# On Gene Regulatory Networks Controlling Flowering Time in *Brassica napus*

Gurpinder Singh Sidhu

A Thesis
submitted to University of East Anglia
for the degree of Doctor of Philosophy

September 2025

## Abstract

Flowering plants respond to multiple environmental and endogenous cues to correctly time the crucial transition from the vegetative to floral state. Most of our knowledge of the gene regulatory network (GRN) controlling flowering time has been derived from the model plant, *Arabidopsis thaliana*. This knowledge needs to be translated to crop plants to support the development of varieties that can be grown in different and rapidly changing climatic conditions. With the dual challenges of more complex genomes and limited prior knowledge in crops, however, this translation is not always straightforward. In this thesis, I present a study of the GRN controlling flowering time in *Brassica napus* (Oilseed rape), an allotetraploid crop that is a close relative of Arabidopsis. Using a comparative transcriptomics approach, I show that the majority of the orthologous gene pairs have similar expression dynamics over plant development between Arabidopsis and *Brassica napus*. However, flowering time genes exhibit a significantly higher rate of divergence in expression patterns from their Arabidopsis orthologues. Despite these divergences, the inferred GRN consisting of these genes exhibits a similar network topology to the network known in Arabidopsis. This is likely due to preferential retention of these genes in higher paralogue numbers, which allows subtle changes in the regulation of individual paralogues, while still conserving the overall regulatory structure through evolutionary time. I discover and present a detailed analysis of one such example where the orthologues of gene, *SOC1*, have similar expression patterns under normal conditions, but have diverged expression patterns under cold temperature conditions, suggesting divergence in regulation among paralogues in response to the temperature change. Altogether, this thesis expands our understanding of the environmental and genetic control of flowering time in *Brassica napus* and provides a study of regulatory divergences among paralogous genes in a polyploid system.

## Acknowledgements

Surprisingly, I have managed to write something that somewhat resembles a PhD thesis. Unsurprisingly, I have had a lot of help from people smarter, kinder and more capable than me. Years ago, Richard allowed me to have my first ever proper research experience as an undergradute that has snowballed into this thesis today. I am thankful to him for all I have learned over these years. His mentorship style and the culture he has cultivated in the group are amazing. I consider myself lucky to have done a PhD under his supervision.

While I credit Richard for helping me develop my computational skills, I would have been lost in the weeds of biology if not for Rachel. She has helped me draw biological inferences from my results at every step of my research and thanks to her, I can also claim to have some lab skills. I am also thankful to Wilfried Haetry, who, as part of my supervisory team, has endured all my main review presentations, while providing feedback that has helped sculpt this work into its current form.

Hugh Woolfenden has been my go-to source for all my bioinformatics and computational queries. His data processing scripts were the first examples of bioinformatics code I had seen and to this day, a lot of my coding practices are derived from his code. I am grateful that he, along with Shannon Woodhouse, also supervised me during the summer school at the height of the pandemic. A big thank you to everyone in the Morris group, and my fellow team Brassica members, Ruth Kristianingsih and Aileen Magilin, for making the student office an amazing place to work.

My friends Jaspreet Singh, Molly Bergum and Markus Dräger have kept me company outside of science; and I am thankful for their friendships. Lastly, I would not have been able to do any of this if not for the efforts of my parents to equip me with the best possible education and cultivating my interest in science. I will be forever thankful to my family for all their hardwork and their support.

And finally, thank you to the music of Sigur Rós, thirteen cans of Coke Zero, Mars bars, Financial Times and the spider webs that kept appearing in my library carrel for keeping me company during the writing process.

## Publications

My PhD work has resulted in a few publications, which are listed below. While all these publications have originated from my work during my PhD, publications directly related to either the results or methods presented in this thesis are marked with †.

† Yalcin HA, Jacott CN, Ramirez-Gonzalez RH, Steuernagel B, **Sidhu GS**, Kirby R, Verbeek E, Schoonbeek H, Ridout CJ, Wells R. A complex receptor locus confers responsiveness to necrosis and ethylene-inducing like peptides in *Brassica napus*. The Plant Journal. 2024. doi: 10.1111/tpj.16760

Jacott CN, Schoonbeek H, **Sidhu GS**, Steuernagel B, Kirby R, Zheng X, Tiedermann A, Macioszek VK, Fell H, Bruce Fitt DL, Mitrousia GK, Stotz HU, Ridout CJ, Wells R. Pathogen lifestyle determines host genetic signature of quantitative disease resistance loci in oilseed rape (*Brassica napus*). Theoretical and Applied Genetics. 2024. doi: 10.1007/s00122–024–04569–1

*Preprints*

Hoerbst F, **Sidhu GS**, Tomkins M, Morris RJ. What is a differentially expressed gene? biorxiv. 2025

Hoerbst F, **Sidhu GS**, Omori T, Tomkins M, Morris RJ. A bayesian framework for ranking genes based on their statistical evidence for differential expression. biorxiv. 2025

Hoerbst F, **Sidhu GS**, Tomkins M, Morris RJ. A Closed-Form Solution to the 2-Sample Problem for Quantifying Changes in Gene Expression using Bayes Factors. arxiv. 2024

*In preparation*

† **Sidhu GS**, Burrows S, Woolfenden H, Wells R, Morris RJ. Network analysis of flowering time genes suggests regulatory changes among *SOC1* orthologues in response to cold in *Brassica napus*.

† Kristianingsih R, Calderwood A, **Sidhu GS**, Woodhouse S, Woolfenden H, Kurup S, Wells R, Morris RJ. Identifying dynamical similarities between sets of gene expression profiles using curve registration.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# The control of flowering time

## Abstract

How do plants know when to flower has been a question under investigation for centuries. While earlier investigations postulated the role of both environmental and endogenous cues, the current molecular biology era is concerned with finding the genetic targets of these factors, and elucidating the complex regulatory network controlling the timing of this transition. In this chapter, I present a brief overview from early forays to the cutting edge of our knowledge in the control of flowering time. Most of our knowledge of genetic regulation of flowering time is in the model plant *Arabidopsis thaliana* with some efforts to translate this knowledge to agronomically important crop plants. *Brassica napus*, the focus of this thesis, is a closely related crop to the model plant and a prime candidate for similar translational research. The chapter swivels from historical narrative to factual summaries of research, with the aim of giving the reader a comprehensive background and overview to understand the results presented in following chapters in the wider context.

## 1.1 It’s all in the timing

On 7 April 1874, Charles Darwin wrote a letter to J.D. Hooker, forwarding some moth larvae in the post, encouraging him to try using them to fertilise the Yucca plants — as “the most wonderful case of fertilisation ever published” — or to throw them away, as he had no plants himself [1].

Charles Valentine Riley, whose work Darwin was alluding to in his letter, in 1873, published one of the first descriptions of the process of pollination in the Yucca plants, which occurs during oviposition of the Yucca moth (now, *Tegiticula yuccasella*) [2]. This mechanism is now studied as the archetypical example in co-evolution, where Yucca plants have synchronised the timing of their flowering, with the emergence of their active pollinators. The Yucca moths posses tentacle like mouth parts that collect pollen and deposit it on another flower, ensuring the provision of seeds for their offspring [3].

We now know that plants do not flower randomly. They have evolved to ensure the success of their reproductive endeavours, and hence ensure that they flower during favourable conditions, often determined by multiple environmental factors. Even in the case of the Yucca-Yucca moth interaction, from the point of view of a Yucca plant, environmental cues such as temperature, precipitation and photoperiod aid in determination of its timing to flower. In fact, a change in these factors can sometimes ‘trick’ the Yucca plants into flowering even when their active pollinators are not around [4].

The transition to flowering, or the floral transition, is one of the multiple developmental transitions that plants undergo in their life cycle. For seed producing spermatophytes, life begins as a seed in embryonic stage. Germination is the first transition that marks the shift to postembryonic stage of growth. The resultant seedling, in the case of angiosperms, undergoes a juvenile vegetative phase before adult vegetative phase, where it transitions to flowering, marking the start of reproductive phase. Plants respond to cues influencing the transition to flowering within the adult vegetative phase, and the transition from this to reproductive phase is termed floral transition [5].

Ambient temperature, photoperiod, the circadian clock and in some plant species, exposure to long term winter cold, are environmental signals that influence the timing of floral transition. Endogenous cues, such as the age of the plant and certain growth hormones, such as gibberellins and sugars also have an effect. A lot of scientific investigations for the past few decades have been focussed on understanding the sensing and integration of these signals, and interplay of genetic interactions that regulate the timing of the switch to flowering [5] [6].

### 1.1.1 Flowering time research: initial forays

In Arlington, Virginia, two scientists working for the USDA, Wightman Garner and Harry Allard, were investigating the timing of blossoming in multiple crops. Among them was ‘Maryland Mammoth’, a tobacco variety that would only switch to flowering after producing about 100 leaves, as opposed to other tobacco plants that transitioned to flowering after only about 20 leaves. This made propagation of this variety difficult, as plants would often die of frost damage before setting any seed. Their reported investigations, published in 1920, are now accredited as the first experiments that determined the role of daylength as a factor in controlling the timing of floral transition. [7]

There already had been postulations about the role of environmental factors, particularly daylength in determining the transition to flowering in plants. For instance, Georg Klebs, in 1913 published “The development of flowering plants” (“Der Entwicklungsgang der Blütenpflanzen” [8]), in which he emphasised the role of daylight duration as a “catalytic factor”,

*“In der freien Natur wird sehr wahrscheinlich die Blütezeit dadurch bestimmt, daß von der Tag- und Nachtgleiche (21. März) ab die Länge des Tages zunimmt, die von einer gewissen Dauer ab die Anlagen der Blüte veranlaßt. Das Licht wirkt wohl nicht als ernährender Faktor, sondern mehr katalytisch”*

*“In nature, the flowering period is most likely determined by the fact that from the equinox (March 21) onwards, the length of the day increases, which, after a certain period, triggers the development of flowering. The light probably acts not as a nourishing factor, but rather as a catalytic factor”*

Garner and Allard were the first to perform controlled experiments that clearly demonstrated the role of light duration on flowering time. They sowed the tobacco plant seeds in two sets, but exposed one set to short day conditions by transferring them to a dark shed in the afternoon. Their reported data showed that by giving plants a short-day treatment of only 5 hours of daylight, the time to flowering was reduced to just around 60 days, from about 155 days in control. They coined the term *photoperiod*, now recognised as an important pathway that controls the timing of flowering transition [9].

Several research studies followed this discovery with aims to understand how photoperiod was perceived by plants [10] and in which tissue it was perceived [11]. The famous 'florigen' hypothesis, first posited by the soviet physiologist, M. Chailakyan [12] [13], stated that the sensing of daylength occured in the leaves, and then the stimulus traveled towards the shoot apex to initiate flower formation. Experiments where leaves of photoperiodically induced donors were grafted onto non-induced receptors to promote flowering were explained by this hypothesis [14], and it still holds merit into the present molecular biology era, so it will come back in further detail in the following sections. For now, we turn our attention to research towards the next important factor influencing the floral transition, temperature.

Temperature, as it turns out, has two ways to influence floral transition. The research linking ambient temperature, the first of the two ways, to plant phenology dates back to early 18th century. In 1735, René A. F. de Réaumur, a French entomologist, argued that plants require a certain amount of heat to reach a given stage of maturity. His thermometric constant, measured in Réaumur scale was termed Réaumur's thermal constant of phenology [15] [16]. While Réaumur's temperature scale is now obsolete, except among a few cheesemakers [17], his ideas formed the basis of further research correlating reaching vegetative maturity (and subsequently flowering) to mean temperatures [18], including the introduction of "Growing Degree Days", as a unit to measure the amount of heat needed for a plant to reach a certain growth stage [19].

The second temperature aligned factor concerns the effect of prolonged exposure to low temperature conditions. Gustav Gaßner, a German botanist, in 1918, published his seminal text on the effects of low temperature on crops, highlighting differences in cold requirements among plants [20]. He classified numerous species either as "biennials" or "winter-annuals", that require long term exposure to cold-treatment or as "spring plants" or "summer annuals" that do not need chilling treatment prior to flowering.

Trofim Lysenko, a (now infamous) Soviet agronomist, studied these cold requirements in cereals and showed that even seeds, without requiring excessive germination, were responsive to this cold treatment under slightly imbibed conditions. This had significant effects on agriculture in the USSR as this allowed the use of sowing machines as excessive germination prior to transplantation was not needed. He termed it 'Jarovisation', from the word 'Jarovoe' for cereals in Russian. This word was translated to English, French and German as 'Vernalisation', from the Latin word *vernum*, meaning 'related to spring' [21]. Vernalisation does not necessarily cause plants to commence flowering, but rather primes the plant to do so. Flowering often occurs well after vernalisation, however, plants maintain a 'memory' of winter as the effects of vernalisation are mitotically stable.

The studies of possible inheritance of this acquired trait were encouraged by Lysenko, as the director of the Institute of Genetics of the Academy of Sciences of the USSR. His 'Lysenkoist doctrine' was poliferated for political purposes and scientists studying genetic inheritance were tragically imprisoned and prosecuted [22], including the well-known geneticist Nikolai Vavilov [23].

The current accepted definition of vernalisation was provided by Chouard in 1960, as 'the acquisition or acceleration of the ability to flower by a chilling treatment' [21]. Grafting experiments further confirmed that this acquisition occurs at the shoot tip, and it remains among the most important areas of flowering time research in present times [24].

While the research on environmental factors driving the timing of flowering transition was being carried out, investigations into endogenous cues were also underway. Eiichi Kurosawa, a Japanese botanist in 1926, noticed that an infection by the fungus *Gibberella fujikuroi* caused rice seedlings to grow taller. They would grow tall enough to not be able to withstand their own weight, without ever transitioning to flowering. However, unlike environmental factors, the investigations into endogenous cues have proven to be more challenging — often leading to conflicting results. Gibberellic acid, the causative chemical from Kurosawa's study was isolated in 1938. While it delayed flowering in rice seedlings, a study in 1952 by Anton Lang reported acceleration instead in *Samolus parviflorus* and *Crepis tectorum* [25] [26]. While Gibberellic acid managed to earn the moniker of 'flowering hormone', studies distingushed it from the hypothesised 'florigen', and few even ascribed its effects as secondary to environmental effects [25]. Carbohydrates, were also shown to increase in concentration within the shoot apex, just prior to the transition to flowering [27]. These endogenous cues likely ensure that the plant flowers when it is at right age, with correct hormone and nutrient levels to support flower formation.

In conclusion, research into flowering time has had a long and eventful history. Centuries of research has shaped our understanding of how plants time this transition. All of this was before the dawn of modern molecular biology techniques, following which, the pace has only increased. We now have knowledge of a number of genes constituting these pathways. The following section moves away from history and reviews the current knowledge of floral transition in the model plant, *Arabidopsis thaliana*.

### 1.1.2 Flowering time control in *Arabidopsis thaliana*

Thale cress or *Arabidopsis thaliana*, owing to its small genome size, short life cycle and ease of cultivation and crossing, is the model in plant research. It is an annual flowering plant with long-days promoting floral transition [28]. In the present era of molecular biology, we have now identified genes that likely constitute photoperiod, vernalisation, autonomous, hormonal, ageing and ambient temperature pathways that control the timing of the shoot apical meristem's transition to flowering [29].

Research using Arabidopsis mutants first unveiled links to genetic loci that affected flowering time of plants [30] [31]. Among the first mutants identified was the *constans* (*co*) mutant. The gene *CONSTANS* (*CO*) is now known to be the central gene within the photoperiod pathway. *CO* encodes a putative zinc finger transcription factor that, research suggests, acts as a bridge between the circadian clock and the control of flowering time [32]. *CO* is controlled by the circadian clock of the plant, and is affected by the expression of circadian genes such as a putative membrane protein encoding gene, *GIGANTEA* (*GI*) and *CYCLING DOF FACTOR 1* (*CDF1*) which encodes for a zinc finger domain containing protein [33]. The regulation of *CO* has been studied extensively and there are multiple upstream regulators and downstream targets of this gene — however, one main target that is of particular interest is a gene that encodes for a phosphatidylethanolamine-binding protein, *FLOWERING LOCUS T* (*FT*).

The 'florigen' hypothesis, as introduced in the previous section, hypothesised the existence of a signal from leaf to apex that induces flowering in plants. As *CO* emerged as a key gene within the photoperiod pathway, it was questioned if *CO* could be the long distance signal that induces flowering [34]. However, research suggests that it regulates a protein, FT, through direct binding, that then acts as the florigen signal [35]. *FT* has been shown to be expressed within the companion cells and its protein then moves to the shoot apical meristem through the phloem [36]. As FT reaches the shoot apical meristem, it forms a complex with a bZIP transcription factor encoded by *FLOWERING LOCUS D* (*FD*), that then activates a meristem identity gene, *APETALA 1 (AP1)* to induce flowering [36]. However, the formation of the FT-FD complex has yet to be demonstrated with the shoot apical meristem [29].

Besides the photoperiod, temperature also regulates the expression of the florigen, *FT*. *SHORT VEGETATIVE PHASE* (*SVP*) encodes the key MADS box regulator constituting the ambient temperature pathway. SVP binds to 'the CArG box', a sequence motif within the *FT* promoter and represses its activity [37]. *SVP* acts along with the protein encoded by *FLOWERING LOCUS C* (*FLC*), as a complex for a lot of its target genes while regulating flowering transition [38]. *FLC* is the central gene within the other temperature related pathway — vernalisation.

As introduced in the previous section, vernalisation refers to the process by which flowering is induced by a long period of exposure to cold temperature. In Arabidopsis, *FRIGIDA* (*FRI*), a gene that encodes a likely nuclear protein, plays a role in determination of vernalisation requirement in plants [39]. *FRI* confers vernalisation requirement through promotion of accumulation of the *FLC* mRNA [40]. *FLC* encodes a MADS box protein that directly blocks the transcription of *FT* and another MADS box floral integrator, *SUPRESSOR OF CONSTANS 1* (*SOC1*) [41]. This hinders the photoperiodic activation of these genes, and hence does not allow the plant to start flowering. Long term exposure to cold epigenetically silences *FLC*. This silencing mechanism, as well as maintenance of the silenced state of *FLC* as 'memory of winter' is a complex process.

The mutants *vrn1* and *vrn2* show completely normal silencing of *FLC*, however, the silencing isn't maintained after return to warm conditions [42] [43]. Later, it was discovered that *FLC* repression is disrupted in *vin3* mutants [44]. So, at a very simple level, VERNALISATION 3 (*VIN3*), a gene that encodes a zinc finger protein, is required for initial silencing of *FLC*, and its state is maintained by *VERNALISATION 1* (*VRN1*) encoded DNA binding protein and another nuclear-localized zinc finger protein encoded by *VERNALISATION 2* (*VRN2*) in the dividing plant cells post vernalisation [45]. Furthermore, research also points to the role of antisense transcription in silencing, and role of autonomous pathway genes in maintainence of the silenced state [29].

Similar to the environmental factors, research has been done to determine the genomic loci involved in the hormonal pathway as well. External application of gibberellic acid (GA) was shown to restore a plant's ability to flower in *ga1–3* loss of function mutants [46]. *GIBBERELLIC INSENSITIVE DWARF 1* (*GID*) is the key gene in this pathway that interacts with a family of proteins called DELLA proteins, to regulate GA signal transduction. *SOC1* is also regulated by the GA pathway [47]. Besides GA, sugars are also known to influence flowering time, but their genetic mechanisms are not well understood yet [29]. Carbohydrate metabolism could be linked to the ageing pathway, as it also affects a novel pathway based on miRNA156, which decreases as the plant ages, independent of photoperiod, vernalisation and GA [48]. These cues act as 'failsafes' and allow the plant to transition to flowering even when environmental cues are not ideal.

The section above, with a high level review of the pathways that constitute flowering time control, shows that the regulatory network controlling flowering is complex and contains multiple genes that often act redundantly. Figure 1.1 provides a graphical overview of the network described. All this knowledge has validated the initial ideas that researchers had in the pre-molecular biology era. Interestingly though, there is now evidence that these different pathways do not act strictly independently — there is an extensive overlap in their function and genes that constitute those pathways. In fact, these pathways act on a few key genes that are termed 'integrator genes'. These are central regulators that switch on the meristem identity genes to produce flowering [49].

*FT*, the already introduced 'florigen' is one such integrator gene. Photoperiod pathway, via *CO* is not the only pathway regulating this gene, as it is also a target of *FLC*. Besides these genes, ambient temperature and hormonal pathways also modulate *FT* expression [51].

*SOC1*, initially discovered as a supressor of *CO* overexpression, is also a target of *FLC* and *SVP* within the shoot apical meristem. It is also affected by GA signalling pathway and its mRNA levels increase alongside miR156 and *SQUAMOSA PROMOTER BINDING PROTEIN-LIKE* (*SPL*) genes that encode for SBP-box containing transcription factors [48]. *SOC1* functions as a promoter of flowering within the shoot apex. It acts together with another MADS box floral integrator gene, *AGAMOUS-LIKE 24* (*AGL24*). Regulated independently of *FLC*, *AGL24* is upregulated during vernalisation [52]. *SOC1* and *AGL24* also regulate each other, forming a heterodimer, to regulate the activity of another integrator gene, *LEAFY (LFY)* [53].

The transcription factor encoded by *LFY*, in addition to its role in regulating

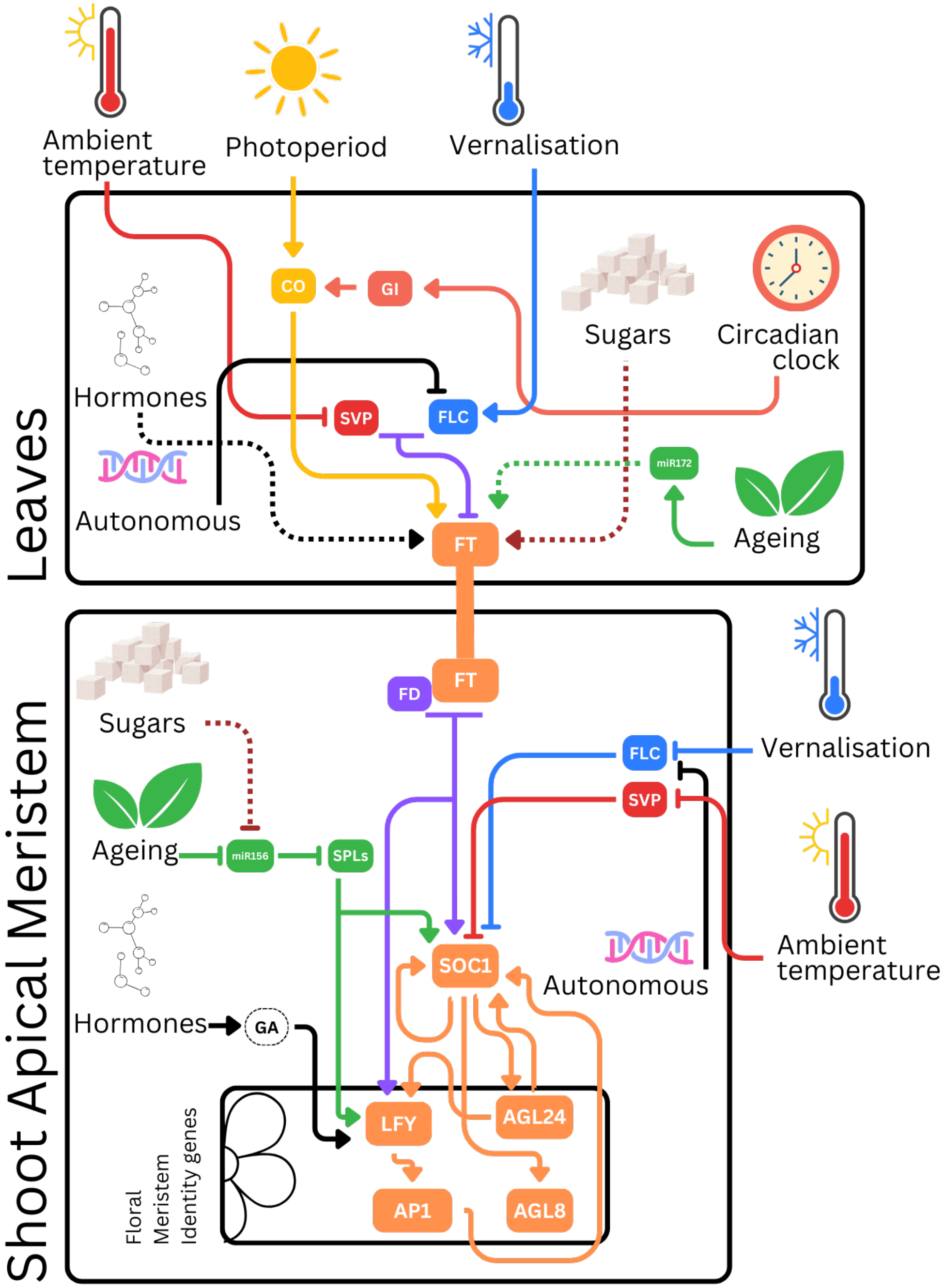


**Figure 1.1: Overview of Gene regulatory network controlling flowering time with key genes in *Arabidopsis thaliana***

Dotted lines indicate putative mechanisms. Interactions between floral integrators are shown in orange. Simplified from the originals by Bouché *et al.* [50] and Srikanth and Schmid [29].

floral transition, also has a role in flower development. In fact, the first recognition of *LFY* in relation to flowering was for its function in flower development, with *lfy* mutants showing leaf-like structures replacing flowers [54]. *LFY* is a positive regulator of flowering, and a regulator as well as a target of *SOC1* [55]. In addition to *SOC1*, it is also regulated by GA and miRNA-156 regulated SPLs [56]. Like *LFY*, *AGL24* is also involved in flower development in addition to its role as an integrator gene in floral transition regulation [57]. *AP1*, mentioned earlier as a downstream target of *FT*, is a MADS box meristem identity gene that also regulates *SOC1* expression and functions alongside floral integrator genes.

In conclusion, early 21st century has seen rapid advancements in identification of genes associated with factors that were shown to be regulating flowering in the centuries prior. Research has shown that it is a network of genes, over 300 in number [58], working in different pathways, with a degree of crosstalk that controls the switch to flowering. While there clearly is more research to be done, the current knowledge in Arabidopsis serves as a starting point to study the regulation in crop species, which is the focus of the section that follows.

### 1.1.3 Flowering time control in crop plants

Garner and Allard's investigations into the effect of photoperiod on flowering stemmed from practical agronomic questions. Just as plants in wild context need to time their floral transition for their reproductive success, synchronised flowering at the correct time is important for agronomic output. Domestication of plants and subsequently, the rise of modern agriculture has witnessed the expansion of growth areas of a lot of crop plants far away from their centres of origin.

This expansion, particularly in latitudinal range, has likely led to artificial selection of traits that have likely led to modifications and divergences in the flowering physiology of crops from their wild ancestors. These selections by humans would likely have taken place based on multiple factors, such as ability to grow in high density monoculture, shoot architecture, shorter growth periods to fit into rotations, delayed bolting to increase certain organ sizes and of course, maximum yield [59]. With an increase in understanding of flowering regulation at a molecular level within Arabidopsis, reseachers have turned their attention to achieve similar understanding in other crops as well [60]. A lot of this research involves identification of orthologues of Arabidopsis flowering time genes within the crop plant of interest, often complicated by more complex genomes present in non-model species.

Research into vernalisation within Wheat (*Triticum aestivum*), the world’s highest acreage crop, has aimed to understand the winter and spring growth habits of its cultivars [61]. Studies have identified genes regulating flowering that share homology with Arabidopsis flowering genes. For instance, a core component of vernalisation pathway in wheat, the gene *VERNALISATION 1* (*VRN1*), encodes a protein homologous to Arabidopsis *AP1* [62]. *VERNALISATION 2* (*VRN2*), a gene homologous to Arabidopsis *COL* (*CO-LIKE*), is downregulated via vernalisation and acts as a floral repressor, comparable to the role played by *FLC* in Arabidopsis [63]. *VERNALISATION 3* (*VRN3*), the third component of this pathway, is similar to the florigen *FT* [64]. Loss of *VRN2* leads to an increase in *VRN1* and *VRN2* levels, leading the cultivar to exhibit spring annual behaviour [65]. The *VRN1*, *VRN2* and *VRN3* pathway is also conserved in another cereal crop, Barley (*Hordeum vulgare*) [64].

Soybean (*Glycine max*) has spread from its origin in the Huang-Huai Valley in Central China, a temperate region to both higher and lower latitude regions [66]. Research has shown that adaptation of cultivars has led to circadian clock genes involved in controlling flowering time being the main targets of domestication. Mutations in flowering suppressor genes within the photoperiod pathway are responsible for adaptation to higher latitudes. *E1*, a legume specific transcription factor as part of the photoperiod pathway [67], *E2*, an orthologue of Arabidopsis *GI* [68], *E3*, an orthologue of Arabidopsis *PHYTOCHROME A* (*PHYA*) [69] and lastly, two orthologues of Arabidopsis *LATE ELONGATED HYPOCOTYL 1* (*LHY1*) [70] [71], central component of circadian clock, have been identified as components of the five flowering suppressors facilitating the adaptation in flowering time. For lower latitudes, loss of an orthologue of Arabidopsis *EARLY FLOWERING 3* (*ELF3*) leads to an extended vegetative phase and higher yields at lower latitudes [72].

Wild rice originated in low latitudes and hence exhibits characteristics of short-day flowering plants [73]. Maize (*Zea mays*) has similar origins in low lands of Mexico [74]. Within cultivated rice, genes homologous to Arabidopsis *CO* and *COL* have facilitated cultivation of rice at higher latitudes [75]. In both rice and maize, orthologues of the florigen, *FT* have been selected for local adaptations [76] [77] [78].

Despite the conserved role of circadian clock genes, photoperiod pathways in multiple crops, some of which have been highlighted above, the *FLC* clade MADS-box genes have not been reported as major flowering genes in crops outside the Brassicaceae family [60]. The following section introduces this family of crops plants and the plant at the centre of this work, *Brassica napus*.

## 1.2 The tale of *Brassica napus*

Gaius Plinius Secundus or Pliny the Elder in his 'Natural history' (Latin: *Naturalis historia*, AD 77–79), the single largest work to survive from the roman period into present day, talks about the wonderful praise showered on 'Brassica' by Cato the Elder (in *De agri cultura*, 160 BC).

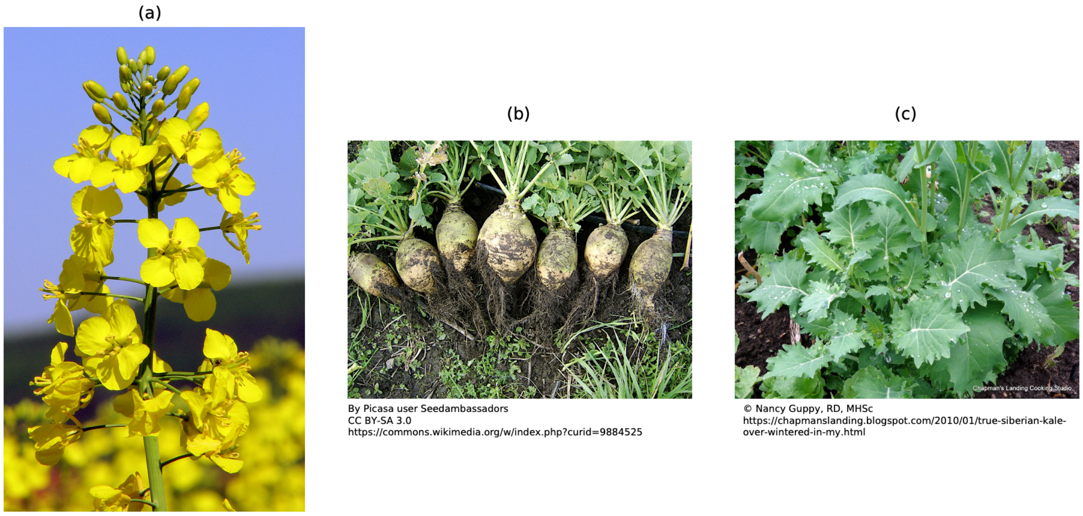
*"...sed Cato brassicae miras canit laudes, quas in medendi loco reddemus. genera eius facit: extentis foliis, caule magno, alteram crispo folio, quam apiacam vocant, tertiam minutis caulibus, lenem, teneram minimeque probat"* (Book 19, Chap. 41)

The 1855 translation of this seminal work [79] into English takes away some epigrammatic beauty,

*"...but Cato, on the other hand, sings the wondrous praises of the cabbage, the medicinal properties of which we shall duly enlarge upon when we come to treat of that subject. Cato distinguishes three varieties of the cabbage; the first, a plant with leaves wide open, and a large stalk; a second, with crisped leaves, to which he gives the name of 'apiaca'; and a third, with a thin stalk, and a smooth, tender leaf, which with him ranks the lowest of all"*

This early description of 'brassicae', later, formally described by Carl Linnaeus as genus *Brassica* [80], notably mentions how different varieties of 'brassicae' (translated to 'cabbage') had different morphologies. The *Brassicaceae* family is noted for its extensive genetic and phenotypic diversity, often within the cultivars of same species [81]. The artificial selection of different parts has created diverse morphologies. For instance, from *Brassica oleracea*, selection for leaves has produced the modern day Kale, for inflorescences has led to Cauliflower and Broccoli and Kohl-rabi is a result of breeding for stems [82].

Woo Jang-chun (or Nagahara U in Japanese) first described the genetic relationship between the six cultivated species of Brassicas [83]. His conclusions are summarised in the now famous 'Triangle of U', shown in Figure 1.3. The diploid



By Tilo Hauke - Self-published work by Tilo Hauke
CC BY-SA 3.0
https://commons.wikimedia.org/w/index.php?curid=777407

By Picasa user Seedambassadors
CC BY-SA 3.0
https://commons.wikimedia.org/w/index.php?curid=9884525

© Nancy Guppy, RD, MHSc
https://chapmanslanding.blogspot.com/2010/01/true-siberian-kale-over-wintered-in-my.html

**Figure 1.2: The diverse cultivated forms of *Brassica napus***

*Brassica napus* has diversity in its cultivated forms. (a) *Brassica napus* subsp. *napus* or Oilseed rape is the most commonly cultivated form. (b) *Brassica napus* subsp. *napobrassica* or Swede is a root vegetable. (c) *Brassica napus* subsp. *pabularia* or Siberian Kale is a leafy vegetable.

species within the genus are *Brassica rapa*, *Brassica nigra* and *Brassica oleracea*, representing the A, B and C genomes respectively. Their hybridisation has led to the formation of three allotetraploids, *Brassica carinata* (BBCC), *Brassica juncea* (AABB) and *Brassica napus* (AACC).

Out of these, *Brassica napus* ranks the high among the most economically important crops in the world [84]. It is grown for its oil which is used for cooking and industrial applications that dramatically increased with the advent of industrial era [85]. *Brassica napus* originated about 7500 years ago due to hybridisation of *Brassica oleracea* and *Brassica rapa* [86]. The origin of *Brassica napus* is an active area of research and no truly wild populations of *Brassica napus* are known. Interestingly, *Brassica napus* has diverse types of cultivated forms, shown in Figure 1.2, ranging from oilseed (subsp. *napus*) to swedes (subsp. *napobrassica*) and leafy (subsp. *pabularia*) types. Hence, it has been posited that the hybridisation events leading to the formation of the allotetraploid occurred multiple times. The most common-type, *Brassica napus* subsp. *napus*, (simply referred to as *Brassica napus* throughout this thesis), also has further diversity in cultivar growth habits, with winter-type, semi-winter and spring-type cultivars, originating at different points due to hybridisation events followed by selection [87]. The duplicated nature of the genome has lead to

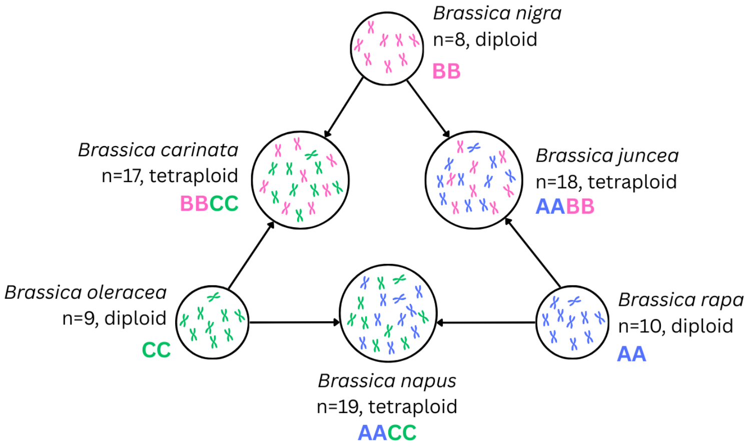


**Figure 1.3: The 'Triangle of U' showing relationships among six species of the genus *Brassica***

Chromosomes from each of the genomes are represented in different colours. The diploid species represent the three vertices of the triangle, while the tetraploid species form the edges.

a remarkable degree of diversity in cultivated forms of *Brassica napus*. In fact, species of this family are closely related and have intertwined origins, with close links to the model plant, *Arabidopsis thaliana*.

The Arabidopsis-Brassica lineage split between 14.5 to 20.4 million years ago [88]. Genomes of diploid Brassicas underwent a triplication event following this split [89]. This would mean that the hybridised tetraploid *Brassica napus* should theoretically have six paralogues for each Arabidopsis orthologue, however on average, there exists only a four-fold difference in the number of genes between the two species on the whole genome level [86]. This indicates gene loss over evolutionary time. The Flowering time genes on the contrary however, have been shown to be preferentially retained [90], with orthologues of genes like *FLC* for example, present in 9 copies in *Brassica napus* [91].

The close evolutionary relationship to the model, together with a complicated polyploid genome structure as result of duplication, deletion and retention of genes makes *Brassica napus* an interesting model to study the effects of polyploidy on floral transition. Furthermore, flowering is a key phase for determination of yield in *Brassica napus*, hence, breeding efforts have aimed to time this crucial transition to maximise growth time while reducing exposure to adverse climate conditions [92]. Environmental conditions such as droughts [93] and winter warming [94] have been shown to negatively impact yeilds in this crop. With most *Brassica napus* growing areas set to face droughts during its flowering time as a result of climate change [95], it is more important than ever to understand how flowering is regulated within this crop to aid in development of new varieties to sustain production.

The following section provides an overview of our current knowledge of flowering time regulation in this crop.

### 1.2.1 Flowering control in *Brassia napus*

Beginning with the photoperiod pathway, just like Arabidopsis, *Brassica napus* is generally a long day flowering plant, though some accessions have been shown to flower at shorter daylengths, with a significant delay [96]. Orthologues of Arabidopsis *CO* in *Brassica napus* were among the first genes to be identified as regulators of flowering time in this crop [97]. With the advent of genomic resources, we now know there are six orthologues of *CO* and four orthologues of *CO-LIKE* genes [98]. While expression studies do not point to any substantial differences between these paralogues [99], our knowledge is still very limited.

Recent research into *SVP*, the key gene of the ambient temperature pathway, has identified four orthologues in *Brassica napus*, and demonstrated that mutants for these copies show accelerated flowering [100]. However, there are no reports on differences in regulation of these different copies. This pathway remains to be dissected within *Brassica napus*.

The vernalisation pathway on the other hand as been extensively studied, as the orthologues of the main vernalisation regulator *FLC* have been a focus of a lot of studies [92] [101]. *Brassica napus* has 9 orthologues of *FLC* in its genome [102], alongside another incomplete paralogue [103]. The *FLC* orthologues show differences in their expression and different copies have different effects on vernalisation, with the *FLC* orthologue on chromosome A10 as having the strongest effect [104]. Some *FLC* orthologues are not downregulated by cold, indicating likely regulatory divergences [103]. A study showed that the total gene expression of all paralogues correlates with vernalisation requirements of different types of cultivars of *Brassica napus*, indicating relaxation of selection pressure on individual orthologues, allowing for divergence in regulation [91]. Studies have also investigated the four identified orthologues of *VIN3*, all of which are upregulated during cold and downregulated on return to warmer conditions [103].

Our understanding of endogenous cues promoting flowering in *Brassica na-*

*pus* remains limited. There are no major reports of studies dissecting the autonomous pathway [92]. Interestingly, most flowering time genes affected by drought stress were found to be orthologues of Arabidopsis genes constituting the ageing and gibberellin pathways. This indicates that these pathways might be responsible for flowering regulation in response to abiotic stresses [105]. Likely due to the lack of miRNA gene annotations, there have not been investigations on miRNA156 and miRNA172 mediated control of flowering time [92].

*FT* is the only integrator gene that has been studied in some detail in *Brassica napus*. Six orthologues of *FT* have been indentified [106]. These orthologues show divergences in their cis-regulatory regions. Two paralogues have been reported to have lost the CArG box motif, the binding site for *FLC*, but contain binding site for *CO* [106] [107]. This could indicate a potential divergence in integration of photoperiod and vernalisation pathways. For orthologues of another integrator, *SOC1*, within *Brassica juncea*, there has been a report of similar differences in promoter regions of the six orthologues, with indications that it could be the case in *Brassica napus* as well [108].

In short, there is a lot still unknown about flowering time regulation in *Brassica napus*. Studies seek to exploit the close relationship with the model plant, Arabidopsis, however, gene duplications complicate knowledge transfer. There is limited data on expression, subfunctionalisation or protein stability for orthologues of most flowering time genes. Since the study of multiple paralogues of genes is a key cornerstone of *Brassica napus* research, the next section briefly reviews these key concepts related to these duplications, before introducing the contents of this thesis.

## 1.3 Polyploidy and the fate of genes

Transposition, tandem gene duplication and whole genome duplications are a few ways that can produce an expanded genetic repertoire — or duplicated genes [109]. Out of these, whole genome duplication is the phenomena that might lead to an organism possessing two or more complete set of chromosomes, termed ‘polyploidy’. Most, if not all, species carry signatures of ancient whole genome duplication events [110]. This change in ‘ploidy’ brings with it an increase in cell size, propensity for tissue differentiation and an effect on numerous physiological events of the cell, many of which are yet to be understood [111].

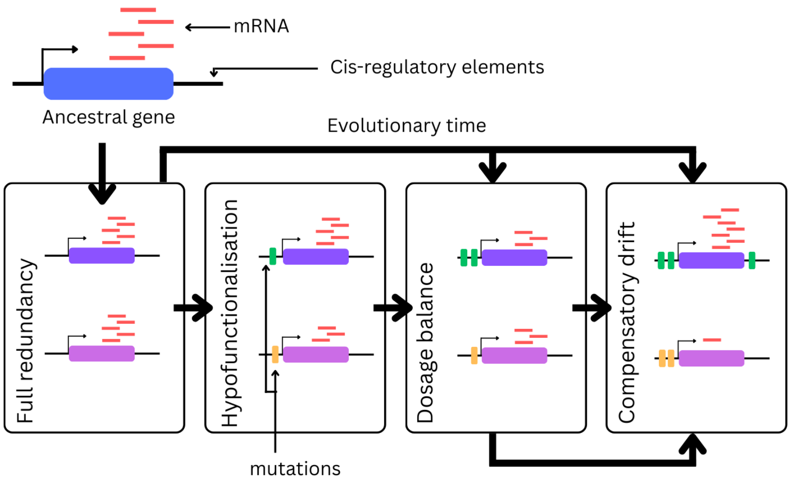


**Figure 1.4: Evolution of genetic redundancy among retained paralogues following duplication**

Immediately after duplication, paralogues are likely fully redundant. Following which, duplicates can drift to reduced expression levels, known as 'hypofunctionalisation'. Dosage sensitive genes can exist in under dosage balance, however, it is prone to decay over evolutionary time. These states are a result of mutations in the cis-regulatory elements for gene paralogues that continue proteins with similar biochemical properties. The duplicated genes are shown in two different colours. Modified from the original by Iohannes and Jackson [115].

The organism at the centre of this work, *Brassica napus*, is a polyploid and this thesis presents a study of a system with multiple duplicated genes, stably present and expressed within its genome. It has been posited that the ultimate fate of a paralogue, or duplicated gene within a genome, is to acquire deleterious mutations so that it becomes nonfunctional, a process termed 'nonfunctionalisation', or gather beneficial mutations that allow it to develop novel functions, a process termed 'neofunctionalisation' [112].

However, redundant genes are often retained in nature — sometimes even preferentially, as is the case with flowering time genes in *Brassica napus* [90]. Comparative analyses of gene retention in angiosperms has suggested that genes involved in regulation, signal transduction and metabolic processes are enriched in duplicates [113]. The key characteristic is that these genes are often involved in dosage dependent stoichiometric reactions. This leads to genes stabilising in 'dosage balance', and any deletions of paralogues could have negative fitness consequences. This leads to 'subfunctionalisation' [114].

Mutations within cis-regulatory elements are thought to be the key reason for the maintainence of the redundancy. As shown in Figure 1.4, paralogues immediately following a duplication event would likely be fully redundant. Over evolutionary time, genes reduce their expression levels, through accumulation of mutations in cis-regulatory elements and might exist in a 'hypofunctionalised' state where duplicates reduce their expression to levels insufficient for function. Overtime, genes may reach a state where they exist in 'dosage balance'. This dosage balance is prone to decay and can lead to 'compensatory drift', where one or more paralogues drifts to lower expression levels, while another to higher levels to compensate. [115].

The key result of paralogues being maintained in a subfunctionalised state is compensation. It could be in the form of full redundancy, where mutants of individual paralogues do not have any phenotype. However, it could also be in a partial or unequal compensation. For example, *AGAMOUS, SHATTERPROOF1, SHATTERPROOF2* and *SEEDSTICK* are flower development genes, acting downstream of floral transition that exhibit partial redundancy [116]. In partial redundancy, mutants in one gene exhibit a milder phenotype, that is enhanced in a double mutant. Within *Brassica napus*, orthologues of *SVP*, show partially redundant behaviour, as time to flowering is reduced further as more paralogues are disrupted [100]. Unequal redundancy however, results in a mutant in one gene exhibiting a phenotype while mutant in another does not. Mutations in the cis-regulatory regions are the cornerstone of maintainence of this subfunctionalised state. As an example, the floral integrator and meristem identity gene, *AP1* exhibits unequal redundancy with its duplicated MADS-box transcription factor copy, *CAULIFLOWER* [117]. Research has shown that following duplication, *AP1* has gained an extra CArG box site in its promoter region, allowing autoregulation and cross regulation by *CAL*, while *CAL* has lost some binding sites. This led to *AP1* emerging as the main orthologue maintaining the function of ancestral gene while *CAL* has undergone some functional divergence.

