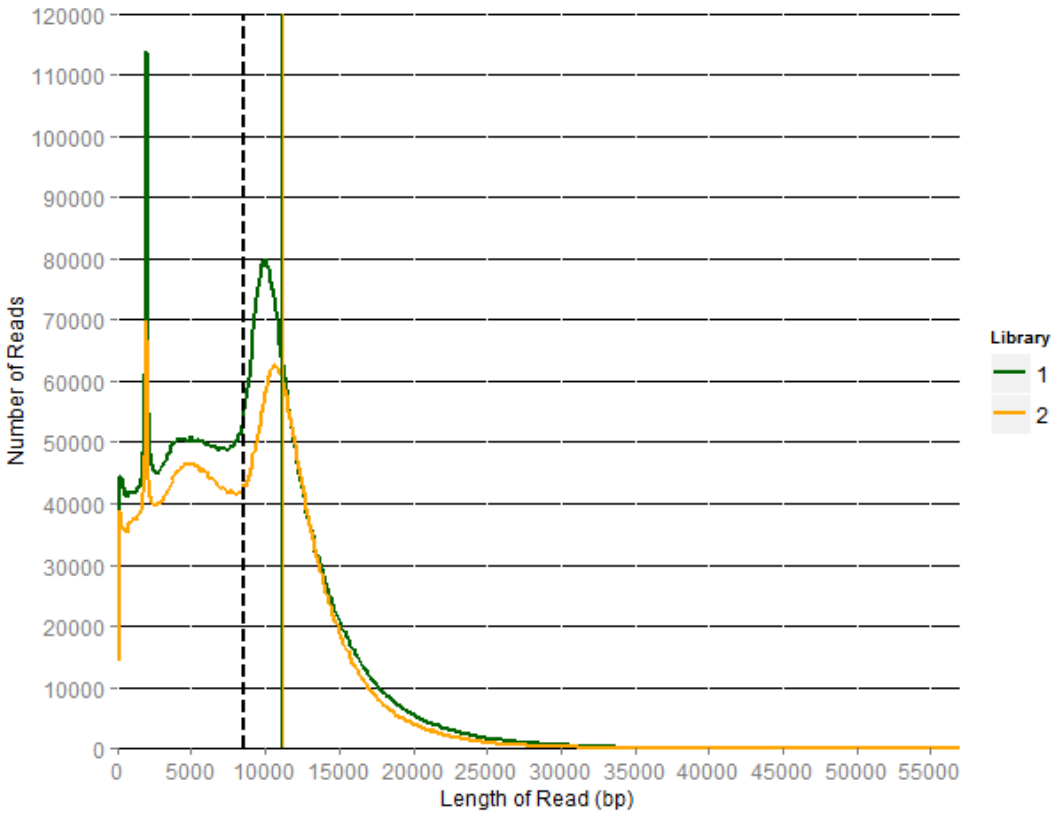


FIGURE 4.8: The distribution of the read length in PacBio Libraries 1 and 2. The mean is indicated with a dashed, vertical black line (as they are indistinguishable at this scale), the n50 is indicated with solid vertical lines.

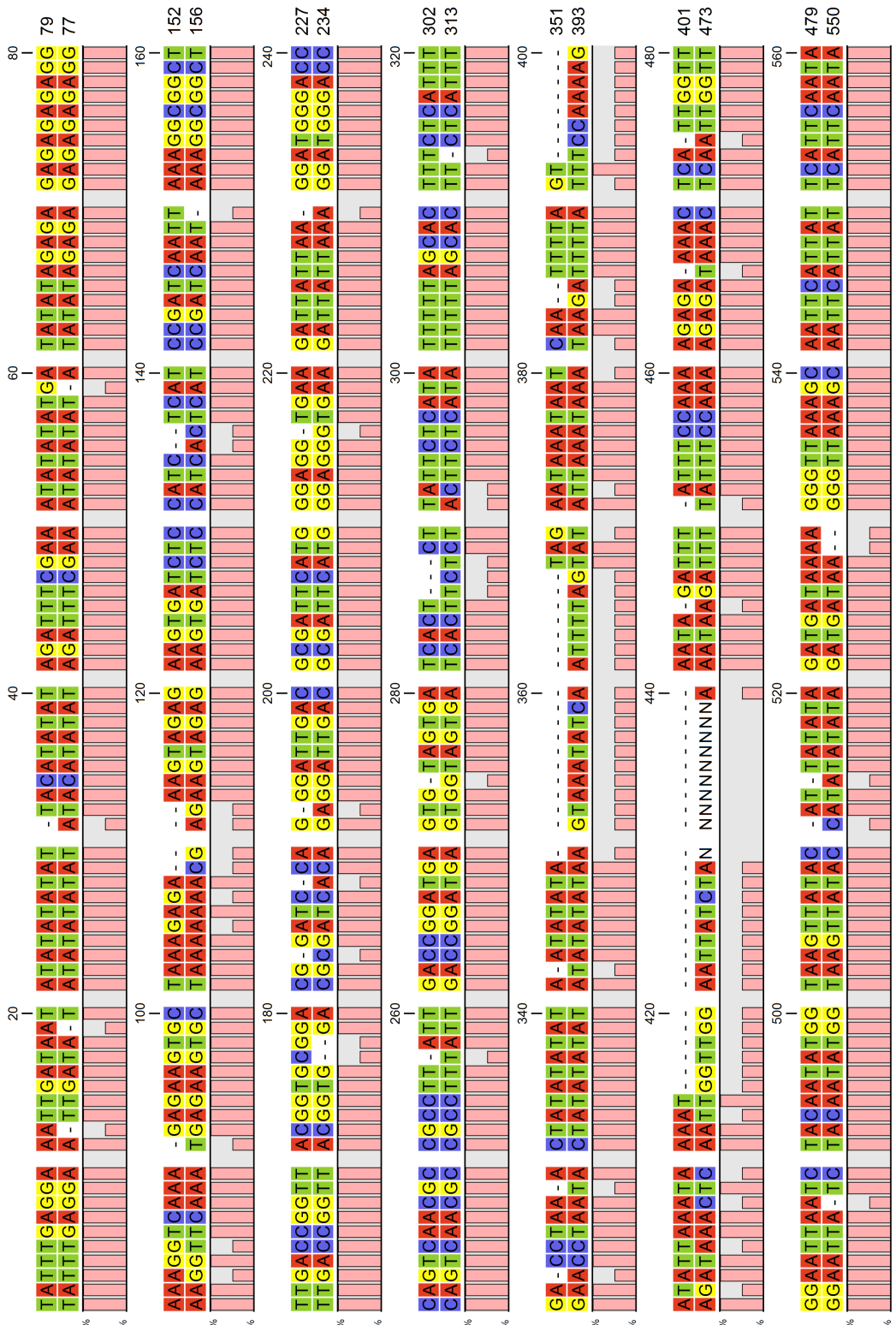


FIGURE 4.9: The first 560 bp alignment between a raw PacBio read from Library 2 and genomic contig CarTin_gx_s317_Scaff077467 from the CSIRO draft safflower genome. The conservation line (pink) indicates where there is either a mismatched nucleotide or a gap in the alignment. The error rate for this PacBio read is approximately 13%.

4.3.4 Library 1: A Draft Safflower Chloroplast

After error correcting reads from the first PacBio genomic library, the assembly produced by PacBiokanga resulted in only 45 contigs, with a final assembly size of just over 2.2 Mbp (Table 4.5). When the largest of these contigs, approximately 1.5 Mbp, was aligned to the NCBI nucleotide database, every alignment on this *de novo* contig was to a chloroplast sequence from another previously characterised plant species. This indicated that there was contamination of this library with a substantial quantity of genetic material from the draft safflower chloroplast.

TABLE 4.5: Attributes of the PacBio Library 1 assembly using PacBiokanga.

	PacBiokanga
Total Size (bp)	**2,202,704**
Contigs	45
Min Length (bp)	10,032
n50 (bp)	20,636
Mean Length (bp)	48,948
Max Length (bp)	1,489,929*

* This contig is referred to as the draft safflower chloroplast genome

Despite contamination, the chloroplast assembly produced from Library 1 was further analysed. When this assembly was compared to the chloroplast genomes of other plant species, the PacBio chloroplast assembly was 1.5 Mbp, approximately ten times the size of a previously assembled safflower chloroplast (Lu et al. 2016) and contrasts to chloroplasts from other species, which are approximately 154 kbp (Sato et al. 1999). When aligned to other chloroplasts, there was no centralised alignment. Instead there were a large number of smaller alignments scattered across the entire length of the *de novo* safflower chloroplast genome (Fig. 4.10).

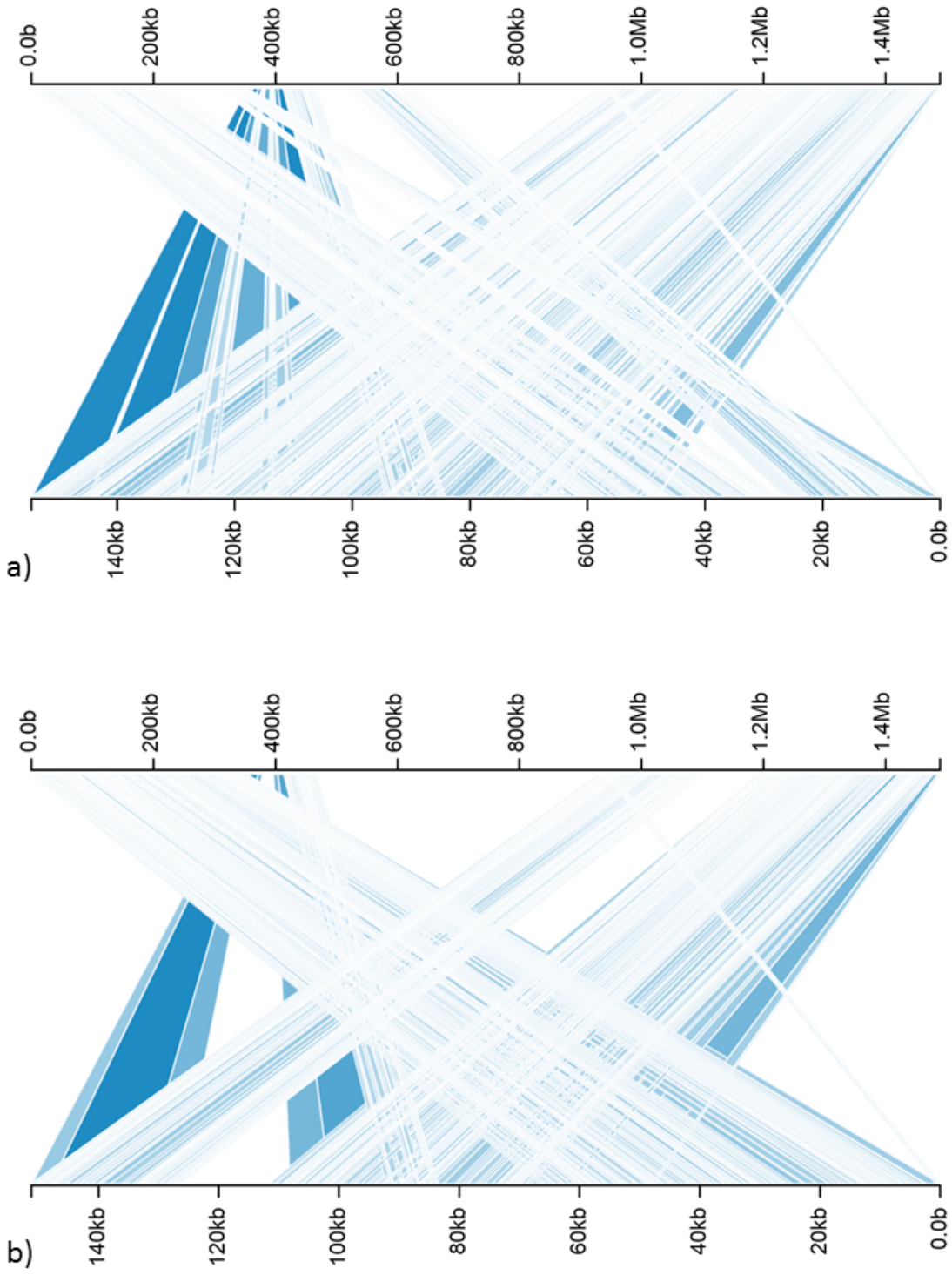


FIGURE 4.10: A visualisation of the alignment of the *de novo* safflower chloroplast and the chloroplast found in *Arabidopsis* (Panel a) and the chloroplast found in sunflower (*Helianthus annus*, Panel b). The blue polygons represents alignments from the *de novo* safflower chloroplast sequence against the chloroplast reference in *Arabidopsis* and sunflower respectively. The shade of blue on the polygon represent the length and score of the alignment. The deeper the blue, the longer and higher scoring the alignment.

To see if there was a misassembly and incorrect repetition of regions of the draft safflower chloroplast, the Hamming Distances were calculated across the length of the contig to determine how unique it was in comparison to other regions (Figs. 4.11 and 4.12). Across the entire length of the safflower PacBio chloroplast genome, there were a large number of unique regions, with only a small region between approximately 355,001 to 430,000 bp containing, in comparison, a substantially smaller quantity of unique regions (Fig. 4.11a). On closer inspection of a region containing stretches with a high Hamming Distance of 275,001 to 355,000 bp (Fig. 4.11b), the calculated Hamming Distances did not always have Illumina reads uniquely aligning at these locations, especially for the smaller stretches. Focusing on a single location on the draft safflower chloroplast between 350,001 and 355,000 bp (Fig. 4.11c), there were Illumina reads present that map over a region of approximately 1,700 bp, but not across the two shorter unique regions located between approximately 353,000 and 353,600 bp. Examining this specific region of high Hamming Distances with three of the seven aligned Illumina read libraries (Fig. 4.12), these reads span across almost all of the unique region, with all uniquely aligning at a very high coverage depth at two specific locations.

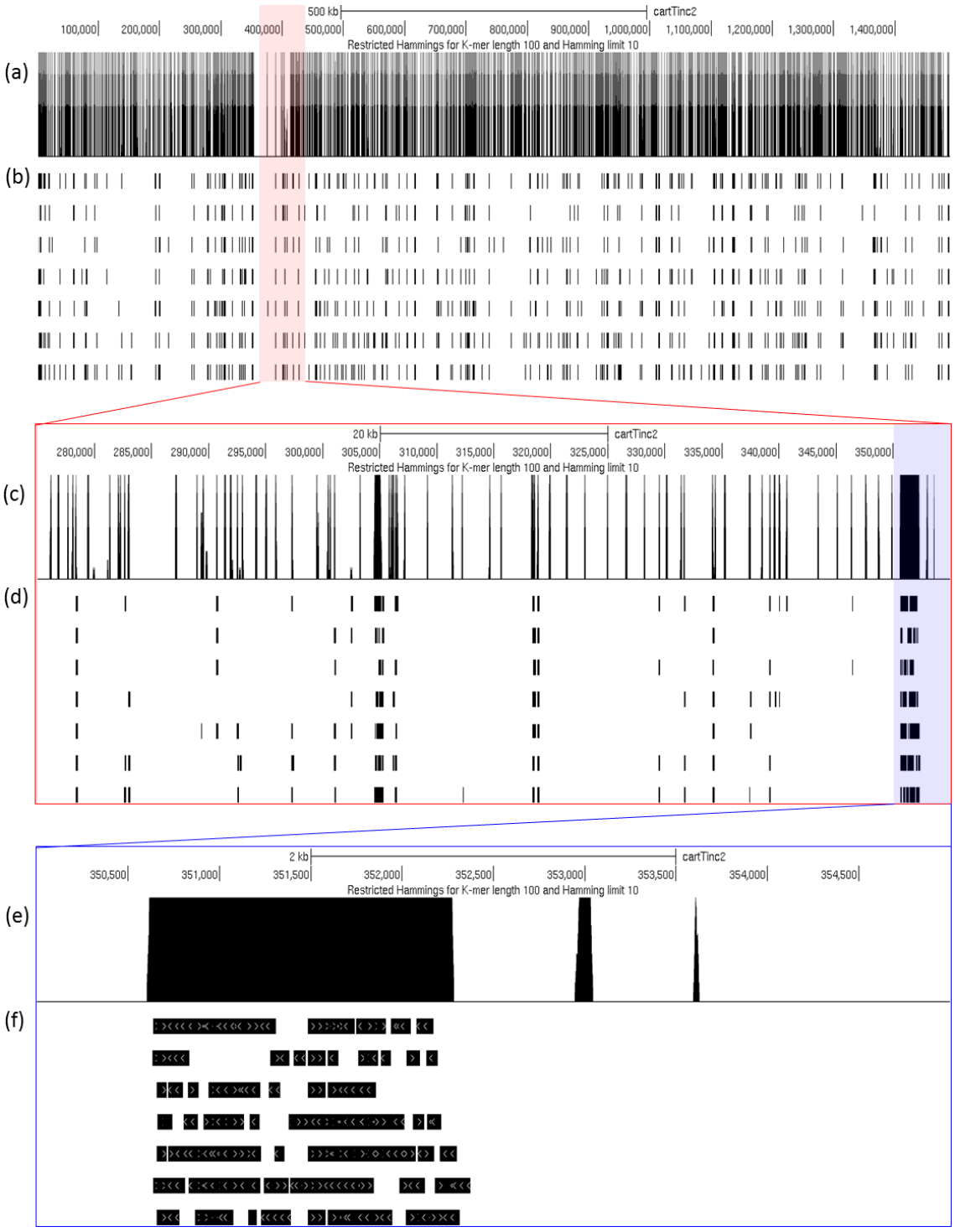


FIGURE 4.11: A visualisation of the assembled genome of the draft safflower chloroplast at three resolutions. Panels (a), (c) and (e) show the Hamming Distances (between zero and > 10) across the assembly. Panels (b), (d) and (f) show the alignment of the seven different 100 bp Illumina read libraries against the assembly. Panels (a) and (b) show the Hamming Distances and short read alignments across the entire assembly, Panels (c) and (d) show the Hamming Distances and short read alignments between 275,001 bp and 355,000 bp, and Panels (e) and (f) show the Hamming Distances and short read alignments between 350,001 and 355,000 bp. Arrows show the directionality of the alignments).

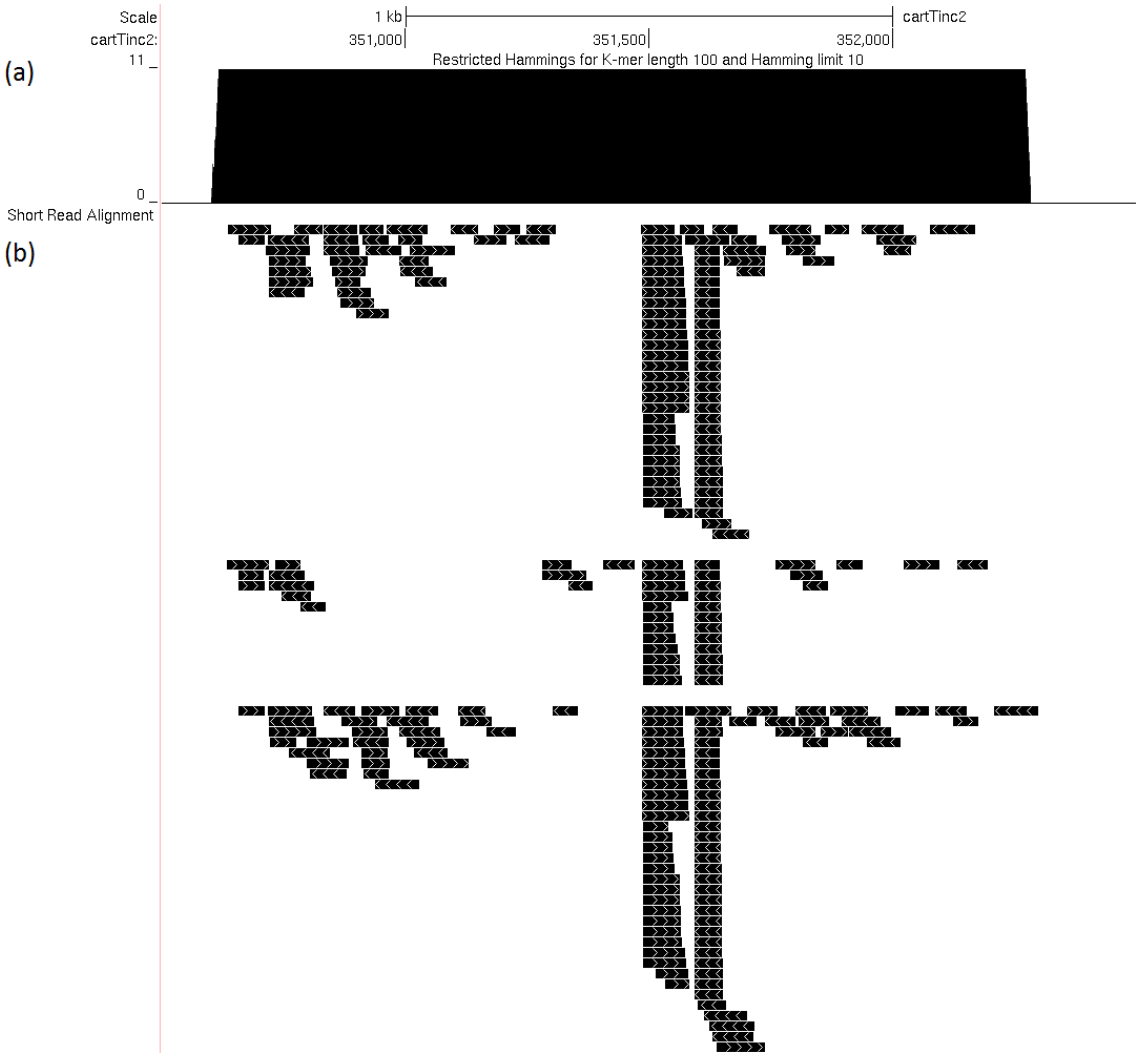


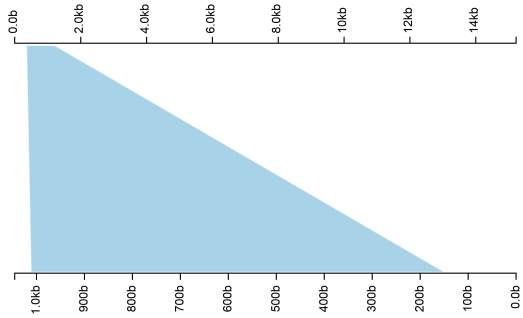
FIGURE 4.12: A high resolution image of the draft safflower chloroplast. Panel (a) shows 350,501 and 352,500 bp, the Hamming Distances, zero to 10+. Panel (b) shows the alignment of short reads from three of the seven Illumina genomic read libraries across a unique region, indicated by the large Hamming Distances. Arrows on the short reads indicate the read orientation. Pair information not shown.

4.3.5 Library 2: A Work in Progress

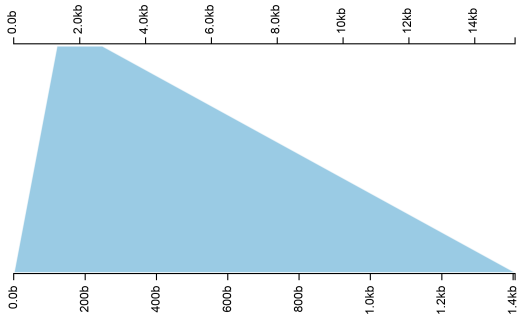
At the time of authoring this Thesis, only a small portion of the reads from PacBio Library 2 had been error corrected (Table 4.6). Nine error corrected PacBio reads from the PacBio genomic library were sampled and aligned to the Illumina safflower genome using BLASTN. A large number of high scoring alignments were reported, with the top six showing multiple different Illumina contigs aligning, with only minimal gaps, in unique and sequential locations along the length of the error corrected PacBio read (Fig. 4.13). The other randomly sampled error corrected reads had a large number of Illumina genomic contigs aligning to them. But few aligned the same Illumina genomic contig to multiple locations on the error corrected PacBio read or had different Illumina contigs reporting multiple high scoring alignments to the same location.

TABLE 4.6: Attributes of the partially error corrected PacBio Library 2, using PacBiokanga for error correction. At the time of analysis, error corrected reads were approximately 4.5% of the total library size.

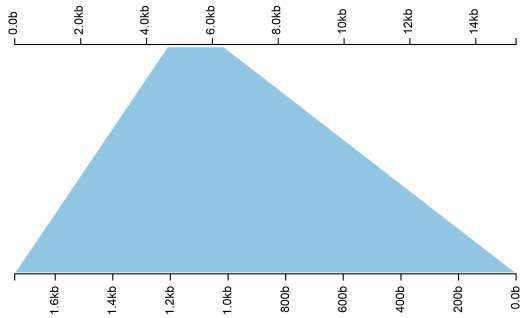
	Error Correction
Total Size (bp)	**2,666,153,537**
Contigs	200974
Min Length (bp)	7,500
n50 (bp)	14,867
Mean Length (bp)	13,266
Max Length (bp)	32,917



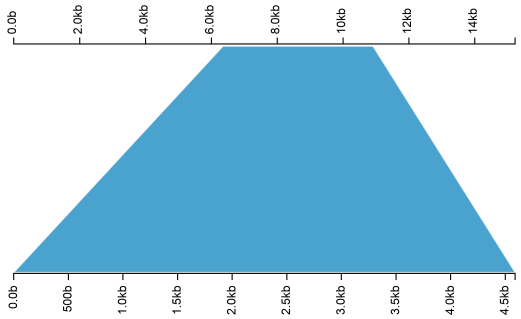
(a) Error corrected PacBio read ecseq3284803_1; (top) aligned to Illumina genomic contig CarTin_gx_s317_Scaff044478 (bottom)



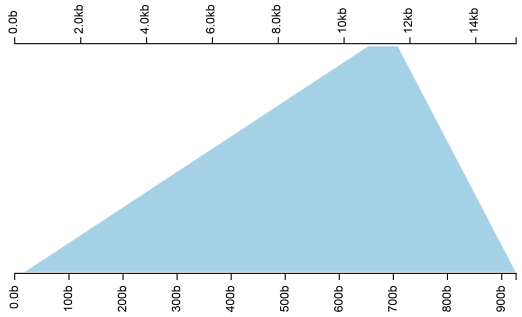
(b) Error corrected PacBio read ecseq3284803_1; (top) aligned to Illumina genomic contig CarTin_gx_s317_Scaff303867 (bottom)



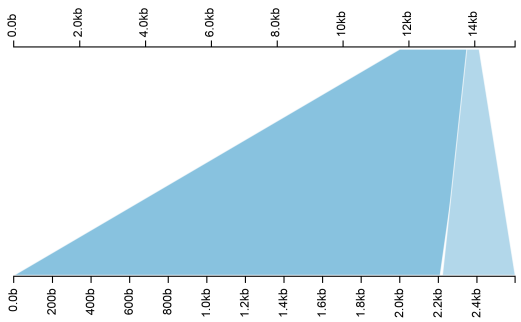
(c) Error corrected PacBio read ecseq3284803_1; (top) aligned to Illumina genomic contig CarTin_gx_s317_Scaff561682 (bottom)



(d) Error corrected PacBio read ecseq3284803_1; (top) aligned to Illumina genomic contig CarTin_gx_s317_Scaff382322 (bottom)



(e) Error corrected PacBio read ecseq3284803_1; (top) aligned to Illumina genomic contig CarTin_gx_s317_Scaff581821 (bottom)



(f) Error corrected PacBio read ecseq3284803_1; (top) aligned to Illumina genomic contig CarTin_gx_s317_Scaff095276 (bottom)

FIGURE 4.13: A randomly selected error corrected PacBio read ecseq3284803_1; 15,231 bp) aligned to the Illumina safflower *de novo* genome. The shade of blue represents the length of the alignment, with a deeper shade of blue representing a longer alignment.

