Effects of gene multiplication on flowering time regulation in spring and winter varieties of *Brassica napus*

David Marc Jones

A thesis presented for the degree of
Doctor of Philosophy

John Innes Centre, UK
University of East Anglia, UK
September 2017

Out of intense complexities, intense simplicities emerge.

Winston S. Churchill, *The World Crisis*

Contents

List of Tables

List of Figures

Acknowledgements

Many thanks to my supervisors, Richard Morris and Judith Irwin, for their support and reassurance throughout my PhD. Richard and I have a worryingly similar sense of humour and it has been reassuring to see that self-deprecation, making awful jokes and forcing tenuous puns do not preclude a career in science, as these are all traits we share. Thank you for introducing me to opera and Fivefingers. Judith's passion for science, but also the way she values her time away from work, has set a fantastic example that I hope to follow for the rest of my career. Her patience, particularly when it comes to myself and qPCR is admirable. I really appreciate the freedom both have given me to explore aspects of this project as I desired, and their constructive inputs and insights on the work throughout my PhD. Working with them, and in their groups, has been a truly enjoyable four years and I couldn't have asked for a better PhD experience. Thanks also to both for comments on this thesis, and my apologies for any worries leading up to the deadline!

Elsewhere at the John Innes Centre, I thank the members, and ex-members, of Richard's group. Many thanks to Vinod Kumar for being on my supervisory team and for interesting discussions and comments throughout my PhD. Thank you to Matthew Hartley and Tjelvar Olsson for informative conversations on computational science and ways to improve my coding. Julie Ellwood's attitude and general approach to life is brilliant, and I feel honoured to have witnessed it first hand.

The Brassica tendency group has grown over the years of my PhD, but my thanks go to the core members who have been there since the beginning: Richard Morris, Judith Irwin, Rachel Wells, Martin Trick, Nick Pullen, Emily Hawkes, and Eleri Tudor. It's hard to imagine the early days of these meetings

when we were almost speaking different languages, but doesn't that go to show how far we've all come? The social occasions we've shared have always been really enjoyable; there aren't many students who can say they've played Rock Band with their PhD supervisor! Thank you in particular to Nick Pullen, who in many ways, feels like an additional supervisor. The drive Nick has for science and his motivation has been inspiring, and I thank him for his continued guidance on aspects of this project and for his friendship.

My thanks go to my peers on the Rotation PhD programme, those that run it, and those labs that I rotated in. Thanks to Sarah O'Connor for letting me join her lab and to Richard Payne for the unenviable task of having to teach a computational biologist his way around said lab. Many thanks to Caroline Dean for my time in her lab and to Susan Duncan for supervising me. Thank you for the fantastic trip to Sweden, and to both Caroline and Susan for discussions about vernalization that have worked their way into this thesis. Nick Brewin, and later, Steph Bornemann both ran the Rotation PhD programme exceedingly well and the Rotation retreats were always a highlight of the year. Many thanks to the Rotation students from my year, Andrew Maclean, Nuno Leitão, Jemima Brinton, and Jie Li, for many social occasions and being great friends.

The cohort of PhD students with whom I've spent my time at the John Innes Centre have been fantastic. Thank you for Friday nights at the Rec, house parties, and a lot of fun. In particular, thanks to Andrew Maclean, Nuno Leitão, and Bethan Edmunds for three years of cohabitation in the Marble Factory; may the strange inside jokes and quote wall live on forever. Thank you also to Daisy Orme and Natasha Senior for an enjoyable final year of student house life. I am exceeding grateful to Rachel Prior both for making me realise the benefits of organized fun and for help getting this thesis bound. Many thanks to the Board Game Group of Chris Judge, Peter Emmrich, Matt Evans, Alex Calderwood, Thom Booth, and Ben Hales. You've helped foster and nourish an addiction that I hope to have for the rest of my life. I'm very lucky to have met Jemima Brinton during my PhD, whom I thank for being there when I need someone, picking me up when I'm down, and for having a great attitude towards life.

Finally, I am indebted to my family for the love and support they give me on a daily basis. It gives me tremendous strength and perseverance, and I will always be grateful. These thanks extend to family members who are no longer with us, but who are never too far from my thoughts.

Marc Jones

John Innes Centre, Norwich

September, 2017

List of publications

This thesis includes material from the following work.

D. Marc Jones, Rachel Wells, Nick Pullen, Martin Trick, Judith A. Irwin, Richard J. Morris. 2017. **Regulatory divergence of flowering time genes in the allopolyploid *Brassica napus*.** bioRxiv doi: 10.1101/178137

Abstract

Brassica napus (oilseed rape) is an economically important crop species that exhibits considerable varietal differences in flowering behaviour. Efforts to translate knowledge of flowering time control from model species are complicated by the evolutionary history of the crop. Whole genome duplication events have resulted in multiple copies of genes being present in the *B. napus* genome. A better understanding of the roles additional gene copies play during the floral transition would aid predictive models in directing future breeding efforts.

As a first step towards unravelling the regulatory network underlying the floral transition in the crop, a transcriptomic time series was conducted and used to investigate gene expression during the floral transition. Expression differences between homologous flowering time genes indicated that duplicated genes occupy separate locations in the gene regulatory network. This suggests the complexity of the regulatory network is vastly increased in *B. napus* relative to model species, and that the duplicated genes are likely to have different roles during the floral transition.

Duplicated genes were observed to diverge in different ways. Loss of regulatory elements surrounding certain *B. napus TFL1* homologues correlated with expression changes, highlighting the importance of cis-regulatory elements in the evolution of gene function. Sequence differences between *B. napus FD* homologues were found to alter the predicted dimerization affinities of the proteins. Expression variation between *B. napus FLC* homologues suggests only some confer a vernalization response, revealing these genes have diverged to have altered sensitivity to cold.

The finding that multiple homologues of the same flowering time gene in *B. napus* are expressed but show different expression dynamics reveals that the floral regulatory network from model species cannot be directly translated, but will require modification. This added complexity likely contributes to the developmental and genetic plasticity that has been exploited in this important crop.

Chapter 1

Introduction

Forecasting future events has been something humans have tried to do for millennia. Cicero in *De Divinatione* discusses the use of animal entrails, bird flight, and movements of celestial objects, to forecast the outcome of battles, trade deals, and crop growth (Cicero, *De Divinatione* 1.10, 1.1). A story recounted by Cicero, and also by Aristotle (Aristotle, *Politics* 1.1259a), tells the tale of Thales of Miletus, a philosopher who used astrology to predict the olive harvest for the following year. Knowing that oil presses would be in great demand during that time, Thales proceeded to rent oil presses months in advance at reduced rates. When the predicted olive harvest came Thales had access to every available press, and was able to profit from his forecast by subletting the presses at high rates. Most modern farmers would likely respond to crop predictions based on the movements of moons and stars with a couple of choice words. The story of Thales' olive presses, however, illustrates how useful crop predictions can be, regardless of where the forecast comes from. Knowing the best varieties to plant and the growth behaviour of crops allows for improved crop management.

Modern methods of predicting crop yields have progressed beyond the study of entrails[1]. However, the genetic links between the model inputs (satellite imagery, meteorological data, a cow's liver) and the outputs (crop yield) are often lacking. Although crop simulation models can be very sophisticated, detail at the genetic level is either not included, or is included empirically[2]. A pragmatic stance may be that if the predictions from such methods are

accurate, then who cares? What this viewpoint ignores is that understanding the underlying mechanisms of how climate and the environment affects crop growth can allow for novel crop varieties to be engineered, either through directed breeding or genetic modification[3]. These varieties could be engineered to suit particular growing seasons or locations.

Despite how potentially useful they could be, mechanistic models of plant growth have received most research effort within model plant species. It is my goal, in this thesis, to tackle the problem of how to adapt models of flowering time for the model plant species *Arabidopsis thaliana* to the crop species *Brassica napus*.

1.1 *Arabidopsis thaliana* as a model for flowering time

Model species have been key to the progression of biology by allowing researchers from all over the world to collaborate and focus research effort on common systems[4]. Although it has been worked on since the turn of the 20$^{\text{th}}$ century, it was not until the 1970s, and the desire for a plant well-suited to molecular genetics, that *Arabidopsis thaliana* (hereafter Arabidopsis) cemented its position as *the* model plant species[5,6]. Arabidopsis makes a good model organism due to a short generation time, a small physical size, and because it produces many seeds from a single, self-pollinated flower. Experimental tools have been developed to facilitate both forward genetics (identifying genotype from phenotype) and reverse genetics (identifying phenotype from genotype). The use of ethyl methanesulphonate to mutagenize Arabidopsis facilitated forward genetic screens to identify plants that are deficient in a phenotype of interest[7]. Such screens allowed the identification of global regulators of floral organ identity[8]. For reverse genetics, transformation methods using *Agrobacterium tumefaciens* were developed in the 1980s, allowing laboratory made genetic constructs to be inserted into the plant[9]. Another factor in the use of Arabidopsis as a model was the availability of a complete genome sequence, which was the first plant genome fully sequenced[10], and the third multicellular organism after *Caenorhabditis elegans*[11] and *Drosophila melanogaster*[12]. This

was in part possible due to the relatively small size of the Arabidopsis genome, which in hindsight also contributed to the success of mutant screens. The availability of these tools for manipulating the genome of Arabidopsis have allowed multiple developmental pathways to be dissected in the plant.

One developmental pathway of particular interest is the transition from vegetative growth to reproductive growth[13]. Timing this transition correctly is extremely important to ensure reproductive success of plants growing in the wild and maximal yields of plants grown as crops. The presence of the above mentioned genetic tools has allowed a deep understanding of the floral transition in Arabidopsis to be attained. Multiple pathways sense a myriad of internal and external cues to ensure that flowering in the plant is properly timed. The variation in floral response between different Arabidopsis accessions has also aided this work, making use of association studies to identify genes that influence the floral response most strongly[14]. There are five main pathways that influence flowering in Arabidopsis. These are the photoperiod pathway, the autonomous pathway, the vernalization pathway, the hormone pathway, and the ageing pathway[15]. All of these pathways converge and are integrated by a central network of genes to ensure that the plant flowers at an optimal time. In this section current knowledge of each of the pathways, and the key genes involved in them, will be summarized.

1.1.1 Floral pathways

The floral pathways can be divided into whether they respond to external (exogenous) or internal (endogenous) cues. The pathways that sense exogenous cues (the photoperiod and vernalization pathways) will be considered first.

The photoperiod pathway allows the plant to sense the day length. This is achieved through close association of the plant's circadian clock and light sensing apparatus. The circadian clock is a regulatory network that maintains a consistent oscillatory signal in the plant[16]. *CONSTANS* (*CO*) encodes a zinc finger transcription factor whose expression is downstream of the circadian clock, with *CO* mRNA accumulating and degrading in a regular manner each day[17,18]. However, CO protein is only able to accumulate when the plant is exposed to light, as it is rapidly degraded during the night[19]. During short days,

CO mRNA accumulates. CO protein is translated, but is rapidly degraded and cannot accumulate. However, during long days, *CO* mRNA is expressed at dusk, allowing CO protein to accumulate. *CO* accelerates the floral transition by binding to the promoter and activating the expression of a floral activator called *FLOWERING LOCUS T* (*FT*)[20–23]. This allows Arabidopsis to sense the day length and flower when the days are long enough.

Arabidopsis plants are capable of exhibiting two main life strategies[24,25]. Summer annual accessions germinate in spring, flower in the summer, and are able to set seed before winter. This is possible in warmer climates, such as central and southern Europe, where the length of summer is long, but in temperate climates, such as northern Europe, a winter annual strategy is followed[26]. These plants germinate in late summer or autumn, remain vegetative during the winter, and flower in the spring. If a plant following a winter annual strategy were to rely solely on the photoperiod pathway to determine flowering time, there is a risk that the day length in autumn would be long enough to activate flowering. This would result in vastly reduced reproductive success for the plant, due to the seed filling period taking place during the photosynthetically poor winter months. The vernalization pathway ensures that winter annual plants remain vegetative until after a period of cold has been experienced by the plant[27]. The vernalization response was found to be largely determined by two genes; *FRIGIDA* (*FRI*) and *FLOWERING LOCUS C* (*FLC*)[28]. *FLC*, a MADS-box containing transcription factor[29], acts antagonistically to CO by binding to the first intron of the *FT* locus[30,31] and repressing *FT* gene expression. In winter annual lines *FLC* is high when the plant germinates. During cold conditions the expression of *FLC* decreases. This repression is mitotically stable. What this means is that a plant that has experienced cold, when returned to growth in warm conditions, will continue to exhibit low *FLC* expression. In this manner *FLC* expression acts as a memory of whether a plant has experienced winter. This temporal separation of when the signal is sensed and when it is responded to is possible through epigenetic changes at the *FLC* locus. Cold treatment results in the expression of *VERNALIZATION INSENSITIVE 3* (*VIN3*), which in turn recruits the Polycomb Repressive Complex 2 (PRC2) to the *FLC* locus. PRC2 changes how the *FLC* DNA is packed in the nucleus, silencing it in a mitotically stable manner. In summer

annual accessions of Arabidopsis the expression of *FLC* is low when the plant germinates, negating the requirement for cold. Whether a plant is a winter or spring accession is largely dependent on the gene *FRI*. *FRI* activates *FLC* expression by stimulating the activity of a histone methyltransferase called *EARLY FLOWERING IN SHORT DAYS*[32,33]. An active allele of both *FRI* and *FLC* is therefore required for a plant to exhibit a vernalization response.

The other floral pathways sense endogenous cues in the plant. The autonomous pathway was named after a collection of mutant lines that flowered late regardless of the photoperiod the plants were grown under; *LUMINIDEPENDENS* (*LD*), *FCA*, *FY*, *FPA*, *FLOWERING LOCUS D* (*FLD*), *FVE*, *FLK*, and *RELATIVE OF EARLY FLOWERING 6* (*REF6*)[34]. However, the late flowering phenotype of these mutants was no longer observed if the plants were vernalized, suggesting that the autonomous and vernalization pathways converge[35]. This convergence was found to occur at *FLC*, with *FRI* and the autonomous pathway activating and repressing *FLC*, respectively, to set intial levels of the gene. In this manner, *FRI* and the autonomous pathway together control the vernalization response of the plant, with the autonomous pathway required in spring annual accessions to repress *FLC* expression[36].

A class of plant hormones called the gibberellins (GA) have been found to be linked to the floral transition, although other classes of plant hormones have also been implicated[37]. Plants that are mutant in the synthesis of GAs have a severe delay in flowering during short days, but show little effect during long days[38]. This indicates that the GA pathway in Arabidopsis is mainly involved with promoting flowering during non-inductive conditions.

Finally, the ageing pathway represses flowering when the plant is juvenile and promotes it when the plant ages. This response is mediated by microRNAs, 18 to 24 nucleotide RNA molecules that do not encode proteins[39]. These small molecules are involved with controlling the regulation of genes across both plant and animal kingdoms. With regard to the floral transition, two families of microRNA are particularly important; *miR156* and *miR172*[40]. The *miR156* family is expressed in the juvenile phase and decreases in expression as the plant ages. Conversely, the *miR172* family accumulates in expression as the plant ages. The *miR156* family targets the *SQUAMOSA PROMOTER BINDING-LIKE* (*SPL*) transcription factors, repressing their expression. The

SPL transcription factors activate the expression of a number of floral activators, namely *FT*, *SUPPRESSOR OF CONSTANS 1* (*SOC1*), *APETALA 1* (*AP1*), and *LEAFY* (*LFY*). Therefore the decrease in expression of the *miR156* family as the plant ages allows these floral activators to be expressed. Another SPL transcription factor target is *miR172*. Hence, the decrease in *miR156* expression results in the increase of *miR172* expression. *miR172* represses the activity of the Arabidopsis *APETALA 2* (*AP2*) family of genes, a set of floral repressors that have found to have binding sites upstream of floral activators. The feedback loop created ensures that the switch from the juvenile to the mature growth phase is stable. The regulation of *miR156* is hypothesized to be regulated by sugar or carbohydrate availability, which is used as a proxy for the age of the plant.

1.1.2 Floral integrators

The pathways described above converge onto a set of floral integrator genes, that mediate the transition to flowering. The core of this network is composed of relatively few transcription factors with multiple regulatory links between them. These feedback loops allow for the signals from the flowering time pathways to be appropriately interpreted and provide robustness to the system[41].

Both the photoperiod and vernalization pathways converge onto the expression of the floral activator *FT*. Grafting experiments in a number of plant species led to the conclusion that a floral inducer, referred to as the florigen, was transported from leaves to the shoot apex to initiate flowering[42,43]. It later emerged that the florigen, initially hypothesised to be a plant hormone, was the protein FT. *FT* is expressed in the phloem companion cells, and the FT protein is transported in the plant vasculature from leaves to the apex to promote flowering[44–46]. The gene was identified from a photoperiod sensitive mutant plant that exhibited delayed flowering when the plants were grown in long days[35]. This photoperiod sensitivity was found to be due to *FT* being directly regulated by the circadian clock gene *CO*[20–22]. The vernalization pathway also influences the expression of *FT*, with *FLC* binding to a site within the first intron of *FT* to repress its expression[30,31]. *FT* activates the expression of three MADS-box containing proteins that promote flowering; *FRUITFULL*

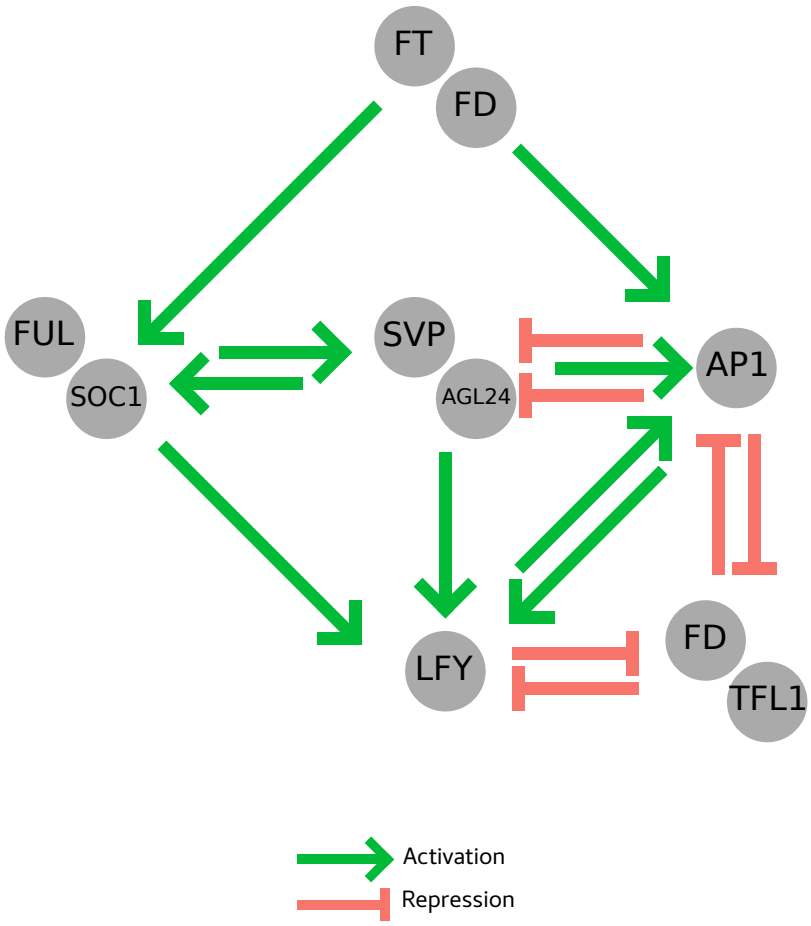


Figure 1.1: The core network of floral integrators.
Many regulatory interactions have been found between the nine floral integrators depicted here. This results in a tightly interconnected gene regulatory network, with many possible feedback loops and control mechanisms. Only key regulatory and protein-protein interactions are depicted here. Adapted from Bouché et al. (2016)[299].

(*FUL*)[47], *SOC1*[48], and *AP1*[49]. These will be discussed in more detail later in this section.

A gene found to act in an antagonistic manner to *FT* in determining the floral transition is *TERMINAL FLOWER 1* (*TFL1*). Wild type Arabidopsis flowers develop in an indeterminate manner[50]. When the transition to flowering occurs, the vegetative meristem converts into an inflorescence meristem that in turn generates the floral structure. Additional inflorescence meristems, and eventually floral meristems, develop on the side of the main inflorescence stem. However, the shoot apical meristem, located at the top of the floral structure, remains as an inflorescence meristem, and hence floral growth is indeterminate in Arabidopsis. Mutants in *TFL1* result in the primary inflorescence meristem converting into a floral meristem, such that the floral structure terminates in a flower as opposed to maintaining an indeterminate state[51]. In addition, *TFL1* null mutant plants also undergo the floral transition earlier than wild type plants[52]. *TFL1*, therefore, is a repressor of the floral state, influencing meristem identity and regulating the timing of the floral transition. The inflorescence meristem identity is maintained by TFL1 protein through limiting the activity of *AP1* and *LFY*[53–55]. In addition to transcriptional repression, TFL1 protein limits the activity of AP1 and LFY protein, as shown by Arabidopsis lines that overexpressed *TFL1* and either *AP1* or *LFY*[55]. Likewise, AP1 and LFY repress *TFL1* gene expression, with the mutual antagonism likely leading to the sharp expression boundaries required to accurately specify floral development[53,56] *TFL1* and *FT* are very closely related proteins, with only 39 amino acid changes that distinguish the two proteins[57]. Indeed, mutations have been found that produce TFL1 proteins that are FT-like and vice versa[57–59].

Despite being central floral integrators, both *FT* and *TFL1* do not possess DNA binding activity themselves and are therefore not transcription factors. Instead, the proteins of both genes interact with the FD protein, a bZIP transcription factor[41,47,49]. *FD* was originally identified as a late flowering mutant[35] found to repress the phenotype of *FT* overexpression lines[49]. The FD protein was confirmed as an interacting partner of FT in a yeast two-hybrid screen[47], and was found to also bind FT *in vitro*[49]. Two lines of evidence point towards FT and FD interacting to promote *AP1* expression. The first is the ectopic expression of *AP1* observed in *FD* overexpression lines that is

dependent on the presence of *FT*[49]. The second line of evidence are chromatin immunoprecipitation experiments conducted in an *FD* overexpression line. Antibodies for the FT protein were used to enrich DNA and *AP1* promoter sequence was found in a *FT* dependent manner[47].

A homeotic mutation in Arabidopsis that severely impacts the transition from vegetative to floral growth is in the *LEAFY* (*LFY*) gene. *LFY* was identified in a mutant screen as a mutant that produced leafy shoots in the place of flowers, with the flowers that were produced often lacking petals and stamens[60]. The gene was found to play a role both in the transition to flowering, but also in specifying the determinacy of the floral meristem[61]. LFY binds to DNA as a dimer, with the cooperative nature of this binding suggested to facilitate a sharp developmental transition[62]. *LFY* has been found to regulate or interact with a number of other genes involved with the floral transition. Increasing *LFY* expression precedes an increase in *AP1* expression[63], with additional evidence suggesting that *AP1* is a direct target of *LFY*[64,65]. Other genes important for flowering that are regulated by *LFY* are *TFL1*[66], *AGAMOUS* (*AG*)[67,68], and *CAULIFLOWER* (*CAL*)[65], with *LFY* itself being regulated by *SOC1* and *AGAMOUS-LIKE 24* (*AGL24*)[69]. In addition, a suite of transcription factors and signalling molecules, both related to flowering time and not, were found to respond to *LFY* activation or have *LFY* binding detected in promoter regions[65,70]. Interactions between *LFY* and the photoperiod pathway[71], and the GA pathway[72,73], suggest that many environmental pathways that regulate flowering converge on *LFY*, underpinning its role as a floral integrator.

AP1 is a MADS-box containing transcription factor[74] important for both controlling meristem identity and floral organ specification. Null mutations in the *AP1* gene result in the mutant plants lacking petals[75], a consequence of the role *AP1* has in specifying floral organ identity. Additionally, the sepals that usually surround flowers in *AP1* mutant plants are instead converted to bracts, with secondary flower buds formed in the axils of each bract[76]. This particular phenotype suggests that *AP1* is important for the conversion of the inflorescence meristem into a floral meristem, as without an active version of the *AP1* protein the floral meristem partially reverts back to an inflorescence meristem[74]. This is also supported by the *AP1* overexpression phenotype, where apical and lateral shoots are converted into flowers[77]. The modulation of meristem activity

by *AP1* is believed to be via the plant hormone cytokinin, with *AP1* affecting both the biosynthesis and degradation pathways of the hormone[78]. 25% of the putative targets of *AP1* are other transcription factors, such as *LFY*, explaining why plants mutant in and overexpressing *AP1* have such dramatic effects on flower development in Arabidopsis[79]. *AP1* and *LFY* double null mutants had a significantly more severe phenotype than either of the single mutants, indicating that these genes seem to act synergistically[80]. In mutant plants lacking *AP1*, *SHORT VEGETATIVE PHASE* (*SVP*), *AGL24*, and *SOC1* become ectopically expressed[81], with further evidence suggesting that *AP1* directly represses the expression of these genes[82]. *SVP* and *AGL24* maintain the vegetative and inflorescence meristems respectively[81]. The expression of *AP1*, therefore, confers a floral state to the meristem.

A gene involved with integration of inputs from an array of different flowering time pathways is *SOC1*[83]. The gene was discovered[84] and rediscovered[85] in Arabidopsis through a number of different experimental methods. The *SOC1* gene was found to be differentially expressed after activation of an inducible CO protein in the absence of protein translation, suggesting *SOC1* is a direct target of CO[20] and thus downstream of the photoperiod pathway. Indeed, *SOC1* gets its name as a mutant in the *SOC1* gene was able to suppress the early flowering phenotype of an Arabidopsis line overexpressing *CO*[84]. The overexpression of *SOC1* in a vernalization requiring line of Arabidopsis was able to overcome the vernalization requirement, suggesting that *SOC1* is also a part of the vernalization pathway[85]. Subsequent analysis has revealed that this regulation is likely to be direct, as a transcription factor motif in the *SOC1* promoter was found to be bound by FLC *in vitro* and required for *SOC1* repression *in vivo*[30,86]. *SOC1* was initially discovered, therefore, as acting downstream of the vernalization and photoperiod flowering pathways, and subsequent investigations have revealed that *SOC1* is involved with additional floral pathways. The rescue of a GA biosynthesis mutant with the treatment of GA causes an increase in the expression of *SOC1*[87]. This finding, in addition to the *SOC1* mutant being less sensitive to the treatment of GA[87], suggests *SOC1* integrates the response to the GA-dependent, hormonal pathway. *SOC1* has also been implicated in the intermittent cold-sensing pathway[88] and the age-dependent flowering pathway[89]. All this evidence points towards *SOC1*

being a central integrator that is the convergence point of a range of flowering time control pathways.

The regulation of *SOC1* is tied to another MADS-box containing flowering time gene, *AGL24*, as both regulate each other in a positive feedback loop[90]. The AGL24 protein was found to be important for the entry of SOC1 protein into the nucleus, with AGL24 and SOC1 binding as a probable dimer at the promoter of *LFY*[69]. *AGL24* seems to act somewhat redundantly with *AP1* and *SVP* to repress certain genes involved with floral organ specification to properly pattern the developing flower[91].

The *SOC1* gene is at least somewhat redundant with the gene *FUL*, suggesting that *FUL*[92], like *SOC1*, is activated by *FT* expression[93]. The gene was characterised as affecting the development of the Arabidopsis seed pod[94], but was also found to act earlier in the reproductive phase by controlling flowering time and meristem identity alongside *AP1* and *SOC1*[95]. Indeed, plants that are mutant in both *FUL* and *SOC1* remain vegetative, and almost resemble perennial plants[92]. SOC1 and FUL interact[96], and have been found to bind to and activate *LFY* expression[97].

Finally, *SVP* is a gene that seems to have a dual role as a floral repressor early in development, and as a floral meristem identity specification gene later in development, with differing target genes[98]. As a floral repressor, it has been found to form a heterodimer with FLC, although lack of SVP does not significantly impact the targets of FLC. This is not mutual, however, as the presence of FLC causes a large effect on the targets of SVP, with the number of targets doubling[99]. As with FLC, SVP has been found to bind at the *FT* locus to delay the floral transition[100]. When the floral transition occurs, however, SVP seems to act redundantly with AP1, AGL24, and SOC1 to maintain an indeterminate meristem[82,91,101,102]. Extensive heterodimer formation has been demonstrated between the MADS-box containing flowering time proteins[96]. It therefore seems likely that the role of SVP changes depending on which proteins it dimerizes with[82,99].

Being sessile organisms, plants need to interpret environmental cues and respond appropriately. The different floral pathways allow for these environmental cues, and for endogenous cues such as age, to be interpreted. The

combined interactions of the floral integrators discussed here allow for these signals to be integrated, providing robustness to the floral transition[41]. The flowering time genes and pathways identified in Arabidopsis have been found to be somewhat conserved in a wide range of crop species[3], leading some to dub Arabidopsis the 'Rosetta stone' of flowering time research[103].

1.2 The origin of *Brassica napus* and why flowering time is important

The *Brassica* genus is in the same taxonomic family as Arabidopsis, the Brassicaceae[104], and comprises a large number of economically important vegetable and oil crops that show broad morphological divergence[105]. Among the *Brassicas* are both diploid and tetraploid species. Diploid species of the *Brassica* genus include *B. rapa* (Chinese cabbage, turnip, and pak choi), *B. oleracea* (kale, cabbage, broccoli, cauliflower, and Brussels sprout), and *B. nigra* (black mustard). A theory proposed by Woo Jan-choon in 1935, that has become known as the triangle of U, posits that ancestors of the above diploid species hybridized to give ancestors of the tetraploid species of *Brassicas*[106]. These tetraploid species are *B. napus* (oilseed rape, swede, kale), *B. carinata* (Ethiopian mustard), and *B. juncea* (Indian mustard). As the tetraploids are the result of interspecies hybridization events they are termed allopolyploids. Progenitors to modern day *B. rapa* and *B. oleracea* plants are thought to have hybridized to form ancestral *B. napus* less than 10,000 years ago[107], with multiple hybridizations having taken place to give the modern *B. napus* gene pool[108]. Rapeseed crops, such as *B. napus*, are the second most cropped oil crop worldwide comprising 13% of the total yield[109], with the oil being used as a vegetable oil and for industrial lubricants. In the UK, 13% of the total area on which crops were grown in 2016 (608,000 hectares) was used for oilseed crops, generating £541 million in income[110]. Oilseed rape is frequently grown in rotation with wheat, with wheat grown in such a way yielding 10% more than wheat grown continuously, on average[111].

Aside from being an economically important crop, the *Brassica* species are also a model for gene retention. The genomes of *Brassica* species have undergone

Table 1.1: Main *Brassica* crops, their common names, and the part of the plant that is consumed.

Table obtained from Cartea et al. (2011)[105].

Species	Group	Common name	Organ consumed
Brassica oleracea	*acephala*	Kale, collards	Leaves
	capitata capitata	Cabbage	Terminal leaf buds (heads)
	capitata sabauda	Savoy cabbage	Terminal leaf buds (heads)
	costata	Tronchuda cabbage	Loose heads
	gemmifera	Brussels sprouts	Vegetative buds
	botrytis botrytis	Cauliflower	Inflorescences
	botrytis italica	Broccoli	Inflorescences
	gongylodes	Kohlrabi	Stem
	albogabra	Chinese kale	Leaves
Brassica rapa	*rapa*	Turnip	Roots
	rapa	Turnip greens	Leaves
	rapa	Turnip tops	Shoots
	chinensis	Pak choi, bok choi	Leaves
	pekinensis	Chinese cabbage, pe-tsai	Leaves
	parachinensis	Choy sum	Leaves
	ruvo	Broccoleto	Shoots
	perviridis	Komatsuna, Tendergreen	Leaves
Brassica napus	*pabularia*	Leaf rape, nabicol	Leaves
	napobrassica	Swede	Roots
Brassica juncea	*rugosa*	Mustard greens	Leaves
	capitata	Head mustard	Heads
	crispifolia	Cut leaf mustard	Leaves

genome duplication events relative to *Arabidopsis* since the two genera diverged 43 million years ago[112]. There is evidence for an ancestor of the *Brassica* lineage being a hexaploid, with estimates of when the genome triplication occurred varying from 7.9 - 14.6 million years ago[113] and 23 million years ago[112]. Subsequent diploidization of this hexaploid ancestor has given us the diploid *Brassica* species we have today. *B. rapa* and *B. oleracea* diverged 0.12 - 3.7 million years ago[114,115], with the process of chromosome rearrangement and loss resulting in a chromosome number of ten for *B. rapa* (A genome)[116] and nine for *B. oleracea* (C genome)[117]. It is thought that the interspecies hybidization events resulting in the allopolyploid *B. napus* occurred less than 10,000 years ago[107]. Both the ancient hexaploid state of the *Brassica* genomes and the interspecies hybridization event mean that *B. napus* has a greatly increased gene number than Arabidopsis (101,040[118] relative to 25,498[10]), with genes in Arabidopsis often having multiple homologues in the *B. napus* genome. Despite large scale genome rearrangements, extensive collinearity between the *Brassica* and Arabidopsis genomes remains[116–118]. This genomic collinearity and relatedness of the two plant species has been exploited to translate research from the model plant to the crop species, as well as investigate the effects of gene duplication.

1.2.1 How does flowering time affect the cultivation of *Brassica* species?

The success of many *Brassica* crops is dependent on their flowering time. The edible component of both broccoli (*B. rapa* var. *botrytis italica*) and cauliflower (*B. rapa* var. *botrytis botrytis*) are the plant inflorescences, and the timing of the floral transition and floral development in general is very important for these crops as a consequence. Using variation in curd formation in cauliflower, a number of potential candidate genes were identified as controlling the response to temperature, with some of these genes being homologues of floral genes in Arabidopsis[119]. With other *Brassica* crops, such as Chinese cabbage (*B. rapa* var. *pekinensis*) the prevention of flowering is desired. Chinese cabbage is grown for its leaves (Table 1.1). If the plant transitions to floral growth, it will bolt, significantly reducing its economic value. The expression of a floral

repressor, a *B. rapa* homologue of *FLC*, was found to correlate with bolting time in different Chinese cabbage lines[120].

B. napus crops are predominantly used as oilseed crops, in which the timing of the floral transition impacts both when the seed filling period begins and how long it progresses. Indeed, the interconnected nature of yield and flowering has been suggested by association studies finding regions of the genome associated with both traits[121,122]. The yield of oilseed rape crops is determined by the number of seeds the plants produce per area over which the crop is grown and the weight of each seed. Numbers of pods and seeds are largely determined during a 3 week phase after flowers have formed[123,124], with the quality of the seed dependent on a period of seed filling. The seed quality is related to temperature, with cooler conditions extending seed filling, and the rate of photosynthesis, with the majority of oil in the seed accumulated during the second half of seed filling[124]. The effect of photosynthesis during the seed filling period is potentially of greater significance in *B. napus* relative to other crops as the remobilization of carbohydrates accumulated before flowering is ~12%, compared to 20 - 50% in wheat[124,125]. Yield of winter oilseed rape has been found to be related to the size of the crop at flowering[123]. This was in turn a function of the length of time between the beginning of spring, when mean growing temperatures exceeded 5 °C, and when the plants flowered in late May. Therefore, the highest yielding years were those where spring was early and flowering late, allowing the longest period of time for growth in this critical period. Similar findings came out of modelling the growth of *B. napus*, with higher yields predicted to be obtained by delaying plant maturity and promoting earlier flowering, to ensure the seed filling period is as long as possible[126] Therefore, when flowering occurs during the growing season, and how that relates to the climate in which the crop is grown, can heavily influence the yield and quality of the crop.

Whether oilseed rape is a spring or winter variety is also important, as different growing regions require different types of crop. In Europe and Asia, winter oilseed rape is predominantly grown, whereas in Australia, Canada, and northern Europe spring types are generally grown[127]. For Canada and northern Europe, the requirement for spring types results form harsh winters that prevent the crop from being overwintered. Therefore, the vernalization requirement of

a variety is important to consider for the planned crop rotation a particular farmer or growing region requires. Additionally, the length and severity of cold required by a variety will dictate whether that variety is suitable to a particular application.

Finally, the availability of pollinators can significantly impact the yield of *B. napus* crops. Preventing pollinators visiting winter oilseed rape plants led to a 27% decrease in the number of seeds produced and a 30% decrease in the seed weight per pod[128]. In addition, the diversity of those pollinators visiting the plants is related to oilseed rape yield[129]. Changes to flowering time will affect the pollinators that are available to the flowering plants. This has been found to profoundly affect the reproductive success of perennial wildflowers[130]. Therefore, as the yield of oilseed rape is influenced by pollinator availability[128,129], the correct timing of flowering is required.

1.2.2 Work on the control of flowering in Brassica species

Extensive work on how the floral response is controlled in Arabidopsis has facilitated understanding of floral control in a range of crop plants[3]. Homologues of the floral genes (section 1.1) have been detected in the genomes of *Brassica* species, which due to the gene multiplication events that have occurred in the Brassica lineage are often present as multiple copies[131]. However, identification of whether these *B. napus* homologues have similar functions to their counterparts in Arabidopsis, and of functional differences between the homologues, is often lacking.

Likely as a result of both spring and winter varieties of *Brassica* crops being of such economic value, the vernalization pathway has arguably been the most well studied flowering pathway in Brassicas. Association studies focussing on mapping the vernalization response in *B. rapa*[132–137], *B. oleracea*[136,138–140], and *B. napus*[134,141–143] have identified regions containing homologues of *FLC* and *FRI* as explaining flowering time variation. These homologues exhibit similar decreases in expression during cold as their Arabidopsis counterpart[120,144] and have been investigated to determine if they have diverged in function or

not. Expression of five different *FLC* homologues in Arabidopsis conferred a vernalization requirement in a rapid-cycling accession of Arabidopsis[145]. Interestingly, the delay in flowering as a result of the transgenic gene varied depending on the homologue, suggesting that the genes have diverged roles in *B. napus*[145]. The results from association studies carried out with different mapping populations have also suggested that the *FLC* copies in *Brassicas* have diverged, with the copies on chromosomes A10 and A2 showing stronger associations with flowering time[137,141]. Similarly, multiple *FLC* homologues from *B. rapa* delayed flowering when overexpressed in both Arabidopsis and Chinese cabbage, suggesting a conservation of function[146]. Such functional conservation is also observed for *FRI*, with *FRI* homologues from *B. oleracea* able to complement an Arabidopsis accession that contains a nonfunctional copy of the gene[147]. Despite all homologues being able to complement Arabidopsis, structure of the *FRI* homologues from *B. oleracea* have diverged with alterations in the number of coiled-coil domains, potentially impacting protein-protein interactions[147]. Therefore, although it has been established that *FLC* and *FRI* seem to be important in the vernalization pathways of both *Brassica* crops and Arabidopsis[28], how the copies of these genes have diverged in *Brassica* is only beginning to be understood.

Genes in other flowering time pathways, and the floral integrators, have also been investigated in *Brassica* species. Genes involved with the circadian clock have been retained in the *B. rapa* genome, suggesting that the dosage of the genes is important for their function[148]. In particular, homologues of the clock sensitive gene *CO* are associated with changes in flowering time in both *B. oleracea*[149] and *B. nigra*[150,151]. Homologues of *TFL1* were identified in *B. rapa*, *B. oleracea*, and *B. napus*, with expression in the flower in the latter species in line with expression of the gene in Arabidopsis[152]. Mutations in the A10 copy of *TFL1* in *B. napus* caused a delay in flowering, affected internode elongation, and resulted in an increase in seed number and weight[153]. Homologues of *FT* in *B. napus* exhibited different expression patterns, with certain copies having a stronger effect on flowering than others[153]. Expression differences were observed between the homologues of *FT* in *B. rapa*, *B. oleracea*, and *B. napus*[154]. Within *B. napus*, one copy was silenced as a result of transposon insertion into the promoter region and the expression of another two

copies was crop type specific[154]. Transposon mediated changes to the expression of an *FT* homologue were also identified in *B. rapa*, resulting in flowering time differences. This suggested that this copy of *FT* has retained a function similar to its counterpart in Arabidopsis[155]. Arrest of floral development is required in broccoli and cauliflower to form the heads correctly. Interestingly, the floral genes predicted to cause the arrest (*LFY*, *AP1*, and *TFL1*) were not implicated, causing the authors to suggest other floral meristem genes are mediating the change relative to Arabidopsis[156]. Links between flowering time and *SOC1* homologues have been identified in *B. rapa*[157], with expression differences detected between the different homolgoues in *B. rapa* and *B. juncea*[157,158].

Despite evidence of flowering time genes homologues having similar roles in *Brassica* species, in-depth analysis of how different homologues are behaving is often lacking. This is not the case for all genes however, with the roles of *FT*, *FLC*, and *FRI* homologues in *Brassica* species being dissected in a copy-specific manner[141,142,147,154]. These investigations have revealed that individual copies have indeed diverged in function and behaviour.

1.3 Modelling flowering time and crops

From simulating cell-signalling dynamics[159], patterning of biological systems[160], up to population level models[161], mathematical models have been able to capture the behaviour of a range of biological processes. Models allow researchers to collect potentially disparate observations together to test if they are consistent with each other. If they are consistent, then the researchers' assumptions about the system are compatible with the data and the model can be used to make predictions. If the model does not capture the behaviour of the system, then clearly the system is more complex than originally thought. Either way, modelling systems can direct future research work and highlight features of the system that might not have been appreciated had a reductionist approach been taken. This section will highlight models of the floral transition that have been developed, as well as how models of crop growth have been used by both the agricultural industry and the scientific community to direct scientific effort and farming practices.

1.3.1 Models of the floral transition

The floral transition is composed of a suite of transcription factors that control the floral development both spatially and temporally (section 1.1.2). As the regulatory interactions between these transcription factors have been elucidated, gene regulatory networks have been used to model the floral transition[41,162]. Gene regulatory networks consist of genes as nodes in the network and the regulatory interactions between those genes as edges of the network[163]. The genes involved in these networks generally encode transcription factors; proteins that have the capacity to alter the transcription of other genes. The network structure results as a consequence of regulatory links between transcription factors. The combination of interactions between transcription factors can lead to complex behaviours that have favourable properties such as noise cancellation, high-pass filters, and low-pass filters[164]. The combination of these, and other, simple regulatory structures allows for complex responses to stimuli to be encoded[165].

As a consequence of their capacity to capture complex behaviours between genes, gene regulatory networks have been employed in many fields of biology[166]. The behaviours captured by the models, and the consequences of those behaviours such as noise cancelling or signal amplification, are often initially unintuitive, highlighting the necessity of the models[164]. An example of particular interest to the work presented here is that of Jaeger et al. (2013), in which the floral transition was modelled[41]. A simplified network of five floral integrators, *FT*, *LFY*, *FD*, *TFL1*, and *AP1* were used as nodes in the network, with edges consisting of regulatory interactions determined genetically and molecularly (section 1.1.2). The model consisted of five gene hubs and was parameterized using the flowering time (measured as the number of rosette and cauline leaves present at flowering) of Arabidopsis single and double mutants in the floral integrators. The model was able to capture a number of dynamics of the floral transition, such as irreversibility and noise filtering. Insights from the model included the observation that the relative levels of *TFL1* and *FT* were important for determining when the floral transition occurred. Additional regulatory interactions involving the regulation of *TFL1* were also proposed as necessary for the maintenance of a high *TFL1* expression state[41].

Valentim et al. (2015) extended the model of Jaeger et al. (2013) by incorporating additional genes and by using expression data to parameterize the model[162]. This meant that, unlike the gene hubs used in the earlier study, the network nodes in the Valentim et al. model better corresponded to the genes themselves. The findings of the study highlight the sometimes unintuitive dynamics that are unveiled when a system is computationally modelled. For example, it was found that mutating *SOC1* has a greater effect on the expression of *AP1* than on *LFY*, which is surprising given that the regulation is indirect and direct respectively.

Although much more simplified than other modelling strategies, a two gene regulatory model of the floral transition in a perennial relative of Arabidopsis, *Arabidopsis halleri*, was capable of accurately modelling the floral transition and the timing of floral reversion back to vegetative growth[167]. By incorporating temperature responsive production and degradation rates of the two genes into the model, the projected effects of climate change on the developmental timings of natural populations of the plants could be predicted.

The model developed by Espinosa-Soto et al. (2004) models a regulatory network of 15 genes involved with the ABCE model of floral patterning[168]. Instead of modelling the expression level of genes as continuous variables, as the other models have done, discrete gene expression levels were used. This simplification allowed the researchers to test every possible initial condition for their model. The properties of the network resulted in the expression of the genes converging to only 10 stable states, which corresponded to the expression profiles of different floral cell lineages in the Arabidopsis flower. In addition, the model was capable of reproducing regulatory effects of knockout and overexpression mutations[168,169].

Extending gene regulatory network based models away from model species, Dong et al. (2012) developed a four gene regulatory model that took structural cues from the network in Arabidopsis to model the floral transition in maize[170]. As with the Jaeger et al. Arabidopsis model, this maize model was parameterized using total leaf number as a proxy measurement for flowering time, and validated using mutants in the genes involved in the network.

All of these examples illustrate the insights that can be obtained from taking into account the regulatory networks that underlie the floral transition.

1.3.2 Models of crop growth and yield prediction

Crop models have been studied and used in the research community for over fifty years[171]. These models aim to explain, or predict, the growth of plant species that are grown as crops. The motivation for using crop models can vary[171]. For the scientific community, crop models allow for the integration of seemingly distinct models of processes. Initial models focussed on modelling photosynthesis[172] have been improved upon, with modern models incorporating processes such as leaf development, light interception, photosynthesis efficiency, and partitioning of biomass within the plant[173]. The other use of crop models is to aid decision making, at a farm, country, and global scale[174,175]. Such models incorporate additional processes, such as nitrogen use efficiency[176] and soil erosion[177], in order to take into account the effect of fertilizer use not only on the crop but also to the wider environment[178,179]. The incorporation of climate and weather data into these models have allowed predictions to be made about the effects of climate change on crop growth and yield. Using this methodology with multiple models allowed Rosenzweig et al. (2014) to predict that low latitude areas would be most affected by climate change in terms of crop yield for four different crop types[180]. Ultimately crop models at this scale can be used to predict harvesting dates of some crops, allowing sowing dates to be optimized and allowing the supply of the crop to be more accurately estimated[181]. For example, the use of climate forecasting was used in the sugar industry to improve water use efficiency at the farm level, while also benefiting industries further down the sugar supply line through enhanced scheduling[182].

Crop models can be split into two types; process-led models and statistical models[2,183,184]. In process-led models, the inputs to processes and how those outputs interact are explicitly modelled, and are used to help understand plant-environment interactions. The effects of changing inputs can be tracked through the model, and stability analysis can be conducted to determine which input parameters the model is particularly sensitive to[185]. The advantages of modelling processes explicitly is that, generally, the predictions that the

model can make are more accurate. Specifically, the ability of the model to extrapolate and make predictions about future events is improved by effectively giving the model an understanding of how the crop plants under study will respond to particular inputs[2]. The downside of such models is that they often have many parameters, that either have to be measured or predicted from training sets of data. This parameterization often requires a lot of data to be collected, which with crop plants may be difficult or costly to do. The complexity of the models will also affect how quickly these parameters can be estimated, and often how long the model will take to run. Once trained, however, the insights from the models can be very precise. Modelling wheat growth in sub-tropical India found yields were very sensitive to temperature, potentially informing the selection of future varieties grown[185]. Modelling the growth of maize, spring wheat, and soybean revealed that an altered planting date combined with alternative varieties could reduce losses due to projected climate change by 18%.

Statistical models, conversely, do not explicitly model processes, and instead attempt to relate model inputs, such as climate data, to model outputs, such as crop yield, in a correlative manner[186]. These models are much simpler, with fewer parameters, than the process-led models. This means the models are faster to run and potentially require less data to parameterize them. This makes statistical models well-suited for use as summary models, that capture the general trends between variables[171]. However, as the models do not interpret the data in terms of plant growth, statistical models are potentially less accurate when extrapolating the data to make predictions. Despite their simplicity, statistical models are still capable of facilitating insight, such as predicting potato yields from satellite imaging and remote sensing data[187].

1.3.3 Integrating the two types of models

A potential short-coming of modelling plant growth responses using models that do not simulate regulatory networks is that regulatory logic may be lost. Different crop varieties and species are often incorporated into crop models through parameter changes[175,178]. However, the regulatory logic of the crop models will remain unchanged. For example, the output from two signalling

pathways may be required simultaneously to activate expression of a particular target pathway. Genetic differences between varieties could potentially alter this logic, resulting in the target pathway being activated if *either* input pathway is active. This could result from differences in promoter binding sites between varieties. Implementing this change in logic, in the APSIM framework for example, would require writing an alternative module that integrated the responses from the input pathways in a different manner[175,188].

Integrating gene regulatory networks into crop models would only be beneficial for processes where the regulatory logic of the system is important. For example, plant developmental processes that have previously been modelled are the circadean clock[189–192], auxin signalling[193–196], floral organ development[168,197–199], and the regulation of flowering time by photoperiod[200,201]. The gene regulatory network modelling studies discussed in the previous section required detailed information for the regulatory connections between genes, and often large numbers of parameters had to be estimated. To have such in-depth models for each regulatory pathway that can be adequately modelled with gene regulatory networks would lead to a vast increase in complexity for crop models that incorporated them. This could be overcome by using the more in-depth regulatory modules to help parameterize the broader crop models, or identify changes in regulatory logic that will influence the results of the model. Some genes may also have pleiotropic effects, influencing multiple pathways. Ordinarily, with crop models that have a modular structure[188], this would require parameters to be changed in each module in which the gene plays a role. Being able to determine which genes are likely to exhibit pleiotropic effects by their location in regulatory networks would allow these parameters to be estimated together, or for those particular modules to be more intimately linked in the model.

A number of the models discussed in the previous section were parameterized or validated using plants that lack parts of the regulatory network. Therefore, aspects of gene regulatory networks such as the presence or absence of nodes and edges could be estimated from both genome sequencing and transcriptome profiling. Sequencing of four *B. napus* varieties with varying flowering times and vernalization requirements uncovered variation in flowering time genes that were mapped onto regulatory networks[131]. This revealed which copies

of the genes were likely to be causative of the phenotypes displayed by the plants. The cost of sequencing now facilitates variety-specific genome sequences to be generated[202], as has been done with Arabidopsis[203]. Whereas crop models currently require crop growth data in order to parameterize models to particular varieties, future models may be able to combine sequencing data with gene regulatory networks to aid the process of parameterization. Regulatory networks therefore have the capacity to act as a bridge to allow sequencing data to be incorporated into crop models. The difficulty arises in translating knowledge of regulatory networks that have been elucidated in model organisms, the challenges of which will be discussed in the following section.

1.4 Challenges of knowledge transfer from Arabidopsis to Brassicas

The central challenge of moving gene regulatory networks from Arabidopsis to *B. napus* is a consequence of the genome multiplication events that have occurred in the crop[107,112,113,118]. Genome multiplication events have contributed to adaptive radiations[204], speciation[205], and increases in organism complexity, as a result of the additional copies of genes introduced. The presence of additional copies reduces the selective pressure on genes, allowing mutations to occur in genes with limited phenotypic effects. Over time these mutations can result in genes acquiring novel functions (neofunctionalization), losing a subset of their original function (subfunctionalization), or becoming nonfunctional[206]. In this way, genome multiplication events provide evolutionary 'raw material'. A major challenge when translating knowledge from Arabidopsis to *B. napus*, therefore, is to determine how copies of a gene have diverged, and whether the function of the gene in the model plant can be used to infer the function of genes in the crop.

This problem is exacerbated when it comes to regulatory networks. If a whole genome duplication occurs, not only is a transcription factor present as multiple copies but so are its targets, leading to a huge increase in the number of possible regulatory links. If we take the total number of regulatory interactions present

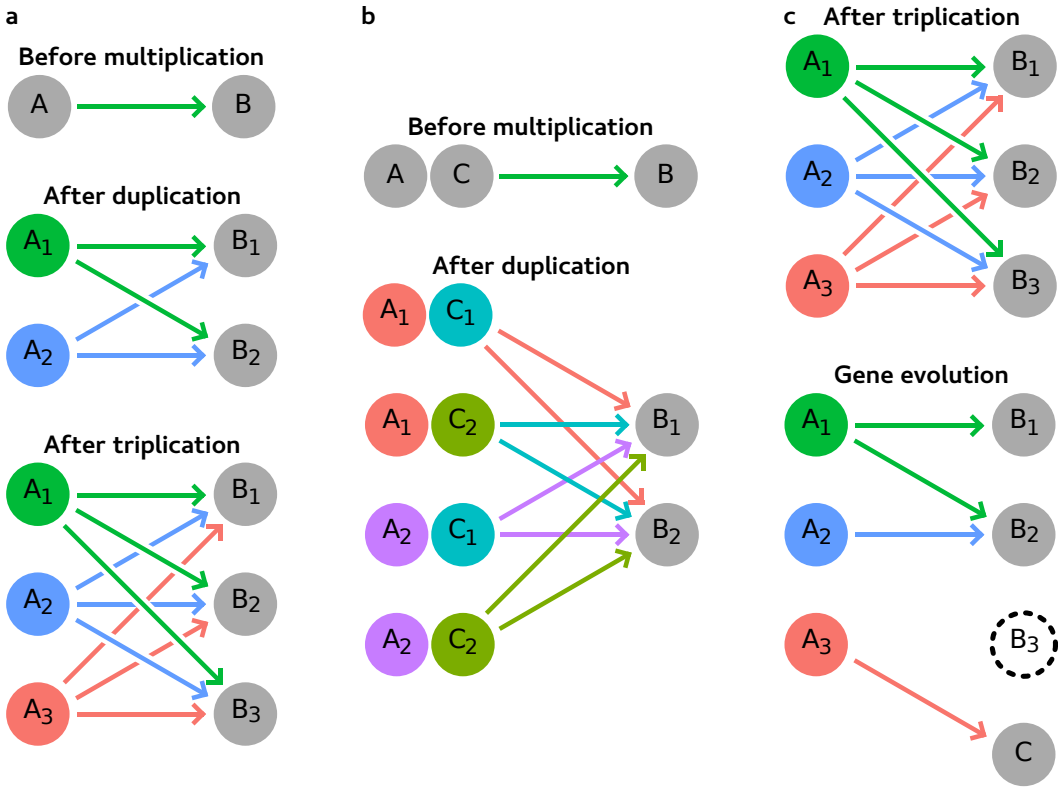


Figure 1.2: Whole genome multiplications lead to a vast increase in the number of regulatory interactions.

a Regulatory links (arrows) between transcription factors (A_x) and their targets (B_y) increase in a quadratic manner following successive multiplication events. **b** The increase in the number of regulatory links is cubic for dimers, where A_x and C_z are able to form dimers. **c** Over evolutionary time, regulatory links may be lost (A_2 to B_1), novel regulatory links may form (A_3 to C), and genes may be lost (B_3).

between genes in an organism to be n, a genome duplication event will cause this number to increase to $4n$. For a genome triplication, this number increases to $9n$ (Figure 1.2a). In general, the number of regulatory links after a genome multiplication event, assuming no dimerization of transcription factors, will be nm^2, where m is the number of times the genome was multiplied. If the original regulatory interaction before the multiplication involved a complex of proteins as the regulator, however, the number of potential regulatory interactions post-multiplication is greater than nm^2. In the case of dimers, using the same definitions of n and m as given above, the increase in the number of regulatory links after a multiplication event is nm^3 (Figure 1.2b). For a complex of p proteins, the number of regulatory links present after a multiplication event is $nm^{(p+1)}$. Therefore, taking a regulatory network elucidated in and validated using Arabidopsis and using it to make predictions for *Brassica* crops is problematic. Without knowing which copies of genes have diverged in function and which have retained their function, all copies of each gene would have to be used in the model. The resulting model would be unwieldy to use and would offer very little insight. It is therefore pertinent to understand how copies of genes have diverged before using the regulatory network from Arabidopsis to aid the construction of *Brassica* regulatory networks.

This thesis will investigate the divergence of gene copies in *B. napus* on a genome-wide scale, with a particular focus on the flowering time genes. This was accomplished by generating a transcriptomic time series collected before, during and after the floral transition.

The first chapter explains how the data was collected and motivates the experimental design decisions taken. Using only data from a spring *B. napus* variety, I reveal that flowering time genes have been preferentially retained in the *B. napus* genome. Widespread divergence in the pattern of regulation between copies of homologous genes suggests that this could have contributed to the observed retention. An in-depth assessment of regulatory divergences between key floral integrators is conducted. The chapter concludes with two case studies investigating sequence divergence for *B. napus* homologues of *TFL1* and *FD*. In the case of *TFL1*, the sequence divergence correlates with regulatory divergence, whereas the sequence divergence in *FD* potentially influences the molecular function of the gene.

The second chapter focusses on a winter variety of *B. napus*. The effects of a vernalization requirement on the global transcriptional landscape are studied by assessing the extent of variety-specific expression. Regulatory divergence in the genes involved in the vernalization pathway are assessed and compared to the expression of the same genes in the spring variety. The comparison between a spring and winter variety allows the vernalization response to be assessed for each copy. Finally, the effects of a cold requirement for flowering on the expression of floral integrators are studied to determine if certain copies are more vernalization sensitive than others.

The final chapter details a web resource, created to allow the dataset collected to be interrogated in a user-friendly and intuitive manner. The dataset can be searched using Arabidopsis gene names to identify *B. napus* homologues and displays the expression patterns of these homologues in both varieties and in both tissues sampled. Alternatively, *B. napus* genes can be searched using sequence homology. Although flowering time genes are the focus of this thesis, the approach taken in the first two chapters to assess regulatory divergence can be carried out using any gene family. The creation of this website allows researchers to study their own genes of interest without the need to download large datasets or carry out laborious read alignment.

Chapter 2

Homologue divergence in a spring variety

2.1 Introduction

The fate of duplicated genes following duplication has been studied in a range of species[207–210], and in a range of theoretical contexts[211–216]. Ultimately, duplicated genes need to provide an advantage to the organism or they will be lost[215]. Early discussions suggested that duplicated genes become mutated and acquire novel, evolutionarily advantageous functions, a process termed neofunctionalization[211]. However, as deleterious mutations occur more frequently than beneficial mutations[217], under this model the expected rate of gene retention following duplication is very low[218], with the majority of duplicated genes acquiring mutations that lead to them being silenced[212]. To account for this, the duplication-degeneration-complementation (DDC) hypothesis[213], posited that multiple copies of genes are maintained through a partitioning of ancestral gene functions among the duplicated genes, a process termed subfunctionalization. Another method of subfunctionalization has been described as escape from adaptive conflict[216]. In this scenario, multiple functions of a gene cannot be mutually optimized, with enhancement of one function occurring at the detriment of the other. Upon gene duplication, selection will favour each gene becoming adaptively specialized to a particular function, leading to subfunctionalization.

A further method of gene retention in a genome following gene duplication is gene redundancy. Redundancy can be defined as genetic redundancy, in which gene loss is compensated for by another gene, or functional redundancy, in which two genes may be functionally similar but loss of one of the copies can still result in deleterious phenotypes manifesting. Genetic redundancy led to the the idea of responsive backup circuits, in which duplicated genes are retained in the genome to provide robustness to gene loss, but also buffer against stochastic effects during development[219,220]. Functional redundancy can be explained by the gene dosage hypothesis, which posits that duplicate genes are retained to maintain the correct protein stoichiometry[214,221–224]. Such dosage effect may result if the gene product acts as part of a protein complex, where an incorrect stoichiometry of proteins can lead to deleterious phenotypes[222]. Interestingly the type of duplication event is predicted to influence whether dosage effects result in gene retention, or favour gene loss. The two main classes of gene duplication event are small scale duplications and whole genome duplications[225,226]. After whole genome duplication events the original protein stoichiometry is maintained. In this scenario, selection will tend to retain dosage sensitive genes in the genome[222,224,227]. Conversely, small scale duplications of dosage sensitive genes lead to incorrect protein stoichiometry, with selection favouring loss of gene duplicates[214]. Evidence from many species are consistent with gene dosage effects maintaining duplicate genes in the genome[228–230]. An interesting observation from these species, and from simulation studies[231], is that certain classes of genes are found to be retained in the genome. This includes genes whose products tend to form protein complexes, such as proteins involved with signal transduction, transcriptional regulation, protein binding and modification, and kinase activity. In *Saccharomyces cerevisiae*, genes retained following whole genome duplication are also genes found to have phenotypic effects when silenced or overexpressed, indicative of the genes being dosage sensitive[210]. An expectation of the gene dosage hypothesis, observed in *S. cerevisiae*[227], is that genes maintained via gene dosage will tend to be co-regulated[224,227]. Assessing the contribution of each of these potential methods of gene retention can therefore be achieved by studying the retention and developmental expression patterns of homologous genes across the entire genome.

Extensive numbers of genes have been lost from the *B. napus* genome, which can be simply assessed by comparing gene numbers with Arabidopsis. One would expect, given the hexaploid *Brassica* ancestor[112,113] and the interspecies hybridization to give *B. napus*[107], a six-fold difference between the number of genes in the *B. napus* genome and the number in the Arabidopsis genome. That the actual fold difference is closer to four (101,040[118] relative to 25,498[10]) illustrates the extent of gene loss in *B. napus*. Despite this, in line with expectations from the gene dosage hypothesis, duplicated genes associated with the circadian clock are retained in the *B. rapa* genome[148]. This observation, and the fact that the majority of flowering time genes in Arabidopsis are transcription factors that form protein complexes[103], suggests that gene dosage effects may be influencing the retention of flowering time genes in *Brassica* genomes.

In order to investigate gene retention in *B. napus*, particularly of the flowering time genes, a transcriptomic time series experiment was designed and the data collected. This chapter will introduce this dataset and the quality control checks performed on it. Global trends in the data reveal the tissue specificity of the expression data and the behaviour of key developmental pathways and protein families. The expression data collected supports the observation of preferential retention of flowering time genes in the *B. napus* genome. Comparative analysis and clustering techniques revealed that the regulation of flowering time genes has diverged, potentially influencing the retention of the genes in the genome. The regulatory divergence observed in key floral integrators provides evidence for some of these genes aquiring novel functions in the plant. Finally, sequence divergence between *B. napus* homologues of two floral integrators, *TFL1* and *FD*, is discussed. In the case of *TFL1*, using knowledge of cis-regulatory elements downstream of the Arabidopsis *TFL1* gene, sequence variation is identified that correlates with the observed regulatory divergence. In contrast, the sequence divergence identified between the *B. napus* homologues of *FD* genes is within the coding region, and is predicted to cause differences in dimerization affinity between the homologues. These case studies highlight that, in addition to potential gene dosage effects, regulatory divergence (*TFL1*) and sequence divergence (*FD*) may also influence gene retention.

2.2 Transcriptome time series design, quality control, and trends

To assess regulatory divergence at the level of the whole genome, a transcriptomic time series was collected for *B. napus*. In order to focus on divergence between *B. napus* homologues of flowering time genes, the time series was collected during the floral transition. As both the leaf and the apex are key organs in the regulation of flowering time[13,15], these two tissue types were sampled at each time point. As a vernalization requirement is a key agronomic trait for *Brassica* crops[127], both a winter and a spring variety were grown. Comparing the expression of genes between winter and spring varieties has been used to as a method of determining vernalization responsive genes[232]. Indeed, regulatory divergence between potential vernalization sensitive genes may only be apparent when making this type of comparison. Once the samples were collected, a number of downstream quantification and quality control steps were necessary to ensure the reliability of the data. This section will discuss how the samples were collected, justifications for the experimental design, and the downstream analysis steps carried out. General regulatory trends observed in the data are also presented. Decisions regarding the design of the experiment, and the sample collection, were made in collaboration with Dr. Rachel Wells, Dr. Nick Pullen, Dr. Martin Trick, Dr. Judith A. Irwin, and Prof. Richard J. Morris[1].

2.2.1 Experimental design and sample collection

In order to investigate the control mechanisms for flowering, suitable tissues were sampled from *B. napus* plants. Two key tissues in which floral genes are expressed are the apical meristem and the leaves[13,15]. Due to the role leaves play in light capture and plant primary metabolism, samples from that tissue allow for the circadian clock[17] and photoperiod pathways[233] to be studied. The expression of *FLC* in plant vasculature also implicates the tissue in the vernalization pathway[29,31]. In addition, the leaf is the site of *FT* expression, with FT

[1]Preprint paper available at https://doi.org/10.1101/178137 and Appendix C.

protein transported to act at the apical meristem[44–46]. The majority of the floral integrators (section 1.1.2) are expressed in the apex[20,20,47,49,55,74,80,85,233,234]. However, genes involved with the vernalization[29,31] and ageing[235] flowering time pathways also have been shown to be expressed in the apex.

To ensure biologically equivalent tissue was collected at each time point, the first true leaf (the first leaf formed after the cotyledons open) was sampled. An alternative would have been to sample the most recently opened true leaf. To sample biologically equivalent new true leaves, one would ideally collect the tissue a fixed number of days after leaf opening. However, as the sampling dates were based on floral development (discussed below), the age of the new true leaves when sampled would likely not be consistent within or between varieties. In addition, determining whether a leaf has fully opened introduces subjectivity into the sampling. Therefore, the first true leaves were sampled at each time point. A consequence of collecting the first true leaf is the tissue ageing over the course of the time series. As plant age plays a role in promoting flowering[39,235], sampling an ageing tissue can potentially allow the role of the ageing pathway to be assessed.

Sampling biologically equivalent apex tissue required removing as much of the surrounding leaf and stem tissue as possible. As angiosperms develop, two collections of stem cells give rise to the entire plant[236]. The shoot apical meristem generates the above ground organs of the plant, forming leaves, stems, and floral structures, while the root meristem forms the below ground organs. The shoot apical meristem itself is composed of a mass of stem cells surrounded by leaf primordia, with floral integrators expressed within the meristem itself[237]. To ensure that the apex samples were enriched for the meristem tissue, the surrounding leaf and stem tissue was removed by hand dissection using a razor blade. Although the method does not achieve the spatial resolution achievable with laser microdissection[238], it is still able to suitably enrich for apex tissue (Figure 6.3; Appendix A). Measuring gene expression in biologically equivalent leaf and apex tissue allowed for the genes from key flowering pathways to be studied throughout the floral transition.

To capture transitions in gene expression relevant to flowering time genes, the time points during development at which plant tissue was sampled were carefully chosen. A schematic of the sampling scheme is displayed in Figure 2.1. As

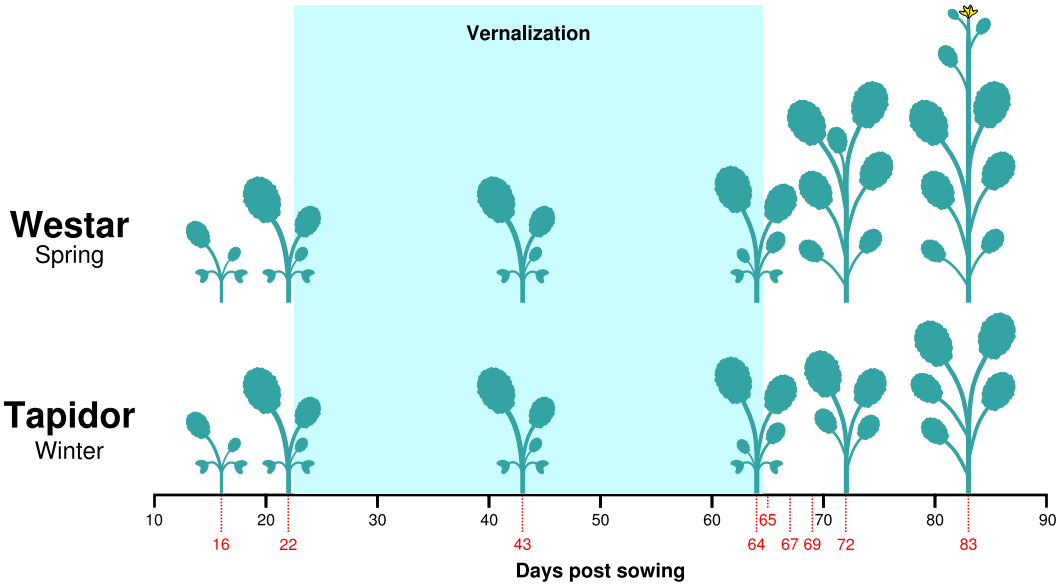


Figure 2.1: The sampling scheme for the transcriptome time series. Red numbers displayed below the bottom axis indicate the time points at which the plants were sampled. The representations of the plants indicate the approximate number of full leaves at those time points.

with previous studies investigating the vernalization response[143,232,239], spring and winter varieties of oilseed rape, called Westar and Tapidor respectively, were grown. Seeds from both varieties were sown and plants grown under long day conditions (16 hours of light) with controlled temperatures of 18 °C during the day and 15 °C at night. The six week vernalization treatment involved growing the plants in short day conditions (8 hours of light) at 5 °C. Plants were sampled at approximately 12:00 each day. During long and short days, 12:00 corresponded to the midpoint of the day. Although this means that the time since dawn was different depending on the day length, the proportion of day elapsed was the same in both conditions. Sampling at a proportionally similar time each day minimizes the noise due to circadian rhythms, as the oscillations have been observed to become entrained to light and dark cycles[240]. Vernalization was necessary in order to accelerate the onset of flowering for the winter variety, Tapidor. Although the spring variety does not have a vernalization requirement, plants still exhibit a vernalization response[241]. To investigate the facultative vernalization response of the spring variety, and to

ensure that data from the two varieties would be comparable, both Westar and Tapidor plants were vernalized. To establish an appropriate pre-vernalization baseline of gene expression, two time points were sampled before the cold; one after two weeks of growth and another the day before the plants were transferred into vernalization. A potential confounding factor the vernalization treatment introduces is a change in both temperature (15 - 18 °C to 5 °C) and day length (16 hours to 8 hours). Both changes in growth conditions were required to make the vernalization treatment as physiologically accurate as possible, as short days accompany the cold temperatures of winter. However, transcriptional changes due to altered photoperiod[19,242] and temperature[243,244] have the potential to obscure the response of genes to vernalization. To differentiate between expression responses that result from these different flowering pathways, two time points were sampled during the vernalization period; one halfway through the treatment, after three weeks of cold, and another the day before the treatment ended, after six weeks of cold. From results in Arabidopsis, it is known that gene expression responds to changes in photoperiod[19,242] and ambient temperature[243,244] on the order of hours or days. The vernalization response, however, changes gene expression over the course of weeks[29,245]. The mid-vernalization time point allowed for these two transcriptional time scales to be resolved, while the time point at the end of cold acts as a reference point for the transcriptional changes that occurred post-cold. Sampling after the vernalization period was much more frequent as rapid developmental changes were expected to occur after the plants were returned to warmer temperatures and long day conditions[19,242]. Tissue was collected 1, 3, and 5 days post-vernalization to capture these expected shifts in the transcriptome. To ensure that the developmental time period sampled for each variety was comparable, the final two time points were sampled when the plants had flower buds visible from above (BBCH stage 51[246]). For the spring variety Westar this developmental stage was reached 8 days post-vernalization, whereas for Tapidor the final time point was sampled at 19 days post-vernalization. Therefore, although the age of the spring and winter plants at the final relevant time point (when the plants reached BBCH stage 51[246]) differed, the developmental time period sampled for the two varieties is very comparable.

2.2.2 Reference genome sequence and gene models

In order to carry out RNA-Seq, short reads obtained from the sequencing run have to be aligned to a suitable reference sequence[247]. For *B. napus*, three different reference sequences are available. The set of *B. napus* unigenes is a community resource generated using expressed sequence tags from *B. napus*, *B. oleracea*, and *B. rapa*[248]. The aim with the unigene construction was to resolve gene models of orthologous genes, such as homoeologous genes on the A and C genome, and paralogous genes, which arose from the ancestral genome triplication event in the *Brassica* lineage[112,113]. The pan-transcriptome resource is in many ways an updated version of the unigenes, utilizing published coding DNA sequences (CDS) for *B. napus*, *B. oleracea*, and *B. rapa*[249]. To generate the resource, CDS models from the two diploid species were aligned to their respective reference genomes. Gene models from the *B. napus* reference genome[118] were then compared to the CDS models from the diploid species, and any *B. napus* gene models that did not match any CDS model from the diploid species was added to the pan-transcriptome[249]. The final main reference available was the *B. napus* reference genome sequence itself, sequenced from a European winter variety of oilseed rape called Darmor-*bzh*[118]. While the unigenes and the pan-transcriptome consist of tens of thousands of individual gene models, the reference genome consists of genomic sequence arranged into chromosomes. The advantage of such a reference is that gene models can be viewed in a genomic context. In addition, the Tuxedo suite of tools used to perform the quantification can more readily estimate total gene expression, combining the expression from all isoforms of a gene, when a genomic reference is used[250]. To take advantage of these benefits, the *B. napus* genome sequence was used as the reference sequence for the transcriptomic time series.

The Tuxedo suite of RNA-Seq tools is able to predict gene models from RNA-Seq reads without prior knowledge of gene models[250]. This is possible due to TopHat aligning reads in a splice-aware manner[251], allowing the intron structure and the splice variants of genes to be discovered. Aligning RNA-Seq reads obtained from the time series samples to the *B. napus* genome sequence using the Tuxedo suite resulted in two problems. The first manifested in instances when neighbouring genes were oriented on opposite strands with transcription

occurring in the direction of the other gene. Due to transcriptional read-through, reads were obtained that spanned the gap between the genes, causing the prediction algorithm to combine the genes into a single gene model. These chimeric gene models resulted in aberrant expression traces being generated. The other problem arose as a result of genes that had undergone tandem multiplication events, such that multiple copies of the gene were located relatively close to each other in the genome. In these cases, reads that spanned across two exons would occasionally be aligned partially to one gene in the tandem array and partially to another. This lead to large gene models being predicted that spanned multiple genes in the tandem array. The chimeric gene models created as a result of these issues lead to additional reads mapping to these gene models, affecting the expression level quantification.

To address these issues, predetermined gene models were used to quantify gene expression. The Darmor-*bzh* reference genome was published with gene models predicted by *ab initio* gene prediction, RNA-Seq data, and mapping *A. thaliana*, *B. rapa*, *B. oleracea*, and *Oryza sativa* protein sequences to the genome[118]. These different sources of data were combined using the software GAZE[252] and were weighted differently based on the researchers' confidence in the data[118]. However, weighting the data sources introduces subjectivity to the gene models. To overcome this problem, and to maximise the number of genes included in the transcriptomic time series, an approach utilizing short reads obtained from the time series samples was taken. The gene model prediction software AUGUSTUS[253] was used to combine evidence of gene models from the RNA-Seq data directly into the Hidden Markov model based prediction process. RNA-Seq reads from both tissues and varieties across the entire time series were pooled and used to aid the prediction of exon-intron boundaries. While the Darmor-*bzh* gene models were also directed by transcriptomic data, the short reads used were obtained from roots, stems, leaves, and flower buds in low and high nitrogen conditions[118]. Potentially important floral genes that are expressed during the period of development addressed in the current study, therefore, may not have been represented in this dataset. By using the short reads from the transcriptomic time series to aid the generation of gene models, however, this problem is mitigated.

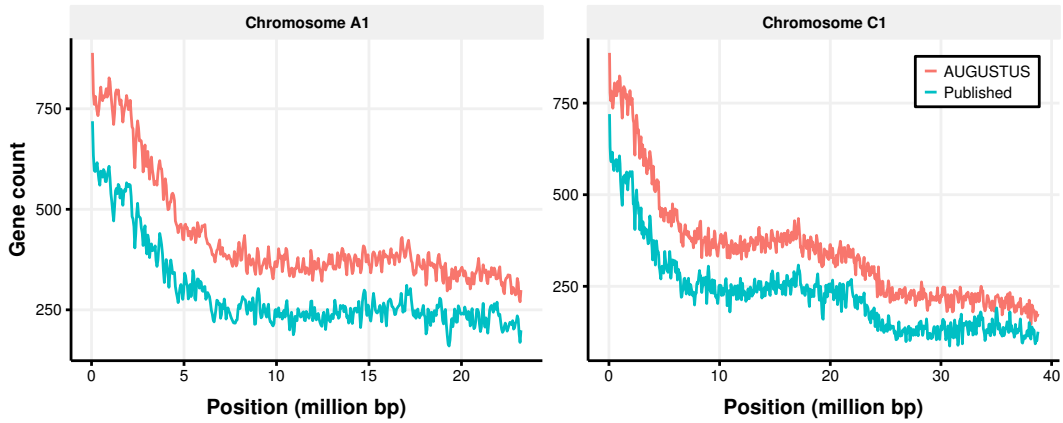


Figure 2.2: Gene density is increased consistently across chromosomes with the AUGUSTUS derived gene models relative to the published gene models.
Gene count is calculated using a 100 kbp sliding window across the chromosome. The patterns shown here are representative of the patterns seen across all chromosomes.

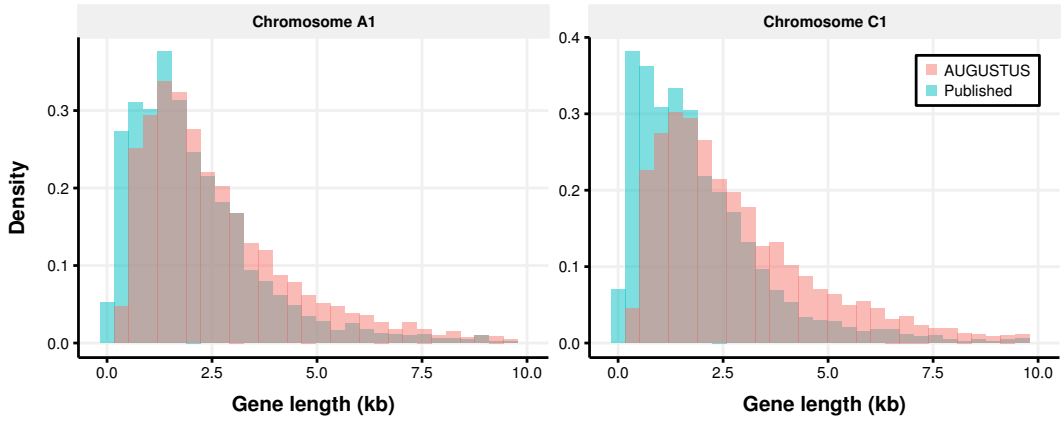


Figure 2.3: AUGUSTUS derived gene models tend to be longer than published gene models.
Gene length is calculated as the length of the unprocessed mRNA transcript. The patterns shown here are representative of the patterns seen across all chromosomes within a genome.

The number of gene models obtained from AUGUSTUS[253] was 155,648, while the number of published gene models for the *B. napus* reference sequence is 101,040[118]. To investigate whether the gene models were distributed in the same way across the genome, the density of genes across the genome was calculated for both sets (Figure 2.2). The gene density across the chromosomes is correlated between the two sets of gene models (Figure 2.2). This result indicates that similar proportions of genes are located in the same regions of the genome in both gene model sets, despite the AUGUSTUS-derived models exhibiting greater gene density. As gene density is greater for the AUGUSTUS-derived models, one may expect that the length of the gene models would be reduced due to models being split. In order to test this, distributions of gene model lengths were calculated. The AUGUSTUS-derived gene models (mean length of 363 bases) are on average longer than the gene models published with the Darmor-*bzh* genome sequence (mean length of 245 bases; Figure 2.3). Taken together, these results suggest that the AUGUSTUS-derived gene models better represent the genes present in the *B. napus* genome. This is due to the greater number of AUGUSTUS-derived gene models relative to the published gene models, that are not a consequence of gene models becoming split. Additionally, the AUGUSTUS-derived gene models were able to resolve chimeric gene models formed as a result of convergent transcription of genes and tandem arrays of similar genes, discussed earlier. As a consequence of these benefits relative to the published gene models, the AUGUSTUS derived gene models were used to guide the RNA-Seq quantification process.

2.2.3 Aligning reads and quantification of expression levels

To quantify gene expression using RNA-Seq, short reads have to be aligned to the chosen reference sequence to allow gene expression levels to be estimated and normalized. There exist a number of different methods for quantifying the expression level of genes using short read data. A frequently used pipeline involves the Tuxedo suite of tools[250]. The pipeline consists of first aligning the short reads using Bowtie[254,255], an alignment algorithm that makes use of the Burrows-Wheeler transform of genomic DNA sequence to allow for very

efficient alignment. Bowtie is used by another part of the Tuxedo suite called TopHat[251]. TopHat is a splice aware aligner; if a particular read does not align to the reference sequence then the read is segmented and the individual segments are aligned separately[251]. In this way, reads that span exon-exon boundaries can be detected, allowing different splice isoforms to be detected and their expression quantified. Finally, once the reads are aligned, Cufflinks is used to quantify gene expression[256]. This is done in a probabilistic manner that takes into account both the error measured from different biological replicates and the uncertainty in read mismapping. The latter arises when reads align with equally high alignment scores in multiple places in the genome. Instead of removing these reads from further analysis, which has the potential to discard a lot of the sequencing data collected, Cufflinks is able to incorporate this uncertainty into the error associated with the expression measurement[256]. A more recent RNA-Seq analysis pipeline involves the pseudoalignment of reads to a reference transcriptome. Kallisto assigns reads to transcripts based on k-mer matching between the read and the transcript[257]. In order to take into account ambiguous read mapping, Kallisto implements a bootstrap technique that resamples the read assignments. This bootstrapping technique is made possible due to the speed with which Kallisto runs and allows for the technical variation within a sequencing run to be estimated[257]. While the speed and technical variation estimation of Kallisto are advantages over the Tuxedo suite, the software requires transcript sequences in order to be run. In the case of *B. napus*, splice isoforms are less well categorized than for other species, such as Arabidopsis. Additionally, the downstream statistics pipeline for Kallisto[258] is designed to carry out differential expression analysis using RNA-Seq data, rather than estimating expression levels taking into account technical and biological noise. Due to these issues with Kallisto, and as the Tuxedo suite is a mature suite previously used in other *B. napus* RNA-Seq studies[259,260], the latter was used to quantify gene expression.

To quantify gene expression for the the transcriptomic time series, short reads were aligned to the *B. napus* reference genome[118] using the AUGUSTUS-derived gene models (discussed in Section 2.2.2). Initially, only short reads from a single sequencing run were available for each sample, with an average of 67 million reads per sample obtained. Of these total reads, 82% were mapped

to the reference genome. The confidence intervals calculated by Cufflinks using this sequencing data, however, were too large to allow confident conclusions to be drawn from the data. A hypothesis for why this was the case is that Cufflinks did not have information from biological repeats to properly calculate confidence intervals. In the absence of multiple measurements for a sample, Cufflinks treats all samples as repeats of each other in order to parametrize the error model used[256]. To test if this was the case, gene expression values were calculated separately for the two tissues. If the large confidence intervals were indeed due to the lack of repeat measurements, it was expected that only using samples of the same tissue type to parameterize the error model would result in smaller confidence intervals being calculated. Performing the analysis in this way lead to a general reduction in the size of the confidence intervals calculated for each expression level estimate (Figure 2.4), while not affecting the expression level estimations for genes (Figure 2.5). This suggests that the initial size of the confidence intervals was indeed because samples from different tissues, different varieties, and different points in development were used to calculate the uncertainty in the data.

Although calculating expression values separately for each tissue results in reduced confidence interval sizes relative to both tissues combined, the intervals calculated were still large. To reduce the size of the intervals further, a second set of samples, constituting a biological replicate, were sequenced. To ensure that the uncertainty in expression levels was calculated accurately in both tissues across the entire time series, samples selected to be in the second sequencing run were chosen to span the entire time series. Samples from every time point were not required, as the Cufflinks algorithm uses samples for which repeat measurements are available to parameterize an error model, that is then applied to samples that are lacking repeat measurements[256]. Additional pools of tissue from the apex and leaf, sampled at days 22, 43, 64, 67, and 72 of the time series, were sequenced with an average of 33 million reads per sample being obtained. The pooled samples were composed of different plant tissue to that sequenced in the first sequencing run, making this data a biological replicate. As with the first sequencing run, an average of 82% of reads mapped to the reference sequence. Incorporating the repeat measurements, while also performing the quantification separately for each tissue, resulted in a large

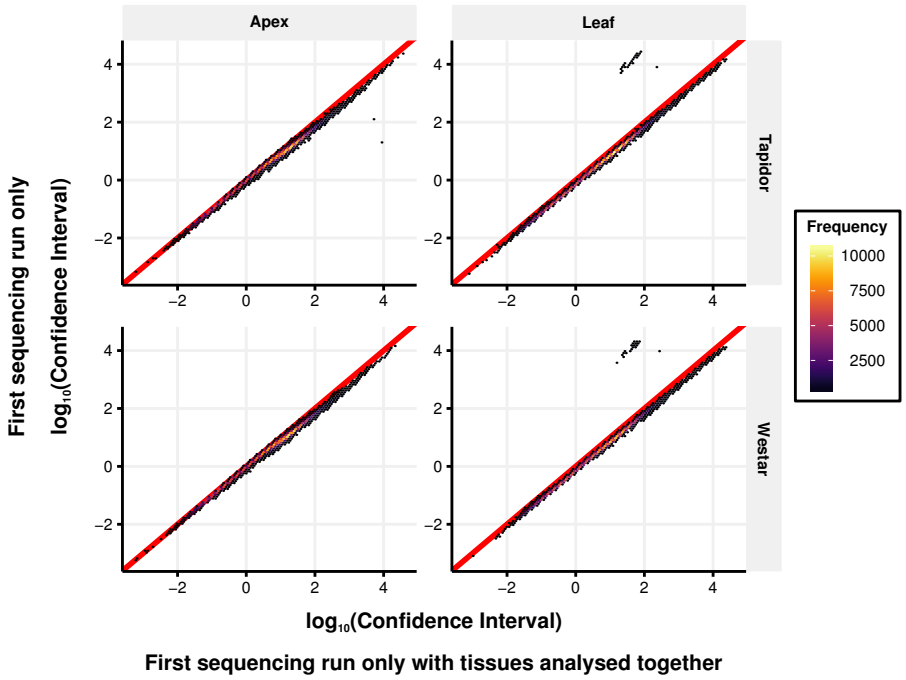


Figure 2.4: Calculating FPKM values for the apex and leaf separately reduces the size of the confidence intervals.

95% confidence intervals were calculated using the same quantification pipeline for both the leaf and the apex samples from the first sequencing run combined (x-axis) or separately (y-axis). The ranges of these intervals were $\log_{10}$ transformed for clarity. That the majority of points lie below the $y = x$ line (red diagonal line) indicates that calculating the confidence intervals separately for each tissue reduces the uncertainty in the expression value measurement. The data is displayed as a two dimensional histogram, where the colour of the hexagonal unit indicates the number of data points mapping to that part of the plot.

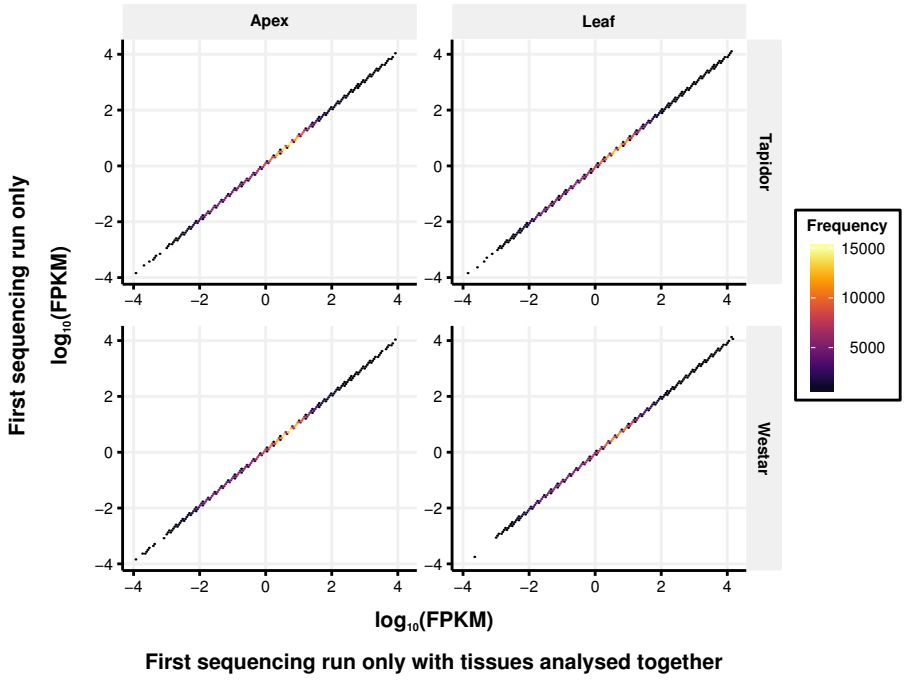


Figure 2.5: Quantifying gene expression for the apex and leaf separately has little effect on FPKM values.

FPKM gene expression values were calculated using the same quantification pipeline for both the leaf and the apex samples from the first sequencing run combined (x-axis) or separately (y-axis). These values were $\log_{10}$ transformed for clarity. That the points lie along the $y = x$ line indicates that both approaches result in similar FPKM values being calculated. The data is displayed as a two dimensional histogram, where the colour of the hexagonal unit indicates the number of data points mapping to that part of the plot.

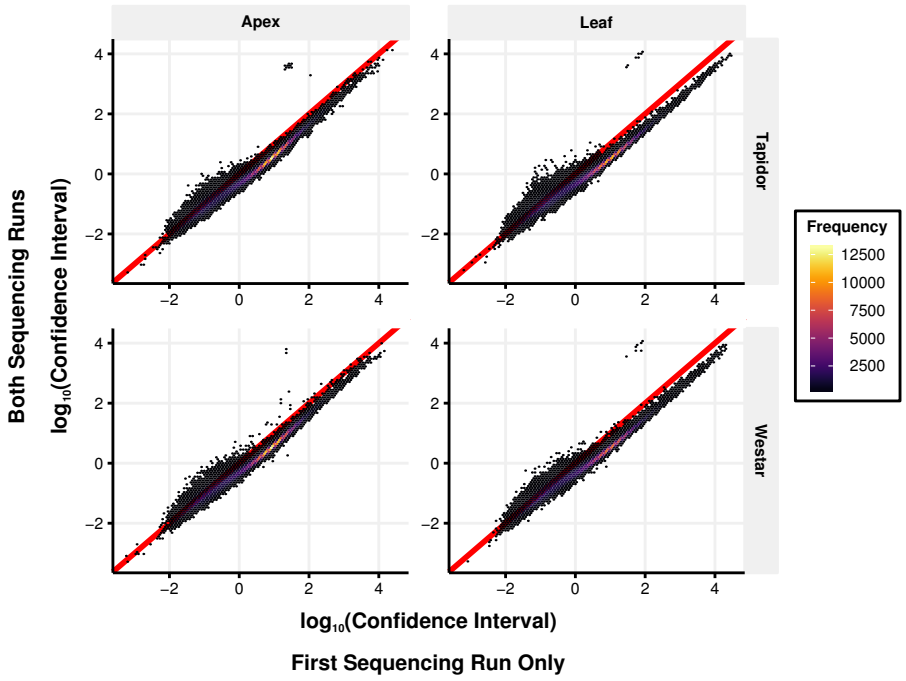


Figure 2.6: Including data from a second sequencing run causes a reduction in the majority of estimated confidence interval sizes.
95% confidence intervals were calculated using the same quantification pipeline for the first sequencing run only (x-axis) or both sequencing runs combined (y-axis). The ranges of these intervals were $\log_{10}$ transformed for clarity. That the majority of points lie below the $y = x$ line (red diagonal line) indicates that calculating the confidence intervals with reads from biological repeats reduces uncertainty in the expression value measurements for the majority of genes. The data is displayed as a two dimensional histogram, where the colour of the hexagonal unit indicates the number of data points mapping to that part of the plot.

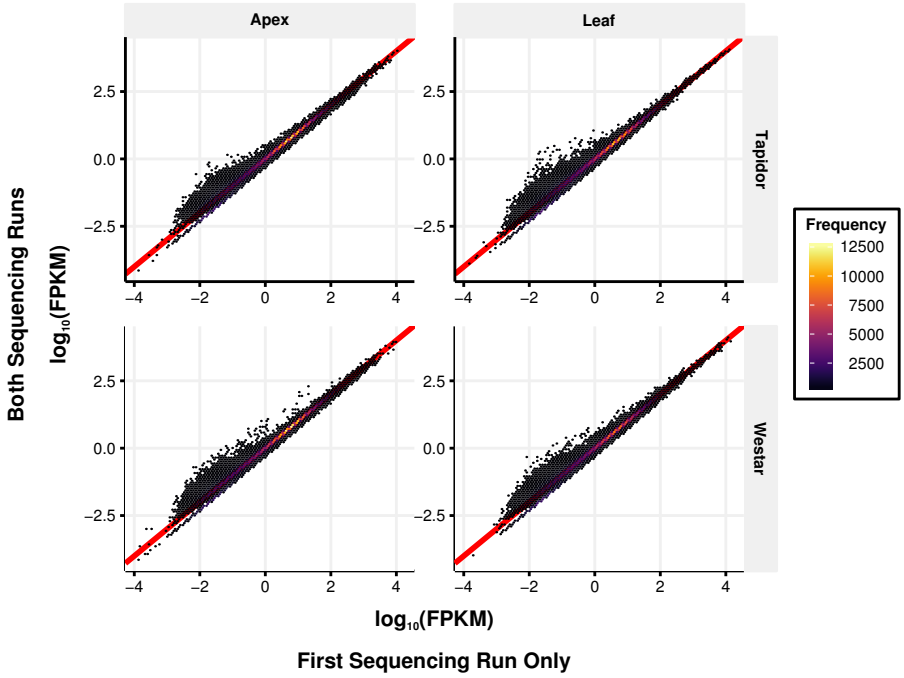


Figure 2.7: Including data from a second sequencing run does not affect the majority of estimated FPKM values.

FPKM gene expression values were calculated using the same quantification pipeline for the first sequencing run only (x-axis) or both sequencing runs combined (y-axis). These values were $\log_{10}$ transformed for clarity. That the highest frequencies of points lie along the $y = x$ line indicates that both approaches result in similar FPKM values being calculated for the majority of genes. The data is displayed as a two dimensional histogram, where the colour of the hexagonal unit indicates the number of data points mapping to that part of the plot.

reduction in confidence interval sizes (Figure 2.6) while having a comparatively small effect on expression levels for the majority of measurements (Figure 2.7). Therefore, the second sequencing run was able to provide enough additional data to reduce the uncertainty in the gene expression level estimations to acceptable levels for further work to be carried out.

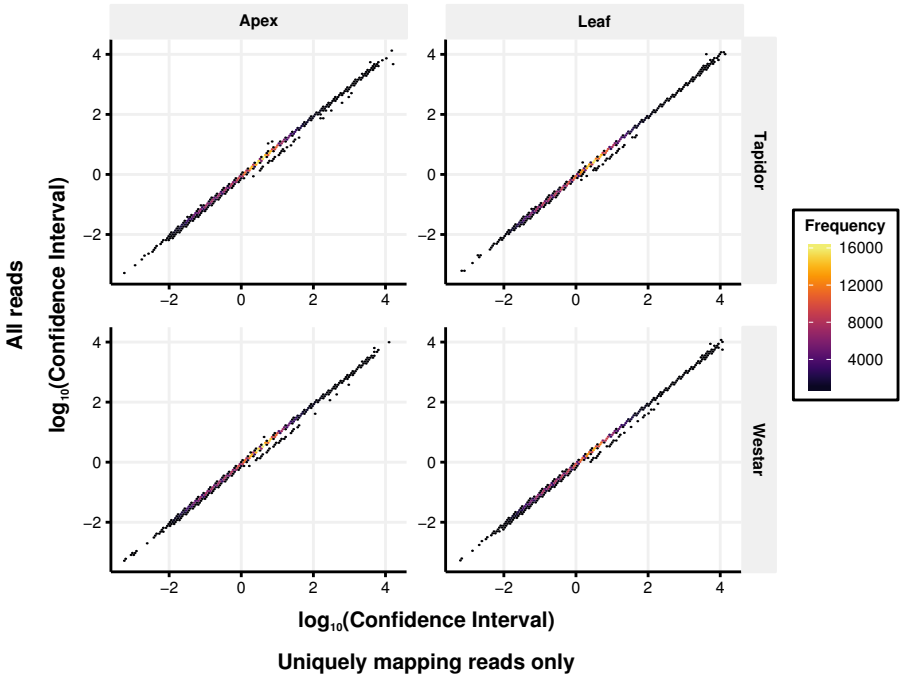


Figure 2.8: Multiply mapping reads have little effect on the estimated confidence interval range.
95% confidence intervals were calculated using the same quantification pipeline for all reads (y-axis) or reads that only align to a single position in the reference sequence (x-axis). The ranges of these intervals were $\log_{10}$ transformed for clarity. That the majority of points lie along the $y = x$ line indicates that both approaches result in similar confidence interval ranges being calculated for the majority of genes. The data is displayed as a two dimensional histogram, where the colour of the hexagonal unit indicates the number of data points mapping to that part of the plot.