Gene duplications can serve as sources of novel variation or provide robustness to the system through compensatory mechanisms. The existence of paralogues, and these compensatory mechanisms, however, create challenges in functional studies, and breeding of crops. The reader of this thesis would encounter comparisons of expression of gene paralogues in the following chapters. The knowledge of which individual paralogues have diverged in their regulation or acquired new roles would provide useful information for applications in breeding and precision gene editing of *Brassica napus* [105].

## 1.4 The current work and its objectives

This final section will introduce you, dear reader, to the work presented in this thesis. I aim to exploit the close relationship between *Brassica napus* and the model plant *Arabidopsis thaliana* to translate knowledge from model to crop. Time to flowering is directly linked to yield and is arguably the most important agronomic trait in *Brassica napus*. Yet, as highlighted in the sections above, very little is known about the regulatory control of flowering transition in this crop. Additionally, *Brassica napus* is an interesting model to study the effect genome duplications and deletions have on the regulatory network of a fundamental process.

RNA-Sequencing (RNA-Seq) provides a powerful way to capture the total transcriptome at any given timepoint for a tissue. Performing RNA-Seq at multiple timepoints throughout plant development provides detailed data regarding function and regulation of genes. Chapter 2 as an extension of this introduction, presents a brief guide to RNA-Seq. As RNA-Seq is used extensively in this work, an introduction covers the basics of the technology and limitations that should be kept in mind while intrepreting the results presented in this work.

Chapter 3 presents the first detailed comparison of orthologues between the model plant *Arabidopsis thaliana* and *Brassica napus*. I use comparative transcriptomics to show that most genes follow similar expression dynamics over the course of plant development, leading up to flowering, in both species. However, significantly lower proportion of orthologues of flowering time genes show this similarity, compared to rest of the genome.

My investigations into this set of genes continue through the inference of Gene Regulatory Networks (GRNs) controlling flowering time *Brassica napus* in Chapter 4. I show that while the GRN follows similar principles of flowering time control already known in Arabidopsis, short-day, cold treatment for plants to undergo vernalisation changes the network topology. I further observe that orthologues of *SOC1*, a floral integrator introduced earlier, differ in their expression dynamics in response to this treatment and have differences in their regulation.

Chapter 5 explores these differenes in expression dynamics of *SOC1* orthologues further. I show that these differences are also present in a cultivar that does not require vernalisation to flower, and are in response to change in temperature. I investigate orthologues of known regulators of *SOC1* to find a plausible explanation for this behaviour.

The thesis closes with a discussion in Chapter 6 about the results and the limitations of this work, and recommendations for further investigations that could help improve our understanding of the regulatory control of flowering time in *Brassica napus*.

# Chapter 2

# Capturing the transcriptome

## Abstract

Measuring gene expression on a whole genome level has been revolutionised by RNA-Seq. First introduced in 2008, this technique involves isolation of RNA from the samples, library preparation to prepare the isolated RNA, sequencing of the libraries and analysis of the data using bioinformatics tools. Multiple tools are often combined to make processing pipelines. RNA-Seq pipelines often differ depending upon the experimental aims, type of sample, sequencing platform being used and the bioinformatics tools and parameters used to process data. This thesis presents multiple RNA-Seq timeseries datasets to study the dynamics of flowering time genes in *Brassica napus*, hence, this chapter aims to provide a general overview of the technique and the various steps involved in obtaining gene expression values from sampled tissues. Limitations and sources of bias that can lead to errors in output from RNA-Seq experiments are also briefly discussed.

## 2.1 Introducing RNA-Seq

Next generation sequencing, introduced at the tail end of the 20th century, enabled the sequencing of the nucleotide makeup of an organism's DNA molecule on massive parallel scale [118]. This revolutionised molecular biology allowing full length genome sequences to be elucidated. However, there still remained a significant challenge: which regions of the genome are actually expressed?

One early approach that tackled this question was expressed sequence tag (EST) or cDNA sequencing [119]. However, it struggled with genes expressed at low expression levels and didn't identify 3' and 5' ends of the coding sequencing with high confidence [120]. DNA microarrays were another attempt, and while they were a great tool to identify low expressed sequences, they couldn't distinguish between similar sequences. Moreover, these methodologies could not identify the 3' and 5' boundaries of exons [121] [122]. Despite their limitations, they were widely adopted by researchers, leading to discoveries of genes involved in various developmental processes, both in Arabidopsis [123] and other crop plants [124] [125].

A wave of publications in 2008 introduced a new technique to tackle this problem. It was termed 'Short Quantitative Random RNA Libraries' or SQRL by one of those papers [126]. While quite descriptive, their nomenclature did not stick. Another paper, published just a few days before, described the method for a sequencing-based approach for global transcriptome mapping as RNA sequencing or 'RNA-Seq' for short [120]. The authors presented a novel approach that involved generating cDNA and subjecting it to Illumina sequencing to annotate exons, 5' and 3' boundaries of the genes, introns and quantify gene expression levels.

Their method consisted of a molecular biology pipeline in which Poly- (A) RNA isolated from yeast cells was used to generate double-stranded cDNA using reverse transcription. The double-stranded cDNA was subjected to fragmentation and sequenced using Illumina sequencers to generate sequences of 35 base pairs in length from either end of each fragment. The sequencing data generated from this step was used as an input for an informatics pipeline. The authors used SOAP [127], an alignment tool to map reads to the yeast genome. 56 % of the total number of reads were mapped at unique regions within the genome, and used to determine a high resolution transcriptome map for yeast.

## 2.2 The RNA-Seq workflow

While RNA-Seq has come a long way since its first introduction in 2008, the workflow still follows the structure introduced in the first publications. A general overview of RNA-Seq process for quantification of gene expression from tissue sample is shown in Figure 2.1.

### Before sequencing

The first step in RNA-Seq workflow is to extract RNA from tissue sample. While the exact protocol depends on the type of tissue and if it is fresh or frozen, all extraction procedures begin with grinding the tissue samples to a fine powder [128] [120]. This mechanical grinding is followed by enzymatic treatments to disrupt the tissue and break down cells to release the RNA. RNA is highly unstable and prone to degradation, hence, proper care needs to be taken for inactivation of RNases [129]. Following the release of RNA, it needs separation from DNA, proteins and other substances suspended in the solution after cell lysis. One method to achieve this is by using phenol/chloroform method, where homogenized solution is phase separated using chloroform. DNA and proteins are separated from RNA which remains in aqueous phase. RNA is precipitated using isopropanol from the solution [130]. Another method, used in commercial RNA-extraction kits uses silica membranes to which RNA binds in presence of chaotropic agents. Washing steps remove contaminants such as proteins and DNA and pure RNA is eluted in RNase-free water [131]. Following extraction, gel electrophoresis, TapeStation or spectrophotometry can be used to check the purity and integrity of the extracted RNA [128].

After isolation, the RNA molecules need to be converted into a form compatible with sequencers. Ribosomal RNA (rRNA) consitutes the majority of total RNA extracted, hence its removal is necessary to enrich the sample in messenger RNA (mRNA) content, which constitutes the RNA that is translated into proteins [132]. Two main strategies are used for mRNA enrichment. Poly- (A) selection exploits the fact that eukaryotic mRNA is polyadenylated following transcription. This method uses oligo-dT beads that bind to poly-A tails. Another strategy is to capture rRNA using complementary oligonucleotides [132]. Following mRNA enrichment, the sequences are fragmented to generate suitable insert sizes for a specific sequencing platform. This can be achieved using

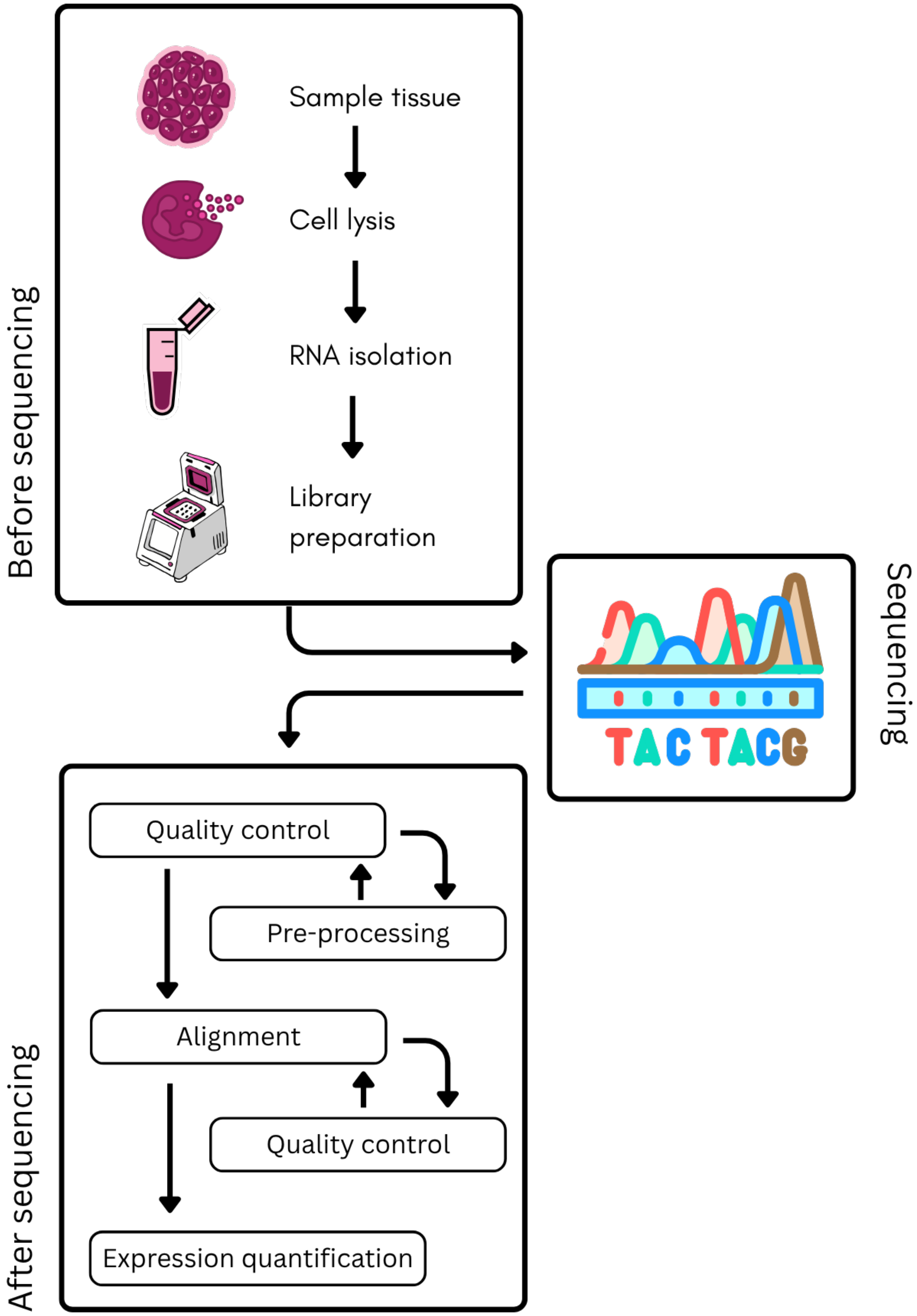


**Figure 2.1: The RNA-Seq workflow for quantification of gene expression**
RNA-Seq workflow can be divided into three major steps, before sequencing, isolated RNA is converted to libraries for sequencing and the data generated is processing using a bioinformatics pipeline.

enzymatic or physical methods like sonication. The fragments are then reverse transcribed into complementary DNA (cDNA) using reverse transcriptase. To these cDNA ends, adapters are ligated which contain binding sites for sequencing primers, barcodes for multiplexing and flow cell binding sites for different sequencing platforms. The adapter-ligated cDNA is amplified by PCR to generate enough concentration required by the sequencing technology [133]. Following QC, libraries are pooled and loaded onto the sequencing platform.

### Sequencing

The choice of sequencing platform is dictated by the aims of the experiment. Illumina offers short-read sequencing platforms such as NovaSeq, NextSeq and MiSeq. These platforms use the sequencing-by-synthesis (SBS) method. While the details of the sequencing chemistry are often proprietary, most SBS technologies use a method where cDNA molecules are distributed to millions of wells or chambers on a solid substrate. These molecules are then subjected to synthesis reactions, in which nucleotides, which are labelled or are incorporated into a chemical reaction for identification, are added sequentially and imaged or detected. These reactions can be massively parallelised to include millions of DNA sequences in one run. Illumina sequencers have low error rates and output short reads up to 150 base pairs in length. This sequencing platform is suitable for gene expression quantification [134] [135].

Long read sequencing platforms are available from Oxford Nanopore Technologies, such as MinION, GridION and PromethION. These platforms perform sequencing by detecting the changes in electrical current as different nucleic acids pass through a nanopore. These methods produce ultra-long reads and enable full length sequencing of transcripts. This allows for precise isoform identification. Isoforms are alternative forms of mRNA originating from the same gene, due to alternative splicing, alternative promoter usage and other post-transcriptional modifications. While these methods have lower throughput and accuracy compared to short-read sequences, they directly sequence RNA molecules, without reverse transcription or amplification. They are also portable, allowing rapid sequencing in field conditions [136] [137].

### After sequencing

The bioinformatics pipeline is used to process the data obtained following sequencing. The sequencing data is in the form of nucleotide 'reads' that need quality control to separate information from noise [138]. Programs such as FastQC [139] allow for quality assesment of raw sequencing data. This initial quality assesment checks for multiple different metrics. Adapter sequences, which are ligated to fragments during sequencing need to be removed before downstream analyses. Overrepresented sequences, duplication levels and the amount of GC content checks can flag any biases during library preparation or sequencing steps. Reads are also checked for sequencing quality throughout their length. Cutadapt [140], Trimmomatic [141] or fastp [142] can be used to remove adapter sequences, short or low quality reads and overrepresented or duplicated sequences from the data. Following quality control, reads can either be aligned to a reference genome or transcriptome or used for assembly of a new transcriptome. Aligners are broadly based on either the Burrows-Wheeler transform [143], such as Bowtie [144] and Burrows-Wheeler aligner (BWA) [145]; or on Needleman-Wunsch or Smith-Waterman algorithms, such as SHRiMP [146] or GNUMAP [147]. RNA-Seq reads span exon-exon junctions, hence require alignment algorithms that can perform splice-aware alignments. HISAT2, based on Burrows-Wheeler Transform for graphs [148] and Pass, based on Needleman-Wunsch and Smith-Waterman algorithms [149] are examples of such aligners. HISAT2 and Pass are de-novo splice aligners, i.e. no prior annotation within the genome is required, and they can be used to detect new splice junctions. Since HISAT2 is the latest available de-novo aligner, that has been successfully used for alignment of short-read RNA-Seq data to reference genomes of species in the *Brassicaceae* family [150], including *Brassica napus* [91], I selected it for the pipeline used in this thesis. Further details are in Chapter 3.

Following alignment, if gene expression quantification is needed, programs such as featurecounts [151], or StringTie [152] can be used to count the number of reads mapped to a particular gene. The amount of reads mapped to each gene can be represented as raw read counts — however, they are often affected by the library size and length of the gene. Hence, normalised gene expression metrics, such as FPKM (Fragments Per Kilobase of exon per Million mapped reads) or TPM (Transcripts per Million mapped reads) are used to normalise for gene length and the total number of reads. While being quite similar, TPM and

FPKM differ in their order of operations. While calculating FPKM, total reads in a sample are divided by 1,000,000 to create a scaling factor and read count attributed to each gene is divided by that scaling factor, this gives 'reads per million'. This 'reads per million' quantity is then divided by the length of the gene in kilobases. For TPM, the read counts are normalised by gene length first before scaling by the scaling factor. This results in the sum of total TPMs to be the same between samples, while this is not true for FPKM [153] [154].

After quantification of reads mapped to a gene, downstream analyses such as comparing gene expression, searching for 'differentially' expressed genes, network analyses and other functional genomics analyses are carried out [138]. However, while performing these analyses, one must be aware of different errors that can be introduced in reads generated in an RNA-Seq experiment. The following section provides a brief introduction to these errors.

## 2.3 Sources of errors in RNA-Seq data

RNA-Seq has a complicated workflow, with many possible sources of bias that can lead to erroneous results in downstream analyses and incorrect interpretations and conclusions from a sequencing experiment.

The choice of sample preservation and RNA extraction methods are the first point where biases can be introduced. Freeze storage in liquid nitrogen or -80 °C is considered standard practice. It has been shown that sample preservation in non-standard mediums, such as formalin-fixed and paraffin-embedded (FFPE), which is done for archival samples, makes nucleotide extraction difficult and leads to poor sequencing libraries [155] [156]. Research has also shown that longer sample processing times, multiple freezing and thawing cycles can degrade the quality of extracted RNA [132]. RNA degrading enzymes (RNases) pose another challenge to extraction of high quality RNA during the isolation and process, which directly affects the quality of RNA libraries [157].

Following extraction, protocols used for library preparation can also introduce deviations. Library preparation techniques that enrich poly- (A) RNA transcripts using primers can induce a 3'-end capture bias [132]. This technique, while is able to enrich the library in mRNA as most eukaryotic mRNA and long non-coding RNAs have a poly A tail [158], it removes all non-poly- (A) RNAs, which includes bacterial mRNA and various long non-coding RNAs. Samples can also be enriched for mRNA by depletion of rRNA. However, it uses the exact sequence content of rRNAs, so only works for species where rRNA complementary probe kits are available commercially [159].

PCR amplification, which is a key step in libary preparation is among the main sources of artefacts in RNAseq. Fragments that are GC-neutral are often preferentially amplified compared to GC-rich or AT-rich sequences. This is especially problematic for organisms with such genomes, such as the AT-rich human malaria parasite [160]. While researchers have proposed workarounds for this issue [161], there are amplification-free RNA library preparation methods, that are useful for samples with acceptable quantity of RNA content [162]. The number of PCR cycles can also produce significant biases, and it has been recommended to perform the amplification steps in as few cycle numbers as possible [163].

Differences in sequencing machines can also induce deviations in the data obtained from RNA-Seq. For instance, Illumina HiSeq platform is prone to substitution bias [164], while single-molecule sequencing platforms such as PacBio single-molecule real-time (SMRT) have high error rates compared to Illumina platforms [165].

The bioinformatics pipelines in RNA-Seq workflows are not immune to biases either. The results between analyses can be affected by choice of different alignment algorithms, normalisation techniques and choice of methods [138]. While RNA-Seq has a few sources of limitations and there is no ‘one size fits all’ approach to analysing RNA-Seq data, proper quality checks for raw reads, alignment and downstream analyses can ensure that any conclusions or inferences drawn from the datasets are not artefacts resulting from these deviations. The methods sections of the following chapters highlight the steps taken to ensure accuracy of results.

# Chapter 3

# On comparison of transcriptome dynamics between *Arabidopsis* and *Brassica napus*

## 3.1 Abstract

*Brassica napus* is among the closest crop relatives to the model plant *Arabidopsis thaliana*. This shared lineage provides an avenue to transfer knowledge from model to crop. Here, using comparative transcriptomics, I show that majority of genes in *Brassica napus* have similar expression dynamics to their Arabidopsis orthologues throughout the plant's development, highlighting the likely conservation of regulatory frameworks between these two species. However, the set of orthologues of Arabidopsis flowering time genes in *Brassica napus* show a higher degree of divergence in their expression dynamics than the rest of the genome. Previous research has shown that flowering time genes are preferentially retained within the *Brassica napus* genome. I further show that this preferential retention in higher copy numbers is correlated with an increased chance of paralogues diverging in their expression patterns.

## 3.2 From Arabidopsis to *Brassica napus*

Arabidopsis and *Brassica napus* go through similar developmental stages during their life cycle [166] [167]. Following germination, they have a period of vegetative growth before the plants undergo the floral transition. Their floral morphologies are also very similar [168]. Due to these shared characteristics and close *Arabidopsis-Brassica* lineage [88], a lot of research in cultivated Brassicas involves comparisons with the model plant, *Arabidopsis thaliana*. This is especially true for *Brassica napus* [98] [97] [102].

This knowledge transfer, however, is complicated due to presence of multiple orthologues of each Arabidopsis gene within the *Brassia napus* genome. *Brassica napus* is an allotetraploid, formed by hybridisation of two diploid progenitors, *Brassica oleracea* and *Brassica rapa*, whose genomes underwent whole genome triplication events. This means that, theoretically, it should have six different orthologues for each Arabidopsis gene. However, research involving comparisons of total gene numbers has revealed that there is only a four-fold difference between *Arabidopsis thaliana* (25,498 genes) and *Brassica napus* (101,040 genes) [86]. This indicates that on a whole genome level, there has been a loss of paralogues — which is the expected outcome following whole genome duplications [115]. This is according to the dogma posited by the theory of duplicate evolution, which states that ultimate fate of paralogous genes is to either accumulate deleterious mutations to become non-functional or acquire mutations to develop novel functions [112].

It is interesting to note that, despite the gene loss on a whole genome level, certain sets of genes could be preferentially retained. For example, in *Brassica rapa*, one of the diploid progenitors of *Brassica napus*, circadian clock genes have been shown to be preferentially retained following genome duplications [169]. For a number of genes involved in the regulation of flowering time in *Brassica napus*, the number of paralogues exceeds the theoretical six. [98] [91] [92]. In fact, flowering time genes in *Brassica napus* have been shown to be preferentially retained compared to rest of the genome [90].

Preferential retention could lead to divergences in expression dynamics. For example, the vernalisation pathway gene, *FLOWERING LOCUS C* (*FLC*) has at least nine orthologues in *Brassica napus* (along with another possible truncated orthologue) [103]. In Arabidopsis, *FLC* is a floral repressor and its expression level decreases as the plant undergoes vernalisation in winter conditions [40]. However, in *Brassica napus*, different paralogues respond differently to vernalisation and total *FLC* dynamics explain the vernalisation requirements for a number of cultivars [91]. Gene dosage hypothesis can provide an explanation for its occurrence [170]. Gene products need to be present in appropriate stoichiometric ratios for efficient function and this balance can be achieved even when paralogues drift in their expression levels — as long as the appropriate protein or mRNA concentrations are maintained on a total level. This leads to selection pressure on total gene product and allows room for divergence for individual paralogues while ensuring their retention within the genome.

This divergence could be a more widespread occurrence among genes with retained paralogues, however, no investigations have been done to compare the dynamics of orthologues between Arabidopsis and *Brassica napus*. While flowering genes have been preferentially retained, it is unknown if they still share the same dynamics with their Arabidopsis orthologues. This comparison of dynamics is complicated by the fact that while these two plants develop through similar morphological states, they develop at very different rates. Hence, the comparison requires using an algorithm that can allow for comparison across timescales.

The following section briefly introduces curve registration, a technique that facilitates that comparison.

### 3.2.1 Introduction to Curve registration

Curve registration is a method to align two individual curves, where one curve serves as the 'reference' while the other, termed as 'query', is transformed such that it maximally coincides with the reference [171]. Figure 3.1 shows the curve registration method as implemented in greatR [172]. greatR uses two parameters, 'stretch' and 'shift' to achieve the transformation of the query curve. It then uses Bayesian model comparison to test the hypotheses that curves can be explained by one underlying model or whether two different models best explain the data.

Mathematically, let *query* ($q$) data be collected at $t_{q,i}$ timepoints and $y_{q,i}$ be the corresponding value at timepoint $i$, where $1 \leq i \leq N_q$, with $N_q$ being the total number of timepoints. Hence, *query* dataset can be denoted as,

$$q = (t_{q,i}, y_{q,i}), 1 \leq i \leq N_q \tag{3.1}$$

Similarly, the *reference* ($r$) dataset, sampled for $N_r$ timepoints, can be denoted as,

$$r = (t_{r,j}, y_{r,j}), 1 \leq j \leq N_r \tag{3.2}$$

greatR fits a cubic B-spline model [173] to the data, with the models for datasets $r$ and $q$ being denoted by $m_r(t)$ and $m_q(t)$, respectively. If two curves are similar, then one common model $m_1$ can explain both datasets, however, if the two datasets are not similar, they would be best explained by two different models. This evaluation occurs after transformation of the $q$ dataset using function,

$$h(t_{q,i}) = \beta_1 + \beta_2 t_{q,i} \tag{3.3}$$

where, $\beta_1$ is the 'shift' in time, and $\beta_2$ is the 'stretch' in timescale between the two datasets.

The two hypotheses hence are:

*H1*: the datasets are best explained by one common model, $m_1$

*H2*: the datasets are best explained by two different models, $m_r$ and $m_q$

These hypotheses are evaluated using the Bayesian Information Criterion,

$$\text{BIC} = -2\mathscr{L} + k \log N_D \tag{3.4}$$

where, $\mathscr{L}$ is the maximum log likelihood (following optimisation), $k$ is the number of parameters and $N_D$ are the parameters for the corresponding models, $m_1$ (*H1*) or $m_q$ and $m_r$ (*H2*). If

$$\text{BIC } (H1) < \text{BIC } (H2), \tag{3.5}$$

then the curves are considered 'registered'. In the context of the data analysis presented in this chapter, the curves correspond to expression dynamics of genes. Hence, if two genes in two different timeseries, are following the same dynamics, they will be determined as 'registered', while if they are not, they will be labelled as 'not registered'.

This framework allows for comparison of dynamics between a gene and its

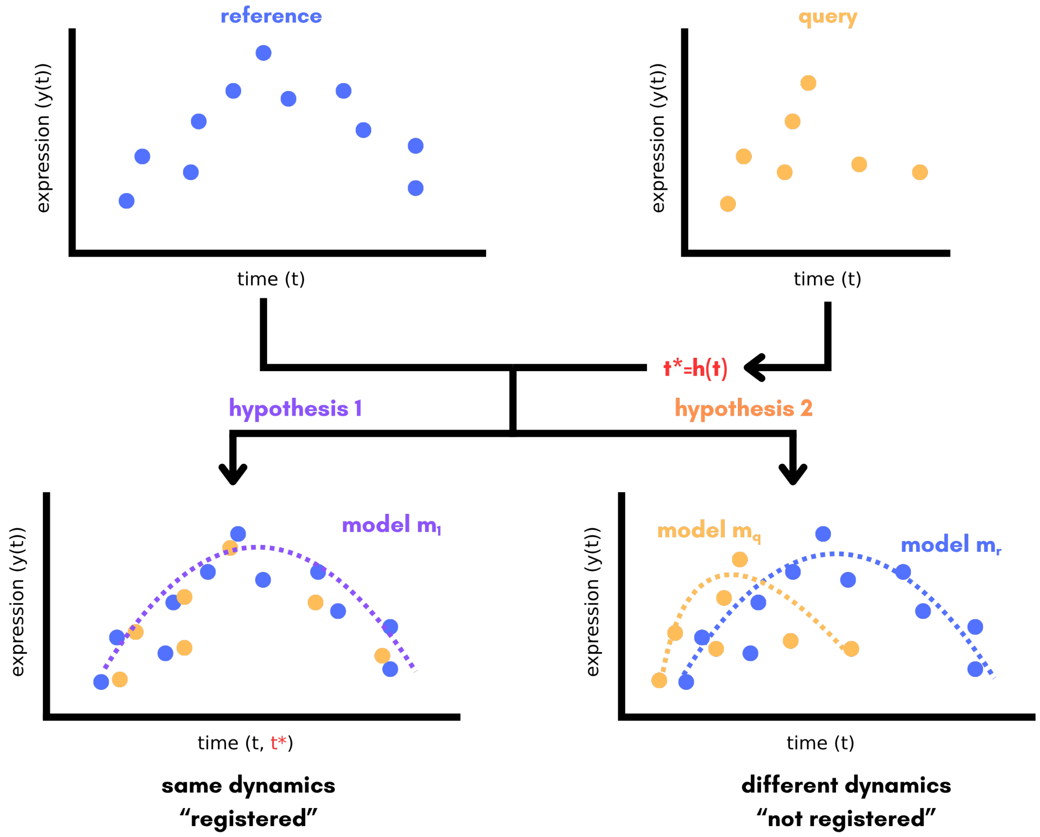


**Figure 3.1: Curve registration method used in greatR**

greatR compares two curves, reference and query. Following transformation ($h(t)$) of the query curve, it statistically tests hypotheses for a single model ($m_1$) or two different models ($m_q$, $m_r$). Figure modified from [172] orthologue, even in plants that develop on different timescales. Hence, opening up the possibility of a detailed comparison of dynamics of genes between the model plant, *Arabidopsis thaliana* and polyploid crop, *Brassica napus*.

The following section details the data collection and techniques used to facilitate a detailed comparison of gene expression dynamics between these two species, with curve registration at the core of the analysis.

## 3.3 Methods

### 3.3.1 Plant material

The data presented in this chapter is collected from two *Brassica napus* cultivars, Stellar and Zhongshuang 11. Stellar is a spring-type cultivar, that does not require vernalisation to undergo floral transition. Zhongshuang 11, abbreviated as 'ZS11', is a Chinese semi-winter-type cultivar that requires vernalisation prior to floral transition.

### 3.3.2 Sampling

Stellar and Zhongshuang 11 (ZS11) plants were sown in cereals mix (40% medium grade peat, 40% sterilised soil, 20% horticultural grit, 1.2 kg/m$^3$ PG mix 14–16–18 + Te base fertilizer, 3 kg/m$^3$ maglime, 300 g/m$^3$ Exemptor). Following germination, each seedling was transplanted to a 5 cm x 5 cm x 4.5 cm cell in a standard 24 cell-tray. Plants were germinated and grown in a Conviron MTPS 144 controlled environment room with Valoya NS1 LED lighting (250 $\mu$mol/m$^2$s) with a 16-hour photoperiod. Temperatures were set at 18 °C during the day and 15 °C during the night, with relative humidity maintained at 70 %. At day 21, ZS11 plants were shifted to 5 °C, 8-hour photoperiod conditions for plants to undergo vernalisation. They were transferred back to normal conditions after 3 weeks at day 42. Stellar does not require vernalisation, hence Stellar plants were not subjected to these conditions.

Each sampled timepoint for cultivars is shown in Figure 3.2. The aim of the sampling was to capture the gene expression leading up to floral transition, hence sampling timepoints are concentrated on days prior to floral transition. Similar RNA-Seq timeseries in the shoot apical meristem leading upto floral transition for the Arabidopsis Col-0 ecotype, was downloaded from NCBI Sequence

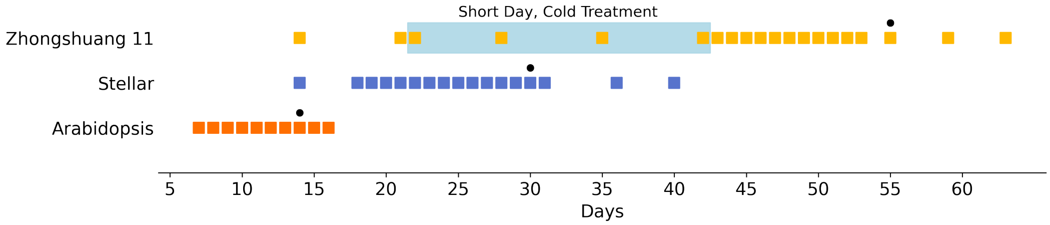


**Figure 3.2: Sampled timepoints for *B. napus* cultivars Stellar, Zhongshuang 11 (ZS11) and Arabidopsis (Col-0)**

'•' indicates the time of floral transition. Shoot apical meristem was sampled for RNA-Seq at the indicated timepoints. 'Days' refers to days post sowing. Arabidopsis data is from [174]

read archive [Project ID: PRJNA268115] [174]. For Stellar timeseries, at each sampled timepoint, three replicates were sampled with three dissected shoot apical meristems pooled for each replicate. The same sampling strategy was employed for ZS11 timeseries, except for timepoint 53, which only has one replicate. Following dissection, samples were immediately frozen in liquid nitrogen. This sampling and subsequent sequencing was done as part of the Biotechnology and Biological Sciences Research Council (BBSRC) grant 'Brassica rapeseed and vegetable optimisation (BRAVO)' (BB/P003095/1).

### 3.3.3 RNA sequencing

Samples frozen in liquid nitrogen were ground to a fine powder for RNA extraction. Manufacturer's instructions were followed for RNA extraction using EZNA ® Plant RNA Kit (Omega Bio-tek Inc.). RNA samples were processed at Novogene, with library preparation using NEB next ultra directional library kit (New England Biolabs), and sequenced using Illumina HiSeq X for ZS11 timeseries and NovaSeq 6000 for Stellar.

### 3.3.4 RNA-Seq data processing

Figure 3.3 shows the RNA-Seq pipeline used for processing of the RNA-Seq data for both Stellar and ZS11 timecourse datasets. The Col-0 dataset, obtained from Klepikova *et al*. [174] was also processed using the same pipeline.

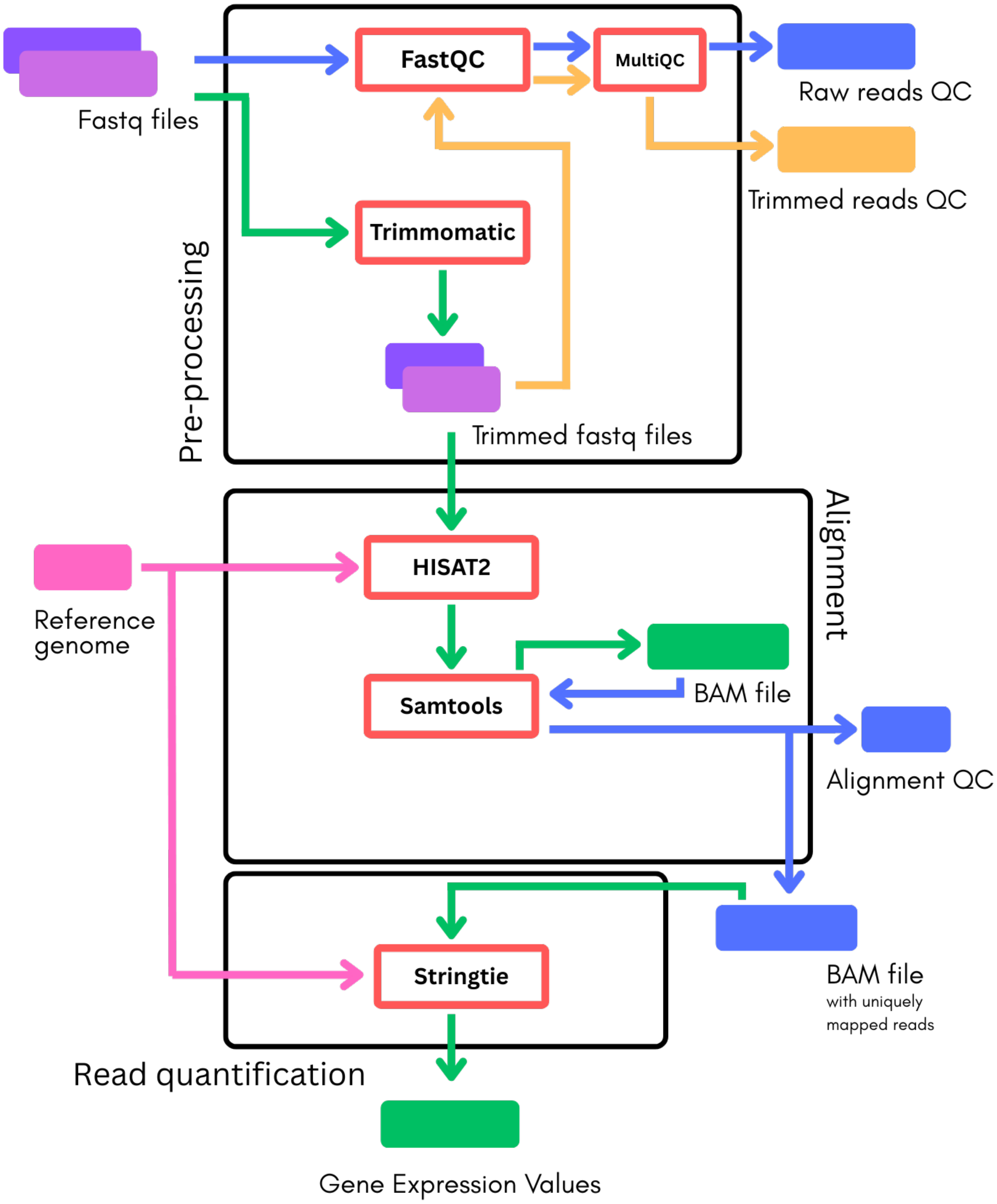


**Figure 3.3: Overview of the pipeline used for processing of paired-end short read RNA-Seq data**

The processing pipleline used for pre-processing, QC, alignment and quantification of mapped reads of the RNA-Seq data obtained following sequencing of the samples. The input is a pair of fastq files (paired-end sequencing data) and the key output is a comma-separated values (csv) file of expression values for genes within the reference genome.

### Pre-processing

FastQC (version 0.11.9) [139] was used for initial quality control of the sequencing data. MultiQC (version 1.29) [175] was used to collate QC data for the whole timeseries. Trimmomatic (version 0.39) [141] was used to remove adapter contamination from reads, with the following flags, 'ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:10:1:true' to remove adapters, 'HEADCROP:15' to trim the first 15 bases of the reads, 'SLIDINGWINDOW:4:15' to trim reads if the average quality within a window of 4 bases falls below 15 and 'MINLEN:50' to drop any reads below 50 base pairs in length.

### Alignment

The trimmed sequencing data for Stellar and ZS11 was aligned to the Darmor v10 reference genome [176]. HISAT2 (version 2.1.0) [148] was used for alignment, with 'RF' strandedness, due to library preparation protocol used by Novogene. Samtools (version 1.9) [177] was used for sorting, indexing and generating a QC report. The resultant BAM files were filtered to only keep uniquely mapped reads.

### Read quantification

StringTie (version 2.1.1) [152] was used for quantification of reads aligned to different genes in the genome. 'Transcripts per million mapped reads' or TPM was selected as the measure for gene expression [178].

## 3.3.5 Gene mappings using reciprocal BLAST

A custom reciprocal BLAST [179] script was used to determine orthologue mappings between Arabidopsis TAIR v10 annotation [180] and *B. napus* Darmor v10 genome [176]. A blastn search was performed using the 'blastn algorithm' from BLAST+ suite of tools [179]. To generate the 'one-to-many' mapping from query genome (here, TAIR v10) to target genome (here, Darmor v10) follows the method outlined in Algorithm 1.

Using this method, 67.88 % of Arabidopsis genes were found to have atleast one orthologue in *Brassica napus*. The mappings were manually referenced against known orthologues for key genes for verification.

**Algorithm 1** ‘One-to-many’ reciprocal blast

```
for all gene $\in$ QueryGenes do
    Hits $\leftarrow$ GetHitsBelowEvalue (gene, TargetGenome, $1e{-}5$)
    for all hit $\in$ Hits do
        BestHit $\leftarrow$ RunBLAST (hit, QueryGenome)
        if BestHit == gene then
            AssignMapping (gene, hit)
        end if
    end for
end for
```

### 3.3.6 Determining orthologues of Arabidopsis flowering time genes and transcription factors

Algorithm 1 was used to search for orthologues for sets of Arabidopsis genes in *Brassica napus* Darmor-v10 reference. The list of 306 flowering time genes was obtained from the FLOR-ID database [50]. Similarly, the list of Arabidopsis transcription factors was obtained from PlantTFDB [181]

### 3.3.7 Curve registration

greatR (version 2.0.0.9000) [172] was used to perform curve registration. This novel R-package was developed by Ruth Kristianingsih improving upon the first use of curve registration to study developmental progression between Arabidopsis and *Brassica rapa* [150]. I contributed to the testing of this improved method and the R-package. For both Stellar and ZS11 registrations against the Col-0 timeseries, Col-0 was used as the query accession. The z-score scaling method was used and overlapping percentage between curves set to 75. The rest of the parameters were set to default, with L-BFGS as the optimiser.

### 3.3.8 Distance between two timecourses

For two timeseries datasets, let $y$ and $z$ be the vectors of expression values for a set of $N$ genes at timepoints $t_i$ and $t_j$. The pairwise distance between two timepoints can be defined as,

$$d(t_i, t_j) = \left[ \sum_{k=1}^{N} (y_k - z_k)^2 \right]^{\frac{1}{2}} \tag{3.6}$$

This calculation can be performed for all possible pairs of timepoints between two timeseries and represented as a 'heatmap' to visualise distance between two timecourses (Results and Discussion; Figure 3.10). The greatR function 'calculate_distance' was used for this calculation. For plotting distances between timeseries following registration, 'match_timepoints' flag was set to 'True' (i.e. $t_i = t_j$).

## 3.4 Results and Discussion

### 3.4.1 RNA-Seq Alignment

**Pre-processing raw reads improves data quality for alignment**

An overview of the various FastQC checks is represented as a heatmap in Figures 3.4 and 3.5. Heatmaps (a) in both figures for Stellar and ZS11 show that the data had some adapter content and overrepresented sequences within the raw sequence files. The report showed a warning for sequence duplication for all samples, however, that is to be expected given the duplicated nature of the *Brassica napus* genome. Sequence files in both datasets also had warnings for the 'Per Base Sequence Content' flag. Figure 3.6 shows an illustrative example from a sample from ZS11 timeseries. The imbalance in proportion of base calls, shown as lines that ideally should run parallel in a large random library, at the beginning of the reads. These errors were corrected by trimming the reads using parameters defined in methods.

Heatmaps (a) in both figures show that trimming did correct for 'Per Base Sequence Content', 'Per sequence GC content', 'Overrepresented sequences' and 'Adapter content flags'. The trimmed sequences, however, have a warning for 'Sequence Length Distribution' — which occurs due to difference in the number of bases trimmed from the start of the reads. The read length differences however, are in the range of 0 to 1 bases, hence would not affect downstream alignment. Similarly, the warnings in 'Per Tile Sequence Quality' flag any deviations in quality of base calls from a flowcell. These deviations are confined to limited instances within the data, as shown in some examples in Figure 3.7 and would not affect downstream alignment.

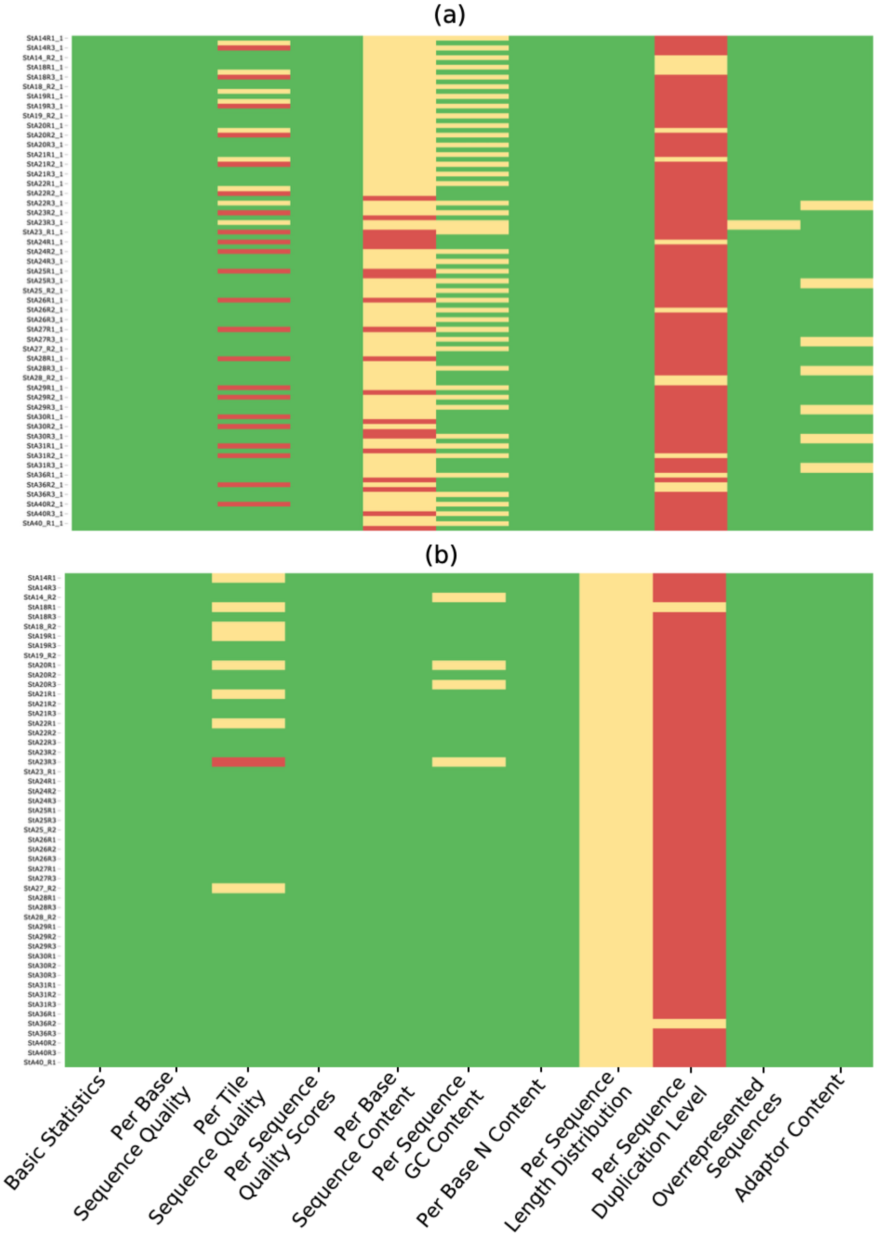


**Figure 3.4: An overview of various QC checks on raw and trimmed RNA-Seq files for Stellar timeseries**

(a) and (b) show a heatmap of QC checks for sequence files from the Stellar timeseries, before and after initial trimming respectively. FastQC shows warnings for sequence duplication across all samples, however, it is expected given the duplicated nature of *B. napus* genome. 'Sequence length distribution' warning occurs post trimming because of variation in number of bases trimmed from start of different reads (in range of 0 to 1 bases), while initially all reads are 150bp in length. Remaining 'Per Tile Sequence Quality' warnings left in samples in (b) were investigated individually. These did not represent any major issues that would impact the quality of downstream alignment.