4.3.6 DNA Based Markers of Vernalisation in Safflower

Of the F_3 families that were examined, 28 had a spring:winter segregation ratio of 11:5, 3:1 or 13:03. Due to the large number of F_3 seeds available from the crossing population, seven of the F_3 families, X017, X030, X100, X142, X181, X246 and X395, were grown again to confirm the 3:1 segregation ratio of spring:winter phenotypes. This time, 100 plants grown in controlled growth cabinet conditions for four weeks. F_3 families X017, X030, X100 and X395, which segregated in the 3:1 ratio, were investigated further for the presence of genetic markers. The segregation ratios for other F_3 families is listed in Appendix H (Table H.1).

A list of SNPs and digest markers were created for all four of the F_3 crossing families, using late elongation behaviour of the winter phenotype as a proxy for the vernalisation response. The analysis (Table 4.7) reported that there were 2,849 digest fragments and 4,047 SNP markers in the samples from X395 (Batch 1). The samples from X017, X030 and X100 (Batch 2) were reported to have 4,491 digest fragments and 2,763 SNPs. After calculating the segregation scores for the early and late elongation phenotypes, Batch 1 reported 67 digest fragments and 83 SNPs and Batch 2 reported 93 digest fragments and 81 SNPs. Across all four F_3 families, some digest fragments and SNPs showed a presence in the early elongation phenotypes, i.e. spring, and an absence in the late elongation phenotype, i.e. winter. When the SNPs and digest fragments from these families were compared to each other, 60 digest fragments and three SNPs were determined to be common across all four F_3 families.

TABLE 4.7: Digest and SNP markers reported by DArT, correlating the elongation phenotype in the F_3 crossing family X395 (Batch 1) with X017, X030 and X100 (Batch 2), and those digest fragments and SNPs that determined to be common across all four F_3 families.

	digest marker		SNP marker	
	X395	X017 X030 X100	X395	X017 X030 X100
Total	2,849	4,491	4,047	2,763
Vernalisation	67	93	83	81
Common	60		3	

Further investigation of the SNP markers showed that for two of them, the genomic sequence fragments containing the SNPs, sequenced by DArT, did not align to any contig found in the CSIRO draft safflower genome. The third SNP marker aligned to a high number of different contigs in the safflower draft genome, but these alignments were short and not high scoring. For this reason, the three SNP fragments common across the four crossing families were not investigated further.

Fifty-three of the digest fragments aligned against the CSIRO draft safflower genome, with three digest fragments (15670156, 15674427 and 15670077) aligning multiple times to different contigs (Appendix I, Table I.1). The remaining SNPs returned single high scoring alignments against 46 different CSIRO draft safflower genome contigs. When these 46 contigs were aligned to the Bowers genetic map, 45 were reported to have a high scoring alignment to a Bowers genomic contig. 28 of these matched a Bowers contig that contained SNPs (Appendix I, Table I.1), with 27 aligning to chromosome 8 (Fig. 4.14) and one aligning to chromosome 4.

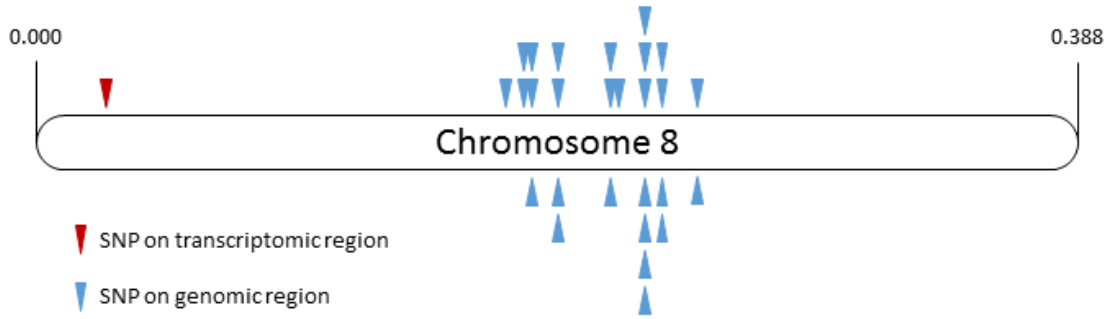


FIGURE 4.14: The genetic map of Chromosome 8 of safflower, ranging from 0.000 to 0.388 cM, showing the number of DArT markers from the CSIRO draft safflower genome that can be placed onto the Bowers genetic map (Bowers et al. 2016).

4.3.6.1 Aligning Differentially Expressed Transcripts onto the Genetic Map

Transcripts that were identified as very significantly and significantly differentially expressed from Experiments 1 and Experiment 2, respectively (Chapter 3) were aligned to the Bowers safflower genome to determine if they could be mapped to specific chromosomes. Of Experiment 1's differentially expressed transcripts, 28 of the 30 contigs found a high scoring alignment from the Bowers genomic contigs (Appendix I, Table I.2). Of these contigs, the five located on Chromosomes 1, 5, 7, 9 and 12 were listed as containing SNPs. The transcript for *CtMADS1* maps to Chromosome 1 and *CtAP1-LIKE* maps to Chromosome 9, while *CtFT-LIKE* and *CtVRN1-LIKE* could not be mapped to the Bowers genetic map at all. Of the differentially expressed transcripts from Experiment 2, all but one returned a high scoring alignment to a Bowers contig. Of the 31 differentially expressed transcripts mapped to Bowers contigs that were listed as

containing SNPs, 14 were excluded, as they did not have an expression profile in the winter time course that changed as the vernalisation treatment period increased (Chapter 3.3.4). Of the remaining 17 transcriptomic contigs, one mapped to Chromosome 4 and one to Chromosome 8 (Fig. 4.14). The 44 other Bower contigs that aligned to the remaining differentially expressed Experiment 2 transcripts were not listed as containing SNPs.

4.4 Discussion

4.4.1 The *De Novo* Assemblies

In the case of the safflower genome constructed with Illumina sequences, the deduplication and overlap filtering processes reduce the amount of computational processing needed for assembly by removing the need to process redundant sequences. Previously, it had been shown that the safflower genome is a diploid, with an estimated haploid genome size of between 1.3 and 1.4 Gbp (Garnatje et al. 2006). After scaffolding with the MP reads and processing with SCUBAT, using the alignment information between the Illumina genome and transcriptome, the result was a haploid draft genome of approximately 1.1 Gbp across just over 900,000 contigs with an n50 of just under 2 kbp representing approximately 80% of the safflower genome. The CEGMA and BUSCO completeness scores for the draft safflower genome was approximately 65% for both, which is substantially lower than completeness scores found in other heavily annotated organisms analysed with BUSCO, such as *Drosophila melanogaster* (98%) and *Caenorhabditis elegans* (85%; Simão et al. 2015). This decreased BUSCO score could be due to the larger genome size of safflower and the fact that the BUSCO reference data set for plants was added much later than the other reference sets. Despite this, it was decided to use this as a reference genome sequence for safflower, as the alignment of *de novo* transcriptomic contigs showed the presence of introns both within and between contigs (section 4.4.2) gave confidence in the utility of the draft genome sequence as a reference.

The draft safflower genome presented in Bowers et al. (2016) is comprised of 2,195,958 contigs covering approximately 65% of the 1.4 Gbp safflower genome, an n50 of 1,976 bp and a mean length of 402 bp with the largest contig being 55,679 bp. When comparing the CSIRO draft safflower genome to the Bowers draft genome (Table 4.8) the only metrics that the Bowers draft genome performed better were the lengths of the longest contigs and the n50. The CSIRO draft safflower genome, while containing less than half the number of contigs than the Bowers genome, the length of these contigs are much longer and, when combined, result in a longer overall assembly. Based on the above metrics, the CSIRO draft safflower genome assembled using Biokanga and scaffolded with SCUBAT is a far better reference than the Bowers genome.

TABLE 4.8: The CSIRO draft safflower genome constructed using Biokanga compared against the draft safflower genome presented in Bowers et al. (2016).

	CSIRO	Bowers
Total Size (bp)	**1,163,499,791**	882,813,871
Contigs	**904,199**	2,195,958
Min Length (bp)	**300**	100
n50 (bp)	**1,940**	1,976
Mean Length (bp)	**1,286**	402
Max Length (bp)	**32,974**	55,679

Confidence in the quality of the CSIRO draft safflower genome was provided by analysis with using CEGMA and BUSCO. Both of these independently developed algorithms reported that approximately 65% of the conserved sequences in each assessment tool (248 sequences in CEGMA, 956 sequences in BUSCO) were found in their entirety in the *de novo* safflower genome. Further, when contigs from the *de novo* transcriptome (Chapter 3) were aligned to the Illumina *de novo* genome, 144,931 out of 146,780 (98.7%) of the assembled transcriptomic contigs were aligned to the *de novo* genome. Because the spring safflower transcriptome was assembled using 'Trinity' (Grabherr et al. 2011) and the spring safflower genome was built with Biokanga, the high similarity between the sequences is not the result of an artefact created by the use of a single assembly algorithm. The usefulness of this assembly has already been demonstrated in other research e.g. to locate and characterise the number of transgenic events in a number of transgenic safflower lines (Wood et al. 2016, in preparation).

It was observed that there are non-trivial differences in the number of Illumina PE reads that align to the CSIRO draft safflower genome (approximately 1.4x and 65x coverage respectively) when the fragment length used to back align the unfiltered reads is changed from a fixed length of 180 bp, the PE library fragment length reported by AGRF, and the fragment length used in the scaffolding step of the assembly, to a range of fragment lengths between 100 and 500 bp. While the sequencing report from AGRF stated the insert size of the paired end fragments were 180 bp, the actual insert size of paired end reads was actually the average length of the insert size, which was approximately between 100 and 500 bp. In this case, two different back alignment patterns were seen when the paired end reads are mapped back to the CSIRO draft safflower genome. While it has not been detrimental to the assembly of the CSIRO draft safflower genome, it has meant that, while accurate, the final scaffolded genome was fragmented across nearly one million contigs. If this assembly were to be revisited in the future, allowing a range of fragment lengths, such as a single standard deviation above and below the mean insert length, as a parameter in the scaffolding stages rather than an inflexible fragment length should decrease fragmentation of the assembly even further by allowing contigs to scaffold, thereby improving the final assembly.

4.4.2 Intron/Exon Boundaries for Genes Annotated in the Vernalisation Response

In the four characterised gene models, created by aligning the four characterised safflower transcriptomic contigs to the CSIRO draft safflower genome, the high scoring alignments showed that all of the transcripts contained an intronic structure commonly seen among other better studied plant systems. When examining the intron/exon structure of *CtFT-LIKE*, there is an intron/exon structure similar to that seen in *AtFT*. The Arabidopsis Information Resource (TAIR10; `www.arabidopsis.org/servlets/TairObject?id=30541&type=locus`) shows that *AtFT* contains four exons, with the second and third exons relatively small in comparison to the first and fourth, and that the first two introns are substantially larger than the third. A similar intron/exon structure was seen in *CtFT-LIKE*. Investigation of these intronic regions may reveal regulatory regions that control the expression of *CtFT-LIKE* during vernalisation, but resolving these intronic regions must happen before this can occur.

Similarly, *CtVRN1-LIKE* was shown to have an intron/exon structure similar to that seen in *HvVRN1* (Trevaskis 2010). While *HvVRN1* has eight exons, with a very large first intron in comparison to the others, *CtVRN1-LIKE* only contains seven exons. The 3′ flanking region of exon 1 and the 5′ flanking region of exon 2 add up to a length of almost 3,800 bp. As this region is unresolved, the total length is most probably even larger. If this large intronic region were to be investigated in *CtVRN1-LIKE*, it should reveal sequences associated with its regulation during vernalisation.

There were no published exon structures available for *CtAP1-LIKE* or *CtMADS1*, nor were there any clear transcript homologues found in *Arabidopsis*. While there were three *Chrysanthemum* sequences that had a high scoring sequence homology to *CtAP1-LIKE*, there were no annotated exon structures associated with these genes. The high scoring alignments of these transcripts against the safflower Illumina genome and the similarities in structure seen between *CtFT-LIKE* and *CtVRN1-LIKE* and their homologues in *Arabidopsis* and barley, respectively, builds confidence in the accuracy of the gene structures presented for *CtAP1-LIKE* and *CtMADS1*, despite not having annotations to compare them to. Similar to *CtFT-LIKE* and *CtVRN1-LIKE*, investigations into the large intronic regions of *CtAP1-LIKE* and *CtMADS1* should reveal regulatory regions.

While four different annotated safflower genes, *CtAP1-LIKE*, *CtMADS1*, *CtFT-LIKE* and *CtVRN1-LIKE*, were analysed in detail the very high number of transcriptomic contigs that align to the CSIRO draft genome means that as more safflower transcripts are annotated, the creation of gene models for these newly annotated transcripts should quickly follow. This intronic information is also an invaluable resource for investigating the regulatory mechanisms of genes. In *AtFLC*, the PHD-PRC2 complex binds to the

first intronic region at the 5′ end of *AtFLC* to regulate its expression (Sheldon et al. 2002). Without the non-protein coding information, this regulatory mechanism could not have been identified as a regulator of *AtFLC* expression. As vernalisation and other desirable safflower traits are experimentally characterised, these intronic regions will become critical to understanding the regulatory mechanisms controlling the expression of these biologically essential genes and, ultimately, the phenotypes that result from this regulation.

In the future, by combining the Illumina and PacBio assemblies together into a single assembly, examination of the non-protein coding regions of the genome, including upstream and downstream UTRs and intronic regions that are proximal and within transcript sequences of interest, will assist in further annotating the safflower genome.

4.4.3 Genetic Markers of the Vernalisation Response in Safflower

There were 27 different digest markers for the vernalisation response identified in this analysis. If each of these fragments were further refined and tested, and their absence confirmed in winter safflower, they will become invaluable for the identifying safflower crosses and cultivars that are responsive to vernalisation. These 27 markers would be immensely useful to help guide the breeding of safflower lines by enabling the detection of vernalisation traits more quickly than using the current visual phenotyping methods.

However, based on the results of Section 4.3.6, the high concentration of markers on Chromosome 8 indicates that this regions plays an important role in the regulation of early flowering and the vernalisation response in safflower, but this would need to be confirmed by comparing it to a safflower cultivar with a different genetic background. This was shown by the 27 digest markers mapping to the Bowers contigs at Chromosome 8. There was a single transcriptomic contig, CarTin_tx_s317_comp1208157_c0_seq1, that mapped to Chromosome 8, though no homologues for this transcript were found within NCBI. Further investigation into the function of this transcript will be crucial for determining if this is, in fact, a trigger for the vernalisation response in safflower.

There was also a single transcript identified in the differential expression analysis in Experiment 2, CarTin_tx_s317_comp31946_c0_seq1, that mapped to Chromosome 4 in the Bowers genetic map. This transcript, while sharing a significant amount of sequence homology with other sequences within NCBI, could not be functionally annotated using this information, as the translated product shared sequence homology with a number of 'uncharacterised proteins'. That it could not be functionally annotated and being the only genetic marker found on Chromosome 4, this transcript was not considered a causal factor of the vernalisation response in safflower.

The first genetic map for safflower, the Bowers genetic map, is limited in that it only covers approximately 15% of the total estimated size of the safflower genome. As a result, transcripts and digest markers not identified in the Bowers genetic map could not be eliminated from the pool of those potentially involved with the vernalisation response in safflower. In contrast, any differentially expressed transcripts that were not mapped to Chromosome 8 but to another Chromosome were able to be eliminated as candidate molecular components of the vernalisation response in winter safflower. *CtMADS1* and *CtAP1-LIKE* were both mapped to Chromosomes 1 and 9 respectively, which eliminates them for consideration as triggers of the vernalisation response in safflower. But *CtFT-LIKE* and *CtVRN1-LIKE* could not be mapped to a chromosome, meaning they must be located beyond the 15% of the safflower genome mapped by Bowers et al. (2016). As the genetic map is extended into these unmapped regions, further investigations could reveal whether *CtFT-LIKE* and *CtVRN1-LIKE* are proximal to this cluster of SNPs in safflower, confirming or refuting their role as triggers of the vernalisation response. The remaining differentially expressed transcripts that are unmapped could still be potential triggers for the vernalisation response in winter safflower, but further testing is needed to confirm or refute their involvement.

One limitation with regard to the generation of these DNA fragments was using the F_3 generation of the crossing population. As mentioned above (see Chapter 2.3.5), due to limited glasshouse space, the F_2 generation of the crossing population was grown in the field, of which an unintended consequence was that members of the crossing population were exposed to vernalisation conditions. This made phenotyping the F_2 population for vernalisation response impossible. In the interests of time and space constraints with the project, the F_3 population was used to identify segregants and generate DNA fragments via DArT. Further investigation of these potential markers in a Recombinant Inbred Line or in a population back crossed with the spring cultivar would allow fine tuning of the population and allow markers that were unrelated to the vernalisation response to be filtered out. While these 27 potential markers are found in four F_3 families, until these potential markers are tested in another genetic background, they remain potential candidates only.

4.4.4 The Curious Case of the Safflower Chloroplast

The published size of the safflower chloroplast assembly is 153,675 bp (Lu et al. 2016), which is similar in size to a vast number of chloroplast genomes found in other species, including *Arabidopsis* (154,478 bp; Sato et al. 1999). After assembly of the first library of PacBio reads using PacBiokanga, the largest contig was approximately 1.5 Mbp, around ten times the size of the published safflower chloroplast genome. Despite this, BLASTN reported significant alignments between the draft safflower chloroplast and the assembled chloroplasts of other species. Using Hamming Distances, the number of changes required for a specified string or sequence to match another, the uniqueness of a sequence of nucleotides can be calculated. The greater the Hamming Distance of a

sequence, the greater the number of changes required for that sequence to match another, and therefore, the more unique that sequence is. In this case, large Hamming Distances indicate that the safflower Illumina genomic reads have mapped at unique regions in the chloroplast genome (Figs. 4.11 and 4.12).

This large chloroplastic contig could be attributed to mistakes in the error correction stage of the PacBiokanga algorithm, resulting in a misassembly. However, the locations of high scoring Hamming Distances scattered frequently throughout the entire chloroplast assembly, combined with the Illumina data uniquely mapping to these locations, implies that the assembly of PacBio Library 1 is correct. Previous literature also suggests the presence of a multinodal structure in the chloroplast (Deng et al. 1989; Bendich 2004). Molecular characterisation, either by the use of PCR primers that bind to similar but unique regions, combined with other methods such as pulse field gel electrophoresis (Oldenburg and Bendich 2004; Oldenburg and Bendich 2016), would give confidence in the accuracy of the multinodal *de novo* assembly of the safflower chloroplast.

The chloroplast is the energy centre of the plant cell and is responsible for the conversion of carbon dioxide to oxygen along with the synthesis of sugars and lipids. If the larger size of the chloroplast is confirmed, the implications for our fundamental understanding of biology will impact a broad range of fields, including GM and novel oil synthesis in crop species (as food and animal feed as well as an oleochemical precursor), through to the way that plants and crops will cope with the challenges associated with climate change. This unexpected result, while warranting closer examination, does not relate to the vernalisation response in safflower, and so was not investigated further.

4.4.5 Future Directions

The ultimate goal with any *de novo* assembly, regardless of the technology used, is the creation of a reference sequence that, as accurately as possible, represents the genetic information encoded within the genome. Using a combination of sequencing technologies was shown to be the most effective method of building an accurate assembly (Fig. 4.15). For example, a transcriptomic contig (*de novo* or Sanger sequenced; Fig. 4.15a) aligned to a *de novo* genome created with Illumina reads reveals the intronic structure of the transcript, as well as the orientation of the genomic contigs in relation to that transcript (Fig. 4.15b). Using this combination of sequencing technologies resulted in the creation of an accurate assembly of safflower genomic sequences, as well as an assembly of sequences originating from the safflower chloroplast. These resources will not only help us understand how the vernalisation response works in winter safflower, but prove to be invaluable for investigating other favourable traits in safflower.

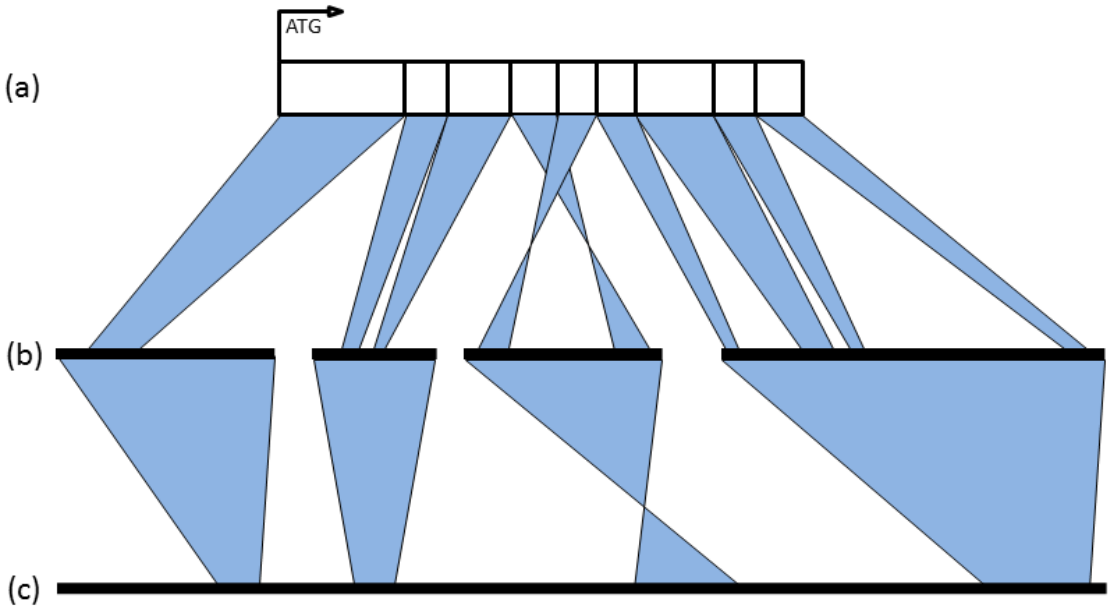


FIGURE 4.15: A hypothetical alignment of different assemblies. The transcriptomic contig (a) aligns across multiple Illumina genomic contigs (b), with orientation information but unresolvable intronic regions between contigs. Using longer length contigs (c) created through assemblies of longer length reads, such as PacBio, or scaffolding Illumina contigs, using Chicago sequencing, will allow these intronic regions to be resolved, as well as annotate upstream or downstream non-protein-coding regions.

The length of the intronic regions between many of the adjacent exons in *CtAP1-LIKE*, *CtMADS1* and *CtVRN1-LIKE*, are unknown, as the transcriptomic regions align to different genomic contigs. The long reads of the PacBio assembly (Fig. 4.15c) can span these unknown regions and resolve the size of these introns, as well as correctly orient the transcript. This has been shown by multiple Illumina genomic contigs aligning to a single error corrected PacBio read (Fig. 4.13). In addition, if the *de novo* PacBio contigs are long enough, upstream promoter, enhancer or silencer regions can also be identified as regulators of the transcript. At the time of authoring this Thesis, the PacBio assembly of the safflower genome was still in progress.

While transforming *CtAP1-LIKE*, *CtMADS1*, *CtFT-LIKE* into *Arabidopsis* or transforming *CtVRN1-LIKE* into barley or *Brachypodium distachyon* may provide some information on the function of the protein products encoded by these transcripts, transforming safflower plants to disrupt these genes, within coding and non-coding regions, is the most effective way of confirming or refuting the role these genes play in the vernalisation response. However, the transformation of safflower is a non-trivial undertaking i.e. a complete project in itself, and was beyond the scope of this thesis.

At the time of authoring this Thesis, a PacBio library was in the process of error correcting reads and the assembly of these reads into a *de novo* genome. The initial results with regard to scaffolding the Illumina genomic library using PacBio data are

encouraging. When completed, this assembly will provide even further scaffolding of the currently fragmented genomic contigs. There are also newer technologies being developed e.g. the Chicago method of linking long range sections of DNA to provide better scaffolding support for *de novo* assemblies (Putnam et al. 2016). Incorporating these differing methods with the high quality *de novo* Illumina assembly will further scaffold the safflower genome.

With regard to the gene models, markers and important locations on Chromosomes as identified by the Bowers SNP-containing contigs, the genomic information presented in this chapter, while being of high quality, is far from complete. The draft genomic assembly in this chapter aligns very well against the transcriptomic information presented previously (Chapter 3) and with the contigs generated as part of the Bowers et al. (2016) study. Despite this, the Illumina genomic contigs have no reference orientation with regard to one another, and while the Bowers SNP contigs provide some information as to where they are located, this only represents approximately 15% of the entire safflower genome. Further scaffolding of the Illumina genome is necessary to identify and annotate not only the intron/exon structure of safflower loci, but to characterise the regulatory sequences, be they proximal or distal, controlling the expression of each locus of interest. Once a longer and higher quality genomic reference sequence is developed, even just scaffolding the existing CSIRO de novo genome, it will become clearer as to where these DNA fragments fall. Further, testing these markers in other wild vernalisation sensitive safflower varieties (especially if they can be sourced from different growing regions) will allow us to guage whether these DNA fragments can be used as markers for a vernalisation response in Safflower.

4.5 Conclusion

Based on the results of the Illumina *de novo* genome assembly alone, a high quality draft genome has been created, covering approximately 80% of the safflower genome. The degree of confidence in this assembly is predicted from the high level of alignment between the genome and transcriptome, generated by using two different algorithms. Despite its fragmented state, this draft genome was successfully used to examine and confirm T-DNA insertion sites in a transgenic safflower, reinforcing confidence that the CSIRO draft safflower genome is suitable for use as a reference in its current state.

The next steps for the safflower genomic reference is to further scaffold contigs and increase their length. The longer scaffolded contigs resulting from an assembly of PacBio data (independently or combined with the Illumina data) will allow better exploration of the genetic map created by Bowers et al. (2016). This improved genome, containing longer contigs, will enable production of a more accurate genetic map and allow more markers to be mapped to the safflower genome. Combined with the

differentially expressed transcripts, makers and transcripts co-located on the same long contigs will identify even more candidates for the molecular basis of the vernalisation response in safflower.

Eventually, a combination of sequencing and a higher coverage genetic map will produce a better reference genome for safflower. This will provide insight into the regulatory mechanisms of safflower and how the expression of transcripts are regulated, whether it be the vernalisation response or some other biological or physiological pathway.

