INVESTIGATION OF TRANSCRIPTOME CHANGES DURING FLOWER DEVELOPMENT USING THIRD–GENERATION SEQUENCING AND INDUCIBLE GENE PERTURBATION APPROACHES

Andrea Finocchio

Supervisor: Professor Frank Wellmer

July 2023

A thesis submitted to The University of Dublin
for the degree of Doctor of Philosophy

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Declaration

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work. I agree to deposit this thesis in the University's open access institutional repository or allow the Library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement. I consent to the examiner retaining a copy of the thesis beyond the examining period, should they so wish (EU GDPR May 2018).

Dedication

If you can dream—and not make dreams your master;
If you can think—and not make thoughts your aim;
If you can meet with Triumph and Disaster
And treat those two impostors just the same;

Se sei capace di sognare—e non fare del sogno il tuo padrone;
Se sei capace di pensare—e non fare del pensiero il tuo fine,
Se sei capace di incontrarti con Trionfo e Disastro
E di trattare quei due impostori allo stesso modo

"If" — by Rudyard Kipling

Voglio ringraziare la mia famiglia: mio padre e mia madre, per avermi insegnato a guardare all' ignoto con curiositá e mai con timore, e i miei due fratelli Gabriele e Luca, per non avermi fatto crescere con la spocchia del figlio unico.

I want to thank my supervisor Professor Frank Wellmer for the support over the years and for always believing in me, and all the members of the Wellmer lab, past and present: Bennett, Kevin, Christina, Caoilfhionn, Joseph, Kate. A thank you also to all the amazing people I have met in the Genetics Department and the entire Mendel Maulers team.

To Bennett, I am sure you are going to be a great dad. You have been the first person in my TCD emergency contact list, the first person I felt I could truly rely on in my new life in Dublin. To this day, that has never changed.

To Kevin, I would have to write another PhD thesis on how it is possible to live with someone in barely liveable Dublin dwellings for four years, through a lockdown, and never have a single argument, just a great time. I am grateful to have shared with you such a big chunk of my life.

To Dáire, the experiences we shared, from the surfing to tag rugby, are

the best memories I have of Ireland. Thank you for everything.

Anthony and Taylor, you are the kindest people I know, thank you for always being there for me.

Una cosa che ho imparato durante questi quattro anni é che:
Trionfo e Disastro, i sogni, non sono il fine della vita, ma parte del viaggio. L' importante é condividerli, o superarli, con le persone che abbiamo a cuore.

There is one thing I have learnt over these four years is this:
Triumph and Disaster, dreams, are not the life's aim, but part of the journey. The key is to share, or overcome them, with the people we care about.

Acknowledgements

This work was funded by:

Contents

Summary

Over the last 30 years, flower development has been intensively studied using a wide range of genetic and molecular methods. Much of this work was guided by the ABC model of floral organ identity specification, which posits that a small number of regulatory genes, termed the floral homeotic genes, act in a combinatorial manner to specify the four different types of floral organs. Although the central tenets of this model have been confirmed in diverse angiosperms, the molecular activities of the transcription factors that are encoded by the floral homeotic genes are not well understood and there are layers of regulation that are still largely unexplored. The development of methods, including new sequencing approaches that move past the measurement of gene expression, and instead shift the focus deeper, to the level of single transcripts or cells, holds the potential for further progress in our understanding of flower development. At the same time, there is now a focus on the application of the knowledge already gained to improve crop plants and to generate new varieties with improved traits. These new varieties are urgently needed as they could aid in alleviating the effects of climate change on crop production and to make agriculture more sustainable. Based on discoveries made in our laboratory, we are currently exploring the potential of trichomes, or hairs, for the protection of crops against insect pests. Specifically, we are attempting to induce trichomes on seed pods of *Brassica* crops to make them more resistant to herbivory. To underpin this translational research and to broaden the knowledge base for the processes under study, my work aimed at investigating the processes of flower development and trichome formation in more detail. To this end, I conducted three largely separate projects using the model plant *Arabidopsis thaliana*: in the first project (Chapter 3), I used cutting-edge long-read direct cDNA sequencing with the Oxford Nanopore MinION to build a new atlas of the Arabidop-

sis transcriptome during flower development, to improve our understanding of complex transcriptome dynamics, and to generate a valuable reference dataset for the scientific community for use in future studies. In the second project (Chapter 4), I studied two key trichome repressors, TRIPTYCHON (TRY) and CAPRICE (CPC) because it has been shown that their inactivation can lead to the formation of trichomes on carpel valves and are a major target for the ongoing translational research in our laboratory. Specifically, I explore the interactome around TRY and CPC using TurboID, a recently developed proximity labeling system, as well as immunoprecipitation coupled with mass spectrometry. Work in the third project (Chapter 5), was centered around the floral homeotic factors APETALA3 (AP3) and AGAMOUS (AG), which are involved in the specification of floral organ identities. Using an inducible gene perturbation approach combined with RNA-sequencing, I explored the gene expression programs acting downstream of these master regulators at different stages of early flower development.

Chapter 1

Introduction: Genetic control of flower development in the model plant *Arabidopsis thaliana*

NB: part of this Introduction is based on a review I co-authored: Goslin, K., Finocchio, A., and Wellmer, F. (2023) *Floral homeotic factors: a question of specificity* Plants – in press.

Arabidopsis thaliana (Arabidopsis, hereafter) is an angiosperm of the Brassicaceae family that has been extensively used over the past 40 years for the study of plant biology. It is a near perfect model organism for research because it is small and easy to maintain in a laboratory, has a short generation time (approximately 8 weeks under laboratory conditions), and is diploid with a small genome (∼125 MB). Furthermore, it is self-fertile and can be readily transformed, allowing the application of a wide range of molecular techniques for analysis. Among the processes that have been most intensely studied in Arabidopsis is flower development. Flowers contain the reproductive organs of angiosperms and are hence central for propagation. They produce much of the food humans and their livestock consume and are therefore pivotal for agriculture and the economy.

Arabidopsis flowers are composed of four different types of floral organs which are arranged in four concentric circles, or whorls. From the outermost

whorl to the center, the organs are sepals, followed by petals, followed by stamens and finally carpels, which fuse to form the gynoecium. Organ numbers are largely invariable in Arabidopsis, with four sepals, four petals, six stamens and two carpels being most frequently observed.

Flowers are formed after the plant has switched from vegetative to reproductive development. Vegetative development commences after seed germination and the appearance of two cotyledons (embryonic leaves) as well as a primary root. True leaves then start to emerge from a pool of undifferentiated cells in the centre of the plantlet called the shoot apical meristem (SAM). The initiation of leaves occurs sequentially and the position of leaves on the stem follows a flat spiral organization resulting in the formation of a rosette. After several weeks, the rosette reaches its maximum size, and plant transitions into the reproductive phase. This step is characterized by the elongation of the stem (a process called bolting) and the transformation of the SAM into an inflorescence meristem (IM) [Boyes et al. 2001]. While the stem grows in length the inflorescence meristem produces lateral meristems, called floral meristems (FM), in a spatial arrangement that follows the so-called Golden Angle of 137.5°. The flower buttress starts to emerge from the FM in the first 24 hours, the sepal primordia arise after two days followed at 4 days by petals and stamens primordia and later by the gynoecium. After around 13 days, all floral organs are completely formed, and the floral bud opens in anthesis [Smyth et al. 1990; Alvarez-Buylla et al. 2010].

1.1 Genetic pathways inducing flower development in *Arabidopsis*

The SAM is composed of 3 layers termed L1, L2 and L3. The outermost layers L1 and L2 divide anticlinally (the plane of cell division is at a right angle to the plant surface) and differentiate into the epidermis and subepidermis, respectively. In contrast to the organized growth of the L1 and L2 layers, cells in the L3 divide without a clear pattern in the direction of cell division. Stem cell identity in the SAM is maintained by an interplay between the homeodomain transcription factor WUSCHEL (WUS), and the protein products of three *CLAVATA* genes (*CLV1*, *CLV2* and *CLV3*). *WUS* and *CLV* genes interact in a negative feedback loop: WUS, which is expressed in the cells of the Organizing Center underneath the stem cells, can move into

the stem cells where it promotes transcription of *CLV3*. CLV3 encodes a small protein, which is proteolytically cleaved into a short biologically active peptide. This peptide is secreted from the stem cells and serves as a ligand for a number of membrane-bound receptors including CLV1, CLV2 and its partner, CORYNE. This mechanism confines *WUS* expression to the L3 layer and *CLV3* to the outer layers, at the very top of the apex, ensuring proper balance between the different cell groups of the meristem [Schoof et al. 2000].

Cells located in the periphery of the inflorescence meristem can differentiate and be incorporate into floral primordia. The specification of these floral primodia depends on the activities of several genes of which *LEAFY (LFY), APETALA1 (AP1) and TERMINAL FLOWER1 (TFL1)* are arguably the most important ones. *LFY* and *AP1* promote floral meristem identity and antagonize *TFL1*, which is expressed in the IM and prevents floral meristem identity.

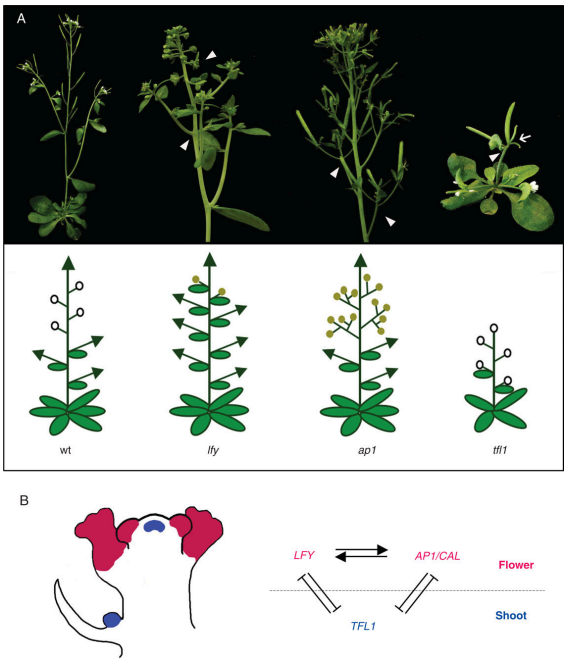


Figure 1.1: (A) phenotype of mutants *leafy* (*lfy*), *apetala1* (*ap1*) and *terminal flower1* (*tfl1*); (B) genetic interactions between *TFL1*, *AP1/CAL* (*CAULIFLOWER*) and *LFY*, colors show their role in the SAM (blue – undifferentiated) or in the FM (red – differentiated) [Benlloch et al. 2007].

LFY encodes a plant-specific transcription factor. There is only one *LFY*

gene in Arabidopsis, which is strongly expressed in young FMs at the early stages of flower development [Weigel et al. 1992]. The *lfy* mutant develops structures with shoot characteristics instead of flowers, but this phenotype is less pronounced in the proximity of the apical region of the stem, because of the independent activation of other genes, like *AP1* [Huala and Sussex 1992; Schultz and Haughn 1991]. When *LFY* is constitutively expressed, it triggers an early determination of all shoots into flowers, an indication that *LFY* is both necessary and sufficient for FM differentiation and has a central role in timing the floral transition in Arabidopsis [Blázquez et al. 1997]. *LFY* overexpression upregulates *AP1* that in turn activates the transcription of other genes required for flower meristem identity specification; moreover, *LFY* regulates flower development by acting on the auxin transport through the regulation of *PIN-FORMED1* (*PIN1*). The creation of local auxin maxima in the SAM triggers the FM differentiation and PIN1 has a central role in the process by tuning auxin concentrations accordingly [Benková et al. 2003].

AP1 encodes a MADS-domain transcription factor that, in conjunction with *LFY*, maintains flower identity. Its expression is distributed evenly in young floral meristems but becomes confined to sepals and petals at later stages. AP1 promotes the floral transition by binding several repressors of floral organ formation and suppressing their expression [Kaufmann et al. 2010].

The expression pattern of *AP1* follows that of *LFY*, but with a slightly delayed timing [Mandel et al. 1992]. The loss-of-function mutant *ap1* forms a high number of ectopic flowers resulting in an increased inflorescence branching. Sepals are rarely seen in *ap1* mutants and petals are homeotically transformed into sepal-like organs, from which, at the axels, new *ap1* flowers develop displaying the same phenotype [Mandel and Yanofsky 1995]. Constitutive expression of *AP1* under the *Cauliflower Mosaic Virus 35S* promoter causes early flowering shoots, a phenotype that resembles a *tfl1* mutant [Mandel and Yanofsky 1995].

AP1 and LFY act in concert to promote the floral transition but interestingly, LFY promotes *TFL1* expression when AP1 is not present, revealing another layer of complexity that is still largely unfathomed [Goslin et al. 2017].

CAULIFLOWER (*CAL*) is a gene that is partially redundant with *AP1*.

Both genes are the result of a recent duplication event that only occurred in the *Brassicaeceae*. While *cal* single mutants are indistinguishable from the wild type, in the *ap1 cal* double mutants the inflorescence is not able to differentiate floral meristems for an extended period of time.

The pools of cells that do not acquire floral meristem identity remain in an "in-between state" that is similar to the inflorescence meristem. These meristems develop additional meristems on their flanks with the same inflorescence meristem-like identity. The result of this is a strong over-proliferation of inflorescence meristems, which ultimately lead to an appearance that resembles the curds of a cauliflower [Mandel and Yanofsky 1995; Bowman et al. 1993].

During the rosette formation, *TFL1* is responsible for the IM identity and acts antagonistically to *LFY* and *AP1*. Unlike *LFY* and *AP1*, *TFL1* doesn't encode for a transcription factor but belongs to the phosphatidylethanolamine binding protein (PEBP) family whose members have various functions in controlling growth and differentiation.

TFL1 is expressed at low levels in the SAM, during the rosette formation. When the plant bolts, *TFL1* expression builds up ensuring the maintenance of the IM and confining cell pools undergoing floral transition at its flanks [Ratcliffe et al. 2000]. The *tfl1* mutant shows a similar phenotype to the ectopic overexpression of *AP1*: all shoot meristems are converted early in floral meristems, cauline leaves subtend flowers and IMs are converted into terminal flowers [Shannon and Meeks-Wagner 1991]. The constitutive expression of *TFL1* from a constitutive promoter leads to a strong delay in development, slowing down the progression of the shoot apex [Telfer et al. 1997]. *TFL1* has a prominent role in the IM identity and retards the phase transition by repressing *LFY* and *AP1*. Consistent with this observation, in *tfl1* mutants *LFY* and *AP1* expression is found in the IM and vice versa [Bradley et al. 1997; Weigel et al. 1992].

After the establishment of floral meristem identity through the interplay of the regulators discussed above, several floral homeotic genes start being expressed in specific regions of the floral meristem. The activation of these genes is mainly controlled by LFY and AP1, although many other regulators are also involved to ensure the expression of the floral homeotic genes in partially overlapping domains., which mark the four floral whorls.

1.2 The ABC model of floral identity specification

Once floral meristems have been specified, organ primordia arise in the different floral whorls. The specification of these primordia into different floral organ types (i.e., sepals, petals, stamens and carpels) directly depends on the activities of the floral homeotic genes, whose activities are summarised by the so-called ABC model of floral organ identity specification [Coen and Meyerowitz 1991a; Bowman et al. 2012]. This model posits that the floral homeotic genes act in a combinatorial manner to specify the four different types of floral organs. Specifically, according to the ABC model, the formation of sepals is controlled by A class genes; petal development by the combined activities of A and B class genes; stamen formation by B and C class genes; and carpel development by C class gene activity alone (Figure 1.2).

In Arabidopsis, *AP1* and *APETALA2 (AP2)* are considered as A class genes, *APETALA3 (AP3)* and *PISTILLATA (PI)* as B class genes, and *AGAMOUS (AG)* as a C class gene. In addition to postulating the combinatorial activity of the floral homeotic genes, the ABC also predicts an antagonistic effect between A and C class genes to ensure that their functions are non-overlapping.

Over the past 30 years, the central tenets of this model have been confirmed in diverse angiosperms. However, the model has also been expanded to incorporate D class genes, which are involved in specifying ovule identity, and E class genes, which are required for the activities of the A, B, C and D class genes [Bowman et al. 2012; Causier et al. 2010] (Figure 1.2). Molecular cloning of the floral homeotic genes showed that, with exception of *AP2*, all of them encode members of the family of MADS-domain transcription factors (named after four of the first family members identified, namely MCM1 from the yeast *Saccharomyces cerevisiae*, AGAMOUS from the plant *Arabidopsis thaliana*, DEFICIENS from the plant *Antirrhinum majus*, and SERUM RESPONSE FACTOR from the human *Homo sapiens*). The MADS-domain protein family is much enlarged in land plants relative to other eukaryotes (with more than 100 members in Arabidopsis [Jin et al. 2017]) and comprises many key regulators of reproductive development involved not only in floral organ specification and development, but also in processes such as flowering time control and fruit formation [Smaczniak et al. 2012a]. Transgenic and mutant studies have shown that the transcription factors encoded by the

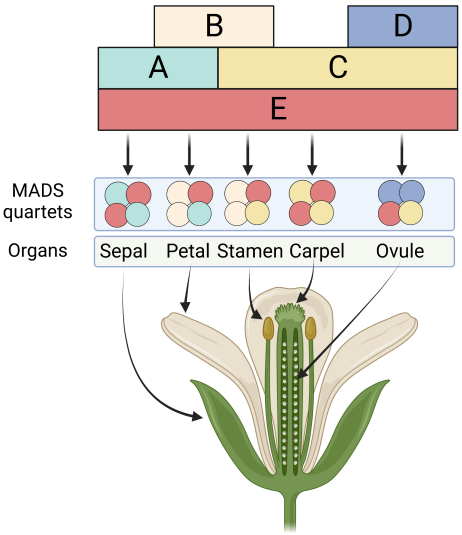


Figure 1.2: The ABCDE model of floral organ identity specification. The identity of the different floral organs is specified by the combinatorial activity of A, B, C, D and E class genes (as indicated). The MADS-domain transcription factors encoded by these genes act together in different tetrameric complexes ("quartets") to control the developmental programs needed for the formation of sepals, petals, stamens, carpels, and ovules. Colors indicate the composition of the different MADS-domain protein quartets. Figure created with `BioRender.com`.

floral homeotic genes are necessary and sufficient to define the identity of lateral organs as specific floral organs (summarised in Wellmer et al. [2014] and Yan et al. [2016]). Furthermore, gene perturbation experiments showed that these floral organ identity factors not only mediate floral organ specification at the earliest stages of flower development but are required throughout most of floral morphogenesis to control organ growth and differentiation [Ito et al. 2007; Wuest et al. 2012; Ó'Maoiléidigh et al. 2013; Gómez-Mena et al. 2005; Wellmer et al. 2004; Alves-Ferreira et al. 2007]. This finding is in agreement with the prolonged expression of the corresponding genes during flower development [O'Maoileidigh et al. 2014]. Though initially thought to act as homodimers or as heterodimers in specific combinations [Folter et al. 2005], it was later predicted and then shown that the floral organ identity factors can form tetramers to give rise to different regulatory complexes that medi-

ate the specification of the different floral organ types [Theißen et al. 2016; Theißen 2001; Honma and Goto 2001; Smaczniak et al. 2012a] (Figure 1.2). It is currently unknown whether the formation of tetramers is necessary for floral organ identity factor function, or whether dimers are sufficient for the regulation of at least some target genes. However, it has been shown recently that in Arabidopsis, tetramerization of the C class factor AG and the E class factor SEPALLATA3 (SEP3) is required for the control of floral meristem determinacy [Hugouvieux et al. 2019], suggesting that tetramers of floral organ identity factors may carry out most, if not all, of their functions. The molecular mechanism(s) underlying the activities of these complexes are, however, still not well understood.

1.3 DNA-binding preferences as a determinant of specificity

It has been shown that homo- and heterodimers of MADS-domain transcription factors bind to so-called CArG-box sequences (consensus sequence: $5'$CC[A/T]$_6$GG-$3'$) and that their preferences for and affinities to individual CArG-box motifs vary depending on the composition of a given dimer [Gramzow and Theissen 2010]. The sequences flanking the actual CArG-box also appear to play a role in determining the DNA-binding specificities of these transcription factors both *in vitro* and *in vivo* [Aerts et al. 2018; Smaczniak et al. 2012b; Käppel et al. 2021]. Because there are many MADS-domain transcription factors encoded in plant genomes (see above), which can often interact if co-expressed, there is likely a vast array of dimers with different DNA-binding preferences, providing a rich pool for regulatory diversity. Chromatin immunoprecipitation assays coupled to next-generation sequencing (ChIP-seq) revealed that several thousand sites are bound by each of the floral organ identity factors in the Arabidopsis genome [Wuest et al. 2012; Ó'Maoiléidigh et al. 2013; Kaufmann et al. 2010; Pajoro et al. 2014; Skene and Henikoff 2017; Kaufmann et al. 2009]. As expected, the bound regions often contain CArG-box sequences, but there are also sites where no such sequences can be identified, suggesting that in some cases, the floral organ identity factors bind to other motifs. In agreement with the idea that these transcription factors can act together in complexes, large overlaps between their genome-wide DNA binding patterns were detected [Ó'Maoiléidigh et al.

2013]. However, surprisingly, this is also true for the A-class regulators AP1 and AG, which are not expressed in the same cells [Yanofsky et al. 1990; Gustafson-Brown et al. 1994; Mandel et al. 1992] and have not been recovered as part of the same protein complex in proteomics experiments [Smaczniak et al. 2012a]. Thus, it appears that these floral organ identity factors can bind to many of the same sites in the Arabidopsis genome in spite of their clearly distinct functions in flower development. This finding was confirmed and further expanded on in a study comparing ChIP-seq datasets for eight MADS-domain proteins, including the floral organ identity factors but also several regulators of flowering time [Aerts et al. 2018]. It was shown that the binding patterns of these MADS-domain transcription factors show considerable overlap but that there are also clear differences, with some sites being targeted by only one of the factors analysed. Based on these observations, it can be asked whether, or to what extent, the different functions of the floral organ identity factors (and, by extension, those of other MADS-domain proteins) are determined by their inherent DNA-binding activities. In classic experiments, in which the amino-terminal half of the MADS-domain of different floral organ identity factors was replaced by the corresponding sequences of human MADS-domain proteins, it was shown that the expression of the resulting chimeric proteins led to gain-of-function phenotypes similar to those observed in lines overexpressing the unmodified floral organ identity factors [Riechmann and Meyerowitz 1997]. Because it was observed that the DNA-binding preferences of the chimeric proteins were altered in vitro when compared to the unmodified transcription factors, it was concluded that the functions of the floral organ identity factors may not depend primarily on their precise DNA-binding specificities. The results of recent studies have challenged this view. For example, a comparison of data from high-throughput *in vitro* DNA-binding assays and genome-wide localization data showed that different floral organ identity factor complexes exhibit distinct DNA-binding preferences [Smaczniak et al. 2017]. Furthermore, these preferences could be linked to specific sets of targets and developmental programs during flower development, suggesting that differences in DNA-binding preferences account for at least some of the specific functions carried out by different floral organ identity factor complexes. What could the molecular basis of this proposed mechanism be? It was shown recently that the so-called Intervening (I) domain, which follows

the amino-terminal DNA-binding MADS-domain, is involved in determining the DNA-binding specificities and dimerization properties of floral organ identity factors and other plant MADS-domain proteins [Lai et al. 2021]. Notably, a domain swap experiment showed that the I domain of AP1 is sufficient to confer AP1-like activity on AG. It was therefore suggested that the I domain confers, to a large extent, functional identity to the floral organ identity factors. Expanding this work to other pairs of MADS-domain proteins should reveal whether the I domain is a universal specificity determinant of the plant MADS-domain protein family. In summary, there is evidence that differences in the DNA-binding preferences of individual MADS-domain transcription factors (or MADS-domain protein dimers) account, at least in part, for their functional specificities. At the same time, the closely-related floral organ identity factors show considerable overlaps in their genome-wide binding patterns, raising the question of whether their DNA-binding activities alone are sufficient to determine their clearly separable functions.

1.4 Genes acting downstream of the floral homeotic factors

As described above, genome-wide localization studies showed that the floral homeotic factors bind to thousands of sites in the Arabidopsis genome. At the same time, it was found that a perturbation of their activities does affect the expression of only a small subset of these genes. Among the bona fide target genes of the floral homeotic factors (i.e., genes that are bound by them and that transcriptionally respond to their perturbation) are many that encode proteins with known or likely regulatory functions. Thus, the floral homeotic factors appear to control flower development to a large extent by controlling the expression of other regulatory genes. One example for this is the gene *NOZZLE/SPOROCYTELESS (NZZ/SPL)*, which is a key regulator of sporogenesis and is activated by the C function regulator AG [Ito et al. 2004; Wuest et al. 2012]. Importantly, *NZZ/SPL*, once activated, appears to act independently of AG, suggesting that floral homeotic factor such as AG act in part through the activation of master regulators of specific processes during flower morphogenesis. AG also mediates the termination of floral meristems by suppressing regulators of meristem function such as *WUS* and *CLV3* [Liu et al. 2011]. At least part of this process is mediate by the

C2H2 zinc-finger transcription factor KNUCKLES (KNU), whose expression is activated by AG shortly before the floral meristems ceases its activity.

The floral homeotic factors have also been shown to regulate genes involved in phyohormone signaling pathways. These include both auxin and cytokinin, which play opposite roles during development. Auxin promotes organ differentiation, with organ primordia differentiation starting from auxin maxima. Cytokinin promotes the maintenance of meristems and cells in active division. *AG* expression is also promoted by type-B ARR proteins which are cytokinin response factors [Gómez-Felipe et al. 2021]. In a feedback loop, *AG* promotes the expression of type-A ARR genes, which repress cytokinin production.

In summary, each floral homeotic factor regulates the expression of hundreds of genes of which many have regulatory functions. Most of these target genes act during early stages of flower development and relatively little is known about how the expression programmes that act downstream of the floral homeotic factors change during the course of flower development (see Chapter 5 for further details).

1.5 Floral organs as modified leaves

In his seminal book "The Metamorphosis of Plants", published in 1790, Johann Wolfgang von Goethe proposed that floral organs are derived from leaf-like structures. Support of this hypothesis came from the results of genetic experiments that showed that Arabidopsis plants carrying mutations in A, B and C class genes, or in four redundant E class genes, formed leaf-like structures in place of floral organs [Pelaz et al. 2000; Ditta et al. 2004]. This work suggested that the floral homeotic factors are key to the transformation process. In support of this idea, constitutive expression of different combinations of floral homeotic factors was shown to lead to the conversion of rosette leaves into different types of floral organs [Honma and Goto 2001]. How the floral homeotic factors mediate this transformation process is however not well understood. The above-mentioned genome-wide analyses showed that many genes with known function in leaves act downstream of the floral homeotic factors, suggesting that they modulate, and partially suppress, the genetic programme for leaf development. One example where this process has been studied in detail is the suppression of trichome formation in flow-

ers. Trichomes, or hairs, form on the epidermis of rosette leaves and stems and are thought to aid in the defence against herbivores and insect pests, the regulation of transpiration, and the reflection of UV light [Wang et al. 2021]. Floral organs in Arabidopsis, except for the abaxial side of sepals, are completely devoid of trichomes. Because several regulators of trichome formation were found to act downstream of the floral homeotic factors, it was hypothesized that the latter may directly suppress trichome formation on floral organs [O'Maoileidigh et al. 2013]. To test this, the activities of B and/or C class genes were knocked down through the induction of artificial microRNA (amiRNAs) expression [Ó'Maoiléidigh et al. 2018]. This led to the formation of trichomes on carpel valves and in some cases on the anthers of stamens, especially when the knockdowns were done in a 'sensitized background' where trichome repressor genes had been mutated. These results showed that the floral homeotic factors suppress a key aspect of leaf formation by targeting important regulatory genes. It appears likely, that regulators of other processes in leaf development are similarly under direct control of the floral homeotic factors.

1.6 Modulating trichome initiation as a possible new avenue for crop protection

It has been shown recently that the formation of trichomes on carpel valves occurs naturally in some Arabidopsis accessions as a result of mutations in regulators of trichome formation [Arteaga et al. 2022]. It has been suggested that this phenotype evolved in arid climates to reduce transpiration in fruits, but it may also aid in the defence against insect pests. In fact, seed pods of the genus Sinapis, including *Sinapis alba* (white mustard), which are closely related to Arabidopsis, are densely covered in trichomes, and it has been shown that this provides some degree of protection against damage by herbivores [Levin 1973]. In contrast, seed pods of many crop plants, including those of the genus Brassica (e.g., cauliflower, broccoli, rapeseed and turnip), are glabrous which makes them susceptible to pests such as pod borers. One possible new avenue for crop protection may therefore be to engineer new crop varieties that form trichomes on seed pods. This idea was to be tested under the research grant funded by Science Foundation Ireland and the Environmental Protection Agency that supported my work. Specifically, this

project aimed at translating the results outlined above on the regulation of trichome formation in Arabidopsis flowers, to Brassica crops. To this end, mutants in floral homeotic and trichome repressor genes were isolated in *Brassica rapa* and phenotypically characterized.

1.7 Aims of the thesis

Translational research is made possible through discoveries stemming from fundamental research. To underpin the work on floral trichomes outlined above, my work aimed at investigating the processes of flower development and trichome formation in more detail. To this end, I focused on three separate areas of research:

a) The transcriptome dynamics during early flower development
(see Chapter 3)

b) The interactomes of the trichome repressors TRIPTYCHON and CAPRICE
(see Chapter 4)

c) The gene expression programmes acting downstream of the floral homeotic factors at different stages of early flower development
(see Chapter 5)

The specific aims for each of these subprojects are outlined at the beginning of the three Results chapters that are part of this thesis.

Chapter 2

Materials and Methods

2.1 Bacterial strains

***Escherichia coli* strain STBL2 (ThermoFisher):** Genotype F^- *endA1 glnV44 thi-1 recA1 gyrA96 relA1* $\Delta(lac - proAB)mcrA\Delta(mcrBC - hsdRMS - mrr)\lambda^-$

***Agrobacterium tumefaciens* strain C58 pGV2260:** Described in [McBride and Summerfelt 1990]

2.1.1 Transgenic lines generation

All the vectors used to transform the transgenic lines in this work were generated with the GreenGate system (see section 2.5 for more details).

Lines for the study of PI isoforms

To generate the transgenic lines for the study of PI isoforms, I first cloned the entry constructs necessary for the GreenGate cloning. For the *PI* promoter, I used an entry vector (pKG17) containing a region of 1500 bp upstream of *PI*, constructed by my colleague Dr. Kevin Goslin (pKG17). The genomic sequence of *PI* was amplified from Col-0 genomic DNA with **C** entry overhangs using primers AF60-AF61. The two *PI* isoforms were amplified from cDNA of FIS flowers. *PI.1* was amplified using the primers AF68-AF69, while the *PI.2* isoform was cloned with an overlapping PCR. Two fragments, overlapping on the retained intron, were amplified with the

primers pairs AF68-AF73 and AF69-AF72, and then the whole *PI.2* sequence was amplified with AF68-AF69 with an overlapping PCR. To add the **C** entry overhangs, both *PI.1* and *PI.2* amplicons were reamplified with AF60-AF61. After the sequences were inserted into the respective entries, I used the Q5® site-directed mutagenesis kit (NEB #E0554S) to introduce silent mutations that eliminate the *Bsa*I site present in the third exon of *PI*, which would have interfered with the downstream cloning step. The *StrepII-V5* and *V5-StrepII* tags contain two *Bsa*I sites each, and they are also very short sequences. I decided to order them from a DNA synthesis service (IDT gBlock™). The **B** entry overhangs were then added to *StrepII-V5* by PCR with primer pair AF62-AF63, and the **D** overhangs to *V5-StrepII* with primer pair AF64-AF65.

The entry plasmids were combined to create vectors pAF1, pAF2, pAF5, pAF6, pAF9 and pAF10 (see Table 2.3). The vectors were transformed in Agrobacterium strains which were then used to transform plants by floral-dip (see section 2.9). Arabidopsis mutant lines were selected using the glufosinate-ammonium herbicide BASTA® and the presence of the construct was validated by PCR amplication (see Table 2.1).

2.1.2 Lines for the TRY and CPC mass spectrometry experiments

To generate the transgenic lines for the study of *TRY* and *CPC* isoforms, I first cloned the entry constructs necessary for the GreenGate cloning. The promoters of *TRY* and *CPC* were PCR-amplified from a region spanning ∼1,500 bp upstream of the transcribed regions, using primers pairs AF25-AF26 and AF21-AF22, respectively, adding the correct overhangs for the **A** entry. *TRY* and *CPC* genomic sequences were amplified with **C** overhangs using the AF17-AF18 and AF19-AF20 primer pairs, respectively. The TurboID sequence was amplified from the Addgene plasmid #107177, with the **D** overhangs, a 4 AA C-terminal linker and a CMyc tag, using the primers AF27-AF30. The *GFP* sequence was amplified with the primers AF51-AF52 to add the **D** entry overhangs. The entry plasmids were combined to create vectors PL184-PL200, PL199 and PL200 (see Table 2.3). The vectors were transformed in Agrobacterium strains which were then used to transform plants by floral-dip (see section 2.9). Arabidopsis mutant lines were selected using the glufosinate-ammonium herbicide BASTA® and the presence of the

construct was validated by PCR amplification (see Table 2.1).

2.1.3 Plant strains

Table 2.1: Plant strains used in this study

Line	Genotype	Background
Col-0		
cpc	*cpc-1*	WS
try	*try-29760*	Col-0
FIS	*pAP1:AP1-GR ap1 cal*	L-*er*
#329	*p35S:GR-LhG4/6xpOp:AP3-amiRNA*	*pAP1:AP1-AR ap1 cal*
#330	*p35S:GR-LhG4/6xpOp:AG-amiRNA*	*pAP1:AP1-AR ap1 cal*
AF2	*pTRY::NLS-GFP-TurboID-Myc*	Col-0
AF4	*pCPC::NLS-GFP-TurboID-Myc*	Col-0
AF5	*pTRY::gTRY-TurboID-Myc*	*try29760*
AF6	*pCPC::gCPC-TurboID-Myc*	*cpc-1*
AF11	*pTRY::gTRY-GFP*	*try29760*
AF12	*pCPC::gCPC-GFP*	*cpc-1*
AF13	*pUBQ10::StrepII-V5-gPI*	
AF14	*pUBQ10::gPI-V5-StrepII*	
AF15	*pPI::StrepII-V5-gPI*	Col-0
AF16	*pPI::gPI-V5-StrepII*	Col-0
AF17	*pPI::PI.1*	Col-0
AF18	*pPI::PI.2*	Col-0
AF19	*pPI::GFP-gPI*	Col-0
AF20	*pPI::gPI-GFP*	Col-0

The T-DNA insertion lines *try-29760* (#N6518) and *cpc-1* (#N6399) were ordered from Nottingham Arabidopsis Stock Centre (NASC). The floral induction system line (*pAP1:AP1-GR ap1 cal*) used in Chapter 3 was published by our lab in O'Maoileidigh et al. [2014]. The plant lines used in knockdown experiments in Chapter 5 were generated by Dr. Bennett Thomson during his PhD in the lab [Thomson 2018]. The rest of the lines were generated in this study using the GreenGate cloning system and Arabidopsis plant transformation by floral-dip (see section 2.1.1).