A potential issue with RNA-Seq are reads mapping equally likely to multiple positions in the genome. To alleviate this problem, previous studies investigating the differential expression of paralogous genes have only used reads that map

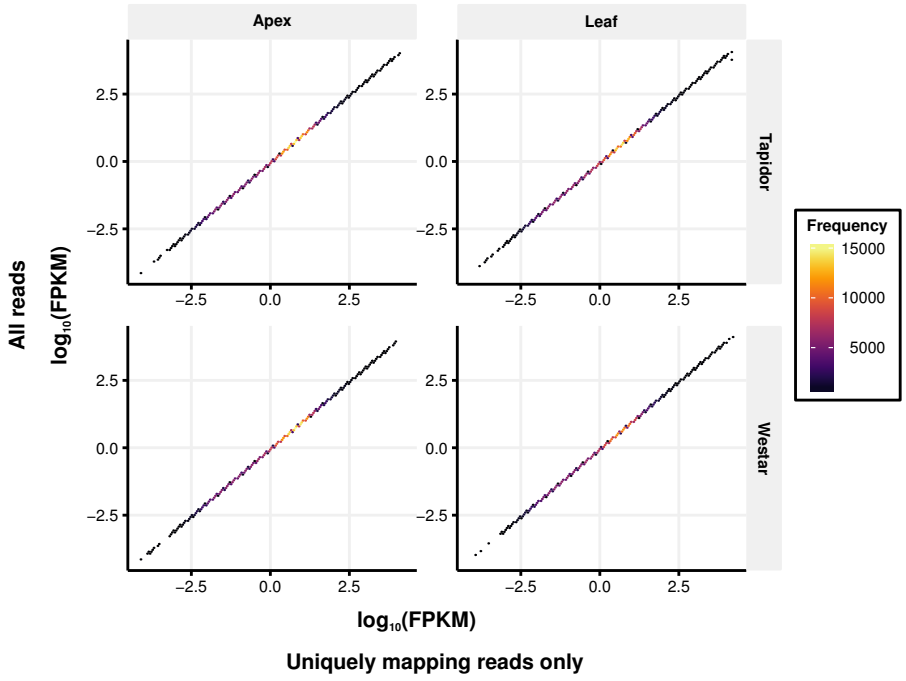


Figure 2.9: Reads aligning to multiple regions of the genome have little effect on the estimated gene expression levels.

FPKM gene expression values were calculated using the same quantification pipeline for all reads (y-axis) or reads that only align to a single position in the reference sequence (x-axis). These values were $\log_{10}$ transformed for clarity. That the points lie along the $y = x$ line indicates that both approaches result in similar FPKM values being calculated for the majority of genes. The data is displayed as a two dimensional histogram, where the colour of the hexagonal unit indicates the number of data points mapping to that part of the plot.

to single positions in the genome to calculate expression levels[261]. Cufflinks is able to incorporate the uncertainty introduced by reads mapping to multiple locations into the calculation of expression level uncertainty[256]. However, the high amount of duplicated sequence in the *B. napus* genome[118] may result in high uncertainty in the calculated expression levels. To investigate whether this was the case, the effect on gene expression levels of reads aligning to multiple positions in the genome was assessed. Of the reads mapped to the genome, 14% were mapped to multiple positions in the genome, with 0.3% in the first sequencing run and 0.4% in the second sequencing run mapping to over twenty positions. To test if reads mapping to multiple locations would affect the expression levels calculated by Cufflinks, the expression level quantification was repeated with these reads removed. Comparisons of FPKM values and confidence interval ranges both reveal very little effect when reads that map to multiple positions in the genome are excluded from the analysis (Figures 2.9 and 2.8). This result demonstrates that reads mapping to multiple positions in the genome are not adversely affecting the calculation of expression levels and are therefore included in the expression level quantification used throughout this study.

2.2.4 Self-organizing map based clustering of expression data

Having constructed the transcriptomic time series, validation was conducted to determine if expected trends were observed in the dataset. In order to assess trends in the data, gene expression profiles across time were clustered using self-organizing maps (SOMs). SOMs adaptively take into account the variation present in the data to ensure that the dataset is properly represented. When used to cluster time series data, each cluster represents a particular expression profile across time, with genes exhibiting a similar expression profile assigned to that cluster. Due to the process by which SOMs are trained to the dataset (Figure 2.10), neighbouring clusters will tend to have similar expression profiles to each other. If particular parts of the dataset are more dense, in terms of the number of data points present, then the training process will explore that part

[2]https://commons.wikimedia.org/wiki/File:Somtraining.svg

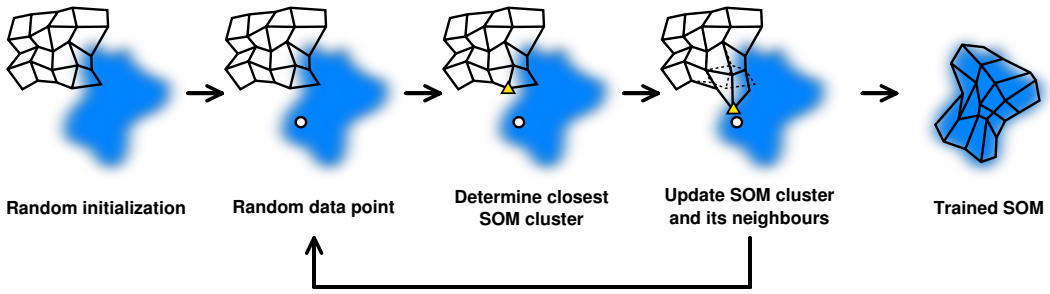


Figure 2.10: Self-organizing maps (SOMs) are trained to represent multidimensional datasets.

SOMs are randomly initiated. Clusters are assigned neighbours based on their Euclidean distances from one another, such that neighbouring clusters have a lower Euclidean distance between them. During the training process, the SOM (black grid) is trained to represent the dataset (blue shape). The training process begins by selecting a random data point. The SOM cluster closest to that data point (yellow triangle), determined by Euclidean distance, is translated closer to the data point. At the same time, the neighbouring clusters are also translated, although to a lesser extent. Another data point is selected and the process repeats. The training process continues until the SOM accurately represents the dataset. Image adapted from a diagram by Mcld[2], distributed under a CC BY-SA 3.0 license.

of the dataset more, leading to a higher density of clusters in that area. The ratio of grid dimensions are set as the same ratio as the eigenvalues of the first two principal components of the data, to maximise the variation captured by the SOM (Section 6.7; Methods). These properties lead to a clustering method that allows for the time series data to be summarized and visualized in an intuitive manner. Only SOMs generated using data from Westar are displayed here, with SOMs generated using data from Tapidor discussed elsewhere in the thesis (Section 3.2.2).

Within the SOM generated using the transcriptomic time series from the apex (Figure 2.11), there are two regions that have a high number of genes mapped to them, represented by clusters 19 and 46. The expression profile of cluster 19 is low at the start of the time series, increases during the cold, and returns to pre-cold levels when the plants are transferred back to growth in warmer conditions. The other region of the map with a high number of genes mapped to it are the clusters located towards the centre of the map, represented by cluster 46. These clusters exhibit an expression pattern that remains largely constant throughout the developmental time series, with an increase in expression towards the final time point (Figure 2.11). These findings suggest that in the apex a large number of genes are responding to the change in growth conditions in the vernalization treatment, that is, short days and 5 °C temperatures. The large number of genes that increase in expression at the final time point may be due to flower buds being formed in the apex, which would require the coordinated expression of many genes.

To determine whether trends similar to the apex would also be observed in the leaf transcriptome, a SOM was generated for the leaf transcriptome time series (Figure 2.12). High numbers of genes mapped to three regions of the leaf SOM; represented by clusters 19, 82, and 99. Cluster 82 exhibits an expression profile that is high initially, decreases during the vernalization period, and remains lowly expressed when plants are returned to warmer growth conditions. This suggests that a large number of genes are becoming stably repressed during the cold period, which may be due to a vernalization response or to effects resulting from the leaf ageing during the time series. Clusters 19 and 99 exhibit similar expression profiles as clusters 46 and 19 from the apex-derived SOM (Figure 2.11). This suggests that, as with the apex-derived SOM, that a large

subset of genes are responding to growth in the cold, short day conditions of the vernalization treatment, while another subset are potentially responding to age effects and the floral transition.

SOMs have been used in previous investigations to cluster gene expression traces[262] and distil general trends from time series expression data[263]. To validate that the transcriptome time series accurately captures important expression profiles, SOMs were used to cluster data from the Westar leaf and apex samples. Both of the SOMs for the leaf and apex reveal that a large number of genes exhibited transcriptional responses to the change in growth conditions that occur when the plants are grown in short days at 5 °C. Transcriptional changes occurring as a result of photoperiod and temperature changes have been observed in Arabidopsis[19,242–244] and ryegrass[264]. That similar expression changes are observed for the *B. napus* transcriptome time series suggests that key expression differences have indeed been captured by the experiment. This result also highlights the importance of subjecting both the spring and winter varieties to vernalization. As discussed in section 2.2.1, studying transcriptional effects of vernalization requires differentiating between vernalization responsive genes and genes that are affected by ambient temperature and photoperiod changes[232]. That a vernalization responsive cluster and a cold treatment responsive cluster are identified in the leaf SOM suggest this differentiation is possible. In addition, that a vernalization responsive cluster is observed in the leaf in Westar, a spring variety, suggests that genes controlling the vernalization response in Westar[241] and the vernalization requirement in Tapidor can be disentangled. Finally, many genes increase in expression towards the final time point in both tissues. This suggests that the transcriptional changes that accompany the transition to floral growth have been captured by the transcriptome time series.

2.2.5 Gene ontology term enrichment

To further investigate general trends that the SOM clustering reveals, enrichment analyses were carried out for gene ontology (GO) terms of interest. Co-expressed genes may be part of the same developmental pathway, or may be co-expressed as a consequence of the way the experiment was designed,

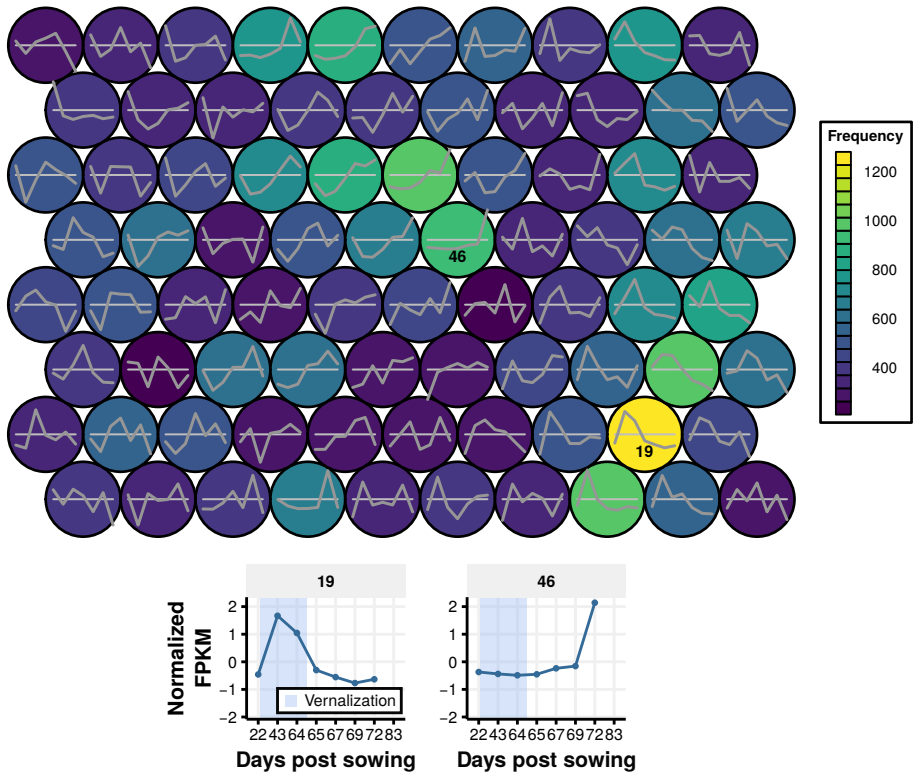


Figure 2.11: SOM generated using the apex transcriptome time series in Westar.

The size of the SOM was chosen such that it captured ~85% of the global squared distance from the mean (Section 6.7; Methods). The grey lines within each SOM cluster indicate the normalized expression profile that particular cluster represents. The SOM is toroidal, such that clusters on the top and bottom rows are adjacent, as are clusters on the left and right hand columns. The colour of the cluster represents the number of genes mapped to that particular cluster. The graphs under the plot correspond to clusters 19 and 46, that represent areas of the SOM with high numbers of genes.

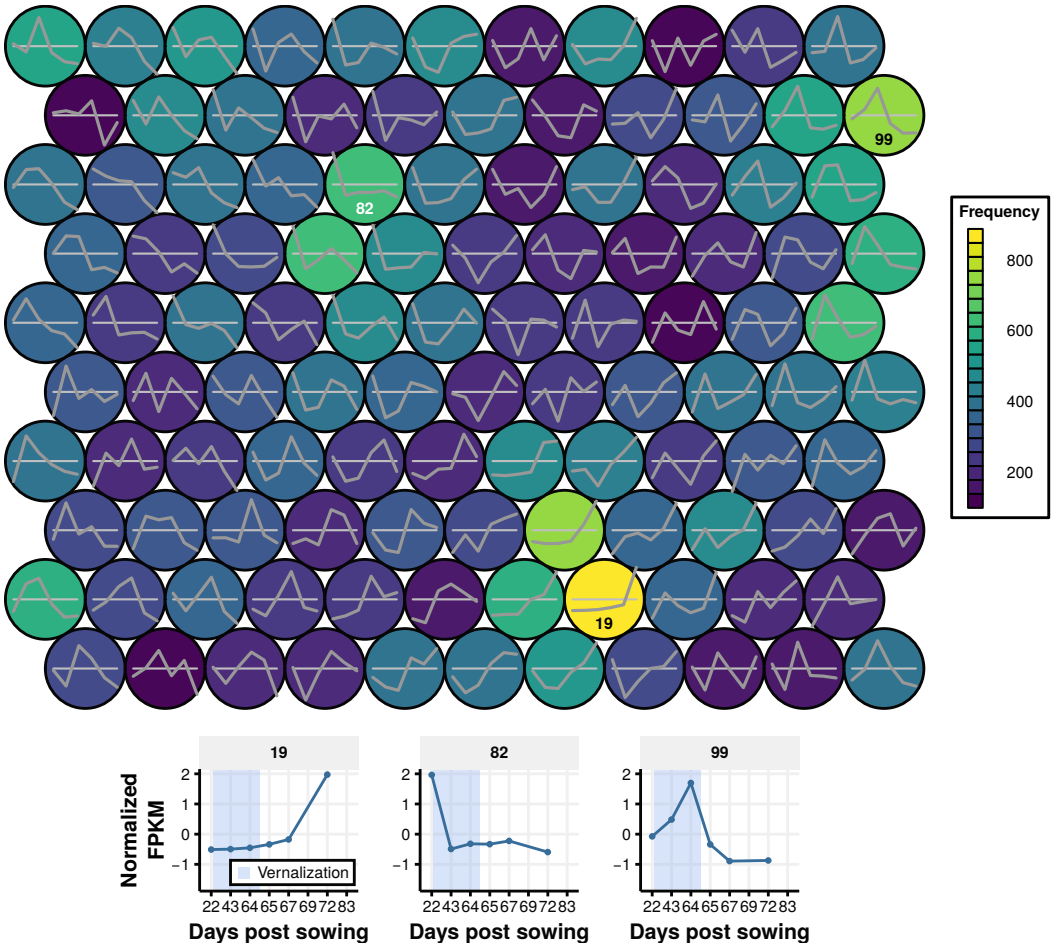


Figure 2.12: SOM generated using the leaf transcriptome time series in Westar. The size of the SOM was chosen such that it captured ~85% of the global squared distance from the mean (Section 6.7; Methods). The grey lines within each SOM cluster indicate the normalized expression profile that particular cluster represents. The SOM is toroidal, such that clusters on the top and bottom rows are adjacent, as are clusters on the left and right hand columns. The colour of the cluster represents the number of genes mapped to that particular cluster. The graphs under the plot correspond to clusters 19, 82, and 99, that represent areas of the SOM with high numbers of genes.

such as simultaneous changes in growth conditions[265]. GO term enrichment is one method of determining whether the observed clustering is biologically meaningful or a technical artefact. GO terms are a precise, fixed vocabulary for describing where in an organism a gene acts, the molecular function of that gene, and the biological process the gene is involved in. When GO gene annotations are available for a particular organism, the proportion of genes annotated with a particular GO term across the entire genome can be determined. If a significantly higher proportion of genes within a subset of genes are annotated with a GO term than would be expected given the across genome proportion, then that subset of genes is said to be enriched for that GO term. To understand the expression dynamics of key developmental pathways during the transcriptomic time series, GO term enrichment was carried out using the clusters identified in the SOM analysis (section 2.2.4).

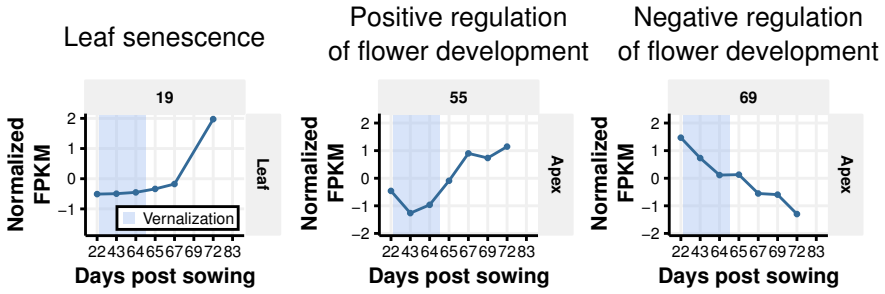


Figure 2.13: Normalized expression profiles for SOM clusters enriched for leaf senescence and regulation of flower development.
Normalized expression profiles for SOM clusters that are significantly enriched for each GO term and that also contain the most *B. napus* genes annotated with that GO term are displayed. The expression patterns of genes associated with "leaf senescence" in the leaf and regulation of flower development in the apex are consistent with phenotypic observations from those tissues.

To establish that GO term enrichment analysis would provide reliable results, and to further validate the transcriptomic time series, the enrichment of GO terms associated with phenotypic observations were tested. During the time series, the first true leaf was sampled at every time point (section 2.2.1). As a consequence, the leaf tissue sampled was older at later time points, and some of the first true leaves had begun to visibly senesce by the final time

point. To test if this resulted in a change in the transcriptome in the leaf, SOM clusters enriched for GO terms associated with "leaf senescence" were identified. The most highly enriched cluster identified in the leaf data for the term "leaf senescence" exhibits an expression pattern that gradually increases across the entire time series, with a large increase in expression at the final time point (Figure 2.13). This suggests that genes associated with leaf senescence are co-expressed in *B. napus*, a finding also observed in the transcriptome of senescing Arabidopsis leaves[266]. The time points selected for the time series were chosen to allow the progression of the floral transition to be investigated (section 2.2.1). An expectation arising from this would be that GO terms relating to flower development would exhibit expression changes across the time series. To test whether this is the case, clusters enriched for the GO terms "positive regulation of flower development" and "negative regulation of flower development" were identified in the apex-derived SOM. The expression of genes annotated with the GO term "positive regulation of flower development" increased during the time series, while genes associated with the "negative regulation of flower development" decreased in expression across the time series in the apex (Figure 2.13). These responses are consistent with phenotypic observations that flower buds were visible from above (BBCH stage 51[246]) at the final time point in the series. An additional observation for the expression traces of the cluster enriched for genes associated with the positive regulation of flower development is the slight decrease in expression during the vernalization treatment (Figure 2.13). As will be discussed later in this chapter when the behaviour of key floral integrators are investigated (Section 2.4.1), this is likely a result of the short day conditions the plants were grown in not being conducive to flowering.

Having established that clustering expression profiles from the transcriptomic time series resulted in biologically relevant groupings of genes, the enrichment of other GO terms was investigated. Controlling the cell cycle is an integral aspect of growth that plants need to tightly control. In terms of flowering, a sudden burst in the expression of genes controlling the cell cycle was observed during the floral transition in the shoot apical meristem of Arabidopsis[267]. This behaviour was hypothesised to be a result of large scale meristem reorganization initiated by the floral transition. In the apex-derived SOM, there are two main

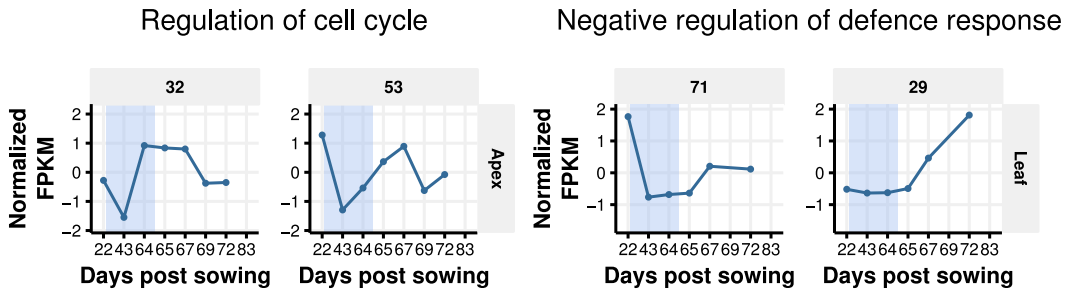


Figure 2.14: Normalized expression profiles for SOM clusters enriched for regulation of cell cycle and defence response.
Normalized expression profiles for the top two SOM clusters that are significantly enriched for each GO term. The expression profiles of genes involved with regulating the cell cycle in the apex decrease during the cold treatment, suggesting that the cold temperature may involve a change in the rate of cell division. The response of SOMs enriched for negative regulation of defence response in the leaf suggest interplay between defence responses, cold, and flowering.

clusters enriched for the GO term "regulation of cell cycle". Both clusters exhibit high expression prior to the cold and a decrease in expression during the cold (Figure 2.14). Immediately after cold the expression traces of these SOM clusters peak before returning to lower expression levels. The peak in expression after the vernalization period is in line with the findings discussed for Arabidopsis[267]. The decrease in expression during the vernalization period suggests that the cell cycle is responding to growth at lower temperatures. This result is in agreement with observations from maize leafs, where the cell cycle duration increased during growth in cold conditions and cell cycle related genes exhibited differential expression[268].

The interactions between plant defence response, flowering, and temperature are beginning to be revealed in model species[244,269]. The energetic costs of growth and the maintenance of an active immune response in the plant have to be balanced to ensure robust development[270–272]. In Arabidopsis, mutants in a particular negative regulator of defence had reduced seed production, indicating that negative regulation of defence during the reproductive phase of plant development is important[273]. The *PIF4* transcription factor in Arabidopsis is

important for the thermal acceleration of flowering[244], but also mediates the balance between growth and pathogen immunity at different temperatures[269]. At low temperatures, immune responses are upregulated and growth is inhibited, while at warmer temperatures the immune response is downregulated, with growth and flowering promoted. The expression profiles of SOM clusters enriched for genes with the GO term "negative regulation of defence response" reflect this (Figure 2.14). Cluster 71 in the leaf-derived SOM exhibits high expression initially, with a rapid reduction in expression during the cold. Upon return to warmer growth conditions, the expression increases. The other cluster enriched for genes involved with down-regulating plant defence responses is cluster 29. This cluster is not affected by the cold treatment, but exhibits a steady increase in expression after the treatment. Both of these observations point towards the *B. napus* defence response being modulated by temperature and flowering in a similar manner to that observed in Arabidopsis.

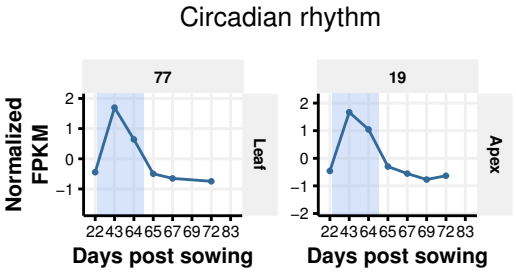


Figure 2.15: Normalized expression profiles for SOM clusters enriched for genes associated with the circadian rhythm.
Normalized expression profiles for the top two SOM clusters that are significantly enriched for the GO term "circadian rhythm" in both tissues in Westar. Both expression profiles increase during the cold treatment, suggesting a response to the change in photoperiod or cold experienced during the vernalization treatment.

To ensure the vernalization treatment was physiologically accurate, plants were subjected to growth in short days at 5 °C. The spring variety, Westar, was subjected to the vernalization treatment alongside the winter variety, Tapidor, to allow for the transcriptomic effects of photoperiod and ambient temperature changes to be differentiated from the effects of vernalization (Section 2.2.1). To investigate the effects of this treatment, SOM clusters enriched for the

GO term "circadian rhythm" were determined. The most highly enriched clusters in both the leaf and the apex of Westar exhibit very similar expression traces (Figure 2.15). Both undergo increases in expression during the cold treatment, with expression returning to pre-treatment levels on the first day of growth post-treatment. This suggests that the altered photoperiod during the vernalization period results in changes to the circadian clock, potentially due to the clock becoming entrained to the different light regime[16].

Although GO term enrichment is a relatively high level analysis that does not investigate the gene level responses across the transcriptomic time series, it is still a useful analysis for investigating the overall behaviour of key developmental pathways. The results presented here reveal a number of general trends that are in agreement with observations in Arabidopsis. The response of the cell cycle and the defence response genes to the period of cold the plants were subjected to is in line with findings from Arabidopsis[267,269]. In the case of the behaviour of defence genes, the observation that the response seems to be conserved between Arabidopsis and *B. napus* may have a future agronomic benefit. The expression response of genes associated with the circadian rhythm validates the experimental design decision to sample two time points during the vernalization treatment. If a single time point was sampled, the observed expression differences as a result of the changing photoperiod would be indistinguishable from effects due to a vernalization response.

2.2.6 Protein domain enrichment

Proteins are modular in structure, composed of protein domains that are often responsible for the molecular activity of the protein[274,275]. As a result, particular classes of protein are associated with certain biological pathways or activities. This is especially true with transcription factors, with different transcription factor domains in Arabidopsis binding to distinct recognition sequences[276] and thus having distinct sets of target genes. Investigating the expression of particular transcription factor families across development can reveal the roles they play in development[277]. In order to take a similar approach using the transcriptomic time series, *B. napus* gene models were annotated with protein domains using previously published tools (Section 6.11; Methods).

Two case studies that illustrate the insights such an analysis facilitates are MADS-box and AP2 domain containing proteins.

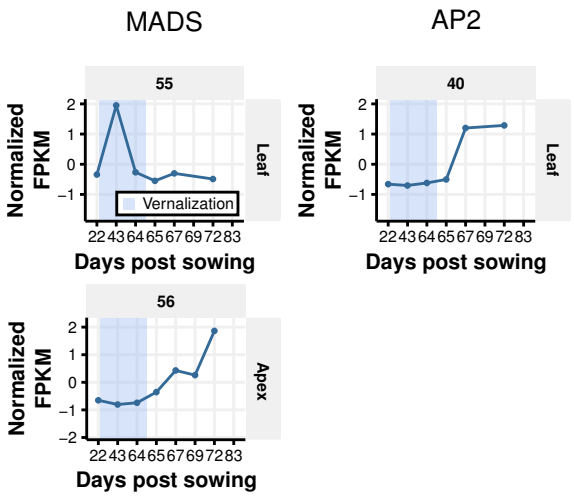


Figure 2.16: Normalized expression profiles for SOM clusters enriched for MADS and AP2 protein domains in the leaf and apex tissue of Westar.
Normalized expression profiles for SOM clusters that are significantly enriched for each protein domain and that also contain the most *B. napus* genes annotated with that protein domain are displayed. The expression patterns of MADS-box containing genes exhibit different patterns of expression in each tissue, suggesting that the proteins play tissue-specific roles in development. The expression profile of AP2 containing genes suggests that the proteins play a role late in development in the leaf.

The MADS-box domain is a protein domain that is conserved across a diverse array of species. Indeed, the MADS-box takes its name from the *MINICHROMOSOME MAINTENANCE 1* genes in yeast, the *AGAMOUS* gene in Arabidopsis, *DEFICIENS* in *Antirrhinum majus* and serum response factor in humans[278]. In Arabidopsis, MADS-box containing genes have been found to control a wide range of roles related to flowering[279]. To determine the regulation of this important family of proteins in *B. napus*, the clusters enriched for genes containing the domain were found (Section 6.11; Methods). In the leaf samples, 35 *B. napus* genes with detectable MADS-box domains are expressed, whereas 85 were expressed in the apex. The expression profiles for the SOM clusters most highly enriched for MADS-box containing proteins are quite

different between the leaf and apex (Figure 2.16). The leaf cluster peaks in expression during cold, with expression at the other time points, before and after cold, being somewhat similar. The SOM cluster enriched in the apex exhibits an expression trace that is lowly expressed before and during cold but steadily increases after the cold to peak expression at the final time point. To investigate why SOM clusters with such different expression profiles were enriched for MADS-box containing genes, the MADS-box containing genes within each cluster were scrutinised further. The MADS-box containing genes mapping to cluster 55 in the leaf-derived SOM correspond to genes involved with the control of flowering time such as *SVP*, *FLC*, *SOC1*, and *AGL24*[29,81,83]. The genes mapping to cluster 56 in the apex-derived SOM, in contrast, include the meristem identity controlling genes *AP1* and *FUL* and genes which are involved with the ABCE model of flower morphology control[8,280]. All four of the gene classes of the model are represented; A class (*AP1*), B class (*AP3* and *PI*), C class (*AG*), and E class (*SEP1*, *SEP2*, and *SEP4*). Therefore, the MADS-box containing genes within these clusters represent different functional classes of MADS-box genes. The upregulation of floral identity genes in the apex at the end of the time series is consistent with the plants beginning to flower at the final time point. The regulation of the MADS-box containing genes in the leaf is likely related to the regulatory effects of the circadian rhythm (Figure 2.15), as the expression of *SVP*, *SOC1*, and *AGL24* are all influenced by the photoperiod pathway[20,69,90,281].

In addition to *AP1*, another A class meristem identity gene important for the specification of flower organ identity is the homeotic gene *APETALA2* (*AP2*)[282]. The function of the gene is dependent upon a 68 amino acid repeated motif called the AP2 domain[283]. This domain has been found to be present in a wide range of plant transcription factors that have been divided into three families; Ethylene Responsive Factors (ERF), AP2 and RAV families[284]. These proteins are involved in a wide range of developmental processes as well as regulating metabolism and stress responses[284]. Investigating SOM clusters enriched for genes containing the AP2 domain reveals cluster 40 in the leaf-derived SOM as being highly enriched. The expression trace of cluster 40 is low initially and during the cold treatment, with a large increase in expression at the penultimate and final time points (Figure 2.16). This suggests that the

AP2 containing genes contained in this cluster are involved with leaf senescence (Figure 2.13). This is consistent with the observation that the majority of AP2 domain containing genes within cluster 40 are members of the ERF family. Genes in this family are frequently induced in response to stresses, and as their name suggests, are responsive to plant hormones associated with stress; ethylene, jasmonic acid and abscisic acid[284]. The role ethylene plays in leaf senescence[285] also strengthens the hypothesis that the AP2 domain containing genes within this cluster are mediating this response.

2.2.7 Conclusions

To investigate regulatory changes during floral development in *B. napus*, a transcriptomic time series experiment was designed to dissect the roles of different flowering time pathways. Sampling from both the leaf and the apex allows a much richer view into flowering time control[13,15] as both tissues are involved with different aspects of regulation. Developmentally similar tissues were sampled from both a winter and a spring variety in order to generate the time series. Comparing these two varieties allows vernalization responsive genes to be elucidated[232]. This is particularly important given the agronomic importance of the vernalization response to the growth of *Brassica* crops[127]. The reference sequence and downstream expression analysis pipeline used to analyse the short read data were chosen in order to make best use of the data. The final dataset is of good quality, with uncertainty estimates that allow for the similarity of expression traces across time to be quantified in a statistically sound manner.

Initial analysis of the transcriptomic time series was focused on validating the responses of key developmental pathways. In order to carry this out, SOMs were generated to cluster the expression profiles across time. Two main expression responses were observed in both the apex and leaf of the spring variety; a response to the changing growth conditions of the vernalization treatment and an increase in expression towards the end of the time series. Analysis of GO terms suggest that the transcriptomic response to the vernalization treatment is in part a response to the change in photoperiod, as would be expected given results from Arabidopsis[16]. As the photoperiod pathway is a key floral

pathway[15,17,18], the expression of flowering time genes during the time series should be viewed with this response in mind. The large number of genes in both tissues increasing in expression towards the end of the time series seem to be the result of different developmental pathways. In the leaf, the response of genes annotated with the GO term "leaf senescence" (Figure 2.13) and genes containing the AP2 protein domain (Figure 2.16) suggest that leaf ageing is a strong influence on transcriptional responses in the tissue. In contrast, the increase at the final time point in the apex seems to be linked to floral development (Figures 2.13 and 2.16). Interestingly, MADS-box containing genes known to repress each other are co-expressed in the SOM cluster enriched for MADS-box containing genes (Figure 2.16). For example, *AG* represses the expression of *AP1* in the inner two whorls of the flower[54], while *AP2* limits the expression domain of *AG*[286]. This co-expression illustrates that the dissection of the apex is not able to resolve the distinct expression zones in the apex[13]. The alignment of key developmental pathways with phenotypic observations and expectations from model species demonstrates that the transcriptomic time series is able to capture biologically relevant changes in expression.

2.3 Regulatory divergence at the whole genome scale

The effects of polyploidy on gene expression are varied and seemingly influenced by the species and the time since hybridization[287]. Immediately following hybridization, large transcriptional changes are observed in polyploids[288,289]. In synthetic Arabidopsis allopolyploids, Wang et al. (2006)[290] observed different contributions to the transcriptome from the different constituent genomes, consistent with extensive gene silencing following polyploidy[291]. These results from Arabidopsis allopolyploids demonstrate a major way in which gene expression can vary after polyploidy: genome dominance. Genome dominance is observed when the combined gene expression of gene pairs from the two constituent genomes of a polyploid are consistently biased towards a particular genome[292,293]. These expression inequalities may influence the evolution of the polyploid, with results in maize revealing that gene loss favours copies

that contribute less to overall expression[294]. In cotton (*Gossypium raimondii*) 99.4% of ~2,000 gene pairs exhibited biased expression in at least one of the three tissues tested[261]. Interestingly, this bias was found to be tissue specific, suggesting that homologous genes may diverge to become tissue specific over evolutionary time[261,292].

In order to investigate global differences in expression between the genomes of *B. napus*, the expression of genes on the separate genomes were compared using the transcriptomic time series. The genome of origin seems to influence the expression of genes in the *B. napus* genome, with different patterns of expression bias observed at the genome-wide level relative to homoeologue level comparisons. Investigating the retention of genes reveals that flowering time genes have been retained in the *B. napus* genome, and that this is also observed among the subset of expressed genes. This suggests that the retained gene copies may be functional. Determining expression pattern divergence among flowering time gene homologues in *B. napus* reveals that the majority exhibit regulatory divergence. This suggests that regulatory divergence has contributed to the retention of flowering time genes in *B. napus*, although this has occurred alongside potential gene dosage effects.

2.3.1 Genome level expression differences between the A and C genomes

Previous studies of gene expression in polyploid species have generally focussed on comparing the expression of genes on different genomes to determine whether gene expression is biased[288,290,295–297]. To determine whether such a bias was observed in the expression data from the transcriptomic time series, density plots of the gene expression data for each of the two genomes was generated (Figure 2.17). Different regions of the density curves will hereafter be referred to as very low (below -1), low (between -1 and 0), high (between 0 and 1), and very high (above 1), relating to the expression of genes within those regions. The A genome has a greater proportion of genes in the high expression region relative to the C genome (Figure 2.17a). Conversely, for genes in the very low expression region, the opposite trend is observed (Figure 2.17a). Similar patterns are observed when only *B. napus* genes exhibiting sequence

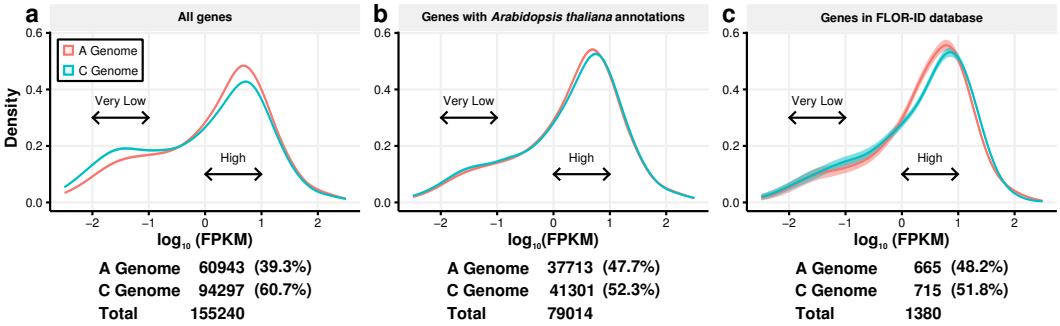


Figure 2.17: The *B. napus* A and C genomes show different overall patterns of gene expression.

Density plots of transformed expression levels ($\log_{10}(FPKM)$) calculated using different subsets of genes. The expression data was sampled 1000 times using a Gaussian error model. The density plot of $\log_{10}(FPKM)$ values was calculated for each sample. The mean density and the 95 % confidence interval estimated using the 1000 samples is displayed. Tabulated below each density plot are the number of *B. napus* genes used to calculate the density plot, separated by their genome of origin. The data used to generate the density plots consisted of expression data from: **a** all annotated *B. napus* genes, **b** *B. napus* genes that show sequence conservation to an annotated Arabidopsis gene, and **c** *B. napus* genes that show sequence conservation to an annotated Arabidopsis gene that is present in the FLOR-ID database[299]. These plots are generated using apex expression data from the time point taken at day 22, but are representative of the density plots obtained for all time points across both tissue types sampled (Figure 6.4; Appendix A).

conservation to an annotated Arabidopsis gene are considered (Figure 2.17b) and when *B. napus* flowering time homologues are considered (Figure 2.17c). However the differences between the density plots are less apparent when these subsets are taken. Interestingly, the proportions of genes represented from each genome change when these subsets of genes are taken. When no subset is taken, approximately 40% of *B. napus* gene models are located on the A genome. When subsets are taken, however, the percentage of genes on the A genome is 48% in both cases (Figure 2.17). This difference reveals that there are more genes on the C genome that do not show sequence similarity to an Arabidopsis gene.

Table 2.1: Number of genes expressed two-fold higher than their homoeologue for all homoeologue pairs.

Homoeologue pairs were determined and filtered at each time point for those which both had expression levels above 2 FPKM. The number and percentage of these genes expressed two-fold higher than their homoeologue is indicated. The geometric mean of the fold difference of the C genome gene relative to the A genome homoeologue for all homoeologue pairs is 1.12 in the apex and 1.11 in the leaf.

Days post sowing	Apex			Leaf		
	Both expressed	A genome two-fold higher	C genome two-fold higher	Both expressed	A genome two-fold higher	C genome two-fold higher
22	7313	596 (8.1 %)	1113 (15.2 %)	6294	620 (9.9 %)	1066 (16.9 %)
43	7389	597 (8.1 %)	1132 (15.3 %)	6176	626 (10.1 %)	1133 (18.3 %)
64	7325	602 (8.2 %)	1085 (14.8 %)	6307	597 (9.5 %)	1021 (16.2 %)
65	7243	609 (8.4 %)	1120 (15.5 %)	6182	601 (9.7 %)	993 (16.1 %)
67	7299	601 (8.2 %)	1135 (15.6 %)	6257	603 (9.6 %)	1046 (16.7 %)
69	7342	594 (8.1 %)	1130 (15.4 %)	-	-	-
72	7449	612 (8.2 %)	1119 (15.0 %)	6237	601 (9.6 %)	1054 (16.9 %)

To compare expression changes between the A and C genomes at the gene level, as has been done previously[298], a list of homoeologues was generated by genomic synteny and sequence similarity, following a published method[118]. Pairs of homoeologues were classified as exhibiting biased expression in the direction of a particular genome if the gene on that genome had an FPKM

Table 2.2: Number of genes expressed two-fold higher than their homoeologue for all flowering time gene homoeologue pairs.

As for Table 2.1, calculated using homoeologue pairs that showed sequence similarity to Arabidopsis flowering time genes from the FLOR-ID database[299]. The geometric mean of the fold difference of the C genome gene relative to the A genome homoeologue for all flowering time homoeologue pairs is 1.10 in the apex and 1.04 in the leaf.

Days Post Sowing	Apex			Leaf		
	Both Expressed	A Genome two-fold higher	C Genome two-fold higher	Both Expressed	A Genome two-fold higher	C Genome two-fold higher
22	136	11 (8.1 %)	19 (14.0 %)	109	8 (7.3 %)	14 (12.8 %)
43	149	15 (10.1 %)	24 (16.1 %)	118	12 (10.2 %)	16 (13.6 %)
64	147	12 (8.2 %)	20 (13.6 %)	114	11 (9.6 %)	13 (11.4 %)
65	145	13 (9.0 %)	25 (17.2 %)	108	10 (9.3 %)	16 (14.8 %)
67	138	14 (10.1 %)	19 (13.8 %)	112	7 (6.3 %)	12 (10.7 %)
69	139	11 (7.9 %)	18 (12.9 %)	-	-	-
72	142	15 (10.6 %)	21 (14.8 %)	112	5 (4.5 %)	14 (12.5 %)

expression value at least two-fold higher than the gene on the other genome. Biased expression occurs in the direction of both genomes, although there is a clear preference, with approximately double the number of pairs exhibiting biased expression towards the C genome rather than the A genome (16.9% towards the C genome relative to 9.7% towards the A genome in the apex, and 15.2% compared to 8.2% in the leaf; Table 2.1). This pattern is consistent with the findings of Chalhoub et al. (2014)[118], and is maintained across the entire time series and for both tissue types sampled (Figure 6.4; Appendix A). Although more pairs of homoeologues show biased expression towards the C genome rather than the A genome, the pairs biased toward the A genome may exhibit larger fold differences. If the overall expression of homoeologues was balanced between the two genomes in this way, the geometric mean of the fold differences of the C genome genes relative to their A genome homoeologues should equal unity. Calculating the geometric mean reveals a value above 1 (Table 2.1) demonstrating that, on average, expression is biased towards the C genome. When pairs of homoeologues identified as *B. napus* flowering time genes are tested in the same way, patterns are largely maintained although are less consistent across the time series due to fewer genes being considered (Table 2.2).

Investigating expression differences between the two genomes of *B. napus* reveals expression bias, although the direction of the bias depends on the scale at which it is considered. The results from the genome level analysis suggest an expression bias towards the A genome, while the homoeologue level results suggest bias towards the C genome. This discrepancy may be due to genes with low expression levels tending to lack homoeologue pair information (Figure 6.5; Appendix A). It is interesting that the bias towards the A genome observed at the genome scale is less apparent when *B. napus* genes lacking sequence conservation to an Arabidopsis gene are removed. This potentially indicates a higher proportion of silenced or pseudogenes on the C genome, consistent with the higher DNA methylation levels and transposon density observed in the C genome[118].

2.3.2 Tissue-specific expression is biased towards the apex

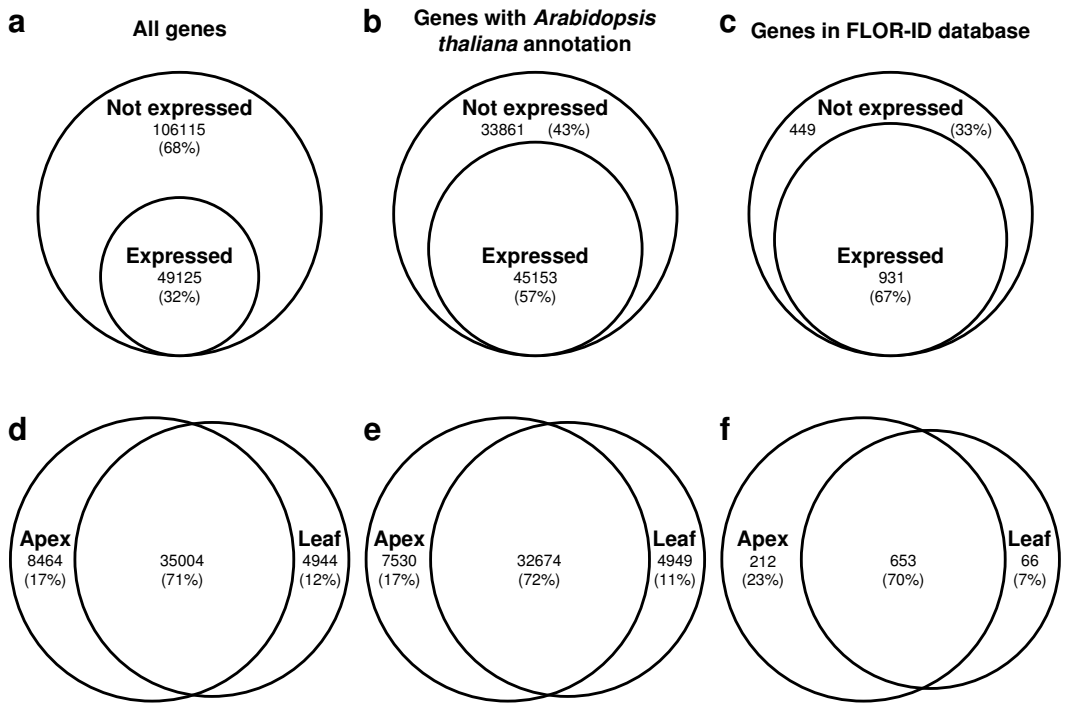


Figure 2.18: The majority of annotated *B. napus* genes are not expressed. **a-c** Euler diagrams indicating the percentage of genes that are expressed and those that are not in the developmental time series. A gene was regarded as expressed if the expression level of the gene exceeded 2.0 FPKM at at least one time point in either the leaf or apex sample. **d-f** Venn diagrams indicating the number of expressed genes showing tissue-specific expression. **a and d** All annotated *B. napus* genes; **b and e** Only *B. napus* genes with an identified Arabidopsis homologue are considered; **c and f** Only *B. napus* genes with an identified Arabidopsis homologue that is in the FLOR-ID database[299] are considered.

The genome level analysis uncovered biased expression between the two genomes of *B. napus*. In order to investigate other forms of expression bias in the data, the number of genes exhibiting tissue-specific expression during the transcriptome time series was assessed. Genes were classified as expressed during the time series if the expression of the gene exceeded 2.0 FPKM at at least one

time point. By this definition, 32% of annotated *B. napus* genes were classified as expressed in the time series (Figure 2.18). This percentage increases to 57% and 67% when only *B. napus* genes with Arabidopsis homologues or *B. napus* flowering time genes were considered, respectively. The finding that there are many lowly expressed *B. napus* genes that lack an Arabidopsis homologue is consistent with the results presented in section 2.3.1. Potentially these lowly expressed genes that lack sequence similarity to annotated Arabidopsis genes are pseudogenes. Taking all *B. napus* genes, regardless of whether they have an Arabidopsis homologue or not, reveals that of the 49,125 genes that are expressed during the developmental time series, 17% are expressed specifically in the apex and 12% are expressed specifically in the leaf, with the remaining 71% of genes expressed in both tissues (Figure 2.18d). These percentages remain largely unchanged when *B. napus* genes lacking an Arabidopsis homologue are removed (Figure 2.18e). For flowering time genes the percentage of genes exhibiting tissue-specific expression shifts towards the apex. Of the 931 expressed *B. napus* flowering time genes, 23% are specifically expressed in the apex and 7% of genes are leaf specific (Figure 2.18). This analysis reveals that the majority of genes do not exhibit tissue-specific expression. Of those that do, there are more genes specifically expressed in the apex than the leaf, perhaps as a result of the apex undergoing a greater developmental change during the time series than the leaf. The percentage of genes exhibiting tissue-specific expression changes depending on the gene subset considered, with *B. napus* flowering time genes having 76% of tissue-specific genes expressed in the apex compared to 63% for all genes. This supports the hypothesis that, for the transcriptomic time series collected in this study, it is the apex transitioning from vegetative to floral growth that results in the observed percentage of genes expressed in an apex-specific manner being higher relative to the leaf.

2.3.3 Multiple copies of flowering time genes have been retained in the *B. napus* genome

Genes that have undergone duplication in the genome and have been subsequently retained are either under a selective pressure to be maintained or have not yet been lost in the population due to genetic drift[212,215]. To investigate

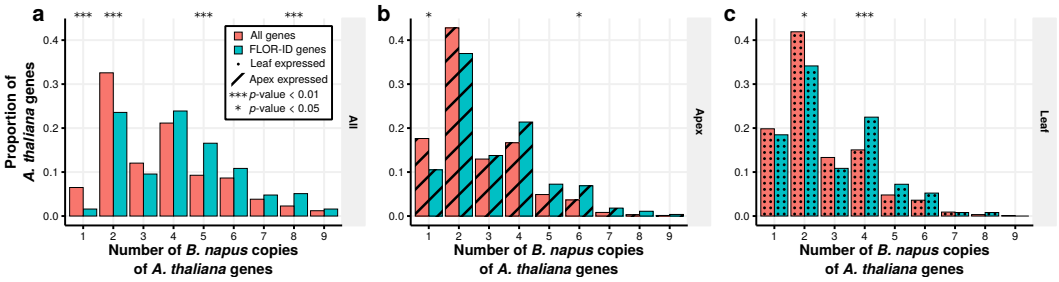


Figure 2.19: Multiple *B. napus* flowering time gene homologues are expressed during the floral transition.

This plot shows the proportions of Arabidopsis genes that have particular numbers of homologues identified and expressed in *B. napus*. *B. napus* genes were considered to be expressed if their maximal expression level within a tissue across the time series was above 2.0 FPKM. False discovery corrected p-values were computed by taking 1000 samples of genes from the All distribution. The mean and standard deviation of these samples were used to perform a two-tailed test of observing a proportion as extreme as the FLOR-ID value. **a** *B. napus* genes that show sequence conservation to an annotated Arabidopsis gene. **b** *B. napus* genes expressed in the apex tissue that show sequence conservation to an annotated Arabidopsis gene. **c** *B. napus* genes expressed in the leaf tissue that show sequence conservation to an annotated Arabidopsis gene.

whether the flowering time genes have been retained in the genome, distributions of Arabidopsis gene copies were calculated. These distributions were derived by assigning *B. napus* genes to the Arabidopsis gene with the highest sequence similarity, then counting the number of copies of each Arabidopsis gene in the *B. napus* genome. This was done separately for all Arabidopsis genes and for the subset of genes identified as being involved with flowering[299] and the distributions compared. Significant differences between the distributions are observed at low copy numbers, with there being fewer Arabidopsis flowering time genes with one or two copies in *B. napus* than expected given the distribution for all genes (Figure 2.19a). At higher copy numbers, a significantly higher proportion of Arabidopsis flowering time genes have five and eight *B. napus* copies relative to the distribution for all genes. To determine if this pattern was also true for expressed *B. napus* genes, similar distributions were generated for expressed genes in the apex (Figure 2.19b) and leaf (Figure 2.19c). These distributions also reveal a shift towards the expression of a higher number of flowering time gene copies relative to the whole genome. In general, flowering time genes tend to have a lower proportion of genes expressed at low copy numbers (three and below) and higher proportions at higher copy numbers relative to the whole genome. This is indicative of the flowering time genes in *B. napus* having been retained in the genome following the genome multiplication events that have occurred throughout the evolutionary history of *B. napus*. In addition, that these patterns are also observed for expressed genes suggests that the retained flowering time genes are functional.

2.3.4 Expression divergence in the number of expressed copies of annotated genes

The distributions of *B. napus* homologue number suggest that genes involved with the regulation of flowering time have been retained in the genome. Investigating the regulatory divergence between these homologues can provide clues as to the evolutionary forces maintaining them in the genome[219,227]. In order to assess regulatory divergence of *B. napus* genes in a binary manner (expressed versus not expressed), the number of annotated *B. napus* homologues of Arabidopsis genes were compared to the number of those genes expressed during

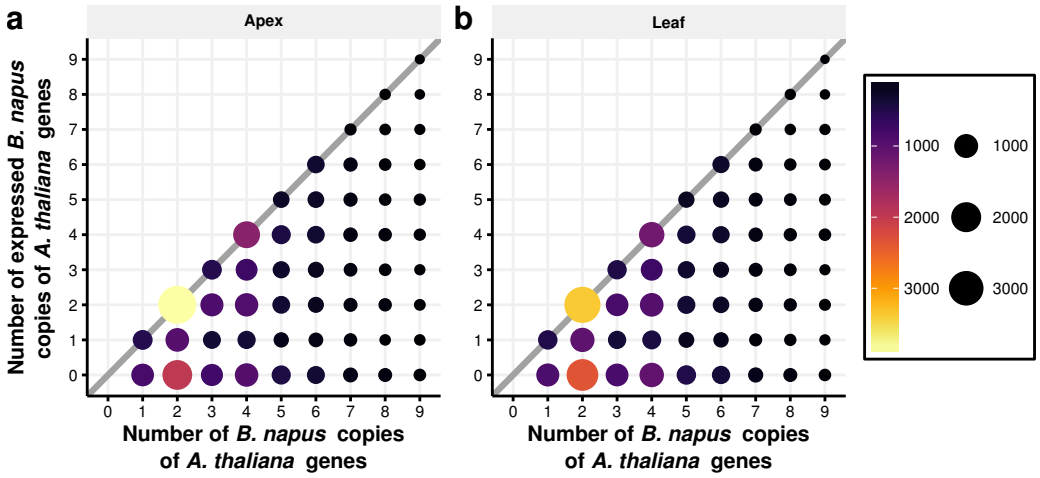


Figure 2.20: Not all copies of genes are expressed in *B. napus*.
Copies of Arabidopsis genes were identified in the *B. napus* gene models through sequence similarity. These copies were regarded as expressed if their maximum expression level during the entire time series exceeded 2.0 FPKM. The size and colour of the cirlces indicates the number of data points at that position in the graph.

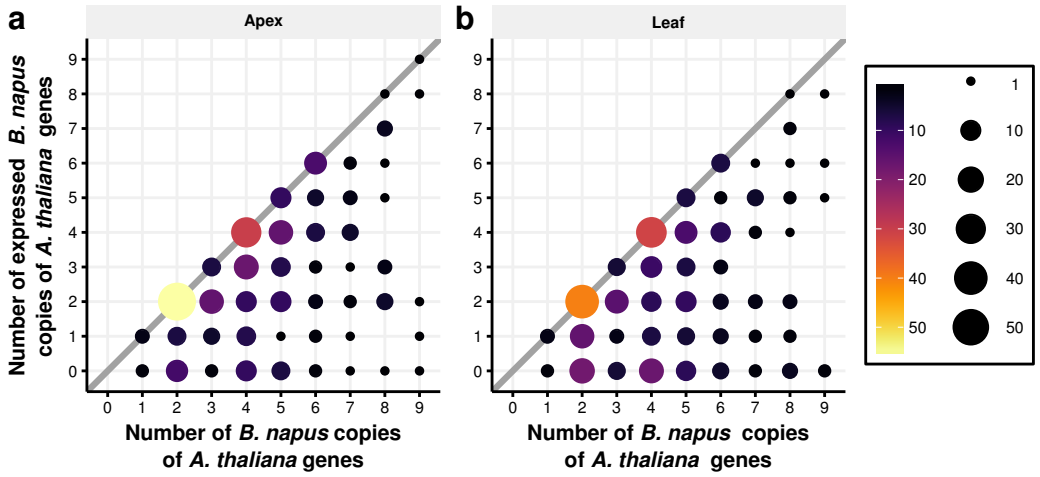


Figure 2.21: Not all copies of flowering time genes are expressed in *B. napus*.
As for figure 2.20, but only using *B. napus* genes that have sequence similarity to annotated Arabidopsis flowering time genes in the FLOR-ID database[299].

the transcriptomic time series (Figures 2.20 and 2.21). In both the apex and the leaf, the majority (66% in the apex, 70% in the leaf) of Arabidopsis genes have at least one *B. napus* homologue that does not exhibit expression during the time series (Figure 2.20). The percentage of Arabidopsis flowering time genes that have at least one homologue that is not expressed are similar to the results observed genome-wide (61% in the apex, 69% in the leaf; Figure 2.21). This indicates widespread expression divergence among *B. napus* homologues during the transcriptomic time series, with the majority of Arabidopsis genes having at least one homologue that is not expressed in the two tissues sampled.

2.3.5 Expressed copies of flowering time genes exhibit regulatory divergence during the floral transition

In order to further investigate regulatory divergence between *B. napus* homologues of Arabidopsis genes, the behaviour of genes across the time series was studied. Different hypotheses for the retention of duplicated genes predict different patterns of co-regulation between these genes[213,219,224,227]. Therefore, by comparing the temporal expression patterns between genes, the mechanism of retention for the flowering time genes can be investigated. In order to do this, Weighted Gene Co-expression Network Analysis (WGCNA) was used to identify regulatory modules[265]. WGCNA uses normalized expression profiles across time to cluster genes based on their temporal expression profiles. Thus, genes that are co-regulated will be assigned to the same cluster, whereas genes that have diverged in their temporal expression will be assigned to different clusters. To assess regulatory divergence between *B. napus* homologues, the number of *B. napus* homologues of an Arabidopsis gene were compared to the number of WGCNA clusters those homologues occupy (Figure 2.22). Assuming that gene dosage leads to co-regulation of duplicated genes[227], the null hypothesis is that all *B. napus* homologues of an Arabidopsis gene would be assigned to the same regulatory module (dashed line in Figure 2.22). However, if regulatory divergence is observed with at least one homologue this null hypothesis is inaccurate, with the extreme situation being that every *B. napus* homologue occupies a separate WGCNA cluster (solid diagonal line in Figure 2.22). Most *B. napus* homologues exhibit regulatory divergence (69% in the

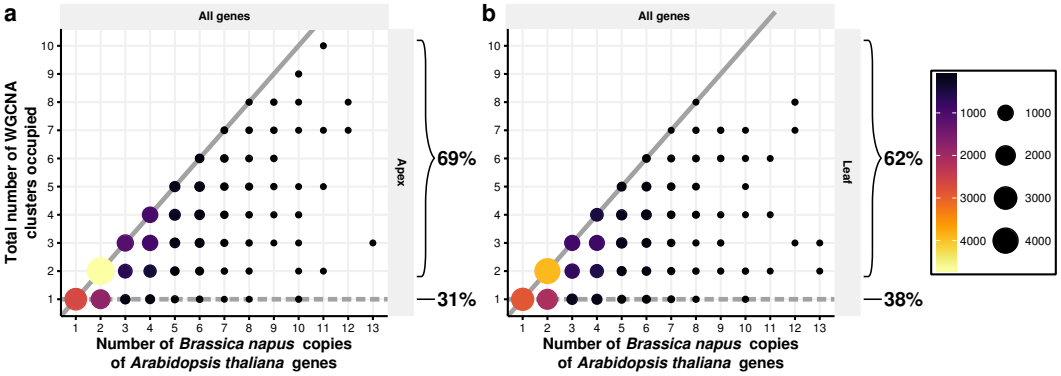


Figure 2.22: The majority of gene homologues in *B. napus* are assigned to different regulatory modules.

Regulatory module assignments for the apex, **a**, and leaf, **b**. The size and colour of the circles indicate the number of data points at that position in the graph. The thick lines on each graph represent two potential extremes. The dashed line represents the null hypothesis that all *B. napus* copies of an Arabidopsis gene are assigned to the same WGCNA cluster. The solid line represents the Arabidopsis genes that have *B. napus* copies that are each assigned to separate WGCNA clusters. The percentages indicated on the graph indicate the percentage of data points that agree, and the percentage that do not agree, with the null hypothesis. Only *B. napus* genes with expression above 2.0 FPKM in at least one time point in the transcriptomic time series and sequence conservation to an annotated Arabidopsis gene were used.

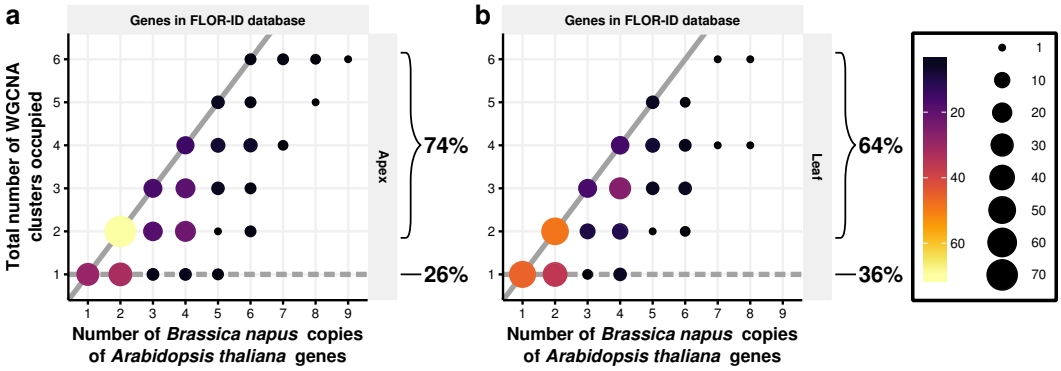


Figure 2.23: The majority of flowering time gene homologues in *B. napus* are assigned to different regulatory modules.
As for figure 2.22, but only using *B. napus* genes that have sequence similarity to annotated Arabidopsis flowering time genes in the FLOR-ID database[299].

apex, 62% in the leaf) which does not conform to the null hypothesis derived from dosage balance arguments. This pattern is also observed when just *B. napus* flowering time genes are considered (Figure 2.23). These findings reveal that the majority of *B. napus* genes have diverged from the expression patterns of their homologues, calling into question the extent to which gene dosage effects have maintained these duplicate genes in the genome.

The regulatory divergence determined using the WGCNA was assessed in a binary manner; *B. napus* genes are either assigned to the same cluster or not. However, this approach does not quantify the similarity between profiles. The consequence of this is genes that exhibit expression profiles that could be assigned to multiple regulatory modules will only be assigned to a single module. In addition, the WGCNA approach does not account for the uncertainty in the RNA-Seq data when determining module assignment. To overcome these issues, a SOM-based sampling approach was taken to assess regulatory divergence between *B. napus* flowering time homologues (Figure 2.24a). This method accounts for the uncertainty in the RNA-Seq data by sampling from the data. By counting the number of sampling iterations in which two genes cluster together, relative to the total number of sampling iterations, empirical probabilities of two expression traces mapping to the same SOM cluster are generated (Figure 2.24a). These probabilities are

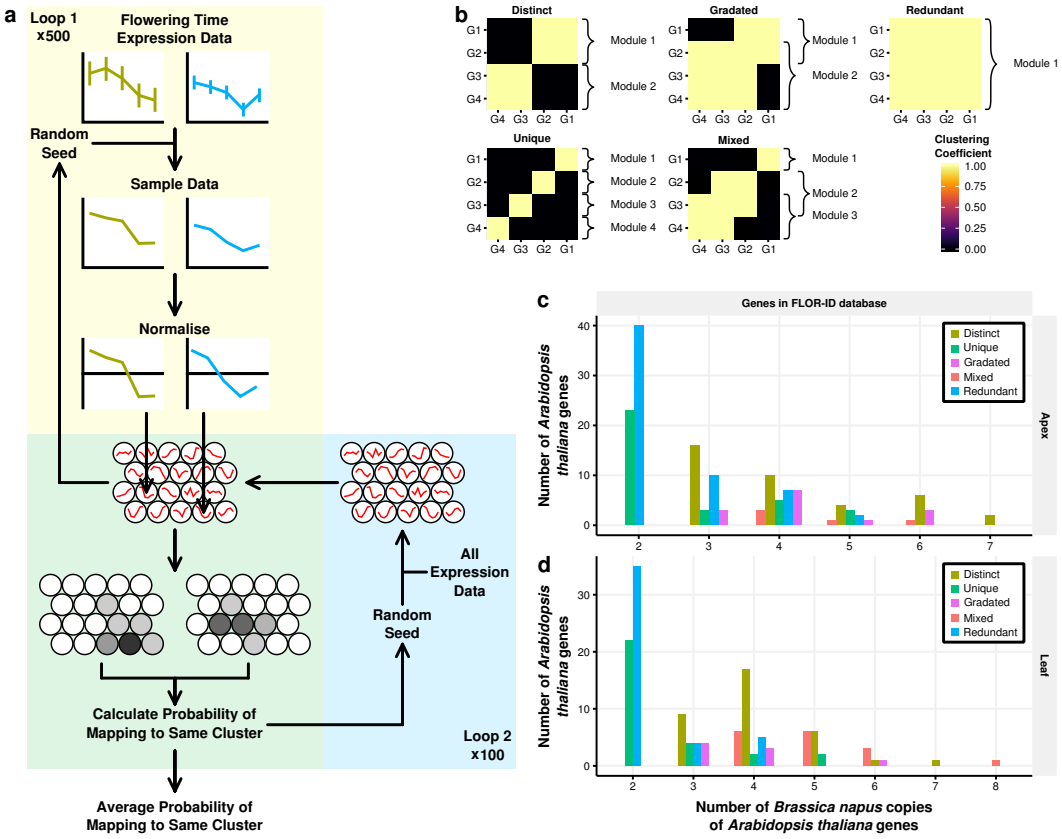


Figure 2.24: Self-organizing map (SOM) based assessment of expression trace divergence uncovers widespread regulatory divergence and subtle patterns of divergence.

a A schematic of the SOM based clustering approach. The approach consists of two overlapping sampling loops. In loop 1, expression data from flowering time gene copies is sampled assuming a Gaussian error model. Sampled expression traces are zero mean and unit variance normalized and mapped to the SOM. This procedure is repeated 500 times to give two density plots of where in the SOM the copies map. These density plots are used to calculate the probability of the copies mapping to the same SOM cluster. As SOM clustering has a random component, loop 2 consists of regenerating the SOM using all expression data and calculating the probability of copies clustering to the same cluster 100 times. Using this, an average probability of mapping to the same cluster is calculated. *Continued on Page 99.*

Continued from Page 98. **b** Representations of the five patterns of regulatory module assignment detected by the SOM based method. High clustering coefficients between two different genes indicates that those genes have similar expression traces. Clustering coefficients between a gene and itself represent how robustly a gene maps to the SOM. A *distinct* pattern indicates multiple regulatory modules being identified, with no gene occupying more than one module. A *gradated* pattern represents multiple regulatory modules being detected, but genes occupy multiple modules. *Redundant* patterns occur when only one regulatory module is detected, and all copies of a gene are assigned to that module. *Unique* patterns are a special case of a *distinct* pattern where each copy of a gene is assigned to a different regulatory module. *Mixed* patterns consist of a mixture of *distinct* and *gradated* patterns, where the gene assignment of some modules overlap while others do not show overlap. When assessing the regulatory module assignment, gene copies that do not robustly map to the SOM are removed. **c and d** The relationships between the number of expressed *B. napus* copies of Arabidopsis flowering time genes and the number of different types of regulatory module assignment patterns exhibited by those gene copies. This relationship is calculated using expression data from the apex (**c**) and the leaf (**d**).

normalized to give a clustering coefficient (Methods; section 6.7). The higher the coefficient, the higher the probability of two expression traces mapping to the same cluster. *B. napus* copies of Arabidopsis genes are grouped into regulatory modules based on the clustering coefficients, with copies that have high clustering coefficients between them being assigned to the same regulatory module. Unlike some methods of clustering gene expression profiles, genes have the potential to be assigned to multiple regulatory modules. This allows more subtle patterns of divergence to be detected. There are five different possible patterns of regulatory module assignment using the SOM-based resampling method (Figure 2.24b). A *distinct* pattern represents the identification of multiple regulatory modules whose membership does not overlap. *Gradated* patterns indicate that multiple regulatory modules were identified, but the membership of those modules overlap. *Redundant* patterns occur when all *B. napus* copies of an Arabidopsis gene are assigned to the same regulatory

module. The *unique* pattern is a special case of the *distinct* pattern, where only one gene is assigned to each identified regulatory module. Finally, the *mixed* pattern is observed when at least three regulatory modules are identified, with some genes assigned to multiple regulatory modules and others not. The benefit of allowing genes to occupy multiple regulatory modules is that subtle patterns can be detected. For example, copies exhibiting *gradated* patterns of regulatory module assignment exhibit intransitivity; although gene A and gene B are in the same regulatory module, and gene B and gene C are in the same regulatory module, gene A and gene C are not necessarily mapped to the same module. In this case, given that gene A and gene C are not in the same module, it is clear that gene B exhibits a regulatory trace that is intermediate between gene A and gene C.

To assess the extent of regulatory divergence among *B. napus* flowering time gene homologues using the SOM-based method, the regulatory module assignments were quantified. As with the WGCNA-based approach, the null hypothesis considered was that of genes exhibiting co-regulation. In the SOM-based analysis, this hypothesis corresponds to observing a *redundant* regulatory module assignment. Data from the developmental time series reveals that as the number of *B. napus* copies of an Arabidopsis gene increases, the occurrence of *redundant* patterns decreases in both the apex and the leaf (Figures 2.24c and 2.24d). When three or more copies of a gene are present, regulatory module patterns other than *redundant* are observed in the majority of cases in both tissues, with no redundant patterns seen above 5 copies in the apex or 4 copies in the leaf. *Unique* patterns were also observed less frequently at higher numbers of copies, suggesting that as the number of homologues increases, the more likely it is that at least two homologues exhibit similar expression profiles. Therefore, as with the results from the WGCNA analysis, the null hypothesis ceases to be true for flowering time genes with five or more copies in the *B. napus* leaf (Figure 2.24d) or six or more copies in the apex (Figure 2.24c). An advantage that the SOM-based analysis has compared to the WGCNA-based analysis is that the method allows for the detection of *mixed* and *gradated* patterns. In the apex and leaf, *mixed* and *gradated* patterns are seen at a lower frequency than *distinct* patterns. This reveals that genes with intermediary regulatory behaviour are observed less frequently than

genes exhibiting greater divergence in their expression profiles. Gene copies with intermediate regulatory behaviour may indicate that particular copies are more susceptible to regulatory cross-talk than others.

An interesting observation from the SOM-based analysis is the relatively large number of *distinct* patterns observed at four gene copies (Figures 2.24c and 2.24d). To test if this was due to homoeologous genes displaying similar expression profiles, homoeologue information was incorporated into the analysis. For the genes for which homoeologue information was available, the majority (76% in apex, 72% in leaf) of genes are in the same regulatory module as their homoeologue. More generally, for all expression traces, of 85 pairs of homoeologues expressed in the apex, 67 (79%) are found in the same regulatory module. In the leaf, 53 of 69 (77%) of expressed homoeologous pairs are found in the same module, with 29 of the co-regulated pairs being common between the two tissues. The percentage of Arabidopsis genes with at least two expressed homologues in the apex (leaf) exhibiting each of the regulatory module assignments are 25% (26%) *distinct*, 9% (6%) *gradated*, 23% (23%) *unique*, 39% (33%) *redundant*, and 3% (6%) *mixed*. This reveals that although extensive regulatory divergence is observed, homoeologous genes still tend to exhibit similar expression profiles. This suggests that since the formation of *B. napus* 10,000 years ago[107], the majority of homoeologous genes have not diverged in their expression.

2.3.6 Conclusions

To investigate whether flowering time genes have been retained in the *B. napus* genome, and the mechanisms by which these gene copies have been retained, the expression of *B. napus* gene homologues were compared during the transcriptomic time series. Analysis of the expression levels of all genes revealed that, on average, the A genome has a greater proportion of highly expressed genes relative to the C genome. That this observation becomes less apparent when *B. napus* genes lacking sequence conservation to an Arabidopsis gene are removed suggests that the C genome contains a greater number of pseudogenes; gene models detected by the gene prediction algorithm but that are transcriptionally silenced. This supports observations that the C genome

contains a higher density of transposons and higher DNA methylation levels than the A genome[118]. At the homoeologue level, biased gene expression was observed towards both genomes, although a higher number of homoeologue pairs were biased towards the C genome. This is also consistent with previous observations[118], although that biases are observed in both directions proves inconclusive for determining whether one genome is dominant over the other.

Investigating the expression of flowering time genes in *B. napus* reveals that these genes exhibit higher retention in the genome relative to the genome-wide trend (Figure 2.19). The majority of Arabidopsis genes have at least one *B. napus* homologue that lacks expression during the transcriptomic time series (Figure 2.19). This is consistent with the idea of responsive backup circuits, which posits that duplicate genes can be retained in the genome, with one copy only expressed when the other copy becomes non-functional as a result of mutation[219,220]. Alternatively, the *B. napus* homologues lacking expression in the transcriptomic time series may be expressed at a point in developmental not represented by the time series, or expressed in a different tissue. To further investigate regulatory divergence between homologues, WGCNA- and SOM-based clustering approaches were employed to quantify the extent of divergence between expressed *B. napus* homologues. The WGCNA-based analysis revealed extensive regulatory divergence for all genes, including the subset of flowering time genes. The SOM-based approach confirmed the observation of flowering time genes exhibiting regulatory divergence in a manner robust to the calculated experimental uncertainty. Additionally, the SOM-based analysis reveals that some copies of flowering time genes exhibit a *gradated* patterns of regulatory module assignment, representing subtle differences in regulation. This may be the result of regulatory cross-talk between the copies, or represents subtle functional differences that have consequences for the control of flowering time in *B. napus*. The regulatory divergence observed for the flowering time genes is counter to the expectations of a gene dosage model for their retention; namely co-regulation[224,227]. As the spatiotemporal expression pattern of a gene plays a crucial role in its function, this also suggests functional divergence of *B. napus* flowering time gene homologues. This would therefore suggest that mechanisms other than gene dosage, such as subfunctionalization or

neofunctionalization, have also contributed to flowering time gene retention in *B. napus*[206,213,219,220,229].

2.4 Regulatory divergence of key floral integrators

The main floral pathways that influence flowering are the photoperiod pathway, the autonomous pathway, the vernalization pathway, the hormone pathway, and the ageing pathway[15]. The signals from these pathways are integrated by a central decision network of floral integrators (Section 1.1.2; Figure 1.1). Despite the importance of this network for determining the timing of the floral transition in Arabidopsis[41], work investigating homologues of these floral integrators in *Brassica* species is relatively scarce, especially when compared to the available literature concerning the vernalization pathway in *Brassica* crops (section 1.2.2). The work that is available reveals that the key Arabidopsis floral integrators are present as multiple copies in the *B. napus* genome, and that sequence variation exists both between different varieties and between homologues[131,152]. For *TFL1*, *FT*, and *SOC1*, sequence variation between copies has been related to functional differences between the copies, such as changes in expression pattern and different effects on plant phenotype[153–155,157,158]. However, although these studies have identified expression pattern differences between *B. napus* homologues of floral integrators, none have determined which copies exhibit expression consistent with the regulatory interactions identified in Arabidopsis. In addition, only in the case of *SOC1* homologues has the tissue-specific expression of the different copies been assessed[158]. This is of particular interest given results from Arabidopsis that suggest that duplicated regulatory networks will tend to diverge and form parallel networks that are distinct in terms of their spatiotemporal expression[229].

To investigate whether *B. napus* homologues of the floral integrators have diverged in *B. napus*, the expression profiles of these genes were assessed in the transcriptomic time series for the spring variety Westar. Every Arabidopsis floral integrator considered has at least one copy in *B. napus* that exhibits an expression profile consistent with the expression pattern expected from

observations in Arabidopsis. However, regulatory divergence is also observed among the integrators, with the degree of divergence varying based on the gene. Analysing the regulatory patterns exhibited by *BnSOC1*[3] genes suggests that some copies respond to the vernalization treatment, while others do not. This provides evidence that these genes have subfunctionalized to become responsive to particular inputs. *BnLFY* genes, however, seem to be acting in a redundant manner, suggesting that dosage effects may influence the retention of the additional *BnLFY* genes in the genome. In order to focus this analysis, only the floral integrator hubs included in the model of the floral transition by Jaeger et al. (2013)[41] will be considered.

2.4.1 *FLOWERING LOCUS T*

FT is a floral activator that is induced in long day conditions to promote flowering[20–22]. In Arabidopsis, *FT* is primarily expressed in the phloem companion cells, with the FT protein transported in the plant vasculature to the apex to initiate flowering[42,44–46]. It is likely that this mechanism of *FT* action is conserved in *B. napus*[300]. Although the leaf is the primary expression domain of *FT*, expression of the gene has also been observed in the shoot apex and the hypocotyl of long day grown plants[22,233], although the biological relevance of these observations is unknown. In contrast to other studies that found six copies of *FT* in *B. napus*[153,301], only four copies of *BnFT* were found in the transcriptomic time series, situated on chromosomes A2, A7, C2, and C6. In previous studies, two additional copies were found on A7 and C6, with these copies located in inverted blocks of duplicated sequence[301]. Potentially the additional copies of *BnFT* are not present in the Darmor-*bzh* reference genome as a result of genome assembly error, caused by the inverted blocks failing to be resolved.

As *FT* is primarily expressed in the leaf in Arabidopsis[42,44–46], the expression of the gene in this tissue was analysed. The four *BnFT* homologues exhibit a *gradated* pattern of regulatory module assignment with two regulatory modules

[3]Gene abbreviations prefixed by two letters indicate homologues of Arabidopsis genes in other organisms. The first letters of the genus and species of the organism are used. For example, *BnSOC1* refers to *B. napus* homologues of the Arabidopsis gene *SOC1*.

(Figure 2.25). All four *BnFT* genes exhibit moderate expression prior to cold treatment. During vernalization, *BnFT* gene expression decreases to very low values, with expression increasing when plants are returned to growth in warm, long day conditions. Between the penultimate and final time points, the A7 and C6 copies exhibit a significant decrease in their expression, while the A2 and C2 copies do not. This decrease in expression is not as severe for the *BnFT.A7* gene, resulting in the gene being assigned to both regulatory modules (Figure 2.25). In the leaf, therefore, *BnFT.A2* and *BnFT.C2* both exhibit a divergent expression trace to *BnFT.C6*, but *BnFT.A7* shows similarities in its expression trace with all homologues. This suggests subtle regulatory divergence between the copies of *BnFT*. Comparing the magnitude of expression, the A genome copies of *BnFT* are more highly expressed than the copies on the C genome. *BnFT.A2* is generally five-fold more highly expressed across the time series relative to *BnFT.C2.Random*[4], while *BnFT.A7* is two- to three-fold more highly expressed than *BnFT.C6*. This genome of origin bias suggests that the A genome copies potentially influence flowering to a greater extent than the C genome copies.

To determine whether the *BnFT* genes exhibit tissue-specific expression, the expression of these four genes was analysed in the apex samples. In the apex, only two of the *BnFT* genes are expressed; *BnFT.A7* and *BnFT.C6* (Figure 2.26). As opposed to the expression pattern observed in the leaf (Figure 2.25), the expression of both copies begins lowly expressed, gradually increasing during the time series until decreasing at the final time point. The magnitude of expression of both copies is similar. These findings suggest that the *BnFT* genes may indeed have diverged in their spatial expression domains, with *BnFT.A7* and *BnFT.C6* exhibiting expression in both the leaf and the apex, whereas *BnFT.A2* and *BnFT.C2.Random* are only expressed in the leaf. In addition, the expression of the *BnFT* genes in the apex does not seem to be as responsive to the cold treatment as the copies in the leaf, suggesting that

[4]The *B. napus* reference genome[118] constructed sequence scaffolds that were joined to generate 19 pseudochromosomes. Scaffolds that mapped to a pseudochromosome but could not be oriented were denoted 'random'. Unmapped scaffolds that could be assigned to the A or C genome were denoted 'Ann' and 'Cnn' respectively. Scaffolds that were not mapped during any of these steps were denoted 'Unn'. Throughout this work, similar notation is used to indicate the scaffold on which the gene is located.

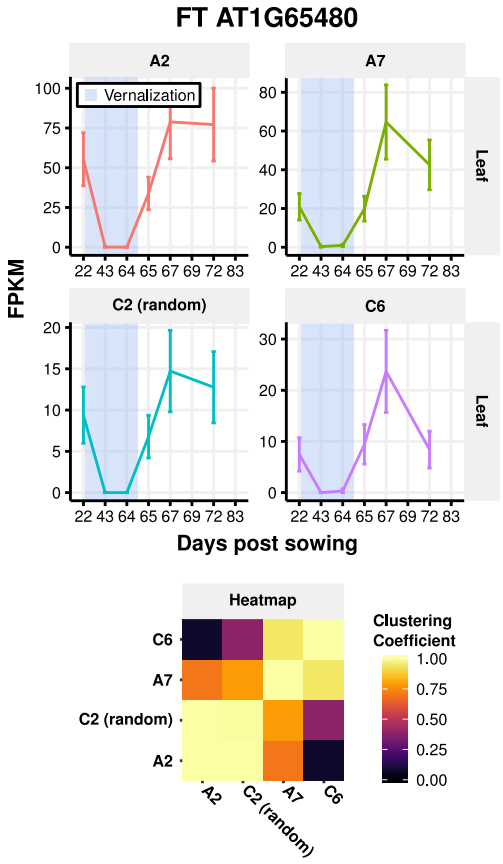


Figure 2.25: Expression traces for the *BnFT* genes in the Westar leaf. The expression values in FPKM and the 95% confidence intervals of those expression values as computed by Cufflinks are displayed. A heatmap of the clustering coefficients calculated by the SOM based method (Section 2.3.5) is also displayed. The expression patterns between the four genes are similar, yet diverge at the final time point, with the A7 and C6 copies decreasing in expression while the A2 and C2 copies do not.

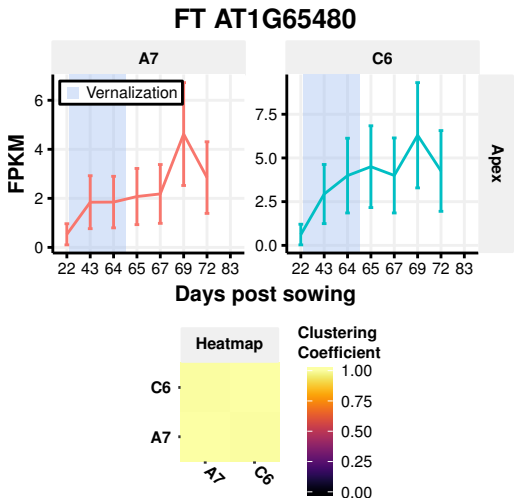


Figure 2.26: Expression traces for the *BnFT* genes in the Westar apex. The expression values in FPKM and the 95% confidence intervals of those expression values as computed by Cufflinks are displayed. A heatmap of the clustering coefficients calculated by the SOM based method (Section 2.3.5) is also displayed. The A7 and C6 copies exhibit very similar expression traces, increasing gradually during the time series.

potentially different pathways are regulating the expression of *BnFT* genes in the apex relative to the leaf.

Taking the results from the two tissues together reveals that the A2 and C2 copies of *BnFT* exhibit similar expression profiles, which are distinct to those of the A7 and C6 copies. In the leaf, the factor differentiating these sets of copies is the expression of the genes at the end of the time series. In Arabidopsis, *FT* increases in expression during long days that are inductive to flowering[302]. Assuming the same is true in *B. napus*, the decrease in expression of *BnFT.A7* and *BnFT.C6* is unexpected. A potential explanation could be that the *BnFT* genes have diverged in their target genes. *FT* activates the expression of MADS-box containing genes in Arabidopsis to promote flowering[47–49]. However, some MADS-box containing genes have dual roles in floral development, influencing both the floral transition and floral organ identity[75,303] *AGL24*, for example, promotes the formation of the inflorescence meristem, but is repressed at later points to allow the meristem to differentiate into floral organs[303]. It is conceivable that the A7 and C6 copies of *BnFT* influence the expression of genes that need to be repressed to allow floral development to occur, while the A2 and C2 copies do not.

The differences in the magnitude of expression reveal that the A genome copies are more highly expressed than the C genome copies. Although the magnitude of expression is not necessarily an indication of the role that gene plays in the plant, it is interesting to note that variation in *BnFT.A2*, the most highly expressed copy in the leaf, was found to be associated with variation in flowering time[301]. It is therefore possible that the expression differences observed between the *BnFT* genes do indeed influence the effect the genes have on the floral transition.

The decrease in expression of all *BnFT* genes in the leaf during vernalization is likely a consequence of the change in photoperiod. The vernalization treatment consisted of short day conditions (8 hours of light) at 5 °C. When Arabidopsis plants, grown in long day, floral inductive conditions, are transferred to short day growth conditions, *FT* expression decreases[302]. As *B. napus* also requires long days for the induction of flowering[304], the expression of *BnFT* during the vernalization period is consistent with a photoperiod driven repression. An alternative explanation could be that the *BnFT* genes are responding to

temperature during the vernalization period, given that both the ambient temperature response[244] and the vernalization response[30] have been implicated in the control of *FT* in Arabidopsis. However, the ambient temperature pathway generally responds to less severe changes in temperature[305], and a *BnFLC* gene with an expression profile consistent with *BnFT* repression during the cold is not present in Westar (Figure 3.15). This suggests that all four copies of *BnFT* are influenced by the photoperiod pathway in the leaf.

Finally, the copies exhibit further regulatory divergence in terms of tissue-specific expression, with A7 and C6 being the only *BnFT* genes expressed in the apex. A potential explanation for observing these expression patterns could be from residual leaf and stem tissue surrounding the apex due to the dissection procedure (section 2.2.1). However, that the expression profiles are different in the apex relative to the leaf, and that *BnFT.A2*, the most highly expressed copy in the leaf, is not observed in the apex implies this is not the case. Although expression of *FT* has been detected in the apex in Arabidopsis[22,233], it has been shown that *FT* mRNA is not required in the apex for its role in promoting the floral transition[22,45,49]. This suggests that the *BnFT.A7* and *BnFT.C6* may have a functional role in the apex that is not related to the floral transition. The lack of a response to vernalization for the *BnFT* genes in the apex may be due to the leaf being the primary plant organ that senses photoperiod signals[17,18,20–22]. Therefore, potentially the Arabidopsis *FT* gene has an heretofore unknown function in the apex that is unrelated to flowering and is conserved in the A7 and C6 copies of *FT* in *B. napus*.

2.4.2 *APETALA 1*

The transcription factor *AP1* controls both meristem identiy and floral organ specification[74]. In Arabidopsis, *AP1* mRNA is uniformly expressed in the floral meristem and is later localized to the sepals and petals[74]. No *AP1* RNA was detected in Arabidopsis roots, stems, leaves, or inflorescence meristems[74], suggesting the shoot apex is the primary domain of *AP1* expression. Seven copies of *BnAP1* are found in the transcriptomic time series on chromosomes

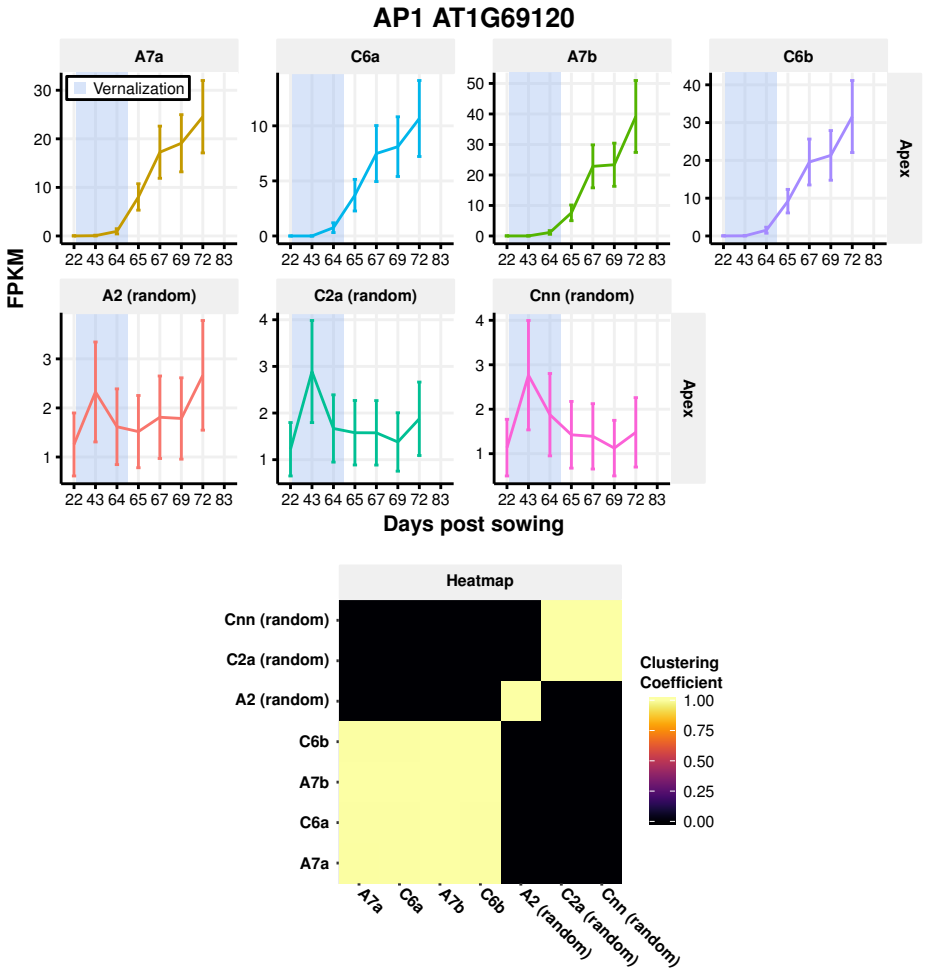


Figure 2.27: Expression traces for the *BnAP1* genes in the Westar apex. The expression values in FPKM and the 95% confidence intervals of those expression values as computed by Cufflinks are displayed. A heatmap of the clustering coefficients calculated by the SOM based method (Section 2.3.5) is also displayed. The expression profiles of the four A7 and C6 copies are very similar to each other. The remaining copies exhibit similar expression profiles, although *BnAP1.A2.Random* diverges in expression relative to the C2 and Cnn copies towards the end of the time series.

A2, C2, Cnn, two copies on A7, and two copies on C6. All copies are only expressed in the apex tissue, in line with expectations from Arabidopsis[74].

The *BnAP1* genes exhibit a *distinct* regulatory module assignment, with three patterns of regulation (Figure 2.27). The two A7 and two C6 copies display low expression initially and during the cold, with a steady and gradual increase until the final time point. The A2, C2a[5], and Cnn copies show somewhat similar expression traces, which diverge at the final time point. All three exhibit an increase in expression at the midpoint of the vernalization treatment, with a return to pre-treatment expression levels by the end of cold. The C2a and Cnn copies maintain this expression level until the end of the time series, while the A2 copy exhibits a slight increase in expression at the final time point. In terms of the magnitude of expression, the two pairs of homoeologues on A7 and C6 have expression levels an order of magnitude higher than the other copies. Comparing the magnitude of expression between the genes located on the same chromosome reveals that the copy located further along the chromosome is more highly expressed on both chromosome A7 and C6.

The expression of the A7 and C6 copies is most similar to the expression pattern of *AP1* in Arabidopsis, with expression lacking in inflorescence meristems and present in floral meristems, increasing as the meristem increases in size[74]. This suggests that these copies are acting redundantly to promote floral meristem identity. The magnitude differences observed between copies located on the same chromosome suggests that the genetic factors controlling this difference may have been established in an ancestral Brassica before *B. rapa* and *B. oleracea* diverged 0.12 - 3.7 million years ago[114,115]. The expression patterns of the A2, C2, and Cnn copies of *BnAP1* respond to growth in short days and cold temperatures, which is not typical of *AP1* expression in Arabidopsis. A potential explanation is provided by the expression profiles of *BnSVP* genes in *B. napus* (Figure 6.6; Appendix A). The A4, C4, and Ann copies of *BnSVP* all exhibit a similar expression response during the vernalization period as A2, C2, and Cnn. As *AP1* and *SVP* form dimers[91] in Arabidopsis, potentially this response is a consequence of those interactions. It should be noted, however, that the expression levels of *BnAP1.A2*, *BnAP1.C2a.Random*, and

[5]When multiple homologous gene models are located to the same chromosome, letters are appended to the chromosome to allow the gene models to be distinguished.

BnAP1.Cnn.Random are very low relative to the A7 and C6 copies, suggesting their expression in the apex may not have as much of a regulatory effect as the more highly expressed copies.

2.4.3 *SUPPRESSOR OF OVEREXPRESSION OF CO 1*

SOC1 is a gene in Arabidopsis involved with integrating the inputs from the photoperiod[20], vernalization[85,86], hormone[87], and age-dependent[89] floral pathways. Expression of *SOC1* has been detected in the shoot apical meristem, leaves, stem, and roots of Arabidopsis plants[20,85], but not in vegetative meristems[306]. The role of *SOC1* in flowering is primarily mediated by its expression in the apex, although expression of the gene in the vasculature has also been found to mediate an effect on the floral transition[31]. A number of regulatory interactions govern the expression of *SOC1* in Arabidopsis. *SOC1* and *AGL24* regulate each other in a positive feedback loop[90], while *FT*, *CO*, and *FLC* have been implicated in *SOC1* upregulation during a shift from growth in short day to long day conditions[307]. Mutant analysis suggested a hierarchy of regulation such that *FT* regulates *SOC1*, which in turn regulates *LFY*[48]. In *B. napus* we find six copies of *BnSOC1* expressed in both the apex and the leaf samples, located on chromosomes A3, A4, A5, Cnn, and two copies on C4.

As *SOC1* has been found to act in the apex[31,90], the expression of the *BnSOC1* genes were assessed in this tissue. In the apex, a *distinct* regulatory module assignment is observed (Figure 2.28). The *BnSOC1.A3.Random* copy and *BnSOC1.A4* copy exhibit different expression profiles relative to every other *BnSOC1* gene with the other four gene exhibiting similar expression profiles. There are two time points in development where the expression of the *BnSOC1* genes increase. These time points are day 43, during the cold treatment, and at day 69 post-sowing. However, the increase at these time points are only observed in some of the copies. The four copies that demonstrate similar expression profiles (*BnSOC1.A4*, *BnSOC1.A5*, *BnSOC1.Cnn*, and *BnSOC1.C4.Random*) exhibit an increase in expression at both of these time points. Interestingly, the relative expression between these peaks varies

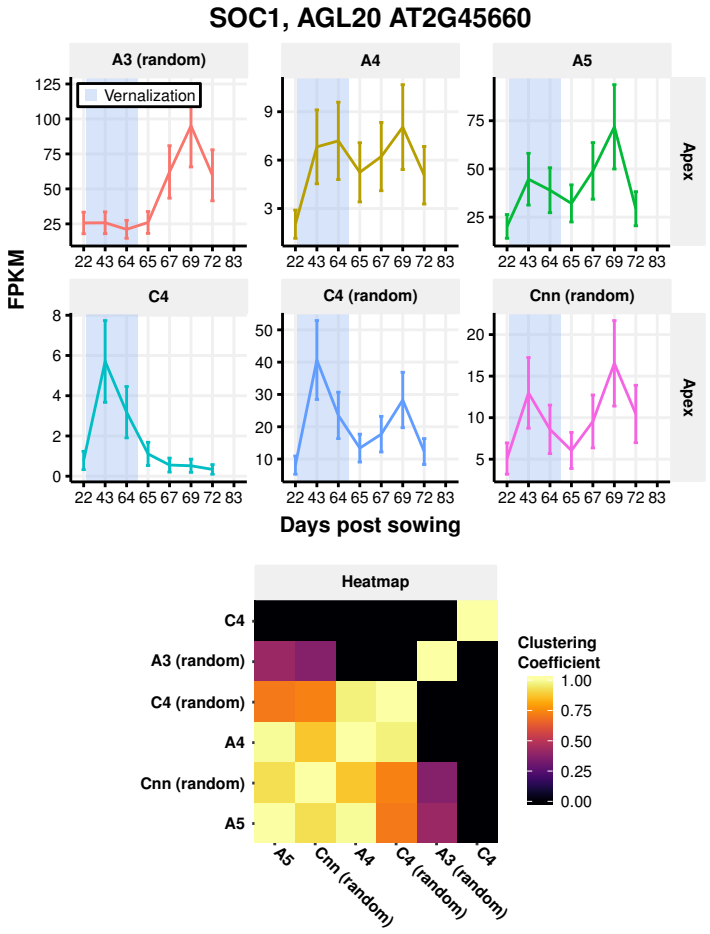


Figure 2.28: Expression traces for the *BnSOC1* genes in the Westar apex. The expression values in FPKM and the 95% confidence intervals of those expression values as computed by Cufflinks are displayed. A heatmap of the clustering coefficients calculated by the SOM based method (Section 2.3.5) is also displayed. Expression profiles of *BnSOC1.A4*, *BnSOC1.A5*, *BnSOC1.C4.Random*, and *BnSOC1.Cnn.Random* are similar, increasing both during vernalization and towards the end of the time series. The other two copies only exhibit one of these increases, with *BnSOC1.C4* increasing during vernalization and *BnSOC1.A3.Random* increasing towards the end of the time series.

between the copies. The *BnSOC1.A5* copy is expressed ~50% higher at the day 69 time point relative to the time point taken at day 43. Conversely, the same comparison made with the *BnSOC1.C4.Random* gene reveals that the gene is expressed ~25% lower at day 69 relative to day 43 of the time series. The A3 and C4 copies exhibit expression profiles that are divergent from the other four copies. Expression of the *BnSOC1.A3.Random* copy is high but stable during the cold treatment with an increase in expression post-cold peaking at day 69. This is contrasted by the *BnSOC1.C4* copy that peaks in expression at the day 43 time point, then returns to very low expression post-cold. These results suggest that the *BnSOC1* genes respond to the cold treatment and increase in expression during the floral transition. However, the different copies exhibit regulatory divergence in terms of the degree to which they respond to these two signals. When the magnitude of expression between the copies is compared, *BnSOC1.A3*, *BnSOC1.A5*, and *BnSOC1.C4.Random* exhibit the highest expression levels. However, even within these genes, significant divergence is observed with *BnSOC1.A3* and *BnSOC1.A5* expressed approximately two-fold more highly than *BnSOC1.C4.Random*. This suggests regulatory divergence in terms of the magnitude of expression, in addition to expression profile differences.