Chapter 5

Overall Discussion

Across the world, almost all cropped safflower are spring cultivars. The only suggestion of any kind of vernalisation response reported has been in a number of wild 'winter hardy' safflower varieties sourced from eastern China (Johnson et al. 2006). In Chapter 2, when these 'winter hardy' varieties were grown at 25°C alongside an elite spring safflower cultivar, they exhibited both a greater quantity of vegetative material and a significant delayed time to flowering. When exposed to vernalisation conditions, this 'winter hardy' safflower behaved differently to unvernalised winter safflower and elongated in a similar manner to spring safflower (Carapetian 2001). The basis of this project was to document and characterise this vernalisation response and examine the genetic basis underlying the observed phenotypes and transcriptomics that underlay it.

In terms of physiology and expressed phenotype, the vernalisation response in safflower appears to be physiologically similar to that seen in other species. In Chapter 1 and as has been reported for a number of other agronomically important crop species, the mechanisms responsible for the vernalisation response are diverse and unique to each family. After a number of subsequent generations and crosses of both winter and spring varieties, it was determined that the vernalisation response in safflower is:

i) only present in winter safflower

ii) a recessive trait

iii) epigenetic in nature

iv) reset after each generation

v) most pronounced when the plants were vernalised at 8°C, rather than the lower temperatures reported for other species

vi) saturated after 2 weeks in vernalisation conditions of 4°C

vii) linked to either a single gene, or a pair of genes where one gene is dominant and the second gene is recessive.

In this study only the vernalisation response was investigated. In most vernalisation responsive plant species, there is also a response to lengthening daylight. It is this interplay between exposure to vernalisation conditions and an increased exposure to light that determines how a plant responds to winter. While the day length conditions

for safflower was briefly considered in this project in so much as to determine that long day conditions were essential for expression of the vernalisation response, the affect of day length on safflower was beyond the scope of this project. No further investigations into this trait were undertaken. To successfully incorporate a vernalisation response into elite safflower cultivars, their response to day length also needs to be taken into consideration.

The primary focus of this thesis was the investigation of the vernalisation response in two safflower cultivars; one that was sensitive to vernalisation conditions and one that was insensitive. In many plant species, a vernalisation response is coupled with a response to increasing day length, such as the conditions found in late winter and early spring. While increased day length was considered in so much as it was a necessary environmental condition to trigger flowering post vernalisation, no further investigations of this day length response was undertaken. To incorporate the vernalisation response into elite safflower cultivars that are currently vernalisation insensitive, the response of safflower to increasing day length also needs to be taken into consideration. However, the genetic mechanisms controlling the day length response in safflower was outside the scope of this project.

High quality reference sequences, be they genomic or transcriptomic, are paramount for effectively investigating any kind of phenotypic or genetic pathways in any species. In the last ten years, the generation of such sequence-based resources has been facilitated by the decrease in cost of their use, an increase in data yield and a reduction in error rates of the technologies available. When compared to *Arabidopsis* or the cereals, the *Asteraceae* have only a small number of publicly available resources and these are restricted in their scope to specific metabolic pathways, SNP markers for specific crosses, Expressed Sequence Tags (ESTs) or smaller resources such as the chloroplast genome. This aspect of the project aimed to redress this deficiency in available information and create reference assemblies that could not only be used for the characterisation of the gene regulation underpinning the vernalisation response in safflower, but be used to investigate any other metabolic or physiological processes of interest.

In Chapter 3, two transcriptomic references were created, each using different assembly software. The first was assembled from RNA-Seq libraries extracted from 16 different and diverse spring safflower tissues. The second reference was created from vernalised and unvernalised winter safflower. When these references were aligned, the high degree of similarity between the two assemblies gave confidence in their accuracy, rather than being considered an artefact created by the assembly algorithms. When the vernalised and unvernalised winter safflower RNA-Seq reads were aligned to both the winter and spring safflower transcriptomes, (30 transcripts vs. 20 transcripts respectively), three of the winter safflower transcripts were not found in the list of 30

differentially expressed spring safflower transcripts. Although NCBI did not annotate these three winter safflower transcripts as being involved with the vernalisation response, their differential expression in winter but not spring safflower identifies each as a potentially important candidate requiring further experimental characterisation.

A genomic reference was also assembled using Illumina short read libraries created from spring safflower DNA. When the spring safflower transcriptome was aligned against the genomic Illumina assembly, a high degree of similarity was reported. Specifically, 98.7% of transcripts from spring safflower were mapped to one or more positions on the CSIRO draft safflower genome. This not only increased the confidence in the quality of the assemblies, but the establishment of an intron/exon structure of a number of genes was possible.

Four different transcripts were identified as being differentially expressed in the vernalisation response, *CtAP1-LIKE*, *CtMADS1*, *CtFT-LIKE* and *CtVRN1-LIKE*. These transcripts were differentially expressed in both experiments and were constitutively expressed in spring safflower. Using RT-qPCR, the expression of all four transcripts in Experiment 1 and the expression of *CtFT-LIKE* and *CtMADS1* in Experiment 2 were confirmed as involved in the vernalisation response in safflower.

To further the understanding of the vernalisation response, it will be essential to examine the differentially expressed transcripts and genomic regulatory regions using knock-out and mutant safflower lines. After winter safflower has been exposed to vernalisation conditions, *CtFT-LIKE* and *CtMADS1* shows an increase in expression. While this profile is similar for *CtFT-LIKE* and *AtFT-LIKE*, *AtFLC* decreases in expression after exposure to vernalisation, the opposite to what is seen in *CtMADS1* (under the assumption that *CtMADS1* is a functional homologue of *AtFLC*). Generating random mutations or creating RNAi constructs targeting specific transcripts could cause a winter safflower cultivar, or a vernalisation responsive cross, to express a spring safflower-like phenotype, or vice versa. Further, a transgenic approach would allow for a deeper understanding of the mechanics of vernalisation in safflower and develop a clearer picture of how this trait could be utilised to modify existing cultivars or, possibly used to produce new elite varieties of safflowers.

Despite a high level of confidence in the accuracy of all of the assemblies generated, they are limited by the fragmented state of the CSIRO draft safflower genome. While nearly all the transcriptomic contigs align somewhere on this genomic reference created with Illumina reads, the assembly has an n50 of just under 2,000 bp, with the largest contig being 32,974 bp in length. Even when transcripts align to multiple genomic contigs allowing them to be linked and annotated as an intron, the length of this intronic region can only be estimated. Similarly, only limited information can be gathered on the upstream and downstream regulatory regions that flank individual loci.

Newer technologies, such as the recently developed PacBio and Chicago sequencing methodologies, have the potential to augment Illumina assemblies by scaffolding these fragmented contigs together. Longer contigs will facilitate future research into other aspects of safflower genetics, specifically the identification of regulatory sequences in the non-protein-coding regions of the safflower genome. This knowledge could be used to guide even more research into existing safflower varieties, directing and informing the development of GM cultivars and characterising important and novel traits found in other wild safflower germplasm, thereby improving safflower as an oilseed crop.

Nonetheless, the generation of crossing populations and development of elite lines, via either traditional cross breeding or with GM methods, requires unique and specific markers to guide this process. Markers are essential to identify and track existing and novel traits in any crop or plant breeding strategy, not just for safflower. Twenty seven of the digest markers for the vernalisation response were identified from the genomic reference and were located within very close proximity to each other on chromosome 8. Although these markers were obtained from the crossing families that expressed both early and late elongation phenotypes, this elongation phenotype is directly related to the vernalisation response in safflower. In Chapter 2, it was hypothesised that, while a single gene model fit the physiological segregation of the late elongation trait proxy for vernalisation, a two gene model, where one gene was dominant and other recessive, was a more confident fit with the data. However, this two gene model may not be correct. Clustering of digest markers to one area on a single chromosome and a transcript that also maps to this location implies that a single loci may be responsible for the vernalisation response.

Carapetian (2001) observed that the rosette habit, i.e. elongation, of safflower follows a 3:1 ratio of short rosette (spring):long rosette (winter) behaviour. The characterisation and the segregation of the vernalisation response observed in Chapter 2 indicates a single recessive gene with nearly all the genetic markers found in the Bowers genetic map clustered on chromosome 8, with a single transcript of unknown function. This data combined with, the observations of Carapetian (2001), shows solid evidence that the vernalisation response in safflower is caused by a single recessive locus. To further characterise the molecular mechanisms underpinning the vernalisation response in safflower, it will be necessary to characterise the function of this unknown transcript mapping to chromosome 8 and its relationship to the SNPs located on the same chromosome.

An important limitation with the SNP data is that the crossing families sent to DArT for sequencing were from the F_3 population rather than the F_2. This has most probably fixed a number of recombinant loci in these families, confounding the markers reported by the DArTSeq analysis. Rebuilding a crossing population, using single seed descent winter and spring safflower parents, and repeating the crossing experiment and DArT

marker sequencing using F_2 recombinants, without the need for backcrossing, would provide the clarity needed to discover further SNP markers for the vernalisation response in safflower.

An unexpected result of assembling the first PacBio sequencing library was a single 1.5 Mbp contig that mapped to a number of different chloroplast genomes in other species. Further analysis of this 'chloroplast genome' showed that despite being ten times the size of other published chloroplast genomes, the presence of a large number of unique regions across its entire length and the unique alignment of Illumina genomic reads to these locations, implies an accurate and correct assembly. A number of metabolic processes utilise the chloroplast as a biofactory, especially with regard to fatty acid synthesis (Ohlrogge and Jaworski 1997). If it is confirmed that the chloroplast is indeed much larger than previously described and contains a greater diversity of genic sequences and the associated molecular machinery within, this finding could have major implications on our fundamental understanding of not only the structure of the chloroplast genome, but how a chloroplast operates and functions. Just as importantly, this revised chloroplast genome could be leveraged in the generation of GM crops. While this discovery definitely warrants further investigation, it was outside of the scope of this project.

Finally, there are several broader results that can be taken from this project with regard to safflower as an oilseed crop. One of the more obvious results is that the vernalisation response is a trait that is transferable to elite cultivars via traditional cross breeding methods. By crossing vernalisation responsive safflower varieties with existing cultivars and fixing this trait in a population allows for greater flexibility regarding where and when safflower can be planted in both Australia and globally. Further investigation of the vernalisation response in safflower may result in a mechanism, produced either via traditional cross breeding or GM methods, where the vernalisation response is 'activated' if seeds are planted in late autumn or winter, but behave like a spring safflower when planted after winter or in a non-temperate climate. This could then be combined with desirable GM traits e.g. those found in GLA or SHO elite safflower varieties, to develop a safflower suitable for use as a break crop in a wider range of growing regions in Australia, that by producing a lucrative oil, ensures farmers receive a greater return on investment when the safflower is harvested. Investigations into the interaction between day length and the vernalisation response will be critical to the success of such a mechanism. But the processes and resources described herein are not limited in application to the vernalisation response. Any trait expressed in safflower could be investigated following a similar protocol, for example (but not limited to), the seed oil profile and content, the flower colour and plant spikiness. A critical mass of genetic resources for safflower would assist in the creation of reference resources for the broader *Asteraceae* and help facilitate a variety of research in this family.

This PhD project combined physiological, transcriptomic and genomic methods in an holistic approach to investigate the vernalisation response in safflower (Fig. 5.1). The development of several accurate genetic resources for safflower opens up a number of pathways for investigation and improvement in this oilseed. Further, the information currently seen in safflower may be useful, even critical, in the research of traits and behaviours of other important *Asteraceae*, such as lettuce, sunflower and artichoke. The processes outlined herein, combined with the digital resources produced, may be used to define any expressed trait in any dicot, not just the vernalisation response in safflower. These processes can be exploited to determine the molecular basis of any trait in any uncharacterised diploid genome, with or without an existing reference.

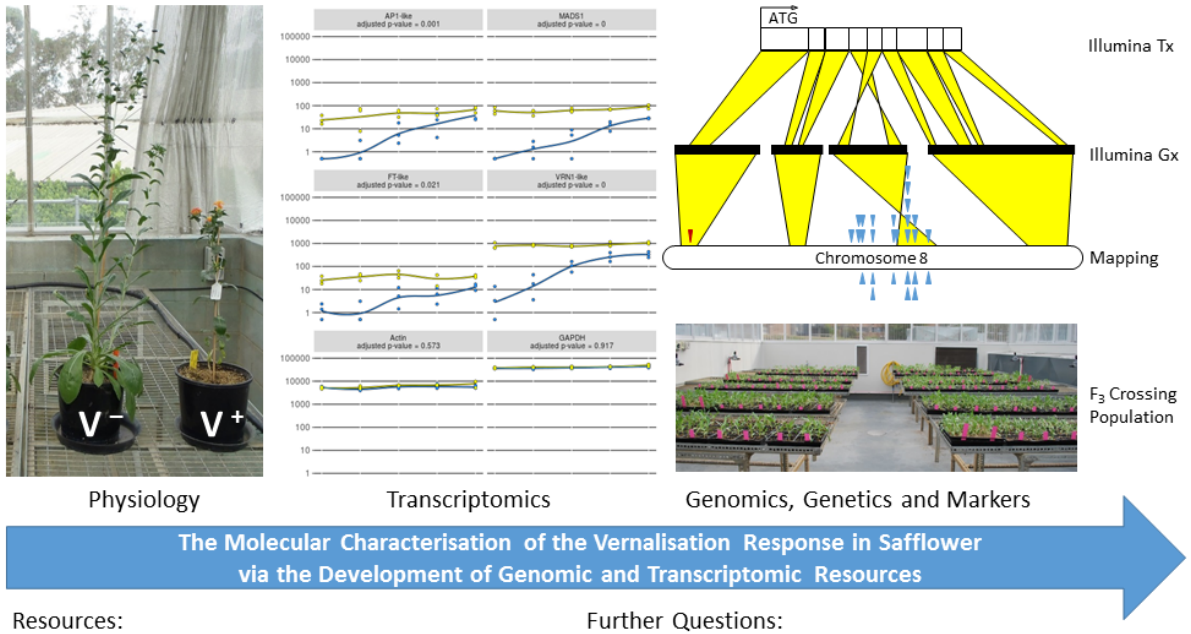


FIGURE 5.1: Summary of the PhD Project "The Molecular Characterisation of Vernalisation in Safflower via the Development of Genomic and Transcriptomic Resources", with resources developed and future questions.

References

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

Airoldi, C. A., McKay, M., and Davies, B. (2015). *MAF2* Is Regulated by Temperature-Dependent Splicing and Represses Flowering at Low Temperatures in Parallel with *FLM*. *PLoS ONE*, 10(5):e0126516.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402.

Amasino, R. (2004). Vernalization, Competence, and the Epigenetic Memory of Winter. *The Plant Cell Online*, 16(10):2553–2559.

Amasino, R. M. (2005). Vernalization and flowering time. *Current Opinion in Biotechnology*, 16(2):154–158.

Andres, F. and Coupland, G. (2012). The genetic basis of flowering responses to seasonal cues. *Nat Rev Genet*, 13(9):627–639.

Angus, J. F., Cunningham, R. B., Moncur, M. W., and MacKenzie, D. H. (1980). Phasic development in field crops. I. Thermal response in the seedling phase. *Field Crop Research*, 3:365–378.

Belide, S., Hac, L., Singh, S. P., Green, A. G., and Wood, C. C. (2011). *Agrobacterium*-mediated transformation of safflower and the efficient recovery of transgenic plants via grafting. *Plant Methods*, 7:12.

Bendich, A. J. (2004). Circular chloroplast chromosomes: The grand illusion. *The Plant Cell*, 16(7):1661–1666.

Berardini, T. Z., Reiser, L., Li, D., Mezheritsky, Y., Muller, R., Strait, E., and Huala, E. (2015). The *Arabidopsis* Information Resource: Making and mining the 'gold standard' annotated reference plant genome. *Genesis*, 53(8):474–485.

Bowers, J. E., Pearl, S. A., and Burke, J. M. (2016). Genetic Mapping of Millions of SNPs in Safflower (*Carthamus tinctorius* L.) via Whole-Genome Resequencing. *G3: Genes | Genomes | Genetics*, 6(7):2203–2211.

Cao, S., Zhou, X.-R., Wood, C. C., Green, A., Singh, S., Liu, L., and Liu, Q. (2013). A large and functionally diverse family of *FAD2* genes in safflower (*Carthamus tinctorius* L.). *BMC Plant Biology*, 13(1):5.

Carapetian, J. (2001). Characterisation and Inheritance of Long Rosette Safflower. In *Vth International Safflower Conference*, pages 67–71, Williston, N.D.

Chao, Y., Yang, Q., Kang, J., Zhang, T., and Sun, Y. (2013). Expression of the alfalfa *FRIGIDA*-Like Gene, *MsFRI-L* delays flowering time in transgenic *Arabidopsis thaliana*. *Molecular Biology Reports*, 40(3):2083–2090.

Chouard, P. (1960). Vernalization and its Relations to Dormancy. *Annual Review of Plant Physiology*, 11(1):191–238.

Corbesier, L., Vincent, C., Jang, S., Fornara, F., Fan, Q., Searle, I., Giakountis, A., Farrona, S., Gissot, L., Turnbull, C., and Coupland, G. (2007). FT Protein Movement Contributes to Long-Distance Signaling in Floral Induction of *Arabidopsis*. *Science*, 316(5827):1030–1033.

De Lucia, F., Crevillen, P., Jones, A. M. E., Greb, T., and Dean, C. (2008). A PHD-Polycomb Repressive Complex 2 triggers the epigenetic silencing of *FLC* during vernalization. *Proceedings of the National Academy of Sciences*, 105(44):16831–16836.

Deng, W., Casao, M. C., Wang, P., Sato, K., Hayes, P. M., Finnegan, E. J., and Trevaskis, B. (2015). Direct links between the vernalization response and other key traits of cereal crops. *Nature Communications*, 6.

Deng, X.-W., Wing, R. A., and Gruissem, W. (1989). The chloroplast genome exists in multimeric forms. *Proceedings of the National Academy of Sciences*, 86(11):4156–4160.

Dennis, E. and Peacock, W. J. (2009). Vernalization in cereals. *Journal of Biology*, 8(6):57.

Dijk, H. V., Boudry, P., McCombre, H., and Vernet, P. (1997). Flowering time in wild beet (*Beta vulgaris* ssp. *maritima*) along a latitudinal cline. *Acta Oecologica*, 18(1):47–60.

Ferrarini, M., Moretto, M., Ward, J. A., Šurbanovski, N., Stevanović, V., Giongo, L., Viola, R., Cavalieri, D., Velasco, R., Cestaro, A., and Sargent, D. J. (2013). An evaluation of the PacBio RS platform for sequencing and de novo assembly of a chloroplast genome. *BMC Genomics*, 14(1):1–12.

Finnegan, E. J. and Dennis, E. S. (2007). Vernalization-Induced Trimethylation of Histone H3 Lysine 27 at FLC Is Not Maintained in Mitotically Quiescent Cells. *Current Biology*, 17(22):1978–1983.

Finnegan, E. J., Kovac, K. A., Jaligot, E., Sheldon, C. C., Peacock, W. J., and Dennis, E. S. (2005). The downregulation of *FLOWERING LOCUS C* (*FLC*) expression in plants with low levels of DNA methylation and by vernalization occurs by distinct mechanisms. *The Plant Journal*, 44(3):420–432.

Fletcher, J. C. (2002). Shoot and Floral Meristem Maintenance in *Arabidopsis*. *Annual Review of Plant Biology*, 53(1):45–66.

Garnatje, T., Garcia, S., Vilatersana, R., and Vallès, J. (2006). Genome Size Variation in the Genus *Carthamus* (Asteraceae, Cardueae): Systematic Implications and Additive Changes During Allopolyploidization. *Annals of Botany*, 97(3):461–467.

Gassner, G. (1918). Beiträge zur physiologiischen Charakteristik sommer- und winterannueller Gewächse, insbesondere der Getreidepflanzen. *Zeitschrift für Botanik*, 10:417–480.

Gecgel, U., Demirci, M., Esendal, E., and Tasan, M. (2007). Fatty Acid Composition of the Oil from Developing Seeds of Different Varieties of Safflower (*Carthamus tinctorius* L.). *Journal of the American Oil Chemists' Society*, 84(1):47–54.

Gehring, M. (2013). Genomic Imprinting: Insights From Plants. *Annual Review of Genetics*, 47(1):187–208.

Gladstones, J. S. and Hill, G. D. (1969). Selection for economic characters in *Lupinus angustifolius* and *L. digitatus*. 2. Time of flowering. *Australian Journal of Experimental Agriculture*, 9(37):213–220.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., and Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, 29(7):644–652.

Gray, S. G. (1942). Increased Earlieness of Flowering in Lettuce Through Vernalisation. *Journal of the Council for Scientific and Industrial Research, Australia*, 15(3):211–212.

Greb, T., Mylne, J. S., Crevillen, P., Geraldo, N., An, H., Gendall, A. R., and Dean, C. (2007). The PHD Finger Protein VRN5 Functions in the Epigenetic Silencing of *Arabidopsis FLC*. *Current Biology*, 17(1):73–78.

Greenup, A. G., Sasani, S., Oliver, S. N., Talbot, M. J., Dennis, E. S., Hemming, M. N., and Trevaskis, B. (2010). *ODDSOC2* Is a MADS Box Floral Repressor That Is Down-Regulated by Vernalization in Temperate Cereals. *Plant Physiology*, 153(3):1062–1073.

Hecht, V., Foucher, F., Ferrándiz, C., Macknight, R., Navarro, C., Morin, J., Vardy, M. E., Ellis, N., Beltrán, J. P., Rameau, C., and Weller, J. L. (2005). Conservation of *Arabidopsis* Flowering Genes in Model Legumes. *Plant Physiology*, 137(4):1420–1434.

Hecht, V., Laurie, R. E., Vander Schoor, J. K., Ridge, S., Knowles, C. L., Liew, L. C., Sussmilch, F. C., Murfet, I. C., Macknight, R. C., and Weller, J. L. (2011). The Pea *GIGAS* Gene is a *FLOWERING LOCUS T* Homolog Necessary for Graft-Transmissible Specification of Flowering but not for Responsiveness to Photoperiod. *The Plant Cell Online*, 23(1):147–161.

Helliwell, C. A., Anderssen, R. S., Robertson, M., and Finnegan, E. J. (2015). How is *FLC* repression initiated by cold? *Trends in Plant Science*, 20(2):76–82.

Işigigür, A., Karaosmanoglu, F., and Aksoy, H. A. (1995). Characteristics of safflower seed oils of turkish origin. *Journal of the American Oil Chemists' Society*, 72(10):1223–1225.

Ivany, L. C., Patterson, W. P., and Lohmann, K. C. (2000). Cooler winters as a possible cause of mass extinctions at the Eocene/Oligocene boundary. *Nature*, 407(6806):887–890.

Jaeger, K. E. and Wigge, P. A. (2007). FT Protein Acts as a Long-Range Signal in Arabidopsis. *Current Biology*, 17(12):1050–1054.

Jaudal, M., Yeoh, C. C., Zhang, L., Stockum, C., Mysore, K. S., Ratet, P., and Putterill, J. (2013). Retroelement insertions at the Medicago *FTa1* locus in spring mutants eliminate vernalisation but not long-day requirements for early flowering. *The Plant Journal*, 76(4):580–591.

Jochinke, D., Wachsmann, N., Potter, T., and Norton, R. (2008). Growing safflower in Australia: Part 1 - History, experiences and current constraints on production. In *7th International Safflower Conference*, page 7.

Johnson, R. C., Dajue, L., and Bradley, V. (2006). Autumn growth and its relationship to winter survival in diverse safflower germplasm. *Canadian Journal of Plant Science*, 86(3):701–709.

Johnson, R. C. and Li, D. (2008). Registration of WSRC01, WSRC02, and WSRC03 Winter-Hardy Safflower Germplasm. *Germplasm*, 2(2):140–142.

Klippart, J. H. (1857). An essay on the origin, growth, diseases, varieties etc of the wheat plant. *Annual Report of the Ohio State Board of Agriculture*, pages 562–720.

Knights, S. (2010). Raising the bar with better safflower agronomy. Technical report, Grains Research and Development Corporation and Australian Oilseeds Federation.

Knowles, P. F. (1949). Dual purpose safflower seed: Produces drying oil for the paint industry and seed meal for livestock and poultry feed. *California Agriculture*, 3(2):11–16.

Knowles, P. F. (1960). Safflower's Native Home. *Crops and Soils*, 12(6):17.

Knowles, P. F. (2012). *Safflower in California: The Paulden F. Knowles personal history of plant exploration and research on evolution, genetics, and breeding*. Department of Plant Sciences, University of California, Davis.

Köhler, C. and Villar, C. B. R. (2008). Programming of gene expression by Polycomb group proteins. *Trends in Cell Biology*, 18(5):236–243.

Landers, K. F. (1995). Vernalization responses in narrow-leafed lupin (*Lupinus angustifolius*) genotypes. *Australian Journal of Agricultural Research*, 46(5):1011–1025.

Laurie, R. E., Diwadkar, P., Jaudal, M., Zhang, L., Hecht, V., Wen, J., Tadege, M., Mysore, K. S., Putterill, J., Weller, J. L., and Macknight, R. C. (2011). The Medicago *FLOWERING LOCUS T* Homolog, *MtFTa1*, Is a Key Regulator of Flowering Time. *Plant Physiology*, 156(4):2207–2224.

Lee, J. and Lee, I. (2010). Regulation and function of *SOC1*, a flowering pathway integrator. *Journal of Experimental Botany*, 61(9):2247–2254.

Levy, Y. Y., Mesnage, S., Mylne, J. S., Gendall, A. R., and Dean, C. (2002). Multiple Roles of *Arabidopsis VRN1* in Vernalization and Flowering Time Control. *Science*, 297(5579):243–246.

Li, H., Dong, Y., Yang, J., Liu, X., Wang, Y., Yao, N., Guan, L., Wang, N., Wu, J., and Li, X. (2012). De Novo Transcriptome of Safflower and the Identification of Putative Genes for Oleosin and the Biosynthesis of Flavonoids. *PLoS ONE*, 7(2):e30987.

Liu, X., Dong, Y., Yao, N., Zhang, Y., Wang, N., Cui, X., Li, X., Wang, Y., Wang, F., Yang, J., Guan, L., Du, L., Li, H., and Li, X. (2015). De novo sequencing and analysis of the safflower transcriptome to discover putative genes associated with safflor yellow in carthamus tinctorius l. *International Journal of Molecular Sciences*, 16(10):25657–25677.

Loman, N. J., Misra, R. V., Dallman, T. J., Constantinidou, C., Gharbia, S. E., Wain, J., and Pallen, M. J. (2012). Performance comparison of benchtop high-throughput sequencing platforms. *Nat Biotech*, 30(5):434–439.

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

Lu, C., Shen, Q., Yang, J., Wang, B., and Song, C. (2016). The complete chloroplast genome sequence of safflower (carthamus tinctorius l.). *Mitochondrial DNA Part A*, 27(5):3351–3353. PMID: 25740214.

Lulin, H., Xiao, Y., Pei, S., Wen, T., and Shangqin, H. (2012). The first illumina-based de novo transcriptome sequencing and analysis of safflower flowers. *PLOS ONE*, 7(6):1–11.