2.2 Seed sterilisation

Seeds were sterilised by exposure to chlorine gas in a sealed container for three hours. The chlorine gas was released by adding 3 mL of concentrated HCl to 100 mL of bleach in a glass beaker. After sterilisation, the seeds were aerated under a sterile laminar flow hood to remove any excess chlorine gas, and then stored in a sterile reaction tube, or immediately sowed on MS agar plates.

2.3 Plants growth conditions

Seeds were planted on soil made up of compost, vermiculite, and perlite (NAD, Dublin) mixed in a 3:1:1 ratio, or on Murashige and Skoog (MS) medium composed of 2.2 g/L MS salts (Sigma-Aldrich) and 6 g/L agar (Fisher Scientific) adjusted to pH 5.7 using KOH. Seeds were then stratified at 4°C for three days before being transferred into a plant growth room. Plants were grown under constant illumination at ∼20°C using cool white light. Arabidopsis pAP1::AP1-GR ap1-1 cal-1 and tobacco (*Nicotiana benthamiana*) plants were grown in a separate room at a lower temperature (∼16°C).

2.4 Genomic DNA preparation

Extraction buffer (EB): 200 mM Tris-HCl pH 7.5, 250 mM NaCl, 25 mM EDTA, 0.5% (w/v) SDS

Tris-EDTA (TE) buffer: 10 mM Tris-HCl pH 7.5, 1 mM EDTA

Genomic DNA was extracted using a previously published protocol [Edwards et al. 1991]. In brief, a small section of leaf tissue was cut out from the plant and ground using a pestle in a 1.5 mL reaction tube filled with 400 µL of EB. The samples were then centrifuged at maximum speed for 1 min to pellet cell debris, and 300 µL of the supernatant was transferred to a new tube containing an equal volume of isopropanol to precipitate DNA. The samples were mixed by vortexing and centrifuged at maximum speed for 3 min to pellet the DNA. The supernatant was decanted, and the pellet was washed once with 70% ethanol. The tubes were then briefly centrifuged

at maximum speed to position the DNA pellet at the bottom of the tube, and then the ethanol was poured on a tissue. The samples were then dried inside a vacuum centrifuge and the pellet resuspended in 50 µL of TE buffer.

2.5 Generation of plant expression vectors

All the plasmids used in this thesis were generated using the GreenGate system [Lampropoulos et al. 2013].

The GreenGate system is a quasi-scarless cloning method which uses a set of pUC19 plasmids modified to contain specific overhangs, or "entry" vectors. These entry vectors are the building blocks that can be combined in a final plasmid inside a pGreen-IIS based backbone, or "destination vector". Each entry vector has two *Bsa*I sites, and a carbenicillin resistance gene for selection. When cut with the *Bsa*I restriction enzyme, the two *Bsa*I sites create overhanging "sticky" ends. The DNA of interest has to be cloned with ends containing the *Bsa*I site and the same overhang sequence found in the entry plasmid of choice. Then, in a single restriction enzyme reaction, followed by ligation, the cloned fragment can be inserted into the entry vector. The vector is then transformed in *E. coli* cells to be multiplied, purified and validated for successful insertion of the DNA fragment. To increase the cloning efficiency, the entry plasmids contain the *ccdB* gene between the *Bsa*I sites, where the new DNA sequence is going to be inserted. The *ccdB* gene produces a toxin that kills the *E. coli* strain regularly used for cloning (*stbl2*), but it can be propagated using resistant strains. A successful insertion will replace the *ccdB* gene and allow the *E. coli* to grow. To grow on LB medium supplied carbenicillin, *E. coli* has to have been trnaformed with an entry vector, which contains the carbenicillin resistance required for its survival. Therefore, the carbenicillin selection ensures that the plasmid be kept around through multiple generations of *E. coli*. Only *E. coli* cells that have been successfully transformed with an entry plasmid with an insert, will have lost the *ccdB* gene and be resistant to carbenicillin thanks to the resistance gene in the entry plasmid. These will make sure that most of the *E. coli* colonies on the plate will contain an entry with the proper insert.

There are a total of six entry plasmids, referred to with the letters A to E, which have complementary overhangs designed to be chained together, maintaining a specific order and direction. During the final reaction, the

destination vector is mixed with the six entries, T4 DNA ligase and *Bsa*I restriction enzyme. The reaction then is cycled between a 37°C, and a 15°C incubation for 30 cycles. During the 37°C step the *Bsa*I cuts the entry plasmids and the destination vector, releasing their inserts. During each cycle, the complementary ends will ligate together. Crucially, in the destination vector the *Bsa*I sites are found inside the insert. If, during ligation, one of the overhangs is ligated with an insert, the plasmid will not close again, because the *Bsa*I sites will be lost. This ensures that the ligation step is irreversible, greatly increasing the cloning efficiency. The destination vector has a spectinomycin resistance that can be used for selection.

2.6 Constructs generated in this study

As outlined in section 2.5, all plasmids generated in this study were made using the GreenGate system. I used a GreenGate kit, which contains a set of entry plasmids for the construction of plant vectors. I used the following GreenGate kit vectors for all the constructs I generated in this work: a destination vector with the spectinomycin resistance at left border (Addgene #pGGZ003), the pea RBCS terminator entry (Addgene #pGGE001) and the entry containing the BASTA resistance cassette with the *Bsa*I site removed (Addgene #pGGF002). The kit also offers useful "dummy" entries, to use when no insert is needed in a specific position. I used the N-tag/B-dummy (Addgene #pGGB003) and C-tag/D-dummy (Addgene #pGGD002) entries. To generate the overexpression constructs for the tobacco agroinfiltration experiments, I used the *pUB10::* promoter entry from the GreenGate kit (Addgene #pGGA006).

2.6.1 Generation of entry plasmids for GreenGate cloning

The purified PCR product, amplified with the correct overhangs, was mixed with the respective entry plasmid at a ∼3:1 molarity ratio, in a 15 µL reaction containing 1 µL of *Bsa*I (NEB #R3733) and 1.5 µL of CutSmart buffer (NEB #B7204). The reaction was incubated at 37°C for 1 hour and then at 65°C for 20 min to inactivate *Bsa*I. After *Bsa*I inactivation, 1 µL of T4 ligase (NEB #M0202), 0.5 µL of CutSmart buffer and 3.5 µL of water were added to the mix and the reaction was incubated for 30 min at room temperature, followed

by 65°C inactivation for 10 min. The entry solution was then transformed in *E. coli stbl2* cells and plated on LB agar with carbenicillin selection.

2.6.2 Generating constructs with GreenGate cloning

GreenGate is based on a series of entry plasmids, that can be combined to form complex constructs.

Once all the necessary entries were obtained, the final construct was generated with the following reaction: 1.5 µL of the entry vectors (100 ng/µl stock), 1 µL of the destination vector (Addgene #pGGZ003), 1 µL of *Bsa*I, 1 µL of T4 ligase, 1.5 µL of CutSmart buffer and 1.5 µL of ATP 10 mM were mixed in a thin-walled PCR tube, mixing by pipetting after each addition. The reaction was then incubated in a thermocycler for 40 cycles, alternating 5 min at 37°C (digestion) and 5 min at 16°C (ligation). The reaction was then transformed in *E. coli stbl2* cells and plated on LB agar with spectinomycin selection.

2.7 *Escherichia coli* tranformation

2.7.1 Generation of competent cells

Competent cells were prepared as described in [Inoue et al. 1990]. Stbl2 cells were streaked on a LB plate and incubated overnight at 37°C. The cells were then inoculated in 2 mL of liquid LB medium and grown overnight at 37°C with agitation. The following day, 200 µL of the culture was used to inoculate 250 mL of SOB medium. The SOB culture was grown at 20°C with 250 rpm shaking until OD_{600} ∼0.7. The culture was then cooled on ice for 10 min and spun down at 4,000 x *g* for 10 min at 4°C. The pellet was resuspended in 80 mL ice-cold TB buffer and centrifuged again at 4,000 x *g* for 10 min at 4°C. The pellet was resuspended in 20 mL ice-cold TB and DMSO was added slowly until a final concentration of 7% (v/v). Cells were incubated on ice for 10 min and then 100 µL aliquots were frozen in liquid nitrogen and stored at -80°C.

2.7.2 Transformation of competent cells

An of 100 µL aliquot of competent cells was thawed on ice and incubated with the 3-5 µL of the purified plasmid while keeping the cells on ice. After

a 20 min incubation the cells were transferred at 42°C for 40 s and quickly cooled on ice for 1 min. One mL of liquid LB medium was immediately added to the cells which were then left at 37°C for 2 hours to recover. After the recovery, the cells were plated on LB medium supplied with the appropriate antibiotics for selection.

2.8 *Agrobacterium tumefaciens* tranformation

The Agrobacterium strain C58 pGV2260, containing the pSOUP helper plasmid, was used for the Agrobacterium-mediated transformation of Arabidopsis plants.

2.8.1 Generation of competent cells

A 5 mL Agrobacterium culture was grown overnight at 28°C. The next day, 0.5 mL of the 5 mL culture was inoculated in 250 mL of LB supplemented with 0.2 g/l $MgSO_4$ and appropriate antibiotics, and grown under 250 rpm shaking 28°C until OD_{600} of ∼1 was reached. The culture was then chilled on ice and spun at 4°C for 10 min at 4,000 x *g*. The pelleted cells were resuspended in 20 mL of ice-cold sterile 10 mM $CaCl_2$. The cells were spun at 4°C for 10 min at 4,000 x *g* and resuspended in 4 mL ice-cold sterile 10 mM $CaCl_2$. One-hundred µL aliquots were snap-frozen in liquid nitrogen and stored at -80°C.

2.8.2 Transformation of competent cells

An 100 µL aliquot of competent cells was thawed on ice and incubated with the 3-5 µL of the purified plasmid for 5 min in liquid nitrogen. The cells were then thawed at room temperature, diluted with 1 mL of LB and incubated at 28°C for 4 h. After incubation, the cells were pelleted at 4,000 x *g* for 5 min and the pellet was suspended in 500 µL LB. A 50 µL aliquot was then plated on LB agar containing carbenicillin and spectinomycin. The plates were then incubated at 28°C for 3 days.

2.9 Agrobacterium mediated plant transformation using the floral dip method

Arabidopsis plants were transformed by floral dip, according to the protocol described in Zhang et al. [2006]. For floral induction system plants, the dipping time was extended to 5 min and performed under a vacuum bell, to increase the penetration of the Agrobacterium solution.

2.10 Agroinfiltration of tobacco leaves

An Agrobacterium strain containing the desired construct was streaked on an LB plate and grown for 2 days. With the use of a sterile swab, the streak was inoculated in 10 mL of liquid LB and grown for 24 hours to an OD_{600} between 0.5 and 0.7. The cells were then spun down at 4,000 x g for 10 min and resuspended in MES medium. Tobacco plants were grown in colder conditions to reduce the amount of wax produced on the leaf surface. Leaves with less wax are easier to infiltrate. Around one week after the onset of the first pair of true leaves, the plants were watered and moved at 20°C for 24 hours, to stimulate the opening of the stomata which again helps with the infiltration process. The tobacco plants were infiltrated with Agrobacterium in MES solution using a 1 mL syringe. After seventy-two hours the leaves were collected and snap frozen and stored at -80°C.

2.11 Induction of flower formation using the floral induction system

Floral induction system plants were grown in colder conditions, which stimulates the production of bigger meristem curds that are less leafy than plants grown at 20°C (see section 2.3). When most of the plants in the trays had bolted, they were moved to 20°C for 24 hours. Flower formation was then induced using either a DEX or a DHT solution (section 2.25), in the case of AP1-GR and AP1-AR based systems, respectively (see Table 2.1) as previously described in O'Maoileidigh et al. [2014].

2.12 *AG* and *AP3* amiRNA mediated knockdowns

The *AG* and *AP3* knockdowns were induced by DEX treatment 24 hours before the target time-point. A set of plants was treated with a mock solution as a control. After 24 hours, meristematic tissue was collected in a 1.5 ml Eppendorf tube with a surgical blade, snap-frozen, and stored at -80°C. The tissue was then ground using a pestel and the RNA was extracted with a column based system and DNA-depleted. One µg of RNA was reverse-transcribed and used in real-time qPCR amplification experiments, to evaluate the strength of the amiRNA mediated knockdowns (see section 2.21). The rest of the RNA was sent out for sequencing.

2.13 Silver staining

The silver staining of polyacrylamide gels was performed following the protocol described in [Blum et al. 1987].

2.14 Protein concentration measurements

2.14.1 Amido black assay

Staining solution: 10% (v/v) acetic acid, 90% (v/v) methanol, amido black 10B

Wash solution: 10% (v/v) acetic acid, 90% (v/v) methanol

The amido black assay was used to measure the concentration of proteins in SDS loading dye. Five µL of each sample were added to 195 µL of water. Eight-hundred µL of amido black staining solution were then added, and the solution was mixed thoroughly by pipetting. The samples were centrifuged at 10,000 x g for 20 min at room temperature and the supernatant discarded. One mL of the wash solution was then added to the pellet and the samples were centrifuged at 10,000 x *g* for 20 min at room temperature. The wash was repeated once. The pellet was then dried under vacuum and resuspended in 0.2 M NaOH. The absorbance of each sample and a set of BSA standards was measured at OD_{600}. The protein concentration was inferred from a standard curve fitted on the readings of the standard samples, which have a known concentration.

2.14.2 Bradford assay

The Bradford assay was used when on samples that didn't contain high concentrations of detergents. The measurements were done by adding 5 µL of sample to 1 mL of Bradford dye reagent (ThermoFisher #J61522-K2). A set of BSA standards and a blank were processed together with the samples. After mixing by vortexing and 2 min of incubation, the absorbance was measured at OD_{600}. The protein concentration was inferred from a standard curve fitted on the readings of the standard samples, which have a known concentration.

2.14.3 Bicinchoninic acid assay (BCA)

The BCA assay was used for samples with high concentration of detergents. The BCA protein assay kit (Pierce™ #23225) was used following the manufacturer's instructions.

2.15 Protein detection by Western blotting

Stacking buffer: 0.5 M Tris pH 6.8, 0.4% (w/v) SDS

Separating buffer: 1.5 M Tris pH 8.8, 0.4% (w/v) SDS

10% Separating gel: 3.5 mL 4x separating buffer, 3.5 mL 40% acrylamide:Bis (29:1), 6.84 water, 20 µL TEMED, 160 µL APS 10% (w/v)

10% Stacking gel: 1.25 mL 2x stacking buffer, 565 µL 40% acrylamide:Bis (29:1), 3.15 µL water, 17.5 µL TEMED, 17.5 µL APS 10% (w/v)

2X Electrophoresis buffer: 15.1 g Tris base, 72.0 g glycine, 5.0 g SDS, add water to 1000 mL. Dilute when needed.

Transfer buffer: 28.8 g glycine, 6.04 g Tris base, 200 mL methanol, 1.6 L water

5X Ponceau staining solution: 1% Ponceau S (w/v), 50% acetic acid (v/v)

Blocking solution: 5% milk in PBS-T (w/v). BSA was used in place of milk when detecting biotin because the biotin in milk could interfere with detection

2.15.1 Protein separation by SDS-PAGE

For the SDS-PAGE, I used 10% polyacrylamide gels that I prepared in house and stored at 4°C the night before the experiment. On the day, the gel was inserted in an electrophoresis rig, which was then filled with electrophoresis buffer. The proteins were loaded alongside 7 µL of PageRuler™ prestained protein ladder (ThermoFisher #26616) and then separated by applying voltage to the rig. In a first phase, 50V were applied to pack the samples on the edge of the separating gel and then 100V for the rest of the run. When finer protein separation was needed, I used NuPAGE™ Bis-Tris pre-cast gels (ThermoFisher #NP0326BOX), which have a gradient of separating gels form 4% to 12% that increase the band resolution. The pre-cast gels were run in NuPAGE™ MES SDS Running Buffer (ThermoFisher #NP0002).

2.15.2 Protein transfer on nitrocellulose membrane

For the immunoblots, the proteins were transferred from the gel to an Amersham™ Protran® nitrocellulose membrane using a wet transfer protocol. First, the excess SDS in the polyacrylamide gel was carefully washed away leaving the gel under gentle shaking in transfer buffer for 20 min. The gel and the membrane were then packed in-between two sheets of Whatman paper on either side, while submerged in transfer buffer, and placed into a transfer rig, which was pre-filled with ice-cold transfer buffer. For the protein transfer, a voltage of 90V was applied for 90 min.

2.15.3 Antibody incubation and protein detection

After disassembling the transfer unit the membrane was quickly washed with 7.5% (v/v) acetic acid and incubated with 1xPonceau staining solution under shaking for 5 min. The membrane was then rinsed with 7.5% acetic acid until the background is clear, and the protein bands are well-defined. The membrane was scanned to record the result of the Ponceau staining as a measure of the protein loadings. The membrane was then rinsed thoroughly with water before blocking. The membrane was blocked by incubating under shaking three times for 5 min in blocking solution. The solution was changed for each incubation. Next, the membrane was incubated overnight at 4°C with the primary antibody in 5 mL of blocking solution.

The next day, the membrane was washed three times for 5 min with

PBS-T while shaking. The membrane was incubated with the appropriate secondary antibody in 50 mL of blocking solution for 2 hours. After three washes with PBS-T, the membrane was ready for the chemiluminescence reaction. The streptavidin-HRP antibody doesn't require a secondary antibody incubation.

2.15.4 Chemiluminescence and protein detection

The membrane was gently dried on filter paper, laid on a piece of cling film and covered with WesternBright® ECL HRP substrate (Advansta #K-12045), keeping the protein side up at all times. The reaction takes approximately 2 min during which the cling film was manually tipped in all directions to spread the solution evenly on the membrane. The membrane was then imaged in a luminescent image analyser.

2.16 Isolation of nuclei from leaf tissue

This protocol was applied to both Arabidopsis and tobacco leaves.

M1: 10 mM sodium phosphate pH 7, 0.1 M NaCl, 1 M 2-methyl 2,4-pentanediol, 10 mM β-mercaptoethanol and cOmplete™ protease inhibitor cocktail

M2: Like the M1 buffer with 10 mM $MgCl_2$ and 0.5% Triton X-100

M3: Like the M1 buffer without 2-methyl 2,4-pentanediol

The leaf tissue was ground in a mortar in liquid nitrogen. After being finely ground, it was left to thaw until the mortar was warm enough to avoid instant freezing of the M1 buffer. When the mortar reached the right temperature, 2 mL M1 buffer were added and the tissue was ground a second time on ice. Then, 8 mL of M1 buffer were added, and the suspension was filtered twice through Miracloth (Millipore #475855) using a syringe and a plunger. The filtered suspension was spun at 1000 x g for 20 min at 4°C. The supernatant was discarded, and the pellet was resuspended in 3 mL of M2 buffer. Subsequently, the suspension was spun again at 1,000 x g for 10 min at 4°C. The supernatant was discarded, and the pellet was resuspended in 1.5 mL of M2 buffer. The suspension was transferred into an Eppendorf tube

and spun again at 1,000 x *g* for 10 min at 4°C. This step was repeated once with the M2 and once with the M3 buffer. The pellet was resuspended in 2x SDS loading dye and boiled at 95°C for 5 min. After boiling, the sample was spun down at maximum speed for 10 min at room temperature, to eliminate any remaining debris. The supernatant was stored at -20°C.

2.17 GFP immunoprecipitation

For the immunoprecipitation of GFP-tagged proteins I used GFP-Trap® agarose beads (Chromotek #gta) following the manufacturer's protocol tailored to Arabidopsis plant samples.

Lysis buffer: 50 mM Tris/HCl pH 7.5, 150 mM NaCl, 1% Triton X-100, 1 mM PMSF or 1x cOmplete,

Wash buffer: Same as lysis buffer

The harvested plant material was ground in liquid nitrogen into a fine powder, resupended in 2 mL of lysis buffer and transferred in a 2 mL Eppendorf tube. The suspension was mixed by vortexing and incubated with slow mixing on a wheel for 30 min at 4°C. After incubation, the sample was spun down at maximum speed for 15 min at 4°C to remove cell debris and the supernatant was filtered through two layers of Miracloth. In a separate tube, 25 µL of agarose beads slurry (one for each sample) was washed with 500 µL of ice-cold lysis buffer and sedimented at 2,500 x *g* for 5 min at 4°C. The supernatant was carefully discarded, leaving the beads covered in buffer to avoid drying. 50 µL of the supernatant (input fraction) of the cleared protein solution was saved for Western blotting, while the rest was transferred into the 1.5 mL Eppendorf tube containing the washed agarose beads. The samples were rotated end-over-end for 1 hour at 4°C and then the beads were sedimented at 2,500 x *g* for 5 min at 4°C. 50 µL of supernatant was saved for Western blotting (unbound fraction). The remaining supernatant was discarded and the beads resuspended in 500 µL wash buffer. The beads were then sedimented at 2,500 x *g* for 5 min at 4°C and the supernatant was discarded. The washes were repeated three times with wash buffer and twice with PBS. After the last wash, 5% of the beads were boiled in 50 µL 2X SDS loading dye for 5 min (eluate fraction). The beads were sedimented

at 2,500 x g for 2 min and the supernatant was stored for Western blotting. The rest of the beads were spun down at 2,500 x g for 5 min at 4°C and then snap-frozen in liquid nitrogen after discarding the supernatant. The samples where processed in batches with one replica of each sample in each batch. The eluate fractions were used for immunoblots, to check the performance of the immunoprecipitation before proceeding with the trypsinization.

2.18 Affinity purification using streptavidin beads

The protocol is adapted from Mair et al. [2019].

Extraction buffer: 50 mM Tris pH 7.5, 150 mM NaCl, 0.1% SDS, 1% Triton-X-100, 0.5% Na-deoxycholate, 1 mM EGTA, 1 mM DTT, 1x cOmplete, 1 mM PMSF

Equilibration buffer: extraction buffer without complete and PMSF

For the affinity purification with streptavidin beads, 2 mL of densely packed ground plant material was resuspended in 2 mL extraction buffer and incubated on a rotor wheel at 4°C for 10 min. One µl lysonase (Millipore #71230) was added to digest cell walls and DNA/RNA and the suspension was incubated on the rotor wheel at 4°C for another 15 min. The extracts were then distributed into 1.5 mL reaction tubes and sonicated 10 times at 4°C for 30 s with 30 s breaks. The suspension was centrifuged at 15000 x g for 15 min at 4°C to remove cell debris. The clear supernatant was then applied to a PD-10 desalting column (GE healthcare #52130800) following the gravity protocol, to remove free biotin in excess. After equilibrating the resin with five volumes of ice-cold equilibration buffer, 2.5 mL of the protein extract was loaded on the column. After the protein extract fully penetrated the resin, 1 mL of equilibration buffer was added. The proteins were eluted by adding 3.5 mL of equilibration buffer. Protein concentration was then measured using a BCA protein assay (section 2.14.3) and then normalised across samples. The samples were then supplied with protease inhibitors and transferred into 5 mL Eppendorf tube. One-hundred µL of Dynabeads™ MyOne™ Streptavidin C1 (Invitrogen #65001) bead slurry was washed with extraction buffer and added to the samples, which were then incubated on a rotor wheel at overnight (16 hours) 4°C.

Subsequently, the beads were washed with 1 mL of the following solutions at room temperature (as described in Cho et al. [2020]): twice with lysis buffer for 2 min, once with 1 M KCl, once with 0.1 M Na_2CO_3 for ∼10 s, once with 2 M urea in 10 mM Tris-HCl pH 8.0 for ∼10 s, and twice with lysis buffer for 2 min. As the beads are paramagnetic, a magnetic rack was used to separate them out of solution after each wash. After the last wash, 2% of the beads were boiled in 50 µL 2x Laemmli buffer supplemented with 20 mM DTT and 2 mM biotin at 95°C for 5 min and stored at -20°C for Western blotting. The rest of the beads was washed twice with PBS buffer. The supernatant was discarded, and the beads were immediately snap frozen and stored at -80°C for trypsinization. The samples were processed in batches with one replica of each sample in each batch. The samples from the third batch 3 were used for immunoblots to assess the pull down efficiency before trypsinization.

2.19 On-beads tryptic digest

All the reagents were mass spectrometry grade and the steps were performed under a laminar flow hood to minimise contamination.

2.19.1 Agarose beads

The protocol used followed the manufacturer's (Chromotek) guidelines, which were adapted from Hubner et al. [2010].

Lysis buffer: 50 mM Tris/HCl pH 7.5, 150 mM NaCl

Wash buffer: Same as lysis buffer

Elution buffer 1 (EB1): 50 mM Tris-HCl pH 7.5, 2 M urea, 5 µg/ml sequencing grade modified trypsin (Promega #V5111), 1 mM DTT

Elution buffer 2 (EB2): 50 mM Tris-HCl pH 7.5, 2 M urea, 5 mM iodoacetamide

The beads were resuspended in 500 µL ice-cold lysis buffer and centrifuged at 2,500 x *g* for 2 min at 4°C, then the supernatant was discarded. The was repeated twice. The beads were resuspended in 500 µL of ice-cold wash buffer and centrifuged at 2,500 x *g* for 2 min at 4°C. The supernatant

was discarded, and the wash repeated once. The beads were then resuspended in 25 µL of EB1 and incubated in a thermomixer with 400 rpm mixing at 30°C for 30 min. After incubation the beads were sedimented 2,500 x *g* for 2 min at 4°C. The supernatant was set aside in a separate vial ("digest") while the beads were resuspended in 50 µL EB2 and sedimented at 2,500 x *g* for 2 min at 4°C. The supernatant was then combined with the previous supernatant ("digest"). The step was repeated once, to obtain a final volume of 125 µL. The digestion was continued in a thermomixer with 400 rpm shaking at 32°C overnight, away from light. The next day, the reaction was stopped by adding 1 µL trifluoroacetic acid. The digest was stored at -80°C and then sent to the mass spectrometry facility.

2.19.2 Streptavidin beads

The protocol for the streptavidin beads the trypsinization step was extended respect to agarose beads, which was shown to greatly improve the peptide recovery [Mair et al. 2019].

Trypsin buffer: 50 mM Tris pH 7.5, 1 M Urea, 1 mM DTT, 0.4 µg trypsin (Promega #V5111)

The frozen streptavidin beads were thawed and washed twice each with 1 mL 50 mM Tris pH 7.5 (transferred to new tube with first wash) and 1 mL 2 M Urea in 50 mM Tris pH 7.5. The buffer was replaced with 80 µL trypsin buffer and the beads were then incubated with shaking at 25°C for 3 hours. The beads were separated on a magnet and the solution was transferred into a fresh tube. The beads were then washed twice with 60 µL 1 M Urea in 50 mM Tris pH 7.5. The supernatant from the washes were combined with the first digest solution for a final volume of 200 µL. The solution was reduced by adding DTT to a final concentration of 4 mM and incubated for 30 min with shaking at 25°C. The eluates were then alkylated by adding iodoacetamide to a final concentration of 10 mM and incubated for 45 min with shaking at 25°C. A supplemental 0.5 µg trypsin was added, and the digest completed overnight at 25°C with shaking and away from light. The digest was acidified by adding formic acid to a final concentration of ∼1% to stop the reaction. The samples were stored at -80°C and then sent to the mass spectrometry facility.

2.20 Mass spectrometry analysis

The mass spectrometry was performed by Dr. Kieran Wynne at the proteomics facility in the Conway Institute of Biomolecular and Biomedical Research, University College Dublin.

2.20.1 LCMSMS Method (Bruker timsTof Pro/Evosep One)

Samples were run on a Bruker timsTof Pro mass spectrometer connected to an Evosep One liquid chromatography system. Tryptic peptides were resuspended in 0.1% formic acid and each sample was loaded on to an Evosep tip. The Evosep tips were placed in position on the Evosep One, in a 96-tip box. The autosampler is configured to pick up each tip, elute and separate the peptides using a set chromatography method (30 samples a day) [Bache et al. 2018]. The mass spectrometer was operated in positive ion mode with a capillary voltage of 1750 V, dry gas flow of 3 l/min and a dry temperature of 180°C. All data was acquired with the instrument operating in trapped ion mobility spectrometry (TIMS) mode. Trapped ions were selected for ms/ms using parallel accumulation serial fragmentation (PASEF). A scan range of 100-1700 m/z was performed at a rate of 5 PASEF MS/MS frames to 1 MS scan with a cycle time of 1.03 s [Meier et al. 2018].

2.20.2 Chromatography Buffers

Buffer B: 99.9% acetonitrile, 0.1% formic acid

Buffer A: 99.9% water, 0.1% formic acid

2.20.3 Maxquant peptide search

The raw data was searched against the *Arabidopsis thaliana* subset of the SwissProt/Trembl database (reviewed) using the search engine Maxquant (release 2.0.3.0) using specific parameters for trapped ion mobility spectra data dependent acquisition (TIMS DDA). Each peptide used for protein identification met specific Maxquant parameters, i.e., only peptide scores that corresponded to a false discovery rate (FDR) of 0.01 were accepted from the Maxquant database search. The normalised protein intensity of each identified protein was used for label free quantitation (LFQ) [Cox et al. 2014].

2.21 RNA extraction

RNA extractions were performed with the GeneJET Plant RNA Purification Kit (ThermoFisher #K0801). Column based RNA extraction systems are not suitable for long-read sequencing transcriptomics because they introduce mechanical shearing. For this reason, the RNA for Nanopore sequencing was extracted with a phenol-chloroform method, adapted from Toni et al. [2018], which preserves RNA integrity and increases yield. Before cDNA synthesis for qPCR or short-read sequencing, the DNA was completely depleted from RNA samples using a DNase-free removal kit (Ambion #AM1906).

The meristematic tissue was ground with a pestle in liquid nitrogen and then the tissue was immediately added directly to a tube containing 1 mL TRIzol™ (ThermoFisher #15596026). I have found empirically, that adding the tissue to the TRIzol reagent, and not vice versa, improves the RNA integrity. The suspension was vortexed thoroughly for 45 s and then left at room temperature for 3 min. Then, 200 µL RNase-free chloroform was added to the tube and mixed vigorously by hand for 15 s. The sample was transferred to a clean Phasemaker™ tube (Invitrogen #A33248), left at room temperature for 3 minutes, and then centrifuged at 1,2000 x *g* for 15 min at 4°C. During the centrifugation step, a new Phasemaker tube was prepared, contaning 200 µL of RNase-free chloroform. At the end of the centrifugation step, the RNA-containing upper aqueous phase was transferred in the new tube containing chloroform and the choloform extraction was then repeated once. After the second extraction, the upper acqueous phase was transferred to a clean 1.5 mL Eppendorf tube containing 500 µL of isopropanol to precipitate the RNA. The solution was mixed inverting the tube by hand 10-20 times and left for 10 min at room temperature. The RNA was pelleted by centrifugation at 1,2000 x *g* for 10 min at 4°C. The supernatant was discarded, and the pellet washed with RNase-free ethanol. The pellet was positioned at the bottom of the tube by centrifugating at 1,2000 x *g* for 10 min at 4°C. The ethanol wash was repeated at least twice, until the tube stopped smelling of phenol. After pulse spinning the samples, the remaining ethanol was removed carefully with a thin pipette tip, and the tubes were left open to air-dry for 5 min at room temperature. To make sure that the ethanol was completely removed, the tubes were heated for 3 min at 65°C with the lid open. The RNA was resuspended in 100 µl of RNase-free water.

To ensure the RNA was completely resuspended, the sample was heated for 2 min at 65°C, vortexed at very low speed, and then placed immediately on ice.

The RNA integrity number (RIN) was computed by running 1 µL of the RNA on an Agilent TapeStation using an RNA ScreenTape cartridge (Agilent #5067-5576). The quantity and purity of the RNA was quantified with a spectrophotometer, aiming for absorbance ratios of $OD_{260/280} > 2.0$ and $OD_{260/230} > 2.0$. For Nanopore sequencing, the total RNA extraction was poly(A)$^+$ selected on the same day.

2.22 Poly(A)$^+$ RNA selection

The poly(A)$^+$ fraction was selected from 100 µL of total RNA solution, using the NEXTflex® Poly(A) Beads (PerkinElmer #NOVA-512980) according to the manufacturer's protocol. The poly(A)$^+$ RNA concentration was then measured using a Qubit™ RNA High Sensitivity assay (Invitrogen #Q32852). To check that the ribosomal RNA was successfully depleted, 2 µL of the RNA were loaded on a High Sensitivity RNA ScreenTape (Agilent #5067-5579) and analysed on a TapeStation. The ribosomal peaks should be absent from the curve computed by the TapeStation software in the poly(A)$^+$ fraction.

2.23 cDNA synthesis by reverse transcription

The RevertAid H Minus Reverse Transcriptase (ThermoFisher #EP0451) was used for cDNA synthesis following the manufacturer's protocol. The reaction was set up using: 1 µg of DNA-depleted RNA in 11.5 µL of RNase-free water, 1 µL of Oligo(dT)18 primer, 4 µL of 5X RT Buffer, 0.5 µL of Ribolock, 2 µL of dNTPs and 1 µL of Revertaid. The reaction was incubated for 1 hour at 42°C followed by 5 min at 70°C for inactivation. 1 µL of the cDNA reaction was used for qPCR.

2.24 Quantitative PCR (qPCR)

For qPCR reactions, the primers were designed to have melting temperatures of ∼60°C, a length between 18-26 nucleotides, an amplicon length between 60-200 nucleotides, a $\Delta G >$ -8 for self and hetero dimers and a $T_m < 40$°C for

hairpins. To optimise these parameters, I used the online tool Oligo Analyzer from the Integrated DNA Technologies (IDT) website. The primers were designed using Primer BLAST (NCBI), which also ensures the specificity of the primer pair. It was not always possible maintain these standards, because of the highly palindromic and repetitive nature of the region around splice sites.