The expression of *SOC1* in the Arabidopsis apex is proposed to occur in a positive feedback loop with the gene *AGL24*[90]. To test if this interaction is also observed in *B. napus*, the expression profiles of *BnAGL24* were compared to those of *BnSOC1*. Four copies of *BnAGL24* are expressed in the apex, situated on chromosomes A1, C1, A3, and C7 (Figure 2.29). The expression of the A1 and C1 genes increases gradually during the time series, decreasing at the final time points. The A3 and C7 copies, however, show an almost inverse expression profile; highly expressed initially with a gradual decrease during the time series. Comparing these expression profiles with those of *BnSOC1* reveals that the expression of the *BnAGL24.A1* and *BnAGL24.C1* genes is consistent with with regulatory feedback with all *BnSOC1* genes except the C4 copy. Likewise, *BnAGL24.A3* and *BnAGL24.C7* potentially regulate all *BnSOC1* genes except *BnSOC1.A3.Random*. The expression profiles of *BnAGL24* suggest, therefore, that the positive feedback loop may exist between these genes in *B. napus*, but copy specificity is observed.

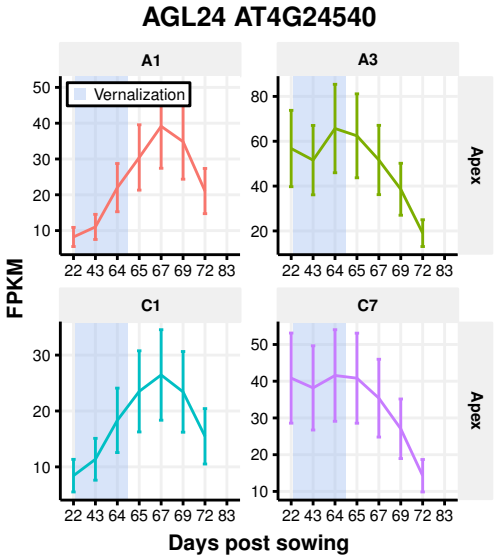


Figure 2.29: Expression traces for the *BnAGL24* genes in the Westar apex. The expression values in FPKM and the 95% confidence intervals of those expression values as computed by Cufflinks are displayed. The A3 and C7 copies exhibit a decrease in expression over the time series while A1 and C1 increase over the time series. Both of these expression traces are consistent with *BnAGL24* interacting with *BnSOC1* genes.

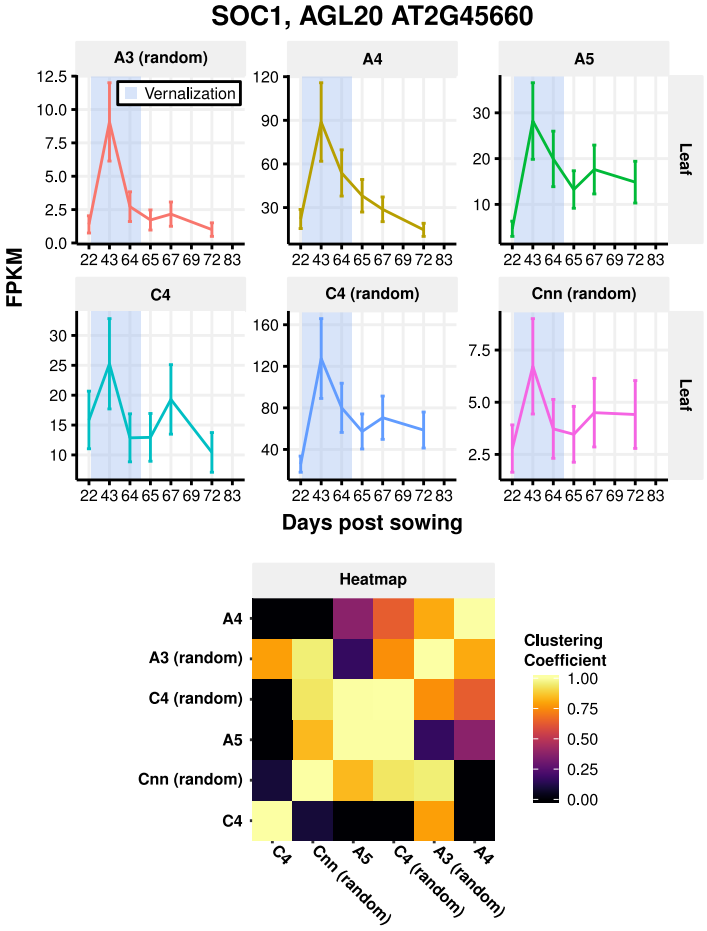


Figure 2.30: Expression traces for the *BnSOC1* genes in the Westar leaf. The expression values in FPKM and the 95% confidence intervals of those expression values as computed by Cufflinks are displayed. A heatmap of the clustering coefficients calculated by the SOM based method (Section 2.3.5) is also displayed. The expression profiles of all *BnSOC1* genes increases during vernalization. The expression profiles exhibit a complex *gradated* pattern of regulatory module assignment, with the difference between pre- and post-cold expression levels being the main differentiator.

To determine if the *BnSOC1* genes exhibit tissue-specific regulatory divergence, the expression of the genes was assessed in the leaf. The same six copies of *BnSOC1* are expressed in the leaf as in the apex. The *BnSOC1* copies in the leaf exhibit a *gradated* regulatory module assignment, suggesting subtle differences between the expression profiles of the *BnSOC1* genes (Figure 2.30). A commonality between the expression patterns is the response to the cold treatment, with all six of the copies peaking in expression at day 43 of the time series, halfway through vernalization. The differentiating factor between the expression profiles of *BnSOC1* genes in the leaf is the difference between the pre- and post-cold expression levels. At one extreme, the *BnSOC1.A5* and *BnSOC1.C4* genes are expressed approximately two-fold higher post-cold relative to before the treatment. This is in contrast to the *BnSOC1.A3* and *BnSOC1.A4* genes, that are expressed at similar levels before and after the treatment. This finding suggests that all copies of *BnSOC1* respond to the cold treatment when it is occurring, but only some copies continue to respond to the treatment when it ends. As observed in the apex, expression magnitude differences are also observed between the copies in the leaf. *BnSOC1.A4* and *BnSOC1.C4.Random* exhibit the highest expression levels, with the next most highly expressed copy, *BnSOC1.A5*, expressed three- to four-fold lower.

These results from both the apex and leaf suggest regulatory divergence of the *BnSOC1* genes, both in terms of expression profile and tissue-specific expression. From Arabidopsis it has been shown that *SOC1* is activated in the apex by the photoperiod pathway downstream of *FT* and *CO*[20,48,84,307,308]. Based on the expression of *BnFT* (Figure 2.25), *BnSOC1.A3.Random* is the only *BnSOC1* gene with an expression pattern consistent with this regulation (Figure 2.28). This is also supported by the *BnSOC1.A3.Random* copy exhibiting the highest expression of all the copies in the apex. Therefore, *BnSOC1.A3.Random* is a good candidate for carrying out the role of *SOC1* in *B. napus*.

All other *BnSOC1* genes in the apex, and all *BnSOC1* genes, including the A3 copy in the leaf, exhibit an increase in expression during the cold treatment. This is interesting given that in Arabidopsis, *SOC1* expression is activated during vernalization by both *FLC* dependent[30,86] and independent[87] pathways. Although Westar is a spring variety, it still exhibits a weak vernalization response[241], and a number of *BnFLC* genes exhibit expression consistent with

BnSOC1 activation in the leaf and apex (Figures 3.15 and 3.11). Therefore, potentially the vernalization response is mediating the cold-induced increase in *BnSOC1* expression. This hypothesis is strengthened by the observation that some *BnSOC1* genes in the leaf do not return to pre-cold levels after the cold, a response that would be expected from vernalization sensitive genes.

Taken together, the transcriptomic time series reveals regulatory divergence between *SOC1* homologues in *B. napus*, which seems to be tissue specific. In the apex, different expression profiles suggest that different copies of *BnSOC1* are sensitive to different environmental inputs. The relative magnitudes of expression between *BnSOC1* genes differ depending on the tissue, with *BnSOC1.A3.Random* and *BnSOC1.A5* copies being most highly expressed in the apex and *BnSOC1.C4.Random* and *BnSOC1.A4* in the leaf. Both of these examples of regulatory divergence suggest that the *BnSOC1* genes have subfunctionalized, both in terms of the inputs they respond to and the tissues in which they are expressed.

2.4.4 *FD*

The FD protein is a bZIP transcription factor that interacts with FT and TFL1 proteins[41,47,49] to mediate their effect on the floral transition. *FD* expression in Arabidopsis is high at the shoot apex and does not exhibit circadian oscillations or photoperiod dependent expression, with *FD* expression decreasing soon after *AP1* expression begins to increase[47,49]. The upregulation of *FD* was found to be mediated by LFY, with two LEAFY binding sites being found in the *FD* promoter[41]. In the transcriptomic time series there are six copies of *BnFD* expressed in the apex, situated on chromosomes A1, A8, Ann, C1, C3, and C7.

The expression of *FD* in Arabidopsis is primarily in the apex[47,49]. Investigating the expression of *BnFD* genes in the apex reveals a *distinct* regulatory module assignment (Figure 2.31). Five of the six copies have similar expression profiles to each other. These copies, consisting of the A8, Ann, C7, C1, and C3 copies, are relatively lowly expressed before and during cold and increase in expression after vernalization. After peaking in expression at day 67 of the time series, these genes decrease in expression. Some slight variation in the expression profiles of these copies is observed at the initial time points,

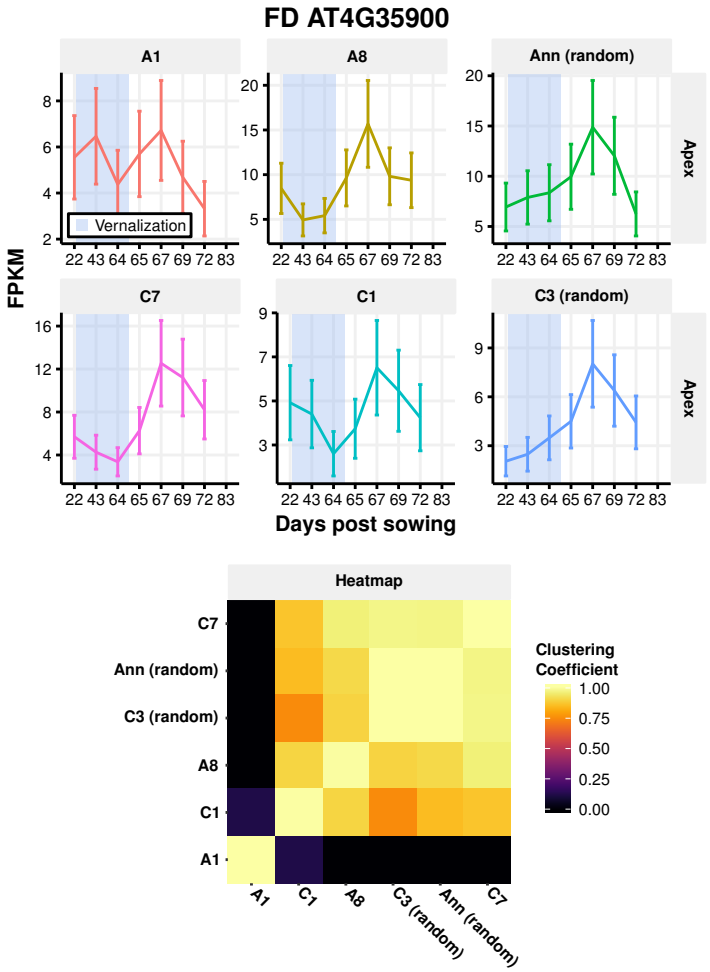


Figure 2.31: Expression traces for the *BnFD* genes in the Westar apex. The expression values in FPKM and the 95% confidence intervals of those expression values as computed by Cufflinks are displayed. A heatmap of the clustering coefficients calculated by the SOM based method (Section 2.3.5) is also displayed. Expression of five *BnFD* genes exhibit similar expression profiles, increasing in expression during the time series until day 67, and then decreasing. *BnFD.A1* exhibits a different response, staying approximately constant in expression throughout the time series.

with *BnFD.C1* exhibiting a decrease during the cold. This is reflected in the slightly lower clustering coefficients between *BnFD.C1* and the other copies assigned to the same regulatory module (Figure 2.31). Whether this difference is biologically relevant, however, would need further validation. Comparing the magnitude of expression between these five copies reveals that the *BnFD.C1* and *BnFD.C3.Random* are more lowly expressed than the other copies. The final copy, *BnFD.A1* exhibits a relatively noisy expression trace throughout the entire time series. This data suggests that, aside from the *BnFD.A1* copy, the *BnFD* genes have not diverged significantly from one another in terms of expression.

The expression of the *BnFD* genes exhibits similarities to the *FD* gene in Arabidopsis; apex-specific expression with an increase in expression during the floral transition[47,49]. The timing of the decrease in *FD* expression after the day 67 time point corresponds with the increase in four *AP1* copies (Figure 2.27), as observed in Arabidopsis[47], and also with the increase in *BnLFY* gene expression (Figure 2.32), consistent with the direct repression of *FD* by *LFY*[41]. Therefore, five of the six *BnFD* copies seem to be regulated in a similar manner to *FD* in Arabidopsis. The expression levels of all six *BnFD* copies are relatively similar in the plant. Both the similar expression patterns and the similar expression magnitudes suggest that the *BnFD* genes may have been maintained in the *B. napus* genome due to gene dosage effects.

2.4.5 *LEAFY*

LFY is a transcription factor that acts synergistically with *AP1*[80] to promote the floral transition and specifiy the determinacy of the floral meristem[61]. The gene is expressed in the floral primordia in Arabidopsis and increases during flower development[80], promoting the expression of other floral integrators such as *AP1*[63–65] and *TFL1*[66]. In the *B. napus* genome, four copies of the gene are found, one on chromosome A6, and three assigned to the C genome but not to a particular chromosome in the Darmor-*bzh* reference genome.

The four copies of *BnLFY* are only expressed in the Westar apex. The four copies of *BnLFY* exhibit a *redundant* regulatory module assignment, with all copies exhibiting low expression initially and increasing in expression after

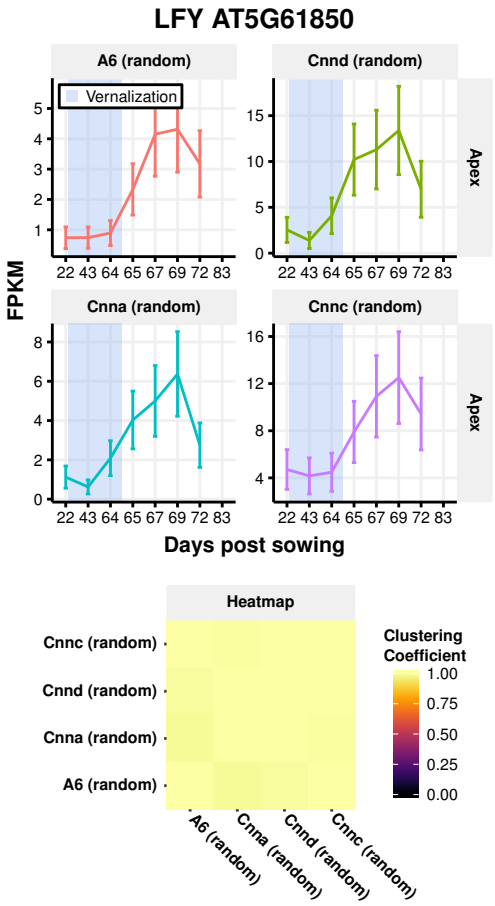


Figure 2.32: Expression traces for the *BnLFY* genes in the Westar apex. The expression values in FPKM and the 95% confidence intervals of those expression values as computed by Cufflinks are displayed. A heatmap of the clustering coefficients calculated by the SOM based method (Section 2.3.5) is also displayed. All copies of *BnLFY* exhibit a similar expression profile, with low initial expression and an increase in expression after vernalization.

vernalization (Figure 2.32). At the final time point, a decrease in expression is observed. This expression profile, increasing during the floral transition, is consistent with the expression of *LFY* in Arabidopsis[71]. Both the expression traces and the apex-specific expression is consistent with the expression of *LFY* in Arabidopsis, with a gradual increase during development until flowering[71,80].

The expression traces of the *BnLFY* genes are consistent with the regulatory interactions observed for *LFY* in Arabidopsis. Five of the six *BnSOC1* genes expressed in the apex exhibit a peak in expression at day 69 (Figure 2.28), in agreement with *LFY* being regulated by *SOC1*[48,69]. The expression of certain *BnAP1* and *BnTFL1* genes is also consistent with *BnLFY* mediated regulation (Figures 2.27, 2.33), as has been observed in Arabidopsis[63–66]. This evidence suggests that the *BnLFY* genes are similarly regulated to their homologue in Arabidopsis, and that the regulatory roles elucidated for *LFY* in Arabidopsis seem to be conserved in *B. napus*.

The co-regulation of the *BnLFY* genes is consistent with the gene balance hypothesis[224,227]. Dosage balance is also consistent with observations in Arabidopsis. The *LFY* null mutation was found to be haploinsufficient under short day conditions[72], while insertion of additional copies of *LFY* into the Arabidopsis genome altered the flowering time of the transformed plants, with an additional shortening of the flowering time observed with each additional copy of *LFY*[71]. These findings suggest that potentially the copies of *BnLFY* have been maintained in the *B. napus* genome as their loss, or an alteration of their expression, results in a change in flowering time. A prediction that arises from this is that a *B. napus* plant lacking a copy of *BnLFY* would have later flowering. *LFY* has a dual role in both determining the timing of the floral transition and mediating correct floral patterning[61]. Assuming that the copies of *BnLFY* are redundant, a single inactive copy could potentially alter flowering time without altering floral patterning, due to the other copies being able to complement the inactive copy. These findings could therefore provide a potential avenue for altering flowering time in *B. napus*.

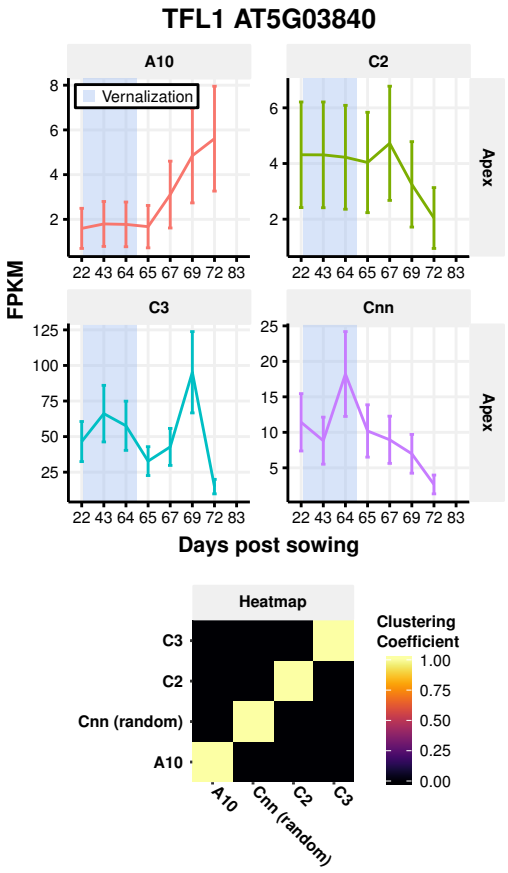


Figure 2.33: Expression traces for the *BnTFL1* genes in the Westar apex. The expression values in FPKM and the 95% confidence intervals of those expression values as computed by Cufflinks are displayed. A heatmap of the clustering coefficients calculated by the SOM based method (Section 2.3.5) is also displayed. The *BnTFL1* genes exhibit total divergence in their expression profiles, with the C genome copies of the gene being more highly expressed than the A genome copies.

2.4.6 *TERMINAL FLOWER 1*

TFL1 acts in an antagonistic manner to *FT* in Arabidopsis[50], with the gene maintaining inflorescence meristem identity by limiting the expression of *AP1* and *LFY*[53–55]. The expression domain is just below the growing meristem at the apex, and also in the axillary meristems[55,234]. The expression is initially low, with an increase when the floral transition occurs[52,55,56,234]. In agreement with previous studies[152,153], four *BnTFL1* genes were identified in the transcriptomic time series on chromosomes A10, C2, C3, and Cnn.

The four *BnTFL1* genes exhibit a *unique* pattern of regulatory module assignment (Figure 2.33), with each gene assigned to a separate module. The *BnTFL1.A10* copy is very lowly expressed initially and remains at that level until after the cold treatment. From the day 67 time point onwards, the expression of this copy increases until the final time point. Conversely, the *BnTFL1.C2* copy effectively exhibits the inverse response, with expression before, during and after the cold treatment being comparatively high before decreasing after the day 67 time point. *BnTFL1.C3* is the most highly expressed copy of *BnTFL1*, with expression levels an order of magnitude higher than the A10 and C2 copies. The expression of the C3 copy increases during vernalization with a return to pre-cold levels when plants are transferred back to warm, long day growth conditions. The copy increases in expression to a peak at day 69 of the time series, before decreasing in expression at the final time point. Finally, the *BnTFL1.Cnn.Random* copy shows a transient peak of expression towards the end of vernalization, with a continued decrease in expression until the final time point thereafter.

The expression profiles of *BnTFL1.A10* and *BnTFL1.C3* are most consistent with the expression of *TFL1* in Arabidopsis, as both show increasing expression during the floral transition[52,55,56,234]. These copies differ in their behaviour during the cold treatment and at the final time point. In Arabidopsis, the floral structure is indeterminate and this requires continued expression of *TFL1* at the apex[52]. This pattern of expression is exhibited most clearly by *BnTFL1.A10*, as *BnTFL1.C3* decreases in expression at the final time point. An explanation for this decrease may be due to *BnTFL1.C3* only

maintaining the inflorescence meristem identity early in development, with this role performed by other genes later in development.

Comparing the expression of *BnTFL1.C3* and *BnTFL1.A10* to *BnAP1* and *BnLFY*, the mutual antagonism observed between these genes in Arabidopsis[53–55,66] is not seen between the *B. napus* homologues of these genes. This is potentially due to the apex sampling procedure (section 2.2.1) not separating the expression domains of these genes[52]. However, it is interesting that both *BnTFL1.C3* and the *BnLFY* genes (Figure 2.32) exhibit a decrease in expression at the final time point, given the mutual antagonism of the genes in Arabidopsis. The reduction in *BnLFY* activity potentially results in less *BnTFL1.C3* being required to maintain the inflorescence meristem state, or vice versa. The regulatory antagonism between *BnTFL1*, *BnAP1*, and *BnLFY* might be manifested in the repression of *BnTFL1.Cnn.Random* and *BnTFL1.C2* towards the end of the time series. The expression profiles of the four *BnTFL1* copies reveals that genes have diverged from each other in terms of regulation, and suggests that dosage effects have not influenced the retention of *BnTFL1* genes in the *B. napus* genome.

2.4.7 Conclusions

The floral integrators in Arabidopsis are integral to the interpretation of environmental signals to accurately coordinate the floral transition[41]. Whether the homologues of these Arabidopsis floral integrators have retained the same function in *B. napus* was previously only understood for relatively few examples[131,152–155,157,158]. This work has been complicated by Arabidopsis floral integrators often having multiple homologous genes in the *B. napus* genome[131]. To investigate whether the homologues of Arabidopsis floral integrators have expression profiles consistent with their function in the model species, the expression of *B. napus* floral genes was assessed in the transcriptomic time series. For all six of the floral integrators examined, at least one *B. napus* homologue exhibited an expression profile consistent with retaining a function similar to its Arabidopsis homologue. This suggests a general conservation of the gene regulatory network in *B. napus* relative to Arabidopsis. Testing these candidates could be achieved by expressing the gene in Arabidopsis mutants for

the gene, as has been done to investigate the efficacy of homologous *B. napus* flowering time genes previously[145].

An advantage of assessing gene expression for all genes simultaneously is that regulatory interactions known to exist between the floral integrators in Arabidopsis can be investigated in *B. napus*. For example, *SOC1* is upregulated by *FT* in Arabidopsis[20,48,84,307,308]. That five of the six *BnSOC1* genes are upregulated during vernalization (Figure 2.28), when all four *BnFT* genes exhibit very low expression (Figure 2.25), indicates that these *BnSOC1* genes are not upregulated as a result of *FT* expression. This in turn makes the one *BnSOC1* gene that does not increase during the cold, *BnSOC1.A3.Random*, the best candidate for exhibiting similar behaviour as *SOC1* in *B. napus*.

Finally, different patterns of divergence suggest different selective pressures may be acting on the *B. napus* floral integrator genes, despite the genes being involved with the same regulatory pathway in Arabidopsis. Co-regulation of floral integrators suggest that gene dosage effects may be playing a role[227]. This is particularly true for *BnLFY*, where dosage effects have also been demonstrated in Arabidopsis[71,72]. However, from the observed divergence it is also clear that subfunctionalization, neofunctionalization, or the evolution of responsive backup circuits have also influenced gene retention[206,213,219,220,229]. These different scenarios could be tested by identifying lines that have non-functional alleles of particular floral integrator genes and investigating how the expression profiles of the remaining floral integrators are different in those lines, or by identifying phenotypic effects of the mutation.

2.5 Sequence divergence between copies of two floral integrators

Comparative analysis of the DNA sequence of homologous genes in *Brassica* crops has been used to reveal divergence between the copies. An analysis of *Brassica* homologues of *FLC* found variation in the promoter of the gene, including some copies lacking a region of the promoter important for the expression of the gene in Arabidopsis[141]. For *FT* homologues in *B. napus*

and *B. oleracea*, a transposable element and a retro-element in the upstream promoter of the gene on chromosome C2 was correlated with a lack of expression relative to the other copies of the gene[154]. Among *BnTFL1* genes, sequence variation was identified within the first intron of the gene and in the 3' regulatory regions[152]. Other studies investigating sequence changes have instead focussed on polymorphisms between varieties, identifying regions of sequence important for gene function[131,140,155,157,158]. A common theme between these analyses is that the amino acid sequences of the analysed homologues are often very similar[141,152,158]. In the case of *BnTFL1* genes, for example, a maximum of 5 amino acid differences between the homologues was identified[152]. However, it has been shown that in Arabidopsis it only takes a single amino acid substitution to confer FT-like function onto TFL1 proteins, and vice versa[59]. Therefore, although the observed differences between *B. napus* genes may be minor, they have the potential to severely impact the function of the gene.

The transcriptomic time series allows sequence differences between *B. napus* floral integrators to be viewed in the context of gene expression during the floral transition. To illustrate how the transcriptomic time series can be used to facilitate insights on sequence divergence, two case studies will be considered. For *BnTFL1* genes, sequence divergence downstream of the gene, in regions identified as cis-regulatory elements, correlates with the expression divergence observed between the genes during the time series. In the case of *BnFD*, sequence polymorphisms within the bZIP domain are predicted to alter the dimerization affinity of the genes. The observed sequence differences in bZIP proteins are also identified in other species, suggesting that this form of divergence is common among duplicated bZIP proteins. Given that the *BnFD* genes are co-regulated during the time series, modelling studies reveal that the observed sequence divergence may impact the expression of genes regulated by FD.

2.5.1 *BnTFL1* cis-regulatory elements

Cis-regulatory elements downstream of the *TFL1* gene in Arabidopsis have been found to direct different aspects of gene regulation[309]. In the study by Serrano-Mislata et al. (2016), regions of sequence conservation between

the Arabidopsis *TFL1* and homologues in *Arabidopsis lyrata*, *Capsella bursa-pastoris*, *B. rapa*, and *Leavenworthia crassa* were identified up- and downstream of the gene. Further analysis of these regions determined that these areas of sequence conservation corresponded to cis-regulatory elements. Interestingly, different regions were found to influence *TFL1* expression in different ways. For example, one region identified 1.0 - 1.3 kilobases (kb) downstream of the gene was required for *TFL1* expression in the vegetative meristem, while another region situated 1.6 - 2.2 kb downstream of the gene was required for gene expression in lateral meristems[309]. These results are particularly interesting given the conservation of these cis-regulatory elements between Arabidopsis and *B. rapa*[309], and previous identification of between homologue variation in the 3' regulatory regions of *BnTFL1* genes[152].

2.5.1.1 Cis-regulatory element variation downstream of *BnTFL1* genes potentially explain observed regulatory divergence

To investigate whether the *BnTFL1* genes in the Darmor-*bzh* reference genome exhibit sequence variation in the 5' and 3' intergenic regions surrounding the genes, sequence conservation between the genes and *TFL1* was calculated. Several conserved regions within the intergenic regions were identified (Figure 2.34a). Serrano-Mislata et al. (2016) identified seven regions of inter-species sequence conservation surrounding the *TFL1* gene (denoted by green letters in figure 2.34a) and five regions that were experimentally verified to influence *TFL1* expression (denoted by blue numerals in figure 2.34a). Focussing the analysis on the five experimentally verified cis-regulatory elements, differences in the extent of sequence conservation within these regions are found between the *BnTFL1* genes. The high sequence conservation in region II and IV of *BnTFL1.C3* and *BnTFL1.A10* suggests these two copies of the gene possess Arabidopsis-like cis-regulatory elements. Conversely, the lack of sequence conservation in these two regions in the *BnTFL1.C2* and *BnTFL1.Cnn.Random* copies suggests these copies are lacking such regulatory sequence. Maximal sequence conservation within region III is below 50% in the *BnTFL1.Cnn.Random* copy, while this value is above 70% for the other three copies (81%, 87%, and 78% for *BnTFL1.A10*, *BnTFL1.C2*, and *BnTFL1.C3* respectively). Interestingly, the area of significant sequence conservation within

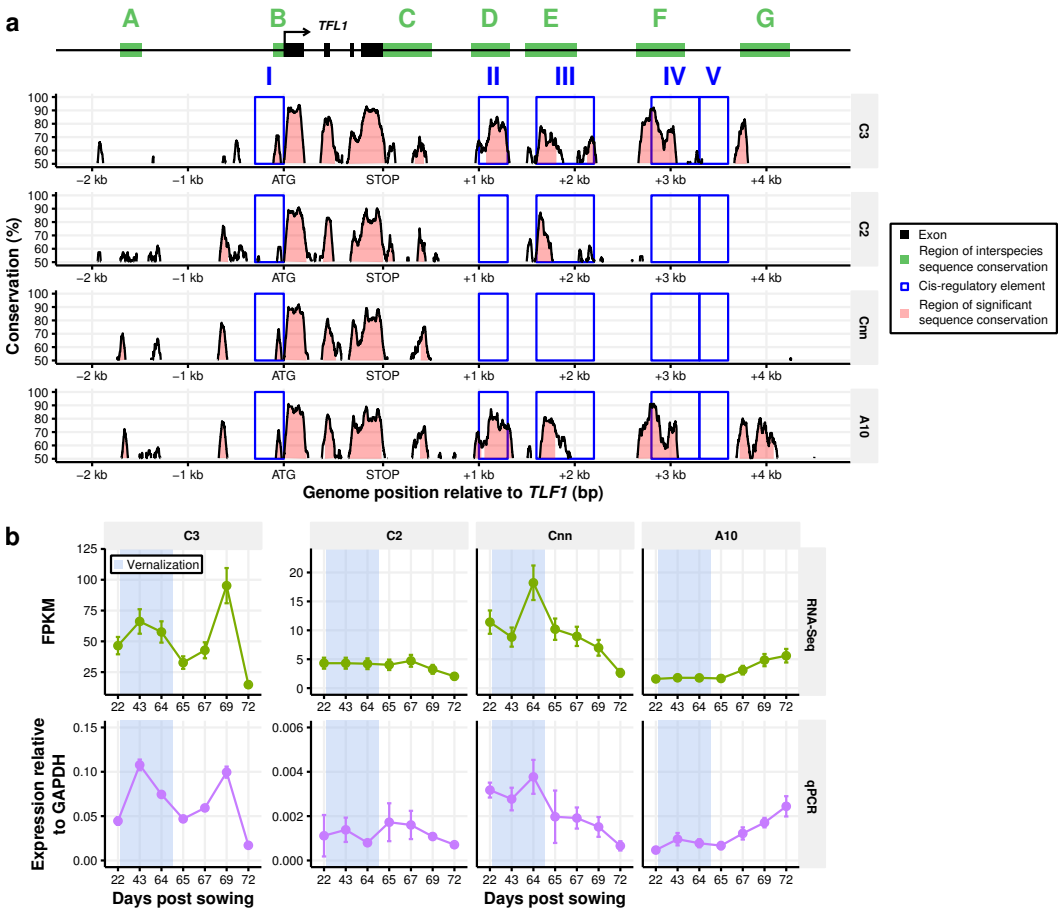


Figure 2.34: Sequence analysis reveals that cis-regulatory modules identified in Arabidopsis are not present downstream of some *BnTFL1* genes.

a The degree of sequence conservation between the *BnTFL1* genes and *TFL1* from Arabidopsis. Sequence alignment and conservation calculations were performed using the mVISTA server[475,476] with a sliding window size of 100 bp. The seven regions of high interspecies sequence conservation (green bars) and the five cis-regulatory regions (blue boxes) identified by Serrano-Mislata et al. (2016) are shown relative to the *TFL1* gene model[309] (black bars). The labelling of these regions follows the same conventions as the previous study. The pink shaded areas under the sequence conservation curves are regions above 70% sequence conservation. Genomic position upstream and downstream of the *TFL1* gene copy are given relative to the ATG and STOP codon sites respectively. *Continued on Page 130.*

Continued from Page 129. **b** The unnormalized expression profiles for the *BnTFL1* genes determined through RNA-Seq and qPCR. The expression values calculated for qPCR are normalized to *GAPDH* with the error determined from two biological replicates (Section 6.9; Methods).

region III in *BnTFL1.C2* (154 bases) and *BnTFL1.A10* (162 bases) is decreased compared to that of *BnTFL1.C3* (273 bases) copies, potentially suggesting the cis-regulatory elements in the former are incomplete. Considering regions identified as conserved across species by Serrano-Mislata et al. (2016), but not experimentally implicated in the regulatory control of *TFL1* (green shading in Figure 2.34a), sequence divergence is observed in region G. *BnTFL1.A10* exhibits high sequence conservation relative to Arabidopsis across this entire region, while *BnTFL1.C3* shows conservation over ~50% of the region. As with regions II and IV, *BnTFL1.C2* and *BnTFL1.Cnn.Random* lack conserved sequence in region G. A region of conservation not annotated in the previous analysis of *TFL1* cis-regulatory elements was also identified. This region, situated ~600 bp upstream of the transcription start site of *TFL1*, shows ~80% sequence conservation relative to Arabidopsis in *BnTFL1.A10*, *BnTFL1.C2*, and *BnTFL1.Cnn.Random*. In *BnTFL1.C3*, sequence conservation in this newly identified region is ~55%. These findings reveal that the *BnTFL1* genes identified in the transcriptomic time series exhibit sequence variation within potential cis-regulatory regions downstream of the gene.

2.5.1.2 Variation in cis-regulatory elements correlates with expression divergence

The experiments conducted to identify the regulatory effects of the cis-regulatory elements downstream of *TFL1* in Arabidopsis consisted of transgenic and mutational studies[309]. Insertion lines that disrupted cis-regulatory elements and transgenic lines transformed with reporter genes whose expression was driven by different combinations of the regulatory elements were used to dissect the role each element played in directing the correct spatiotemporal expression of *TFL1*. A prediction arising from the finding that certain *BnTFL1* genes seemingly lack these downstream

regulatory elements would be that the regulatory divergence observed between the genes (Figure 2.33) is a consequence of variation in cis-regulatory elements. To test this, expression patterns of *TFL1* in the mutant and transgenic lines of Serrano-Mislata et al. (2016) were compared to the expression of the *BnTFL1* genes during the transcriptomic time series. The *BnTFL1* genes that increase in expression during the floral transition (*BnTFL1.C3* and *BnTFL1.A10*) both show high sequence conservation in region II. Conversely, *BnTFL1.C2* and *BnTFL1.Cnn.Random* both lack sequence conservation in region II and are not unregulated during the floral transition. Region II was found to be necessary for the upregulation of *TFL1* during the floral transition in Arabidopsis[309], which correlates with the expression profiles of *BnTFL1* genes during the developmental time series. Another region showing a similar presence-absence pattern between the *BnTFL1* genes as region II is region IV. In Arabidopsis, this region corresponds to a cis-regulatory element responsible for driving the expression of *TLF1* in the inflorescence meristem[309]. Potentially the presence or absence of this region also contributes to the expression differences observed between the *BnTFL1* genes. Region III was found to be important for the expression of *TFL1* in the lateral meristems of the plant[309]. Sequence conservation within region III is below 50% for the *BnTFL1.Cnn.Random* gene. This finding predicts that this particular copy, therefore, would not be expressed in the lateral meristems in *B. napus*.

2.5.1.3 Quantitative PCR validation of *BnTFL1* RNA-Seq expression levels

The above observations of gene expression correlating with the presence and absence of cis-regulatory elements is dependent on the accuracy of the RNA-Seq results. Although findings presented in section 2.2.3 suggest that spurious expression levels as a result of read mismapping are a rare occurrence (Figure 2.9), the expression profiles of the *BnTFL1* genes were confirmed in a copy-specific manner. Quantitative PCR (qPCR) primers were designed to be specific to each of the four copies of *BnTFL1*, and qPCR performed (Section 6.9; Methods). The qPCR results obtained show strong similarity to the expression profiles derived from the RNA-Seq data (Figure 2.34b). As the qPCR primers designed were copy specific, this suggests that the expression

profile divergence observed for *BnTFL1* genes in the RNA-Seq data is not an artefact of read mismapping or incomplete gene models.

Taken together this reveals that the presence and absence of cis-regulatory elements downstream of the *BnTFL1* genes may confer similar regulatory control in *B. napus* as in Arabidopsis. *BnTFL1* genes contain different combinations of cis-regulatory elements, which have the potential to underlie the divergent expression profiles they exhibit.

2.5.2 *FD* dimerization

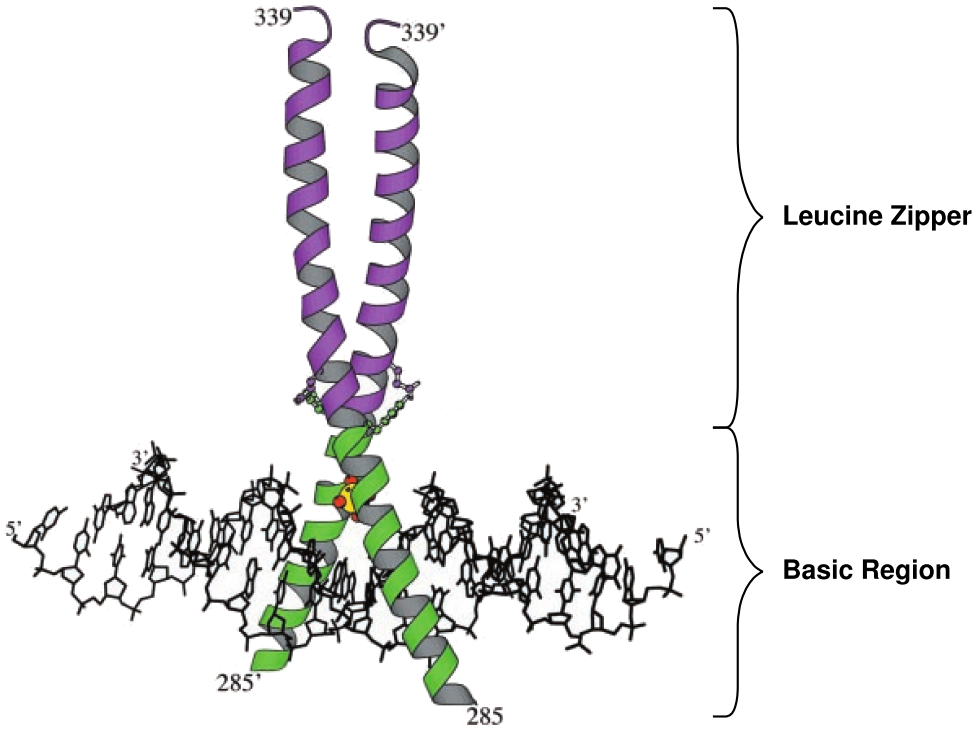


Figure 2.35: Structure of a bZIP transcription factor.
Ribbon diagram of the cAMP responsive element-binding protein bound to DNA. The leucine zipper region (purple) mediates the dimerization of the two monomers. The basic region (green) interacts with the major groove of DNA (black). Figure modified from Schumacher et al. (2000)[319].

The FD protein is a transcription factor that interacts with both FT and TFL1 proteins to mediate their association with DNA[41,49]. The FD protein contains

a basic region leucine zipper (bZIP) domain, making it a member of the bZIP transcription factor family[49]. This family of transcription factors interact with DNA as dimers (Figure 2.35)[310–312]. The structure of bZIP transcription factors consists of a basic region that interacts with the major groove of DNA and mediates the binding of the protein to transcription factor binding sites[310,312]. The dimerization of bZIP monomers is mediated by a coiled-coil structure of two α-helicies known as the leucine zipper[313]. The coiled-coil structure is stabilized by hydrophobic amino acid side chains, such as that of leucine, that form a hydrophobic core to the structure. In addition to the hydrophobic core of the interaction interface, charged amino acid residues adjacent to the core influence the binding of monomers through electrostatic interactions[310,314]. bZIP transcription factors are able to form homodimers, a dimer made from two copies of the same monomer, or heterodimers, where the two monomers are different[315]. Indeed, the dimers formed may influence the DNA sequences bound by the transcription factor, with dimerization acting as a key regulatory mechanism[316]. Changing dimerization and DNA-binding specificity has been found to be important in the evolution of bZIP transcription factor function[317].

Five of the six copies of *BnFD* expressed in the apex in *B. napus* share similar expression profiles (Figure 2.31). As a result, it is likely that their protein products are present in the cell at the same time, and would have the potential to interact to form dimers. Assuming all six BnFD proteins are able to dimerize, a total of 21 different dimer combinations are possible. To determine whether the BnFD proteins are capable of dimerizing, the protein sequences were compared. Between homologue differences in the protein sequence were identified between BnFD proteins, with a number of polymorphic sites identified within the bZIP domain. Amino acid differences observed in the basic region have the potential to influence DNA binding, while differences in the leucine zipper region are predicted to influence the dimerization affinities of the BnFD proteins. The amino acid divergence observed within the leucine zipper region was also found in bZIP proteins of other species, suggesting that this form of divergence is frequently observed among bZIP proteins. Computational modelling of monomer dimerization suggests that the differences in dimerization affinity could represent an interesting regulatory mechanism.

2.5.2.1 Protein sequence divergence exists between the six *BnFD* copies

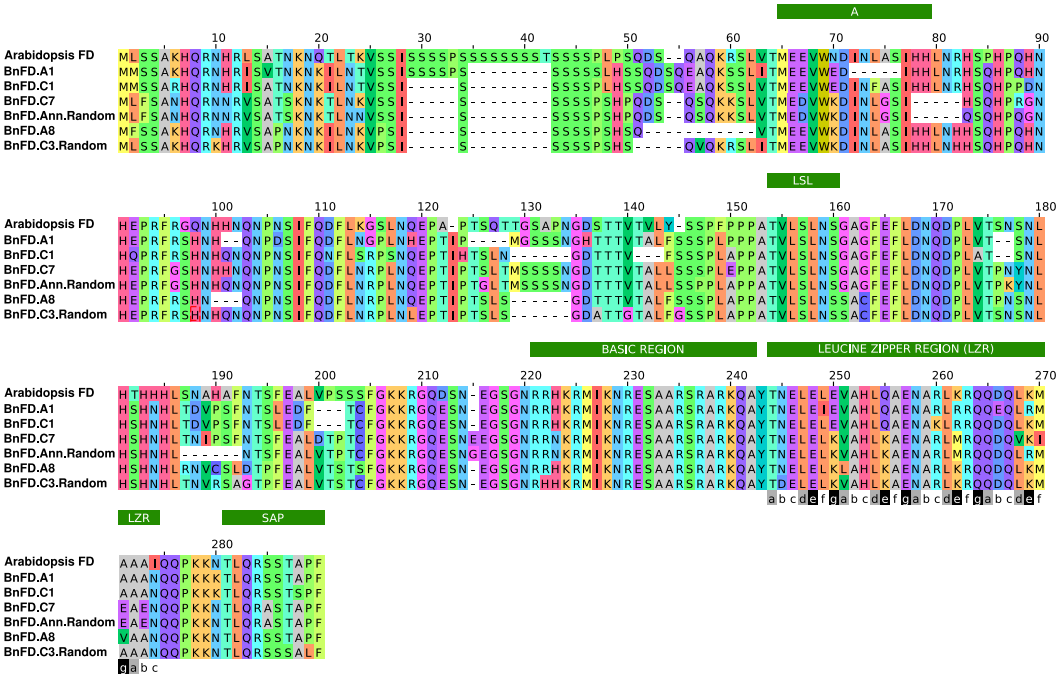


Figure 2.36: Multiple sequence alignment of the Arabidopsis and BnFD proteins The indicated regions of the protein are defined as in Tsuji et al. (2013)[318]. Between copy variation is observed in the A, BASIC, LEUCINE ZIPPER, and SAP regions, in addition to elsewhere in the protein. The heptad structure of the α-helix that makes up the leucine zipper region is displayed below the alignment of that region in the diagram. Amino acid residues located in the hydrophobic core are residues `a` and `d` (black). Amino acid residues capable of forming electrostatic interactions are in positions `e` and `g` (grey), with between copy variation visible in these positions.

In order to assess the extent of amino acid divergence between the six copies of BnFD, their predicted protein sequences were determined and aligned (Figure 2.36). To identify polymorphisms likely to affect the molecular function of the protein, the results of a comparative study of FD-like genes from many species were used[318]. The Arabidopsis FD protein was found to have four conserved regions: the A region, the LSL region, the bZIP region (composed of the basic region and a leucine zipper region), and the SAP region[318]. Focussing

on the same regions in *B. napus* (Figure 2.36) identifies a number of amino acid changes and deletions in the A region, with four different forms of the region present in the six BnFD proteins. Comparing the BnFD proteins to the Arabidopsis FD protein reveals that, in the A region, BnFD.A8 and BnFD.C3 show the greatest amino acid sequence similarity to the Arabidopsis FD protein, with only a single amino acid change present.

The LSL region displays no amino acid variation within the *B. napus* FD proteins or between species. This is consistent with the findings of Tsuji et al. (2013), which suggested the LSL region was indicative of FD-like proteins that played a role in the floral pathway[318].

In the SAP region, there are again a number of amino acid changes between the BnFD proteins (Figure 2.36). Of note is the amino acid polymorphism at position 287 between a threonine and serine. This position in Arabidopsis becomes phosphorylated and is important for the binding of FD to the protein FT in Arabidopsis, as mutation of the threonine to an alanine disrupts complex formation[49]. Changing the threonine to a serine was found to not affect FD binding to FT in Arabidopsis, although potentially different kinases are responsible for the phosphorylation of the different residues[49].

2.5.2.2 Polymorphisms in the DNA binding interface have the potential to affect binding affinities

The basic region of bZIP transcription factors consists of the protein-DNA interaction interface, which forms hydrogen bonds within the major groove of DNA. To investigate whether the amino acid differences observed in the basic region of the BnFD proteins could impact DNA binding, predicted hydrogen bonding was analysed. Within the basic region of the BnFD proteins are two positions that exhibit between copy differences; positions 222 and 223 (Figure 2.36). To investigate the potential effects of these mutations on the DNA binding properties of BnFD, an available crystal structure of a bZIP transcription factor bound to DNA was used (PDB ID: 1DH3; Section 6.13; Methods)[319]. The crystal structure of the mammalian cAMP responsive element-binding protein (CREB) bZIP transcription factor bound to DNA revealed that the arginine in position 222 is important as the amino acid side

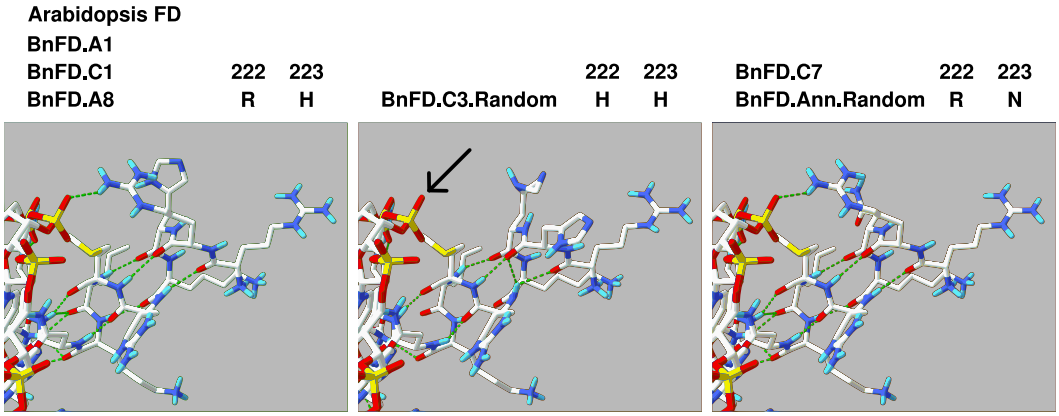


Figure 2.37: Protein structure of the BnFD proteins complexed with DNA reveal different hydrogren bonding.

The protein structure of the CREB protein (PDB ID: 1DH3) from Schumacher et al. (2000)[319] was changed to match the amino acids present in the basic region of BnFD proteins. The single letter codes of the amino acids replaced, and their positions in the amino acid alignment in Figure 2.36, are displayed above each plot. The green dashed lines indicate hydrogen bonding between atoms. The colour scheme for atoms is as follows: white (carbon), dark blue (nitrogen), yellow (phosphorus), red (oxygen), and light blue (hydrogen). Similar hydrogen bonding is observed between the Arabidopsis FD protein, BnFD.A1, BnFD.C1, BnFD.A8, BnFD.C7, and BnFD.Ann.Random. The BnFD.C3.Random protein is predicted to lose hydrogren bonding with the oxygen atom of the DNA backbone indicated with an arrow.

chain forms a hydrogen bond with the DNA backbone[319]. Mapping the amino acids in the basic region from the BnFD proteins onto the crystal structure of the CREB transcription factor revealed that changing the amino acid in position 222 from an arginine to a histidine disrupts hydrogen bond formation between the protein and the DNA (Figure 2.37). Whether a histidine or an asparagine is present in position 223 does not seem to affect the hydrogen bonding in the α-helix or between the protein and DNA (Figure 2.37). Therefore, the amino acid polymorphisms present in the basic region of BnFD proteins potentially affect the DNA binding affinity of the monomers, but only for the BnFD.C3.Random protein.

2.5.2.3 Amino acid differences in the leucine zipper region of BnFD proteins is predicted to alter dimerization affinity

Several amino acid differences between the BnFD proteins occur in the leucine zipper region (Figure 2.38a). To determine whether these differences have the potential to alter the dimerization affinity of the proteins, the amino acid polymorphisms were assessed in the context of the coiled-coil dimerization interface (Figure 2.39). Previous studies of bZIP transcription factors have revealed that amino acid residues in the `e` and `g` positions of the α-helix heptad are important in the determination of dimerization specificity[311,314]. Specifically, when the proteins form a coiled-coil structure, the side chain of an amino acid in the `e` position on one α-helix is able to form electrostatic bonds with the side chain of an amino acid in the `g` position on the other α-helix (Figure 2.39). This is illustrated in the helical wheel representations in Figure 2.39, that represent the positions of amino acids in the coiled-coil. An example of this is residue 250 (in the `g` position of the heptad) which has the capacity to form electrostatic interactions with residue 255 (in the `e` position of the heptad; Figure 2.38a) due to their opposing charges. Therefore, the charges these residues carry is a factor that determines the dimerization affinity between bZIP proteins. Positions 250, 255, 262, and 271 are all in either the `e` or `g` positions of the heptads and show amino acid polymorphisms that alter the charge of the amino acid side chains (Figure 2.38b). The effect this has on the predicted electrostatic interactions is illustrated in Figure 2.39. The BnFD.C1 homodimer and the BnFD.C1-BnFD.C7 heterodimer are both predicted to

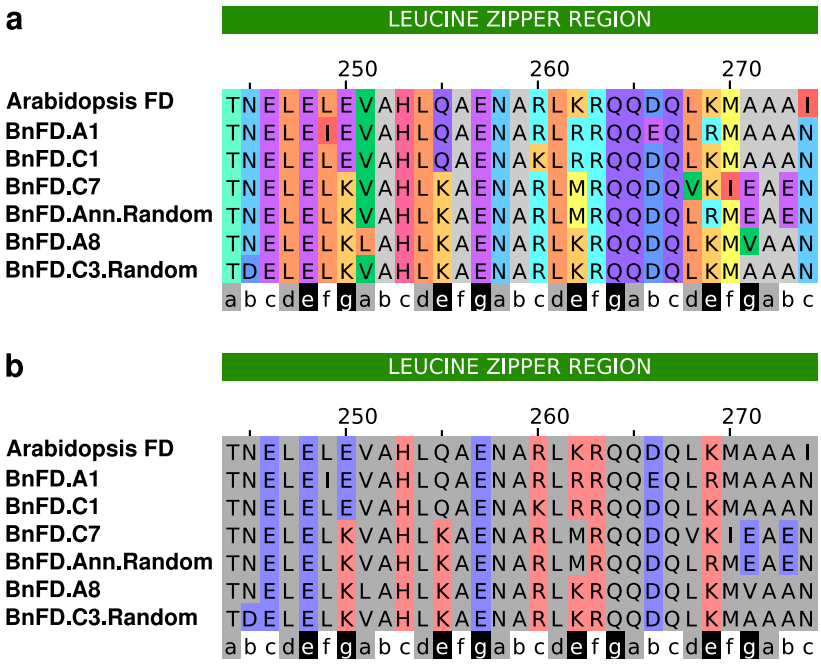


Figure 2.38: Amino acid differences in the leucine zipper region result in differently charged amino acids in the e and g heptad positions.
The amino acid sequence for the Arabidopsis FD protein and the six *B. napus* proteins are displayed. **a** Amino acids are coloured based on their residue type. **b** Amino acids are coloured based on their charge. Blue coloured amino acids have positively charged side chains while the red coloured amino acids have negatively charged side chains.

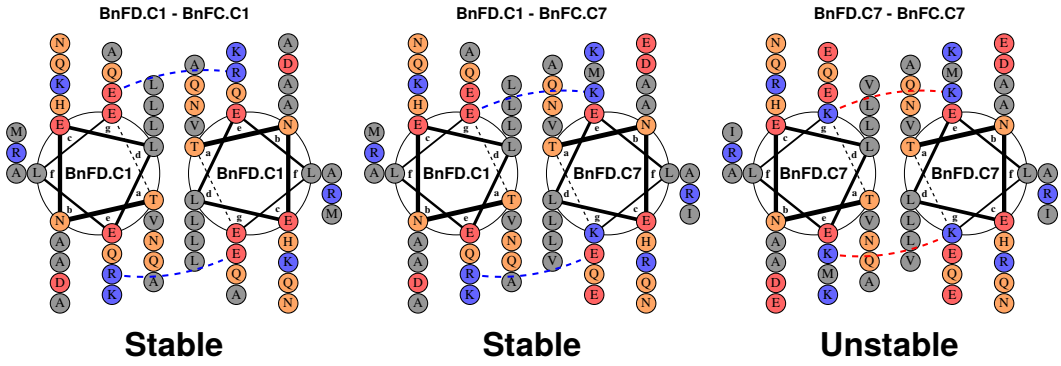


Figure 2.39: Helical wheel representation of the homodimers and heterodimer possible with the BnFD.C1 and BnFD.C7 proteins.

The coiled-coil structures of the leucine zippers are represented as helical wheels. Amino acids, denoted by single letter codes, in the seven positions of the α-helix heptad are displayed, with the columns of amino acids representing the amino acids the entire length of the coiled-coil. The blue coloured amino acids have positively charged side chains, the red coloured amino acids have negatively charged side chains, and the orange amino acids have polar side chains. The blue and red dotted lines between helical wheels indicate attractive and repulsive electrostatic charges between the two helicies respectively. The helical wheels demonstrate that attractive forces are predicted to form between the BnFD.C1 homodimer and the BnFD.C1-BnFD.C7 heterodimer, while a repulsive force is present in the BnFD.C7 homodimer. The helical wheels were drawn using DrawCoil[483] (version 1.0).

have attractive electrostatic interactions between the two monomers, while a repulsive force is predicted for the BnFD.C7 homodimer (Figure 2.39). These polymorphisms suggest that certain dimer combinations of the BnFD proteins will be more favoured than others.

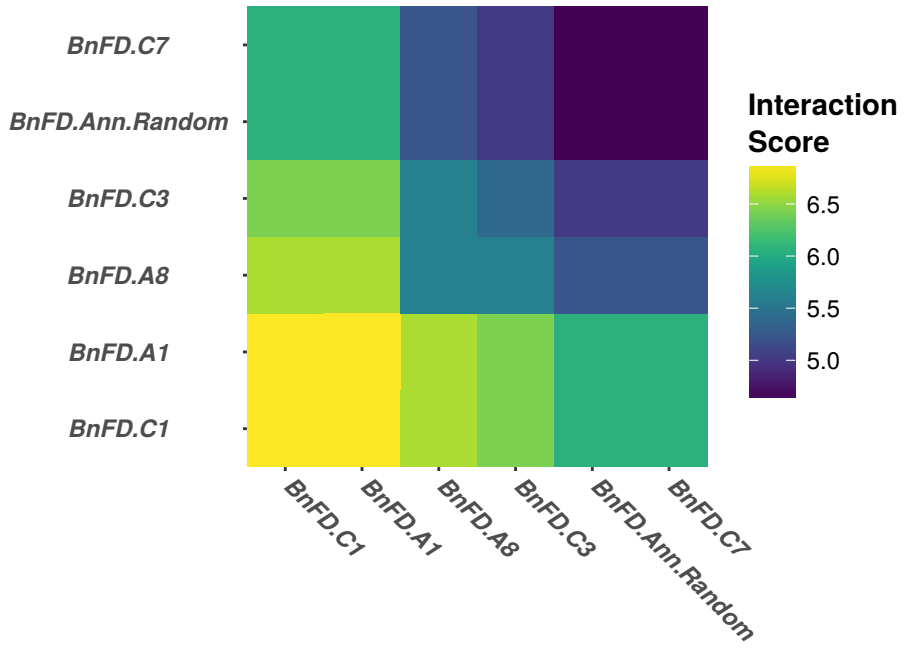


Figure 2.40: Heatmap of the dimerization affinity scores computed between BnFD leucine zipper regions.

The machine learning algorithm developed by Potapov et al. (2015)[320] was used to score the dimerization affinity of the leucine zipper regions of the possible BnFD dimers. Higher scoring dimers are more likely to form than lower scoring dimers. The leucine zipper regions used for the analysis correspond to the region indicated in Figure 2.38a. The heatmap reveals that certain BnFD dimers are predicted to be more likely to occur than others.

The sequence analysis suggests that the amino acid polymorphisms observed in the `e` and `g` positions of the heptad may affect the dimerization affinity of the proteins. To investigate this in a more quantitative manner, a published machine learning algorithm[320] was used to score the potential interaction affinity of pairs of BnFD monomers (Figure 2.40). The interaction scores between the BnFD monomers range from 4.3 to 7.2, with the higher interaction scores indicating a higher likelihood of interaction. To put these scores into

context, the dimerization of the bZIP transcription factors Fos and Jun have been extensively studied in terms of their dimerization affinity[314]. It has been shown that the Fos-Jun heterodimer is more thermally stable than either the Fos homodimer or the Jun homodimer, with the Fos homodimer being particularly unfavourable[314]. Using the machine learning scoring algorithm of Potapov et al. (2015)[320], Fos homodimers score 6.2, Jun homodimers score 6.3 and Fos-Jun heterodimers score 8.8. The score range for Fos and Jun dimers is 2.6, a similar range as that observed for the BnFD proteins. Therefore, the differences in interactions scores observed between the BnFD proteins are large enough to suggest a functional effect. The interaction scores group the six BnFD genes into three interaction groups (Figure 2.40). BnFD.C1 and BnFD.A1 form a group that have a higher affinity for forming dimers between themselves than with the remaining four proteins. BnFD.A8 and BnFD.C3.Random are more likely to form dimers with both BnFD.C1 and BnFD.A1 rather than themselves. Finally, BnFD.Ann.Random and BnFD.C7 have the lowest likelihood to form dimers between themselves relative to the other dimers tested, and have the highest likelihood to form dimers with both BnFD.C1 and BnFD.A1. The machine learning approach predicts that the six copies of BnFD have variation in their dimerization affinities, with four of the six copies predicted to form more stable heterodimers than homodimers. The range of interaction scores predicted for the BnFD proteins is similar in size to the range of interaction scores predicted for the Fos and Jun proteins, suggesting that the predicted differences have the potential to be biologically relevant.

2.5.2.4 Changes in dimerization affinities may be a common way of bZIP proteins diverging

To investigate whether polymorphisms influencing dimerization affinity were a common occurrence in organisms where gene multiplication events have occurred, sequences of *FD* orthologues identified in the EnsemblPlants database[321] were aligned. Only those species containing multiple Arabidopsis *FD* orthologues in the genome are displayed in Figure 2.41. Focussing on the leucine zipper regions of these proteins reveals similar charge influencing polymorphisms in the `e` and `g` heptad positions between the genes within a species.

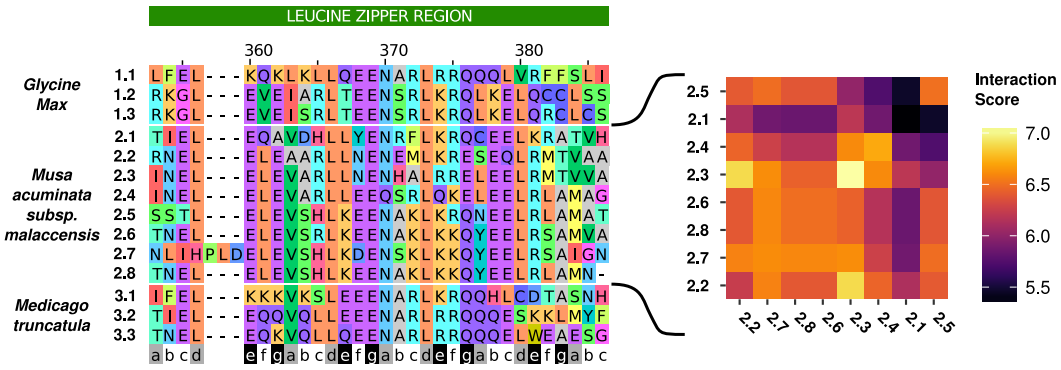


Figure 2.41: Multiple sequence alignment of the leucine zipper region of Arabidopsis *FD* orthologues in *Glycine max*, *Musa acuminata subsp. malaccensis*, and *Medicago truncatula*.

Amino acids are coloured based on their residue type. Several amino acid differences resulting in side chain charge differences are observed in the e and g heptad positions. The effect these changes have on the interaction scores calculated using the method of Potapov et al. (2015)[320] are displayed as a heatmap for the *M. acuminata* orthologues. The gene names are displayed in Table 6.4; Appendix A.

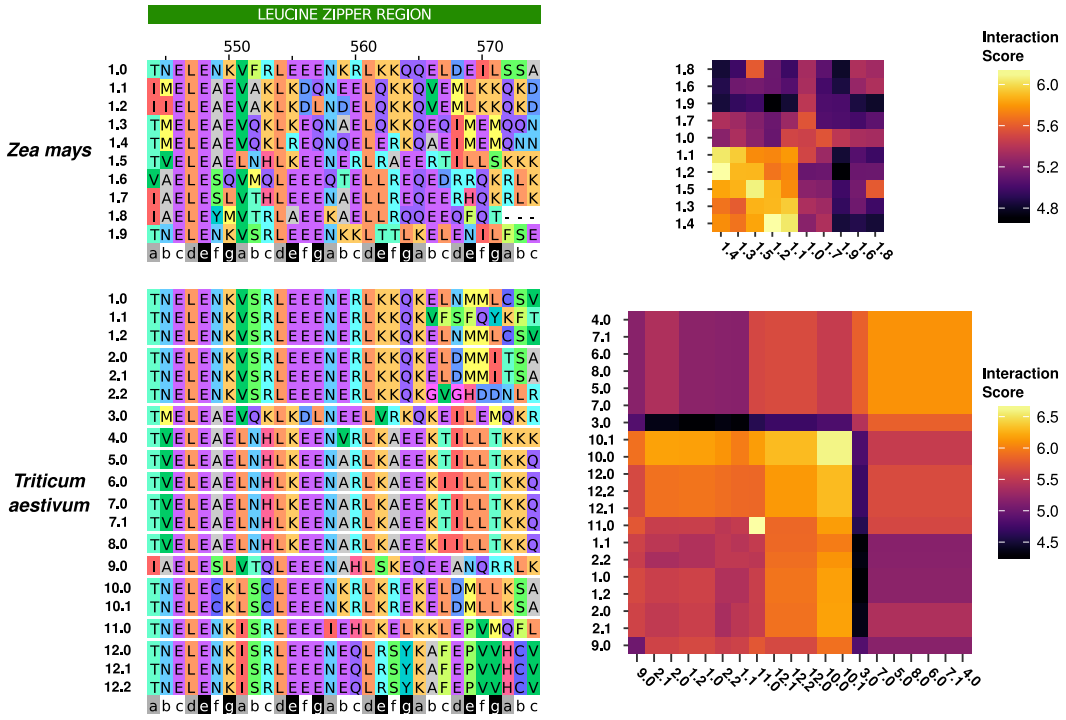


Figure 2.42: Multiple sequence alignment of the leucine zipper regions of the proteins with highest amino acid similarity to Arabidopsis FD from the *Zea mays* and *Triticum aestivum* proteomes.

Amino acids are coloured based on their residue type. Several amino acid differences, which result in side chain charge changes, are observed in the e and g heptad positions. The effect these changes have on the interaction scores calculated using the method of Potapov et al. (2015)[320] are displayed as heatmaps. The *Z. mays* proteins plotted were chosen by selecting the *Z. mays* protein with the highest sequence similarity to the Arabidopsis FD protein, as identified in the EnsemblPlants database[321]. In addition, the paralogues of each gene identified in this way were also included. The *T. aestivum* proteins were identified in the same way, except that in addition to the paralogues, the homoeologues of all proteins were also included. The gene names are displayed in Table 6.5; Appendix A.

Charge influencing polymorphisms in the e and g heptad positions are present in the *Glycine max* orthologues at positions 360, 362 and 381, *Musa acuminata* at positions 362, 367, 374, and 376 and *Medicago truncatula* at positions 360, 362, 367, and 381. Likewise, *Zea mays* and *Triticum aestivum* proteins with high sequence similarity to Arabidopsis *FD* also exhibit polymorphisms in the e and g heptad positions that alter the charge of the amino acid side chain. The machine learning algorithm[320] predicts considerable variation in the dimerization affinity for the identified FD-like proteins, with the range of scores being similar to the range identified for the BnFD proteins. These findings suggest that variation in dimerization affinities between duplicated bZIP proteins is frequently observed in different plant species.

2.5.2.5 Variation in dimerization affinity influences the proportions of hetero- and homodimers expected at steady state

To test potential regulatory repercussions of altered dimerization, a system of ordinary differential equations was used to model the dimerization reactions. Two different monomer types, a and b, were modelled, with the monomers able to form homodimers (aa and bb) and a heterodimer (ab). To investigate how the behaviour of the system depends on the dimerization affinities, three different reaction rates for the homodimerization of the b species were tested; 0.5, 4.0, and 7.0. For each of these rates, the heterodimerization rate for the monomers was varied and the steady state concentrations of the various species calculated. Equal concentrations of each monomer were used as the initial conditions of the model, and the system of equations was numerically solved until a steady state was reached (Figure 2.43; Section 6.14; Methods). When all dimerization rates are 7.0, the steady state concentrations of all dimers are identical (Figure 2.43c). For a b homodimerization rate of 7.0, the two homodimer species have the same steady state concentrations at all heterodimerization rates, as expected given that all dimerization reactions have the same reaction rates. By changing the b homodimerization rate to 0.5, the bb homodimer is disfavoured, with an observed increase in the steady state concentration of the undimerizaed b monomer (Figure 2.43a). This also affects with the steady state concentration of the heterodimer. Above a heterodimer formation rate of ~2.0, the heterodimer becomes more favourable than either of the homodimers.

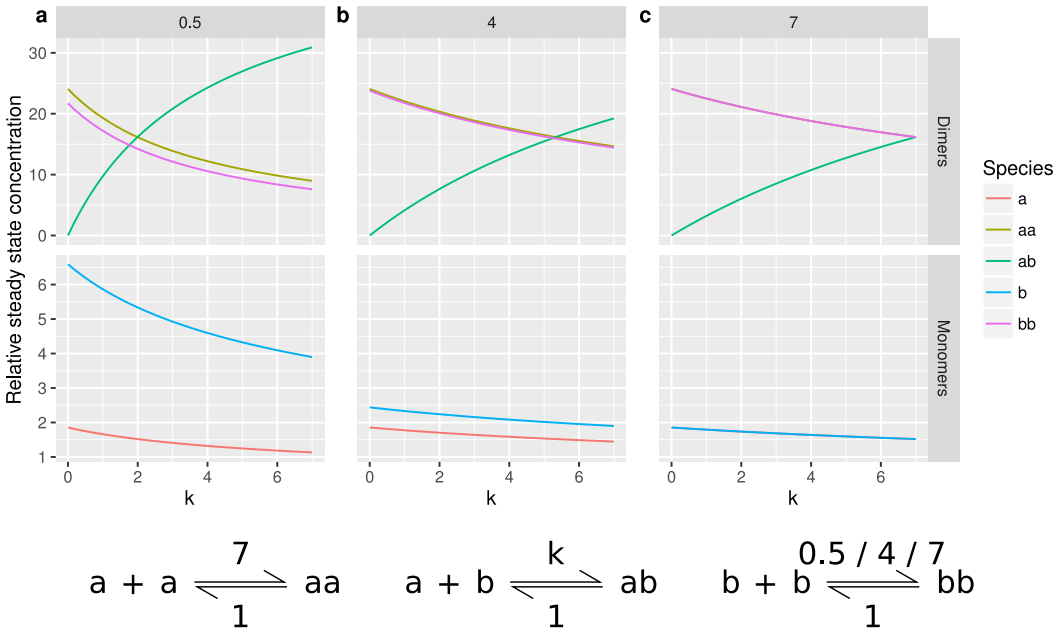


Figure 2.43: Dimerization affinity differences influence the dimer population expected at steady state.

The steady state concentrations of monomers and dimers are displayed. The simulation was run with different `bb` homodimer production rate, either 0.5 (**a**), 4.0 (**b**), or 7.0 (**c**), and was started with equal concentrations of each monomer. The equilibria simulated, with the rate constants used, are displayed below the plot. The x-axis corresponds to the `ab` heterodimer production rate. To generate these results the system of equations were modelled as ordinary differential equations and numerically solved. The concentrations plotted are steady state concentrations of the system. The simulations reveal that as the dimerization affinity of the `bb` dimer decreases, the relative concentrations of the `ab` heterodimer and `b` monomer at steady state increase. The simulations were run as described in Section 6.14; Methods.

The simulation results reveal that an unfavourable `bb` homodimer increases the `b` monomer concentration at steady state. A high relative concentration of the `b` monomer favours the formation of the `ab` heterodimer rather than the `aa` homodimer. This is despite the forward reaction rate of the `aa` homodimer being 2.5 times greater than the `ab` heterodimer formation rate. A similar pattern is observed when less extreme `bb` homodimer formation rate of 4.0 is used in the modelling (Figure 2.43b).

2.5.3 Conclusions

Analysing sequence divergence between *B. napus* homologues of two Arabidopsis floral integrators highlights the potential role both cis-regulatory elements and non-synonymous sequence differences play in gene divergence following duplication. The expression divergence observed between *BnTFL1* genes in the transcriptomic time series suggested that cis-regulatory element changes may have occurred. Comparing the downstream sequence of *BnTFL1* genes with Arabidopsis *TFL1* identified different patterns of sequence conservation for different homologues. These regions of differential sequence conservation were located in regions previously shown in Arabidopsis to contain cis-regulatory elements[309]. *TFL1* expression dynamics in Arabidopsis mutants lacking these cis-regulatory elements[309] were consistent with the expression of *BnTFL1* genes lacking sequence conservation within those elements. A section of sequence downstream of the gene, termed region II, was found to be important for *TFL1* upregulation in the meristem during the floral transition[309]. The *B. napus* *TFL1* homologues observed to increase in expression during the floral transition, *BnTFL1.A10* and *BnTFL1.C3*, exhibit sequence conservation to Arabidopsis *TFL1* within region II. However, *BnTFL1* genes that do not increase in expression during the floral transition, *BnTFL1.C2* and *BnTFL1.Cnn.Random*, do not display such sequence conservation in this region. This conservation suggests that the spatiotemporal domains of expression defined by the cis-regulatory elements is conserved between *B. napus* and Arabidopsis. Although the relationship between the sequence conservation downstream of the *BnTFL1* genes and the expression profiles exhibited by the genes is correlative, it provides a hypothesis to be tested by future studies. This case study is potentially

an example of cis-regulatory element changes driving the development of novel gene functions, as predicted by the DDC model[213] (section 2.1).

For the *BnFD* genes, expression profiles suggest that five of the six genes are co-regulated and potentially form dimers amongst themselves[49,310]. To investigate whether the different copies of the gene could potentially dimerize, the protein sequences of the genes were analysed. Amino acid differences were observed in multiple domains identified as conserved in FD-like proteins from diverse plant species[318]. An amino acid change in the SAP domain in the BnFD.C3.Random protein corresponds to an amino acid that is important for the interaction of the protein with FT[49], suggesting this copy may have altered protein binding. Amino acid differences identified in the DNA binding basic region, when compared to published crystal structures of bZIP transcription factors[319], suggest that the BnFD.C3.Random protein may also exhibit altered DNA binding. However, without characterising this experimentally, it is difficult to determine whether the single amino acid changes observed would have an appreciable effect on DNA binding. A potential improvement on the analysis presented here would be to perform more accurate predictions of hydrogen bond formation[322,323]. Between homologue amino acid differences in the leucine zipper region were predicted to alter the dimerization dynamics between BnFD proteins, with certain dimers predicted to be more likely to occur than others. Investigating *FD* orthologues in other species revealed that variation in dimerization affinity might be a common form of divergence for bZIP transcription factors that are present as multiple copies in the genome. Computational modelling of the dimerization dynamics suggest that having a system of monomers with different dimerization affinities can result in interesting regulatory consequences. However, this is dependent on the dimers formed having different molecular activities.

2.6 Discussion

Polyploidy plays a large factor in the success of both domesticated[324], and wild[325] plants. The gene duplication following whole genome duplication introduces a vast amount of genetic material. The relaxed selective pressures

allow for duplicated genes to acquire new roles, neofunctionalize, become more specialized, subfunctionalize, or be lost or silenced, the latter being the most common outcome for duplicated genes[212]. Despite this, a significant number of genes have been observed to be retained following gene duplication[229]. This has led to the gene dosage hypothesis being proposed, which states that dosage sensitive genes are preferentially retained in the genome following whole genome duplication to maintain the stoichiometry of protein complexes[224,227]. This has been observed in Arabidopsis, with signal transduction and transcription factors being preferentially retained[229] in the Arabidopsis genome following whole genome duplication[10,326,327].

To investigate the factors influencing gene retention in *B. napus*, particularly for the flowering time genes, a transcriptomic time series was developed for *B. napus*. The time series spanned from early growth to flower development, to allow transcriptomic changes during the floral transition to be followed. In order to confirm that the transcriptomic time series was able to capture biologically relevant effects, GO term and protein domain enrichment was performed. GO term analysis revealed transcriptional responses appropriate to the tissue. For example, genes associated with leaf senescence were upregulated in the leaf towards the end of the time series and genes associated with the regulation of flower development responding as expected in the apex (Figure 2.13). The response of the circadian rhythm genes to the vernalization period in both the leaf and apex revealed that the short day conditions of the cold treatment were influencing transcription (Figure 2.15). The sessile nature of plants means that they need to interpret environmental signals and alter their development accordingly. As such, the circadian clock in plants becomes entrained to different light regimes[16], and this effect is likely responsible for the response here. That genes associated with the circadian rhythm respond during the cold treatment needs to be taken into account when considering the expression profiles of other genes in the transcriptomic time series.

2.6.1 Gene retention

Genome dominance, that is, the finding that gene expression is biased towards gene copies from one genome, is a potential method by which gene expression

can diverge[292,293]. The results from the transcriptomic time series reveal that if all genes are considered, the A genome tends to have a higher proportion of genes that are highly expressed whereas the C genome has a higher proportion of lowly expressed genes. Interestingly, this pattern is not observed when pairs of homoeologues are considered, with a greater number of pairs exhibiting bias towards the C genome. This was found to occur independently of tissue, in contrast to previous results in *B. napus* that suggested genome dominance may be tissue and developmental stage specific[328]. While the results from the genome level and homoeologue level analysis may initially seem contradictory, observations in maize suggest that gene loss is biased towards the genome that has reduced homoeologue expression[294]. Therefore, gene loss may have occurred more frequently on the A genome, leading to the proportionally higher expression when all genes are considered. The potential effect of genome biased expression, however, is uncertain. In *Coffea arabica*, differential use of homoeologues was not found to contribute to the ability of plants to tolerate a broader range of growing temperatures than its diploid parents[329]. However, in *Gossypium hirsutum*, differential homoeologue expression was found to be tissue specific[292], suggesting the copies are functionally distinct. The age of the polyploid likely plays a significant role in this, with biased expression being observed more frequently in recent or synthesised allopolyploids rather than natural polyploids[288,289]. As the polyploidy event leading to *B. napus* occurred less than 10,000 years ago[107], relative to the 1 - 2 million years of cotton[289], or the 5 - 12 million years of maize[294], potentially the different homoeologue expression patterns observed between species are a consequence of *B. napus* being a relatively young polyploid.

Investigating the subset of flowering time genes in *B. napus* reveals that these genes seem to be preferentially retained relative to the entire genome. Similar patterns are also observed when just expressed genes are considered, suggesting that the additional copies of flowering time genes are expressed and functional. Given that the majority of flowering time genes are transcription factors, that are involved in highly networked gene regulatory networks[103,299], the gene dosage theory predicts their retention in the genome[224,227]. However, differences in the number of expressed versus annotated genes, the WGCNA-based clustering, and the SOM-based clustering, all suggest that expression

divergence between gene copies is common. For flowering time genes, 61% and 69% of Arabidopsis genes in the apex and leaf, respectively, have at least one *B. napus* gene that lacks expression in the time series. This potentially suggests that these genes are part of responsive backup circuits[219,220]. A prediction from this observation, assuming these genes are part of such backup circuits, is that these copies that are not expressed would be expressed if one of the other expressed copies became silenced[219,220]. A potential method of testing this would be to leverage the variation present among different *B. napus* varieties[330], to identify if homologue preference is observed. The alternative possibility is that the homologues exhibit tissue-specific expression[292], and that this is not captured in the leaf or apex transcriptomes. Determining the expression of these homologues in other tissues besides the apex and leaf would allow this to be tested.

To determine divergence among expressed genes, two clustering approaches were employed. The WGCNA-based clustering approach revealed, both genome wide and among flowering time genes, that expressed homologues have diverged in terms of their expression profiles across the transcriptomic time series. These results were supported by the SOM-based approach, which was used to ensure the observed divergence between flowering time genes was robust to the uncertainty inherent in gene expression data. This suggests that evolutionary mechanisms other than gene dosage have played a role in the retention of flowering time genes in the *B. napus* genome. This is consistent with observations in Arabidopsis that revealed 85% of paralogous regulatory genes exhibit expression that suggests neo- or subfunctionalization[331]. The divergence of expression patterns can be explained by both the DDC and the escape from adaptive conflict hypotheses[213,216]. The former hypothesis would predict that deleterious mutations have arisen in cis-regulatory elements, resulting in divergent expression patterns. The escape from adaptive conflict hypothesis posits that the different expression patterns of the genes may represent an adaptive partitioning of ancestral gene function. Either way, both of these potential hypotheses suggest that changes to cis-regulatory elements, or to other regulatory machinery, have occurred post-duplication. These findings for *B. napus* genes are consistent with findings that suggests regulatory divergence is one of the primary mechanisms by which genes diverge after duplication[332].

The flowering time genes were the focus of this work, and this was aided by curated lists of flowering time genes being available[299]. However, the patterns observed at the whole genome level suggest that homologue divergence is relatively common, with 69% and 62% of Arabidopsis genes in the apex and leaf respectively having at least one *B. napus* homologue located in a different regulatory module. An interesting avenue for future work would be to determine other subsets of genes that seem to be preferentially retained in the genome and determine whether similar expression divergence is observed. Good candidates for such groups of genes would be genes whose products are involved in signal transduction pathways, as these were found to be preferentially retained in Arabidopsis[229].

Although regulatory divergence between homologues is observed for the majority of Arabidopsis genes, many homologues do still exhibit similar expression profiles. The similarity of homoeologue expression patterns among flowering time genes revealed that many are found in the same regulatory module (79% in the apex and 77% in the leaf). Similarity in the expression of some homologues is also observed at the genome wide level, represented in the WGCNA-based analysis as groups of homologues occupying fewer regulatory clusters than the number of homologues present (Figure 2.22). At the individual gene level, the *BnLFY* genes all exhibit similar expression profiles, which is interesting given the dosage sensitivity of the *LFY* gene in Arabidopsis[71,72]. Homologues exhibiting similar expression profiles could represent genes where gene dosage based selection is maintaining them in the genome. However, due to the relatively young age of *B. napus* as a polyploid[107] it is also possible that these genes are redundant and selective pressures have not yet removed the duplicate copies from the genome, as theory would predict[212,215]. An important determinant of whether genetic redundancy is stable in the genome, and how long redundant genes are maintained in the genome if it is not stable, is the mutational rate of the duplicated genes[215]. Although mutational rates have been determined in other organisms[333,334], no such data is available for *B. napus*. If such data were available for *B. napus*, it would strengthen conclusions about whether seemingly redundant genes are in a transient state before being lost from the genome or whether redundancy is being selected for. An additional aspect of this would be the effect artificial selection and breeding has had on the

retention of duplicate genes. Mutational rates can be artificially altered to introduce variation into breeding genotypes[335], and potentially in this scenario of high mutational rates selection for genes that are redundant is favoured.

2.6.2 Floral transition

The floral transition is one of the most important developmental transitions an angiosperm can go through. Floral integrators form a tightly interconnected gene regulatory network that ensures the timing of the floral transition is consistent[299]. Indeed, the structure of this network confers favourable behaviours such as noise filtering of input signals and irreversability[41]. To determine whether this network is conserved in *B. napus*, the expression profiles of the key floral integrator genes were investigated. The tissue specificity of expression, and the expression profiles themselves, were generally consistent with the expression of the genes in Arabidopsis. At least one *B. napus* homologue displayed an expression profile consistent with that expected from Arabidopsis. This suggests a general conservation between the regulatory network underlying flowering in Arabidopsis and *B. napus*, that will aid efforts to translate knowledge from Arabidopsis to the crop.

Of the floral integrators, *BnFT* is the most well studied, potentially because of its proximity to regions of the genome found to be associated with flowering time[135,304]. All four profiles are upregulated after the cold period, as expected from Arabidopsis[20–22]. The A7 and C6 copies were found to exhibit divergence relative to the A2 and C2 copies. In the leaf these copies exhibit a greater fold difference between pre-cold expression levels and post-cold peak expression levels. This is interesting given results from vernalization sensitive lines of *B. napus* that found the A7 and C6 copies were silenced prior to vernalization, whereas the A2 copy was expressed prior to vernalization[154]. Although Westar is a spring variety, a slight vernalization response is still observed and *BnFLC* genes in the variety display expression consistent with being vernalization sensitive[241] (section 3.3.1). This potentially suggests that these copies mediate are vernalization responsive in Westar. However, this response may be variety specific, as other findings from Guo et al. (2014)[153] found *BnFT.A7* and *BnFT.C6* to be most highly expressed when floral buds were visible, which does

not agree with results from the transcriptomic time series. It is also interesting that *BnFT.C2.Random* is found to exhibit expression, given that multiple accounts have reported that the C2 copy of *BnFT* is not expressed[153,154]. This could represent a difference between spring and winter varieties of *B. napus*. The expression of *BnFT.A7* and *BnFT.C6* in the apex is somewhat surprising, given that *FT* in Arabidopsis is not required for the function of the gene to promote flowering[22,45,49]. Although *FT* homologues are expressed in the apex in cabbage (*B. oleracea*)[144], and seem to be involved with the floral transition in the plant, the morphological differences between cabbage and oilseed rape make the findings difficult to compare. Finally, the expression profiles of *BnFT* copies in the leaf suggest that the experimental design decision to subject the spring variety to the same vernalization treatment of the winter variety likely aided in synchronizing the development of the two varieties. The high expression of *BnFT* genes prior to the cold suggests that the Westar plants were capable of flowering prior to the cold treatment, as would be expected of a spring variety. The short day photoperiod of the vernalization treatment seemingly repressed *FT* expression until after the cold, delaying the flowering of the spring variety[302].

For the other floral integrators, less is known about their expression in *B. napus*. However, the expression profiles of all *BnLFY* genes, and the most highly expressed *BnAP1* genes, are consistent with the roles the homologous genes have in Arabidopsis[74,80]. The *BnSOC1* genes exhibit spatial divergence, with *BnSOC1.A3.Random* and *BnSOC1.A5* most highly expressed in the apex and *BnSOC1.A4* and *BnSOC1.C4.Random* most highly expressed in the leaf. This suggests that these copies may have undergone spatial subfunctionalization[206,213]. Further divergence is observed between the *BnSOC1* expression profiles, with some copies responding to vernalization, while others do not. This suggests that different *BnSOC1* genes have diverged to respond to different inputs. In Arabidopsis, *SOC1* is downstream of the FT protein and becomes upregulated in the apex when *FT* is expressed[20,48,84,307,308]. The only copy consistent with this regulatory interaction in the apex is *BnSOC1.A3.Random*, making it a good candidate for maintaining the same role as *SOC1* in Arabidopsis. For the copies of the key floral integrators that exhibit expression consistent with their Arabidopsis counterparts, a similar

approach to the one taken by Tadege et al. (2001)[145] could be employed, with the best candidates transformed into Arabidopsis mutants to determine whether they indeed retain their role or not.

Despite the similarities between the regulation of Arabidopsis and *B. napus* floral genes, divergence is observed between *B. napus* homologues of floral genes. In Arabidopsis, duplicated regulatory networks have been observed to diverge, such that parallel networks that are spatiotemporally distinct are formed[229]. If this was the case with the gene regulatory network underlying flowering in *B. napus*, divergence would be expected for all floral integrators. However, the analysis here reveals that *B. napus* homologues of floral integrators instead exhibit different patterns of regulatory module assignment. At one extreme, *BnLFY* genes seem not to have diverged relative to each other in terms of expression profile, while at the other extreme all four copies of *BnTFL1* exhibit different expression profiles. This suggests that the gene regulatory network underlying flowering has not diverged to form parallel networks in *B. napus*. This is potentially due to differences in the evolutionary time that has elapsed since gene duplication. The gene duplication analysed in the Arabidopsis genome, that lead to the observed formation of parallel networks, occurred 20 - 60 million years ago[212,229,336]. However, the genome triplication event that formed the ancestral hexaploid *Brassica* ancestor occurred 8 - 23 million years ago[112,113], while the interspecies hybridization event to give *B. napus* occurred less than 10,000 years ago[107]. Therefore, potentially not enough evolutionary time has elapsed for this form of divergence to be observed.

BnTFL1 was the only floral integrator where all copies exhibited expression profiles that are completely divergent. To investigate potential explanations for this, the regulatory regions surrounding the gene were investigated. For each *BnTFL1* gene, different patterns of sequence conservation were observed downstream of the gene, with the differences correlating with the expression divergence observed between the genes. This is in agreement with previous investigations of *BnTFL1* genes, that found between copy sequence conservation both within the first intron and downstream of the gene[152]. Serrano-Mislata et al. (2016) identified and characterised cis-regulatory elements downstream of the *TFL1* gene in Arabidopsis[309], that colocalized with the regions displaying sequence conservation differences between *BnTFL1* genes. The expression of

TFL1 in Arabidopsis mutants lacking certain cis-regulatory regions[309] were strikingly similar to the expression profiles of *BnTFL1* in the transcriptomic time series. For example, region II (Figure 2.34) was identified as important for the upregulation for *TFL1* in the Arabidopsis apex during the floral transition[309]. *BnTFL1* genes lacking sequence conservation in that region were not upregulated during flowering, whereas *BnTFL1.C3* and *BnTFL1.A10*, which did exhibit conservation in region II, were upregulated. Although correlative, these findings certainly provide hypotheses for future investigations. The *BnTFL1.Cnn.Random* copy does not exhibit conservation in region III, identified as responsible for expression of *TFL1* in Arabidopsis lateral meristems[309]. Determining whether this copy is indeed lacking expression in the lateral meristem would be one way of testing whether the cis-regulatory elements downstream of *TFL1* genes are conserved between *B. napus* and Arabidopsis. The observed divergence between *BnTFL1* genes is interesting given results from pea (*Pisum sativum*). Three homologues of *TFL1* were identified in the pea genome, which through mutant and expression experiments were determined to have separate functions[337]. One of the homologues was involved with maintaining floral indeterminacy, while the other two genes seemed to regulate flowering time[337]. As the *TFL1* gene is involved with both of these functions in Arabidopsis[51,52], this suggests subfunctionalization has occurred among *TFL1* homologues in pea. Potentially a similar type of functional partitioning is observed among the *BnTFL1* genes. In order to dissect the roles these four copies play in the plant, detailed analysis of their expression domains within the apical structure, combined with the same analysis for *BnAP1* and *BnLFY* genes, would be required. This is due to the mutual antagonism between *TFL1*, *AP1*, and *LFY* and the small zones of the apex in which they are expressed[52–56]. In addition, analyses of *B. napus* plants with null mutations in each of the *BnTFL1* copies will help to determine whether the C3 or the A10 copy of *BnTFL1* has greatest functional similarity to *TFL1* in Arabidopsis, as both show expression patterns that are consistent with the observed regulation of *TFL1*. Transgenic investigations of Arabidopsis could be used to test such hypotheses, such as transforming *tfl1* null mutant Arabidopsis lines with the *BnTFL1* genes. If these insertions also included the downstream intergenic regions, the functional conservation of the cis-regulatory elements could be established.

Due to the co-regulation observed among *BnFD* genes, and the *FD* protein being a bZIP transcription factor that binds to DNA as a dimer[49], the possibility of the different BnFD proteins dimerizing was explored. Sequence variation between the copies was found to alter the predicted amino acid sequence within the dimerization interface, the leucine zipper. These amino acid differences resulted in positively charged amino acid side chains being present in some BnFD proteins, and negatively charged amino acids in others. A published machine learning algorithm was used to assess the probability of dimerization between BnFD monomers[320], which identified that not all possible BnFD dimers were equally likely. For example, the BnFD.A1 and BnFD.C1 homo- and heterodimers are likely to form, while the BnFD.C7 and BnFD.Ann.Random homo- and heterodimers are not. Taken together this suggests that the BnFD proteins have diverged in terms of the dimers that they are able to form. Computational modelling revealed that alterations to dimerization affinities have the potential to affect the proportions of dimers expected to form at steady state, potentially representing a novel method of *FD* target regulation in *B. napus* relative to Arabidopsis. Indeed, a number of examples illustrate that transcription factor dimerization is able to act to regulate gene expression. In mouse, it was found that the helix-loop-helix (HLH) protein Id formed protein-protein interactions with three other HLH proteins (MyoD, E12, and E47) and that the heterodimers involving Id were compromised in their ability to bind to the DNA recognition sequences[338]. In flower development, the ABCE model proposes that the composition of the protein tetramers directs the formation of different floral structures[280]. *BROTHER OF FT AND TFL1* (*BFT*) produces a protein that competes with FT for binding to FD, and this competition mediates the delay in flowering that salt stress induces[339]. Therefore, the *BnFD* proteins have diverged and, in doing so, have potentially expanded the range of signals they are capable of responding to.

A caveat to this analysis would be that the spatial expression domains in which the *BnFD* proteins are expressed may be too small to be resolved by the sampling method used. Therefore, although five of the six *BnFD* genes are assigned to the same regulatory module (Figure 2.31), they may not be expressed in the same cells. If this is the case, then the dimerization dynamics and the potential regulatory consequences of them would not be applicable. To

test whether the different *BnFD* proteins interact *in vivo*, enrichment techniques and proteomics could be used to elucidate the *in vivo* interaction partners of particular proteins. Another potential caveat is that although analysis of other regions of BnFD amino acid sequences identified potentially functional changes[318] (Figure 2.37), it is not known whether different BnFD have differing target sequence preferences or protein-protein interactions. If this was the case, then the hypothesized regulatory effects of different dimerization affinities between BnFD proteins would not be applicable. One way of testing this would be to use transgenic *FD* genes where the two FD protein monomers are forced to dimerize through a linker peptide. Alternatively, a similar approach to that taken to investigate the alternative binding of SVP and FLC homo- and heterodimers could be employed[99]. An aspect of this analysis that is not well understood is whether FD in Arabidopsis binds to DNA as a homodimer with itself or as a heterodimer with another bZIP monomer. If so, the hypothesised dimerization changes observed between BnFD proteins may instead represent divergence for other bZIP proteins, with complementary changes occurring in the interaction partners.

Taken together, these results highlight that both gene dosage and regulatory divergence have played a role in gene retention in *B. napus*. One form of regulatory divergence, that is observed among the *BnSOC1* genes, is a potential divergence in terms of the environmental inputs the genes respond to. The next chapter will introduce data from a winter, vernalization requiring variety of *B. napus*. The effects of a cold requirement on the transcriptome will be assessed, and evidence for the divergence of floral integrators between a spring and winter variety presented.

Chapter 3

Effects of a requirement for cold on regulatory divergence

3.1 Introduction

Being sessile organisms, plants have to time and regulate their development based on seasonal and environmental cues. One of the seasonal cues that plants are capable of responding to is the prolonged cold of winter[27]. In the model species Arabidopsis, accessions are either summer or winter annuals[24,25]. Summer annuals germinate and set seed in the same year by germinating in the spring, flowering in the summer, and setting seed before the winter months. Conversely, winter plants germinate in the autumn, stay vegetative over the winter months, then flower and set seed the following spring or summer. Without experiencing an extended period of cold, winter annual plants may not flower or flowering may be severely delayed. Delaying the floral transition until an extended period of cold is experienced is a vernalization response; an evolutionary adaptation to the climate where the plants are growing[26]. One of the central genes in the vernalization response is *FLC*, a MADS-box containing transcription factor[29]. However, in addition to having a functional *FLC* allele, Arabidopsis plants also require a functional allele of *FRI* to exhibit a vernalization response. Comparing an early flowering accession and a late flowering, vernalization-responsive accession of Arabidopsis revealed an active *FRI* allele as being required for the latter phenotype[340]. However, when

this active *FRI* allele was crossed into another Arabidopsis accession that does not require vernalization, the inheritance of the late flowering phenotype indicated that a locus in addition to *FRI* was required for the late flowering phenotype[341–343]. Through additional studies it was determined that a winter annual life strategy was largely conferred through active alleles of both *FRI* and *FLC*. Sequence polymorphisms in the first intron of *FLC* conferred a summer annual growth habit on some Arabidopsis accessions[344], while different *FLC* alleles were found to alter the length of vernalization required to accelerate flowering[245]. A Swedish variety of Arabidopsis, Lov-1, was found to require a longer period of vernalization to fully repress *FLC* expression relative to other accessions[345]. The *FLC* allele from the Lov-1 accession has a higher optimum vernalization temperature than other tested accessions, and this is proposed to be an adaptation to the snowfall experienced by the plants in their natural region of growth in northern Sweden[346]. Although *FLC* is important for the vernalization response, sequence variation at *FRI* was responsible for ~70% of flowering time variation in a collection of natural Arabidopsis accessions[28]. This result highlights the importance of both genes for conferring a winter growth habit in Arabidopsis.