McKeown, M., Schubert, M., Marcussen, T., Fjellheim, S., and Preston, J. C. (2016). Evidence for an early origin of vernalization responsiveness in temperate pooideae grasses. *Plant Physiology*, 172(1):416–426.

McKinney, H. H. (1940). Vernalization and the growth-phase concept. *The Botanical Review*, 6(1):25–47.

Michaels, S. D. and Amasino, R. M. (2001). Loss of *FLOWERING LOCUS C* Activity Eliminates the Late-Flowering Phenotype of *FRIGIDA* and Autonomous Pathway Mutations but Not Responsiveness to Vernalization. *The Plant Cell Online*, 13(4):935–941.

Michaels, S. D., He, Y., Scortecci, K. C., and Amasino, R. M. (2003). Attenuation of FLOWERING LOCUS C activity as a mechanism for the evolution of summer-annual flowering behavior in Arabidopsis. *Proceedings of the National Academy of Sciences*, 100(17):10102–10107.

Mündel, H.-H. and Bergman, J. (2010). Safflower. In Vollmann, J. and Rajcan, I., editors, *Oil Crops SE - 14*, volume 4 of *Handbook of Plant Breeding*, pages 423–447. Springer New York.

Mylne, J. S., Barrett, L., Tessadori, F., Mesnage, S., Johnson, L., Bernatavichute, Y. V., Jacobsen, S. E., Fransz, P., and Dean, C. (2006). LHP1, the *Arabidopsis* homologue of HETEROCHROMATIN PROTEIN1, is required for epigenetic silencing of *FLC*. *Proceedings of the National Academy of Sciences of the United States of America*, 103(13):5012–5017.

Naim, F., Nakasugi, K., Crowhurst, R. N., Hilario, E., Zwart, A. B., Hellens, R. P., Taylor, J. M., Waterhouse, P. M., and Wood, C. C. (2012). Advanced Engineering of Lipid Metabolism in *Nicotiana benthamiana* Using a Draft Genome and the V2 Viral Silencing-Suppressor Protein. *PLoS ONE*, 7(12):e52717.

Nakano, Y., Kawashima, H., Kinoshita, T., Yoshikawa, H., and Hisamatsu, T. (2011). Characterization of *FLC*, *SOC1* and *FT* homologs in *Eustoma grandiflorum*: effects of vernalization and post-vernalization conditions on flowering and gene expression. *Physiologia Plantarum*, 141(4):383–393.

Nordborg, M. and Bergelson, J. (1999). The Effect of Seed and Rosette Cold Treatment on Germination and Flowering Time in Some *Arabidopsis thaliana* (Brassicaceae) Ecotypes. *American Journal of Botany*, 86(4):470–475.

Notredame, C., Higgins, D. G., and Heringa, J. (2000). T-coffee: a novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology*, 302(1):205–217.

Nykiforuk, C. L., Shewmaker, C., Harry, I., Yurchenko, O. P., Zhang, M., Reed, C., Oinam, G. S., Zaplachinski, S., Fidantsef, A., Boothe, J. G., and Moloney, M. M. (2012). High level accumulation of gamma linolenic acid (C18:3Δ6.9,12 cis) in transgenic safflower (*Carthamus tinctorius*) seeds. *Transgenic Research*, 21(2):367–381.

Ohlrogge, J. B. and Jaworski, J. G. (1997). Regulation of Fatty Acid Synthesis. *Annual Review of Plant Physiology and Plant Molecular Biology*, 48(1):109–136.

Oldenburg, D. J. and Bendich, A. J. (2004). Most chloroplast {DNA} of maize seedlings in linear molecules with defined ends and branched forms. *Journal of Molecular Biology*, 335(4):953 – 970.

Oldenburg, D. J. and Bendich, A. J. (2016). The linear plastid chromosomes of maize: terminal sequences, structures, and implications for dna replication. *Current Genetics*, 62(2):431–442.

Oliver, S. N., Finnegan, E. J., Dennis, E. S., Peacock, W. J., and Trevaskis, B. (2009). Vernalization-induced flowering in cereals is associated with changes in histone methylation at the *VERNALIZATION1* gene. *Proceedings of the National Academy of Sciences*, 106(20):8386–8391.

Owen, F. V., Carsner, E., and Stout, M. (1940). Photothermal induction of flowering in suagr beets. *Journal of Agricultural Research*, 61:101–124.

Parra, G., Bradnam, K., and Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*, 23(9):1061–1067.

Pergola, G. (1992). The need for vernalization in *Eustoma russellianum*. *Scientia Horticulturae*, 51(1–2):123–127.

Périlleux, C., Pieltain, A., Jacquemin, G., Bouché, F., Detry, N., D'Aloia, M., Thiry, L., Aljochim, P., Delansnay, M., Mathieu, A.-S., Lutts, S., and Tocquin, P. (2013). A root chicory MADS box sequence and the *Arabidopsis* flowering repressor *FLC* share common features that suggest conserved function in vernalization and devernalization responses. *The Plant Journal*, 75(3):390–402.

Pilcher, C. D., Wong, J. K., and Pillai, S. K. (2008). Inferring HIV Transmission Dynamics from Phylogenetic Sequence Relationships. *PLoS Med*, 5(3):e69.

Pin, P. A., Benlloch, R., Bonnet, D., Wremerth-Weich, E., Kraft, T., Gielen, J. J. L., and Nilsson, O. (2010). An Antagonistic Pair of *FT* Homologs Mediates the Control of Flowering Time in Sugar Beet. *Science*, 330(6009):1397–1400.

Pin, P. A., Zhang, W., Vogt, S. H., Dally, N., Büttner, B., Schulze-Buxloh, G., Jelly, N. S., Chia, T. Y., Mutasa-Göttgens, E. S., Dohm, J. C., Himmelbauer, H., Weisshaar, B., Kraus, J., Gielen, J. J., Lommel, M., Weyens, G., Wahl, B., Schechert, A., Nilsson, O., Jung, C., Kraft, T., and Müller, A. E. (2012). The Role of a Pseudo-Response Regulator Gene in Life Cycle Adaptation and Domestication of Beet. *Current Biology*, 22(12):1095–1101.

Prothero, D. R. (1994). The Late Eocene-Oligocene Extinctions. *Annual Review of Earth and Planetary Sciences*, 22(1):145–165.

Putnam, N. H., O'Connell, B. L., Stites, J. C., Rice, B. J., Blanchette, M., Calef, R., Troll, C. J., Fields, A., Hartley, P. D., Sugnet, C. W., Haussler, D., Rokhsar, D. S., and Green, R. E. (2016). Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Research*.

Putterill, J., Zhang, L., Yeoh, C. C., Balcerowicz, M., Jaudal, M., and Gasic, E. V. (2013). *FT* genes and regulation of flowering in the legume *Medicago truncatula*. 40(12):1199–1207.

Rappapport, L., Wittwer, S. H., and Tukey, H. B. (1956). Seed Vernalization and Flowering in Lettuce (*Lactuca sativa*). *Nature*, 178(4523):51.

Ratcliffe, O. J., Kumimoto, R. W., Wong, B. J., and Riechmann, J. L. (2003). Analysis of the Arabidopsis *MADS AFFECTING FLOWERING* Gene Family: *MAF2* Prevents Vernalization by Short Periods of Cold. *The Plant Cell Online*, 15(5):1159–1169.

Ream, T. S., Woods, D. P., Schwartz, C. J., Sanabria, C. P., Mahoy, J. A., Walters, E. M., Kaeppler, H. F., and Amasino, R. M. (2014). Interaction of Photoperiod and Vernalization Determines Flowering Time of *Brachypodium distachyon*. *Plant Physiology*, 164(2):694–709.

Reeves, P. A., He, Y., Schmitz, R. J., Amasino, R. M., Panella, L. W., and Richards, C. M. (2007). Evolutionary Conservation of the *FLOWERING LOCUS C*-Mediated Vernalization Response: Evidence From the Sugar Beet (*Beta vulgaris*). *Genetics*, 176(1):295–307.

Reid, J. B. and Murfet, I. C. (1975). Flowering in *Pisum*: the Sites and Possible Mechanisms of the Vernalization Response. *Journal of Experimental Botany*, 26(6):860–867.

Ross, J. J. and Murfet, I. C. (1986). The Mechanism of Action of Vernalization in *Lathyrus odoratus* L. *Annals of Botany*, 57(6):783–790.

Ruelens, P., de Maagd, R. A., Proost, S., Theißen, G., Geuten, K., and Kaufmann, K. (2013). FLOWERING LOCUS C in monocots and the tandem origin of angiosperm-specific MADS-box genes. *Nature Communications*, 4.

Sato, S., Nakamura, Y., Kaneko, T., Asamizu, E., and Tabata, S. (1999). Complete Structure of the Chloroplast Genome of *Arabidopsis thaliana*. *DNA Research*, 6(5):283–290.

Schittenhelm, S. (2001). Effect of sowing date on the performance of root chicory. *European Journal of Agronomy*, 15(3):209–220.

Sheldon, C. C., Conn, A. B., Dennis, E. S., and Peacock, W. J. (2002). Different Regulatory Regions Are Required for the Vernalization-Induced Repression of FLOWERING LOCUS C and for the Epigenetic Maintenance of Repression. *The Plant Cell Online*, 14(10):2527–2537.

Sheldon, C. C., Hills, M. J., Lister, C., Dean, C., Dennis, E. S., and Peacock, W. J. (2008). Resetting of *FLOWERING LOCUS C* expression after epigenetic repression by vernalization. *Proceedings of the National Academy of Sciences*, 105(6):2214–2219.

Sheldon, C. C., Rouse, D. T., Finnegan, E. J., Peacock, W. J., and Dennis, E. S. (2000). The molecular basis of vernalization: The central role of *FLOWERING LOCUS C* (*FLC*). *Proceedings of the National Academy of Sciences*, 97(7):3753–3758.

Silva, I. P. and Jenkins, D. G. (1993). Decision on the Eocene-Oligocene boundary stratotype. *Episodes*, 16(3):379–382.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19):3210–3212.

Simpson, G. G. and Dean, C. (2002). *Arabidopsis*, the Rosetta Stone of Flowering Time? *Science*, 296(5566):285–289.

Smith, J. (1996). *Safflower*. AOCS Press, Champaign, Illinois.

Smith, S. A., Beaulieu, J. M., Stamatakis, A., and Donoghue, M. J. (2011). Understanding angiosperm diversification using small and large phylogenetic trees. *American Journal of Botany*, 98(3):404–414.

Song, J., Angel, A., Howard, M., and Dean, C. (2012). Vernalization – a cold-induced epigenetic switch. *Journal of Cell Science*, 125(16):3723–3731.

Speelman, E. N., Van Kempen, M. M. L., Barke, J., Brinkhuis, H., Reichart, G. J., Smolders, A. J. P., Roelofs, J. G. M., Sangiorgi, F., De Leeuw, J. W., Lotter, A. F., and Sinninghe Damste, J. S. (2009). The Eocene Arctic Azolla bloom: environmental conditions, productivity and carbon drawdown. *Geobiology*, 7(2):155–170.

Stout, M. (1945). Translocation of the reproductive stimulus in sugar beets. *Botanical Gazette*, 107(1):86–95.

Sung, S. and Amasino, R. M. (2004a). Vernalization and epigenetics: how plants remember winter. *Current Opinion in Plant Biology*, 7(1):4–10.

Sung, S. and Amasino, R. M. (2004b). Vernalization in *Arabidopsis thaliana* is mediated by the PHD finger protein VIN3. *Nature*, 427(6970):159–164.

Sung, S., Schmitz, R. J., and Amasino, R. M. (2006). A PHD finger protein involved in both the vernalization and photoperiod pathways in *Arabidopsis*. *Genes & Development*, 20(23):3244–3248.

Trevaskis, B. (2010). The central role of the *VERNALIZATION1* gene in the vernalization response of cereals. 37(6):479–487.

Trevaskis, B., Bagnall, D. J., Ellis, M. H., Peacock, W. J., and Dennis, E. S. (2003). MADS box genes control vernalization-induced flowering in cereals. *Proceedings of the National Academy of Sciences*, 100(22):13099–13104.

Trevaskis, B., Hemming, M. N., Dennis, E. S., and Peacock, W. J. (2007a). The molecular basis of vernalization-induced flowering in cereals. *Trends in Plant Science*, 12(8):352–357.

Trevaskis, B., Hemming, M. N., Peacock, W. J., and Dennis, E. S. (2006). *HvVRN2* Responds to Daylength, whereas *HvVRN1* is Regulated by Vernalization and Developmental Status. *Plant Physiology*, 140(4):1397–1405.

Trevaskis, B., Tadege, M., Hemming, M. N., Peacock, W. J., Dennis, E. S., and Sheldon, C. (2007b). Short Vegetative Phase-Like MADS-Box Genes Inhibit Floral Meristem Identity in Barley. *Plant Physiology*, 143(1):225–235.

United Nations (2016). FAOSTAT.

United States Department of Agriculture (2016). Basic Report: 04511, Oil, safflower, salad or cooking, high oleic (primary safflower oil of commerce).

Vogt, S. H., Weyens, G., Lefèbvre, M., Bork, B., Schechert, A., and Muller, A. E. (2014). The *FLC*-like gene *BvFL1* is not a major regulator of vernalization response in biennial beets. *Frontiers in Plant Science*, 5.

Wade, B. S., Houben, A. J. P., Quaijtaal, W., Schouten, S., Rosenthal, Y., Miller, K. G., Katz, M. E., Wright, J. D., and Brinkhuis, H. (2012). Multiproxy record of abrupt sea-surface cooling across the Eocene-Oligocene transition in the Gulf of Mexico. *Geology*, 40(2):159–162.

Wang, R., Farrona, S., Vincent, C., Joecker, A., Schoof, H., Turck, F., Alonso-Blanco, C., Coupland, G., and Albani, M. C. (2009). *PEP1* regulates perennial flowering in *Arabis alpina*. *Nature*, 459(7245):423–427.

Wang, X. (2010). *The Dawn Angiosperms*. Berlin, Berlin.

Warne, L. G. G. (1947). Vernalization of Lettuce. *Nature*, 159(4027):31–32.

Waycott, W. (1995). Photoperiodic Response of Genetically Diverse Lettuce Accessions. *Journal of the American Society for Horticultural Science*, 120(3):460–467.

Wellensiek, S. J. (1962). Dividing Cells as the Locus for Vernalization. *Nature*, 195(4838):307–308.

Werner, J. D., Borevitz, J. O., Henriette Uhlenhaut, N., Ecker, J. R., Chory, J., and Weigel, D. (2005). *FRIGIDA*-independent variation in flowering time of natural *Arabidopsis thaliana* accessions. *Genetics*, 170(3):1197–1207.

Wetterstrand, K. (2014). DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP).

Wikström, N., Savolainen, V., and Chase, M. W. (2001). Evolution of the angiosperms: calibrating the family tree. *Proceedings of the Royal Society of London B: Biological Sciences*, 268(1482):2211–2220.

Wilfinger, W. W., Mackey, K., and Chomcynski, P. (1997). Effect of ph and ionic strength on the spectrophotometric asssessment of nucleic acid purity. *BioTechniques*, 22(3):474–480.

Wolfe, K. H., Gouy, M., Yang, Y. W., Sharp, P. M., and Li, W. H. (1989). Date of the monocot-dicot divergence estimated from chloroplast DNA sequence data. *Proceedings of the National Academy of Sciences*, 86(16):6201–6205.

Wood, C. C., Robertson, M., Tanner, G., Peacock, W. J., Dennis, E. S., and Helliwell, C. A. (2006). The *Arabidopsis thaliana* vernalization response requires a polycomb-like protein complex that also includes VERNALIZATION INSENSITIVE 3. *Proceedings of the National Academy of Sciences*, 103(39):14631–14636.

Woods, D. P., Ream, T. S., and Amasino, R. M. (2014). Memory of the Vernalized State in Plants including the Model Grass *Brachypodium distachyon*. *Frontiers in Plant Science*, 5.

Yan, L., Fu, D., Li, C., Blechl, A., Tranquilli, G., Bonafede, M., Sanchez, A., Valarik, M., Yasuda, S., and Dubcovsky, J. (2006). The wheat and barley vernalization gene VRN3 is an orthologue of FT. *Proceedings of the National Academy of Sciences*, 103(51):19581–19586.

Yan, L., Loukoianov, A., Blechl, A., Tranquilli, G., Ramakrishna, W., SanMiguel, P., Bennetzen, J. L., Echenique, V., and Dubcovsky, J. (2004). The Wheat *VRN2* Gene Is a Flowering Repressor Down-Regulated by Vernalization. *Science*, 303(5664):1640–1644.

Yan, Y., Shen, L., Chen, Y., Bao, S., Thong, Z., and Yu, H. (2014). A myb-domain protein {EFM} mediates flowering responses to environmental cues in arabidopsis. *Developmental Cell*, 30(4):437 – 448.

Zohary, D. and Hopf, M. (1993). *Domestication of Plants in the Old World*. Oxford Scientific Publications, 2nd ed. edition.

Appendix A

Significantly Differentially Expressed Spring Safflower Transcripts from Experiment 2

<Content starts on the following page>

TABLE A.1: Significantly ($\alpha = 0.05$) differentially expressed spring safflower transcriptomic contigs from Experiment 2, with BLASTP sequence homology results using all six reading frames. Sequences were classified into 4 categories: 1) Annotated and believed to be part of the vernalisation response (in **bold**); 2) Annotated but unclear if they are involved in the vernalisation response; 3) No annotation information, and; 4) Differentially expressed when winter and spring safflower are compared, but no change in the expression of winter safflower throughout the timecourse.

Contig	Mean Counts	Fold Change (log2)	Adjusted p-value	BLASTP Homologue	Category
CarTin_tx_s317_comp33367_c7_seq4	**39**	**-0.27**	**0.000**	**MADS1 (MADS box containing)**	**1**
CarTin_tx_s317_comp33519_c0_seq70	**512**	**-0.29**	**0.000**	**Vernalisation 1 (VRN1)-like**	**1**
CarTin_tx_s317_comp1764285_c0_seq1	65	-0.22	0.000	No Hits Found	4
CarTin_tx_s317_comp1571506_c0_seq1	19	-0.22	0.000	Ribonuclease/endonuclease/exonuclease	4
CarTin_tx_s317_comp26769_c0_seq1	**32**	**-0.31**	**0.001**	**Apetala 1 (AP1)-like**	**1**
CarTin_tx_s317_comp4835_c0_seq1	34	-0.16	0.001	Miraculin	2
CarTin_tx_s317_comp13787_c0_seq1	69	0.13	0.002	Monothiol glutaredoxin-S2	2
CarTin_tx_s317_comp1578437_c0_seq1	36	-0.19	0.002	No Hits Found	4
CarTin_tx_s317_comp2219162_c0_seq1	41	-0.28	0.002	Hypothetical protein/retrotransposable element	4
CarTin_tx_s317_comp10252_c0_seq1	44	-0.12	0.004	No hits found	4
CarTin_tx_s317_comp31946_c0_seq1	1111	0.21	0.004	Uncharacterised/hypothetical protein	3
CarTin_tx_s317_comp360223_c0_seq1	240	0.04	0.006	Uncharacterised/hypothetical protein	4
CarTin_tx_s317_comp23005_c0_seq1	22	-0.14	0.006	Uncharacterised/hypothetical protein	4
CarTin_tx_s317_comp1426363_c0_seq1	69	-0.09	0.008	Terpene/germacrene A synthase	2
CarTin_tx_s317_comp20690_c0_seq1	1115	0.03	0.009	*Homeodomain-like protein/MYB transcription factor*	4
CarTin_tx_s317_comp69290_c0_seq1	162	0.16	0.009	Quinone oxidoreductase-like protein	2
CarTin_tx_s317_comp963682_c0_seq1	142	0.07	0.009	Abscisic acid receptor PYL9-like	2
CarTin_tx_s317_comp80349_c0_seq1	665	-0.07	0.012	Leucine-rich repeat-containing protein	2

TABLE A.1: Significantly ($\alpha = 0.05$) differentially expressed spring safflower transcriptomic contigs from Experiment 2 (continued).

Contig	Mean Counts	Fold Change (log2)	Adjusted p-value	BLASTP Homologue	Category
CarTin_tx_s317_comp14932_c0_seq1	1415	0.07	0.012	Lipid transfer protein	2
CarTin_tx_s317_comp185938_c0_seq1	108	0.07	0.012	No hits found	3
CarTin_tx_s317_comp28184_c0_seq1	772	0.03	0.012	Uncharacterised/hypothetical protein	4
CarTin_tx_s317_comp826687_c0_seq1	42	-0.09	0.012	Carotenoid oxygenase	4
CarTin_tx_s317_comp355653_c0_seq1	39	-0.11	0.012	Hydroxymethylglutaryl-CoA reductase	2
CarTin_tx_s317_comp665796_c0_seq1	31	0.14	0.012	Glutaredoxin	2
CarTin_tx_s317_comp21320_c0_seq1	3665	0.05	0.013	No hits found	4
CarTin_tx_s317_comp123834_c0_seq1	66	0.15	0.014	Zinc finger, RAN binding protein	2
CarTin_tx_s317_comp15252_c0_seq1	96	-0.08	0.014	Uncharacterised/hypothetical protein	3
CarTin_tx_s317_comp29736_c0_seq1	273	-0.06	0.014	Uncharacterised/hypothetical protein	3
CarTin_tx_s317_comp31514_c0_seq2	20	0.15	0.014	Uncharacterised/hypothetical protein	4
CarTin_tx_s317_comp34718_c1_seq1	3404	-0.05	0.014	Plant lipid transfer protein/proline rich protein	4
CarTin_tx_s317_comp411243_c0_seq1	263	-0.10	0.014	Transcription repressor OFP6-like	2
CarTin_tx_s317_comp1528262_c0_seq1	45	0.10	0.014	No hits found	4
CarTin_tx_s317_comp23058_c0_seq3	25	0.15	0.014	Alcohol dehydrogenase/2-alkenal reductase	2
CarTin_tx_s317_comp18241_c1_seq1	233	0.10	0.015	No hits found	3
CarTin_tx_s317_comp2816374_c0_seq1	53	-0.24	0.015	Polyprotein/peroxidase 64	4
CarTin_tx_s317_comp5019_c0_seq1	25	-0.12	0.015	Uncharacterised/hypothetical protein	4
CarTin_tx_s317_comp5087_c0_seq1	209	0.07	0.015	Uncharacterised/hypothetical protein	4
CarTin_tx_s317_comp5321_c0_seq1	2080	0.28	0.016	Mannose-binding lectin	4
CarTin_tx_s317_comp77834_c0_seq1	1071	0.09	0.020	Lactuca sativa O-glucosyl transferase 1	4

TABLE A.1: Significantly ($\alpha = 0.05$) differentially expressed spring safflower transcriptomic contigs from Experiment 2 (continued).

Contig	Mean Counts	Fold Change (log2)	Adjusted p-value	BLASTP Homologue	Category
CarTin_tx_s317_comp11506_c0_seq1	49	0.09	0.021	Uncharacterised/hypothetical protein	3
CarTin_tx_s317_comp182733_c0_seq1	17	0.13	0.021	Alcohol dehydrogenase/2-alkenal reductase	4
CarTin_tx_s317_comp6677_c0_seq1	51	0.08	0.021	CASP-like protein	2
CarTin_tx_s317_comp32761_c0_seq1	**21**	**-0.18**	**0.021**	**Flowering Locus T (FT)-Like**	**1**
CarTin_tx_s317_comp33309_c0_seq11	41	-0.09	0.021	Uncharacterised/hypothetical protein	4
CarTin_tx_s317_comp72407_c0_seq1	62	0.17	0.025	No hits found	3
CarTin_tx_s317_comp870612_c0_seq1	26	0.11	0.025	*Zinc finger, RING/FYVE/PHD-type*	2
CarTin_tx_s317_comp144284_c0_seq1	92	-0.17	0.026	Hypothetical protein/transposase	2
CarTin_tx_s317_comp749263_c0_seq1	143	0.11	0.030	Auxin responsive SAUR protein	4
CarTin_tx_s317_comp32337_c0_seq1	66	-0.08	0.032	Terpene/germacrene A synthase	2
CarTin_tx_s317_comp39512_c0_seq1	50	0.10	0.032	Late embryogenesis abundant protein, LEA-18	4
CarTin_tx_s317_comp1420929_c0_seq1	140	-0.10	0.034	Retrotransposable element/peroxidase 64	4
CarTin_tx_s317_comp31683_c1_seq19	117	0.05	0.037	NAD-dependent epimerase/dehydratase	2
CarTin_tx_s317_comp19470_c0_seq1	76	-0.08	0.038	No hits found	4
CarTin_tx_s317_comp5028_c0_seq1	237	-0.06	0.038	AP2/ERF domain-containing protein	4
CarTin_tx_s317_comp1208157_c0_seq1	16	0.11	0.040	No hits found	3
CarTin_tx_s317_comp1513234_c0_seq1	270	0.26	0.040	No hits found	4
CarTin_tx_s317_comp21975_c0_seq2	80	-0.07	0.040	Uncharacterised/Hypothetical Protein	4
CarTin_tx_s317_comp22584_c0_seq1	374	0.08	0.040	AP2/ERF domain-containing protein	2
CarTin_tx_s317_comp251834_c0_seq1	143	0.06	0.040	Strigalactone esterase DAD2	4
CarTin_tx_s317_comp30776_c0_seq2	40	-0.08	0.040	Uncharacterised/Hypothetical Protein	4

TABLE A.1: Significantly ($\alpha = 0.05$) differentially expressed spring safflower transcriptomic contigs from Experiment 2 (continued).