The cDNA reaction was prepared by mixing 5 µL of 2x SYBR green mastermix (Roche), 1 µL of cDNA, 1 µL of 20 µM primers (2 µL each as the final concentration), 3 µL of molecular biology grade water, and then loaded on the Real-Time PCR thermal cycler. The expression of each gene was calculated from C_p values that were computed using the second-derivative max algorithm, inside the Roche analysis software. The expression was then normalised with an appropriate reference gene using the formula:

$$RelativeExpression = -2^{gC_p - rC_p}$$

Where gC_p is the C_p value of the region amplified from the target gene and rC_p is the C_p value for the region amplified from the reference gene. For qPCR analysis of cDNA generated from Arabidopsis tissue the reference gene *PP2AA3* was used, also known as *REF1* [Wang et al. 2014].

2.25 Library preparation and Nanopore sequencing

The Nanopore sequencing library was prepared from ∼200 ng of poly(A)$^+$ RNA using the direct cDNA sequencing kit (Oxford Nanopore #SQK-DCS109), following the manufacturer's protocol. In brief, the poly(A)$^+$ RNA is reverse transcribed once, using a primer on the poly(A) tail and one on the 5′-cap. These two primers are the markers that allow to separate the reads spanning full-length transcripts from the rest, using `pychopper2`. After the RNA/DNA duplex was cleaned up with AMPure® XP beads (Beckman Coulter #A63880), the RNA strand was digested, and a new DNA strand was synthesised with a second reverse transcription reaction, using a primer complementary to the one used on the 5′-cap in the first reaction. The cDNA was cleaned up and then end-repaired to prepare it for adapter ligation. The nanopore adapters are conjugated with motor proteins that are responsible for the translocation of the DNA molecule through the pore during the sequencing run. After adapter ligation, the library is cleaned up and ready to

be loaded onto the flow cell. The cleanup step after adapter ligation uses a special buffer, which ensures that any non-ligated adapter is washed away from the library. This is critical step. Free adapters are detrimental for the sequencing run because the motor proteins are able to bind the pore and start consuming ATP even without DNA translocation, which depletes the supply of ATP and takes away valuable pores that could be used for sequencing. The Nanopore flow cell is shipped with a storage buffer that slows down the pores' degradation. In a process called flow cell "priming", the storage buffer is washed away, and the flow cell is prepared for the sequencing run by adding a buffer containing the ATP necessary for the motor proteins and positive ions to create the required current for the base detection. The library is then loaded dropwise on the flow cell. The library is mixed with special beads that weight down the DNA, keeping it close to the pores during sequencing. The library was loaded on a 9.4.1 chemistry flow cell (Oxford Nanopore #FLO-MIN106D) and sequenced on a MinION sequencer (Oxford Nanopore #Mk1B).

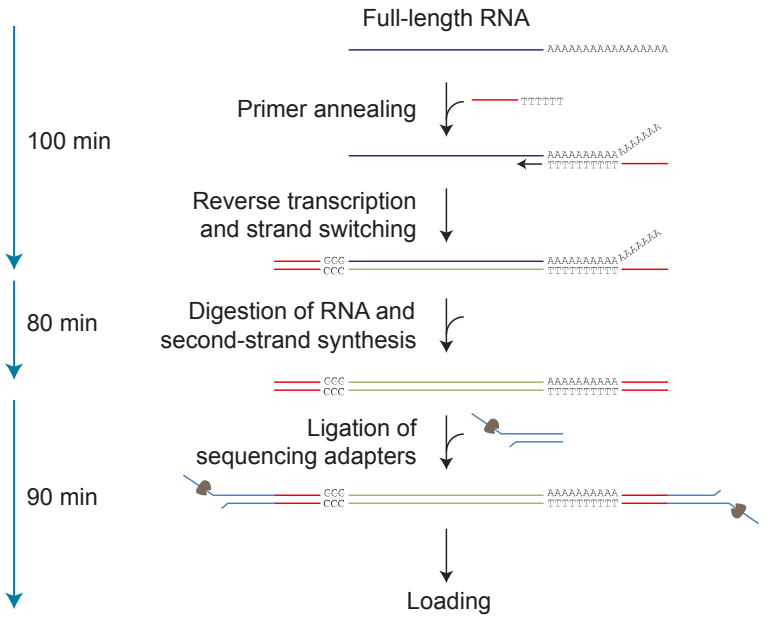


Figure 2.1: Nanopore direct cDNA sequencing protocol overview. Figure obtained from nanoporetech.com.

I made a couple of improvements to the standard protocol. The starting

material I used was 200 ng, more than the suggested 150 ng, which resulted in a more efficient use of the pores on the flow cell when coupled with an increased amount of adapter solution during the adapter ligation step. I obtained the 10 million reads advertised by the manufacturer on more than one occasion (see Table 3.1).

A typical Nanopore run can last 48 hours, after 72 hours the flow cell is usually completely depleted. Not only the pores degrade but also the difference in voltage potential between the two sides of the membrane equilibrates, halting the ions flow, which is necessary to read the bases while they pass through the pore. During the run, the output asymptotically decreases until it comes to an almost complete halt, which usually happens at the 48 hours mark. Most of the sequencing output is produced in the first 6 hours, which means that the pores have to be actively sequencing most of the time to achieve good results. The amount of DNA ends that are ligated with an adapter is one of the most important factors in a Nanopore sequencing run. More "free ends" means a higher chance that a motor protein is tethered to a pore and start sequencing the ligated fragment. That is why I increased the amount of input RNA together with the adapter amount. It is detrimental to have cDNA molecules that don't have an adapter ligated because they end up clogging the membrane, occluding the pores. I also took great care to avoid mechanical stress for the RNA/cDNA fragments. Whenever possible, I avoided pipetting in favour of gently flicking the tube (i.e., when resuspending beads).

2.26 Software and software environments

The Anaconda package manager was used to manage the installation of most of the bioinformatics tools used in this thesis [ana 2020]. I added a file containing my personal working environment to the `GitHub` repository I used to manage the code for this thesis, which makes possible to replicate my setup.

2.26.1 Code and data availability

All the code, datasets and images generated in this thesis are hosted Google Drive at https://drive.google.com/drive/folders/1rNwDOQ4CFcH4FGyWNY-$_r kR5nevCSslH?usp = sharing$.

2.26.2 Alignments visualisation

The read alignments were visualised using the Integrated Genome Browser (IGV), developed at the Broad Institute [Robinson et al. 2011].

2.26.3 Sequence alignment

Multiple sequence alignments were obtained using the `ClustalOmega` algorithm via the official command line interface, while the pairwise alignments was performed using `needle`, an implementation of the NeedleMan-Wunsch alignment algorithm and part of the EMBOSS suite [Sievers et al. 2011; Rice et al. 2000]. Both tools were run with default settings. The plots were produced using the `ggmsa` package in `R`.

2.26.4 3D protein modeling

The 3D model was approximated using the `PHYRE`^{2} web portal [Kelley et al. 2015]. The amino acid sequences were submitted the server, which performed the computation. The predicted structures were provided in `.pdb` format, which were then visualised and superimposed on the `Mol Viewer` web app [Sehnal et al. 2021].

2.26.5 Protein disorder quantification

The per base disorder was estimated using two algorithms, IUPred2A and ANCHOR, through the `Python` package `iupred3.py` [Dosztányi et al. 2009; Mészáros et al. 2018; Erd et al. 2021]. The raw data was then plotted in `R` using `ggplot2` [Wickham 2016].

2.26.6 Cloning and plasmid design

Plasmid sequences were manipulated using the free software ApE [Davis and Jorgensen 2022]. The cloning of the entries into the final destination vectors were run first *in silico* using Benchling® and then translated to make sure that the final construct was in frame and would produce the expected protein.

2.27 Growth media and buffers

Luria-Bertani medium (LB): 10 g/L bacto-tryptone, 5 g/L bacto-yeast extract, 10 g/L NaCl. 1.5% (w/v) agar was added to make LB agar plates.

Murashige and Skoog (MS): 2.2g/L Murashige and Skoog (MS) powder pH adjusted to 5.7 with KOH. 0.6% (w/v) agar was added to make MS agar plates.

Super Optimal Broth (SOB): 2% (w/v) tryptone, 0.5% (w/v) yeast extract, 0.05% (w/v) NaCl, 10 mM MgSO4, 10 mM $MgCl_2$

Plant soil: A mix of compost, vermiculite, and perlite at a ratio of 3:1:1 (volume) was sterilised for 60 min at 120°C and cooled down before planting seeds.

Infiltration medium: 10 mM MES pH 5.6, 10 mM $MgCl_2$, 150 µM acetosyringone

Dexamethasone (DEX) solution: 10 µM dexamethasone (in ethanol), 0.015% Silwet L-77

Dihydrotestosterone (DHT) solution: 500 µM dihydrotestosterone (in ethanol), 2.5 mg/ml (2-Hydroxypropyl)-β-cyclodextrin, 0.015% Silwet L-77

TB buffer: 10 mM PIPES/KOH pH 6.7, 15 mM CaCl2, 250 mM KCl, 55 mM $MnCl_2$

PBS buffer: 8 g/L NaCl, 0.2 g/L KCl, 1.15 g/L Na_2HPO_4, 0.2 g/L KH_2PO_4

PBS-T: Like PBS but with 0.05% Tween 20 (v/v)

2X SDS loading Dye (LD): 25% (v/v) 4x stacking buffer, 20% (v/v) glycerol, 4% (v/v) SDS, 2% (v/v) β-mercaptoethanol, 1% (w/v) bromophenol blue

2.28 Antibiotics

Carbenicillin: 100 mg/µl carbenicillin in sterile water (used as 1000X dilution)

Spectinomycin: 100 mgµl spectinomycin in sterile water (used as 1000X dilution)

All the antibiotics were filter-sterilised after preparation.

2.29 Hardware

Stereomicroscope: SZM-LED1 stereo microscope (OPTIKA)

Sonicator: Bioruptor® Pico sonication system (Diagenode)

Tapestation: Agilent 2200 TapeStation device

Qubit: Qubit fluorometer version 4 (Invitrogen #Q33238)

Real-time PCR thermal cycler: Lightcycler 480 (Roche)

Luminescent image analyser: FujiFilm LAS-3000

2.30 Appendix

Table 2.2: Antibodies

Antibody	Code	Secondary	Chapter
αV5	Invitrogen #R96025	αMouse-HRP	3
αCMyc	Sigme-Aldrich #M5546	αMouse-HRP	4
αGFP	Roche #1181446001	αRabbit-HRP	4
streptavidin-HRP	Invitrogen #S911	*Not needed*	4

Note:
The table describes the primary antibodies used for Western blotting, and the respective secondary antibodies, if any.
HRP: horse radish peroxidase (for chemiluminescence).

Table 2.3: Plasmids generated for this thesis

Code	Construct	Chapter
pAF13	**B** entry GFP N-TAG	3
pAF14	**B** entry StrepII-V5 N-TAG	3
pAF18	**C** entry PI.1 mut CDS	3
pAF19	**C** entry PI.2 mut CDS	3
pAF20	**C** entry gPI(NOSTOP) mut CDS	3
pAF21	**C** entry gPI mut CDS	3
pAF22	**D** entry V5-StrepII C-TAG	3
pAF1	pPI::StrepII-V5-gPI	3
pAF2	pPI::gPI-V5-StrepII	3
pAF3	pUBQ10::StrepII-V5-gPI	3
pAF4	pUBQ10::gPI-V5-StrepII	3
pAF5	pPI::GFP-gPI	3
pAF6	pPI::gPI-GFP	3
pAF9	pPI::PI.1	3
pAF10	pPI::PI.2	3
PL188	**A** entry pTRY	4
PL189	**A** entry pCPC	4
PL190	**C** entry gTRY	4
PL191	**C** entry gCPC	4
PL192	**C** entry TurboID-Myc	4
PL204	**D** entry GFP	4
PL184	pTRY::gTRY-TurboID-Myc	4
PL185	pTRY::NLS-GFP-TurboID-Myc	4
PL186	pCPC::gCPC-TurboID-Myc	4
PL187	pCPC::NLS-GFP-TurboID-Myc	4
PL199	TRY::gTRY-GFP	4
PL200	CPC::gCPC-GFP	4

Note:
A, B, C and D in bold face refer to the entry position in the Greengate system.

Table 2.4: Primers for the cloning of GreenGate constructs

Name	Sequence
AF17	aacaggtctcaggctATGGATAACACTGACCGTC
AF18	aacaggtctcactgaGGAAGGATAGATAGAAAAGCG
AF19	aacaggtctcaggctATGTTTCGTTCAGACAAGG
AF20	aacaggtctcactgaTTTCCTAAAAAAGTCTCTTCGT
AF21	aacaggtctcaacctACGCTTTAACCACTTCGTC
AF22	aacaggtctcatgttCGAACTAATCTGAAGACACGA
AF25	aacaggtctcaacctGGGTCGATTCAGATTTATAATAG
AF26	aacaggtctcatgttTACTATTGAAGTAAGAAAAG
AF27	aacaggtctcatcaggcgccgccATGAAAGACAATACTGTGCC
AF30	aacaggtctcagcagCTACAGATCTTCTTCAGAAATAAGTTTTT∼ ∼GTTCgcctgaCTCGAGTGCGGCCG
AF51	aacaggtctcatcaggcgccgccATGGTGAGCAAGGGCGAG
AF52	aacaggtctcagcagCTACTTGTACAGCTCGTCCATG
AF60	aacaggtctcaggctccATGGGTAGAGGAAAGATCGAGA
AF61	aacaggtctcactgaATCGATGACCAAAGACATAATCTTTT
AF62	aacaggtctcaaacaATGTGGTCGCATCCTCAAT
AF63	aacaggtctcaagccGCCTGAGCCGCCC
AF64	aacaggtctcatcagccGGCGGCTCAGG
AF65	aacaggtctcagcagTCATTTTTCAAATTGAGGATGCGA
AF68	CGAGACCACCAGATGGAGATCCT
AF69	AGTGAGTTGCCGTTGCTCCT
AF72	CACATGAATATATTTAACTTCTTATTTGGTCTGCTTGG
AF73	CCAAGCAGACCAAATAAGAAGTTAAATATATTCATGTG
AF74	AAGTCCGAGAtCACCAGATGG
AF75	TGTCGAGGCCATGTTCAAT

Table 2.5: Primers for qPCR

Primer	Sequence	Description
AF97	TGAGTTGCCGTTGCTCCTC	rv PI
AF98	CCGAGACCACCAGATGGA	fw PI on splice site
AF99	CTTCTTATTTGGTCTGCTTGGC	fw PI in intron
AF103	GACACATGCTTGCTGGAACC	rv ROC1 after exon
AF104	TGTAGCTTGGGGGGTTTGC	fw ROC1 on splice site
AF105	TGCTCAATGAAAGATGACCACT	fw ROC1 in exon
AF109	TGATTTCACTTGTGTTCTGTTGAGT	fw GIF1 in intron
AF110	GCCGAGAATCAAGCAAGGCT	fw GIF1 on splice site
AF111	ATGCACACTTGGTGGCTGAG	rv GIF1 after splice site
AF112	TGATGTTTTTGTGTGGTGTGAA	fw SEP3(first intron) in intron
AF113	CGAGCATGCTTCGGACACTG	fw SEP3(first intron) on splice site
AF114	GGGTTCTGGTGCTCCATAGT	rv SEP3(first intron) after splice site
AF115	TTATGGAGCAGGTATCAGGGG	fw SEP3(last intron) in intron
AF116	GATCGGGTATCAGGGGCAAC	fw SEP3(last intron) on splice site
AF117	GGTGTCATAAGGTAACCAACCCA	rv SEP3(last intron) after splice site
AF118	TGAACGTTTCAAAATGCGGT	fw PIN6 in intron
AF119	CAGACCTACTCAGTACATTGGTT	fw PIN6 on splice site
AF120	AGGCCCAAGAGGACGTAGTA	rv PIN6 after splice site
AF121	CTCACCTGCCAATTAATATCACACA	rv YAB3 in intron
AF122	GAGGAACACTCACAGCAAGGA	rv YAB3 on splice site
AF123	CATGTCGTCCTCCTCAGCTC	fw YAB3 before splice site
AF124	CACAACAATCTTCACTAACATACCG	rv REM1 in intron
AF125	CCGTCTTCTCCGATATGTTCTCTAT	rv REM1 on splice site
AF126	GTAGCTGTGGTGTCCCCATC	fw REM1 before splice site
AF127	ACAAGAGGGGATGAACTGAGG	rv CCL(first intron) in intron
AF128	GCTGGAAAAACAGAAAGGATGGA	rv CCL(first intron) on splice site
AF129	CTTCCTCACGCAACCAAAGC	fw CCL(first intron) before splice site
AF130	TGTTTCCAAGTCCAAAAATCACA	rv CCL(last intron) in intron
AF131	AAACCGTTTTGGGGCTGTTA	rv CCL(last intron) on splice site
AF132	AGCTCGATACAAAGCCCGTC	fw CCL(last intron) before splice site
DM534	ACCAGATTCTTCGTGCAAAGATAGCTG	fw on AG
DM535	AAGCTGCTCGTAGTTAGATCCTCCTG	rv on AG
BT744	TCCGGACTCAGATCAAGCA	fw on AP3
BT745	TTCCATTTCATCCTCAAGACG	rv on AP3
DM242	AAGCGGTTGTGGAGAACATGATACG	fw on REF1
DM243	TGGAGAGCTTGATTTGCGAAATACCG	rv on REF1

Chapter 3

Unraveling the transcriptome dynamics during floral transition using long-read Nanopore sequencing

3.1 Introduction

3.1.1 The role of alternative splicing in the regulation of development

Since the advent of DNA microarray technology more than 20 years ago, numerous transcriptomics papers have been published on flower development using initially primarily the model plant Arabidopsis but in recent years also many other angiosperms of diverse groups. Gene expression data for a variety of tissues and developmental stages is now available, and this extensive collection of data can be explored using powerful databases such as those provided by The Arabidopsis Information Resource (TAIR).

From this data, as well as from the results of other global approaches such as genome-wide localization studies, which have been conducted for many key floral regulators, we have gained a much better understanding of the gene regulatory networks underlying flower development. However, the data available to date still do not describe the full complexity of the floral transcriptome. For example, information on cell-type specific gene expression

during flower development is limited although new approaches such a single-cell sequencing may be able to fill this knowledge gap in the near future. Other areas of floral transcriptomics that are currently not well explored include the role of alternative splicing and the functions of short and long non-coding RNAs (lncRNAs) in flower development.

An analysis of alternative splicing, in particular, holds the potential to provide new insights into the genetic control of flowering. As in animals, many plant genes give rise to pre-mRNAs that are spliced into one or several different mRNA isoforms. While in humans, more than 90% of multi-exon genes undergo alternative splicing, this number is thought to be smaller in plants [Baralle and Giudice 2017; Kim et al. 2014; Pan et al. 2008]. In the model plant Arabidopsis the proportion of genes with alternative splicing is estimated to be around 60%, with the most prominent event being intron-retention in 40% of the cases [Zhang et al. 2017].

Alternative spliced isoforms from a given gene can then be translated into proteins with different functions (e.g., in humans, approximately 37% of alternatively spliced genes generate multiple protein isoforms), substantially widening the proteome landscape. However, a sizeable portion of alternative spliced mRNAs never undergo translation because they are degraded by nonsense mediated decay, for example, when there is an early stop codon caused by a mispliced intron [Baralle and Giudice 2017; Yang et al. 2016].

Alternative splicing has been shown to play key roles in a multitude of different processes in eukaryotes, including the response to stress and the regulation development. It is thought that alternative splicing can serve as a "quick" and flexible way for organisms to adapt their proteome or transcript levels to a variety of situations [Verta and Jacobs 2022; Kelemen et al. 2013].

Transcriptomics analysis in mice brain cells revealed that splicing events act as "switches" during early and late developmental stages of neuronal cells. Similar mechanisms were observed for erythropoiesis, liver and heart development [Bhate et al. 2015; Cheng et al. 2014; Giudice et al. 2014; Gupta et al. 2018; Su et al. 2018; Weyn-Vanhentenryck et al. 2018].

In plants, alternative splicing has been studied extensively in relation to the response to environmental cues and biotic/abiotic stresses [Filichkin et al. 2015; Herz et al. 2019; Huertas et al. 2019; James et al. 2012]. A striking example is *FLOWERING LOCUS M (FLM)*, which encodes a MADS-domain transcription factors and is involved in the temperature-dependent control of

flowering time in Arabidopsis. It was shown that FLM transcripts undergo alternative splicing and that different ambient temperatures lead to the predominant accumulation of one or another spliceform. The proteins synthesized from these alternative transcripts interact with another MADS-domain protein, SHORT VEGETATIVE PHASE (SVP), resulting in complexes that either promote or repress flowering depending on the FLM isoform incorporated into the dimer [Capovilla et al. 2017; Sureshkumar et al. 2016]. The role of alternative splicing in plant development is less well studied but transcriptomics datasets of Arabidopsis root tissues suggest that changes in the proportion of different isoforms, or differential transcript usage, is more prominent in older differentiated tissues when compared to younger, undifferentiated ones, located proximal near the root apex. Thus, alternative splicing and differential transcript usage may play important roles in developmental processes in plants [Li et al. 2016].

3.1.2 The roles of long non-coding RNAs in the regulation of development

Another regulatory mechanism for gene expression that has recently gained much attention is based on long non-coding RNAs (lncRNAs). LncRNAs are non-coding transcripts typically defined as being longer than 200 bp (to distinguish them from microRNAs). LncRNAs can be catagorized based on their location in the genome, and their orientation of transcription. They are called intergenic or intragenic lncRNAs depending on whether they are located in-between genes or within a gene locus, respectively. Furthermore, for intragenic lncRNAs, one can distinguish between sense and antisense lncRNAs depending on whether they are transcribed in the same (sense) or opposite (antisense) direction of the gene locus they are on.

LncRNAs have multiple functional mechanisms: they can act as scaffolds, decoys or signals, target genomic regions in cis or trans, and contrast microRNA activity by antisense interference. They seem to play important roles during development and were found to be putative targets for cancer treatments. The first published lncRNA, *X inactive specific transcript (Xist)*, acts in trans to silence the X chromosome in *Drosophila melanogaster* by chromatin remodeling, possibly via recruitment of the Polycomb Repressive Complex 2 (PRC2) [Wutz and Jaenisch 2000]. In primates, *lncND (NeuroDevelopment)* has been shown to regulate neuronal development acting

as a microRNA "sponge", sequestering microRNA miR-143-3p, thus delineating a new brain development regulatory mechanism [Rani et al. 2016]. Many lncRNAs, as well as other protein coding genes, when actively transcribed, can affect the expression of neighboring genes in *cis*, a mechanism that is conserved in mammalian genomes [Engreitz et al. 2016]. The study of these cis-regulatory networks is a fertile field that has expanded greatly in the last few years. One example is the *HOX* gene family, which regulate early morphogenesis in animals along the anterior-posterior axis. Over two-hundred non-coding RNAs (ncRNAs) were identified that are transcribed from the four *HOX* gene clusters in humans, many of which were shown to affect the expression of neighboring genes [Rinn et al. 2007]. The expression of one of these ncRNAs, *HOX Transcript Antisense RNA (HOTAIR)*, promotes cancer metastasis by reprogramming chromatin state and H3K27Me3 distribution. Furthermore, in cells overexpressing *HOTAIR*, the PRC2 complex occupancy on chromatin was shown to resemble the one of embryonic fibroblasts, increasing cancer invasiveness [Gupta et al. 2010].

A broad transcriptomics study analysed the expression pattern of lncRNAs across several developmental time-points in seven major organs, from early organogenesis to adulthood, in human, macaque, mouse, rat, rabbit, opossum, and chicken [Sarropoulos et al. 2019]. Many of the lncRNAs investigated exhibited dynamic patterns of expression, suggesting that they may have functional roles in the control of development. Furthermore, there was a progressive transition from conserved broadly expressed lncRNAs at early time points, to lineage and organ-specific lncRNAs at later stages of development. Although similar have yet to be conducted in plants, there is growing evidence for the differential expression of lncRNAs during development. Once floral meristems have been specified, organ primordia arise in the different floral whorls. The specification of these primordia into different floral organ types (i.e., sepals, petals, stamens and carpels) directly depends on the activities of the floral homeotic genes, whose activities are summarised by the so-called ABC model of floral organ identity specification [Jiao and Meyerowitz 2010; Tian et al. 2019]. Another example for the function of lncRNAs in plants comes from rice where *LRK Antisense Intergenic RNA (LAIR)*, is transcribed from the antisense strand of the *LEUCINE-RICH REPEAT RECEPTOR KINASE (LRK)* gene cluster. The *LRK* cluster maps to a Quantitative Trait Locus that regulates grain yield. Indeed, overexpres-

sion of LAIR was shown to lead to an upregulation of *LRK* genes, increasing grain yield [Wang et al. 2018]. Thus, lncRNAs provide a potential source of genetic variability that could be harnessed to improve crop plants.

Some of the best studied lncRNAs in plants regulate development in response to environmental cues. *COLD ASSISTED INTRONIC NONCODING RNA (COLDAIR)* is an intronic lncRNA transcribed from the floral repressor *FLOWERING LOCUS C (FLC)*. During vernalization, *COLDAIR* expression is upregulated and leads to chromatin remodeling at the *FLC* locus. Specifically, repressive histone H3 lysine-27 tri-methylation marks (H3K27me3) are deposited by PRC2 (see above), leading to *FLC* silencing and flowering transition [Kim et al. 2017]. *COLDWRAP* and *COOLAIR* are additional lncRNAs involved in *FLC* regulation: they are transcribed from a cryptic *FLC* promoter and act together with *COLDAIR* in the response to cold temperatures [Csorba et al. 2014; Kim and Sung 2017].

A lncRNA involved in the response to internal cues is *AUXIN REGULATED PROMOTER LOOP (APOLO)*, which is upregulated in response to auxin, a key phytohormone that is involved in most, if not all, processes during plant development. *APOLO* transcription promotes the formation of a repressive chromatin-loop at the promoter of the *PINOID (PID)* gene, which encodes a known positive regulator of cellular auxin efflux [Ariel et al. 2014].

The examples listed above suggest that alternative splicing and lncRNAs play important and possibly widespread roles in the control of development. However, their functions and the scale to which they contribute to developmental processes has yet to be fully understood. A key factor hampering the analysis of alternative splicing and lncRNAs is the short read length produced by current second-generation DNA sequencing approaches. Third-generation sequencing methods such as Oxford Nanopore that can produce long reads promise to solve these problems and open new possibilities to accurately quantify the expression of single transcripts, isoforms, or intragenic lncRNAs. In the following sections, I will briefly compare and contrast the strengths and limitations of second and third generation sequencing approaches for the analysis of alternative splicing and lncRNA expression.

3.1.3 Second-generation transcriptomics approaches

The advent of high-throughput second generation sequencing technologies resulted in a huge leap forward in our understanding of the gene expression programmes underlying the development of organisms. Several companies have entered the market with their own patented sequencing technologies, and although they substantially vary in the mechanism they employ, at the core, they all utilise the same 'shotgun' sequencing approach. The shotgun approach relies on an initial fragmentation of the DNA into small pieces, usually around 200 bp, before the sequencing step. The high number of DNA fragments obtained allows sequencing to occur in parallel on a massive scale. This increased throughput comes at the cost of losing information contained in long DNA/RNA fragments. For example, if the expression of a gene leads to multiple RNA isoforms, it is not trivial to reconstruct them from the short reads generated by second-generation sequencing approaches as individual reads cannot be readily attributed to one isoform or another. For the same reason, read counting to assess expression of individual isoforms can likewise present its own challenges. Ad-hoc bioinformatics tools can in part overcome these shortcomings, but the inherent limitations of the technology remain the main source of errors in transcriptome assemblies and transcript expression estimates. There are currently two main approaches for transcriptome assembly for second generation sequencing technologies:

Genome-guided assembly As suggested by the name of the approach, reads are first aligned to the genome sequence and then isoforms are identified using primarily reads that span exon-exon junctions. This is usually the best performing algorithm when a reliable genome assembly is available [Behera et al. 2021].

De novo assembly De novo assembly is used when a genome assembly is not available. The reads are aligned directly against each other and the transcriptome is build up from that information alone. Because of the lack of a reference sequence, this approach is inherently more error-prone than a genome-guided assembly.

Counting the reads that belong to a given transcript has a similar set of challenges. During the read counting phase, reads are assigned to an isoform and the results are then used to estimate the expression of a given

transcript. To map a read to a specific transcript, one can either align reads first to the genome and then assign them to the isoform they most likely belong to. Alternatively, in a strategy termed "pseudo-alignment", reads are aligned directly to the transcripts, without the need for a genome assembly. Both these strategies suffer from the same limitation: there is a sizeable number of reads that cannot be unambiguously assigned to a given transcript. This problem is exacerbated when transcriptomes are complex such in the case of Homo sapiens. Bioinformatics tools adapted to this problem have been developed that employ probabilistic models to smooth out the bias coherently. One example for this are expectation maximization algorithms to assign fractional count to all the transcripts they align to. The size of the fractional count assigned to a given transcript is calculated using only the reads that can be unambiguously assigned to the sequence. These methods improve the reliability of the read counting, but they are still noisy which disproportionately affects lowly expressed transcripts that don't have a big enough sample size to employ expectation maximization algorithms reliably. Synthetic and real-world benchmarks suggest that is very hard to achieve a reliable analysis of such lowly expressed transcripts (such as lncRNAs) using current methods, even when including more than 30 replicates [Zheng et al. 2018]. This problem can't be fully overcome as it is inherent to the short read sequencing approach.

A new generation of aligners such as `Kallisto` or `Salmon` does not rely on a reference genome, instead reads are aligned directly to transcript sequences, which results in much faster computation times when compared to standard aligners with comparable precision [Bray et al. 2016; Patro et al. 2017]. Due to their fast alignments, most of these so-called "pseudo-aligners" employ bootstrapping, usually around 100 times, to estimate the alignment confidence. The bootstrapped replicates can be incorporated downstream to aid the expression quantification and precisely quantify technical variance [Pimentel et al. 2017]. The downside of pseudo-aligning reads is that the approach is highly dependent on the available annotation. Also, there is still the need for an upstream transcriptome assembly if the aim of a study is not only the quantification of expression but also transcript discovery.

3.1.4 Third-generation transcriptomics approaches

Third-generation sequencing approaches are poised to expand the boundaries of transcriptomics, albeit with some limitations at the current state of development. Abandoning the "shotgun" sequencing approach in favor of full-length transcript sequencing enables the resolution of complex long-range splicing events both at the assembly and the counting step. The current major roadblocks to the wide adoption of third-generation sequencers are their, compared to second-generation methods, lower precision and throughput.

The two leading third-generation sequencing technologies are PacBio and Oxford Nanopore. Both appoaches, which rely on different principles, have their respective advantages and shortcomings [Weirather et al. 2017]. PacBio sequencers have obtained higher precision at the cost of lower read lengths than Oxford Nanopore sequencing. In contrast, Oxford Nanopores offer a theoretically unrestricted read length and much higher throughput at but the technology has been plagued by low base calling accuracy (80-90%). In recent years both technologies have been significantly improved, and Oxford Nanopore sequencing has dramatically improved in precision. While still lagging behind PacBio, Nanopore sequencers can reach now up to 99% median precision. The higher output and long read length offers obvious advantages in differential expression studies [Byrne et al. 2017; Seki et al. 2019; Zhao et al. 2019]. Another advantage of Oxford Nanopores is the fact that the procedure does not require any amplification step, neither during the library preparation step nor the sequencing, that could introduce a bias into the sequencing the results.

3.2 Project aims

To date, several studies have employed second-generation sequencing to characterise the transcriptomes of flowers of different stages, specific types of floral organs, and even of different domains within the floral meristem [Pajoro et al. 2014; O'Maoileidigh et al. 2013; Ó'Maoiléidigh et al. 2018; Kaufmann et al. 2009; 2010; Mantegazza et al. 2014; Zhang et al. 2015; Ryan et al. 2015a]. The resulting datasets have led to new progress in the field and offered snapshots into the gene expression programmes underlying flower development. However, the limitations of current short-read sequencing technologies (as discussed above) have not allowed researchers to fully investigate

complex phenomena such as alternative splicing and lncRNA expression. To overcome these limitations, my work aimed at increasing the resolution of transcriptomics in flowers by using long-read sequencing technology. The Oxford Nanopore sequencer is the most affordable long-read sequencer on the market. It is not limited by the length of nucleic acids it can sequence which makes the sequencing of full-length transcripts finally possible. For this project, I aimed at exploring differential expression at the level of single transcript isoforms during the onset of flower formation. A major challenge when using Oxford Nanopore for direct cDNA sequencing is the amount of polyA-selected RNA needed as starting material. This is especially problematic when working with small numbers of undifferentiated cells as in the case of early flower development. Furthermore, the collection of tissue of specific stages from Arabidopsis is difficult given the small size of the plants. To overcome these limitations, I sought to employ a floral induction system, which is widely used in flower development research and allows the collection of large numbers of developmentally synchronized floral buds for analysis. Using this line, I aimed at characterizing floral transcriptomes at very early stages of development and to identify stage-specific transcriptome differences, as well as differences in alternative splicing and lncRNA expression.

3.3 Results

3.3.1 Experimental design

As outlined above, the collection of sufficient tissue from early floral buds of Arabidopsis for analysis is challenging as these flowers are minute. What's more, every bud in an inflorescence is at a different developmental stage as floral primordia are initiated sequentially, one at a time. To circumvent this limitation and obtain sufficient RNA for Nanopore sequencing, I employed a floral induction system (FIS) which allows the collections of hundreds of synchronized floral buds from a single plant. This system is based on an *ap1 cal* double mutant. As described in Chapter 1.1, flower formation in these plants is blocked and instead they accumulate inflorescence-like meristem at their apex. When a fusion protein between AP1 and the hormone-binding domain of the rat glucocorticoid receptor (denoted, AP1-GR) is expressed in this background, it is normally retained in the cytoplasm by chaperone proteins such as hsp70. However, when plants are treated with the steroid hormone dexamethasone (DEX), the fusion protein is released from its cytoplasmic retention, enters the nucleus and activates flower formation. Each apex will then contain a pool of hundreds of synchronized floral buds, making it easier to precisely sample large amounts of tissue of specific stages. This method overcomes both the issue with the starting amount of tissue and developmental timing issue, allowing for greater precision and reproducibility.

A schematic representation of the experimental design I employed is outlined in Figure 3.1. FIS plants were grown until they had bolted. Meristematic tissue was then collected from a subset of randomly selected plants to serve as an untreated control sample. The remaining plants were treated with DEX to induce flower formation. Four and 8 days after induction (DAI), floral buds that were forming on the FIS plants were then collected. At 4 DAI, floral buds are approximately at stage 5-6 of development, while at 8 DAI, they have reached approximately stage 9 and thus a more intermediate stage of development [Ryan et al. 2015a]. Thus, the samples collected (i.e., untreated, 4 DAI, and 8 DAI) covered many important processes during early flower development, including the specification and termination of floral meristems, the formation of organ primordia, and the specification and development of floral organs.