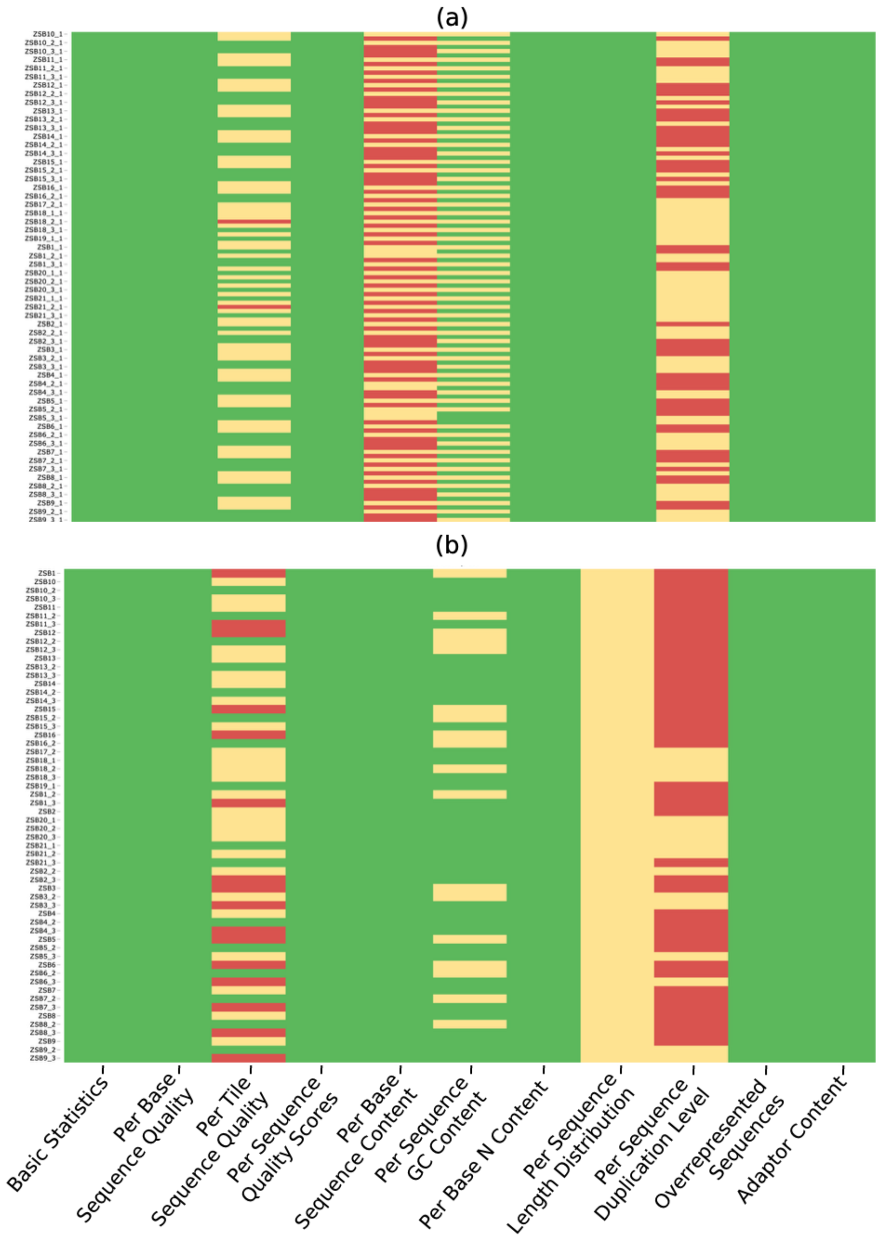
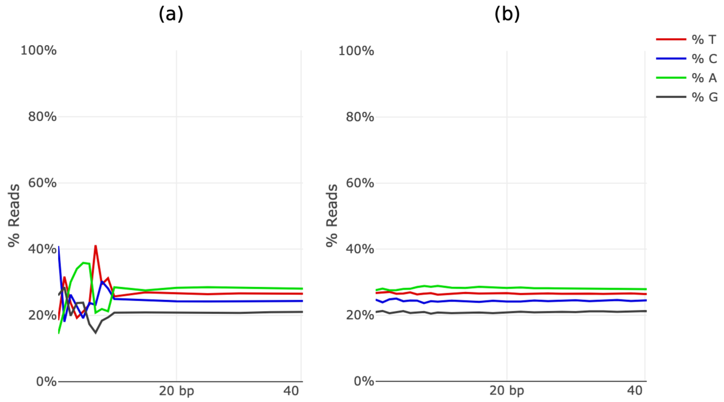


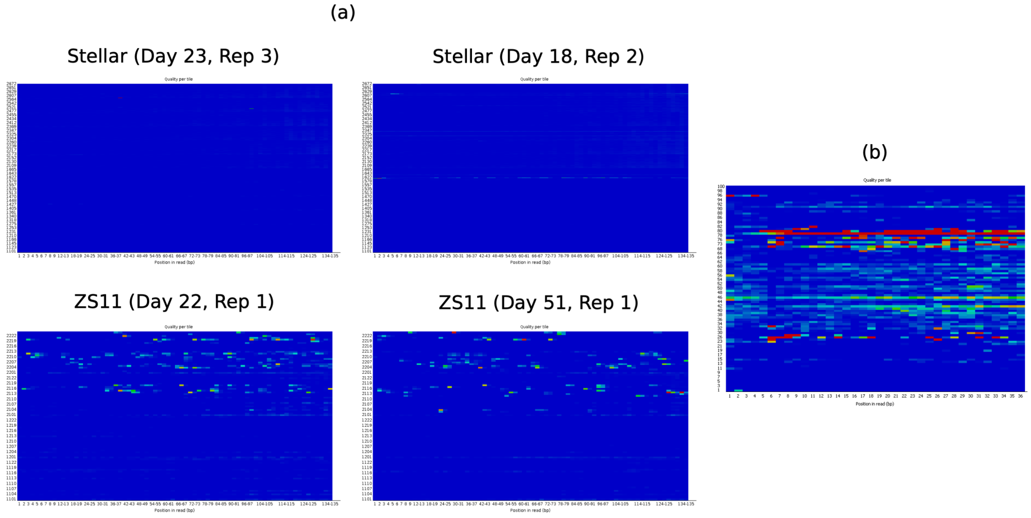
**Figure 3.5: An overview of various QC checks on raw and trimmed RNA-Seq files for ZS11 timeseries**

(a) and (b) show a heatmap of QC checks for sequence files from the ZS11 timeseries, before and after initial trimming respectively. FastQC shows warnings for sequence duplication across all samples, however, it is expected given the duplicated nature of *B. napus* genome. 'Sequence length distribution' warning occurs post trimming because of variation in number of bases trimmed from start of different reads (in range of 0 to 1 bases), while initially all reads are 150bp in length. Remaining 'Per Tile Sequence Quality' warnings left in samples in (b) were investigated individually. These did not represent any major issues that would impact the quality of downstream alignment.



### Figure 3.6: Example of a sample with 'Per Base Sequence Content' warning from the ZS11 timeseries

The lines in the plots show the proportion of each base position at the start of the reads in a sequence file for which each of the four normal bases were called. In a random library, there is little to no difference between proportions of different bases and the lines are expected to be roughly parallel. (a) Deviations are present at the start of the reads in this sample. These deviations can be a result of different library preparation techniques. (b) Trimming few bases from the start of reads leads to correct proportions for downstream alignment.



### Figure 3.7: Examples of samples with Per Tile Sequence Quality warning from both Stellar and ZS11 timeseries

The plot shows deviations from the average quality of each tile. The colours are from a cold to hot scale, with hotter colours indicating worse than average quality for that tile, while cooler colours indicate the opposite. (a) Samples from Stellar and ZS11 timeseries, showing that 'Per Tile Sequence Quality' is not consistently bad across any tile, indicated by very low number of red tiles. (b) Plot from FastQC documentation, that shows an example when 'Per Tile Sequence Quality' would be concern for quality.

**HISAT2 is able to map the majority of reads uniquely to the Darmor v10 reference**

Tables 3.1 and 3.2 show the alignment summary statistics for both Stellar and ZS11 timeseries. The average number of raw reads per sample were similar in both Stellar and ZS11 at 29,464,770 and 29,871,173 reads respectively. ZS11 timeseries had a slightly higher average alignment rate at 91.21 % compared to 89.99 % for Stellar and the alignment rate is consistent across all the samples. For both cultivars, majority of reads were uniquely mapped, at 73.6 % and 86.10 % for Stellar and ZS11 respectively. This shows that the aligner is able to map pairs of reads with high confidence across the genome, and even by filtering for uniquely mapped read pairs, the majority of the information is retained, and we can obtain a high confidence measure for gene expression in both timeseries.

Figure 3.8 shows the PCA plots, based on final gene expression values for a final check. Similar timepoints and their replicates cluster together for both timeseries, indicating that the data is fit for downstream analyses. However, in ZS11 timeseries, replicate 2 of timepoint 48 clustered with samples from timepoints 59 and 63. Exercising abundant caution, that replicate was removed from downstream analysis.

### 3.4.2 Curve registration facilitates comparison of gene expression across timescales

Having a detailed RNA-Seq timeseries data for both *Brassica napus* and Arabidopsis, along with a mathematical framework for comparison of this data across timescales provides the ability to answer the questions mentioned in the introduction of this chapter.

As mentioned earlier, there has been research highlighting divergence in dynamics among *FLC* paralogues in *Brassica napus*. Using curve registration, I can compare the dynamics of each of those paralogues to their Arabidopsis orthologue. Figure 3.9 shows the comparison of dynamics of *FLC* paralogues in Stellar with the *FLC* orthologue in Arabidopsis. It shows that only three paralogues (A03p04430.1, C03p04920.1 and C03p19690.1) follow the same dynamics as the Arabidopsis *FLC*, decreasing in expression as plant approaches floral transition. Hence, these paralogues are considered ‘registered’ to the Arabidopsis *FLC*. Among the ‘non-registered’ paralogues, for instance, there are paralogues such as C02p04280.1 and C09p67380.1 which have opposite dynamics to the

| Timepoint (Days) | Replicate | Total Reads (Raw) | Total Reads (QC passed) | Mapped Reads (Total) | Mapped Reads (Unique) | Mapped Reads (Multi) | Alignment Rate (Total, %) | Alignment Rate (Unique, %) |
|---|---|---|---|---|---|---|---|---|
| 14 | 1 | 29466748 | 28758229 | 26250311 | 21450623 | 4799688 | 91.28 | 74.59 |
| 14 | 2 | 24666990 | 23939901 | 22244659 | 18205509 | 4039150 | 92.92 | 76.05 |
| 14 | 3 | 24219439 | 23540721 | 20695184 | 16720576 | 3974608 | 87.91 | 71.03 |
| 18 | 1 | 27746132 | 26994954 | 24039328 | 19672135 | 4367193 | 89.05 | 72.87 |
| 18 | 2 | 33326240 | 32381227 | 30014138 | 24840379 | 5173759 | 92.69 | 76.71 |
| 18 | 3 | 27514526 | 26844350 | 23813388 | 19441293 | 4372095 | 88.71 | 72.42 |
| 19 | 1 | 30157849 | 29446093 | 25690963 | 20873799 | 4817164 | 87.25 | 70.89 |
| 19 | 2 | 27995140 | 27481926 | 25596698 | 21166876 | 4429822 | 93.14 | 77.02 |
| 19 | 3 | 25437387 | 24794130 | 22219958 | 18092271 | 4127687 | 89.62 | 72.97 |
| 20 | 1 | 30370848 | 29405460 | 26043924 | 21347396 | 4696528 | 88.57 | 72.60 |
| 20 | 2 | 30099757 | 29388288 | 27204557 | 22345602 | 4858955 | 92.57 | 76.04 |
| 20 | 3 | 27622931 | 26912498 | 24237255 | 19923191 | 4314064 | 90.06 | 74.03 |
| 21 | 1 | 30529239 | 29805535 | 26592738 | 21588327 | 5004411 | 89.22 | 72.43 |
| 21 | 2 | 29794316 | 29136562 | 26831615 | 21922714 | 4908901 | 92.09 | 75.24 |
| 21 | 3 | 36161871 | 35106563 | 30734332 | 25166413 | 5567919 | 87.55 | 71.69 |
| 22 | 1 | 32216530 | 31394820 | 27383769 | 21996099 | 5387670 | 87.22 | 70.06 |
| 22 | 2 | 27599676 | 26998808 | 24780379 | 20283358 | 4497021 | 91.78 | 75.13 |
| 22 | 3 | 30980902 | 29687457 | 25480607 | 21084379 | 4396228 | 85.83 | 71.02 |
| 23 | 1 | 26562465 | 25762479 | 23996188 | 19771662 | 4224526 | 93.14 | 76.75 |
| 23 | 2 | 28097869 | 27432127 | 25438002 | 20846655 | 4591347 | 92.73 | 75.99 |
| 23 | 3 | 29856661 | 28439215 | 24767469 | 20450559 | 4316910 | 87.09 | 71.91 |
| 24 | 1 | 24302179 | 23720233 | 22227620 | 18323458 | 3904162 | 93.71 | 77.25 |
| 24 | 2 | 25725616 | 25002390 | 22554092 | 18347106 | 4206986 | 90.21 | 73.38 |
| 24 | 3 | 34036988 | 33161191 | 29517525 | 24103629 | 5413896 | 89.01 | 72.69 |
| 25 | 1 | 25441290 | 24739891 | 22925859 | 18769779 | 4156080 | 92.67 | 75.87 |
| 25 | 2 | 25364313 | 24703104 | 22616725 | 18484703 | 4132022 | 91.55 | 74.83 |
| 25 | 3 | 34674556 | 33787709 | 28811580 | 23528903 | 5282677 | 85.27 | 69.64 |
| 26 | 1 | 27883642 | 27133476 | 24622563 | 20091788 | 4530775 | 90.75 | 74.05 |
| 26 | 2 | 26767602 | 26016735 | 24070235 | 19627849 | 4442386 | 92.52 | 75.44 |
| 26 | 3 | 30754646 | 29893009 | 26058349 | 21372331 | 4686018 | 87.17 | 71.50 |
| 27 | 1 | 29092425 | 28297242 | 25591451 | 20748923 | 4842528 | 90.44 | 73.32 |
| 27 | 2 | 27636798 | 26948440 | 24168101 | 19537868 | 4630233 | 89.68 | 72.50 |
| 27 | 3 | 27314762 | 26627251 | 23070553 | 18893948 | 4176605 | 86.64 | 70.96 |
| 28 | 1 | 30368320 | 29648974 | 27310613 | 22373761 | 4936852 | 92.11 | 75.46 |
| 28 | 2 | 25388514 | 24712203 | 22165757 | 18035580 | 4130177 | 89.70 | 72.98 |
| 28 | 3 | 35986575 | 34972251 | 29791105 | 24352870 | 5438235 | 85.18 | 69.63 |
| 29 | 1 | 29019928 | 28365045 | 26366028 | 21633502 | 4732526 | 92.95 | 76.27 |
| 29 | 2 | 26945703 | 26330055 | 23975406 | 19563169 | 4412237 | 91.06 | 74.30 |
| 29 | 3 | 32549747 | 31559450 | 26980658 | 22142982 | 4837676 | 85.49 | 70.16 |
| 30 | 1 | 31955891 | 31242074 | 28555816 | 23213624 | 5342192 | 91.40 | 74.30 |
| 30 | 2 | 28568289 | 27908034 | 25435294 | 20680682 | 4754612 | 91.14 | 74.10 |
| 30 | 3 | 35836730 | 34892582 | 29785513 | 24392437 | 5393076 | 85.36 | 69.91 |
| 31 | 1 | 32110567 | 31256305 | 29136417 | 23968576 | 5167841 | 93.22 | 76.68 |
| 31 | 2 | 25037174 | 24462520 | 21993226 | 17886998 | 4106228 | 89.91 | 73.12 |
| 31 | 3 | 36175210 | 35194846 | 29940097 | 24560176 | 5379921 | 85.07 | 69.78 |
| 36 | 1 | 29711573 | 28943845 | 26905093 | 22156305 | 4748788 | 92.96 | 76.55 |
| 36 | 2 | 24666008 | 24092183 | 21873027 | 17825171 | 4047856 | 90.79 | 73.99 |
| 36 | 3 | 36570825 | 35493107 | 31326379 | 25658076 | 5668303 | 88.26 | 72.29 |
| 40 | 1 | 24759987 | 24082704 | 22577507 | 18681864 | 3895643 | 93.75 | 77.57 |
| 40 | 2 | 29149451 | 28423345 | 26149052 | 21360606 | 4788446 | 92.00 | 75.15 |
| 40 | 3 | 38484447 | 37477496 | 33408680 | 27256602 | 6152078 | 89.14 | 72.73 |

**Table 3.1: Alignment statistics for Stellar RNA-Seq timeseries**

Samples in Stellar timeseries show consistent alignment rate with an average of 89.99 %. In every sample, majority of reads are uniquely mapped, with an average of 73.60 %.

| Timepoint (Days) | Replicate | Total Reads (Raw) | Total Reads (QC passed) | Mapped Reads (Total) | Mapped Reads (Unique) | Mapped Reads (Multi) | Alignment Rate (Total, %) | Alignment Rate (Unique, %) |
|---|---|---|---|---|---|---|---|---|
| 14 | 1 | 29088022 | 27897273 | 25583153 | 24099747 | 1483406 | 91.70 | 86.39 |
| 14 | 2 | 28113626 | 26910158 | 24628289 | 23327684 | 1300605 | 91.52 | 86.69 |
| 14 | 3 | 29143760 | 27188491 | 24907004 | 23558444 | 1348560 | 91.61 | 86.65 |
| 21 | 1 | 28816758 | 27624035 | 25312574 | 23909766 | 1402808 | 91.63 | 86.55 |
| 21 | 2 | 29025505 | 27629407 | 25342085 | 24070885 | 1271200 | 91.72 | 87.12 |
| 21 | 3 | 34400390 | 32980128 | 30013865 | 28324762 | 1689103 | 91.01 | 85.88 |
| 22 | 1 | 32080073 | 31112362 | 27798546 | 26254399 | 1544147 | 89.35 | 84.39 |
| 22 | 2 | 26212390 | 25208164 | 22532362 | 21283801 | 1248561 | 89.39 | 84.43 |
| 22 | 3 | 29393360 | 27024586 | 24903455 | 23506869 | 1396586 | 92.15 | 86.98 |
| 28 | 1 | 33070956 | 31984101 | 29270737 | 27710522 | 1560215 | 91.52 | 86.64 |
| 28 | 2 | 33129847 | 31946712 | 28259389 | 26641333 | 1618056 | 88.46 | 83.39 |
| 28 | 3 | 27929090 | 26007458 | 23916788 | 22649543 | 1267245 | 91.96 | 87.09 |
| 35 | 1 | 30649665 | 29666122 | 27092814 | 25591679 | 1501135 | 91.33 | 86.27 |
| 35 | 2 | 32248865 | 30990176 | 27578385 | 25419504 | 2158881 | 88.99 | 82.02 |
| 35 | 3 | 25308905 | 23573540 | 21676941 | 20516806 | 1160135 | 91.95 | 87.03 |
| 42 | 1 | 30734708 | 29641959 | 27200527 | 25742116 | 1458411 | 91.76 | 86.84 |
| 42 | 2 | 28531932 | 27283456 | 24738225 | 23451098 | 1287127 | 90.67 | 85.95 |
| 42 | 3 | 30365866 | 28332979 | 25971772 | 24570151 | 1401621 | 91.67 | 86.72 |
| 43 | 1 | 31928008 | 30943833 | 28129984 | 26599705 | 1530279 | 90.91 | 85.96 |
| 43 | 2 | 32435585 | 31268231 | 28227875 | 26748477 | 1479398 | 90.28 | 85.55 |
| 43 | 3 | 29673339 | 27712760 | 25315800 | 23990479 | 1325321 | 91.35 | 86.57 |
| 44 | 1 | 29199045 | 28253568 | 26015920 | 24600419 | 1415501 | 92.08 | 87.07 |
| 44 | 2 | 28853179 | 27651679 | 25153458 | 23809685 | 1343773 | 90.97 | 86.11 |
| 44 | 3 | 30495075 | 28434393 | 26124353 | 24739848 | 1384505 | 91.88 | 87.01 |
| 45 | 1 | 30894186 | 29926563 | 27340435 | 25842840 | 1497595 | 91.36 | 86.35 |
| 45 | 2 | 32955158 | 31541393 | 28850701 | 27324648 | 1526053 | 91.47 | 86.63 |
| 45 | 3 | 28508077 | 26512504 | 24408144 | 23068291 | 1339853 | 92.06 | 87.01 |
| 46 | 1 | 26539306 | 25685541 | 23234577 | 21920074 | 1314503 | 90.46 | 85.34 |
| 46 | 2 | 29176090 | 27886901 | 25153426 | 23778029 | 1375397 | 90.20 | 85.27 |
| 46 | 3 | 30679980 | 28627216 | 26336814 | 24891047 | 1445767 | 92.00 | 86.95 |
| 47 | 1 | 29970992 | 29032463 | 26198828 | 24701889 | 1496939 | 90.24 | 85.08 |
| 47 | 2 | 27657857 | 26578286 | 23700782 | 22370736 | 1330046 | 89.17 | 84.17 |
| 47 | 3 | 30616638 | 28556301 | 26207885 | 24783939 | 1423946 | 91.78 | 86.79 |
| 48 | 1 | 28597461 | 27681949 | 25411134 | 24040700 | 1370434 | 91.80 | 86.85 |
| 48 | 2 | 31485319 | 29806178 | 27228160 | 25735433 | 1492727 | 91.35 | 86.34 |
| 48 | 3 | 32006349 | 29922957 | 27347500 | 25810830 | 1536670 | 91.39 | 86.26 |
| 49 | 1 | 27407943 | 26454157 | 24248082 | 22917575 | 1330507 | 91.66 | 86.63 |
| 49 | 2 | 32028586 | 30742737 | 27916630 | 26383301 | 1533329 | 90.81 | 85.82 |
| 49 | 3 | 29304806 | 27215277 | 25108385 | 23651732 | 1456653 | 92.26 | 86.91 |
| 50 | 1 | 28434393 | 27450312 | 25132533 | 23757674 | 1374859 | 91.56 | 86.55 |
| 50 | 2 | 31354056 | 29997949 | 27454904 | 25971424 | 1483480 | 91.52 | 86.58 |
| 50 | 3 | 32271589 | 31197519 | 27021341 | 25361830 | 1659511 | 86.61 | 81.29 |
| 51 | 1 | 30790378 | 29781969 | 27166546 | 25648793 | 1517753 | 91.22 | 86.12 |
| 51 | 2 | 29549960 | 28381763 | 25799253 | 24329650 | 1469603 | 90.90 | 85.72 |
| 51 | 3 | 33613132 | 31316736 | 28764374 | 27184063 | 1580311 | 91.85 | 86.80 |
| 52 | 1 | 27696028 | 25951320 | 23863478 | 22493976 | 1369502 | 91.95 | 86.68 |
| 52 | 2 | 26962046 | 25066060 | 22950750 | 21641821 | 1308929 | 91.56 | 86.34 |
| 52 | 3 | 25048637 | 23566957 | 21616755 | 20324936 | 1291819 | 91.72 | 86.24 |
| 53 | 1 | 31131400 | 28854059 | 26673889 | 25083148 | 1590741 | 92.44 | 86.93 |
| 55 | 1 | 29464384 | 28166854 | 25956339 | 24428436 | 1527903 | 92.15 | 86.73 |
| 55 | 2 | 28439600 | 26507046 | 24457825 | 23055147 | 1402678 | 92.27 | 86.98 |
| 55 | 3 | 25729035 | 23318243 | 21238502 | 19880148 | 1358354 | 91.08 | 85.26 |
| 59 | 1 | 25899547 | 24297331 | 22201414 | 20880705 | 1320709 | 91.37 | 85.94 |
| 59 | 2 | 26310042 | 24747388 | 22717900 | 21425940 | 1291960 | 91.80 | 86.58 |
| 59 | 3 | 31038946 | 29438724 | 27148577 | 25538293 | 1610284 | 92.22 | 86.75 |
| 63 | 1 | 28465340 | 27453957 | 25237685 | 23814645 | 1423040 | 91.93 | 86.74 |
| 63 | 2 | 37935025 | 36281692 | 33102666 | 31302369 | 1800297 | 91.24 | 86.28 |
| 63 | 3 | 33737794 | 32226702 | 29482717 | 27727418 | 1755299 | 91.49 | 86.04 |

**Table 3.2: Alignment statistics for ZS11 RNA-Seq timeseries**

Samples in ZS11 timeseries show consistent alignment rate with an average of 91.21 %. In every sample, majority of reads are uniquely mapped, with an average of 86.10 %.

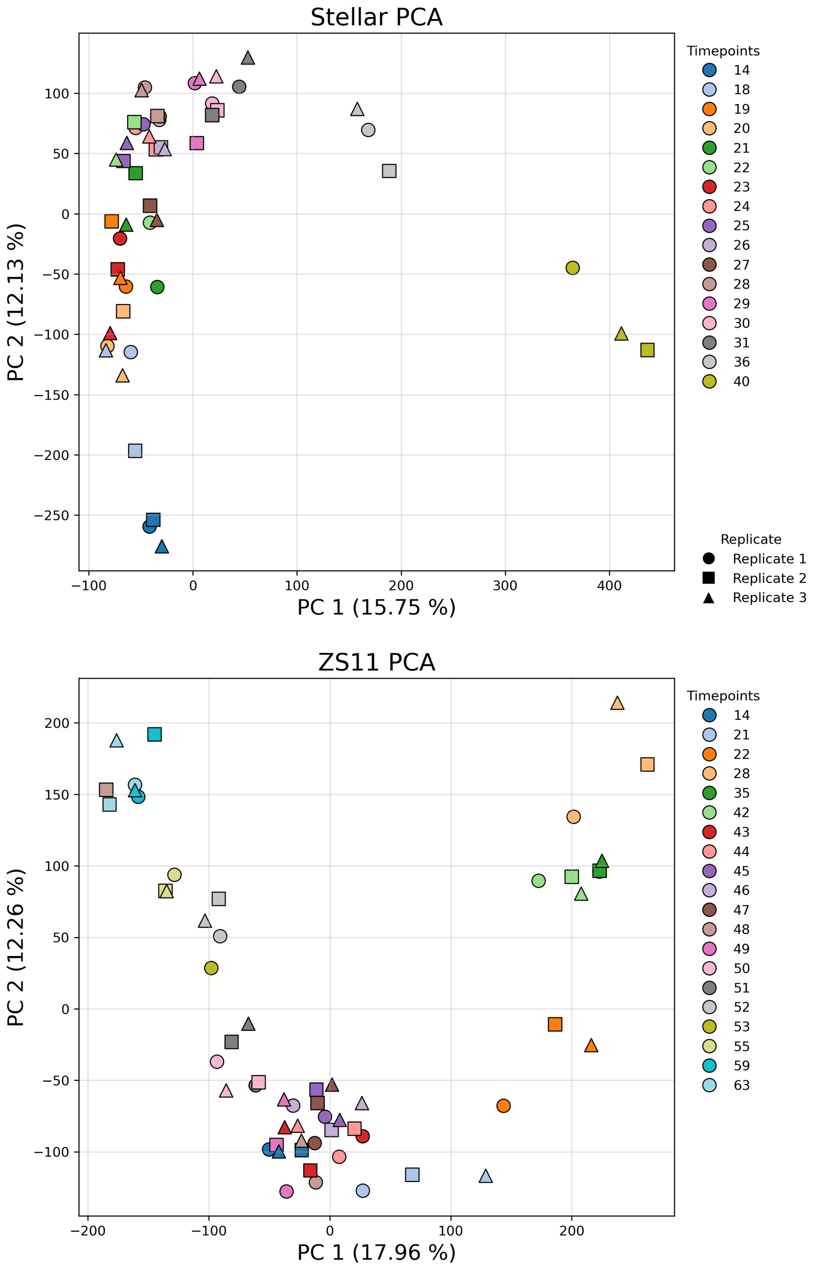


**Figure 3.8: Sense checking the timeseries samples based on gene expression values obtained after RNA-Seq analysis using PCA plots**

For both (a) Stellar and (b) ZS11 timeseries, the gene expression values generated following the RNA-Seq analysis are able to resolve timepoints and the replicates seem consistent. Replicate 2 from timepoint 48 in ZS11 timeseries clusters with timepoints 59 and 63, hence it was removed from further downstream analyses.

Arabidopsis *FLC*. Using this new methodology, I am able to reproduce the results corroborating a previous study by Calderwood *et al.* [91] where they observed variations in expression of *FLC* orthologues under cold conditions. They also performed detailed phylogenetic analysis using coding sequences and attribute their expression differences to relaxed selection pressure on individual orthologues. Reader is referred to their publication for further detail. For the purposes of this study, this result shows that curve registration is able to highlight these differences and in this case, agrees with already published research.

This approach, however, can be scaled to a much larger set of genes, and Figure 3.10 shows that comparison of timeseries is only possible following curve registration. The heatmaps show pairwise distances between timepoints of two timeseries, along the x and y-axis. This comparison is for the set of *Brassica napus* orthologues of 306 genes involved in flowering time in Arabidopsis. Due to the 'one-to-many' mapping in *Brassica napus*, this results in a list of 1103 genes.

If two timeseries 'progress' similarly, the distances in expression levels of orthologue pairs of genes would be minimum along the diagonal, shown by the dark blue colour (Figure 3.10 (a)). Without using curve registration, Arabidopsis and *Brassica napus* appear to have no such similarity in progression (Figure 3.10 (b)). For example, the timepoint when Stellar undergoes floral transition (day 40) would be expected to most similar to day 16 of Col-0 timeseries, as both plants should have same set of orthologues being up and down regulated. This similarity in developmental stages is however, not reflected in transcriptomic data.

Using curve registration on these timeseries, with Stellar as the reference and Col-0 as the query, similarities become clear, and a diagonal is visible on the heatmap (Figure 3.10 (c)). This shows that indeed *Brassica napus* and Arabidopsis have similarities in developmental progression, however, the differences in timescales obfuscate the comparison. Hence, curve registration enables a comparison of timeseries across different timescales.

### 3.4.3 Flowering time genes in *Brassica napus* exhibit more divergence in their expression dynamics than the rest of the genome

Extending the curve registration analysis to all genes, and to the ZS11 timeseries reveals that, for both cultivars, majority of genes show similar dynamics to their

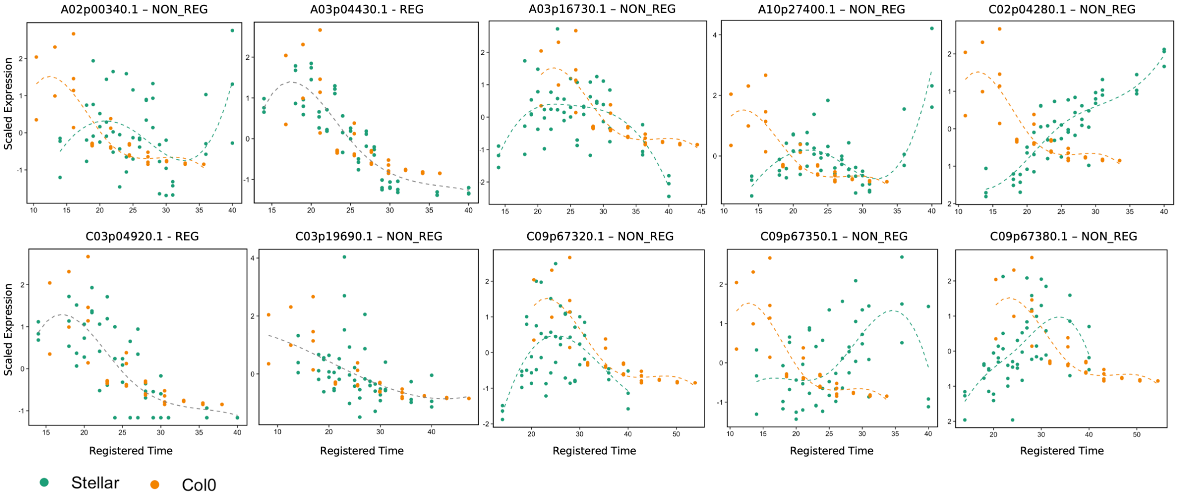


**Figure 3.9: Curve registration highlights the divergence in expression dynamics among *FLC* paralogues in *Brassica napus* cultivar Stellar**

*Brassia napus* Darmor v10 reference genome contains 10 orthologues for the Arabidopsis *FLC*. Only 3 orthologues, A03p04430.1, C03p04920.1 and C03p19690.1 show dynamics similar to the Arabidopsis *FLC* and are hence ‘registered’. The plot titles show the differences in the calculated BIC values, and the corresponding stretch and shift factors. The ‘non-registered’ orthologues have clearly different dynamics to the Arabidopsis *FLC*. For instance, C02p04280.1 and C09p67380.1 increase in expression as the plant approaches floral transition, which is exactly opposite to the dynamics of the Arabidopsis *FLC* orthologue.

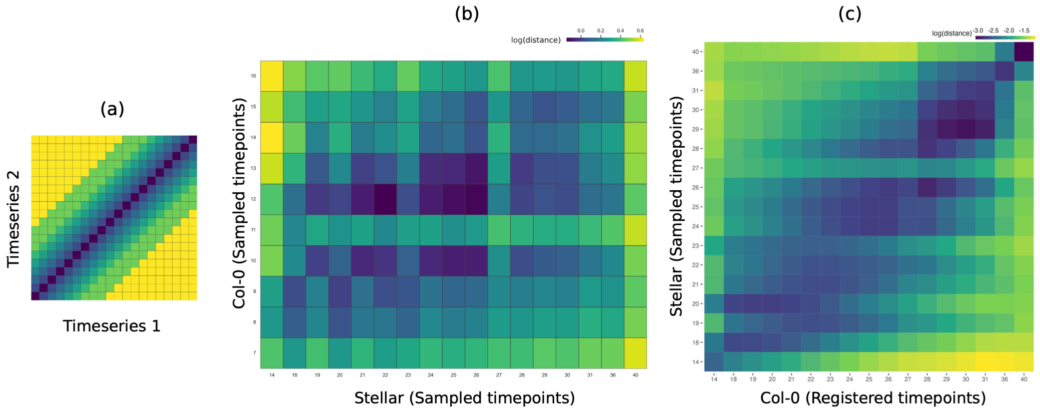


**Figure 3.10: Curve registration facilitates comparison of RNA-Seq time-courses across timescales**

The heatmaps show pairwise distance between gene expression at all timepoints between the Stellar and Col-0 timeseries for flowering time genes. (a) A comparison of pairwise distances two timeseries with similar progression will result in a plot with minimum distances along the diagonal, represented by dark blue colours while timepoints further away would be maximally distant, shown by green-yellow colours. (b) Pairwise distances with just normalised expression data (scaled to be between 0 and 1) between Stellar and Col-0 timeseries. This comparison does not show any similarity in progression between the two timecourses. (c) Using greatR to register the Col-0 timeseries against the reference Stellar, reveals a diagonal, showing that there are similarities between the two timeseries.

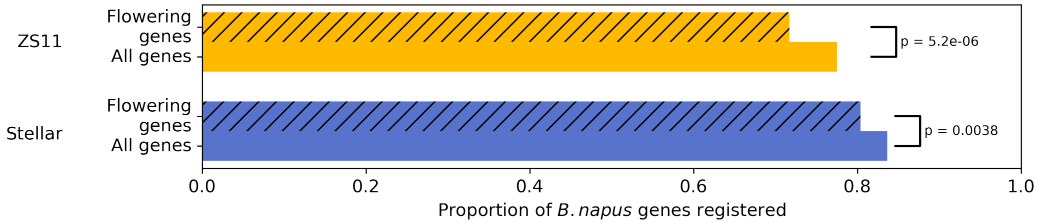


**Figure 3.11: Significantly lower proportion of *B.napus* flowering time genes are registered to their Arabidopsis orthologues**

The bars show the proportion of *Brassica napus* genes registered to their corresponding Arabidopsis orthologue. For the set of flowering genes, this proportion is significantly lower than the rest of the genome. p-values were calculated using the proportions z-test from statsmodels Python package [182].

Arabidopsis orthologues (Figure 3.11). 60,438 genes out of 72,239 expressed genes in Stellar and 55,332 genes out of 71,382 expressed genes in ZS11 were registered to their corresponding Arabidopsis orthologue. A gene in considered expressed if the standard deviation of its expression values is greater than 0, as per the default filtering criteria used by greatR. This shows that a lot of genes have conserved dynamics, hence could be performing similar functions during *Brassica napus* development as their orthologues in Arabidopsis. The proportion is slightly lower in ZS11, which could be because the plants underwent vernalisation before floral transition, while the Col-0 timeseries did not. This environmental variation could lead to changes in gene expression that would have not occured in Col-0 in absence of that treatment.

For both cultivars however, the subset of flowering time genes, show a drop in the proportion of genes registered. Only 887 out of the 1103 expressed orthologues to Arabidopsis flowering genes in Stellar and 789 out of 1100 in ZS11 were registered. While this drop in proportion is statistically significant based on proportions z-test, I performed bootstrapping analysis to demonstrate that this drop is not reproducible just by sampling a random set of genes from the genome. Shown in Figure 3.12, samples of size equal to the number of flowering time orthologues expressed in both timecourses, (1,103 and 1,100 for Stellar and ZS11 respectively) were randomly sampled with replacement from the set of all genes 10,000 times. The proportion of flowering genes registered does not fall within the distributions obtained from this analysis, demonstrating that it is not possible to reproduce the proportion figures by randomly sampling a set of genes. This shows that the set of orthologues of flowering time genes have indeed a higher degree of divergence than rest of the genome.

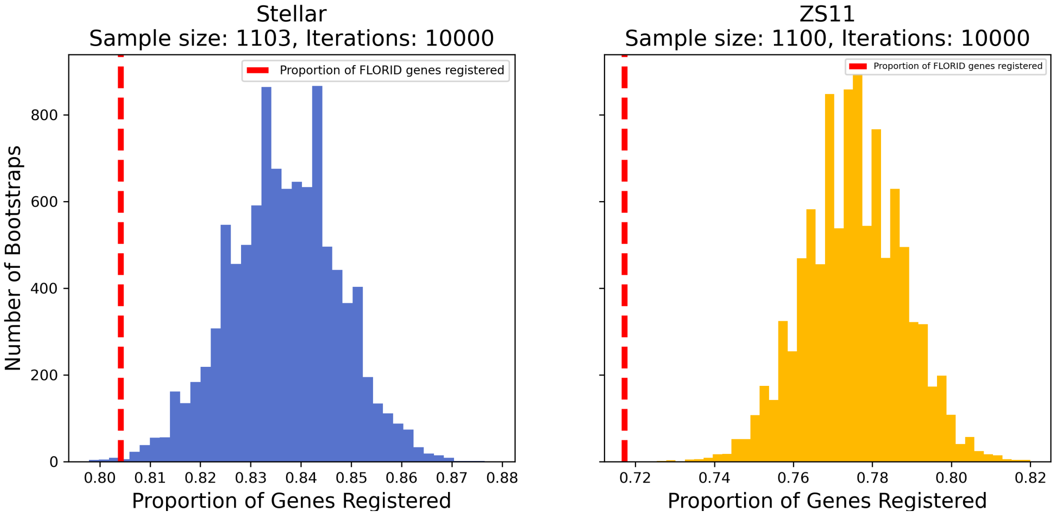


**Figure 3.12: It is not possible to reproduce the lower proportion of registration observed in the set of orthologues of flowering time genes through random sampling**

These results are from a bootstrapping experiment conducted in which samples of genes of sizes 1103 and 1100 for (a) Stellar and (b) ZS11 timeseries were sampled with replacement from a set of all genes for 10,000 iterations. The sample sizes are equal to the number of flowering genes that were expressed in both cultivars, based on filtering criteria of standard deviation $> 0$ used by greatR. The red bar shows the proportion of flowering time orthologues registered, represented on the x-axis. The plots show that the red line falls outside the distributions obtained following sampling, showing that it is unlikely that the lower proportion of registration to Arabidopsis orthologues observed in the set flowering time genes is by random chance.

### 3.4.4 Preferential retention is coupled with a higher divergence in expression dynamics of flowering time gene orthologues in *Brassica napus*

It has been shown that flowering time genes have been preferentially retained in higher number of paralogues in *Brassica napus* genome following duplications [90]. Since, this higher number could theoretically ease selection pressure on individual paralogues of a gene — it increases the chances of paralogues diverging in their regulation, and subsequently, function. As alluded to in the *FLOWERING LOCUS C* (*FLC*) example earlier in Figure 3.9. Hence, I postulated that retention in higher paralogue number is possibly correlated with higher divergence in expression dynamics observed in this set of orthologues.

Figure 3.13 shows that it could be indeed true. The heatmaps show the distribution of the proportion of paralogues of a particular gene in *Brassica napus* registered to their Arabidopsis orthologue on the x-axis against the number of orthologues on the y-axis. For example, in Stellar, for Arabidopsis genes with only one orthologue within the Brassica napus genome, 91 % of Arabidopsis genes have their orthologue registered to them while 9 % do not (Figure 3.13, Stellar (All genes), column 1). Similarly, in ZS11, 72 % of Arabidopsis genes with two orthologues have all their orthologues registered to them while 15 % have half of their orthologues registered, while 13 % have none of their orthologues registered (Figure 3.13, ZS11 (All genes), column 2). The numbers in the heatmaps show proportions out of 1 and Darker colours show higher proportion.

It is clearly visible that majority of Arabidopsis genes have all of their orthologues in *Brassica napus* registered, and this proportion drops as the number of orthologues increases, as denoted by the darker colours visible in the first row of all heatmaps. Notably, this trend starts to diverge in the Stellar (Flowering genes) and ZS11 (Flowering genes) heatmaps as the number of orthologues increases, indicated by darker colours beyond the first row in both heatmaps, compared to the Stellar (All genes) and Z11 (All genes) heatmaps.

Focussing just on the first row, Figure 3.14 (a) shows this drop clearly. For both cultivars, for both all genes and flowering time orthologues, as copy number increases, the proportion of Arabidopsis genes with all of their orthologues registered to them decreases. In agreement with previously published research, I also observed that flowering genes are expressed in higher copy numbers compared to the set of all genes, indicating that they are preferentially retained (Fig-

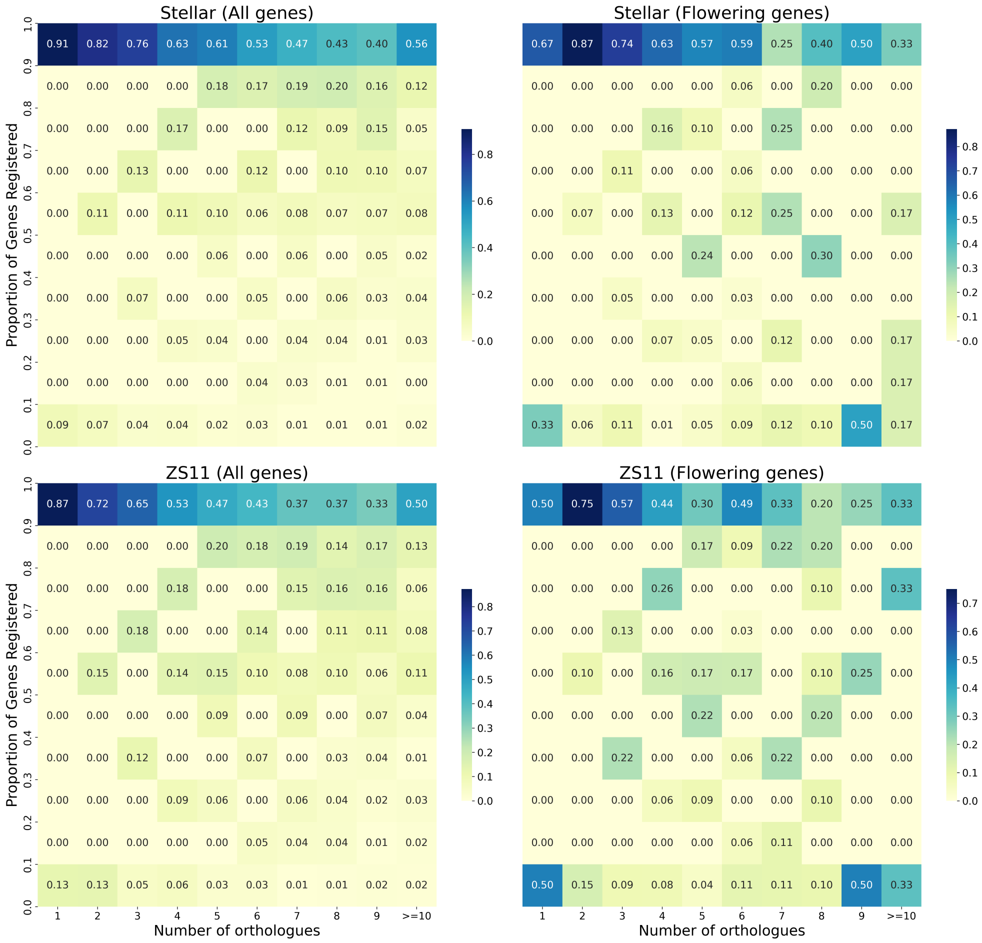


**Figure 3.13: Majority of genes have all the paralogues in *Brassica napus* registered to their Arabidopsis orthologues**

The heatmaps show the proportion of orthologues of Arabidopsis genes registered in *Brassica napus* registered to them (on y-axis) against the number of orthologues (on x-axis). It shows that for both Stellar and ZS11 in both whole genome and flowering time genes majority of genes have all of their orthologues in *Brassica napus* registered, indicated by dark blue colours on the top row of all heatmaps. The numbers in the heatmaps show the proportions. Flowering time genes show a greater drop in these proportions as number of orthologues increases than rest of the genome in both cultivars, indicated by darker blue colours beyond the first row.

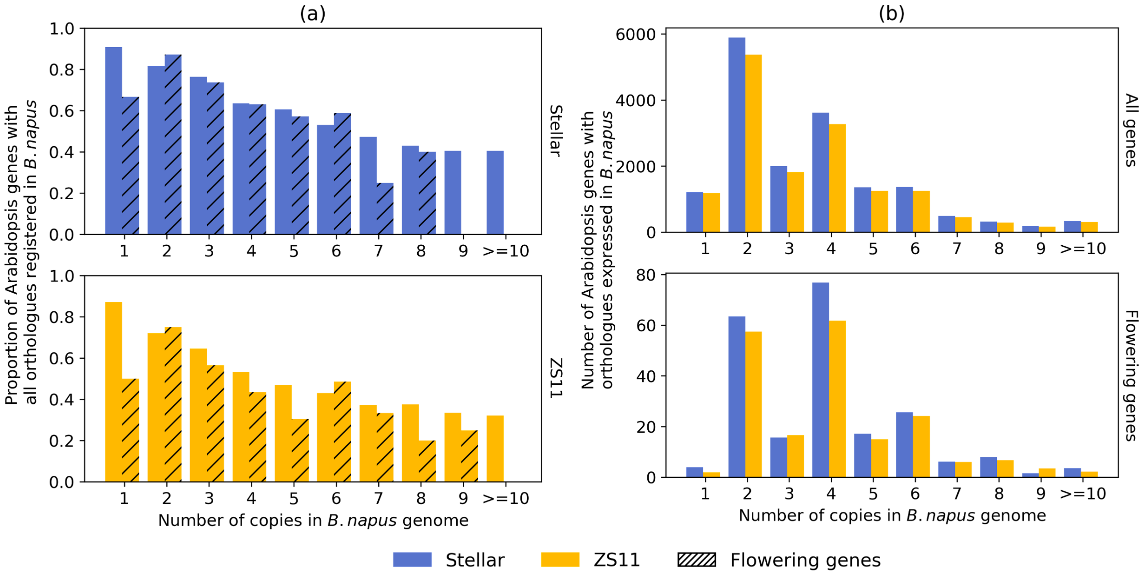


**Figure 3.14: Orthologues of flowering time genes show higher dergree of divergence in their expression dynamics as they are expressed in higher copy number in *Brassica napus***

(a) The bars show the proportion of set of orthologues with all the paralogues having similar dynamics, and hence marked as registered, to their Arabidopsis orthologue. As the number of orthologous copies in *Brassica napus* increases, the proportion of genes with all copies registered decreases. This is true for both the whole genome and flowering time genes for both cultivars. (b) However, flowering time genes are expressed in higher copy number than the set of all genes in *Brassica napus*. All genes are expressed with two as the most common number of paralogues, while this distribution is skewed towards higher number of paralogues for flowering time genes, evident by four as the most common expressed number of paralogues.

ure 3.14 (b)). For all genes, in both cultivars, two is the most common number of paralogues, while for the set of flowering time genes, this is four.