Contig	Mean Counts	Fold Change (log2)	Adjusted p-value	BLASTP Homologue	Category
CarTin_tx_s317_comp310214_c0_seq1	105	0.10	0.040	Auxin responsive SAUR protein	4
CarTin_tx_s317_comp32216_c0_seq1	1235	0.11	0.040	*Zinc finger, RING/FYVE/PHD-type*	2
CarTin_tx_s317_comp32337_c0_seq2	316	-0.08	0.040	Terpene/germacrene A synthase	2
CarTin_tx_s317_comp33670_c1_seq44	68	-0.10	0.040	Hydroxymethylglutaryl-CoA reductase	2
CarTin_tx_s317_comp528341_c0_seq1	71	-0.07	0.040	*Homeodomain-like/MYB-related transcription factor*	2
CarTin_tx_s317_comp6680_c0_seq1	51	-0.08	0.040	kinesin-like protein	4
CarTin_tx_s317_comp7178_c0_seq1	4651	0.07	0.040	Uncharacterised/hypothetical protein	4
CarTin_tx_s317_comp81387_c0_seq1	163	0.08	0.040	Uncharacterised/hypothetical protein	3
CarTin_tx_s317_comp1627019_c0_seq1	22	-0.11	0.041	No hits found	4
CarTin_tx_s317_comp323532_c0_seq1	64	-0.07	0.043	Uncharacterised/hypothetical protein	3
CarTin_tx_s317_comp4835_c0_seq2	188	-0.07	0.044	Miraculin	2
CarTin_tx_s317_comp33541_c1_seq3	196	-0.08	0.044	No hits found	4
CarTin_tx_s317_comp92660_c0_seq1	679	-0.06	0.046	Uncharacterised/hypothetical protein	4

Appendix B

Differential Expression Plots

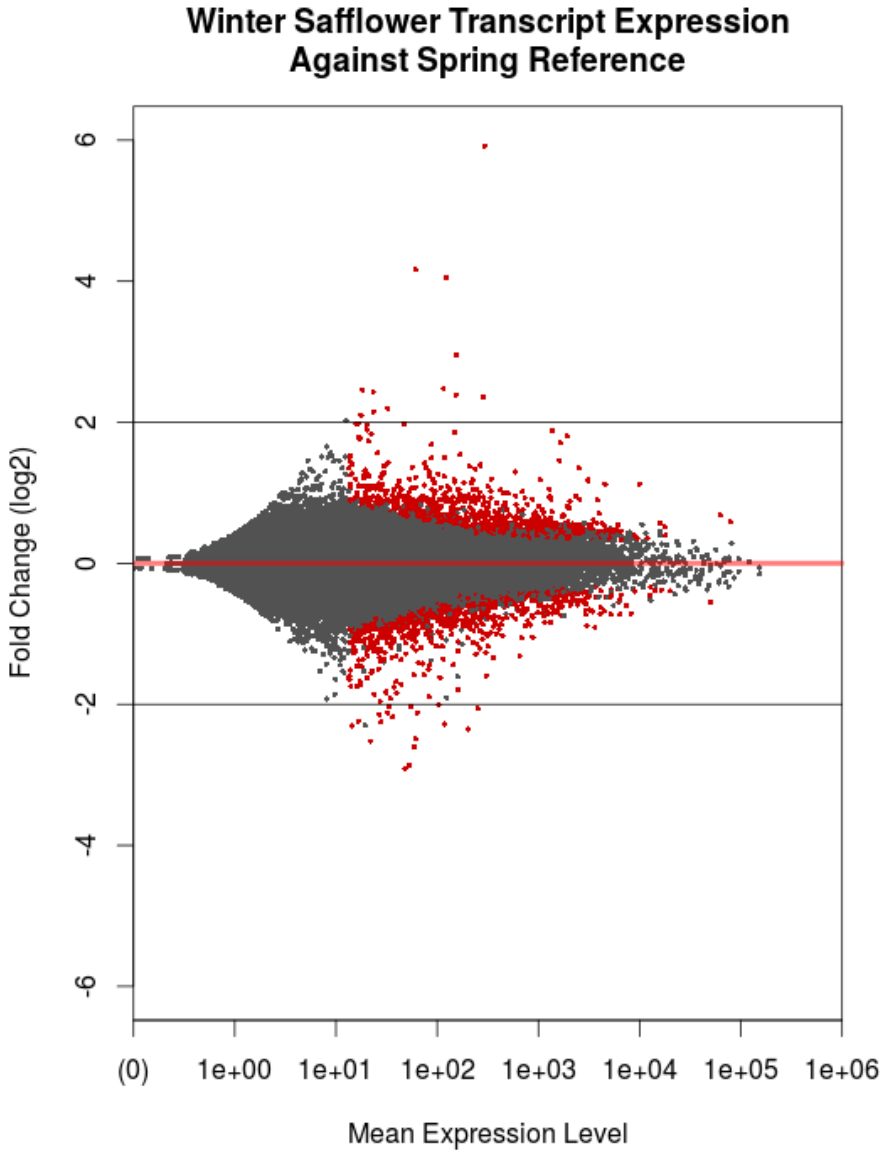


FIGURE B.1: Mean expression counts of every contig in the spring safflower *de novo* transcriptome, vernalised and unvernalised winter safflower. Significantly differentially expressed transcripts ($\alpha = 0.01$) are indicated in red, horizontal lines are cut-offs of absolute two-fold difference in expression between vernalised and non-vernalised transcripts.

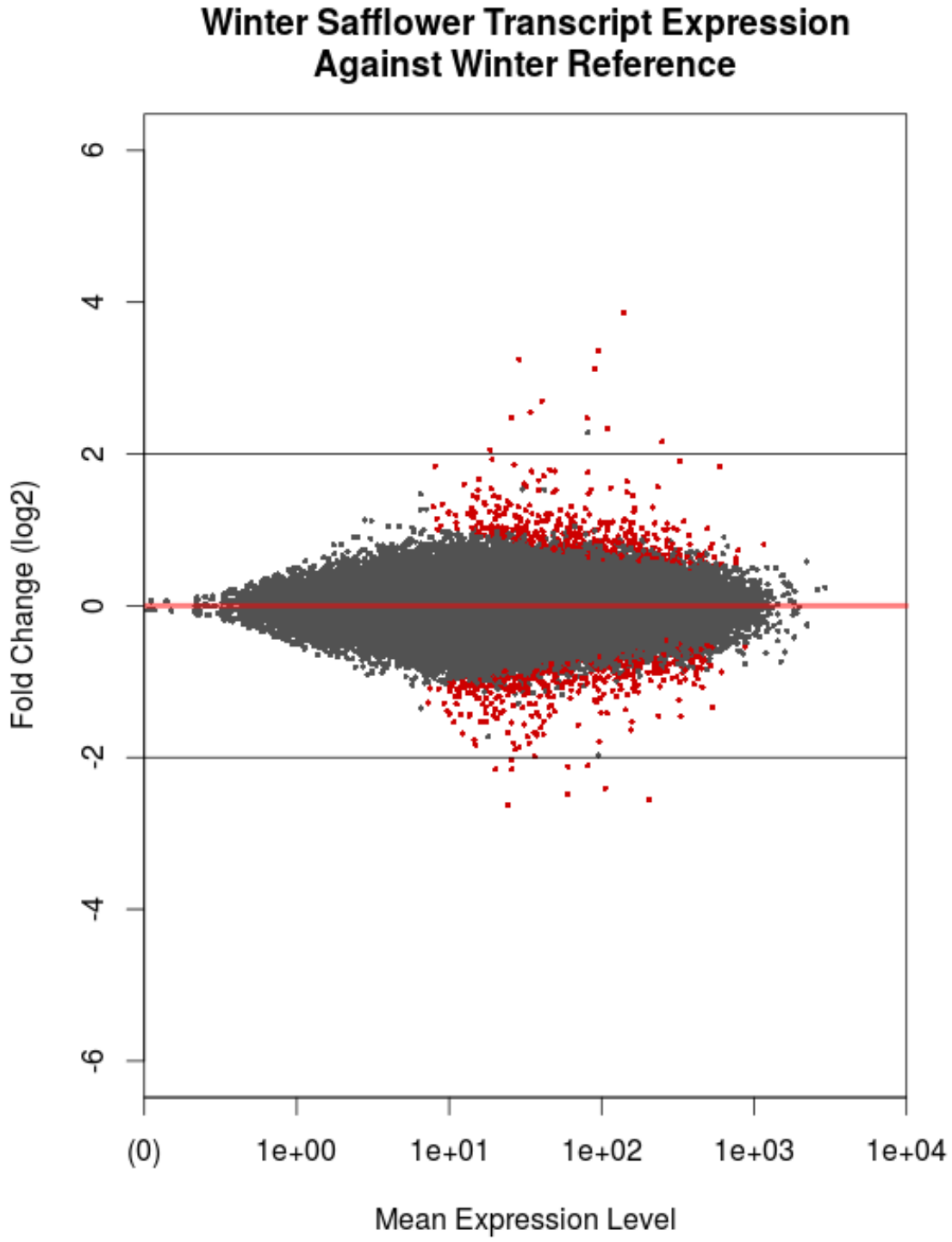


FIGURE B.2: Mean expression counts of every contig in the winter safflower *de novo* transcriptome, vernalised and unvernalised winter safflower. Significantly differentially expressed transcripts ($\alpha = 0.01$) are indicated in red, horizontal lines are cut-offs of absolute two-fold difference in expression between vernalised and non-vernalised transcripts.

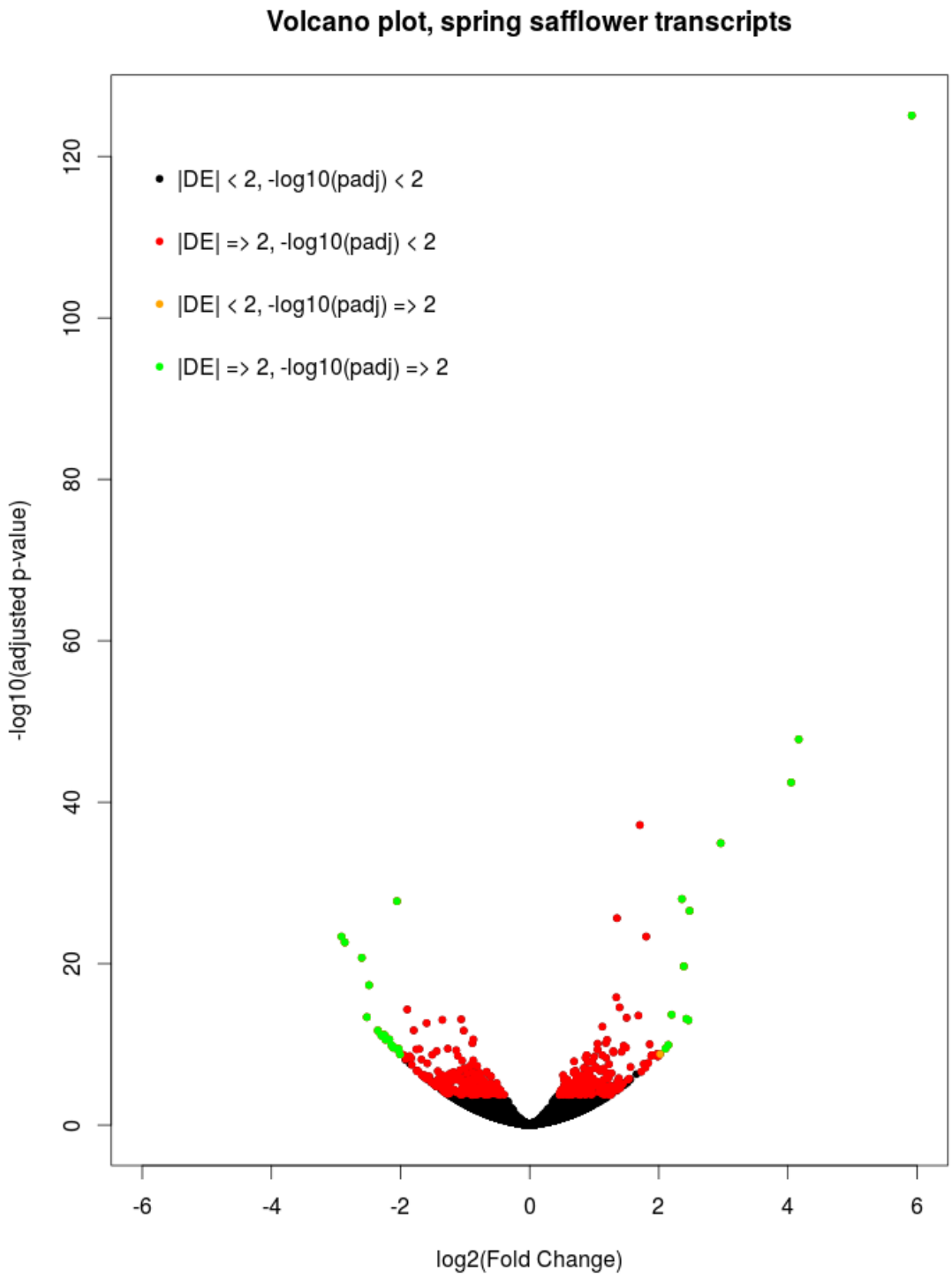


FIGURE B.3: Volcano Plot of the log2 fold change of spring safflower transcripts plotted against the -log10 of the adjusted p-value. Reads are from vernalised and unvernalsied winter safflower (Experiment 1) and aligned against the spring safflower transcriptome. Transcripts with an adjusted p-value =< 0.01 and absolute(log2(FoldChange) > 2 are indicated in green).

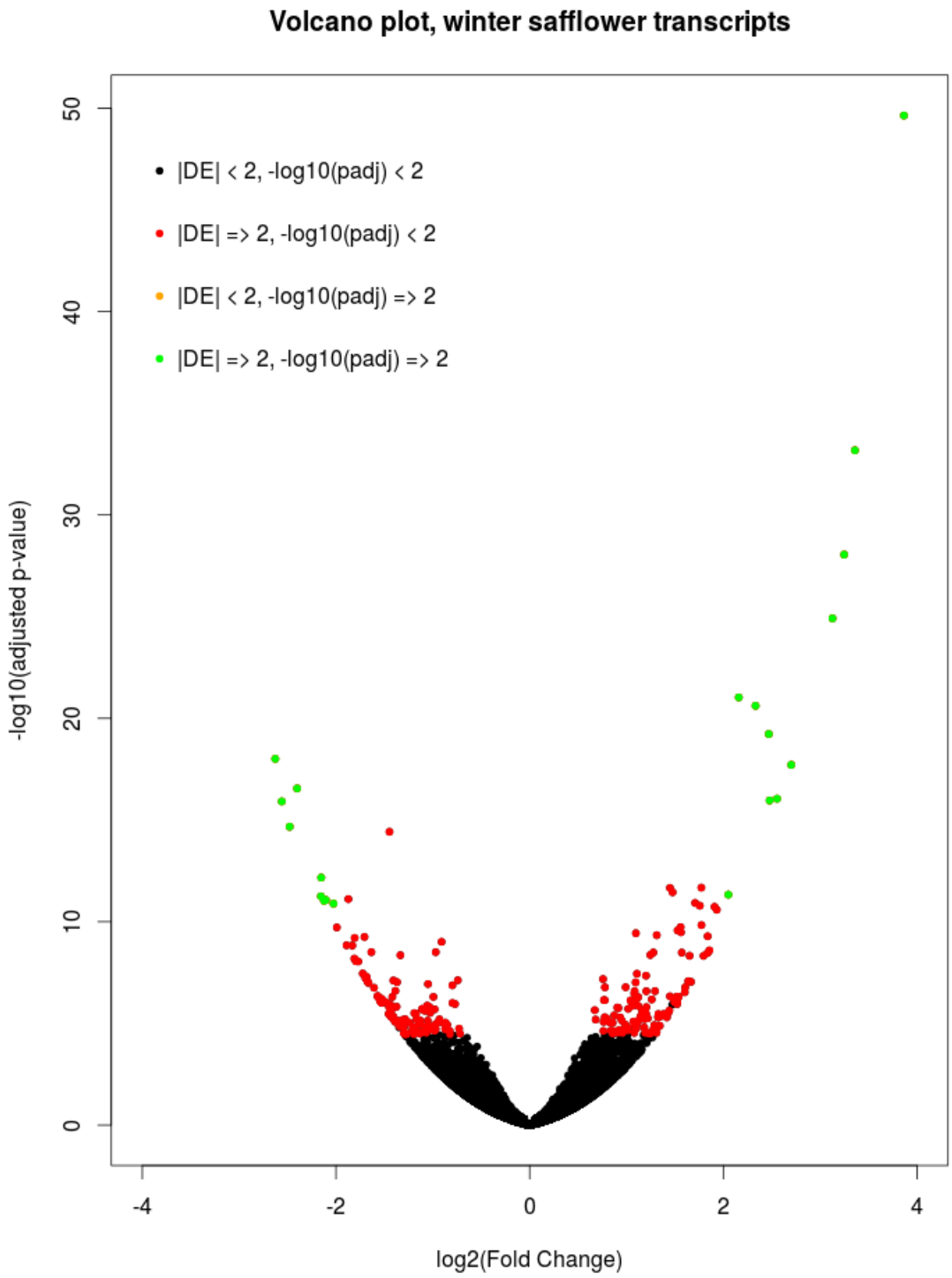


FIGURE B.4: Volcano Plot of the log2 fold change of winter transcripts plotted against the -log10 of the adjusted p-value. Reads are from vernalised and unvernalsied winter safflower (Experiment 1) and aligned against the winter safflower transcriptome. Transcripts with an adjusted p-value =< 0.01 and absolute(log2(FoldChange) > 2 are indicated in green).

Appendix C

Characterised Differentially Expressed Vernalisation Transcripts

C.1 Spring Safflower Vernalisation Transcripts

>CarTin_tx_s317_comp26769_c0_seq1 | CtAP1-like
GAGGATAGAGAACAAAATCAACCGTCAGGTGACTTTCTCCAAACGGAGGTCCG
GCTTGTTGAAGAAAGCTCACGAGATCTCAGTCCTCTGCGATGCAGATGTCGGCC
TCATCATCTTTTCCACTAAAGGAAAGCTCTGCGAGTACGCCACCGATTCTGCAA
TGGAAAGGATTCTTGAGAGGTACGAAAGATACTCTTATGCAGAAATGCAGCTT
ACATCAACCCATAACGAATCACAAGGAAGCTGGACTCTGGAACATGCTAAGCT
GAAAGCTAGAATTGAGATTTTGCAGAAAACTCAAAGGCATTTAATGGGAGAAG
AACTTGACTCATTGAGTCTGAAAGAGCTTCAAAATCTGGAGCAGCAGCTTGATA
ATGCTCTTAAACACATTAGGTTGAAAAAGAATCAAGTGATGATGGAATCAATTT
CCCAGCTCCAAAAGAAGGACAAGGAATTGCTGGATCAAAACAACCTCCTGTCT
AAGCAGATCAAGGAAATGGAGAAAGAAATAGGAGAGCATGATTTAGAGCATC
AAAGCAATGATATTATGGCCTCATGTCAGTTGGGCATCTGTGATGCATACAACG
GGGCCCAGGCGGGAGGAGCAGATGGGGAAGTAGAACAAAACCCGAGACAAG
GTCAAGGTCAAGCATCGACATCAGGGATGCCCGCTTGGATGATTCAACACATG
AACAAGTGAAGTGGCAAGCATATATATATCTTAATGTGTTGTTAATTAAGCGTT
TGGAATGTTGTTGTTACGTGTATAAACGAGAGACATCTATAGCTTTGATCGACC
AACCAGCCATCGATCATGTTTAATGGGTTGCGGGATCAGAAAATAGATGTTAGA
GTAGAAATGTGTACCAAAGCTTATGTGTTCTTTTTGTTGTTGTTGTTGTTGTGGTG
GTGGTGTTTGTGTATATGATGAGGTTCTTTAAAATGGATTTATAGAGATTTCAGA
GTGTTAAATTGGTTGTGAATATGGTGCGTGAATAAAATGTGTAG

>CarTin_tx_s317_comp33367_c7_seq4 | CtMADS1
TTATCCTTTTTATTTATTTCTTTATCTATATCTTTATCTTTATTAAAAAGTCATCAA
GATTATTGTTTAGTTACATTACATCTTCAAATGAAATTAACAAAAAAAATATAT
ATGAAGGGGGAGAAAAAGAAAAGAACCTGGAGAGAGAGGAGGACATGTATTT
AGTTACACAGGCTGCCACGTAAGCTTAATCACATGACACATGGCATGATAGTGT
GAAACGTGGCAAGCAACCTTCTTCTTCTTCTTCTTGCTCCTTGATTACCAGGAAA
TATATTAACTATTAACTGTTAATTTATTTGTATCATTATTATTATTATACATACAA
CAAGACAGTGGAATACTCACACTGCTTAATAACAGAGGATGTAAAGAAGAAA
GCATTCATTCATTACACAATAAAAAACATTCAGCCTCAACCATCATCAGCATTG
CTCTGATTCAGCCTTGCTGCCATTATCTCTCCTACAATTAACTGCTTCTCTTGGCT

CAGCTGCTTTTGCTTCTCTTGCAGGGTCTTCACAGCTCCCAGCATCAACTGTTTCT
TTCTGGTTCTGACTTGTTGGAGAACACCATCGATCTCTTTCTCTAACTGAGTAAG
CCCAGCGATATCCGACCCTTGAATCTTGTGGTCTTCAAGGTGCCGTTGGATCATC
CGCGTTAACTCGTCAGCACTCAATACCTCCCTATACTCCAACGCAAGCTTCTCA
CGTACGGTTCTGTGGACAACTTCTTCTGCATTCTTGTAGTCCTGGTAGCGAGCGA
GGATCCTGGTCATGCTTTCACCGCTGGAGAACTCCGAGAGCCTGTTACTTCCGG
AGAAGATGAAGAGAGCGACGTCAACCTCGCAGAGTACCGACAGCTCGGGAGC
CTTCTTCATGATTCCTTTCCGGCGTTTGGAGAAGGAGACTTGACGGCTGCATTTG
TCTTCGATCCGTTTCAGTTCGACTTTGCCTCTTCCCATCCAAAACCCTAATTGAA
AAAGCTGGTTGGGTTGGTTTACAGAAAGATGCGACTGGGAGAGAGAGAGAGA
GAACAGACTAAACCGGACGTTGCTGCCACGTGGCCCTCGGACTGGGCCCATTTG
AGTCATCCAAGGACTGGGCCGGGCTGCTGGATTAGTGTGGATCAGGGCTGGCTT
TTGTACCGAGGGTTTGTCAAATTATCTGTAATTTAATTTTATTACAATCAAATCA
AATAATCAATTGATCAAATTAGAAATTGAACATCTTACTTCTATTTAAAGAATTT
A A A A T G T T T A A A A A T G C A T T T T T T T G G T G A G C G A G

>CarTin_tx_s317_comp32761_c0_seq1 | CtFT-like
TCCATATTATTATGTTTTTGTATTGTTTCTTATTCCAAATTAACTATGTCTTACAAC
TCTATGATACCTCACACCTAGTGTTCTTATTCCTTAATAACCATTCTAATTAAGTT
AAGAAGGGTTTATATATAGTATGACTTGCATGCCGTCGATCAACATGCGTTTAT
AATACTTGCATGCCGTCGACACGCCTTTGTCTTATCTCCGTCGTCCACCAAAGCC
ACTTTCTCGCTGGCAGTTGAAGTAGACGGCAGCCACCGGGGAACCAAGGTTGT
ACGCTTCCGCAAAGTCTTTTGTGTTGAAGTTCTGGCGCCATCCTGGAGCGTAGAC
GGTTTGTCGACCCAACTGTCGGAACAACACGAAAACCATACGATGAATTCCCA
TTGATGGCCTTGGACTCTCATAGCACACCACTTCTTGACCAAAACGTGTTCCTGT
GGTCGCTGGAATATCAGTCACCAACCAATGTAAATATTCCCTAAGGTTAGGATC
ACTAGGACTAGGAGCATCAGGATTCACCATGACTAAAGTATGAAAGGCACGAA
GGTCGTCACCTCCGATATCAACCCTAGGTTGGCTTACAACCTGAGAGGGCCTTA
ACTCACATCCATTGCTGATTTCCATATCATCATACGATACAGTAAGGTTAATTGA
CCTGCTAAAGTTATCAAGAACATCTCCGATCACTCGTCCAACAACCAACGGTTC
CCTCTCCCTGGGCATCACACAAAAAAAGGTAATTCCCTCAATTTCTTGAGGGGG
T T T C T T T C T T T C T T T C T T

>CarTin_tx_s317_comp33519_c0_seq70 | CtVRN1-like
CCAAACCTAAACAGGCAAAGGAAGAGGTGGGTGTGGGACTGTGGCTGAGCCG
GTGGACCTTTTAAAGGACACGTGGCACCCCTCATCAAAAACATGCCCAAGCTA
AATAACCCCTCCCCGGTCCGGATTACAACGTCCTCGTCGATTCGTCTCCACCTTA
CACTTTTTCTGAAACCCTTCGCCACCGCCGCCCAACCGCCTCGATTCCGGCAGA
ATAAAAAGCGATGGGGAGAGGGAAGGTCGAACTGAAGCGGATCGAAGACAAG
AGCAGCCGGCAGGTCTCCTTCTCCAAGCGCCGGAACGGACTGATGAAGAAGAC
TCACGAGCTGGCGGTGCTGTGCGACGTCGATGTCGCTCTTTTCATCTTCTCCGGC
CGAGGAAGGCTCTACGAGTACTCCACCGGTGAAAGCATGGTTGAGCTTCTCACC

CGCTACCAAAGCCATAAGAAGACAGAGGAACTTGCGCTCGTAAGTCTACACAA
GCAGAAGCTTGATTCTGAAAATGTGTGTACCGCTGATGAGCTGACACGGATGAT
CAAATGGCACCTCGAAGAGAACAATATTAAACTGCTAGATACAACCGGTCTCA
ATCAGCTGGAGCAGCAACTGGATAGCCTTCTCCACCAAGTCAGAACCAGGAAG
AGACAGAAAATGGTGGGAGTTGTGAAGGCCCTACAAGAGGAGGAAATGCAAC
TGAAGAAAGAAAGAAACTTTATGTTGAAGGAGATTATGGCAGCAAGGTCGGAT
GGGATCCATGATGGCTCTGGTGATCCCTCACAGCCACAACTTCCGCCAATGGAC
GCTCAGATGATTATCTGGTGATGAGTGTTTTTTGAAAGAAAAGCAATTCTCGTAT
CTTAGCACTCTGCGAGCTAGTGTTACCACCAATATATATTTATCTATATAGCAGC
AGCTTCGTTCATATTCCATATTCCGGAGAAACCCTTTGCAGGCAATGTCAAGGT
GACAATAGTCCTCCTAGTAGACATGTTTCACCCTGTTATCTTCCCCTGTCTAATA
AGTTAATCCATCGCTGTTTATACGTTTCTAGCAGTGAACCCTACGTCCTTCCGAT
CTCCATCGATCTCCCTAAATCGATGAACAACTAGCCCCGTTGTAAAATATTTGT
ATGGAGCATAATTAATTTTCCATGTATGTGAGTATGTGCTTGTGTCAACCATATC
ATACATCATGTTTCCAGCTTCATTGCAACAAAATATGCATATAATACAGACG

C.2 Winter Safflower Vernalisation Transcripts

>CarTin_tx_WSRC03_Scaff32547 | CtAP1-like
ACCAATTTAACACTCTGAAATCTCTATAAATCCATTTTAAAGAACCTCATCATAT
ACACAAACACAACAACAACAACAAAAAGAACACATAAGCTTTGGTACAC
ATCTCTACTCTAACATCTATTTTCTGATCCCGCAACCCATTAAACATGATCGATG
GCTGGTTGGTCGATCAAAGCTATAGATGTCTCTCGTTTATACACATAACAACAA
CATTCCAAACGCTTAATTAACAACAAATTAAGATATATATGCTTGCCACTTCAC
TTGTTCATGTGTTGAATCATCCAAGCGGGCATCCCTGATGTCGATGCTTGAGATG
TCGATGCTTGACCTTGACCTTGTCTCGGGTTTTGTTCTACTTCCCCATCTGCTCCT
CCCGCCTGGGCCCCGTTGTATGCATCACAGATGCCCAACTGACATGAGGCCATA
ATATCATTGCTTTGATGCTCTAAATCATGCTCTCCTATTTCTTTCTCCATTTCCTTG
ATCTGCTTAGACAGGAGGTTGTTTTGATCCAGCAATTCCTTGTCCTTCTTTTGGA
GCTGGGAAATTGATTCCATCATCACTTGATTCTTTTTCAACCTAATGTTTTTAAGA
GCATTATCAAGCTGCTGCTCCAGATTTTGAAGCTCTTTCAGACTCAATGAGTCAA
GTTCTTCTCCCATTAAATGCCTTTGAGTTTTCTGCAAAATCTCAATTCTAGCTTTC
AGCTTAGCATGTTCCAGAGTCCAGCTTCCTTGTGATTCGTTATGGGTTGATGTAA
GCTGCATTTCTGCATAAGAGTATCTTTCGTACCTCTCAAGAATCCTTTCCATTCTG
TAGGCAGAATCGGTGGCGTACTCGCAGAGCTTTCCTTTAGTGGAAAAGATGATG
AGGCCGACATCTGCATCGCAGAGGACTGAGATCTCGTGAGCTTT