The experiment was repeated several times, resulting in total in four

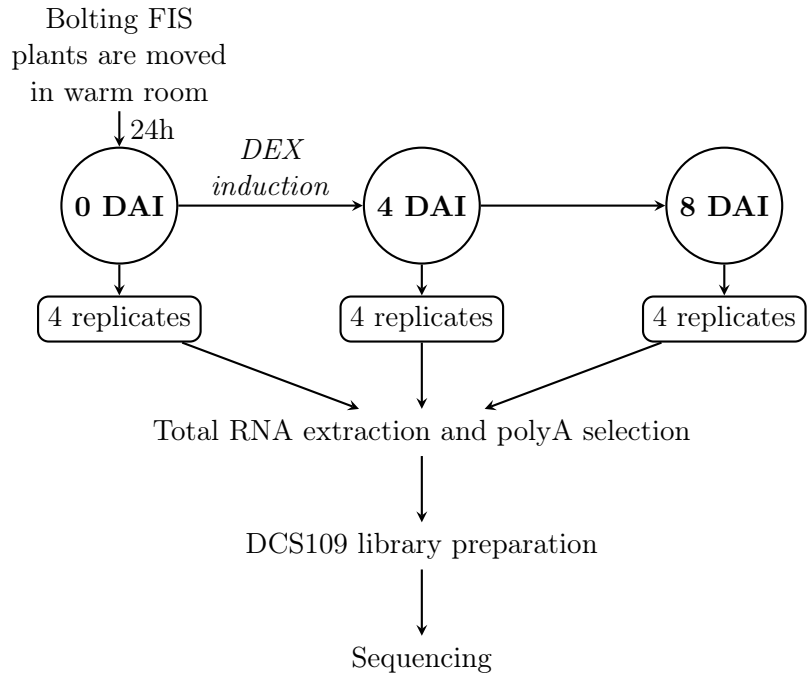


Figure 3.1: Experimental design for the Nanopore transcriptomics experiment. Plants were grown in colder conditions to slow the differentiation of the meristematic tissue in FIS mutants. After bolting, they were moved into a warmer room, with standard *Arabidopsis* growth conditions (See 2.3). DAI: Days After Induction; FIS: Floral Induction System.

biologically independent sets of samples. Total RNA was extracted from the frozen tissue samples with a phenol-chloroform based method adapted from Toni et al. [2018]. Special care was taken to minimise pipetting and vortexing in an effort to avoid mechanical shearing of RNA and cDNA, throughout the entire experiment, from the RNA extraction to the library preparation and sample loading.

The DCS109 direct-cDNA sequencing protocol performs two reverse transcription steps using two primers, one complementary to the 5' cap and one to the polyA tail. The primers add specific sequences to the reads, which bioinformatics tools like `pychopper2` (downloadable from the Nanopore GitHub page) can use to distinguish full length transcripts from degraded RNA.

Nanopore offers the possibility to sequence RNA molecules directly, with custom library prep and flow cells. It avoids the biases that could be introduced by reverse transcription, but at the cost of base calling precision and

most of the throughput. For this project, the use of cDNA libraries was preferred over the RNA direct sequencing, due to the superior throughput and quality. The libraries were prepared and loaded following the manufacturer's specifications and loaded onto a MinION sequencer (see section 2.25).

3.3.2 Bioinformatics pipeline

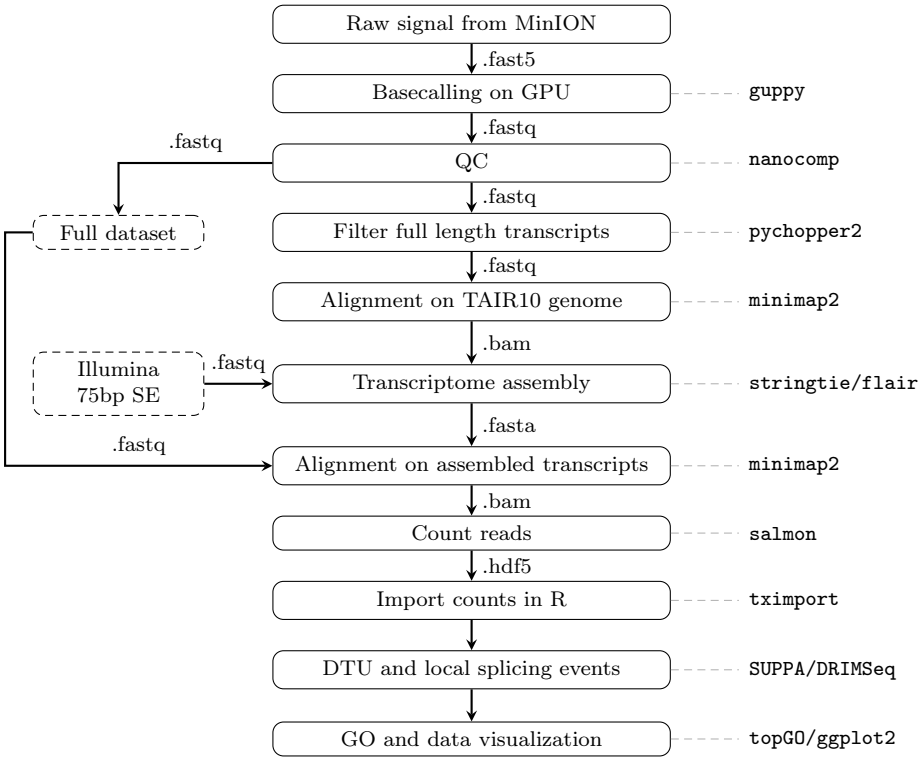


Figure 3.2: Flow chart describing the main steps of the bioinformatics pipeline used to process the raw data from the MinION sequencer and generate final datasets and plots. Individual steps are outlined in the solid boxes. The tools used to perform each step are listed on the right and connected to their corresponding step via dashed, grey lines. The intermediate file types are listed next to the arrows, which connect the steps as input→output. The dashed boxes contain auxiliary datasets such as the Illumina 75bp SE dataset and the full set of Nanopore reads.

During a sequencing run, the MinION sequencer is able to record the current in each pore with high precision and time resolution. Individual bases are called by a machine learning algorithm that segments the signal and infers which base is passing through the pore. For the base calling I

used Guppy, Oxford Nanopore's official tool distribution. Base calling with Guppy in high confidence mode is computationally demanding, which is why it was run separately on a high-end graphics card, allowing to speed up the process up ∼50-fold to reduce the turnaround time to a more manageable ∼10 hours per sample. The fastq files were then processed using `pychopper2` to extract the subset of reads spanning full length transcripts.

Sequencing throughput and quality

For this experiment, the aim was to obtain more than 7 million reads per sample. Coverage calculation are complicated by the read length distribution, a long-tailed binomial from 200 bp to the length of the longer expressed mRNA and a mode of ∼1000 bp. In a conservative estimation, 7 million reads correspond roughly to the same coverage as ∼23 million 150 bp paired-end reads from a second-generation sequencer, which is enough coverage for a transcriptomics experiment in Arabidopsis (Illumina.com). For all samples at 0 and 4 DAI, the desired throughput was achieved with a single run for each sample. The 8 DAI samples were run on slightly older cells that yielded half of the reads, which required pooling the results of 2 or 3 runs to reach past the 7 million target. The relevant run metrics are summarised in Table 3.1.

Each read has an associated quality score, assigned by the base calling software. Following the manufacturer guidelines, reads with a quality score < 7 were discarded, leaving a percentage of "passed" reads that ranged from 74% to 89%. The next step (Figure 3.2) was running the reads through `pychopper2`, a tool developed by Oxford Nanopore that can identify reads covering full-length transcripts by looking for specific sequences that are reverse transcribed at both the 5' and 3' ends during the DCS109 library preparation. Interestingly, the subset of full-length transcript was as low as 35% with only a couple of samples reaching values closer to 50%, meaning that most of the reads come from truncated transcripts (see Table 3.1). The sample 0days_rep2 is an evident outlier, with much better scores across the observed metrics; on the opposite end of the scale, 4days_rep3 has the worst sequencing run, with a low median length which translates in low full-length transcripts recovery at 32.8%.

Table 3.1: Overview of the sequencing runs

Run	Total Reads	Q>7	Pass %	Median Length	Median quality	Full Lenght%
0days_rep1	10668654	8862368	83.1	870	11	35.3
0days_rep2	9958093	8897154	89.3	1333	12.4	49.6
0days_rep3	7120802	6108899	85.8	986	11.4	46.8
0days_rep4	11652907	10175939	87.3	956	10.9	34.5
4days_rep1	9182996	8092560	88.1	953	11.8	39.7
4days_rep2	13243090	11777489	88.9	936	11.8	39.6
4days_rep3	11356218	9655360	85	809	10.8	32.8
4days_rep4	10401030	9150587	88	989	11.2	41.2
8days_rep1	11372412	9160942	80.6	1000	10.8	43.4
8days_rep2	13314917	10159073	76.3	1022	10.6	41
8days_rep3	13440699	10158892	75.6	1048	10.3	36
8days_rep4	11409665	8515959	74.6	1075	10.1	38.9

Note:
Values are colored on a continuous scale from the lowest in the column (lime green) to the highest (dark purple). One run was performed for each sample. In the first column each sample is identified by a code in the format [x]days_rep[y], were x is the DAI timepoint and y is the replica number.

Genome alignment and visual inspection

A further inspection of the results was done by aligning the reads to the Arabidopsis reference genome (TAIR10), both with the whole dataset and the subset of full-length reads isolated by `pychopper2`, and visually inspecting the pileups on a genome browser. An example of the kind of typical Nanopore read pileups is shown in Figure 3.3. As Nanopore sequencing is a new and complex technology, manual inspection can reveal biases difficult to pick up with broader metrics.

Both of the cDNA strands are equally likely to pass through the pore. In the full dataset, the reads are unstranded, which reduces mapping precision. In constrast, the subset of full-length reads is stranded. When `pychopper2` recognises both the 5′ cap and poly-A tail sequences, it reverse complements the reads, if needed, to have the 5′ cap at the 5′ end and the poly-A tail at the 3′ end. Figure 3.3a shows one major problem with long-read sequencing: degradation. When sorting alignments by position, it is obvious that there is a bias in coverage, due to the reads getting progressively more degraded

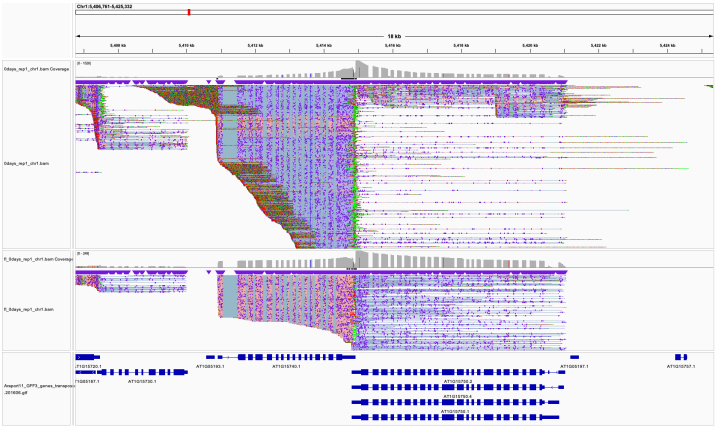
towards the 5′ cap. This likely to be due to degradation present in the starting RNA and not something introduced by the library preparation, but it is difficult to test without a true set of perfect RNA molecules. The problem spills into the set of full-length reads dataset, but to a lower extent. Due to the high error rate of Nanopore reads and the relatively low complexity of the specific DCS109 primers, particularly on the poly-A tail, `pychopper2` is likely classifying as full-length some of the degraded transcripts.

In general, the shorter a read, the harder it is to align to the correct position in the genome. There was a large number of reads mapped to wrong positions, even after full-length read selection (Figure 3.3b). The issue is compounded by the lower sequencing accuracy. Nanopore reads have an associated per base Quality (Q) score, which is assigned by the machine learning algorithm that infers the DNA sequence from the current variations in the pore. The Q-score is not always informative of real world performance; a complementary approach is to look at the identity score of the aligned reads on the reference genome. The reads had a median alignment score of ∼95%, which is low (relative to short-read sequencing platforms) but in line with the manufacturer's specifications.

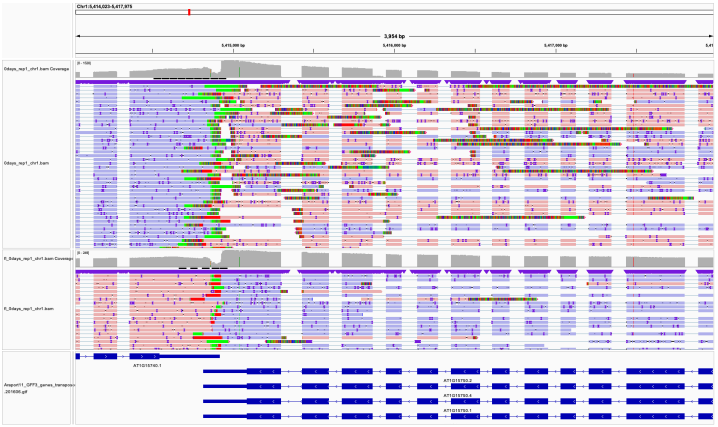
Hybrid transcriptome assembly

Most of the Nanopore reads that don't cover the full length of a transcript can't be easily be dealt with by current transcriptome assemblers. First, I aligned the full-length reads obtained from `pychopper2` to the TAIR10 Arabidopsis genome assembly using `minimap2` in splice mode with the `-x splice` flag, and then used this first alignment in a reference guided transcriptome assembly using `stringtie`. `Stringtie` has recently been updated to add a `-mix` flag which allows to perform a reference-guided hybrid assembly, which uses long and short reads together. Hybrid assembly can improve precision by mitigating the higher error rate of third generation sequencing reads while maintaining the spatial resolution provided by longer reads. Chen et al. [2018] have published a dataset of single-end 75 bp reads from an mRNA sequencing, which uses the same FIS lines and tissue collection timepoints as in this work. These high accuracy short-reads were aligned to the same Arabidopsis genome sequence and used to `stringtie` in combination with the noisy long reads for hybrid assembly.

In parallel, I used another tool, `flair` [Tang et al. 2020], which does end-



(a) Degradation gradient in $3' \rightarrow 5'$ direction.



(b) Alignment artifacts due to degradation products.

Figure 3.3: Spliced alignment of Nanopore reads on the Arabidopsis reference genome. Panel 3.3a offers a wide view on the progressive degradation of the reads sequenced from the DCS109 cDNA sequencing kit while in panel 3.3b the view is zoomed in, to highlight the presence of probable misaligned reads due to degraded fragments. Reads aligned in the forward direction are colored in pink, while reverse reads are colored in purple. Soft clipped bases are shown in different colors (A: green, T: red, G: yellow, C: blue). Mismatched bases are in violet, insertions with violet triangles below the coverage plot, deletions with black horizontal connectors while grey connectors denote a spliced intron. Each panel shows two genome tracks, the top track has the whole dataset read pileup while the bottom one has only the subset of full length reads filtered with `pychopper2`. Each genome track has a coverage plot on the top shown in grey. At the bottom of each panel are shown the transcripts from the Araport11 transcriptome annotation.

to-end analysis of Nanopore sequencing transcriptomics data. `flair` uses `minimap2` to align Nanopore reads to the genome and then implements its own algorithm for transcriptome assembly. The assembly in `flair` is not performed by building a graph of overlapping reads, but by trying to line up the splicing sites opened in the read pile-ups on the reference genome. After correcting the splice sites, it retains only isoforms that have enough supporting reads. The heuristic looks at parameters like the percentage coverage and the number of reads that spans all the isoforms' splice sites. `flair` can also indirectly make use of accurate short-reads: a "junctions" (splice sites) BED file can be supplied to the read correction step, to help with the identification of bona fide splice sites. I used the same Chen et al. [2018] short-reads dataset I used for `stringtie` hybrid assembly and I extracted the junctions using `junctions_from_sam.py`, a helper script that comes with the `flair` installation.

Comparing `stringtie` and `flair` assemblies

When compared against the Araport11 reference annotation of the Arabidopsis transcriptome, which contains 58200 transcripts in 36545 loci, the two approaches produced very different transcriptome assemblies. The transcriptome assembled by `stringtie` predicted 11,025 novel isoforms, considerably fewer than `flair`'s 86,900. Similarly, `stringtie` called 118 novel genes versus `flair`'s 9,051 (see Table 3.2). The numbers in the `flair` assembly seem to be inflated by poor precision and sensitivity (recall). While `stringtie` assembly incorporates almost all the transcripts present in the annotation, `flair` is only able to recall 24,104, or 41.5% of the reference transcripts.

Table 3.2: Comparison of the transcriptome assemblies at the feature level

Tool	Level	Known	Novel
`stringtie`	gene	36186	118
	transcript	57473	11025
`flair`	gene	17696	9051
	transcript	24104	86900

The difference between the tools is exacerbated when further dissecting the assemblies (see Table 3.3). To have a more granular view of the differences between the assemblies, I compared them to the Araport11 reference

annotation using `gffcompare` [Pertea and Pertea 2020]. `Gffcompare` calculates the sensitivity and the precision at the exon level, instead of just comparing the trancripts; ends.

`Stringtie` tends to be more conservative when calling new transcripts and results in higher sensitivity in all the categories evaluated by `gffcompare`, while still maintaining a good accuracy in placing exon boundaries, with a precision of >90%. The lower values of intron chain and transcript precision, around 80% are in line with the amount of new transcripts discovered: `gffcompare` considers true positives as features that are replicated in the reference and false positives as those that are not found, which includes novel transcripts. `Stringtie` is able to maintain the correct exon boundaries and splicing sites for known isoforms while still finding new putative splice isoforms.

`Flair`'s approach is more independent of the provided annotation and this shows in the output. Even when provided with the annotation and a list of junctions in bed format, the sensitivity is lower across the board. A considerable amount of the splice sites are dropped, with exon and intron level precision at 64% and 80% respectively. `Gffcompare` identifies 83% of the introns as already annotated versus only 61% of the exons. Due to the high amount of novel transcripts found by `flair`, the very low precision values in the intron chain and transcript category is expected.

Table 3.3: Comparison of the transcriptome assemblies at the feature level

	`Araport11 vs stringtie2`		`Araport11 vs flair`	
Level	Sensitivity	Precision	Sensitivity	Precision
Base	100	96.6	59.3	77.4
Exon	96.8	91.7	63.8	60.8
Intron	100	94.8	79.6	83.1
Intron chain	100	79.6	50.2	22.6
Transcript	99.1	82.7	41.5	21.9
Locus	99	98.7	48.3	73.1

Note:
Values are colored on a continuous scale from the lowest in the column (lime green) to the highest (dark purple).

In summary, the two tools produce very different assemblies, stemming

from their different approaches: one more conservative and the other more "greedy", resulting in `flair` calling five times more novel transcripts than `stringtie`. Because the two assemblies were generated based on two very different assumptions they could potentially complement each other's shortcomings. I therefore performed the downstream steps of the analysis with both assemblies.

To this end, I translated the coordinate-based description of transcripts in the assemblies (GTF format) to FASTA sequences, which I extracted from the TAIR10 reference genome. I then aligned the whole dataset of `Nanopore` reads to the transcripts in FASTA format using `minimap2` with the `-x map-ont` option, which is the non-spliced alignment mode for Nanopore reads. Subsequently, I used `salmon` to quantify the number of reads belonging to each transcript (for a description of the full pipeline, see Figure 3.2).

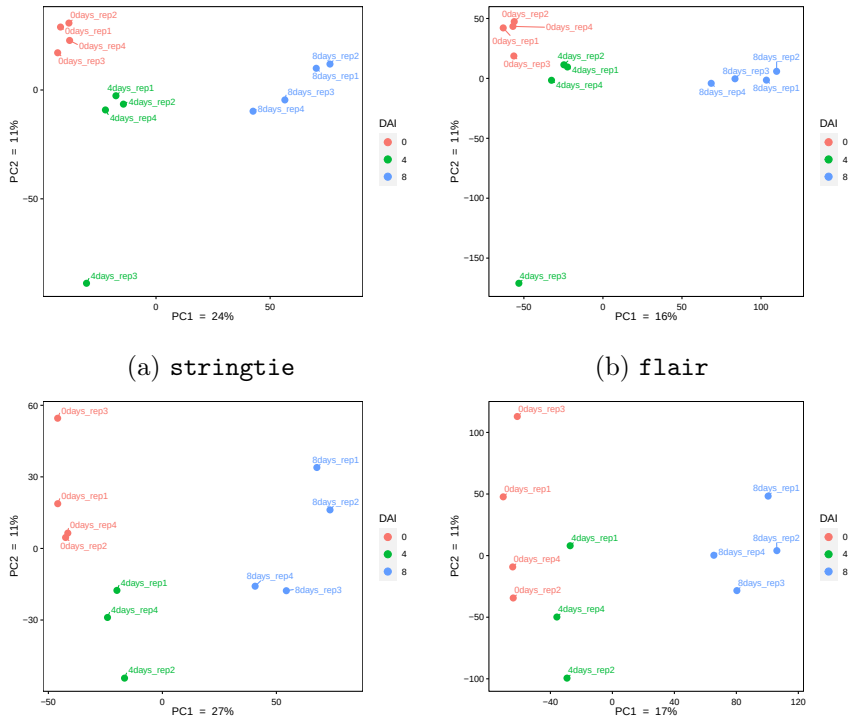
Principal Component Analysis (PCA) of datasets

I imported the `salmon` read counts in `R` with the `tximport` package and then explored the data by plotting the first two principal components (see Figure 3.4). Ideally, the variance between replicates should be lower than the variance between conditions. In a PCA plot, a lower variance between samples visually translates into a tighter clustering of the corresponding data points. Deviations from this pattern can help spot outliers or batch effects.

PCA plots from `stringtie` and `flair` are not significantly different: in both cases the sample 4days_rep3 looked like a clear outlier (see Figure 3.4a and 3.4b). It wasn't immediately obvious why that particular sample had such a different composition with respect to the other replicates. However, looking at the run statistics shown in Table 3.1, it is also the sample with the shortest median read length and the lowest percentage of full length transcripts retrieved by `pychopper2`. This could indicate either a lower RNA integrity or a less efficient translocation of the DNA molecules through the pores during the sequencing run. Poor translocation rate or clogging can lead to premature ejection of DNA from the pores, resulting in truncated transcripts.

Due to the suspected unreliability of this sample and the presence of at least another three replicates for the 4 DAI time-point, I decided to disregard the sample before calculating the differential transcript expression, to avoid skewing the analysis. I generated an updated PCA plot to assess the effect

of eliminating 4days_rep3 and observed a good clustering of replicates (see Figure 3.4c and 3.4d).



(a) `stringtie` (b) `flair`

(c) `stringtie` drop 4days_rep3 (d) `flair` drop 4days_rep3

Figure 3.4: PCA plots on log transformed per transcript read counts. Principal component 1 (PC1) is shown on the x-axis for each panel and the PC2 on the y-axis. The percentage of variance explained by each PC is shown after the equal sign. The top panels 3.4a and 3.4b show the PCA on counts computed on `stringtie` and `flair` transcriptomes respectively. The bottom panels 3.4c and 3.4d how the updated PCA plot after dropping sample 4days_rep3. The read counts used were expressed in transcripts per million, to normalise the read depth across samples so that they would be comparable.

Differential Transcript Usage analysis

To identify significant differences in the expression of isoforms across the time-points taken, I employed the approach of differential transcript usage (DTU). Separately, I also analysed the differential usage of single splicing sites within genes. Considering single splicing sites can help improve sensi-

tivity but narrows the focus of the analysis to local events, missing out on more complex splicing events, where multiple splicing sites are involved. I carried out the DTU analysis in `R` using the package `DRIMSeq`, while I assessed the local splicing site usage using `suppa`, a command line tool written in `Python` [Robinson and Nowicka 2016; Trincado et al. 2018; Alamancos et al. 2015].

Three pairwise comparisons were performed between 0 and 4 DAI, 4 and 8 DAI, 0 and 8 DAI time-points. I summarised the number of significant events for each comparison in Table 3.4. The `DRIMSeq` pipeline for DTU discovery was applied to counts coming from both `flair` and `stringtie` while differential splice site usage is reported only for `stringtie` counts. I tried to use `suppa` for `flair` counts, but the results were too difficult to interpret due to the very large number of transcripts reported by `flair`. The lack of splicing site information coming directly from the Araport11 annotation hampered the analysis and the interpretation of results. Overall, the comparison between 0 and 8 DAI yielded the highest number of significant genes, probably due to the longer interval between the time-points.

Table 3.4: Summary of the total number of significant DTU and local splicing events found in the dataset

		Pairwise comparisons		
Assembly	DTU/Local	DAI0vsDAI4	DAI4vsDAI8	DAI0vsDAI8
`flair`	`DRIMSeq`	15	48	72
`stringtie2`	`DRIMSeq`	20	104	169
	`suppa A3`	58	132	95
	`suppa A5`	48	99	85
	`suppa AF`	12	20	13
	`suppa AL`	2	6	3
	`suppa MX`	2	2	1
	`suppa RI`	91	157	125
	`suppa SE`	24	34	37

Note:
Threshold for DTU events is q-value < 0.05 and for local splicing events is p-value < 0.05. A3: Alternative 3' end; A5: Alternative 5' end; AF: Alternative First exon; AL: Alternative Last exon; MX: Mutually Exclusive exons; RI: Retained Intron; SE: Skipped Exon.

Before interpreting the results, it is important to understand how `DRIMSeq`

and `suppa` work. `Suppa` extracts all the possible splicing events from a given annotation and then bins the isoform counts in the two possible event outcomes. The output is a list of all the splicing events that show a significant structural change between conditions. `DRIMSeq` takes a different approach by fitting a Dirichlet-multinomial model on the isoform expression proportions relative to each gene. It then computes a q-value for each gene, which quantify the posterior expectation of finding a set of proportions in condition B given the observed proportions in condition A. In practical terms, this means that unlike `suppa`, `DRIMSeq` will not output information about the splicing events, but only if a given gene presents a significant change in isoform proportions. `DRIMSeq` can identify complex transcript rearrangements, but won't report which isoforms are differentially used or which splicing event is responsible of the difference.

Gene Ontology enrichment

To better interpret the results, I computed the Gene Ontology (GO) enrichment for `stringtie` results in `R` using the package `topGO`. I chose two statistical tests to find significant enrichments: the classical one-side Fisher exact test, or hypergeometric test, and the Kolgorov-Smirnov (KS) test. The two tests have both advantages and disadvantages. The hypergeometric test compares the number of significant genes that belong to a GO category to the number of genes that would be expected to be found by chance. The KS test compares the observed distribution of p-values with a distribution of p-values that could arise from random data and finds the GO terms that deviate significantly from a random distribution. It follows that the Fisher test is very dependent on an arbitrary threshold of significance that splits the dataset in significant and non-significant data points, while the KS test doesn't have to make this assumption. The hypergeometric test is still by far the most widely used in the literature because of its interpretability and ease of use. I used gene sets from each of the pairwise comparisons to calculate GO enrichments. The resulting set of p-values was then binned by k-means clustering (see Figure 3.5 and 3.6).

The 0 vs 8 DAI comparison yielded the highest number of enriched GO terms, as expected, due to the greater amount of significant genes (see Table 3.4). GO terms associated with the response to various abiotic stresses, splicing regulation and aging are prevalent in the 0 vs 8 DAI comparison. No-

tably, terms related to flower development and hormonal control were found in the 0 vs 4 and 4 vs 8 DAI comparisons. Significant categories for the Fisher test are *leaf development*, *response to cold* (which contains genes like *FLC*) and *regulation of polar auxin transport* in the 0 vs 4 comparison, while *flower development*, *photosynthesis* and *anther wall tapetum development* at later stages, in 4 and 8 DAI comparison.

Due to the large number of GO terms retrieved by the KS test, I increased the stringency of the p-value cutoff from 0.05 to 0.01. As with the Fisher test, most of the significant terms were found in the 0 vs 8 DAI comparison. Most of the terms are related to *abiotic stress response*, *regulation of splicing*, *metabolic process* and *pattern specification*. A few GO terms were significant in all comparisons, notably *leaf development* and *response to cold*. The rest of the categories, which account for almost half of the KS results, were found in the 4 to 8 DAI comparison. At later development stages, splicing seems to affect development related genes, which belong to GO categories like *root development*, *regulation of developmental process*, *pollen development*, *fruit development*, *flower development*, *floral organ development* and *cell division*.

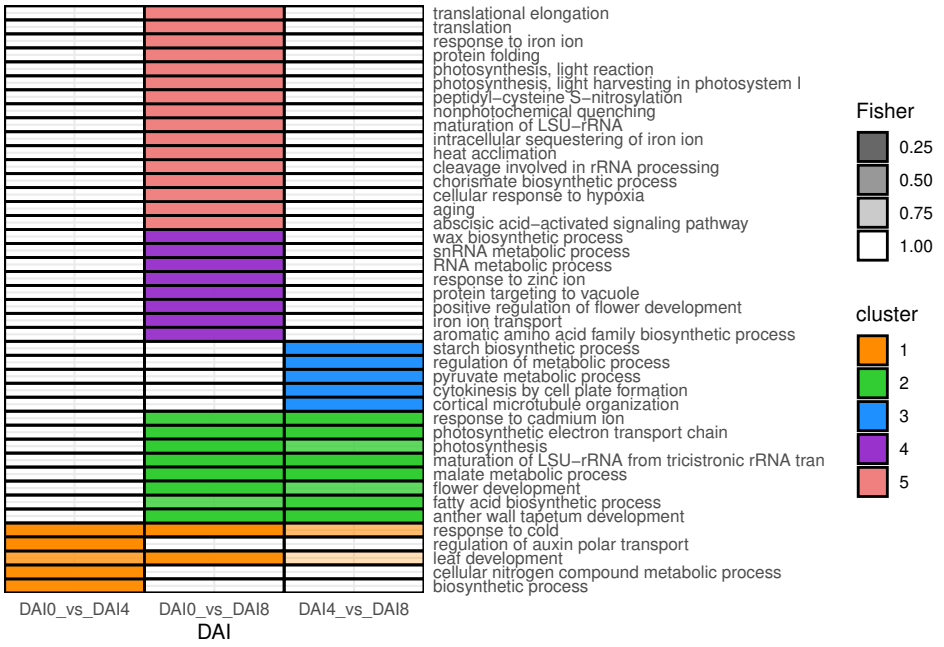


Figure 3.5: GO terms with a significant enrichment in DTU genes identified by `topGO` with Fisher's exact test. The cells are colored in base of which cluster they belong to and shaded in base of the p-value they are representing. A lower p-value translates in a more intense color. Each row of cells represents one GO term, the description which is reported on the right side of the plot. Each column of cells refers to one of the three pairwise alignments performed in `DRIMSeq`.

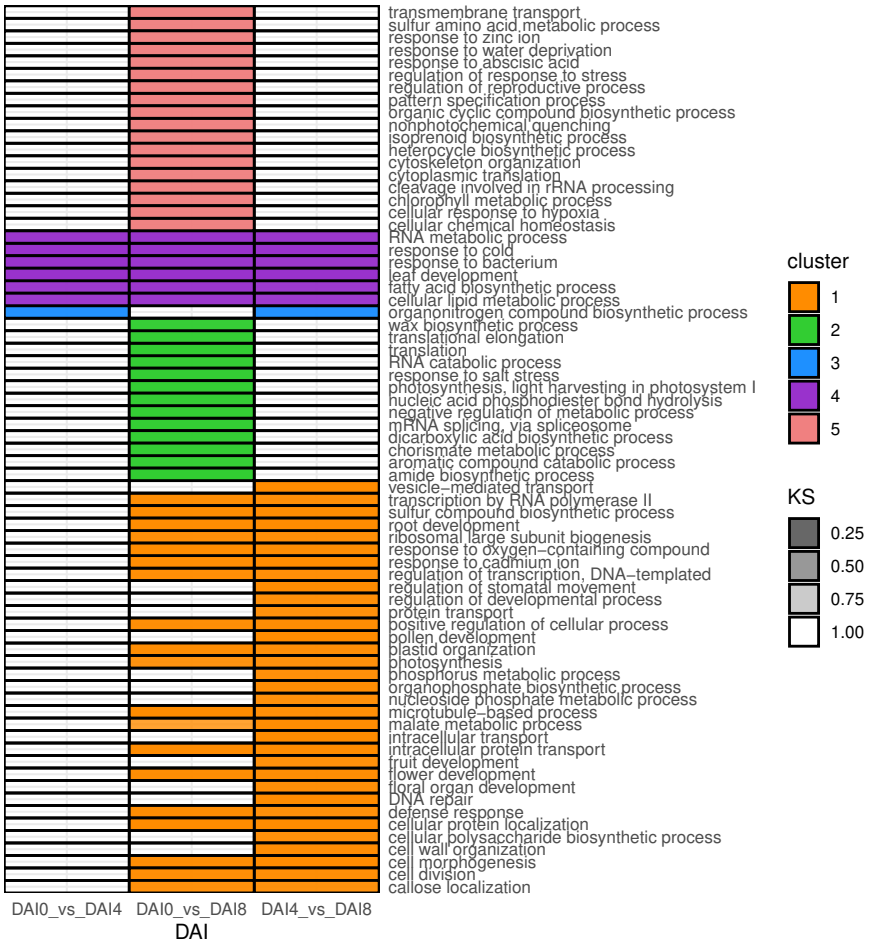


Figure 3.6: GO terms with a significant enrichment in DTU genes identified by `topGO` with the Kolgorov-Smirnov (KS) test. The cells are colored in base of which cluster they belong to and shaded in base of the p-value they are representing. A lower p-value translates in a more intense color. Each row of cells represents one GO term, the description which is reported on the right side of the plot. Each column of cells refers to one of the three pairwise alignments performed in `DRIMSeq`.

3.3.3 Candidate genes for further analysis

Long-read transcriptome sequencing is a very new technology, and assessing the reliability of the results is complicated by the lack of "truth sets" of known biological transcripts. I therefore chose a subset of genes involved in flower development to test the results of the analyses performed with `suppa` and `DRIMSeq`, respectively. To this end, I focused on a total of 10 splicing events across 8 genes and using qPCR, tested both the occurrence of an isoform switch at a specific location and the accuracy of the expression quantification calculated by the transcriptomics pipeline (see Table 3.5).