Hence, on a whole genome level, as the number of retained paralogues increases, the propensity for divergence increases and since flowering time genes are preferentially retained and expressed in higher number of paralogues, I observed that divergence shown in Figures 3.11 and 3.12.

### 3.4.5 Higher proportion of flowering time gene orthologues being transcription factors does not explain their divergence in expression dynamics

It has been hypothesised that one reason behind preferential retention is the fact that a large percentage of flowering time genes act as transcription factors,

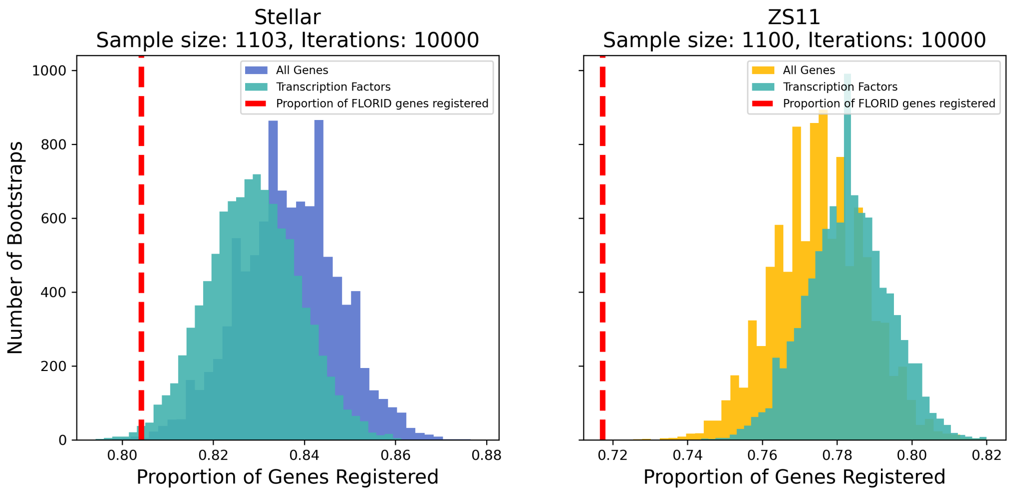


**Figure 3.15: Higher proportion of flowering time genes being transcription factors does not explain their lower proportion of gene registrations**

Samples of genes of sizes 1103 and 1100 for (a) Stellar and (b) ZS11 timeseries were sampled with replacement from a set of orthologues of Arabidopsis transcription factors in *Brassica napus* for 10,000 iterations. The distribution of proportion of registered genes thus obtained is shown in green, over the distribution generated by the same bootstrapping technique but from the set of whole genome. For both cultivars, the distributions sampled from set of transcription factors and all genes have large overlaps, however, the proportion of genes registered for flowering time genes is significantly than both.

hence are critical for multiple regulatory interactions [49] [90]. This could cause selection pressure on maintenance of stoichiometric balance on their gene products, favouring retention of paralogues and creating conditions ripe for increased divergence.

I observed that indeed a higher proportion of flowering time gene orthologues in *Brassia napus* are actually orthologues of Arabidopsis transcription factors. 32.27 % of all expressed flowering time genes in both Stellar and ZS11 are orthologues of Arabidopsis transcription factors, compared to just 7.6% for the whole of the *Brassica napus* genome. In order to test the hypotheses that flowering time genes being transcription factors correlates with higher degree of divergence, I performed a bootstrapping experiment.

Similar to analyses presented in Figure 3.12, I sampled sets of 1103 and 1100 genes out of all expressed transcription factors in Stellar and ZS11 respectively and calculated the proportion of genes registered. The results, shown in Figure 3.15, show that the distributions thus obtained have a high degree of overlap with the previous distributions generated from sampling from all genes for both cultivars. This shows that even when sampling from a smaller set of genes that are orthologues of transcription factors, flowering time gene orthologues have a higher degree of divergence and there appears to be no correlation between the two facts.

## 3.5 Conclusion

These sets of results expand on our current knowledge of flowering time gene orthologues in *Brassica napus* genome. Previous research had shown that orthologues of Arabidopsis flowering time genes are preferentially retained within Brassica napus.

Here, I have analysed a detailed gene expression timecourse for this crop. I show that curve registration can be used to compare dynamics of orthologues genes between species developing on different timescales. This comparison, between the model plant Arabidopsis and *Brassica napus* highlights that while majority of genes have conserved expression dynamics between them, flowering time genes show significantly higher divergence than rest of the genome. Statistical testing as well as bootstrapping experiments show that this divergence is significant. I show that as the number of retained paralogues increases, the propensity for divergence increases and since flowering time genes are present in higher paralogue numbers, preferential retention appears correlated with this higher divergence. Flowering time genes are also more likely to be transcription factors, however, this does not apprear to be the likely reason for this divergence in expression dynamics.

It is however, at this moment, unknown if this preferential retention of genes and higher divergence in their dynamics has any effect on regulatory interactions that from the gene regulatory network that controls flowering time within *Brassica napus*. This question is explored in the next chapter.

# Chapter 4

# On gene regulatory networks controlling flowering time in *Brassica napus*

## 4.1 Abstract

Research in Arabidopsis has shown that the timing of flowering in plants is regulated by a complex network of genes and transcription factors that respond to environmental cues. *Brassica napus* contains multiple orthologues of these genes, which have been shown to be preferentially retained in the genome. Previously, I showed that this preferential retention is coupled with a higher degree of divergence in expression dynamics among these retained genes compared to the rest of the genome. Here, I investigate the effects of these divergences on regulatory interactions between orthologues of flowering time genes. I show that the gene regulatory network retains the general structure observed in Arabidopsis, with nodes representing orthologues of floral integrator genes occupying central positions in the network. However, short-day, cold treatment likely induces changes this structure, leading to a network of locally clustered nodes. I also observe major differences in expression dynamics among orthologues of floral integrator *SOC1*, which is unique among orthologues of floral integrators.

## 4.2 Gene regulatory networks and how to infer them

The regulatory control of flowering time is a complex process controlled by multiple genes and transcription factors acting synergistically in response to external environmental factors [49]. As outlined in Chapter 1, research into the regulation of flowering time began as hypothesised individual factors which were later identified as gene products with the advent of modern biology. It was soon uncovered that these genes regulate each other and form a network, where some genes are directly affected by the external environment, while others are present downstream of these genes, regulated indirectly by transcription factors rather than directly by environmental cues.

It is now known that the flowering time gene regulatory network in Arabidopsis consists of multiple genes, constituting different pathways that converge on a few downstream regulators known as ‘floral integrator’ genes. These floral integrator genes turn on the meristem identity genes, which initiate the transition of the shoot apical meristem from a vegetative state to flowering (Figure 1.1) [49] [5].

The *Brassica napus* genome contains multiple orthologues for each Arabidopsis gene involved in the regulation of flowering time and flowering time genes have been preferentially retained in the genome through evolution [90]. I have shown that coupled with this preferential retention, flowering time genes in *Brassica napus* have diverged in their expression dynamics from their Arabidopsis orthologues at a higher proportion than rest of the genome (Chapter 3; Figures 3.11, 3.12). Given that genes are present in higher paralogue number, the dosage balance hypotheses implies that it likely relaxes selection pressure on individual paralogues [170]. Hence, it can be hypothesised that the network of interactions between these flowering time gene orthologues, despite the differences in expression dynamics, could still be similar to the one known in Arabidopsis, i.e. a tightly-knit network converging on orthologues of floral integrator genes, due to orthologues being preferentially retained.

This chapter presents network analyses of orthologues of flowering time genes based on their expression dynamics in *Brassica napus*. The following section briefly introduces methods to infer putative regulatory interactions between genes.

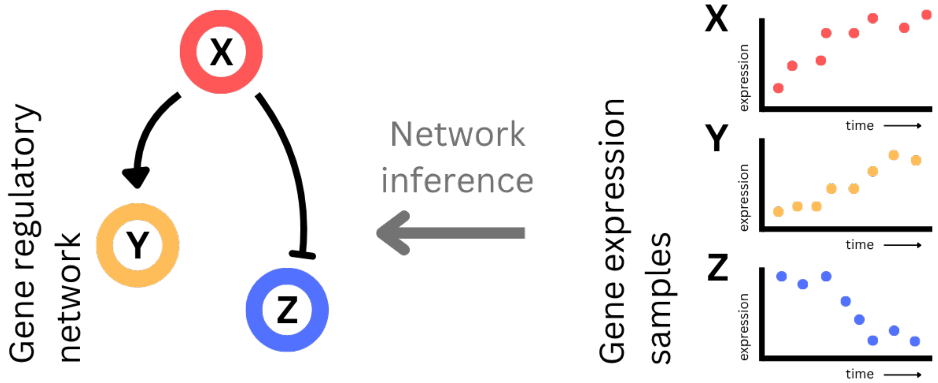


**Figure 4.1: Network inference involves reverse engineering the gene regulatory network from expression data**

In the gene regulatory network shown, X is a transcription factor that controls the expression dynamics for two genes Y and Z. As shown by two different arrowheads, X promotes Y and suppresses Z. This network structure is reflected in the dynamics of timeseries gene expression values for the three gene constituting this network, with expression of Y and Z increasing and decreasing respectively as expression of X increases.

### 4.2.1 Gene regulatory networks from transcriptomic timeseries

The regulation of expression by transcription factors includes control over the transcription of the gene into mRNA and translation of mRNA into protein via gene regulatory networks. Transcription factors bind to specific DNA motifs, often upstream of the target genes to influence the expression, by either promoting (or up-regulating) the gene's expression or by suppressing (or down-regulating) the transcription of the gene. As a result, the dynamics in transcriptome measurements gathered over time using RNASeq are an output of these gene regulatory networks (Figure 4.1). Network inference aims to reverse engineer the topologies of the networks from these outputs [183]. This allows us to disentangle suspected regulatory relationships among genes, including identification of highly connected genes, or 'hubs'.

There are multiple mathematical frameworks to perform this reverse engineering. A set of ordinary differential equations, for example, can be used to represent a deterministic model relating the change in mRNA concentrations of one gene to another gene in the system. This framework has been used to create a model that relates temperature data to flowering in Arabidopsis [184]. In the GRN, temperature, as the environmental cue, influences the expression of *VERNALISATION INSENSITIVE 3* (*VIN3*) gene which represses expression of *FLOWERING LOCUS C* (*FLC*). Repression of *FLC* promotes expression of *FLOW-*

*ERING LOCUS T* (*FT*), which after reaching a defined threshold concentration triggers bolting in plants. This network, while a clear simplification of the real regulatory network, provides a mechanistic understanding of how temperature affects flowering time, and by changing a few parameters is able to simulate annual, biennial and perennial behaviour. While ordinary differential equations provide a deterministic and mechanistic model of the regulatory process, the disadvantage is that it is not a feasible strategy for creating networks from a large set of genes or systems that are not well characterised.

Data driven approaches are more suitable for systems with a large number of uncharacterised genes. Sometimes referred to as 'Non-parametric approaches', they can be used to learn non-linear interactions between genes from the input data to infer gene regulatory networks in systems with a large number of genes and little prior knowledge [185]. This class of methods, however, has its own limitations. Unlike a system of ordinary differential equations, they do not provide mechanistic insights into the system being studied. Furthermore, the inference of regulatory networks from expression data is a hard problem to solve, as a review published in 2023 remarked, that GRN inference for large scale data often *struggles to reach high performance in real-world studies* [186].

Even using noise-free datasets with multiple replicates, network inference is still a challenging problem to solve as there can be multiple network structures that can produce the same output. As illustrated in Figure 4.2, Network structures A, B and C are some of the possible structures that can produce the timeseries output for the three genes X, Y and Z. Just based on timeseries data output, a network inference technique cannot meaningfully disentangle the true network structure within a biological system. This example assumes a three gene system in isolation, the possibilities become infinitely large in real biological datasets, even more so in polyploid systems, *for researchers audacious enough to study them*.

However, when used in conjunction with other data, they are still a *practically useful* [183] tool to analyse large transcriptomic timeseries datasets. Networks are useful to identify the overall regulatory structures between two datasets or conditions, investigate groups of genes that cluster together and can provide insights into regulatory divergences that can, and should, be further investigated.

With these limitations in mind, I conducted network inference using two frameworks in this class of network inference methods, a Gaussian process re-

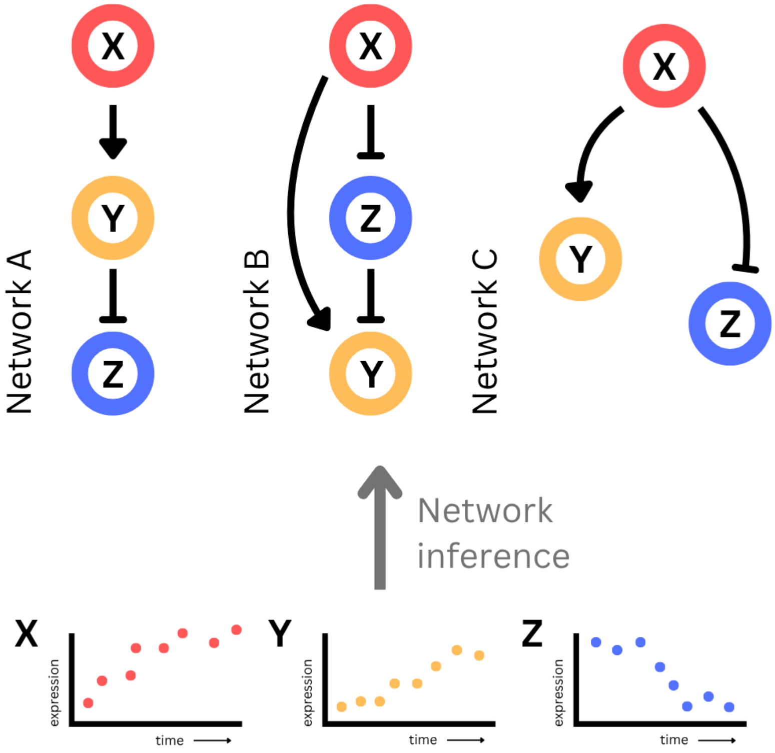


**Figure 4.2: Network inference is a hard problem to solve even with perfect, noise-free data**

The output expression of genes X, Y and Z can be produced by all three networks A, B and C. It is not possible, just based on transcriptomic data, to determine which network structure resembles the true regulatory framework within the biological system under investigation.

gression based and a Random Forest regression based method. Introduced in the following sections, these are two network inference approaches that are based on different mathematical frameworks. While these two approaches have been compared with other techniques using similar mathematical frameworks, there exists no direct comparison between these two techniques. Hence, I performed a comparison between the two techniques to select the better performing one on my data.

### 4.2.2 Gaussian process regression based network inference

In 2009, Äjiö and Lähdesmäki proposed a non-parametic method to infer regulatory functions from transcriptomic data [187]. The method employs Gaussian process regression [188] to infer these regulatory functions. The advantages of this framework are that the method makes minimal assumptions about the data and employs non-parametric modelling to infer GRNs from timeseries and steady-state data. The use of Gaussian processes also allows for quantification of uncertainty, by assuming normally distributed noise and learning its characteristics from the data.

If the set of parents $X_G$ for a gene $G$ are known *a priori*, the task becomes to infer a non-linear function $f(.)$ between a matrix of input expression values for the set of parents, $X_G(t-1)$ at timepoint $t-1$ and an output expression value $G(t)$ at the next timepoint $t$. Mathematically,

$$G(t) = f(X_G(t-1)) + \epsilon \tag{4.1}$$

where $\epsilon$ is Gaussian noise. This regression task can be achieved by assigning a Gaussian process prior for the non-linear function $f(.)$,

$$f(x) \sim GP(m(x), k(x, x')) \tag{4.2}$$

where $m(x)$ is the mean function and $k(x, x')$ is the covariance matrix. Gaussian processes provide a non-parametic prior distribution over functions and are generalisations of multinomial Gaussian distributions. For scaled expression data, the mean function $m(x)$ is zero.

The covariance matrix is based on covariance functions [189]. Äjiö and Lähdesmäki used the Matérn 3/2 covariance function,

$$k(x,x') = \sigma(1+\sqrt{3}\sqrt{u^T P^{-1} u})e^{(-\sqrt{3}\sqrt{u^T P^{-1} u})} \tag{4.3}$$

where, $u = x - x'$, $P =$ diag $(l^2)$, $l$ is a length-scale parameter and $\sigma$ is an additional variance parameter. If $\Theta_j$ represents the learnable parameters of the Gaussian process model $M_G$, for a parental set $j$, for gene $G$, marginal likelihood maximisation,

$$\Theta_j|M_G = \arg_\Theta \max(p(X_G|G,\Theta)), \forall j \tag{4.4}$$

can be achieved using optimisation. This calculation requires a separate Gaussian process model for each possible set of parents for gene $G$, which scales super-exponentially with the number of possible parents. By limiting the maximum number of parents, the scaling becomes polynomial [185].

Other algorithms based on this framework exist. For example, Causal Structure Inference [190] uses the same framework, however, it differs in its use of a squared exponential covariance function and instead of separately maximising $\Theta_j$ for each model, it uses an Expectation-maximisation approach to infer the parameter sets jointly.

### 4.2.3 Random Forest based network inference

Random Forest [191] is an alternative to Gaussian process for regression and many network inference methods are based on it. OutPredict [192] is among the latest methods in this class. It uses expression values from all transcription factors (TFs) as feature data to learn a model, based on Random forest, to predict the expression of the target gene, $G$. All expression values except the last one are used as the training data. It can incorporate both timeseries and steady-state data, and can also incorporate prior information, if available.

Following model training, it calculates the mean-squared error (MSE) between the model's prediction of the last timepoint and the actual value for all genes in the data. The method then compares it to the MSE value based on a naïve pen-ultimate prediction, where the last timepoint of the training data (i.e. second to last timepoint in the timeseries) is used as the prediction for the test timepoint (the last timepoint in the timeseries). A lower model MSE compared to penultimate approach MSE shows that the trained model can predict better than a naïve penultimate prediction approach.

A trained OutPredict model is then used to calculate an importance score, which is a product of the reduction of variance Random Forest importance measure and weight given by priors, if prior information was provided. The weight is set to 1 otherwise. For each decision node $d$,

$$I(d) = [S_{\text{num}}\sigma_y(S) - S_{l_{\text{num}}}\sigma_y(S_l) - S_{r_{\text{num}}}\sigma_y(S_r)]w_{X_i,G} \tag{4.5}$$

where $S$ is the subset of samples that are below decision node $d$ in the tree, $S_l$ and $S_r$ are subsets of samples on left and right branches on the node. $\sigma_y$ is the variance of the target gene in a given subset while $S_{\text{num}}$, $S_{l_{\text{num}}}$ and $S_{r_{\text{num}}}$ are the number of training samples in each subset. $w_{X_i,G}$ is the prior weight from a given feature (TF) to a target gene $G$. As models for each gene are independent, OutPredict runs these calculations in parallel.

For the final importance score, $s_i$ for the TF $X_i$, which can be used as an edge weight in the network is given by,

$$s_i = \frac{1}{T}\sum_{D_i} I(d) \tag{4.6}$$

where $T$ is the number of trees and $D_i$ is the set of nodes that branch based on $X_i$.

A python implementation of OutPredict was available, while I implemented the Gaussian process regression method. The following section details the methods before presenting the results.

## 4.3 Methods

### 4.3.1 Creating a list of transcription factors

To create a high confidence list of transcription factors (TFs) in *Brassica napus* Darmor v10 reference genome [176], protein sequences of orthologues of known Arabidopsis TFs were used as input for the DeepTFactor tool [193]. DeepTF uses a pre-trained machine learning model to classify protein sequences as ‘TF’ and ‘non-TF’. If an orthologue of an Arabidopsis TF was also predicted to be a TF by DeepTFactor, it was included in the high confidence list. 549 transcription factors were identified and used for network inference. The list of Arabidopsis transcription factors was obtained from PlantTFDB [181] (https://planttfdb.gao-lab.org/). Orthologue mapping was done using the reciprocal BLAST method described in Chapter 3 (Algorithm 1) and orthologues of flowering time genes were also identified as described in Chapter 3.

### 4.3.2 Network inference

The Gaussian process regression based network inference algorithm was implemented using GPflow (version 2.5.2) [194], based on TensorFlow (version 2.11.1) [195]. For inference, the gene expression values were scaled to be between 0 and 1, and BFGS optimiser from SciPy (version 1.11.1) [196] was used.

The available OutPredict implementation was used for network inference for Arabidopsis and *Brassica napus* timeseries datasets, as introduced in Chapter 3 (summarised in Figure 3.2). Default parameters were used for all inferred networks.

### 4.3.3 Thresholds used for filtering data

The input expression data for OutPredict requires filtering for low expressed genes. So, only genes with an average expression value across timeseries greater than 0.001 TPM and at least 30 % of the timepoints with expression $>$ 0 TPM were selected for network inference. Out of 1,232 orthologues of flowering time genes in *Brassica napus* Darmor v10 genome reference, 980 and 976 genes passed this criteria in Steller and ZS11 timeseries respectively and were used as input for network inference.

The inferred networks were filtered to remove the likely false positive edges. A cutoff of 0.001 importance score was applied and self-loops removed. The inferred networks following this filtration had 16,791 and 20,314 edges for Stellar and ZS11 timeseries respectively (980 and 976 nodes respectively). These networks were still too large for visualisation and likely still contain a lot of false positive edges. To create 'high confidence' networks, top 500 edges based on importance scores, were selected. These networks had 304 and 377 nodes representing individual genes for Stellar and ZS11 respectively and were used for visualisations presented in results.

### 4.3.4 Visualisation of networks

Networkx (version 3.1) [197] was used to visualise the networks, with node positions decided using spring layout method. Seed was set to 123 for reproducibility. Node sizes were determined by degree centrality (number of edges connected to the node) of each node.

### 4.3.5 Average clustering coefficient analyses

Networkx (version 3.1) was used for calculation of average clustering coefficient. Higher clustering coefficient value for a network indicates structure with a single tight cluster of nodes, while a lower value either means a sparsly connected network, with possible tighter local connections instead.

For an unweighted and undirected graph, for a node $u$, clustering coefficient $C$, is defined as,

$$C = \frac{T(u)}{\deg(u)(\deg(u) - 1)} \tag{4.7}$$

where $T(u)$ is the number of triangles through node $u$ and $\deg(u)$ is the degree (number of edges connected to the node) of the node $u$. The mean value for the whole network is the average clustering coefficient.

### 4.3.6 Identification of communities

Detection of 'k' communities in a network was done using the fluid communities algorithm [198]. A community refers to a cluster of nodes within a network. The seed value was set to 246 for reproducibility. The value of 'k' was determined using maximum modularity criteria [199]. For a network, the modularity $Q$ is defined as,

$$Q = \sum_{c=1}^{n} \left[ \frac{L_c}{m} - \gamma \left( \frac{k_c}{2m} \right)^2 \right] \tag{4.8}$$

where the sum iterates over the number of communities, $m$ is the of the edges, $L_c$ is the number of intracommunity edges for community $c$, $k_c$ is the sum of degree of nodes in community $c$, and $\gamma$ is the resolution parameter, set to 1 by default.

### 4.3.7 Rank-Biased Overlap analysis

Rank-Biased Overlap (RBO) is a method to compare two ranked lists [200]. I wrote a python implementation of the algorithm and used it for comparison of ranked lists of regulators, based on OutPredict importance scores, for individual genes in the inferred networks.

For two lists, RBO is defined as,

$$RBO(L,S,p) = \frac{(1-p)}{p}\left(\sum_{d=1}^{l}\frac{X_d}{d}p^d + \sum_{d=s+1}^{l}\frac{X_s(d-s)}{sd}p^d\right) + \left[\frac{X_l - X_s}{l} + \frac{X_s}{s}\right]p^l \tag{4.9}$$

where $L$ is the longer list and $S$ is the shorter list with lengths $l$ and $s$ respectively. $X_d$ is the intersection of the two lists upto depth $d$. $p \in (0,1)$ determines the contribution of $d$ ranks to the RBO measure. Higher $p$, more top weighted the calculated RBO value is. For calculations in this chapter, $p$ was set to 0.9.

## 4.4 Results and Discussion

### 4.4.1 OutPredict outperforms on Arabidopsis data

Evaluation of performance of Gene regulatory network techniques is often done on simulated datasets [185] [183] [186], however, those datasets are simplified models of the real-world data, often better than available biological datasets and the performance of a network inference method might not generalise to other datasets [201]. Additionally, since the two network inference methods that I selected are based on two different mathematical frameworks, I did not find any direct comparisons between Gaussian processes regression (GPR) network inference technique and OutPredict.

Hence, I decided to use the Arabidopsis timeseries data, as described in Figure 3.2, to evaluate the performance of the two techniques. This dataset is similar to the *Brassica napus* datasets with a key advantage that experimentally verified regulators for a number of flowering genes are known. I selected AT2G45660 (*SOC1*) gene and evaluated the methods on the ranking of known regulators relative to all other flowering time genes. The list of 49 known regulators was obtained from FLOR-ID database [50].

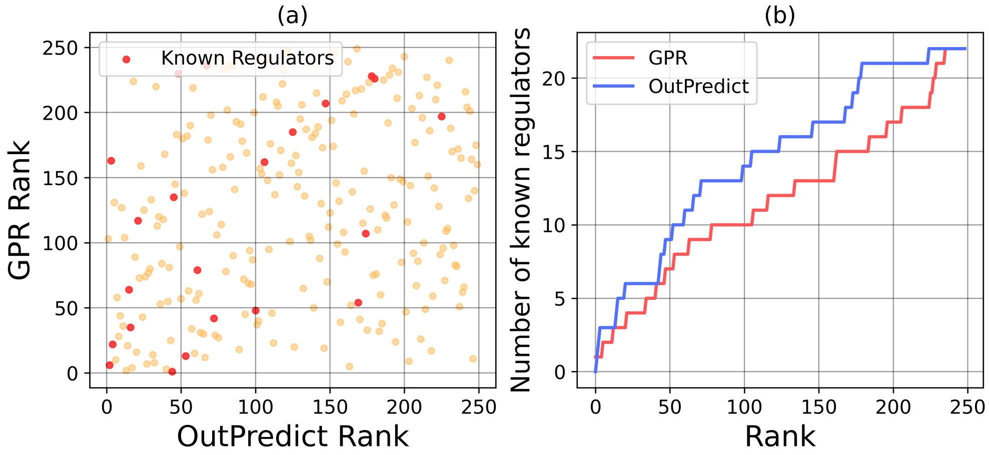


**Figure 4.3: OutPredict ranks known regulators higher than other genes compared to Gaussian process regression (GPR)**

(a) Pairwise comparison of ranks by OutPredict and GPR for each gene, shown as yellow dots in the plot, with known regulators highlighted in red. If the final rankings were in agreement, the scatter plot would have been along the diagonal, but the dispersion shows that the outputs from two frameworks have little agreement. (b) OutPredict ranks known regulators higher than other genes compared to GPR and at any cut-off rank would include higher number of known regulators in the resultant set, shown by blue line being higher than the red one at every value on the x-axis.

Figure 4.3 (a) shows the comparison of ranks of all genes, with known regulators highlighted in red. The scatter plot shows that both techniques have little agreement in their rankings of putative regulators. Regulators that are ranked highly by GPR can be ranked very low by Outpredict and vice versa. This is true for known regulators as well. These observations are broadly in line with other comparisons and illustrative of the differences in the frameworks for network inference and the difficult nature of this reverse engineering problem. However, as shown in Figure 4.3 (b), OutPredict ranks the known regulators higher than other genes compared to GPR. The top ranked gene by GPR is a known regulator, however, if any cutoff is introduced, the resultant set of regulators from OutPredict ranking would contain a higher number of known regulators compared to GPR. For instance, the top 100 genes ranked by GPR included 10 known regulators while it was marginally higher at 14 for OutPredict.

Even more importantly, however, OutPredict completes its network inference much faster than GPR for the same amount of data. This limits the use of GPR to infer regulators of a small number of target genes while OutPredict can handle much larger number of genes and infer a network of interactions between — which is ideal for a set of orthologues of flowering time genes. Due to these two advantages, I selected OutPredict to perform network inference on *Brassica napus* datasets.

### 4.4.2 Network inference in Stellar highlights similarities between *Arabidopsis* and *Brassica napus*

Stellar, is a spring type *Brassica napus* cultivar, described in Chapter 3, Figure 3.2. A set of 980 orthologues of Arabidopsis flowering time genes were used to infer a network of interactions between them. The resultant network of 16,791 edges was filtered to keep top 500 edges based on importance scores. Figure 4.4 shows the largest interconnected component from this network. The sizes of the nodes are proportional to their degree centrality. This means that more important nodes i.e. nodes with more edges with other nodes are larger. The figure also labels ten nodes with the highest degree centrality, along with the pathways their Arabidopsis orthologues are known to be involved in. Table 4.1 shows top 25 nodes within the Stellar network, along with their Arabidopsis orthologue and pathways the orthologue is involved in. The information about each node's constituent pathway is from the FLORID database [50].

It is clear from the structure of the network that it converges on a few key nodes, visualised in the centre with the highest degree centrality. Orthologues of genes from *SQUAMOSA PROMOTER BINDING PROTEIN-LIKE* (*SPL*) family, *SPL3* and *SPL5* are among the key nodes. In Arabidopsis, these genes have been shown to be up-regulated during photoperiod induced flowering [202] and are regulated by microRNAs, specifically miRNA-156, which is involved in the ageing pathway [203] [204]. These genes have also been shown to be part of a module with *SUPRESSOR OF OVEREXPRESSION OF CONSTANS 1* (*SOC1*) that incorporates signals from gibberellic acid, the endogenous hormonal pathway [205]. These genes encode a transcription factor that regulates floral integrator genes, *APETALA 1* (*AP1*), *LEAFY* (*LFY*) and *AGAMOUS-LIKE 8* (*AGL8*) [206]. *AGL8* is also known as *FRUITFUL* (*FUL*). These three genes, along with *SOC1* and *FLOWERING LOCUS T* (*FT*) are termed as floral integrators within this chapter. The *SPL* genes are present alongside orthologues of floral integrators *SOC1*, *AP1* and *AGL8* in the network as the most important nodes in the network.

This shows that the flowering time network in *Brassica napus* maintains a similar overall structure to the Arabidopsis network. Floral integrator genes, or genes directly regulating them are identified as nodes with high degree central-

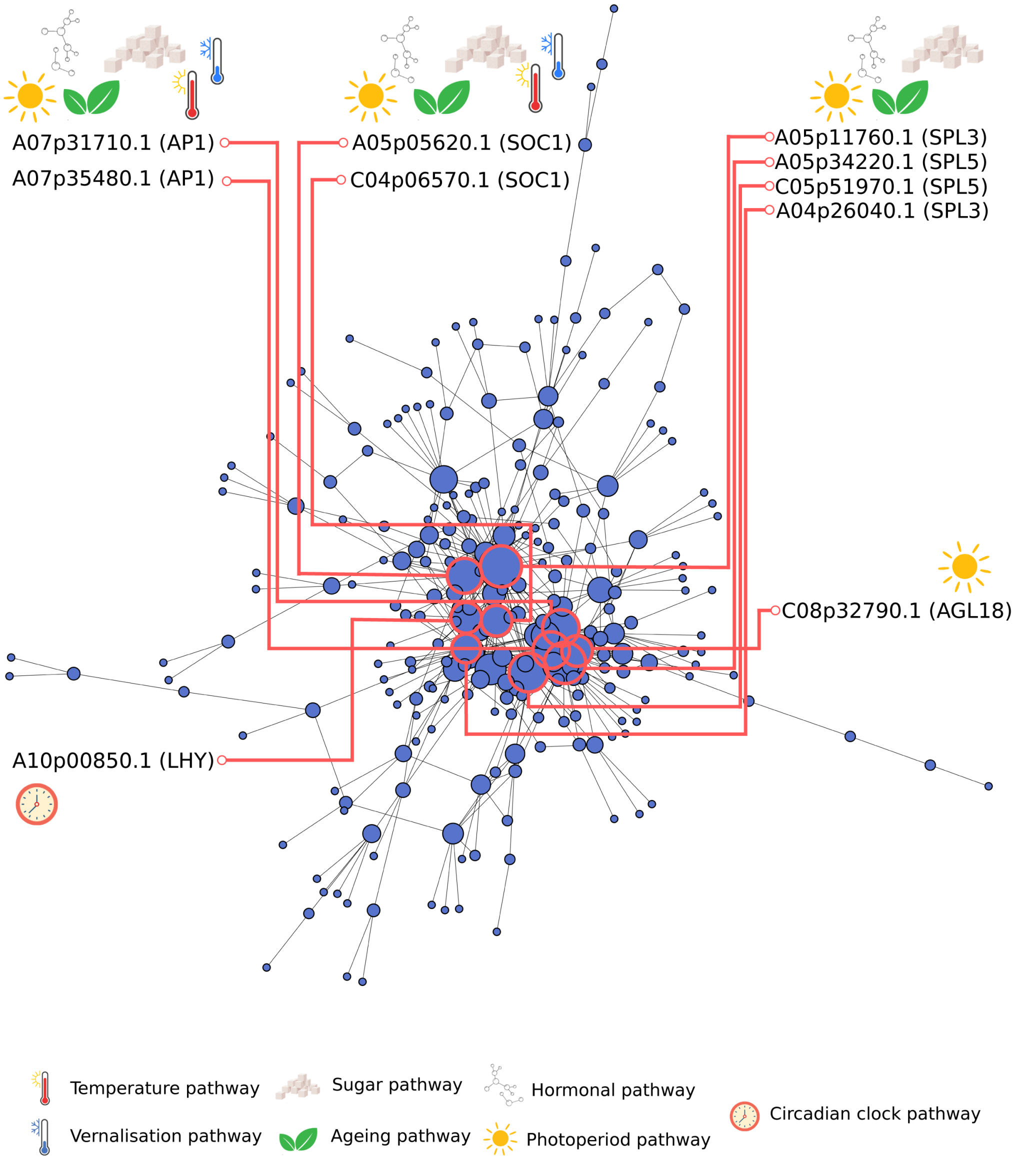


**Figure 4.4: Gene regulatory network of flowering time genes in *Brassica napus* cv. Stellar**

The largest interconnected component from the filtered network inferred from Stellar timeseries data. Each node is a gene, with its size proportional to its degree centrality. Top ten nodes by degree centrality are labelled alongwith their Arabidopsis orthologue and the pathways they are known to be involved in. It shows that the network is similar in structure to Arabidopsis where floral integrators constitute the important downstream nodes.

| *Brassica napus* gene | Centrality rank | Arabidopsis id | Symbol | Pathway (s) |
|---|---|---|---|---|
| A05p11760.1 | 1 | AT2G33810 | SPL3 | Aging, Photoperiod, Sugar, Hormone |
| A05p34220.1 | 2 | AT3G15270 | SPL5 | Aging, Photoperiod, Sugar, Hormone |
| C05p51970.1 | 3 | AT3G15270 | SPL5 | Aging, Photoperiod, Sugar, Hormone |
| A07p35480.1 | 4 | AT1G69120 | AP1 | Floral integrator |
| A05p05620.1 | 5 | AT2G45660 | SOC1 | Floral integrator |
| A07p31710.1 | 6 | AT1G69120 | AP1 | Floral integrator |
| A10p00850.1 | 7 | AT1G01060 | LHY | Circadian clock |
| C08p32790.1 | 8 | AT3G57390 | AGL18 | Photoperiod |
| C04p06570.1 | 9 | AT2G45660 | SOC1 | Floral integrator |
| A04p26040.1 | 10 | AT2G33810 | SPL3 | Aging, Photoperiod, Sugar, Hormone |
| C07p44700.1 | 11 | AT5G60910 | AGL8 | Floral integrator |
| A03p18910.1 | 12 | AT2G33810 | SPL3 | Aging, Photoperiod, Sugar, Hormone |
| C06p40760.1 | 13 | AT1G69120 | AP1 | Floral integrator |
| A09p47980.1 | 14 | AT3G54990 | SMZ | Aging, Hormone, Photoperiod, Vernalisation |
| A05p10920.1 | 15 | AT2G34720 | NF-YA4 | Photoperiod, Hormone |
| A01p38590.1 | 16 | AT3G15270 | SPL5 | Aging, Photoperiod, Sugar, Hormone |
| C01p05690.1 | 17 | AT4G32980 | ATH1 | Photoperiod |
| C05p00880.1 | 18 | AT1G01060 | LHY | Circadian clock |
| C04p01110.1 | 19 | AT2G46830 | CCA1 | Circadian clock |
| A01p35910.1 | 20 | AT3G18550 | BRC1 | Photoperiod, Meristem identity |
| A05p18570.1 | 21 | AT1G53160 | SPL4 | Aging, Photoperiod, Sugar, Hormone |
| C09p52620.1 | 22 | AT5G60120 | TOE2 | Aging, Photoperiod, Temperature |
| A10p17270.1 | 23 | AT5G60120 | TOE2 | Aging, Photoperiod, Temperature |
| C09p04070.1 | 24 | AT3G28910 | ATMYB30 | Autonomous |
| C04p63380.1 | 25 | AT2G33810 | SPL3 | Aging, Photoperiod, Sugar, Hormone |

**Table 4.1: Top 25 nodes by degree centrality in the network inferred from Stellar timeseries**

ity. Only two orthologues of genes, *LATE ELONGATED HYPOCOTYL* (*LHY*) and *AGAMOUS-LIKE* (*AGL18*), which are involved in circadian clock [207] and photoperiod pathway [208] respectively, are present in top nodes but have not been shown to integrate signals from multiple pathways in Arabidopsis. This is also true if the set is expanded to the top 25 genes (Table 4.1).

Figure 4.5 shows that this is true for all nodes in the network. The red line in the plot shows that genes that are orthologues of Arabidopsis genes involved in multiple pathways are few in number, however, these few nodes have high degree centrality. This again highlights that the network converges on a few key genes. The only gene that is involved in nine different pathways is *FLOWERING LOCUS T* (*FT*). *FT*, introduced in Chapter 1, is the florigen gene that is expressed in leaves before its translated protein is transported to the shoot apical meristem. Hence, the expression of its orthologues is not detected in RNASeq timeseries of the shoot apical meristem and the nodes are not present as important nodes in this network.

Hence, flowering time gene network, on a high level, appears conserved between Arabidopsis and *Brassica napus* as stipulated by the gene dosage hypotheses, however, individual genes can still have diverged regulation. This is explored in sections that follow.

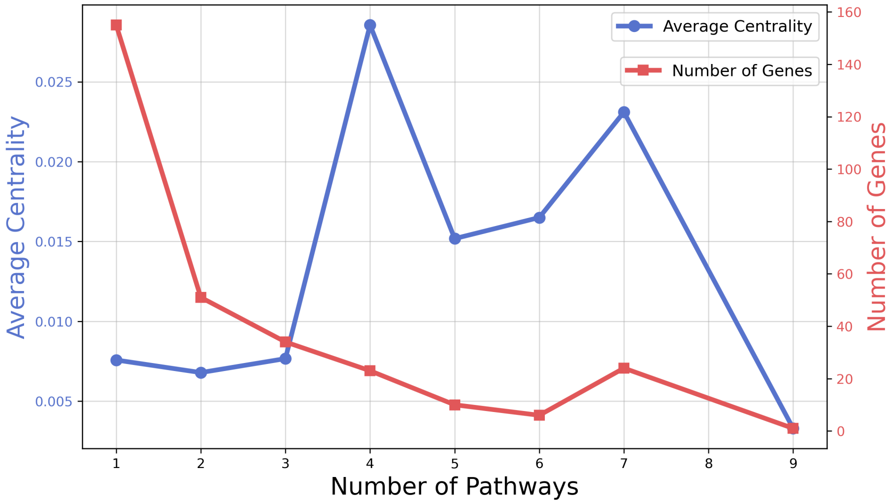


**Figure 4.5: Orthologues of genes involved in mutliple pathways have higher degree centrality**

The line shown in blue (left y-axis) shows that average degree centrality of nodes increases as the number of pathways that an orthologue of the gene in Arabidopsis is involved in increases. This highlights the overall similarities in the network of flowering time genes in *Brassica napus* and Arabidopsis. The red line shows that this increase happens even if the number of genes decreases, indicating the convergent nature of the network on few key regulators. The only gene that is involved in nine pathways in *FLOWERING LOCUS T* (*FT*), whose orthologues are not expressed in shoot apical meristem (*FT* protein is transported to shoot), hence are not inferred as important nodes in the network.

### 4.4.3 Short-day, cold treatment likely induces changes in the network structure in *Brassica napus*

Zhongshuang 11 (ZS11) is a semi-winter type *Brassica napus* cultivar. It was given a 3 weeks long short-day, cold temperature treatment for the plants to undergo vernalisation before floral transition, as shown in Figure 3.2. The effect of this treatment is possibly visible in the inferred network from this timeseries, shown in Figure 4.6, as unlike the inferred network from Stellar timeseries, this network does not seem to converge on a few key genes in the centre. This is evident from the lack of any nodes in the centre of the visualised network. If the network inference is performed using data from timepoints after the plants had finished vernalising, the inferred network returns to a structure similar to the one observed in the network inferred from Stellar timeseries, as shown in Figure 4.7. This shows that data from timeperiod when plants were subjected to cold, short day conditions for vernalisation is causing the observed differences in the inferred network structures between these two cultivars.

There are differences in genes that are identified as top nodes based on degree centrality in the network. Unlike the inferred network from Stellar timeseries, network in Figure 4.6 only has two orthologues of a floral integrator, *SOC1* among nodes with high centrality, in contrast to the network in Figure 4.4, where majority of nodes were either floral integrators or genes directly regualting them. These differences also exist when the comparison is extended to the set of top 25 genes, shown in Table 4.2. Network inferred from data from only post vernalisation timepoints does have orthologues of a key floral integrator *LFY* among its top nodes in addition to *SOC1*.

Orthologues of Arabidopsis circadian clock pathway genes associated with flowering time in response to temperature changes, *REVEILLE 2* (*RVE2*) (or *CIRCADIAN 1* (*CIR1*)) [209] and *PHYTOCLOCK 1* (*PCL1*) [210] are among the most important nodes in the network inferred from ZS11 data. An orthologue of *FLOWERING LOCUS C*, a key gene from vernalisation pathway; and *ALTERED PHLOEM DEVELOPMENT* (*APL*) and *HEME ACTIVATOR PROTEIN HOMOLOG 2B* (*HAP2B*), upstream regulators of *FT* [211] [212], are identified as key genes as well.

A further observation that can be made from this visualisation of a limited section of the whole large network is that nodes representing genes constituting similar pathways are present in clusters. For instance, orthologues of floral

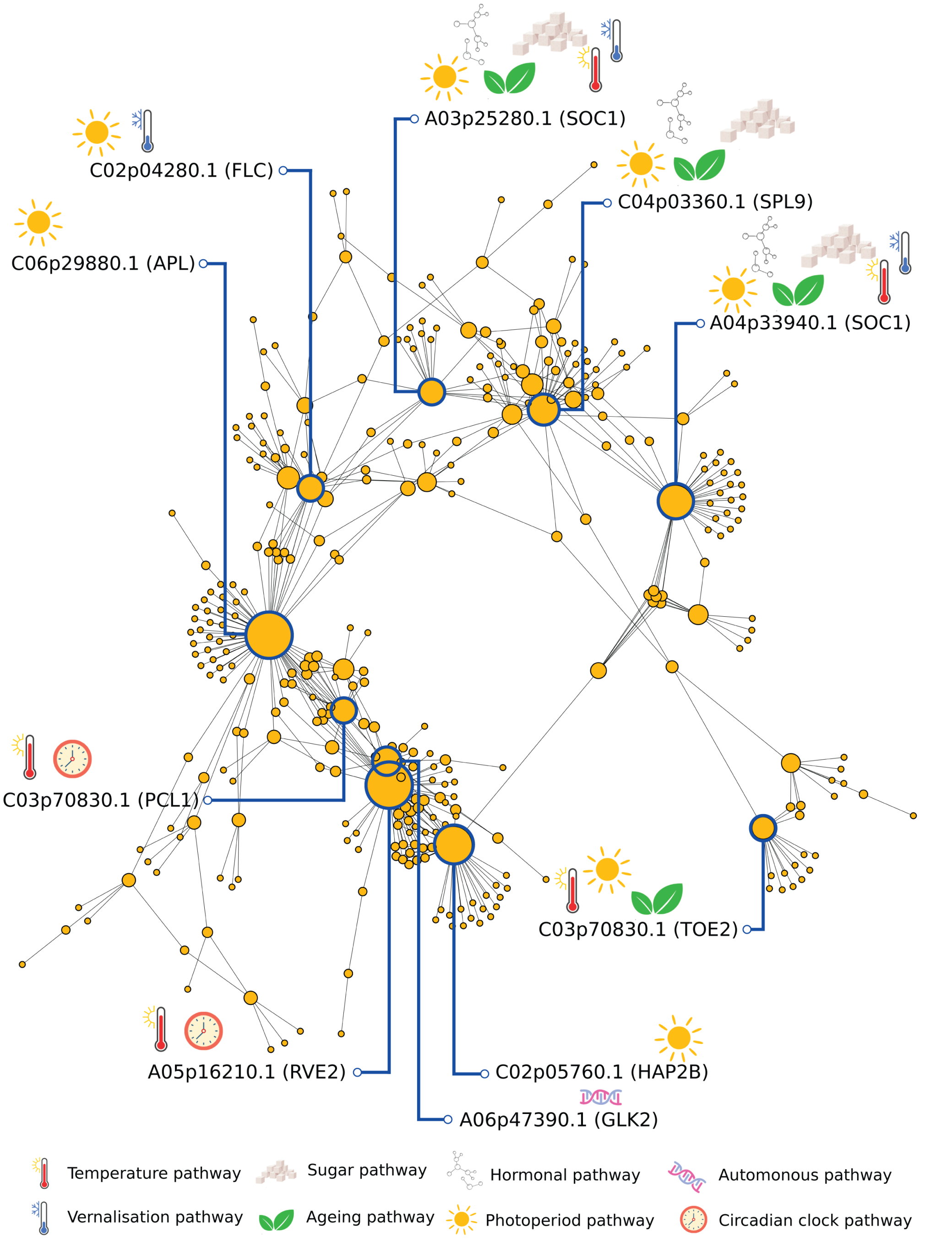


**Figure 4.6: Gene regulatory network of flowering time genes in *Brassica napus* cv. ZS11**

The largest interconnected component from the filtered network inferred from ZS11 timeseries data. Top ten nodes by degree centrality are labelled alongwith their Arabidopsis orthologue and the pathways they are known to be involved in. It shows that the network does not follow the same structure as network inferred from Stellar data, as nodes with high degree centrality are not floral integrators and the network does not converge towards a single central hub.