>CarTin_tx_WSRC03_Scaff20021 | CtMADS1
AAGACAGTGGAATACTCACACTGCTTAATAACAGAGGATGTAAAGAAGAAAG
CATTCATTCATTACACAATAAAAAACATTCAGCCTCAACCATCATCAGCATTGC
TCTGATTCAGCCTTGCTGCCATTATCTCTCCTACAATTAACTGCTTCTCTTGGCTC
AGCTGCTTTTGCTTCTCTTGCAGGGTCTTCACAGCTCCCAGCATCAACTGTTTCTT

TCTGGTTCTGACTTGTTGGAGAACACCATCGATCTCTTTCTCTAACTGAGTAAGC
CCAGCGATATCCGACCCTTGAATCTTGTGGTCTTCAAGGTGCCGTTGGATCATCC
GCGTTAACTCGTCAGCACTCAATACCTCCCTATACTCCAACGCAAGCTTCTCAC
GTACGGTTCTGTGGACAACTTCTTCTGCATTCTTGTAGTCCTGGTAGCGAGCGAG
GATCCTGGTCATGCTTTCACCGCTGGAGAACTCCGAGAGCCTGTTACTTCCGGA
GAAGATGAAGAGAGCGACGTCAACCTCGCAGAGTACCGACAGCTCGGGAGCC
TTCTTCATGATTCCTTTCCGGCGTTTGGAGAAGGAGACTTGACGGCTGCATTTGT
CTTCGATCCGTTTCAGTTCGACTTTGCCTCTTCCCATCCAAAACCCTAATTGAAA
A A G C T G G T A T T G A A A A A G C T G G T

>CarTin_tx_WSRC03_Scaff23886 | CtMADS1
TTCTTCTTGCTCCTTGATTACCAGGAAATATATTAACTATTAACTGTTAATTTATT
TGTATCATTATTATTATTATACATACAACAAGACAGTGGAATACTCACACTGCTT
AATAACAGAGGATGTAAAGAAGAAAGCATTCATTCATTACACAATAAAAAAC
ATTCAGCCTCAACCATCATCAGCATTGCTCTGATTCAGCCTTGCTGCCATTATCT
CTCCTACAATTAACTGCTTCTCTTGGCTCAGCTGCTTTTGCTTCTCTTGCAGGGTC
TTCACAGCTCCCAGCATCAACTGTTTCTTTCTGGTTCTGACTTGTTGGAGAACAC
CATCGATCTCTTTCTCTAACTGAGTAAGCCCAGCGATATCCGACCCTTGAATCTT
GTGGTCTTCAAGGTGCCGTTGGATCATCCGCGTTAACTCGTCAGCACTCAATAC
CTCCCTATACTCCAACGCAAGCTCACGTACGGTTCTGTGGACAACTTCTTCTGCA
TTCTTGTAGTCCTGGTAGCGAGCGAGGATCCTGGTCATGCTTTCACCGCTGGAG
AACTCCGAGAGCCTGTTACTTCCGGAGAAGATGAAGAGAGCGACGTCAACCTC
GCAGAGTACCGACAGCTCGGGAGCCTTCTTCATGATTCCTTTCCGGCGTTTGGA
G A A G G A G A C T T G A C G G C T G C A T T T G T C T T C G A T C C

>CarTin_tx_WSRC03_Scaff57705 | CtFT-like
GTTTTTGTATTGTTTCTTATTCCAAATTAACTATGTCTTACAACTCTATGATACCT
CACACCTAGTGTTCTTATTCCTTAATAACCATTCTAATTAAGTTAAGAAGGGTTT
ATATATAGTATGACTTGCATGCCGTCGATCAACATGCGTTTATAATACTTGCATG
CCGTCGACACGCCTTTGTCTTATCTCCGTCGTCCACCAAAGCCACTTTCTCGCTG
GCAGTTGAAGTAGACGGCAGCCACCGGGGAACCAAGGTTGTACGCTTCCGCAA
AGTCTTTTGTGTTGAAGTTCTGGCGCCATCCTGGAGCGTAGACGGTTTGTCGACC
CAACTGTCGGAACAACACGAAAACCATACGATGAATTCCGATTGATGGCCTTG
GACTCTCATAGCACACCACTTCTTGACCAAAACGTGTTCCTGTGGTCGCTGGAA
TATCAGTCACCAACCAATGTAAATATTCCCTAAGGTTAGGATCACTAGGACTAG
GAGCATCAGGATTCACCATGACTAAAGTATGAAAGGCACGAAGGTCGTCACCT
CCGATATCAACCCTAGGTTGGCTTACAACCTGAGAGGGCCTTAACTCACATCCA
TTGCTGATTTCCATATCATCATACGATACAGTAAGGTTAATTGACCTGGTAAAGT
TATCAAGAACATCTCCGATCACTCGTCCAACAACCAACGGTTCCCTCTCCCTGG
GCATCACACAAAAAAAGGTAATTCCCTCAATTTCTTGAGGGGGTTTCTTTCTTTC
T T T

>CarTin_tx_WSRC03_Scaff43593 | CtVRN1-like
GCACATACTCACATACATGGAAAATTAATTATGCTCCATACAAATATTTTACAA
CGGGGCTAGTTGTTCATCGATTTAGGGAGATCGATGGAGATCGAAAGGACGTA
GGGTTCACTGCTAGAAACGTATAAACAGCGATGGATTAACTTATTAGACAGGG
GAAGATAACAGGGTGAAACATGTCTACTAGGAGGACTATTGTCACCTTGACATT
GCCTGCAAAGGGTTTCTCCGGAATATGGAATATGAACGAAGCTGCTGCTATATA
GATAAATATATATTGGTGGTAACACTAGCTCGCAGAGTGCTAAGATACGAGAA
TTGCTTTTCTTTCAAAAAACACTCATCACCAGATAATCATCTGAGCGTCCATTGG
CGGAAGTTGTGGCTGTGAGGGATCACCAGAGCCATCATGGATCCCATCCGACCT
TGCTGCCATAATCTCCTTCAACATAAAGTTTCTTTCTTTCTTCAGTTGCATTTCCT
CCTCTTGTAGGGCCTTCACAACTCCCACCATTTTCTGTCTCTTCCTGGTTCTGACT
TGGTGGAGAAGGCTATCCAGTTGCTGCTCCAGCTGATTGAGACCGGTTGTATCT
AGCAGTTTAATATTGTTCTCTTCGAGGTGCCATTTGATCATCCGTGTCAGCTCAT
CAGCGGTACACACATTTTCAGAATCAAGCTTCTGCTTGTGTAGACTTACGAGCG
CAAGTTCCTCTGTCTTCTTATGGCTTTGGTAGCGGGTGAGAAGCTCAACCATGCT
TTCACCGGTGGAGTACTCGTAGAGCCTTCCTCGGCCGGAGAAGATGAAAAGAG
CGACATCGACGTCGCACAGCACCGCCAGCTCGTGAGTCTTCTTCATCAGTCCGT
TCCGGCGCTTGGAGAAGGAGACCTGCCGGCTGCTCTTGTCTTCGATCCGCTTCA
GTTCGACCTTCCCTCTCCCCATCGCTTTTTATTCTGCCGGAATCGAGGCGGTTGG
TGGGCGGCGGTGGCGAAGGGTTTCAGAAAAAGTGTAAGGTGGAGACGAATCG
ACGAGG

Appendix D

Winter safflower transcripts

>CarTin_tx_WSRC03_Scaff61146
AACCACTTGTTGGCGGCTTAGCCCTCCAGTGGTTGCTGGAGAGCCAGATGGAGG
TCGACAGGCAGGTTGCGAGTCGGTGGCACAACCTTTGAGCTTGAAGTTGCTGTA
CCTACCAACAAAAGGTTGATACTCGTAATTTGCTTTGTATCTACCTTTTTCAGTC
GCCCATGAGGATGCATCCCATATTGATCCGTACATGTACATGGGTCTTAAAGGA
AATGTGGTGTCGCTTTTTCTCGGGTACCTTCTAATTGGCACATCGTCCACAAAAA
ATATAATCTCTTTGGGTGTCCATAACATCGCATATTGGTGAAAATCCTCGGTGGG
ATCGAACCAAAGGTGGAACTGTTGTTCTCTTCCTATGATGTTTCCGTCACCAC
>CarTin_tx_WSRC03_Scaff62404
ACTTCTAGAAAAGCAGGATTTTTATTAACACAAGCAGAAACACAGGAGATGCG
ACGAAAAGACGACGAAACGAAATGCACTATGCAGGATGTGACGAAAAGTCGA
CAAAACGGACAACGTACGGTACTAGCTAGTAGCTAGATGCAGACAGACAACTT
AATTATTACAGACACTTAAAACTACATACCACCCCCATATGCAAACATCAAACT
AAAACCAAACACTCCATCATTCAGCTCTATTTCTACTTTGTGGTGTACTTGGATT
CGTCGATCGGAATCCCTTCCTCCTCCGCCCGTTCACCGCCGGACTTGTCGGTGGT
GCTCAGCCCGCCCTTCTTCCCCATCTCCTGGTACCCTTCCGTCCCAAGCTGATCC
TTTCTCGTCTGTCCTCCACGGCTTCGACCTTCAGCGAGGCGTGCCTGGGCGTCGA
GACTTTTGCCACGGGTACCACCGGGAATGACGGTCTCGCCTTGAGCCGCTCGTT
GGTCGAGTTCCCTTTTCTCCTCCTCCGAGATCTGCTCCTGTGGCCTCCTTGATTGT
TGCTGCTGCGATGCCATTTTCTTCACACTTTCCCTTATTTTGTAATTAACTGAATG
TAAAGCTTGATTCTTTGC
>CarTin_tx_WSRC03_Scaff65369
CAATCTCCATTTTACACGACAAAAGAAATTTAACACATAAATAATCATGGAGC
ATTCTCAAGACGATGACGATTAATTTGTTGTAATAGTAGTCGTAAAGCAAACAA
TCACATACTTGCACGAAACAAATTCTCTAAGGAACCGTTGGTTTTGGAATCTCA
GGGACCGTGGGCTTTGGAATTTCTGGAACAATGGTTTTGGGAATTTCTGGAACT
GTTGGCTTTGGAATCTCAGGAAGAGTGGGCTTTGGAATCTCGGGTACCATGGGC
TTCGGAATCTCTGGAACAATGGGTTTGGGAATTTCTGGAAGTGTTGGCTTCGGA
ATCTCGGGTACCGTGGGCTTCGGGATCTCTGGAACAATGGGTTTGGGAATCTCG
GGTAGTGTGGGCTTCGGAATCTCGGGTACCGTGGGCTTTGGAATCTCTGGAACA

Appendix E

Multiple Sequence Alignments of Annotated Safflower Transcripts

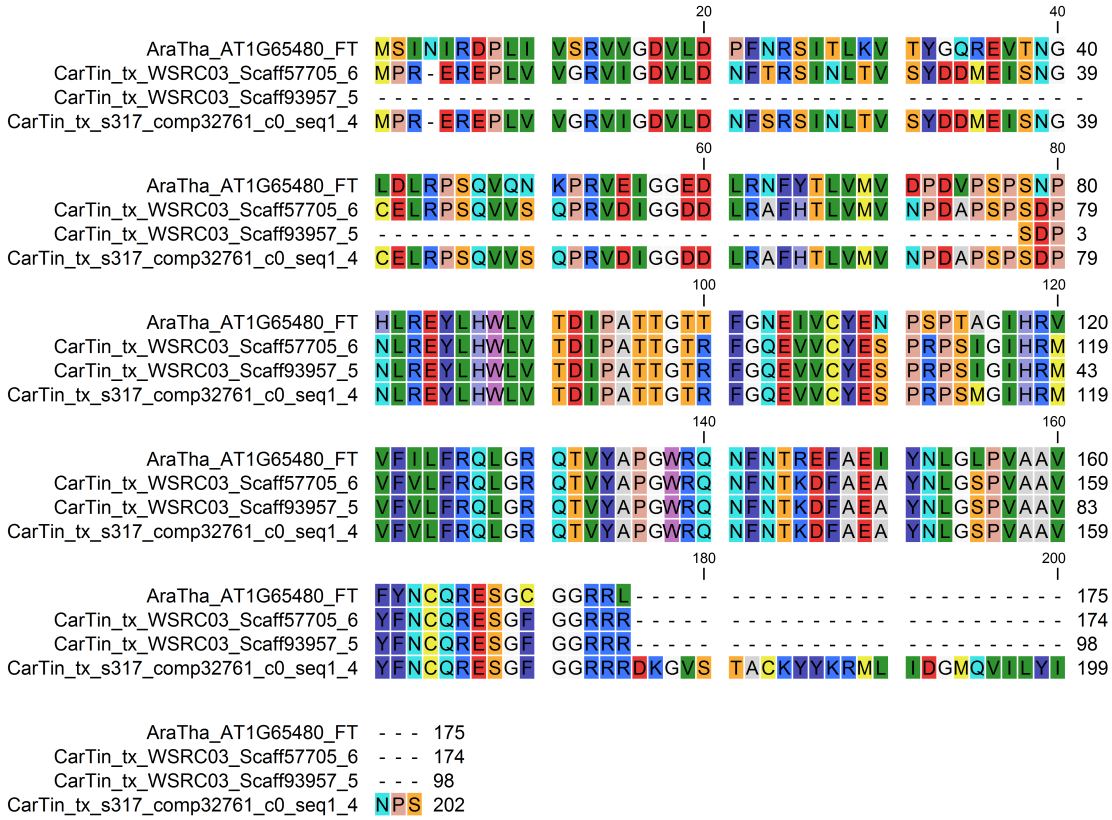


FIGURE E.1: Multiple Sequence Alignment of amino acid sequences with homology to *Ct*FT-LIKE. Multiple sequence alignments were generated with T-Coffee. CarTin_tx_s317_comp32761_c0_seq1 originated from spring safflower, and CarTin_tx_WSRC03_Scaff57705 and CarTin_tx_WSRC03_Scaff93957 originate from winter safflower.

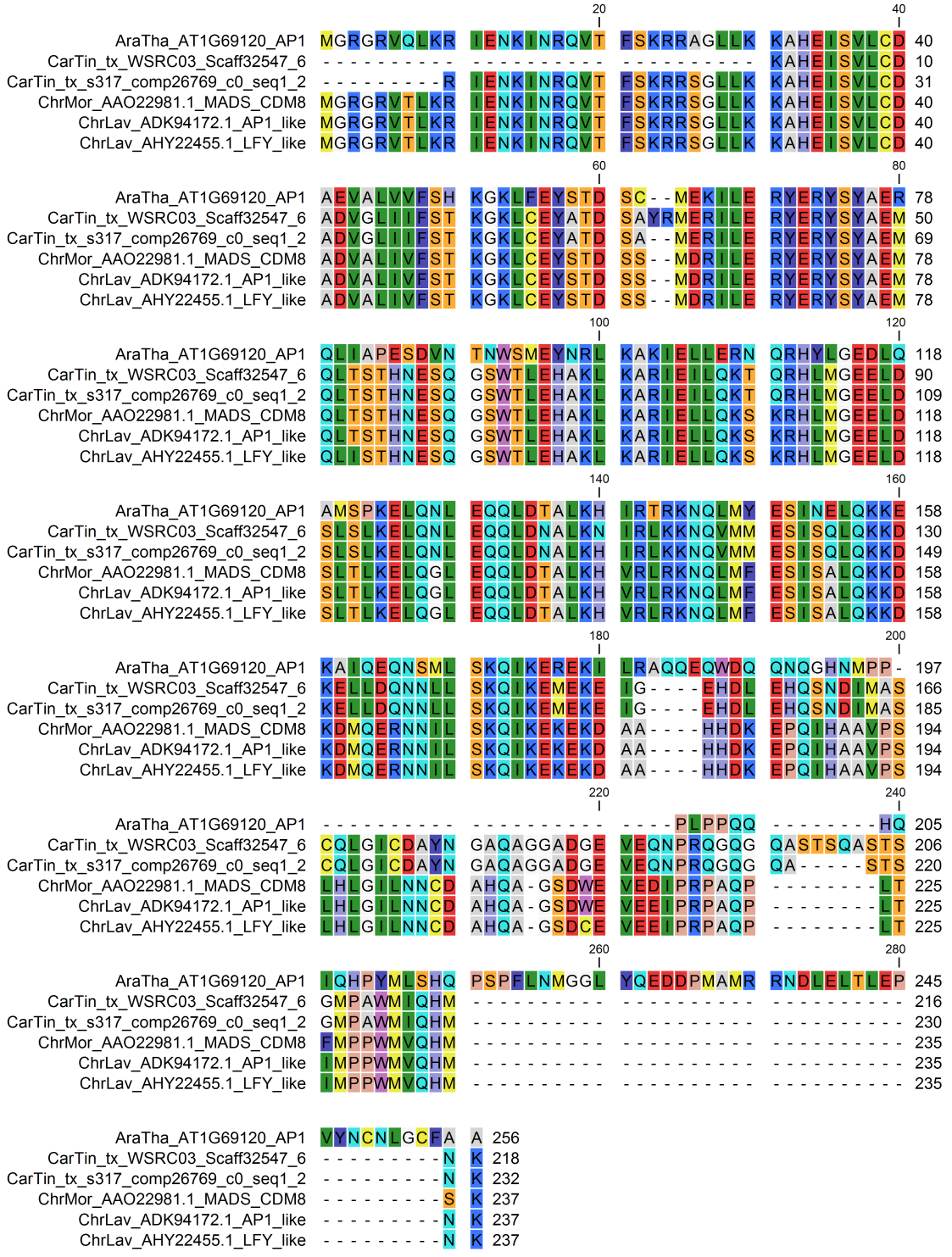


FIGURE E.2: Multiple Sequence Alignment of amino acid sequences with homology *Ct*AP1-LIKE. Multiple sequence alignments were generated with T-Coffee. CarTin_tx_s317_comp26769_c0_seq1 and CarTin_tx_WSRC03_Scaff32547, originate from spring and winter safflower *de novo*transcriptomes, respectively. *Chrysanthemum*

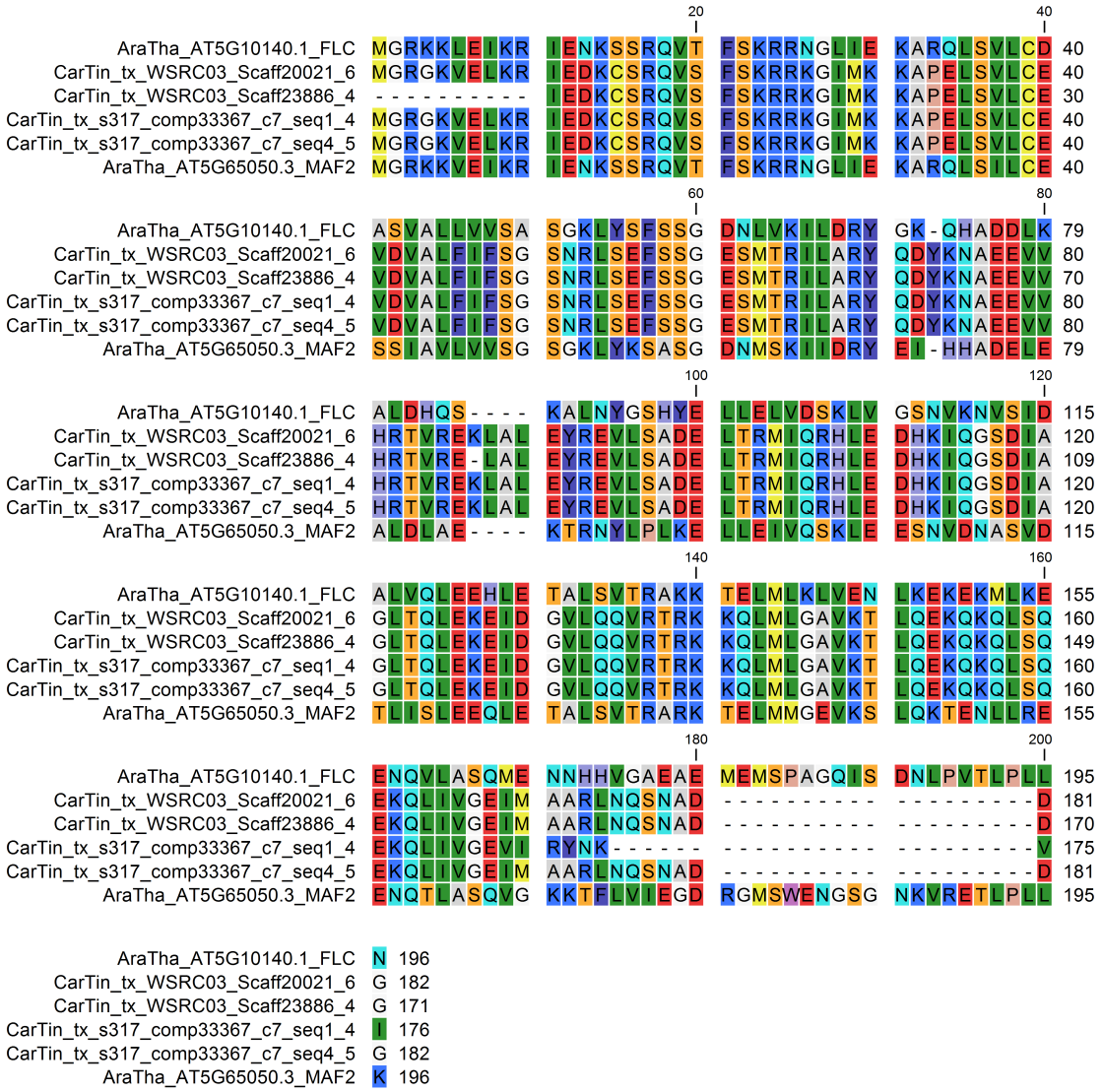


FIGURE E.3: Multiple Sequence Alignment of amino acid sequences with homology *Ct*MADS1 and other MADS-box containing sequences. Multiple sequence alignments were generated with T-Coffee. CarTin_tx_s317_comp33367_c7_seq4 and CarTin_tx_s317_comp33367_c7_seq1 originate from spring safflower, and CarTin_tx_WSRC03_Scaff20021 and CarTin_tx_WSRC03_Scaff23886 originate from winter safflower.

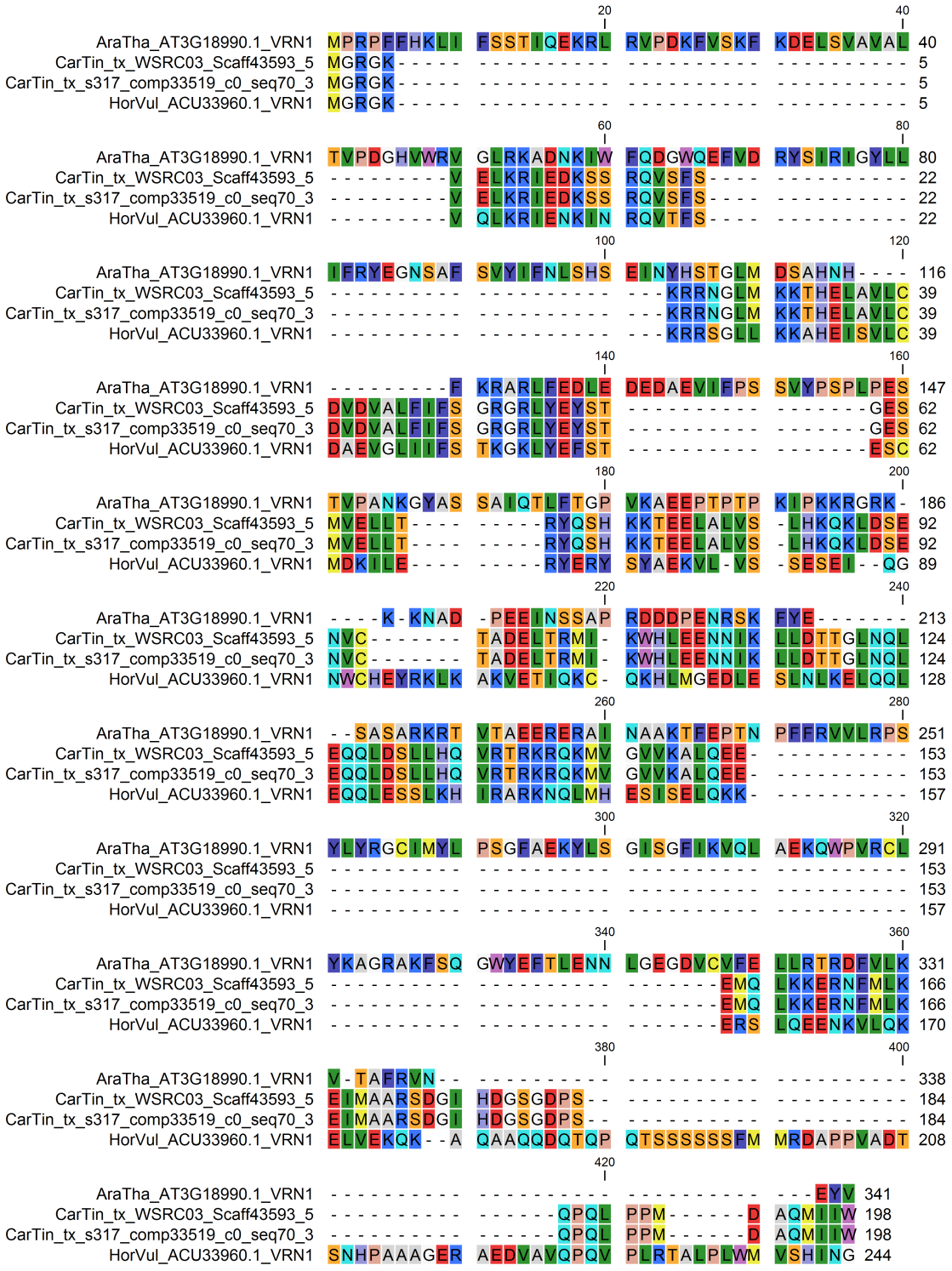


FIGURE E.4: Multiple Sequence Alignment of amino acid sequences with homology to *Ct*VRN1-LIKE. Multiple sequence alignments were generated with T-Coffee. CarTin_tx_s317_comp33519_c0_seq70 originated from spring safflower, and CarTin_tx_WSRC03_Scaff43593 originated from winter safflower.

Appendix F

PCR Primers

TABLE F.1: PCR Primers used in RT-qPCR experiments.