FLC is a floral inhibitor[29] controlled by both the autonomous and vernalization flowering time pathways, that binds to and represses the expression of *FT*[30] in addition to other floral integrators[31]. The autonomous pathway increases the expression of *FLC* while the vernalization pathway represses expression of the gene[347–350]. The expression of the gene was found to decrease during vernalization in a quantitative manner, with the more cold the plant experienced, the less the gene was expressed[29]. The repression of *FLC* during the cold is mediated by a host of different mechanisms that result in epigenetic silencing of the locus. A long non-coding RNA expressed from the antisense strand at the *FLC* locus is one of the first processes that occur during vernalization[351]. Recruitment of Polycomb repressive complex 2 (PRC2) follows. PRC2 mediates changes to the methylation state of histones, leading to a change in the chromatin structure at the *FLC* locus, repressing its expression[352–355]. The recruitment of PRC2 to the *FLC* locus during cold is proposed to involve the product of the *VERNALIZATION INSENSITIVE 3* (*VIN3*) gene[356]. *VIN3* is a plant homoeodomain-finger (PHD) protein upregulated during exposure to

cold[357]. PHD-finger containing proteins mediate histone interactions[358], and it is thought that the VIN3 protein directs the PRC2 complex to the *FLC* locus to induce epigenetic silencing of the gene. These epigenetic changes are stable across mitotic divisions, allowing the perception of the cold to impact development months after the environmental signal has been perceived[359]. The response of *FLC* at the level of the locus is digital in nature (the locus is either active or repressed in individual cells), despite showing a quantitative response to cold at the cell population level[360,361].

A vernalization requirement is a key agronomic trait of *B. napus*, with spring varieties constituting the majority of oilseed rape growth in Canada, Australia, and Northern Europe and winter varieties being grown in Europe and Asia[127]. Understanding the requirement for cold is therefore a key part of any analysis of flowering time control in the crop. Characterisation of *Brassica* homologues of genes in the vernalization pathway suggest conservation of the pathway in these crops[120,144,147]. Four copies of *FLC* are present in the *B. rapa* genome[137], four copies in the *B. oleracea* genome[138], and nine copies in *B. napus*[141]. For *FRI*, two copies have been identified in both *B. rapa* and *B. oleracea*[147,362] and four copies in *B. napus*[142]. Divergence of *Brassica FLC* (*BnFLC*) and *FRI* (*BnFRI*) homologues have been revealed in a number of different studies and different *Brassica* crops. One way in which the genetics of the floral response have been dissected in *Brassica* crops is through the use of association studies. These studies find correlations between genetic variation and phenotypic variation to try and identify regions of the genome that underlie the phenotypic difference. Mapping populations are generated by breeding two lines together that exhibit phenotypic differences. For example, a Doubled Haploid (DH) mapping population generated by crossing Ningyou7, a Chinese semi-winter *B. napus* variety with a slight vernalization response, and Tapidor, the winter variety used in this study, identified genomic regions associated with flowering time that contained *FLC* homologues on chromosomes A10 and A3[141,143]. Interestingly, the region on A10 was only associated with unvernalized flowering time as opposed to vernalized flowering time, leading the authors to suggest this locus is one of the determinants of whether a *B. napus* variety is a spring or a winter variety[141]. The *FLC* copy on A2 has also been linked to the vernalization response in *B. napus*[134]. Using a mapping population

derived from two spring varieties of *B. napus* found regions containing *FLC* on chromosomes on A3 and C2 associated with flowering time, suggesting the effect of certain *FLC* copies on flowering time is variety dependent[141,363,364]. Functional divergence of *B. napus FLC* homologues has been suggested using transgenic studies. Different copies of *BnFLC* were found to delay flowering to different extents when expressed in Arabidopsis, indicating conservation in function between the species but divergence in the efficacy of the homologues at repressing the floral transition[145]. In *B. rapa*, as in *B. napus*[134], an *FLC* copy on chromosome A2 emerged as a candidate underlying flowering time variation[132–135,365]. In addition, Schranz et al. (2002) found that *FLC* copies on A10, A2, and A3 in *B. rapa* influence flowering time[137]. The C2 copy of *B. oleracea* seems to influence flowering time to a greater extent than the other copies in the species. A nucleotide difference at the C2 copy of *FLC* in *B. oleracea* reduced the sensitivity of the gene to the environment, resulting in later heading date[140]. Variation in this same homologue was found to account for the majority of flowering time variation in cauliflower[139], and was identified as associated with vernalization response in another population[138]. Divergence at the protein structure level has been found between *FRI* homologues in *B. oleracea*[147]. That associations to flowering time variation differ between vernalization pathway gene homologues in *Brassica* crops suggests that the copies have diverged, with this being confirmed molecularly in some cases. However, the roles of copies that do not seem to influence flowering, or influence flowering to a lesser extent, remain elusive.

A potential avenue of subfunctionalization, a partitioning of the roles of an ancestral gene, is spatial subfunctionalization[206,213]. For example, an ancestral gene may be expressed in both leaves and roots when present as a single copy. Following a gene duplication event, however, evolutionary forces may lead to the presence of leaf-specific and root-specific gene homologues. This form of subfunctionalization is an expectation from the duplication-degeneration-complementation model[213]. This model for gene evolution posits that after gene duplication, mutations disrupting cis-regulatory elements, which control the spatiotemporal expression of genes, are likely to be neutral. This is because gene copies without the mutation would complement the copy with the mutation. If different cis-regulatory elements are required for gene expression in different

tissues, then over time mutations in these cis-regulatory elements would result in tissue-specific expression of the genes. This method of subfunctionalization is of particular interest in the context of vernalization, as evidence from a range of sources has found that vernalization acts at both the shoot apex and at the leaves. Localized cooling experiments in celery (*Apium graveolens*) found that the shoot apex was the site at which vernalization acted in the plant[366]. Similar cooling and grafting experiments also identified the apex as the organ at which vernalization was sensed in *Thlaspi arvense*, a member of the Brassicaceae family like Arabidopsis and the *Brassica* species[367]. However, the authors noted that other tissues, such as the leaves, were still capable of responding to vernalization[367]. In another Brassicaceae family plant, *Lunaria biennis*, plants regenerated from a cutting of vernalized leaves were competent to flower without experiencing cold, indicating that vernalization had occurred in the leaves[368]. Further work indicated that, as opposed to particular tissues, mitotically dividing cells were required for vernalization to take place[369,370]. Results from other species have reinforced that vernalization can be sensed in a range of tissues, such as flower buds[371] and roots[360,372], with the general consensus being that the location at which vernalization is sensed in a plant is likely to be species specific[373]. One of the most thorough assessments of the role of *FLC* at both the apex and the leaves was performed by Searle et al. (2006). By expressing *FLC* in a tissue-specific manner, the authors were able to deduce that *FLC* has a dual role in Arabidopsis[31]. Not only does *FLC* induce floral signals in the leaf, through the derepression of *FT*, the product of the gene also acts on floral integrators in the apex, making the regulatory network competent to respond to the signal coming from the leaf[31]. Assessing divergence of *FLC* copies in *B. napus* is of particular interest given the "flowering rheostat" model of *FLC*[373]. This model is based on observations that the delay of flowering mediated by *FLC* is dosage dependent[341,347]. Low or no expression of *FLC* results in a summer annual growth habit, whereas additional copies of *FLC* expressed in Arabidopsis resulted in the plants exhibiting a biennial life strategy[373]. A key question, then, is whether the additional *FLC* copies have been maintained in the *B. napus* genome to maintain gene balance (discussed in chapter 2), or have they diverged to have different expression domains and different effects on the floral transition?

In this chapter, I will discuss the expression of floral integrators and key vernalization pathway genes in the *B. napus* winter variety Tapidor. The expression of these genes will be compared to the expression of the same genes in Westar in order to address three lines of investigation. By interpreting global differences between the spring and winter variety, the effect of a requirement for cold on the overall transcriptional landscape is assessed. This revealed that a requirement for cold has a global effect on the entire transcriptome, delaying expression responses relative to the spring variety. In addition, the apical transcriptome is determined more by the developmental stage of the plant, whereas the leaf transcriptome seems to be more a consequence of plant age. The second line of investigation concerns understanding the divergence of vernalization pathway genes. Specifically, which vernalization pathway genes are candidates for mediating the vernalization requirement in Tapidor and is there tissue specificity in expression. Finally, by determining which floral integrator genes are expressed differently in Tapidor relative to Westar, particular copies of floral integrators are assessed for their vernalization sensitivity. This reveals *FT* and *TFL1* homologues to be most differently expressed, although *B. napus* homologues of other floral integrators also exhibit different patterns of expression in the winter compared to the spring variety. This provides some evidence that certain floral integrators have diverged to become biased towards, or more sensitive to, particular inputs.

The comparisons made in this chapter were made between a winter and a spring variety of *B. napus*. However, the results should not be considered as general differences between spring and winter varieties of *B. napus*, as the observed differences may instead be due to variety-specific divergence. This limitation is a consequence of only one winter variety and one spring variety being compared in the study. Despite this, the results still lead to hypotheses which can be tested in a larger panel of *B. napus*, and in some cases are consistent with the current literature.

3.2 A requirement for cold affects global expression patterns

The effect of cold periods on plant transcriptomes has been investigated in lilies[374,375], barley[376], radish[377], and *Brachypodium distachyon*[378]. These studies generally compare gene expression before, during, and immediately after vernalization to identify cold and vernalization responsive genes[374,376–378], although others focus solely on gene expression during vernalization[375]. These studies were designed to identify vernalization responsive genes, and therefore lacked longer term effects of the cold requirement on the transcriptome. For example, these studies are not able to assess whether a vernalization requirement delays development in a global fashion, or whether it delays the floral transition in a more specific manner. Equally, no attempt was made by these studies to try and assess whether the effect of vernalization on the transcriptome is tissue specific. The study by Paina et al. (2014) was an improvement in this regard. Using ryegrass, an experimental design very similar to the one used to generate the transcriptome time series described in this thesis was employed[264]. Leaf tissue was collected once before vernalization, three times during, and twice post-vernalization, and apex tissue was sampled at the end of vernalization and twice post-vernalization. Tissue was collected from both a vernalization insensitive and a vernalization requiring line. The ryegrass vernalization response was found to have links to the photoperiod pathway and carbohydrate metabolism[264]. However, the final time point in the series was sampled only seven days after vernalization in both varieties sampled, limiting the ability of the study to assess how development was delayed. In addition, the relatively few time points for the apex samples restricted the scope of the study when assessing the extent of tissue specificity[264].

In order to assess the global impact of a cold requirement on the *B. napus* transcriptome, the transcriptomic time series (Section 2.2) was used. Comparisons between Westar and Tapidor reveal that Tapidor has an expanded set of genes expressed in a variety-specific manner in the leaf, potentially representing an expanded sensory capability in the winter variety. Clustering results find that a requirement for cold delays developmental transcriptome responses in a global manner, although photoperiod responses seem unchanged. Finally,

correlation analysis between time points and between varieties suggests that while the apex transcriptome is largely defined by the developmental stage of the plant, the leaf transcriptome is instead influenced more by the age of the plant. Therefore, although the leaf seems to have an expanded set of expressed genes in the winter variety, the apex transcriptome is more responsive to the vernalization signal than the leaf transcriptome.

3.2.1 Variety-specific expression is biased towards Tapidor in the leaf

As flowering in Tapidor is dependent on experiencing a period of cold, the plant potentially has an increased ability to sense its environment relative to a spring variety. This expansion in sensory ability could be mediated through the expression of additional genes in the winter variety relative to the spring variety. To investigate this, the overlap between the expressed *B. napus* genes in each variety was calculated. In the leaf, 4% to 6% of genes exhibit spring-specific expression, whereas 8% to 9% show winter-specific expression (Figure 3.1). The bias towards Tapidor increases when floral genes are considered; there are 43% more Tapidor-specific genes than Westar-specific when all *B. napus* genes are considered (Figure 3.1a), 53% when only *B. napus* homologues of Arabidopsis genes are considered (Figure 3.1b), and 88% when *B. napus* floral genes are considered (Figure 3.1c). This bias was not observed to the same extent in the apex, where all *B. napus* genes and *B. napus* genes with identified Arabidopsis homologues only showed 2% and 3%, respectively, more Tapidor-specific genes relative to Westar, with 12% more among floral genes. There therefore seems to be a consistent bias, across all gene subsets considered, towards Tapidor having more variety-specific genes expressed in the leaf.

The bias towards Tapidor-specific expression observed from the overlaps of expressed genes (Figure 3.1) does not take into account homologue relationships. For example, within a set of *B. napus* genes homologous to the same Arabidopsis gene, variety-specific expression of one homologue towards one variety and another homologue towards the other variety would result in the same number of homologues being expressed in each variety. This phenomenon will be described as compensatory expression of homologues. In order to investigate

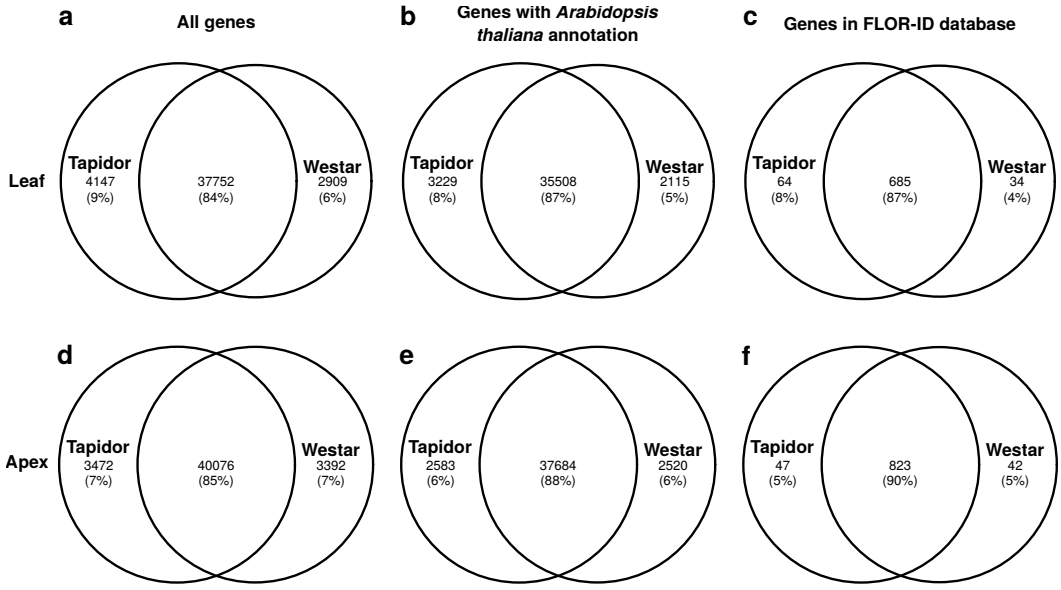


Figure 3.1: Overlap between varieties in the sets of expressed genes. *B. napus* genes were regarded as expressed if their maximal expression level across the transcriptomic time series was greater than, or equal to, 2 FPKM. The overlaps in the leaf reveal a greater number of variety-specific expression in Tapidor, with 43 - 88% more genes than Westar. This is the case regardless of the gene subset taken. This finding is not as evident in the apex. The gene subsets used to calculate the overlaps in each case are: **a** and **d** All *B. napus* genes; **b** and **e** *B. napus* genes with identifiable Arabidopsis homologues; **c** and **f** *B. napus* genes that show sequence similarity to Arabidopsis genes in the FLOR-ID database of floral genes[299]. Percentages have been rounded to the closest integer.

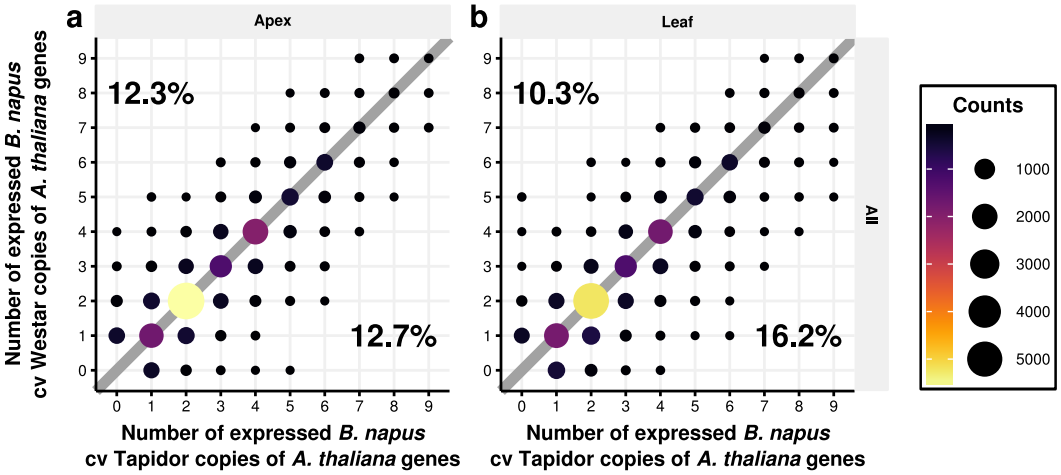


Figure 3.2: Relationship between the number of expressed copies of Arabidopsis genes in Tapidor relative to Westar.

The number of expressed copies of an Arabidopsis gene in *B. napus* was determined as the number of homologues that had a maximal expression value above or equal to 2 FPKM at at least one time point in the time series. The size and colour of the circles indicate the number of data points at that position. Points on the diagonal, grey line represent Arabidopsis genes that have equal numbers of homologues expressed in both Tapidor and Westar. The left most percentage within each graph represent the number of Arabidopsis genes that have more homologues expressed in Westar, whereas the right most percentage is the corresponding percentage for Tapidor. In both the apex (**a**) and the leaf (**b**) there are more Arabidopsis genes with more copies expressed in Tapidor relative to Westar. Using a chi-squared goodness-of-fit test (using the `chisq.test` function in the R statistical programming language[467]), reveals that the bias towards Tapidor is not significant in the apex (p-value of 0.359) but is significant in the leaf (p-value of $<$ 2.2e-16), assuming a 0.05 significance threshold, with the null hypothesis assuming equal numbers of points on each side of the diagonal, grey line.

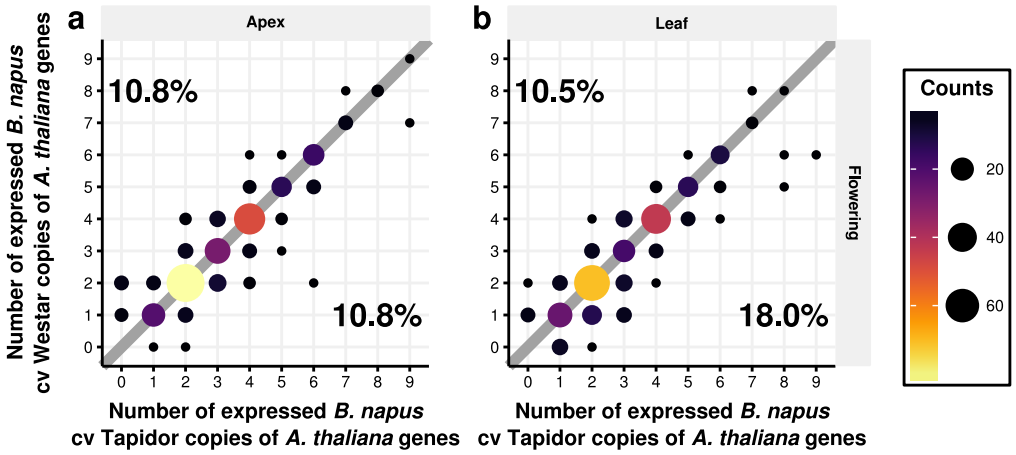


Figure 3.3: Relationship between the number of expressed copies of Arabidopsis floral genes in Tapidor relative to Westar.

The number of expressed copies of an Arabidopsis floral gene in *B. napus* was determined as the number of homologues that had a maximal expression value above or equal to 2 FPKM at at least one time point in the time series. The size and colour of the circles indicate the number of data points at that position. Points on the diagonal, grey line represent Arabidopsis floral genes that have equal numbers of homologues expressed in both Tapidor and Westar. The left most percentage within each graph represent the number of Arabidopsis genes that have more homologues expressed in Westar, whereas the right most percentage is the corresponding percentage for Tapidor. In the apex (**a**) there are equal numbers of Arabidopsis floral genes on both sides of the diagonal, whereas in the leaf (**b**) there are more Arabidopsis genes with more copies expressed in Tapidor relative to Westar. The observed difference in the leaf (**b**) is significant (p-value of 0.026), based on the same statistical test described for Figure 3.2.

whether this form of compensation takes place, the number of Tapidor expressed and the number of Westar expressed copies of each Arabidopsis gene were compared (Figure 3.2). In the apex, 12.3% of Arabidopsis genes have more copies expressed in Westar relative to Tapidor, while 12.7% show the converse relationship (Figure 3.2a). However, the percentages calculated using expression data from the leaf (Figure 3.2b) reveal a higher percentage of Arabidopsis genes have a greater number of homologues expressed in Tapidor (16.2%) relative to Westar (10.3%). Assuming the null hypothesis of no bias towards either variety, the observed difference is significant (Figure 3.2). Within the range of 0 to 9 expressed *B. napus* homologues, the maximal difference in the number of expressed homologues between varieties is 5 (Figure 3.2). Percentages of Arabidopsis genes exhibiting different numbers of expressed homologues in each variety are higher than the percentages of *B. napus* genes exhibiting variety-specific expression (Figure 3.1). For example, 10.3% of Arabidopsis genes have more homologues expressed in the leaf in Westar relative to Tapidor (Figure 3.2b), whereas 5% of *B. napus* genes are expressed specifically in Westar (Figure 3.1b). Given that the mapping of Arabidopsis genes to *B. napus* is one-to-many, this suggests that *B. napus* genes exhibiting variety-specific expression are generally well distributed among different Arabidopsis genes.

To test if the retention of flowering time genes would affect the observation of Arabidopsis genes tending to have more expressed homologues in Tapidor leaf tissue, this was tested using a subset of flowering time genes. In the apex (Figure 3.3a) a higher percentage of Arabidopsis genes have the same number of homologues expressed in both varieties (78.4%) relative to the global percentage (75.0%). This suggests that the functions of multiple copies of flowering time genes may tend to be more conserved between varieties than the rest of the genes in the genome, although further validation would be required. An alternative explanation is that compensatory expression of homologues occurs more frequently among floral genes. The observed bias towards Arabidopsis genes having more expressed genes in Tapidor is slightly exaggerated when floral genes are considered separately (Figure 3.3b). These findings reveal that flowering time genes exhibit less variety-specific expressed homologue counts in the apex, yet the bias towards Arabidopsis genes having more expressed homologues in the winter variety is slightly exaggerated.

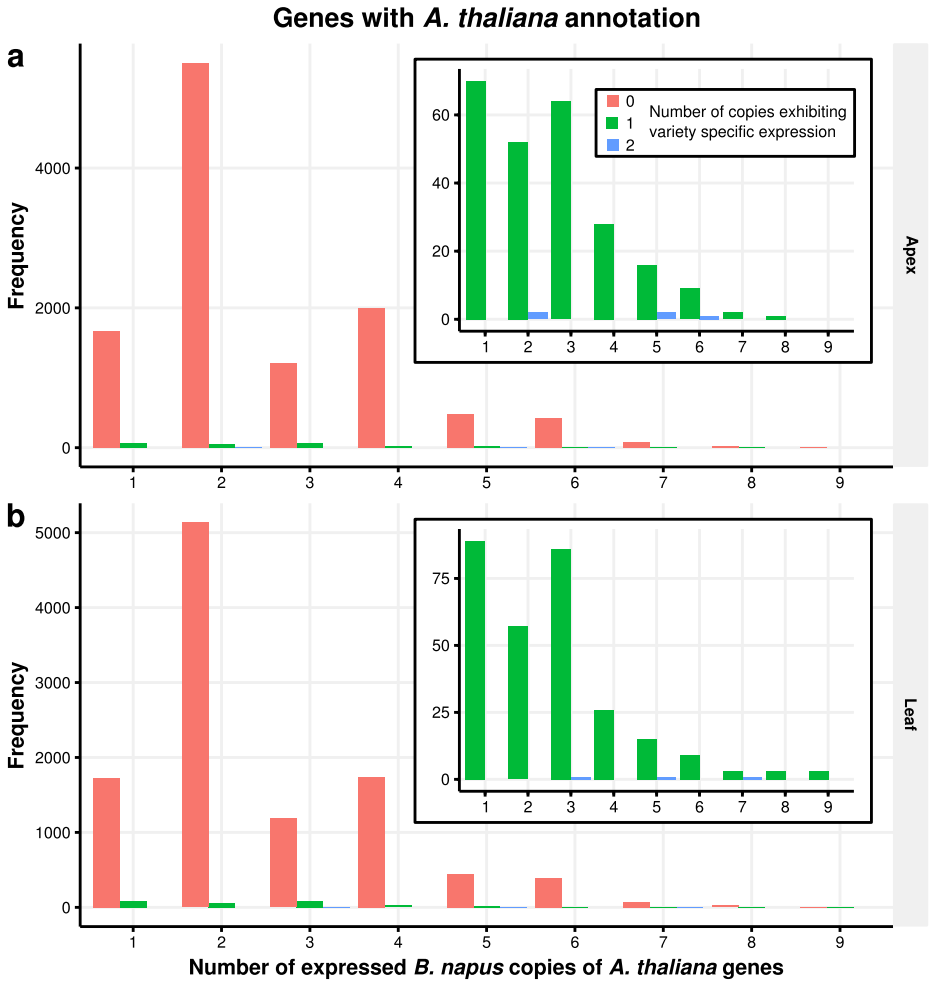


Figure 3.4: Extent of compensatory homologue expression.
Only Arabidopsis genes that have the same number of homologues expressed in both Tapidor and Westar (points that lie on the diagonal grey line in Figure 3.2) are considered. These are separated by those that have 0, 1, or 2 homologues that exhibit compensatory expression behaviour. The inset displays the same data as the main figure, but without the bars corresponding to Arabidopsis genes with zero homologues that exhibit compensatory behaviour. Very few instances of compensation are observed between homologues in both the apex, **a**, and the leaf, **b**.

The occurrence of compensatory expression between homologues could represent a form of varietal differentiation. The extent of compensatory expression was assessed among Arabidopsis genes that have the same number of copies expressed in both *B. napus* varieties (75.0% for the apex, 73.5% for the leaf; diagonal grey lines in Figure 3.2). For the vast majority of cases (98% in the apex, 97% in the leaf) the same complement of gene copies were expressed in both varieties (Figure 3.4). The maximal number of copies showing compensatory variety-specific expression is two, which represents instances where six copies of the gene are expressed across both varieties, four in each. However, the instances of this are low.

Similar patterns are observed with the floral genes, with 98% of genes in both tissues having the same complement of gene copies expressed in both varieties (Figure 3.5). These results indicate that the occurrence of compensatory homologue expression is comparatively rare, with floral genes having little effect on this pattern.

Taken together these results illustrate that variety-specific expression of *B. napus* genes occurs, although the majority of genes do not exhibit it. In Tapidor, there are more *B. napus* genes expressed in a variety-specific manner in the leaf relative to the apex (Figure 3.1), with the differences between varieties increasing when a subset of floral genes are taken. At the Arabidopsis gene level, approximately a quarter of Arabidopsis genes exhibit differences in the number of *B. napus* homologues expressed in each variety (Figure 3.2). Once again, the bias towards Tapidor-specific expression in the leaf is maintained (Figure 3.2). This tissue dependent bias towards Tapidor having a greater number of expressed homologues in the leaf raises the possibility that the additional copies are required for processes occurring in the leaf in Tapidor that are not occurring in Westar. In addition, *B. napus* homologues of the same Arabidopsis gene compensating for each other between varieties is a relatively rare occurrence (Figure 3.4). This suggests that this potential form of varietal divergence does not play a role in phenotypic differences between the varieties, or if it does play a role, that it is the effect of relatively few genes.

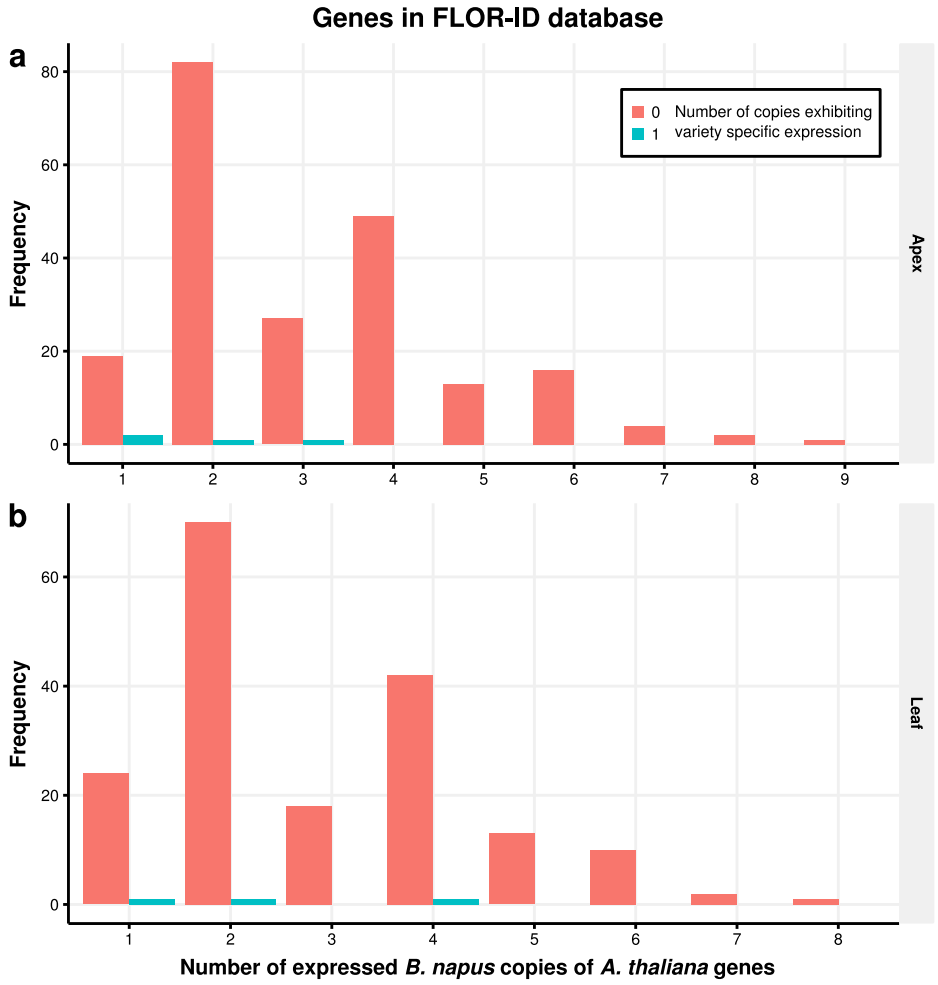


Figure 3.5: Extent of compensatory homologue expression among floral genes. Only Arabidopsis flowering time genes that have the same number of homologues expressed in both Tapidor and Westar (points that lie on the diagonal grey line in Figure 3.3) are considered. These are separated by those that have 0 or 1 homologues that exhibit compensatory expression behaviour. The inset displays the same data as the main figure, but without the bars corresponding to Arabidopsis flowering time genes with zero homologues that exhibit compensatory behaviour. Very few instances of compensation are observed between homologues in both the apex, **a**, and the leaf, **b**.

3.2.2 Self-organizing maps reveal that a cold requirement delays developmental transcriptional programs

To understand whether a vernalization requirement has large scale effects on gene expression, self-organizing maps (SOMs) were used to cluster gene expression profiles across time. This was done to determine if the vernalization response acts through relatively few genes that have a large effect on flowering, or by affecting gene expression on a global scale. SOMs were employed to allow broad comparisons in regulatory patterns to be made between the two varieties. The SOM clusters to which most genes mapped in Tapidor showed remarkable similarity to the SOM clusters containing most genes in the spring variety (section 2.2.1). In the apex (Figure 3.6) clusters 88 and 98 both exhibit increased expression during the vernalization period, with expression returning to pre-cold levels after the treatment. This expression trace closely follows that of cluster 19 from the Westar apex SOM (Figure 2.11). Likewise, cluster 46 from both the Tapidor and Westar apex SOMs (Figure 3.6 and 2.11) exhibit relatively constant expression during the entire time series, with expression increasing significantly between the penultimate and final time points. However, although a similar pattern is observed, the final time point in the Tapidor time series (83 days of growth) does not occur at the same time as the final time point of the Westar time series (72 days of growth). Therefore, the upregulation of genes in Tapidor cluster 46 is delayed relative to Westar. The most highly enriched GO terms for this cluster relate to carpel, gynoecium, and floral whorl development, which is consistent with the vernalization response delaying flowering in the winter variety. Clusters 88 and 98 are both enriched for the GO term "circadian rhythm". That the expression of these clusters is very similar to clusters in Westar suggests that the vernalization requirement does not influence the expression of genes associated with the circadian rhythm or the photoperiod flowering pathway.

Similarities to Westar were also observed in the SOM generated using the leaf transcriptomes from Tapidor, with two clusters having many genes mapped to them (Figure 3.7). Cluster 25 exhibits an increase in expression during the vernalization treatment (Figure 3.7), similarly to cluster 99 in the Westar

leaf SOM (Figure 2.12). Both clusters are enriched for GO terms linked to translation and protein biosynthesis, suggesting that the response to cold in the leaf requires the synthesis of novel cellular components. The other cluster with a large number of genes mapped to it in the Tapidor leaf SOM is cluster 59, which exhibits a slight increase in expression post-cold and a large increase at the final time point (Figure 3.7). This is a similar expression trace to that exhibited by cluster 19 in the Westar leaf SOM (Figure 2.12). The GO terms enriched in these two clusters relate to responding to cell stress, ageing and cell death. As with the apex, therefore, it seems that a requirement for cold delays the expression of genes that are expressed later in development but does not affect genes expressed as a result of the cold treatment.

In order to compare transcriptional responses between tissues, comparisons between the apex and leaf SOMs were made. By comparing expression differences between the tissues in both varieties, it allows for differences that are biologically relevant, and not the result of biological noise, to be highlighted. Of the clusters to which most genes are mapped in all SOMs generated, there is consistently a cluster with an expression pattern that increases during the vernalization treatment, with expression returning to pre-cold levels after the treatment. However, tissue-specific subtleties exist between the expression traces for these clusters. In the apex, the peak expression value during the cold is observed at the day 43 time point in both Westar (cluster 19; Figure 2.11) and Tapidor (cluster 88; Figure 3.6), with expression decreasing slightly at the day 64 time point before returning to pre-cold levels after the treatment. However, the response in the leaf is more gradual, with expression increasing during the cold treatment and peaking at the day 64 time point in both Westar (cluster 99; Figure 2.12) and Tapidor (cluster 25; Figure 3.7). A potential explanation is the difference in mitotic activity between the two tissues[373]. A mitotically active tissue, such as the apex, potentially responds to environmental stimuli more quickly than tissues where cell division is not as prolific, such as the leaf. The mitotic activity of a tissue has been proposed to influence the ability of that tissue to become vernalized[369,370]. The slower response to the cold treatment in the leaf may therefore be due to the lack of cell division inhibiting the rate at which vernalization directed transcriptional changes occur. The GO terms also suggest differences in the genes mapped to these clusters, with

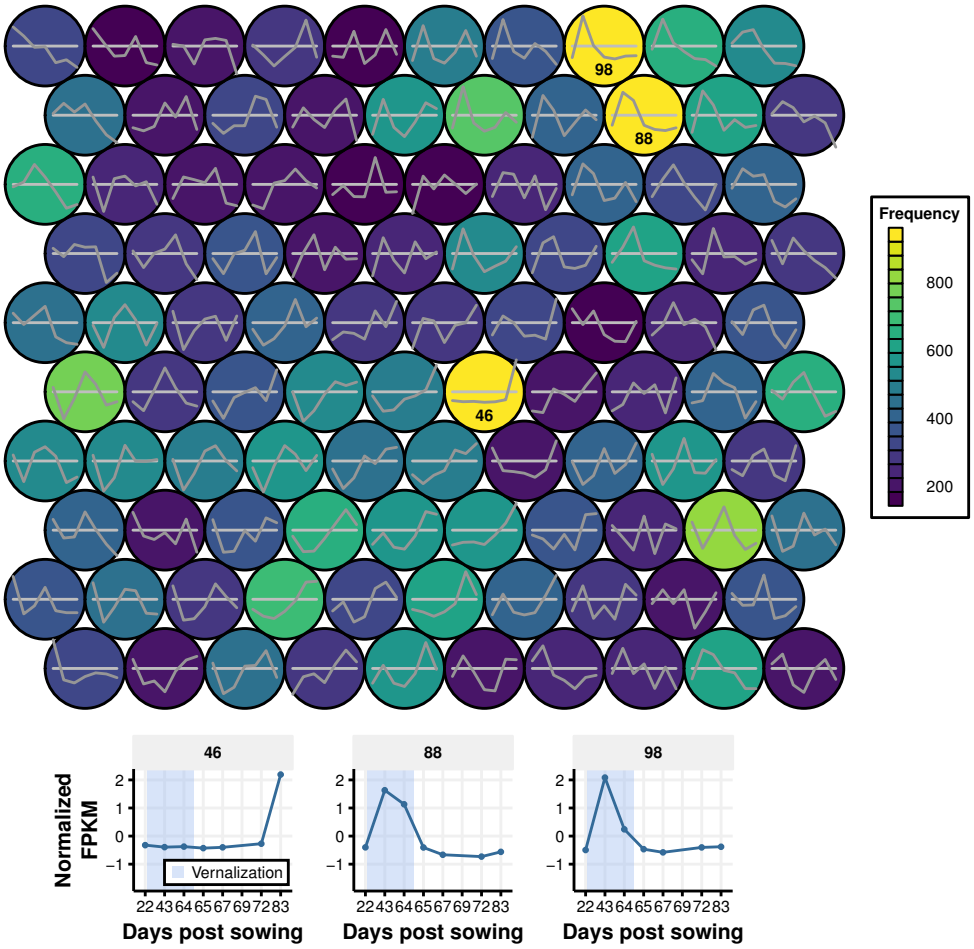


Figure 3.6: Self-organizing map of the apex transcriptome in Tapidor. Gene expression patterns were normalized to zero mean, unit variance across time and clustered. Nodes (coloured circles) are situated adjacent to nodes with a similar expression pattern. The nodes on the edges of the map are adjacent to the nodes on the opposing side of the map, such that the map, when viewed in three dimensions, would form a toroid. The colour of the circle indicates the number of genes mapped to that particular node. The three clusters with the most genes mapped to them are 46, exhibiting an increase in expression at the final time point, and 88 and 98, with the expression pattern of both clusters increasing during the vernalization treatment.

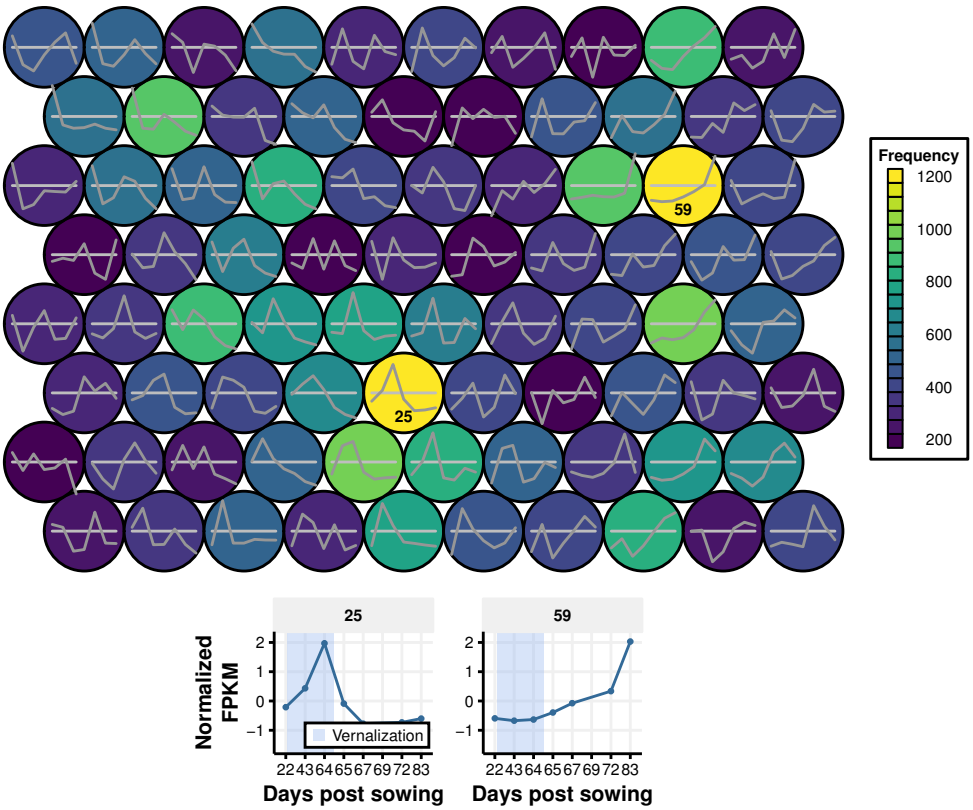


Figure 3.7: Self-organizing map of the leaf transcriptome in Tapidor. Gene expression patterns were normalized to zero mean, unit variance across time and clustered. Nodes (coloured circles) are situated adjacent to nodes with a similar expression pattern. The nodes on the edges of the map are adjacent to the nodes on the opposing side of the map, such that the map, when viewed in three dimensions, would form a toroid. The colour of the circle indicates the number of genes mapped to that particular node. The two clusters with the most genes mapped to them are 59, exhibiting an increase in expression after the vernalization treatment, and 25, the expression pattern of which increases during the vernalization treatment.

the apex clusters enriched for circadian rhythm genes and the leaf clusters enriched for genes associated with translation. It may, therefore, be that these clusters actually represent different ensembles of genes, with the transcriptional program in the apex responding to photoperiod changes and expression in the leaf responding to a requirement for novel cellular components.

3.2.3 Correlation analysis suggests apex and leaf transcriptomes behave differently during plant development

The SOM analysis revealed that a vernalization requirement delays the upregulation of genes associated with flower development in the apex, which is expected. However, it also seems to delay the upregulation of stress, cell death, and age related genes in the leaf, suggesting that a vernalization requirement delays development more generally than just delaying the floral transition. To investigate how the timing of transcriptomic changes compare between the two varieties, Pearson correlation coefficients were calculated between time points. Pearson correlation coefficients are calculated by determining how linear the relationship is between the FPKM values from one sample and the FPKM values from another sample. A coefficient of 1 indicates a positive, linear relationship between the gene FPKM values between samples, whereas a coefficient of 0 indicates that a linear relationship is not present. The coefficients were calculated both within and across varieties for each tissue; the within variety comparisons allow for the timing of transcriptional changes to be determined while the across variety comparisons allow for differences in these timings, if they exist, to be assessed.

The first observation that stands out is the baseline similarity in expression values between samples. The lowest correlation coefficient observed is 0.4, which is found between the day 43 and day 83 samples within the Tapidor leaf (Figure 3.9). That there is this basal level of correlation between the samples suggests that many genes are regulated similarly in both varieties. Calculating correlation values between tissues results in coefficients that are much lower, with means of 0.35 (Westar) and 0.31 (Tapidor), suggesting that the basal level

of correlation observed between varieties is a consequence of tissue-specific gene expression.

An expectation of a correlation analysis such as this is that time points within a variety would tend to be most similar to temporally proximal time points, with similarity decreasing as time passes. This is based on the assumption that global transcriptional changes take time to orchestrate. Such behaviour is observed between samples from the same variety, with the patterns being observed most clearly post-cold, from the day 65 time point onwards (Figures 3.8 and 3.9). For example, in both tissues from Tapidor the day 22 time point is most highly correlated with the day 65 time point, with the size of the coefficient decreasing as time progresses. In addition, adjacent time points post-cold are generally highly correlated (Figures 3.8 and 3.9), suggesting that the transcriptional time series captures dynamic changes in expression. This pattern is not as clear in Westar however, with all three time points sampled immediately after cold in the apex (day 65, 67, and 69; Figure 3.8a) and the two post-cold time points in the leaf (day 65 and day 67; Figure 3.9a) being highly correlated. This indicates that large scale changes in transcription were only observed between the day 69 and day 72 time points in both tissues in the Westar samples. This is in contrast to Tapidor, where transcriptional changes occurred more slowly post-cold (Figures 3.8c and 3.9c). The cold treatment results in a transcriptome distinct from the other time points. In both varieties, and in both tissues, the day 43 time point (half way through the vernalization treatment) has the highest correlation with the other time point taken during cold; the day 64 time point sampled the day before plants were removed from cold. This is also exemplified by the day 22 time point exhibiting highest correlation with the day 65 time point in both varieties and tissues; the first time point sampled after the plants were removed from the cold treatment. This reveals both that the cold treatment has a large effect on the transcriptome, and that the transcriptome, at a global level, responds quickly to removal from cold by returning to a largely similar state as pre-cold.

The most striking result from this analysis is in the comparisons between varieties for both tissues (Figures 3.8b and 3.9b). In the leaf, the highest correlation coefficients are between samples taken at the same time point (Figure 3.9b). The exception to this is the day 83 time point from Tapidor,

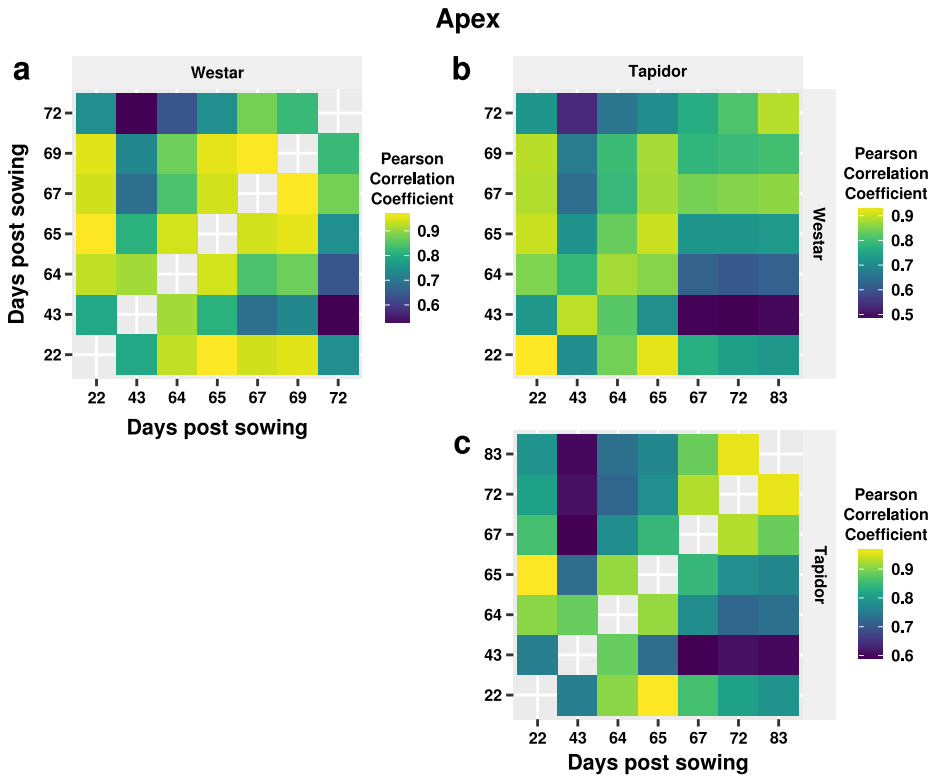


Figure 3.8: Pearson correlation coefficients between apex samples. Coefficients were calculated between the transcriptomes of all apex samples, with Westar-Westar (**a**), Westar-Tapidor (**b**), and Tapidor-Tapidor (**c**) comparisons scaled individually. Coefficients between like samples (diagonal lines in **a** and **c**) have been removed for clarity. The higher the coefficient value, the more similar two samples are. It should be noted that although there are seven time points for Westar and Tapidor, the final two time points in Westar (69 and 72 days) are different to the final two time points in Tapidor (72 and 83 days)

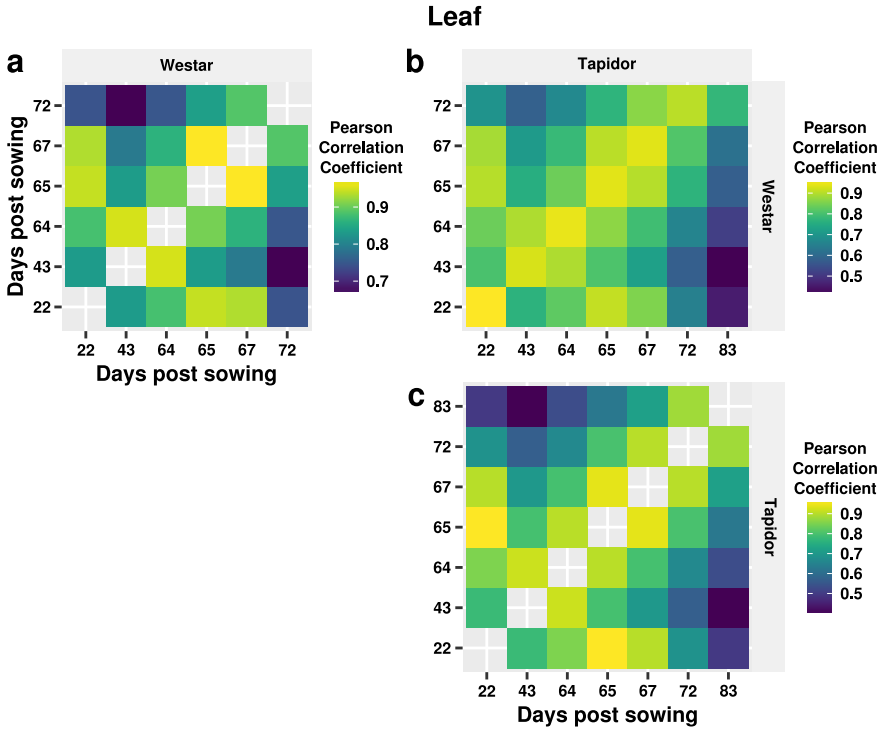


Figure 3.9: Pearson correlation coefficients between leaf samples. Coefficients were calculated between the transcriptomes of all leaf samples, with Westar-Westar (**a**), Westar-Tapidor (**b**), and Tapidor-Tapidor (**c**) comparisons scaled individually. Coefficients between like samples (diagonal lines in **a** and **c**) have been removed for clarity. The higher the coefficient value, the more similar two samples are. The additional time point in Tapidor results in the rectangular Westar-Tapidor comparison heatmap.

as there is no corresponding sample taken for Westar. This trend, however, does not apply to the entire time series in the apex samples. The highest correlation coefficients for the Tapidor samples at day 22, day 43, and day 64 are the Westar samples from the corresponding time points (Figure 3.8b). The day 65 time point in Tapidor is most correlated with the day 22 time point in Westar, although the day 65 time point has the second highest coefficient. This is likely due to the confounding effects of day 22 and day 65 time points being highly correlated within variety. After the day 65 time point, however, the most highly correlated sample does not correspond to the samples taken on the same day. The day 67 and day 72 samples from Tapidor are most highly correlated with the day 67 time point in Westar. The two final time points are also most highly correlated, despite the Tapidor sample being sampled 83 days post-sowing and the Westar sample 72 days post-sowing (Figure 3.8b). Taken together these two results suggest that different factors are influencing the transcriptome in each tissue. The equivalent time points being most highly correlated in the leaf suggests that the age of the leaf is having the largest effect on the transcriptome. That there is a time delay between the most highly correlated samples in the apex suggests that age does not influence the transcriptome in the apex as strongly as the leaf. Instead, the pattern of correlation coefficients suggests that developmental stage influences the transcriptome in the apex. This is seen most clearly at the final time point, which was sampled such that the two varieties were at a similar developmental stage (BBCH stage 51[246]).

3.2.4 Conclusions

To investigate whether a cold requirement impacts the *B. napus* transcriptome at the global level, or as a more focussed effect, the transcriptomes from both Tapidor and Westar across the time series were compared. Analysis of variety-specific expression of *B. napus* genes, and of variety-specific numbers of expressed homologues for Arabidopsis genes, reveals that there are more Tapidor-specific *B. napus* genes than Westar-specific in the leaf. The leaf is the plant organ at which photoperiod is interpreted[17,18,20–22] and also plays a role in sensing the vernalization response[29,31,368]. An expanded set of genes expressed

exclusively in Tapidor could represent increased sensory machinery in the leaf in the winter variety relative to the spring variety, in line with the increased vernalization sensitivity of Tapidor. This is especially interesting given that *FLC* has been found to influence the activity of the circadian clock[379]. In addition, it is interesting that the percentage of *B. napus* genes expressed in both varieties (Figure 3.1) is larger than the percentage of genes expressed in both tissues in Westar (Figure 2.18). This reveals that the occurrence of variety-specific expression is lower than tissue-specific expression within a variety, suggesting that the tissue dissection was successful at enriching for apex tissue.

The results from the SOM clustering reveal that there is delayed upregulation of genes associated with flower development in the apex of Tapidor, the variety with a vernalization requirement. This is fully expected given the role the vernalization pathway plays in repressing the floral transition. What the correlation analysis uncovers, however, is that global transcriptional responses are also delayed in Tapidor relative to Westar. Therefore, in the apex, vernalization seems to have a large effect on the transcriptome to delay development. This suggests the developmental stage of the plant has a large effect on the transcriptome in the apex. Vernalization also seems to delay the upregulation of genes associated with stress responses, cell death, and aging at the final time point in the Tapidor leaf samples, relative to the Westar leaf samples. However, the correlation analysis suggests that the transcriptome of the leaf is affected more by the age of the tissue, rather than the developmental stage of the plant as a whole like the apex. This is likely a result of the first true leaf being sampled throughout the experiment (discussed in section 2.2.1). The observed delay in the upregulation of genes at the final time point in Tapidor is therefore likely to be an artefact of the expression profile normalization procedure.

3.3 *B. napus* vernalization pathway regulatory divergence

The vernalization response is arguably one of the most investigated floral pathways in *Brassica* crops[120,137,138,141,144,147], likely due to its agronomic importance[120,127]. Work has also been motivated by *FLC* and *FRI* homologues in *Brassica* crops being found in regions of the genome statistically associated with flowering time variation[132–143]. Molecular characterisation has also identified the importance of vernalization pathway genes, with between variety polymorphisms at the *FLC* locus responsible for heading data variation in *B. oleracea*[140]. However, aside from the association studies, the interactions between the copies of vernalization genes *in planta* have not been assessed. Even within the association studies, although large phenotypic effects were attributed to certain vernalization gene homologues, more subtle variation attributable to other homologues might be masked. The importance of considering the effects of multiple homologues on the floral transition is perfectly exemplified with *FLC*. Not only have the dosage effects of the gene been revealed[373], but the long non-coding RNA expressed from the *FLC* locus also has the potential of acting in trans to influence the expression of other *FLC* loci in the genome[351,380].

To investigate whether *B. napus* homologues of Arabidopsis vernalization pathway genes are mediating the difference in vernalization requirement between Tapidor and Westar, the behaviour of the genes was assessed in the transcriptomic time series. From analysing the expression of *FLC*, *FRI*, and PRC2 component genes, in both Westar and Tapidor, *BnFLC* genes emerge as being the most likely candidates for mediating the difference in flowering time between Tapidor and Westar. Specifically, *BnFLC* genes on A10, A2, and A3 show variety-specific responses, suggesting these copies are responsible for the requirement for cold that Tapidor plants exhibit in order to flower. *BnFLC* genes on chromosomes C2 and A3 exhibit cold induced silencing of expression in both varieties, suggesting that these copies are responsible for the vernalization response observed in Westar[241]. No apparent tissue specificity was present between the *BnFLC* genes, suggesting that spatial subfunctionalization[206,213] has not taken place.

3.3.1 *FLOWERING LOCUS C*

The product of the *FLC* gene in Arabidopsis is the central regulator of the vernalization pathway[27,373]. Given that *FLC* copy number in Arabidopsis impacts floral growth in a dosage dependent manner[373] and that the gene product seems to have contrasting roles in both the leaf and apex[31], a key question is whether the copies in *B. napus* exhibit regulatory divergence. In order to assess whether this is the case, it is important to consider the expression of the gene in both the apex and the leaf, as it has been found that *FLC* plays roles in both[31]. In the developmental time series, across both tissues and varieties, ten copies of *BnFLC* are expressed above 2.0 FPKM at at least one time point; four on the A genome and six on the C genome. The complement of copies expressed, however, varies based on the tissue and variety investigated.

In the Tapidor apex, nine copies of *BnFLC* are expressed, and exhibit a *mixed* pattern of regulatory module assignment, indicating regulatory divergence (Figure 3.10). The largest of these regulatory modules consist of all A genome *BnFLC* genes and the two expressed *BnFLC.C3* copies. These genes are grouped together as all exhibit a decrease in expression during the vernalization period, with expression remaining low post cold. This pattern of expression mirrors that of *FLC* from Arabidopsis[29]. *BnFLC.C2* also decreases during the vernalization period, but the repression is not stable, and reactivation is observed post-cold (Figure 3.10). This pattern of *FLC* expression has been observed in Arabidopsis lines that have not been given sufficient cold exposure to result in the stable repression of *FLC*[245,345]. A partial reactivation of *BnFLC.A3b*, *BnFLC.C3b*, and *BnFLC.C3c* results in the clustering coefficient of these copies, and the *BnFLC.C2* copy, being high. Finally, the *BnFLC.C9a* and *BnFLC.C9b* have similar expression traces in the time series, with both genes having relatively low expression before and during cold, with an increase in expression post-cold. Interestingly, the *BnFLC.C9b* gene seems to increase during cold, an expression behaviour that has not been reported for *FLC* in Arabidopsis.

Seven of the nine *BnFLC* copies expressed in the Tapidor apex are also expressed in the Westar apex, with the *BnFLC.A10* and *BnFLC.C3c* copies

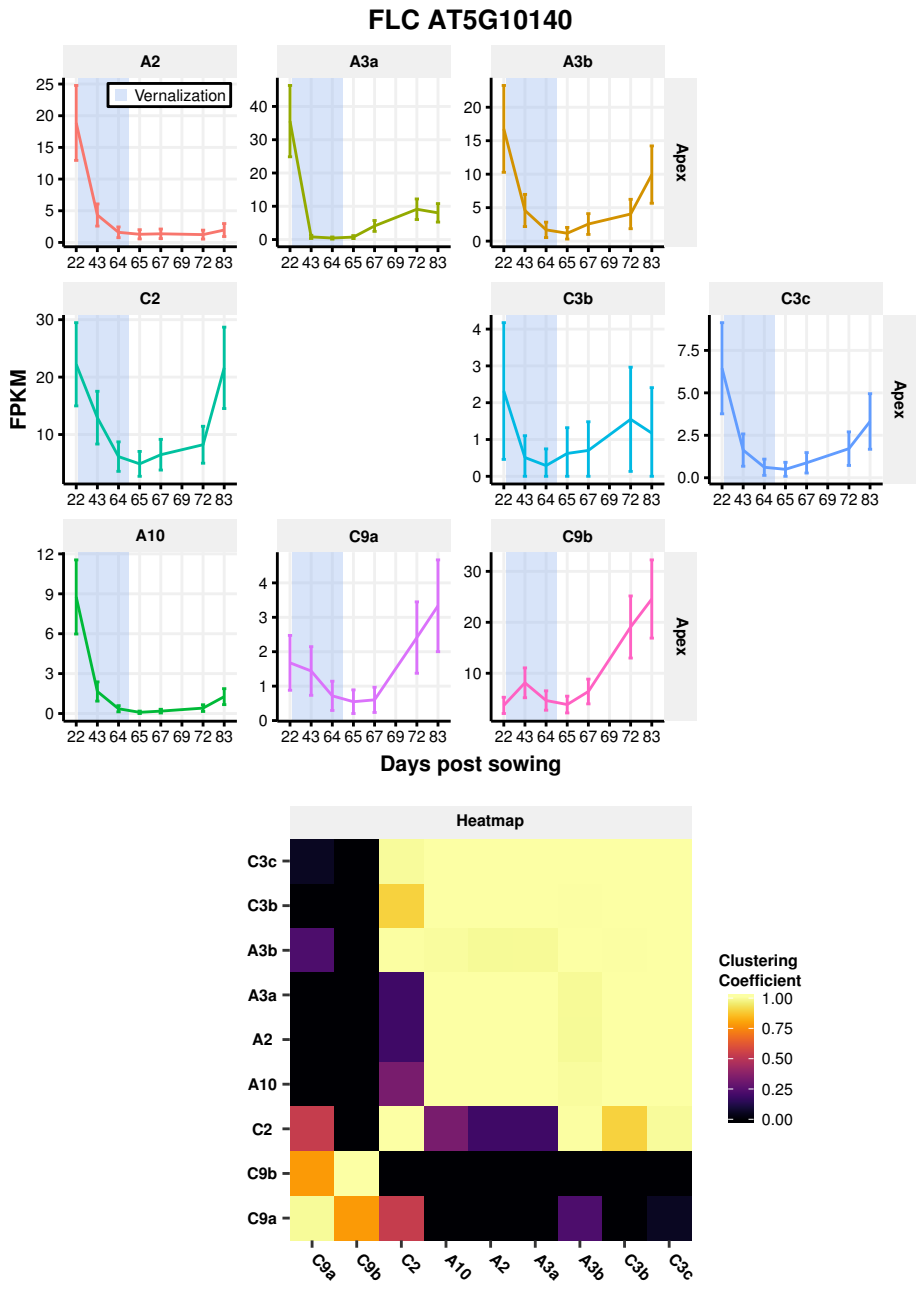


Figure 3.10: Expression traces for the *BnFLC* genes in the apex of Tapidor. The expression values and the 95% confidence intervals of those expression values as computed by Cufflinks are displayed. A heatmap of the clustering coefficients calculated by the SOM based method (discussed in section 2.2.4) quantifies the similarity between the expression profiles. The A2, A10, A3 and C3 copies show very similar expression traces. *Continued on Page 187.*

Continued from Page 186. The C2 copy behaves similarly to the A3b and C3 copies. The C9 copies are similar to each other, but have expression profiles that are different from the other *BnFLC* copies.

lacking expression in the spring variety relative to the winter variety (Figure 3.11). Given that the A10 copy was relatively highly expressed in the winter variety, this supports findings that this copy is the main copy driving the requirement for cold in *B. napus*[134,141–143]. All copies except the *BnFLC.C9b* copy decrease in expression during the vernalization period, with expression remaining low after the treatment, resulting in high clustering coefficients between these copies. The *BnFLC.C3b* copy shows slight reactivation after the cold treatment, leading to it having lower clustering coefficients relative to the other genes in the regulatory module (Figure 3.11). As was the case in Tapidor, *BnFLC.C9b* shows a markedly different expression trace, exhibiting a slight increase in expression halfway through vernalization, with a further increase in expression post-cold.

Analysis of the expression traces in the apex in both Tapidor and Westar reveals that all *BnFLC* genes, except *BnFLC.C9b*, decrease in expression during the cold treatment. The A10 and C3c copy are expressed in the winter variety, yet lack expression in the spring variety. Some copies exhibit reactivation after the cold-induced decrease, suggesting that the vernalization treatment was not sufficient to stably silence those copies. Interestingly, this reactivation seems to be variety-specific for some genes, with the A3b, C2, and C9a copies exhibiting reactivation in the winter and not the spring variety.

The expression response of *FLC* in Arabidopsis is quantitative at the tissue level[29,361]. The magnitude of expression is therefore an important aspect of *FLC* regulation. Comparing the expression traces of the A genome copies of *BnFLC* expressed in both varieties in the apex (Figure 3.12) revealed that the A2 and A3b copies are initially expressed at significantly lower levels in the spring variety relative to the winter variety. Therefore, although these copies both exhibit a decrease in expression during the cold treatment in both varieties, the absolute difference in expression level is greatest in the winter variety. The A3a copy of *BnFLC* shows remarkably similar expression traces

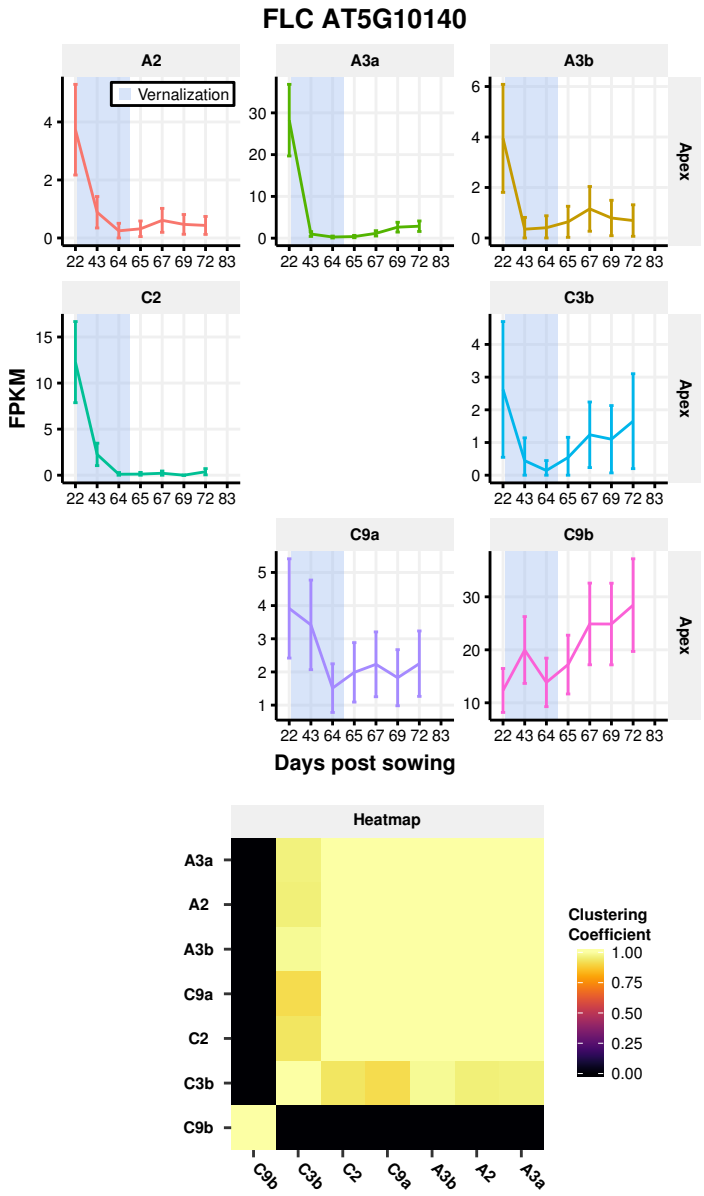


Figure 3.11: Expression traces for the *BnFLC* genes in the apex of Westar. The expression values and the 95% confidence intervals of those expression values as computed by Cufflinks are displayed. A heatmap of the clustering coefficients calculated by the SOM based method (discussed in section 2.2.4) quantifies the similarity between the expression profiles. All copies, except the C9b copy, show similar expression traces.

and expression levels throughout the developmental time series in both varieties. Differences in the magnitude of expression are also observed for the C genome copies of *BnFLC* (Figure 3.13). Like the A2 and A3b copies, *BnFLC.C2* is more highly expressed pre-cold in the winter variety relative to the spring variety, suggesting a role in delaying the floral transition in Tapidor. The C3b copy of *BnFLC* shows very similar expression patterns and levels across the entire time series in both varieties, and is very lowly expressed in general. This suggests that it does not contribute to the differences in flowering observed between the two varieties. Finally, the C9 copies of *BnFLC* are frequently more highly expressed in the spring variety relative to the winter variety. This is especially true for *BnFLC.C9b*, where the expression level of the gene in the spring variety is approximately three-fold higher than the winter variety at the beginning of the time series. That these copies are more highly expressed in the spring variety indicates that these copies likely do not play a role in delaying the floral transition, unlike the role of *FLC* in Arabidopsis.

The expression of *BnFLC* genes in the apex reveals that all but one homologue decrease in expression during the cold treatment, in line with expectations from Arabidopsis[29]. That some copies exhibit reactivation in the winter, and not the spring variety, suggests that potentially the length of cold was not sufficient to stably repress the expression of those copies. Comparing the magnitude of expression between the copies reveals that the A2, A3b, A10, and C2 copies seem to be more highly expressed in Tapidor at the beginning of the time series relative to Westar. That these copies exhibit stable decreases in expression during cold treatment and are highly expressed in Tapidor makes them good candidates for being responsible for the delay in flowering in the winter variety. The A3a copy, however, exhibits cold induced stable repression in both varieties, indicating that it is potentially responsible for the vernalization response of Westar[241]. Finally, the C9b copy is more highly expressed in the spring variety relative to the winter variety and is not repressed during the cold treatment. This indicates that this particular copy of *BnFLC* has diverged significantly in its regulation relative to Arabidopsis *FLC*, and is likely not involved with mediating the vernalization response in the *B. napus* varieties considered here.

To assess whether the *BnFLC* genes exhibited tissue-specific expression, the expression of the genes was analysed in the leaf tissue. In the Tapidor leaf

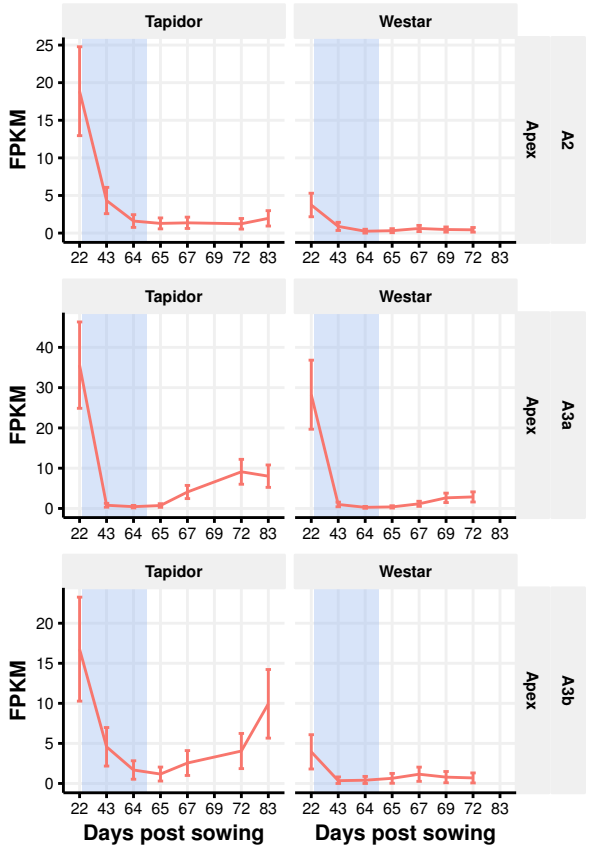


Figure 3.12: Expression traces for the A genome *BnFLC* genes commonly expressed in the apex of both varieties.

The expression values and the 95% confidence intervals of those expression values as computed by Cufflinks are displayed. The A2 and A3b copies exhibit varietal differences in the magnitude of expression at the pre-cold time point, in line with these copies delaying the floral transition in Tapidor relative to Westar. The A3a copy is similarly expressed in both varieties, suggesting it does not contribute to the observed delay.

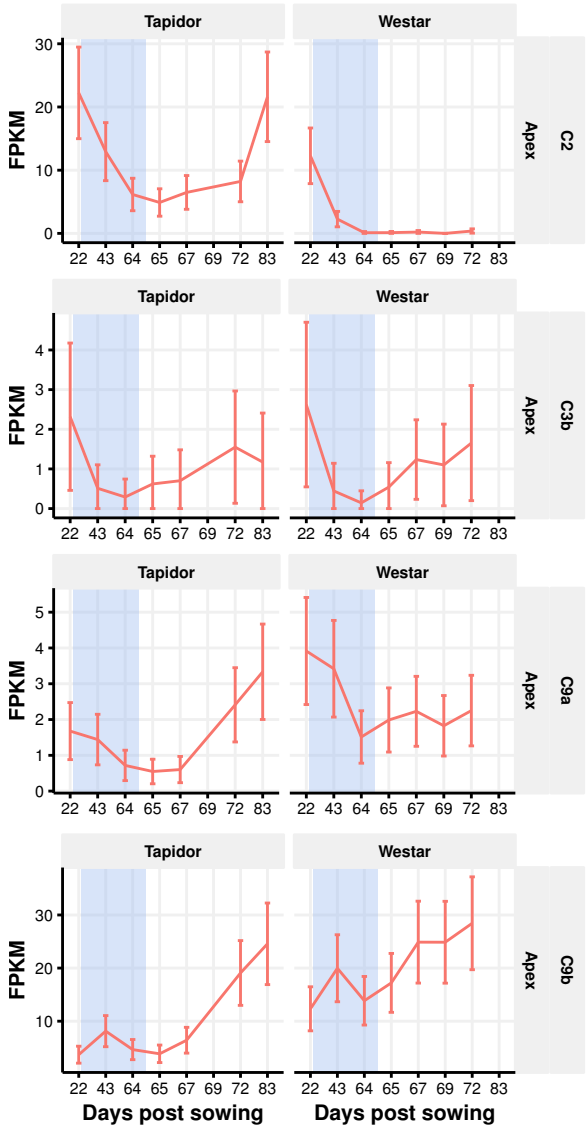


Figure 3.13: Expression traces for the C genome *BnFLC* genes commonly expressed in the apex of both varieties.

The expression values and the 95% confidence intervals of those expression values as computed by Cufflinks are displayed. Variety-specific differences in the magnitude of expression at the pre-cold time point are consistent with a role in the vernalization response. In contrast, the expression of the C9b copies is frequently higher across the time series in the spring variety, suggesting that these copies do not delay the floral transition.

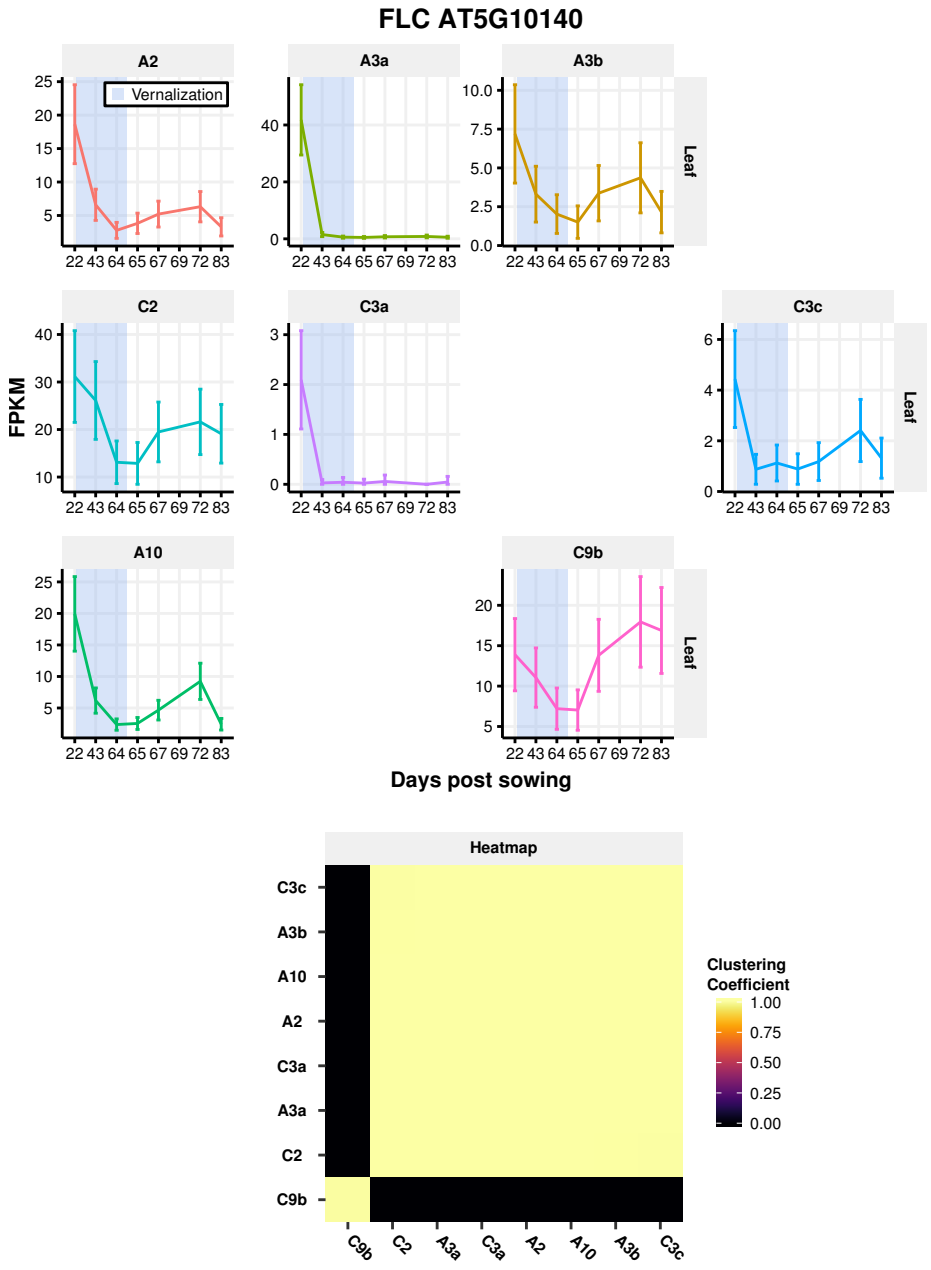


Figure 3.14: Expression traces for the *BnFLC* genes in the leaf of Tapidor. The expression values and the 95% confidence intervals of those expression values as computed by Cufflinks are displayed. A heatmap of the clustering coefficients calculated by the SOM based method (discussed in section 2.2.4) quantifies the similarity between the expression profiles. All copies, except C9b, have similar expression profiles as determined by the clustering coefficients.

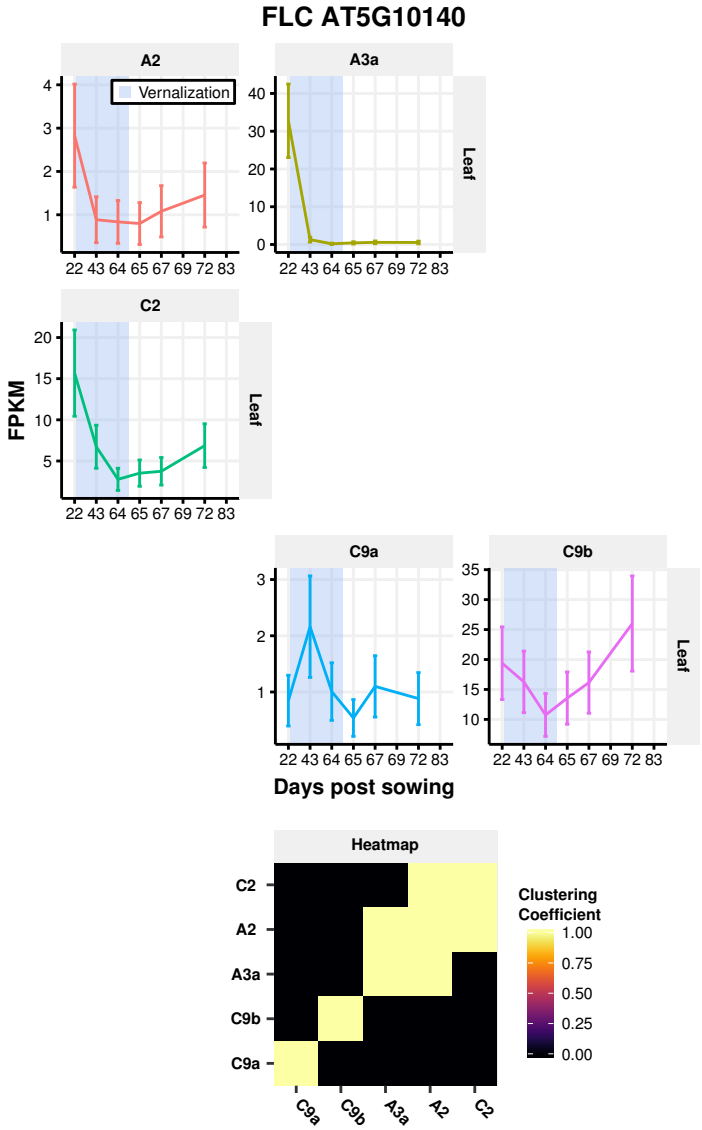


Figure 3.15: Expression traces for the *BnFLC* genes in the leaf of Westar. The expression values and the 95% confidence intervals of those expression values as computed by Cufflinks are displayed. A heatmap of the clustering coefficients calculated by the SOM based method (discussed in section 2.2.4) quantifies the similarity between the expression profiles. A *mixed* pattern of clustering coefficients is observed, with C9 copies being in separate regulatory modules and A2, C2, and A3a exhibiting a gradiant of similarity.

samples eight copies of *BnFLC* are detected as expressed; all four copies from the A genome, *BnFLC.C2*, *BnFLC.C3a*, *BnFLC.C3c*, and *BnFLC.C9b* (Figure 3.14). The *BnFLC.C3a* copy is expressed in the leaf and not the apex, whereas two *BnFLC* genes (*BnFLC.C9a* and *BnFLC.C3b*) are expressed in the apex and not in the leaf. However, the expression of these genes in their respective tissues is close to the 2.0 FPKM threshold used to determine if genes are expressed or not, suggesting that the presence or absence of these genes in the set of expressed genes is more heavily influenced by noise relative to the other copies. This suggests that *FLC* homologues in *B. napus* have not diverged in terms of spatial expression domains. The genes expressed in the leaf have a *distinct* regulatory module assignment, with all the genes except *BnFLC.C9b* being assigned to the same regulatory module. The seven genes assigned to the largest regulatory module all exhibit a decrease in expression during the vernalization period to very low levels. In the case of the A3a and C3a copies, this repression is very stable, whereas the other genes show a slight reactivation of expression that peaks at the day 72 time point before decreasing at the final time point. The *BnFLC.C9b* copy also decreases in expression during the vernalization period, although the repression is not stable, with the expression level increasing post-cold. This reveals that in the Tapidor leaf, all expressed copies of *BnFLC* exhibit a cold-induced repression in expression, which in the case of *BnFLC.C9b* is not stable.

Fewer *BnFLC* copies are expressed in the leaf in Westar relative to Tapidor. The A10, A3b, and C3 copies are not expressed in the spring variety, whereas the C9a copy is expressed in Westar and not Tapidor (Figure 3.15). That A10 and C3c show variety-specific expression in both the apex and leaf indicates that these copies may delay the floral transition in the winter variety. In the Westar leaf, a *mixed* pattern of regulatory module assignment is observed, with four modules identified (Figure 3.15). The *BnFLC.A2* and *BnFLC.C2* copies form one module, with both exhibiting decreases during the cold, with partial reactivation post-cold. *BnFLC.A2* is in another regulatory module with *BnFLC.A3a*, with the latter rapidly decreasing in expression in response to cold and staying repressed after the cold treatment. This intransitivity is likely due to the combination of two differences between the *BnFLC.A3a* and *BnFLC.C2* expression traces. The rate of decrease during the cold is

more rapid in the *BnFLC.A3a* copy relative to the *BnFLC.C2* copy, with the former having a near zero expression level at the day 43 time point, taken halfway through the cold treatment. The other behaviour that differs is the post-cold treatment, with the *BnFLC.C2* copy showing partial reactivation unlike the *BnFLC.A3a* copy. Different rates of *FLC* silencing and different reactivation dynamics are also observed as natural variation in *FLC* expression for Arabidopsis[245], suggesting that the variation observed in the leaf tissue in Westar between the A2, A3a, and C2 copies of *BnFLC* may have biological consequences. The two *BnFLC* copies on the C9 chromosome are located in regulatory modules that are unique to them. The *BnFLC.C9a* copy shows a partial increase in expression halfway through the vernalization treatment, but returns to pre-cold expression levels towards the end of cold and after the treatment. Although the *BnFLC.C9a* copy is expressed in the leaf in Westar and not Tapidor, the expression of the gene only marginally exceeds the 2.0 FPKM at a single time point, suggesting that its effect on flowering, if any, will be minimal. Like the A2, C2, and A3a copies of *BnFLC*, *BnFLC.C9b* shows a decrease during the cold treatment, but also displays a reactivation after the cold treatment.

The expression traces of *BnFLC* genes in the leaf, like those from the apex, reveal that the majority of copies respond to cold treatment by decreasing in expression. Interestingly, the prevalence of *BnFLC* copies exhibiting reactivation in the leaf is less than in the apex, potentially indicating that the apex in *B. napus* is perennial in nature. The only copy that exhibits a significant change in expression pattern between tissues is *BnFLC.C9b*. In the apex this copy does not exhibit cold-induced silencing, whereas it does in the leaf. This suggests the C9b copy of *BnFLC* exhibits tissue-specific regulation in both varieties.

Comparing between varieties (Figure 3.16), similar differences in the magnitude of *BnFLC* gene expression were observed in the leaf as they were for the apex. Although *BnFLC.A2* demonstrates a similar response to cold in both Tapidor and Westar, expression is approximately four-fold higher in Tapidor across the entire time series. Likewise, the C2 copy is expressed two-fold higher in Tapidor at the first time point relative to Westar, and remains more highly expressed across the entire time series (Figure 3.13). The differences in the

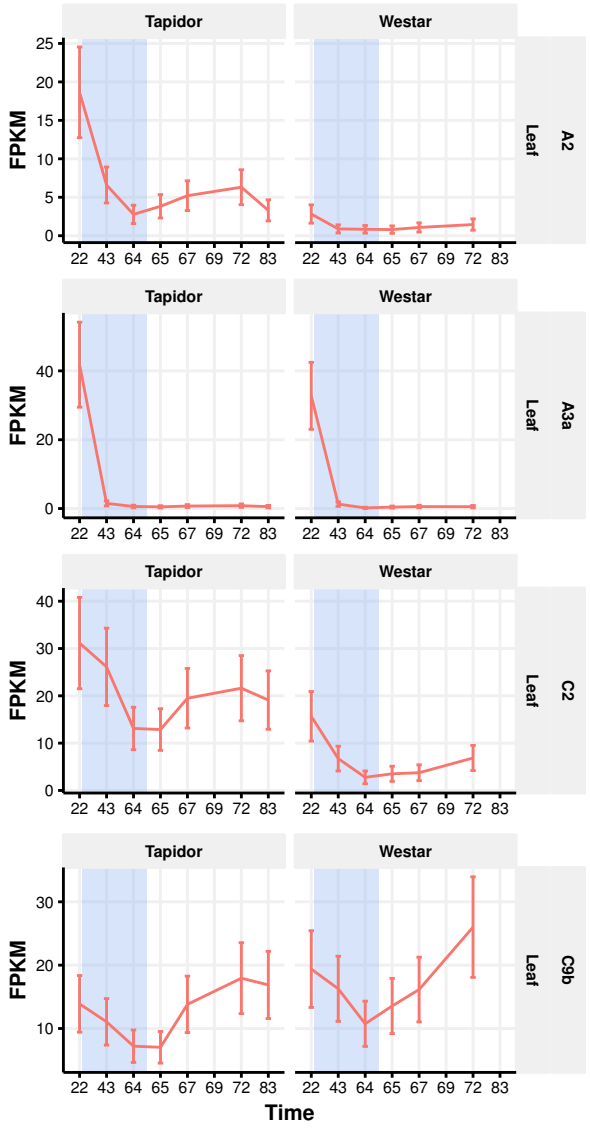


Figure 3.16: Expression traces for the *BnFLC* genes commonly expressed in the leaf of both varieties.

The expression values and the 95% confidence intervals of those expression values as computed by Cufflinks are displayed. Variety-specific differences in the magnitude of expression at the pre-cold time point for *BnFLC.A2* and *BnFLC.C2* are consistent with a role in the vernalization response. The response of the A3a copy to cold treatment is similar in both varieties. In contrast, the expression of the C9b copy is more highly expressed across the time series in the spring variety, suggesting that this copy is not involved with delaying the floral transition.

magnitude of expression in the leaf between the two varieties for the A2, A3b, A10, and C2 copy of *BnFLC* are consistent with the genes delaying the floral transition in the winter variety. *BnFLC.A3a* shows similar expression levels in both the winter and spring variety in the leaf, suggesting that this copy is not responsible for the delayed floral transition in Tapidor. Finally, the *BnFLC.C9b* copy is more highly expressed in Westar throughout the entire time series, suggesting that is does not play a role in the vernalization response in *B. napus*.

Taken together, this evidence indicates that the majority of *BnFLC* genes have retained a regulatory response to cold, and do not exhibit significant tissue specificity in their expression. In the apex in both varieties, all but one *BnFLC* gene decreases in expression during the cold. The same is also true in the leaf, although the *BnFLC* gene that does not exhibit a cold-induced decrease is very lowly expressed and unlikely to have a significant role in the plant. Within the apex in Tapidor, certain copies exhibit regulatory divergence by reactivating in expression after cold. This behaviour mirrors that of Arabidopsis accessions that have not received enough cold to fully repress *FLC*[245,345] and *FLC* homologues in perennial relatives of Arabidopsis[167,381]. This suggests that these particular copies of *BnFLC* have not received sufficient cold to be fully repressed, or the apex in Tapidor is somewhat perennial in nature. The copy that exhibits most regulatory divergence is *BnFLC.C9b*. In the apex this copy increases in expression during and after vernalization, while in the leaf cold-induced silencing is observed but is not stable. This suggests that this particular copy does not influence the vernalization response in *B. napus*, and has therefore acquired a separate function in the plant.

Comparing the magnitude of expression between varieties reveals copies of *BnFLC* that are likely to be mediating the delay in flowering in Tapidor relative to Westar. The A2, A3b, A10, C2, and C3c copies of *BnFLC* are all expressed more highly in the winter variety than in the spring. That all of these copies also exhibit cold induced silencing makes them good candidates for mediating the delay in flowering in the winter variety. Of particular interest are A2 and A10, as the silencing of these copies is more stable post-cold than the others. This suggests that one or both of these copies controls the vernalization requirement of Tapidor, that is, the expression of these copies

has to be repressed in order for the plants to flower. The other *BnFLC* copies, A3b, C2, and C3c, may mediate the vernalization response, in that they delay flowering when expressed, but do not have to be repressed for the plant to undergo the floral transition. Comparing the magnitude of expression between varieties also suggests that *BnFLC.C9b* is not involved with delaying flowering time, as the gene is more highly expressed in Westar relative to Tapidor.

Comparing expression data between the apex and the leaf reveals some tissue-specific expression. More copies of *BnFLC* exhibit reactivation in the Tapidor apex (Figure 3.10) relative to the Tapidor leaf (Figure 3.14). This supports the hypothesis that the Tapidor apex may have perennial characteristics. In *Arabidopsis halleri*, a perennial relative of Arabidopsis, the expression of a *FLC* homologue was found to reactivate in young leaves[381]. It is therefore likely that *BnFLC* reactivation is not observed in the leaf as the first true leaf was sampled throughout the time series, such that the age related effects and leaf senescence result in the lack of expression. *BnFLC.C9b* undergoes cold-induced silencing in the leaves of both varieties, but does not do so in the apex. In addition, in Tapidor samples, *BnFLC.A3b* is expressed at approximately the same level as *BnFLC.A2* in the apex, whereas in the leaf the A3b copy is expressed ~2.5-fold lower than the A2 copy. These findings suggests that some copies of *BnFLC* are expressed in a tissue-specific manner. In the case of *BnFLC.A3b*, potentially its effect on a vernalization response is mediated predominantly in the apex. This is interesting given the different roles *FLC* has in the apex and leaf in Arabidopsis[31].

3.3.2 Polycomb repressive complex 2 proteins

Most *BnFLC* genes become silenced during cold in a similar manner to *FLC* in Arabidopsis (section 3.3.1). As the Polycomb repressive complex 2 (PRC2) proteins are integral to this repression, homologues of the genes were investigated to understand whether expression divergence between the genes could influence the response to cold in *B. napus*. First identified in *D. melanogaster*, Polycomb group (PcG) proteins regulate gene expression in both animal and plant kingdoms[382,383]. The PcG proteins form multiple families of protein complexes that possess different biochemical activities[384]. PRC2 is one such

complex that is involved with chromatin compaction through the methylation of lysine 27 of histone protein H3[383]. PRC2 is composed of four core units: Enhancer of zeste (E[z]), which confers the histone methyltransferase activity to the complex[385]; Suppressor of zeste (Su[z]12); Extra sex combs (Esc), and Nucleosome remodelling factor 55 (Nurf55)[382]. In Arabidopsis, there are three identified E[z] homologues, three Su[z]12 homologues, five Nurf55 homologues, and one Esc homologue[382,386], leading to a much more complex role for PRC2 during development[386,387]. Despite this complexity, it seems that one particular combination of PRC2 proteins is involved with vernalization[388,389]. *VRN2* is the Su[z]12 homologue in Arabidopsis that associates with the Arabidopsis homologues of Esc (*FERTILIZATION INDEPENDENT ENDOSPERM 1*; *FIE1*), E[z] (*SWINGER*; *SWN*), and Nurf55 (*MULTICOPY SUPRESSOR OF IRA1*; *MSI1*)[388,389]. The gene was identified in a mutant screen for plants that had an impaired vernalization response[390]. In addition, in *Medicago truncatula* a mutant in a homologue of *VRN2* was found to disrupt the vernalization response in the plant[391]. In order to assess whether regulatory divergence among components of the PRC2 could be influencing the vernalization response in *B. napus*, expression of *VRN2*, *SWN*, *MSI1*, and *FIE1* homologues was analysed. As very little regulatory and between variety divergence was observed for *SWN* and *FIE1 B. napus* homologues, the analysis of those genes can be found in Appendix B.

Two *B. napus* homologues of *VRN2* are expressed in both the leaf (Figure 3.18) and apex (Figure 3.17). The expression of the genes does not change dramatically across the time series in either tissue or variety, although all copies of the gene exhibit a slight increase in expression during the vernalization treatment. The magnitude of expression is largely similar between varieties also, suggesting that expression differences in *BnVRN2* genes does not influence the different vernalization requirements of Tapidor and Westar. However, in the apex *BnVRN2.A8* (Figure 3.17) is more highly expressed than the *BnVRN2.C8.Random* copy, whereas in the leaf this relationship is reversed (Figure 3.18). This potentially indicates that the two homologues of *VRN2* have undergone spatial subfunctionalization in *B. napus*. The expression of *VRN2* in *A. thaliana* was found to be relatively unaltered by vernalization, being consistently expressed throughout development[388]. The increase in

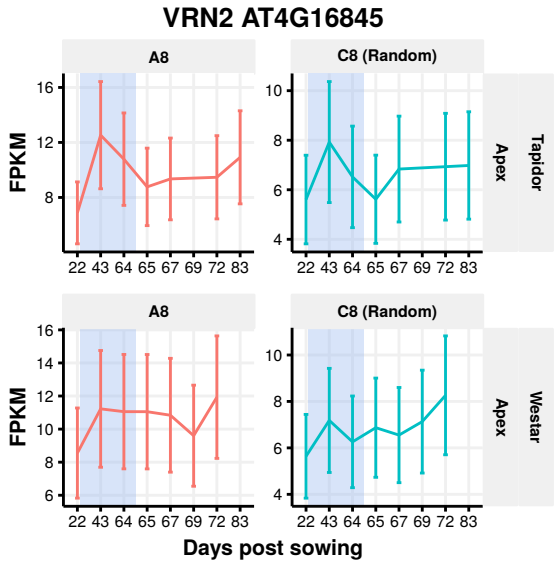


Figure 3.17: Expression traces for the *BnVRN2* genes in the apex.
The expression values and the 95% confidence intervals of those expression values as computed by Cufflinks are displayed. Within a variety, the two homoeologues retain similar expression profiles.