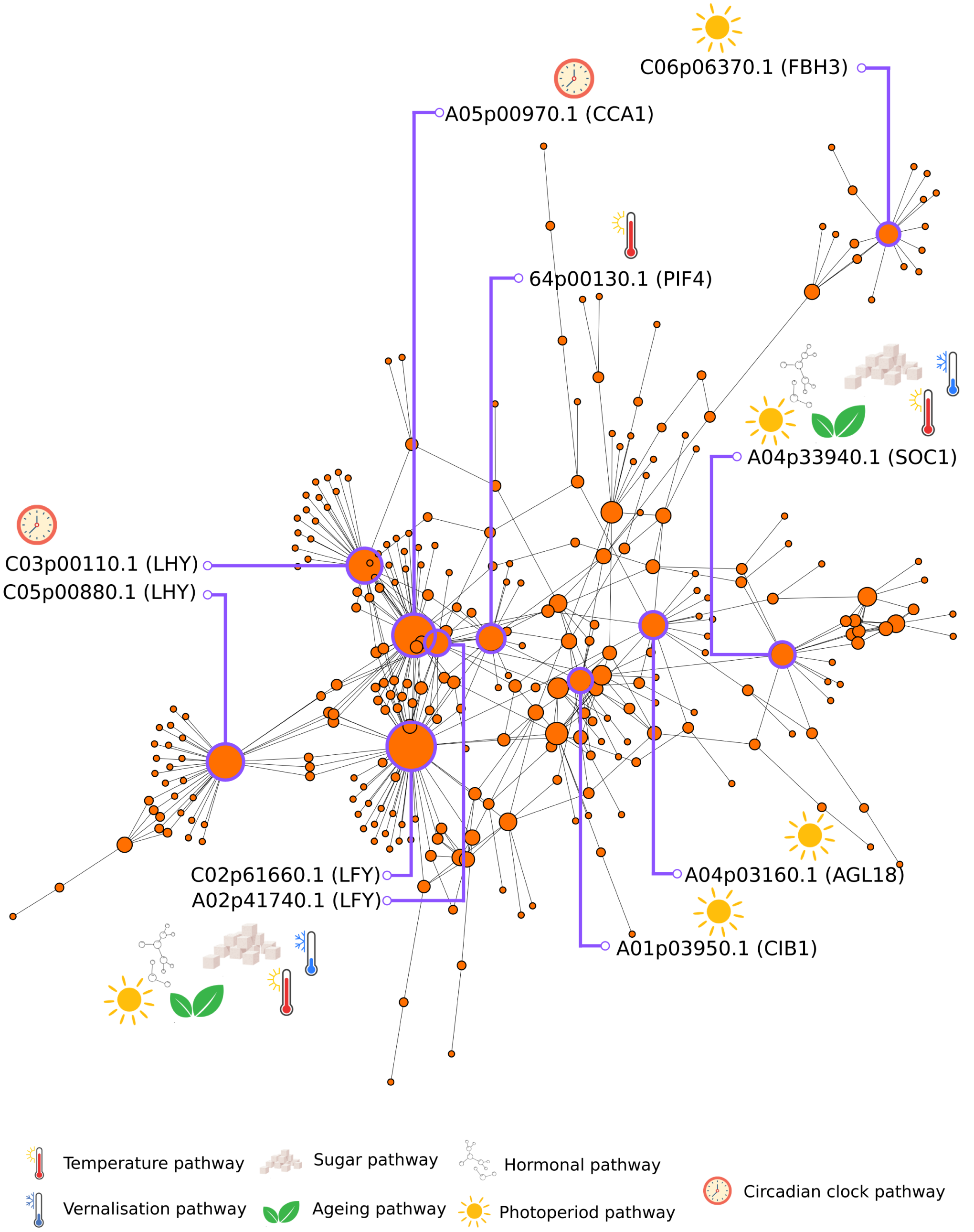


**Figure 4.7: Gene regulatory network of flowering time genes inferred only using data from post vernalisation timepoints in *Brassica napus* cv. ZS11**

By using only datapoints post vernalisation, the network structure resembles a single global cluster again. However, there are still differences to the network inferred from Stellar, with most important nodes not representing floral integrator genes.

| *Brassica napus* gene | Centrality rank | Arabidopsis id | Symbol | Pathway (s) |
|---|---|---|---|---|
| C06p29880.1 | 1 | AT1G79430 | APL | Photoperiod |
| A05p16210.1 | 2 | AT5G37260 | RVE2 | Temperature, Circadian clock |
| C02p05760.1 | 3 | AT5G12840 | HAP2A | Photoperiod |
| A04p33940.1 | 4 | AT2G45660 | SOC1 | Floral integrator |
| C04p03360.1 | 5 | AT2G42200 | SPL9 | Hormone, Aging, Sugar |
| A06p47390.1 | 6 | AT5G44190 | GLK2 | Autonomous |
| C02p04280.1 | 7 | AT5G10140 | FLC | Photoperiod, Vernalization |
| A03p25280.1 | 8 | AT2G45660 | SOC1 | Floral integrator |
| C03p70830.1 | 9 | AT3G46640 | PCL1 | Temperature, Circadian clock |
| A02p06580.1 | 10 | AT5G60120 | TOE2 | Temperature, Aging, Photoperiod |
| A10p27400.1 | 11 | AT5G10140 | FLC | Photoperiod, Vernalization |
| A02p41740.1 | 12 | AT5G61850 | LFY | Floral intergrator |
| A06p20850.1 | 13 | AT3G46640 | PCL1 | Temperature, Circadian clock |
| A01p26150.1 | 14 | AT3G46640 | PCL1 | Temperature, Circadian clock |
| C04p63380.1 | 15 | AT2G33810 | SPL3 | Hormone, Aging, Photoperiod, Sugar |
| C07p05420.1 | 16 | AT2G17770 | ATBZIP27 | Temperature, Hormone, Aging, Photoperiod, Sugar |
| A09p47980.1 | 17 | AT3G54990 | SMZ | Temperature, Hormone, Aging, Photoperiod, Vernalization |
| A05p18570.1 | 18 | AT1G53160 | SPL4 | Hormone, Aging, Photoperiod, Sugar |
| A09p35240.1 | 19 | AT1G30970 | SUF4 | Vernalisation |
| A03p16730.1 | 20 | AT5G10140 | FLC | Photoperiod, Vernalization |
| A03p04430.1 | 21 | AT5G10140 | FLC | Photoperiod, Vernalization |
| C03p65260.1 | 22 | AT5G61850 | LFY | Floral integrator |
| A05p05620.1 | 23 | AT2G45660 | SOC1 | Floral integrator |
| C03p31440.1 | 24 | AT2G28550 | RAP2.7 | Temperature, Hormone, Aging, Photoperiod, Vernalization |
| C04p69560.1 | 25 | AT2G42200 | SPL9 | Hormone, Aging, Sugars |

**Table 4.2: Top 25 nodes by degree centrality in the network inferred from ZS11 timeseries**

| *Brassica napus* gene | Centrality rank | Arabidopsis id | Symbol | Pathway (s) |
|---|---|---|---|---|
| C02p61660.1 | 1 | AT5G61850 | LFY | Floral integrator |
| A05p00970.1 | 2 | AT2G46830 | CCA1 | Circadian clock |
| C05p00880.1 | 3 | AT1G01060 | LHY | Circadian clock |
| C03p00110.1 | 4 | AT1G01060 | LHY | Circadian clock |
| 64p00130.1 | 5 | AT2G43010 | PIF4 | Temperature |
| A04p03160.1 | 6 | AT3G57390 | AGL18 | Photoperiod |
| A01p03950.1 | 7 | AT4G34530 | CIB1 | Photoperiod |
| A02p41740.1 | 8 | AT5G61850 | LFY | Floral integrator |
| A04p33940.1 | 9 | AT2G45660 | AGL20 | Floral integrator |
| C06p06370.1 | 10 | AT1G51140 | FBH3 | Photoperiod |
| C04p06570.1 | 11 | AT2G45660 | AGL20 | Floral integrator |
| C04p69560.1 | 12 | AT2G42200 | SPL9 | Hormone, Aging, Sugar |
| C04p03360.1 | 13 | AT2G42200 | SPL9 | Hormone, Aging, Sugar |
| A04p32040.1 | 14 | AT2G42200 | SPL9 | Hormone, Aging, Sugar |
| A08p16750.1 | 15 | AT4G34530 | CIB1 | Photoperiod |
| C03p85700.1 | 16 | AT4G34530 | CIB1 | Photoperiod |
| A07p03180.1 | 17 | AT2G17770 | BZIP27 | Temperature, Hormone, Aging, Photoperiod, Sugar |
| A04p26040.1 | 18 | AT2G33810 | SPL3 | Hormone, Aging, Photoperiod, Sugar |
| C03p64450.1 | 19 | AT5G62430 | CDF1 | Photoperiod |
| C09p58740.1 | 20 | AT5G18240 | MYR1 | Photoperiod |
| A02p42300.1 | 21 | AT5G62430 | CDF1 | Photoperiod |
| A01p26970.1 | 22 | AT1G54440 | RRP6L1 | Autonomous |
| C02p14050.1 | 23 | AT5G57660 | ATCOL5 | Photoperiod |
| A07p19090.1 | 24 | AT2G28550 | RAP2.7 | Temperature, Hormone, Aging, Photoperiod, Vernalization |
| C06p29880.1 | 25 | AT1G79430 | APL | Photoperiod |

**Table 4.3: Top 25 nodes by degree centrality in the network inferred using data from post vernalisation timepoints from ZS11 timeseries**

integrators (and the closely related *SPL9* gene) from a cluster (or a network community) of nodes towards the top right of the network in Figure 4.6, while genes that are influenced to a greater extent by temperature and photoperiod changes form a community further away from those nodes. Even within this community, *FLC* orthologue and its adjacent genes are separate from nodes representing orthologues of *PCL1* and *RVE2*. This is different from network inferred from Stellar data, as that network is structured more akin to a single community of nodes (Figure 4.4).

I calculated the average clustering coefficient, which is the mean of the number of triangles each node is part of in a network. If a network has higher clustering coefficient, it means the network has a single, sparse global cluster while a lower clustering coefficient in comparison indicates more a community structure made up of smaller, local clusters. For the networks inferred from Stellar and ZS11, the average clustering coefficients are 0.11 and 0.05 respectively. This clearly indicates that the network from ZS11 data is split into communities of nodes compared to Stellar, which follows a global structure similar to the flowering time network known in Arabidopsis.

### 4.4.4 Orthologues of floral integrator *SOC1* form part of separate communities in ZS11 network

To explore and infer the community structure of the whole network inferred from ZS11 data (976 genes; 20,314 edges), I employed the fluid commmunities [198] community detection algorithm. The maximum modularity method [199] partitioned the network into three communities with 374, 341 and 261 genes respectively. Figure 4.8 shows that division into three commmunities has the maximum modularity value.

Floral integrator genes, namely, *SOC1*, *AP1*, *LFY*, *AGL8/FUL*, *FT* and *AGL24* are key downstream genes within the flowering time gene regulatory network. Presence of their orthologues in certain community of the network can provide insight into the role of that subnetwork within the whole network. Figure 4.9 shows scaled expression profiles of all the genes that constiute each of the three communities. Against that background of all genes, shown in grey, expression profiles of individual orthologues for each floral integrator are highlighted.

Figure 4.9 shows that most of the orthologue genes of floral integrators are present in community 2, indicating the communities 1 and 3 likely contain genes

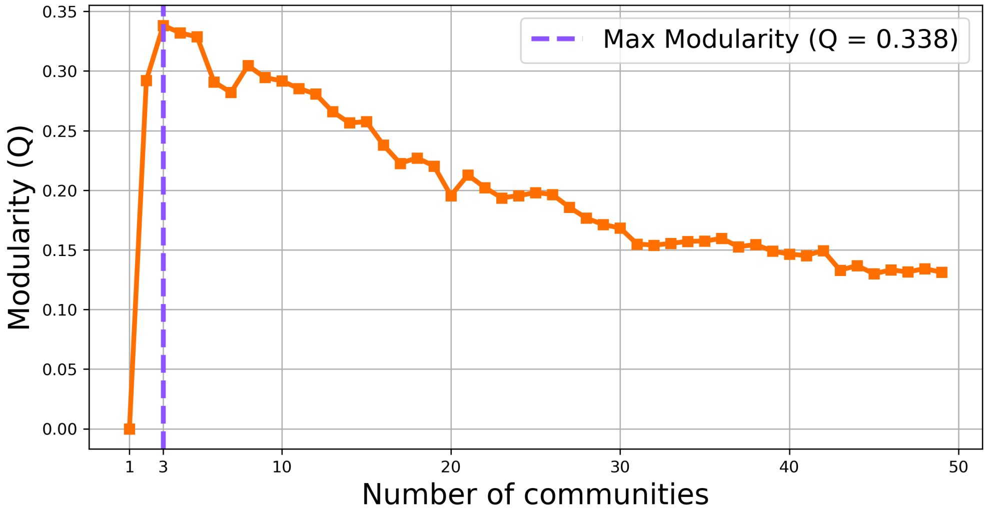


**Figure 4.8: ZS11 network can be divided into three communities using the maximum modularity method**

The plot shows the calculation of the modularity metric ($Q$) for dividing the network inferred from ZS11 data into various number of communities. The maximum $Q$ occurs when number of communities equals 3, so it was selected as the optimum number.

that constitute vernalisation or temperature related flowering control. Based on gene expression profiles, communities 1 and 3 consist of genes that undergo changes to their gene expression between days 21 and 42, which were the days plants were subjected to cold, short day conditions. Three orthologues of *AGL8*, one orthologue of *AGL24*, four orthologues of *LFY*, two orthologues of *FT*, seven orthologues of *AP1* and three orthologues of *SOC1* are present in community 2, along with genes that undergo a change in expression dynamics after vernalisation.

Interestingly, orthologues of *SOC1* are distributed in all three communities and show the most diversity in their expression patterns during vernalisation. The three orthologues of this gene present in communities 1 and 3, show substantial upregulation when plants undergo vernalisation, with their maximum expression occuring during this period. Orthologues of *SOC1* are the only set of paralogues of floral integrators that are present in all three identified communities in the network inferred from ZS11 data.

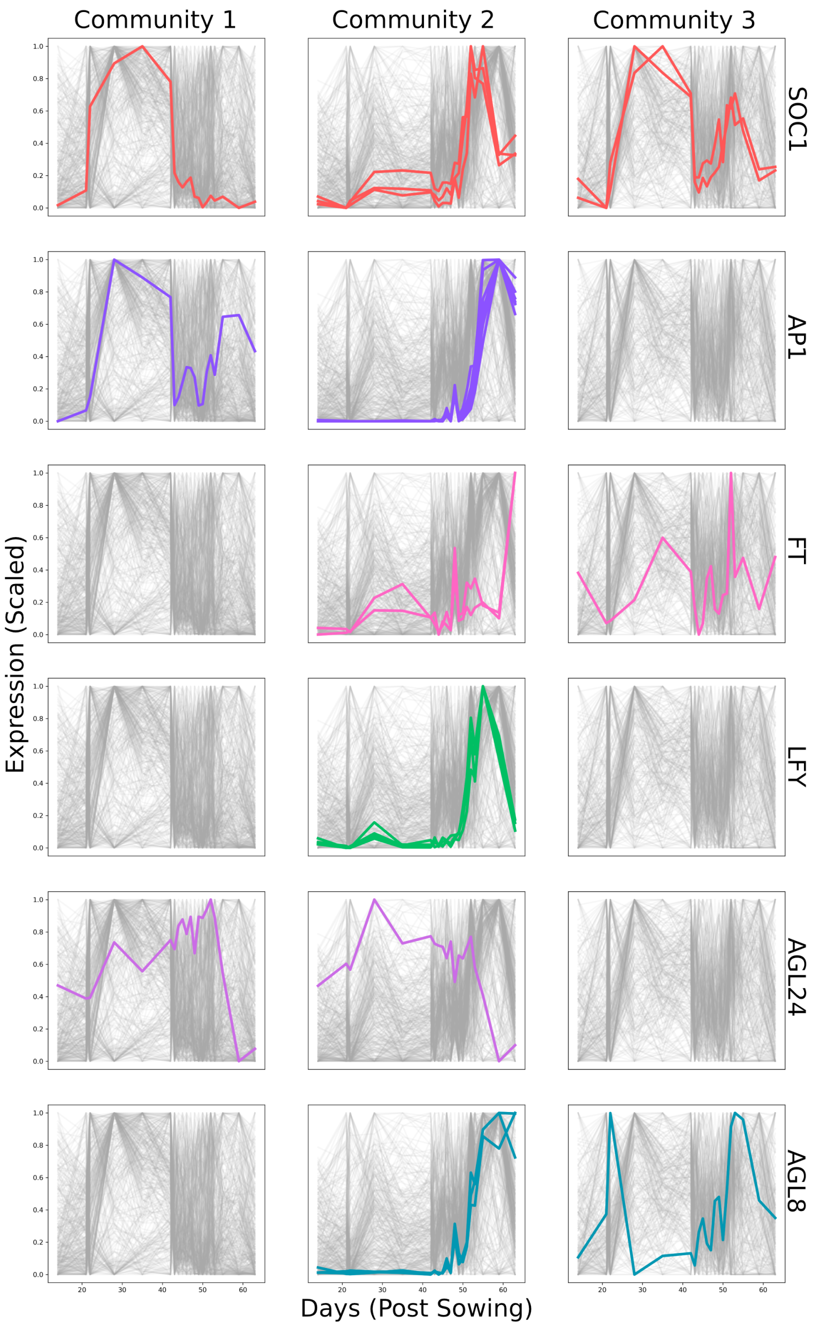


**Figure 4.9: Expression profiles of genes in the three identified communities in the network inferred from ZS11 data**

Each plot shows, in grey, expression profiles of all genes that are part of the respective community in the network. The three commmunities have 374, 341 and 261 genes respectively. These profiles show that genes in community 1 and 3 undergo expression changes during the vernalisation timeperiod, while genes in community 2 undergo expression changes post vernalisation. The coloured expressed profiles in each row highlight orthologues of floral integrator genes. Most orthologues are present in community 2, however, orthologues of *SOC1* appear in all three communities, indicating regulatory divergences.

### 4.4.5 Orthologues of *SOC1* are unique among floral integrators to have major differences in regulation between the two cultivars

I investigated if these differences exhibited by *SOC1* in ZS11 are also present in the network inferred from Stellar timeseries data. From the two inferred *Brassica napus* networks, I created a ranked list based on importance scores of all putative regulators for each orthologue of floral integrators *SOC1*, *AP1*, *LFY* and *AGL8*. Since *AGL24* only has two orthologues and mRNA expression of *FT* genes is very low to have information to infer their regulators in both networks, they were excluded from this analysis.

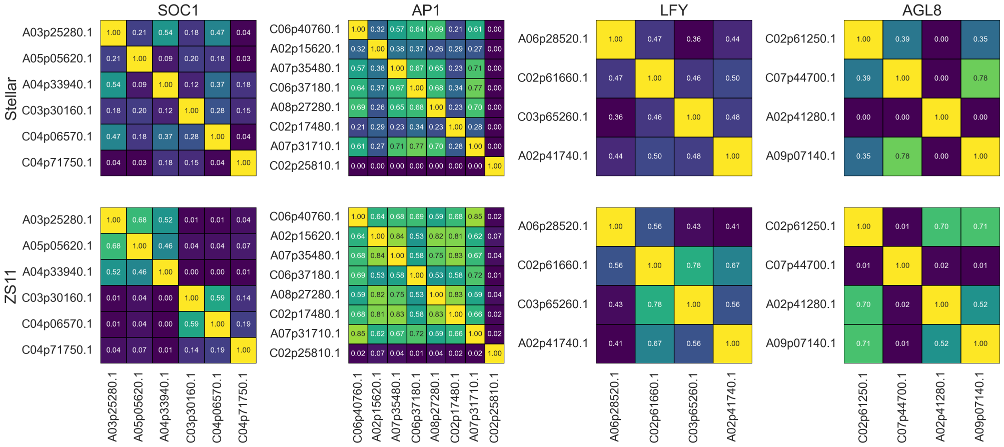


**Figure 4.10: Orthologues of *SOC1* are unique among floral integrators to have major differences in upstream regulators of different paralogues between the two *Brassica napus* networks of flowering time genes**

The heatmaps show pairwise rank-biased overlap values for regulators of orthologues of floral integrators *SOC1*, *AP1*, *LFY* and *AGL8*. Each cell in the heatmap represents the overlap between the ranking of upstream regulators of the two genes, with diagonals, representing comparison of a gene with itself, having the maximum value of 1.0. The first row corresponds to data from network inferred from Stellar data, while second row is for ZS11. The heatmaps show that unlike for *AP1*, *LFY* and *AGL8*, orthologues of *SOC1* show differences in regulation of paralogues in the network inferred from ZS11 timeseries. These differences are not as stark in the network inferred from stellar timeseries data.

I used Rank-Biased Overlap (RBO) to calculate the similaritites between ranked lists of putative regulators in pairwise comparisons among paralogues. The results, shown as heatmaps in Figure 4.10, show that orthologues of *SOC1* are unique among orthologues of floral integrators to have differences in the regulation of paralogues between networks inferred from Stellar and ZS11 timeseries data. Out of the six paralogues, in ZS11, A03p25280.1, A05p05620.1

and A04p33940.1 show very high similarities in the ranked list of their upstream regulators, however, they share no similarities with C03p30160.1 and C04p06570.1, which have very high overlaps in their ranked list of upstream regulators. These differences are not as strong in Stellar as A03p25280.1 has simlar inferred regulators as A04p33940.1, C04p06570.1 and A05p05620.1.

Furthermore, these differences among orthologues are not as strong in any other floral integrators. While there are minor differences in regulation among paralogues of *AP1*, *LFY* and AGL8, they do not exhibit cultivar specific differences observed in *SOC1* orthologues.

### 4.4.6 ***SOC1*** **orthologues exhibit different dynamics between Stellar and ZS11 timeseries**

A comparison of expression across time of orthologues of *SOC1* between Stellar and ZS11 shows the differences between the two cultivars. Shown in Figure 4.11, all six orthologues follow similar expression dynamics in Stellar, where their expression is minimum at the beginning of the timeseries. It increases as the plant develops, reaches maxima around floral transition and declines afterwards. In ZS11 however, these genes show different expression profiles. A03p25280.1 and A04p33940.1 show little change in expression levels as plants experience short day, cold conditions to undergo vernalisation. A05p05620.1 shows some upregulation, while, C03p30160.1, C04p06570.1 and C04p71550.1 have their maximum expression during vernalisation period. Following vernalisation, only A03p25280.1, A04p33940.1 and A05p05620.1 are upregulated during timepoints leading up to the floral transition. C03p30160.1 and C04p06570.1 also increase their expression levels, however, it is lower than their expression levels during the vernalisation period.

While the different paralogues also exhibit variation in expression levels, however, this is consistant between the two timeseries. The paralogues on chromosome A03 (A03p25280.1) and A05 (A05p05620.1) have the highest expression values in both Stellar and ZS11, while the ones on C04 (C04p71550.1) and A04 (A04p33940.1) have the lowest expression values.

This analysis shows that indeed *SOC1* orthologues have different dynamics in ZS11 timeseries, while their expression patterns are similar in Stellar. It is however, unclear if this is due to the short-day, cold treatment or due to differences between the two different cultivars.

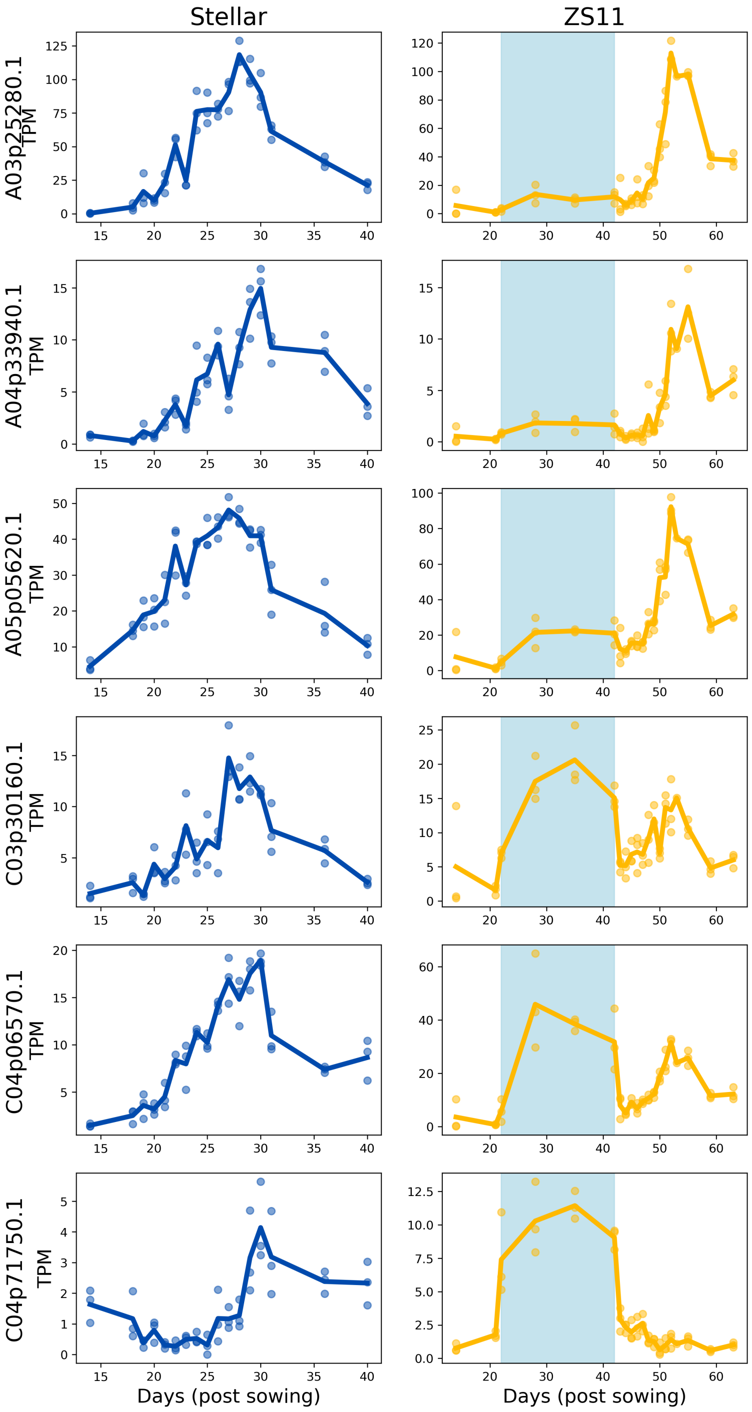


**Figure 4.11: Differences in expression of *SOC1* orthologues in Stellar and ZS11 timeseries**

Expression dynamics of six orthologues of *SOC1* in *Brassica napus* show divergent expression dynamics in ZS11, particularly during the short-day, cold treatment. The expression dynamics of the six genes is similar in Stellar.

## 4.5 Conclusion

In this chapter, I have further analysed the timeseries data presented in Chapter 3. Despite flowering time genes in *Brassica napus* having diverged in their expression dynamics when compared to their Arabidopsis orthologues, the inferred regulatory networks capture many of the same interactions that are known to be present in Arabidopsis. I show that the regulatory network inferred from Stellar forms a single global cluster, converging on orthologues of Arabidopsis floral integrators. The inferred regulatory network from ZS11 data however, does not resemble this structure, with the network divided into more local clusters of nodes instead. If network inference is done without data from the timeperiod during vernalisation, the network structure reverts back to a single cluster structure like the network inferred from Stellar. This shows the effect short-day, cold treatment has on gene dynamics, and in turn on the inferred network form that data.

Further analyses of the community structure in the network inferred from ZS11 data reveals that while most orthologues of floral integrators are present in one community within the network, orthologues of floral integrator *SOC1* appear to be present in different communities. *SOC1* orthologues are also unique among floral integrators to have very different regulation between networks inferred from Stellar and ZS11 timeseries. Their expression profiles do show differences in dynamics between Stellar and ZS11, however, it is unclear if this is due to the effect of the short day, cold treatment given to ZS11 plants, or the *SOC1* paralogues have diverged in regulation between the two cultivars.

The next chapter presents my investigations into these observed differences in regulation of *SOC1* paralogues.

# Chapter 5

# On orthologues of *SUPRESSOR OF OVEREXPRESSION OF CONSTANS 1* (*SOC1*) in *Brassica napus*

## 5.1 Abstract

The analysis of inferred networks in the semi-winter *Brassica napus* cultivar, ZS11, suggests divergence in regulation of the six orthologues of the floral integrator, *SOC1*. This is evident by the differences in their expression dynamics compared to one another in ZS11 following the start of the short-day, cold treatment for vernalisation. On the other hand, their expression dynamics and regulation appear similar in the spring type cultivar, Stellar. It is unclear whether it is the difference in cultivars that leads to these different dynamics or the environmental changes. In this chapter, I grow Stellar under the same conditions as ZS11 and show that this divergence in expression dynamics occurs in response to short-day, cold conditions. Promoter regions of *SOC1* genes show variation in binding sites for a known repressor, *CIRCADIAN CLOCK ASSOCIATED 1* (*CCA1*), that I show is upregulated under cold temperature. Based on these results I hypothesise that differences in *CCA1* binding to promoter regions of *SOC1* orthologues in *Brassica napus* leads to their divergent expression dynamics.

## 5.2 A hitchhiker's guide to *SOC1*

*"Suppressor mutation 3 is recessive and is located in a region ∼16 cM from a simple sequence length polymorphism marker nga168 on chromosome 2... The fpa mutation, which causes late flowering is the only flowering-time locus previously shown to be located in this region... experiments demonstrate that suppressor mutation 3 is not an allele of fpa"* (Extract from Hitoshi *et al*. [213])

Researchers in the Coupland group at John Innes Centre were investigating mutants that suppress the effect of overexpression of *CONSTANS* (*CO*), a photoperiod pathway gene that they had previously characterised [32]. They identified four mutations, designated as suppressor mutations 1 to 4 where flowering was delayed compared to *CO* overexpressed plants. Suppressor mutations 1 and 4 were identified as alleles of *ft*, corresponding to then recently cloned gene, *FLOWERING LOCUS T* (*FT*), introduced earlier in this thesis as the 'florigen' [35]. Suppressor mutation 2 was an allele of *fwa*, a previously identified mutation [30], later shown to correspond to the transcription factor, *FLOWERING WAGENINGEN* (*FWA*) [214]. Suppressor mutation 3, however, was identified as novel, defining a new flowering time gene, hence it was named as *suppressor of overexpression of constans 1* (*soc1*) [213].

In another experiment by the same group, to identify early targets of *CO*, they found a MADS-box transcription factor mRNA in the library enriched for cDNAs of genes that are immediate targets of *CO* [215]. First designated as *AGAMOUS-LIKE 20* (*AGL20*) in the study, the authors showed that it corresponded to the locus identified in the *soc1* mutation, and *AGL20* was renamed to *SOC1*. *SOC1* mRNA was not detected in the shoot apical meristem of plants grown under short-day conditions until they were close to floral transition, indicating that *CO* promotes flowering in part by activation of *SOC1*. The mRNA analysis also hinted at a redundant role for *SOC1* in the regulation of floral identity. Their results indicated that *SOC1*, along with *FT*, was also activated via a pathway independent of *CO*, suggesting that *SOC1* could be regulated by more than one pathway [215] [216].

These set of publications introduced *SOC1*, which is now known to be a central gene in the regulatory network controlling flowering time [50].

### 5.2.1 Introducing *SOC1*

The gene *SOC1* is present on chromosome 2 of the Arabidopsis genome. As shown in Figure 5.1, it is present on the negative strand, and according to the TAIR10 genome release [180], consists of seven coding sequence regions, encoding for a MADS-box transcription factor protein of 214 amino acids in length [53].

MADS-box proteins are named after the conserved MADS-box sequence motif, which is an acronym for MINICHROMOSOME MAINTENANCE 1, AGAMOUS, DEFICIENS and SERUM RESPONSE FACTOR, four proteins discovered with this domain [217]. Characteristic of MADS-box proteins, the SOC1 protein follows the MIKC-type structure. The first 57 amino acids from the N-terminal end in the SOC1 protein constitute the MADS-box motif, designated as the M domain, followed by Intervening (I), Keratin-like (K) and carboxyl-terminal (C) domains [53]. Research has shown that SOC1 forms a heterodimer with the AGAMOUS-LIKE 24 (AGL24) protein and is translocated to the nucleus. The M and I domains are required for this translocation [53].

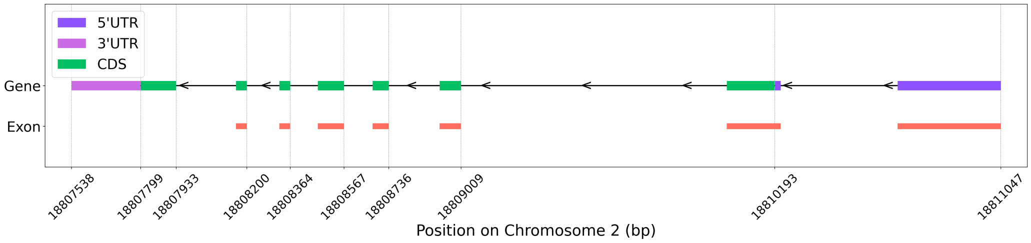


**Figure 5.1: Structure of the *SOC1* gene in Arabidopsis**

The *SOC1* gene structure in Arabidopsis TAIR10 genome annotation [180]. The gene is on the negative strand of chromosome 2 with transcription start site at position 18810193. It consists of seven coding sequence regions flanked by 5' and 3' untranslated regions.

### 5.2.2 Regulation and function of *SOC1*

*SOC1* expression has been detected in both the leaves and the shoot apical meristem [216]. Within the shoot apical meristem, its expression levels change as the plant develops. *SOC1* expression peaks just prior to the floral transition, indicating its role as a promoter of floral transition [215]. An early flowering phenotype has been shown in the *SOC1* overexpressor [216] [218], while *soc1* single mutants are late-flowering under both short-day and long-day conditions [216].

*SOC1* plays the role of a floral integrator, hence it is directly and indirectly regulated by a number of different genes. Under long-day conditions, *SOC1* is up-regulated by CONSTANS (CO). This up-regulation depends on the presence of *FLOWERING LOCUS T* (*FT*) encoded protein [219].

The *SOC1* promoter contains seven identified CArG box sites which serve as the binding sites for other MADS-box transcription factors [220]. Protein encoded by *FLOWERING LOCUS C* (*FLC*), a key gene within the vernalisation pathway, directly binds to one of these sites in the promoter to repress *SOC1* [221]. *SHOOT VEGETATIVE PHASE* (*SVP*) also encodes a protein that binds to a CArG box motif on the *SOC1* promoter [222]. Interestingly, while FLC and SVP proteins act as a dimer to regulate the expression of their common targets, there is evidence that they act independently to repress *SOC1* expression [38]. FLM$\beta$, a protein formed by alternative splicing of the mRNA transcribed from *FLOWERING LOCUS M* (*FLM*) under cold temperature acts together with SVP to form a FLM-SVP complex that also directly binds to the *SOC1* promoter to repress its expression [223]. *CIRCADIAN CLOCK ASSOCIATED 1* (*CCA1*), a gene part of the circadian clock pathway, encodes a protein that directly binds to the *SOC1* promoter to repress its activity [224]. Transcription factor encoded by *SQUAMOSA PROMOTER BINDING PROTEIN-LIKE* (*SPL9*), an aging pathway gene has been shown to bind to GTAC sites in the *SOC1* promoter and promote its expression [48].

The SOC1 protein also represses its own expression in combination with MADS domain proteins encoded by genes *APETALA* (*AP1*) and *AGAMOUS* (*AG*) as part of a negative autoregulatory loop, while *SEPALLATA 3* (*SEP3*) has been shown to repress it independently after floral transition [220]. Another floral meristem identity gene, *APETALA 2* (*AP2*) encodes a protein that binds to the *SOC1* promoter and represses its expression, likely following floral transition [225].

*SOC1* expression is induced by AGAMOUS-LIKE 24 (AGL24), a MADS-box transcription factor, which also functions as a floral integrator [218]. FRUITFUL or AGAMOUS-LIKE 8 (FUL/AGL8), another floral integrator that acts in a redundant manner with SOC1 to promote flowering also binds to the *SOC1* promoter to promote its expression [226]. LFAFY (LFY), also a floral integrator, has been shown to bind to the *SOC1* promoter [55]. LFY likely forms part of a feedback loop to repress SOC1 expression following floral transition.

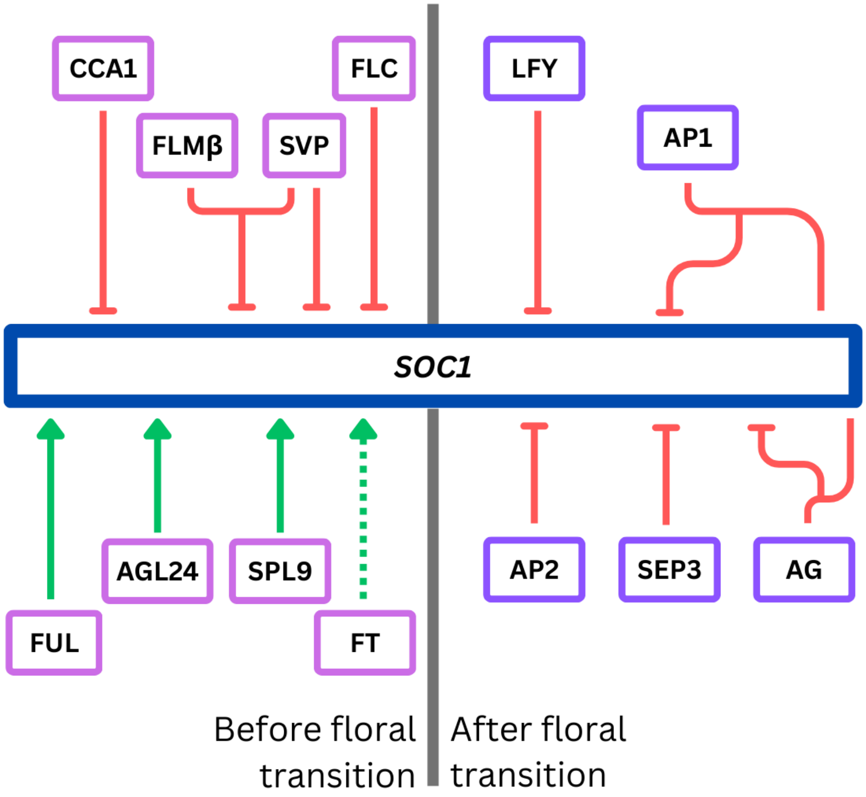


**Figure 5.2: Transcription factors known to regulate *SOC1* by directly binding its promoter**

A diagram to show the known transcription factors that bind to the *SOC1* promoter to either repress or promote its activity. Some transcription factors keep *SOC1* repressed prior to floral transition, while positive regulators of flowering have been shown to promote its expression. *SOC1* is repressed following floral transition via transcription factors that act independently and in a feedback loop with SOC1. FT acts through indirect mechanisms, hence, it is shown with a dotted line.

| TF | Effect | Binding region | Binding site |
|---|---|---|---|
| FLC | Repression | 940–950 | CTTATTTTGG |
| SVP | Repression | 920–1033 | CCAAAAATAGC |
| FLM$\beta$ | Repression | 639–969 | TCTTTTCTATTTTATTTCTTG |
| CCA1 | Repression | 1044–1153 | AAAAATCT |
| FUL | Promotion | 657–752 | CATCTATTTGTGTGT |
| AGL24 | Promotion | 1149–1261 | CTATATTGG |
| SPL9 | Promotion | 209–463 | GTAC |

**Table 5.1: Transcription factors regulating *SOC1* expression prior to floral transition through direct binding**

Transcription factors (TFs) that have been shown to directly bind and regulate *SOC1* expression prior to floral transition and their binding sequences. Binding positions are positions on the negative strand with start codon = 1 in TAIR10 genome annotation. Modified from the original by Sri *et al*. [108].

Besides transcription factors directly binding to the *SOC1* promoter, expression of *SOC1* is also under epigenetic [227] [228] and post transcriptional regulation [229]. In conclusion, the regulation of *SOC1* is complex, and it plays a central role in the gene regulatory network controlling flowering time. Further research in Arabidopsis would likely unravel even more regulatory links. Figure 5.2 depicts all the transcription factors that are known to regulate *SOC1* through direct binding as discussed in this section. Table 5.1 summarises the subset of transcription factors that have known binding sequences within the *SOC1* promoter and regulate its expression leading up-to floral transition.

### 5.2.3 The orthologues of *SOC1* in *Brassica napus*

A review of flowering time genes in *Brassica napus* [92], published at the end of 2020, remarked that "*the second most important floral integrator, SUPRESSOR OF OVEREXPRESSION OF CONSTANS 1 (SOC1) is not well studied in [Brassica napus]. No data is available . . . for its six copies*". While there are a lot of unknowns about the six orthologues of *SOC1* in *Brassica napus*, two studies provide some information about this key regulator. A study by Sri *et al*., presented an analysis of the promoter regions of *SOC1* in *Brassica* spp [108]. The authors focussed on *Brassica juncea*, an allotetraploid relative of *Brassica napus* with a commmon progenitor diploid species (Chapter 1; Figure 1.3). Their analysis revealed that despite high sequence identity among coding sequences of *SOC1* orthologues, their expression levels were different across tissues in *Brassica juncea*. The authors showed that divergences in transcription factor binding sites are present across Brassica species, including *Brassica napus* and could be responsible for divergences in expression and regulation of *SOC1* orthologues.

The second report, by Matar *et al*., showed that not all orthologues of *SOC1* are involved in floral transition under cold conditions [230]. They studied the winter-type *Brassica napus* cultivar, Express, and observed that only two paralogues, one on chromosome A05 and another on chromosome C04 had an up-regulated expression as plants transitioned to flowering under cold, long-day conditions. The authors later also reported that winter-type *Brassica napus* cultivars have a 598-bp InDel variation in the promoter region of the *SOC1* orthologue on A05 chromosome [231], again suggesting divergences in upstream promoter regions among orthologues of *SOC1* in *Brassica napus*.

In Chapter 4, I showed that *SOC1* paralogues have divergences in upstream regulators that causes diverged expression patterns over time in the shoot apical meristem in the semi-winter cultivar, Zhongshuang 11 (ZS11) (Figures 4.10 and 4.11). These differences in regulation and expression pattern, however, were not present in the spring type cultivar, Stellar. This necessitates further investigation into whether this is a cultivar specific behaviour or caused by the short-day, cold vernalisation treatment given to ZS11 plants. Based on the results presented in the above-mentioned studies, it can be hypothesised that this is an environmental effect and if the spring-type cultivar is subjected to the same treatment, *SOC1* orthologues would diverge in their expression patterns. It can be further hypothesised that this expression divergence could be due to differences in binding sites for transcription factors that directly affect *SOC1* repression or up-reguation.

In the following sections, I present an RNA-Seq timeseries experiment to investigate expression of *SOC1* orthologues in Stellar under short-day, cold conditions. I also investigate expression patterns of known upstream regulators as summarised in Table 5.1 and perform another RNA-Seq experiment to understand this divergence in gene expression further.

## 5.3 Methods

### 5.3.1 Plant growth and sampling

All plants were sown in cereals mix (40% medium grade peat, 40% sterilised soil, 20% horticultural grit, 1.2 kg/m$^3$ PG mix 14–16–18 + Te base fertilizer,

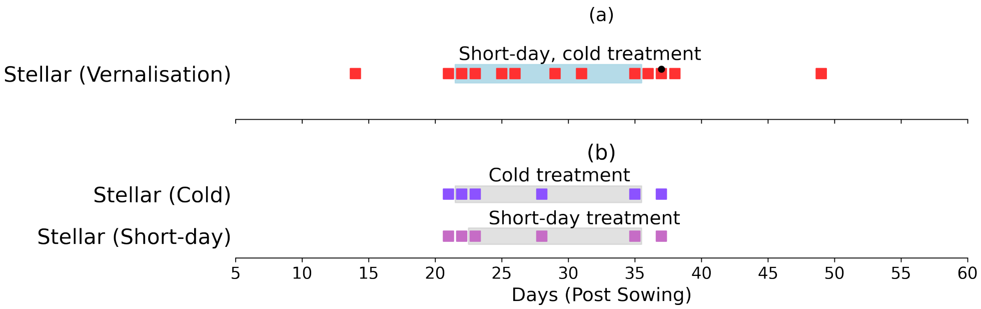


**Figure 5.3: Sampled timepoints for *B. napus* cv Stellar under different conditions**

Shoot apical meristem was sampled for RNA-Seq at the timepoints, indicated by a square. ‘Days’ refers to days post sowing. (a) The time of floral transition is indicated by ‘•’. Short-day, cold treatment refers to 5°C, 8 h photoperiod. (b) Cold treatment refers to 5°C, 16 h photoperiod. Short-day treatment refers to 18°C/15°C, 8 h photoperiod. Timing of the floral transition was not monitored.

3 kg/m$^3$ maglime, 300 g/m$^3$ Exemptor). Following germination, each seedling was transplanted to a 5 cm x 5 cm x 4.5 cm cell in a standard 24 cell-tray. Relative humidity was maintained at 70 %.