Gene	Primer	Direction	Sequence
CtActin-LIKE	Actin-sun-ACT1-s1	Forward	5′-ACCACAGGTATTGTGCTGGATTC-3′
	Actin-sun-ACT1-a1	Reverse	5′-CACCAATTGTGATGACTTGTCCAT-3′
CtAP1-LIKE	qCtAP1f1	Forward	5′-AGGAAATGGAGAAAGAAATAGGAG-3′
	qCtAP1r1	Reverse	5′-GTTTTGTTCTACTTCCCCATCTG-3′
CtMADS1	qCtFLCf1	Forward	5′-GCATTCTTGTAGTCCTGGTAGCGA-3′
	qCtFLCr1	Reverse	5′-ACAAATGCAGCCGTCAAGTCTC-3′
	qCtFLCf2	Forward	5′-CTCCAACAAGTCAGAACCAGAAAG-3′
	qCtFLCr2	Reverse	5′-CAGCCTCAACCATCATCAGC-3′
CtFT-LIKE	qCtFTf1	Forward	5′-GACTAAAGTATGAAAGGCACGAAG-3′
	qCtFTr1	Reverse	5′-TTTTGTGTGATGCCCAGGGGAGAG-3′
	qCtFTf2	Forward	5′-AGCGACCACAGGAACACG-3′
	qCtFTr2	Reverse	5′-GGAACAAACACGAAAACCATACG-3′
CtVRN1-LIKE	qCtVRN1f1	Forward	5′-GTCCTCGTCGATTCGTCTCCAC-3′
	qCtVRN1r1	Reverse	5′-CTCTTGTCTTCGATCCGCTTCAG-3′

Appendix G

Read Alignments

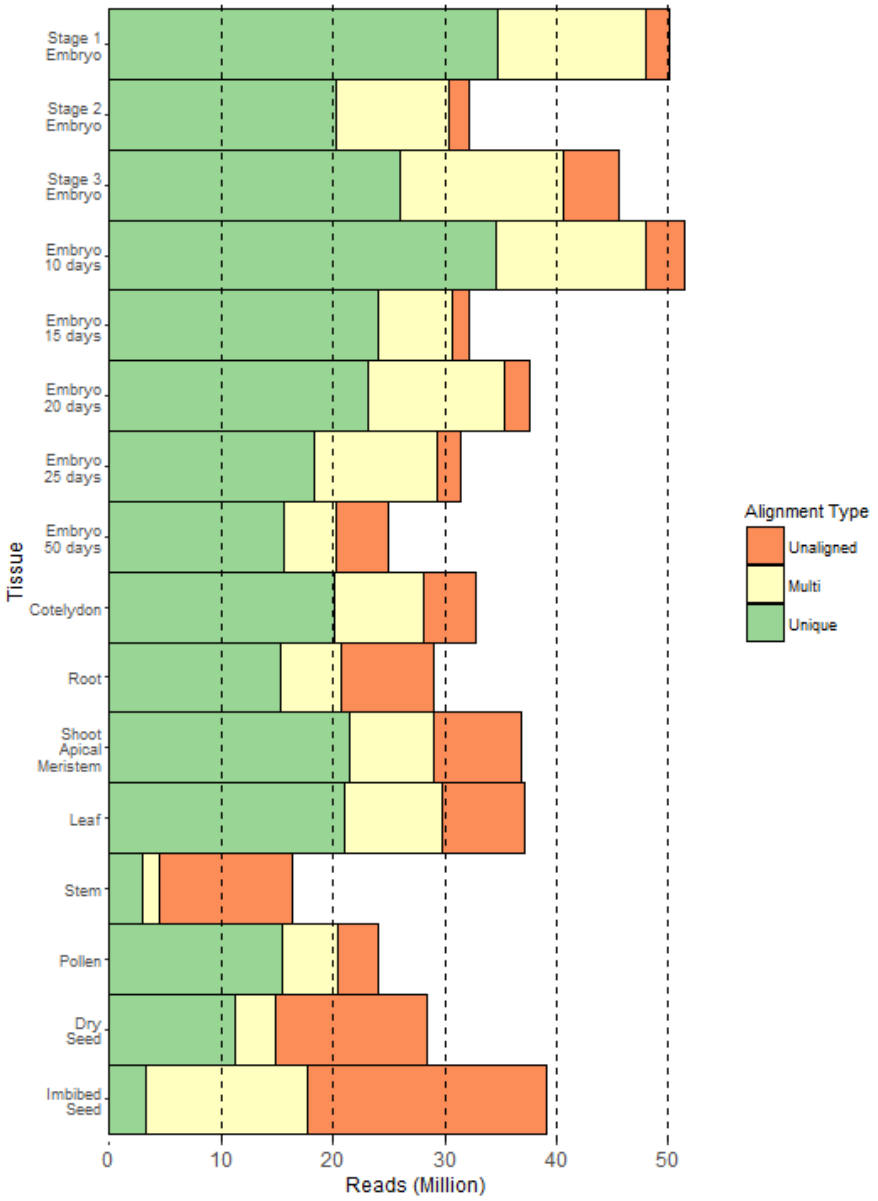


FIGURE G.1: Back alignment of short reads generated from spring safflower tissues.

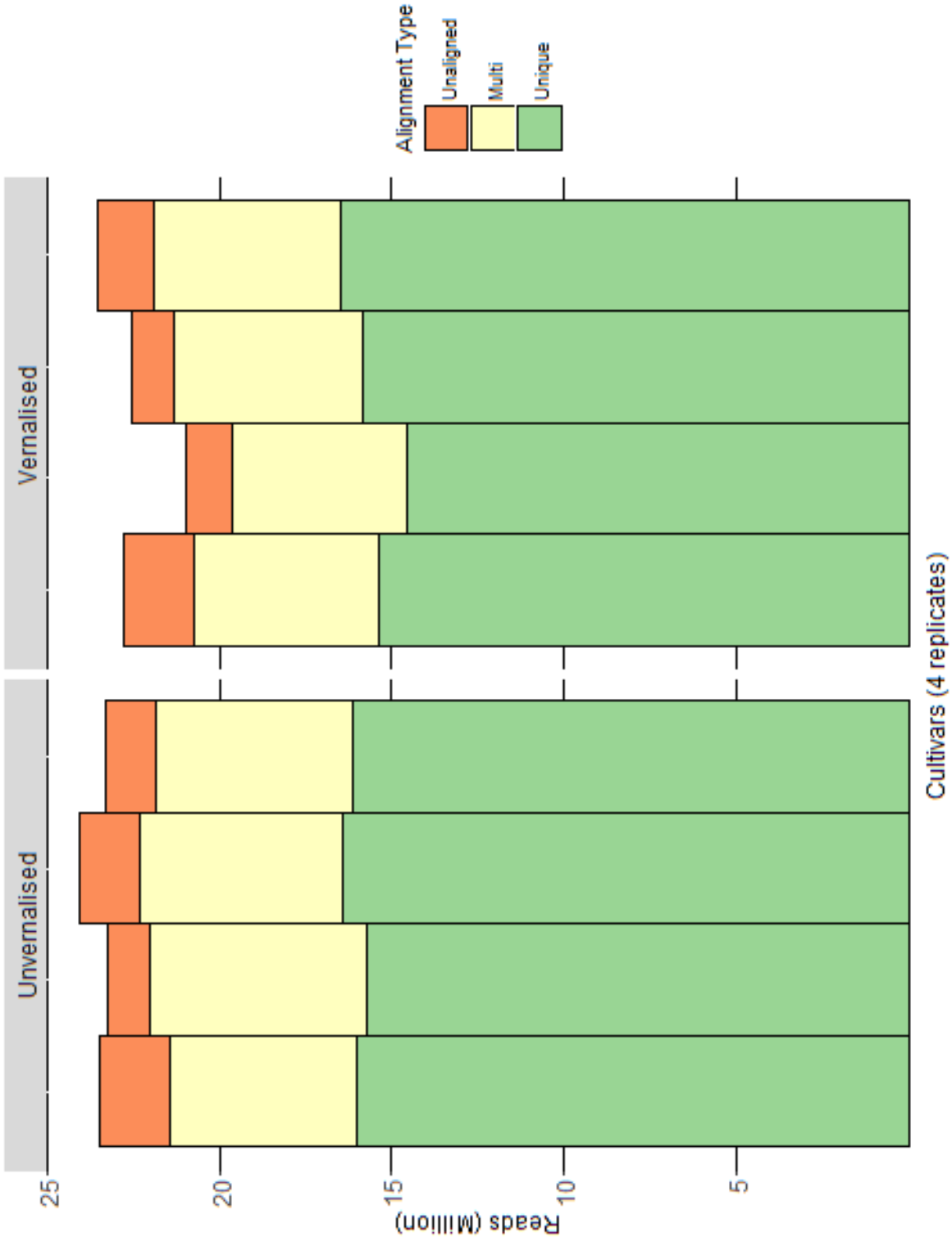


FIGURE G.2: Reads generated for each replicate for vernalised and unvernalised winter safflower aligned against the de novo transcriptome.

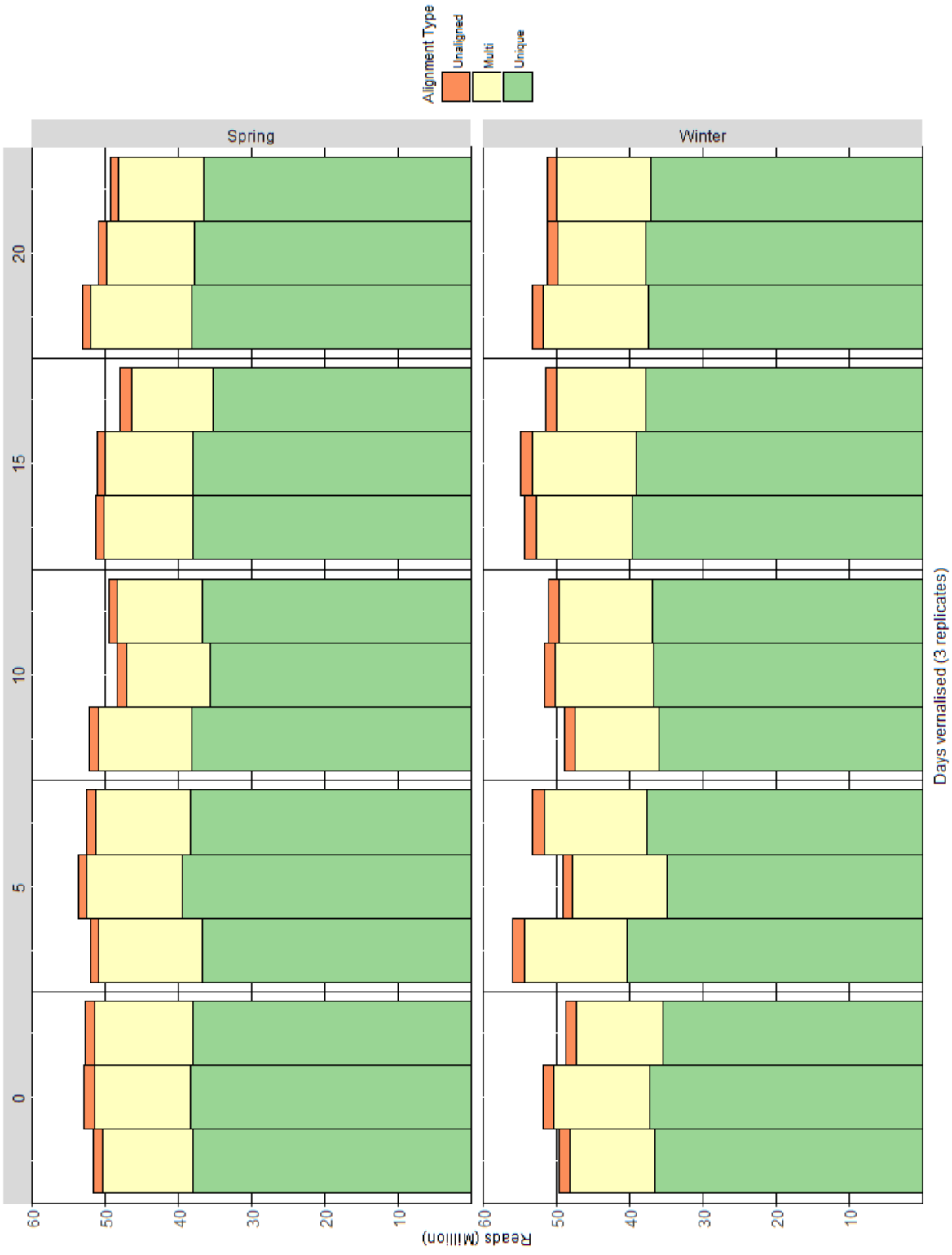


FIGURE G.3: Reads generated for each replicate for winter and spring safflower as time in vernalisation conditions is extended. Reads were aligned against the *de novo* safflower transcriptome.

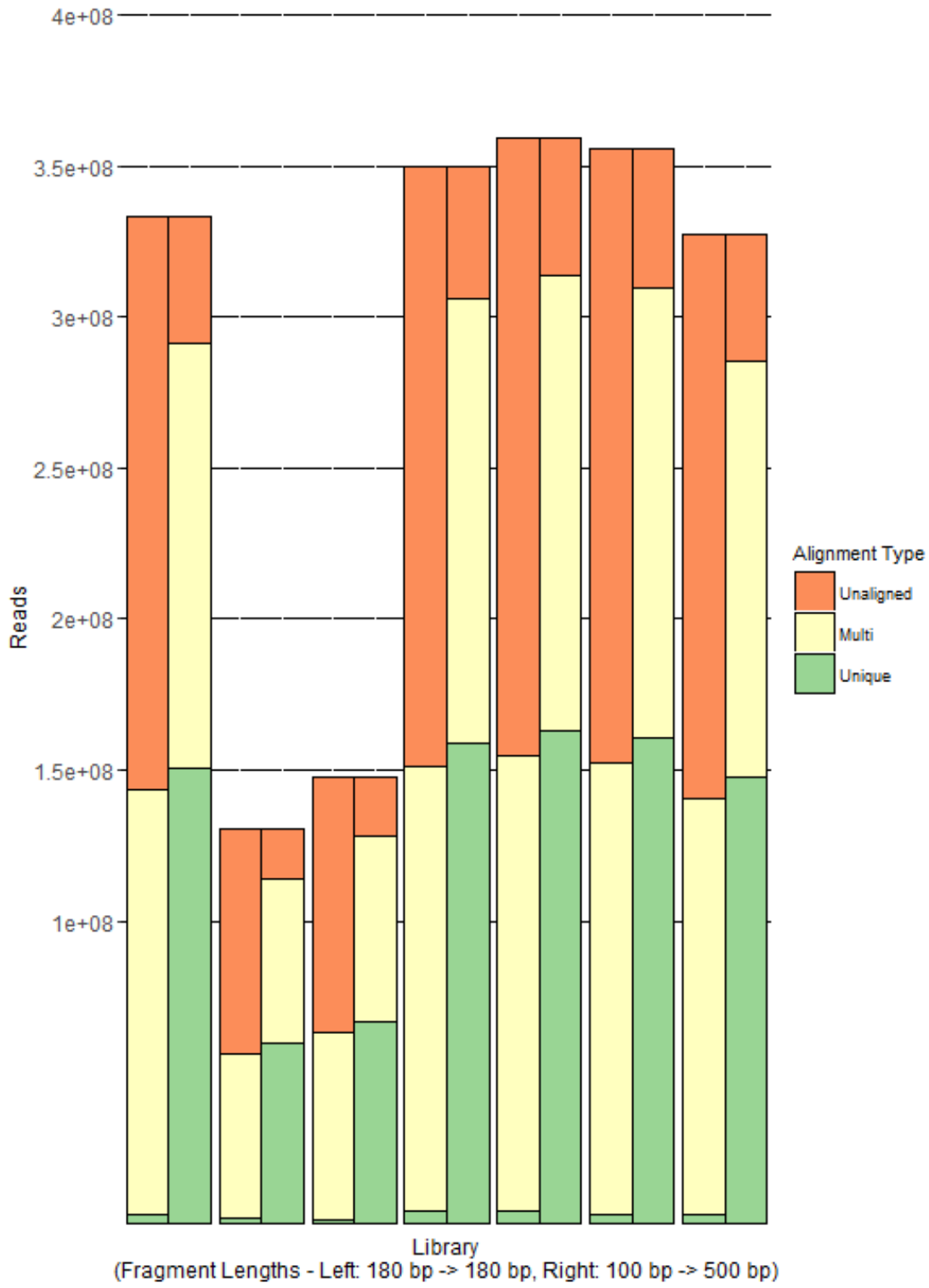


FIGURE G.4: Back alignments of each unfiltered Illumina genomic library against the *de novo* spring safflower genome.

Appendix H

Spring:Winter Segregation Ratios for Crossing Population

TABLE H.1: The segregation ratios of the F_3 population of crossed safflower plants. Crosses in **bold** were used to further investigate the segregation ratios. Crosses in ***bold italic*** were used to generate the genetic markers.

Cross	Alive	Elongated	% Elongated	Approx Ratio (S:W)
Winter-05	21	0	0.00	00:16
Winter-06	24	0	0.00	00:16
Winter-09	22	0	0.00	00:16
Winter-12	24	0	0.00	00:16
X028	22	0	0.00	00:16
X029	22	0	0.00	00:16
X055	12	0	0.00	00:16
X076	15	0	0.00	00:16
X079	17	0	0.00	00:16
X087	12	0	0.00	00:16
X101	11	0	0.00	00:16
X102	0	0	0.00	00:16
X148	18	0	0.00	00:16
X178	23	0	0.00	00:16
X214	0	0	0.00	00:16
X262	22	0	0.00	00:16
X295	22	0	0.00	00:16
X296	7	0	0.00	00:16
X310	5	0	0.00	00:16
X397	5	0	0.00	00:16
X052	12	1	0.08	01:15
X140	12	1	0.08	01:15
X011	16	2	0.13	02:14
X062	11	2	0.18	03:13
X306	20	4	0.20	03:13
X384	15	3	0.20	03:13
X007	15	4	0.27	04:12

TABLE H.1: The segregation ratios of the F_3 population of crossed safflower plants. Crosses in **bold** were used to further investigate the segregation ratios. Crosses in ***bold italic*** were used to generate the genetic markers, continued.

Cross	Alive	Elongated	% Elongated	Approx Ratio (S:W)
X026	10	3	0.30	05:11
X036	9	3	0.33	05:11
X075	3	1	0.33	05:11
X034	8	3	0.38	06:10
X302	5	2	0.40	06:10
X357	12	5	0.42	06:10 or 07:09
X374	7	3	0.43	07:09
X084	9	4	0.44	07:09
X366	15	7	0.47	07:09
X031	12	6	0.50	08:08
X043	16	8	0.50	08:08
X088	2	1	0.50	08:08
X136	10	5	0.50	08:08
X185	12	6	0.50	08:08
X407	4	2	0.50	08:08
X391	13	7	0.54	09:07
X168	9	5	0.56	09:07
X002	5	3	0.60	10:06
X119	5	3	0.60	10:06
X147	5	3	0.60	10:06
X354	18	11	0.61	10:06
X393	18	11	0.61	10:06
X137	21	13	0.62	10:06
X016	16	10	0.63	10:06
X049	8	5	0.63	10:06
X242	8	5	0.63	10:06
X228	11	7	0.64	10:06
X022	3	2	0.67	10:06
X099	15	10	0.67	10:06
X349	15	10	0.67	10:06
X350	3	2	0.67	10:06
X083	16	11	0.69	11:05
X314	10	7	0.70	11:05
X017	7	5	0.71	11:05
X246	7	5	0.71	11:05
X030	20	15	0.75	12:04
X044	8	6	0.75	12:04

TABLE H.1: The segregation ratios of the F_3 population of crossed safflower plants. Crosses in **bold** were used to further investigate the segregation ratios. Crosses in ***bold italic*** were used to generate the genetic markers, continued.

Cross	Alive	Elongated	% Elongated	Approx Ratio (S:W)
X100	12	9	0.75	12:04
X177	8	6	0.75	12:04
X222	12	9	0.75	12:04
X307	8	6	0.75	12:04
X308	16	12	0.75	12:04
X321	4	3	0.75	12:04
X335	12	9	0.75	12:04
X340	4	3	0.75	12:04
X325	13	10	0.77	12:04
X312	18	14	0.78	12:04
X395	9	7	0.78	12:04
X143	14	11	0.79	13:03
X239	14	11	0.79	13:03
X142	19	15	0.79	13:03
X317	19	15	0.79	13:03
X023	10	8	0.80	13:03
X041	10	8	0.80	13:03
X181	10	8	0.80	13:03
X251	5	4	0.80	13:03
X131	17	14	0.82	13:03
X264	6	5	0.83	13:03
X280	12	10	0.83	13:03
X362	20	17	0.85	14:02
X113	14	12	0.86	14:02
X224	14	12	0.86	14:02
X042	16	14	0.88	14:02
X047	16	14	0.88	14:02
X248	8	7	0.88	14:02
X404	16	14	0.88	14:02
X056	9	8	0.89	14:02
X135	18	16	0.89	14:02
X172	9	8	0.89	14:02
X237	18	16	0.89	14:02
X244a	9	8	0.89	14:02
X343	9	8	0.89	14:02
X139	19	17	0.89	14:02
X150	10	9	0.90	14:02

TABLE H.1: The segregation ratios of the F_3 population of crossed safflower plants. Crosses in **bold** were used to further investigate the segregation ratios. Crosses in ***bold italic*** were used to generate the genetic markers, continued.

Cross	Alive	Elongated	% Elongated	Approx Ratio (S:W)
X039	12	11	0.92	15:01
X045	12	11	0.92	15:01
X171	12	11	0.92	15:01
X256	12	11	0.92	15:01
X065	13	12	0.92	15:01
X160	15	14	0.93	15:01
X032	20	19	0.95	15:01
X104	20	19	0.95	15:01
X338	20	19	0.95	15:01
X001	9	9	1.00	16:00
X018	16	16	1.00	16:00
X027	18	18	1.00	16:00
X033	6	6	1.00	16:00
X050	5	5	1.00	16:00
X057	3	3	1.00	16:00
X059	13	13	1.00	16:00
X063	3	3	1.00	16:00
X066	2	2	1.00	16:00
X071	21	21	1.00	16:00
X090	2	2	1.00	16:00
X093	9	9	1.00	16:00
X098	11	11	1.00	16:00
X107	5	5	1.00	16:00
X121	18	18	1.00	16:00
X122	6	6	1.00	16:00
X184	14	14	1.00	16:00
X191	13	13	1.00	16:00
X200	15	15	1.00	16:00
X201	5	5	1.00	16:00
X219	8	8	1.00	16:00
X223	3	3	1.00	16:00
X225	15	15	1.00	16:00
X238	16	16	1.00	16:00
X250	15	15	1.00	16:00
X260	15	15	1.00	16:00
X266	6	6	1.00	16:00
X269	11	11	1.00	16:00

TABLE H.1: The segregation ratios of the F_3 population of crossed safflower plants. Crosses in **bold** were used to further investigate the segregation ratios. Crosses in ***bold italic*** were used to generate the genetic markers, continued.

Cross	Alive	Elongated	% Elongated	Approx Ratio (S:W)
X273	7	7	1.00	16:00
X276	18	18	1.00	16:00
X282	3	3	1.00	16:00
X301	9	9	1.00	16:00
X304	16	16	1.00	16:00
X311	15	15	1.00	16:00
X324	15	15	1.00	16:00
X338	9	9	1.00	16:00
X344	15	15	1.00	16:00
X345	10	10	1.00	16:00
X365	9	9	1.00	16:00
X389	13	13	1.00	16:00
X400	13	13	1.00	16:00
Spring-05	21	21	1.00	16:00
Spring-06	18	18	1.00	16:00
Spring-09	24	24	1.00	16:00
Spring-12	23	23	1.00	16:00

Appendix I

Molecular Markers of the Vernalisation Response in Safflower

TABLE I.1: Illumina genomic contigs, containing digest markers, that align to SNP-containing Bowers contigs.

			Bower SNP details		
DArT Marker	Gx Contig	Bower Contig	Location	SNPs	Chr
15670003	CarTin_gx_s317_Scaff412625	scaffold249746	4.023	28	4
15670051	scaffold_m2223	scaffold121178	8.232	80	8
15670077	scaffold_m10377	scaffold195586	8.226	140	8
15670097	CarTin_gx_s317_Scaff288935	scaffold291664	8.214	17	8
15670104	CarTin_gx_s317_Scaff337648	scaffold205299	8.214	101	8
15670206	CarTin_gx_s317_Scaff252984	scaffold294958	8.226	58	8
15670233	CarTin_gx_s317_Scaff443845	scaffold79875	8.195	81	8
15670331	CarTin_gx_s317_Scaff535465	scaffold204017	8.195	91	8
15670420	CarTin_gx_s317_Scaff208900	scaffold304153	8.232	133	8
15670459	CarTin_gx_s317_Scaff148691	scaffold284638	8.232	137	8
15670753	CarTin_gx_s317_Scaff392143	scaffold294958	8.226	58	8
15671289	scaffold_m2223	scaffold121178	8.232	80	8
15671381	CarTin_gx_s317_Scaff098481	scaffold274501	8.247	18	8
15672254	CarTin_gx_s317_Scaff104607	scaffold32825	8.226	102	8
15672412	CarTin_gx_s317_Scaff080043	scaffold287485	8.214	116	8
15672774	CarTin_gx_s317_Scaff443845	scaffold79875	8.195	81	8
15672806	CarTin_gx_s317_Scaff552779	scaffold287394	8.226	87	8
15672960	CarTin_gx_s317_Scaff535383	scaffold169209	8.226	20	8
15673464	CarTin_gx_s317_Scaff392143	scaffold294958	8.226	58	8
15673754	scaffold_m2013	scaffold292154	8.183	44	8
15673763	CarTin_gx_s317_Scaff133836	scaffold251991	8.176	29	8
15673847	scaffold_m2013	scaffold292154	8.183	44	8
15673852	CarTin_gx_s317_Scaff543512	scaffold181097	8.216	81	8
15673882	CarTin_gx_s317_Scaff315205	scaffold244361	8.193	49	8
15673900	scaffold_m11467	scaffold304463	8.181	46	8
15673963	scaffold_m11467	scaffold304463	8.181	46	8
15674164	CarTin_gx_s317_Scaff099929	scaffold37632	8.247	93	8
15674288	scaffold_m2013	scaffold292154	8.183	44	8

TABLE I.1: Illumina genomic contigs, containing digest markers, that align to SNP-containing Bowers contigs (continued).