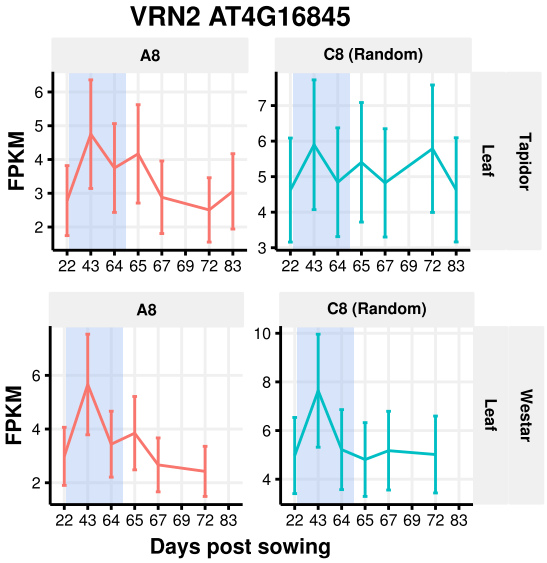


Figure 3.18: Expression traces for the *BnVRN2* genes in the leaf.
The expression values and the 95% confidence intervals of those expression values as computed by Cufflinks are displayed. Within a variety, the two homoeologues retain similar expression profiles.

expression observed for the *BnVRN2* genes during the cold is an indication, therefore, that the *BnVRN2* genes may be more cold responsive than the gene in Arabidopsis. This is supported by results from *Medicago truncatula*, where a *VRN2* homologue was found to increase in expression during the cold and influence the timing of the floral transition when mutated[391].

MSI1 is part of a family of WD40 repeat proteins that bind to histones and are thought to act as a protein scaffold[392]. *MSI1* is involved with the vernalization response in Arabidopsis[393] and has been found to be important for the regulation of plant homeotic genes in the apex[394]. The gene is expressed in many tissues, and when expression is impaired a number of floral and developmental processes are affected[394,395].

In total there are six expressed copies of *BnMSI1* in both Tapidor and Westar; two from the A genome and four from the C genome. In the leaf, three copies are expressed; the A2, C2, and C3a copies, although the C2 is so lowly expressed it will not be discussed further. The A3 and C3a copies exhibit very similar expression profiles to each other and between varieties, with a transient increase in expression during the vernalization period (Figure 6.7). This suggests that *BnMSI1.A3* and *BnMSI1.C3a* are cold-responsive, and potentially play a role in the vernalization response. In the apex, six copies of *BnMSI1* are expressed; in addition to the three expressed in the leaf there are also copies expressed from the A10, C5, and C9 chromosomes (Figure 3.19). Unlike in the leaf, *MSI1* homologues either do not respond to the cold, or exhibit a decrease in expression during vernalization. Therefore, copies of *MSI1* in *B. napus* seem to be cold-responsive in a tissue-specific manner. Considering the magnitude of expression, *BnMSI1.A3* and *BnMSI1.C3a* are the most highly expressed copies in each tissue. Interestingly, these copies exhibit expression magnitude differences between varieties in both tissues. For example, the maximal expression value for *BnMSI1.A3* in Tapidor apex is three- to four-fold higher than the expression maxima in Westar, in both tissues. Therefore, regulatory divergence between *BnMSI1* genes is present, with the A3 and C3a copies being most highly expressed. Between varieties, these two copies exhibit differences in expression magnitude that could potentially influence the floral transition.

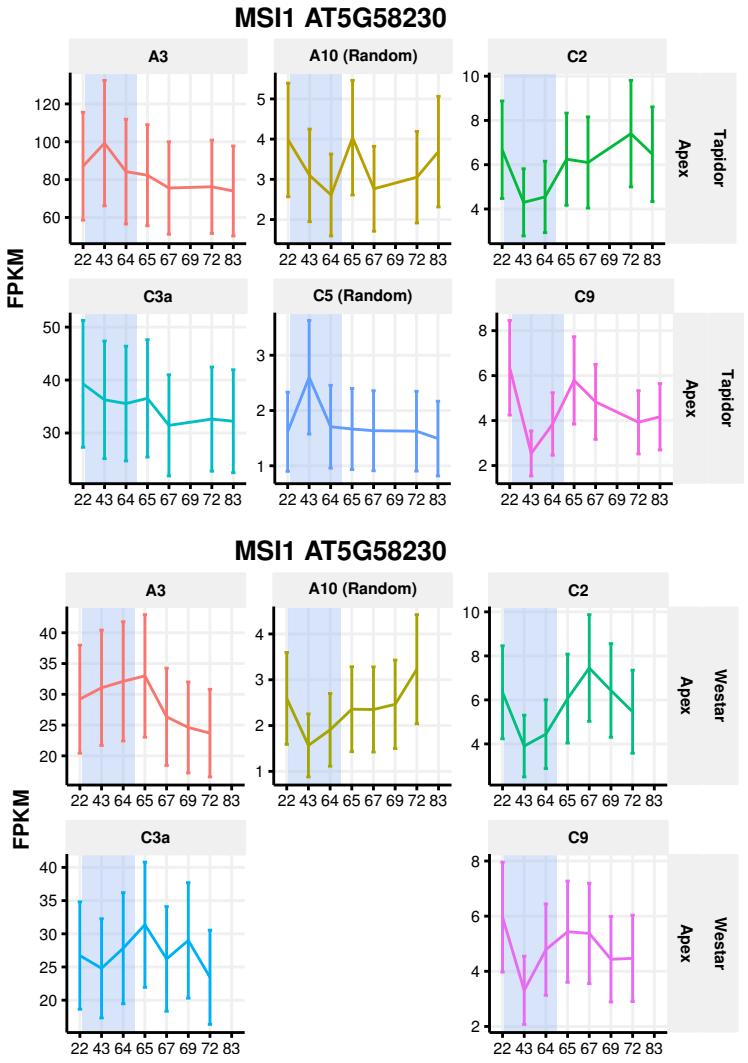


Figure 3.19: Expression traces for the *BnMSI1* genes in the apex. The expression values and the 95% confidence intervals of those expression values as computed by Cufflinks are displayed. Largely similar patterns of expression are observed between the two varieties, although the A3 and C3a copies are much more highly expressed in Tapidor.

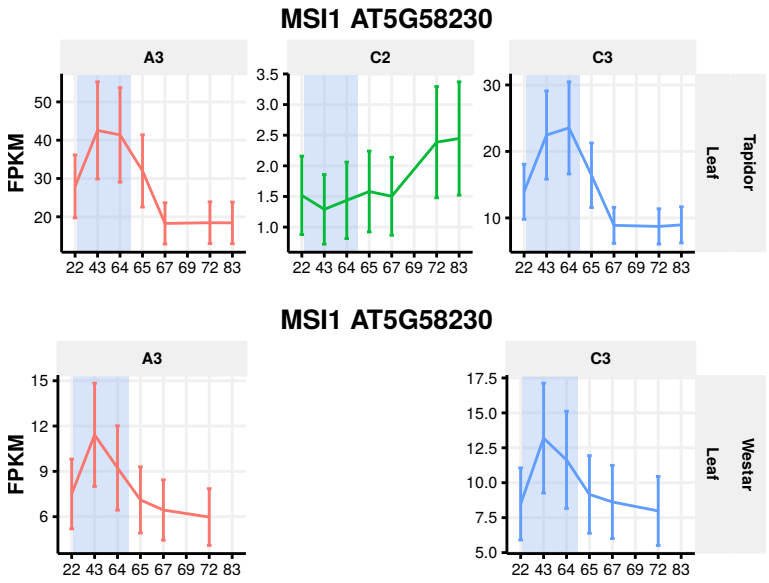


Figure 3.20: Expression traces for the *BnMSI1* genes in the leaf. The expression values and the 95% confidence intervals of those expression values as computed by Cufflinks are displayed. Largely similar patterns of expression are observed between the two varieties, although the A3 and C3a copies are much more highly expressed in Tapidor.

Among the PRC2 components found to be involved with the vernalization response in Arabidopsis, only homologues of *VRN2* and *MSI1* exhibited regulatory divergence. The *BnVRN2* genes have diverged in terms of spatial expression domains, with the A8 copy more highly expressed in the apex, and the C8 copy more highly expressed in the leaf. This spatial divergence may represent subfunctionalization, with each copy having become specialized towards the requirements of each tissue. Although there is variation in the magnitude of expression for the *BnMSI1* genes between the varieties, this variation does not account for the altered vernalization requirement between the varieties. If *BnMSI1* was repressing *BnFLC*, which would be expected given that the PRC2 complex is involved with the silencing of *FLC* in Arabidopsis[396], the higher expression of *BnMSI1* in Tapidor would result in lower *BnFLC* expression. This is not observed (Figures 3.12, 3.13, and 3.16), suggesting that potentially the higher expression of *BnMSI1.A3* and *BnMSI1.C3a* in Tapidor relative to Westar has another role. Potentially the higher expression of *BnMSI1* is required in Tapidor to sensitize the system, such that when cold is sensed Polycomb based silencing responds quickly. Alternatively, the high expression of *BnMSI1* may be repressing genes other than *BnFLC* copies, such as floral activators as has been shown in Arabidopsis[394].

3.3.3 PHD finger containing proteins

Proteins containing plant homeodomain (PHD)-finger proteins have been found to mediate histone interactions[358] and hence induce structural changes to chromatin. In Arabidopsis, a PHD finger protein was found in a mutant screen for plants insensitive to vernalization[357]. *VERNALIZATION INSENSITIVE 3* (*VIN3*) is required for both *FLC*-dependent and *FLC*-independent vernalization, and changes to the expression of *VIN3* result in histone modifications at the *FLC* locus. These modifications were found to be a consequence of PRC2 activity, with *VIN3* associating with the complex during vernalization[356]. Further work identified additional PHD-finger proteins that associate with the PRC2 implicated with vernalization, namely, *VIN3-LIKE1* (*VIL1*), and *VIL2*[386]. With *VIN3*, these *VIL* proteins form a family of proteins called the (*VERNALIZATION5*/*VIN3-LIKE*) VEL family[397]. In line with their

roles with the vernalization PRC2 complex, these three PHD-finger proteins have been found to associate[353,398]. In addition to the vernalization pathway, *VIL1* and *VIL2* have been found to influence the photoperiod flowering pathway[398,399]. As a result of the key roles these genes play in mediating the vernalization response, their expression profiles in the two *B. napus* varieties were investigated. As very little regulatory and between variety divergence was observed for *VIL1* and *VIL2 B. napus* homologues, the analysis of those genes can be found in Appendix B.

Three copies of *BnVIN3* are expressed across both tissues and varieties; one copy on the A2, A3, and C2 chromosomes. In both the apex (Figure 3.21) and the leaf (Figure 3.22) the expression pattern of the gene exhibits an increase during the vernalization treatment and returns to low temperatures post-cold. This is in line with the expression of *VIN3* in Arabidopsis[357]. Comparing the magnitude of expression, between variety differences are present, but only for certain copies. In the apex, *BnVIN3.A2* and *BnVIN3.A3* are two- to three-fold more highly expressed during the cold treatment in Tapidor compared to Westar, whereas the C2 copy is similarly expressed in both (Figure 3.21). In the leaf, only the A3 copy exhibits similar differences in the magnitude of expression between varieties, with the A2 and C2 copy being more similarly expressed (Figure 3.22).

Copies of *BnVIN3* exhibit between variety expression that is consistent with *VIN3* being required to direct the repression of *FLC* during cold. The higher expression of *BnVIN3.A2* in apex tissue, and the higher expression of *BnVIN3.A3* in both tissues, in the winter variety relative to the spring variety, may be required in order to repress the more transcriptionally active *BnFLC* copies in Tapidor. In addition, this between variety divergence is tissue specific, with both A2 and A3 exhibiting higher expression magnitudes in the apex samples and only A3 in the leaf samples.

3.3.4 FRIGIDA

Despite variation at *FRI* accounting for the majority of flowering time variation in Arabidopsis[28], the spatiotemporal expression profile of the gene has not been well elucidated. What is known, however, is that mutations that disrupt

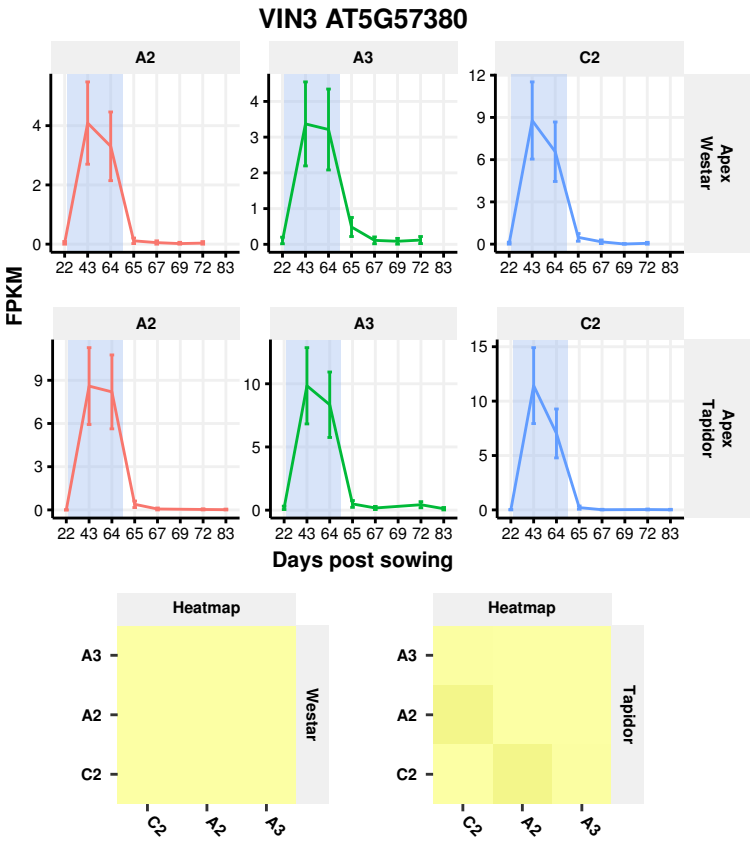


Figure 3.21: Expression traces for the *BnVIN3* genes in the apex.
The expression values and the 95% confidence intervals of those expression values as computed by Cufflinks are displayed. A heatmap of the clustering coefficients calculated by the SOM based method (discussed in section 2.2.4) quantifies the similarity between the expression profiles. An upregulation of expression during the vernalization treatment is observed in all copies and in both varieties.

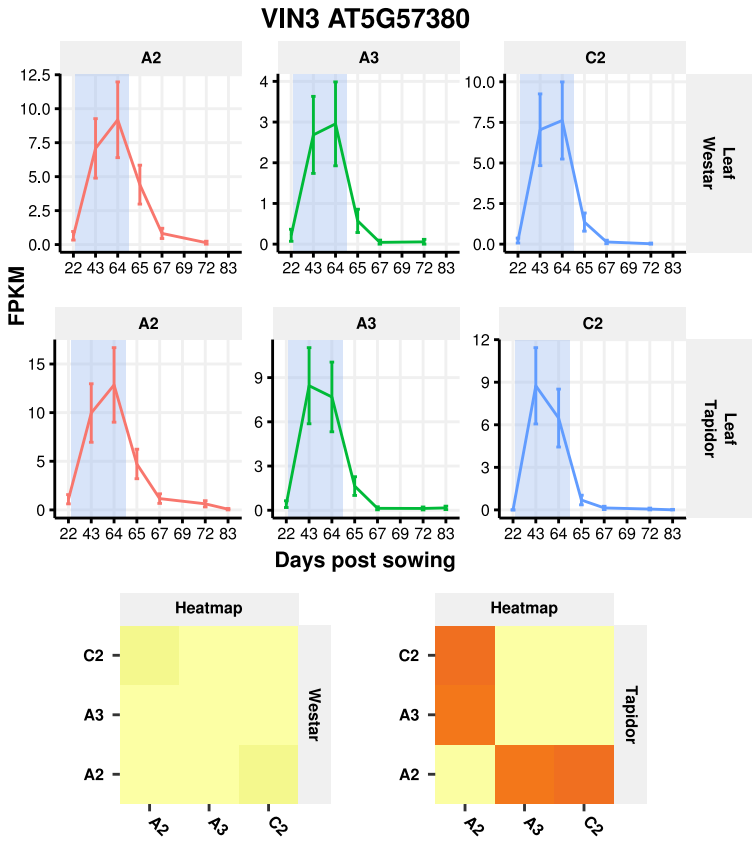


Figure 3.22: Expression traces for the *BnVIN3* genes in the leaf.
The expression values and the 95% confidence intervals of those expression values as computed by Cufflinks are displayed. A heatmap of the clustering coefficients calculated by the SOM based method (discussed in section 2.2.4) quantifies the similarity between the expression profiles. An upregulation of expression during the vernalization treatment is observed in all copies and in both varieties.

the expression of the *FRI* gene causes early flowering through *FLC* expression being lowly expressed[28,400–402].

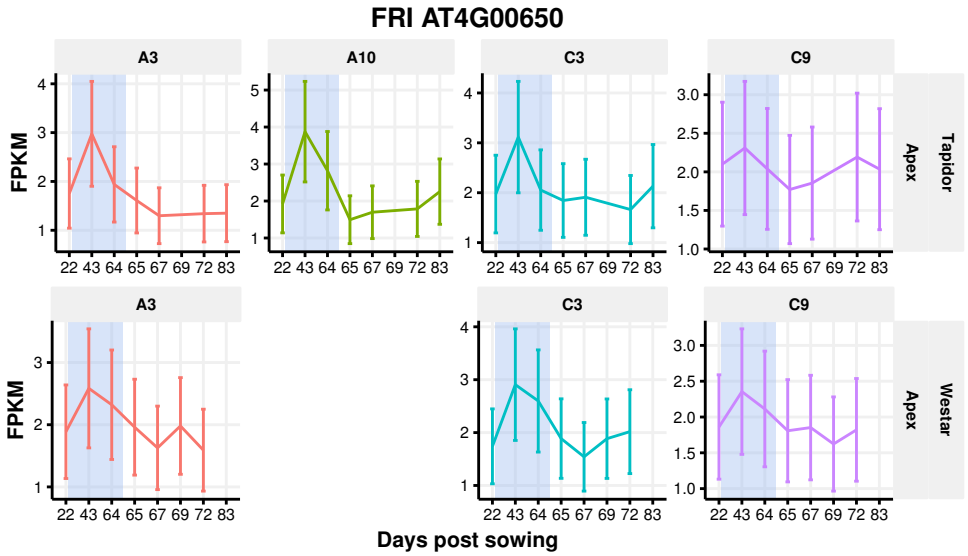


Figure 3.23: Expression traces for the *BnFRI* genes in the apex. The expression values and the 95% confidence intervals of those expression values as computed by Cufflinks are displayed. Expression of all copies are very low, with the A10 copy being expressed below the 2.0 FPKM threshold to be regarded as expressed.

The expression profiles of *BnFRI* genes in the apex (Figure 3.23) and leaf (Figure 3.24) exhibit strong similarities, suggesting that the *BnFRI* genes have not diverged in terms of expression domain. Slight expression increases are observed during cold for most copies in both the apex and leaf, with this not being the case for the C9 copy in the leaf (Figure 3.24). Comparing the magnitudes of expression between varieties reveals *BnFRI.A10* is the only copy that exhibits clear differences. The copy of *BnFRI* on A10 is more highly expressed in the winter variety, consistent with this copy being potentially responsible for the higher expression of *BnFLC* genes in the winter variety (section 3.3.1).

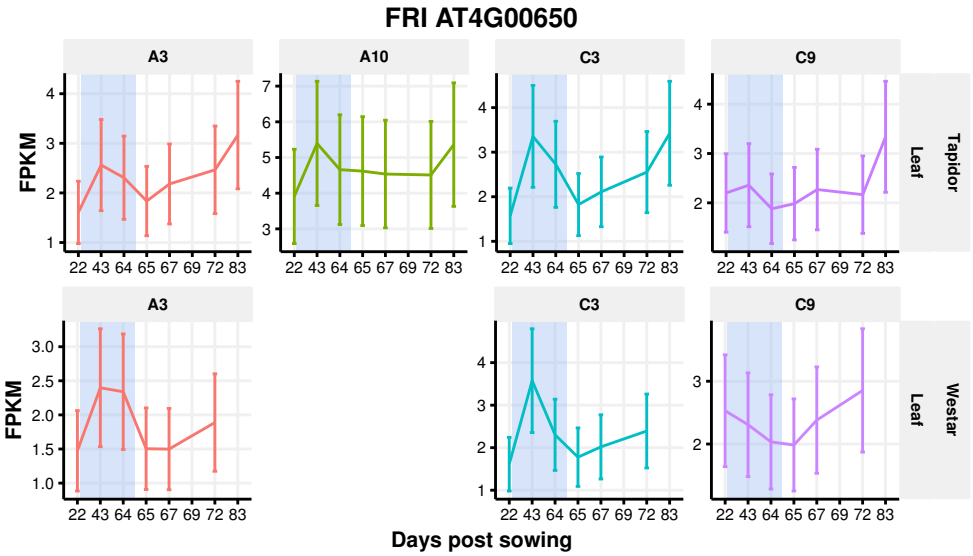


Figure 3.24: Expression traces for the *BnFRI* genes in the leaf. The expression values and the 95% confidence intervals of those expression values as computed by Cufflinks are displayed. Expression of all copies are very low, with the A10 copy being expressed below the 2.0 FPKM threshold to be regarded as expressed.

3.3.5 Conclusions

Analysing expression differences between *B. napus* homologues of genes involved with the Arabidopsis vernalization pathway identified a number of candidate genes that may be responsible for the delay in flowering observed in Tapidor. Among the *BnFLC* genes, the A2 and A10 copies seem most likely to mediate the cold requirement of Tapidor in order to flower. Both copies are lowly expressed in the spring variety throughout the time series, while in the winter variety the copies are more highly expressed initially and are stably repressed by the vernalization treatment. Analysis of the other key vernalization gene from Arabidopsis, *FRI*, identified the *BnFRI.A10* gene as exhibiting variety-specific expression. Given that alleles of *FRI* that fail to confer a vernalization requirement in Arabidopsis are the result of promoter deletions that result in low expression[28,400–402], it seems feasible that the observed difference in *BnFRI.A10* could play a role in the differences between Tapidor and Westar. Finally, components of the PRC2-PHD complex were more highly expressed in Tapidor than in Westar. While this initially seems counterintuitive, given that the complex is involved in the repression of *FLC*, it makes sense when thought of in terms of the products of these genes mediating the response to vernalization. Potentially more protein is required as more loci require repression in Tapidor compared to Westar. Alternatively, having high levels of these proteins available may increase the sensitivity of the system to cold. Having a sensitive system may be more important in Tapidor, which requires cold to flower, than in Westar.

As with the floral integrators in Westar (section 2.4), regulatory divergence is observed among the homologues of vernalization genes. *BnFLC* copies on chromosomes A3b, C2, and C3c are not stably repressed by cold in the apex and reactivate in expression after vernalization, while others remain lowly expressed. This suggests that different copies have different sensitivities to cold, the ramifications of which will be discussed at the end of this chapter. One of the most diverged *BnFLC* genes in terms of regulation is *BnFLC.C9b*, which exhibits divergence between varieties and tissues. Given that MADS-box containing genes have a wide range of roles and functions in plants[279], it is conceivable that *BnFLC.C9b* has diverged to have a role not involved with the

vernalization response. A number of the vernalization genes have tissue-specific expression, with *BnMSI1* genes exhibiting expression responses to cold in the leaf, and not the apex, and *BnVRN2* genes potentially partitioning their expression between the apex and leaf. This suggests that different vernalization responsive genes may be regulating the response in different tissues. The vernalization response in Arabidopsis is involved in both generating signals in the leaves and affecting how those signals are perceived in the apex[31]. Decoupling these two processes by having copies specialized towards each role could allow for greater robustness and flexibility in the system.

3.4 Floral integrator expression divergence in a winter variety

A potential avenue for the production of *B. napus* varieties with altered flowering time is via changes to the regulation of floral integrators. This is evidenced by studies that characterised the phenotypes of Arabidopsis plants constituently expressing floral genes, with plants frequently exhibiting alterations to flowering time and flower morphology[20,22,66,77,85,93,233]. This is supported by findings in Arabidopsis where natural variation at the *CO* promoter impacts flowering time[403], while variation at the *FT* orthologue in perennial ryegrass has also been found to be associated with flowering time differences[404]. Therefore, different alleles or altered regulation of floral integrators could potentially be contributing to the delay in flowering observed in the winter variety.

The altered expression of particular floral integrators could be due to an increased sensitivity to the vernalization response. In their analysis of gene expression divergence in Arabidopsis, Blanc and Wolfe (2004) discussed the concerted divergence of gene expression[229]. Concerted divergence involves the parallel divergence of duplicated genes that are in the same interaction network, resulting in two versions of the network expressed in a spatiotemporally distinct manner. A potential scenario in such a situation is that each network becomes specialized towards a particular role. This could occur when multiple signalling pathways are integrated by the network, with the diverged

networks becoming specialized towards particular inputs. In the case of the regulatory network underlying flowering, duplication and subsequent loss or modification of cis-regulatory elements[405–407] could result in certain copies of the floral integrators becoming more sensitive to particular inputs, such as the photoperiod, vernalization, or ageing pathways. This is particularly interesting given the regulatory divergence observed in the *BnFLC* genes in Tapidor (section 3.3), as different homologues of floral integrators may be influenced by different *BnFLC* homologues.

To determine if any of the duplicated floral integrators in *B. napus* have diverged to become more sensitive to the vernalization response, the expression of these genes was compared between Westar and Tapidor. The greatest difference was observed for *BnFT* and *BnTFL1* gene expression, with the expression of *BnFT* being consistent with *BnFLC* mediated repression as observed in Arabidopsis[30,31]. The regulation of *BnAP1* and *BnFD* homologues are also altered in the winter variety. As observed at the global level, the vernalization requirement seems to delay the upregulation of many of the floral integrators. However, despite differences in timing, the expression behaviours of the majority of floral integrators in the winter variety are in agreement with the spring variety, suggesting that these genes are not responsible for the flowering time differences observed between the varieties. The differences identified, however, provide potential future avenues for dissecting the flowering response in *B. napus*.

3.4.1 A vernalization requirement delays the upregulation of floral integrators during the floral transition

At the global level, vernalization delayed the increase in expression of genes involved with flower development in the apex (section 3.2.2). As many of the floral integrators increased in expression during the floral transition (section 2.4), the expression of these genes was investigated to determine if vernalization delays their upregulation also. For the *BnFT* and *BnAP1* genes, a post-cold increase is seen in the first time point sampled after the vernalization treatment in the spring variety (Figures 2.25 and 2.27), whereas the increase in the winter

variety is only seen at the final time point (Figures 3.25 and 3.28). Likewise, *BnLFY* and *BnSOC1* genes peak in expression at the day 69 time point in spring (Figures 2.32, 2.28, and 2.30) and the day 83 time point in winter (Figures 3.31, 3.29 and 3.30). Finally, the *BnFD* genes peak at day 67 in the spring (Figure 3.27) and day 72 in the winter variety (Figure 2.31). The later upregulation of floral integrators in the winter variety during the time series relative to the spring variety is consistent with the vernalization response acting to repress the floral transition.

3.4.2 Between variety regulatory divergence in all *BnFT* and *BnTFL1* genes and select homologues of *BnFD* and *BnAP1*

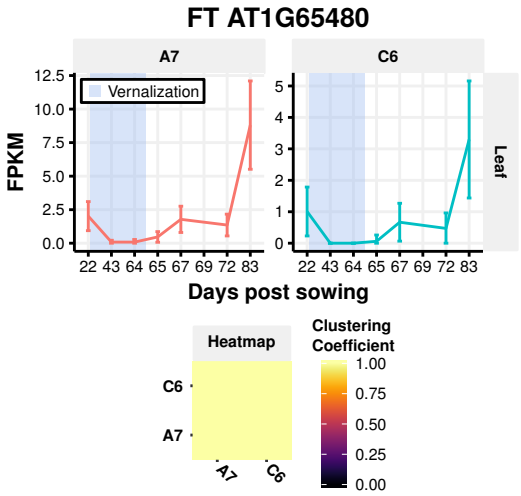


Figure 3.25: Expression traces for the *BnFT* genes in the leaf of Tapidor. The expression values and the 95% confidence intervals of those expression values as computed by Cufflinks are displayed. A heatmap of the clustering coefficients calculated by the SOM based method (discussed in section 2.2.4) quantifies the similarity between the expression profiles.

One of the ways that *FLC* acts as a floral repressor is through the repression of *FT* expression in the leaf[30,31]. To investigate how *BnFT* genes were affected by a requirement for cold, their expression was investigated. In the Westar leaf, four *BnFT* copies are expressed, exhibiting high expression before and

after the cold treatment (Figure 2.4.1). In contrast in the winter variety, the *BnFT.A2* and *BnFT.C2.Random* copy are not expressed above 2.0 FPKM at any point during the time series. In terms of the expression traces, *BnFT.A7* and *BnFT.C6* in the spring variety decrease in expression during the cold treatment, returning to pre-cold expression levels at the first time point after the cold treatment. In the winter variety, however, the expression of *BnFT.A7* and *BnFT.C6* is low initially and remains low until the final time point, at which point the genes increase in expression. The high level of *BnFT* expression before cold in the spring variety, correlates with low level of many *BnFLC* copies (Figure 3.16). Likewise, the low levels of *BnFT* before cold correlate with high levels of *BnFLC* in the winter variety. These observations are consistent with certain *BnFLC* copies maintaining their repressive effect on *BnFT* expression.

In terms of the magnitude of expression, the maximal expression level of the A7 and C6 copies of *BnFT* are six- to eight-fold lower in the winter variety, while the A2 and C2 copies are not observed above the 2.0 FPKM expression threshold. This could suggest that the requirement for cold maintains the expression of these genes at a lower level. However, it should also be noted that the lower expression in the winter variety may also result from the effect of leaf senescence impacting the expression. This is supported by the correlation analysis, that suggested the developmental stage of the plant influenced the first true leaf to a lesser extent than the apex (section 3.2.3). Regardless, that the A2 and C2 copies are not observed above 2.0 FPKM is particularly striking given that *BnFT.A2* is the copy with the highest maximal expression level in Westar. In addition, while the spring variety had low, but detectable, expression of *BnFT.A7* and *BnFT.C6* in the apex, no such expression is observed in the winter variety. Taken together, this suggests that the vernalization response has a greater effect on the expression of the A2 and C2 copies of *BnFT* than on the A7 and C6 copies, although lower expression is observed for all copies in both tissues in the winter variety.

Although *TFL1* and *FT* are very highly related structurally[57–59] their regulation is quite distinct. For example, the vernalization flowering pathway has not been found to influence the expression of *TFL1*, whereas it has for *FT*[30,31]. Despite this, copies of *BnTFL1* display large differences in regulation between the two

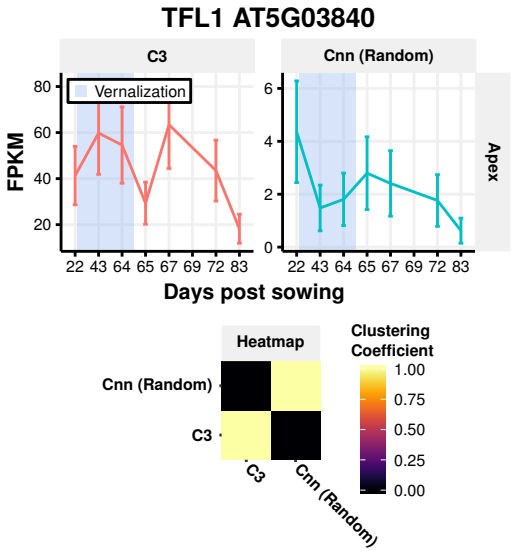


Figure 3.26: Expression traces for the *BnTFL1* genes in the apex of Tapidor. The expression values and the 95% confidence intervals of those expression values as computed by Cufflinks are displayed. A heatmap of the clustering coefficients calculated by the SOM based method (discussed in section 2.2.4) quantifies the similarity between the expression profiles, which in this case demonstrates that the two regulatory profiles have diverged.

varieties (Figures 2.33 and 3.26). Comparing regulatory patterns, *BnTFL1.C3* displays a somewhat similar expression trace in the two varieties, in that a cold-induced increase and a post-cold increase in expression is observed. However, the post-cold peak in expression occurs earlier in the winter variety at the day 67 time point, as opposed to day 69 in the spring variety. In section 2.4.6 I discuss the decrease in expression of *BnTFL1.C3* at the final time point in light of the expression of *BnLFY* and *BnAP1*, suggesting that the increase in expression of these genes results in the observed decrease in *BnTFL1.C3* expression. However, for this to be the case the expression of *BnLFY* and *BnAP1* genes in the winter variety would have to increase in expression earlier in the time series, rather than later as observed (Figure 3.31 and 3.28). Another potential explanation for the earlier peak in expression may be a result of the day 69 samples from Tapidor not being included in the sequencing. Therefore, the expression of *BnTFL1.C3* may in fact be very similar in both varieties, with the different timings of the floral transition between the two varieties having little effect on the expression of the gene. Performing qPCR on the full set of samples taken from Tapidor, as was done in Westar (Figure 2.34), would allow for this to be tested. The expression magnitude of the other *BnTFL1* copies is reduced in the winter variety relative to the spring variety, with *BnTFL1.Cnn.Random* being the only other copy expressed above 2.0 FPKM at at least one time point. In addition to being lowly expressed, *BnTFL1.Cnn.Random* lacks the peak in expression during the cold treatment in the winter compared to the spring variety. The cold response therefore reduces the expression of all copies of *BnTFL1*, and also influences the timing of regulatory changes. However, the regulatory divergence observed between the homologues is present in both varieties.

In addition to *FT*, *SOC1* is another floral integrator directly regulated by *FLC*[30,31,86]. As already discussed (section 3.4.1) the upregulation of *BnSOC1* genes post-cold treatment is delayed in the winter variety. However, an additional manner in which the *BnSOC1* genes have diverged between varieties is the expression magnitude. In both the apex (Figure 3.29) and the leaf (Figure 3.30), the maximal expression levels of the *BnSOC1* genes are two- to four-fold lower in the winter variety than in the spring. This suggests that the vernalization requirement results in suppression of *BnSOC1* expression for all

copies, while the general pattern of expression (whether the copy is expressed during cold, or increases after cold, or both) is maintained.

As both *BnFT* and *BnTFL1* genes exhibit altered expression in the winter variety, the expression of *BnFD* copies was investigated as the product of the *FD* gene interacts with both FT and TFL1 in Arabidopsis[41,47,49]. Within the apex, five of the six *BnFD* copies display similar expression patterns between Tapidor and Westar. However, the A1 copy shows markedly different regulation in the winter variety relative to the spring (Figure 3.27). *BnFD.A1* exhibits an expression pattern similar to the A8 and C7 copies, resulting in the three genes sharing a regulatory module. This is in stark contrast to the spring variety, where the A1 copy has a regulatory pattern completely distinct from the other copies (Figure 2.31). This change causes the *BnFD* genes in Tapidor to have a *gradated* pattern of regulatory module assignment, whereas in Westar they exhibited a *distinct* pattern. The magnitude of expression is also different, with *BnFD.A1* achieving the highest maximal expression value of all the other copies in Tapidor apex samples, whereas in Westar *BnFD.A1* was one of the most lowly expressed copies. In addition, whereas no *BnFD* copy was expressed above the 2.0 FPKM threshold in the leaf in Westar, the *BnFD.A1* copy was expressed in the leaf in Tapidor. All of these observations suggest that the *BnFD.A1* has a different function in the winter variety as opposed to the spring variety.

Finally, another notable difference observed between varieties was the expression of the *BnAP1* copy on A2. The *BnAP1* genes in the spring variety were divided into three regulatory modules; one displaying an increase post-cold, one showing a transient increase during vernalization, and one displaying partial behaviour of both (Figure 2.27). *BnAP1.A2.Random* was uniquely assigned to the latter module in the spring variety. In the winter variety, the post-vernalization increase of *BnAP1.A2.Random* is exaggerated, with the magnitude of expression at the end of the time series being similar to the *BnAP1.C6a* copy. The transient increase during the vernalization period is still observed in the A2 copy in the winter variety. However, as a result of the increase during vernalization being relatively slight in comparison to the increase at the final time point, the gene is assigned to the same regulatory module as the A7 and C6 copies. The *BnAP1.A2.Random* gene in Tapidor,

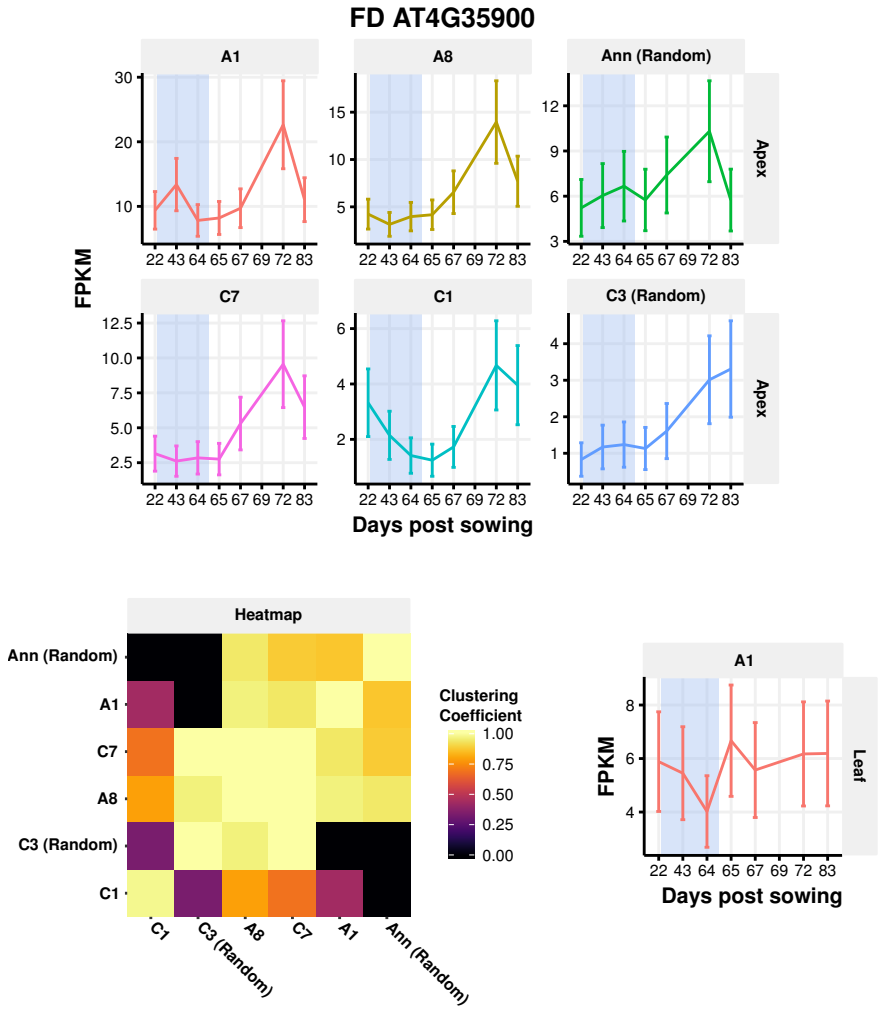


Figure 3.27: Expression traces for the *BnFD* genes in Tapidor.
The expression values and the 95% confidence intervals of those expression values as computed by Cufflinks are displayed. A heatmap of the clustering coefficients calculated by the SOM based method (discussed in section 2.2.4) quantifies the similarity between the expression profiles. Regulatory divergence between some copies is observed early in the time series, although all copies increase in expression after the vernalization treatment.

therefore, behaves more similarly to the A7 and C6 copies and to *AP1* in Arabidopsis, than it does in the spring variety.

3.4.3 Similarities in floral integrator regulation between varieties

One of the key behaviours observed in many of the expression traces of the floral integrators in Westar was an increase in expression after the vernalization treatment . All expressed *BnLFY* copies, the A7 and C6 copies of *BnFT* and *BnAP1* in the apex, five of the six *BnSOC1* and *BnFD* copies in the apex, and the C3 and A10 copies of *BnTFL1* all exhibit a post-cold increase in the spring variety (Section 2.4). To determine if this regulation is maintained in the winter variety, the expression of these copies was investigated. With the exception of *BnTFL1.A10*, these copies are expressed in the winter variety and increase in expression after the cold treatment (Figures 3.25, 3.28, 3.29, 3.27, 3.31, and 3.26). As a consequence of this similarity, many of the same genes are assigned to the same regulatory modules in Tapidor as they are in Westar. All *BnLFY* copies again have a *redundant* pattern of regulatory module assignment in both varieties (Figures 3.31 and 2.32). Likewise, the A7 and C6 copies of both *BnFT* and *BnAP1* display similar expression profiles in both varieties. Therefore, a vernalization requirement does not seem to completely abolish the upregulation of floral integrators during the floral transition. This suggests that the *BnFLC* copies that exhibit expression reactivation post-cold do not repress any of the floral integrators that display upregulation in both varieties.

SOC1 is a direct target of *FLC* in Arabidopsis[30,31]. However, homologues of this gene do not seem to be impacted by the vernalization response in Tapidor. In the apex in Westar, the *BnSOC1* genes exhibit peaks in expression during and after the vernalization treatment, with the ratio of expression magnitudes between these peaks varying between the copies (section 2.4.3). In both varieties, the C4 and A3 copies exhibit the most extreme of these ratios, with *BnSOC1.A3.Random* peaking in expression post-cold and the *BnSOC1.C4* copy peaking during the cold (Figures 2.28 and 3.29). That these observations are not altered in the winter variety suggests that the effect of a cold requirement impacts the regulation of all *BnSOC1* genes similarly.

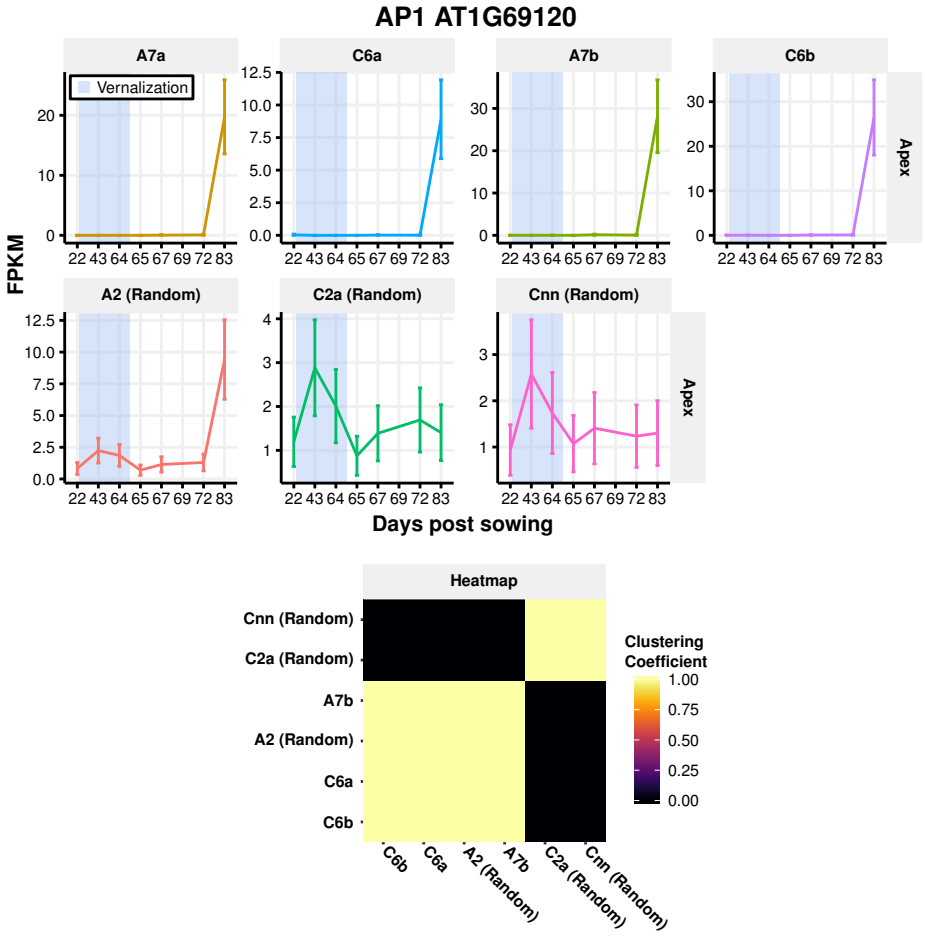


Figure 3.28: Expression traces for the *BnAP1* genes in the apex of Tapidor. The expression values and the 95% confidence intervals of those expression values as computed by Cufflinks are displayed. A heatmap of the clustering coefficients calculated by the SOM based method (discussed in section 2.2.4) quantifies the similarity between the expression profiles. The A7 and C6 copies and the A2 copy have high similarity between their expression profiles, while the C2a and Cnn copies have very low expression and do not show regulatory similarity to the other *BnAP1* copies.

Expression magnitude differences between copies are also maintained in the winter variety. Of the A7 and C6 copies of *BnAP1*, *BnAP1.C6a* has the lowest maximum expression in both varieties. *BnAP1.C2a.Random* and *BnAP1.Cnn.Random* are both lowly expressed copies that display a slight peak during vernalization in both varieties (Figure 3.28). The tissue-specific differences in *BnSOC1* expression observed in the spring variety are conserved in the winter variety, with *BnSOC1.A3.Random*, and *BnSOC1.A4.Random* and *BnSOC1.A4* being most highly expressed in the apex and leaf respectively.

It might be expected, given how genes have diverged in Arabidopsis[229], that certain homologues of floral integrators would be more vernalization responsive than others. If this was the case, one would expect the regulatory divergence between homologues in Tapidor to be greater than that observed in Westar. However, the expression of the floral integrator homologues in Tapidor reveals that this is not the case.

3.4.4 Conclusions

When regulatory or protein interaction networks are duplicated in whole genome multiplication events it has been found that the duplicated networks can diverge into distinct networks[229]. When this occurs, it is possible that the networks will diverge to be more or less sensitive to particular environmental inputs. To investigate whether this has occurred with the regulatory network underlying flowering in *B. napus*, the expression of the floral integrators was compared between varieties. The vernalization response does not result in all floral integrators exhibiting increased regulatory divergence, as might be observed if a particular set of floral integrators had increased vernalization sensitivity. Instead, the main difference between the varieties is a delay in the increase of floral activators post-cold in the winter variety, suggesting that the vernalization requirement is acting to repress the floral transition through influencing the expression of all homologues. This is in line with the findings at the global level, where vernalization was found to delay development (section 3.2.2).

Although there is not evidence for a vernalization-specific regulatory network, certain *B. napus* homologues of floral integrators do exhibit different regulatory

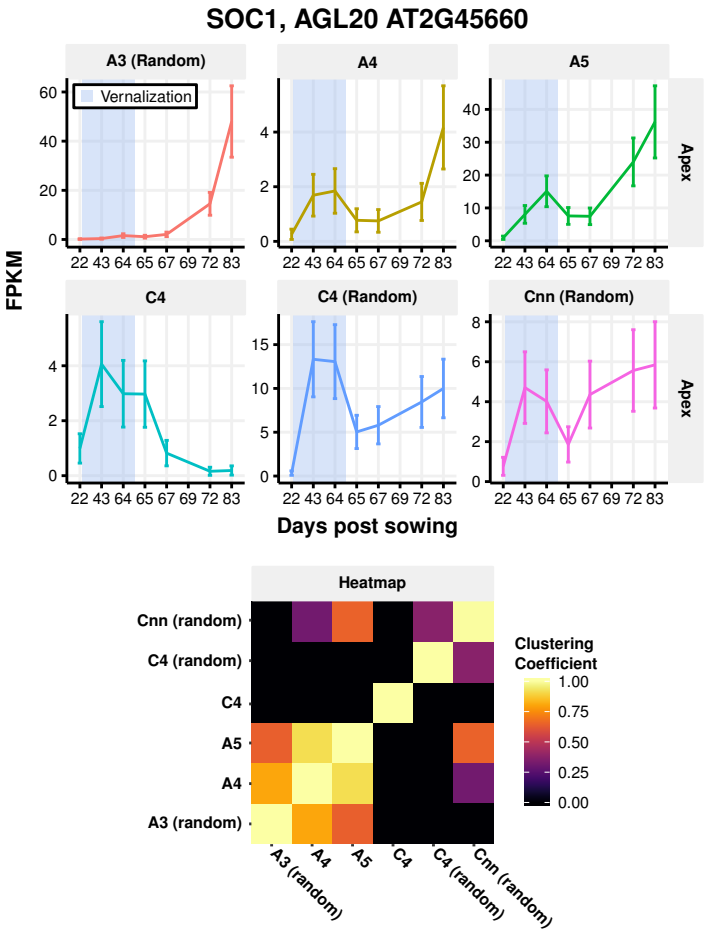


Figure 3.29: Expression traces for the *BnSOC1* genes in the apex of Tapidor. The expression values and the 95% confidence intervals of those expression values as computed by Cufflinks are displayed. A heatmap of the clustering coefficients calculated by the SOM based method (discussed in section 2.2.4) quantifies the similarity between the expression profiles. Regulatory divergence between the copies is observed, both in terms of expression pattern and the magnitude of expression.

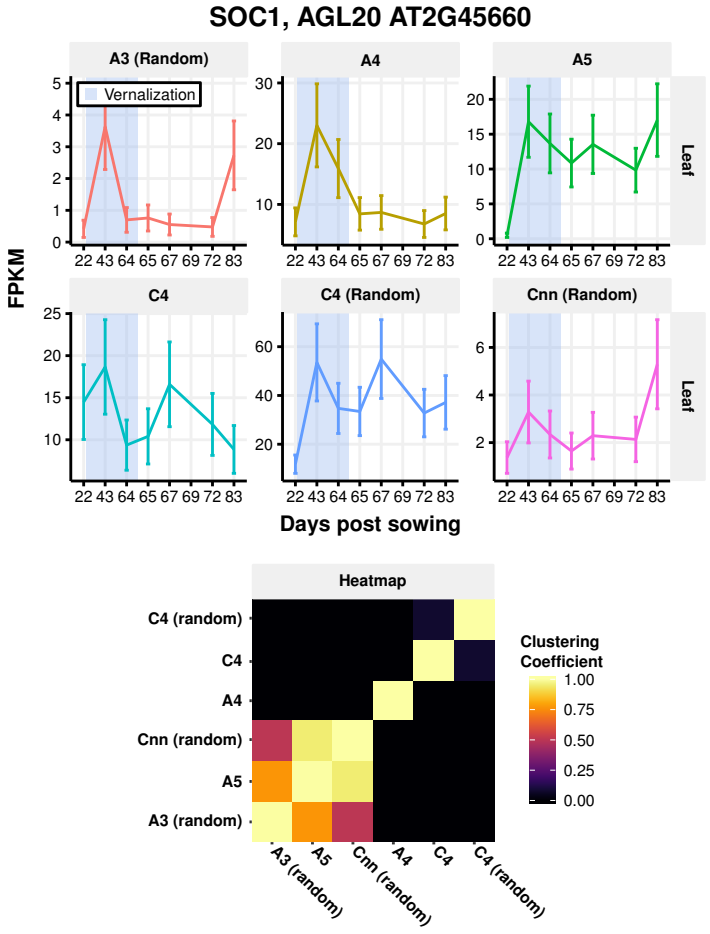


Figure 3.30: Expression traces for the *BnSOC1* genes in the leaf of Tapidor. The expression values and the 95% confidence intervals of those expression values as computed by Cufflinks are displayed. A heatmap of the clustering coefficients calculated by the SOM based method (discussed in section 2.2.4) quantifies the similarity between the expression profiles. Regulatory divergence between the copies is observed, both in terms of expression pattern and the magnitude of expression.

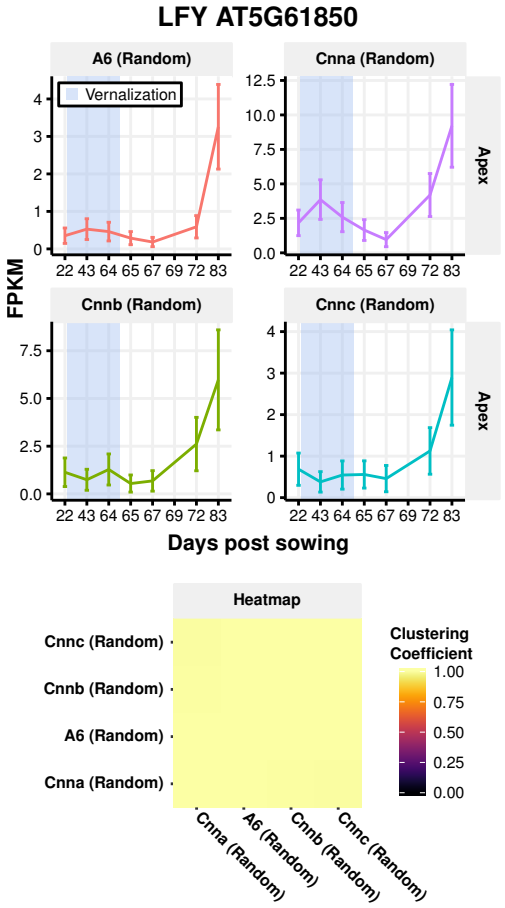


Figure 3.31: Expression traces for the *BnLFY* genes in the apex of Tapidor. The expression values and the 95% confidence intervals of those expression values as computed by Cufflinks are displayed. A heatmap of the clustering coefficients calculated by the SOM based method (discussed in section 2.2.4) quantifies the similarity between the expression profiles. All four copies exhibit very similar expression profiles.

behaviour in the winter variety. Analysis of *BnFT* copies suggests that certain copies are more vernalization sensitive than others, with the A2 and C2 copies exhibiting a more severe reduction in expression across the entire time series relative to the A7 and C6 copies. These findings contradict with results from other studies of vernalization requiring lines of *B. napus*, where the A7 and C6 copies were silenced prior to cold, while the A2 copy was expressed[154]. The response of the *BnFT* genes to vernalization, therefore, may be variety specific. That A2 and C2 are not expressed at all, even when the Tapidor plants have undergone the floral transition, potentially suggests that the A2 and C2 copies are not required for the floral transition in *B. napus*. This analysis will require further validation, as potentially the reduced expression of *BnFT* genes in the Tapidor leaf is a consequence of the leaves being older when *BnFT* expression increases in Tapidor relative to Westar. Repression is also observed for the *BnSOC1* genes. Although the copies maintain their general expression profiles, the expression magnitude of the copies is greatly reduced in the winter variety. Although the *BnTFL1* genes have not been found to have a vernalization requirement in Arabidopsis, these genes also show reduced expression in the winter compared to the spring *B. napus* variety. This may be a result of the highly interconnected nature of the floral integrators, such that the reduction in expression is indirect (Figure 1.1). Alternatively, *BnFLC* copies may directly regulate the expression of *BnTFL1* genes, representing an additional manner in which the requirement for cold can alter flowering behaviour. Finally, single copies of *BnAP1* and *BnFD* have altered expression patterns in the winter compared to the spring. In both cases, the homologues with novel regulation in the winter variety acquire expression profiles similar to other homologues. This suggests that these two genes have potentially lost regulatory elements in Westar. Without further work it is difficult to determine whether the differences observed between varieties are due to the vernalization requirement or due to between variety differences. Studies that introgress *BnFLC* genes from Tapidor to Westar would be able to discriminate between these two possibilities. Alternatively, assessing the expression of the floral integrators could be done in a larger collection of *B. napus* varieties, to determine more consistent differences between winter and spring varieties.

3.5 Discussion

A vernalization requirement is of great agronomic importance for the growth of *Brassica* crops[127] and has a large effect on the floral transition[27]. To understand the vernalization response in *B. napus*, a time series of transcriptomes were compared between a winter variety, Tapidor, and a spring variety, Westar. Comparing the number of expressed genes between varieties revealed that Tapidor had a greater number of *B. napus* genes exhibiting variety-specific expression in the leaf compared to Westar. This difference was also observed when *B. napus* genes were grouped based on sequence conservation to Arabidopsis genes, with Arabidopsis genes tending to have more expressed homologues in the leaf in Tapidor than in Westar. A potential hypothesis to explain this observation is that an increased number of proteins are required in the leaf in Tapidor. Being the organ that intercepts the majority of light, the leaf senses photoperiod signals[17,18,20–22]. Combined with the expression of *FLC* in the vasculature, and the movement of FT protein from leaves to the apex[408,409], this positions the leaf as the organ that mediates the vernalization and photoperiod response in Arabidopsis. The increased number of variety-specific genes expressed in the leaf in Tapidor could potentially represent an expansion of this sensory machinery to allow the plant to respond to vernalization.

Correlation analysis of the leaf and apex revealed that the transcriptomes develop similarly in both varieties, but the rates of change are dependent on the tissue. In the first true leaf, samples grown for the same number of days displayed the greatest similarity in terms of correlation between varieties. This was not the case in the apex, where developmentally similar samples from each variety exhibited the greatest similarity between their transcriptomes. This suggests that the leaf transcriptome is influenced by the age of the tissue, whereas the apex transcriptome is influenced by the developmental stage of the plant. This is counter to findings in Arabidopsis, where the onset of the floral transition was found to correlate strongly with the start of leaf senescence among a group of both early and late flowering accessions[410]. Unfortunately, concurrent transcriptomic analysis of apex and leaf samples are not available in order to determine whether these phenotypic observations translate to expression differences. However, analysis of apex and leaf transcriptomes

individually support the observations in *B. napus*. Transcriptome analysis of laser dissected Arabidopsis meristems identified a set of genes, enriched for roles in floral development, that are upregulated during long days[411]. The expression of these genes correlated with commitment of the apex to flower. Conversely, analysis of the Arabidopsis leaf transcriptome from early growth stages to senescence revealed that diverse biological processes were more likely to have have correlated expression during senescence than during early development[266]. The authors concluded that this was due to the transcriptional changes during leaf senescence being tightly coordinated to maximise the remobilization of resources from leaves to developing tissues. A potential explanation for why the transcriptomes of the leaf samples remain synchronized, despite the plants being at different developmental stages, is due to artificial selection for regular leaf senescence. As both varieties used are oilseed rape varieties, remobilization of resources from old leaves may have been selected for. This might be especially relevant for oilseed crops, where the formation of the pod canopy blocks light to older leaves.

Investigating the expression of *B. napus* homologues of vernalization pathway genes implicates certain copies of *BnFLC* as mediating the vernalization response in Tapidor. During the cold, the expression of *FLC* in vernalization requiring lines decreases, whereas in Arabidopsis spring accessions the expression of *FLC* is low throughout development[29]. Two *BnFLC* copies were found that were lowly expressed in Westar and became stably repressed in Tapidor during cold; the A10 and A2 copies. This finding confirms results from association studies, that found regions containing these genes to be associated with flowering time. Using a *B. napus* Doubled Haploid mapping population between Ningyou7, a Chinese semi-winter variety, with a slight vernalization response, and Tapidor (TNDH population), a region on A10 was associated with flowering time variation in unvernalized conditions[141,143]. As this region was not associated with flowering time variation when the plants were vernalized, it led the authors to propose *BnFLC.A10* as the copy conferring a vernalization requirement in *B. napus*[141]. The A2 copy has also been found to be associated with flowering time in *B. napus* and *B. rapa*[132–135,365]. Interestingly, the effects of A10 and A2 on flowering were found to be additive in *B. napus*, suggesting that both copies are delaying the floral transition to

some extent[137]. The other *BnFLC* copy identified in the TNDH population as being associated with flowering is A3b[141]. In the transcriptome time series, *BnFLC.A3b* is expressed approximately four-fold lower in Westar relative to Tapidor before cold, and displays a cold-induced decrease in Tapidor. However, the repression of the gene is not stable, and reactivation of expression is observed. Expression reactivation is also observed post-cold in *BnFLC.C2*, while stable repression is observed in Westar. Reactivation of *FLC* expression is observed in Arabidopsis when vernalization sensitive lines are not given adequate vernalization[245,345,346]. This suggests that these particular copies have not received adequate cold in order to become fully repressed. These copies do not, therefore, need to be fully repressed in the apex for the plants to flower. This is consistent with findings from the TNDH mapping population, where *BnFLC.A3b* was detected in both vernalized and unvernalized conditions[141]. Another association study utilizing a mapping population created using two spring lines (Skipton/Ag-Spectrum DH), that nonetheless exhibited slight vernalization responses, identified a region containing *BnFLC.C2* as being associated with flowering time[141,363,364]. That these *BnFLC* are associated with flowering time in unvernalized growth conditions, and with a mapping population of two spring parents, suggests that the A3b and C2 copies do not confer a vernalization requirement, and may instead modulate the response to cold. This is also in line with results from *B. rapa*[137] and *B. oleracea*[140], that also implicated A3 and C2 homologues of *FLC* with flowering time. Despite being a spring variety, Westar has been found to respond to a vernalization treatment with accelerated flowering[241]. Two *BnFLC* genes have high expression in Westar and exhibit stable, cold-induced repression; *BnFLC.A3a* and the aforementioned *BnFLC.C2*. In addition to a region containing *BnFLC.C2*, a region containing *BnFLC.A3a* was also associated with flowering in the Skipton/Ag-Spectrum DH mapping population[141,363,364]. It therefore seems likely that *BnFLC.A3a* and *BnFLC.C2* confer a weak vernalization response in Westar. *BnFLC.A3a* is expressed at a similar level in both Tapidor and Westar. This suggests that the delay to flowering in Tapidor resulting from the expression of *BnFLC.A3a* could be epistatic to the delay conferred by other copies, such as the copies on chromosomes A2 or A10. Finally, although divergence is observed in other vernalization pathway genes, the significance of the differences is difficult to judge based on our current mechanistic understanding.

Increased expression of *VIN3* and *MSI1* homologues in Tapidor compared to Westar may allow Tapidor to respond more dynamically to cold, or alternatively may be required to repress the higher levels of *BnFLC*. This is supported by findings from *Arabidopsis arenosa* where higher expression of *VIN3* during cold was observed in vernalization-requiring accessions relative to a rapid-cycling accession[232]. For *BnFRI* genes, the lack of *BnFRI* expression in the spring variety could potentially explain the reduced expression of *BnFLC* genes in Westar. However, this would require validating, especially as previous work on *FRI* homologues in *Brassicas* have not found *BnFRI.A10* to be associated with flowering time[142,143,147].

That *BnFLC.C2* exhibits reactivation in the winter and not the spring is interesting given findings from Arabidopsis. In Arabidopsis the pre-vernalization expression level of *FLC* was found to not correlate with the vernalization response for different Arabidopsis accessions[245,412]. Instead, variation in the efficiency of *FLC* silencing accounted for the observed natural variation in vernalization response[245]. For *BnFLC.C2*, differences in both the initial expression value of the gene and the extent of silencing are present between varieties. Tapidor has higher expression of the gene initially, and although the vernalization treatment causes a decrease in expression, the silencing is not stable. The gene in Westar, however, is more lowly expressed initially and becomes stability expressed after cold treatment. The reactivation of the copy in Tapidor parallels the reactivation of *FLC* in a Swedish variety of Arabidopsis, Lov-1. The *FLC* in this accession required 9 weeks of cold at 5 °C to become fully silenced, as opposed to 4 weeks for a common laboratory strain, Col-FRI[345,346]. This difference was found to be an adaptive response, with the Lov-1 copy having a different optimal vernalization temperature than Col-FRI[346]. Applying this to the differences in expression of *BnFLC.C2* between varieties poses two hypotheses. The first is that the basal level of *BnFLC.C2* silencing in Westar is higher, resulting in a shorter vernalization period being required for stable silencing of the gene. Alternatively, the optimum temperature at which the *BnFLC.C2* copy is repressed might be different.

Therefore, it seems that the *BnFLC* genes have diverged to either require different lengths of cold to become stably silenced, or have different optimal temperatures at which silencing occurs. Having multiple copies of *FLC* with

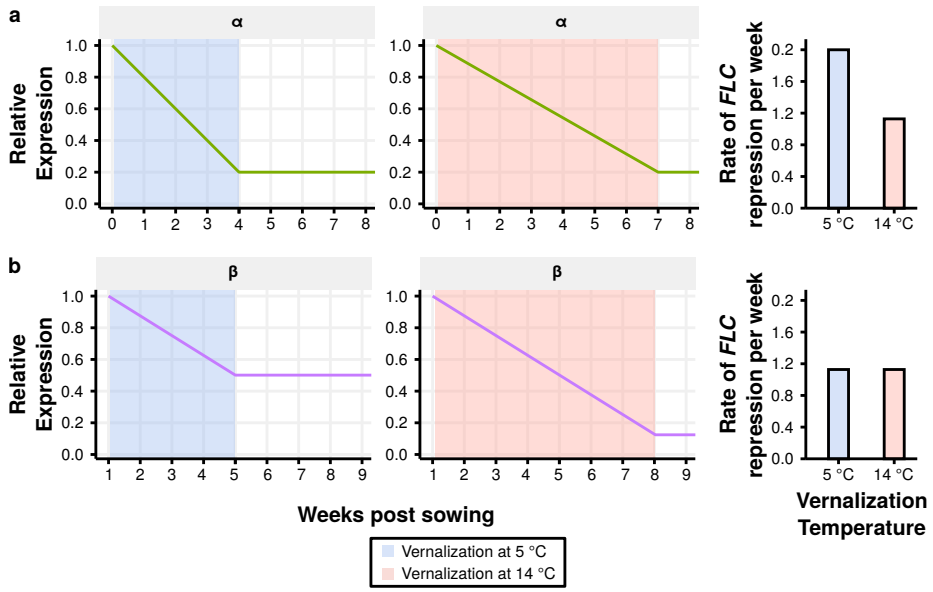


Figure 3.32: The "underdetermined system" hypothesis.
Different *FLC* copies (α and β) with different sensitivities to cold allow for both the length and the severity of cold to be determined. **a** *FLC*-α is repressed more strongly at 5 °C relative to 14 °C. This difference results in the expression level of *FLC*-α being the same after both a four week vernalization period at 5 °C and a seven week vernalization period at 14 °C. **b** *FLC*-β repression occurs at the same rate at both 5 °C and 14 °C. This results in the expression level of the gene being different at the end of a four week vernalization period at 5 °C compared to the expression level after a seven week vernalization period at 14 °C.

potentially different requirements for cold has interesting implications for the vernalization pathway in *B. napus*, relative to Arabidopsis. From experiments in Arabidopsis, vernalization has been shown to be a quantitative response; the more cold experienced, the more *FLC* expression is repressed[29,245]. In addition, the severity of cold influences the rate of vernalization, where severity of cold is used to refer to the temperature used for the vernalization treatment. In thorough experiments using five different Arabidopsis accessions, different lengths of cold, and different vernalization temperatures, Duncan et al. (2015) revealed that the efficacy of vernalization was dependent on all three factors; genotype, length of cold, and severity of cold[346]. The interaction between the length of cold and the severity of cold leads to a hypothesis for the retention of *FLC* copies in *B. napus*, and their apparent divergence in vernalization response; the "underdetermined system" hypothesis. This hypothesis comes directly from observations of *BnFLC* regulatory behaviour presented in this work. The central idea that this hypothesis puts forward is that additional copies of *FLC* could allow regulatory responses to respond separately to the length of cold and to the vernalization temperature. Consider the level of *FLC* expression (FLC_t) after a vernalization period of length t as a function of the initial *FLC* expression level (FLC_0), the length of vernalization, and the rate of *FLC* repression ($f(T)$), which is itself a function of temperature (T). Assuming *FLC* expression decreases in a linear fashion at a constant temperature, FLC_t can be expressed as:

$$FLC_t = FLC_0 - tf(T)$$

Assuming that the initial level of *FLC* is the same for all plants of the same genotype, the level of *FLC* expression after cold is dependent solely on the length and the severity of cold. Therefore, with only a single *FLC* locus, a plant is not able to distinguish between a long, mild period of cold and a short, severe period of cold (Figure 3.32a). This is the case in the Var2-6 Arabidopsis accession, where 6 weeks of cold at 5 °C and 12 weeks of cold at 14 °C resulted in the plants flowering at approximately the same time[346]. As there is only one equation, and two unknowns, the system is underdetermined. The presence of an additional *FLC* copy, with a different sensitivity to the severity of cold (represented by the rate of *FLC* repression having a different relationship with

temperature; $g(T)$), would provide an additional equation, allowing both the length of cold and the temperature experienced to be determined from the expression levels of both *FLC* copies (Figure 3.32a). This hypothesis assumes that *FLC* repression is the sole mediator of the vernalization response, which it is not[350], and that *BnFLC* copies have different molecular activities, such as different target genes, in order to enact different transcriptional programmes. It also assumes that it would be beneficial to the plant to respond to the length and severity of cold separately. However, as genes in the vernalization response have been found to have pleiotropic effects, such as on plant architecture[413], this seems likely. Regardless it demonstrates a potential use for additional *FLC* copies; to allow the length of cold and the severity of cold to be dissected. Testing the "underdetermined system" hypothesis could be done by performing similar vernalization experiments as Duncan et al. (2015) with *B. napus*[346].

Taking the expression of *BnFLC* genes (section 3.3.1) and the floral integrators (sections 2.4 and 3.4) together, the effects of a vernalization requirement on the transcription of floral integrators can begin to be dissected. Two genes directly repressed by *FLC* in Arabidopsis are *SOC1* and *FT*[85,308]. In line with this, both of these genes exhibit lower expression in Tapidor relative to Westar. Not only is the magnitude of expression lower in Tapidor, but the regulatory profiles vary also. In Westar, the expression of the *BnFT* genes were initially high at the first time point, suggesting that the plants were competent to flower, but had not yet undergone the floral transition. However, the *BnFT* copies expressed in Tapidor only increased at the final time point. Therefore, it seems unlikely that the *BnFLC.A3a* gene, that is expressed similarly in both Tapidor and Westar, influences the expression of *BnFT* genes. The expression of *SOC1* in Arabidopsis is directly repressed by *FLC* expression, particularly in the apex[30,31,86]. In addition to the vernalization pathway, the expression of *SOC1* is also regulated by the photoperiod pathway[20,84]. The interaction of the vernalization and photoperiod pathways on the expression of *SOC1* was found to be additive[307]. In a transcriptomic analysis of the Arabidopsis apex, the upregulation of *SOC1* during the floral transition was found to occur in the presence and absence of *FLC*. However, the overall expression of *SOC1* was much lower in lines containing an active *FLC* allele[307]. This same additive interaction is observed for all *BnSOC1* genes in both the

apex and the leaf in *B. napus*. Finally, in the same manner as *BnFT* and *BnSOC1*, the expression levels of *BnTFL1* were lower in Tapidor than in Westar. This is interesting given the relatedness of *BnFT* and *BnTFL1*[57–59], despite *TFL1* not previously being implicated as a direct *FLC* target. Indeed the *TFL1* gene is not identified when the binding of *FLC* is assessed in a genome-wide manner[99,414]. This therefore suggests that the *BnTFL1* genes may be downregulated indirectly by *BnFLC* genes. In addition, as opposed to the post-cold upregulation of *BnTFL1.C3* being delayed in the winter variety, as was seen consistently with other genes exhibiting such regulatory behaviour, it occurred days before the spring variety. An explanation for this difference is the sampling intervals used to generate the developmental time series. Potentially the dynamics are similar between the winter and the spring, and are missed in the transcriptomic time series due to the time period between sampling dates changing. A more biologically relevant explanation is due to the role of *TFL1* as a repressor of floral development in the shoot meristem in Arabidopsis[50]. The earlier upregulation in the winter variety may therefore occur to maintain the indeterminate nature of the shoot apex, as the plants were not then sufficiently induced to flower.

A study conducted in Arabidopsis found that when genes that interact in a regulatory manner are duplicated, the expression of the genes tends to diverge and form distinct regulatory networks[229]. It is then possible that each of these networks becomes specialized towards particular roles. In the case of *B. napus*, multiple copies of floral integrators may have resulted in multiple parallel regulatory networks forming and becoming specialized to particular inputs or locations. However, in general the vernalization requirement in Tapidor seems to influence the expression of all copies of a floral integrator. Although exceptions exist in the A1 copy of *BnFD* and the A2 copy of *BnAP1*, it is difficult to determine whether this represents a difference due to a vernalization requirement or a difference due to varietal divergence. Testing this would require analysing the expression of these potential vernalization sensitive homologues in a larger panel of *B. napus* lines.

By comparing the expression of *BnFLC* homologues between varieties and between tissues, biologically relevant differences were identified. These results highlight the benefits of being able to make these kinds of expression profile

comparisons. The next chapter will introduce a tool developed to allow such comparisons to be quickly and easily made. The web application, dubbed the Oilseed Rape Developmental Expression Resource, allows the vast dataset collected in this study to be searched and plotted to facilitate comparisons between genes and homologues.

Chapter 4

Data dissemination using a web based application

4.1 Introduction

Genome-wide expression analysis has been a key tool in the "-omics" era of science, facilitating top-down approaches to identify candidate genes and understanding developmental processes[415]. Microarrays were the initial method used to assess genome-wide gene expression[416]. This technology quantified gene expression through hybridization of fluorescent labelled transcripts to pre-designed probes, printed onto a slide. In recent years, RNA-Seq has largely replaced microarrays as the standard for conducting transcriptomic analysis[417]. RNA-Seq has many advantages over microarrays due to a higher detection sensitivity and a broader dynamic range[418]. In addition, as probes do not need to be designed, RNA-Seq does not require prior knowledge of the sample. This makes it an ideal tool for the investigation of non-model systems[419]. For example, before a genome sequence was available for *B. napus*, RNA-Seq was used across a population of *B. napus* varieties to identify genes whose expression correlated with glucosinolate content of the seed[330]. Due to the breadth of data generated during a transcriptomic study, an important consideration for RNA-Seq studies is making the data available for other researchers to use. Doing so facilitates meta-analysis[420], and is particularly relevant for large datasets

that have the potential to provide insights beyond the original motivation for collecting the data.

Repositories exist for expression data[421–423] allowing data to be downloaded and analysed by others. However, this requires a certain level of technical skill, providing a barrier to entry that slows efforts to investigate genes of interest. Alternatively, large scale repositories and tools are available that process the data and are able to visualize many different experiments and experimental designs[424–427]. These tools facilitate meta-analysis of many disparate datasets, although as a consequence the visualizations are often simplified. Other projects are much more focussed in their scope, providing a frontend to a single particular dataset. The "Electronic Fluorescent Pictograph" browser displays microarray data from a variety of Arabidopsis organs at many developmental stages[415] as a pictorial heatmap[428]. This provides a very intuitive method of interrogating this large dataset, albeit at the cost of flexibility in terms of the types of dataset that can be visualized in this way. For *Brassica* crops, although centralized respositories exist, none currently support the submission and visualization of gene expression data. The Brassica database, BRAD, is a repository of genetic data for *Brassica* crops[429], while synteny and gene homology data is available as part of the EnsemblPlants database[321]. In addition, trait and genotype data can be submitted to the Brassica Information Portal, facilitating programmatic access to this data and enabling meta-analyses to be conducted[430]. As a consequence, no resource or service is currently suitable for the appropriate visualization of time series expression data for *B. napus*.

To address this need, the Oilseed Rape Developmental Expression Resource (ORDER) was developed to allow the transcriptomic time series dataset to be queried and visualized in an intuitive manner. An extensible database structure was employed to allow future studies to be easily integrated into the website. Querying the database using Arabidopsis gene identifiers identifies all *B. napus* genes exhibiting sequence similarity, allowing the expression of homologues to be compared. In order to plot the expression profiles of *B. napus* genes that lack sequence conservation to an annotated Arabidopsis gene, a sequence based search function is also available. To demonstrate the utility of the website, two use cases are discussed. The first uses the adaptive plotting functions available to compare the expression of *B. napus* homologues of *AGL24* and

AP1, identifying expression traces consistent with an antagonistic regulatory relationship between the genes. The second uses the sequence similarity based search function to investigate microRNA expression during the time series. The functionality of this web-based application was written to be as reusable as possible, and could therefore be easily incorporated into other tools.

4.2 Website structure and user interface

The success of any web-based application is dependent on how data is stored and retrieved on the server, and how users interface with that data on their devices. If the underlying data is stored inefficiently or in a convoluted manner the website is difficult to maintain, while an unintuitive interface leads to users not being able to use the service effectively. ORDER was designed as a community resource with the primary objective of allowing users to quickly and easily search the *B. napus* transcriptomic time series to study expression dynamics of their genes of interest. To increase the potential impact on the community, a secondary objective was to make the website easily extensible, to allow data from future studies to be incorporated with minimal code changes. To achieve these goals, the database structure was carefully chosen to allow the data to be efficiently searched and subsets taken. The website functionality was implemented to provide access to the entire dataset and to make it as user-friendly as possible to search for relevant genes.

4.2.1 Database structure

How the data is stored affects the efficiency with which it can be searched and processed. The database software stores the transcriptome time series information with each gene as a single contained object (Figure 4.1). This object includes basic information, such as the Cufflinks[250] assigned gene name, which chromosome the gene is on, and where on that chromosome the gene is. A list of gene expression measurements is also associated with each gene. Each measurement within this list comprises an individual time point in the time series. The time points contain information on the gene expression value

Figure 4.1: Schematic of how the database is structured.
On the left of the figure is a single entry in the database, with one entry present for each *B. napus* gene. This is the entry for a *B. napus* gene that shows sequence conservation with *FLC*. As each measurement of gene expression contains metadata, the database can be easily extended with information from additional time points, tissues, and accessions.

and associated metadata, such as the size of the confidence interval for the expression value, the time point at which that value was measured, and the *B. napus* variety and tissue from which the sample was taken. Structuring the measurements as such allows the website to be extensible, as additional measurements can be added to the list and annotated with applicable metadata without having to change measurements already in the list. The final component of a gene entry is the homology information. This is precomputed for each gene using sequence conservation (section 6.3). The Arabidopsis Genome Initiative (AGI) identifier and the gene symbol information allow users to search for *B. napus* genes. As many *B. napus* genes are reported in terms of the Arabidopsis gene to which they exhibit sequence conservation[145,152,153] this seems a reasonable method by which to search for relevant *B. napus* genes. The Highest Scoring Pair (HSP) information is used to rank which Arabidopsis genes have the highest sequence conservation to the *B. napus* gene. The flexibility of this database structure allows for additional gene expression data to be easily added to entries in the database, making the data storage easy to manage and extensible.

4.2.2 Website functionality

An important aspect of any large dataset is how to focus analysis to areas of interest. Therefore, providing methods for users to search the database is essential. In addition to pages introducing the dataset and describing how to use the search functions of the website, there are three pages that allow users to explore the dataset; a page for searching using sequence similarity to Arabidopsis genes, a page for searching using sequence similarity to a user submitted sequence, and a page displaying a table of the genomic locations of the identified genes and additional sequence similarity information.

The Search page (Figure 4.2) allows users to search using sequence similarity to Arabidopsis genes, and displays the expression values over time for the selected genes. *B. napus* genes showing homology to the selected Arabidopsis genes are displayed below the search box as a checklist. Clicking on a *B. napus* gene causes its developmental expression trace to be plotted automatically. Additionally, hovering over the each gene in the checklist displays the chromosome the gene

Figure 4.2: Screenshot of the Search page.

The search page allows for Arabidopsis gene identifiers and names to be used to search the transcriptome time series dataset. *B. napus* genes that share sequence conservation to the Arabidopsis gene are displayed in the bar on the right. Selecting a particular *B. napus* gene plots the expression profile in all tissues and varieties.

is located on. Generated plots can be manipulated to facilitate comparisons and provide visual clarity. Selecting the checkbox to flip the facet labels will plot the four graphs with the varieties as the rows and the tissues as columns. This allows more meaningful comparisons between the two varieties when investigating the timing of expression changes during development. Plotting expression traces for many homologues simultaneously on the graph can reduce the clarity of the plot. To mitigate this, the drawing of error bars can be toggled and hovering over gene names in the plot legend highlights the expression trace of that gene in the graph. The interval of time plotted can be controlled with the slider located under the search box, to generate plots focused on a particular period of development. Finally, the generated plot image, the cDNA sequences of the selected genes, and the raw expression levels can all be downloaded from this page.

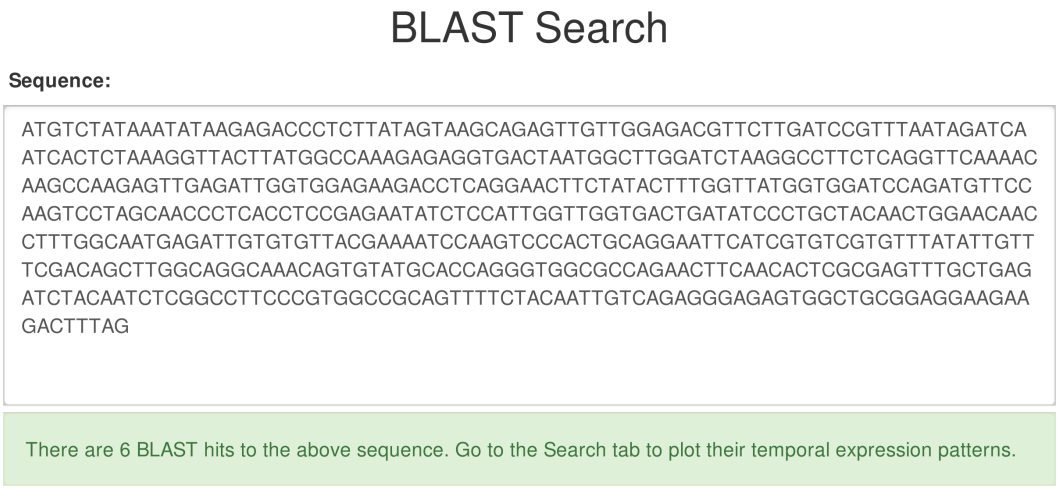


Figure 4.3: Screenshot of the BLAST Search page.

Inserting a nucleotide sequence into the search box prompts the server to perform a search for *B. napus* genes that exhibit sequence conservation. The result of the search is displayed on the sequence search page, and the identified *B. napus* genes are displayed on the Search page to allow users to plot the relevant expression profiles.

49% of the 155,240 gene models identified in the dataset do not show suitable homology to an Arabidopsis gene. In order to allow these genes to be searched, ORDER contains a search tool that uses the BLAST algorithm to identify

B. napus genes displaying sequence conservation to user submitted sequence (Figure 4.3). The number of *Brassica napus* genes found is displayed on the BLAST Search page (Figure 4.3). In order to plot the expression patterns of the discovered group of genes, the user returns to the Search page and selects the checkboxes corresponding to the identified genes. This search function allows users to access the entire dataset agnostic to whether the gene or sequence of interest is found in the Arabidopsis genome.

Determining the genomic location of *B. napus* genes is important in order to compare results to other work, such as association studies. In order to compare the results identified using ORDER and previous publications, it is therefore important to allow users to determine where in the genome their genes of interest are located. To facilitate this, ORDER generates an information table for the genes which are selected on the Search page (Figure 4.4). This table contains the chromosome on which the genes are located as well as their start and end positions on that chromosome. The Arabidopsis gene to which the selected *B. napus* gene shows homology is also displayed, along with the percentage sequence identity, score and length of the sequence identified by the BLAST algorithm as being similar between the two genes. In addition, other Arabidopsis genes identified as having similarity to the selected *B. napus* gene by the BLAST algorithm can be viewed. The colour of the rows in the sub-table correspond to the selected Arabidopsis gene on the Search page. If the selected Arabidopsis gene matches the gene in that row of the table exactly, or is a slice isoform of that gene, then the row will be coloured green or orange respectively. This colouration is also used on the Search page, to help determine the genes most likely to be homologues of the Arabidopsis gene entered in the search box. Other community resources are integrated on this page. The *B. napus* gene name is a hyperlink that takes the user to the position of the gene in a genome browser of the *B. napus* genome[118], while the Arabidopsis AGI identifier takes the user to the gene's entry on The Arabidopsis Information Resource (TAIR)[431].

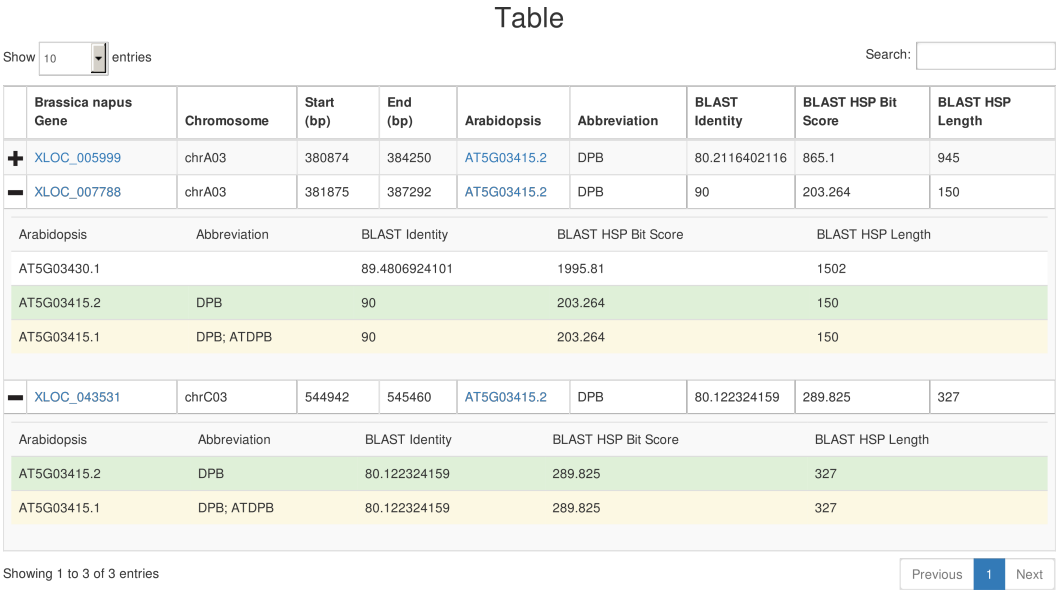


Figure 4.4: Screenshot of the Table page.

Selecting *B. napus* genes on the Search page creates a row in the table on this page. Displayed on each row is the Cufflinks[256] assigned gene name, the chromosome and chromosome position where the gene is located, details about the Arabidopsis gene to which the *B. napus* gene exhibits sequence conservation, and details about the degree of sequence conservation information. Additional sequence similarity information can be accessed by clicking the + symbol on the left of the table. Due to the many-to-many mapping of *B. napus* genes to Arabidopsis genes, a colour code is used. In this case, the user has searched for *B. napus* genes exhibiting homology to the Arabidopsis gene *DPB*. The *B. napus* gene XLOC_043531 shows highest sequence conservation to *DPB*, and is coloured green (Figure 4.2). XLOC_007788, however, shows greatest sequence similarity to the Arabidopsis gene AT5G03430, rather than *DPB*, and is coloured white. Genes that are coloured yellow (Figure 4.2) display greatest similarity to the gene searched for, although to a different splice isoform than the one the user searched for.

4.2.3 Website implementation

The website makes use of the Bootstrap framework for the user interface. The Bootstrap framework provides a clean, clear interface that is suitable for different devices. As a result, ORDER is equally usable on computers and tablets. Much of the responsive elements of the website utilize Javascript with jQuery, with the plotting making use of the D3.js library. ORDER is hosted on a CentOS (version 7.1.1503) server with Apache (version 2.4.6) as the web server used. The database used is MongoDB (version 2.6.11) with the server code written in Python (version 2.7.5), making use of the Flask web development framework.

4.3 Use cases

To demonstrate the utility of ORDER for exploring the transcriptomic time series, two examples of using the website will be outlined. The first uses the Arabidopsis homology based search function to compare the expression of *B. napus AGL24* and *AP1* homologues, identifying expression profiles consistent with the repression of *AGL24* by *AP1*. The second investigates the expression of precursors for the age-related flowering pathway microRNAs, which have to be identified using the sequence conservation based search. The graphs of gene expression profiles are downloaded directly from ORDER, and therefore accurately represent the visualizations available on the resource.

4.3.1 Regulatory interactions between floral integrators

The ability to plot the expression profiles of multiple genes simultaneously facilitates similar analysis as that conducted in section 2.4. A floral integrator not discussed in detail in that section was *AGL24*. *AGL24* is expressed in the vegetative meristem and promotes the floral transition, with mutants lacking *AGL24* displaying delayed flowering and overexpression of the gene causing earlier flowering[432,433]. Plants overexpressing *AGL24* also display a partial reversion of floral meristems into inflorescence shoots, suggesting that the gene helps to maintain the meristem in an inflorescent state[303]. Therefore, although

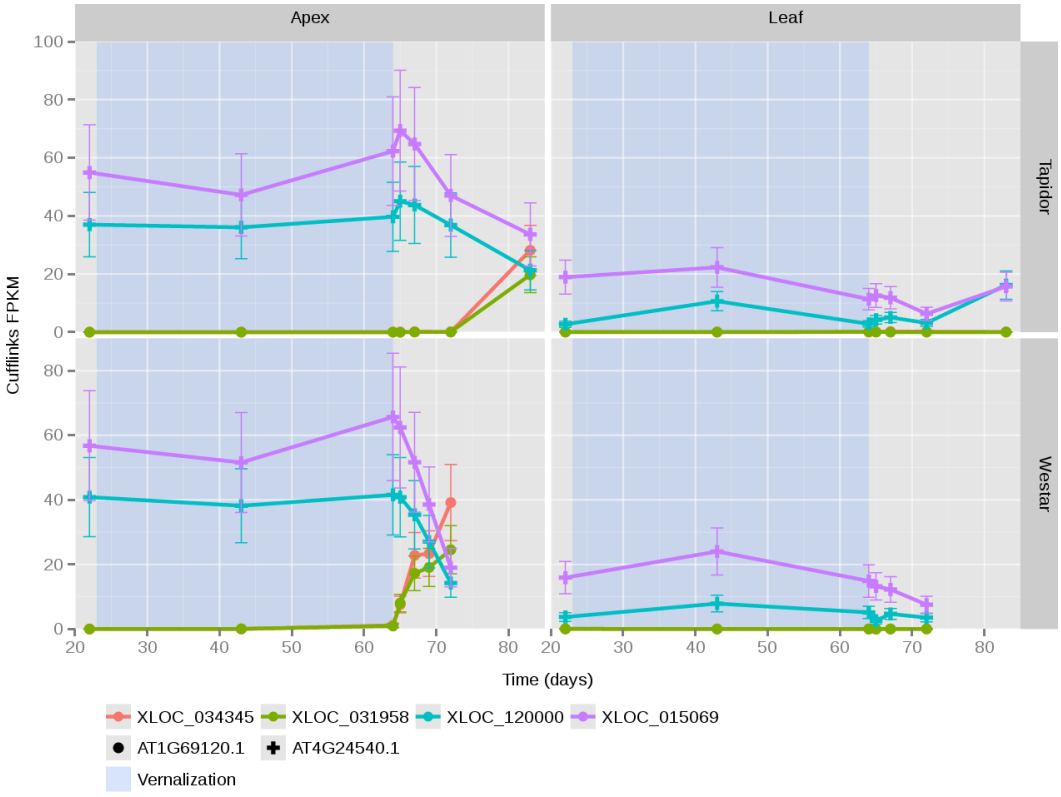


Figure 4.5: Expression profiles of *BnAGL24* and *BnAP1* genes reveals potential repression.

The expression values and the 95 % confidence intervals of those expression values as computed by Cufflinks are displayed. The expression profiles of *B. napus* homologues of *ALG24* (AT4G24540.1) and *AP1* (AT1G69120.1) are plotted. In this figure, the tissue and variety divisions have been swapped relative to figure 4.2 using the plotting controls. Plotting the figure in this manner allows for the timing of the expression changes to be more easily compared between varieties. In the apex the expression of *BnAGL24* genes (XLOC_015069 and XLOC_120000) decreases after the cold treatment, with the expression of *BnAP1* genes (XLOC_034345 and XLOC_031958) increasing.

the gene initially promotes the floral transition, expression of the gene has to be downregulated as the flower develops to prevent floral reversion[303]. This repression is mediated directly by AP1[81,82,303].

To determine whether such repression is observed in the trancriptomic time series, ORDER was used to plot the expression profiles of *B. napus* homologues of *AGL24* and *AP1* (Figure 4.5). As previously discussed (sections 2.4.2 and 3.4.1), four copies of *BnAP1* become upregulated during the floral transition in the apex. When plotted simultaneously, the increasing expression of *BnAP1* genes is concurrent with the decrease in expression of two *BnAGL24* genes in the apex of both varieties (Figure 4.5). Although purely correlative, these expression profiles are consistent with the repression of *BnAGL24* homologues by BnAP1, as findings from Arabidopsis would suggest[81,82,303]. That the expression level of the *BnAGL24* genes begins to decrease before *BnAP1* genes begin to increase suggests that other proteins may also be playing a role in the repression of *BnAGL24* in *B. napus*. Comparing between the two varieties, a delay in the timing of the expression changes is observed in Tapidor, as was observed for all of the floral integrators previously discussed (section 3.4.1).

4.3.2 Expression profiles of microRNA precursors

The age-dependent flowering pathway in Arabidopsis is mediated by microRNAs (miRNAs)[39,434]. The *miR156* and *miR172* families of miRNAs in Arabidopsis have contrasting expression patterns in that *miR156* family miRNAs are expressed highly at the beginning of development and decrease in expression as the plant ages, while the *miR172* family miRNAs are lowly expressed initially and increase during development[40]. To understand whether similar miRNA species could regulate a similar ageing pathway in *B. napus*, the expression profiles of the two families were plotted using ORDER. The Arabidopsis AGI identifiers for these miRNAs did not yield a hit in the database, which meant that an approach such as that taken for the *AGL24* and *AP1* homologues above could not be taken. MicroRNAs are 18 - 24 nucleotides in length, but these sequences are derived from longer precursor sequences that form step-loop structures before being processed to form miRNAs[40]. When the stem-loop precursor sequences of *miR156a* and *miR172a*[435–440], representative members

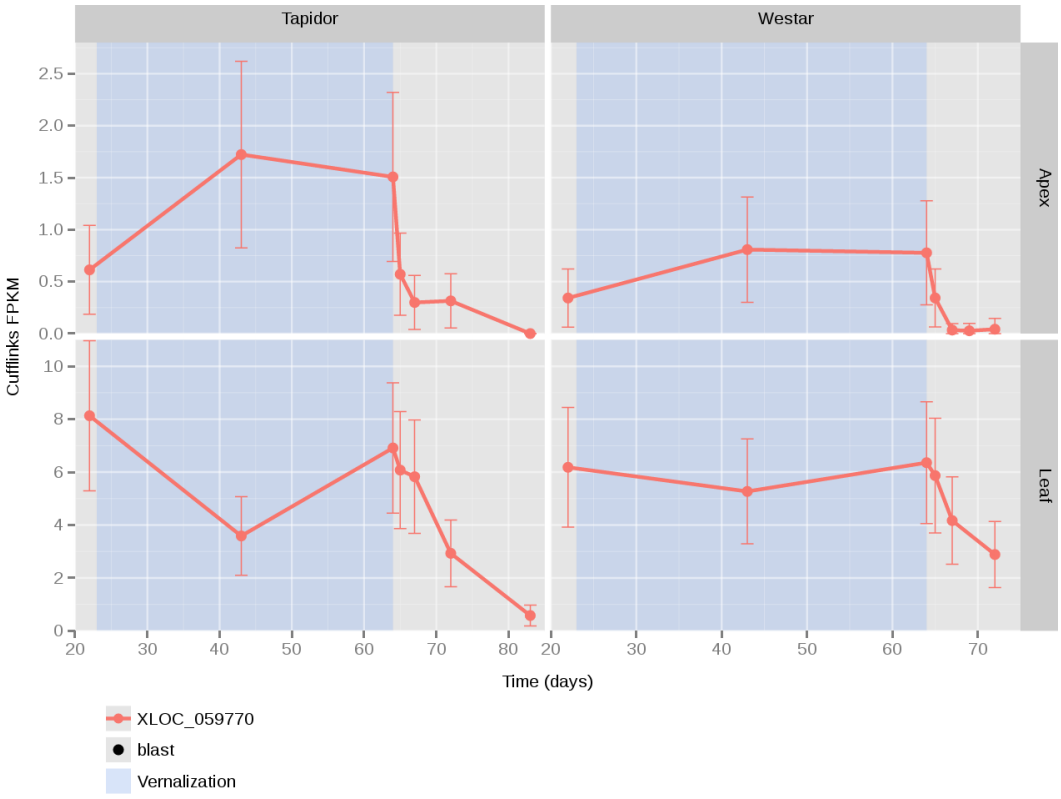


Figure 4.6: Expression patterns of the most highly expressed *B. napus* gene showing sequence similarity to the Arabidopsis *miR156* precursor
The expression values and the 95 % confidence intervals of those expression values as computed by Cufflinks are displayed. Expression in the leaf is relatively high before in both varieties, but decreases after the cold treatment.