For the Stellar under short-day, cold treatment timeseries, plants were grown in 18°C/15°C, 16 h photoperiod conditions for three weeks, following which, plants were shifted to 5°C, 8 h photoperiod conditions for vernalisation treatment for two weeks. Following this treatment, plants were shifted back to 18°C/15°C, 16 h photoperiod conditions. They were germinated and grown in a Conviron MTPS 144 controlled environment room with Valoya NS1 LED lighting (250 $\mu$mol/m$^2$s). Sampled timepoints are shown in Figure 5.3 (a). Three biological replicates, each consiting of three pooled shoot apical meristems were taken at each sampled timepoint. Shoot apical meristems were monitored under the microscope to detect if they had undergone floral transition.

For the photoperiod and temperature change timeseries experiment, Stellar plants were grown in Hettich 1700 plant growth cabinet under 18°C/15°C, 16 h photoperiod conditions for three weeks, following which, half of the plants were shifted to 5°C, 16 h photoperiod conditions for the cold treatment. The other half of plants were shifted to 18°C/15°C, 8 h photoperiod conditions for the short-day treatment. Following two weeks of these two treatments, plants were shifted back to 18°C/15°C, 16 h photoperiod conditions. Figure 5.3 (b) shows the sampled timepoints. Three biological replicates, each consiting of three pooled shoot apical meristems were taken at each sampled timepoint.

### 5.3.2 RNA extraction and sequencing

Manufacturer's instructions were followed for RNA extraction from shoot apical meristem samples using the EZNA ® Plant RNA Kit (Omega Bio-tek Inc.). RNA samples were processed at Novogene, with library preparation using the NEB next ultra directional library kit (New England Biolabs). This protocol is the same as used in Chapter 3.

### 5.3.3 Processing sequencing data

FastQC (version 0.11.9) [139] was used for initial quality control of the sequencing data. MultiQC (version 1.29) [175] was used for data collation. Trimmomatic (version 0.39) [141] was used to remove any adapter contamination from reads, with the following flags, 'ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:10:1:true' to remove adapters, 'HEADCROP:15' to trim the first 15 bases of the reads, 'SLIDINGWINDOW:4:15' to trim reads if the average quality within a window of 4 bases falls below 15 and 'MINLEN:50' to drop any reads below 50 base pairs in length. The trimmed sequencing data was aligned to the Darmor-v10 reference genome [176]. HISAT2 (version 2.1.0) [148] was used for alignment. Samtools (version 1.9) [177] was used for sorting, indexing and generating a QC report. The resultant BAM files were filtered to only keep uniquely mapped reads. StringTie (version 2.1.1) [152] was used for quantification of reads aligned to different genes in the genome. 'Transcripts per million mapped reads' or TPM was selected as the measure for gene expression [178]. These steps are the same as detailed in Chapter 3.

### 5.3.4 Multiple Sequence Alignment

For promoter alignment, sequences up to 2000 base pairs upstream of the gene sequences were isolated from the TAIR 10 [180] and Darmor-v10 [176] genome assemblies for Arabidopsis and *Brassica napus* respectively. Protein sequences were also obtained from the same assemblies. MAFFT online server [232] was used to perform the alignment in both cases using the FFT-NS-i method, with automatic inference of sequence direction. Jalview (version 2.11.5.0) [233] was used for visualation and calculation of pairwise identities between sequences from the alignments.

## 5.4 Results and Discussion

### 5.4.1 Observations on the floral transition and RNA-Seq data

Stellar, the spring type *Brassica napus* cultivar, introduced in Chapter 3, was subjected to two weeks of short-day, cold treatment. Plants were grown in 18°C/15°C, 16 h photoperiod conditions for three weeks, following which, they were shifted to 5°C, 8 h photoperiod conditions for short-day, cold treatment. After two weeks, plants were shifted back to the initial conditions. Throughout this timeperiod, the shoot apical meristems were dissected and regularly monitored for the floral transition. Figure 5.4 shows images of the shoot apical meristem under a light microscope. The meristem looks clearly vegetative until day 34, indicated by the smooth, dome-like appearance and absence of any floral primodia. On day 36, one day after plants were shifted out of short-day, cold conditions, some floral primodia appear visible. These become clear on day 37, marking the transition from vegetative to floral state for the shoot apical meristem. The short-day, cold treatment delays this floral transition as Stellar has been recorded to undergo floral transition at day 30 under normal conditions (Chapter 3; Figure 3.2).

The RNA-Seq data obtained from this experiment was processed using the pipeline described in Chapter 3, Figure 2.1. Figure 5.5 shows the quality control checks done on individual sequence files post trimming. There seems to be no adapter contamination and the quality scores are also within acceptable limits. The data does have sequence duplication warnings, however, it is expected as *Brassica napus* genome is highly duplicated. The sequence length distribution warning is expected after trimming as trimmed reads have variations in length. These warnings are similar to the QC warnings reported for the data presented earlier in Chapter 3. As discussed in Chapter 3 these variations, do not affect the downstream alignment. The sequence files here also have the same warnings for GC content and Per tile sequence quality. These are also similar to the previous data (See Chapter 3; Figure 3.6) and are within acceptable levels for gene expression quantification.

Table 5.2 shows the alignment statistics for each replicate in the timeseries. On average, 31,393,388 reads passed the QC and 27,204,976 average reads were mapped as pairs to the genome with a consistent mean alignment rate of ∼86 %. Out of the total mapped reads, ∼82 % of reads were uniquely mapped

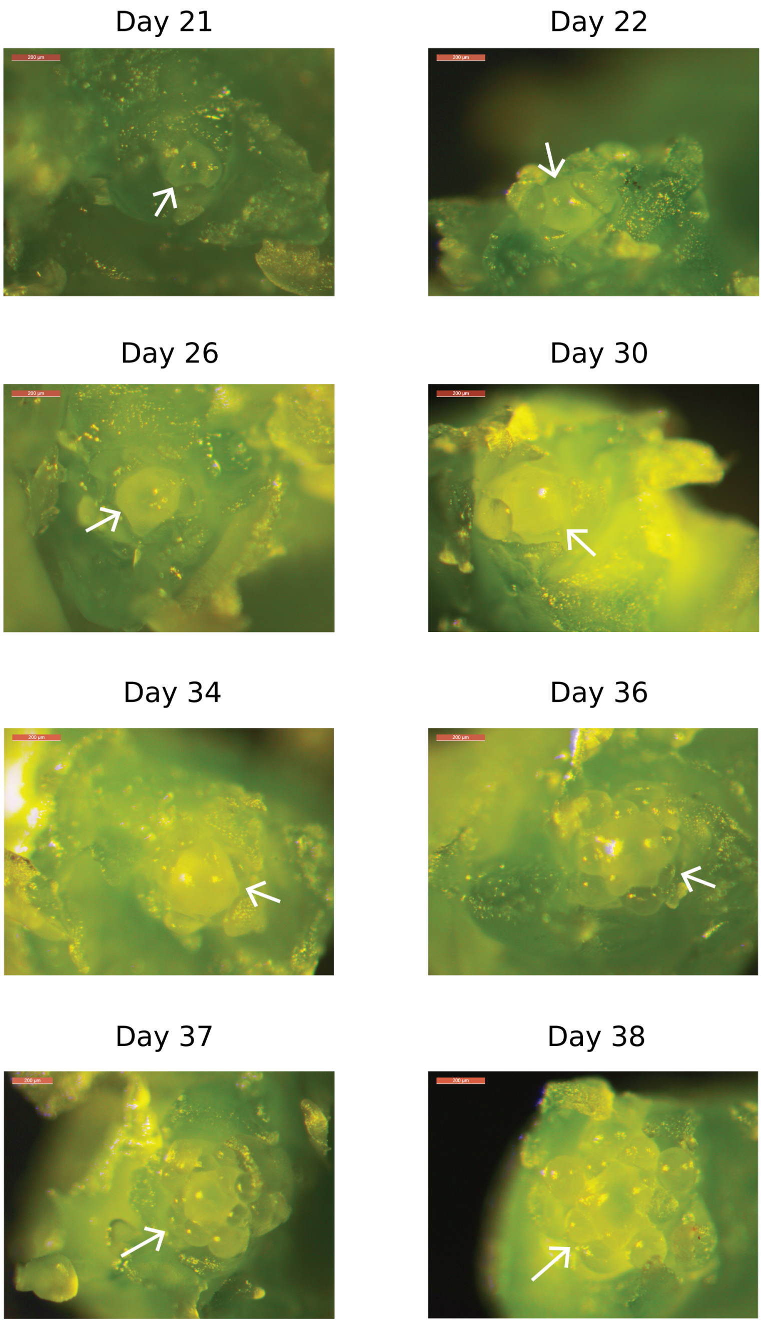


**Figure 5.4: The Stellar apex turns floral at day 37 after two weeks of short-day, cold treatment**

Images of dissected shoot apical meristems under a light microscope at various time-points. The shoot apical meristem, indicated by an arrow, appears vegetative until day 37 when floral primodia become clearly visible, marking the floral transition.

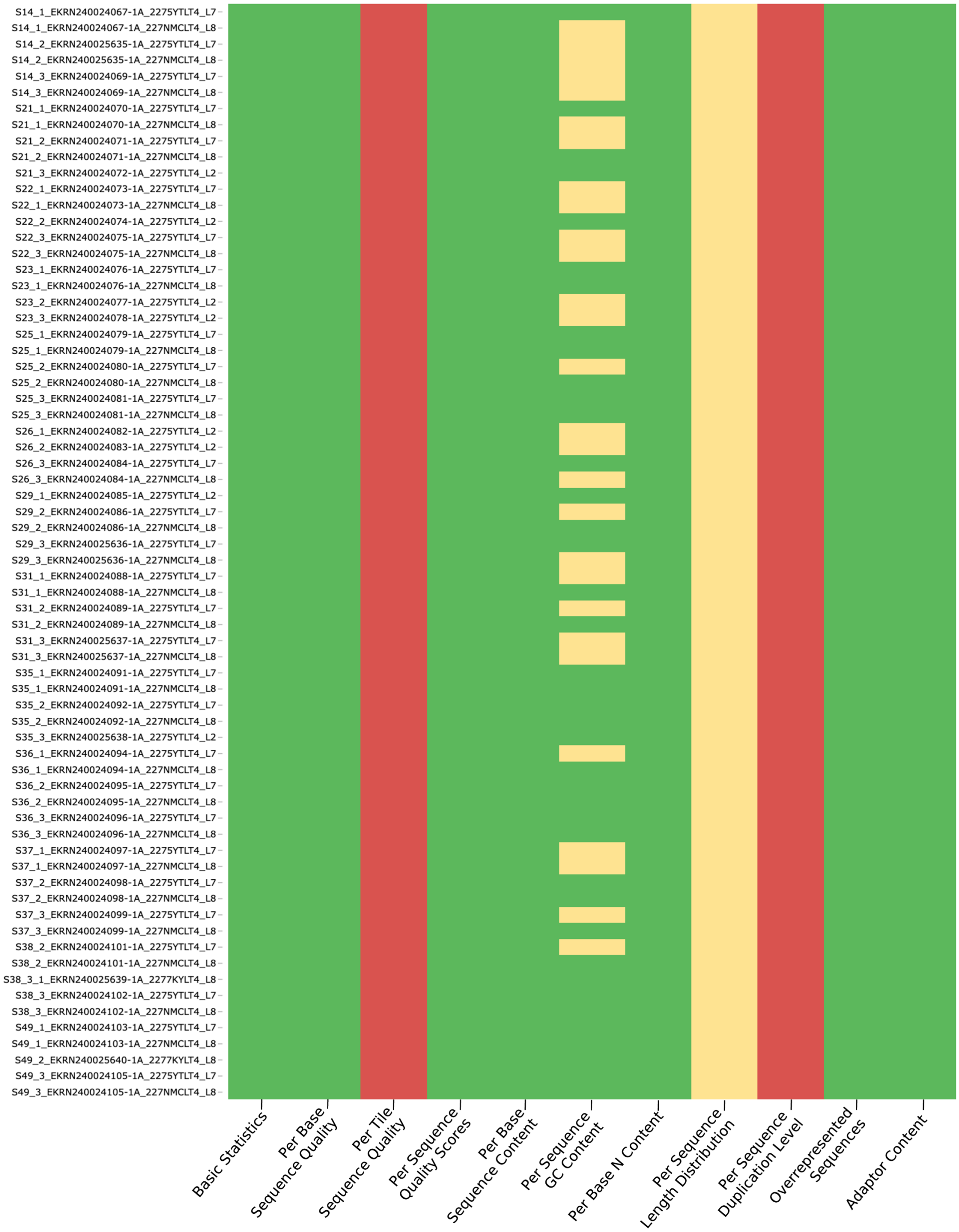


**Figure 5.5: Heatmaps showing an overview of various QC checks on trimmed RNA-Seq files for Stellar timeseries**

The QC analysis shows absence of any adapter contamination or overrepresented sequences. The sequence duplication warning is expected due to duplicated nature of the *Brassica napus* genome. Samples with Per Tile Sequence Quality along with few samples with GC content warning were further investigated, as detailed in the checks done for the data in Chapter 3. Similarly, this data was within the acceptable standards for gene expression.

| Timepoint (Days) | Replicate | Total Reads (Raw) | Total Reads (QC passed) | Mapped Reads (Total) | Mapped Reads (Unique) | Mapped Reads (Multi) | Alignment Rate (Total, %) | Alignment Rate (Unique, %) |
|---|---|---|---|---|---|---|---|---|
| 14 | 1 | 30178662 | 29868960 | 25868162 | 24293235 | 1574927 | 86.61 | 81.33 |
| 14 | 2 | 30140586 | 29888488 | 25516565 | 24016142 | 1500423 | 85.37 | 80.35 |
| 14 | 3 | 31748305 | 31428381 | 27343573 | 25746431 | 1597142 | 87.00 | 81.92 |
| 21 | 1 | 29453769 | 29132822 | 25625086 | 24188455 | 1436631 | 87.96 | 83.03 |
| 21 | 2 | 35325843 | 34856207 | 30053285 | 28505678 | 1547607 | 86.22 | 81.78 |
| 21 | 3 | 32163441 | 31832268 | 27882753 | 26228398 | 1654355 | 87.59 | 82.40 |
| 22 | 1 | 32875775 | 32540353 | 28405314 | 26813377 | 1591937 | 87.29 | 82.40 |
| 22 | 2 | 35123821 | 34761904 | 31066745 | 29369855 | 1696890 | 89.37 | 84.49 |
| 22 | 3 | 32612796 | 32296901 | 28680141 | 27066045 | 1614096 | 88.80 | 83.80 |
| 23 | 1 | 30663396 | 30384318 | 26052468 | 24608155 | 1444313 | 85.74 | 80.99 |
| 23 | 2 | 31002636 | 30685102 | 26172107 | 24686139 | 1485968 | 85.29 | 80.45 |
| 23 | 3 | 31404525 | 31092666 | 26774977 | 25276725 | 1498252 | 86.11 | 81.29 |
| 25 | 1 | 31420572 | 31100933 | 26519075 | 25019533 | 1499542 | 85.27 | 80.45 |
| 25 | 2 | 32254357 | 31938148 | 27027742 | 25367460 | 1660282 | 84.63 | 79.43 |
| 25 | 3 | 35829873 | 35535358 | 30423833 | 28725428 | 1698405 | 85.62 | 80.84 |
| 26 | 1 | 37030518 | 36648458 | 31157873 | 29329925 | 1827948 | 85.02 | 80.03 |
| 26 | 2 | 34547279 | 34195054 | 29974216 | 28363693 | 1610523 | 87.66 | 82.95 |
| 26 | 3 | 32330566 | 32020559 | 27814166 | 26263701 | 1550465 | 86.86 | 82.02 |
| 29 | 1 | 38570606 | 38177086 | 33503006 | 31707733 | 1795273 | 87.76 | 83.05 |
| 29 | 2 | 26883341 | 26635929 | 23433201 | 22219480 | 1213721 | 87.98 | 83.42 |
| 29 | 3 | 30944267 | 30623811 | 25982853 | 24524892 | 1457961 | 84.85 | 80.08 |
| 31 | 1 | 30822223 | 30531993 | 26865782 | 25422575 | 1443207 | 87.99 | 83.27 |
| 31 | 2 | 30714067 | 30399753 | 26314654 | 24881642 | 1433012 | 86.56 | 81.85 |
| 31 | 3 | 31481518 | 31145993 | 26558079 | 25074272 | 1483807 | 85.27 | 80.51 |
| 35 | 1 | 30896082 | 30581703 | 26182804 | 24737789 | 1445015 | 85.62 | 80.89 |
| 35 | 2 | 32213032 | 31924309 | 27553472 | 26032475 | 1520997 | 86.31 | 81.54 |
| 35 | 3 | 29067759 | 28754558 | 25031415 | 23713149 | 1318266 | 87.05 | 82.47 |
| 36 | 1 | 30792159 | 30477227 | 26290131 | 24822900 | 1467231 | 86.26 | 81.45 |
| 36 | 2 | 32171627 | 31841500 | 27630040 | 26162538 | 1467502 | 86.77 | 82.16 |
| 36 | 3 | 31476922 | 31159577 | 26827466 | 25363068 | 1464398 | 86.10 | 81.40 |
| 37 | 1 | 30686278 | 30436063 | 25733572 | 24272413 | 1461159 | 84.55 | 79.75 |
| 37 | 2 | 29738063 | 29394686 | 25813098 | 24465113 | 1347985 | 87.82 | 83.23 |
| 37 | 3 | 30944889 | 30617821 | 26785196 | 25327299 | 1457897 | 87.48 | 82.72 |
| 38 | 1 | 31179305 | 30923316 | 27045157 | 25568310 | 1476847 | 87.46 | 82.68 |
| 38 | 2 | 28385363 | 28113180 | 24412029 | 23093501 | 1318528 | 86.83 | 82.14 |
| 38 | 3 | 29623101 | 29349248 | 25981098 | 24551562 | 1429536 | 88.52 | 83.65 |
| 49 | 1 | 31767498 | 31458136 | 27385413 | 25821238 | 1564175 | 87.05 | 82.08 |
| 49 | 2 | 31832896 | 31577498 | 27234232 | 25668075 | 1566157 | 86.25 | 81.29 |
| 49 | 3 | 30319830 | 30011880 | 26073288 | 24634983 | 1438305 | 86.88 | 82.08 |

**Table 5.2: Alignment statistics for the Stellar (Vernalisation) RNA-Seq timeseries**

Samples of the Stellar (vernalisation) timeseries (Figure 5.3 (a)) show consistent alignment rates with an average of 86.8 %. In every sample, the majority of reads are uniquely mapped, with an average of 81.84 % out of total mapped reads. These reads were used for gene expression quantification.

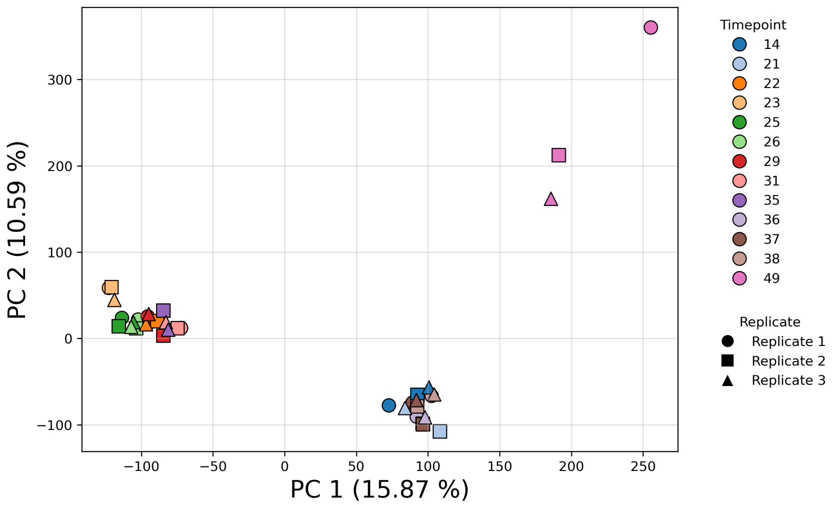


**Figure 5.6: Principal Component Analysis (PCA) of timeseries samples based on gene expression**

PCA analysis shows that sample replicates for each timepoint cluster together, indicative of consistency between the replicates. Timepoints that were collected when the plants were given short-day, cold treatment cluster away from the other timepoints along both principal components. Samples from timepoint 49 are present further away from other samples due to the large time difference.

across samples on average. These uniquely mapped reads were used for expression quantification.

Principal component analysis of the samples based on gene expression shows consistency between replicates for every timepoint, as shown in Figure 5.6. Samples taken from days when plants were subjected to the short-day, cold treatment (day 22–35) are clustered further away from other timepoints, suggesting a major effect of the environmental treatment on gene expression in the shoot apical meristem. Replicates from timepoint 49 are clustered further away due to a difference in development stage. The plants are likely closer to producing floral buds at that timepoint than floral transition. Crucially, this analysis shows that replicates for each timepoint are consistent, and the data can be used to make inferences based on gene expression.

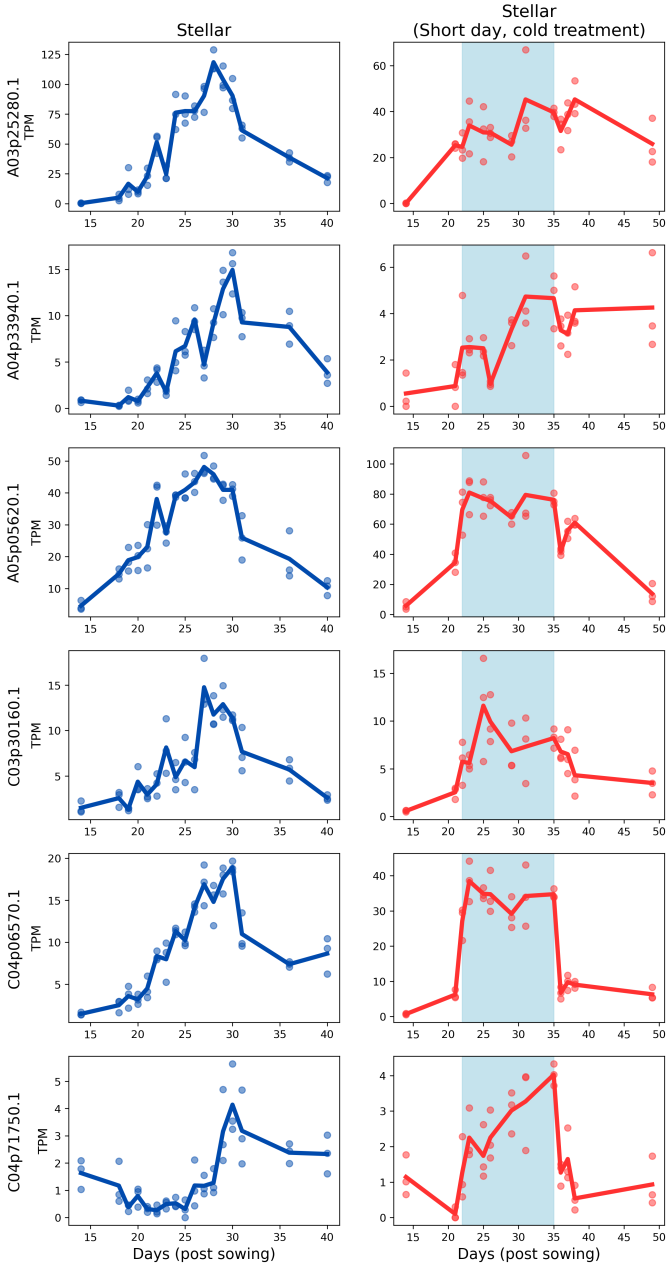
### 5.4.2 *SOC1* orthologues have divergent responses to short-day, cold treatment

As presented in Chapter 4, orthologues of Arabidopsis *SOC1* in the semi-winter type *Brassica napus* cultivar ZS11, were clustered in different communities within the inferred network (Chapter 4; Figure 4.9), had differences in inferred regulators (Chapter 4; Figure 4.10) and showed divergent expression dynamics, particularly when plants underwent vernalisation in short-day, cold conditions (Chapter 4; Figure 4.11). I did not observe these differences in the data or the inferred network from spring-type *Brassica napus* cultivar, Stellar.

In order to ascertain if this was an environmental effect or due to differences between the two genotypes, I sampled shoot apical meristems over time from Stellar with a two-week short-day, cold treatment similar to that of the timeseries sampled from ZS11. Figure 5.7 shows the expression of the orthologues of *SOC1* in *Brassica napus* cv. Stellar from the previously sampled timeseries (detailed in Chapter 3; Figure 3.2) and the timeseries with short-day, cold treatment, as detailed in Figure 5.3 (a). The *SOC1* orthologues have very clear differences in dynamics between the two timeseries, indicating that the differences in *SOC1* regulation are indeed in response to short-day, cold treatment rather than due to differences between the two cultivars.

The *SOC1* orthologue on the chromosome A03, A03p25280.1, was the highest expressed copy when Stellar and ZS11 plants underwent floral transition (Chapter 4; Figure 4.11). However, this paralogue is not up-regulated under short-day, cold conditions and is no longer the highest expressed copy when the Stellar plants undergo floral transition under vernalisation conditions (Figure 5.7 (b)). Interestingly, its expression after the plants were taken out of vernalisation conditions returns to become the highest, as indicated by the samples on day 49.

The *SOC1* orthologue on chromosome A05, A05p05620.1 also had high expression levels, but lower than A03p25280.1 in the previously sampled Stellar and ZS11 timeseries. However, during the short-day, cold conditions, it was up-regulated and had higher expression than A03p25280.1. These levels however, are down-regulated following return to normal growth conditions with A03p25280.1 having higher expression than A05p05620.1 at day 49. In the report by Matar *et al*. [230], where a winter-type *Brassica napus* cv. Express was given long-day, cold treatment until it underwent floral transition, A05p05620.1 was the highest expressed *SOC1* orthologue, followed by its C genome homeologue, C04p06570.1. This suggests that A03p25280.1 is the main paralogue driving floral transition under normal growth conditions, while under cold stress conditions, A05p05620.1 is responsible for floral transition. When the cold



## Figure 5.7: *SOC1* orthologues exhibit divergent expression dynamics in response to short-day, cold conditions in Stellar

*SOC1* orthologues have similar expression dynamics in Stellar under normal conditions, however, when plants are shifted to short-day, cold conditions for vernalisation, paralogues show differences in their expression profiles. The short-day, cold conditions are indicated by the blue colour. The solid line indicates the mean expression with individual replicates shown as scatter points.

stress conditions are no longer present, the regulatory framework works to down regulate A05p05620.1 expression. This inference of course depends on the condition that both orthologues are performing the same function, analogous to their Arabidopsis *SOC1* orthologue. It is also possible that their functions have diverged.

The expression of the third A genome *SOC1* orthologue, A04p33940.1, is low in all sampled datasets. The three C-genome paralogues, C03p30160.1, C04p06570.1 and C04p71550.1 have their expression maxima during the period of short-day, cold treatment. This is similar to the dynamics observed in ZS11 timeseries, shown in Figure 4.11, where the same three paralogues had rapid regulation after plants were shifted to short-day, cold conditions; with their maximum expression occuring within that period.

This data shows that indeed orthologues of *SOC1* have divergence in their regulation, and this divergence is only exhibited under short-day cold conditions. The individual paralogues that are up-regulated under short-day, cold conditions could be performing functions that are different to other paralogues. We do not have any structural information for Arabidopsis *SOC1*, however, differences in protein sequences compared to the Arabidopsis *SOC1* could still indicate functional divergence among its orthologues in *Brassica napus*.

Presented results so far provide a strong indication that there are regulatory differences among orthologues of *SOC1* in *Brassica napus*. An investigation of expression patterns of their upstream regulators in response to change in environmental conditions and a comparison of the upstream promoter regions for *SOC1* can help determine these regulatory differences causing these observed differences in expression dynamics.

The following sections present results from these investigations.

### 5.4.3 *SOC1* orthologues have conserved protein sequences

The *SOC1* gene in Arabidopsis encodes a MADS-box transcription factor of 214 amino acids, consisting of MADS-box (M), Intervening (I), Keratin-like (K) and the carboxyl-terminal (C) domains. The MADS-box motif, which consists of the first 57 amino acids of the protein sequence, is essential for its translocation to nucleus and its function [53]. To further investigate if different paralogues have differences in amino acids that would indicate a loss or change of function, I performed multiple sequence alignment of the protein sequences of the Arabidopsis *SOC1* and the six *Brassica napus* orthologues.

| Orthologue | Identity with Arabidopsis SOC1 | |
|---|---|---|
| | Full protein | MADS domain |
| A03p25280.1 | 94.39 % | 100.0 % |
| A05p05620.1 | 95.33 % | 98.25 % |
| A04p33940.1 | 92.52 % | 100.0 % |
| C03p30160.1 | 92.06 % | 100.0 % |
| C04p06570.1 | 95.33 % | 98.25 % |
| C04p71550.1 | 92.66 % | 100.0 % |

**Table 5.3: Identity of protein sequences of *SOC1* orthologues in *Brassica napus* to Arabidopsis SOC1**

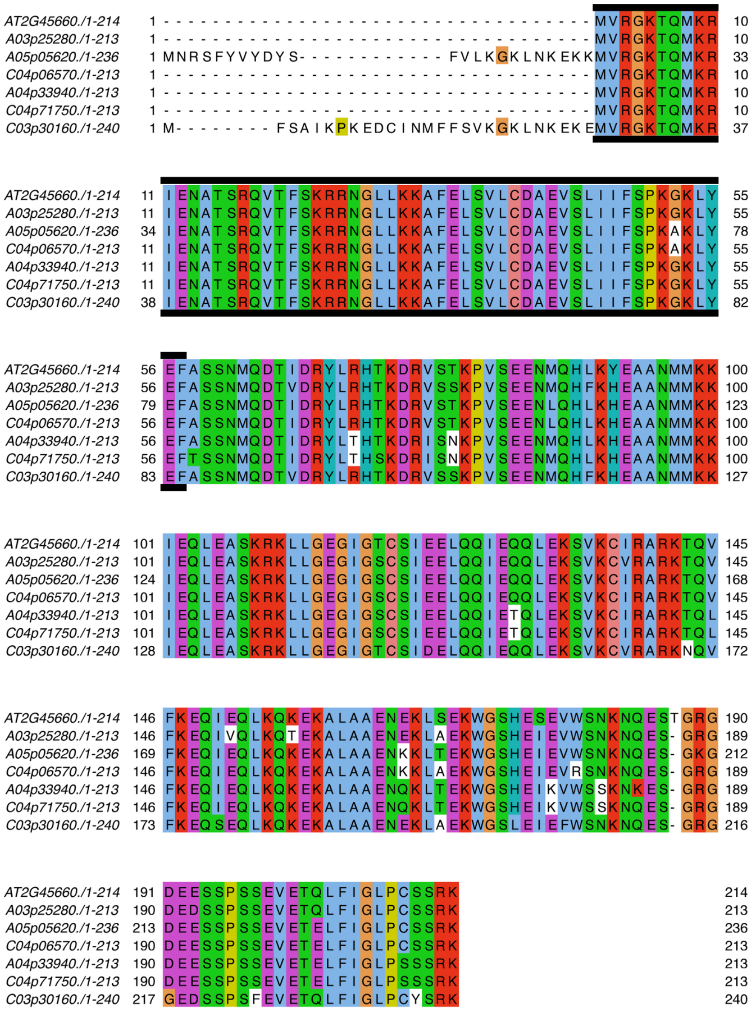
All *SOC1* proteins in *Brassica napus* have highly conserved sequences. The MADS domain, the key functional domain consisting of the first 57 amino acids in the Arabidopsis SOC1, is also highly conserved, except in two paralogues with a substitution from Glycine (G) to Alanine (A) at position 52.

The alignment, presented in Figure 5.8, shows that the protein sequences among the six paralogues and the Arabidopsis orthologue are highly conserved. There are unaligned residues at the beginning of the genes A05p05620.1 and C03p30160.1. These residues could either be due to inconsistencies in the Darmor-v10 gene annotation or a result of sequence variation. There are mismatches towards the carboxyl-terminal end too, however, as summarised in Table 5.3, the sequences consistently have greater than 90% indentity to the Arabidopsis *SOC1* protein. All sequences have identical MADS-domains, except A05p05620.1 and C04p06570.1, which have a Glycine (G) to Alanine (A) substitution at position 52 relative to the Arabidopsis sequence. Alanine has a bulkier methyl group compared to Glycine, however, structural information is needed to determine any effects on the *SOC1* protein and its function.

The current analysis hence, shows that despite the different expression patterns, there is no evidence for functional divergence between the *SOC1* paralogues in *Brassica napus*.

### 5.4.4 *CCA1*, a *SOC1* repressor, is up-regulated under short-day, cold treatment

To determine if any known regulators could be causing the expression divergence, I investigated the effect of short-day, cold treatment on the upstream



```
AT2G45660./1-214   1 ----------------------------------MVRGKTQMKR 10
A03p25280./1-213   1 ----------------------------------MVRGKTQMKR 10
A05p05620./1-236   1 MNRSFYVYDYS-----------FVLKGKLNKEKKMVRGKTQMKR 33
C04p06570./1-213   1 ----------------------------------MVRGKTQMKR 10
A04p33940./1-213   1 ----------------------------------MVRGKTQMKR 10
C04p71750./1-213   1 ----------------------------------MVRGKTQMKR 10
C03p30160./1-240   1 M-------FSAIKPKEDCINMFFSVKGKLNKEKEMVRGKTQMKR 37

AT2G45660./1-214  11 IENATSRQVTFSKRRNGLLKKAFELSVLCDAEVSLIIFSPKGKLY 55
A03p25280./1-213  11 IENATSRQVTFSKRRNGLLKKAFELSVLCDAEVSLIIFSPKGKLY 55
A05p05620./1-236  34 IENATSRQVTFSKRRNGLLKKAFELSVLCDAEVSLIIFSPKAKLY 78
C04p06570./1-213  11 IENATSRQVTFSKRRNGLLKKAFELSVLCDAEVSLIIFSPKAKLY 55
A04p33940./1-213  11 IENATSRQVTFSKRRNGLLKKAFELSVLCDAEVSLIIFSPKGKLY 55
C04p71750./1-213  11 IENATSRQVTFSKRRNGLLKKAFELSVLCDAEVSLIIFSPKGKLY 55
C03p30160./1-240  38 IENATSRQVTFSKRRNGLLKKAFELSVLCDAEVSLIIFSPKGKLY 82

AT2G45660./1-214  56 EFASSNMQDTIDRYLRHTKDRVSTKPVSEENMQHLKYEAANMMKK 100
A03p25280./1-213  56 EFASSNMQDTVDRYLRHTKDRVSSKPVSEENMQHFKHEAANMMKK 100
A05p05620./1-236  79 EFASSNMQDTIDRYLRHTKDRVSTKPVSEENLQHLKHEAANMMKK 123
C04p06570./1-213  56 EFASSNMQDTIDRYLRHTKDRVSTKPVSEENLQHLKHEAANMMKK 100
A04p33940./1-213  56 EFASSNMQDTIDRYLTHTKDRISNKPVSEENMQHLKHEAANMMKK 100
C04p71750./1-213  56 EFTSSNMQDTIDRYLTHSKDRISNKPVSEENMQHLKHEAANMMKK 100
C03p30160./1-240  83 EFASSNMQDTVDRYLRHTKDRVSSKPVSEENMQHFKHEAANMMKK 127

AT2G45660./1-214 101 IEQLEASKRKLLGEGIGTCSIEELQQIEQQLEKSVKCIRARKTQV 145
A03p25280./1-213 101 IEQLEASKRKLLGEGIGSCSIEELQQIEQQLEKSVKCVRARKTQV 145
A05p05620./1-236 124 IEQLEASKRKLLGEGIGSCSIEELQQIEQQLEKSVKCIRARKTQV 168
C04p06570./1-213 101 IEQLEASKRKLLGEGIGSCSIEELQQIEQQLEKSVKCIRARKTQV 145
A04p33940./1-213 101 IEQLEASKRKLLGEGIGSCSIEELQQIETQLEKSVKCIRARKTQL 145
C04p71750./1-213 101 IEQLEASKRKLLGEGIGSCSIEELQQIETQLEKSVKCIRARKTQL 145
C03p30160./1-240 128 IEQLEASKRKLLGEGIGTCSIDELQQIEQQLEKSVKCVRARKNQV 172

AT2G45660./1-214 146 FKEQIEQLKQKEKALAAENEKLSEKWGSHESEVWSNKNQESTGRG 190
A03p25280./1-213 146 FKEQIVQLKQTEKALAAENEKLAEKWGSHEIEVWSNKNQES-GRG 189
A05p05620./1-236 169 FKEQIEQLKQKEKALAAENKKLTEKWGSHEIEVWSNKNQES-GKG 212
C04p06570./1-213 146 FKEQIEQLKQKEKALAAENKKLAEKWGSHEIEVRSNKNQES-GRG 189
A04p33940./1-213 146 FKEQIEQLKQKEKALAAENQKLTEKWGSHEIKVWSSKNKES-GRG 189
C04p71750./1-213 146 FKEQIEQLKQKEKALAAENQKLTEKWGSHEIKVWSSKNQES-GRG 189
C03p30160./1-240 173 FKEQSEQLKQKEKALAAENEKLAEKWGSLEIEFWSNKNQES-GRG 216

AT2G45660./1-214 191 DEESSPSSEVETQLFIGLPCSSRK 214
A03p25280./1-213 190 DEDSSPSSEVETQLFIGLPCSSRK 213
A05p05620./1-236 213 DEESSPSSEVETELFIGLPCSSRK 236
C04p06570./1-213 190 DEESSPSSEVETELFIGLPCSSRK 213
A04p33940./1-213 190 DEESSPSSEVETELFIGLPSSSRK 213
C04p71750./1-213 190 DEESSPSSEVETELFIGLPSSSRK 213
C03p30160./1-240 217 GEDSSPSFEVETQLFIGLPCYSRK 240
```

**Figure 5.8: Protein sequences for Arabidopsis *SOC1* and its *Brassica napus* orthologues are highly conserved**

Multiple sequence alignment shows that all *SOC1* proteins in *Brassica napus* have highly conserved sequences. The MADS domain, the key functional domain highlighted in the figure, is also highly conserved, except in two paralogues with one substitution from Glycine (G) to Alanine (A) at position 52.

regulators of *SOC1* that are known to directly bind to the *SOC1* promoter before floral transition. These proteins, narrowed down from the wider list of putative, indirect and post-floral transition regulators, are listed in Table 5.1. I investigated expression profiles of orthologues *FLMβ*, *SVP*, *SPL9*, *FLC* and *CCA1* — the upstream regulators of *SOC1* that are not already classed as floral integrators, and were not already studied in Chapter 4. Orthologues of *SVP* and *SPL9* showed no change in expression dynamics in response to environmental conditions. *FLC* orthologues were silenced and expression of *FLMβ* orthologues was not detected in the data. However, orthologues of one of these upstream regulators, *CIRCADIAN CLOCK ASSOCIATED 1* (*CCA1*) are, interestingly, only expressed under short-day, cold conditions in *Brassica napus*.

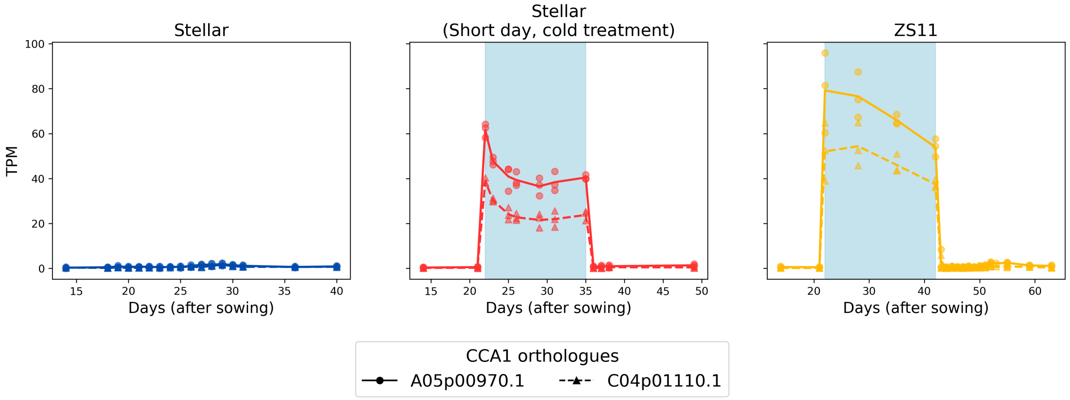


**Figure 5.9: Orthologues of *CCA1* are upregulated under short-day, cold conditions**

The expression data shows that orthologues of *CCA1* in *Brassica napus*, in both cultivars, are significantly up-regulated when plants are subjected to short-day, cold conditions.

There are two orthologues of *CCA1* in *Brassica napus*. The expression of these orthologues is very low in Stellar under normal conditions compared to when Stellar is subjected to short-day, cold conditions, as shown in Figure 5.9. In both ZS11 and Stellar, *CCA1* orthologues are rapidly up regulated as soon as environmental conditions change. A study by Lu *et al*. [224] reported significant repression in expression of *SOC1* in a *CCA1* overexpressor line. They further reported, using ChIP analysis that *CCA1* directly binds to a binding site upstream of *SOC1*. Analysis of *SOC1* promoter sequences in *Brassica napus* by Sri *et al*. [108] indicates that *SOC1* promoters might have undergone divergence in their binding sites for these regulators. In their analysis, they also determined the coordinates of CCA1 binding sequence 'AAAAATCTT' in the Arabidopsis *SOC1* promoter region.

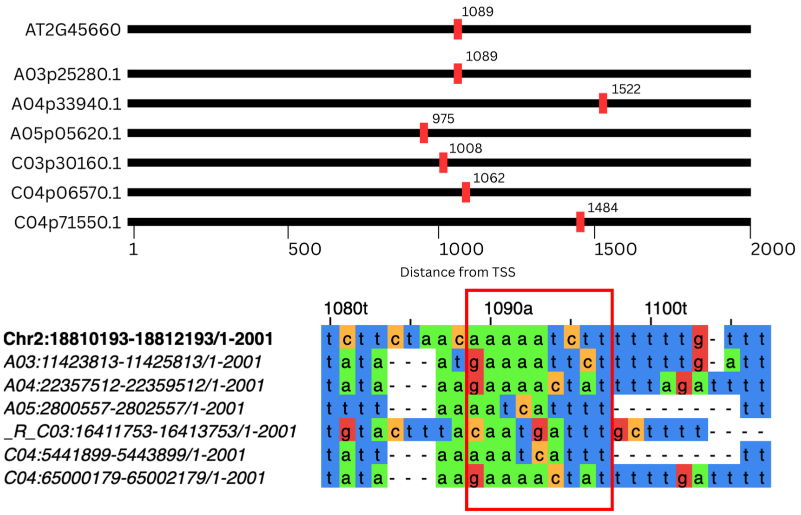

AT2G45660
1089
A03p25280.1
1089
A04p33940.1
1522
A05p05620.1
975
C03p30160.1
1008
C04p06570.1
1062
C04p71550.1
1484
1
500
1000
1500
2000
Distance from TSS
1080t
1090a
1100t
Chr2:18810193-18812193/1-2001
A03:11423813-11425813/1-2001
A04:22357512-22359512/1-2001
A05:2800557-2802557/1-2001
_R_C03:16411753-16413753/1-2001
C04:5441899-5443899/1-2001
C04:65000179-65002179/1-2001
t c t t c t a a c a a a a a t c t t t t t t t g - t t t
t a t a - - - a t g a a a a t t c t t t t t t g - a t t
t a t a - - - a a g a a a a c t a t t t t a g a t t t t
t t t t - - - a a a a t c a t t t t - - - - - - - - t t
t g t a c t t t a c a a t g a t t t g c t t t t - - - -
t a t t - - - a a a a a t c a t t t - - - - - - - - t t
t a t a - - - a a g a a a a c t a t t t t t g a t t t t

**Figure 5.10: Variations in the binding sequence for CCA1 in the promoter regions of *SOC1* orthologues in *Brassica napus***

Multiple sequence alignment of upstream regions of Arabidopsis *SOC1* and the six *SOC1* orthologues shows that variation is present in the CCA1 binding site in the upstream regions of *Brassica napus* orthologues. The labels show the exact coordinates from TAIR10 reference [180] (Chr2; for Arabidopsis) and Darmor-v10 reference [176] (A03,A04,A05,C03 (reverse complement) and C04) for the upstream regions. Notably, the trailing 'TTCT' region of the binding site has been disrupted in all sequences, with the promoter for the A03 orthologue being the most similar to the Arabidopsis reference. The postions of the sites from the transcription start site (TSS) are shown in red.

Building on these reports, I isolated 2000 bp upstream promoter regions of Arabidopsis *SOC1* (from TAIR10 reference [180]) and the six *Brassica napus* orthologues (from Darmor-v10 reference [176]) and performed multiple sequence analysis. I was not able to locate the *CCA1* binding site using position weight matrix in Arabidopsis promoter, hence, I used the multiple sequence alignment strategy also used in analysis by Sri *et al.* [108]. Focussing on the site with the CCA1 binding sequence, shown in Figure 5.10, all *SOC1* promoters have differences in the binding site compared to the Arabidopsis promoter. The sequences have variations in the trailing 'TCTT' sequence in particular, with A03p25280.1 *SOC1* orthologue having the closest match to the Arabidopsis site. The variations in this site indicate that CCA1 binding could be disrupted.

While the transcription factor binding site disruption on its own is not strong evidence, together with the expression pattern of *CCA1* orthologues, it can be hypothesised that *CCA1* up-regulation is likely responsible for differences in *SOC1* expression while the plants undergoes vernalisation. Futhermore, the variations in binding sites are linked with differences in expression dynamics of the *SOC1* orthologues. The A03p25280.1 promoter has the binding site most similar to the Arabidopsis sequence and is kept repressed under short-day, cold conditions while all other paralogues have varying degrees of upregulation coupled with variations in the CCA1 binding site.

However, vernalisation conditions involve changing both photoperiod and temperature. *CCA1* is controlled by the circadian clock, hence either of these environmental changes (or their combination), could be the reason for its upregulation. Although, to control for the effects of the circadian clock, all samples in the data presented so far were collected after 60 % of the day had passed — it is not possible to determine from the current data if it is the change in photoperiod or the change in temperature that leads to the changes in expression dynamics for the *CCA1* and, subsequently the *SOC1* orthologues. Hence, further investigation to unpick these two variables is necessary.

*CCA1* has been reported to undergo alternative splicing under low temperatures [234] and has been postulated to play a role in cold acclimation. Hence, it can be hypothesised that the changes in expression dynamics in *CCA1* and *SOC1* orthologues are due to temperature changes and not due to a change in photoperiod.

The following section tests this hypothesis.