DTU event visualisation

Visualisation is key when screening gene loci for DTU because there are three important angles that have to be considered at the same time: the normalised counts for each isoform, the proportion of isoform counts given the total counts for the gene and the structure of each isoform in the assembly. I developed a custom `R` script to plot this information in a coherent manner aligning classic `ggplot2` boxplots with a modified version of the isoform plots in the `ggtranscript` package. Plots for each of the gene candidates are shown in Figure 3.7. The plots allowed me to dissect the alternative splicing events for each gene, which has to be done manually at least for `DRIMSeq` results.

Among the 8 genes I chose for further analysis were both known and novel splice event (see Table 3.5). PISTILLATA (PI) is a master regulator of flower development (see Chapter 1). It has an intron retention in the third intron that adds an early stop codon, which in turn deletes part of a domain that has been shown to be important for the formation of higher order complexes [Yang and Jack 2004; Yang et al. 2003a]. *GFR1-Interacting Factor 1 (GIF1)* encodes a transcriptonal coactivator which promotes cell proliferation and lateral organ growth. *GIF1* has a retained second intron in a novel isoform, which severly disrupts the amino acid sequence. *CCR-LIKE (CCL)* is involved in the response to the circadian clock. The pipeline identified DTU evidence for two novel isoforms, which retain the first and second intron respectively. *PIN-FORMED 6 (PIN6)* codes for a well-studied auxin efflux carrier. It has a splice isoform that retains the last intron, which disrupts the amino acid sequence in the last transmembrane domains. *REGULATOR OF*

CULLINS 1(ROC1) encodes a component of an SCF ubiquitination complex which mediates auxin response. *ROC1* has a predicted DTU event involving exon skipping. Exon skipping events are rare in plants transcriptomes, where intron retention events are the most common form of splicing. The skipped exon cuts off part of the zinc finger domain, which is likely to affect the functionality of the protein. *YABBY3 (YAB3)* encodes a transcription factor involved in abaxial identity specification in leaves and fruits. The retention of the first intron in *YAB3* completely changes the amino acid sequence resulting in a truncated protein. *REPRODUCTIVE MERISTEM 1 (REM1)* is part of the large *REM* gene family. It is specifically expressed in reproductive meristems and was shown to be involved in male and female gametophyte development. *REM1* is the only gene among those selected for analysis that was found only using the `flair` assembly. *REM1* has a splice isoform with a retained second intron that has an increased expression at the 8 DAI time-point. *SEPALLATA3 (SEP3)* is a member of the MADS-box gene family, its central role flower development was already described in section 1.2. The DTU analysis predicted two novel retention events in the first and last intron of the gene. Splicing events changing *SEP3* amino acid sequence could play a role in regulating the formation of higher order complexes with other MADS-domain transcription factors, affecting flower development.

The genes described above were selected to test the different DTU testing frameworks used for this work. The qPCR experiments had two aims: firstly, I wanted to confirm that the pipeline can produce reliable calls for new transcript isoforms, and secondly, I wanted to test if it is able to properly quantify the expression at the isoform level. It is impossible to experimentally check the expression of only one single isoform with high confidence, because it is unknown how many and which isoforms are expressed by any given gene at a time, and also because complex isoform switches can involve multiple splicing sites. The qPCR experiments were therefore limited to quantifying isoform usage at one site splicing site at a time.

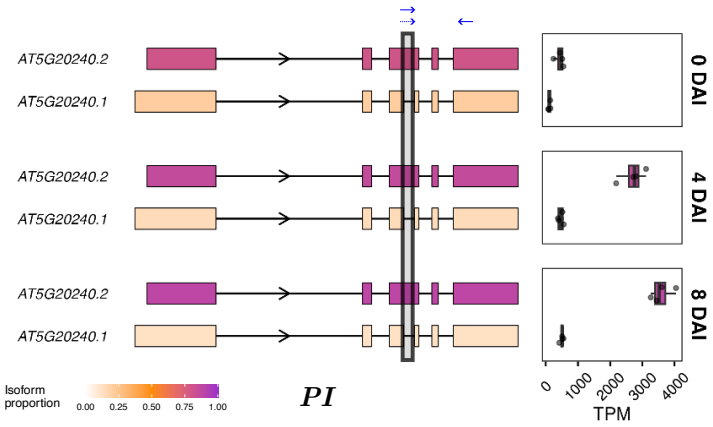
For the qPCR experiments, I designed for each slicing event to be tested two pairs of primers. One primer was used in both primer pairs and flanked the spliced region, while the second and third primer selectively bound either the reference isoform or the alternative isoform. For example, if the predicted splicing event is an intron retention, one primer will flank the re-

Table 3.5: Local splicing events analysed by qPCR

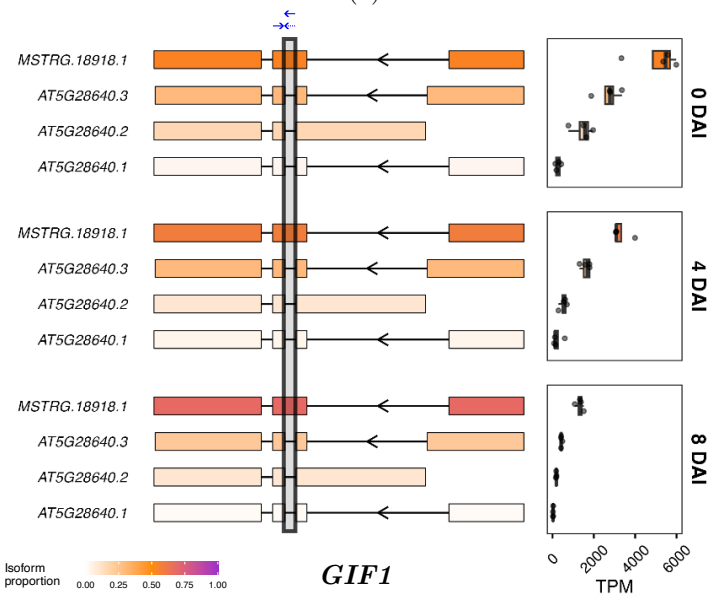
Gene Name	Reference Isoform	Found in	Description	Known Feature?
PI	*AT5G20240.1*	`drimseq`	Retain third intron	Yes
GIF1	*AT5G28640.1*	`drimseq+flair`	Retain second intron	No
SEP3	*AT1G24260.2*	`suppa`	Retain first intron	No
SEP3	*AT1G24260.2*	`suppa`	Retain last intron	No
PIN6	*AT1G77110.1*	`suppa`	Retain last intron	Yes
ROC1	*AT5G20570.2*	`drimseq+suppa`	Skip fourth exon	Yes
YAB3	*AT4G00180.1*	`drimseq+suppa`	Retain first intron	Yes
REM1	*AT4G31610.2*	`flair`	Retain second intron	No
CCL	*AT3G26740.1*	`drimseq`	Retain first intron	No
CCL	*AT3G26740.1*	`drimseq`	Retain last intron	No

Note:
DTU genes chosen from the output of `stringtie`, `SUPPA` and `flair` for further analysis are reported in the first column. The "Found in" column containas information about were the DTU gene was found and the "Description" column which type of splicing event is being amplified by qPCR.The last column defines a feature as "known" if it's present in the annotation. The "Reference Isoform" column shows which isoform is considered the "reference" according to the TAIR10 database.

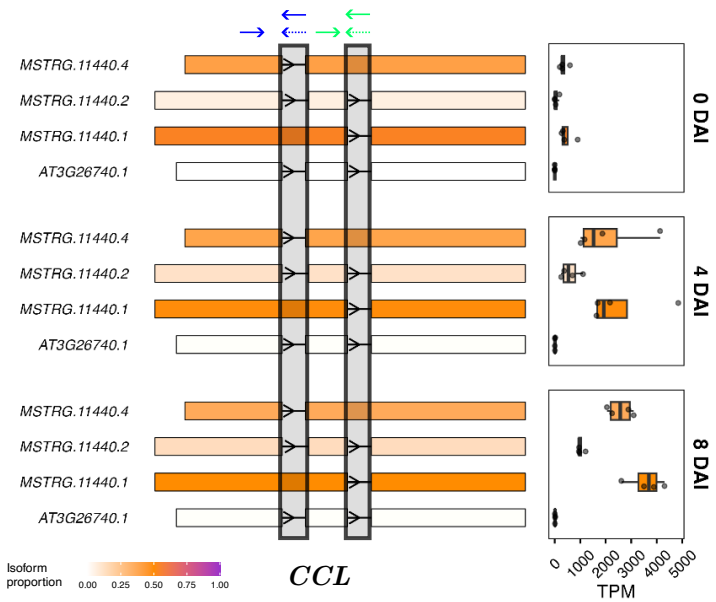
gion, possibly on one of the two exons in the immediate vicinity, while the other two will be designed inside the intron or directly on the splice site. A schematic representation of the primer pairs used for each gene can be found in Figure 3.7. Great care was taken to ensure that the primers pairs would selectively amplify only the intended targets and had good predicted efficiency (see section 2.24).


AT5G20240.2
AT5G20240.1
AT5G20240.2
AT5G20240.1
AT5G20240.2
AT5G20240.1
0 DAI
4 DAI
8 DAI
Isoform proportion
0.00 0.25 0.50 0.75 1.00
PI
0 1000 2000 3000 4000
TPM

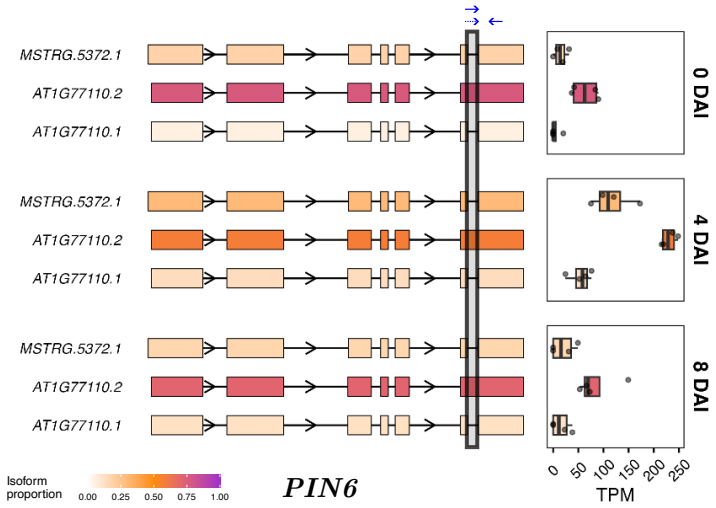
(a)


MSTRG.18918.1
AT5G28640.3
AT5G28640.2
AT5G28640.1
MSTRG.18918.1
AT5G28640.3
AT5G28640.2
AT5G28640.1
MSTRG.18918.1
AT5G28640.3
AT5G28640.2
AT5G28640.1
0 DAI
4 DAI
8 DAI
Isoform proportion
0.00 0.25 0.50 0.75 1.00
GIF1
0 2000 4000 6000
TPM

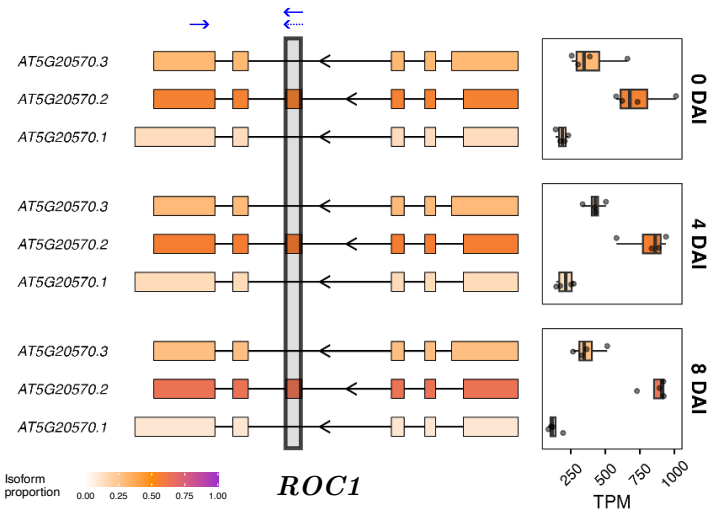
(b)


MSTRG.11440.4
MSTRG.11440.2
MSTRG.11440.1
AT3G26740.1
0 DAI
4 DAI
8 DAI
Isoform proportion
0.00 0.25 0.50 0.75 1.00
CCL
0 1000 2000 3000 4000 5000
TPM

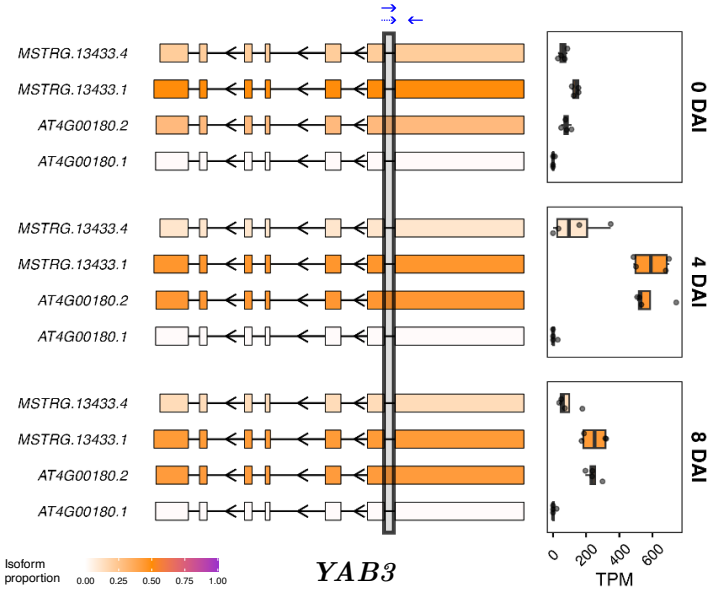
(c)


MSTRG.5372.1
AT1G77110.2
AT1G77110.1
0 DAI
4 DAI
8 DAI
Isoform proportion
0.00 0.25 0.50 0.75 1.00
PIN6
0 50 100 150 200 250
TPM

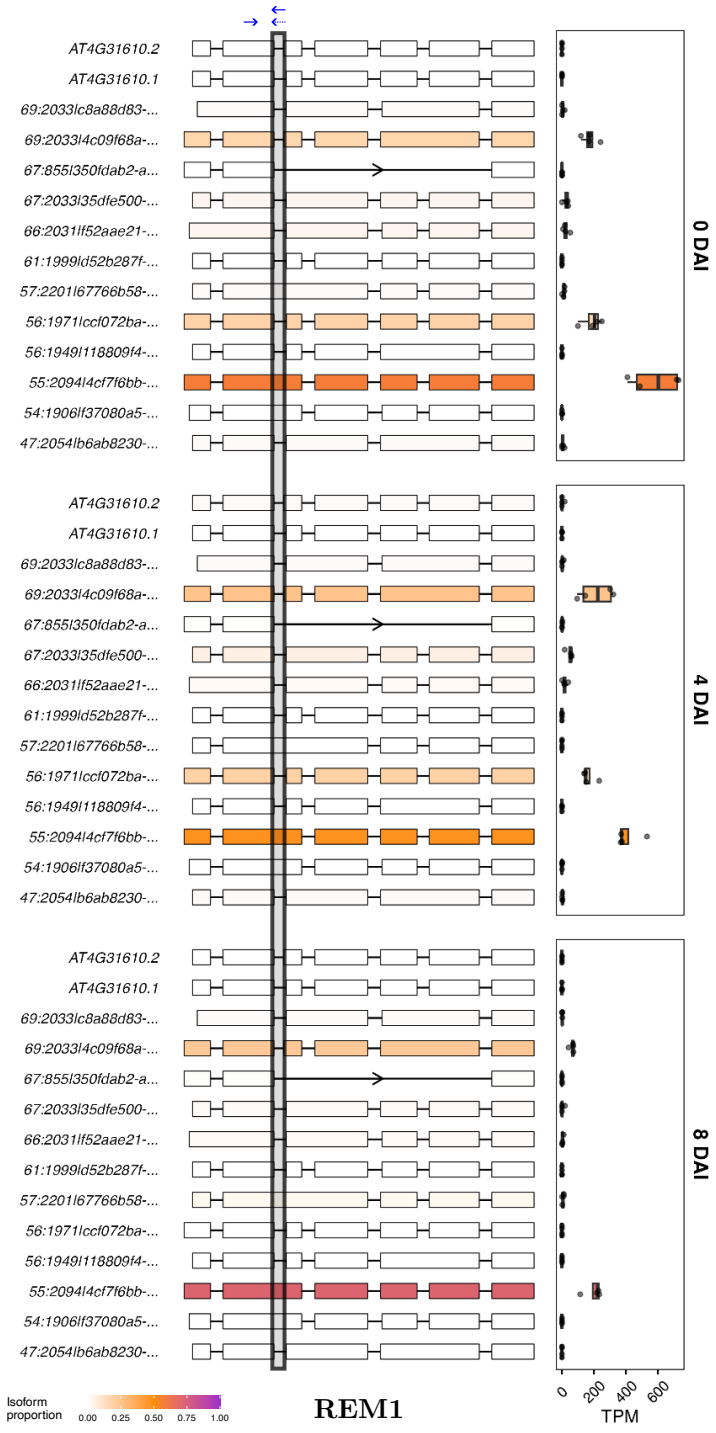
(d)


AT5G20570.3
AT5G20570.2
AT5G20570.1
AT5G20570.3
AT5G20570.2
AT5G20570.1
AT5G20570.3
AT5G20570.2
AT5G20570.1
0 DAI
4 DAI
8 DAI
Isoform proportion
0.00 0.25 0.50 0.75 1.00
ROC1
250 500 750 1000
TPM

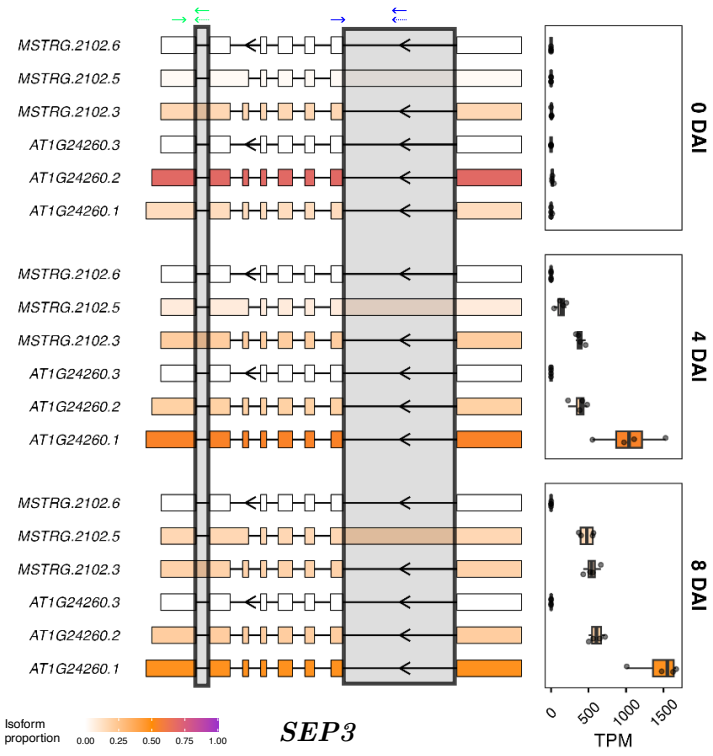
(e)


MSTRG.13433.4
MSTRG.13433.1
AT4G00180.2
AT4G00180.1
MSTRG.13433.4
MSTRG.13433.1
AT4G00180.2
AT4G00180.1
MSTRG.13433.4
MSTRG.13433.1
AT4G00180.2
AT4G00180.1
0 DAI
4 DAI
8 DAI
Isoform proportion
0.00 0.25 0.50 0.75 1.00
YAB3
0 200 400 600
TPM

(f)


AT4G31610.2
AT4G31610.1
69:2033|c8a88d83-...
69:2033|4c09f68a-...
67:855|350fdab2-a...
67:2033|35dfe500-...
66:2031|f52aae21-...
61:1999|d52b287f-...
57:2201|67766b58-...
56:1971|ccf072ba-...
56:1949|118809f4-...
55:2094|4cf7f6bb-...
54:1906|f37080a5-...
47:2054|b6ab8230-...
0 DAI
4 DAI
8 DAI
Isoform proportion
0.00 0.25 0.50 0.75 1.00
REM1
0 200 400 600
TPM

(g)



(h)

Figure 3.7: Representation of the DTU genes chosen for further analysis (see Table 3.5). The plots are composed by two main panels: a representation of the isoforms expressed from each gene is shown on the left, while the boxplots on the right shows the TPM counts for each isoform. Both the isoforms and the boxplots are coloured in base of the respective isoform proportion. The figures are divided in three rows, one for each DAI time-point, which is reported on the right side. Grey boxes delimit each of the local splicing events that were analysed by qPCR. The primers used in the qPCR amplification are shown on the top in blue or green if more than one splicing event was analysed for the same gene. The "flanking" primer is shown as a solid arrow, while the primer binding the splice site and in the spliced region are represented by a solid or dashed arrow respectively.

DTU event validation

The expression of each splice variant was normalised against the *REF1* reference gene [Czechowski et al. 2005] (see Figure 3.8). Surprisingly, the results of the qPCR experiments showed substantial differences to the predicted expression of isoforms across time-points. In contrast, the expression of the alternative isoforms could be detected in the majority of cases. *SEP3* showed the strongest signal for the alternative isoform for both the splicing events tested, followed by *PIN6*. For other genes, including *YAB3*, *CCL* and *PI*, weak expression of alternative isoforms could be detected. Thus, while predicted splicing events could be confirmed in many cases, the differential expression data from the long-read sequencing did generally not match the results from the qPCR experiments.

3.3.4 Investigating alternative splicing of *PISTILLATA*

To begin to understand the role of the alternative splicing events observed by Nanopore sequencing, I focused on the above-mentioned *PI* intron retention isoform because it directly related to the laboratory's work on flower development The reference isoform *PI.1* and the intron retention isoform *PI.2* differ in their C-terminus. *PI.2* has an early stop codon that is predicted to translate into a truncated protein of amino acid residues. PI.1 is 208 aa long and PI.2 is 150 aa. Notably, this truncation would eliminate domains that have been shown to be involved in the formation of higher order MADS-domain protein complexes [Yang and Jack 2004; Yang et al. 2003a]. Thus, expression of this isoform could have substantial effects on the regulation of flower development.

Transcripts with early stop codons can be recognised by specific proteins in the nucleus which could target it for and immediate degradation or blocking it from to exiting the nucleus [Hwang and Kim 2013; Nickless et al. 2017]. To test whether the *PI.2* isoform is translated, I designed an experiment in which a genomic *PI* sequence is fused in frame with either an N-terminal or a C-terminal epitope tag. The idea behind this design was that a C-terminal tag would be part of the fully spliced PI.1 protein, but wouldn't be translated in the case of PI.2 because translation is halted at the premature stop codon. In contrast, translation of the N-terminal tag would be unaffected by the premature stop codon. This difference would give PI a unique signa-

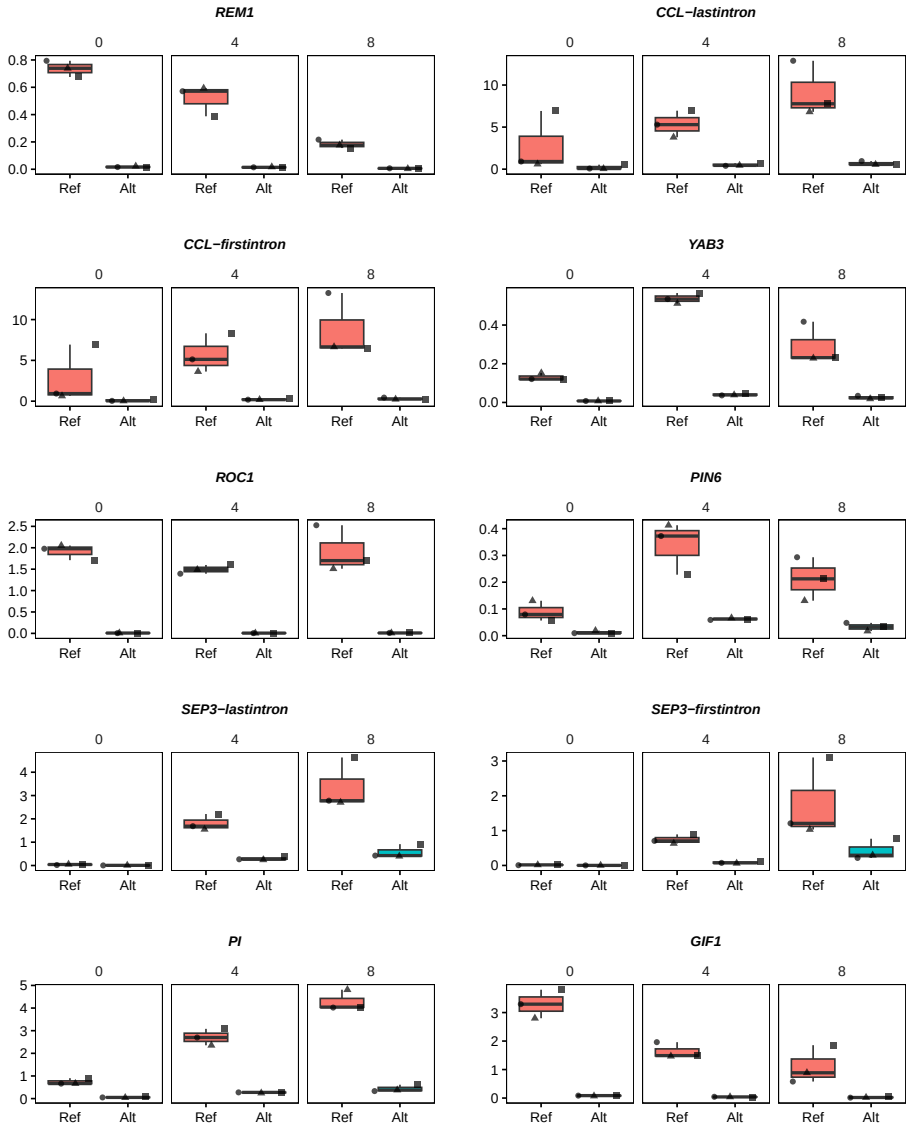


Figure 3.8: Validation of DTU gene candidates by qPCR. The y-axis shows the expression the reference (Ref) and alternative (Alt) splice isoforms relative to the *REF1* gene [Czechowski et al. 2005]. Each plot is split in three subplots, one for each DAI time-point (0, 4 and 8). The primers design is shown in Figure 3.7. The data points are shaped in base of the replica they represent: 1: △, 2 : □, 3 : ○.

ture that can be used for detection by Western blotting. As an epitope, I used the small peptide V5 to which I added a StrepII tag for future protein

purification experiments.

To test accumulation of the resulting $^{\text{StrepII}-\text{V5}}$PI and PI$^{\text{V5}-\text{StrepII}}$ fusion proteins, I expressed them transiently under control of a *UBQ10* promoter (see sections 2.10 and 2.6). To this end, the constructs were transformed into Agrobacterium cells and then infiltrated into tobacco leaves. After 72 hours, the leaves were collected and snap frozen. Total protein preparations were used for Western blotting using an αV5 antibody (see Figure 3.9). Wild-type leaves and leaves infiltrated with a Mock solution not containing Agrobacterium were used as negative controls.

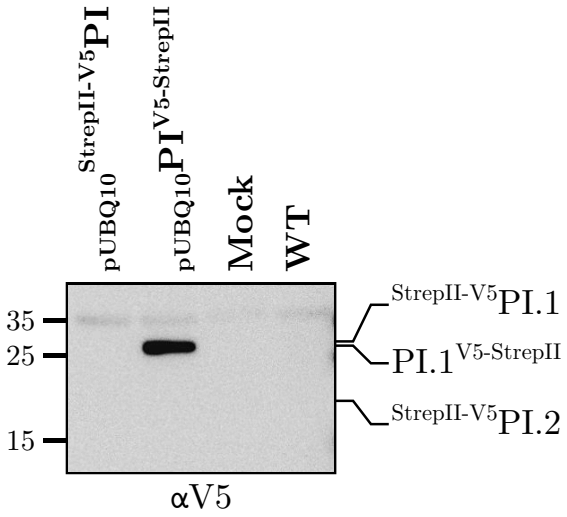


Figure 3.9: Results of immuno blots with total protein extracts from tobacco leaves. Leaves were infiltrated with $_{\text{pUBQ10}}{}^{\text{StrepII}-\text{V5}}$PI, $_{\text{pUBQ10}}$PI$^{\text{V5}-\text{StrepII}}$, a Mock solution or not infiltrated. The expected migration distance of the V5-tagged peptides are shown on the right of the blot while on the left are the protein ruler bands in kDa. PI.1 refers to the canonical isoform while PI.2 to the isoform with intron retention and early stop codon.

One band corresponding in size to PI.1 was detected in the $_{\text{pUBQ10}}$PI$^{\text{V5}-\text{StrepII}}$ lane and the negative controls did not result in any specific signals, as expected. Surprisingly, the $_{\text{pUBQ10}}{}^{\text{StrepII}-\text{V5}}$PI sample also did not yield any specific bands. As explained above, I designed this construct to detect both PI.1 and PI.2 simultaneously. Even if PI.2 was not expressed PI.1 should still be detected as in the $_{\text{pUBQ10}}$PI$^{\text{V5}-\text{StrepII}}$ sample. To understand this result, I first tested all plasmids I used for generating the construct by sequencing and enzymatic digestion and found them to be in order. I next considered that the failure to detect both protein isoforms in the experiment could be

due to reduced protein stability of the N-terminally tagged version. Because PI forms and obligate heterodimer with AP3, I hypothesized that the newly translated PI protein may need AP3 for stability and for translocation into the nucleus, where it could be protected from rapid degradation. Current and previous group members had reported similar difficulties with detecting AP3 and PI especially with heterologous expression in tobacco. I therefore repeated the experiment with the same PI constructs while co-expressing an $AP3^{GFP}$ fusion protein using a *pUBQ10::AP3-GFP* construct. To this end, I co-infiltrated with AP3 and PI constructs into tobacco leaves. In this experiment, I used in addition single construct infiltrations for ${}_{pUBQ10}{}^{StrepII-V5}PI$, ${}_{pUBQ10}PI^{V5-StrepII}$, and ${}_{pUBQ10}AP3^{GFP}$ as controls. I then repeated the Western blotting with the expanded set of samples (Figure 3.10).

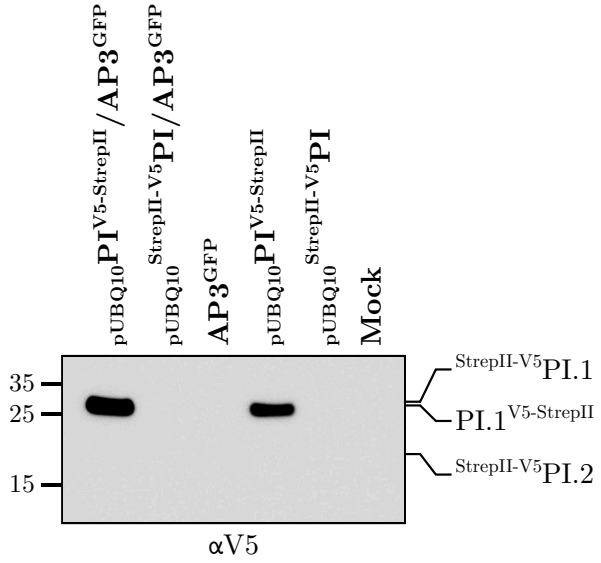


Figure 3.10: Results of immuno blots with total protein extracts from tobacco leaves. Leaves were infiltrated with ${}_{pUBQ10}{}^{StrepII-V5}PI$, ${}_{pUBQ10}PI^{V5-StrepII}$ or a Mock solution. The WT sample comes from non infiltrated tobacco leaves. The expected migration distance of the V5-tagged peptides are shown on the right of the blot while on the left are shown the migration distances of the protein ruler in kDa. PI.1 refers to the canonical isoform while PI.2 to the isoform with intron retention and early stop codon. The experimental setup was identical to the experiment shown in Figure 3.9 but with the addition of the $AP3^{GFP}$ (*pUBQ10::AP3-GFP*) construct. $AP3^{GFP}$ was infiltrated in tobacco leaves on its own or coinfiltrated with ${}_{pUBQ10}{}^{StrepII-V5}PI$, ${}_{pUBQ10}PI^{V5-StrepII}$.

As in the previous experiment, the only bands detected corresponded to PI.1 when expressed with a C-terminal tag either alone or together with AP3. Thus, co-expression with AP3 did not lead to an accumulation of the N-terminally tagged PI protein.

I next considered the possibility that the $^{\mathrm{StrepII-V5}}$PI transcript was unstable and did not accumulate in the transformed cells. To test this hypothesis, I performed a semiquantitative PCR to quantify the transcript levels of the two tagged PI constructs. For this experiment, I extracted RNA from the same tissue from which I extracted total protein for Western blotting. Two specific primer pairs were designed to amplify each of the PI constructs. The semiquantitative PCR amplification was halted at 20, 25 and 30 cycles and samples were taken and analysed by agarose gel electrophoresis (see Figure 3.11).

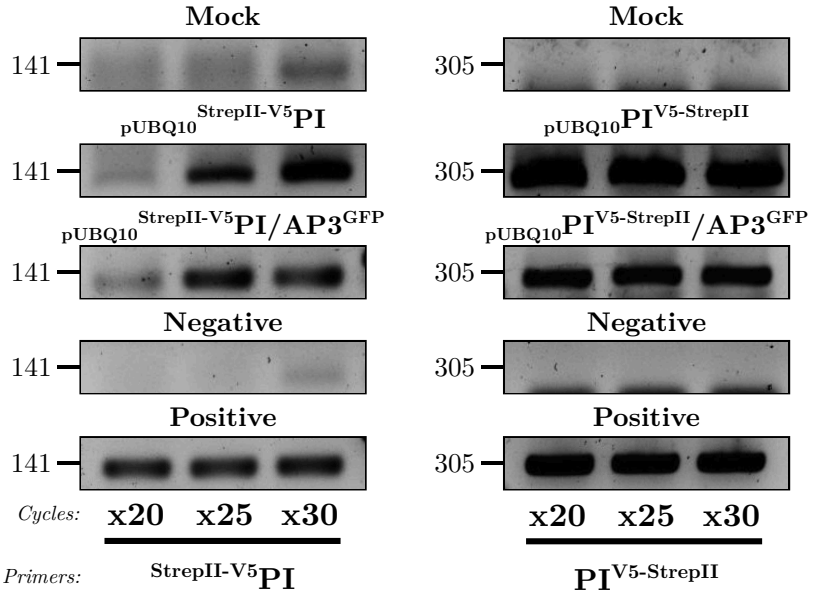


Figure 3.11: qPCR on cDNA extracted from tobacco leaves infiltrated with the constructs $_{\mathrm{pUBQ10}}{}^{\mathrm{StrepII-V5}}$PI, $_{\mathrm{pUBQ10}}$PI$^{\mathrm{V5-StrepII}}$ and AP3$^{\mathrm{GFP}}$(*pUBQ10::AP3-GFP*). Three PCRs were done at 20, 25 and 30 cycles respectively, with primers specific for $^{\mathrm{StrepII-V5}}$PI or PI$^{\mathrm{V5-StrepII}}$. The expected product size in base pairs is shown on the left of each gel. The number of cycles and the primer pairs used are found on the bottom of the panel. The positive controls are plasmids containing the respective PI constructs and the negative is a no-template control. The mock sample refers to cDNA extracted from tobacco leaves infiltrated without Agrobacterium.