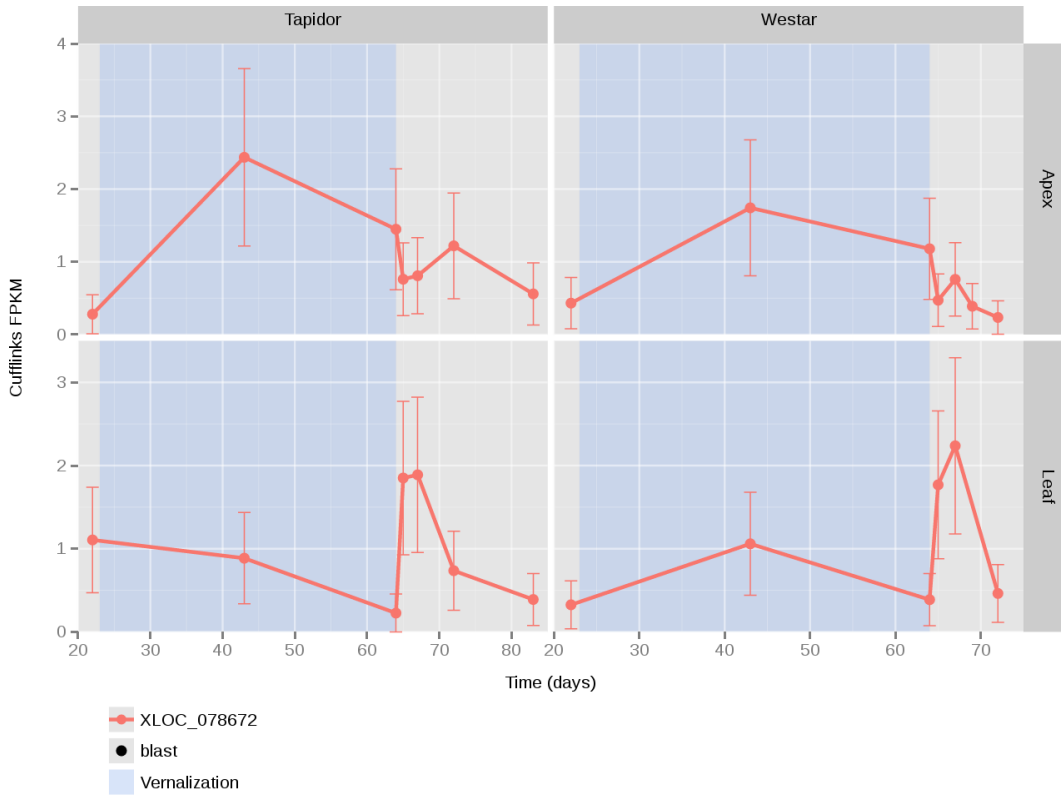


Figure 4.7: Expression patterns of the only *B. napus* gene showing sequence similarity to the Arabidopsis *miR172* precursor

The expression values and the 95 % confidence intervals of those expression values as computed by Cufflinks are displayed. Expression is very low in both tissues.

of their respective families, were used to query the ORDER database using the BLAST Search function, nine and one *B. napus* genes displayed sequence similarity respectively. The most highly expressed *B. napus* homologue of *miR156a*, displays relatively high expression in the leaf tissue in both varieties at the start of the time series (Figure 4.6). After the cold treatment, the expression of the gene decreases in both varieties (Figure 4.6). Such an expression profile is consistent with the expression of the *miR156* family in Arabidopsis[434], suggesting that *B. napus* has a similar age-dependent flowering pathway. In the apex tissue in both varieties, the gene exhibits expression values below 2.0 FPKM, and is therefore regarded to not be expressed. The single *B. napus* homologue of *miR172* is expressed very lowly in both the apex and the leaf tissue, barely being expressed above the 2.0 FPKM expression threshold (Figure 4.7). Therefore, although the expression of the *miR156* precursor suggests *B. napus* shares a similar age-dependent flowering pathway with Arabidopsis, a highly expressed *miR172* precursor could not be identified. The lack of *miR172* could potentially be the result of the sequencing depth not being adequate to detect the transcript, or alternatively due to the *B. napus* ageing pathway being mechanistically distinct to the pathway elucidated in Arabidopsis. This is suggested by the *miR172* family of miRNAs being under-represented in the *B. napus* genome relative to other families[441].

4.4 Conclusions and future directions

The objective of ORDER was to facilitate access to the transcriptomic time series dataset, allowing users to easily search for *B. napus* gene of interest and plot their expression profiles. The dual search functions allow full access to the dataset, allowing users to search using homology to Arabidopsis genes or homology to user submitted sequences. Examples of using ORDER to investigate regulation of floral development were given, emphasizing the requirement for both search methods in order to access the dataset. Finally, the database structure is such that new data can be easily added to the database and be plotted alongside the transcriptomic time series data currently collected.

Future developments to the website would focus on better integration with other *Brassica* crop resources and improved tools for data analysis. The EnsemblPlants database[321] contains a wealth of data about synteny between different *Brassica* species. Integrating that data would allow users to search the database using gene identifiers from any *Brassica* crop. Having access to this data would also allow users to make interesting comparisons, such as compare the expression profiles of homoeologues more easily. Although ORDER is currently a standalone application, the plotting functions and server code could easily be integrated into a larger resource, such as the Brassica Information Portal[430]. As both the database code and the plotting functions are written to be agnostic to the input data, this would allow user submitted transcriptomic time series data to be uploaded and available to search. In addition to better integration with current resources, improvements could be made to the way users query the data. Currently, users search the database based on the cDNA sequence of the gene. However, for some use cases it may be more useful to search the dataset using the shape of the expression profile, or using genomic location. One example would be a user investigating genes that exhibit similar expression profiles as their gene of interest. This could be achieved in ORDER by integrating an interactive SOM plot of the transcriptome time series (section 2.2.4), allowing users to search through genes located in the same SOM cluster as their gene of interest. The genomic location based search would be useful for association studies, where researchers have identified a region of the genome associated with a particular trait and wish to narrow down which genes or gene in the interval could potentially be responsible. Combining these search methods would allow the dataset to be divided even further, narrowing down candidate gene lists. Finally, although ORDER was constructed as an interface to a particular dataset, that does not limit the scope of its impact. Much like how the Arabidopsis "Electronic Fluorescent Pictograph" browser[428] has lead to similar browsers for other species[442], ORDER could become a template for how gene expression time series experiments are made available to the research community.

Chapter 5

Discussion

5.1 Chapter summaries

5.1.1 Floral gene retention and divergence in a spring variety

Whole genome duplication events have occurred throughout angiosperm development[443]. Following whole genome duplication, many genes are expected to be lost due to chromosome rearrangements and fusions[444]. Those genes that remain may partition multiple gene functions between duplicates[213,216], acquire novel functions[211], or retain the same function and be retained in the genome to maintain the correct gene dosage[221,224]. Determining the expression profiles of duplicated genes can provide clues as to which of these processes have led to certain genes being retained in the genome.

In *B. napus*, duplicated genes from ancient whole genome duplications[112,113], and more recent polyploidy[107] are present in the genome. Understanding the genetic pathways controlling flowering in *B. napus* would allow for breeding varieties with improved flowering behaviour[3]. However, the combinatorial explosion of regulatory possibilities that result from these duplicated genes complicate efforts to translate knowledge about the floral transition from Arabidopsis where much is known[15], to *B. napus*. To elucidate the extent of gene divergence in *B. napus*, particularly of the flowering time genes, a

transcriptomic time series was conducted and is presented in chapter 2. While the true extent of gene divergence is difficult to determine without functional data concerning protein activity, taking a genome-wide approach results in more general, widely applicable, conclusions to be drawn. Such a dataset allowed the changes to the transcriptome to be followed across the floral transition in two tissues, the leaf and apex. Expressed flowering time genes were found to be retained in the genome at a higher rate than other genes in the genome. The extensive regulatory divergence observed suggested that other processes, other than gene dosage effects, have contributed to the retention of genes in *B. napus*.

The floral transition in Arabidopsis is controlled by a tightly interconnected regulatory network, consisting of multiple feedback loops to result in irreversible, robust flowering[41,299]. At least one *B. napus* homologue, and often more, exhibited expression profiles consistent with the homologous gene in Arabidopsis, suggesting a general conservation between the expression domains of the genes in Arabidopsis and *B. napus*. A dramatic pattern of tissue-specific expression of homologues was observed for *BnSOC1* genes, suggesting a partitioning of spatial expression domains between different homologues. The expression profiles also suggest that different *BnSOC1* genes have different sensitivities to environmental signals. Given the role *SOC1* plays in integrating multiple environmental signals in Arabidopsis[83], the work presented here suggests that *BnSOC1* genes are an almost archetypal example of the ways subfunctionalization can manifest.

Changes to cis-regulatory elements represent a way by which the expression of genes can diverge[213]. The expression profiles of *BnTFL1* were shown to correlate with the presence and absence of regions of sequence conservation with Arabidopsis *TFL1* downstream of the gene. These regions of sequence conservation occur in areas of downstream sequence identified as cis-regulatory elements in Arabidopsis[309]. The similarities between the expression profiles of particular *BnTFL1* genes and the expression domains the regulatory elements defined in Arabidopsis suggest conservation between how the *BnTFL1* genes and the *TFL1* in Arabidopsis are regulated. Although correlative, this analysis suggests that cis-regulatory element changes may represent an important driver of subfunctionalization in *B. napus*. It also highlights how studies

dissecting regulatory elements in a model species can lead to insights in a crop. As the regulatory elements controlling other genes are elucidated to the level of *TFL1*[309] it will be interesting to investigate whether the other cases of regulatory divergence observed in *B. napus* flowering time genes can be explained with regulatory element changes.

The results from analysis of *BnFD* sequence divergence demonstrate that determining gene divergence from expression data leads to an underestimation of the true divergence present. Sequence differences between *BnFD* homologues are predicted to alter dimerization affinities of BnFD proteins, resulting in certain BnFD dimers being more likely to occur than others. This may be a common method of bZIP divergence, as similar differences between FD-like proteins in a range of plant species are observed. A simple computational model was used to understand the consequences of different dimerization affinities between BnFD proteins. The simulations highlighted that novel regulatory behaviours are possible as a result of different dimerization affinities. While the results are currently only theoretical, dimerization has been shown to facilitate gene regulatory logic[315] and is a factor influencing the evolution of bZIP transcription factors[317]. However, without further data it is difficult to conclude whether the observed differences are biologically relevant. This would require determining if different BnFD dimers possess different activities, such as different preferences in binding sites. The changes observed may also represent a form of complementary change, whereby BnFD proteins are diverging simultaneously with binding partners. As data on FD protein interactions become available it will be interesting to revisit these results.

The questions of gene retention raised in this chapter have to be viewed in the context of *B. napus* being a crop, grown under artificial selection. Although gene redundancy is not necessarily stable in natural conditions[212,215], it may be selected for in an agricultural setting where consistency is paramount. It is suggested that polyploidy represents a method of fixing heterosis, or hybrid vigour, in a crop[324,445–447]. The regulatory divergence observed suggests that polyploidy may indeed lead to a 'Swiss Army knife' of similar genes being retained in the genome, each adapted to particular growth conditions, tissue, or stage of development, which can be expressed as and when it is needed.

5.1.2 Effects of a requirement for cold on regulatory divergence

An important agronomic trait of *B. napus* is whether the plant requires a period of cold in order to flower. Varieties that do not require cold are often grown in Canada and Northern Europe, where harsh winters would damage crops grown during winter, whereas varieties that do require cold in order to flower are grown in Europe and Asia[127]. This requirement for cold is called vernalization, and is a pathway that is well understood at the molecular level in the model species Arabidopsis[27]. The pathway is arguably the most well understood flowering time pathway in *Brassica* crops with an array of different studies finding *B. napus* vernalization gene homologues associated with flowering and exhibiting sequence divergence[137,141,147]. How a requirement for cold influences the overall transcriptome, however, and whether the *BnFLC* genes influence the expression of certain floral integrators more than others was not known.

In chapter 3 the effects of a requirement for cold on the transcriptome of *B. napus* were assessed by comparing a winter variety, Tapidor, and spring variety, Westar. The potential importance of the leaf during vernalization in *B. napus* was revealed through an expansion of expressed gene number in Tapidor relative to Westar. As the action of *FLC*, a key vernalization sensitive gene, acts at both the leaf and the apex[31], exploring the biological significance of this increased gene set expressed in the leaf in the winter variety will be a central question motivating future work. Correlation analysis suggested that different factors influence the transcriptome depending on the tissue, with the leaf seemingly influenced by plant age and the apex by developmental stage. This is counter to expectations from Arabidopsis[410], but may represent an instance of artificial selection for leaf senescence to allow metabolites to be remobilized from leaves to the growing flowers and seeds, leading to yield increases. Taken together, these findings suggest that the vernalization response may be affecting both the signals the leaf transmits to the apex and the way the apex interprets those signals, consistent with the role of *FLC* in Arabidopsis[31].

Considering genes involved with the vernalization response, copies of *BnFLC* were found to exhibit varietal differences and expression profiles consistent with these genes mediating the vernalization response in the winter variety. The

two best candidates for conferring the vernalization response in Tapidor, based on their expression profiles, are *BnFLC.A2* and *BnFLC.A10*, consistent with previous studies[137]. Not all *BnFLC* genes were found to respond similarly to cold. *BnFLC.A3b* and *BnFLC.C2* are not stably silenced in Tapidor, revealing that expression of these genes does not prevent flowering. Potentially these *BnFLC* genes may require longer periods of cold to become fully repressed, in a similar manner to certain *FLC* alleles in Arabidopsis[345,346]. The experimental design decision to subject a spring variety to vernalization may initially seem strange. However, doing so allowed candidate *BnFLC* genes for the mild vernalization response in Westar to be identified[241]. The theoretical consequences of having differently tuned *FLC* homologues were considered, and proposed to allow plants to disentangle the length of vernalization and the temperatures experienced during cold. Other genes involved with the vernalization response exhibited differences in the magnitude of expression between varieties. However, the consequences of these observed differences are difficult to assess.

The expression differences of *BnFT* and *BnSOC1* genes between varieties were consistent with the effects of *BnFLC* mediated repression, in line with findings from Arabidopsis[85,308]. Two genes exhibited very different expression profiles in Tapidor relative to Westar. *BnFD.A1* and *BnAP1.A2* exhibited expression divergence in the spring variety, but the expression profiles of these genes in Tapidor were more consistent with the other homologues of those genes. Determining whether these differences influence the vernalization response, or represent differences due to variety, would require a more thorough assessment of transcriptomic changes involving multiple winter and spring lines. Either way, given the results from chapter 2, the altered expression of a single *BnFD* may have large impacts on the BnFD dimers observed.

It should be emphasized that the differences observed between Tapidor and Westar represent a single comparison between a winter variety and a spring variety. Therefore, the findings presented in chapter 3 should not be extrapolated to other *B. napus* varieties to explain the differences between all spring and winter varieties. However, the results do highlight potential candidate genes consistent with the literature.

5.1.3 Data dissemination using a web application

An important part of the scientific process is the sharing of data. Sharing data allows others in the field to more readily consider their results in light of previous studies. This is of particular relevance to extremely large transcriptomic datasets, where the scale of the data makes it infeasible for every gene to be investigated by a single group of researchers. The ready availability of large datasets such as this allow for a division of labour, with insights on particular genes made by experts.

The transcriptomic time series presented in this work is of general interest to any Brassica researcher investigating genes expressed during the floral transition. In chapter 4 a web resource that facilitates access to the dataset is described. The search features allow researchers to find genes of interest and plot expression profiles in an intuitive manner. To ensure the resource is as generally useful as possible, the database structure and plotting features facilitate the easy inclusion of additional data.

5.2 Outlooks and limitations

A number of observations from the transcriptomic time series, as well as limitations of the dataset, pose interesting avenues for future work.

The way the plants were grown and tissues sampled influenced the transcriptomic time series obtained. The *A. thaliana* shoot apical meristem is composed of a relatively small subset of cells and is on the order of 100 μm in size[448,449]. Within this small collection of cells, transcriptionally distinct zones are present[13]. The floral repressor *TFL1* and the floral activators *AP1* and *LFY* mutually antagonize each other's expression[54,55], leading to sharp boundaries between expression domains. This is proposed to be important for accurately defining regions of floral development[52,53,56]. In the *B. napus* transcriptome time series, all of these genes increase in expression during the floral transition post-cold, which you would not expect if mutual antagonism was taking place (Figures 2.27, 2.32, and 2.33). While it is possible that the genes have diverged entirely in their function, this seems unlikely given

the observed conservation in flowering time control genes between *B. napus* and Arabidopsis. This suggests that although the dissection of the apical region was adequate to enrich for apically expressed genes (section 2.4), these distinct expression domains were all sampled together. While this does not limit the use of the data to assess functional divergence, it is an important caveat as the time series is not able to capture the antagonistic regulatory interactions expected between these flowering time genes. High resolution laser microdissection of apical meristems, however, is able to accurately separate these domains[450]. Conducting laser microdissection of *B. napus* apices during the floral transition, followed by assessing gene expression, would allow these transcriptional domains to be resolved.

This idea of unique expression domains can be taken further: single-cell transcriptomics. An example of where understanding expression dynamics at the cell resolution is required is the expression of the floral repressor *FLC* in Arabidopsis. *FLC* is expressed and silenced in a cell-specific manner, such that each particular cell is either expressing *FLC*, or it is not[360,361]. However, when whole plant or leaf samples are assayed for *FLC*, a quantitative, analogue response is observed[245], as a result of averaging at the tissue level. This will be important when assessing genes that seemingly have the same expression profile in the transcriptomic time series. Although regulatory divergence was observed between flowering time genes, there are still a significant number of homologues that exhibit similar expression profiles. This can be visualized in Figure 2.23 as any point that does not lie on the diagonal line that represents complete regulatory divergence between *B. napus* homologues of an Arabidopsis gene, and in expression profiles of homologues such as *BnLFY* (Figure 2.32), *BnAP1* (Figure 2.27), and *BnFLC* (Figure 3.10), to name a few. Potentially, these seemingly co-regulated homologues are actually expressed in a cell-specific manner, with only a single homologue expressed per cell. This is consistent with the framework of responsive backup circuits, that proposes that duplicated genes may autoregulate each other to provide genetic backup and regulatory robustness[219,220]. This theory is particularly attractive given that a number of MADS-box containing genes involved with floral development have been found to autoregulate their own expression in Arabidopsis[451–453]. If such regulatory interactions were present between different homologues, then potentially the

cell-specific 'decision' of which homologue to express would be a stochastic process. Testing such a hypothesis could be achieved by using single-cell RNA-Seq to determine cell-to-cell variability in homologue expression[454].

An aspect of sampling which potentially limits the transcriptional time series in terms of the developmental responses it can be used to investigate is the change in temperature and photoperiod during the vernalization period. Changing both growth variables is necessary in order for the vernalization treatment to be as physiologically accurate as possible. However, this results in transcriptional changes due to cold stress[455,456] and photoperiod[19,242] to be observed simultaneously (section 2.2.5). Thus, in the current study, these pathways cannot be disentangled. In order to allow these pathways to be studied during the floral transition, a staggered vernalization treatment could be given, with a change in photoperiod occurring before a change in growth temperature.

The results from BnFD proteins suggest that changing dimer specificity may be a way in which genes diverge after duplication (section 2.5.2). Another family of transcription factors that bind to DNA as dimers are the MADS-box domain containing proteins[279,457,458]. This family of proteins are of particular interest because of the roles they play in the floral transition and floral development[279]. Indeed, the dimerization dynamics of the proteins have been highlighted as influencing the function of the proteins. SVP-FLC heterodimers bind different target sequences than SVP homodimers[99], while the function of AP1 protein changes based on its interaction partners, with the gene regulating floral meristem identity when complexed with AGL24 or SVP, and controlling sepal and petal identity when complexed with SEPALLATA proteins[82]. Indeed, interaction maps of the floral MADS-box containing proteins suggest a multitude of interactions are possible[96]. However, compared to the literature available on bZIP dimerization[314,459], the understanding of what controls the dimerization preferences of MADS-box containing proteins is lacking. This makes computationally predicting whether different homologues of MADS-box containing genes in *B. napus* have diverged in terms of interaction partner difficult. To test this, a yeast two-hybrid approach, such as that used to construct the Arabidopsis MADS-box transcription factor interaction map[96], could be used with *B. napus* genes as bait. Alternatively, the machine learning

algorithm developed by Potapov et al. (2015)[320] and used in chapter 2 to score BnFD interactions was trained using results from a protein microarray analysis of bZIP protein interactions[459]. Potentially a similar approach could be used to not only quantify dimerization differences between *B. napus* MADS-box homologues, but also develop a scoring algorithm for MADS-box protein dimerization.

The assessment of gene function from expression data has certain caveats associated with it. The function of a gene is a product of two things; the molecular activity of the protein the gene encodes and the spatiotemporal expression pattern of the gene. Two genes may encode identical proteins, but if they act in different tissues, or act at different points in development, for example, they have different functions. Likewise, two genes that are co-expressed may encode proteins with different molecular activities. Therefore, the level of divergence estimated from the transcriptomic time series is an underestimation of the true divergence that is present between duplicated genes in *B. napus*. This is demonstrated in the *BnFD* results, where despite similar gene expression the BnFD proteins seem to have diverged in terms of dimerization affinity. While limited in this way, the transcriptome is able to assess divergence genome-wide. In contrast, assessing changes in protein function genome-wide is more difficult, but as knowledge of protein structure and how that relates to function increases these types of studies will become possible. The results from *BnFD* also demonstrate this, as without the prior knowledge of bZIP dimerization preferences[314,320], the insights made here would not have been possible.

The gene regulatory network for flowering in Arabidopsis was elucidated over decades of molecular and genetic studies[15,299]. However, computational approaches exist that allow gene regulatory networks to be inferred from time series data[460–463]. Using the transcriptome time series to elucidate such regulatory networks would be a potential avenue for future work. Indeed, collecting transcriptomic data from additional tissues and additional developmental phases would allow for specific regulatory networks to be generated for each tissue and transition. The expression of floral integrators observed in the transcriptomic time series supports the notion that tissue-specific expression of homologues is possible in *B. napus* (Section 2.4). Understanding the tissue

specificity of different homologues may allow more directed breeding efforts[464]. An example of how this could be used is the floral repressor *FLC*. In addition to its key role in the vernalization pathway[27], the gene also plays a role in regulating seed germination[465]. If different homologues of *FLC* were found to be specific to particular pathways[229], breeding efforts could more readily make changes to one pathway while minimizing pleiotropic effects on the other.

Identifying regulatory networks in different tissues and developmental transitions is one of the approaches being undertaken as part of the Biotechnology and Biological Sciences Research Council's (BBSRC) Brassica Rapeseed And Vegetable Optimization (BRAVO) project. BBSRC BRAVO was built upon observations from this work of certain *B. napus* homologues exhibiting divergence in responses to particular regulatory or environmental inputs. By generating transcriptomic time series for multiple *B. napus* varieties, across a number of developmental transitions, the project aims to construct variety-specific and transition-specific gene regulatory networks to better understand the role of duplicated flowering time genes in *B. napus*. The insights and data generated as a result of BBSRC BRAVO should lead to a much better understanding of flowering time gene function in *B. napus*, and will allow a number of predictions and hypotheses made in this work to be revisited.

5.3 Concluding thoughts

The original aim of this project was to determine the extent to which the regulatory network underlying flowering in Arabidopsis could be applied to *B. napus*. The intention was to use the transcriptomic time series to reduce the complexity of the network by grouping similarly expressed flowering time gene homologues together as a single network node. The work presented in this thesis, however, revealed that regulatory divergence between homologous genes is frequently observed in *B. napus*. This introduced the challenge of how to deal with the combinatorial explosion of regulatory possibilities and to reduce the model to a computationally tractable system. The obstacle of additional regulatory complexity caused by multiple gene copies changed the

direction of the project to instead investigate how the dynamics of these genes have diverged.

This study represents the first study in *B. napus* to follow the transcriptome before, during, and after vernalization. Arabidopsis floral genes were found to be retained in the genome more frequently than expected, with the patterns of regulation suggesting different selective pressures are acting on the genes. Analysis of both the leaf and apex transcriptomes revealed that these tissues are distinct in their transcriptional responses, and identified cases where floral gene homologues have diverged in terms of their spatial expression domains. The importance of cis-regulatory elements in the evolution of duplicated genes is highlighted, and represents an example of how research in model species can begin to be translated to a crop species. The findings that similarly expressed genes exhibit functionally relevant sequence differences calls into question the very assumption on which the original project aim was based, namely, that similarly expressed genes can be considered as a single node in the network.

Despite the value of this work in elucidating regulatory divergence between gene homologues, a key question remains: how much closer are we to a regulatory network of the *B. napus* floral transition? This project emphasizes the problems inherent to determining a simple gene regulatory network of a crop that has experienced multiple rounds of gene duplication. Instead, the complexity of polyploid networks could be approximated by sub-networks based on modules with little regulatory dependence. This subsetting will require a better understanding of how the multiple copies have diverged, both in gene expression and protein activity, and provides a clear direction for future work in *B. napus*.

Chapter 6

Methods

Some of these methods are included in a paper written in collaboration with Dr. Rachel Wells, Dr. Nick Pullen, Dr. Martin Trick, Dr. Judith A. Irwin, and Prof. Richard J. Morris[1].

6.1 Plant growth and sample preparation

B. napus cv. Westar and *B. napus* cv. Tapidor plants were sown on the 7th May 2014 in cereals mix. Plants were grown in unlit glasshouses in Norwich, UK, with glasshouse temperatures set at 18 °C during the day and 15 °C at night. The sunrise during the sampling period was approximately 05:00, while sunset was approximately 21:00. On day 22 of growth, plants were transferred to a 5 °C, short day (8 hour) growth chamber to undergo vernalization. The lights in the growth chamber turned on at 08:00 and turned off at 16:00 each day. After a 42 day period of vernalization, plants were transferred back to unlit glasshouses and grown until the plants flowered. The first true leaf of each plant and shoot apices were sampled at 22, 43, 64, 65, 67, 69, and 72 days after sowing (Table 6.1). First true leaves were cut and immediately frozen in liquid nitrogen. The growing shoot apices were dissected using razor blades on a dry ice chilled tile before transfer to liquid nitrogen. Samples were pooled and ground in preparation for RNA extraction. For apex tissue, ~0.1 g of

[1]Preprint paper available at https://doi.org/10.1101/178137 and Appendix C.

Table 6.1: Sampling and sequencing scheme for the transcriptomic time series. Numbers in the rightmost two columns indicate the number of biological pools sampled for that time point within each tissue.

Date Sampled	Days Post Sowing	Days Vernalized	Days Post Vernalization	Tapidor		Westar	
				Leaf	Apex	Leaf	Apex
2014-05-23	16	0	0	-	-	-	-
2014-05-29	22	0	0	2	2	2	2
2014-06-19	43	21	0	2	2	2	2
2014-07-10	64	42	0	2	2	2	2
2014-07-11	65	42	1	1	1	1	1
2014-07-13	67	42	3	2	2	2	2
2014-07-15	69	42	5	-	-	-	1
2014-07-18	72	42	8	2	2	2	2
2014-07-29	83	42	19	2	2	-	-

apices were ground as a pool. At the early time points, as the apices were smaller, this mass of tissue equated to approximately 20 plant apices, while at later time points approximately 10 apices were pooled. For leaf samples, between 6 - 10 leaf samples from separate plants were pooled and ground. RNA extraction and DNase treatment was performed following the method provided with the E.Z.N.A® Plant RNA Kit (R6827-01; Omega Bio-tek Inc., USA). Library preparation and RNA sequencing was carried out by the Earlham Institute (Norwich, UK). Initial quality control of the RNA was carried out using the Quant-iT™ RNA Assay Kit (Q-33140; Thermo Fisher Scientific, USA) and the Quant-iT™ DNA Assay Kit (high sensitivity; Q-33120; Thermo Fisher Scientific, USA), and was quantified using a Tecan plate reader. RNA quality was further tested using the PerkinElmer GX, with high sensitivity DNA reagents and high sensitivity chips (5067-4626; PerkinElmer Inc., USA). Library preparation was carried out according to the TruSeq RNA protocol v2 (15026495 Rev. F; Illumina Inc., USA). Biotin beads were used to extract polyadenylated mRNA from the samples. The mRNA was fragmented and first strand cDNA was synthesized from random hexamer primers. Adapters were ligated to the DNA fragments, and the ligated products underwent bead-based size selection using Beckman Coulter XP beads (A63880; Beckman Coulter

Inc., USA). PCR was used to enrich for DNA fragments that had adapter molecules on both ends. RNA-Seq was performed on RNA samples from six time points for leaf tissue and seven time points from apex tissue. 100bp, single end reads were generated using an Illumina HiSeq2500, with an average of 67 million reads per sample (Table 6.2). To assess biological variation, a second RNA sample for five time points in both the leaf and apex were sequenced at a lower average coverage of 33 million reads per sample (Table 6.1).

6.2 Gene model prediction and read alignment

The gene model prediction software AUGUSTUS[253] (version 3.2.2) was used to determine gene models for the Darmor-bzh reference genome. TopHat[251] (version 2.0.13) aligned RNA-Seq reads from across the entire time series were combined and filtered using the filterBam tool provided with AUGUSTUS. AUGUSTUS used the filtered reads to aid the estimation of intron locations. Arabidopsis derived parameters provided with the AUGUSTUS software were used to predict *B. napus* gene models in the Darmor-*bzh* genome, with default parameters used otherwise. RNA-Seq reads were aligned and expression levels quantified using the Tuxedo suite of software following the published workflow[250]. TopHat[251] (version 2.0.13) with the `b2-very-sensitive`, `transcriptome-only`, and `prefilter-multihits` parameters set was used to align reads to the Darmor-*bzh* reference sequence, using the AUGUSTUS derived gene models to determine the location of gene models. Cufflinks[256] (version 2.2.1) was used to quantify the expression levels of *B. napus* genes. Data normalisation using `cuffnorm` was performed separately for leaf and apex tissue samples. Aside from the named parameters, default values were used.

Table 6.2: Sequencing statistics for the two sequencing runs carried out to generate the developmental transcriptome

Continued on Page 267.

Variety	Tissue	Days Post Sowing	Biological Replicate 1				Biological Replicate 2			
			Total Reads (millions)	Mapped Reads (millions)	Multiple Mapping Reads (millions)	Above 20 Mappings Reads	Total Reads	Mapped Reads	Multiple Mapping Reads	Above 20 Mappings Reads
Tapidor	Apex	22	78.4	65.9 (84.0%)	8.8 (13.3%)	22.7 (0.3%)	35.1	29.5 (83.9%)	4.0 (13.5%)	7.8 (0.3%)
		43	69.5	56.4 (81.1%)	7.3 (12.9%)	20.1 (0.4%)	32.2	26.3 (81.6%)	3.5 (13.2%)	6.1 (0.2%)
		64	69.3	56.6 (81.7%)	7.3 (12.8%)	18.0 (0.3%)	34.9	28.9 (83.0%)	3.9 (13.3%)	3.7 (0.1%)
		65	70.6	58.7 (83.1%)	7.6 (13.0%)	21.1 (0.4%)	-	-	-	-
		67	80.2	67.6 (84.4%)	9.1 (13.5%)	41.4 (0.6%)	34.2	28.7 (83.9%)	4.0 (13.8%)	4.7 (0.2%)
		72	54.2	45.2 (83.4%)	5.9 (13.1%)	18.3 (0.4%)	27.3	22.4 (82.2%)	3.0 (13.2%)	6.6 (0.3%)
		83	66.2	55.0 (83.0%)	7.3 (13.2%)	62.5 (1.1%)	31.3	25.9 (82.6%)	3.4 (13.3%)	10.0 (0.4%)
	Leaf	22	66.2	55.8 (84.3%)	8.5 (15.3%)	11.4 (0.2%)	32.9	27.8 (84.6%)	4.2 (15.3%)	3.8 (0.1%)
		43	58.1	48.4 (83.2%)	7.1 (14.7%)	12.0 (0.2%)	32.8	26.8 (81.9%)	3.9 (14.5%)	3.9 (0.1%)
		64	63.5	53.5 (84.2%)	7.4 (13.9%)	7.7 (0.1%)	22.9	19.2 (83.8%)	2.7 (14.1%)	2.3 (0.1%)
		65	74.1	62.5 (84.3%)	8.9 (14.3%)	11.4 (0.2%)	-	-	-	-
		67	79.7	64.5 (80.9%)	9.3 (14.5%)	8.1 (0.1%)	25.5	21.2 (83.0%)	3.1 (14.6%)	4.4 (0.2%)
		72	56.5	47.1 (83.3%)	6.6 (14.0%)	7.9 (0.2%)	30.4	25.3 (83.2%)	3.8 (14.9%)	3.5 (0.1%)
		83	58.4	48.1 (82.4%)	6.5 (13.5%)	8.3 (0.2%)	42.3	34.2 (80.9%)	4.8 (14.2%)	4.6 (0.1%)
Westar	Apex	22	75.6	61.8 (81.8%)	8.3 (13.4%)	20.7 (0.3%)	41.9	34.3 (81.9%)	4.7 (13.8%)	7.8 (0.2%)
		43	71.5	56.8 (79.4%)	7.4 (13.1%)	17.8 (0.3%)	31.7	25.3 (79.8%)	3.4 (13.6%)	5.3 (0.2%)
		64	70.5	57.4 (81.4%)	7.5 (13.0%)	21.6 (0.4%)	28.7	23.3 (81.2%)	3.2 (13.8%)	149.4 (6.4%)
		65	67.6	54.6 (80.7%)	7.2 (13.2%)	26.5 (0.5%)	-	-	-	-
		67	78.6	63.5 (80.8%)	8.4 (13.2%)	36.3 (0.6%)	30.5	25.1 (82.3%)	3.5 (13.9%)	5.6 (0.2%)
		69	66.2	54.4 (82.2%)	7.3 (13.5%)	30.7 (0.6%)	-	-	-	-
		72	59.7	48.6 (81.4%)	6.4 (13.2%)	35.2 (0.7%)	31.5	25.8 (81.8%)	3.6 (14.1%)	4.5 (0.2%)
	Leaf	22	68.2	54.7 (80.2%)	8.4 (15.4%)	9.5 (0.2%)	33.9	28.0 (82.5%)	4.4 (15.7%)	3.7 (0.1%)
		43	50.5	41.5 (82.1%)	6.2 (15.0%)	11.1 (0.3%)	33.0	26.4 (80.1%)	4.0 (15.1%)	4.6 (0.2%)
		64	73.9	60.7 (82.1%)	8.8 (14.4%)	10.2 (0.2%)	35.5	29.1 (82.1%)	4.3 (14.8%)	3.7 (0.1%)
		65	45.7	37.6 (82.2%)	5.5 (14.6%)	5.4 (0.1%)	-	-	-	-
		67	81.8	67.1 (82.1%)	10.0 (14.9%)	9.4 (0.1%)	35.7	28.8 (80.7%)	4.4 (15.4%)	3.5 (0.1%)
		72	49.0	40.3 (82.1%)	5.8 (14.5%)	5.8 (0.1%)	32.2	26.2 (81.2%)	3.9 (15.1%)	3.9 (0.1%)

Continued from Page 266. Reads were mapped to the Darmor-*bzh* reference genome using TopHat[251]. The percentage of mapped reads is given as the percentage of the total reads. Multiply mapped reads are defined as reads that mapped to multiple places in the genome with an equal probability. The percentages of multiply mapped reads and the percentage of reads mapping to more than 20 position in the genome are calculated as a total of the reads that were mapped to the genome, and not a percentage of the total reads.

6.3 Identification of sequence similarity between *B. napus* and Arabidopsis gene models

The BLAST algorithm, using the `blastn` binary provided by NCBI[466] (version 2.2.30+) was used to identify sequence similarity between the AUGUSTUS[253] derived gene models and the published Arabidopsis gene models downloaded from TAIR[431] (version 10). The `blastn` algorithm was run using default parameters, with an e-value threshold of 10^{-50} used to identify sequence similarity between the AUGUSTUS derived *B. napus* gene models and published Arabidopsis gene models. For the analysis conducted in this study, only the most highly scoring `blastn` hit was used to identify *B. napus* copies of Arabidopsis genes.

6.4 Between genome expression comparison

Density plots of $\log_{10}$ transformed FPKM values were calculated and visualised using the R statistical programming language[467]. The subsets of *B. napus* genes used showed sequence similarity to at least one published Arabidopsis gene model downloaded from TAIR[431] (version 10), and sequence similarity to an Arabidopsis gene in the FLOR-ID database[299] (accessed 2016-08-19). The expression fold change for homoeologue pairs was calculated using untransformed FPKM values (Tables 2.1 and 2.2). The geometric mean of the fold change across all n homoeologous gene pairs was calculated as $\sqrt[n]{\prod_{g=1}^{n} \frac{FPKM_{C,g}}{FPKM_{A,g}}}$

where $FPKM_{X,g}$ is the FPKM value of the X genome copy of the homologue pair g.

6.5 Homoeologue pair identification

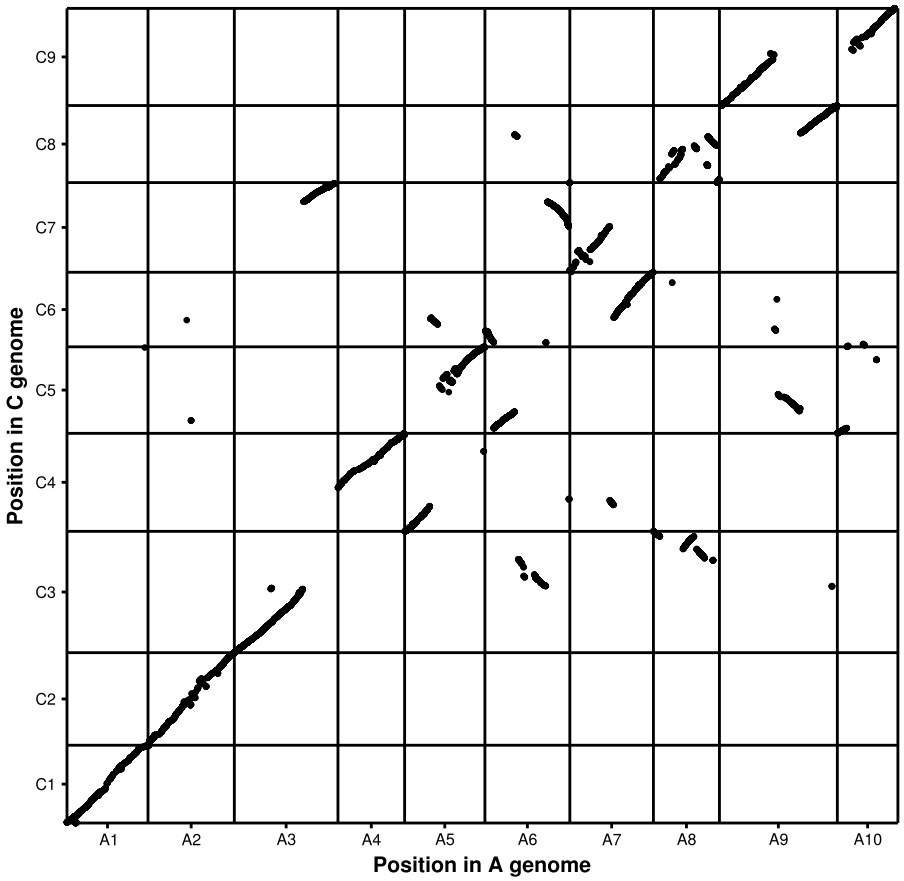


Figure 6.1: Locations of identified homoeologues pairs in the *B. napus* genome The locations of these pairs give a representation of the chromosomal rearrangements that have occurred between the A and C genomes.

The method outlined by Chalhoub et al. (2014)[118] was used to identify pairs of homoeologues between the A and C genomes[118]. The Darmor-*bzh* reference genome was divided into the A and C genomes, removing the reference pseudo-chromosomes which consist of sequence that is unassigned to a specific chromosome. The separated genomes were uploaded to the CoGe portal[468] and

the SynMap tool[469] was used to identify regions of syntenic genes between the two genomes. Chains of syntenic genes were identified using DAGchainer[470], allowing a maximum 20 gene distance between two matches and with a minimum number of 4 aligned pairs constituting a syntenic block. A 1:1 synteny screen was performed using the QUOTA-ALIGN[471] procedure. The synteny screen is necessary to distinguish homoeologous regions of the genome and paralogous regions which are the result of genome multiplication events which occurred prior to the interspecies hybridisation event in the evolutionary history of *B. napus*. Once syntenic genes were identified using SynMap, a reciprocal sequence similarity filter was applied using the BLAST algorithm. The `blastn` algorithm was used with default parameters and a 10^{-50} e-value threshold to assess sequence similarity, and only homoeologue pairs which were reciprocal best hits in this analysis were considered. This resulted in 14427 homoeologous pairs distributed across the entire *B. napus* genome (Figure 6.1).

6.6 Weighted gene co-expression network analysis

The weighted gene co-expression network analysis was carried out using the WGCNA library[265] (version 1.51) available for the R statistical programming language[467] (version 3.2.2). Due to the size of the dataset, WGCNA was performed on clustered data. The expression data was first filtered and normalised for each tissue separately. Any genes with a maximum FPKM value across the time series of less than 2.0 were removed. For the remaining genes, the expression across time was normalised to have a mean of 0.0 and a variance of 1.0. Using the normalised expression values, hierarchical clustering was conducted separately on the leaf and apex data using Euclidean distances between expression traces and a complete agglomeration method. The hierarchical tree was cut into H numbers of clusters and the ratio $\frac{\sum_{c=1}^{H} N_c(\bar{x}_c - \bar{x})^2}{\sum_{g=1}^{N}(x_g - \bar{x})^2}$ was calculated for each tree cut height, where N is the total number of genes, N_c is the total number of genes assigned to cluster c, x_g is the expression vector for gene g, $\bar{x}_c$ is the mean expression vector for genes assigned to cluster c, and $\bar{x}$ is the global mean of all expression vectors. The expression vectors

are defined as $\bar{x}_g = (\widehat{FPKM}_{g,22}, \widehat{FPKM}_{g,43}, \cdots, \widehat{FPKM}_{g,72})$ where $\widehat{FPKM}_{g,t}$ represents the normalised FPKM level of gene g at time point t, with all time points included in the vector. A ratio of ~0.98 was chosen as a good balance between the number of clusters and how well the clusters represented the expression data. This ratio corresponded to 2683 clusters for leaf tissue and 6692 clusters for apex tissue in Westar. WGCNA[265] was carried out using the mean expression vectors for the 6692 apex clusters and the 2683 leaf clusters. Based on the assumption of a scale-free network structure, a soft threshold of 30 was used for both the apex and leaf samples. A minimum regulatory module size of 30 was used and modules with similar eigengene values were merged to give the final regulatory modules used for regulatory module assignment.

6.7 Self-organising maps and the identification of regulatory modules

Self-organising maps (SOM) were generated using the `kohonen` library[472] available for the R statistical programming language[467]. As with the WGCNA analysis, the data was filtered and normalised prior to carrying out the SOM analysis. The number of nodes used in the SOM was chosen based on the ratio $\frac{\sum_{c=1}^{S} N_c(\bar{x}_c - \bar{x})^2}{\sum_{g=1}^{N}(x_g - \bar{x})^2}$ where N is the total number of genes, S is the total number of SOM nodes, N_c is the total number of genes assigned to SOM node c, x_g is the expression vector for gene g, x_c is the expression vector for SOM node c, and $\bar{x}$ is the global mean of all expression vectors. A value of S was chosen such that the above ratio was ~0.85 for both tissues. To adequately capture the variation present in the data, the dimensions of the SOM were set as the ratio between the first two principal component eigenvalues of the data, as has been done previously[473].

To assign probabilities of genes clustering to the same SOM cluster, a resampling procedure was employed (Figure 2.24). Expression values were resampled assuming a Gaussian noise model, using the true expression value as the mean of the distribution and the true expression value uncertainty calculated by Cufflinks as the distribution variance. The resampled expression values for each gene, within each tissue, were normalised to a mean expression of 0.0

with a variance of 1.0 across the time series and assigned to a SOM cluster based on a minimal Euclidean distance. This sampling loop was repeated 500 times, and the SOM clusters to which the genes of interest mapped were recorded. From this process, an empirical probability of mapping to each SOM cluster was calculated for each gene of interest. The probability of two genes mapping to the same SOM cluster was then calculated as $\sum_{c=1}^{S} \frac{n_{g_1,c} n_{g_2,c}}{250000}$ where S is the total number of SOM clusters, and $n_{g_i,c}$ is the number of times gene i mapped to SOM cluster c. As the SOM training process begins from a random starting point, some SOMs were found to better discriminate between the expression traces of some pairs of genes than other SOMs. To overcome this, the probability of two genes of interest mapping to the same SOM cluster was calculated for 100 different SOMs. This probability was averaged to give the average probability of two genes of interest mapping to the same SOM cluster.

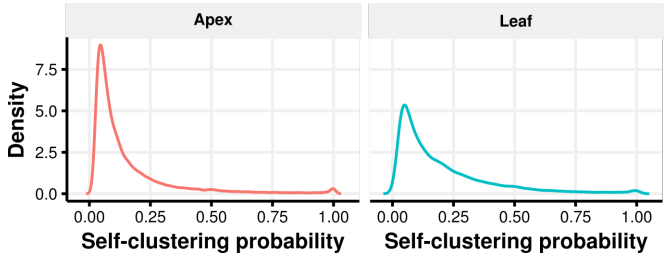


Figure 6.2: A bimodal distribution of self-clustering probabilities necessitates the use of a threshold to visualise the probabilities
The density curves presented here represent the self-clustering probabilities calculated from a single SOM. The clustering coefficient threshold was taken by determining the self-clustering probability that corresponded to the peak of the density curve. This threshold was calculated for each SOM and averaged to give the final thresholds for the apex (0.053) and the leaf (0.056).

The probability of mapping to the same cluster can also be calculated for a single gene of interest by calculating $\sum_{c=1}^{S} \left(\frac{n_{g_1,c}}{500}\right)^2$. This value is a measure of how consistently a gene maps to the same SOM cluster, giving an indication of the uncertainty in the expression values calculated for that gene. Plotting a distribution of these self-clustering probabilities (Figure 6.2) reveals a bimodal distribution with maxima at ~0.05 and ~1.0. To aid with visualising the average probabilities of two genes mapping to the same SOM cluster, as a

consequence of this bimodality, a soft threshold based on a cumulative Gaussian density function was applied. The resulting value is referred to as a clustering coefficient. Clustering coefficients were calculated as $\frac{1}{2}\left[1+\text{erf}\left(\frac{\mu_{p_{g_1,g_2}}-\theta}{\sigma_{p_{g_1,g_2}}\sqrt{2}}\right)\right]$ where erf is the error function defined as $\text{erf}(x)=\frac{1}{\sqrt{\pi}}\int_{-x}^{x}e^{-t^2}\text{d}t$, $\mu_{p_{g_1,g_2}}$ is the average probability of genes g_1 and g_2 mapping to the same cluster, $\sigma_{p_{g_1,g_2}}$ is the standard deviation of the probabilities calculated from the 100 different SOMs used in the sampling procedure, and θ is the tissue-specific threshold. A threshold of 0.053 (apex) or 0.056 (leaf) was used in Westar. This threshold was calculated by taking the self-clustering probability that corresponded to the maximum of the density curve (Figure 6.2) for each SOM and averaging them. An automated approach was taken to quantify the pattern of clustering coefficients between copies of the same gene. Clustering coefficients were subjected to a binary filter, such that coefficients above 0.5 were set to 1 and those below set to 0. Regulatory modules were defined as groups of genes where the binary clustering coefficients between all genes were 1. Based on the membership of these groups, patterns were assigned as *distinct*, *unique*, *gradated*, *mixed*, or *redundant*.

6.8 Sequence conservation analysis of *BnTFL1* genes

Sequence upstream and downstream of the Arabidopsis *TFL1* gene was extracted from the AtGDB TAIR9/10 v171 Arabidopsis genome assembly located on PlantGDB[474]. *BnTFL1* sequence was extracted from the Darmor-*bzh* reference genome sequence[118]. Regions of conserved sequence were identified using mVISTA from the VISTA suite of tools[475,476]. The alignment algorithm used was AVID[477], which performed global pair-wise alignments for all sequences. Percentage sequence conservation was calculated using a 100bp sliding window.

Table 6.3: *BnTFL1* and *BnGAPDH* qPCR primer sequences.

Gene	Forward Primer (5' - 3')	Reverse Primer (5' - 3')	Amplicon Length
BnTFL1.A10	GTCTCCAATGGCCATGAGT	GTGCCGGGGATGTTCATG	179
BnTFL1.Cnn.Random	GTCATGAACATCCCCGGC	GATCATTCTCGATCGCAAATTCA	196
BnTFL1.C2	CTGATGTTCCAGGTCCTAGC	TGGGGAGATATCGATAACATGTC	197
BnTFL1.C3	GAGGTGGTGAGCTATGAGTTG	CTGGGCGTTAAAGAAGACAGCA	189
GAPDH	AGAGCCGCTTCCTTCAACATCATT	TGGGAACACGGAAGGACATTCC	112

6.9 Quantitative PCR of BnTFL1 homologues

Reverse transcription quantitative PCR (RT-qPCR) was carried out on copies of *BnTFL1* using custom designed primers (Table 6.9). The SuperScript® III First-Strand Synthesis System (Thermo Fisher Scientific Inc., USA) was used to generate cDNA, with 2 μg of RNA used as input. The RNA was extracted as described above, with all Westar apex samples, from both biological replicates, being used. Each RT-qPCR reaction consisted of 5 μl LightCycler® 480 SYBR Green I Master (Roche Molecular Systems Inc., USA), 4 μl cDNA, 0.125 μl of the forward and reverse primers at a concentration of 10 μM and 0.75 μl water. Quantification was performed on a LightCycler® 480 (Roche Molecular Systems Inc., USA). The RT-qPCR cycle consisted of a 95 °C denaturation step for 5 minutes followed by 50 quantification cycles. Each cycle consisted of 15 seconds at 95 °C, 20 seconds at 58 °C, 30 seconds at 72 °C. Fluorescence was quantified at 75 °C as the temperature was ramping from 72 °C to 95 °C.

6.10 Gene Ontology term enrichment

Gene Ontology (GO) term enrichment was performed using custom scripts written in the R statistical programming language[467]. *B. napus* genes were first annotated with GO terms using homology to Arabidopsis genes. The Arabidopsis GO terms used were from the `org.At.tair.db` library[478] (version 3.2.3). The GO terms associated with the Arabidopsis gene with the highest sequence similarity to each *B. napus* gene, as determined by `blastn`[466] (version 2.2.30+), were assigned to each *B. napus* gene. The `topGO` library[479] (version 2.22.0) was used to perform the GO term enrichment. The parameters used to

generate the `topGO` data structure were `BP` for the `ontology` parameter and a `nodeSize` of `10`. For the enrichment test, the `classic` algorithm was used with the statistic parameter set to `fisher`. The significance threshold used was 0.01.

6.11 Protein domain enrichment

The `rpstblastn` binary provided by NCBI[466] (version 2.2.30+), was run with the Conserved Domain Database[480] (accessed 2015-04-25) to identify conserved protein domains in the *B. napus* gene models identified by AUGUSTUS. An e-value of 0.01 was used, and the `rpsbproc` utility used to filter the results by removing overlapping domain identifications. The `fisher.test` function in R[467] was used to perform Fisher's exact test to test for enrichment of protein domains of interest, with a `greater` alternative hypothesis. The significance threshold used was 0.01.

6.12 BnFD probability of dimerization calculation

The protein sequence of *BnFD* genes was determined by performing DNA sequence alignment to the Arabidopsis *FD* gene using the MUSCLE multiple sequence alignment tool[481] within AliView[482] (version 1.16). Intron-exon boundaries were manually assessed and the DNA sequence translated within AliView. DrawCoil[483] (version 1.0) was run with default parameters to generate the helical wheel diagrams depicted in figure 2.39. The trained scoring script described in Potapov et al. (2015)[320] (Amy E. Keating, personal communication, 2016-05-10) was run with every combination of BnFD dimer.

6.13 BnFD DNA binding predictions

The protein structure of the CREB protein (PDB ID: 1DH3) from Schumacher et al. (2000)[319] was downloaded. Based on sequence alignment, the amino acids in positions 286 and 287 of the crystal structure were modified to match the BnFD protein amino acids in those positions. For Arabidopsis FD, BnFD.A1, BnFD.C1, and BnFD.A8, an arginine was used in position 286 and a histidine in position 287. For BnFD.C7 and BnFD.Ann.Random, an arginine was used in position 286 and an asparagine used in position 287. For BnFD.C3.Random histidines were used in both positions. These modified structures were imported into Jmol[484] and the commands `minimize ADDHYDROGENS` and `calculate HBONDS` were used consecutively to predict hydrogen bonding.

6.14 Mathematical modelling of BnFD dimerization dynamics

To model the dynamics of BnFD dimerization, the law of mass action was assumed. Concentrations of monomers and dimers were modelled using the following system of equations:

$$\mathrm{a} + \mathrm{a} \underset{k_{-aa}}{\overset{k_{+aa}}{\rightleftharpoons}} \mathrm{aa}$$

$$\mathrm{a} + \mathrm{b} \underset{k_{-ab}}{\overset{k_{+ab}}{\rightleftharpoons}} \mathrm{ab}$$

$$\mathrm{b} + \mathrm{b} \underset{k_{-bb}}{\overset{k_{+bb}}{\rightleftharpoons}} \mathrm{bb}$$

$$\frac{d[\mathrm{a}]}{dt} = k_{-ab}[\mathrm{ab}] + 2k_{-aa}[\mathrm{aa}] - k_{+ab}[\mathrm{a}][\mathrm{b}] - 2k_{+aa}[a]^2$$

$$\frac{d[\mathrm{b}]}{dt} = k_{-ab}[\mathrm{ab}] + 2k_{-bb}[\mathrm{bb}] - k_{+ab}[\mathrm{a}][\mathrm{b}] - 2k_{+bb}[b]^2$$

$$\frac{d[\mathrm{aa}]}{dt} = k_{+aa}[\mathrm{a}]^2 - k_{-aa}[\mathrm{aa}]$$

$$\frac{d[\mathrm{ab}]}{dt} = k_{+ab}[\mathrm{a}][\mathrm{b}] - k_{-ab}[\mathrm{ab}]$$

$$\frac{d[\mathrm{bb}]}{dt} = k_{+bb}[\mathrm{b}]^2 - k_{-bb}[\mathrm{bb}]$$

Where [x] is the concentration of the monomer `x`, [yz] is the concentration of the dimer `yz`, k_{+yz} is the forward reaction rate for the creation of dimer `yz`, and k_{-yz} is the reverse reaction rate for the destruction of dimer `yz`. Initial concentrations used were 50 for each monomer, and 0 for each dimer. The constant reaction rates used were:

$k_{+aa} = 7$

$k_{-aa} = 1$

$k_{-ab} = 1$

$k_{-bb} = 1$

The value of k_{+bb} was either 0.5, 4, or 7, depending on the simulation run. Values of k_{+ab} were increased from 0 to 7 in 0.2 increments. At each increment, the simulation was run until equilibrium and the steady state concentrations recorded. These simulations were performed using the `deSolve` library[485] (version 1.13) using the R statistical programming language[467].

6.15 Correlation analysis

The correlation analysis used expression levels for all genes. The `cor` function in the R statistical programming language[467] was used to calculate Pearson correlation coefficients between time points using vectors of FPKM values from each time point.

Bibliography

1. Di Paola, A., Valentini, R. & Santini, M. An overview of available crop growth and yield models for studies and assessments in agriculture. *Journal of the Science of Food and Agriculture*. **96,** 709–714 (2016).

2. Whisler, F. D. *et al.* Crop simulation models in agronomic systems. *Advances in Agronomy*. **40,** 141–208 (1986).

3. Jung, C. & Müller, A. E. Flowering time control and applications in plant breeding. *Trends in Plant Science*. **14,** 563–573 (2009).

4. Flavell, R. Role of model plant species in *Plant Genomics*. 1–18 (Humana Press, 2009).

5. Koornneef, M. & Meinke, D. The development of Arabidopsis as a model plant. *The Plant Journal*. **61,** 909–921 (2010).

6. Redei, G. P. *Arabidopsis* as a genetic tool. *Annual Review of Genetics*. **9,** 111–127 (1975).

7. Kim, Y., Schumaker, K. S. & Zhu, J.-K. EMS mutagenesis of *Arabidopsis* in *Arabidopsis Protocols*. 101–103 (Humana Press, 2006).

8. Weigel, D. & Meyerowitz, E. M. The ABCs of floral homeotic genes. *Cell*. **78,** 203–209 (1994).

9. Nester, E. W. *Agrobacterium*: Nature's genetic engineer. *Frontiers in Plant Science*. **5,** 730 (2015).

10. The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*. **408,** 796–815 (2000).

11. The C. elegans Sequencing Consortium. Genome sequence of the nematode *C. elegans*: A platform for investigating biology. *Science*. **282,** 2012–2018 (1998).

12. Adams, M. D. *et al.* The genome sequence of *Drosophila melanogaster*. *Science*. **287,** 2185–2195 (2000).

13. Andrés, F. & Coupland, G. The genetic basis of flowering responses to seasonal cues. *Nature Reviews Genetics*. **13,** 627–639 (2012).

14. Boss, P. K., Bastow, R. M., Mylne, J. S. & Dean, C. Multiple pathways in the decision to flower: Enabling, promoting, and resetting. *The Plant Cell*. **16,** S18–S31 (2004).

15. Srikanth, A. & Schmid, M. Regulation of flowering time: All roads lead to Rome. *Cellular and Molecular Life Sciences*. **68,** 2013–2037 (2011).

16. McClung, C. R. Plant circadian rhythms. *The Plant Cell*. **18,** 792–803 (2006).

17. An, H. *et al.* CONSTANS acts in the phloem to regulate a systemic signal that induces photoperiodic flowering of *Arabidopsis*. *Development*. **131,** 3615–3626 (2004).

18. Suárez-López, P. *et al. CONSTANS* mediates between the circadian clock and the control of flowering in *Arabidopsis*. *Nature*. **410,** 1116–1120 (2001).

19. Valverde, F. *et al.* Photoreceptor regulation of CONSTANS protein in photoperiodic flowering. *Science*. **303,** 1003–1006 (2004).

20. Samach, A. *et al.* Distinct roles of CONSTANS target genes in reproductive development of *Arabidopsis*. *Science*. **288,** 1613–1616 (2000).

21. Yoo, S. K. *et al. CONSTANS* activates *SUPPRESSOR OF OVEREXPRESSION OF CONSTANS 1* through *FLOWERING LOCUS T* to promote flowering in Arabidopsis. *Plant Physiology*. **139,** 770–778 (2005).

22. Kobayashi, Y., Kaya, H., Goto, K., Iwabuchi, M. & Araki, T. A pair of related genes with antagonistic roles in mediating flowering signals. *Science*. **286,** 1960–1962 (1999).

23. Tiwari, S. B. *et al.* The flowering time regulator CONSTANS is recruited to the *FLOWERING LOCUS T* promoter via a unique cis-element. *The New Phytologist.* **187,** 57–66 (2010).

24. Shindo, C., Bernasconi, G. & Hardtke, C. S. Natural genetic variation in Arabidopsis: Tools, traits and prospects for evolutionary ecology. *Annals of Botany.* **99,** 1043–1054 (2007).

25. Koornneef, M., Alonso-Blanco, C. & Vreugdenhil, D. Naturally occurring genetic variation in *Arabidopsis thaliana*. *Annual Review of Plant Biology.* **55,** 141–172 (2004).

26. Thompson, L. The spatiotemporal effects of nitrogen and litter on the population dynamics of *Arabidopsis thaliana*. *Journal of Ecology.* **82,** 63–68 (1994).

27. Song, J., Irwin, J. & Dean, C. Remembering the prolonged cold of winter. *Current Biology.* **23,** R807–R811 (2013).

28. Shindo, C. *et al.* Role of *FRIGIDA* and *FLOWERING LOCUS C* in determining variation in flowering time of Arabidopsis. *Plant Physiology.* **138,** 1163–1173 (2005).

29. Michaels, S. D. & Amasino, R. M. *FLOWERING LOCUS C* encodes a novel MADS domain protein that acts as a repressor of flowering. *The Plant Cell.* **11,** 949–956 (1999).

30. Helliwell, C. A., Wood, C. C., Robertson, M., James Peacock, W. & Dennis, E. S. The Arabidopsis FLC protein interacts directly *in vivo* with *SOC1* and *FT* chromatin and is part of a high-molecular-weight protein complex. *The Plant Journal.* **46,** 183–192 (2006).

31. Searle, I. *et al.* The transcription factor FLC confers a flowering response to vernalization by repressing meristem competence and systemic signaling in *Arabidopsis*. *Genes & Development.* **20,** 898–912 (2006).

32. Hyun, K.-g., Noh, Y.-S. & Song, J.-J. *Arabidopsis* FRIGIDA stimulates EFS histone H3 Lys36 methyltransferase activity. *Plant Cell Reports.* **36,** 1183–1185 (2017).

33. Kim, S. Y. *et al.* Establishment of the vernalization-responsive, winter-annual habit in *Arabidopsis* requires a putative histone H3 methyl transferase. *The Plant Cell.* **17,** 3301–3310 (2005).

34. Simpson, G. G. The autonomous pathway: Epigenetic and post-transcriptional gene regulation in the control of *Arabidopsis* flowering time. *Current Opinion in Plant Biology.* **7,** 570–574 (2004).

35. Koornneef, M., Hanhart, C. J. & Veen, J. H. van der. A genetic and physiological analysis of late flowering mutants in *Arabidopsis thaliana*. *Molecular and General Genetics MGG.* **229,** 57–66 (1991).

36. Amasino, R. M. & Michaels, S. D. The timing of flowering. *Plant Physiology.* **154,** 516–520 (2010).

37. Davis, S. J. Integrating hormones into the floral-transition pathway of *Arabidopsis thaliana*. *Plant, Cell & Environment.* **32,** 1201–1210 (2009).

38. Wilson, R. N., Heckman, J. W. & Somerville, C. R. Gibberellin is required for flowering in *Arabidopsis thaliana* under short days. *Plant Physiology.* **100,** 403–408 (1992).

39. Spanudakis, E. & Jackson, S. The role of microRNAs in the control of flowering time. *Journal of Experimental Botany.* **65,** 365–380 (2014).

40. Yamaguchi, A. & Abe, M. Regulation of reproductive development by non-coding RNA in Arabidopsis: To flower or not to flower. *Journal of Plant Research.* **125,** 693–704 (2012).

41. Jaeger, K. E., Pullen, N., Lamzin, S., Morris, R. J. & Wigge, P. A. Interlocking feedback loops govern the dynamic behavior of the floral transition in *Arabidopsis*. *The Plant Cell.* **25,** 820–833 (2013).

42. Aksenova, N. P., Milyaeva, E. L. & Romanov, G. A. Florigen goes molecular: Seventy years of the hormonal theory of flowering regulation. *Russian Journal of Plant Physiology.* **53,** 401–406 (2006).

43. Lang, A., Chailakhyan, M. K. & Frolova, I. A. Promotion and inhibition of flower formation in a dayneutral plant in grafts with a short-day plant and a long-day plant. *Proceedings of the National Academy of Sciences.* **74,** 2412–2416 (1977).

44. Jaeger, K. E. & Wigge, P. A. FT protein acts as a long-range signal in *Arabidopsis*. *Current Biology*. **17,** 1050–1054 (2007).

45. Mathieu, J., Warthmann, N., Küttner, F. & Schmid, M. Export of FT protein from phloem companion cells is sufficient for floral induction in *Arabidopsis*. *Current Biology*. **17,** 1055–1060 (2007).

46. Notaguchi, M. *et al.* Long-distance, graft-transmissible action of *Arabidopsis* FLOWERING LOCUS T protein to promote flowering. *Plant and Cell Physiology*. **49,** 1645–1658 (2008).

47. Wigge, P. A. *et al.* Integration of spatial and temporal information during floral induction in *Arabidopsis*. *Science*. **309,** 1056–1059 (2005).

48. Moon, J., Lee, H., Kim, M. & Lee, I. Analysis of flowering pathway integrators in *Arabidopsis*. *Plant and Cell Physiology*. **46,** 292–299 (2005).

49. Abe, M. *et al.* FD, a bZIP protein mediating signals from the floral pathway integrator FT at the shoot apex. *Science*. **309,** 1052–1056 (2005).

50. Alvarez, J., Guli, C. L., Yu, X.-H. & Smyth, D. R. *Terminal flower*: A gene affecting inflorescence development in *Arabidopsis thaliana*. *The Plant Journal*. **2,** 103–116 (1992).

51. Shannon, S. & Meeks-Wagner, D. R. A mutation in the Arabidopsis *TFL1* gene affects inflorescence meristem development. *The Plant Cell*. **3,** 877–892 (1991).

52. Bradley, D., Ratcliffe, O., Vincent, C., Carpenter, R. & Coen, E. Inflorescence commitment and architecture in *Arabidopsis*. *Science*. **275,** 80–83 (1997).

53. Shannon, S. & Meeks-Wagner, D. R. Genetic interactions that regulate inflorescence development in Arabidopsis. *The Plant Cell*. **5,** 639–655 (1993).

54. Gustafson-Brown, C., Savidge, B. & Yanofsky, M. F. Regulation of the Arabidopsis floral homeotic gene *APETALA1*. *Cell*. **76,** 131–143 (1994).

55. Ratcliffe, O. J., Bradley, D. J. & Coen, E. S. Separation of shoot and floral identity in *Arabidopsis*. *Development*. **126,** 1109–1120 (1999).

56. Liljegren, S. J., Gustafson-Brown, C., Pinyopich, A., Ditta, G. S. & Yanofsky, M. F. Interactions among *APETALA1*, *LEAFY*, and *TERMINAL FLOWER1* specify meristem fate. *The Plant Cell*. **11,** 1007–1018 (1999).

57. Ho, W. W. H. & Weigel, D. Structural features determining flower-promoting activity of *Arabidopsis* FLOWERING LOCUS T. *The Plant Cell*. **26,** 552–564 (2014).

58. Ahn, J. H. *et al.* A divergent external loop confers antagonistic activity on floral regulators FT and TFL1. *The EMBO Journal*. **25,** 605–614 (2006).

59. Hanzawa, Y., Money, T. & Bradley, D. A single amino acid converts a repressor to an activator of flowering. *Proceedings of the National Academy of Sciences of the United States of America*. **102,** 7748–7753 (2005).

60. Schultz, E. A. & Haughn, G. W. LEAFY, a homeotic gene that regulates inflorescence development in Arabidopsis. *The Plant Cell*. **3,** 771–781 (1991).

61. Huala, E. & Sussex, I. *LEAFY* interacts with floral homeotic genes to regulate Arabidopsis floral development. *The Plant Cell*. **4,** 901–913 (1992).

62. Hamès, C. *et al.* Structural basis for LEAFY floral switch function and similarity with helix-turn-helix- proteins. *The EMBO Journal*. **27,** 2628–2637 (2008).

63. Hempel, F. D. *et al.* Floral determination and expression of floral regulatory genes in *Arabidopsis*. *Development*. **124,** 3845–3853 (1997).

64. Wagner, D., Sablowski, R. W. M. & Meyerowitz, E. M. Transcriptional activation of APETALA1 by LEAFY. *Science*. **285,** 582–584 (1999).

65. William, D. A. *et al.* Genomic identification of direct target genes of LEAFY. *Proceedings of the National Academy of Sciences of the United States of America*. **101,** 1775–1780 (2004).

66. Weigel, D. & Nilsson, O. A developmental switch sufficient for flower initiation in diverse plants. *Nature*. **377,** 495–500 (1995).

67. Weigel, D. & Meyerowitz, E. M. Activation of floral homeotic genes in *Arabidopsis*. *Science*. **261,** 1723–1726 (1993).

68. Hong, R. L., Hamaguchi, L., Busch, M. A. & Weigel, D. Regulatory elements of the floral homeotic gene *AGAMOUS* identified by phylogenetic footprinting and shadowing. *The Plant Cell.* **15,** 1296–1309 (2003).

69. Lee, J., Oh, M., Park, H. & Lee, I. SOC1 translocated to the nucleus by interaction with AGL24 directly regulates *LEAFY*. *The Plant Journal.* **55,** 832–843 (2008).

70. Winter, C. M. *et al.* LEAFY target genes reveal floral regulatory logic, cis motifs, and a link to biotic stimulus response. *Developmental Cell.* **20,** 430–443 (2011).

71. Blazquez, M. A., Soowal, L. N., Lee, I. & Weigel, D. *LEAFY* expression and flower initiation in *Arabidopsis*. *Development.* **124,** 3835–3844 (1997).

72. Okamuro, J. K., Boer, B. G. W. den, Lotys-Prass, C., Szeto, W. & Jofuku, K. D. Flowers into shoots: Photo and hormonal control of a meristem identity switch in Arabidopsis. *Proceedings of the National Academy of Sciences.* **93,** 13831–13836 (1996).

73. Eriksson, S., Böhlenius, H., Moritz, T. & Nilsson, O. GA_4 is the active gibberellin in the regulation of *LEAFY* transcription and *Arabidopsis* floral initiation. *The Plant Cell.* **18,** 2172–2181 (2006).

74. Alejandra Mandel, M., Gustafson-Brown, C., Savidge, B. & Yanofsky, M. F. Molecular characterization of the *Arabidopsis* floral homeotic gene *APETALA1*. *Nature.* **360,** 273–277 (1992).

75. Koornneeff, M., Dellaert, L. W. M. & Veen, J. H. van der. EMS- and relation-induced mutation frequencies at individual loci in *Arabidopsis thaliana* (L.) Heynh. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis.* **93,** 109–123 (1982).

76. Irish, V. F. & Sussex, I. M. Function of the apetala-1 gene during Arabidopsis floral development. *The Plant Cell.* **2,** 741–753 (1990).

77. Mandel, M. A. & Yanofsky, M. F. A gene triggering flower formation in *Arabidopsis*. *Nature.* **377,** 522–524 (1995).

78. Han, Y., Zhang, C., Yang, H. & Jiao, Y. Cytokinin pathway mediates *APETALA1* function in the establishment of determinate floral meristems in

Arabidopsis. *Proceedings of the National Academy of Sciences*. **111,** 6840–6845 (2014).

79. Kaufmann, K. *et al.* Orchestration of floral initiation by APETALA1. *Science*. **328,** 85–89 (2010).

80. Weigel, D., Alvarez, J., Smyth, D. R., Yanofsky, M. F. & Meyerowitz, E. M. *LEAFY* controls floral meristem identity in Arabidopsis. *Cell*. **69,** 843–859 (1992).

81. Liu, C. *et al.* Specification of *Arabidopsis* floral meristem identity by repression of flowering time genes. *Development*. **134,** 1901–1910 (2007).

82. Gregis, V., Sessa, A., Colombo, L. & Kater, M. M. *AGAMOUS-LIKE24* and *SHORT VEGETATIVE PHASE* determine floral meristem identity in Arabidopsis. *The Plant Journal*. **56,** 891–902 (2008).

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

84. Onouchi, H., Igeño, M. I., Périlleux, C., Graves, K. & Coupland, G. Mutagenesis of plants overexpressing *CONSTANS* demonstrates novel interactions among Arabidopsis flowering-time genes. *The Plant Cell*. **12,** 885–900 (2000).

85. Lee, H. *et al.* The AGAMOUS-LIKE 20 MADS domain protein integrates floral inductive pathways in *Arabidopsis*. *Genes & Development*. **14,** 2366–2376 (2000).

86. Hepworth, S. R., Valverde, F., Ravenscroft, D., Mouradov, A. & Coupland, G. Antagonistic regulation of flowering-time gene *SOC1* by CONSTANS and FLC via separate promoter motifs. *The EMBO Journal*. **21,** 4327–4337 (2002).

87. Moon, J. *et al.* The *SOC1* MADS-box gene integrates vernalization and gibberellin signals for flowering in *Arabidopsis*. *The Plant Journal*. **35,** 613–623 (2003).

88. Seo, E. *et al.* Crosstalk between cold response and flowering in *Arabidopsis* is mediated through the flowering-time gene *SOC1* and its upstream negative regulator *FLC*. *The Plant Cell*. **21,** 3185–3197 (2009).

89. Wang, J.-W., Czech, B. & Weigel, D. miR156-regulated SPL transcription factors define an endogenous flowering pathway in *Arabidopsis thaliana*. *Cell*. **138,** 738–749 (2009).

90. Liu, C. *et al.* Direct interaction of AGL24 and SOC1 integrates flowering signals in *Arabidopsis*. *Development*. **135,** 1481–1491 (2008).

91. Gregis, V., Sessa, A., Dorca-Fornell, C. & Kater, M. M. The Arabidopsis floral meristem identity genes AP1, AGL24 and SVP directly repress class B and C floral homeotic genes. *The Plant Journal*. **60,** 626–637 (2009).

92. Melzer, S. *et al.* Flowering-time genes modulate meristem determinacy and growth form in *Arabidopsis thaliana*. *Nature Genetics*. **40,** 1489–1492 (2008).

93. Teper-Bamnolker, P. & Samach, A. The flowering integrator FT regulates *SEPALLATA3* and *FRUITFULL* accumulation in *Arabidopsis* leaves. *The Plant Cell*. **17,** 2661–2675 (2005).

94. Gu, Q., Ferrandiz, C., Yanofsky, M. F. & Martienssen, R. The *FRUITFULL* MADS-box gene mediates cell differentiation during *Arabidopsis* fruit development. *Development*. **125,** 1509–1517 (1998).

95. Ferrandiz, C., Gu, Q., Martienssen, R. & Yanofsky, M. F. Redundant regulation of meristem identity and plant architecture by *FRUITFULL*, *APETALA1* and *CAULIFLOWER*. *Development*. **127,** 725–734 (2000).

96. Folter, S. de *et al.* Comprehensive interaction map of the Arabidopsis MADS box transcription factors. *The Plant Cell*. **17,** 1424–1433 (2005).

97. Balanzà, V., Martínez-Fernández, I. & Ferrándiz, C. Sequential action of *FRUITFULL* as a modulator of the activity of the floral regulators *SVP* and *SOC1*. *Journal of Experimental Botany*. **65,** 1193–1203 (2014).

98. Gregis, V. *et al.* Identification of pathways directly regulated by SHORT VEGETATIVE PHASE during vegetative and reproductive development in *Arabidopsis*. *Genome Biology*. **14,** R56 (2013).

99. Mateos, J. L. *et al.* Combinatorial activities of SHORT VEGETATIVE PHASE and FLOWERING LOCUS C define distinct modes of flowering regulation in *Arabidopsis*. *Genome Biology*. **16,** 31 (2015).

100. Jang, S., Torti, S. & Coupland, G. Genetic and spatial interactions between *FT*, *TSF* and *SVP* during the early stages of floral induction in Arabidopsis. *The Plant Journal*. **60,** 614–625 (2009).

101. Immink, R. G. H. *et al.* Characterization of SOC1's central role in flowering by the identification of its upstream and downstream regulators. *Plant Physiology*. **160,** 433–449 (2012).

102. Liu, C., Xi, W., Shen, L., Tan, C. & Yu, H. Regulation of floral patterning by flowering time genes. *Developmental Cell*. **16,** 711–722 (2009).

103. Simpson, G. G. & Dean, C. *Arabidopsis*, the Rosetta Stone of flowering time? *Science*. **296,** 285–289 (2002).

104. Al-Shehbaz, I. A. A generic and tribal synopsis of the Brassicaceae (Cruciferae). *Taxon*. **61,** 931–954 (2012).

105. Cartea, M., Lema, M., Francisco, M. & Velasco, P. Basic information on vegetable Brassica crops in *Genetics, Genomics and Breeding of Vegetable Brassicas*. 1–33 (Science Publishers, 2011).

106. U, N. Genome analysis in *Brassica* with special reference to the experimental formation of *B. napus* and peculiar mode of fertilization. *Japanese Journal of Botany*. **7,** 389–452 (1935).

107. Rana, D. *et al.* Conservation of the microstructure of genome segments in *Brassica napus* and its diploid relatives. *The Plant Journal*. **40,** 725–733 (2004).

108. Allender, C. J. & King, G. J. Origins of the amphiploid species *Brassica napus* L. investigated by chloroplast and nuclear molecular markers. *BMC Plant Biology*. **10,** 54 (2010).

109. EST: Oilcrops, oils and meals market assessment. Available at: http://www.fao.org/economic/est/est-commodities/oilcrops/oilcrops-oils-and-meals-market-assessment/en. (Accessed: 6th September 2017).

110. Agriculture in the United Kingdom 2016. Available at: https://www.gov.uk/government/statistics/agriculture-in-the-united-kingdom-2016. (Accessed: 25th August 2017).

111. Christen, O. & Sieling, K. The effect of different preceding crops on the development, growth and yield of winter barley. *Journal of Agronomy and Crop Science*. **171,** 114–123 (1993).

112. Beilstein, M. A., Nagalingum, N. S., Clements, M. D., Manchester, S. R. & Mathews, S. Dated molecular phylogenies indicate a Miocene origin for *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences of the United States of America*. **107,** 18724–18728 (2010).

113. Lysak, M. A., Koch, M. A., Pecinka, A. & Schubert, I. Chromosome triplication found across the tribe *Brassiceae*. *Genome Research*. **15,** 516–525 (2005).

114. Cheung, F. *et al.* Comparative analysis between homoeologous genome segments of *Brassica napus* and its progenitor species reveals extensive sequence-level divergence. *The Plant Cell*. **21,** 1912–1928 (2009).

115. Inaba, R. & Nishio, T. Phylogenetic analysis of *Brassiceae* based on the nucleotide sequences of the S-locus related gene, *SLR1*. *Theoretical and Applied Genetics*. **105,** 1159–1165 (2002).

116. The Brassica rapa Genome Sequencing Project Consortium *et al.* The genome of the mesopolyploid crop species *Brassica rapa*. *Nature Genetics*. **43,** 1035–1039 (2011).

117. Liu, S. *et al.* The *Brassica oleracea* genome reveals the asymmetrical evolution of polyploid genomes. *Nature Communications*. **5,** 3930 (2014).

118. Chalhoub, B. *et al.* Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome. *Science*. **345,** 950–953 (2014).

119. Matschegewski, C. *et al.* Genetic variation of temperature-regulated curd induction in cauliflower: Elucidation of floral transition by genome-wide association mapping and gene expression analysis. *Crop Science and Horticulture*. **6,** 720 (2015).

120. Li, Z. *et al.* Molecular cloning and characterization of an anti-bolting related gene (*BrpFLC*) from *Brassica rapa* ssp. *Pekinensis*. *Plant Science*. **168,** 407–413 (2005).

121. Schiessl, S., Iniguez-Luy, F., Qian, W. & Snowdon, R. J. Diverse regulatory factors associate with flowering time and yield responses in winter-type *Brassica napus*. *BMC Genomics*. **16,** 737 (2015).

122. Shi, J. *et al.* Unraveling the complex trait of crop yield with quantitative trait loci mapping in *Brassica napus*. *Genetics*. **182,** 851–861 (2009).

123. Mendham, N. J., Shipway, P. A. & Scott, R. K. The effects of delayed sowing and weather on growth, development and yield of winter oil-seed rape (*Brassica napus*). *The Journal of Agricultural Science*. **96,** 389–416 (1981).

124. Oilseed rape guide. Available at: https://cereals.ahdb.org.uk/publications/2015/june/10/oilseed-rape-guide.aspx. (Accessed: 25th August 2017).

125. Habekotté, B. Quantitative analysis of pod formation, seed set and seed filling in winter oilseed rape (*Brassica napus* L.) under field conditions. *Field Crops Research*. **35,** 21–33 (1993).

126. Habekotté, B. Options for increasing seed yield of winter oilseed rape (*Brassica napus* L.): A simulation study. *Field Crops Research*. **54,** 109–126 (1997).

127. Friedt, W. & Snowdon, R. Oilseed rape in *Oil Crops*. 91–126 (Springer, New York, NY, 2009).

128. Stanley, D. A., Gunning, D. & Stout, J. C. Pollinators and pollination of oilseed rape crops (*Brassica napus* L.) in Ireland: Ecological and economic incentives for pollinator conservation. *Journal of Insect Conservation*. **17,** 1181–1189 (2013).

129. Zou, Y. *et al.* Wild pollinators enhance oilseed rape yield in small-holder farming systems in China. *BMC Ecology*. **17,** 6 (2017).

130. Rafferty, N. E. & Ives, A. R. Pollinator effectiveness varies with experimental shifts in flowering time. *Ecology*. **93,** 803–814 (2012).

131. Schiessl, S., Samans, B., Hüttel, B., Reinhard, R. & Snowdon, R. J. Capturing sequence variation among flowering-time regulatory gene homologs in the allopolyploid crop species *Brassica napus*. *Frontiers in Plant Science*. **5,** 404 (2014).

132. Zhao, J. *et al. BrFLC2* (*FLOWERING LOCUS C*) as a candidate gene for a vernalization response QTL in *Brassica rapa*. *Journal of Experimental Botany*. **61,** 1817–1825 (2010).

133. Kole, C., Quijada, P., Michaels, S. D., Amasino, R. M. & Osborn, T. C. Evidence for homology of flowering-time genes *VFR2* from *Brassica rapa* and *FLC* from *Arabidopsis thaliana*. *Theoretical and Applied Genetics*. **102,** 425–430 (2001).

134. Osborn, T. C. *et al.* Comparison of flowering time genes in *Brassica rapa*, *B. napus* and *Arabidopsis thaliana*. *Genetics*. **146,** 1123–1129 (1997).

135. Lou, P. *et al.* Quantitative trait loci for flowering time and morphological traits in multiple populations of *Brassica rapa*. *Journal of Experimental Botany*. **58,** 4005–4016 (2007).

136. Axelsson, T., Shavorskaya, O. & Lagercrantz, U. Multiple flowering time QTLs within several *Brassica* species could be the result of duplicated copies of one ancestral gene. *Genome*. **44,** 856–864 (2001).

137. Schranz, M. E. *et al.* Characterization and effects of the replicated flowering time gene *FLC* in *Brassica rapa*. *Genetics*. **162,** 1457–1468 (2002).

138. Okazaki, K. *et al.* Mapping and characterization of *FLC* homologs and QTL analysis of flowering time in *Brassica oleracea*. *Theoretical and Applied Genetics*. **114,** 595–608 (2007).

139. Ridge, S., Brown, P. H., Hecht, V., Driessen, R. G. & Weller, J. L. The role of *BoFLC2* in cauliflower (*Brassica oleracea* var. *botrytis* L.) reproductive development. *Journal of Experimental Botany*. **66,** 125–135 (2015).

140. Irwin, J. A. *et al.* Nucleotide polymorphism affecting *FLC* expression underpins heading date variation in horticultural brassicas. *The Plant Journal*. **87,** 597–605 (2016).

141. Zou, X. *et al.* Comparative analysis of *FLC* homologues in Brassicaceae provides insight into their role in the evolution of oilseed rape. *PLoS ONE*. **7,** e45751 (2012).

142. Wang, N. *et al.* Flowering time variation in oilseed rape (*Brassica napus* L.) is associated with allelic variation in the *FRIGIDA* homologue *BnaA.FRI.a*. *Journal of Experimental Botany*. **62,** 5641–5658 (2011).

143. Long, Y. *et al.* Flowering time quantitative trait loci analysis of oilseed Brassica in multiple environments and genomewide alignment with Arabidopsis. *Genetics*. **177,** 2433–2444 (2007).

144. Lin, S.-I. *et al.* Differential regulation of *FLOWERING LOCUS C* expression by vernalization in cabbage and Arabidopsis. *Plant Physiology*. **137,** 1037–1048 (2005).

145. Tadege, M. *et al.* Control of flowering time by *FLC* orthologues in *Brassica napus*. *The Plant Journal*. **28,** 545–553 (2001).

146. Kim, S.-Y. *et al.* Delayed flowering time in *Arabidopsis* and *Brassica rapa* by the overexpression of *FLOWERING LOCUS C* (*FLC*) homologs isolated from Chinese cabbage (*Brassica rapa* L. ssp. *pekinensis*). *Plant Cell Reports*. **26,** 327–336 (2007).

147. Irwin, J. A. *et al.* Functional alleles of the flowering time regulator *FRIGIDA* in the *Brassica oleracea* genome. *BMC Plant Biology*. **12,** 21 (2012).

148. Lou, P. *et al.* Preferential retention of circadian clock genes during diploidization following whole genome triplication in *Brassica rapa*. *The Plant Cell*. **24,** 2415–2426 (2012).

149. Bohuon, E. J. R. *et al.* The association of flowering time quantitative trait loci with duplicated regions and candidate loci in *Brassica oleracea*. *Genetics*. **150,** 393–401 (1998).

150. Lagercrantz, U., Putterill, J., Coupland, G. & Lydiate, D. Comparative mapping in *Arabidopsis* and *Brassica*, fine scale genome collinearity and congruence of genes controlling flowering time. *The Plant Journal*. **9,** 13–20 (1996).

151. Österberg, M. K., Shavorskaya, O., Lascoux, M. & Lagercrantz, U. Naturally occurring indel variation in the *Brassica nigra COL1* gene is associated with variation in flowering time. *Genetics*. **161,** 299–306 (2002).

152. Mimida, N., Sakamoto, W., Murata, M. & Motoyoshi, F. *TERMINAL FLOWER 1*-like genes in *Brassica* species. *Plant Science*. **142,** 155–162 (1999).

153. Guo, Y., Hans, H., Christian, J. & Molina, C. Mutations in single *FT*- and *TFL1*-paralogs of rapeseed (*Brassica napus* L.) and their impact on flowering time and yield components. *Frontiers in Plant Science*. **5,** 282 (2014).

154. Wang, J. *et al.* Promoter variation and transcript divergence in Brassicaceae lineages of *FLOWERING LOCUS T*. *PLOS ONE*. **7,** e47127 (2012).

155. Zhang, X. *et al.* A transposon insertion in *FLOWERING LOCUS T* is associated with delayed flowering in *Brassica rapa*. *Plant Science*. **241,** 211–220 (2015).

156. Duclos, D. V. & Björkman, T. Meristem identity gene expression during curd proliferation and flower initiation in *Brassica oleracea*. *Journal of Experimental Botany*. **59,** 421–433 (2008).

157. Franks, S. J. *et al.* Variation in the flowering time orthologs *BrFLC* and *BrSOC1* in a natural population of *Brassica rapa*. *PeerJ*. **3,** e1339 (2015).

158. Sri, T., Mayee, P. & Singh, A. Sequence and expression variation in *SUPPRESSOR of OVEREXPRESSION of CONSTANS 1* (*SOC1*): Homeolog evolution in Indian Brassicas. *Development Genes and Evolution*. **225,** 287–303 (2015).

159. Kholodenko, B. N. Cell-signalling dynamics in time and space. *Nature Reviews Molecular Cell Biology*. **7,** 165–176 (2006).

160. Turing, A. M. The chemical basis of morphogenesis. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*. **237,** 37–72 (1952).

161. Wangersky, P. J. Lotka-Volterra population models. *Annual Review of Ecology and Systematics*. **9,** 189–218 (1978).

162. Valentim, F. L. *et al.* A quantitative and dynamic model of the Arabidopsis flowering time gene regulatory network. *PLOS ONE*. **10,** e0116973 (2015).

163. Karlebach, G. & Shamir, R. Modelling and analysis of gene regulatory networks. *Nature Reviews Molecular Cell Biology.* **9,** 770–780 (2008).

164. Tyson, J. J., Chen, K. C. & Novak, B. Sniffers, buzzers, toggles and blinkers: Dynamics of regulatory and signaling pathways in the cell. *Current Opinion in Cell Biology.* **15,** 221–231 (2003).

165. Alon, U. Network motifs: Theory and experimental approaches. *Nature Reviews Genetics.* **8,** 450–461 (2007).

166. Emmert-Streib, F., Dehmer, M. & Haibe-Kains, B. Gene regulatory networks and their applications: Understanding biological and medical problems in terms of networks. *Frontiers in Cell and Developmental Biology.* **2,** 38 (2014).

167. Satake, A. *et al.* Forecasting flowering phenology under climate warming by modelling the regulatory dynamics of flowering-time genes. *Nature Communications.* **4,** 2303 (2013).

168. Espinosa-Soto, C., Padilla-Longoria, P. & Alvarez-Buylla, E. R. A gene regulatory network model for cell-fate determination during *Arabidopsis thaliana* flower development that is robust and recovers experimental gene expression profiles. *The Plant Cell.* **16,** 2923–2939 (2004).

169. Chaos, Á. *et al.* From genes to flower patterns and evolution: Dynamic models of gene regulatory networks. *Journal of Plant Growth Regulation.* **25,** 278–289 (2006).

170. Dong, Z. *et al.* A gene regulatory network model for floral transition of the shoot apex in Maize and its dynamic modeling. *PLOS ONE.* **7,** e43450 (2012).

171. Bouman, B. A. M., Keulen, H. van, Laar, H. H. van & Rabbinge, R. The 'School of de Wit' crop growth simulation models: A pedigree and historical overview. *Agricultural Systems.* **52,** 171–198 (1996).

172. de Wit, C. T. Photosynthesis of leaf canopies. *Agricultural Research Report No. 663.* (1965).

173. Dingkuhn, M., Vries, F. W. T. P. D. & Miezan, K. M. Improvement of rice plant type concepts: Systems research enables interaction of physiology and

breeding in *Systems approaches for agricultural development*. 19–35 (Springer, Dordrecht, 1993).

174. Stone, R. C. & Meinke, H. Operational seasonal forecasting of crop performance. *Philosophical Transactions of the Royal Society B: Biological Sciences*. **360,** 2109–2124 (2005).

175. Boote, K. J., Jones, J. W. & Pickering, N. B. Potential uses and limitations of crop models. *Agronomy Journal*. **88,** 704–716 (1996).

176. Shibu, M. E., Leffelaar, P. A., Keulen, H. van & Aggarwal, P. K. LINTUL3, a simulation model for nitrogen-limited situations: Application to rice. *European Journal of Agronomy*. **32,** 255–271 (2010).

177. Kersebaum, K. C. Modelling nitrogen dynamics in soil–crop systems with HERMES. *Nutrient Cycling in Agroecosystems*. **77,** 39–52 (2007).

178. Keating, B. A. *et al.* An overview of APSIM, a model designed for farming systems simulation. *European Journal of Agronomy*. **18,** 267–288 (2003).

179. Deryng, D., Sacks, W. J., Barford, C. C. & Ramankutty, N. Simulating the effects of climate and agricultural management practices on global crop yield. *Global Biogeochemical Cycles*. **25,** GB2006 (2011).

180. Rosenzweig, C. *et al.* Assessing agricultural risks of climate change in the 21^{st} century in a global gridded crop model intercomparison. *Proceedings of the National Academy of Sciences*. **111,** 3268–3273 (2014).

181. Horie, T. A model for evaluating climatic productivity and water balance of irrigated rice and its application to Southeast Asia. *The Southeast Asian Studies*. **25,** 62–74 (1987).

182. Everingham, Y. L. *et al.* Enhanced risk management and decision-making capability across the sugarcane industry value chain based on seasonal climate forecasts. *Agricultural Systems*. **74,** 459–477 (2002).

183. Penning de Vries, F. W. T., Jansen, D. M., Berge, H. F. M. ten & Bakema, A. *Simulation of ecophysiological processes of growth in several annual crops*. (1989).

184. Marcelis, L. F. M., Heuvelink, E. & Goudriaan, J. Modelling biomass production and yield of horticultural crops: A review. *Scientia Horticulturae.* **74,** 83–111 (1998).

185. Aggarwal, P. K., Kalra, N., Singh, A. K. & Sinha, S. K. Analyzing the limitations set by climatic factors, genotype, water and nitrogen availability on productivity of wheat I. The model description, parametrization and validation. *Field Crops Research.* **38,** 73–91 (1994).

186. Lobell, D. B. & Burke, M. B. On the use of statistical models to predict crop yield responses to climate change. *Agricultural and Forest Meteorology.* **150,** 1443–1452 (2010).

187. Al-Gaadi, K. A. *et al.* Prediction of potato crop yield using precision agriculture techniques. *PLOS ONE.* **11,** e0162219 (2016).

188. Jones, J. W., Keating, B. A. & Porter, C. H. Approaches to modular model development. *Agricultural Systems.* **70,** 421–443 (2001).

189. Chow, B. Y. *et al.* Transcriptional regulation of *LUX* by CBF1 mediates cold input to the circadian clock in *Arabidopsis*. *Current Biology.* **24,** 1518–1524 (2014).

190. Gould, P. D. *et al.* Network balance via CRY signalling controls the *Arabidopsis* circadian clock over ambient temperatures. *Molecular Systems Biology.* **9,** 650 (2013).

191. Pokhilko, A. *et al.* The clock gene circuit in *Arabidopsis* includes a repressilator with additional feedback loops. *Molecular Systems Biology.* **8,** 574 (2012).

192. Schmal, C., Reimann, P. & Staiger, D. A circadian clock-regulated toggle switch explains *AtGRP7* and *AtGRP8* oscillations in *Arabidopsis thaliana*. *PLOS Computational Biology.* **9,** e1002986 (2013).

193. Grieneisen, V. A., Xu, J., Marée, A. F. M., Hogeweg, P. & Scheres, B. Auxin transport is sufficient to generate a maximum and gradient guiding root growth. *Nature.* **449,** 1008–1013 (2007).

194. Jönsson, H., Heisler, M. G., Shapiro, B. E., Meyerowitz, E. M. & Mjolsness, E. An auxin-driven polarized transport model for phyllotaxis. *Proceedings of the National Academy of Sciences.* **103,** 1633–1638 (2006).

195. Mourik, S. van *et al.* Simulation of organ patterning on the floral meristem using a polar auxin transport model. *PLOS ONE.* **7,** e28762 (2012).

196. Péret, B. *et al.* Sequential induction of auxin efflux and influx carriers regulates lateral root emergence. *Molecular Systems Biology.* **9,** 699 (2013).

197. Mendoza, L. & Alvarez-Buylla, E. R. Dynamics of the genetic regulatory network for *Arabidopsis thaliana* flower morphogenesis. *Journal of Theoretical Biology.* **193,** 307–319 (1998).

198. Sánchez-Corrales, Y.-E., Álvarez-Buylla, E. R. & Mendoza, L. The *Arabidopsis thaliana* flower organ specification gene regulatory network determines a robust differentiation process. *Journal of Theoretical Biology.* **264,** 971–983 (2010).

199. Mourik, S. van *et al.* Simulation of organ patterning on the floral meristem using a polar auxin transport model. *PLOS ONE.* **7,** e28762 (2012).

200. Song, Y. H., Smith, R. W., To, B. J., Millar, A. J. & Imaizumi, T. FKF1 conveys timing information for CONSTANS stabilization in photoperiodic flowering. *Science.* **336,** 1045–1049 (2012).

201. Salazar, J. D. *et al.* Prediction of photoperiodic regulators from quantitative gene circuit models. *Cell.* **139,** 1170–1179 (2009).

202. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics.* **17,** 333–351 (2016).

203. Weigel, D. & Mott, R. The 1001 genomes project for *Arabidopsis thaliana.* *Genome Biology.* **10,** 107 (2009).

204. McGrath, C. L. & Lynch, M. Evolutionary significance of whole-genome duplication in *Polyploidy and Genome Evolution.* 1–20 (Springer, Berlin, Heidelberg, 2012).

205. Werth, C. R. & Windham, M. D. A model for divergent, allopatric speciation of polyploid pteridophytes resulting from silencing of duplicate-gene expression. *The American Naturalist.* **137,** 515–526 (1991).

206. Conant, G. C. & Wolfe, K. H. Turning a hobby into a job: How duplicated genes find new functions. *Nature Reviews Genetics.* **9,** 938–950 (2008).

207. Pires, J. C. *et al.* Flowering time divergence and genomic rearrangements in resynthesized *Brassica* polyploids (Brassicaceae). *Biological Journal of the Linnean Society.* **82,** 675–688 (2004).

208. Chaudhary, B. *et al.* Reciprocal silencing, transcriptional bias and functional divergence of homeologs in polyploid cotton (Gossypium). *Genetics.* **182,** 503–517 (2009).

209. Buggs, R. J. A. *et al.* Transcriptomic shock generates evolutionary novelty in a newly formed, natural allopolyploid plant. *Current Biology.* **21,** 551–556 (2011).

210. Wapinski, I., Pfeffer, A., Friedman, N. & Regev, A. Natural history and evolutionary principles of gene duplication in fungi. *Nature.* **449,** 54–61 (2007).

211. Ohno, D. S. The creation of a new gene from a redundant duplicate of an old gene in *Evolution by Gene Duplication.* 71–82 (Springer Berlin Heidelberg, 1970).

212. Lynch, M. & Conery, J. S. The evolutionary fate and consequences of duplicate genes. *Science.* **290,** 1151–1155 (2000).

213. Force, A. *et al.* Preservation of duplicate genes by complementary, degenerative mutations. *Genetics.* **151,** 1531–1545 (1999).

214. Veitia, R. A. & Potier, M. C. Gene dosage imbalances: Action, reaction, and models. *Trends in Biochemical Sciences.* **40,** 309–317 (2015).

215. Nowak, M. A., Boerlijst, M. C., Cooke, J. & Smith, J. M. Evolution of genetic redundancy. *Nature.* **388,** 167–171 (1997).

216. Des Marais, D. L. & Rausher, M. D. Escape from adaptive conflict after duplication in an anthocyanin pathway gene. *Nature.* **454,** 762–765 (2008).

217. Lynch, M. & Walsh, B. *Genetics and analysis of quantitative traits*. (Sinauer, 1998).

218. Walsh, J. B. How often do duplicated genes evolve new functions? *Genetics*. **139,** 421–428 (1995).

219. Kafri, R., Bar-Even, A. & Pilpel, Y. Transcription control reprogramming in genetic backup circuits. *Nature Genetics*. **37,** 295–299 (2005).