### 5.4.5 Changes in expression dynamics of orthologues of *CCA1* and *SOC1* are temperature mediated and independent of photoperiod change

In order to investigate if the changes in expression of *CCA1* and *SOC1* orthologues in *Brassica napus*, were caused by the change in temperature or the change in photoperiod, two separate RNA-Seq timeseries, one under short-day, normal temperature conditions and the other under cold, long-day conditions, were sampled separately as shown in Figure 5.3. For the first three weeks, plants were grown under normal temperature (18 °C/15 °C) and long-day (16 hours) photoperiod conditions. Following that plants were divided in two groups and subjected to short-day (8 hours) photoperiod with normal temperature conditions and cold temperature (5 °C) with long-day (16 hour) photoperiod conditions separately. Plants were given this treatment for two-weeks and the sampled shoot apical meristems were subjected to RNA-Seq. The alignment statistics, along with details of the sampled timepoints in the two datasets are summarised in Table 5.4.

| Timepoint (Days) | Replicate | Conditions | Total Reads (Raw) | Total Reads (QC passed) | Mapped Reads (Total) | Mapped Reads (Unique) | Mapped Reads (Multi) | Alignment Rate (Total, %) | Alignment Rate (Unique, %) |
|---|---|---|---|---|---|---|---|---|---|
| 21 | 1 | 18/15 °C, 16h | 26408972 | 26116762 | 23079155 | 21949803 | 1129352 | 88.37 | 84.04 |
| 21 | 2 | 18/15 °C, 16h | 29257738 | 28948795 | 25650667 | 24320601 | 1330066 | 88.61 | 84.01 |
| 21 | 3 | 18/15 °C, 16h | 28134831 | 27819064 | 24756800 | 23554830 | 1201970 | 88.99 | 84.67 |
| 22 | 1 | 18/15 °C, 8h | 29520012 | 29197554 | 25830916 | 24567306 | 1263610 | 88.47 | 84.14 |
| 22 | 2 | 18/15 °C, 8h | 35112824 | 34689491 | 30760967 | 29285028 | 1475939 | 88.68 | 84.42 |
| 22 | 3 | 18/15 °C, 8h | 35909757 | 35457101 | 31259873 | 29731571 | 1528302 | 88.16 | 83.85 |
| 23 | 1 | 18/15 °C, 8h | 31265438 | 30906338 | 27397303 | 26044327 | 1352976 | 88.65 | 84.27 |
| 23 | 2 | 18/15 °C, 8h | 29887458 | 29557426 | 26244064 | 24791509 | 1452555 | 88.79 | 83.88 |
| 23 | 3 | 18/15 °C, 8h | 31708146 | 31345768 | 27828879 | 26496094 | 1332785 | 88.78 | 84.53 |
| 28 | 1 | 18/15 °C, 8h | 30934706 | 30558054 | 26985907 | 25595606 | 1390301 | 88.31 | 83.76 |
| 28 | 2 | 18/15 °C, 8h | 27929931 | 27564519 | 24481470 | 23202946 | 1278524 | 88.82 | 84.18 |
| 28 | 3 | 18/15 °C, 8h | 30220406 | 29872652 | 26610218 | 25135200 | 1475018 | 89.08 | 84.14 |
| 35 | 1 | 18/15 °C, 8h | 33621146 | 33220961 | 29864743 | 28322503 | 1542240 | 89.90 | 85.25 |
| 35 | 2 | 18/15 °C, 8h | 32976845 | 32689066 | 29469286 | 28001953 | 1467333 | 90.15 | 85.66 |
| 35 | 3 | 18/15 °C, 8h | 31289234 | 30913863 | 27677951 | 26283065 | 1394886 | 89.53 | 85.02 |
| 37 | 1 | 18/15 °C, 16h | 27496683 | 27146294 | 24267295 | 23018286 | 1249009 | 89.39 | 84.79 |
| 37 | 2 | 18/15 °C, 16h | 32331501 | 31935020 | 28734200 | 27271037 | 1463163 | 89.98 | 85.40 |
| 37 | 3 | 18/15 °C, 16h | 37927362 | 37449228 | 33400899 | 31681960 | 1718939 | 89.19 | 84.60 |
| 22 | 1 | 5 °C, 16h | 30739487 | 30410580 | 27053587 | 25755505 | 1298082 | 88.96 | 84.69 |
| 22 | 2 | 5 °C, 16h | 30048219 | 29705879 | 26463637 | 25233533 | 1230104 | 89.09 | 84.94 |
| 22 | 3 | 5 °C, 16h | 35905818 | 35489341 | 31826553 | 30213908 | 1612645 | 89.68 | 85.14 |
| 23 | 1 | 5 °C, 16h | 37898621 | 37472593 | 33146158 | 31521672 | 1624486 | 88.45 | 84.12 |
| 23 | 2 | 5 °C, 16h | 35049527 | 34674034 | 30490549 | 29002808 | 1487741 | 87.93 | 83.64 |
| 23 | 3 | 5 °C, 16h | 33429065 | 33013191 | 27957743 | 26603351 | 1354392 | 84.69 | 80.58 |
| 28 | 1 | 5 °C, 16h | 38291212 | 37834882 | 33087184 | 31334112 | 1753072 | 87.45 | 82.82 |
| 28 | 2 | 5 °C, 16h | 30252551 | 29863873 | 26709085 | 25440420 | 1268665 | 89.44 | 85.19 |
| 28 | 3 | 5 °C, 16h | 30229217 | 29863011 | 26651599 | 25253858 | 1397741 | 89.25 | 84.57 |
| 35 | 1 | 5 °C, 16h | 31769644 | 31364530 | 28194509 | 26811361 | 1383148 | 89.89 | 85.48 |
| 35 | 2 | 5 °C, 16h | 27647080 | 27299647 | 24418218 | 23141758 | 1276460 | 89.45 | 84.77 |
| 35 | 3 | 5 °C, 16h | 30443905 | 29997079 | 26582363 | 25265910 | 1316453 | 88.62 | 84.23 |
| 37 | 1 | 18/15 °C, 16h | 36691375 | 36214062 | 32285550 | 30717887 | 1567663 | 89.15 | 84.82 |
| 37 | 2 | 18/15 °C, 16h | 28559766 | 28211752 | 25464681 | 24179292 | 1285389 | 90.26 | 85.71 |
| 37 | 3 | 18/15 °C, 16h | 27110640 | 26788149 | 24266084 | 23074382 | 1191702 | 90.59 | 86.14 |

**Table 5.4: Alignment statistics for Stellar RNA-Seq timeseries under cold and short-day conditions**

Samples in Stellar timeseries show consistent alignment rate with an average of 88.93 %. In every sample, majority of reads are uniquely mapped, with an average of 84.46 %.

The expression dynamics of the *CCA1* orthologues are shown in Figure 5.11. The two *CCA1* orthologues show upregulated expression in the plants that were subjected to a change in temperature while keeping the photoperiod constant. This behaviour is similar to the upregulation in expression observed, in both cultivars, under vernalisation conditions where both the temperature and photoperiod were changed in Figure 5.9. No such upregulation is observed when the photoperiod is changed while keeping the temperature the same. The expression of the two genes is low, as observed in Stellar under normal growth conditions, as shown earlier in Figure 5.9. This shows that the temperature change, independent of photoperiod change, causes upregulation in *CCA1* orthologues.

The differences in the expression dynamics of the *SOC1* orthologues also occur when only temperature is changed as shown in Figure 5.11. In plants that

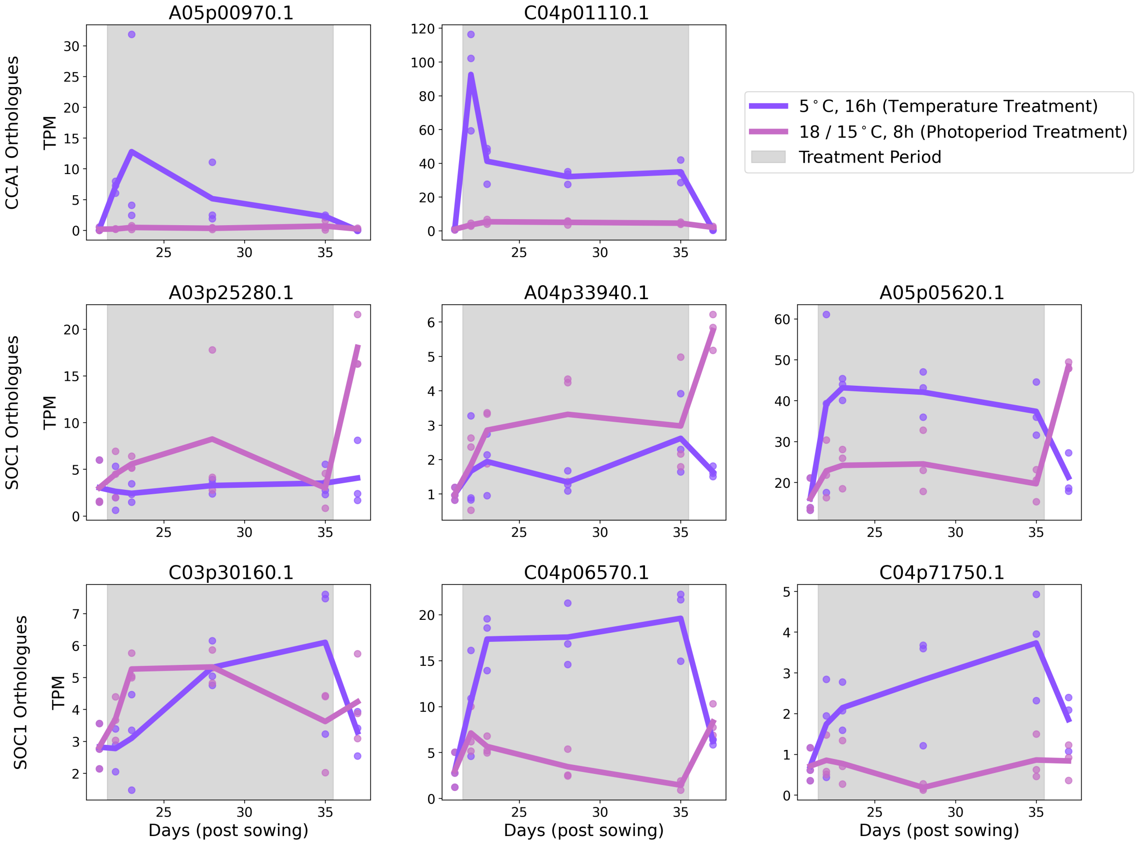


**Figure 5.11: *CCA1* orthologues are up-regulated and *SOC1* orthologues change their expression dynamics in response to change in temperature**

Expression of both *CCA1* orthologues is upregulated under cold treatment compared to short-day treatment. The lines show mean expression while individual scatter points show expression values from samples. *SOC1* orthologues show no change in their expression patterns under short-day treatment, while under cold treatment the paralogues have diverged expression dynamics. A03p25280.1 is clearly downregulated compared to its other paralogues, most of which are upregulated with a change in temperature.

were shifted to cold, long-day conditions, the *SOC1* orthologues show similar differences in dynamics to the ones observed in both cultivars when the plants were subjected to cold, short-day vernalisation conditions (Figure 5.7 and Chapter 4; Figure 4.11). As observed in earlier datasets, the A03p25280.1 is the highest expressed gene among these paralogues under normal conditions — however, its expression levels are the lowest when plants are subjected to cold. All other genes show up-regulation when plants are subjected to cold temperature, compared to no effect on their expression dynamics under short-day treatment. This establishes the fact that differences in the expression of *SOC1* orthologues is temperature mediated and not in response to a change in photoperiod.

Together the Figures 5.11 (a) and (b) show a correlation (also shown earlier using Figures 5.9 and 5.7) between an upregulation in the expression of the orthologues of *CCA1* and divergence in expression dynamics of its downstream regulated *SOC1* orthologues. These results, together with observed variations in the CCA1 binding sequence (Figure 5.10), suggest a divergence in the regulatory control of the *SOC1* orthologues in response to cold temperature.

## 5.5 Conclusion

In this chapter I have presented investigations into the divergence in the expression dynamics and regulation among *SOC1* orthologues in *Brassica napus*, first uncovered in Chapter 4 using network analysis. In the semi-winter type cultivar, ZS11, orthologues of *SOC1*, a key downstream floral integrator, exhibited divergent expression dynamics leading upto the floral transition. This was unique among the orthologues of known floral integrators and also suggested regulatory divergences among the paralogues in the inferred network. These differences appeared to be due to upregulation in expression under vernalisation conditions and were not observed in the spring type cultivar, Stellar.

In order to determine whether this was in response to vernalisation conditions or due to differences in the two cultivars, I grew Stellar under vernalisation conditions. By subjecting the plants to the same short-day, cold treatment same as ZS11, I showed that this divergence is in response to change in environmental conditions and not due to cultivar differences. I then investigated these paralogues further for possible functional and regulatory divergences. All six genes have highly conserved protein sequences, along with a conserved functional MADS-box domain, hence providing no evidence to suggest any functional differences between these *SOC1* orthologues. Investigations into expression patterns of known regulators acting immediately upstream of *SOC1* in the regulatory network revealed a correlation between up-regulation of the orthologues of a *SOC1* repressor, *CCA1* and the divergences in dynamics of *SOC1* orthologues under vernalisation conditions. My investigations, along with an already published study [108], also hint at variations in binding site for the CCA1 transcription factor in the promoter regions of these paralogues.

However, since under vernalisation conditions, both temperature and photoperiod are changed, it wasn't possible based on sampled data to determine the factor causing these changes and if they were correlated. So, I grew Stellar plants under two different conditions by changing photoperiod and temperature separately. The results show that upregulation in the expression of the orthologues of *CCA1* and divergence in expression dynamics of its downstream regulated *SOC1* orthologues occurs due to a change in temperature, and not photoperiod and is correlated.

In conclusion, these results show that there are regulatory differences among orthologues of *SOC1* in *Brassica napus*. These differences are linked to the divergences in *CCA1* mediated repression of *SOC1* in response to cold conditions. Further investigations into the variations observed in the binding sites and binding assays to determine differences in CCA1 binding to the *SOC1* orthologue promoter sequences would help validate these findings.

# Chapter 6

# The discussion

## 6.1 A brief recap of everything

This thesis began with a summary of some of the first documented efforts of researchers to understand the factors that influenced the timing of blossom in agriculturally important crops. The initial findings stemmed from efforts to reduce the number of days that certain varieties of crop plants took to reach the stage where they could produce blossom [9]. With the advent of molecular biology, efforts moved into finding genetic targets linked to the previously identified factors in a simple and quick to cultivate model system, *Arabidopsis thaliana* [29]. Research into the control of flowering time in Arabidopsis has revealed a gene regulatory network, consisting of multiple genes that form distinct pathways that converge on a few floral integrator genes. These floral integrator genes then turn on meristem identity genes that cause the plants to transition into flowering [49]. Research efforts have aimed to transfer this knowledge to agronomically important crops [60] [64] [68] [76]. Chapter 1 also introduced *Brassica napus*, a close relative to the model plant Arabidopsis and the plant at the centre of this work. The chapter finished with a brief introduction to the concepts of polyploidy, an essential part of the introduction to *Brassica napus* research as it is an allotetraploid, formed by hybridisation of two diploid species, *Brassica oleracea* and *Brassica rapa* [86]. A comparison of multiple orthologues of Arabidopsis flowering time genes and the gene regulatory network controlling flowering time in *Brassica napus* is the main focus of thesis.

Chapter 2, an extension of the introduction section, provided a quick introduction to RNA-Seq, the technique that is used to quantify gene expression values from shoot apical meristems sampled across development of *Brassica napus* plants.

Stellar, a spring type cultivar, was grown under 16 hour daylight conditions with temperatures of 18 °C during daytime and 15 °C at night. Shoot apical meristems were sampled on various days until the plants turned floral at day 30. Zhongshuang 11 (ZS11), a chinese semi-winter type cultivar, requires vernalisation to flower. The plants were initially grown in same conditions as Stellar, however, they were shifted at day 21 to short-day, 8 hour daylight with 5 °C constant temperature conditions for vernalisation. At day 35, plants were shifted back to normal conditions and underwent floral transition at day 55. Chapter 3 presented the first results of the comparison of timeseries obtained using RNA-Seq from *Brassica napus* with a similar timeseries for *Arabidopsis thaliana*, Col-0 from Klepikova *et al*. [174]. I used curve registration [150] to compare expression dynamics between orthologous pairs of genes. The results show that the majority of genes have similar expression dynamics between Arabidopsis and *Brassica napus*, highlighting the similarities between the two closely related species. However, for the subset of genes known to be involved in flowering time, significantly higher proportion of genes exhibit differences in expression dynamics instead. I also show, corroborating a previous report [90], that flowering time genes are preferentially retained in higher number of paralogues than rest of the genes in the genome. This suggests that the individual paralogues are under lower selection pressure, allowing for more plasticity and divergence in expression dynamics from their Arabidopsis orthologues. This has been shown, for example by Calderwood *et al*. [91] in the case of *FLOWERING LOCUS C*. They showed that the 9 individual paralogues in *Brassia napus* have different expression dynamics, however, their sum total amount corresponds to vernalisation requirements of different cultivar types. The results from this chapter suggest that individual paralogues diverging could be a more widespread phenomenon among orthologues of flowering time genes.

Following on from these observations, in Chapter 4, I present my investigations into the interactions between the set of flowering time genes in *Brassica napus* using network inference [192]. The inferred networks from the expression dynamics show that despite the differences in dynamics of individual paralogues, the inferred network topology is similar to the gene regulatory network known in Arabidopsis. The inferred network converges on a set of genes that are orthologues of floral integrators *APETALA 1* (*AP1*), *LEAFY* (*LFY*), *AGAMOUS-LIKE 8* (*AGL8*) and *SUPPRESSOR OF OVEREXPRESSION OF CONSTANS 1* (*SOC1*) and other genes that directly regulate them such as *SQUAMOSA PROMOTER*

*BINDING PROTEIN-LIKE* (*SPL*) family. Orthologues of genes that are involved in multiple pathways in Arabidopsis also have higher degree centrality in *Brassica napus*, further suggesting conservation of flowering control between the two species. However, interestingly, the inferred network from the ZS11 data is organised into smaller community clusters rather than converging on few orthologues of floral integrators. I show that this is due to the expression changes induced by the short-day, cold treatment that the plants experienced. Further investigations show that, orthologues of *SUPPRESSOR OF OVEREXPRESSION OF CONSTANS 1* (*SOC1*) show divergent dynamics and regulatory links in the inferred networks. This is unique among the set of floral integrator orthologues in *Brassica napus*.

Chapter 5 focusses on the six orthologues of Arabidopsis *SOC1* in *Brassica napus*. It was unclear whether the observed differences in dynamics and regulation of these paralogues was a characteristic of the semi-winter cultivar ZS11, as I observed no such differences among these genes in the data sampled from the spring type cultivar, Stellar. By growing Stellar plants under the same environmental conditions as ZS11, I show that the differences in the dynamics of these paralogues occur in response to the short-day, cold treatment. I investigated these six paralogues for any functional or regulatory divergences in detail. The genes have conserved protein sequences and my analyses resulted in no indications to suggest any functional divergences. Investigations into expression patterns of known regulators acting immediately upstream of SOC1 in the regulatory network revealed a correlation between up-regulation of the orthologues of a SOC1 repressor, *CIRCADIAN CLOCK ASSOCIATED 1* (*CCA1*) and the divergences in dynamics of SOC1 orthologues under vernalisation conditions. I also found variations in the sequence of the binding site for CCA1 in promoter regions of the six paralogues. This, in conjunction with a similar study published by Sri *et al*. [108] hints at loss of binding of CCA1 transcription factor to promoter of some *SOC1* orthologues. There are other studies that also suggest variations in promoter regions of *SOC1* orthologues [231] [230].

However, since vernalisation conditions involved changing both temperature and photoperiod it wasn't possible based on sampled data to determine the factor causing these changes. Furthermore, *CCA1* is a gene regulated by the circadian clock, so either of these factors can induce changes in its expression, which might not be correlated to changes in *SOC1* expression. By growing Stellar plants under two different conditions of changing photoperiod and temperature separately, I have shown that expression of the orthologues of CCA1 and divergence in expression dynamics of its downstream regulated SOC1 orthologues occurs due to a change in temperature, not photoperiod and is correlated.

## 6.2 A brief summary of everything

In summary, the results presented in this thesis show that orthologues of Arabidopsis genes in polyploid *Brassica napus* retain similar expression dynamics, and likely regulatory frameworks on a whole genome level. The gene regulatory network controlling flowering time also has similaritites to Arabidopsis, however, flowering time genes have been preferentially retained within the *Brassica napus* genome. This allows individual paralogues to diverge in regulation, while still maintaining the required structure for regulatory framework. This allows plasticity in the system, aiding in adaptations. The results presented in first half of the thesis — Chapter 3 and the first half of Chapter 4 — presents results that support this idea. Flowering time genes are preferentially retained, have higher degree of divergence among individual paralogues yet the inferred gene regulatory network follows same principles known from Arabidopsis.

The second half of the thesis — from the second half of Chapter 4 and Chapter 5 — presents one example of such regulatory divergence in detail. The six orthologues of the key floral integrator gene *SUPPRESSOR OF OVEREXPRESSION OF CONSTANS 1* (*SOC1*) in *Brassica napus* have similar expression dynamics under normal growth conditions. They have different expression levels, likely due to dosage balance (reviewed in Chapter 1) but the genes have their maximum expression coinciding with the floral transition. Under cold temperature however, the expression dynamics change and copies are rapidly up-regulated. Assuming all the *SOC1* orthologues are still performing the same function as the Arabidopsis gene, this suggests subtle changes to the regulatory framework, likely meaning some *SOC1* orthologues are responsible for floral transition under cold conditions, while other paralogues are more important under warmer temperatures. This suggests that different paralogues could be integrating signals from different pathways of the GRN. The same inference can be drawn from the data reported in a study by Matar *et al*. [230] where the authors studied floral transition in a winter-type *Brassia napus* under cold temperatures and reported that expression of only two orthologues coincided with floral transition under cold. Further experiments would be required to test this theory of subfunctionalisation between paralogues.

## 6.3 A brief note on contributions

This work presents a first system level analysis of flowering time regualtion in *Brassica napus*. Using transcriptomic timecourses, I highlight the conserved similarities between the diploid model plant, and the allotetraploid crop system. I observe that the sets of genes, in this case flowering time genes, that are preferentially retained in higher number of paralogues also exhibit greater divergence in their expression dynamics. This indicates that gene duplications are maintained in nature through neo- and sub-functionalisations. This allows the genes to diverge in expression dynamics while still maintaining similar overall regulatory structure. Through the results presented from my analysis of the orthologues of the gene *SOC1*, I also show that these changes can be subtle, and only observed when system is perturbed with a change in environmental conditions. This highlights that these subfunctionalisations could be useful in responses and adaptations to environmental conditions and diversification.

As highlighted in the previous chapters, along with reviews from the researchers working in *Brassica napus*, it is necessary to understand what functional each individual paralogue of a gene is performing to enable targeted breeding efforts to improvement of this crop. The data, networks and curve registration presented in this thesis would be a valuable resource for further *Brassica napus* crop improvement research.

## 6.4 A brief note on limitations

While the results and conclusions presented in this thesis have been drawn from careful analysis of data and build on results from previous reports, however, it would not be very scientific of me to report these findings to the reader without a rigorous mention of its limitations. The following sections outline these issues that could challenge the results presented in this thesis. The first section covers technical limitations, where I am able to provide steps I have taken to try to ensure that the conclusions drawn are supported by the data. The second section discusses blindspots in the data that require further investigations and validations.

### 6.4.1 Ensuring accuracy with technical limitations

The conclusions drawn in this thesis rely on bulk RNA-Seq data. The limitations of the technology and sources of potential biases have already been reviewed in Chapter 2. The polyploid nature of *Brassica napus* adds another potential source of bias in values of gene expression obtained after bioinformatic analysis of the RNA-Seq data. Inaccuracies in these results would lead to inaccuarte conclusions. Hence, steps have been taken to address these limitations in the data analysis. Exact version numbers and parameter values for software packages are reported to ensure the reproducibility of results. Detailed quality control analyses were performed, and reported in the results to carefully assess the data obtained from RNA-Seq. Only uniquely mapped reads were used to reduce any false positive mappings potentially skewing the gene expression values. While, this might cause the obtained expression values for genes to be lower than the true expression levels, a lot of downstream analyses presented in this thesis rely on the trend of expression values over time rather than absolute values. The time series nature of data also allows for additional quality control to ensure consistency between timepoints. Principal component analysis accompanies every time series dataset presented in this thesis to ensure that replicates for each timepoint are consistent and there is a clear trend where samples closer in time cluster together compared to timepoints further away over the course of plant development. As a bonus, the effects of different environmental treatments are also captured using this analysis.

A technique used throughout this thesis is the comparison of orthologues of genes between *Arabidopsis thaliana* and *Brassica napus*. The key limitation here is that there exists no standard list of mapped orthologues between the two species on a whole genome level. I used a reciprocal BLAST search based strategy to create the ‘one-to-many’ mapping between *Arabidopsis thaliana* and *Brassica napus* gene models. The key adavantage to this method is that I have the ability to outline the exact method used to ensure transparency in reporting. While there exists no ‘ground truth’ to access the accuracy of this technique, it correctly identified known orthologues of all key genes discussed in Chapters 4 and 5. Furthermore, conclusions from the overall analysis presented in Chapter 3 where flowering time genes were found to be preferentially retained corroborates a previous report by Jones *et al*. [90] where the authors used a synteny based method to map orthologues between *Arabidopsis thaliana* and an older genome assembly of *Brassica napus*.

Finally, as highlighted in Chapter 4, network inference from expression dynamics is a challenging problem, even with perfect data. I have been reasonably cautious of making direct inferences from presence or absence of individual edges in the inferred networks. Filtration thresholds were applied, and clearly reported in the methods, to make inferences about the general topology of the networks and study the different clusters formed by nodes due to the environment treatment. The observation about potential differences in regulation of *SOC1* orthologues is confirmed by further data and analyses reported in Chapter 5.

### 6.4.2 The gaps in our knowledge and how to fill them

This work uses bulk RNA-Seq data to quantify gene expression and infer networks from the expression dynamics of the genes. This technique can only quantify average RNA content for whole tissue, which in this case is the shoot apical meristem. Gene regulatory networks, however, act within individual cells. There are reports in Arabidopsis that highlight the heterogeneity of cells based on their transcriptome within the vegetative shoot apical meristem [235] and models that show spatial differentiation among expressed genes within the shoot apical meristem [236]. Therefore, it is likely there are differences in regulatory networks between cell clusters and current networks cannot unpick this information. Single cell RNA-Seq can be used to infer more accurate networks from expression data as it can provide single-cell resolution to improve the networks inferred from bulk RNA-Seq presented in this thesis. Movement of proteins within cells is another aspect that requires further investigation. Using spatial transcriptomics, for example, using probes specific for different *SOC1* orthologues can unpick if the different paralogues with distinct expression patterns have distinct spatial expression patterns too.

More detailed sampling and increasing the number of replicates would also improve the networks presented. The timescale for gene regulatory interactions between a transcription factor and a promoter sequence is in the order of minutes [237]. The data used to infer the networks presented here is not detailed enough to capture these causal interactions between genes. A more detailed timeseries sampling focussing on, for instance, the period when plants are shifted to cold temperature conditions can provide more granular information about gene expression dynamics.

Using multi-modal data can supplement the data and conclusions presented in this work. ATAC-Seq [238] can be used to determine chromatin accessiblity of binding site specific sequences, which can be used in addition to expression dynamics to infer regulatory interactions between genes. There is a report that shows that *SOC1* promoter undergoes demethylation before floral transition that leads to upregulation of *SOC1* via the Gibberellic Acid (GA) pathway in Arabidopsis [227]. The work presented here has not considered these epigenetic factors, however, they could play a role in the differential upregulation of *SOC1* if there are differences in methylation status under cold temperatures. The methylation status of different *SOC1* promoters under cold temperature can be investigated using methyl ATAC-Seq [239].

Furthermore, regulatory interactions between genes occur at the protein level, as transcription factors are proteins that bind to the promoter regions of genes. Correlation between the expression level of a gene and the amount of translated protein, while often assumed, isn't always present [240]. This proteomic information is not incorporated in the inferred networks presented here. Factors like post-transcriptional regulation, accessiblity and disruption of binding sites can cause dissrepencies between expression and protein levels of a gene and its regulators. In the absence of all this information, the networks presented in Chapter 4 should only be treated as tentative hypotheses, with further data and multiple investigations required to validate any inferences drawn.

While Arabidopsis is a well studied model, we do not know everything. There is no structural information available for the SOC1 protein. I use sequence similarity to reach a conclusion that there is no evidence for functional divergence between these proteins, however, in the absence of structural information it is difficult to ascertain the impact of the mutations identified in the sequences (Chapter 5; Figure 5.8). Studies on SOC1 proteins are needed to verify this conclusion.

Using gene editing to knockout individual *SOC1* paralogues can also shed light on their conservation of function as floral integrators and investigate the differences in their expression levels observed in data. If deletion of a paralogue with high expression levels causes up regulation of low expressed paralogues in compensation to allow plant to floral transition under normal timeframe, it would confirm conservation of function among the paralogues. Agrobacterium mediated gene editing has been proposed in *Brassica napus* [241] and used to perform knockout study for orthologues of *SVP* [100] but its efficiency varies between cultivars. As a change in single amino-acid could lead to differences in *SOC1* function, functional investigations such as complementation assays by expressing different SOC1 proteins in Arabidopsis or gel shift assay to compare SOC1 binding to DNA motifs could be utilised instead of gene editing.

Absence of any information on possibility of splice variants formed under cold conditions in *SOC1* is another blindspot. Some flowering time genes in Arabidopsis are known to form splice variants under low temperatures that can cause functional divergence [223] [234]. It is not possible based on the presented short-read RNA-Seq data to investigate these different isoforms of *SOC1* genes and there are no such reports in Arabidopsis. Long read RNA-Seq is a potential solution to study isoforms of different paralogues under cold conditions.

*CCA1* upregulation and variation binding sites provides some explanation for the divergence in expression of *SOC1* orthologues in *Brassica napus*, however, *CCA1* has only been shown to bind to the *SOC1* promoter *in-vitro* under normal conditions [224]. There are no reports of *CCA1* binding under cold temperature conditions to the *SOC1* promoter to repress its activity in Arabidopsis. The differences in binding of the *CCA1* transcription factor to promoter regions of *SOC1* orthologues hence needs further validation. Chromatin Immunoprecipitation Sequencing (ChIP-Seq) can be used to access association of *CCA1* protein [242] to different promoter sequences of *SOC1* orthologues in *Brassica napus*. Additionally, the binding site comparison was done in Darmor v10 reference genome, and *SOC1* sequences in individual cultivars could show meaningful differences compared to the reference.

It is still unknown what causes *SOC1* up-regulation under cold temperatures. There are other factors, like the gibberellin pathway [243] that can up-regulate *SOC1*, however, not enough information is available to investigate this in *Brassica napus*. Additionally, other genes like *FLC* and *SVP* also repress *SOC1* expression [38] however, the *SOC1* promoter has multiple CArG box sites where *FLC* and *SVP* can bind. Based on the current data and *in silico* techniques, it is not possible to investigate either of these two additional hypotheses.

In conclusion, this thesis expands our understanding of flowering control in a polyploid system and describes an example of subtle re-wiring of the regulatory networks controlling paralogous genes. However, like every scientific work ever published, there remain gaps to fill and enough unknowns to keep researchers brave enough to study *Brassica napus* busy for a while.

# References

1. Darwin C. Letter to JD Hooker, April 7, 1874. A Calendar of the Correspondence of Charles Darwin. The Press Syndicate of the Univ of Cambridge, Cambridge;.

2. Riley CV. On a New Genus in the Lepidopterous Family Tineidae, with Remarks on the Fertilization of Yucca. Transactions of the Academy of Science of St Louis. 1873;3:55–64.

3. Smith CI, Leebens-Mack JH. 150 Years of Coevolution Research: Evolution and Ecology of Yucca Moths (Prodoxidae) and Their Hosts. Annual Review of Entomology. 2024;69(1):375–391. doi:10.1146/annurev-ento-022723-104346.

4. Brenskelle L, Barve V, Majure LC, Guralnick RP, Li D. Analyzing a Phenological Anomaly in Yucca of the Southwestern United States. Scientific Reports. 2021;11(1). doi:10.1038/s41598-021-00265-y.

5. Bäurle I, Dean C. The Timing of Developmental Transitions in Plants. Cell. 2006;125(4):655–664. doi:10.1016/j.cell.2006.05.005.

6. Bernier G, Périlleux C. A Physiological Overview of the Genetics of Flowering Time Control: Flowering Time Control. Plant Biotechnology Journal. 2005;3(1):3–16. doi:10.1111/j.1467-7652.2004.00114.x.

7. Kobayashi Y, Weigel D. Move on up, It’s Time for Change—Mobile Signals Controlling Photoperiod-Dependent Flowering. Genes & Development. 2007;21(19):2371–2384. doi:10.1101/gad.1589007.

8. Klebs G. Über das Verhältnis der Außenwelt zur Entwicklung der Pflanzen: eine theoretische Betrachtung. Sitzber Akad Wiss Heidelberg. 1913;5:1–47. doi:10.11588/DIGLIT.37628.

9. Garner WW, Allard HA. Effect of the Relative Length of Day and Night and Other Factors of the Environment on Growth and Reproduction in Plants. Journal of Agricultural Research. 1920;18(11):553–806.

10. Pittendrigh CS. Circadian Rhythms and the Circadian Organization of Living Systems. Cold Spring Harbor Symposia on Quantitative Biology. 1960;25(0):159–184. doi:10.1101/SQB.1960.025.01.015.

11. Evans LT. Flower Induction and the Florigen Concept. Annual Review of Plant Physiology. 1971;22(1):365–394. doi:10.1146/annurev.pp.22.060171.002053.

12. Chailakhyan M. New Facts in Support of the Hormonal Theory of Plant Development. Doklady Akademii nauk SSSR. 1936;13:79–83.

13. Romanov GA. Mikhail Khristoforovich Chailakhyan: The Fate of the Scientist under the Sign of Florigen. Russian Journal of Plant Physiology. 2012;59(4):443–450. doi:10.1134/S1021443712040103.

14. King RW, Zeevaart JAD. Floral Stimulus Movement in *Perilla* and Flower Inhibition Caused by Noninduced Leaves. Plant Physiology. 1973;51(4):727–738. doi:10.1104/pp.51.4.727.

15. Wang JY. A Critique of the Heat Unit Approach to Plant Response Studies. Ecology. 1960;41(4):785–790. doi:10.2307/1931815.

16. Réaumur RAF. Observation Du Thermointre, Faites a Paris Pendant l'annee 1735, Comparees Avec Celles Qui Ont Été Faites Sous La Ligne, à l'Isle de France, à Alger et En Quelques-Unes de Nos Isles de l'Anmerique. Mem Acad des Sci, Paris. 1735; p. 545.

17. Reaumur Thermometer for Parmigiano-Reggiano Cheese; 2007.

18. Lehenbauer PA. Growth of Maize Seedlings in Relation to Temperature. vol. 1 of Physiological Researches. University of Illinois; 1914.

19. Mcmaster G. Growing Degree-Days: One Equation, Two Interpretations. Agricultural and Forest Meteorology. 1997;87(4):291–300. doi:10.1016/S0168-1923(97)00027-0.

20. Gaßner G. Beiträge Zur Physiologischen Charakteristik Sommer- Und Winterannueller Gewächse, Insbesondere Der Getreidepflanzen. Zeitschrift fuer Botanik. 1918;10:417–480.

21. Chouard P. Vernalization and Its Relations to Dormancy. Annual Review of Plant Physiology. 1960;11(1):191–238. doi:10.1146/annurev.pp.11.060160.001203.

22. Caspari EW, Marshak RE. The Rise and Fall of Lysenko: Spectacular Successes of Western Biology Initiate a Reorientation of Russian Biology along Western Lines. Science. 1965;149(3681):275–278. doi:10.1126/science.149.3681.275.

23. Soyfer VN. The Consequences of Political Dictatorship for Russian Science. Nature Reviews Genetics. 2001;2(9):723–729. doi:10.1038/35088598.

24. Amasino R. Vernalization, Competence, and the Epigenetic Memory of Winter. The Plant Cell. 2004;16(10):2553–2559. doi:10.1105/tpc.104.161070.

25. Lang A. THE EFFECT OF GIBBERELLIN UPON FLOWER FORMATION. Proceedings of the National Academy of Sciences. 1957;43(8):709–717. doi:10.1073/pnas.43.8.709.

26. Lang A. Gibberellin-like Substances in Photoinduced and vegetativeHyoscyamus Plants. Planta. 1960;54(5):498–504. doi:10.1007/BF01990006.

27. Bodson M, Outlaw WH. Elevation in the Sucrose Content of the Shoot Apical Meristem of *Sinapis Alba* at Floral Evocation. Plant Physiology. 1985;79(2):420–424. doi:10.1104/pp.79.2.420.

28. Meyerowitz EM. Prehistory and History of Arabidopsis Research. Plant Physiology. 2001;125(1):15–19. doi:10.1104/pp.125.1.15.

29. Srikanth A, Schmid M. Regulation of Flowering Time: All Roads Lead to Rome. Cellular and molecular life sciences: CMLS. 2011;68(12):2013–2037. doi:10.1007/s00018-011-0673-y.

30. Rédei GP. SUPERVITAL MUTANTS OF ARABIDOPSIS. Genetics. 1962;47(4):443–460. doi:10.1093/genetics/47.4.443.

31. Koornneef M, Hanhart CJ, Van Der Veen JH. A Genetic and Physiological Analysis of Late Flowering Mutants in Arabidopsis Thaliana. Molecular and General Genetics MGG. 1991;229(1):57–66. doi:10.1007/BF00264213.

32. Putterill J, Robson F, Lee K, Simon R, Coupland G. The CONSTANS Gene of Arabidopsis Promotes Flowering and Encodes a Protein Showing Similarities to Zinc Finger Transcription Factors. Cell. 1995;80(6):847–857. doi:10.1016/0092-8674(95)90288-0.

33. Fornara F, Panigrahi KCS, Gissot L, Sauerbrunn N, Rühl M, Jarillo JA, et al. Arabidopsis DOF Transcription Factors Act Redundantly to Reduce CONSTANS Expression and Are Essential for a Photoperiodic Flowering Response. Developmental Cell. 2009;17(1):75–86. doi:10.1016/j.devcel.2009.06.015.

34. An H, Roussot C, Suárez-López P, Corbesier L, Vincent C, Piñeiro M, et al. CONSTANS Acts in the Phloem to Regulate a Systemic Signal That Induces Photoperiodic Flowering of Arabidopsis. Development. 2004;131(15):3615–3626. doi:10.1242/dev.01231.

35. Kardailsky I, Shukla VK, Ahn JH, Dagenais N, Christensen SK, Nguyen JT, et al. Activation Tagging of the Floral Inducer *FT*. Science. 1999;286(5446):1962–1965. doi:10.1126/science.286.5446.1962.

36. Corbesier L, Vincent C, Jang S, Fornara F, Fan Q, Searle I, et al. FT Protein Movement Contributes to Long-Distance Signaling in Floral Induction of *Arabidopsis*. Science. 2007;316(5827):1030–1033. doi:10.1126/science.1141752.

37. Lee JH, Yoo SJ, Park SH, Hwang I, Lee JS, Ahn JH. Role of SVP in the Control of Flowering Time by Ambient Temperature in Arabidopsis. Genes & Development. 2007;21(4):397–402. doi:10.1101/gad.1518407.

38. Mateos JL, Madrigal P, Tsuda K, Rawat V, Richter R, Romera-Branchat M, et al. Combinatorial Activities of SHORT VEGETATIVE PHASE and FLOWERING LOCUS C Define Distinct Modes of Flowering Regulation in Arabidopsis. Genome Biology. 2015;16(1):31. doi:10.1186/s13059-015-0597-1.

39. Napp-Zinn K. Vernalization - Environmental and Genetic Regulation. Manipulation of flowering. 1987; p. 123–132.

40. Michaels SD, Amasino RM. FLOWERING LOCUS C Encodes a Novel MADS Domain Protein That Acts as a Repressor of Flowering. The Plant Cell. 1999;11(5):949–956. doi:10.1105/tpc.11.5.949.

41. Hepworth SR. Antagonistic Regulation of Flowering-Time Gene SOC1 by CONSTANS and FLC via Separate Promoter Motifs. The EMBO Journal. 2002;21(16):4327–4337. doi:10.1093/emboj/cdf432.

42. Gendall AR, Levy YY, Wilson A, Dean C. The VERNALIZATION 2 Gene Mediates the Epigenetic Regulation of Vernalization in Arabidopsis. Cell. 2001;107(4):525–535. doi:10.1016/S0092-8674(01)00573-6.

43. Levy YY, Mesnage S, Mylne JS, Gendall AR, Dean C. Multiple Roles of Arabidopsis VRN1 in Vernalization and Flowering Time Control. Science. 2002;doi:10.1126/science.1072147.

44. Sung S, Amasino RM. Vernalization in Arabidopsis Thaliana Is Mediated by the PHD Finger Protein VIN3. Nature. 2004;427(6970):159–164. doi:10.1038/nature02195.

45. Lucia FD, Crevillen P, Jones AME, 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. 2008;105(44):16831–16836. doi:10.1073/pnas.0808687105.

46. Wilson RN, Heckman JW, Somerville CR. Gibberellin Is Required for Flowering in Arabidopsis Thaliana under Short Days. Plant Physiology. 1992;100(1):403–408. doi:10.1104/pp.100.1.403.

47. Harberd NP, Belfield E, Yasumura Y. The Angiosperm Gibberellin-GID1-DELLA Growth Regulatory Mechanism: How an “Inhibitor of an Inhibitor” Enables Flexible Response to Fluctuating Environments. The Plant Cell. 2009;21(5):1328–1339. doi:10.1105/tpc.109.066969.

48. Wang JW, Czech B, Weigel D. miR156-Regulated SPL Transcription Factors Define an Endogenous Flowering Pathway in Arabidopsis Thaliana. Cell. 2009;138(4):738–749. doi:10.1016/j.cell.2009.06.014.

49. Simpson GG, Dean C. *Arabidopsis* , the Rosetta Stone of Flowering Time? Science. 2002;296(5566):285–289. doi:10.1126/science.296.5566.285.

50. Bouché F, Lobet G, Tocquin P, Périlleux C. FLOR-ID: An Interactive Database of Flowering-Time Gene Networks in *Arabidopsis Thaliana*. Nucleic Acids Research. 2016;44(D1):D1167–D1171. doi:10.1093/nar/gkv1054.

51. Song YH, Ito S, Imaizumi T. Flowering Time Regulation: Photoperiod- and Temperature-Sensing in Leaves. Trends in Plant Science. 2013;18(10):575–583. doi:10.1016/j.tplants.2013.05.003.

52. Michaels SD, Ditta G, Gustafson-Brown C, Pelaz S, Yanofsky M, Amasino RM. *AGL24* Acts as a Promoter of Flowering in *Arabidopsis* and Is Positively Regulated by Vernalization. The Plant Journal. 2003;33(5):867–874. doi:10.1046/j.1365-313X.2003.01671.x.

53. Lee J, Oh M, Park H, Lee I. SOC1 Translocated to the Nucleus by Interaction with AGL24 Directly Regulates *LEAFY*. The Plant Journal. 2008;55(5):832–843. doi:10.1111/j.1365-313X.2008.03552.x.

54. Weigel D, Alvarez J, Smyth DR, Yanofsky MF, Meyerowitz EM. LEAFY Controls Floral Meristem Identity in Arabidopsis. Cell. 1992;69(5):843–859. doi:10.1016/0092-8674(92)90295-N.

55. Moyroud E, Minguet EG, Ott F, Yant L, Posé D, Monniaux M, et al. Prediction of Regulatory Interactions from Genome Sequences Using a Biophysical Model for the *Arabidopsis* LEAFY Transcription Factor. The Plant Cell. 2011;23(4):1293–1306. doi:10.1105/tpc.111.083329.

56. Eriksson S, Böhlenius H, Moritz T, Nilsson O. GA4 Is the Active Gibberellin in the Regulation of *LEAFY* Transcription and *Arabidopsis* Floral Initiation. The Plant Cell. 2006;18(9):2172–2181. doi:10.1105/tpc.106.042317.

57. Yu H, Ito T, Wellmer F, Meyerowitz EM. Repression of AGAMOUS-LIKE 24 Is a Crucial Step in Promoting Flower Development. Nature Genetics. 2004;36(2):157–161. doi:10.1038/ng1286.

58. Bouché F, D'Aloia M, Tocquin P, Lobet G, Detry N, Périlleux C. Integrating Roots into a Whole Plant Network of Flowering Time Genes in Arabidopsis Thaliana. Scientific Reports. 2016;6(1):29042. doi:10.1038/srep29042.

59. Gaudinier A, Blackman BK. Evolutionary Processes from the Perspective of Flowering Time Diversity. New Phytologist. 2020;225(5):1883–1898. doi:10.1111/nph.16205.

60. Maple R, Zhu P, Hepworth J, Wang JW, Dean C. Flowering Time: From Physiology, through Genetics to Mechanism. Plant Physiology. 2024;195(1):190–212. doi:10.1093/plphys/kiae109.

61. Kippes N, Guedira M, Lin L, Alvarez MA, Brown-Guedira GL, Dubcovsky J. Single Nucleotide Polymorphisms in a Regulatory Site of VRN-A1 First Intron Are Associated with Differences in Vernalization Requirement in Winter Wheat. Molecular Genetics and Genomics. 2018;293(5):1231–1243. doi:10.1007/s00438-018-1455-0.

62. Konopatskaia I, Vavilova V, Kondratenko EY, Blinov A, Goncharov NP. VRN1 Genes Variability in Tetraploid Wheat Species with a Spring Growth Habit. BMC Plant Biology. 2016;16(S3):244. doi:10.1186/s12870-016-0924-z.

63. Yan L, Loukoianov A, Blechl A, Tranquilli G, Ramakrishna W, San-Miguel P, et al. The Wheat *VRN2* Gene Is a Flowering Repressor Down-Regulated by Vernalization. Science. 2004;303(5664):1640–1644. doi:10.1126/science.1094305.

64. Yan L, Fu D, Li C, Blechl A, Tranquilli G, Bonafede M, et al. The Wheat and Barley Vernalization Gene *VRN3* Is an Orthologue of *FT*. Proceedings of the National Academy of Sciences. 2006;103(51):19581–19586. doi:10.1073/pnas.0607142103.

65. Dixon LE, Karsai I, Kiss T, Adamski NM, Liu Z, Ding Y, et al. *VERNALIZATION1* Controls Developmental Responses of Winter Wheat under High Ambient Temperatures. Development. 2019;146(3):dev172684. doi:10.1242/dev.172684.