The primer pair designed for $_{\text{pUBQ10}}{}^{\text{StrepII}-\text{V5}}\text{PI}$ seemed to be slightly less efficient than the primer pair for $_{\text{pUBQ10}}\text{PI}^{\text{V5}-\text{StrepII}}$ pair, while also showing a shallow amplification signal in the Mock and no-template controls at 30 cycles. An estimate of the primer efficiency was done by looking at the positive controls, which consist of ∼1 ng of the respective plasmids. In the samples infiltrated with $_{\text{pUBQ10}}{}^{\text{StrepII}-\text{V5}}\text{PI}$ and $_{\text{pUBQ10}}\text{PI}^{\text{V5}-\text{StrepII}}$ alone, or co-infiltrated with *pUBQ10::AP3-GFP*, the abundance of $_{\text{pUBQ10}}{}^{\text{StrepII}-\text{V5}}\text{PI}$ transcript product was lower than in $_{\text{pUBQ10}}\text{PI}^{\text{V5}-\text{StrepII}}$. However, these differences did not appear to be substantial and could be due to the observed differences in primer efficiency. Thus, the complete absence of the N-terminally tagged PI protein in the immunoblots was not caused by a lack of transcription of the corresponding construct. It appears then that somehow the presence of an epitope tag at the N-terminus interfered with protein accumulation.

3.3.5 LncRNA prediction

LncRNAs, especially when intragenic, are often elusive to short read sequencing. LncRNAs are difficult to assemble, hard to univocally assign to one isoform and often have low expression levels [Uszczynska-Ratajczak et al. 2018]. Longer reads can help to obtain a more accurate assembly and better sensitivity for lowly expressed lncRNA transcripts. To investigate lncRNA transcription in the floral time-course experiment described above, I extracted potential lncRNAs from the `stringtie2` assembly with a custom `Python` script (`lncRNA_finder.py`). To this end, I developed an analysis pipeline based on key characteristics of lncRNAs as described in the literature to date [Rai et al. 2018; Khemka et al. 2016]. First, I eliminated from the dataset all the transcripts shorter than 200 bp. Next, I used the position of the transcripts respect to the Araport11 reference annotation to filter out uninteresting transcripts. The relative position was calculated by `gffcompare` (see section 3.3.2), and classified using letters of the alphabet. Putative lncRNAs belong to one of the categories among `[j, o, i, x, u]`: When comparing `stringtie` annotation to the Araport11 reference, `gffcompare` classifies each transcript in the `stringtie` dataset by their match with the Araport11 reference, in base of the similarity between intron chains of overlapping transcripts. All the different kinds of lncRNA transcripts known to date (see section 3.1.2) belong to at least one of these categories:

`j`: multi-exon with at least one junction match (intragenic)

`o`: other or same strand overlap with reference exons (intragenic)

`i`: fully contained within a reference intron (intragenic, intronic or "lincRNA")

`x`: exonic overlap on the opposite strand (intragenic)

`u`: none of the above (no overlap with known transcipts, intergenic)

Next, I mapped each one of the remaining transcripts on the full Brassicaceae proteome (taxid = 3700) with `BLASTX`, using an evalue cutoff of 10^{-4}. The transcripts with hits that had an alignment length of more than 40 aa, a percent identity higher than 35% and an alignment corresponding to at least 35% of the query sequence were carried on to the next step. Finally, the coding potential of each transcript was computed using `CPC2`, a coding potential calculator [Kang et al. 2017] that predicts how likely a given sequence is to encode for a functional protein. The transcripts that were considered to be "coding" by `CPC2` were excluded from the dataset. Finally, I annotated the list of remaining transcripts, the putative lncRNAs, with a custom script (`lncRNA_annotate.R`) which connects each lncRNA with its underlying locus, if any, and then fetches a description from the Ensembl database. I identified a total of 53 putative lncRNAs. Nineteen of these RNAs are transcribed from known non-coding loci, 31 from unknown loci and 3 from protein coding genes. Out of the 3 coding genes, only one gene has an experimental annotation and encodes for a ribosomal protein *RPL41G*, the others are generically annotated as putative proteins.

To test the method, I also analysed known Arabidopsis lncRNAs that had been experimentally-validated: *COOLAIR (AT5G01675.1)* [Hawkes et al. 2016], *FLORE (AT1G69572.1)* [Henriques et al. 2017], *ELENA1 (AT4G16355.1)* [Seo et al. 2017]. Out of the three known lncRNA, only *ELENA1* was predicted as non-coding by the pipelines, while the other two had at lest a significant hit in the Brassicacee protein database.

Table 3.6: Putative novel lncRNAs found in the `stringtie` dataset

Gene	Transcript	GeneName	TranscriptBiotype
AT1G04425	*MSTRG.330.1*		ncRNA
	MSTRG.330.2		ncRNA
	MSTRG.330.3		ncRNA
	MSTRG.330.4		ncRNA
	MSTRG.330.5		ncRNA
	MSTRG.330.6		ncRNA
	MSTRG.330.8		ncRNA
	MSTRG.330.9		ncRNA
MSTRG.2191	*MSTRG.2191.1*		
MSTRG.2677	*MSTRG.2677.1*		
AT1G34418	*MSTRG.2851.1*		ncRNA
MSTRG.2935	*MSTRG.2935.1*		
MSTRG.2991	*MSTRG.2991.1*		
MSTRG.3070	*MSTRG.3070.2*		
MSTRG.3297	*MSTRG.3297.1*		
MSTRG.3404	*MSTRG.3404.1*		
MSTRG.4073	*MSTRG.4073.1*		
MSTRG.5553	*MSTRG.5553.1*		
MSTRG.5992	*MSTRG.5992.1*		
MSTRG.6107	*MSTRG.6107.1*		
MSTRG.6125	*MSTRG.6125.1*		
MSTRG.6158	*MSTRG.6158.1*		
MSTRG.6342	*MSTRG.6342.2*		
AT2G40205	*MSTRG.8321.2*	RPL41G	protein coding
AT2G09195	*MSTRG.8370.1*		lncRNA
MSTRG.8836	*MSTRG.8836.1*		
MSTRG.8998	*MSTRG.8998.1*		
AT3G04854	*MSTRG.9434.2*		protein coding
AT3G05152	*MSTRG.9468.1*		ncRNA
MSTRG.9638	*MSTRG.9638.1*		
MSTRG.10075	*MSTRG.10075.1*		

Table 3.6: Putative novel lncRNAs found in the `stringtie` dataset *(continued)*

Gene	Transcript	GeneName	TranscriptBiotype
MSTRG.10959	*MSTRG.10959.1*		
MSTRG.11157	*MSTRG.11157.1*		
MSTRG.11666	*MSTRG.11666.1*		
AT3G59765	*MSTRG.13071.1*		ncRNA
	MSTRG.13071.4		ncRNA
MSTRG.13556	*MSTRG.13556.1*		
AT4G06701	*MSTRG.13871.1*		ncRNA
	MSTRG.13871.2		ncRNA
	MSTRG.13871.3		ncRNA
	MSTRG.13871.7		ncRNA
AT4G10843	*MSTRG.14051.7*		protein coding
AT4G06080	*MSTRG.14244.1*		lncRNA
MSTRG.15912	*MSTRG.15912.1*		
MSTRG.16081	*MSTRG.16081.1*		
AT4G09240	*MSTRG.16269.4*		lncRNA
MSTRG.16297	*MSTRG.16297.1*		
MSTRG.16959	*MSTRG.16959.1*		
	MSTRG.16959.2		
	MSTRG.16959.3		
MSTRG.17859	*MSTRG.17859.1*		
MSTRG.19035	*MSTRG.19035.1*		
MSTRG.19086	*MSTRG.19086.1*		

3.4 Discussion

Third-generation transcriptomics is still in the early stages of its development. The possible advantages that longer read sequencing can bring to an analysis, allowing precise transcriptome assemblies and better read alignments, can't be overstated. The aim of this project was to use the state-of-the-art Oxford Nanopore third-generation sequencing technology to explore transcriptome changes during early floral development. Moving at the cutting edge of transcriptomics has proven more challenging than expected for two main reasons. Firstly, even with the recent improvements, Nanopore reads are still error-prone, with a modal value of alignment identity around ∼95%. Secondly, the bioinformatics tools available for third-generation sequencing transcriptomics are few and mostly not optimised for long-reads. This issue is particularly noticeable in the assembly process, where Nanopore sequencing is really supposed to outperform other approaches. Tools such as `stringtie` have been built for short-reads data and have later been adapted to noisier long-reads but do not take full advantage of longer reads. There is a flourishing number of de novo transcriptome assembly tools but currently no analysis software can use long reads properly when an annotated reference genome is available for genome-guided assembly (see section 3.1.3).

3.4.1 DTU genes and enriched GO categories

The GO analysis of the significant DTU genes outlined in 3.3.2 suggested that splicing might be involved in flower development, particularly at later stages of development. Enriched GO terms related to flower development were found prevalently across the 4-8 DAI time-points. Notably however, the majority of the GO terms were related to a number of biotic and abiotic stresses. As described in section 3.1.2, splicing in plants seem to be heavily involved in stress response, with many such examples in known literature. Moreover, Many of those contain very few genes, which can easily skew the calculation. It is therefore more likely to find enrichment in such GO categories. It is worth noting that some of the genes in stress related categories are also important flower development genes. A prime example is *FLC*, which is the central regulator of vernalization-dependent flowering. *FLC* is part of both the *response to cold* and *flower development* GO categories. Finally, the plants might have suffered mild stresses in the growth room, which could

have further increased the number of DTU events in genes involved in stress response.

3.4.2 Shortcomings of Nanopore sequencing

In section 3.3.2, I described the major challenges of Nanopore sequencing, namely the high error rate, degradation products being mistaken for full-length transcripts, the unstranded nature of the library and misplaced reads. The degraded and misplaced reads trick the assemblers into annotating these artifacts as actual transcripts and as a consequence, the reads are realigned on a corrupted annotation with a portion of them being assigned to false transcripts. In the end, the read counts can be severely skewed and this affects the statistical power of the DTU analysis.

To date, there isn't a good program to perform adapter trimming. The only one currently available is a third party tool, `porechop`, which has been unsupported since 2018. In principle, the software `guppy` performs adapter trimming during the basecalling step, but in practice, the adapters do not seem to be trimmed efficiently, with a large number of soft clipped bases present at most transcripts' ends (see Figure 3.3a). Passing reads through available tools, like `cutadapt` or the above cited `porechop` didn't seem to achieve the goal, with either too much or no trimming at all. Adapter trimming is particularly difficult for Nanopore reads because of the comparatively low quality of the base calling, which is even poorer at the read ends. The downside of using machine learning to call bases from the pore signal is that it is very dependent on the training set. In the case of `guppy`, the performance is optimal at specific DNA translocation speeds namely 450 bp/s. When the DNA speed through the pore deviates from the speed the model expects, the basecalling quality drops considerably. This usually happens in two cases: at the read ends, because it takes ∼200 bp for the DNA speed to "stabilise", and in long stretches of repetitive bases, which can translocate faster causing sudden spikes in speed. The full-length reads I used for the transcriptome assembly didn't suffer from this issue as much because `pychopper2` is able to use the DCS109 primers as anchor points for trimming (see Figure 3.3a).

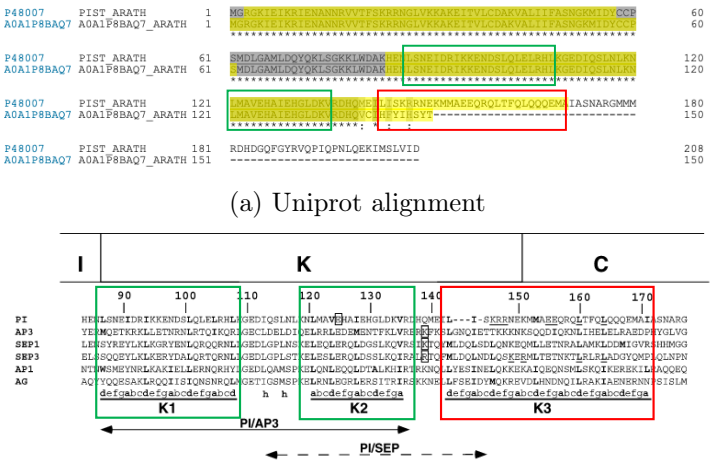
The pipeline I designed for this work was largely based on the benchmarks and guidelines outlined in Soneson et al. [2019], which is currently the best overview of bioinformatics methods for long-read transcriptomics. `Flair` was also incorporated in the pipeline with mixed results. It is diffi-

cult to assess the performance of these tools due to the lack of real truth-set. In their benchmark, Oxford Nanopore mostly uses mixture of perfect RNAs made *in vitro*, like SIRV, which fail to capture the complexity of true biological transcripts. For example, palindromic or repetitive sequences can cause significant problems during reverse transcription, by tricking the polymerase in transcribing erroneous sequences or stalling it with secondary structures. Arabidopsis introns, which are rich in A and T bases, likely affecting translocation speed. Splicing sites tend to be palindromic which can create secondary structures and stall sequencing, increasing truncated transcripts. Furthermore, degraded RNA is present in even the highest grade RNA samples but current bioinformatics tools fail to properly account for that.

3.4.3 PI.2 could be incapable of high order interactions

PI transcript dynamics could have the potential to add a layer of regulation during flower development by finely tuning PI's ability to form high-order complexes with other MADS-domain transcription factors. The PI.2 isoform retains the third intron that introduces an early stop codon. While the translated protein has not been described in literature, the translation should generate a PI protein that lacks much of its C-terminus, possibly affecting its ability to interact with other proteins (Figure 3.12).

It has been shown that PI binding to AP3 and SEP3 strongly depends on different stretches of its protein sequence, mainly in the K domain. AP3 has higher affinity with the K1 and K2 domains, while SEP3 has more affinity for the K2 and K3 domains, specifically for the interhelical region between K2 and K3 [Yang and Jack 2004; Yang et al. 2003a]. SEP3 is often associated with the formation of higher-order complexes of MADS-domain proteins. The *pi-5* mutant has a E125K mutation that falls in the K2 domain, lowering PI's affinity with AP3 and SEP3. The mutant exhibits defects in the second whorl, where sepals develop in place of petals, but third whorl stamens are most often normal [Yang et al. 2003b]. Even though the Nanopore sequencing didn't seem to accurately quantify the expression of PI.2, the isoform is already present in the Araport11 reference annotation and the isoform expression is detected albeit at low levels (Figure 3.8). In light of the literature I decided to further investigate this particular splicing event because it could have interesting effects regulating the formation of higher



(a) Uniprot alignment

(b) PI interaction domains, adapted from [Yang and Jack 2004]

Figure 3.12: A schematic representation of the predicted translation of PI protein isoforms. The three K domains are framed by colored square: K1 and K2 are framed by green square because they are translated in both isoforms, while K3 has a red square, because it is located after the early stop codon introduced by the retained intron.

order MADS-domain protein complexes.

3.4.4 PI N-terminal tagged protein is not translated

The experiment with the constructs I constructed to study PI splicing led to inconclusive results as described in section 3.3.4. There were two expected scenarios for the Western blot's results. If the transcript with the retained intron made the transcript unstable, there should be only one band corresponding to the reference protein in the N-terminally tagged PI, otherwise both a reference band and a lower band corresponding to the truncated isoform should be detected. In any case, a reference band should be present, as it happens when PI is tagged at the C-terminus (Figure 3.9). One hypothesis was that the N-terminally tagged version of the protein can't properly translocate to the nucleus in tobacco leaves, due to the absence of its partner protein AP3. Co-infiltrating PI constructs with a construct for AP3 expression didn't seem to help to stabilise PI, with $^{\text{StrepII}-\text{V5}}$PI still being undetectable (see Figure 3.10). The constructs' expression in the tobacco samples was quantified, ruling out a problem with transcript stability (Figure 3.11).

More experiments are needed to find the root cause of the issue. It is known that N-terminal tags don't work well for MADS-domain proteins, probably because they interfere with the MADS domain which is required for DNA-binding. Although I was aware of this limitation before commencing the experiment, the goal here was to test for protein expression and not to have a fully functional protein. The smallest epitope tag available (V5) was fused to PI, in place e.g. of a bulkier GFP tag, to minimise disruption to protein stability. It's possible that protein stability is severely disrupted in N-tagged versions of PI, regardless of the tag size, or that it is only stable when expressed in Arabidopsis flowers due to the presence of necessary cofactors. I generated ${}_{\mathrm{pPI}}\mathrm{PI}^{\mathrm{StrepII-V5}}$ and ${}_{\mathrm{pPI}}\mathrm{PI}^{\mathrm{V5-StrepII}}$ lines to repeat similar experiments in Arabidopsis. I am growing the second generation of plants at the moment of writing.

3.4.5 LncRNA prediction performance

In section 3.3.5, I used the newly assembled annotation, which contains both known and novel transcripts, to predict new lncRNAs. Known non-coding transcripts are a good control for the prediction method. Although the pipeline identified 53 putative lncRNAs, including 19 known lncRNAs, it was only able to predict two out of three experimentally validated lncRNA. There are no clear guidelines in what consists a reliable lncRNA discovery pipeline because the definition and scope of this transcript class is very broad. It is only known what lcnRNAs aren't: they aren't coding and no longer than 200 bp. There is an extensive body of work on lncRNA discovery described in the literature, from more simple filtering pipelines to machine learning models. In this work I used both approaches: I filtered out possible coding lncRNAs by using a series of heuristic cutoffs and then calculated a "coding score" using a predictive model trained on real data. This approach could be too restrictive, as suggested by the lack of well known lncRNAs among the predicted lncRNAs. Recently, a plethora of machine learning approaches for lncRNAs identification have been published. The models are trained by supervised learning, to classifiy sets of protein coding and non-coding sequences. This approach is highly dependent on the sequences used in the training set. I tried all the available models to date and either they failed to run, even after reasonable debugging, or they weren't trained on plant data [da Costa Negri et al. 2018; Simopoulos et al. 2018; Cao et al.

2020; Ammunét et al. 2022]. Training a new model is a project in itself and unfortunately was well beyond the scope of this work.

In recent years, new lncRNA databases have become available which collect and annotate lncRNAs from a variety of plant species by mining publicly RNA-Seq datasets. The two main databases are GreeNC and PLncDB, which were both recently updated with the introduction of thousands of new lncRNA transcripts [Marsico et al. 2022; Jin et al. 2021]. These resources will be useful to orthogonally validate the lncRNAs found in this work and quantify the conservation of their sequence and position (synteny) across species. LncRNA can regulate microRNA levels by acting either as microRNA precursors or microRNA "sinks" (see section 3.1.2). Both GreeNC and PLncDB contain precomputed information relating the lncRNAs in the database and the landscape of known microRNAs, which offers both a direct resource and a blueprint to develop new pipelines to characterise the lncRNAs described in this work.

3.4.6 Future perspectives

The next step will be to test if the N-terminally tagged PI protein that gave inconclusive results in Tobacco leaves, can be detected in stably transformed Arabidopsis lines. Transgenic lines are being selected at the moment and will be ready for analysis soon. I also generated more lines where the two PI's protein isoforms were expressed under the *PI* promoter in a *pi-1* mutant background. It will be interesting to see if both isoforms can complement a full loss of function phenotype and what will be the resulting phenotype. As explained in section 3.4.3, PI.2 is missing part of the domain that interacts with SEP3, which is involved in formation of MADS-domain protein tetramers. This could affect PI activity in subtle ways, for example leading to partial complementation in which the phenotype of certain organs is completely rescued in certain organs but not others. This partial complementation could resemble the phenotype of *pi-5* (see section 3.4.3), which also as a defective protein-protein interaction due to a mutation in the K-domain [Yang et al. 2003b]. This work could provide new insights into the role of high-order protein-protein interactions in floral organ formation.

3.5 Conclusion and final remarks

Long-read sequencing has the potential to revolutionise transcriptomics, in principle allowing the detection of new splicing events and lncRNA that cannot be readily identified using short read sequencing. In this project, I therefore explored the potential of Nanopore sequencing for the analysis of gene expression during Arabidopsis flower development. I found that especially problems with base calling accuracy and the current lack of tailor-made bioinformatics tools for data analysis hamper the use of this technology at the present time. Still, by adapting existing analysis tools into a coherent pipeline, I succeeded in identifying new splicing events as well as new candidates for lncRNAs. What's more, the datasets I have generated can be re-analysed whenever improved software tools become available. For now, however, I aim at publishing the dataset to make it available to a broader scientific community. It would be very valuable for anybody studying flower development and can be selectively mined to unravel complex splicing events at loci of interest and to explore the expression dynamics of lncRNAs.

Chapter 4

Exploring the interactomes of the trichome repressors TRIPTYCHON and CAPRICE

4.1 Introduction

As outlined in Chapter 1, trichomes, or hairs, play important roles in the defence of plants against herbivores. Trichome formation is therefore a target trait for breeding efforts to generate new crop cultivars with increased resistance to insect pests [Fambrini and Pugliesi 2019]. A recent paper from our laboratory showed that trichomes could be initiated on the seed pods of Arabidopsis through loss-of-function mutations in the key trichome repressors *TRIPTYCHON (TRY)* and *CAPRICE (CPC)* when they were introgressed into a mild *ag* mutant background [Ó'Maoiléidigh et al. 2018]. To test whether this phenotype can also be induced in crop plants, efforts are underway in our laboratory to translate the research already conducted in Arabidopsis to Brassica crops. In addition to being targets for crop improvement, trichomes are interesting from a developmental biology perspective as they constitute an excellent model system for studying cell type specification and differentation. In fact, there is a large variety of trichomes in the plant world, with sometimes complex architectures. However, it is still largely unknown how a plant cell undergoes the structural changes required to develop into a trichome. Moreover, the pattern of trichome initiation on the plant is highly organized to ensure that trichomes grow only in specific positions and

at a certain density. In the sections below, I will briefly review the current model for trichome development in Arabidopsis and describe some of the knowledge gaps that remain to be filled.

4.1.1 The regulatory mechanism for trichome formation

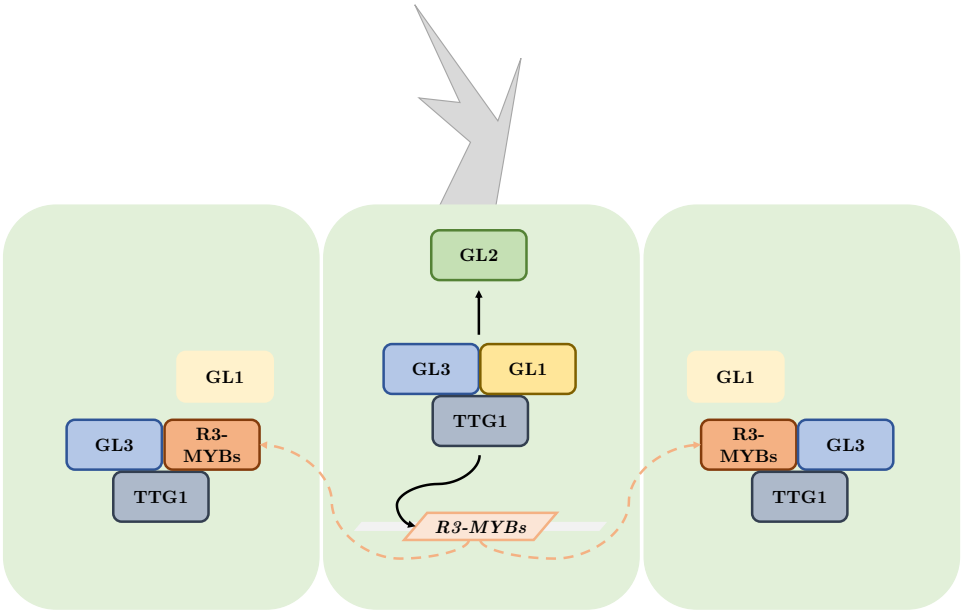


Figure 4.1: Schematic representation of the main pathway regulating trichome initiation and development. The trichome initiation complex, composed by GL3, TTG1 and GL1, activates the expression of GL2 and a number of genes from the *R3-MYB* family of trichome repressors. GL2 then initiates the formation of a trichome (cell shown in the center). The R3-MYB proteins travel from the trichome-forming cell to neighboring cells and compete with GL1 for binding of the trichome initiation complex, repressing its activity and hence trichome development.

Genetic and biochemical analyses showed that trichome initiation in Arabidopsis is mediated by a transcriptional activator complex composed of three proteins: the R2R3-MYB protein GLABRA1 (GL1); a basic helix-loop-helix (bHLH) protein, either GLABRA3 (GL3) or ENHANCER OF GLABRA3 (EGL3); and the WD40 repeat (WD40) protein TRANSPARENT TESTA GLABRA1 (TTG1) [Morohashi et al. 2007; Payne et al. 2000; Szymanski and Marks 1998]. This complex promotes expression of the homeodomain transcription factor GLABRA2 (GL2), which directs trichome formation [Lee

and Schiefelbein 2002; Ohashi et al. 2002; Rerie et al. 1994].

This model of trichome initiation was recently revised to incorporate new results on the dynamic composition of the activator complex. It was shown through yeast two-hybrid, Bimolecular Fluorescence Complementation (BiFC) and Förster Resonance Energy Transfer (FRET) assays that while GL3 interacts with both TTG1 and GL1, TTG1 and GL1 do not directly interact with each other.

Additionally, the results of yeast three-hybrid assays investigating the interaction of GL3 with either GL1 or TTG1 suggested that GL3 interaction with either of these proteins may be mutually exclusive [Digiuni et al. 2008; Pesch et al. 2015]. Thus, trichome initiation may depend on the activities and the interplay of two regulatory sub-complexes, both involving GL3.

A key question in the study of trichomes is how trichomes are initiated with specific spacing and density on the surface of organs such as leaves, where some epidermal cells develop into trichomes, while trichome formation is suppressed in neighboring cells. Particle bombardment studies elegantly demonstrated that TTG1, but not GL3 or GL1, can move from cell to cell. Furthermore, it was shown that *TTG1* expression is ubiquitous throughout leaf development while a fusion protein between TTG1 and YELLOW FLUORESCENT PROTEIN (YFP) was found to be localised only to trichome cells [Balkunde et al. 2011; Bouyer et al. 2008]. These experimental findings, supported by computational simulations, led to the formulation of a model for trichome spacing in which the depletion of TTG1 around mature trichomes suppresses the formation of new trichomes by removing an important component of the trichome activator complex [Bouyer et al. 2008].

It is important to note that many of the genes involved in trichome development also play roles in the development of root hairs. In fact, the main regulators for trichome and root hair initiation are same, with the notable exception of GL1, whose function in roots is carried out by a homologous R2R3-MYB protein, WEREWOLF (WER) [Song et al. 2011; Tominaga et al. 2007]. Interestingly, GL2 has opposite activities in leaves and roots: it promotes trichome formation in leaves but suppresses hair development in roots [Schiefelbein 2003; Rerie et al. 1994].

4.1.2 The R3-MYB gene family of trichome repressors

Another important component of the regulatory mechanism controlling trichome initiation are repressors of trichome formation. These are encoded by members of the R3-MYB gene family, which comprises 7 genes in Arabidopsis: the above-mentioned *TRY* and *CPC*, *TRICHOMELESS1 (TCL1)*, *TCL2*, *ENHANCER OF TRY AND CPC1 (ETC1)*, *ETC2* and *ETC3*. The proteins encoded by these genes can compete with the structurally related GL1 for binding to GL3/EGL3 and hence disrupt the function of the trichome activator complex [Esch et al. 2003; Kurata et al. 2005; Tominaga-Wada and Wada 2016; Tominaga et al. 2007; Wang et al. 2008; Zhang 2003].

As mentioned in section 4.1.1, *TRY* and *CPC* are arguably the two most important trichome repressors in the family. When ubiquitously expressed, they lead to a completely glabrous plant. GUS reporter gene assays showed a very similar pattern of expression for both genes: ubiquitous in young leaves and then increasingly more localised to trichome cells as leaves mature [Pesch et al. 2013; Schellmann et al. 2002]. Imaging of fluorescent reporters for TRY and CPC suggested that the proteins are expressed in mature trichomes (after activation by components of the trichome activation complex; see below) and then travel to neighboring cells, preventing the differentiation of new trichomes in the vicinity of developing trichomes [Digiuni et al. 2008; Schellmann et al. 2002].

Although the amino acid sequences of TRY and CPC exhibit a high degree of similarity, the phenotypes of null mutants are strikingly different: while *cpc* mutants exhibit higher trichome density, trichome density in *try* mutants is unaffected. Instead, a small number (∼4%) of trichome initiation sites form clusters composed of multiple trichomes [Schellmann et al. 2002]. What's more, in *try cpc* double mutants, the number of trichome initiation sites is lower than in *cpc* single mutants, but the cluster frequency dramatically increases relative to *try* single mutants to around 76% of initiation sites [Schellmann et al. 2002].

Although there is limited understanding of the molecular basis behind the phenotypic variations observed in *try* and *cpc* mutants, it has been demonstrated that the TRY protein, when expressed under the *CPC* promoter, was able to fully rescue the clustering phenotype of a *try* mutant. In contrast to this, the expression of *CPC* under the *TRY* promoter did not lead to a res-

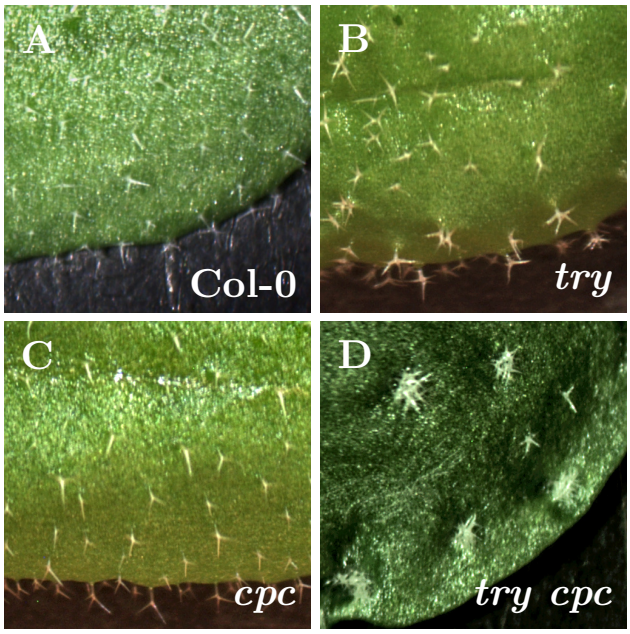


Figure 4.2: Trichome phenotypes of *try* and *cpc* single mutant and of *try cpc* double-mutant plants. Leaves of the different mutants and of a wild-type plant (accession: Col-O) are shown. Pictures were kindly provided by Caoilfhionn Breathnach. *try: try-29760*, *cpc: cpc-1*, *try cpc*: *try-29760 cpc-1*.

cue of the clustering phenotype [Pesch and Hülskamp 2011]. These promoter swap experiments indicate that the TRY protein has specific properties in preventing cluster formation in comparison to CPC.

TRY and CPC proteins have been shown to differ in a number of other aspects. For example, while TRY and CPC were shown to interact with GL3 in transcriptional activation assays in protoplasts, CPC expression resulted in a stronger transcriptional response than TRY [Wang et al. 2008]. CPC was also shown to inhibit the binding of GL1 to GL3 more efficiently than TRY in a yeast three-hybrid system [Wester et al. 2009].

A difference in binding activity for GL3 may also result in differences in how the corresponding proteins move from cell to cell. TRY and CPC are relatively small proteins, well below the size exclusion limit of plasmodesmata (∼40 kDa). It is therefore not surprising that they are capable of cell-to-cell movement; however, it has been shown that they can undergo cell-to-cell movement even when their size is increased by translational fusions with several copies of GREEN FLUORESCENT PROTEIN (GFP). This

suggests that they can somehow overcome the size exclusion limit of plasmodesmata [Kurata et al. 2005]. Specifically, it has been shown that when expressed under its native promoter, CPC can travel freely in roots from hairless cells to hair cells. Strikingly, CPC movement can be detected even when the protein is fused to a 4xGFP protein tag [Kurata et al. 2005; Oparka et al. 1999]. When expressed under the same *CPC* promoter, free GFP (27 kDa) moves freely between the cells but a fusion between two GFP proteins (54 kDa) does not, suggesting that CPC can increase the size exclusion limit up to at least 119 kDa, i.e., the size of the CPC-4xGFP protein [Kurata et al. 2005]. How CPC is able to overcome the normal size exclusion limit, and why this might be required, is currently not understood.

It has also been suggested that the regulation of *TRY* and *CPC* by different, and competing, components of the trichome initiation complex may be the cause of these differences. Specifically, GL1 may prevent the activation of *TRY* by the TTG1-GL3 complex and, vice versa, the activation of *CPC* by GL1-GL3 might be repressed by TTG1 [Pesch et al. 2015]. However, this model does not agree with the available literature for *TRY* activation: *TRY* expression is not detected in *gl1* mutants and chromatin immunoprecipition (ChIP) experiments have shown that GL1 can bind to the *TRY* promoter but is not necessary for GL3 binding [Digiuni et al. 2008; Morohashi et al. 2007; Morohashi and Grotewold 2009].

4.1.3 TurboID proximity labeling

As discussed above, determining the distinct roles of TRY and CPC in repressing trichome formation has proven difficult. One way to investigate the proteins' functions and possible differences is to identify their interaction partners. Many different methods have been developed over the years to identify possible interaction partners of proteins. These include, for example, yeast two-hybrid screens and immunoprecipitation coupled to mass spectrometry (IP-MS). While these methods have been successfully employed in a wide range of studies, they have several disadvantages that may preclude the proteome-wide identification of interaction partners: yeast two-hydrid screens, when used for plant proteins, mean employing a heterologous system and testing for interactions in the yeast nucleus without consideration for the intracellular localization of the proteins of interest. In contrast, IP-MS experiment can be performed *in planta* reducing the likelihood of experi-

mental artefacts. However, in this case, interaction partners can only be recovered if they interact stably – weak or transient interactions are much more difficult to detect using this method. To overcome these limitations, new methods have been developed in recent years that can not only detect weak and transient interactions but also proteins that do not directly interact with the protein of interest but are in proximity to it.