220. Kafri, R., Levy, M. & Pilpel, Y. The regulatory utilization of genetic redundancy through responsive backup circuits. *Proceedings of the National Academy of Sciences*. **103,** 11653–11658 (2006).

221. Veitia, R. A. Gene dosage balance in cellular pathways: Implications for dominance and gene duplicability. *Genetics*. **168,** 569–574 (2004).

222. Veitia, R. A. Nonlinear effects in macromolecular assembly and dosage sensitivity. *Journal of Theoretical Biology*. **220,** 19–25 (2003).

223. Veitia, R. A., Bottani, S. & Birchler, J. A. Cellular reactions to gene dosage imbalance: Genomic, transcriptomic and proteomic effects. *Trends in Genetics*. **24,** 390–397 (2008).

224. Birchler, J. A. & Veitia, R. A. Gene balance hypothesis: Connecting issues of dosage sensitivity across biological disciplines. *Proceedings of the National Academy of Sciences*. **109,** 14746–14753 (2012).

225. Gu, X., Wang, Y. & Gu, J. Age distribution of human gene families shows significant roles of both large- and small-scale duplications in vertebrate evolution. *Nature Genetics*. **31,** 205–209 (2002).

226. Hakes, L., Pinney, J. W., Lovell, S. C., Oliver, S. G. & Robertson, D. L. All duplicates are not equal: The difference between small-scale and genome duplication. *Genome Biology*. **8,** R209 (2007).

227. Papp, B., Pál, C. & Hurst, L. D. Dosage sensitivity and the evolution of gene families in yeast. *Nature*. **424,** 194–197 (2003).

228. Blomme, T. *et al.* The gain and loss of genes during 600 million years of vertebrate evolution. *Genome Biology*. **7,** R43 (2006).

229. Blanc, G. & Wolfe, K. H. Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. *The Plant Cell.* **16,** 1679–1691 (2004).

230. Seoighe, C. & Gehring, C. Genome duplication led to highly selective expansion of the *Arabidopsis thaliana* proteome. *Trends in Genetics.* **20,** 461–464 (2004).

231. Maere, S. *et al.* Modeling gene and genome duplications in eukaryotes. *Proceedings of the National Academy of Sciences of the United States of America.* **102,** 5454–5459 (2005).

232. Baduel, P., Arnold, B., Weisman, C. M., Hunter, B. & Bomblies, K. Habitat-associated life history and stress-tolerance variation in *Arabidopsis arenosa*. *Plant Physiology.* **171,** 437–451 (2016).

233. Kardailsky, I. *et al.* Activation tagging of the floral inducer *FT*. *Science.* **286,** 1962–1965 (1999).

234. Conti, L. & Bradley, D. TERMINAL FLOWER1 is a mobile signal controlling *Arabidopsis* architecture. *The Plant Cell.* **19,** 767–778 (2007).

235. Wu, G. *et al.* The sequential action of miR156 and miR172 regulates developmental timing in *Arabidopsis*. *Cell.* **138,** 750–759 (2009).

236. Barton, M. K. Twenty years on: The inner workings of the shoot apical meristem, a developmental dynamo. *Developmental Biology.* **341,** 95–113 (2010).

237. Pidkowich, M. S., Klenz, J. E. & Haughn, G. W. The making of a flower: Control of floral meristem identity in *Arabidopsis*. *Trends in Plant Science.* **4,** 64–70 (1999).

238. Elhiti, M. *et al.* Gene expression analysis in microdissected shoot meristems of *Brassica napus* microspore-derived embryos with altered *SHOOT-MERISTEMLESS* levels. *Planta.* **237,** 1065–1082 (2013).

239. Qiu, D. *et al.* A comparative linkage map of oilseed rape and its use for QTL analysis of seed oil and erucic acid content. *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik.* **114,** 67–80 (2006).

240. Salomé P. A. & Mcclung C. R. What makes the Arabidopsis clock tick on time? A review on entrainment. *Plant, Cell & Environment.* **28,** 21–38 (2004).

241. Murphy, L. A. & Scarth, R. Vernalization response in spring oilseed rape (*Brassica napus* L.) cultivars. *Canadian Journal of Plant Science.* **74,** 275–277 (1994).

242. Mockler, T. *et al.* Regulation of photoperiodic flowering by *Arabidopsis* photoreceptors. *Proceedings of the National Academy of Sciences of the United States of America.* **100,** 2140–2145 (2003).

243. Kumar, S. V. & Wigge, P. A. H2A.Z-containing nucleosomes mediate the thermosensory response in *Arabidopsis*. *Cell.* **140,** 136–147 (2010).

244. Kumar, S. V. *et al.* Transcription factor PIF4 controls the thermosensory activation of flowering. *Nature.* **484,** 242–245 (2012).

245. Shindo, C., Lister, C., Crevillen, P., Nordborg, M. & Dean, C. Variation in the epigenetic silencing of *FLC* contributes to natural variation in *Arabidopsis* vernalization response. *Genes & Development.* **20,** 3079–3083 (2006).

246. Lancashire, P. D. *et al.* A uniform decimal code for growth stages of crops and weeds. *Annals of Applied Biology.* **119,** 561–601 (1991).

247. Conesa, A. *et al.* A survey of best practices for RNA-Seq data analysis. *Genome Biology.* **17,** 13 (2016).

248. Trick, M. *et al.* A newly-developed community microarray resource for transcriptome profiling in *Brassica* species enables the confirmation of *Brassica*-specific expressed sequences. *BMC Plant Biology.* **9,** 50 (2009).

249. He, Z. *et al.* Construction of *Brassica* A and C genome-based ordered pan-transcriptomes for use in rapeseed genomic research. *Data in Brief.* **4,** 357–362 (2015).

250. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-Seq experiments with TopHat and Cufflinks. *Nature Protocols.* **7,** 562–578 (2012).

251. Kim, D. *et al.* TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology.* **14,** R36 (2013).

252. Howe, K. L., Chothia, T. & Durbin, R. GAZE: A generic framework for the integration of gene-prediction data by dynamic programming. *Genome Research.* **12,** 1418–1427 (2002).

253. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. *Bioinformatics.* **24,** 637–644 (2008).

254. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods.* **9,** 357–359 (2012).

255. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology.* **10,** R25 (2009).

256. Trapnell, C. *et al.* Differential analysis of gene regulation at transcript resolution with RNA-Seq. *Nature Biotechnology.* **31,** 46–53 (2013).

257. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-Seq quantification. *Nature Biotechnology.* **34,** 525–527 (2016).

258. Pimentel, H. J., Bray, N., Puente, S., Melsted, P. & Pachter, L. Differential analysis of RNA-Seq incorporating quantification uncertainty. *bioRxiv.* 058164 (2016). doi:10.1101/058164

259. Xu, H.-M. *et al.* Transcriptome analysis of *Brassica napus* pod using RNA-Seq and identification of lipid-related candidate genes. *BMC Genomics.* **16,** 858 (2015).

260. Chan, A. C. *et al.* Tissue-specific laser microdissection of the *Brassica napus* funiculus improves gene discovery and spatial identification of biological processes. *Journal of Experimental Botany.* **67,** 3561–3571 (2016).

261. Renny-Byfield, S. *et al.* Ancient gene duplicates in *Gossypium* (cotton) exhibit near-complete expression divergence. *Genome Biology and Evolution.* **6,** 559–571 (2014).

262. Payne, R. M. E. *et al.* An NPF transporter exports a central monoterpene indole alkaloid intermediate from the vacuole. *Nature Plants.* **3,** 16208 (2017).

263. Kim, D. hyun, Grün, D. & Oudenaarden, A. van. Dampening of expression oscillations by synchronous regulation of a microRNA and its target. *Nature Genetics.* **45,** 1337–1344 (2013).

264. Paina, C., Byrne, S. L., Domnisoru, C. & Asp, T. Vernalization mediated changes in the *Lolium perenne* transcriptome. *PLOS ONE.* **9,** e107365 (2014).

265. Langfelder, P. & Horvath, S. WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics.* **9,** 559 (2008).

266. Woo, H. R. *et al.* Programming of plant leaf senescence with temporal and inter-organellar coordination of transcriptome in Arabidopsis. *Plant Physiology.* **171,** 452–467 (2016).

267. Klepikova, A. V., Logacheva, M. D., Dmitriev, S. E. & Penin, A. A. RNA-Seq analysis of an apical meristem time series reveals a critical point in *Arabidopsis thaliana* flower initiation. *BMC Genomics.* **16,** 466 (2015).

268. Rymen, B. *et al.* Cold nights impair leaf growth and cell cycle progression in maize through transcriptional changes of cell cycle genes. *Plant Physiology.* **143,** 1429–1438 (2007).

269. Gangappa, S. N., Berriri, S. & Kumar, S. V. PIF4 coordinates thermosensory growth and immunity in *Arabidopsis*. *Current Biology.* **27,** 243–249 (2017).

270. Denancé, N., Sánchez-Vallet, A., Goffner, D. & Molina, A. Disease resistance or growth: The role of plant hormones in balancing immune responses and fitness costs. *Plant Cell Biology.* **4,** 155 (2013).

271. Hua, J. Modulation of plant immunity by light, circadian rhythm, and temperature. *Current Opinion in Plant Biology.* **16,** 406–413 (2013).

272. Alcázar, R. & Parker, J. E. The impact of temperature on balancing immune responsiveness and growth in *Arabidopsis*. *Trends in Plant Science.* **16,** 666–675 (2011).

273. Shi, Z., Maximova, S., Liu, Y., Verica, J. & Guiltinan, M. J. The salicylic acid receptor NPR3 is a negative regulator of the transcriptional defense response during early flower development in *Arabidopsis*. *Molecular Plant*. **6,** 802–816 (2013).

274. McLachlan, A. D. Protein structure and function. *Annual Review of Physical Chemistry*. **23,** 165–192 (1972).

275. Ouzounis, C. A., Coulson, R. M. R., Enright, A. J., Kunin, V. & Pereira-Leal, J. B. Classification schemes for protein structure and function. *Nature Reviews. Genetics*. **4,** 508–519 (2003).

276. Franco-Zorrilla, J. M. *et al.* DNA-binding specificities of plant transcription factors and their potential to define target genes. *Proceedings of the National Academy of Sciences*. **111,** 2367–2372 (2014).

277. Adryan, B. & Teichmann, S. A. The developmental expression dynamics of *Drosophila melanogaster* transcription factors. *Genome Biology*. **11,** R40 (2010).

278. Schwarz-Sommer, Z., Huijser, P., Nacken, W., Saedler, H. & Sommer, H. Genetic control of flower development by homeotic genes in *Antirrhinum majus*. *Science*. **250,** 931–936 (1990).

279. Ng, M. & Yanofsky, M. F. Function and evolution of the plant MADS-box gene family. *Nature Reviews Genetics*. **2,** 186–195 (2001).

280. Krizek, B. A. & Fletcher, J. C. Molecular mechanisms of flower development: An armchair guide. *Nature Reviews Genetics*. **6,** 688–698 (2005).

281. Andrés, F. *et al. SHORT VEGETATIVE PHASE* reduces gibberellin biosynthesis at the *Arabidopsis* shoot apex to regulate the floral transition. *Proceedings of the National Academy of Sciences*. **111,** E2760–E2769 (2014).

282. Bowman, J. L., Smyth, D. R. & Meyerowitz, E. M. Genes directing flower development in Arabidopsis. *The Plant Cell*. **1,** 37–52 (1989).

283. Jofuku, K. D., Boer, B. G. den, Montagu, M. V. & Okamuro, J. K. Control of Arabidopsis flower and seed development by the homeotic gene APETALA2. *The Plant Cell*. **6,** 1211–1225 (1994).

284. Licausi, F., Ohme-Takagi, M. & Perata, P. APETALA2/Ethylene Responsive Factor (AP2/ERF) transcription factors: Mediators of stress responses and developmental programs. *New Phytologist.* **199,** 639–649 (2013).

285. Ferrante, A. & Francini, A. Ethylene and leaf senescence in *Ethylene Action in Plants.* (ed. Khan, D. N. A.) 51–67 (Springer Berlin Heidelberg, 2006).

286. Drews, G. N., Bowman, J. L. & Meyerowitz, E. M. Negative regulation of the Arabidopsis homeotic gene *AGAMOUS* by the *APETALA2* product. *Cell.* **65,** 991–1002 (1991).

287. Osborn, T. C. *et al.* Understanding mechanisms of novel gene expression in polyploids. *Trends in Genetics.* **19,** 141–147 (2003).

288. Yoo, M.-J., Szadkowski, E. & Wendel, J. F. Homoeolog expression bias and expression level dominance in allopolyploid cotton. *Heredity.* **110,** 171–180 (2013).

289. Flagel, L. E. & Wendel, J. F. Evolutionary rate variation, genomic dominance and duplicate gene expression evolution during allotetraploid cotton speciation. *New Phytologist.* **186,** 184–193 (2010).

290. Wang, J. *et al.* Genomewide nonadditive gene regulation in Arabidopsis allotetraploids. *Genetics.* **172,** 507–517 (2006).

291. Comai, L. *et al.* Phenotypic instability and rapid gene silencing in newly formed Arabidopsis allotetraploids. *The Plant Cell.* **12,** 1551–1567 (2000).

292. Adams, K. L., Cronn, R., Percifield, R. & Wendel, J. F. Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. *Proceedings of the National Academy of Sciences.* **100,** 4649–4654 (2003).

293. Buggs, R. J. A. *et al.* The legacy of diploid progenitors in allopolyploid gene expression patterns. *Phil. Trans. R. Soc. B.* **369,** 20130354 (2014).

294. Schnable, J. C., Springer, N. M. & Freeling, M. Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proceedings of the National Academy of Sciences.* **108,** 4069–4074 (2011).

295. Akhunova, A. R., Matniyazov, R. T., Liang, H. & Akhunov, E. D. Homoeolog-specific transcriptional bias in allopolyploid wheat. *BMC Genomics.* **11,** 505 (2010).

296. Bardil, A., Almeida, J. D. de, Combes, M. C., Lashermes, P. & Bertrand, B. Genomic expression dominance in the natural allopolyploid *Coffea arabica* is massively affected by growth temperature. *New Phytologist.* **192,** 760–774 (2011).

297. Ilut, D. C. *et al.* A comparative transcriptomic study of an allotetraploid and its diploid progenitors illustrates the unique advantages and challenges of RNA-seq in plant species. *American Journal of Botany.* **99,** 383–396 (2012).

298. Grover, C. E. *et al.* Homoeolog expression bias and expression level dominance in allopolyploids. *New Phytologist.* **196,** 966–971 (2012).

299. Bouché, F., Lobet, G., Tocquin, P. & Périlleux, C. FLOR-ID: An interactive database of flowering-time gene networks in *Arabidopsis thaliana*. *Nucleic Acids Research.* **44,** D1167–D1171 (2016).

300. Giavalisco, P., Kapitza, K., Kolasa, A., Buhtz, A. & Kehr, J. Towards the proteome of *Brassica napus* phloem sap. *PROTEOMICS.* **6,** 896–909 (2006).

301. Wang, J. *et al.* The evolution of *Brassica napus FLOWERING LOCUS T* paralogues in the context of inverted chromosomal duplication blocks. *BMC Evolutionary Biology.* **9,** 271 (2009).

302. Corbesier, L. *et al.* FT protein movement contributes to long-distance signaling in floral induction of *Arabidopsis*. *Science.* **316,** 1030–1033 (2007).

303. Yu, H., Ito, T., Wellmer, F. & Meyerowitz, E. M. Repression of AGAMOUS-LIKE 24 is a crucial step in promoting flower development. *Nature Genetics.* **36,** 157–161 (2004).

304. Nelson, M. N. *et al.* Quantitative trait loci for thermal time to flowering and photoperiod responsiveness discovered in summer annual-type *Brassica napus* L. *PLOS ONE.* **9,** e102611 (2014).

305. Capovilla, G., Schmid, M. & Posé, D. Control of flowering by ambient temperature. *Journal of Experimental Botany.* **66,** 59–69 (2015).

306. Borner, R. *et al.* A MADS domain gene involved in the transition to flowering in *Arabidopsis*. *The Plant Journal*. **24,** 591–599 (2000).

307. Schmid, M. *et al.* Dissection of floral induction pathways using global expression analysis. *Development*. **130,** 6001–6012 (2003).

308. Michaels, S. D., Himelblau, E., Kim, S. Y., Schomburg, F. M. & Amasino, R. M. Integration of flowering signals in winter-annual Arabidopsis. *Plant Physiology*. **137,** 149–156 (2005).

309. Serrano-Mislata, A. *et al.* Separate elements of the *TERMINAL FLOWER 1* cis-regulatory region integrate pathways to control flowering time and shoot meristem identity. *Development*. **143,** 3315–3327 (2016).

310. Miller, M. The importance of being flexible: The case of basic region leucine zipper transcriptional regulators. *Current protein & peptide science*. **10,** 244–269 (2009).

311. Ellenberger, T. Getting a grip on DNA recognition: Structures of the basic region leucine zipper, and the basic region helix-loop-helix DNA-binding domains. *Current Opinion in Structural Biology*. **4,** 12–21 (1994).

312. Busch, S. J. & Sassone-Corsi, P. Dimers, leucine zippers and DNA-binding domains. *Trends in Genetics*. **6,** 36–40 (1990).

313. Landschulz, W. H., Johnson, P. F. & McKnight, S. L. The DNA binding domain of the rat liver nuclear protein C/EBP is bipartite. *Science*. **243,** 1681–1688 (1989).

314. John, M., Briand, J. P., Granger-Schnarr, M. & Schnarr, M. Two pairs of oppositely charged amino acids from Jun and Fos confer heterodimerization to GCN4 leucine zipper. *Journal of Biological Chemistry*. **269,** 16247–16253 (1994).

315. Amoutzias, G. D., Robertson, D. L., Van de Peer, Y. & Oliver, S. G. Choose your partners: Dimerization in eukaryotic transcription factors. *Trends in Biochemical Sciences*. **33,** 220–229 (2008).

316. Klemm, J. D., Schreiber, S. L. & Crabtree, G. R. Dimerization as a regulatory mechanism in signal transduction. *Annual Review of Immunology*. **16,** 569–592 (1998).

317. Amoutzias, G. D. *et al.* One billion years of bZIP transcription factor evolution: Conservation and change in dimerization and DNA-binding site specificity. *Molecular Biology and Evolution.* **24,** 827–835 (2007).

318. Tsuji, H., Nakamura, H., Taoka, K.-i. & Shimamoto, K. Functional diversification of FD transcription factors in rice, components of florigen activation complexes. *Plant and Cell Physiology.* **54,** 385–397 (2013).

319. Schumacher, M. A., Goodman, R. H. & Brennan, R. G. The structure of a CREB bZIP · somatostatin CRE complex reveals the basis for selective dimerization and divalent cation-enhanced DNA binding. *Journal of Biological Chemistry.* **275,** 35242–35247 (2000).

320. Potapov, V., Kaplan, J. B. & Keating, A. E. Data-driven prediction and design of bZIP coiled-coil interactions. *PLOS Comput Biol.* **11,** e1004046 (2015).

321. Kersey, P. J. *et al.* Ensembl Genomes 2016: More genomes, more complexity. *Nucleic Acids Research.* **44,** D574–D580 (2016).

322. McDonald, I. K. & Thornton, J. M. Satisfying hydrogen bonding potential in proteins. *Journal of Molecular Biology.* **238,** 777–793 (1994).

323. Luscombe, N. M., Laskowski, R. A. & Thornton, J. M. Amino acid-base interactions: A three-dimensional analysis of protein-DNA interactions at an atomic level. *Nucleic Acids Research.* **29,** 2860–2874 (2001).

324. Dubcovsky, J. & Dvorak, J. Genome plasticity a key factor in the success of polyploid wheat under domestication. *Science (New York, N.Y.).* **316,** 1862–1866 (2007).

325. Beest, M. te *et al.* The more the better? The role of polyploidy in facilitating plant invasions. *Annals of Botany.* **109,** 19–45 (2012).

326. Blanc, G., Barakat, A., Guyot, R., Cooke, R. & Delseny, M. Extensive duplication and reshuffling in the Arabidopsis genome. *The Plant Cell.* **12,** 1093–1101 (2000).

327. Blanc, G., Hokamp, K. & Wolfe, K. H. A recent polyploidy superimposed on older large-scale duplications in the *Arabidopsis* genome. *Genome Research.* **13,** 137–144 (2003).

328. Chen, Z. J. & Pikaard, C. S. Transcriptional analysis of nucleolar dominance in polyploid plants: Biased expression/silencing of progenitor rRNA genes is developmentally regulated in *Brassica*. *Proceedings of the National Academy of Sciences*. **94,** 3442–3447 (1997).

329. Combes, M.-C., Dereeper, A., Severac, D., Bertrand, B. & Lashermes, P. Contribution of subgenomes to the transcriptome and their intertwined regulation in the allopolyploid *Coffea arabica* grown at contrasted temperatures. *New Phytologist*. **200,** 251–260 (2013).

330. Harper, A. L. *et al.* Associative transcriptomics of traits in the polyploid crop species *Brassica napus*. *Nature Biotechnology*. **30,** 798–802 (2012).

331. Duarte, J. M. *et al.* Expression pattern shifts following duplication indicative of subfunctionalization and neofunctionalization in regulatory genes of *Arabidopsis*. *Molecular Biology and Evolution*. **23,** 469–478 (2006).

332. De Smet, R. & Van de Peer, Y. Redundancy and rewiring of genetic networks following genome-wide duplication events. *Current Opinion in Plant Biology*. **15,** 168–176 (2012).

333. Hodgkinson, A. & Eyre-Walker, A. Variation in the mutation rate across mammalian genomes. *Nature Reviews Genetics*. **12,** 756–766 (2011).

334. Nishant, K. T., Singh, N. D. & Alani, E. Genomic mutation rates: What high-throughput methods can tell us. *BioEssays : news and reviews in molecular, cellular and developmental biology*. **31,** 912–920 (2009).

335. Oladosu, Y. *et al.* Principle and application of plant mutagenesis in crop improvement: A review. *Biotechnology & Biotechnological Equipment*. **30,** 1–16 (2016).

336. Simillion, C., Vandepoele, K., Montagu, M. C. E. V., Zabeau, M. & Peer, Y. V. de. The hidden duplication past of *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences*. **99,** 13627–13632 (2002).

337. Foucher, F. *et al.* *DETERMINATE* and *LATE FLOWERING* are two *TERMINAL FLOWER1/CENTRORADIALIS* homologs that control two distinct phases of flowering initiation and development in pea. *The Plant Cell*. **15,** 2742–2754 (2003).

338. Benezra, R., Davis, R. L., Lockshon, D., Turner, D. L. & Weintraub, H. The protein Id: A negative regulator of helix-loop-helix DNA binding proteins. *Cell.* **61,** 49–59 (1990).

339. Ryu, J. Y. *et al.* The *Arabidopsis* floral repressor BFT delays flowering by competing with FT for FD binding under high salinity. *Molecular Plant.* **7,** 377–387 (2014).

340. Lee, I., Bleecker, A. & Amasino, R. Analysis of naturally occurring late flowering in *Arabidopsis thaliana*. *Molecular and General Genetics MGG.* **237,** 171–176 (1993).

341. Lee, I., Michaels, S. D., Masshardt, A. S. & Amasino, R. M. The late-flowering phenotype of *FRIGIDA* and mutations in *LUMINIDEPENDENS* is suppressed in the Landsberg *erecta* strain of *Arabidopsis*. *The Plant Journal.* **6,** 903–909 (1994).

342. Clarke, J. H. & Dean, C. Mapping *FRI*, a locus controlling flowering time and vernalization response in *Arabidopsis thaliana*. *Molecular and General Genetics MGG.* **242,** 81–89 (1994).

343. Koornneef, M., Blankestijn-de Vries, H., Hanhart, C., Soppe, W. & Peeters, T. The phenotype of some late-flowering mutants is enhanced by a locus on chromosome 5 that is not effective in the Landsberg erecta wild-type. *The Plant Journal.* **6,** 911–919 (1994).

344. Michaels, S. D., He, Y., Scortecci, K. C. & Amasino, R. M. 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,** 10102–10107 (2003).

345. Coustham, V. *et al.* Quantitative modulation of Polycomb silencing underlies natural variation in vernalization. *Science.* **337,** 584–587 (2012).

346. Duncan, S. *et al.* Seasonal shift in timing of vernalization as an adaptation to extreme winter. *eLife.* **4,** e06620 (2015).

347. Sheldon, C. C. *et al.* The *FLF* MADS box gene: A repressor of flowering in Arabidopsis regulated by vernalization and methylation. *The Plant Cell.* **11,** 445–458 (1999).

348. Koornneef, M., Alonso-Blanco, C., Vries, H. B.-d., Hanhart, C. J. & Peeters, A. J. M. Genetic interactions among late-flowering mutants of Arabidopsis. *Genetics*. **148,** 885–892 (1998).

349. Sanda, S. L. & Amasino, R. M. Interaction of *FLC* and late-flowering mutations in *Arabidopsis thaliana*. *Molecular and General Genetics MGG*. **251,** 69–74 (1996).

350. Michaels, S. D. & Amasino, R. M. 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*. **13,** 935–941 (2001).

351. Swiezewski, S., Liu, F., Magusin, A. & Dean, C. Cold-induced silencing by long antisense transcripts of an *Arabidopsis* Polycomb target. *Nature*. **462,** 799–802 (2009).

352. Pien, S. *et al.* ARABIDOPSIS TRITHORAX1 dynamically regulates *FLOWERING LOCUS C* activation via histone 3 lysine 4 trimethylation. *The Plant Cell*. **20,** 580–588 (2008).

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

354. Zhao, Z., Yu, Y., Meyer, D., Wu, C. & Shen, W.-H. Prevention of early flowering by expression of *FLOWERING LOCUS C* requires methylation of histone H3 K36. *Nature Cell Biology*. **7,** 1256–1260 (2005).

355. Yang, H., Howard, M. & Dean, C. Antagonistic roles for H3K36me3 and H3K27me3 in the cold-induced epigenetic switch at *Arabidopsis FLC*. *Current Biology*. **24,** 1793–1797 (2014).

356. Wood, C. C. *et al.* 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,** 14631–14636 (2006).

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

358. Wysocka, J. *et al.* A PHD finger of NURF couples histone H3 lysine 4 trimethylation with chromatin remodelling. *Nature.* **442,** 86–90 (2006).

359. Sheldon, C. C., Conn, A. B., Dennis, E. S. & Peacock, W. J. Different regulatory regions are required for the vernalization-induced repression of *FLOWERING LOCUS C* and for the epigenetic maintenance of repression. *The Plant Cell.* **14,** 2527–2537 (2002).

360. Berry, S., Hartley, M., Olsson, T. S. G., Dean, C. & Howard, M. Local chromatin environment of a Polycomb target gene instructs its own epigenetic inheritance. *eLife.* **4,** e07205 (2015).

361. Angel, A., Song, J., Dean, C. & Howard, M. A Polycomb-based switch underlying quantitative epigenetic memory. *Nature.* **476,** 105–108 (2011).

362. Fadina, O. A., Pankin, A. A. & Khavkin, E. E. Molecular characterization of the flowering time gene *FRIGIDA* in *Brassica* genomes A and C. *Russian Journal of Plant Physiology.* **60,** 279–289 (2013).

363. Raman, R. *et al.* Molecular mapping of qualitative and quantitative loci for resistance to *Leptosphaeria maculans* causing blackleg disease in canola (*Brassica napus* L.). *Theoretical and Applied Genetics.* **125,** 405–418 (2012).

364. Raman, H. *et al.* Genetic and physical mapping of flowering time loci in canola (*Brassica napus* L.). *Theoretical and Applied Genetics.* **126,** 119–132 (2013).

365. Xiao, D. *et al.* The *Brassica rapa flc* homologue *FLC2* is a key regulator of flowering time, identified through transcriptional co-expression networks. *Journal of Experimental Botany.* **64,** 4503–4516 (2013).

366. Curtis, O. F. & Chang, H. T. The relative effectiveness of the temperature of the crown as contrasted with that of the rest of the plant upon the flowering of celery plants. *American Journal of Botany.* **17,** 1047–1048 (1930).

367. Metzger, J. D. Localization of the site of perception of thermoinductive temperatures in *Thlaspi arvense* L. *Plant Physiology.* **88,** 424–428 (1988).

368. Wellensiek, S. J. Leaf vernalization. *Nature*. **192,** 1097–1098 (1961).

369. Wellensiek, S. J. Dividing cells as the locus for vernalization. *Nature*. **195,** 307–308 (1962).

370. Wellensiek, S. J. Dividing cells as the prerequisite for vernalization. *Plant Physiology*. **39,** 832–835 (1964).

371. Pierik, R. L. M. The induction and initiation of flowerbuds *in vitro* in tissues of *Lunaria annua* L. *Naturwissenschaften*. **53,** 45–45 (1966).

372. Pierik, R. L. M. The induction and initiation of flowerbuds in vitro in root tissues of *Cichorium intybus* L. *Naturwissenschaften*. **53,** 387–387 (1966).

373. Michaels, S. D. & Amasino, R. M. Memories of winter: Vernalization and the competence to flower. *Plant, Cell & Environment*. **23,** 1145–1153 (2000).

374. Li, W., Liu, X. & Lu, Y. Transcriptome comparison reveals key candidate genes in response to vernalization of Oriental lily. *BMC Genomics*. **17,** 664 (2016).

375. Villacorta-Martin, C. *et al.* Whole transcriptome profiling of the vernalization process in *Lilium longiflorum* (cultivar White Heaven) bulbs. *BMC Genomics*. **16,** 550 (2015).

376. Greenup, A. G. *et al.* Transcriptome analysis of the vernalization response in barley (*Hordeum vulgare*) seedlings. *PLOS ONE*. **6,** e17900 (2011).

377. Liu, C., Wang, S., Xu, W. & Liu, X. Genome-wide transcriptome profiling of radish (*Raphanus sativus* L.) in response to vernalization. *PLOS ONE*. **12,** e0177594 (2017).

378. Huan, Q., Mao, Z., Zhang, J., Xu, Y. & Chong, K. Transcriptome-wide analysis of vernalization reveals conserved and species-specific mechanisms in *Brachypodium*. *Journal of Integrative Plant Biology*. **55,** 696–709 (2013).

379. Edwards, K. D. *et al. FLOWERING LOCUS C* mediates natural variation in the high-temperature response of the *Arabidopsis* circadian clock. *The Plant Cell*. **18,** 639–650 (2006).

380. Hawkes, E. J. *et al. COOLAIR* antisense RNAs form evolutionarily conserved elaborate secondary structures. *Cell Reports*. **16,** 3087–3096 (2016).

381. Aikawa, S., Kobayashi, M. J., Satake, A., Shimizu, K. K. & Kudoh, H. Robust control of the seasonal expression of the *Arabidopsis FLC* gene in a fluctuating environment. *Proceedings of the National Academy of Sciences.* **107,** 11632–11637 (2010).

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

383. Margueron, R. & Reinberg, D. The Polycomb complex PRC2 and its mark in life. *Nature.* **469,** 343–349 (2011).

384. Müller, J. & Verrijzer, P. Biochemical mechanisms of gene regulation by polycomb group protein complexes. *Current Opinion in Genetics & Development.* **19,** 150–158 (2009).

385. Czermin, B. *et al. Drosophila* Enhancer of Zeste/ESC complexes have a histone H3 methyltransferase activity that marks chromosomal Polycomb sites. *Cell.* **111,** 185–196 (2002).

386. Hennig, L. & Derkacheva, M. Diversity of Polycomb group complexes in plants: Same rules, different players? *Trends in Genetics.* **25,** 414–423 (2009).

387. Pien, S. & Grossniklaus, U. *Polycomb* group and *trithorax* group proteins in *Arabidopsis. Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression.* **1769,** 375–382 (2007).

388. Gendall, A. R., Levy, Y. Y., Wilson, A. & Dean, C. The *VERNALIZATION 2* gene mediates the epigenetic regulation of vernalization in *Arabidopsis. Cell.* **107,** 525–535 (2001).

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

390. Chandler, J., Wilson, A. & Dean, C. *Arabidopsis* mutants showing an altered response to vernalization. *The Plant Journal.* **10,** 637–644 (1996).

391. Jaudal, M. *et al. MtVRN2* is a Polycomb *VRN2-like* gene which represses the transition to flowering in the model legume *Medicago truncatula. The Plant Journal.* **86,** 145–160 (2016).

392. Hennig, L., Bouveret, R. & Gruissem, W. MSI1-like proteins: An escort service for chromatin assembly and remodeling complexes. *Trends in Cell Biology*. **15,** 295–302 (2005).

393. Derkacheva, M. *et al.* *Arabidopsis* MSI1 connects LHP1 to PRC2 complexes. *The EMBO Journal*. **32,** 2073–2085 (2013).

394. Hennig, L., Taranto, P., Walser, M., Schönrock, N. & Gruissem, W. *Arabidopsis* MSI1 is required for epigenetic maintenance of reproductive development. *Development*. **130,** 2555–2565 (2003).

395. Köhler, C. *et al.* *Arabidopsis* MSI1 is a component of the MEA/FIE *Polycomb* group complex and required for seed development. *The EMBO Journal*. **22,** 4804–4814 (2003).

396. Bastow, R. *et al.* Vernalization requires epigenetic silencing of *FLC* by histone methylation. *Nature*. **427,** 164–167 (2004).

397. Mylne, J., Greb, T., Lister, C. & Dean, C. Epigenetic regulation in the control of flowering. *Cold Spring Harbor Symposia on Quantitative Biology*. **69,** 457–464 (2004).

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

399. Kim, D.-H. & Sung, S. The Plant Homeo Domain finger protein, VIN3-LIKE 2, is necessary for photoperiod-mediated epigenetic regulation of the floral repressor, MAF5. *Proceedings of the National Academy of Sciences*. **107,** 17029–17034 (2010).

400. Gazzani, S., Gendall, A. R., Lister, C. & Dean, C. Analysis of the molecular basis of flowering time variation in Arabidopsis accessions. *Plant Physiology*. **132,** 1107–1114 (2003).

401. Le Corre, V., Roux, F. & Reboud, X. DNA polymorphism at the *FRIGIDA* gene in *Arabidopsis thaliana*: Extensive nonsynonymous variation is consistent with local selection for flowering time. *Molecular Biology and Evolution*. **19,** 1261–1271 (2002).

402. Johanson, U. *et al.* Molecular analysis of *FRIGIDA*, a major determinant of natural variation in *Arabidopsis* flowering time. *Science.* **290,** 344–347 (2000).

403. Rosas, U. *et al.* Variation in *Arabidopsis* flowering time associated with *cis*-regulatory variation in *CONSTANS*. *Nature Communications.* **5,** 3651 (2014).

404. Skøt, L. *et al.* Allelic variation in the perennial ryegrass *FLOWERING LOCUS T* gene is associated with changes in flowering time across a range of populations. *Plant Physiology.* **155,** 1013–1022 (2011).

405. Arsovski, A. A., Pradinuk, J., Guo, X. Q., Wang, S. & Adams, K. L. Evolution of cis-regulatory elements and regulatory networks in duplicated genes of Arabidopsis. *Plant Physiology.* **169,** 2982–2991 (2015).

406. Teichmann, S. A. & Babu, M. M. Gene regulatory network growth by duplication. *Nature Genetics.* **36,** 492–496 (2004).

407. Tsai, Z. T.-Y. *et al.* Evolution of cis-regulatory elements in yeast de novo and duplicated new genes. *BMC Genomics.* **13,** 717 (2012).

408. Zeevaart, J. A. D. Florigen coming of age after 70 years. *The Plant Cell.* **18,** 1783–1789 (2006).

409. Zeevaart, J. A. Leaf-produced floral signals. *Current Opinion in Plant Biology.* **11,** 541–547 (2008).

410. Levey, S. & Wingler, A. Natural variation in the regulation of leaf senescence and relation to other traits in Arabidopsis. *Plant, Cell & Environment.* **28,** 223–231 (2005).

411. Torti, S. *et al.* Analysis of the *Arabidopsis* shoot meristem transcriptome during floral transition identifies distinct regulatory patterns and a leucine-rich repeat protein that promotes flowering. *The Plant Cell.* **24,** 444–462 (2012).

412. Werner, J. D. *et al. FRIGIDA*-independent variation in flowering time of natural *Arabidopsis thaliana* accessions. *Genetics.* **170,** 1197–1207 (2005).

413. Scarcelli, N., Cheverud, J. M., Schaal, B. A. & Kover, P. X. Antagonistic pleiotropic effects reduce the potential adaptive value of the FRIGIDA locus.

Proceedings of the National Academy of Sciences of the United States of America. **104,** 16986–16991 (2007).

414. Deng, W. *et al.* FLOWERING LOCUS C (FLC) regulates development pathways throughout the life cycle of *Arabidopsis*. *Proceedings of the National Academy of Sciences of the United States of America*. **108,** 6680–6685 (2011).

415. Schmid, M. *et al.* A gene expression map of *Arabidopsis thaliana* development. *Nature Genetics*. **37,** 501–506 (2005).

416. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*. **270,** 467–470 (1995).

417. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: A revolutionary tool for transcriptomics. *Nature reviews. Genetics*. **10,** 57–63 (2009).

418. Zhao, S., Fung-Leung, W.-P., Bittner, A., Ngo, K. & Liu, X. Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PLOS ONE*. **9,** e78644 (2014).

419. Ekblom, R. & Galindo, J. Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity*. **107,** 1–15 (2011).

420. Sudmant, P. H., Alexis, M. S. & Burge, C. B. Meta-analysis of RNA-Seq expression data across species, tissues and studies. *Genome Biology*. **16,** 287 (2015).

421. Barrett, T. & Edgar, R. Gene expression omnibus (GEO): Microarray data storage, submission, retrieval, and analysis. *Methods in enzymology*. **411,** 352–369 (2006).

422. Leinonen, R., Sugawara, H. & Shumway, M. The Sequence Read Archive. *Nucleic Acids Research*. **39,** D19–D21 (2011).

423. Kolesnikov, N. *et al.* ArrayExpress update–simplifying data submissions. *Nucleic acids research*. **43,** D1113–6 (2015).

424. Petryszak, R. *et al.* Expression Atlas update–an integrated database of gene and protein expression in humans, animals and plants. *Nucleic Acids Research*. **44,** D746–D752 (2016).

425. Kapushesky, M. *et al.* Gene Expression Atlas at the European Bioinformatics Institute. *Nucleic Acids Research.* **38,** D690–D698 (2010).

426. Kapushesky, M. *et al.* Gene Expression Atlas update–a value-added database of microarray and sequencing-based functional genomics experiments. *Nucleic Acids Research.* **40,** D1077–D1081 (2012).

427. Borrill, P., Ramirez-Gonzalez, R. & Uauy, C. expVIP: A customisable RNA-Seq data analysis and visualisation platform. *Plant Physiology.* **170,** 2172–2186 (2016).

428. Winter, D. *et al.* An 'Electronic Fluorescent Pictograph' browser for exploring and analyzing large-scale biological data sets. *PLOS ONE.* **2,** e718 (2007).

429. Cheng, F. *et al.* BRAD, the genetics and genomics database for Brassica plants. *BMC Plant Biology.* **11,** 136 (2011).

430. Eckes, A. H. *et al.* Introducing the Brassica Information Portal: Towards integrating genotypic and phenotypic Brassica crop data. *F1000Research.* **6,** 465 (2017).

431. Berardini, T. Z. *et al.* The Arabidopsis Information Resource: Making and mining the 'gold standard' annotated reference plant genome. *Genesis (New York, N.Y. : 2000).* **53,** 474–485 (2015).

432. Yu, H., Xu, Y., Tan, E. L. & Kumar, P. P. AGAMOUS-LIKE 24, a dosage-dependent mediator of the flowering signals. *Proceedings of the National Academy of Sciences.* **99,** 16336–16341 (2002).

433. Michaels, S. D. *et al.* *AGL24* acts as a promoter of flowering in *Arabidopsis* and is positively regulated by vernalization. *The Plant Journal.* **33,** 867–874 (2003).

434. Wu, G. & Poethig, R. S. Temporal regulation of shoot development in *Arabidopsis thaliana* by *miR156* and its target *SPL3*. *Development.* **133,** 3539–3547 (2006).

435. Reinhart, B. J., Weinstein, E. G., Rhoades, M. W., Bartel, B. & Bartel, D. P. MicroRNAs in plants. *Genes & Development.* **16,** 1616–1626 (2002).

436. Griffiths-Jones, S. The microRNA registry. *Nucleic Acids Research.* **32,** D109–D111 (2004).

437. Griffiths-Jones, S., Grocock, R. J., Dongen, S. van, Bateman, A. & Enright, A. J. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Research.* **34,** D140–D144 (2006).

438. Griffiths-Jones, S., Saini, H. K., Dongen, S. van & Enright, A. J. miRBase: Tools for microRNA genomics. *Nucleic Acids Research.* **36,** D154–D158 (2008).

439. Kozomara, A. & Griffiths-Jones, S. miRBase: Integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Research.* **39,** D152–D157 (2011).

440. Kozomara, A. & Griffiths-Jones, S. miRBase: Annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Research.* **42,** D68–D73 (2014).

441. Shen, E. *et al.* Identification, evolution, and expression partitioning of miRNAs in allopolyploid *Brassica napus*. *Journal of Experimental Botany.* **66,** 7241–7253 (2015).

442. Hawkins, C. *et al.* An eFP browser for visualizing strawberry fruit and flower transcriptomes. *Horticulture Research.* **4,** 17029 (2017).

443. Soltis, D. E. *et al.* Polyploidy and angiosperm diversification. *American Journal of Botany.* **96,** 336–348 (2009).

444. Mandáková, T., Heenan, P. B. & Lysak, M. A. Island species radiation and karyotypic stasis in *Pachycladon* allopolyploids. *BMC Evolutionary Biology.* **10,** 367 (2010).

445. Comai, L. The advantages and disadvantages of being polyploid. *Nature Reviews Genetics.* **6,** 836–846 (2005).

446. Paterson, A. H. Polyploidy, evolutionary opportunity, and crop adaptation. *Genetica.* **123,** 191–196 (2005).

447. Yao, H., Gray, A. D., Auger, D. L. & Birchler, J. A. Genomic dosage effects on heterosis in triploid maize. *Proceedings of the National Academy of Sciences.* **110,** 2665–2669 (2013).

448. Laufs, P., Grandjean, O., Jonak, C., Kiêu, K. & Traas, J. Cellular parameters of the shoot apical meristem in Arabidopsis. *The Plant Cell.* **10,** 1375–1390 (1998).

449. Williams, L., Grigg, S. P., Xie, M., Christensen, S. & Fletcher, J. C. Regulation of *Arabidopsis* shoot apical meristem and lateral organ formation by microRNA *miR166g* and its *AtHD-ZIP* target genes. *Development.* **132,** 3657–3668 (2005).

450. Iii, L. B. *et al.* Microdissection of shoot meristem functional domains. *PLOS Genetics.* **5,** e1000476 (2009).

451. Samach, A., Kohalmi, S. E., Motte, P., Datla, R. & Haughn, G. W. Divergence of function and regulation of class B floral organ identity genes. *The Plant Cell.* **9,** 559–570 (1997).

452. Goto, K. & Meyerowitz, E. M. Function and regulation of the *Arabidopsis* floral homeotic gene *PISTILLATA*. *Genes & Development.* **8,** 1548–1560 (1994).

453. Gómez-Mena, C., Folter, S. de, Costa, M. M. R., Angenent, G. C. & Sablowski, R. Transcriptional program controlled by the floral homeotic gene *AGAMOUS* during early organogenesis. *Development.* **132,** 429–438 (2005).

454. Svensson, V. *et al.* Power analysis of single-cell RNA-sequencing experiments. *Nat Meth.* **14,** 381–387 (2017).

455. Polisensky, D. H. & Braam, J. Cold-shock regulation of the Arabidopsis *TCH* genes and the effects of modulating intracellular calcium levels. *Plant Physiology.* **111,** 1271–1279 (1996).

456. Leyva-Pérez, M. de la O. *et al.* Early and delayed long-term transcriptional changes and short-term transient responses during cold acclimation in olive leaves. *DNA Research: An International Journal for Rapid Publication of Reports on Genes and Genomes.* **22,** 1–11 (2015).

457. Kaufmann, K., Melzer, R. & Theißen, G. MIKC-type MADS-domain proteins: Structural modularity, protein interactions and network evolution in land plants. *Gene.* **347,** 183–198 (2005).

458. Gramzow, L. & Theissen, G. A hitchhiker's guide to the MADS world of plants. *Genome Biology.* **11,** 214 (2010).

459. Newman, J. R. S. & Keating, A. E. Comprehensive identification of human bZIP interactions with coiled-coil arrays. *Science.* **300,** 2097–2101 (2003).

460. Penfold, C. A., Millar, J. B. A. & Wild, D. L. Inferring orthologous gene regulatory networks using interspecies data fusion. *Bioinformatics.* **31,** i97–i105 (2015).

461. Penfold, C. A., Buchanan-Wollaston, V., Denby, K. J. & Wild, D. L. Nonparametric Bayesian inference for perturbed and orthologous gene regulatory networks. *Bioinformatics.* **28,** i233–i241 (2012).

462. Penfold, C. A. & Wild, D. L. How to infer gene networks from expression profiles, revisited. *Interface Focus.* **1,** 857–870 (2011).

463. Sima, C., Hua, J. & Jung, S. Inference of gene regulatory networks using time-series data: A survey. *Current Genomics.* **10,** 416–429 (2009).

464. Cai, X., Das, S. & Welch, S. A novel strategy for plant breeding based on simulations of gene network models. *International Journal of Bioinformatics Research and Applications.* **9,** 517–533 (2013).

465. Chiang, G. C. K., Barua, D., Kramer, E. M., Amasino, R. M. & Donohue, K. Major flowering time gene, *FLOWERING LOCUS C*, regulates seed germination in *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences.* **106,** 11661–11666 (2009).

466. Camacho, C. *et al.* BLAST+: Architecture and applications. *BMC Bioinformatics.* **10,** 421 (2009).

467. R Core Team. *R: A language and environment for statistical computing.* (R Foundation for Statistical Computing, 2017).

468. Lyons, E. & Freeling, M. How to usefully compare homologous plant genes and chromosomes as DNA sequences. *The Plant Journal.* **53,** 661–673 (2008).

469. Lyons, E., Pedersen, B., Kane, J. & Freeling, M. The value of nonmodel genomes and an example using SynMap within CoGe to dissect the hexaploidy that predates the rosids. *Tropical Plant Biology.* **1,** 181–190 (2008).

470. Haas, B. J., Delcher, A. L., Wortman, J. R. & Salzberg, S. L. DAGchainer: A tool for mining segmental genome duplications and synteny. *Bioinformatics.* **20,** 3643–3646 (2004).

471. Tang, H. *et al.* Screening synteny blocks in pairwise genome comparisons through integer programming. *BMC Bioinformatics.* **12,** 102 (2011).

472. Wehrens, R. & Buydens, L. M. C. Self- and super-organising maps in R: The kohonen package. *J. Stat. Softw.* **21,** 5 (2007).

473. Vesanto, J., Himberg, J., Alhoniemi, E. & Parhankangas, J. Self-organizing map in Matlab: The SOM Toolbox in *In Proceedings of the Matlab DSP Conference.* 35–40 (2000).

474. Duvick, J. *et al.* PlantGDB: A resource for comparative plant genomics. *Nucleic Acids Research.* **36,** D959–D965 (2008).

475. Frazer, K. A., Pachter, L., Poliakov, A., Rubin, E. M. & Dubchak, I. VISTA: Computational tools for comparative genomics. *Nucleic Acids Research.* **32,** W273–W279 (2004).

476. Mayor, C. *et al.* VISTA : Visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics.* **16,** 1046–1047 (2000).

477. Bray, N., Dubchak, I. & Pachter, L. AVID: A global alignment program. *Genome Research.* **13,** 97–102 (2003).

478. Carlson, M. *Org.At.tair.db: Genome-wide annotation for Arabidopsis.* (2017).

479. Alexa, A. & Rahnenfuhrer, J. *TopGO: Enrichment analysis for gene ontology.* (2010).

480. Marchler-Bauer, A. *et al.* CDD/SPARCLE: Functional classification of proteins via subfamily domain architectures. *Nucleic Acids Research.* **45,** D200–D203 (2017).

481. Edgar, R. C. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research.* **32,** 1792–1797 (2004).

482. Larsson, A. AliView: A fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics.* **30,** 3276–3278 (2014).

483. Grigoryan, G. & Keating, A. E. Structural specificity in coiled-coil interactions. *Current Opinion in Structural Biology.* **18,** 477–483 (2008).

484. Jmol: An open-source Java viewer for chemical structures in 3D.

485. Soetaert, K., Petzoldt, T. & Setzer, R. W. Solving differential equations in R: Package deSolve. *Journal of Statistical Software.* **33,** 1–25 (2010).

486. Greb, T. *et al.* The PHD finger protein VRN5 functions in the epigenetic silencing of *Arabidopsis FLC*. *Current Biology.* **17,** 73–78 (2007).

487. Kim, D.-H. & Sung, S. Coordination of the vernalization response through a *VIN3* and *FLC* gene family regulatory network in *Arabidopsis*. *The Plant Cell.* **25,** 454–469 (2013).

Appendix A

Supplementary figures and tables for Chapter 2.

Table 6.4: Gene names for Figure 2.41.

Identifier	Gene name
1.1	GLYMA02G05100.1
1.2	GLYMA04G02420.1
1.3	GLYMA06G02470.2
2.1	GSMUA_Achr4P05090_001
2.2	GSMUA_Achr9P21040_001
2.3	GSMUA_Achr2P03490_001
2.4	GSMUA_Achr5P11220_001
2.5	GSMUA_Achr5P11470_001
2.6	GSMUA_Achr5P17850_001
2.7	GSMUA_Achr4P29580_001
2.8	GSMUA_Achr2P11200_001
3.1	AET03736
3.2	KEH21752
3.3	AES95190

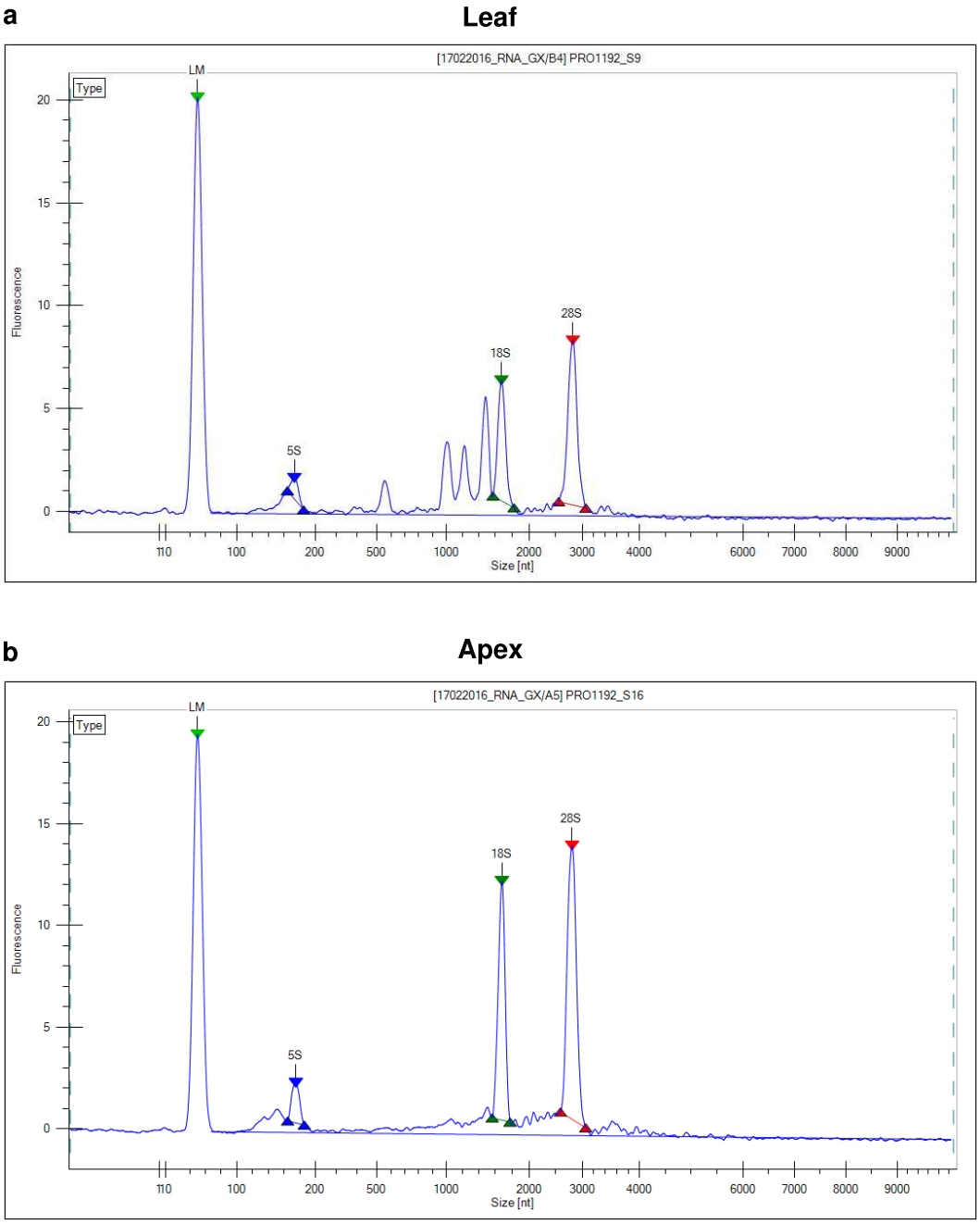


Figure 6.3: Quality assurance plots from the RNA samples submitted for sequencing.

These plots are generated by a Bioanalyzer (Agilent Technologies) to assess the quality of RNA prior to sequencing. The peaks of fluorescence correspond to particular sizes of RNA molecule being present in the sample. The 5S, 18S, and 28S are peaks due to ribosomal RNA. *Continued on Page 326.*

Table 6.5: Gene names for Figure 2.42.

Identifier	Gene name
1.0	TRAES3BF001900070CFD_g
1.1	Traes_3AL_58F294736
1.2	Traes_3DL_20ED2EA4C
2.0	Traes_1BL_DE2CF9613
2.1	Traes_1AL_1FFBFB058
2.2	Traes_1DL_D9BA83221
3.0	Traes_5BL_DE53199D3
4.0	TRAES3BF099600130CFD_g
5.0	TRAES3BF111600130CFD_g
6.0	TRAES3BF099600200CFD_g
7.0	TRAES3BF111600160CFD_g
7.1	Traes_3AL_FC5523394
8.0	TRAES3BF111600080CFD_g
9.0	TRAES3BF019000220CFD_g
10.0	Traes_5BL_FB4EDEA83
10.1	Traes_5DL_73CE92096
11.0	Traes_2BS_84FB90D88
12.0	Traes_4BL_4C9A415F3
12.1	Traes_4DL_F38ED7FB6
12.2	Traes_4AS_F9C171219
1.0	GRMZM2G161009
1.1	GRMZM2G033413
1.2	GRMZM2G008166
1.3	GRMZM2G157722
1.4	GRMZM2G002075
1.5	GRMZM2G168079
1.6	GRMZM2G132868
1.7	GRMZM5G858197
1.8	GRMZM2G438293
1.9	GRMZM2G159134

Continued from Page 324. These plots show that in the leaf (**a**) additional peaks are observed in the range 900 - 1600 nucleotides compared to the apex (**b**). These peaks are likely due to chloroplast RNA. That they are absent in the apex sample suggests the dissection protocol was able to adequately remove the surrounding leaf tissue from the apex.

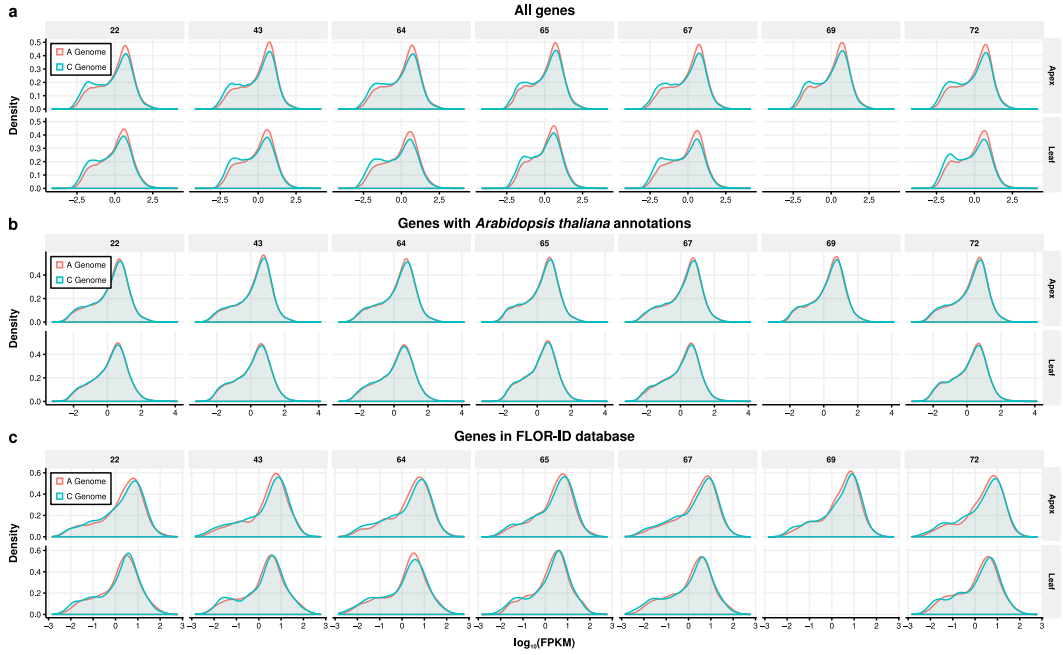


Figure 6.4: Expression differences between A and C genomes are consistent across different tissues and time points.
Density plots of transformed expression levels ($\log_{10}$(FPKM)) calculated using different subsets of genes. The data used to generate the density plots consisted of expression data from: **a** all annotated *B. napus* genes, **b** *B. napus* genes that show sequence conservation to an annotated Arabidopsis thaliana gene, and **c** *B. napus* genes that show sequence conservation to an annotated Arabidopsis gene that is present in the FLOR-ID database[299].

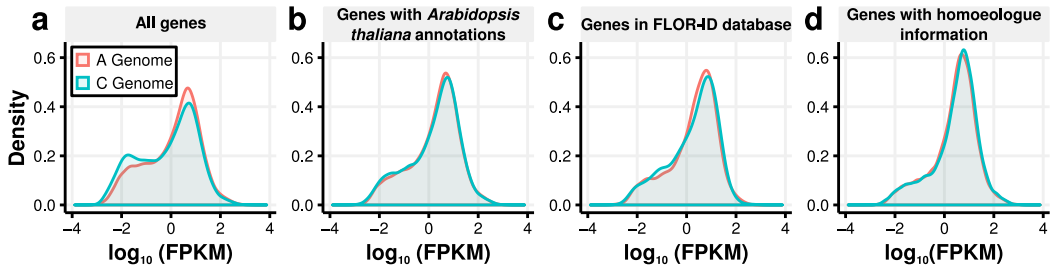


Figure 6.5: Genes for which homoeologue information is available have fewer genes within the very low region of expression.
Density plots of transformed expression levels ($\log_{10}$(FPKM)) calculated using different subsets of genes. The data used to generate the density plots consisted of expression data from: **a** all annotated *B. napus* genes, **b** *B. napus* genes that show sequence conservation to an annotated Arabidopsis gene, **c** *B. napus* genes that show sequence conservation to an annotated Arabidopsis gene that is present in the FLOR-ID database[299], and **d** *B. napus* genes for which homoeologue information is available. These plots are generated using apex expression data from the time point taken at day 22.

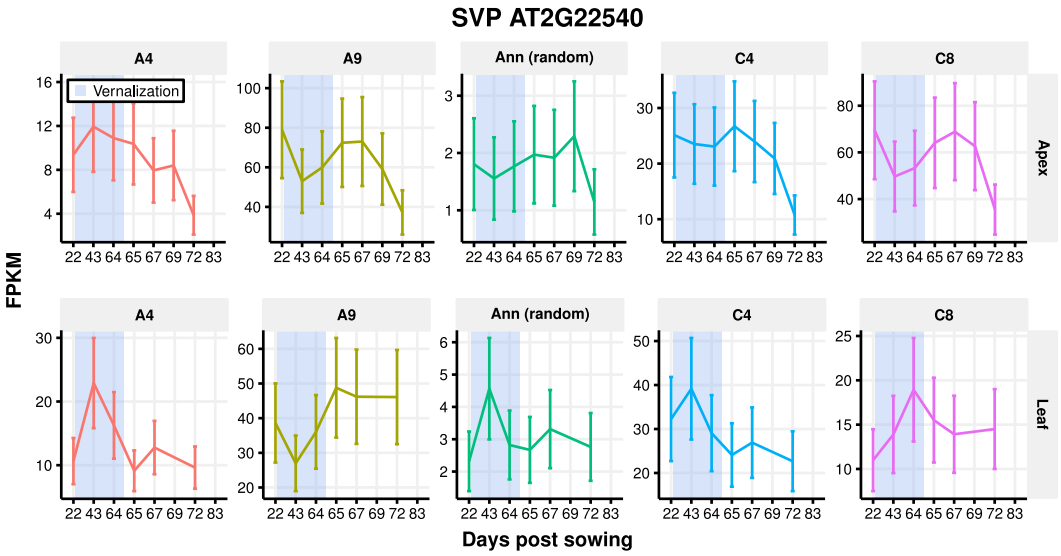


Figure 6.6: Expression traces for the *BnSVP* genes in Westar.
The expression values in FPKM and the 95 % confidence intervals of those expression values as computed by Cufflinks are displayed.

Appendix B

Supplementary analysis of PRC2 and PHD proteins for Chapter 3.

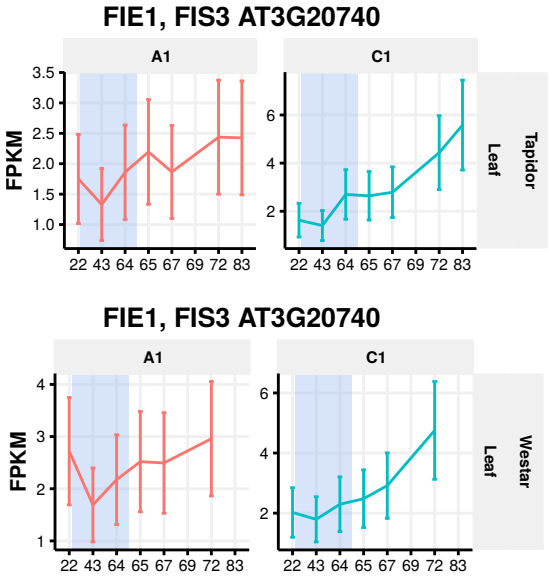


Figure 6.7: Expression traces for the *BnFIE1* genes in the leaf.
The expression values and the 95 % confidence intervals of those expression values as computed by Cufflinks are displayed.

The homologue of *D. melanogaster* Esc, *FIE1*, is the only annotated Arabidopsis homologue of the gene, and is a component of all identified PRC2 complexes in the plant[382,387]. In *B. napus* there are three copies of the gene expressed; the A1 and C1 copies are expressed in both tissues while the Ann copy is only expressed in the apex (Figure 6.8). Although the copies are expressed very similarly in both varieties, the genes show tissue-specific expression. In the leaf, *BnFIE1.A1* is relatively lowly expressed and exhibits a gradual increase across development. *BnFIE1.C1* is more highly expressed than the A1 copy, and shows

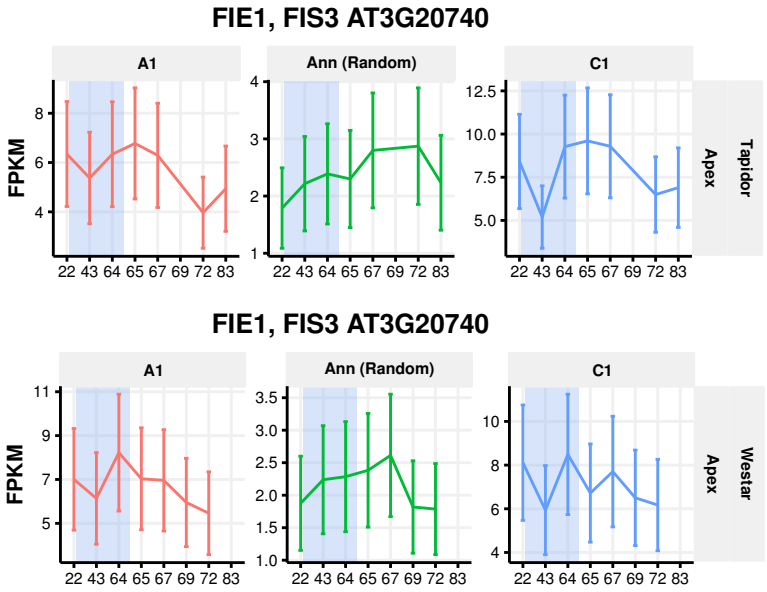


Figure 6.8: Expression traces for the *BnFIE1* genes in the apex. The expression values and the 95 % confidence intervals of those expression values as computed by Cufflinks are displayed.

more pronounced increase in expression during the time series. In the apex all copies of *BnFIE1* are expressed more highly (Figure 6.8). The expression profiles exhibited by the A1 and C1 copies decrease across development, in contrast to their behaviour in the leaf. Although *BnFIE1.Ann.Random* is above the expression threshold, the expression is 3 to 4 fold lower relative the the A1 and C1 copies, suggesting this copy does not play as important a role in the function of PRC2 in *B. napus*.

The histone methyltransferase *SWN* is associated with the PRC2 complex that influences the vernalization response in Arabidopsis. As with *B. napus* copies of *VRN2*, the *BnSWN* genes are relatively consistent in their expression in both varieties and tissues, with slight increases in expression during the cold treatment (Figure 6.9). Although an additional *BnSWN* copy is expressed above the expression threshold in the leaf (Figure 6.9) the expression is very low and only just above the 2.0 FPKM threshold. Due to this low expression the relevance of the gene to the vernalization response is likely to be low.

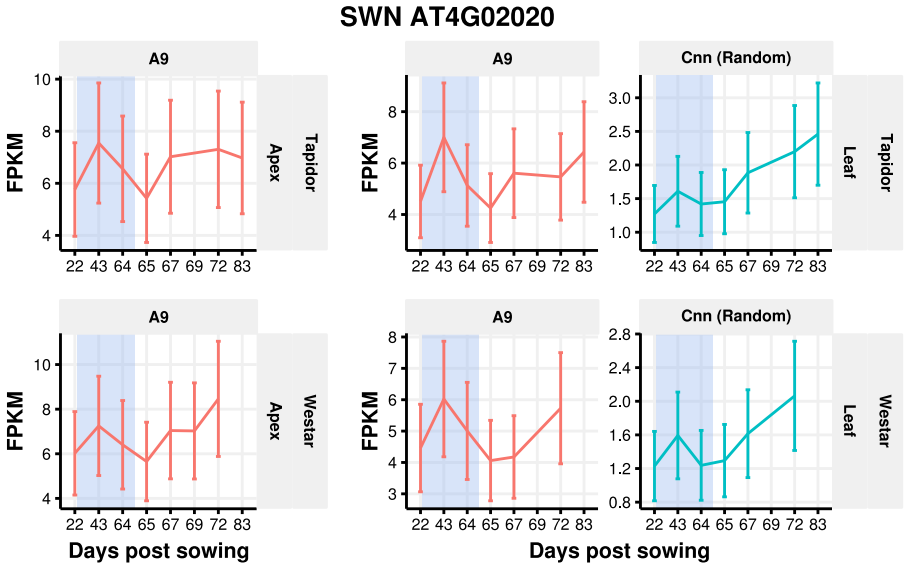


Figure 6.9: Expression traces for the *BnSWN* genes.
The expression values and the 95 % confidence intervals of those expression values as computed by Cufflinks are displayed.

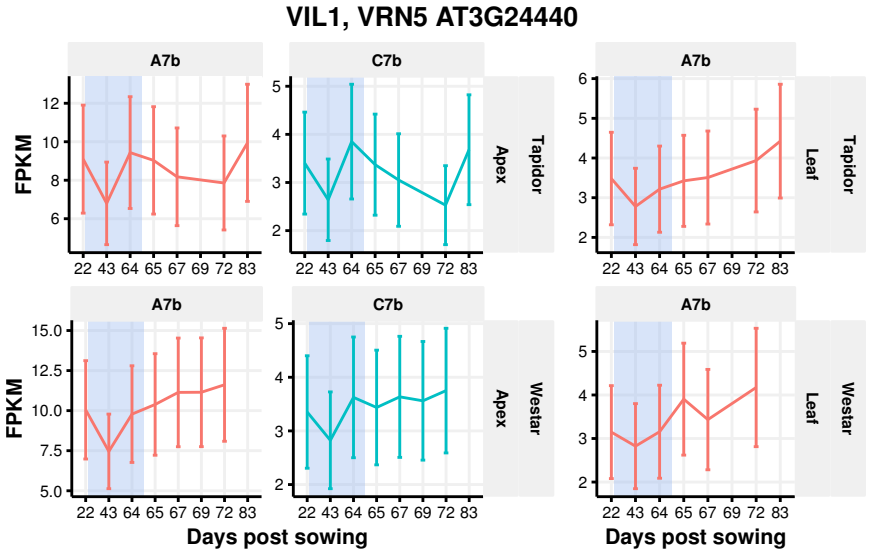


Figure 6.10: Expression traces for the *BnVIL1* genes.
The expression values and the 95 % confidence intervals of those expression values as computed by Cufflinks are displayed.

VIL1 shows only a slight decrease in expression during the vernalization period, with relatively constant expression after the cold (Figure 6.10). This is the case in both varieties, with the magnitude of expression of both copies being very similar in both the winter and the spring. The A7b copy is expressed in both the apex and the leaf, while the C7b copy is only detected in the apex, suggesting potential tissue-specific expression of the copies. The expression patterns of *BnVIL1* deviate from that of *VIL1* in Arabidopsis, in that the expression was found to increase during short day growth[398]. However, the expression in both leaf and apex is consistent with results from the model species[486].

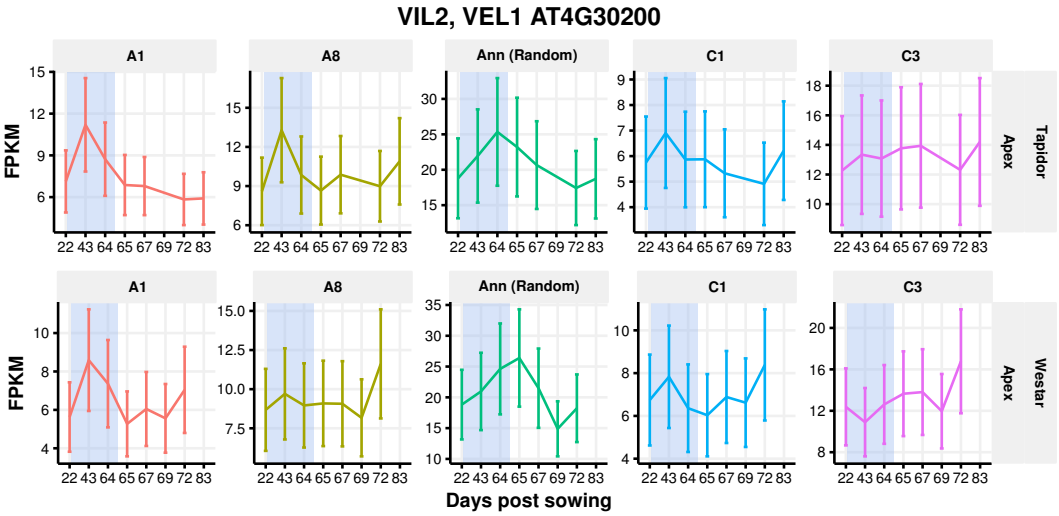


Figure 6.11: Expression traces for the *BnVIL2* genes in the apex. The expression values and the 95 % confidence intervals of those expression values as computed by Cufflinks are displayed.

VIL2 has been found to be associated with the vernalization associated PRC2[386], although down regulation of the gene did not affect flowering time of vernalized plants[486]. However, an increase in expression of the gene during vernalization has also been reported[487], making the role of the gene during vernalization somewhat ambiguous. Five copies of the gene are expressed in *B. napus*, all of which show remarkable similarities in both expression profile and magnitude in the apex (Figure 6.11) and leaf (Figure 6.12).

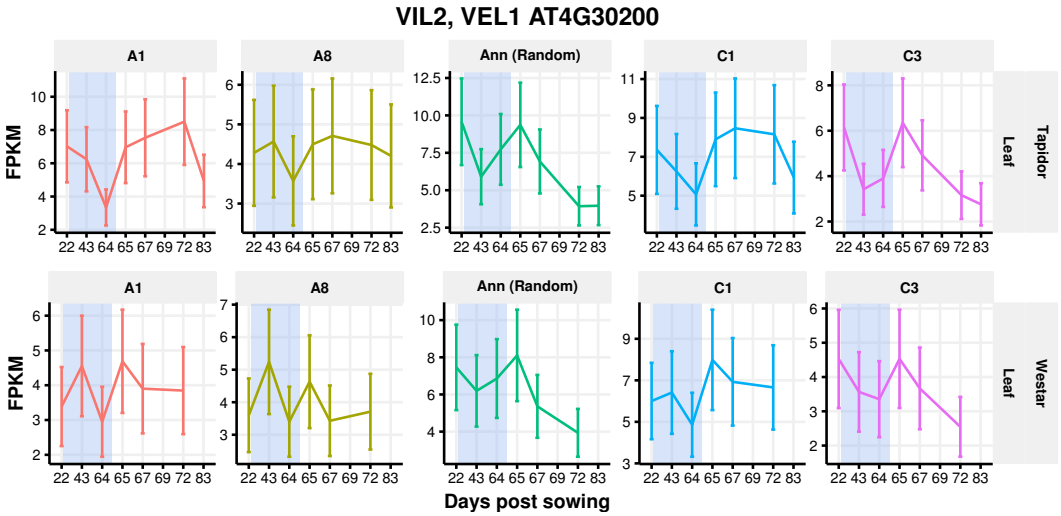


Figure 6.12: Expression traces for the *BnVIL2* genes in the leaf. The expression values and the 95 % confidence intervals of those expression values as computed by Cufflinks are displayed.

Interestingly, three of the copies exhibit tissue-specific responses to the cold treatment. The A1, A8, and Ann, and C1 copies all exhibit increases during the vernalization period in the apex (Figure 6.11), yet show expression decreases in the leaf (Figure 6.12). As the behaviours of these genes are so similar between the varieties, however, it is unlikely that they directly contribute to the flowering time differences observed between Westar and Tapidor.

Appendix C

D. Marc Jones, Rachel Wells, Nick Pullen, Martin Trick, Judith A. Irwin, Richard J. Morris. 2017. **Regulatory divergence of flowering time genes in the allopolyploid *Brassica napus*.** bioRxiv doi: 10.1101/178137

Regulatory divergence of flowering time genes in the allopolyploid *Brassica napus*

D. Marc Jones[1,2], Rachel Wells[1], Nick Pullen[1], Martin Trick[2], Judith A. Irwin[1*], Richard J. Morris[1,2†]

[1] *Crop Genetics, John Innes Centre, Norwich Research Park, Norwich. NR4 7UH. United Kingdom.*

[2] *Computational and Systems Biology, John Innes Centre, Norwich Research Park, Norwich. NR4 7UH. United Kingdom.*

[*]Corresponding author: judith.irwin@jic.ac.uk

[†]Corresponding author: richard.morris@jic.ac.uk

Abstract

Polyploidy is a recurrent feature of eukaryotic evolution and has been linked to increases in complexity, adaptive radiation and speciation. Within angiosperms, such events occur repeatedly in many plant lineages. We investigated the role of duplicated genes in the regulation of flowering in *Brassica napus*. This relatively young allotetraploid represents a snapshot of evolution and artificial selection in progress. In line with the gene balance hypothesis, we find preferential retention of expressed flowering time genes relative to the whole genome. Furthermore, gene expression dynamics across development reveal diverged regulation of many flowering time gene copies. This finding supports the concept of responsive backup circuits being key for the retention of duplicated genes. A case study of *BnaTFL1* reveals differences in cis-regulatory elements downstream of these genes that could explain this divergence. Such differences in the regulatory dynamics of duplicated genes highlight the challenges for translating gene networks from model to more complex polyploid crop species.

Many economically important crops exhibit extensive gene multiplication as a result of recent or ancestral polyploidy[1], for example wheat (*Triticum aestivum*)[2], cotton (*Gossypium hirsutum*)[3], and oilseed rape (OSR, *Brassica napus*)[4]. The presence of multiple copies of a gene relaxes natural and artificial selective pressures on any one individual copy, facilitating the emergence of novel gene functions[5]. The resulting increase in variation can be exploited to breed crop varieties with desirable phenotypes[6]. The presence of multiple orthologues, however, hinders efforts to translate knowledge of gene function and, in particular of regulatory networks, from model to crop species. This is a consequence of not knowing which orthologue, if any, retains the same function as the corresponding gene in the model species, whether ancestral functions have been partitioned between them, or if a novel function has been acquired[7].

The evolutionary fate of gene copies arising from a gene duplication event has been studied in a range of species[8–11]. There are two main classes of gene duplication events: small scale duplications and whole genome duplications (WGD)[5,7,12–14]. These two types of duplication event can lead to different outcomes for gene copies[13]. Whilst gene redundancy has been reported to be evolutionarily unstable[7,15], it is frequently observed[12,16–18]. A proposed driver for the retention of duplicate genes is the maintenance of gene dosage, known as the gene balance hypothesis[14,19–23]. Such dosage constraints may result if the gene product acts as part of a protein complex, where an incorrect stoichiometry of proteins can lead to the appearance of deleterious phenotypes[14]. WGDs maintain the original stoichiometry, resulting in duplicated, dosage sensitive gene orthologues being retained[14,20,23]. Conversely, small scale duplication of individual genes without their partners disrupts protein stoichiometry and disfavours gene retention[19]. Simulations of the dynamics of gene duplication events suggest that genes whose products form protein complexes, such as those associated with kinase activity,

transcription, protein binding and modification, and signal transduction, are preferentially retained in the genome for longer when copied in whole genome relative to small scale duplications[19,24]. Data from a range of species are consistent with gene dosage balance[25–29], including studies focusing on gene retention in the Arabidopsis genome[12,24]. In *Saccharomyces cerevisiae*, genes retained following a WGD are enriched for those that in diploids have haploinsufficiency or overexpression phenotypes, suggesting that the dosage of these genes is important[9]. One expectation of the gene balance hypothesis, illustrated in *S. cerevisiae*[20], is that duplicated genes are more likely to be co-regulated[20,23]. This co-regulation fits with the concept of buffering against stochastic effects in development[30,31]. Studying the regulation of duplicated genes can therefore provide clues for understanding their retention in the genome.

The *Brassica* genus contains several diploid crop species derived from ancestors that underwent a genome triplication event 5 to 28 million years ago[32–34]. OSR is an allopolyploid resulting from the interspecific hybridisation of two diploid species, *Brassica rapa* and *Brassica oleracea*[4]. An important agronomic trait for all Brassica crops is flowering time[35–38], as different growing regions require varieties with very different phenologies. Flowering time has been extensively studied in the model species Arabidopsis[39–41], revealing that flowering time genes are involved in multiple interactions and that many are transcription factors[41,42]. Thus, following the gene balance hypothesis, in a polyploid such as OSR, we would expect orthologues of Arabidopsis flowering time genes to have been preferentially retained relative to other genes in the genome, analogous to previous results that show preferential retention of genes involved with the circadian rhythm in paleopolyploid *B. rapa*[43]. That aspects of flowering time control are conserved between the Arabidopsis and OSR[37,44,45] makes OSR an interesting and agronomically important model to investigate the evolution of gene function following

gene multiplication.

Here we show that data from a transcriptomic time series (global gene expression in the first true leaf and shoot apex prior to and during the floral transition in OSR) support the prediction of preferential retention for flowering time genes in the genome (Figure 1). Through comparative gene expression and cluster analysis we demonstrate that the regulation of many flowering time gene homologues has diverged, suggesting this may be important for their retention. As an exemplar, using knowledge of cis-regulatory elements downstream of the Arabidopsis *TERMINAL FLOWER 1* (*AtTFL1*) gene, we identify sequence variation that correlates with regulatory differences observed for orthologues of *AtTFL1* in OSR. This case study highlights the importance of homologue expression dynamics in characterising gene regulation. The differences in *BnaTFL1* expression dynamics between homologues suggests that, in addition to proposed gene dosage effects, regulatory divergence may be important for gene retention.

Results

OSR exhibits genome level expression bias across tissue types

Previous reports have demonstrated genome dominance in polyploids[46–48]. To test whether this is the case for OSR, we collected gene expression data through the vegetative to reproductive transition in a doubled haploid (DH) line derived from the spring OSR variety Westar (Figure 2). We compared global expression differences between the A and C genomes in the apex and the first true leaf across all time points (Figure 3; Supplementary Figure 1). We find that the A genome has a greater proportion of highly expressed genes than the C genome. Conversely, for genes showing very low expression

we find the opposite relationship (Figure 3a). Similar distributions are found but are less pronounced when only OSR genes showing sequence conservation to annotated Arabidopsis genes are considered (Figure 3b) and when the sample is further restricted to OSR flowering time genes (Figure 3c). In contrast to the tissue-specific genome bias demonstrated in cotton[49], our results are consistent across the two tissue types and throughout the time series (Supplementary Figure 1).

To investigate A and C genome expression at the gene level, we compared pairs of homoeologous genes that we identified using synteny and sequence similarity[4]. We classified a homoeologous pair as showing biased expression toward one genome if that gene has an expression level (measured in Fragments Per Kilobase of transcript per Million mapped reads, FPKM) at least two-fold higher than its homoeologue. At the individual gene level, biased expression was observed towards both genomes, but with 1.5 to 2.0 times as many genes showing bias towards the C rather than the A genome (16.9% towards the C genome relative to 9.7% towards the A genome in the apex, and 15.2% compared to 8.2% in the leaf; Table 1). This pattern is consistent with the findings of Chalhoub et al. (2014) and is maintained across all time points (Supplementary Table 2). The distributions of fold expression changes reveal that homoeologous gene pairs exhibiting a 2 to 8-fold change are primarily responsible for the observed bias (Supplementary Figure 2). Therefore, the homoeologue-level analyses reveal expression bias towards both the A and C genomes that are consistent across the tissue types tested and result in an absence of genome dominance (Supplementary Table 2). At the whole genome level, however, we observe a bias towards the A genome. This discrepancy may be due to genes with low expression levels tending to lack homoeologue pair information (Supplementary Figure 3). Alternatively, this bias may reflect a known higher incidence of homoeologous exchanges in which C genome copies of individual genes are replaced

by their A genome counterparts[50].

OSR expresses a higher number of flowering time gene homologues relative to the whole genome

To test the prediction that flowering time genes are preferentially retained relative to the whole genome (Figure 1), we evaluated whether this was the case for genes expressed during the floral transition. A gene was considered to be expressed if the maximal expression level during the developmental time series was equal to or exceeded 2 FPKM, with leaf and shoot apex tested separately. We assessed the distributions of annotated (Figure 4a) and expressed OSR flowering time genes (Figure 4b and 4c). In both leaf and shoot apex (Figure 4b and 4c), a shift towards the expression of a higher number of flowering time gene copies relative to the whole genome can be observed. To test whether this observation was caused by the retention of circadian genes, as has been reported in *B. rapa*[43], we repeated this analysis after removing this set of genes and found that the pattern remained (Supplementary Figure 4). This confirms the preferential retention of flowering time genes in OSR and suggests that the multiple orthologues of Arabidopsis flowering time genes retained in the genome could be functional.

Analyses of gene expression differences reveals regulatory divergence of retained flowering time genes in OSR

Having shown that genes involved in the control of flowering time are retained as multiple homologues in the OSR genome we next investigated their regulatory control. We first examined global gene tissue specificity and found that of the 45,048 genes expressed across the developmental time series, 16% show apex specific expression and 11% show

leaf specific expression, with the rest (73%) exhibiting expression in both tissues (Supplementary Figure 6). Focussing on annotated orthologues of Arabidopsis flowering time genes, 61% have at least one orthologue in OSR that is not expressed in the apex, compared to 69% in the first true leaf (Figure 5).

We next used Weighted Gene Co-expression Network Analysis (WGCNA) to identify regulatory modules. WGCNA uses normalised expression data to cluster genes together based on their temporal expression profiles rather than expression levels *per se*. We used these cluster assignments to assess the regulatory control of flowering time gene homologues. Based on the premise of tight co-regulation of dosage-sensitive or functionally redundant genes[20,31], our null hypothesis is that all OSR orthologues of an Arabidopsis flowering time gene will have similar expression patterns, leading to orthologues being in the same regulatory module (dashed lines in Figure 6). We found that most OSR flowering time genes (74% in apex, 64% in leaf) do not conform to this null hypothesis (Figure 6). Thus, analysis of both the overall level of expression in both leaf and shoot apex and WGCNA reveal regulatory divergence between retained homologues of flowering time genes in OSR, suggesting regulatory variation between homologues.

Self-organising map based clustering captures different patterns of regulatory divergence for OSR orthologues of the flowering time genes *AtTFL1*, *AtFT*, and *AtLFY*

To further assess differences in regulation between gene homologues we analysed the divergence of expression over time. Whilst WGCNA assigns expression profiles to regulatory modules, the similarity between profiles is not quantified and genes that could

be assigned to multiple regulatory modules are only assigned to a single module. Furthermore, WGCNA does not account for uncertainty in the RNA-Seq data in the assignment of regulatory modules. To address these issues, we employed a self-organising map (SOM) based sampling approach to assess expression profile divergence (Supplementary Figure 8). Figure 7a illustrates the five possible patterns of regulatory module assignment: (1) a *distinct* pattern of multiple regulatory modules with genes assigned to a single module; (2) a *gradated* pattern of multiple modules where gene membership of individual modules overlap; (3) a *unique* pattern (a special case of the *distinct* pattern) where each copy of a gene is assigned to a different module; (4) a *redundant* pattern where all genes are assigned to the same regulatory module; (5) a *mixed* pattern with some modules showing overlap in gene membership and others not. This approach allows us to robustly analyse expression similarity. Of 85 pairs of homoeologues expressed in the apex, 67 (79%) are found in the same regulatory module. In the leaf, 53 of 69 (77%) of expressed homoeologous pairs are found in the same module, with 29 of the co-regulated pairs being common between the two tissues (Additional File 1). The percentage of Arabidopsis genes with at least two expressed homologues in the apex (leaf) exhibiting each of the regulatory module assignments are 25% (26%) *distinct*, 9% (6%) *gradated*, 23% (23%) *unique*, 39% (33%) *redundant*, and 3% (6%) *mixed* (Supplementary Figure 8).

To investigate further we chose three central Arabidopsis flowering time genes *AtLFY*, *AtFT* and *AtTFL1*. These genes form key hubs in the regulatory network responsible for the switch to flowering in rapid cycling Arabidopsis[51]. Each of these genes has four expressed orthologues in OSR with *BnaTFL1* and *BnaLFY* expressed in the apex and *BnaFT* expressed in leaf tissue. SOM analysis revealed that orthologues of *AtLFY*, *AtFT* and *AtTFL1* in OSR exhibit three different patterns of regulatory module assignment;

redundant, *gradated* and *unique* respectively.

Homologues of *BnaLFY* exhibit a *redundant* pattern of regulatory module assignment, with each of the expression profiles in the apex showing low expression initially and an increase after the vernalisation period (Figure 7d), analogous to observations of *AtLFY* expression in Arabidopsis[52]. Co-regulation of *BnaLFY* homologues is consistent with the gene balance hypothesis[20,23] and is supported by *AtLFY* displaying dosage sensitivity[52,53].

The four *BnaFT* homologues exhibit a *gradated* pattern with two modes of regulation (Figure 7c). The expression of all homologues of *BnaFT* decreases during vernalisation and returns to pre-vernalisation levels when the plants are returned to growth in warm, long day conditions. The *BnaFT* expression profiles diverge at the final time point (day 72) with the A7 and C6 homoeologues showing a pronounced decrease in expression between days 67 and 72. The decrease in expression of *BnaFT.A7* is not as marked as that of its homoeologue, resulting in its assignment to both regulatory modules. The *BnaFT* homologues expressed in the leaf therefore exhibit a gradient of regulatory responses, with *BnaFT.A2* and *BnaFT.C2* having divergent expression traces relative to *BnaFT.C6*, but with *BnaFT.A7* showing similarities to all homologues.

OSR orthologues of *AtTFL1* are an example of *unique* regulatory module assignment with each of the four *BnaTFL1* genes assigned to different modules (Figure 7b). *BnaTFL1.A10* is expressed before and during cold with an immediate increase in expression when the plants are returned to growth in warm, long day conditions. *BnaTFL1.C2* also shows stable expression before and during cold but in contrast to *BnaTFL1.A10* decreases in expression when the plants are returned to warm, long day conditions. *BnaTFL1.C3* exhibits reduced expression levels post-cold with a transient peak of expression at day 69. The fourth homologue (mapped to the Darmor-*bzh* C genome and with greatest sequence identity to *BolTFL1.C9* from the EnsemblPlants database[54]) shows increased

expression during cold followed by a steady decrease when plants are returned to warm, long day conditions. These four expression profiles are *unique* as shown in the clustering coefficient heatmap (Figure 7b). Homologues *BnaTFL1.A10* and *BnaTFL1.C3* exhibit expression profiles with the greatest similarity to *AtTFL1*[55] as both show increasing expression during the floral transition.

AtLFY, *AtFT* and *AtTFL1* integrate environmental signals to determine the timing of the floral transition[56–60]. That individual orthologues of these genes in OSR show different patterns of regulatory module assignment suggests that the selective pressures acting on them are different, even though they belong to the same regulatory pathway in Arabidopsis. This result mirrors findings in Arabidopsis where it was found that less than half of gene pairs derived from the most recent duplication still retained significantly correlated expression profiles[12,26].

Patterns of intergenic sequence conservation surrounding *BnaTFL1* genes provide a potential explanation for the observed regulatory divergence

Downstream regulatory sequences of *AtTFL1* in Arabidopsis have been shown to be important for spatiotemporal control of expression[61]. We therefore investigated whether similar variation could explain the *distinct* pattern of regulation displayed by the four *BnaTFL1* orthologues. We analysed sequence conservation between OSR and Arabidopsis in the 5' and 3' intergenic regions surrounding *BnaTFL1*, identifying several conserved regions (Figure 8). Focussing on areas previously identified as *AtTFL1* cis-regulatory elements in Arabidopsis[61], we find variation in the degree of sequence conservation between *BnaTFL1* orthologues (Figure 8a). Sequence conservation within

regions II and IV of *BnaTFL1.A10* and *BnaTFL1.C3* suggests Arabidopsis-like cis-regulatory elements are present downstream of these genes. These *BnaTFL1* orthologues, that increase in expression during the floral transition, show high sequence conservation in region II. Conversely, *BnaTFL1.Cnn* and *BnaTFL1.C2,* which are not upregulated during the floral transition, lack sequence conservation in this region. Region II was found to be necessary for the upregulation of *AtTFL1* during the floral transition in Arabidopsis[61], which correlates with this result. Region IV may also be involved in the observed expression trace divergence between *BnaTFL1* homologues, as this region was found to be important for the expression of *AtTFL1* in the inflorescence meristem.

Sequence conservation within region III is below 50% in *BnaTFL1.Cnn*, whilst for the other three homologues it is 81%, 87%, and 78% for *BnaTFL1.A10*, *BnaTFL1.C2*, and *BnaTFL1.C3,* respectively. Interestingly, the range of significant sequence conservation in *BnaTFL1.C2* (154 bases) and *BnaTFL1.A10* (162 bases) is decreased compared to that of *BnaTFL1.C3* (273 bases), potentially suggesting the cis-regulatory elements in the former two copies are incomplete.

Serrano-Mislata et al. (2016)[61] identified additional regions conserved across species that were not experimentally implicated in the regulatory control of *AtTFL1* (green shading in Figure 8). We observe sequence divergence in one of these regions, region G. Interestingly it is *BnaTFL.A10* and *BnaTFL1.C3,* which exhibit expression profiles most like that of *AtTFL1,* that show sequence conservation in this region. *BnaTFL1.A10* exhibits high sequence conservation relative to Arabidopsis across this entire region, while *BnaTFL1.C3* shows conservation over ~50% of the region. As with regions II and IV, *BnaTFL1.C2* and *BnaTFL1.Cnn* lack conserved sequence in region G. We also identified a region of conservation not annotated in the previous analysis of *AtTFL1* cis-regulatory elements. This region, situated ~600 bp upstream of the transcription start site

of *AtTFL1,* shows ~80% sequence conservation relative to Arabidopsis in *BnaTFL1.A10*, *BnaTFL1.C2* and *BnaTFL1.Cnn*. In *BnaTFL1.C3*, sequence conservation in this region is ~55%.

To confirm the expression differences we observe between the *BnaTFL1* orthologues we performed copy-specific RT-qPCR across the developmental time series (Figure 8b). The RT-qPCR results show good correspondence with the RNA-Seq results, confirming our findings. Thus, using sequence conservation we determine the presence/absence of cis-regulatory elements downstream of the *BnaTFL1* genes that may confer similar regulatory control in OSR as in Arabidopsis. *BnaTFL1* orthologues contain different combinations of cis-regulatory elements, which have the potential to underlie the divergent expression traces they exhibit.

Discussion

WGD events are thought to have occurred in most, if not all, angiosperm lineages[62] and are well documented in the Brassicaceae[33,63] Whole genome triplication[32–34] and interspecific hybridisation events[4] have resulted in extensive gene multiplication in Brassica species relative to the Arabidopsis lineage. WGD is considered a driving force in angiosperm diversification[64], introducing genetic redundancy and allowing the evolution of novel gene function and new interactions, leading to neo- and subfunctionalisation. WGDs are usually followed by a process of "diploidisation"[65] that includes genome downsizing[66], chromosome rearrangement and number reduction[67], and gene loss[68]. So, whilst many additional gene copies gained from WGD are likely to be lost over time, the analysis of genomic sequences has revealed that a significant number of duplicated genes are nevertheless present in the genomes of many species[12,16–18]. For

instance, in the Arabidopsis lineage around 30% to 37% of homoeologous gene duplicates have been retained[25,69]. Based on such observations, modelling studies have determined conditions under which duplicated genes can become evolutionary stable[14,30]. These ideas have given rise to the gene balance hypothesis, which states that dosage sensitive genes are preferentially retained in the genome after WGD, but tend to be lost after local duplication events[20,23]. Kinases, transcription factors and proteins that form part of a complex fall into this category. From the gene balance hypothesis, we might therefore expect that highly networked genes such as those that regulate flowering time[40,41,70] have been preferentially retained in the genome.

This study determines the expression profiles of OSR genes prior to and during the floral transition. We compared expression profiles across development to infer whether orthologues of Arabidopsis flowering time genes retain similar patterns of regulation. Whilst our analysis reveals that a significant proportion of duplicated genes in OSR have divergent regulation (Figures 5 and 6, Supplementary Figures 8b and 8c), it shows that the more recently combined homoeologues are frequently found in the same regulatory module (79% in the apex and 77% in the leaf). The finding of homoeologues tending to be co-regulated in allotetraploid OSR is intriguing, given the comparatively recent origin. An analysis of 2,000 pairs of paralogous genes in *Gossypium raimondii*, resulting from a 5- to 6-fold ploidy increase ~60 Mya, revealed more than 92% of gene pairs exhibited expression divergence[71]. Most of these gene pairs show complementary expression patterns in different tissues, consistent with the idea of responsive backup circuits[15,31]. It is therefore tempting to speculate that regulation of homoeologues in OSR is still in flux with near-complete divergence a likely consequence of “diploidisation” across much longer timeframes. This hypothesis is supported by the finding that in recently synthesised allotetraploid cotton, most homoeologues display similar expression patterns

across multiple tissue types[49] while in allotetraploid upland cotton (*G. hirsutem;* which arose 1-2 Mya) 24% of homoeologues show diverged expression patterns. Recent genomic studies also support the idea that the OSR genome is in flux[50,72], potentially in response to artificial selection for agronomically important traits.