66. Sedivy EJ, Wu F, Hanzawa Y. Soybean Domestication: The Origin, Genetic Architecture and Molecular Bases. New Phytologist. 2017;214(2):539–553. doi:10.1111/nph.14418.

67. Xia Z, Watanabe S, Yamada T, Tsubokura Y, Nakashima H, Zhai H, et al. Positional Cloning and Characterization Reveal the Molecular Basis for Soybean Maturity Locus *E1* That Regulates Photoperiodic Flowering. Proceedings of the National Academy of Sciences. 2012;109(32). doi:10.1073/pnas.1117982109.

68. Watanabe S, Xia Z, Hideshima R, Tsubokura Y, Sato S, Yamanaka N, et al. A Map-Based Cloning Strategy Employing a Residual Heterozygous Line Reveals That the *GIGANTEA* Gene Is Involved in Soybean Maturity and Flowering. Genetics. 2011;188(2):395–407. doi:10.1534/genetics.110.125062.

69. Watanabe S, Hideshima R, Xia Z, Tsubokura Y, Sato S, Nakamoto Y, et al. Map-Based Cloning of the Gene Associated With the Soybean Maturity Locus *E3*. Genetics. 2009;182(4):1251–1262. doi:10.1534/genetics.108.098772.

70. Li MW, Liu W, Lam HM, Gendron JM. Characterization of Two Growth Period QTLs Reveals Modification of *PRR3* Genes During Soybean Domestication. Plant and Cell Physiology. 2019;60(2):407–420. doi:10.1093/pcp/pcy215.

71. Lu S, Dong L, Fang C, Liu S, Kong L, Cheng Q, et al. Stepwise Selection on Homeologous PRR Genes Controlling Flowering and Maturity during Soybean Domestication. Nature Genetics. 2020;52(4):428–436. doi:10.1038/s41588-020-0604-7.

72. Lu S, Zhao X, Hu Y, Liu S, Nan H, Li X, et al. Natural Variation at the Soybean J Locus Improves Adaptation to the Tropics and Enhances Yield. Nature Genetics. 2017;49(5):773–779. doi:10.1038/ng.3819.

73. Huang X, Kurata N, Wei X, Wang ZX, Wang A, Zhao Q, et al. A Map of Rice Genome Variation Reveals the Origin of Cultivated Rice. Nature. 2012;490(7421):497–501. doi:10.1038/nature11532.

74. Matsuoka Y, Vigouroux Y, Goodman MM, Sanchez G J, Buckler E, Doebley J. A Single Domestication for Maize Shown by Multilocus Microsatellite Genotyping. Proceedings of the National Academy of Sciences. 2002;99(9):6080–6084. doi:10.1073/pnas.052125199.

75. Xue W, Xing Y, Weng X, Zhao Y, Tang W, Wang L, et al. Natural Variation in Ghd7 Is an Important Regulator of Heading Date and Yield Potential in Rice. Nature Genetics. 2008;40(6):761–767. doi:10.1038/ng.143.

76. Komiya R, Ikegami A, Tamaki S, Yokoi S, Shimamoto K. *Hd3a* and *RFT1* Are Essential for Flowering in Rice. Development. 2008;135(4):767–774. doi:10.1242/dev.008631.

77. Ogiso-Tanaka E, Matsubara K, Yamamoto Si, Nonoue Y, Wu J, Fujisawa H, et al. Natural Variation of the RICE FLOWERING LOCUS T 1 Contributes to Flowering Time Divergence in Rice. PLoS ONE. 2013;8(10):e75959. doi:10.1371/journal.pone.0075959.

78. Guo L, Wang X, Zhao M, Huang C, Li C, Li D, et al. Stepwise Cis-Regulatory Changes in ZCN8 Contribute to Maize Flowering-Time Adaptation. Current Biology. 2018;28(18):3005–3015.e4. doi:10.1016/j.cub.2018.07.029.

79. Bostock J, Riley HT. Pliny the Elder, The Natural History. Perseus Digital Library; 1855.

80. Linnaeus C. Species Plantarum. vol. 1. London; 1753.

81. Cheng F, Wu J, Wang X. Genome Triplication Drove the Diversification of Brassica Plants. Horticulture Research. 2014;1(1). doi:10.1038/hortres.2014.24.

82. Parkin IA, Koh C, Tang H, Robinson SJ, Kagale S, Clarke WE, et al. Transcriptome and Methylome Profiling Reveals Relics of Genome Dominance in the Mesopolyploid Brassica Oleracea. Genome Biology. 2014;15(6). doi:10.1186/gb-2014-15-6-r77.

83. U N. Genome Analysis in Brassica with Special Reference to the Experimental Formation of B. Napus and Peculiar Mode of Fertilization. Japan J Bot. 1935;7:389–452.

84. Carré P, Pouzet A. Rapeseed Market, Worldwide and in Europe. OCL. 2014;21(1):D102. doi:10.1051/ocl/2013054.

85. Faculty of Food Technology and Biotechnology, University of Zagreb, Pierottijeva 6, 10000 Zagreb, Croatia, Bušić A, Kundas S, Belarussian National Technical University, Power Plant Construction and Engineering Services Faculty, Nezavisimosti Ave 150, 220013 Minsk, Belarus, Morzak G, Belarussian National Technical University, Mining Engineering and Engineering Ecology Faculty, Nezavisimosti Ave 65, 220013 Minsk, Belarus, et al. Recent Trends in Biodiesel and Biogas Production. Food Technology and Biotechnology. 2018;56(2). doi:10.17113/ftb.56.02.18.5547.

86. Chalhoub B, Denoeud F, Liu S, Parkin IAP, Tang H, Wang X, et al. Early Allopolyploid Evolution in the Post-Neolithic *Brassica Napus* Oilseed Genome. Science. 2014;345(6199):950–953. doi:10.1126/science.1253435.

87. Lu K, Wei L, Li X, Wang Y, Wu J, Liu M, et al. Whole-Genome Resequencing Reveals Brassica Napus Origin and Genetic Loci Involved in Its Improvement. Nature Communications. 2019;10(1):1154. doi:10.1038/s41467-019-09134-9.

88. Yang YW, Lai KN, Tai PY, Li WH. Rates of Nucleotide Substitution in Angiosperm Mitochondrial DNA Sequences and Dates of Divergence Between Brassica and Other Angiosperm Lineages. Journal of Molecular Evolution. 1999;48(5):597–604. doi:10.1007/pl00006502.

89. Arias T, Beilstein MA, Tang M, McKain MR, Pires JC. Diversification Times among *Brassica* (Brassicaceae) Crops Suggest Hybrid Formation after 20 Million Years of Divergence. American Journal of Botany. 2014;101(1):86–91. doi:10.3732/ajb.1300312.

90. Jones DM, Wells R, Pullen N, Trick M, Irwin JA, Morris RJ. Spatio-Temporal Expression Dynamics Differ between Homologues of Flowering Time Genes in the Allopolyploid Brassica Napus. The Plant Journal. 2018;96(1):103–118. doi:10.1111/tpj.14020.

91. Calderwood A, Lloyd A, Hepworth J, Tudor EH, Jones DM, Woodhouse S, et al. Total FLC Transcript Dynamics from Divergent Paralogue Expression Explains Flowering Diversity in Brassica Napus. New Phytologist. 2021;229(6):3534–3548. doi:10.1111/nph.17131.

92. Schiessl S. Regulation and Subfunctionalization of Flowering Time Genes in the Allotetraploid Oil Crop Brassica Napus. Frontiers in Plant Science. 2020;11:605155. doi:10.3389/fpls.2020.605155.

93. Hatzig SV, Nuppenau JN, Snowdon RJ, Schießl SV. Drought Stress Has Transgenerational Effects on Seeds and Seedlings in Winter Oilseed Rape (Brassica Napus L.). BMC Plant Biology. 2018;18(1):297. doi:10.1186/s12870-018-1531-y.

94. Brown JKM, Beeby R, Penfield S. Yield Instability of Winter Oilseed Rape Modulated by Early Winter Temperature. Scientific Reports. 2019;9(1):6953. doi:10.1038/s41598-019-43461-7.

95. Lu J, Carbone GJ, Grego JM. Uncertainty and Hotspots in 21st Century Projections of Agricultural Drought from CMIP5 Models. Scientific Reports. 2019;9(1):4922. doi:10.1038/s41598-019-41196-z.

96. Rahman H, Bennett RA, Kebede B. Molecular Mapping of QTL Alleles of Brassica Oleracea Affecting Days to Flowering and Photosensitivity in Spring Brassica Napus. PLOS ONE. 2018;13(1):e0189723. doi:10.1371/journal.pone.0189723.

97. Robert LS, Robson F, Sharpe A, Lydiate D, Coupland G. Conserved Structure and Function of the Arabidopsis Flowering Time Gene CONSTANS in Brassica Napus. Plant Molecular Biology. 1998;37(5):763–772. doi:10.1023/a:1006064514311.

98. Schiessl S, Samans B, HÃ¼ttel B, Reinhard R, Snowdon RJ. Capturing Sequence Variation among Flowering-Time Regulatory Gene Homologs in the Allopolyploid Crop Species Brassica Napus. Frontiers in Plant Science. 2014;5. doi:10.3389/fpls.2014.00404.

99. Jian H, Zhang A, Ma J, Wang T, Yang B, Shuang LS, et al. Joint QTL Mapping and Transcriptome Sequencing Analysis Reveal Candidate Flowering Time Genes in Brassica Napus L. BMC Genomics. 2019;20(1):21. doi:10.1186/s12864-018-5356-8.

100. Ahmar S, Zhai Y, Huang H, Yu K, Hafeez Ullah Khan M, Shahid M, et al. Development of Mutants with Varying Flowering Times by Targeted Editing of Multiple SVP Gene Copies in Brassica Napus L. The Crop Journal. 2022;10(1):67–74. doi:10.1016/j.cj.2021.03.023.

101. Fletcher RS, Mullen JL, Heiliger A, McKay JK. QTL Analysis of Root Morphology, Flowering Time, and Yield Reveals Trade-Offs in Response to Drought in Brassica Napus. Journal of Experimental Botany. 2015;66(1):245–256. doi:10.1093/jxb/eru423.

102. Zou X, Suppanz I, Raman H, Hou J, Wang J, Long Y, et al. Comparative Analysis of FLC Homologues in Brassicaceae Provides Insight into Their

Role in the Evolution of Oilseed Rape. PLoS ONE. 2012;7(9):e45751. doi:10.1371/journal.pone.0045751.

103. Schiessl SV, Quezada-Martinez D, Tebartz E, Snowdon RJ, Qian L. The Vernalisation Regulator FLOWERING LOCUS C Is Differentially Expressed in Biennial and Annual Brassica Napus. Scientific Reports. 2019;9(1):14911. doi:10.1038/s41598-019-51212-x.

104. Tadege M, Sheldon CC, Helliwell CA, Stoutjesdijk P, Dennis ES, Peacock WJ. Control of Flowering Time by *FLC* Orthologues in *Brassica Napus*. The Plant Journal. 2001;28(5):545–553. doi:10.1046/j.1365-313X.2001.01182.x.

105. Schiessl SV, Quezada-Martinez D, Orantes-Bonilla M, Snowdon RJ. Transcriptomics Reveal High Regulatory Diversity of Drought Tolerance Strategies in a Biennial Oil Crop. Plant Science. 2020;297:110515. doi:10.1016/j.plantsci.2020.110515.

106. Wang J, Long Y, Wu B, Liu J, Jiang C, Shi L, et al. The Evolution of Brassica Napus FLOWERING LOCUST Paralogues in the Context of Inverted Chromosomal Duplication Blocks. BMC Evolutionary Biology. 2009;9(1):271. doi:10.1186/1471-2148-9-271.

107. Raman H, Raman R, Qiu Y, Yadav AS, Sureshkumar S, Borg L, et al. GWAS Hints at Pleiotropic Roles for FLOWERING LOCUS T in Flowering Time and Yield-Related Traits in Canola. BMC Genomics. 2019;20(1):636. doi:10.1186/s12864-019-5964-y.

108. Sri T, Gupta B, Tyagi S, Singh A. Homeologs of Brassica SOC1, a Central Regulator of Flowering Time, Are Differentially Regulated Due to Partitioning of Evolutionarily Conserved Transcription Factor Binding Sites in Promoters. Molecular Phylogenetics and Evolution. 2020;147:106777. doi:10.1016/j.ympev.2020.106777.

109. Wendel JF. Genome Evolution in Polyploids. Plant Molecular Biology. 2000;42(1):225–249.

110. Van De Peer Y, Mizrachi E, Marchal K. The Evolutionary Significance of Polyploidy. Nature Reviews Genetics. 2017;18(7):411–424. doi:10.1038/nrg.2017.26.

111. Fox DT, Soltis DE, Soltis PS, Ashman TL, Van De Peer Y. Polyploidy: A Biological Force From Cells to Ecosystems. Trends in Cell Biology. 2020;30(9):688–694. doi:10.1016/j.tcb.2020.06.006.

112. Ohno S. Evolution by Gene Duplication. Berlin, Heidelberg: Springer Berlin Heidelberg; 1970.

113. Jia Y, Xu M, Hu H, Chapman B, Watt C, Buerte B, et al. Comparative Gene Retention Analysis in Barley, Wild Emmer, and Bread Wheat Pangenome Lines Reveals Factors Affecting Gene Retention Following Gene Duplication. BMC Biology. 2023;21(1):25. doi:10.1186/s12915-022-01503-z.

114. Birchler JA, Bhadra U, Bhadra MP, Auger DL. Dosage-Dependent Gene Regulation in Multicellular Eukaryotes: Implications for Dosage Compensation, Aneuploid Syndromes, and Quantitative Traits. Developmental Biology. 2001;234(2):275–288. doi:10.1006/dbio.2001.0262.

115. Iohannes SD, Jackson D. Tackling Redundancy: Genetic Mechanisms Underlying Paralog Compensation in Plants. New Phytologist. 2023;240(4):1381–1389. doi:10.1111/nph.19267.

116. Pinyopich A, Ditta GS, Savidge B, Liljegren SJ, Baumann E, Wisman E, et al. Assessing the Redundancy of MADS-box Genes during Carpel and Ovule Development. Nature. 2003;424(6944):85–88. doi:10.1038/nature01741.

117. Kempin SA, Savidge B, Yanofsky MF. Molecular Basis of the *Cauliflower* Phenotype in *Arabidopsis*. Science. 1995;267(5197):522–525. doi:10.1126/science.7824951.

118. Giani AM, Gallo GR, Gianfranceschi L, Formenti G. Long Walk to Genomics: History and Current Approaches to Genome Sequencing and Assembly. Computational and Structural Biotechnology Journal. 2020;18:9–19. doi:10.1016/j.csbj.2019.11.002.

119. Adams MD, Kerlavage AR, Fleischmann RD, Fuldner RA, Bult CJ, Lee NH, et al. Initial Assessment of Human Gene Diversity and Expression Patterns Based upon 83 Million Nucleotides of cDNA Sequence. Nature. 1995;377(6547 Suppl):3–174.

120. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, et al. The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. Science. 2008;320(5881):1344–1349. doi:10.1126/science.1158441.

121. Kapranov P, Cawley SE, Drenkow J, Bekiranov S, Strausberg RL, Fodor SPA, et al. Large-Scale Transcriptional Activity in Chromosomes 21 and 22. Science. 2002;296(5569):916–919. doi:10.1126/science.1068597.

122. Bertone P, Stolc V, Royce TE, Rozowsky JS, Urban AE, Zhu X, et al. Global Identification of Human Transcribed Sequences with Genome Tiling Arrays. Science. 2004;306(5705):2242–2246. doi:10.1126/science.1103388.

123. Rounsley SD, Glodek A, Sutton G, Adams MD, Somerville CR, Venter JC, et al. The Construction of Arabidopsis Expressed Sequence Tag Assemblies (A New Resource to Facilitate Gene Identification). Plant Physiology. 1996;112(3):1177–1183. doi:10.1104/pp.112.3.1177.

124. Houde M, Belcaid M, Ouellet F, Danyluk J, Monroy AF, Dryanova A, et al. Wheat EST Resources for Functional Genomics of Abiotic Stress. BMC genomics. 2006;7:149. doi:10.1186/1471-2164-7-149.

125. Bombarely A, Merchante C, Csukasi F, Cruz-Rus E, Caballero JL, Medina-Escobar N, et al. Generation and Analysis of ESTs from Strawberry (Fragaria Xananassa) Fruits and Evaluation of Their Utility in Genetic and Molecular Studies. BMC genomics. 2010;11:503. doi:10.1186/1471-2164-11-503.

126. Cloonan N, Forrest ARR, Kolle G, Gardiner BBA, Faulkner GJ, Brown MK, et al. Stem Cell Transcriptome Profiling via Massive-Scale mRNA Sequencing. Nature Methods. 2008;5(7):613–619. doi:10.1038/nmeth.1223.

127. Li R, Li Y, Kristiansen K, Wang J. SOAP: Short Oligonucleotide Alignment Program. Bioinformatics (Oxford, England). 2008;24(5):713–714. doi:10.1093/bioinformatics/btn025.

128. Behnam B, Bohorquez-Chaux A, Castaneda-Mendez OF, Tsuji H, Ishitani M, Becerra Lopez-Lavalle LA. An Optimized Isolation Protocol Yields High-quality RNA from Cassava Tissues ( *Manihot Esculenta* Crantz). FEBS Open Bio. 2019;9(4):814–825. doi:10.1002/2211-5463.12561.

129. Riesgo A, Pérez-Porro AR, Carmona S, Leys SP, Giribet G. Optimization of Preservation and Storage Time of Sponge Tissues to Obtain Quality mRNA for Next-generation Sequencing. Molecular Ecology Resources. 2012;12(2):312–322. doi:10.1111/j.1755-0998.2011.03097.x.

130. Toni LS, Garcia AM, Jeffrey DA, Jiang X, Stauffer BL, Miyamoto SD, et al. Optimization of Phenol-Chloroform RNA Extraction. MethodsX. 2018;5:599–608. doi:10.1016/j.mex.2018.05.011.

131. Yang F, Wang G, Xu W, Hong N. A Rapid Silica Spin Column-Based Method of RNA Extraction from Fruit Trees for RT-PCR Detection of Viruses. Journal of Virological Methods. 2017;247:61–67. doi:10.1016/j.jviromet.2017.05.020.

132. Shi H, Zhou Y, Jia E, Pan M, Bai Y, Ge Q. Bias in RNA-seq Library Preparation: Current Challenges and Solutions. BioMed Research International. 2021;2021:6647597. doi:10.1155/2021/6647597.

133. Kumar R, Ichihashi Y, Kimura S, Chitwood DH, Headland LR, Peng J, et al. A High-Throughput Method for Illumina RNA-Seq Library Preparation. Frontiers in Plant Science. 2012;3. doi:10.3389/fpls.2012.00202.

134. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. Accurate Whole Human Genome Sequencing Using Reversible Terminator Chemistry. Nature. 2008;456(7218):53–59. doi:10.1038/nature07517.

135. Kukurba KR, Montgomery SB. RNA Sequencing and Analysis. Cold Spring Harbor Protocols. 2015;2015(11):pdb.top084970. doi:10.1101/pdb.top084970.

136. Lin B, Hui J, Mao H. Nanopore Technology and Its Applications in Gene Sequencing. Biosensors. 2021;11(7):214. doi:10.3390/bios11070214.

137. Wang Y, Zhao Y, Bollas A, Wang Y, Au KF. Nanopore Sequencing Technology, Bioinformatics and Applications. Nature Biotechnology. 2021;39(11):1348–1365. doi:10.1038/s41587-021-01108-x.

138. Chen JW, Shrestha L, Green G, Leier A, Marquez-Lago TT. The Hitchhikers' Guide to RNA Sequencing and Functional Analysis. Briefings in Bioinformatics. 2023;24(1):bbac529. doi:10.1093/bib/bbac529.

139. Andrews S. FastQC: A Quality Control Tool for High Throughput Sequence Data; 2015.

140. Martin M. Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads. EMBnetjournal. 2011;17(1):10. doi:10.14806/ej.17.1.200.

141. Bolger AM, Lohse M, Usadel B. Trimmomatic: A Flexible Trimmer for Illumina Sequence Data. Bioinformatics. 2014;30(15):2114–2120. doi:10.1093/bioinformatics/btu170.

142. Chen S, Zhou Y, Chen Y, Gu J. Fastp: An Ultra-Fast All-in-One FASTQ Preprocessor. Bioinformatics. 2018;34(17):i884–i890. doi:10.1093/bioinformatics/bty560.

143. Burrows M, Wheeler DJ. A Block-Sorting Lossless Data Compression Algorithm. ystems Reseach Center,: Digital Equipment Corporation; 1994. 124.

144. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and Memory-Efficient Alignment of Short DNA Sequences to the Human Genome. Genome Biology. 2009;10(3):R25. doi:10.1186/gb-2009-10-3-r25.

145. Li H, Durbin R. Fast and Accurate Short Read Alignment with Burrows–Wheeler Transform. Bioinformatics. 2009;25(14):1754–1760. doi:10.1093/bioinformatics/btp324.

146. Rumble SM, Lacroute P, Dalca AV, Fiume M, Sidow A, Brudno M. SHRiMP: Accurate Mapping of Short Color-space Reads. PLoS Computational Biology. 2009;5(5):e1000386. doi:10.1371/journal.pcbi.1000386.

147. Clement NL, Snell Q, Clement MJ, Hollenhorst PC, Purwar J, Graves BJ, et al. The GNUMAP Algorithm: Unbiased Probabilistic Mapping of Oligonucleotides from next-Generation Sequencing. Bioinformatics. 2010;26(1):38–45. doi:10.1093/bioinformatics/btp614.

148. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-Based Genome Alignment and Genotyping with HISAT2 and HISAT-genotype. Nature Biotechnology. 2019;37(8):907–915. doi:10.1038/s41587-019-0201-4.

149. Campagna D, Albiero A, Bilardi A, Caniato E, Forcato C, Manavski S, et al. PASS: A Program to Align Short Sequences. Bioinformatics. 2009;25(7):967–968. doi:10.1093/bioinformatics/btp087.

150. Calderwood A, Hepworth J, Woodhouse S, Bilham L, Jones DM, Tudor E, et al. Comparative Transcriptomics Reveals Desynchronisation of Gene Expression during the Floral Transition between Arabidopsis and *Brassica Rapa* Cultivars. Quantitative Plant Biology. 2021;2:e4. doi:10.1017/qpb.2021.6.

151. Liao Y, Smyth GK, Shi W. featureCounts: An Efficient General Purpose Program for Assigning Sequence Reads to Genomic Features. Bioinformatics. 2014;30(7):923–930. doi:10.1093/bioinformatics/btt656.

152. Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie Enables Improved Reconstruction of a Transcriptome from RNA-seq Reads. Nature Biotechnology. 2015;33(3):290–295. doi:10.1038/nbt.3122.

153. Zhao Y, Li MC, Konaté MM, Chen L, Das B, Karlovich C, et al. TPM, FPKM, or Normalized Counts? A Comparative Study of Quantification Measures for the Analysis of RNA-seq Data from the NCI Patient-Derived Models Repository. Journal of Translational Medicine. 2021;19(1):269. doi:10.1186/s12967-021-02936-w.

154. Zhao S, Ye Z, Stanton R. Misuse of RPKM or TPM Normalization When Comparing across Samples and Sequencing Protocols. RNA (New York, NY). 2020;26(8):903–909. doi:10.1261/rna.074922.120.

155. Groelz D, Sobin L, Branton P, Compton C, Wyrich R, Rainen L. Non-Formalin Fixative versus Formalin-Fixed Tissue: A Comparison of Histology and RNA Quality. Experimental and Molecular Pathology. 2013;94(1):188–194. doi:10.1016/j.yexmp.2012.07.002.

156. Hedegaard J, Thorsen K, Lund MK, Hein AMK, Hamilton-Dutoit SJ, Vang S, et al. Next-Generation Sequencing of RNA and DNA Isolated from Paired Fresh-Frozen and Formalin-Fixed Paraffin-Embedded Samples of Human Cancer and Normal Tissue. PLoS ONE. 2014;9(5):e98187. doi:10.1371/journal.pone.0098187.

157. Adams RL. The Biochemistry of the Nucleic Acids. Springer Science & Business Media; 2012.

158. Sultan M, Amstislavskiy V, Risch T, Schuette M, Dökel S, Ralser M, et al. Influence of RNA Extraction Methods and Library Selection Schemes on RNA-seq Data. BMC Genomics. 2014;15(1):675. doi:10.1186/1471-2164-15-675.

159. O'Neil D, Glowatz H, Schlumpberger M. Ribosomal RNA Depletion for Efficient Use of RNA-Seq Capacity. Current Protocols in Molecular Biology. 2013;103(1). doi:10.1002/0471142727.mb0419s103.

160. Oyola SO, Otto TD, Gu Y, Maslen G, Manske M, Campino S, et al. Optimizing Illumina Next-Generation Sequencing Library Preparation for Extremely at-Biased Genomes. BMC Genomics. 2012;13(1):1. doi:10.1186/1471-2164-13-1.

161. Aird D, Ross MG, Chen WS, Danielsson M, Fennell T, Russ C, et al. Analyzing and Minimizing PCR Amplification Bias in Illumina Sequencing Libraries. Genome Biology. 2011;12(2):R18. doi:10.1186/gb-2011-12-2-r18.

162. Mamanova L, Andrews RM, James KD, Sheridan EM, Ellis PD, Langford CF, et al. FRT-seq: Amplification-Free, Strand-Specific Transcriptome Sequencing. Nature Methods. 2010;7(2):130–132. doi:10.1038/nmeth.1417.

163. Acinas SG, Sarma-Rupavtarm R, Klepac-Ceraj V, Polz MF. PCR-Induced Sequence Artifacts and Bias: Insights from Comparison of Two 16S rRNA Clone Libraries Constructed from the Same Sample. Applied and Environmental Microbiology. 2005;71(12):8966–8969. doi:10.1128/AEM.71.12.8966-8969.2005.

164. Minoche AE, Dohm JC, Himmelbauer H. Evaluation of Genomic High-Throughput Sequencing Data Generated on Illumina HiSeq and Genome Analyzer Systems. Genome Biology. 2011;12(11):R112. doi:10.1186/gb-2011-12-11-r112.

165. Carneiro MO, Russ C, Ross MG, Gabriel SB, Nusbaum C, DePristo MA. Pacific Biosciences Sequencing Technology for Genotyping and Variation Discovery in Human Data. BMC Genomics. 2012;13(1):375. doi:10.1186/1471-2164-13-375.

166. Lancashire PD, Bleiholder H, Boom TVD, Langelüddeke P, Stauss R, Weber E, et al. A Uniform Decimal Code for Growth Stages of Crops and Weeds. Annals of Applied Biology. 1991;119(3):561–601. doi:10.1111/j.1744-7348.1991.tb04895.x.

167. Boyes DC, Zayed AM, Ascenzi R, McCaskill AJ, Hoffman NE, Davis KR, et al. Growth Stage–Based Phenotypic Analysis of Arabidopsis: A Model for High Throughput Functional Genomics in Plants. The Plant Cell. 2001;13(7):1499–1510. doi:10.1105/TPC.010011.

168. Nikolov LA. Brassicaceae Flowers: Diversity amid Uniformity. Journal of Experimental Botany. 2019;70(10):2623–2635. doi:10.1093/jxb/erz079.

169. Lou P, Wu J, Cheng F, Cressman LG, Wang X, McClung CR. Preferential Retention of Circadian Clock Genes during Diploidization Following Whole Genome Triplication in *Brassica Rapa*. The Plant Cell. 2012;24(6):2415–2426. doi:10.1105/tpc.112.099499.

170. Birchler JA, Riddle NC, Auger DL, Veitia RA. Dosage Balance in Gene Regulation: Biological Implications. Trends in Genetics. 2005;21(4):219–226. doi:10.1016/j.tig.2005.02.010.

171. Liu X, Müller HG. Modes and Clustering for Time-Warped Gene Expression Profile Data. Bioinformatics. 2003;19(15):1937–1944. doi:10.1093/bioinformatics/btg257.

172. Kristianingsih R. greatR: Gene Registration from Expression and Time-Courses in R; 2024.

173. Prautzsch H, Böhm W, Paluszny M. Bézier and B-Spline Techniques. Mathematics and Visualization. Berlin, Heidelberg: Springer; 2002.

174. Klepikova AV, Logacheva MD, Dmitriev SE, Penin AA. RNA-seq Analysis of an Apical Meristem Time Series Reveals a Critical Point in Arabidopsis Thaliana Flower Initiation. BMC Genomics. 2015;16(1):466. doi:10.1186/s12864-015-1688-9.

175. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: Summarize Analysis Results for Multiple Tools and Samples in a Single Report. Bioinformatics. 2016;32(19):3047–3048. doi:10.1093/bioinformatics/btw354.

176. Rousseau-Gueutin M, Belser C, Da Silva C, Richard G, Istace B, Cruaud C, et al. Long-Read Assembly of the *Brassica Napus* Reference Genome Darmor-bzh. GigaScience. 2020;9(12):giaa137. doi:10.1093/gigascience/giaa137.

177. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve Years of SAMtools and BCFtools. GigaScience. 2021;10(2):giab008. doi:10.1093/gigascience/giab008.

178. Wagner GP, Kin K, Lynch VJ. Measurement of mRNA Abundance Using RNA-seq Data: RPKM Measure Is Inconsistent among Samples. Theory in Biosciences = Theorie in Den Biowissenschaften. 2012;131(4):281–285. doi:10.1007/s12064-012-0162-3.

179. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: Architecture and Applications. BMC Bioinformatics. 2009;10(1):421. doi:10.1186/1471-2105-10-421.

180. Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, et al. The Arabidopsis Information Resource (TAIR): Improved Gene Annotation and New Tools. Nucleic Acids Research. 2012;40(D1):D1202–D1210. doi:10.1093/nar/gkr1090.

181. Jin J, Tian F, Yang DC, Meng YQ, Kong L, Luo J, et al. PlantTFDB 4.0: Toward a Central Hub for Transcription Factors and Regulatory Interactions in Plants. Nucleic Acids Research. 2017;45(D1):D1040–D1045. doi:10.1093/nar/gkw982.

182. Seabold S, Perktold J. Statsmodels: Econometric and Statistical Modeling with Python; 2010.

183. Bansal M, Belcastro V, Ambesi-Impiombato A, Di Bernardo D. How to Infer Gene Networks from Expression Profiles. Molecular Systems Biology. 2007;3(1):78. doi:10.1038/msb4100120.

184. Satake A. Diversity of Plant Life Cycles Is Generated by Dynamic Epigenetic Regulation in Response to Vernalization. Journal of Theoretical Biology. 2010;266(4):595–605. doi:10.1016/j.jtbi.2010.07.019.

185. Penfold CA, Wild DL. How to Infer Gene Networks from Expression Profiles, Revisited. Interface Focus. 2011;1(6):857–870. doi:10.1098/rsfs.2011.0053.

186. Marku M, Pancaldi V. From Time-Series Transcriptomics to Gene Regulatory Networks: A Review on Inference Methods. PLOS Computational Biology. 2023;19(8):e1011254. doi:10.1371/journal.pcbi.1011254.

187. Äijö T, Lähdesmäki H. Learning Gene Regulatory Networks from Gene Expression Measurements Using Non-Parametric Molecular Kinetics. Bioinformatics. 2009;25(22):2937–2944. doi:10.1093/bioinformatics/btp511.

188. Rasmussen CE, Williams CKI. Gaussian Processes for Machine Learning. MIT Press; 2006.

189. Wang J. An Intuitive Tutorial to Gaussian Process Regression. arXiv. 2020;doi:10.48550/ARXIV.2009.10862.

190. Penfold CA, Shifaz A, Brown PE, Nicholson A, Wild DL. CSI: A Nonparametric Bayesian Approach to Network Inference from Multiple Perturbed Time Series Gene Expression Data. Statistical Applications in Genetics and Molecular Biology. 2015;14(3):307–310. doi:10.1515/sagmb-2014-0082.

191. Tin Kam Ho. Random Decision Forests. In: Proceedings of 3rd International Conference on Document Analysis and Recognition. vol. 1. Montreal, Que., Canada: IEEE Comput. Soc. Press; 1995. p. 278–282.

192. Cirrone J, Brooks MD, Bonneau R, Coruzzi GM, Shasha DE. OutPredict: Multiple Datasets Can Improve Prediction of Expression and Inference of Causality. Scientific Reports. 2020;10(1):6804. doi:10.1038/s41598-020-63347-3.

193. Kim GB, Gao Y, Palsson BO, Lee SY. DeepTFactor: A Deep Learning-Based Tool for the Prediction of Transcription Factors. Proceedings of the National Academy of Sciences. 2021;118(2):e2021171118. doi:10.1073/pnas.2021171118.

194. Matthews AGdG, van der Wilk M, Nickson T, Fujii Keisuke, Boukouvalas A, León-Villagrá P, et al. GPflow: A Gaussian Process Library Using TensorFlow. Journal of Machine Learning Research. 2017;18(40):1–6.

195. Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, et al.. TensorFlow: Large-scale Machine Learning on Heterogeneous Systems; 2015.

196. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods. 2020;17:261–272. doi:10.1038/s41592-019-0686-2.

197. Hagberg AA, Schult DA, Swart PJ. Exploring Network Structure, Dynamics, and Function Using NetworkX. In: Varoquaux G, Vaught T, Millman J, editors. Proceedings of the 7th Python in Science Conference. Pasadena, CA USA; 2008. p. 11–15.

198. Parés F, Garcia-Gasulla D, Vilalta A, Moreno J, Ayguadé E, Labarta J, et al.. Fluid Communities: A Competitive, Scalable and Diverse Community Detection Algorithm; 2017.

199. Newman MEJ. Equivalence between Modularity Optimization and Maximum Likelihood Methods for Community Detection. Physical Review E. 2016;94(5):052315. doi:10.1103/PhysRevE.94.052315.

200. Webber W, Moffat A, Zobel J. A Similarity Measure for Indefinite Rankings. ACM Transactions on Information Systems. 2010;28(4):1–38. doi:10.1145/1852102.1852106.

201. Marbach D, Prill RJ, Schaffter T, Mattiussi C, Floreano D, Stolovitzky G. Revealing Strengths and Weaknesses of Methods for Gene Network Inference. Proceedings of the National Academy of Sciences. 2010;107(14):6286–6291. doi:10.1073/pnas.0913357107.

202. Schmid M, Uhlenhaut NH, Godard F, Demar M, Bressan R, Weigel D, et al. Dissection of Floral Induction Pathways Using Global Expression Analysis. Development. 2003;130(24):6001–6012. doi:10.1242/dev.00842.

203. Wang JW, Schwab R, Czech B, Mica E, Weigel D. Dual Effects of miR156-Targeted*SPL*Genes and*CYP78A5/KLUH*on Plastochron Length and Organ Size in*Arabidopsis Thaliana*. The Plant Cell. 2008;20(5):1231–1243. doi:10.1105/tpc.108.058180.

204. Wu G, Poethig RS. Temporal Regulation of Shoot Development in*Arabidopsis thaliana*by*miR156*and Its target*SPL3*. Development. 2006;133(18):3539–3547. doi:10.1242/dev.02521.

205. Jung JH, Ju Y, Seo PJ, Lee JH, Park CM. The SOC1-SPL Module Integrates Photoperiod and Gibberellic Acid Signals to Control Flowering Time in Arabidopsis. The Plant Journal. 2012;69(4):577–588. doi:10.1111/j.1365-313x.2011.04813.x.

206. Yamaguchi A, Wu MF, Yang L, Wu G, Poethig RS, Wagner D. The MicroRNA-Regulated SBP-Box Transcription Factor SPL3 Is a Direct Upstream Activator of LEAFY, FRUITFULL, and APETALA1. Developmental Cell. 2009;17(2):268–278. doi:10.1016/j.devcel.2009.06.007.

207. Mizoguchi T, Wheatley K, Hanzawa Y, Wright L, Mizoguchi M, Song HR, et al. LHY and CCA1 Are Partially Redundant Genes Required to Maintain Circadian Rhythms in Arabidopsis. Developmental Cell. 2002;2(5):629–641. doi:10.1016/s1534-5807(02)00170-3.

208. Fernandez DE, Wang CT, Zheng Y, Adamczyk BJ, Singhal R, Hall PK, et al. The MADS-Domain Factors AGAMOUS-LIKE15 and AGAMOUS-LIKE18, along with SHORT VEGETATIVE PHASE and AGAMOUS-LIKE24, Are Necessary to Block Floral Gene Expression during the Vegetative Phase. Plant Physiology. 2014;165(4):1591–1603. doi:10.1104/pp.114.242990.

209. Zhang X, Chen Y, Wang ZY, Chen Z, Gu H, Qu LJ. Constitutive Expression of *CIR1* (*RVE2*) Affects Several Circadian-regulated Processes and Seed Germination in Arabidopsis. The Plant Journal. 2007;51(3):512–525. doi:10.1111/j.1365-313x.2007.03156.x.

210. Hazen SP, Schultz TF, Pruneda-Paz JL, Borevitz JO, Ecker JR, Kay SA. *LUX ARRHYTHMO* Encodes a Myb Domain Protein Essential for Circadian Rhythms. Proceedings of the National Academy of Sciences. 2005;102(29):10387–10392. doi:10.1073/pnas.0503029102.

211. Abe M, Kaya H, Watanabe-Taneda A, Shibuta M, Yamaguchi A, Sakamoto T, et al. FE, a Phloem-specific Myb-related Protein, Promotes Flowering through Transcriptional Activation of FLOWERING LOCUS *T* and FLOWERING LOCUS *T* INTERACTING PROTEIN *1*. The Plant Journal. 2015;83(6):1059–1068. doi:10.1111/tpj.12951.

212. Wenkel S, Turck F, Singer K, Gissot L, Le Gourrierec J, Samach A, et al. CONSTANS and the CCAAT Box Binding Complex Share a Functionally Important Domain and Interact to Regulate Flowering of*Arabidopsis*. The Plant Cell. 2006;18(11):2971–2984. doi:10.1105/tpc.106.043299.

213. Onouchi H, Igeño MI, Périlleux C, Graves K, Coupland G. Mutagenesis of Plants Overexpressing *CONSTANS* Demonstrates Novel Interactions among Arabidopsis Flowering-Time Genes. The Plant Cell. 2000;12(6):885–900. doi:10.1105/tpc.12.6.885.

214. Soppe WJJ, Jacobsen SE, Alonso-Blanco C, Jackson JP, Kakutani T, Koornneef M, et al. The Late Flowering Phenotype of Fwa Mutants Is Caused by Gain-of-Function Epigenetic Alleles of a Homeodomain Gene. Molecular Cell. 2000;6(4):791–802. doi:10.1016/S1097-2765(05)00090-0.

215. Samach A, Onouchi H, Gold SE, Ditta GS, Schwarz-Sommer Z, Yanofsky MF, et al. Distinct Roles of CONSTANS Target Genes in Reproductive Development of *Arabidopsis*. Science. 2000;288(5471):1613–1616. doi:10.1126/science.288.5471.1613.

216. Lee H, Suh SS, Park E, Cho E, Ahn JH, Kim SG, et al. The AGAMOUS-LIKE 20 MADS Domain Protein Integrates Floral Inductive Pathways in *Arabidopsis*. Genes & Development. 2000;14(18):2366–2376. doi:10.1101/gad.813600.

217. Gramzow L, Ritz MS, Theißen G. On the Origin of MADS-domain Transcription Factors. Trends in Genetics. 2010;26(4):149–153. doi:10.1016/j.tig.2010.01.004.

218. Liu C, Chen H, Er HL, Soo HM, Kumar PP, Han JH, et al. Direct Interaction of *AGL24* and *SOC1* Integrates Flowering Signals in *Arabidopsis*. Development. 2008;135(8):1481–1491. doi:10.1242/dev.020255.

219. Yoo SK, Chung KS, Kim J, Lee JH, Hong SM, Yoo SJ, et al. *CONSTANS* Activates *SUPPRESSOR OF OVEREXPRESSION OF CONSTANS 1* through *FLOWERING LOCUS T* to Promote Flowering in Arabidopsis. Plant Physiology. 2005;139(2):770–778. doi:10.1104/pp.105.066928.

220. Immink RGH, Posé D, Ferrario S, Ott F, Kaufmann K, Valentim FL, et al. Characterization of SOC1's Central Role in Flowering by the Identifi-

cation of Its Upstream and Downstream Regulators. Plant Physiology. 2012;160(1):433–449. doi:10.1104/pp.112.202614.

221. Helliwell CA, Wood CC, Robertson M, James Peacock W, Dennis ES. 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. 2006;46(2):183–192. doi:10.1111/j.1365-313X.2006.02686.x.

222. Li D, Liu C, Shen L, Wu Y, Chen H, Robertson M, et al. A Repressor Complex Governs the Integration of Flowering Signals in Arabidopsis. Developmental Cell. 2008;15(1):110–120. doi:10.1016/j.devcel.2008.05.002.

223. Posé D, Verhage L, Ott F, Yant L, Mathieu J, Angenent GC, et al. Temperature-Dependent Regulation of Flowering by Antagonistic FLM Variants. Nature. 2013;503(7476):414–417. doi:10.1038/nature12633.

224. Lu SX, Webb CJ, Knowles SM, Kim SHJ, Wang Z, Tobin EM. CCA1 and ELF3 Interact in the Control of Hypocotyl Length and Flowering Time in Arabidopsis. Plant Physiology. 2012;158(2):1079–1088. doi:10.1104/pp.111.189670.

225. Yant L, Mathieu J, Dinh TT, Ott F, Lanz C, Wollmann H, et al. Orchestration of the Floral Transition and Floral Development in *Arabidopsis* by the Bifunctional Transcription Factor APETALA2. The Plant Cell. 2010;22(7):2156–2170. doi:10.1105/tpc.110.075606.

226. 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. 2014;65(4):1193–1203. doi:10.1093/jxb/ert482.

227. Hou X, Zhou J, Liu C, Liu L, Shen L, Yu H. Nuclear Factor Y-mediated H3K27me3 Demethylation of the SOC1 Locus Orchestrates Flowering Responses of Arabidopsis. Nature Communications. 2014;5(1):4601. doi:10.1038/ncomms5601.

228. Heo JB, Sung S, Assmann SM. Ca2+-Dependent GTPase, Extra-large G Protein 2 (XLG2), Promotes Activation of DNA-binding Protein Related to Vernalization 1 (RTV1), Leading to Activation of Floral Integrator Genes and Early Flowering in Arabidopsis. Journal of Biological Chemistry. 2012;287(11):8242–8253. doi:10.1074/jbc.M111.317412.

229. Song HR, Song JD, Cho JN, Amasino RM, Noh B, Noh YS. The RNA Binding Protein ELF9 Directly Reduces *SUPPRESSOR OF OVEREXPRESSION OF CO1* Transcript Levels in *Arabidopsis* , Possibly via Nonsense-Mediated mRNA Decay. The Plant Cell. 2009;21(4):1195–1211. doi:10.1105/tpc.108.064774.

230. Matar S, Kumar A, Holtgräwe D, Weisshaar B, Melzer S. The Transition to Flowering in Winter Rapeseed during Vernalization. Plant, Cell & Environment. 2021;44(2):506–518. doi:10.1111/pce.13946.

231. Matar S, Melzer S. A 598-Bp InDel Variation in the Promoter Region of Bna.SOC1.A05 Is Predominantly Present in Winter Type Rapeseeds. Frontiers in Plant Science. 2021;12:640163. doi:10.3389/fpls.2021.640163.

232. Katoh K, Rozewicki J, Yamada KD. MAFFT Online Service: Multiple Sequence Alignment, Interactive Sequence Choice and Visualization. Briefings in Bioinformatics. 2019;20(4):1160–1166. doi:10.1093/bib/bbx108.

233. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. Jalview Version 2—a Multiple Sequence Alignment Editor and Analysis Workbench. Bioinformatics. 2009;25(9):1189–1191. doi:10.1093/bioinformatics/btp033.

234. Park MJ, Seo PJ, Park CM. *CCA1* Alternative Splicing as a Way of Linking the Circadian Clock to Temperature Response in Arabidopsis. Plant Signaling & Behavior. 2012;7(9):1194–1196. doi:10.4161/psb.21300.

235. Zhang TQ, Chen Y, Wang JW. A Single-Cell Analysis of the Arabidopsis Vegetative Shoot Apex. Developmental Cell. 2021;56(7):1056–1074.e8. doi:10.1016/j.devcel.2021.02.021.

236. Uzair M, Urquidi Camacho RA, Liu Z, Overholt AM, DeGennaro D, Zhang L, et al. An Updated Model of Shoot Apical Meristem Regulation by ERECTA Family and CLAVATA3 Signaling Pathways in *Arabidopsis*. Development. 2024;151(12):dev202870. doi:10.1242/dev.202870.

237. Meeussen JVW, Lenstra TL. Time Will Tell: Comparing Timescales to Gain Insight into Transcriptional Bursting. Trends in Genetics. 2024;40(2):160–174. doi:10.1016/j.tig.2023.11.003.

238. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of Native Chromatin for Fast and Sensitive Epigenomic Profiling of Open Chromatin, DNA-binding Proteins and Nucleosome Position. Nature Methods. 2013;10(12):1213–1218. doi:10.1038/nmeth.2688.

239. Spektor R, Tippens ND, Mimoso CA, Soloway PD. Methyl-ATAC-seq Measures DNA Methylation at Accessible Chromatin. Genome Research. 2019;29(6):969–977. doi:10.1101/gr.245399.118.

240. Koussounadis A, Langdon SP, Um IH, Harrison DJ, Smith VA. Relationship between Differentially Expressed mRNA and mRNA-protein Correlations in a Xenograft Model System. Scientific Reports. 2015;5(1):10775. doi:10.1038/srep10775.

241. Bhalla PL, Singh MB. Agrobacterium-Mediated Transformation of Brassica Napus and Brassica Oleracea. Nature Protocols. 2008;3(2):181–189. doi:10.1038/nprot.2007.527.

242. Muhammad II, Kong SL, Akmar Abdullah SN, Munusamy U. RNA-seq and ChIP-seq as Complementary Approaches for Comprehension of Plant Transcriptional Regulatory Mechanism. International Journal of Molecular Sciences. 2019;21(1):167. doi:10.3390/ijms21010167.

243. Moon J, Suh SS, Lee H, Choi KR, Hong CB, Paek NC, et al. The *SOC1* MADS-box Gene Integrates Vernalization and Gibberellin Signals for Flowering in *Arabidopsis*. The Plant Journal. 2003;35(5):613–623. doi:10.1046/j.1365-313X.2003.01833.x.