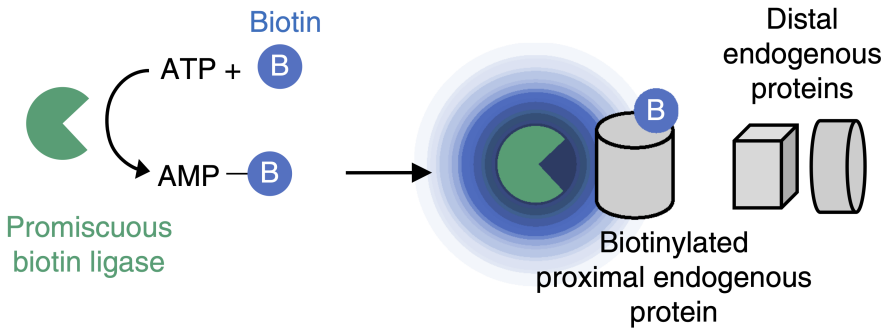


Figure 4.3: Proximity labeling by TurboID. The TurboID biotin ligase has promiscuous activity and in presence of biotin, biotinylates all proteins that are in its vicinity. The approach can therefore be used to identify direct interactors of a protein of interest as well as nearby proteins.

In these proximity labeling approaches, a promiscuous labeling enzyme is fused with the protein of interest, to tag all proteins that come within a few nanometers of the fusion proteins (see Figure 4.3). The tagged proteins can be then isolated and analysed. Since its introduction, proximity labeling has become a valuable alternative to affinity purification/mass-spectrometry based interactome mapping.

BioID, based on a promiscuous biotin ligase from *E. coli*, was the first proximity labeling system to be developed and has been used extensively over the past 10 years. The fraction of proteins biotinylated by BioID can be readily isolated leveraging the highly stringent interaction between biotin and streptavidin, greatly reducing non-specific binding compared to a conventional affinity purification [Li et al. 2019; Roux et al. 2012]. Due to the very short distance required, the labeling is highly specific, reducing false positives; furthermore, BioID proximity labeling is applicable to both soluble and insoluble protein fractions widening the range of cellular proteins that can be pulled down. To date, over 100 proximity labeling studies have

been published, primarily using cell cultures that can be readily treated with biotin to activate the BioID tag. Only recently proximity labeling has been applied to a wider range of model systems [Trinkle-Mulcahy 2019; H Nakamura 2013; Lin et al. 2017].

Even though BioID circumvented many of the limitations of other methods for the identification of interacting proteins, it has its own shortcomings. In particular, its relatively slow labelling kinetics, requiring an incubation of at least 16-24 hours to achieve labeling saturation for common downstream analysis, made it unsuitable for most *in vivo* applications. To improve the BioID method, a new version of the promiscuous biotinylase was engineered by yeast display-based directed evolution. The catalytic activity of the improved enzyme, termed TurboID, is an order of magnitude higher than that of BioID, reducing the incubation time needed for labelling in mammalian cells from >16 hours to approximately 10 minutes.

Because of its improved properties, TurboID opened up new possibilities for proteomic studies in plants. Unlike animal tissues, plant tissues are generally harder to treat, cytoplasmic volume is much lower relative to cell wall mass and the presence of endogenous biotin can interfere with protein detection and identification. Additionally, while the optimal temperature for BioID activity is 37°C and thus too high for *in vivo* applications in plants, TurboID assays can be efficiently conducted at lower temperatures (around 25°C). The first studies employing TurboID in plants have been published in recent years [Khan et al. 2018; Mair et al. 2019; Zhang et al. 2019; 2020; 2019; Arora et al. 2020; Kim et al. 2023], demonstrating the potential of the method for plant research.

4.2 Project aims

As outlined above, the regulation of *TRY* and *CPC* expression has been studied in some detail, and their genetic interactions with other trichome development genes have been elucidated. However, it is still unclear why these very similar proteins have different functions in the regulation of trichome initiation, as indicated by their mutant phenotypes. To address this open question, I hypothesized that the properties of the proteins and specifically, differences in their interaction partners, account for their functional differences. To test this, my work aimed initially at exploring the structural

features of TRY and CPC through the use of silico analysis. Furthermore, I wanted to characterize and compare the interactomes of TRY and CPC on a global scale. To this end, I aimed at applying two different experimental strategies: firstly, a conventional co-immunoprecipitation approach coupled to mass spectrometry and secondly, the above-mentioned TurboID proximity labeling technique. The latter work would require the establishment of the method in our laboratory and to generate a vector system suitable for the study of interacting proteins by TurboID.

4.3 Results

4.3.1 Generating vectors for the rapid assembly of TurboID constructs for proximity labeling in plants

NB: this section is based on a paper I co-authored:
K. Goslin and A. Finocchio and F. Wellmer (2021) *A golden gate-based plasmid library for the rapid assembly of biotin ligase constructs for proximity labelling* bioRxiv

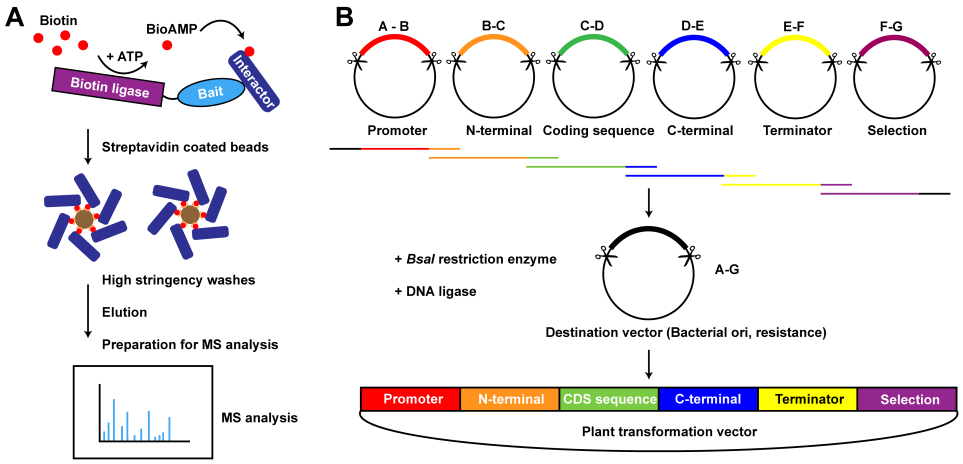


Figure 4.4: From Goslin et al. [2021]: (**A**) Activity of a promiscuous biotin ligase results in the generation and release of reactive bioAMP and the biotinylation of proximal proteins. Biotinylated proteins can be isolated by affinity purification using streptavidin coated beads. High stringency washes are performed to remove background. Following elution of proteins from the beads peptides can be identified by mass spectrometry. (**B**) The Green Gate system of cloning utilises a Golden Gate assembly approach. Entry vectors are first generated that contain single elements flanked by BsaI recognition sites and one of seven different overhangs (A-G). Upon digestion with BsaI individual elements are released from the entry vectors. Annealing of complementary overhangs and T4 ligase catalyzed DNA ligation results in the assembly of a final expression construct. Panel adapted from [Lampropoulos et al. 2013].

In order to generate expression vectors for the analysis of the TRY and CPC interactomes, I made use of the GreenGate cloning assembly kit [Lampropoulos et al. 2013]. This is a set of plant-specific vectors that utilises the Golden Gate cloning method. In this method, entry vector level plasmids are first generated that contain single elements of the final desired expression construct. These elements are flanked by *Bsa*I recognition sites and one of seven different overhangs (**A**-**G**). Upon digestion with *Bsa*I, individual elements are released from the entry vectors. Annealing of complementary overhangs and T4 ligase catalyzed DNA ligation results in the assembly of a final expression construct. To facilitate the rapid cloning of TurboID expression constructs in the laboratory, my colleague Dr. Kevin Goslin and I constructed a set of entry level vectors (see Figure 4.4 or Figure 1 in Goslin et al. [2021]). The toolkit we generated contains the TurboID protein in different entry vectors, allowing them to be used in combination with the GreenGate system. To make the vectors more versatile, we fused different signals for intracellular localization to TurboID to target it to different compartments. These included the SV40 nuclear localization sequence (NLS), the HIV-1 nuclear export signal (NES), and the endoplasmic reticulum targeting signal (ER). Furthermore, we produced a parallel set of plasmids for MiniTurbo, a version of TurboID that is less efficient, but much smaller, and can be used in experiments where a bulkier tag could negatively affect the bait protein's functionality.

To confirm that the localization signals fused to the TurboID constructs resulted in the localization of the protein to the desired subcellular compartment, we transiently expressed the fusion proteins in tobacco leaves. These constructs comprised different TurboID-GFP constructs fused with localization signals for the nucleus, for nuclear export and for transport to the endoplasmic reticulum (ER). Protein expression was checked via immunoblotting and the protein's localization was determined using confocal imaging [Goslin et al. 2021]. Furthermore, we confirmed that TurboID fusion proteins generated with these elements were active by co-expressing AP3-GFP and PI-TurboID-3xMyc constructs in tobacco leaves followed by co-immunoprecipitation and Western blotting. AP3 and PI are obligate heterodimers in flowers and are therefore a good control for proximity labeling [Riechmann et al. 1996]. Both proteins were detected in the eluate fraction by αGFP and streptavidin-HRP antibodies, which suggests that the

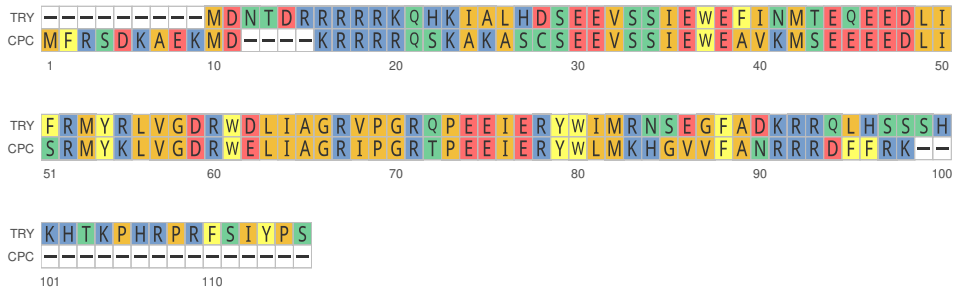
TurboID fusion protein is active and was able to biotinylate PI when it came in contact with AP3 [Goslin et al. 2021].

4.3.2 The R3-MYB family protein sequence

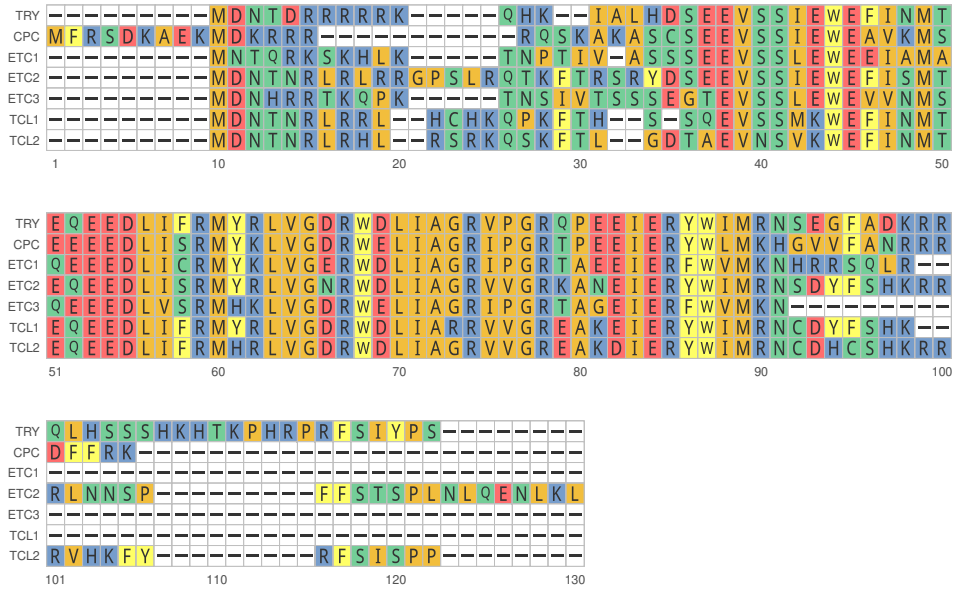
To gain a better understanding of TRY and CPC protein features, I conducted an analysis *in silico*. I downloaded the protein sequences for the seven R3-MYB genes in Arabidopsis using BioMart, a graphical interface provided by the Ensembl database [Kinsella et al. 2011]. I then aligned either the protein sequences of the whole R3-MYB family, or just TRY and CPC together. This highlighted conserved regions and differences in the alignments that informed the construct design and the interpretation of the IP-MS results.

The pairwise alignment between TRY and CPC (see Figure 4.5a) showed a highly conserved protein core between position 30 to 80, which hosts the single MYB domain (defined in PROSITE rule PRU00133). Around the MYB domain there are key differences in more unconserved regions of the proteins. CPC has a N-terminal extension of 9 amino acids, while TRY has a longer tail that extends the protein by 17 amino acids at the C-terminus. As described in section 4.1.2, the N-terminus of CPC hosts a peculiar motif that is likely to be involved in cell-to-cell movement and size exclusion limit modulation. This sequence is not present in TRY which instead has a longer C-terminal that forms a highly disordered tail, as shown in section 4.3.5. TRY's C-terminus has all the hallmarks of intrinsically disordered regions. TRY C-terminus contains three prolines in close proximity and a higher proportion of polar and charged amino acids, which in all probability make the peptide very unstable.

In a next step, I aligned the sequences of the whole R3-MYB family with the ClustalOmega algorithm (see Figure4.5b). The most conserved region was, unsurprisingly, the core MYB domain. CPC's N-terminal motif wasn't found in any of the other R3-MYBs and is unique to CPC. The disordered tail in TRY is not conserved in sequence but ETC2, and to a lesser extent TCL2, have similar peptide sequence at the C-terminus, with the same hallmarks of disordered peptides described above.



(a) Pairwise alignment between TRY and CPC amino acid sequences performed with the Needleman-Wunsch algorithm.



(b) Multiple sequences alignment between the components of the R3-MYB family amino acid sequences performed with the ClustalOmega algorithm.

Figure 4.5: Pairwise protein alignment of TRY and CPC (see Figure 4.5a), and multiple protein sequence alignment of the R3-MYB family (see Figure 4.5b).

4.3.3 R3-MYB family DNA sequence homology

I computed a similarity matrix, expressing the genetic distances in the R3-MYB family, from multiple sequence alignments of the DNA sequences (see Figure 4.6). DNA sequences mutate at a faster rate, which makes genetic distances more informative when calculating evolutionary divergence. For the alignment I ran the ClustalOmega algorithm on coding sequences for each representative gene model in the R3-MYB family. To better visu-

alise the evolutionary relationships, I computed a phylogenetic tree based on the matrix of genetic distances using the neighbor-joining method (see Figure 4.7) [Saitou and Nei 1987]

THe R3-MYB family is divided in two distinct clades. One clade contains the two closely related *ETC1* and *ETC3*, and *CPC*. The other members of the R3-MYB family *TRY*, *ETC2*, *TCL1* and *TCL2*, have greater evolutionary distances and fall in different clades. Interestingly, in agreement with the alignments in section 4.3.2, *TRY* and *CPC* have diverged considerably, while conserving the MYB domain.

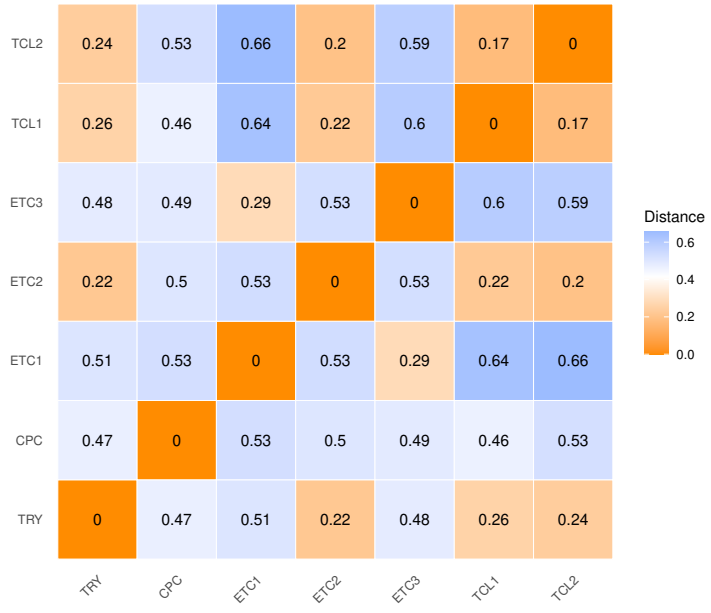


Figure 4.6: Heatmap of the genetic distances between the R3-MYB family members. Values go from 1.00, meaning most distant, to 0.00, the closest. The genetic distance is also represented with a color gradient, from orange, most distant, to light blue. The midpoint for the shading (white) doesn't consider the 0.00 value for self comparisons, to avoid skewing the color gradient and aid visualization.

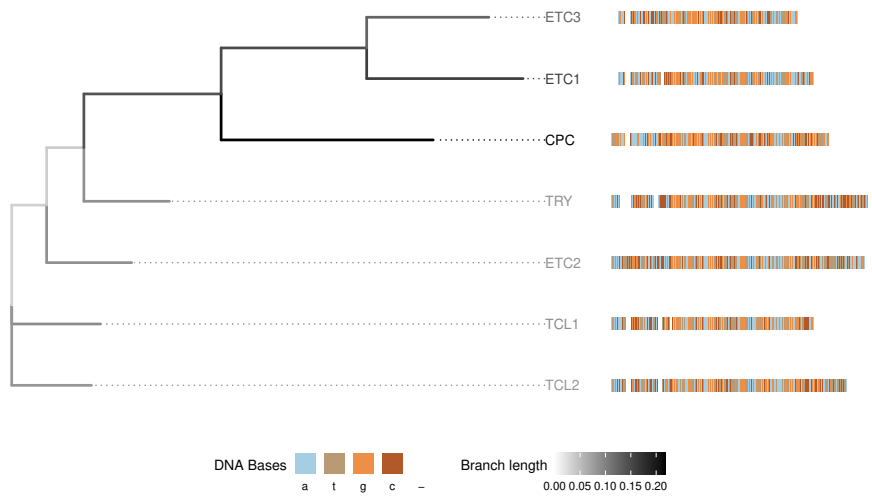


Figure 4.7: Phylogenetic tree representing the hierarchy evolutionary distances between the R3-MYB family members. On the right of the tree, a simplified representation of the multiple sequence alignment. Each DNA base is represented in a different color. The branch length represents the degree of evolutionary divergence.

4.3.4 Protein structure prediction

To evaluate how much the differences between TRY and CPC proteins contribute to the final folded protein, a prediction of the 3D structure of the folded proteins was obtained through the PHYRE^2 model (see section 2.26.4). Predicting protein folding from a sequence of amino acids is still an unsolved problem in computational biology due to the massive search space of all the possible conformations of any given protein.

In very general terms, tools like PHYRE^2 try to infer most of the structure by homology modeling, looking for previously known motifs to approximate the final structure and cut the computational complexity by orders of magnitude. This approach is not guaranteed to give accurate results, but is nevertheless useful to highlight conformational features that could guide the discovery of new protein functions.

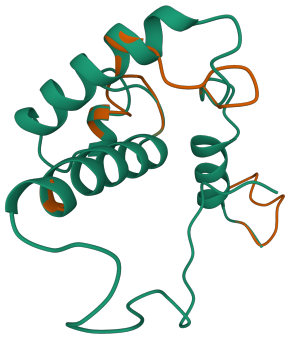
Recently, a new protein folding algorithm has taken the world by storm, AlphaFold [Jumper et al. 2021]. AlphaFold uses new machine learning models based on transformers, and a combination of supervised and self-learning, to achieve, on average, the best protein folding predictions of any tool. I used

AlphaFold to predict the structures of TRY and CPC, which were almost identical to those predicted by PHYRE^2. This is perhaps not surprising as both tools try to match known structures in a database, to the protein sequence that is provided by the user. AlphaFold does that far better than any other model, but the fundamental mechanism is not different. It follows that these tools can not predict structures that have not yet been described (i.e., protein motifs that are difficult to crystallize). Interestingly however, AlphaFold gives a confidence score to each amino acid in the structure, which is very low for TRY's disordered tail, probably because it is almost impossible to crystallize unfolded motifs and is therefore missing from protein structure databases.

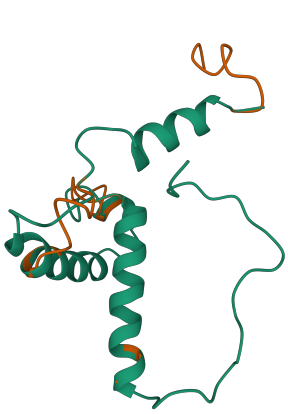
In a next step, I superimposed the 3D structure predictions for TRY and CPC by minimizing the distance between their respective amino acid chains (see Figure 4.8). The structural differences I outlined in section 4.3.2 can be recognised in the folded proteins. CPC unique N-terminus is clearly visible in orange, facing upwards in Figure 4.8b and 4.8c. The motif assumes a helical shape but the confidence in the prediction is not very high. Low confidence can be due to both intrinsic disorder or to the motif not being well described in the available databases. TRY's disordered tail is colored in green and it is downward facing in Figure 4.8b and 4.8c. PHYRE^2 predicts a highly disordered structure for TRY C-terminus, due to the hallmarks of intrinsically disordered peptides described in section 4.3.2. Prolines form a rigid -65 °angle that increase structural stiffness and forces abrupt corners along the amino acid chain. Prolines are often found far apart or at the beginning of α-helices or β-sheets, where a rigid angle helps protein folding. The effect of multiple prolines in close proximity can be seen on the TRY tail where the three rigid angles constrict the protein into an unfolded state. A relatively low concentration of hydrophobic residues further exacerbates this instability by lowering the cost of maintaining the chain in an unfolded state, by increasing affinity with water.

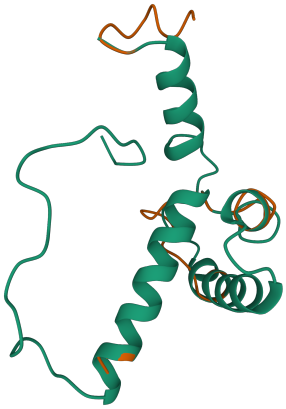
4.3.5 Disorder analysis

In light of the results of the *in silico* analysis, I wanted to further investigate the disordered structures present in some of the R3-MYBs. To this end, I calculated per amino acid disorder scores using the IUPred2A and ANCHOR algorithms (see section 2.26.5). Both methods compute the estimate disorder



(a)





(b) (c)

Figure 4.8: Superimposition of the predicted folded TRY and CPC proteins. The folded protein prediction, was obtained using the PHYRE2 web portal [Kelley et al. 2015]. TRY is colored in green and CPC is colored in orange. The figures offer different angles of the same structure. Figure 4.8b and 4.8c show the proteins N-terminus and C-terminus while panel 4.8a has a better angle of the proteins' core.

at each amino acid position, by sliding a fixed size window on the sequence, but they calculate different disorder scores:

IUPred2A identifies intrinsically disordered regions (IDPs) that are unlikely to reach a well-defined tertiary structure under native conditions.

ANCHOR estimates how much the gain in stability will be for segments of the protein that can't fold on their own and would benefit from the

interaction with a globular partner. It relies on IUPred2A's algorithm to define disordered regions.

Both algorithms consider a stretch of amino acids scoring above 0.5 as being disordered. IUPred2A computed disorder scores above 0.5 for ETC1, ETC3, TRY and CPC N-termini. Strikingly, TRY was the only member of the family that had high disorder scores assigned to the C-terminus (see Figure 4.9a). ANCHOR scores also had a sharp increase towards TRY's C-terminus, while the other R3-MYBs had very low scores in the region (see Figure 4.9b). ETC3 was the only notable expection, with disorder scores briefly peaking above 0.5. As expected, the conserved R3-MYB motif had low disorder in all family members.

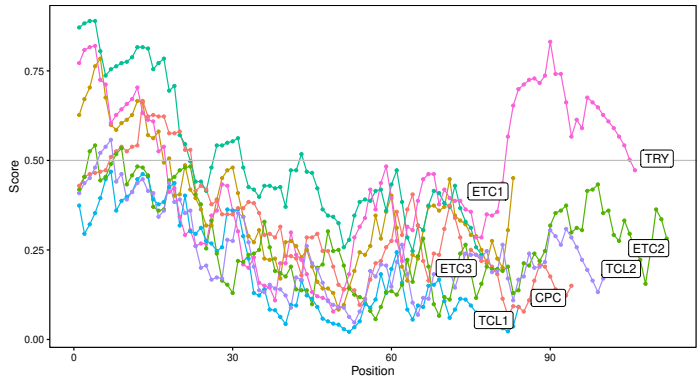
I then more closely analysed the protein features identified in sections 4.3.2 and 4.3.4. IUPred2A scores shown in Figure 4.10a are low for CPC's N-terminal motif but increased above threshold when the sequence is more conserved across the R3-MYBs N-termini. The TRY C-terminus had high ANCHOR scores, suggesting that the intrinsic disorder might cause the protein to be more prone to find an interaction partner that helps stabilise its structure.

4.3.6 Constructs design

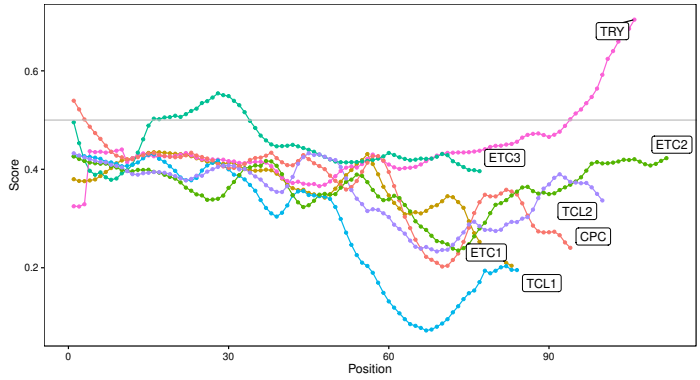
As stated in the project aims, the main objective of this work was to explore the interactome around TRY and CPC using the TurboID proximity labeling method followed by IP-MS. To complement the TurboID labeling experiment, I also generated stable lines expressing GFP-tagged versions of TRY and CPC which would facilitate IP-MS experiments. I generated a total of six new constructs for IP-MS (see section 2.6 and Table 4.1).

Table 4.1: Overview of the constructs used for the IP-MS experiments

Name	Construct	Description
$_{pTRY}TRY^{TID}$	*pTRY::gTRY-TurboID-CMyc*	TRY TurboID
$_{pTRY}GFP^{TID}$	*pTRY::NLS-GFP-TurboID-CMyc*	TRY TurboID control
$_{pCPC}CPC^{TID}$	*pCPC::gCPC-TurboID-CMyc*	CPC TurboID
$_{pCPC}GFP^{TID}$	*pCPC::NLS-GFP-TurboID-CMyc*	CPC TurboID control
$_{pTRY}TRY^{GFP}$	*pTRY::gTRY-GFP*	TRY GFP IP
$_{pCPC}CPC^{GFP}$	*pCPC::gCPC-GFP*	CPC GFP IP



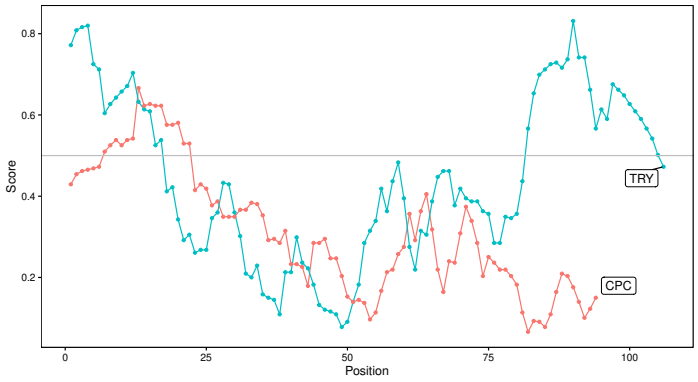
(a) Disorder scores calculated with the IUPred2A algorithm.



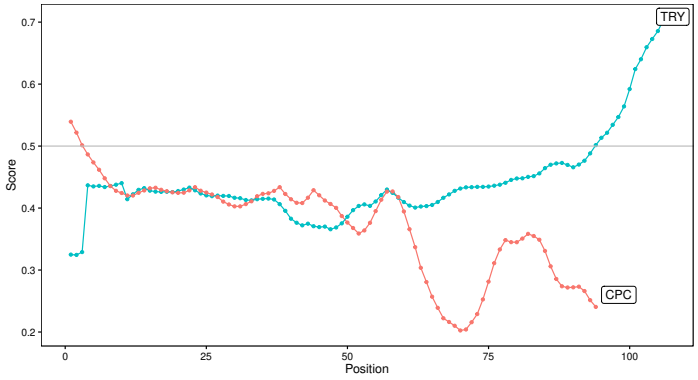
(b) Disorder scores calculated with the ANCHOR algorithm.

Figure 4.9: Disorder analysis of the 7 members of the R3MYB protein family. In both figures the x-axis indicates the amino acid position for each protein. The disorder score is plotted on the y-axis. The grey horizontal line at 0.5 marks the disorder score threshold. A threshold of 0.5 is the default value implementation of both IUPred2A and ANCHOR.

First, I generated two constructs for biotin labeling for TRY and CPC, respectively. One construct had TRY or CPC fused to the TurboID biotinylase described in section 4.1.3. The other construct, which served as a control, had a GFP protein fused to TurboID. TurboID fused with a generic protein like GFP allowed the identification of non-specific labeling events. I cloned the TurboID sequence with an additional single cMyc tag for Western blotting. Since TRY and CPC can travel to the nucleus as well as to neighboring cells, I added an NLS signal to the control constructs, to capture specific interactions in the nucleus. I further generated two constructs for



(a) Disorder scores calculated with the IUPred algorithm.



(b) Disorder scores calculated with the ANCHOR algorithm.

Figure 4.10: Same as Figure 4.9 but only TRY and CPC are shown to reduce clutter. In both panels the x-axis indicates the amino acid position for each protein. The disorder score is plotted on the y-axis. The grey horizontal line at 0.5 marks the disorder score threshold. A threshold of 0.5 is the default value implementation of both IUPred2A and ANCHOR.

the GFP-IP experiments. TRY and CPC, respectively, were tagged with a GFP at the C-terminus. As a control for the GFP-IP I used a *p35S::GFP* mutant line which constitutively overexpressed free GFP (see section 2.30). To generate these constructs I used the entire genomic sequence of *TRY* and *CPC*, together with 1,500 bp of their respective promoters. The *RBCS* terminator from the GreenGate system was used as the terminator sequence for all the constructs. I then transformed the four vectors into the Arabidopsis accession Col-0 using the floral dip method (see section 2.9).

4.3.7 Tobacco infiltration with TurboID constructs

Before transforming Arabidopsis, I evaluated the TurboID constructs' activity and stability in leaves of tobacco (*Nicothiana benthamiana*) plants by transiently expressing the constructs following agroinfiltration (see section 2.10). The leaves were infiltrated either with a solution containing Agrobacterium carrying one of the constructs or with a mock solution. Seventy-two hours after the infiltration, the leaves were collected, flash frozen and ground in liquid nitrogen. Total protein extracts were then obtained by boiling the ground tissue in Laemmli buffer and used for Western blotting with the αMyc antibody. Using this approach I was not able to detect any of the constructs (see Figure 4.11).

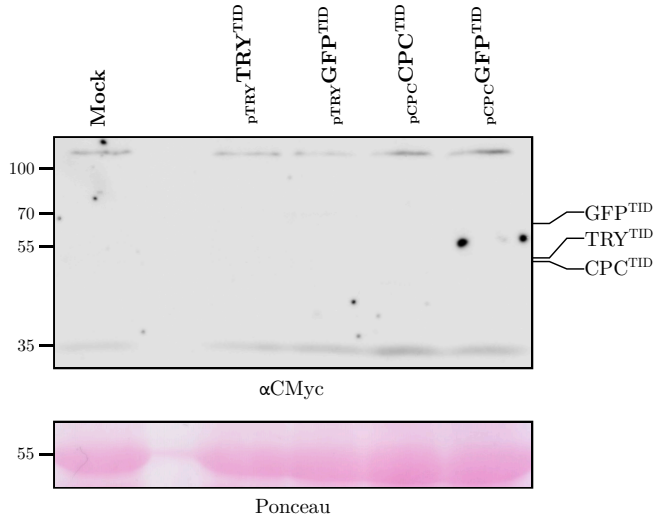


Figure 4.11: Western blots on total protein extracts from tobacco leaves infiltrated with TRY^{TID}, CPC^{TID}, GFP^{TID} constructs. The expected migration height of the CMyc-tagged proteins is shown on the right. Expected weights: TRY^{TID}: 50 kDa, CPC^{TID}: 49 kDa, GFP^{TID}: 65 kDa. The molecular weight size markers in kDa are indicated on the left of each panel. The protein loadings are shown in the Ponceau staining.

Failing to detect the constructs in tobacco does not necessarily mean that they are not expressed. For example, the protein concentration could be too low in the extract for Western blotting as a result of using endogenous promoters rather than strong constitutive promoters. Because the TRY and CPC fusion proteins were translocated to the nucleus, I repeated the experiment on nuclear extracts (see section 2.16), to concentrate the proteins and

improve the signal in the Western blot. Concentrating the proteins allowed a more effective Western blotting and all the expected proteins were detected (see Figure 4.12). $_{pCPC}GFP^{TID}$ was the only protein that produced a fainter band and also another band slightly below the expected weight. The lower band could have been a non-specific target of the αMyc antibody, since it was also present in the lane loaded with the mock control sample.