Gene expression can be controlled through a range of mechanisms. This study highlights the potential role cis-regulatory elements may play in the divergence of gene regulation. Expression divergence of *AtTFL1* orthologues in OSR correlates with the presence and absence of sequence conservation within regions downstream of the gene. Serrano-Mislata et al. (2016) identified these regions as cis-regulatory elements and dissected their roles in the spatiotemporal regulation of *AtTFL1*[61]. *AtTFL1* expression dynamics exhibited by Arabidopsis mutants lacking the identified cis-regulatory elements show striking similarities to those of *BnaTFL1* orthologues lacking sequence similarity to the elements. This suggests conserved function of cis-regulatory elements between Arabidopsis and OSR and highlights that such variation can potentially drive the regulatory divergence of gene homologues. Although the patterns of sequence conservation downstream of *AtTFL1*[61] are retained in OSR orthologues (Figure 8), we have not demonstrated that the changes in these cis-regulatory elements are causative. The differences in region II correlate with the up-regulation of *BnaTFL1* at the floral transition. This region is not conserved in *BnaTFL1.Cnn*, which also lacks high levels of sequence conservation in region III. The latter is associated with the expression of *AtTFL1* in Arabidopsis lateral meristems[61] and thus predicts that *BnaTFL1.Cnn* is not expressed in this tissue.

We have shown that gene dosage and regulatory divergence may have contributed to the over-retention of flowering time genes in OSR. Without biochemical data on the proteins encoded by the genes, we are not able to distinguish whether homologues with diverged

expression patterns have maintained their original molecular functions (redundant), specialised such that the initial function is split between gene duplicates (subfunctionalisation), or developed a novel function (neofunctionalisation). However, following the responsive backup circuit concept, we would expect them to have significant functional overlap.

The presence of multiple gene homologues within crop species complicates the translation of regulatory networks from models to polyploid crops, hampering breeding and selection strategies. Knowledge of functional divergence will support future breeding efforts by allowing more targeted, homologue-specific crop improvement strategies. Detailed knowledge of the function of specific copies of genes, their regulation and importantly how this functionality is combined to determine crop plasticity will be key for targeted approaches for crop improvement.

Methods

Plant growth and sample preparation

Brassica napus cv. Westar plants were sown on the 7th May 2014 in cereals mix. Plants were grown in unlit glasshouses in Norwich, UK, with glasshouse temperatures set at 18 °C during the day and 15 °C at night. On the day 22, plants were transferred to a 5 °C, short day (8 hour) vernalisation room. Although Westar is classed as a spring cultivar of OSR, it may still show a mild response to the vernalisation period. After a 42-day period in the vernalisation room, plants were transferred back to unlit glasshouses and grown until the plants flowered.

The first true leaf of each plant and shoot apices were sampled at 22, 43, 64, 65, 67, 69,

and 72 days after sowing (Supplementary Table 1). First true leaves were cut and immediately frozen in liquid nitrogen. The growing shoot apices were dissected using razor blades on a dry ice chilled tile before transfer to liquid nitrogen.

Samples were pooled and ground in preparation for RNA extraction. For apex tissue, ~0.1 g of apices were ground as a pool. At the early time points, as the apices were smaller, this mass of tissue equated to approximately 20 plant apices, while at later time points approximately 10 apices were pooled. For leaf samples, between 6-10 leaf samples from separate plants were pooled and ground. RNA extraction and DNase treatment was performed following the method provided with the E.Z.N.A® Plant RNA Kit (Omega Bio-tek Inc., USA).

Library preparation and RNA sequencing was carried out by the Earlham Institute (Norwich, UK). RNA-Seq was performed on RNA samples from six time points for leaf tissue and seven time points from apex tissue. 100bp, single end reads were generated using an Illumina HiSeq2500, with an average of 67 million reads per sample (Supplementary Table 4). To assess biological variation, a second RNA sample for five time points in both the leaf and apex were sequenced at a lower average coverage of 33 million reads per sample. Supplementary Table 1 summarises the sampling scheme and indicates the time points for which a second pool of samples was sequenced.

Gene model prediction and read alignment

Gene models are available for the Darmor-*bzh* reference genome sequence[32] but we leveraged our sequencing data to obtain improved predictions for splice junctions. The gene model prediction software AUGUSTUS[73] (version 3.2.2) was used to determine gene models for the Darmor-*bzh* reference genome. Tophat[74] (version 2.0.13) aligned RNA-Seq reads from across the entire time series were combined and filtered using the

`filterBam` tool provided with AUGUSTUS. AUGUSTUS used the filtered reads to aid the estimation of intron locations. Arabidopsis-derived parameters provided with the AUGUSTUS software were used to predict OSR gene models in the Darmor-*bzh* genome, with default parameters used otherwise.

RNA-Seq reads were aligned and expression levels quantified using the Tuxedo suite of software following the published workflow[75]. Tophat[74] (version 2.0.13) with the `b2-very-sensitive`, `transcriptome-only`, and `prefilter-multihits` parameters set was used to align reads to the Darmor-*bzh* reference sequence, using the AUGUSTUS derived gene models to determine the location of gene models. Cufflinks[76] (version 2.2.1) was used to quantify the expression levels of OSR genes. Data normalisation using `cuffnorm` was performed separately for leaf and apex tissue samples. Aside from the named parameters, default values were used.

Identification of sequence similarity between OSR and Arabidopsis gene models

The BLAST algorithm, using the `blastn` binary provided by NCBI[77] (version 2.2.30+) was used to identify sequence similarity between the AUGUSTUS[73] derived gene models and the published Arabidopsis gene models downloaded from TAIR (version 10). The `blastn` algorithm was run using default parameters, with an e-value threshold of 10^{-50} used to identify sequence similarity between the AUGUSTUS derived OSR gene models and published Arabidopsis. For the analysis conducted in this study, only the most highly scoring `blastn` hit was used to identify OSR copies of Arabidopsis genes.

Between genome expression comparison

Density plots of $\log_{10}$ transformed FPKM values were calculated and visualised using the

R statistical programming language[78]. The subsets of OSR genes used showed sequence similarity to at least one published Arabidopsis gene model downloaded from TAIR[79] (version 10), and sequence similarity to an Arabidopsis gene in the FLOR-ID database[40] (accessed 2016-08-19).

The expression fold change for homoeologue pairs was calculated using untransformed FPKM values. The geometric mean of the fold change across all *n* homoeologous gene pairs was calculated as $\sqrt[n]{\prod_{g=1}^{n} \frac{FPKM_{C,g}}{FPKM_{A,g}}}$ where $FPKM_{X,g}$ is the FPKM value of the *X* genome copy of the homologue pair *g*.

Homoeologue pair identification

The method outlined by Chalhoub et al. (2014) was used to identify pairs of homoeologues between the A and C genomes[4]. The Darmor-*bzh* reference genome was divided into the A and C genomes, removing the reference pseudo-chromosomes which consist of sequence that is unassigned to a specific chromosome. The separated genomes were uploaded to the CoGe portal[80] and the SynMap tool[81] was used to identify regions of syntenic genes between the two genomes. Chains of syntenic genes were identified using DAGchainer[82], allowing a maximum 20 gene distance between two matches and with a minimum number of 4 aligned pairs constituting a syntenic block. A 1:1 synteny screen was performed using the QUOTA-ALIGN[83] procedure. The synteny screen is necessary to distinguish homoeologous regions of the genome and paralogous regions which are the result of genome multiplication events which occurred prior to the interspecies hybridisation event in the evolutionary history of OSR. Once syntenic genes were identified using SynMap, a reciprocal sequence similarity filter was applied using the BLAST algorithm. The `blastn` algorithm was used with default parameters and a 10^{-50} e-value threshold to assess sequence similarity, and only homoeologue pairs which

were reciprocal best hits in this analysis were considered. This resulted in 14427 homoeologous pairs distributed across the entire OSR genome (Supplementary Figure 9).

Weighted gene co-expression network analysis

The weighted gene co-expression network analysis was carried out using the WGCNA library[84] (version 1.51) available for the R statistical programming language[78] (version 3.2.2). Due to the size of the dataset, WGCNA was performed on clustered data. The expression data was first filtered and normalised for each tissue separately. Any genes with a maximum FPKM value across the time series of less than 2.0 were removed. For the remaining genes, the expression across time was normalised to have a mean of 0.0 and a variance of 1.0. Using the normalised expression values, hierarchical clustering was conducted separately on the leaf and apex data using Euclidean distances between expression traces and a complete agglomeration method. The hierarchical tree was cut into H numbers of clusters and the ratio $\frac{\sum_{c=1}^{H} N_c(\bar{x}_c-\bar{x})^2}{\sum_{g=1}^{N}(x_g-\bar{x})^2}$ was calculated for each tree cut height, where N is the total number of genes, N_c is the total number of genes assigned to cluster c, x_g is the expression vector for gene g, $\bar{x}_c$ is the mean expression vector for genes assigned to cluster c, and $\bar{x}$ is the global mean of all expression vectors. The expression vectors are defined as $x_g = (\widehat{FPKM}_{g,22}, \widehat{FPKM}_{g,43}, \dots, \widehat{FPKM}_{g,72})$ where $\widehat{FPKM}_{g,t}$ represents the normalised FPKM level of gene g at time point t, with all time points included in the vector. A ratio of ~0.98 was chosen as a good balance between the number of clusters and how well the clusters represented the expression data. This ratio corresponded to 2683 clusters for leaf tissue and 6692 clusters for apex tissue.

WGCNA[84] was carried out using the mean expression vectors for the 6692 apex clusters and the 2683 leaf clusters. Based on the assumption of a scale-free network structure, a soft threshold of 30 was used for both the apex and leaf samples. A minimum regulatory

module size of 30 was used and modules with similar eigengene values were merged to give the final regulatory modules used for regulatory module assignment.

Self-organising maps and the identification of regulatory modules

Self-organising maps (SOM) were generated using the kohonen library[85] available for the R statistical programming language[78]. As with the WGCNA analysis, the data was filtered and normalised prior to carrying out the SOM analysis. The number of nodes used in the SOM was chosen based on the ratio $\frac{\sum_{c=1}^{S} N_c(x_c-\bar{x})^2}{\sum_{g=1}^{N}(x_g-\bar{x})^2}$ where N is the total number of genes, S is the total number of SOM nodes, N_c is the total number of genes assigned to SOM node c, x_g is the expression vector for gene g, x_c is the expression vector for SOM node c, and $\bar{x}$ is the global mean of all expression vectors. A value of S was chosen such that the above ratio was ~0.85 for both tissues. To adequately capture the variation present in the data, the dimensions of the SOM were set as the ratio between the first two principle component eigenvalues of the data, as has been done previously[86].

To assign probabilities of genes clustering to the same SOM cluster, a resampling procedure was employed (Supplementary Figure 8a). Expression values were sampled assuming a Gaussian noise model, using the expression value as the mean of the distribution and the expression value uncertainty calculated by Cufflinks as the distribution variance. The sampled expression values for each gene, within each tissue, were normalised to a mean expression of 0.0 with a variance of 1.0 across the time series and assigned to a SOM cluster based on a minimal Euclidean distance. This sampling loop was repeated 500 times, and the SOM clusters to which the genes of interest mapped were recorded. From this process, an empirical probability of mapping to each SOM cluster was calculated for each gene of interest. The probability of two genes mapping to

the same SOM cluster was then calculated as $\sum_{c=1}^{S} \frac{n_{g_1,c} n_{g_2,c}}{250000}$ where S is the total number of SOM clusters, and $n_{g_i,c}$ is the number of times gene g_i mapped to SOM cluster c. As the SOM training process begins from a random starting point, some SOMs were found to better discriminate between the expression traces of some pairs of genes than other SOMs. To overcome this, the probability of two genes of interest mapping to the same SOM cluster was calculated for 100 different SOMs. This probability was averaged to give the average probability of two genes of interest mapping to the same SOM cluster.

The probability of mapping to the same cluster can also be calculated for a single gene of interest by calculating $\sum_{c=1}^{S} \left(\frac{n_{g_1,c}}{500}\right)^2$. This value is a measure of how consistently a gene maps to the same SOM cluster, giving an indication of the uncertainty in the expression values calculated for that gene. Plotting a distribution of these self-clustering probabilities (Supplementary Figure 10) reveals a bimodal distribution with maxima at ~0.05 and ~1.0. To aid with visualising the average probabilities of two genes mapping to the same SOM cluster, as a consequence of this bimodality, a soft threshold based on a cumulative Gaussian density function was applied. The resulting value is referred to as a clustering coefficient in the main text. Clustering coefficients were calculated as $\frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{\mu_{p_{g_1,g_2}} - \theta}{\sigma_{p_{g_1,g_2}}\sqrt{2}}\right)\right]$ where erf is the error function defined as $\mathrm{erf}(x) = \frac{1}{\sqrt{\pi}} \int_{-x}^{x} e^{-t^2}\, \mathrm{d}t$, $\mu_{p_{g_1,g_2}}$ is the average probability of genes g_1 and g_2 mapping to the same cluster, $\sigma_{p_{g_1,g_2}}$ is the standard deviation of the probabilities calculated from the 100 different SOMs used in the sampling procedure, and θ is the tissue specific threshold. A threshold of 0.053 (apex) or 0.056 (leaf) was used. This threshold was calculated by taking the self-clustering probability that corresponded to the maximum of the density curve (Supplementary Figure 10) for each SOM and averaging them.

An automated approach was taken to quantify the pattern of clustering coefficients between copies of the same gene. Clustering coefficients were subjected to a binary filter, such that coefficients above 0.5 were set to 1 and those below set to 0. Regulatory modules were defined as groups of genes where the binary clustering coefficients between all genes were 1.

Sequence conservation analysis of orthologues of *AtTFL1* in OSR

Sequence upstream and downstream of the *AtTFL1* gene was extracted from the AtGDB TAIR9/10 v171 Arabidopsis genome assembly located on PlantGDB[87] and from the Darmor-*bzh* reference genome sequence[4]. Regions of conserved sequence were identified using mVISTA from the VISTA suite of tools[88,89]. The alignment algorithm used was AVID[90], which performed global pair-wise alignments for all sequences. Percentage sequence conservation was calculated using a 100bp sliding window.

Quantitative PCR of *BnaTFL1* homologues

Reverse transcription quantitative PCR (RT-qPCR) was carried out on copies of *TFL1* using custom designed primers (Supplementary Table 3). The SuperScript® III First-Strand Synthesis System (Thermo Fisher Scientific Inc., USA) was used to generate cDNA, with 2 µg of RNA used as input. The RNA was extracted as described above. Each RT-qPCR reaction consisted of 5 µl LightCycler® 480 SYBR Green I Master (Roche Molecular Systems Inc., USA), 4 µl cDNA, 0.125 µl of the forward and reverse primers at a concentration of 10 µM and 0.75 µl water. Quantification was performed on a LightCycler® 480 (Roche Molecular Systems Inc., USA). The RT-qPCR cycle consisted of a 95 °C denaturation step for 5 minutes followed by 50 quantification cycle. Each cycle consisted of 15 seconds at 95 °C, 20 seconds at 58 °C, 30 seconds at 72 °C. Fluorescence was quantified at 75 °C as the temperature was ramping from 72 °C to 95

°C.

Data availability

All sequencing reads collected as part of this study have been made available in the NCBI Sequence Read Archive under the BioProject number PRJNA398789.

References

1. Renny-Byfield, S. & Wendel, J. F. Doubling down on genomes: Polyploidy and crop plants. *Am. J. Bot.* **101,** 1711–1725 (2014).
2. International Wheat Genome Sequencing Consortium. A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. *Science* **345,** 1251788 (2014).
3. Li, F. *et al.* Genome sequence of cultivated Upland cotton (Gossypium hirsutum TM-1) provides insights into genome evolution. *Nat. Biotechnol.* **33,** 524–530 (2015).
4. Chalhoub, B. *et al.* Early allopolyploid evolution in the post-Neolithic Brassica napus oilseed genome. *Science* **345,** 950–953 (2014).
5. Conant, G. C. & Wolfe, K. H. Turning a hobby into a job: How duplicated genes find new functions. *Nat. Rev. Genet.* **9,** 938–950 (2008).
6. Otto, S. P. The Evolutionary Consequences of Polyploidy. *Cell* **131,** 452–462 (2007).
7. Lynch, M. & Conery, J. S. The Evolutionary Fate and Consequences of Duplicate Genes. *Science* **290,** 1151–1155 (2000).
8. Chaudhary, B. *et al.* Reciprocal Silencing, Transcriptional Bias and Functional Divergence of Homeologs in Polyploid Cotton (Gossypium). *Genetics* **182,** 503–

517 (2009).

9. Wapinski, I., Pfeffer, A., Friedman, N. & Regev, A. Natural history and evolutionary principles of gene duplication in fungi. *Nature* **449,** 54–61 (2007).
10. Pires, J. C. *et al.* Flowering time divergence and genomic rearrangements in resynthesized Brassica polyploids (Brassicaceae). *Biol. J. Linn. Soc.* **82,** 675–688 (2004).
11. Buggs, R. J. A. *et al.* Transcriptomic Shock Generates Evolutionary Novelty in a Newly Formed, Natural Allopolyploid Plant. *Curr. Biol.* **21,** 551–556 (2011).
12. Blanc, G. & Wolfe, K. H. Functional Divergence of Duplicated Genes Formed by Polyploidy during Arabidopsis Evolution. *Plant Cell* **16,** 1679–1691 (2004).
13. Hakes, L., Pinney, J. W., Lovell, S. C., Oliver, S. G. & Robertson, D. L. All duplicates are not equal: the difference between small-scale and genome duplication. *Genome Biol.* **8,** R209 (2007).
14. Veitia, R. A. Nonlinear effects in macromolecular assembly and dosage sensitivity. *J. Theor. Biol.* **220,** 19–25 (2003).
15. Kafri, R., Levy, M. & Pilpel, Y. The regulatory utilization of genetic redundancy through responsive backup circuits. *Proc. Natl. Acad. Sci.* **103,** 11653–11658 (2006).
16. Blomme, T. *et al.* The gain and loss of genes during 600 million years of vertebrate evolution. *Genome Biol.* **7,** R43 (2006).
17. Simillion, C., Vandepoele, K., Montagu, M. C. E. V., Zabeau, M. & Peer, Y. V. de. The hidden duplication past of Arabidopsis thaliana. *Proc. Natl. Acad. Sci.* **99,** 13627–13632 (2002).
18. Freeling, M. Bias in Plant Gene Content Following Different Sorts of Duplication: Tandem, Whole-Genome, Segmental, or by Transposition. *Annu. Rev. Plant Biol.*

60, 433–453 (2009).

19. Veitia, R. A. & Potier, M. C. Gene dosage imbalances: action, reaction, and models. *Trends Biochem. Sci.* **40,** 309–317 (2015).
20. Papp, B., Pál, C. & Hurst, L. D. Dosage sensitivity and the evolution of gene families in yeast. *Nature* **424,** 194–197 (2003).
21. Veitia, R. A. Gene Dosage Balance in Cellular Pathways: Implications for Dominance and Gene Duplicability. *Genetics* **168,** 569–574 (2004).
22. Veitia, R. A., Bottani, S. & Birchler, J. A. Cellular reactions to gene dosage imbalance: genomic, transcriptomic and proteomic effects. *Trends Genet.* **24,** 390–397 (2008).
23. Birchler, J. A. & Veitia, R. A. Gene balance hypothesis: Connecting issues of dosage sensitivity across biological disciplines. *Proc. Natl. Acad. Sci.* **109,** 14746–14753 (2012).
24. Maere, S. *et al.* Modeling gene and genome duplications in eukaryotes. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 5454–5459 (2005).
25. Bomblies, K. & Madlung, A. Polyploidy in the Arabidopsis genus. *Chromosome Res.* **22,** 117–134 (2014).
26. Bowers, J. E., Chapman, B. A., Rong, J. & Paterson, A. H. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* **422,** 433–438 (2003).
27. Makino, T. & McLysaght, A. Ohnologs in the human genome are dosage balanced and frequently associated with disease. *Proc. Natl. Acad. Sci.* (2010). doi:10.1073/pnas.0914697107
28. Freeling, M. *et al.* Many or most genes in Arabidopsis transposed after the origin of the order Brassicales. *Genome Res.* **18,** 1924–1937 (2008).

29. Birchler, J. A., Hiebert, J. C. & Paigen, K. Analysis of Autosomal Dosage Compensation Involving the Alcohol Dehydrogenase Locus in Drosophila Melanogaster. *Genetics* **124,** 677–686 (1990).
30. Nowak, M. A., Boerlijst, M. C., Cooke, J. & Smith, J. M. Evolution of genetic redundancy. *Nature* **388,** 167–171 (1997).
31. Kafri, R., Bar-Even, A. & Pilpel, Y. Transcription control reprogramming in genetic backup circuits. *Nat. Genet.* **37,** 295–299 (2005).
32. The Brassica rapa Genome Sequencing Project Consortium *et al.* The genome of the mesopolyploid crop species Brassica rapa. *Nat. Genet.* **43,** 1035–1039 (2011).
33. Beilstein, M. A., Nagalingum, N. S., Clements, M. D., Manchester, S. R. & Mathews, S. Dated molecular phylogenies indicate a Miocene origin for Arabidopsis thaliana. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 18724–18728 (2010).
34. Lysak, M. A., Koch, M. A., Pecinka, A. & Schubert, I. Chromosome triplication found across the tribe Brassiceae. *Genome Res.* **15,** 516–525 (2005).
35. Nelson, M. N. *et al.* Quantitative Trait Loci for Thermal Time to Flowering and Photoperiod Responsiveness Discovered in Summer Annual-Type Brassica napus L. *PLoS ONE* **9,** e102611 (2014).
36. Raman, H. *et al.* Genetic and physical mapping of flowering time loci in canola (Brassica napus L.). *Theor. Appl. Genet.* **126,** 119–132 (2013).
37. Guo, Y., Harloff, H.-J., Jung, C. & Molina, C. Mutations in single FT- and TFL1-paralogs of rapeseed (Brassica napus L.) and their impact on flowering time and yield components. *Plant Genet. Genomics* **5,** 282 (2014).
38. Irwin, J. A. *et al.* Nucleotide polymorphism affecting FLC expression underpins heading date variation in horticultural brassicas. *Plant J.* **87,** 597–605 (2016).
39. Andrés, F. & Coupland, G. The genetic basis of flowering responses to seasonal

cues. *Nat. Rev. Genet.* **13,** 627–639 (2012).

40. Bouché, F., Lobet, G., Tocquin, P. & Périlleux, C. FLOR-ID: an interactive database of flowering-time gene networks in Arabidopsis thaliana. *Nucleic Acids Res.* **44,** D1167–D1171 (2016).
41. Simpson, G. G. & Dean, C. Arabidopsis, the Rosetta Stone of Flowering Time? *Science* **296,** 285–289 (2002).
42. Ratcliffe, O. J. & Riechmann, J. L. Arabidopsis transcription factors and the regulation of flowering time: a genomic perspective. *Curr. Issues Mol. Biol.* **4,** 77–91 (2002).
43. Lou, P. *et al.* Preferential Retention of Circadian Clock Genes during Diploidization following Whole Genome Triplication in Brassica rapa. *Plant Cell* **24,** 2415–2426 (2012).
44. Tadege, M. *et al.* Control of flowering time by FLC orthologues in Brassica napus. *Plant J.* **28,** 545–553 (2001).
45. Robert, L. S., Robson, F., Sharpe, A., Lydiate, D. & Coupland, G. Conserved structure and function of the Arabidopsis flowering time gene CONSTANS in Brassica napus. *Plant Mol. Biol.* **37,** 763–772 (1998).
46. Parkin, I. A. *et al.* Transcriptome and methylome profiling reveals relics of genome dominance in the mesopolyploid Brassica oleracea. *Genome Biol.* **15,** R77 (2014).
47. Schnable, J. C., Springer, N. M. & Freeling, M. Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proc. Natl. Acad. Sci.* **108,** 4069–4074 (2011).
48. Yoo, M.-J., Szadkowski, E. & Wendel, J. F. Homoeolog expression bias and expression level dominance in allopolyploid cotton. *Heredity* **110,** 171–180 (2013).
49. Adams, K. L., Cronn, R., Percifield, R. & Wendel, J. F. Genes duplicated by

polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. *Proc. Natl. Acad. Sci.* **100,** 4649–4654 (2003).

50. He, Z. *et al.* Extensive homoeologous genome exchanges in allopolyploid crops revealed by mRNAseq-based visualization. *Plant Biotechnol. J.* **15,** 594–604 (2017).
51. Jaeger, K. E., Pullen, N., Lamzin, S., Morris, R. J. & Wigge, P. A. Interlocking Feedback Loops Govern the Dynamic Behavior of the Floral Transition in Arabidopsis. *Plant Cell Online* **25,** 820–833 (2013).
52. Blazquez, M. A., Soowal, L. N., Lee, I. & Weigel, D. LEAFY expression and flower initiation in Arabidopsis. *Development* **124,** 3835–3844 (1997).
53. Okamuro, J. K., den Boer, B. G. W., Lotys-Prass, C., Szeto, W. & Jofuku, K. D. Flowers into shoots: Photo and hormonal control of a meristem identity switch in Arabidopsis. *Proc. Natl. Acad. Sci.* **93,** 13831–13836 (1996).
54. Kersey, P. J. *et al.* Ensembl Genomes 2016: more genomes, more complexity. *Nucleic Acids Res.* **44,** D574–D580 (2016).
55. Ratcliffe, O. J., Bradley, D. J. & Coen, E. S. Separation of shoot and floral identity in Arabidopsis. *Development* **126,** 1109–1120 (1999).
56. Kobayashi, Y., Kaya, H., Goto, K., Iwabuchi, M. & Araki, T. A Pair of Related Genes with Antagonistic Roles in Mediating Flowering Signals. *Science* **286,** 1960–1962 (1999).
57. Kardailsky, I. *et al.* Activation Tagging of the Floral Inducer FT. *Science* **286,** 1962–1965 (1999).
58. Huala, E. & Sussex, I. M. LEAFY Interacts with Floral Homeotic Genes to Regulate Arabidopsis Floral Development. *Plant Cell Online* **4,** 901–913 (1992).
59. Schultz, E. A. & Haughn, G. W. LEAFY, a Homeotic Gene That Regulates

Inflorescence Development in Arabidopsis. *Plant Cell Online* **3,** 771–781 (1991).

60. Shannon, S. & Meeks-Wagner, D. R. A Mutation in the Arabidopsis TFL1 Gene Affects Inflorescence Meristem Development. *Plant Cell Online* **3,** 877–892 (1991).
61. Serrano-Mislata, A. *et al.* Separate elements of the TERMINAL FLOWER 1 cis-regulatory region integrate pathways to control flowering time and shoot meristem identity. *Development* **143,** 3315–3327 (2016).
62. Soltis, D. E. *et al.* Polyploidy and angiosperm diversification. *Am. J. Bot.* **96,** 336–348 (2009).
63. Hohmann, N., Wolf, E. M., Lysak, M. A. & Koch, M. A. A Time-Calibrated Road Map of Brassicaceae Species Radiation and Evolutionary History. *Plant Cell Online* **27,** 2770–2784 (2015).
64. Jiao, Y. *et al.* Ancestral polyploidy in seed plants and angiosperms. *Nature* **473,** 97–100 (2011).
65. Mandáková, T., Heenan, P. B. & Lysak, M. A. Island species radiation and karyotypic stasis in Pachycladonallopolyploids. *BMC Evol. Biol.* **10,** 367 (2010).
66. Leitch, I. J. & Bennett, M. D. Genome downsizing in polyploid plants. *Biol. J. Linn. Soc.* **82,** 651–663 (2004).
67. Lysak, M. A. *et al.* Mechanisms of chromosome number reduction in Arabidopsis thaliana and related Brassicaceae species. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 5224–5229 (2006).
68. Edger, P. P. & Pires, J. C. Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes. *Chromosome Res.* **17,** 699 (2009).
69. Paterson, A. H. Polyploidy, evolutionary opportunity, and crop adaptation. *Genetica* **123,** 191–196 (2005).
70. Higgins, J. A., Bailey, P. C. & Laurie, D. A. Comparative Genomics of Flowering

Time Pathways Using Brachypodium distachyon as a Model for the Temperate Grasses. *PLoS ONE* **5,** e10065 (2010).

71. Renny-Byfield, S. *et al.* Ancient Gene Duplicates in Gossypium (Cotton) Exhibit Near-Complete Expression Divergence. *Genome Biol. Evol.* **6,** 559–571 (2014).
72. Bayer, P. E. *et al.* Assembly and comparison of two closely related Brassica napus genomes. *Plant Biotechnol. J.* (2017). doi:10.1111/pbi.12742
73. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24,** 637–644 (2008).
74. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14,** R36 (2013).
75. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7,** 562–578 (2012).
76. Trapnell, C. *et al.* Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol.* **31,** 46–53 (2013).
77. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10,** 421 (2009).
78. R Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, 2017).
79. Berardini, T. Z. *et al.* The Arabidopsis Information Resource: Making and Mining the 'Gold Standard' Annotated Reference Plant Genome. *Genes. N. Y. N 2000* **53,** 474–485 (2015).
80. Lyons, E. & Freeling, M. How to usefully compare homologous plant genes and chromosomes as DNA sequences. *Plant J.* **53,** 661–673 (2008).
81. Lyons, E., Pedersen, B., Kane, J. & Freeling, M. The Value of Nonmodel Genomes

and an Example Using SynMap Within CoGe to Dissect the Hexaploidy that Predates the Rosids. *Trop. Plant Biol.* **1,** 181–190 (2008).

82. Haas, B. J., Delcher, A. L., Wortman, J. R. & Salzberg, S. L. DAGchainer: a tool for mining segmental genome duplications and synteny. *Bioinformatics* **20,** 3643–3646 (2004).
83. Tang, H. *et al.* Screening synteny blocks in pairwise genome comparisons through integer programming. *BMC Bioinformatics* **12,** 102 (2011).
84. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9,** 559 (2008).
85. Wehrens, R. & Buydens, L. M. C. Self- and Super-organizing Maps in R: The kohonen Package. *J. Stat. Softw.* **21,** 1–19 (2007).
86. Vesanto, J., Himberg, J., Alhoniemi, E. & Parhankangas, J. Self-Organizing Map in Matlab: the SOM Toolbox. in *In Proceedings of the Matlab DSP Conference* 35–40 (2000).
87. Duvick, J. *et al.* PlantGDB: a resource for comparative plant genomics. *Nucleic Acids Res.* **36,** D959–D965 (2008).
88. Frazer, K. A., Pachter, L., Poliakov, A., Rubin, E. M. & Dubchak, I. VISTA: computational tools for comparative genomics. *Nucleic Acids Res.* **32,** W273–W279 (2004).
89. Mayor, C. *et al.* VISTA : visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics* **16,** 1046–1047 (2000).
90. Bray, N., Dubchak, I. & Pachter, L. AVID: A Global Alignment Program. *Genome Res.* **13,** 97–102 (2003).

Acknowledgements

We thank Kirsten Bomblies for critical comments on the manuscript. DMJ acknowledges support from the John Innes Foundation for the JIC Rotation PhD Programme. RW, MT, and JAI are grateful for support from BBSRC's Institute Strategic Programme on Growth and Development (BB/J004588/1) and Genes in the Environment (BB/P013511/1). RJM acknowledges support from BBSRC's Institute Strategic Programme on Biotic Interaction underpinning Crop Productivity (BB/J004553/1) and Plant Health (BB/P012574/1).

Author Contributions

JAI and RJM conceived the project. DMJ, NP, RW, MT, JAI and RJM designed the experiments that were carried out by RW with support from NP, DMJ and JAI. The sequence analysis was carried out by DMJ with help from MT. DMJ performed all transcriptomic time-series analyses and produced all the figures. DMJ drafted the manuscript which was planned by DMJ, NP, JAI and RJM. All authors contributed to writing the manuscript.

Competing Financial Interests

The authors declare that they have no competing interests.

Figure Legends

Figure 1 – Arabidopsis flowering time genes have been maintained in the OSR genome at a higher copy number relative to other Arabidopsis genes.

Annotated OSR genes were assigned to an Arabidopsis gene by taking the highest scoring

BLAST result. The proportions were calculated by counting the number of Arabidopsis genes with a particular number of identified OSR copies and dividing by the total number of Arabidopsis genes represented by at least one gene in OSR. The FLOR-ID distribution is calculated using a subset of 315 Arabidopsis genes annotated as being involved with flower development or flowering time control in the FLOR-ID database[40]. False discovery rate corrected *p*-values were calculated by taking 1000 samples of 315 Arabidopsis genes from the 20882 represented in the All distribution. The mean and standard deviation of these samples were used to perform a two-tailed test of observing a proportion as extreme as the FLOR-ID value.

Figure 2 – Tissue samples were collected for RNA-Seq at selected points through development

Plants were grown as detailed in the Methods. Tissue was sampled on the days indicated with red dotted lines and numbers. The plant silhouettes represent the approximate number of full leaves at the indicated points in development.

Figure 3 – The A and C genomes of OSR show different patterns of gene expression.

Density plots of transformed expression levels ($\log_{10}$(FPKM)) calculated using different gene subsets. The expression data was sampled 1000 times using a Gaussian error model. The density plot of $\log_{10}$(FPKM) values was calculated for each sample. The mean density and the 95% confidence interval estimated using the 1000 samples is displayed. Tabulated below each density plot are the number of OSR genes used to calculate the density plot, separated by their genome of origin. The data used to generate the density plots consisted of expression data from: **a** all annotated OSR genes, **b** OSR genes that

show sequence conservation to an annotated Arabidopsis gene, and **c** OSR genes that show sequence conservation to an annotated Arabidopsis gene that is present in the FLOR-ID database[40]. These plots are generated using apex expression data from the time point taken at day 22, but are representative of the density plots obtained for all time points across both tissue types sampled (Supplementary Figure 1).

Figure 4 – Multiple OSR flowering time gene homologues are expressed during the floral transition.

The proportions of Arabidopsis genes that have particular numbers of homologues identified and expressed in OSR. OSR genes were considered to be expressed if their maximal expression level within a tissue across the time series was above 2.0 FPKM. False discovery corrected *p*-values are computed in the same way as Figure 1 using subsets of genes. **a** OSR genes that show sequence conservation to an annotated Arabidopsis gene. **b** OSR genes expressed in the apex tissue that show sequence conservation to an annotated Arabidopsis gene. **c** OSR genes expressed in the leaf tissue that show sequence conservation to an annotated Arabidopsis gene.

Figure 5 – Not all annotated OSR orthologues of Arabidopsis genes are expressed.

Expression data from the apex, **a**, and leaf, **b**, show that not all OSR copies of Arabidopsis genes were expressed in the developmental transcriptome time series. The size and colour of the circles indicate the number of data points at that position in the graph. The thick diagonal line indicates Arabidopsis genes that have OSR orthologues that are all expressed during the developmental transcriptome. Only OSR genes that show sequence conservation to an annotated Arabidopsis genes present in the FLOR-ID database[40] were

used to generate these results. A similar graph generated using all OSR genes that show sequence conservation to an annotated Arabidopsis gene is shown in Supplementary Figure 5.

Figure 6 – The majority of flowering time gene homologues in OSR are assigned to different regulatory modules.

Regulatory module assignments for the apex, **a**, and leaf, **b**. The size and colour of the circles indicates the number of data points at that position in the graph. The thick lines on each graph represent two potential extremes. The dashed line represents the null hypothesis that all OSR copies of an Arabidopsis gene are assigned to the same WGCNA cluster. The solid line represents the Arabidopsis genes that have OSR copies that are each assigned to separate WGCNA clusters. The percentages indicated on the graph indicate the percentage of data points that agree, and the percentage that do not agree, with the null hypothesis. Only OSR genes with expression above 2.0 FPKM in at least one time point in the developmental time series and sequence conservation to an annotated Arabidopsis gene were used. A similar graph generated using all OSR genes that show sequence conservation to an annotated Arabidopsis gene is shown in Supplementary Figure 7.

Figure 7 - The OSR orthologues of *AtTFL1*, *AtFT*, and *AtLFY* show different patterns of regulation.

a Representations of the five patterns of regulatory module assignment detected by the SOM based method. High clustering coefficients between two different genes indicates that those genes have similar expression traces. Clustering coefficients between a gene

and itself represent how robustly a gene maps to the SOM. A *distinct* pattern indicates multiple regulatory modules being identified, with no gene occupying more than one module. A *gradated* pattern represents multiple regulatory modules being detected, but genes occupy multiple modules. *Redundant* patterns occur when only one regulatory module is detected, and all copies of a gene are assigned to that module. *Unique* patterns are a special case of *distinct* pattern where each copy of a gene is assigned to a different regulatory module. *Mixed* patterns consist of a mixture of *distinct* and *gradated* patterns, where the gene assignment of some modules overlap while others do not show overlap. When assessing the regulatory module assignment, gene copies that do not robustly map to the SOM are removed. **b, c** and **d** Expression traces across the developmental time series were normalised to a mean value of 0.0 FPKM and unit variance across the time series. The shading indicates time points during which the plants were grown in cold conditions. Regulatory module assignment heatmaps calculated using the SOM based method for the OSR copies of *TFL1*, *FT*, and *LFY* are also displayed. Both the expression traces and the clustering coefficients are apex derived for *TFL1* (**b**) and *LFY* (**d**) and leaf derived for *FT* (**c**).

Figure 8 - Sequence analysis reveals that cis-regulatory modules identified in Arabidopsis are not present downstream of some copies of *TFL1* in OSR.

a The degree of sequence conservation between the OSR copies of *TFL1* and *AtTFL1*. Sequence alignment and conservation calculations were performed using the mVISTA server[88,89] with a sliding window size of 100bp. The seven regions of high interspecies sequence conservation (green bars) and the five cis-regulatory regions (blue boxes) identified[61] by Serrano-Mislata et al. are shown relative to the *AtTFL1* gene model (black bars). The labelling of these regions follows the same conventions as the previous study.

The pink shaded areas under the sequence conservation curves are regions above 70% sequence conservation. Genomic position upstream and downstream of the *TFL1* gene copies are given relative to the ATG and STOP codon sites respectively. **b** The unnormalised expression traces for the *BnaTFL1* genes determined through RNA-Seq and RT-qPCR. The expression values calculated for RT-qPCR are normalised to *GAPDH* with the error determined from two biological replicates (Methods).

Tables

Days post sowing	Apex			Leaf		
	Both expressed	A genome 2-fold higher	C genome 2-fold higher	Both expressed	A genome 2-fold higher	C genome 2-fold higher
22	7313	596 (8.1%)	1113 (15.2%)	6294	620 (9.9%)	1066 (16.9%)
43	7389	597 (8.1%)	1132 (15.3%)	6176	626 (10.1%)	1133 (18.3%)
64	7325	602 (8.2%)	1085 (14.8%)	6307	597 (9.5%)	1021 (16.2%)
65	7243	609 (8.4%)	1120 (15.5%)	6182	601 (9.7%)	993 (16.1%)
67	7299	601 (8.2%)	1135 (15.6%)	6257	603 (9.6%)	1046 (16.7%)
69	7342	594 (8.1%)	1130 (15.4%)	-	-	-
72	7449	612 (8.2%)	1119 (15.0%)	6237	601 (9.6%)	1054 (16.9%)

Table 1 – Number of genes expressed 2-fold higher than their homoeologue for all homoeologue pairs.

Homoeologue pairs[4] were determined and filtered at each time point for those which both had expression levels above 2 FPKM. The number and percentage of these genes expressed 2-fold higher than their homoeologue is given. Despite some pronounced differences at the gene level, at the genome level the overall expression change is modest: The geometric mean of the fold difference of the C genome gene relative to the A genome homoeologue for all homoeologue pairs is 1.12 in the apex and 1.11 in the leaf.

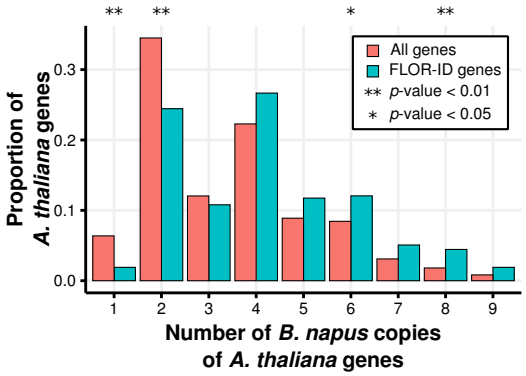

**
**
*
**
All genes
FLOR-ID genes
** p-value < 0.01
* p-value < 0.05
Proportion of
A. thaliana genes
0.0
0.1
0.2
0.3
1
2
3
4
5
6
7
8
9
Number of B. napus copies
of A. thaliana genes

Figure 1

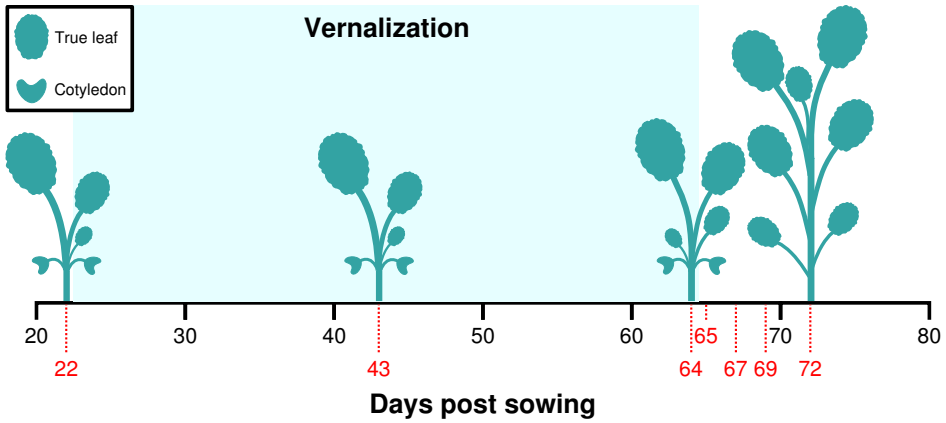

True leaf
Cotyledon
Vernalization
20
30
40
50
60
65
70
80
22
43
64
67
69
72
Days post sowing

Figure 2

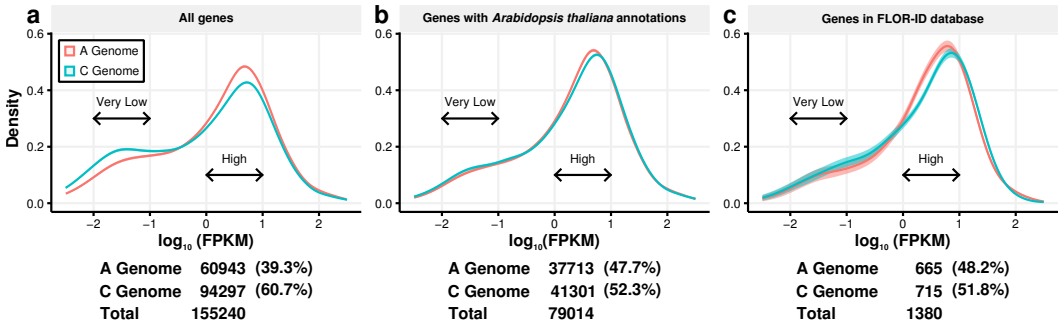

a
All genes
A Genome
C Genome
Very Low
High
Density
log10 (FPKM)
A Genome 60943 (39.3%)
C Genome 94297 (60.7%)
Total 155240
b
Genes with Arabidopsis thaliana annotations
Very Low
High
log10(FPKM)
A Genome 37713 (47.7%)
C Genome 41301 (52.3%)
Total 79014
c
Genes in FLOR-ID database
Very Low
High
log10 (FPKM)
A Genome 665 (48.2%)
C Genome 715 (51.8%)
Total 1380

Figure 3

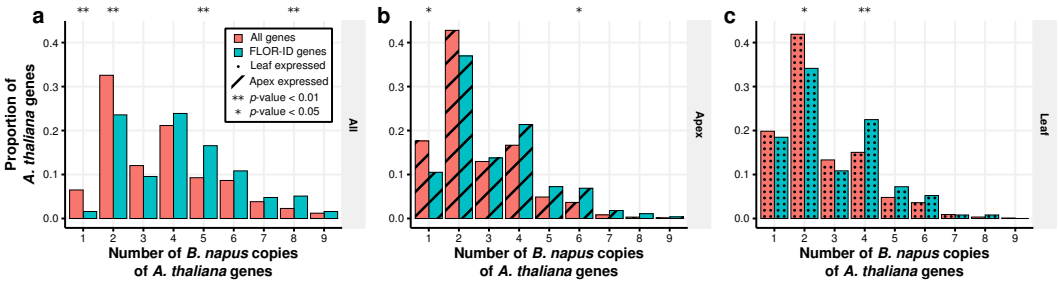

a
b
c
**
**
**
**
*
*
*
**
All genes
FLOR-ID genes
Leaf expressed
Apex expressed
** p-value < 0.01
* p-value < 0.05
Proportion of A. thaliana genes
0.0
0.1
0.2
0.3
0.4
1
2
3
4
5
6
7
8
9
Number of B. napus copies of A. thaliana genes
All
Apex
Leaf

Figure 4

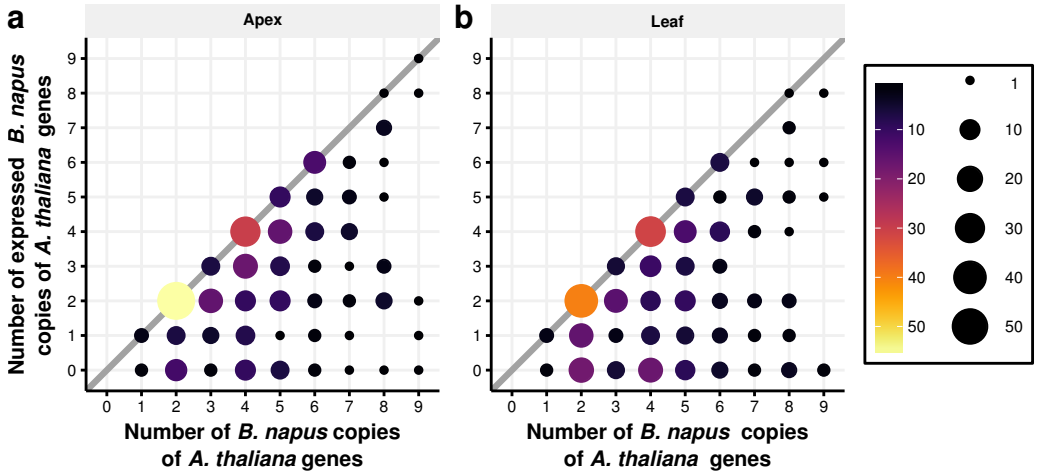

a
Apex
b
Leaf
Number of expressed B. napus copies of A. thaliana genes
Number of B. napus copies of A. thaliana genes
Number of B. napus copies of A. thaliana genes
0
1
2
3
4
5
6
7
8
9
10
20
30
40
50

Figure 5

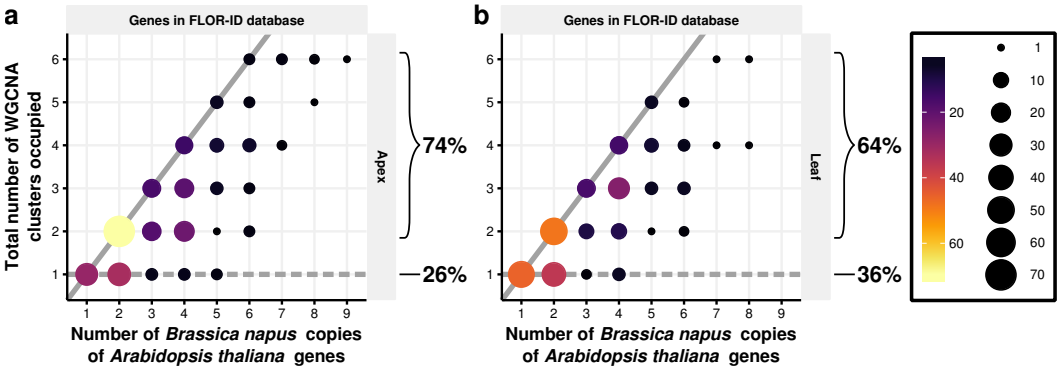

a
Genes in FLOR-ID database
Total number of WGCNA clusters occupied
Apex
74%
26%
Number of Brassica napus copies of Arabidopsis thaliana genes
b
Genes in FLOR-ID database
Leaf
64%
36%
Number of Brassica napus copies of Arabidopsis thaliana genes

Figure 6

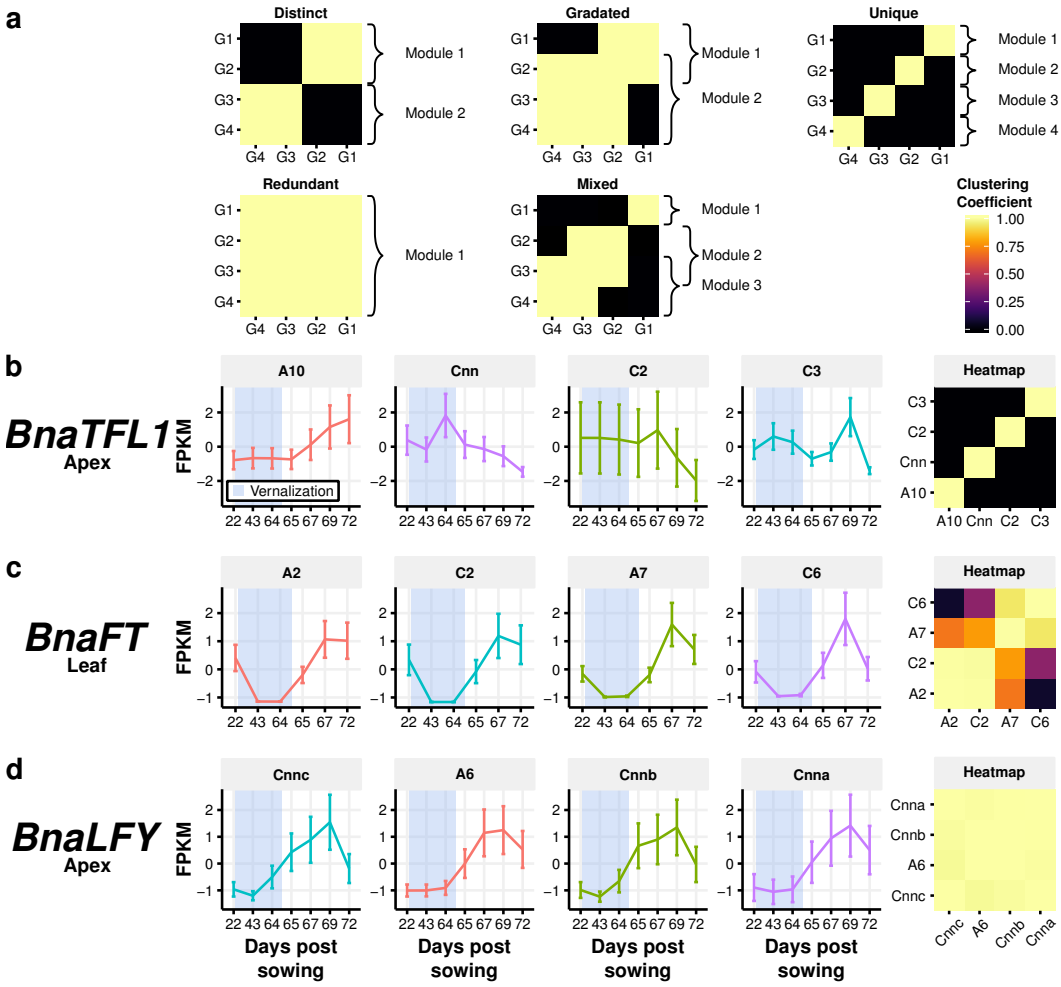

a
Distinct
Module 1
Module 2
Gradated
Module 1
Module 2
Unique
Module 1
Module 2
Module 3
Module 4
Redundant
Module 1
Mixed
Module 1
Module 2
Module 3
Clustering Coefficient
1.00
0.75
0.50
0.25
0.00
b
BnaTFL1
Apex
A10
Cnn
C2
C3
Heatmap
FPKM
Vernalization
c
BnaFT
Leaf
A2
C2
A7
C6
Heatmap
d
BnaLFY
Apex
Cnnc
A6
Cnnb
Cnna
Heatmap
Days post sowing

Figure 7

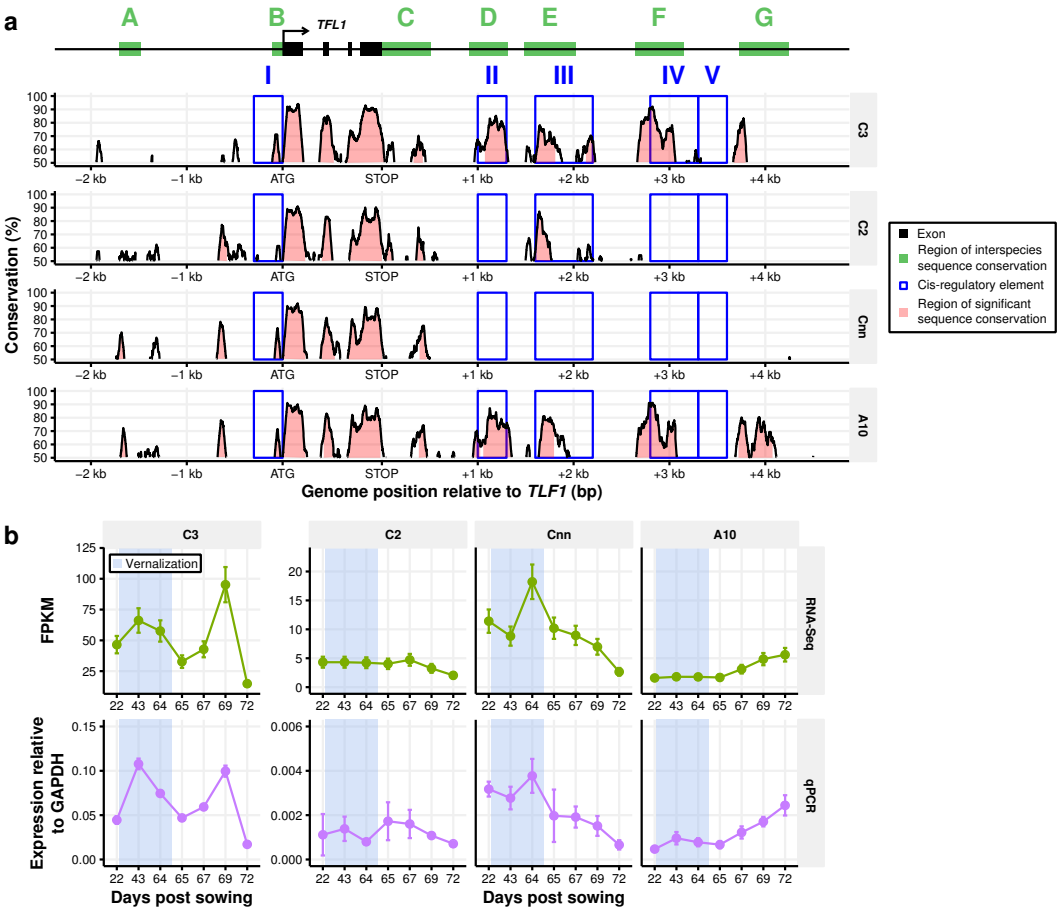

a
A
B
TFL1
C
D
E
F
G
I
II
III
IV
V
C3
C2
Cnn
A10
Conservation (%)
−2 kb
−1 kb
ATG
STOP
+1 kb
+2 kb
+3 kb
+4 kb
Genome position relative to TLF1 (bp)
Exon
Region of interspecies sequence conservation
Cis-regulatory element
Region of significant sequence conservation
b
Vernalization
FPKM
RNA-Seq
Expression relative to GAPDH
qPCR
Days post sowing

Figure 8

Supplementary materials for

Regulatory divergence of flowering time genes in the allopolyploid *Brassica napus*

D. Marc Jones[1,2], Rachel Wells[1], Nick Pullen[1], Martin Trick[2], Judith A. Irwin[1*], Richard J. Morris[1,2†]

[1] *Crop Genetics, John Innes Centre, Norwich Research Park, Colney Lane, Norwich. NR4 7UH. United Kingdom.*

[2] *Computational and Systems Biology, John Innes Centre, Norwich Research Park, Colney Lane, Norwich. NR4 7UH. United Kingdom.*

[*]Corresponding author: judith.irwin@jic.ac.uk

[†]Corresponding author: richard.morris@jic.ac.uk

Supplementary results

A self-organising map based approach corroborates the finding that *Brassica napus* copies of *Arabidopsis thaliana* flowering time genes have diverged in their regulation.

A self-organising map (SOM) based approach was employed to detect regulatory divergence between *B. napus* copies of flowering time genes. The advantage of this approach, over the WGCNA approach discussed in the main text, is that the regulatory module assignments are not binary, allowing for more subtle patterns to be detected. A SOM is a construct that groups together expression traces into clusters. The sampling procedure (Supplementary Figure 8a) returns an empirical probability of two expression traces mapping to the same SOM cluster. In addition, clustering probabilities can be calculated for a single gene which represent the uncertainty in the expression measurements quantified for that gene. In this case the clustering probability calculated is referred to as a self-clustering probability. Visualising the clustering probabilities determined by the SOM based method is complicated by the bimodal distribution the probabilities follow. Supplementary Figure 10 reveals a peak in self-clustering probabilities at 0.05 but also at ~1.0. This bimodal structure is a result of some genes only being expressed at a single time point. When these genes are resampled, their normalised expression trace remains the same, leading to a high self-clustering probability. To visualise probabilities from across this distribution, a soft threshold is applied to the probabilities. After the threshold is applied, the higher the clustering coefficient, the more similar two expression traces will tend to be. Genes are assigned to regulatory modules using heatmaps of clustering coefficients. The different patterns of regulatory module assignment are described in the main text.

This method was applied to *B. napus* flowering time genes. The occurrences of the different regulatory module assignment patterns were counted for both apex (Supplementary Figure 8b) and leaf (Supplementary Figure 8c) expression data. The null hypothesis used in the WGCNA analysis was that copies of genes would not show expression divergence (dashed lines in Figure 6, main text). The *redundant* pattern in the SOM analysis is equivalent to this null hypothesis (Figure 7a, main text). Like the results from the WGCNA analysis, this null hypothesis is not true for any flowering time genes with five or more copies in the *B. napus* leaf (Supplementary Figure 8c) or six or more copies in the apex (Supplementary Figure 8b). As with the *redundant* pattern, the *unique* pattern of regulatory module assignment becomes

less frequent as the number of *B. napus* copies of a gene increases (Supplementary Figure 8b and 8c). This agrees with the WGNCA analysis, where the number of genes lying on the solid line in Figure 6 in the main text (equivalent to the *unique* pattern in the SOM analysis) decreases at higher numbers of copies.

WGCNA cannot detect *gradated* and *mixed* patterns patterns of regulatory module assignment . In the apex and leaf, *mixed* and *gradated* patterns are seen at a lower frequency than *distinct* patterns, revealing that genes exhibiting intermediary regulatory behaviour relative to the other copies of that gene are observed less frequently than genes occupying distinct regulatory modules. Gene copies with intermediate regulatory behaviour may indicate that somecopies are more susceptible to regulatory cross-talk than others. The low number of *gradated* patterns observed when three genes copies are present in both tissues suggests that these genes tend to have expression traces that are detectably different to one another. *Distinct* patterns are more prevalent than *unique* patterns at three gene copies; the majority contain one copy with an expression trace divergent to the expression traces of the other two copies.

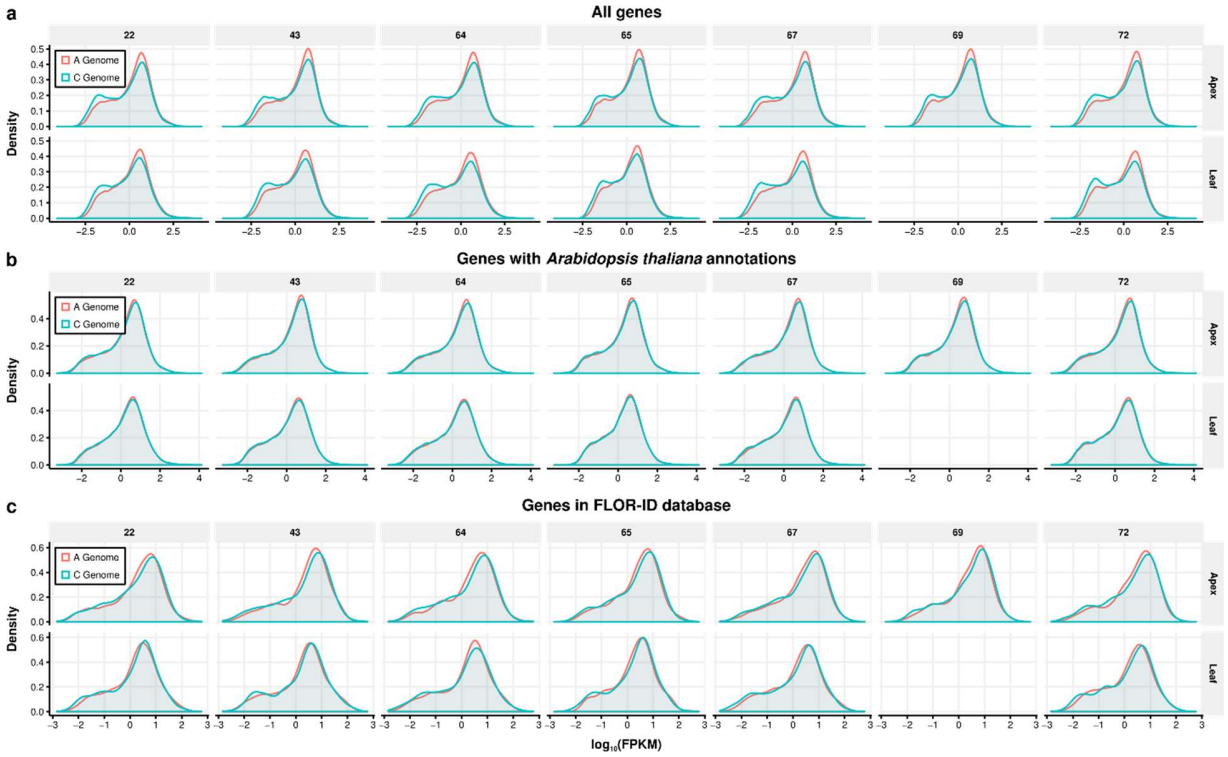
We could integrate homoeologue information for the three copy genes exhibiting a *distinct* pattern of regulatory module assignment to ask whether genes tended to be within the same regulatory modules as their homoeologue. In the apex, this is the case, with 59% of genes located in the same module. More generally, we find that of the genes in the apex (leaf) where homoeologue information is available, 69% (64%) of genes are assigned to the same module as the homoeologue, 18% (19%) of genes are assigned to a different module and 12% (16%) of genes have homoeologues which cannot be clustered. Homoeologues that cannot be clustered arise when the clustering coefficient calculated using the self-clustering probability of a gene is below 0.5, or the homoeologue is not expressed in that tissue.

We then asked whether the relatively large number of *distinct* patterns at four gene copies was due to homoeologous copies of genes displaying similar expression traces. For the genes for which homoeologue information was available, we find the majority (76% in apex, 72% in leaf) of genes are in the same regulatory module as their homoeologue.

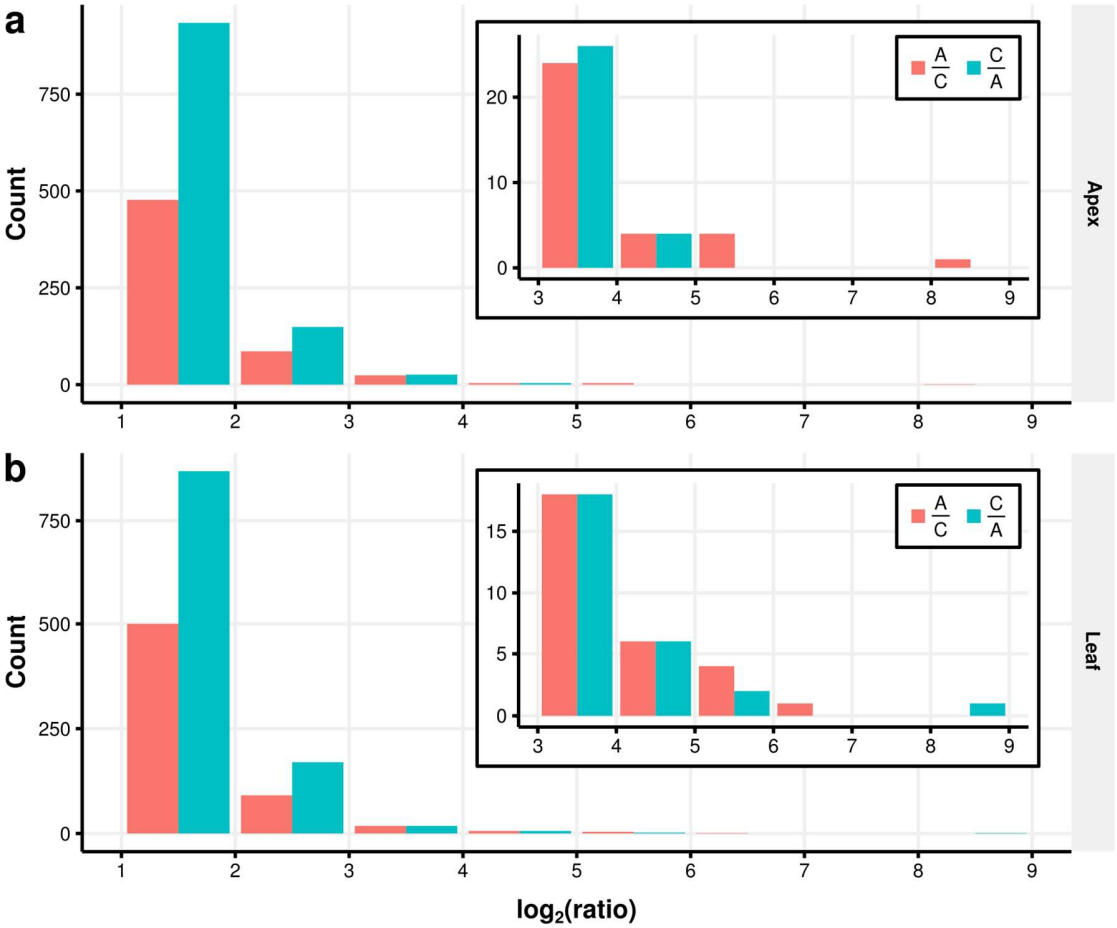
The SOM analysis corroborates many of the key findings of the WGCNA analysis in a manner which takes into account the uncertainty in our data. Namely, that expression divergence between copies is widespread and that as the number of copies of a gene in the

genome increases, the likelihood of observing regulatory divergence between those copies increases. Additionally, the SOM analysis reveals that some copies of flowering time genes exhibit a *gradated* pattern of regulatory module assignment, representing subtle differences in regulation. This may be the result of regulatory cross-talk between the copies, or represents subtle functional differences that have consequences for the control of flowering time in *Brassica napus*.

Supplementary figures

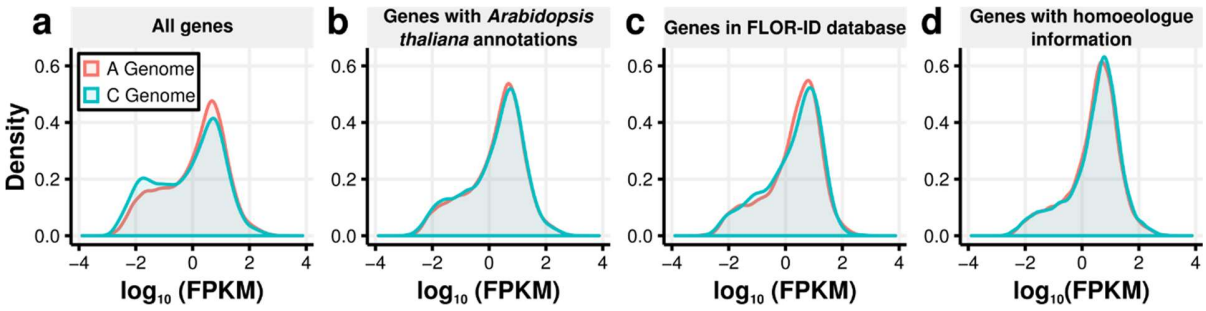


Supplementary figure 1 – Expression differences between A and C genomes are consistent across different tissues and time points.
Density plots of transformed expression levels (log_{10}(FPKM)) calculated using different subsets of genes. The data used to generate the density plots consisted of expression data from: **a** all annotated *Brassica napus* genes, **b** *B. napus* genes that show sequence conservation to an annotated *Arabidopsis thaliana* gene, and **c** *B. napus* genes that show sequence conservation to an annotated *A. thaliana* gene that is present in the FLOR-ID database[1].



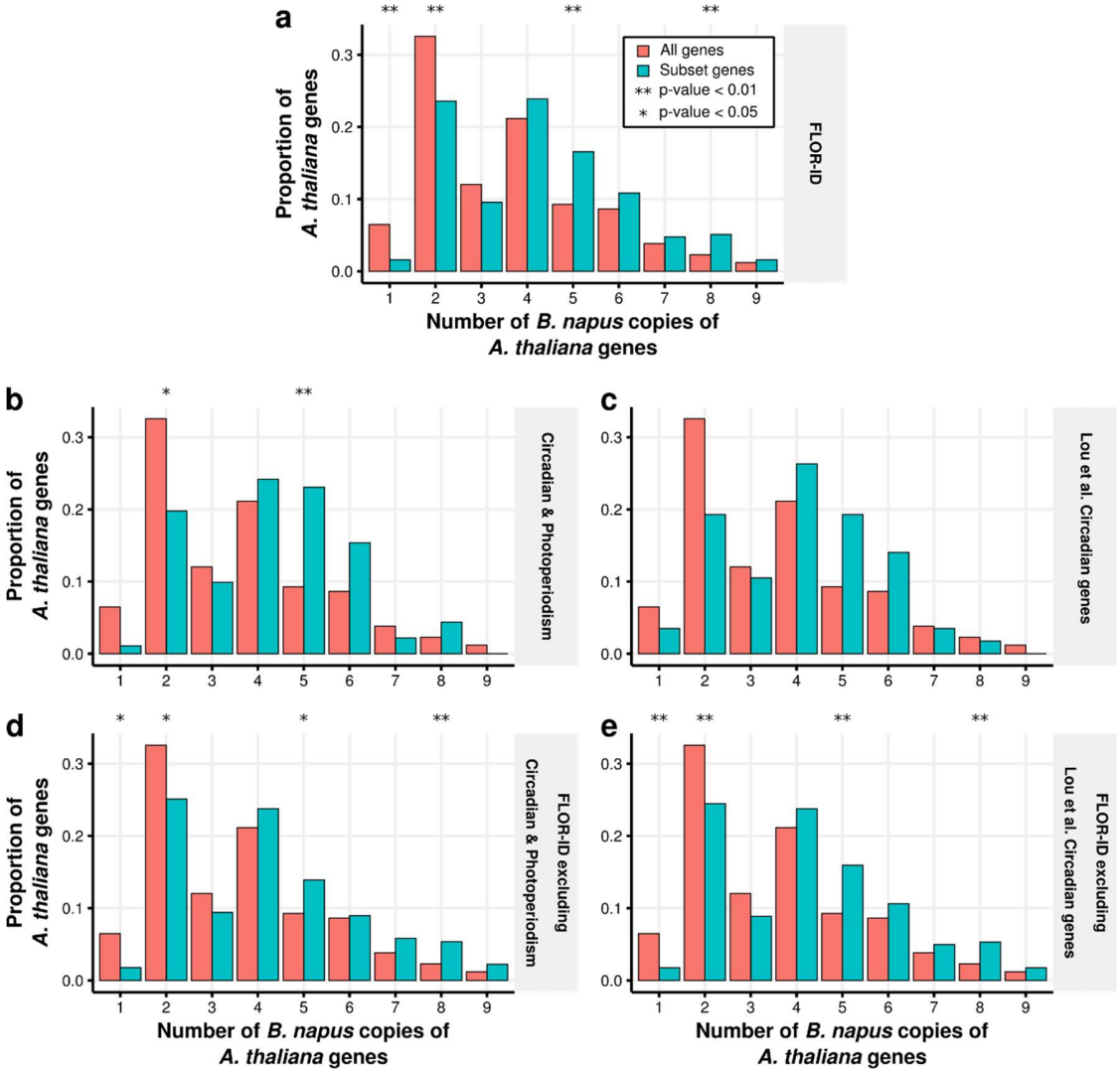
Supplementary figure 2 – Distributions of the fold expression differences between homoeologue pairs exhibiting biased expression

Homoeologue pairs are defined as exhibiting biased expression towards a particular genome if the gene on that genome has an FPKM level at least 2-fold higher than its homoeologue. The fold differences in FPKM level between homoeologues were calculated and $\log_2$ transformed. The values were binned and the number of pairs in each bin are plotted. If the homoeologue pairs exhibit biased expression towards the A genome, then the fold ratio was calculated with the A genome homoeologue FPKM value as the numerator (red bars). Likewise, if the pairs exhibit biased expression towards the C genome then the fold ratio was calculated with the C genome homoeologue FPKM value as the numerator (blue bars). The FPKM values from the day 22 time point were used. The inset of each graph corresponds to the counts above a $\log_2$(ratio) value of 3 plotted on a different count scale.



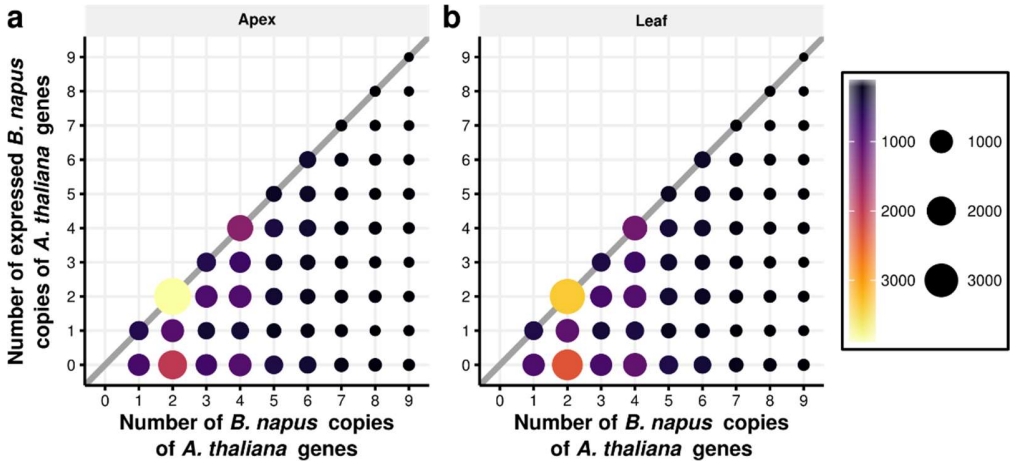
Supplementary figure 3 – Genes for which homoeologue information is available have fewer genes within the very low region of expression

Density plots of transformed expression levels (log_{10}(FPKM)) calculated using different subsets of genes. The data used to generate the density plots consisted of expression data from: **a** all annotated *B. napus* genes, **b** *B. napus* genes that show sequence conservation to an annotated *A. thaliana* gene, **c** *B. napus* genes that show sequence conservation to an annotated *A. thaliana* gene that is present in the FLOR-ID database[1], and **d** *B. napus* genes for which homoeologue information is available. These plots are generated using apex expression data from the time point taken at day 22.



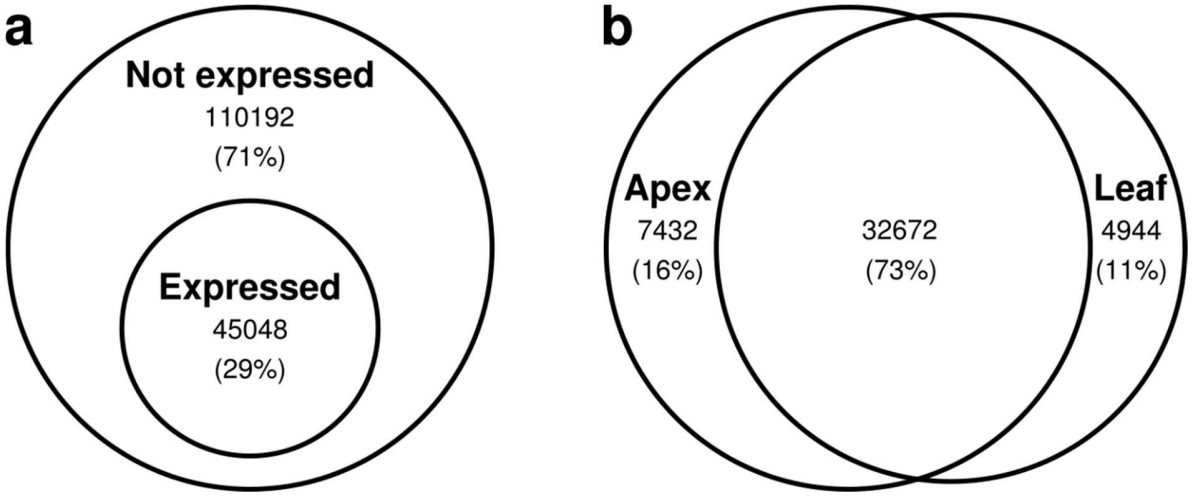
Supplementary figure 4 – The observed retention of flowering time genes is not explained by genes associated with the circadian rhythm alone

The proportions of Arabidopsis genes that have particular numbers of homologues identified in OSR, comparing all genes to a number of different gene subsets. False discovery corrected *p-values* are computed in the same way as Figure 1 in the main text. The gene subsets compared to all genes in each of the plots are as follows: **a** All FLOR-ID genes[1]. **b** FLOR-ID genes annotated as involved with the "Circadian" or "Photoperiodism" pathways. **c** The list of circadian genes used by Lou et al. (2012) to demonstrate gene retention in *B. rapa*[2]. **d** FLOR-ID genes with genes annotated as involved with the "Circadian" or "Photoperiodism" pathways removed. **e** FLOR-ID genes with genes used in the study by Lou et al. (2012) removed[2].

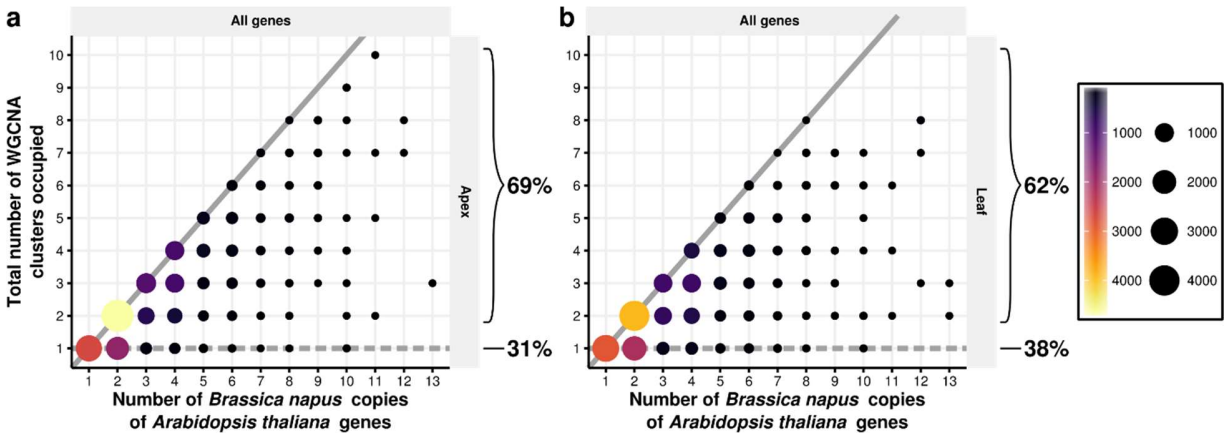


Supplementary figure 5 – Not all annotated *B. napus* copies of *A. thaliana* genes are expressed.

a and **b** depict the relationships when expression data from the apex and leaf are used respectively. The size and colour of the circles indicates the number of data points at that position in the graph. The thick diagonal line indicates *A. thaliana* genes that have *B. napus* orthologues that are all expressed during the developmental transcriptome. All *B. napus* genes that show sequence conservation to an annotated *A. thaliana* gene were used to generate these results.

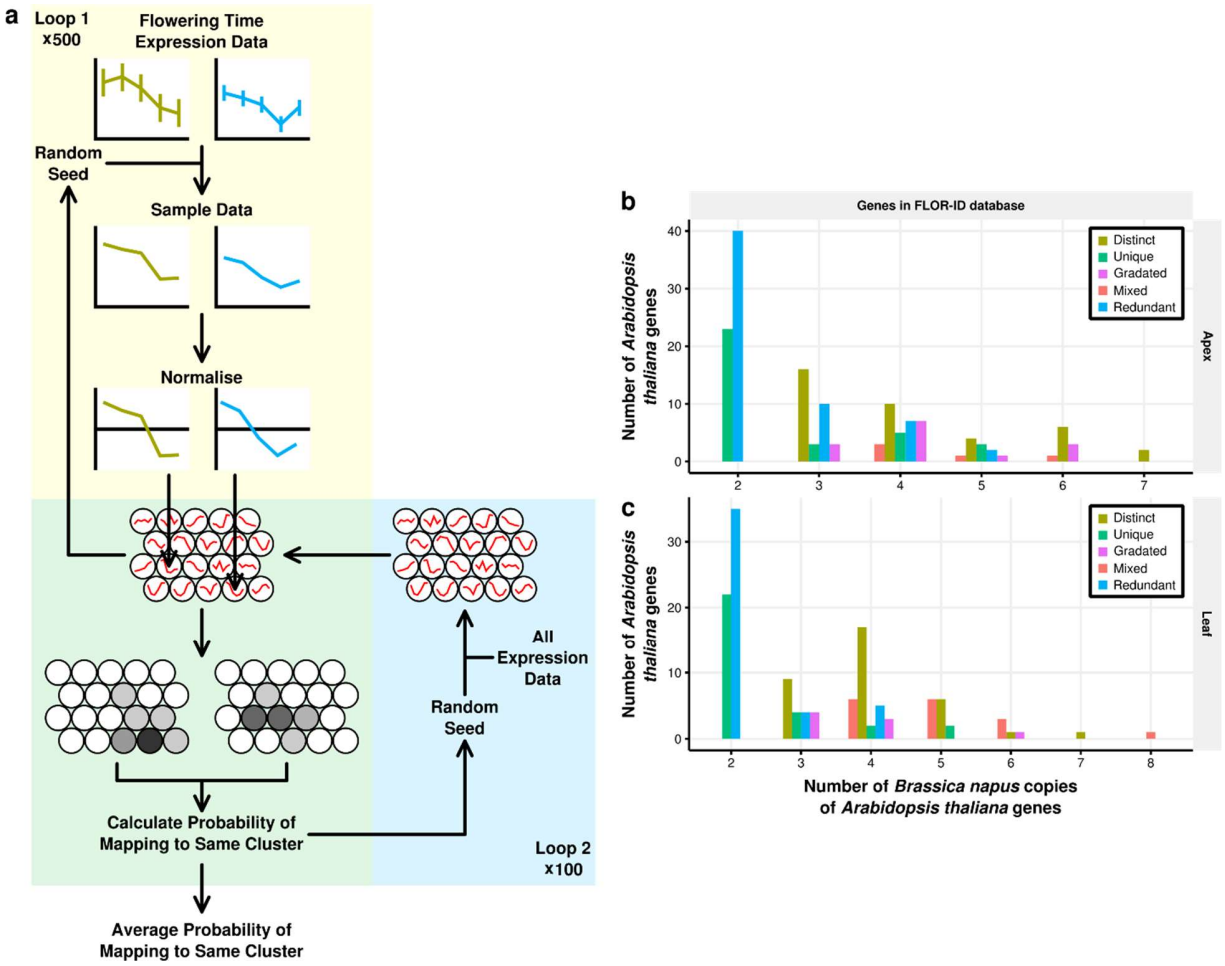


Supplementary figure 6 – Euler and Venn diagrams showing the percentage of expressed genes and the percentage of genes expressed in the apex and leaf samples
Brassica napus genes were classified as expressed if the expression of the genes exceeded 2.0 FPKM at at least one time point during the developmental time series. **a** Genes expressed in at least one tissue of the *Brassica napus* genes compared to the number of annotated genes in the Darmor-*bzh* reference genome. **b** The number of genes expressed specifically in the apex and the leaf and the number of genes that are expressed in both tissues.



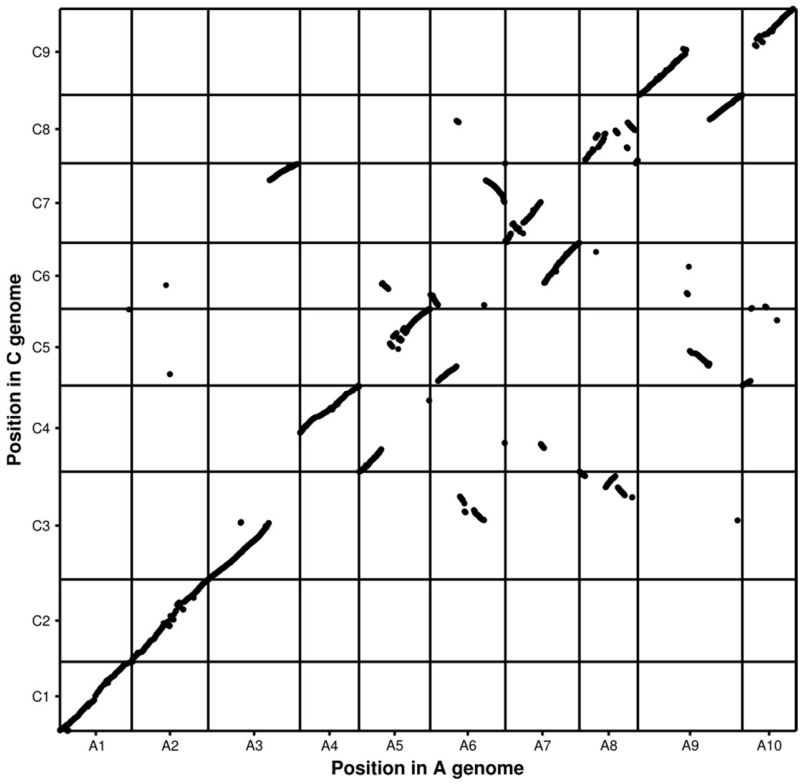
Supplementary figure 7 – Many gene copies are assigned to different regulatory modules in *B. napus.*

B. napus genes were included in this analysis when they i) Have expression above 2.0 FPKM in at least one time point in the developmental time series, and ii) Show sequence conservation to an annotated *A. thaliana* gene. **a** and **b** depict the relationships when expression data from the apex and leaf are used respectively. The size and colour of the circles indicates the number of data points at that position in the graph. The thick lines on each graph represent two potential extremes. The dashed line represents the null hypothesis that all *B. napus* copies of an *A. thaliana* gene are assigned to the same WGCNA cluster. The solid line represents the *A. thaliana* genes that have *B. napus* copies that are each assigned to separate WGCNA clusters. The percentages indicated on the graph indicate the percentage of data points which agree and the percentage which do not agree with the null hypothesis. All *B. napus* genes showing sequence conservation to an annotated *A. thaliana* gene were used to generate these results.



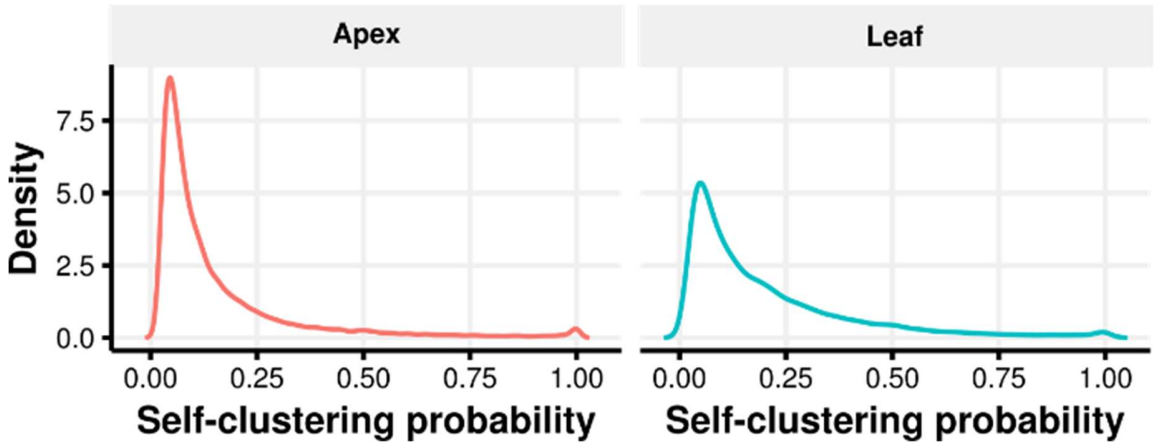
Supplementary figure 8 – Self-organising map (SOM) based assessment of expression trace divergence uncovers widespread regulatory differences and subtle patterns of divergence.

a Schematic of the SOM based clustering approach. The approach consists of two overlapping sampling loops. In loop 1, expression data from flowering time gene copies is sampled assuming a Gaussian error model. Sampled expression traces are zero mean and unit variance normalised and mapped to the SOM. This procedure is repeated 500 times to give two density plots of where in the SOM the copies map. These density plots are used to calculate the probability of the copies mapping to the same SOM cluster. As SOM clustering has a random component, loop 2 consists of regenerating the SOM using all expression data and calculating the probability of copies clustering to the same cluster 100 times. Using this, an average probability of mapping to the same cluster is calculated. **b & c** The relationships between the number of expressed *B. napus* copies of *A. thaliana* genes and the number of different types of regulatory module assignment patterns exhibited by those gene copies. This relationship is calculated using expression data from the apex (**b**) and the leaf (**c**). The different regulatory patterns are illustrated and explained in Figure 7 of the main text.



Supplementary figure 9 – Locations of identified homoeologues pairs in the *B. napus* genome

Homoeologue pairs were identified as detailed in the main text (Methods). The locations of these pairs give a representation of the chromosomal rearrangements that have occurred between the A and C genomes.



Supplementary figure 10 – A bimodal distribution of self-clustering probabilities necessitates the use of a threshold to visualise the probabilities
Self-clustering probabilities are calculated as detailed in the main text (Methods). The density curves presented here represent the self-clustering probabilities calculated from a single SOM. The clustering coefficient threshold was taken by determining the self-clustering probability that corresponded to the peak of the density curve. This threshold was calculated for each SOM and averaged to give the final threshold: apex threshold = 0.053; leaf threshold = 0.056.

Supplementary tables

Date sampled	Days post sowing	Days vernalised	Days post vernalisation	Tissue Type	
				Leaf	Apex
2014-05-29	22	0	-	2	2
2014-06-19	43	21	-	2	2
2014-07-10	64	42	-	2	2
2014-07-11	65	42	1	1	1
2014-07-13	67	42	3	2	2
2014-07-15	69	42	5	0	1
2014-07-18	72	42	8	2	2

Supplementary table 1 – Sampling and sequencing scheme for the developmental time series
The numbers in the rightmost two columns indicate the number of biological pools sampled for that time point within each tissue.

Days post sowing	Apex			Leaf		
	Both expressed	A genome 2-fold higher	C genome 2-fold higher	Both expressed	A genome 2-fold higher	C genome 2-fold higher
22	136	11 (8.1%)	19 (14.0%)	109	8 (7.3%)	14 (12.8%)
43	149	15 (10.1%)	24 (16.1%)	118	12 (10.2%)	16 (13.6%)
64	147	12 (8.2%)	20 (13.6%)	114	11 (9.6%)	13 (11.4%)
65	145	13 (9.0%)	25 (17.2%)	108	10 (9.3%)	16 (14.8%)
67	138	14 (10.1%)	19 (13.8%)	112	7 (6.3%)	12 (10.7%)
69	139	11 (7.9%)	18 (12.9%)	-	-	-
72	142	15 (10.6%)	21 (14.8%)	112	5 (4.5%)	14 (12.5%)

Supplementary table 2 – Number of genes expressed 2-fold higher than their homoeologue for all flowering time gene homoeologue pairs.

As for Table 1 in the main text, calculated using homoeologue pairs which showed sequence similarity to *A. thaliana* flowering time genes from the FLOR-ID database[1]. The geometric mean of the fold difference of the C genome gene relative to the A genome homoeologue for all flowering time homoeologue pairs is 1.10 in the apex and 1.04 the leaf.

Gene	Forward Primer (5’ – 3’)	Reverse Primer (5’ – 3’)	Amplicon Length
TFL1 A10	GTCTCCAATGGCCATGAGT	GTGCCGGGGATGTTCATG	179
TFL1 Cnn	GTCATGAACATCCCCGGC	GATCATTCTCGATCGCAAATTCA	196
TFL1 C2	CTGATGTTCCAGGTCCTAGC	TGGGGAGATATCGATAACATGTC	197
TFL1 C3	GAGGTGGTGAGCTATGAGTTG	CTGGGCGTTAAAGAAGACAGCA	189
GAPDH	AGAGCCGCTTCCTTCAACATCATT	TGGGAACACGGAAGGACATTCC	112

Supplementary table 3 – qPCR primer sequences

Tissue	Days post sowing	Sequencing Run 1				Sequencing Run 2			
		Total reads (millions)	Mapped reads (millions / percentage of total)	Multiply mapping reads (millions / percentage of mapped)	Reads mapped to over 20 positions (ten thousand / percentage of mapped)	Total reads (millions)	Mapped reads (millions / percentage of total)	Multiply mapping reads (millions / percentage of mapped)	Reads mapped to over 20 positions (ten thousand / percentage of mapped)
Apex	22	75.6	61.8 (81.8%)	8.3 (13.4%)	20.7 (0.3%)	41.9	34.3 (81.9%)	4.7 (13.8%)	7.8 (0.2%)
Apex	43	71.5	56.8 (79.4%)	7.4 (13.1%)	17.8 (0.3%)	31.7	25.3 (79.8%)	3.4 (13.6%)	5.3 (0.2%)
Apex	64	70.5	57.4 (81.4%)	7.5 (13.0%)	21.6 (0.4%)	28.7	23.3 (81.2%)	3.2 (13.8%)	149.4 (6.4%)
Apex	65	67.6	54.6 (80.7%)	7.2 (13.2%)	26.5 (0.5%)	NA	NA	NA	NA
Apex	67	78.6	63.5 (80.8%)	8.4 (13.2%)	36.3 (0.6%)	30.5	25.1 (82.3%)	3.5 (13.9%)	5.6 (0.2%)
Apex	69	66.2	54.4 (82.2%)	7.3 (13.5%)	30.7 (0.6%)	NA	NA	NA	NA
Apex	72	59.7	48.6 (81.4%)	6.4 (13.2%)	35.2 (0.7%)	31.5	25.8 (81.8%)	3.6 (14.1%)	4.5 (0.2%)
Leaf	22	68.2	54.7 (80.2%)	8.4 (15.4%)	9.5 (0.2%)	33.9	28.0 (82.5%)	4.4 (15.7%)	3.7 (0.1%)
Leaf	43	50.5	41.5 (82.1%)	6.2 (15.0%)	11.1 (0.3%)	33	26.4 (80.1%)	4.0 (15.1%)	4.6 (0.2%)
Leaf	64	73.9	60.7 (82.1%)	8.8 (14.4%)	10.2 (0.2%)	35.5	29.1 (82.1%)	4.3 (14.8%)	3.7 (0.1%)
Leaf	65	45.7	37.6 (82.2%)	5.5 (14.6%)	5.4 (0.1%)	NA	NA	NA	NA
Leaf	67	81.8	67.1 (82.1%)	10.0 (14.9%)	9.4 (0.1%)	35.7	28.8 (80.7%)	4.4 (15.4%)	3.5 (0.1%)
Leaf	72	49	40.3 (82.1%)	5.8 (14.5%)	5.8 (0.1%)	32.2	26.2 (81.2%)	3.9 (15.1%)	3.9 (0.1%)

Supplementary table 4 – Sequencing statistics for the two sequencing runs carried out to generate the developmental transcriptome

Reads were mapped to the Darmor-*bzh* reference genome[3] using TopHat[4] as described in the main text (Methods). The percentage of mapped reads is given as the percentage of the total reads. Multiply mapped reads are defined as reads that mapped to multiple places in the genome with an equal probability. The percentages of multiply mapped reads and the percentage of reads mapping to more than 20 position in the genome are calculated as a total of the reads that were mapped to the genome, and not a percentage of the total reads.

References

1. Bouché, F., Lobet, G., Tocquin, P. & Périlleux, C. FLOR-ID: an interactive database of flowering-time gene networks in Arabidopsis thaliana. *Nucleic Acids Res.* **44,** D1167–D1171 (2016).
2. Lou, P. *et al.* Preferential Retention of Circadian Clock Genes during Diploidization following Whole Genome Triplication in Brassica rapa. *Plant Cell* **24,** 2415–2426 (2012).
3. Chalhoub, B. *et al.* Early allopolyploid evolution in the post-Neolithic Brassica napus oilseed genome. *Science* **345,** 950–953 (2014).
4. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14,** R36 (2013).