			Bower SNP details		
DArT Marker	Gx Contig	Bower Contig	Location	SNPs	Chr
15670042	CarTin_gx_s317_Scaff672541	scaffold298738	-	-	-
15670046	scaffold_j4031_1	scaffold183080	-	-	-
15670074	CarTin_gx_s317_Scaff183601	scaffold224050	-	-	-
15670092	CarTin_gx_s317_Scaff338782	scaffold214128	-	-	-
15670156	CarTin_gx_s317_Scaff772142	scaffold305865	-	-	-
15670156	CarTin_gx_s317_Scaff741309	scaffold305865	-	-	-
15670381	CarTin_gx_s317_Scaff149844	scaffold163347	-	-	-
15670413	CarTin_gx_s317_Scaff441989	scaffold9348	-	-	-
15670488	CarTin_gx_s317_Scaff212876	scaffold307078	-	-	-
15670794	scaffold_m12933	scaffold293246	-	-	-
15671365	CarTin_gx_s317_Scaff108829	scaffold264546	-	-	-
15671492	CarTin_gx_s317_Scaff006660	scaffold56588	-	-	-
15671506	CarTin_gx_s317_Scaff096913	scaffold212415	-	-	-
15671557	scaffold_j4542_1	scaffold261826	-	-	-
15672092	CarTin_gx_s317_Scaff281948	scaffold190633	-	-	-
15672333	CarTin_gx_s317_Scaff298365	scaffold263929	-	-	-
15672795	CarTin_gx_s317_Scaff176345	scaffold236786	-	-	-
15672830	CarTin_gx_s317_Scaff513467	scaffold125300	-	-	-
15672914	CarTin_gx_s317_Scaff110928	scaffold307101	-	-	-
15673174	CarTin_gx_s317_Scaff108829	scaffold264546	-	-	-
15673452	CarTin_gx_s317_Scaff298365	scaffold263929	-	-	-
15673744	CarTin_gx_s317_Scaff177699	scaffold238864	-	-	-
15673829	CarTin_gx_s317_Scaff021050	scaffold137352	-	-	-
15674427	CarTin_gx_s317_Scaff616284	C18745599	-	-	-
15674427	CarTin_gx_s317_Scaff340871	scaffold252481	-	-	-
15671902	CarTin_gx_s317_Scaff301582	(no hits)	-	-	-
15672626	CarTin_gx_s317_Scaff301582	(no hits)	-	-	-

TABLE I.2: Differentially expressed transcripts from Experiment 1 that map to SNP-containing Bowers contigs. Annotated transcripts (*CtMADS1*, *CtAP1-LIKE*, *CtFT-LIKE* and *CtVRN1-LIKE*) are identified in bold, in the order they appear.

		Bower SNP details		
Tx Contig	Bower Contig	Location	SNPs	Chr
CarTin_tx_s317_comp33367_c7_seq4	scaffold301099	1.254	84	1
CarTin_tx_s317_comp67497_c0_seq1	scaffold196782	5.161	65	5
CarTin_tx_s317_comp5504_c0_seq1	scaffold218356	7.282	34	7
CarTin_tx_s317_comp26769_c0_seq1	scaffold174835	9.114	146	9
CarTin_tx_s317_comp29294_c0_seq1	scaffold293503	12.228	102	12
CarTin_tx_s317_comp145452_c0_seq1	scaffold140293	-	-	-
CarTin_tx_s317_comp147113_c0_seq1	scaffold31055	-	-	-
CarTin_tx_s317_comp176356_c0_seq1	scaffold52640	-	-	-
CarTin_tx_s317_comp188549_c0_seq1	scaffold191283	-	-	-
CarTin_tx_s317_comp26440_c0_seq1	scaffold129631	-	-	-
CarTin_tx_s317_comp26483_c0_seq1	scaffold140693	-	-	-
CarTin_tx_s317_comp26483_c0_seq2	scaffold257398	-	-	-
CarTin_tx_s317_comp26765_c0_seq1	scaffold238177	-	-	-
CarTin_tx_s317_comp28573_c0_seq1	C19218436	-	-	-
CarTin_tx_s317_comp31932_c0_seq1	scaffold302811	-	-	-
CarTin_tx_s317_comp32578_c0_seq1	scaffold45682	-	-	-
CarTin_tx_s317_comp32761_c0_seq1	scaffold37410	-	-	-
CarTin_tx_s317_comp33519_c0_seq70	scaffold305810	-	-	-
CarTin_tx_s317_comp34117_c0_seq1	scaffold236174	-	-	-
CarTin_tx_s317_comp34793_c0_seq1	scaffold291987	-	-	-
CarTin_tx_s317_comp366899_c0_seq1	scaffold166622	-	-	-
CarTin_tx_s317_comp44200_c0_seq1	scaffold76747	-	-	-
CarTin_tx_s317_comp46857_c0_seq1	scaffold189677	-	-	-
CarTin_tx_s317_comp4818_c0_seq1	scaffold140685	-	-	-
CarTin_tx_s317_comp487373_c0_seq1	scaffold63248	-	-	-
CarTin_tx_s317_comp541778_c0_seq1	scaffold274008	-	-	-
CarTin_tx_s317_comp561354_c0_seq1	scaffold302160	-	-	-
CarTin_tx_s317_comp7986_c0_seq1	scaffold181735	-	-	-
CarTin_tx_s317_comp4179_c0_seq1	(no hit)	-	-	-
CarTin_tx_s317_comp14924_c0_seq1	(no hit)	-	-	-

TABLE I.3: Differentially expressed transcripts from Experiment 2 that map to SNP-containing Bowers contigs.

		Bower SNP details		
Gx Contig	Bower Contig	Location	SNPs	Chr
CarTin_tx_s317_comp33367_c7_seq4	scaffold301099	1.254	84	1
CarTin_tx_s317_comp39512_c0_seq1	scaffold24595	1.167	86	1
CarTin_tx_s317_comp7178_c0_seq1	C19269295	1.077	18	1
CarTin_tx_s317_comp15252_c0_seq1	scaffold128684	2.011	77	2
CarTin_tx_s317_comp23005_c0_seq1	scaffold5474	2.063	91	2
CarTin_tx_s317_comp33309_c0_seq11	scaffold281533	2.116	128	2
CarTin_tx_s317_comp5019_c0_seq1	scaffold5474	2.063	91	2
CarTin_tx_s317_comp1426363_c0_seq1	scaffold294911	3.081	40	3
CarTin_tx_s317_comp33541_c1_seq3	scaffold305768	3.127	60	3
CarTin_tx_s317_comp31946_c0_seq1	scaffold144329	4.034	38	4
CarTin_tx_s317_comp123834_c0_seq1	scaffold234471	5.261	42	5
CarTin_tx_s317_comp1528262_c0_seq1	scaffold128000	5.141	59	5
CarTin_tx_s317_comp31683_c1_seq19	scaffold301641	6.216	69	6
CarTin_tx_s317_comp21975_c0_seq2	scaffold291039	7.104	18	7
CarTin_tx_s317_comp69290_c0_seq1	scaffold16939	7.017	97	7
CarTin_tx_s317_comp72407_c0_seq1	scaffold16939	7.017	97	7
CarTin_tx_s317_comp1208157_c0_seq1	scaffold267838	8.027	17	8
CarTin_tx_s317_comp144284_c0_seq1	scaffold109973	9.061	71	9
CarTin_tx_s317_comp144284_c0_seq1	scaffold132706	9.061	75	9
CarTin_tx_s317_comp26769_c0_seq1	scaffold174835	9.114	146	9
CarTin_tx_s317_comp11506_c0_seq1	scaffold228490	10.172	5	10
CarTin_tx_s317_comp19470_c0_seq1	scaffold232709	10.337	179	10
CarTin_tx_s317_comp5028_c0_seq1	scaffold94286	10.182	5	10
CarTin_tx_s317_comp80349_c0_seq1	scaffold150047	10.092	81	10
CarTin_tx_s317_comp92660_c0_seq1	scaffold164136	10.243	245	10
CarTin_tx_s317_comp10252_c0_seq1	scaffold45372	11.204	9	11
CarTin_tx_s317_comp20690_c0_seq1	scaffold301512	11.015	85	11
CarTin_tx_s317_comp28184_c0_seq1	scaffold66481	12.248	116	12
CarTin_tx_s317_comp32216_c0_seq1	scaffold199708	12.27	23	12
CarTin_tx_s317_comp32337_c0_seq1	scaffold163741	12.243	126	12
CarTin_tx_s317_comp870612_c0_seq1	scaffold108886	12.27	44	12
CarTin_tx_s317_comp13787_c0_seq1	scaffold303323	-	-	-
CarTin_tx_s317_comp1420929_c0_seq1	scaffold161503	-	-	-
CarTin_tx_s317_comp14932_c0_seq1	C19277643	-	-	-
CarTin_tx_s317_comp14932_c0_seq1	scaffold23410	-	-	-
CarTin_tx_s317_comp1513234_c0_seq1	scaffold166631	-	-	-
CarTin_tx_s317_comp1571506_c0_seq1	scaffold301245	-	-	-

TABLE I.3: Differentially expressed transcripts from Experiment 2 that map to SNP-containing Bowers contigs (continued).

		Bower SNP details		
Gx Contig	Bower Contig	Location	SNPs	Chr
CarTin_tx_s317_comp1578437_c0_seq1	scaffold304021	-	-	-
CarTin_tx_s317_comp1764285_c0_seq1	scaffold304021	-	-	-
CarTin_tx_s317_comp18241_c1_seq1	scaffold137811	-	-	-
CarTin_tx_s317_comp182733_c0_seq1	scaffold265053	-	-	-
CarTin_tx_s317_comp185938_c0_seq1	scaffold241327	-	-	-
CarTin_tx_s317_comp21320_c0_seq1	scaffold63340	-	-	-
CarTin_tx_s317_comp2219162_c0_seq1	scaffold161503	-	-	-
CarTin_tx_s317_comp22584_c0_seq1	scaffold145754	-	-	-
CarTin_tx_s317_comp23058_c0_seq3	scaffold265053	-	-	-
CarTin_tx_s317_comp251834_c0_seq1	scaffold216889	-	-	-
CarTin_tx_s317_comp2816374_c0_seq1	scaffold161503	-	-	-
CarTin_tx_s317_comp29736_c0_seq1	scaffold305426	-	-	-
CarTin_tx_s317_comp30776_c0_seq2	scaffold293651	-	-	-
CarTin_tx_s317_comp310214_c0_seq1	scaffold10544	-	-	-
CarTin_tx_s317_comp31514_c0_seq2	scaffold288256	-	-	-
CarTin_tx_s317_comp32337_c0_seq2	scaffold76446	-	-	-
CarTin_tx_s317_comp323532_c0_seq1	scaffold248293	-	-	-
CarTin_tx_s317_comp32761_c0_seq1	scaffold37410	-	-	-
CarTin_tx_s317_comp33519_c0_seq70	scaffold305810	-	-	-
CarTin_tx_s317_comp33670_c1_seq44	scaffold17398	-	-	-
CarTin_tx_s317_comp34718_c1_seq1	scaffold280846	-	-	-
CarTin_tx_s317_comp355653_c0_seq1	scaffold17398	-	-	-
CarTin_tx_s317_comp360223_c0_seq1	scaffold274537	-	-	-
CarTin_tx_s317_comp411243_c0_seq1	scaffold91647	-	-	-
CarTin_tx_s317_comp4835_c0_seq1	C19352417	-	-	-
CarTin_tx_s317_comp4835_c0_seq2	C19352417	-	-	-
CarTin_tx_s317_comp5087_c0_seq1	C17788474	-	-	-
CarTin_tx_s317_comp528341_c0_seq1	scaffold260330	-	-	-
CarTin_tx_s317_comp5321_c0_seq1	C18929862	-	-	-
CarTin_tx_s317_comp665796_c0_seq1	scaffold303323	-	-	-
CarTin_tx_s317_comp6677_c0_seq1	scaffold134363	-	-	-
CarTin_tx_s317_comp6680_c0_seq1	scaffold236737	-	-	-
CarTin_tx_s317_comp749263_c0_seq1	scaffold198555	-	-	-
CarTin_tx_s317_comp77834_c0_seq1	C19267335	-	-	-
CarTin_tx_s317_comp81387_c0_seq1	scaffold129644	-	-	-
CarTin_tx_s317_comp826687_c0_seq1	scaffold39752	-	-	-
CarTin_tx_s317_comp963682_c0_seq1	scaffold230894	-	-	-

TABLE I.3: Differentially expressed transcripts from Experiment 2 that map to SNP-containing Bowers contigs (continued).

		Bower SNP details		
Gx Contig	Bower Contig	Location	SNPs	Chr
CarTin_tx_s317_comp1627019_c0_seq1	(no hits)	-	-	-

Appendix J

Software Parameters (Assembly)

J.1 Safflower Transcriptome (Spring Reference)

J.1.1 Trinity (Inchworm, Chrysalis, Butterfly)

Version: v2012-06-08
Inchworm
Kmer - 25 bp
Min Length - 25 bp

Chrysalis
Min_glue: 2
Min_iso_ratio: 0.05
Glue_factor: 0.05
Weldmer_size: 48
Min: 200
Dist: 500
Max_reads: 20000000
Max_mem_reads: 1000000
Paired

Butterfly
(Defaults)

J.1.2 Biokanga 'Align'

Version - v3.8.1
Processing mode is : 'Standard alignment sensitivity'
Processing in standard basespace mode
alignments are to : either Watson '+' and Crick '-' strands
No PCR differential amplification artefact reduction
trim 5' ends raw reads by : 0
trim 3' ends raw reads by : 0
maximum aligner induced substitutions : 10 subs per 100bp of actual read length
minimum Hamming edit distance : 1
maximum number, precentage of length if read length > 100, of indeterminate 'N's : 1
minimum 5' and 3' flank exacts : 0

Raw read quality scores are : 'Ignore'
output format is : 'CSV match loci only'
process for: 'Single ended reads'
Process multiple alignment reads by: 'slough all reads which match to multiple loci'
Offset read start sites when processing site octamer preferencing: -4
Allow microInDels of upto this inclusive length: 0
Check for chimeric sequences in reads of at least this percentage length: 50
Maximum RNA-seq splice junction separation distance: 0
Minimum read coverage at loci before processing for SNP: No SNP processing
QValue controlling FDR (Benjamini-Hochberg) SNP prediction : No SNP processing
Min percentage non-ref bases at putative SNP loci : No SNP processing
Only accept reads which uniquely match a single loci

J.2 Safflower Transcriptome (Winter cultivar)

J.2.1 Biokanga 'Assemb'

Version - 3.5.3
End trimming by: 0bp
Accept input sequences, after any trimming, which are at least: 90bp
PE to SE end trimming by: 10bp
Allow SE conversion into PE: 'No'
Process sequences as strand specific: No
Process sequences as always single end: No
Initial minimal SE overlap required to merge SEs: 150
Final minimal SE overlap required to merge SEs: 25
Initial minimal sum of PE end overlaps required to merge PEs: 150
Final minimal sum of PE end overlaps required to merge PEs: 35
Minimal overlap of PE1 onto PE2 required to merge as SE: 20
Limit number of de Novo assembly processing passes to: 50
No intermediate assemblies output to file
Allow max induced substitutions per 100bp overlapping sequence fragments: 1
Allow max induced substitutions end 12bp of overlaps: 0
Threhold reduction steps: 5
Remaining steps before excessive PE end length checking: 2
5′ PE1 is : 'Sense' and 3′ PE2 is : 'Antisense'

J.2.2 Biokanga 'Scaffold'

Version - 3.5.3
Processing mode is : 'Output scaffold multifasta with edge report'
Allow max induced substitutions per 100bp overlapping sequence fragments: 0
Allow max induced substitutions end 12bp of overlaps: 0

Minimum PE insert size: 110
Maximum PE insert size: 1000
Minimum reported scaffolded sequence length: 300
5′ PE1 is : ′Sense′ and 3′ PE2 is : ′Antisense′

J.2.3 Biokanga ′Scaffold′

Version - 3.5.3
Alignment processing is : ′Standard alignment processing′
Sensitivity is : ′Standard alignment sensitivity′
Core extension score threshold : Auto
Core length : Auto (13)
Core delta : Auto (7)
Maximum depth to explore over-occurring seed K-mers : Auto (1500)
Minimum path score : Auto (130)
Minimum percentage of query sequence aligned : 25
maximum number of highest scoring paths per query : 10

J.3 Safflower Genome (Illumina)

J.3.1 Biokanga - ′Assemb′ (PE)

Version - v3.1.1
Processing mode is : ′standard de Novo assemble′
Allow SE conversion into PE: ′No′
Process sequences as strand specific: No
Process sequences as always single end: No
Initial minimal SE overlap required to merge SEs: 180
Final minimal SE overlap required to merge SEs: 60
Initial minimal sum of PE end overlaps required to merge PEs: 180
Final minimal sum of PE end overlaps required to merge PEs: 70
Minimal overlap of PE1 onto PE2 required to merge as SE: 40
Limit number of de Novo assembly processing passes to: 50
No intermediate assemblies output to file
Allow max induced substitutions per 100bp overlapping sequence fragments: 1 Allow max induced substitutions end 12bp of overlaps: 0
Threhold reduction steps: 5
Remaining steps before excessive PE end length checking: 2
5′ PE1 is : ′Sense′ and 3′ PE2 is : ′Antisense′

J.3.2 Biokanga - 'Assemb' (MP)

Version - v3.1.1
Processing mode is : 'standard de Novo assemble'
Allow SE conversion into PE: 'No'
Process sequences as strand specific: No
Process sequences as always single end: No
Initial minimal SE overlap required to merge SEs: 180
Final minimal SE overlap required to merge SEs: 60
Initial minimal sum of PE end overlaps required to merge PEs: 180
Final minimal sum of PE end overlaps required to merge PEs: 70
Minimal overlap of PE1 onto PE2 required to merge as SE: 40
Limit number of de Novo assembly processing passes to: 50
No intermediate assemblies output to file
Allow max induced substitutions per 100bp overlapping sequence fragments: 1
Allow max induced substitutions end 12bp of overlaps: 0
Threhold reduction steps: 5
Remaining steps before excessive PE end length checking: 2
5' PE1 is : 'Antisense' and 3' PE2 is : 'Sense'

J.3.3 Biokanga - 'Scaffold' (PE)

Version - v3.1.1
Processing mode is : 'Output scaffold multifasta with edge report'
Allow max induced substitutions per 100bp overlapping sequence fragments: 1
Allow max induced substitutions end 12bp of overlaps: 0
Minimum PE insert size: 180
Maximum PE insert size: 180
Minimum reported scaffolded sequence length: 300
5' PE1 is : 'Sense' and 3' PE2 is : 'Antisense'

J.3.4 Biokanga - 'Scaffold' (MP)

Version - v3.1.1
Processing mode is : 'Output scaffold multifasta with edge report'
Allow max induced substitutions per 100bp overlapping sequence fragments: 1
Allow max induced substitutions end 12bp of overlaps: 0
Minimum PE insert size: 5000
Maximum PE insert size: 15000
Minimum reported scaffolded sequence length: 300
5' PE1 is : 'Antisense' and 3' PE2 is : 'Sense'

J.3.5 Biokanga - 'Blitz'

Version - v3.9.8
Alignment processing is : 'Standard alignment processing'
Sensitivity is : 'Very high alignment sensitivity - caution: very slow'
Core extension score threshold : 16
Core length : Auto (12 reported and used)
Core delta : Auto (3 reported and used)
Maximum depth to explore over-occurring seed K-mers : 15000
Minimum path score : Auto (120 reported and used)
Minimum percentage of query sequence aligned : 5
maximum number of highest scoring paths per query : 10
alignments are to : Watson '+' and Crick '-' strands

J.3.6 Biokanga - 'Align' (fixed insert length)

Version - 3.9.8
Processing mode is : 'Standard alignment sensitivity'
Processing in standard basespace mode
alignments are to : either Watson '+' and Crick '-' strands
No PCR differential amplification artefact reduction
trim 5' ends raw reads by : 0
trim 3' ends raw reads by : 0
maximum aligner induced substitutions : 3 subs per 100bp of actual read length
minimum Hamming edit distance : 1
maximum number, precentage of length if read length > 100, of indeterminate 'N's : 1
minimum 5' and 3' flank exacts : 0
Raw read quality scores are : 'Ignore'
output format is : 'SAM Toolset Format, accepted aligned reads only'
If number of target sequences no more than this threshold then write all sequence names to SAM header: 10000
process for: 'Paired end reads with both ends uniquely aligned within the targeted genome'
Accept as paired if observed insert size is between 180 and 180
Accept as paired if 5' and 3' are are same strand: 'No'
Experimental: Output PE insert length distributions for each transcript or contig : 'No'
Experimental: Process PEs for spanning of circularised fragments: 'No'
Output paired end sequence length distribution to file: 'none specified'
Process multiple alignment reads by: 'slough all reads which match to multiple loci'
Offset read start sites when processing site octamer preferencing: -4
Allow microInDels of upto this inclusive length: 0

Check for chimeric sequences in reads of at least this percentage length: 50
Maximum RNA-seq splice junction separation distance: 0
Only accept reads which uniquely match a single loci

J.3.7 Biokanga - 'Align' (varied insert length)

Version - 3.9.8
Processing mode is : 'Standard alignment sensitivity'
Processing in standard basespace mode
alignments are to : either Watson '+' and Crick '-' strands
No PCR differential amplification artefact reduction
trim 5' ends raw reads by : 0
trim 3' ends raw reads by : 0
maximum aligner induced substitutions : 3 subs per 100bp of actual read length
minimum Hamming edit distance : 1
maximum number, precentage of length if read length > 100, of indeterminate 'N's : 1
minimum 5' and 3' flank exacts : 0
Raw read quality scores are : 'Ignore'
output format is : 'SAM Toolset Format, accepted aligned reads only'
If number of target sequences no more than this threshold then write all sequence names to SAM header: 10000
process for: 'Paired end reads with both ends uniquely aligned within the targeted genome'
Accept as paired if observed insert size is between 100 and 500
Accept as paired if 5' and 3' are are same strand: 'No'
Experimental: Output PE insert length distributions for each transcript or contig : 'No'
Experimental: Process PEs for spanning of circularised fragments: 'No'
Output paired end sequence length distribution to file: 'none specified'
Process multiple alignment reads by: 'slough all reads which match to multiple loci'
Offset read start sites when processing site octamer preferencing: -4
Allow microInDels of upto this inclusive length: 0
Check for chimeric sequences in reads of at least this percentage length: 50
Maximum RNA-seq splice junction separation distance: 0
Only accept reads which uniquely match a single loci

J.4 Safflower Chloroplast (PacBio)

J.4.1 PacBiokanga - 'Ecreads' (Pass 1)

Version - 1.8.1
Overlap processing: 'Sense only'
Use seed cores of this length when identifying putative overlapping sequences: 14bp
Require at least this many seed cores between overlapping sequences: 10

Offset cores by this many bp: 2
Maximum seed core depth: 15000
SW score for matching bases: 3
SW mismatch penalty: 7
SW gap opening penalty: 4
SW gap extension penalty: 1
SW gap extension penalty only applied for gaps of at least this size: 2
classify overlaps as artefactual if sliding window of 1Kbp over any overlap deviates by more than this percentage: 50
Minimum PacBio sequence length: 7500bp
Maximum PacBio sequence length: 35000bp
Minimum PacBio overlap required for error correction contribution: 5000
Trimming error corrected PacBio sequences until mean 100bp score at least: 3
Error corrected and trimmed PacBio sequences must be at least this long: 3000

J.4.2 PacBiokanga - 'Ecreads' (Pass 2)

Version - 1.8.1
Overlap processing: 'Sense and antisense'
Use seed cores of this length when identifying putative overlapping sequences: 14bp
Require at least this many seed cores between overlapping sequences: 10
Offset cores by this many bp: 2
Maximum seed core depth: 15000
SW score for matching bases: 3
SW mismatch penalty: 7
SW gap opening penalty: 4
SW gap extension penalty: 1
SW gap extension penalty only applied for gaps of at least this size: 2
classify overlaps as artefactual if sliding window of 500bp over any overlap deviates by more than this percentage: 50
Minimum PacBio sequence length: 3000bp
Maximum PacBio sequence length: 35000bp
Minimum PacBio overlap required for error correction contribution: 3000
Trimming error corrected PacBio sequences until mean 100bp score at least: 3
Error corrected and trimmed PacBio sequences must be at least this long: 2500

J.4.3 PacBiokanga - 'Ecreads' (Build Overlap File)

Version - 1.9.2
Overlap processing: 'Sense and antisense' Use seed cores of this length when identifying putative overlapping sequences: 35bp Require at least this many seed cores between overlapping sequences: 30 Offset cores by this many bp: 10 Maximum seed core depth: 15000 SW score for matching bases: 1 SW mismatch penalty: 10 SW gap

opening penalty: 12 SW gap extension penalty: 6 SW gap extension penalty only applied for gaps of at least this size: 1 classify overlaps as artefactual if sliding window of 500bp over any overlap deviates by more than this percentage: 20 Minimum error corrected sequence length: 5000bp Maximum error corrected sequence length: 35000bp Minimum overlap required: 1500

J.4.4 PacBiokanga - 'Contigs'

Version - 1.2.4
Minimum individual input sequence length: 5000bp
Minimum sequence overlap required to merge into single config: 5000
Minimum 1Kbp normalised overlap score: 980
Accepting orphan sequences: 'No'

J.4.5 PacBiokanga - 'Eccontigs'

Version - 1.2.4
Use seed cores of this length when identifying putative overlapping sequences: 35bp
Require at least this many seed cores between overlapping sequences: 30
Offset cores by this many bp: 10
Maximum seed core depth: 10000
SW score for matching bases: 1
SW mismatch penalty: 10
SW gap opening penalty: 12
SW gap extension penalty: 6
SW gap extension penalty only applied for gaps of at least this size: 1
classify overlaps as artefactual if sliding window of 1Kbp over any overlap deviates by more than this percentage: 10
Minimum contig sequence length: 10000bp
Minimum high confidence sequence length: 1000bp

J.5 Safflower Genome (PacBio)

J.5.1 PacBiokanga - 'Ecreads' (Pass 1)

Version - 1.9.2
Overlap processing: 'Sense and antisense'
Use seed cores of this length when identifying putative overlapping sequences: 14bp
Require at least this many seed cores between overlapping sequences: 10
Filtering PacBio reads for near homopolymer runs which are at least this length: 16bp
Offset cores by this many bp: 2
Maximum seed core depth: 15000
SW score for matching bases: 3

SW mismatch penalty: 7
SW gap opening penalty: 4
SW gap extension penalty: 1
SW gap extension penalty only applied for gaps of at least this size: 2
classify overlaps as artefactual if sliding window of 500bp over any overlap deviates by more than this percentage: 50
Minimum PacBio sequence length for error correction: 9000bp
Maximum PacBio sequence length: 35000bp
Minimum PacBio overlap required for error correction contribution: 5000
Trimming error corrected PacBio sequences until mean 50bp score at least: 3
Error corrected and trimmed PacBio sequences must be at least this long: 7500

J.5.2 PacBiokanga - 'Ecreads' (Build Overlaps - EC read samples)

Version - 1.9.2
Overlap processing: 'Sense and antisense' Use seed cores of this length when identifying putative overlapping sequences: 35bp Require at least this many seed cores between overlapping sequences: 30 Offset cores by this many bp: 10 Maximum seed core depth: 15000 SW score for matching bases: 1 SW mismatch penalty: 10 SW gap opening penalty: 12 SW gap extension penalty: 6 SW gap extension penalty only applied for gaps of at least this size: 1 classify overlaps as artefactual if sliding window of 500bp over any overlap deviates by more than this percentage: 20 Minimum error corrected sequence length: 5000bp Maximum error corrected sequence length: 35000bp Minimum overlap required: 1500