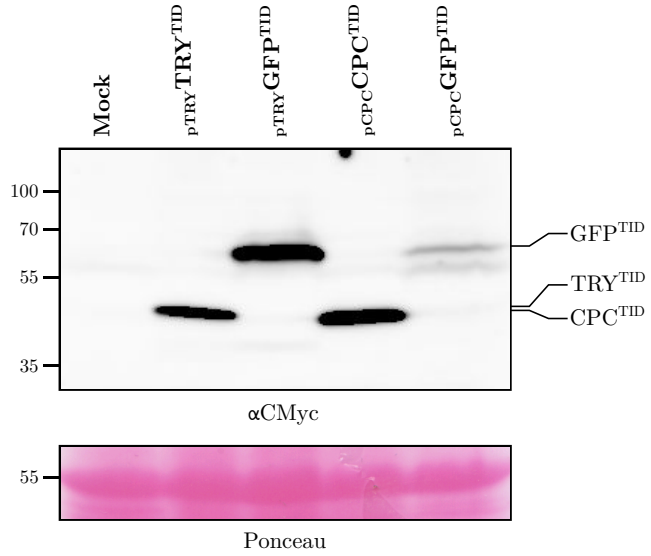


Figure 4.12: Results of Western blots on nuclear protein extracts from tobacco leaves infiltrated with TRY^{TID}, CPC^{TID}, GFP^{TID} constructs. The expected migration height of the CMyc-tagged proteins is shown on the right. Expected weights: TRY^{TID}: 50 kDa, CPC^{TID}: 49 kDa, GFP^{TID}: 65 kDa. The molecular weight size markers in kDa are indicated on the left of each panel. The loadings are shown in the Ponceau staining.

After verifying the viability of the constructs in tobacco, I analysed the TurboID biotinylation activity. For Western blotting I used a streptavidin-HRP antibody that directly binds the biotin on labeled proteins. I wanted to check that the TurboID was still able to label proteins *in planta* and measure the effect of the biotin treatment versus the level of biotin that is already present in plant cells. Seventy-two hours after infiltrating tobacco leaves with Agrobacterium, I cut the leaves into smaller pieces. I then infiltrated them under vacuum for an hour, with either water or a 250 µM biotin solution.

There was an evident difference in the amount of biotin detected between mock-treated and biotin-treated samples (see Figure 4.13). In the mock-treated samples, only proteins that are normally biotinylated in cells,

such as biotin-dependent carboxylases, would have been detected. In contrast, when TurboID is present, the number and intensity of bands increased dramatically, showing that TurboID was able to label a wide variety of proteins. The labeling happened already in water-treated leaves, probably because TurboID was able to utilise the endogenous biotin that is present in the cell. Interestingly, the labeling efficiency increased from water-treated to biotin-treated leaves in the presence of TRY^{TID} and CPC^{TID}, but the difference was less pronounced in the two lanes loaded with the GFP^{TID} controls (Figure 4.13).

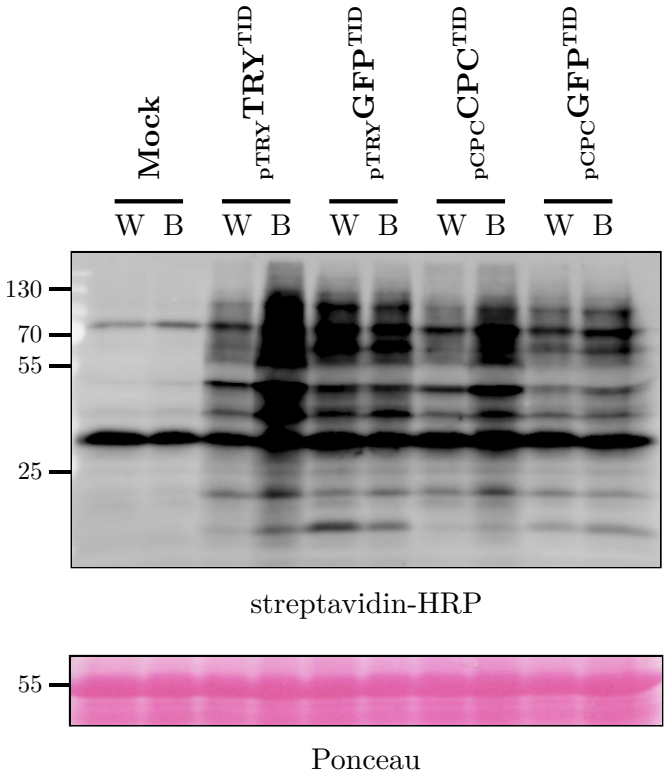


Figure 4.13: Results of Western blots on total protein extracts from tobacco leaves infiltrated with TRY^{TID}, CPC^{TID}, GFP^{TID} constructs. Water and biotin treatments are labeled with a W or a B respectively. The expected migration weights in kDa are indicated on the left of each panel. The protein loading is shown by the Ponceau staining.

4.3.8 Selecting Arabidopsis lines for IP-MS

I transformed the constructs into Arabidopsis wild-type plants (accession: Col-0) and also into *try* and *cpc* loss-of-function mutants. All vectors contained a resistance gene against the herbicide ammonium glufosinate ("BASTA"), which was then used to select for transformants. I verified the presence of the construct in the transformants by PCR and then collected seeds from each plant separately. The progeny was again selected using BASTA and the seeds of 5-6 of the remaining plants were collected. By looking at the proportion of survivors to BASTA selection in the progenies, I identified which parent was homozygous for the transgene.

Subsequently, I tested for TurboID protein expression by Western blotting and I used the lines that showed the highest expression in the IP-MS experiments. To this end, I sampled leaves from 14-day-old seedlings which I then flash froze in liquid nitrogen. I ground the frozen leaves in a mortar and used the total protein fraction for Western blotting (see Figure 4.14). To increase sensitivity, I used the *try* $_{\text{pTRY}}\text{TRY}^{\text{TID}}$ and *cpc* $_{\text{pCPC}}\text{CPC}^{\text{TID}}$ lines, where the endogenous *TRY* and *CPC* genes are not inactive.

Due to the weak signal in the *cpc* $_{\text{pCPC}}\text{CPC}^{\text{TID}}$ line, I repeated the Western blot with the same αMyc antibody but exposed the membrane for a much longer period of time. With the longer exposure, I was able to detect all of the constructs at the expected migration distances in all the lines I selected for the IP-MS experiments.

I also tested leaves from homozygous $_{\text{pTRY}}\text{TRY}^{\text{GFP}}$ and $_{\text{pCPC}}\text{CPC}^{\text{GFP}}$ lines by Western blotting using a monoclonal αGFP antibody (see Figure 4.16). Bands of the expected weight were visible in all lines apart from $_{\text{pTRY}}\text{TRY}^{\text{GFP}}$ line 4.4, which did not express TRY^{GFP} and was discarded. Two bands corresponding to lower weight products were also detected in all the lines, which probably resulted from cleavage at the proteins' C-termini. The lower of the two cleavage products was likely GFP which migrates at around 27.1 kDa. The higher band could also have been GFP but generated from a cleavage further upstream from the TRY or CPC C-termini.

Interestingly, it appeared that TRY^{GFP} protein was cleaved at a higher rate than CPC^{GFP}, which could be due to TRY's longer, disorder tail. *try* $_{\text{pTRY}}\text{TRY}^{\text{GFP}}$ line 2.2 and *cpc* $_{\text{pCPC}}\text{CPC}^{\text{GFP}}$ line 4.4 were used for the GFP-IP followed by mass spectrometry.

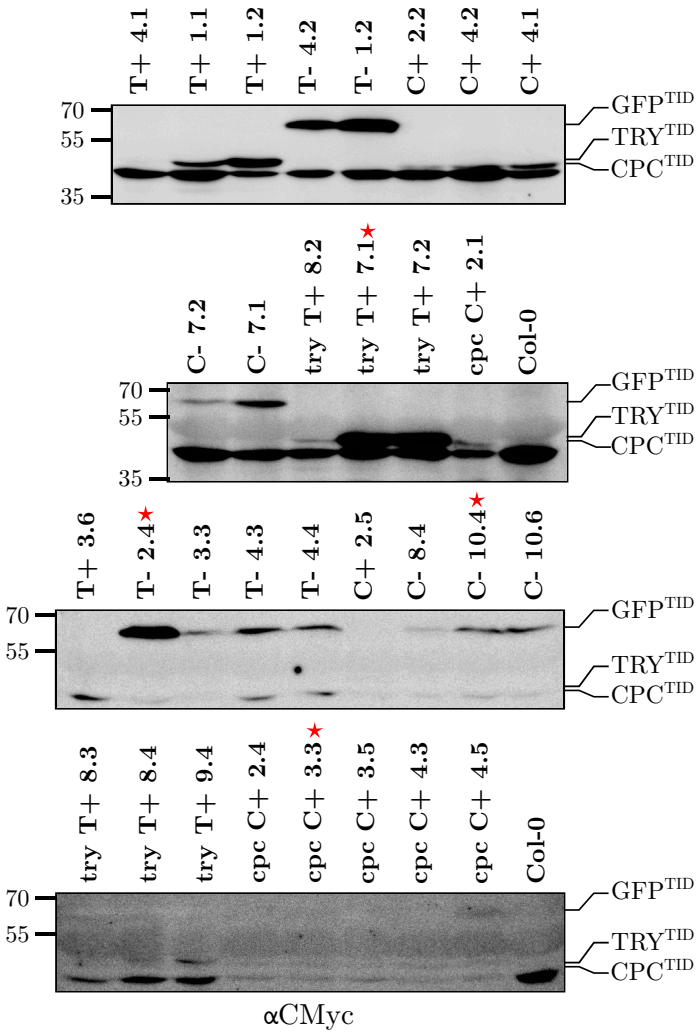


Figure 4.14: Results of Western blots on leaves of 14-day-old seedlings from TurboID lines. Leaves from 14-days-old seedlings were frozen in liquid nitrogen, ground in a mortar, and the total protein fraction was extracted boiling in Laemmli buffer for western blotting. A red star highlights the lines that were used for the IP-MS experiment. 126 µg were loaded for each sample after amido black quantification. The expected product sizes are indicated on the right side of the panels. The molecular weight size markers in kDa are indicated on the left of each panel. T+: *try* $_{pTRY}$TRY^{TID}, T-: $_{pTRY}$GFP^{TID}, C+: *cpc* $_{pCPC}$CPC^{TID}, C-: $_{pCPC}$GFP^{TID}.

4.3.9 Rescuing *try* and *cpc* mutants

To assess whether TRY^{GFP}, CPC^{GFP}, TRY^{TID} and CPC^{TID} are fully functional in Arabidopsis, I compared the phenotypes of the plants carrying the

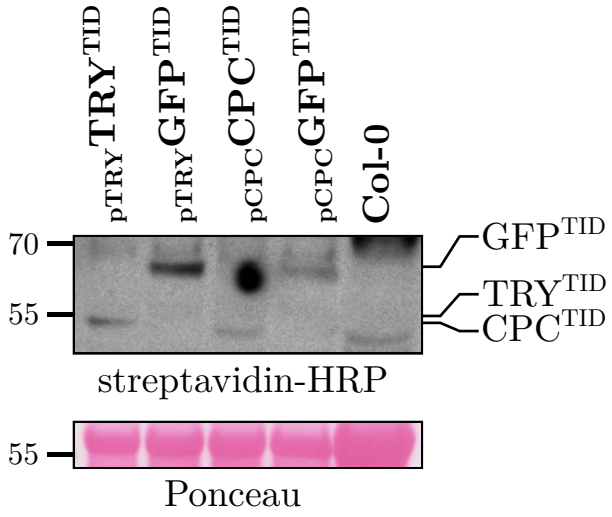


Figure 4.15: Results of Western blotting on leaves of 14-day-old seedlings from TurboID lines. Leaves were frozen in liquid nitrogen, ground in a mortar, and the total protein fraction was extracted boiling in Laemmli buffer. The Western blot membrane was exposed for three hours. Approximately equal protein loading in each lane was confirmed by Ponceau staining.

above-mentioned constructs to those of wild-type, *try* and *cpc* mutant plants. The 4$^{\text{th}}$ leaf was sampled from eight 14-day-old seedlings and photographed with the stereo microscope using a ruler for size reference. The photos were processed in `ImageJ` where I counted the number of trichomes and trichome clusters, and measured the leaf area. Due to the peculiar phenotypes of *try* and *cpc* phenotypes (see section 4.1.2), I used trichome density to verify the rescue of *cpc*, and the density of trichome clusters to verify the rescue of *try*.

I found that the constructs were functional in the trichome development pathway and able to repress trichome development in *try* and *cpc* mutant backgrounds (see Figure 4.17). However, I observed a dosage effect in that the rescue did not lead to the same density of trichomes or trichome clusters observed in the wild type. Instead, expressing the constructs in *try* and *cpc* mutant backgrounds led to leaves that were almost completely glabrous. TRY$^{\text{GFP}}$ and TRY$^{\text{TID}}$ expression prevented the formation of trichome clusters on *try* leaves and led to a visible decrease in the number of branched trichomes. Plants expressing CPC$^{\text{GFP}}$ and CPC$^{\text{TID}}$ in a *cpc* mutant background had fully glabrous leaves.

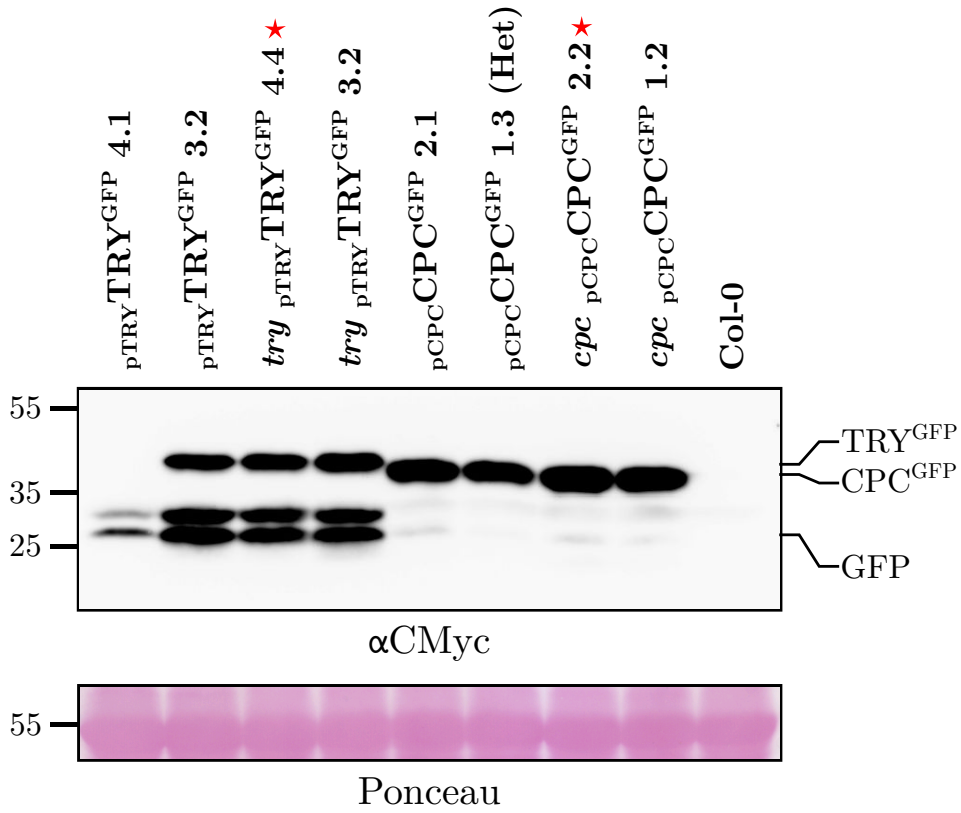
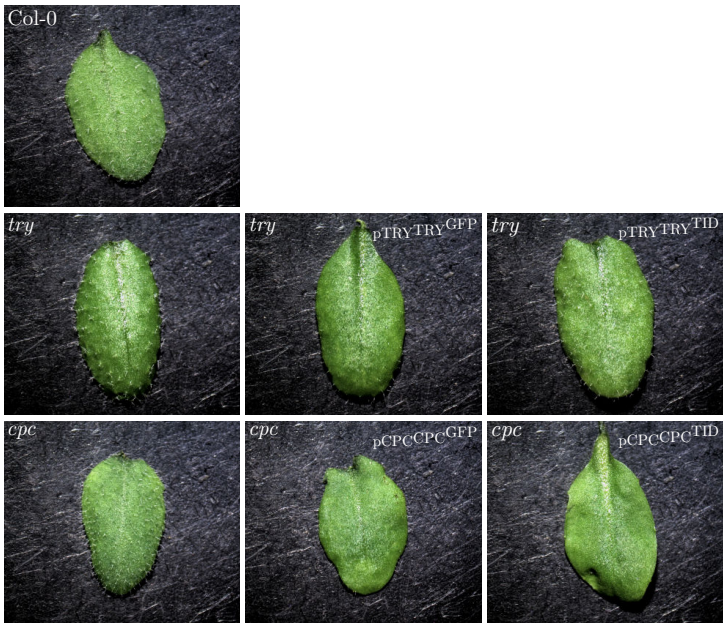


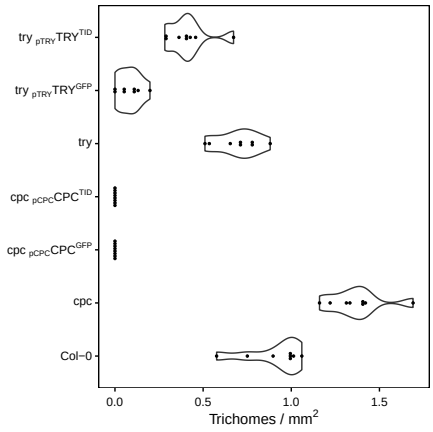
Figure 4.16: Results of Western blots on leaves of 14-day-old seedlings from GFP lines. Leaves from 14-days-old seedlings were frozen in liquid nitrogen, ground in a mortar, and the total protein fraction was extracted boiling in Laemmli buffer for western blotting. A red star highlights the lines that were used for the IP-MS experiment. 128 µg were loaded for each sample after amido black quantification. The expected product sizes are indicated on the right side of the panels, together with GFP, a possible cleavage product. The molecular weight size markers in kDa are indicated on the left of each panel. Approximately equal protein loading in each lane was confirmed by Ponceau staining.

4.3.10 IP-MS of biotin-labeled TRY and CPC

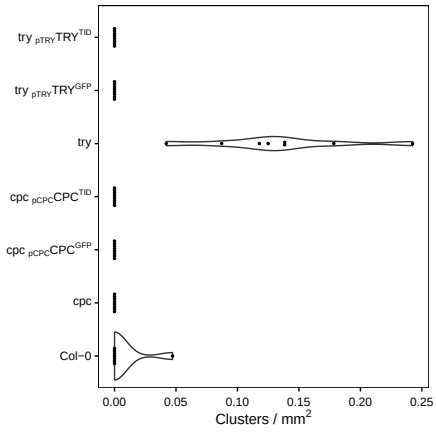
As outlined in section 4.2, the main aim of the project was to take a snapshot of the interactome around TRY and CPC, using both co-immunoprecipitation or biotin labeling, followed by mass spectrometry. To keep the two experiments comparable I used the same plant growth conditions, treatments and tissue collection for both experiments (see Figure 4.18). I grew ∼90 seedlings for each line (*try* $_{\text{pTRY}}$TRY$^{\text{TID}}$, $_{\text{pTRY}}$GFP$^{\text{TID}}$, *cpc* $_{\text{pCPC}}$CPC$^{\text{TID}}$



(a) Comparison of the 4th leaves. The genotype of each line is shown in the top left corner while the transformed constructs are shown in the top right corner of the photos.



(b) Trichome density expressed in mm^2.



(c) Cluster density expressed in mm^2.

Figure 4.17: Overview of the trichome phenotypes of *try* and *cpc* knock out mutants and the respective rescue lines expressing tagged (GFP or TurboID) TRY or CPC. Arabidopsis plants of accession Col-0 were used as wild-type control. The numbers plotted in Figures 4.17b and 4.17c were obtained by counting the number of trichomes, trichome clusters and measuring the area of the 4th leaf of 8 plants.

and $_{pCPC}GFP^{TID}$) on three $\frac{1}{2}$ MS plates on top of a Whatman paper disk, which prevented damaging the roots during the transfer to liquid MS medium. After 14 days, the seedlings were carefully transferred into a liquid MS medium. The MS medium was then supplemented with biotin to trigger the TurboID labeling process. To study the effect of proteasome inhibition, I added either MG132, a potent reversible inhibitor of proteasome activity dissolved in DMSO, or as a control, an equivalent volume of DMSO to the medium. The seedlings were left shaking in the growth room for 3 hours and then the labeling reaction was immediately quenched by 3 washes in cold water. The washes also helped to remove the excess biotin left on the seedlings. The seedlings were then quickly dried on tissue paper, divided in equal amounts for 3 replicates, flash frozen in liquid nitrogen and stored at -80°C.

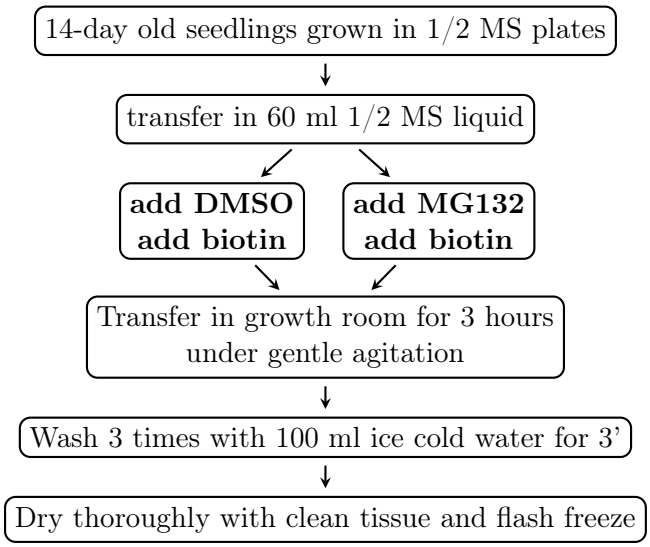


Figure 4.18: Experimental design for the streptavidin beads IP experiment.

The protocol for affinity purification with streptavidin beads and trypsin digestion for mass spectrometry was adapted from Mair et al. [2019] and is described in section 2.18. I eluted the proteins from 5% of the beads for Western blotting with a streptavidin-HRP antibody before trypsinisation and mass spectrometry. The eluates resulted in a number of separate bands, which confirmed the presence of biotin labeled proteins on the beads (see Figure 4.19). Notably, bands corresponding to the predicted constructs weights are visible in each lane, presumably as a consequence of self-biotinylation by

the TurboID tag.

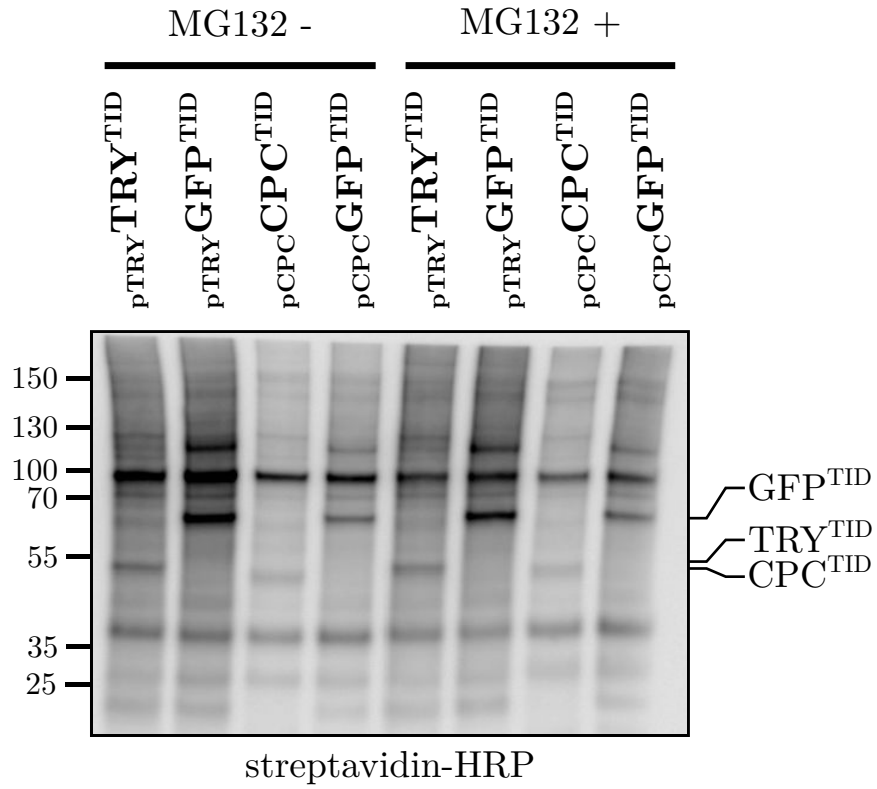


Figure 4.19: Results of Western blotting on proteins affinity-purified from TurboID lines for mass spectrometry using streptavidin beads. The samples shown correspond to the third replica. The treatment (MG132 or mock) is shown on top of the sample names. The expected product sizes are indicated on the right side of the panels. The molecular weight size markers in kDa are indicated on the left of each panel.

4.3.11 IP-MS on GFP-tagged TRY and CPC

To evaluate the performance of the immunoprecipitation of TRY^{GFP} and CPC^{GFP}, I performed Western blotting and silver staining to detect interactors that had have been co-immunoprecipitated.

I used collected, flash froze and ground leaf material from 14-days-old seedlings. I then used ∼500 mg of frozen tissue as starting material for the immunoprecipitation (see section 2.17). I retained samples from three key steps of the protocol to assess the efficiency immunoprecipitation. The first set of samples was collected from the supernatant after the wheel incubation

(the input), a second set was collected from the supernatant after incubation with αGFP beads (the unbound fraction) and a third set after eluting the proteins from the beads (the eluate). The samples were then split in two and used for both Western blotting with αGFP monoclonal antibodies and a silver staining. I used leaves from a *35S::GFP* ($_{p35s}$GFP) line as a control.

Free GFP was detected in $_{p35s}$GFP sample in the input and eluate, but not in the unbound fraction (see Figure 4.20). This indicated that the beads were able to bind all the GFP present in the input. TRY^{GFP} and CPC^{GFP} could only be detected after elution, forming a band of the expected size and two more bands at around the size of GFP, likely from cleaved products. This is very similar to what I already described in section 4.3.8. TRY^{GFP} was cleaved at a higher rate than CPC^{GFP} as shown by higher intensity of the cleaved product bands respect to the band that corresponds to the uncleaved protein.

I wasn't able to detect any proteins in the eluates by gel electrophoresis and silver staining (see Figure 4.20). The silver staining worked in principle, as indicated by the strong staining of the input and unbound fractions, suggesting that the sensitivity of the staining was not sufficient to detect the proteins present at low concentrations in the eluates. At the same time, free GFP should have probably been detected by the staining because of its high level of expression. Although I could not directly verify presence of interactors in the eluate fraction by silver staining, I considered the results of the Western blotting sufficient to proceed with the IP-MS experiment.

To collect samples for the GFP immunoprecipitation and mass spectrometry, I followed the same steps as described in Figure 4.2 in order to keep the conditions similar to those used for the biotin labeling experiments with TurboID. The only difference was that the seedlings were not biotin treated. The IP, followed by trypsinisation, was performed using GFP-Trap agarose beads, as per the manufacturer's protocol (see section 2.17). I kept 5% of the GFP-Trap agarose beads to evaluate the efficiency of the immunoprecipitation before proceeding with trypsinization and mass spectrometry. Before the final PBS wash, these beads were boiled in Laemmli buffer to elute the proteins, which were then analysed by Western blotting using a monoclonal αGFP antibody (see Figure 4.21).

As already described above (see Figures 4.16 and 4.20) I found evidence of GFP cleavage in the samples. TRY^{GFP} seemed to be cleaved at a higher

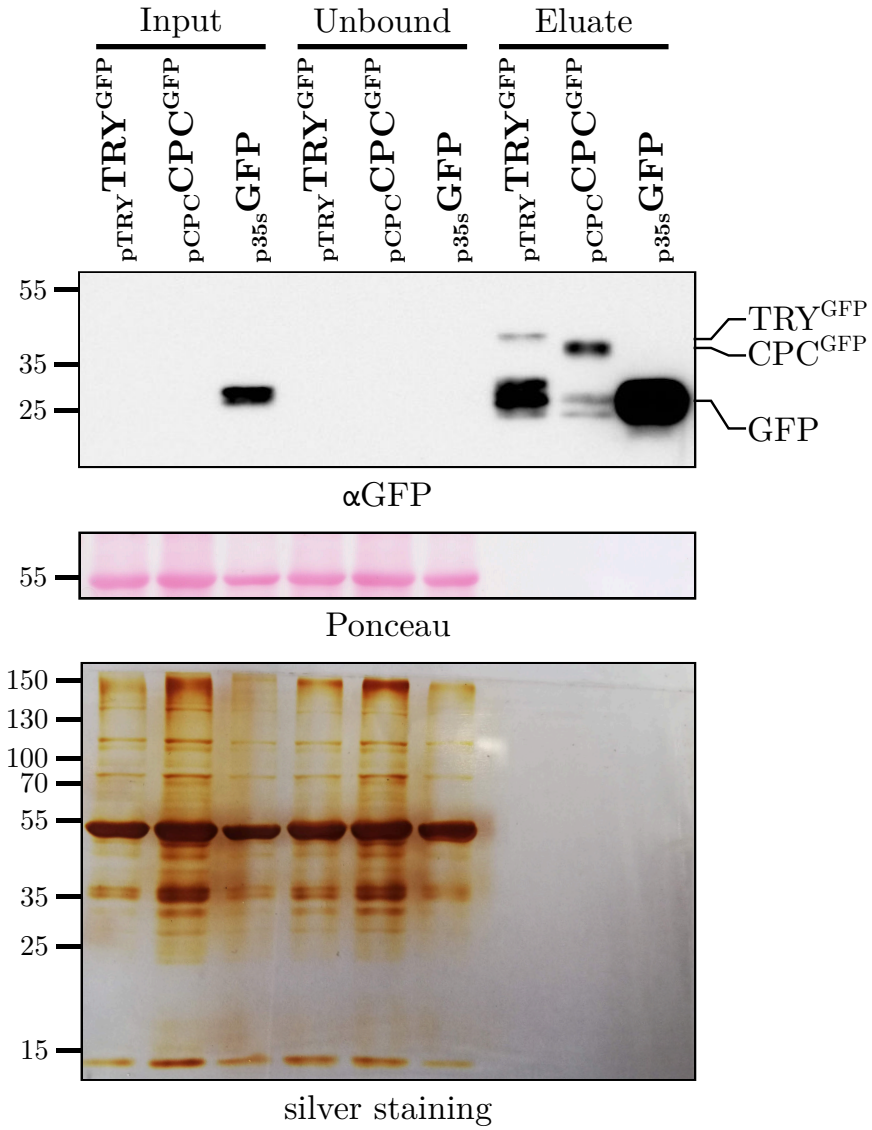


Figure 4.20: Efficiency of the GFP immunoprecipitation protocol. The Western blot with an αGFP monoclonal antibody is shown on top of the figure. The protein loadings are shown below the Western blot by the Ponceau staining. The result of the silver staining is shown on the bottom of the figure. The molecular weight size markers in kDa are indicated on the left. Expected product sizes for each construct are shown on the right-hand side of the blot.

rate than CPC^{GFP} and had one strong band at around the size of GFP (27.1 kDa) and a fainter band right above (~30 kDa). CPC^{GFP} was cleaved at a much lower rate, with the full construct being the most abundant protein. Interestingly, the 30 kDa cleaved product above GFP weight identified

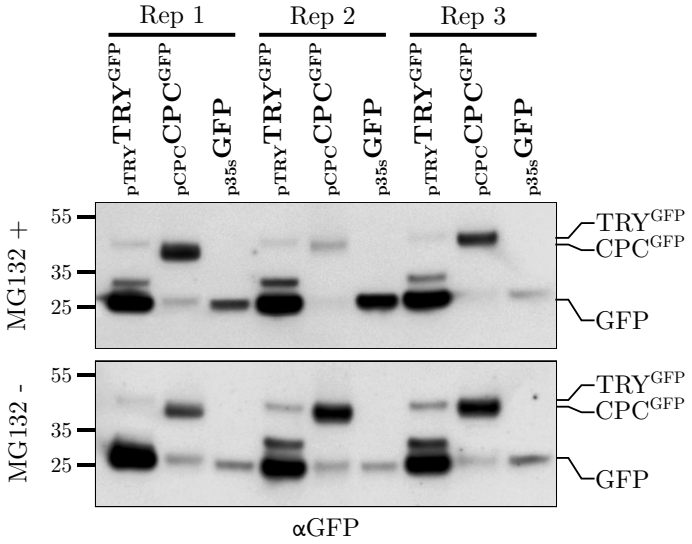


Figure 4.21: Western blot on proteins co-immunoprecipitated from GFP-tagged lines for mass spectrometry using GFP-Trap agarose beads. The expected product sizes are indicated on the right side of the panels. The molecular weight size markers in kDa are indicated on the left of each panel.

in $_{pTRY}TRY^{GFP}$ was not present in any of the $_{pCPC}CPC^{GFP}$ samples. The $_{p35s}GFP$ control produced a clear GFP band at the predicted size. The Western blot didn't show any difference between MG132 and DMSO treatments.

4.3.12 Results of Mass Spectrometry analyses

In order to generate the peptide samples required for mass-spectrometry analysis I carried out the trypsinization of samples in two batches. This was required because the trypsinization protocols differ between co-immunoprecipitated and biotin labeled proteins. Protein-protein interactions must be carefully preserved in co-immunoprecipitated samples because interactors are not attached directly to the beads, but are bound to the bait protein. When the interactors are biotin labeled, they are directly bound to the streptavidin beads by a much stronger bond. It has been shown that streptavidin beads require two trypsinization steps to achieve optimal results [Mair et al. 2019]. Proteins can be labeled at multiple sites and free biotin might still be present in solution. This makes the beads prone to clumping which can be mitigated by increasing the trypsinization time.

I sent the trypsinized samples to Dr. Kieran Wynne at the proteomics fa-

cility in University College Dublin who kindly performed the mass spectrometry quantification (see section 2.20) and provided me with a list of protein hits called by the software `MaxQuant`. Each protein hit had an associated label-free quantification (LFQ) value, which is a measure of the abundance of the detected protein in each sample. To find proteins that were enriched in the samples containing the bait proteins (TRY or CPC) compared to the control, I used the `R` package `DEP`. Using `DEP` I was able to obtain the lists of enriched proteins for TRY and CPC, which are putative interactors. The relationship between protein enrichment and significance (expressed terms of q-values) is summarised in volcano plots (see Figures 4.22 and 4.23). As expected, the bait proteins TRY and CPC were in most experiments among the most highly enriched and significant protein hits, indicating that they have been successfully recovered.

When compared to the results of the TurboID experiments, which yielded in excess of 1,000 candidate interactors, the GFP-IP samples resulted in relatively few enriched proteins and the enrichment values for them had much higher variance. As previously discussed (see section 4.1.3), the TurboID labeling system is highly reactive and so a higher sensitivity should be expected in this case. The extremes in the number of significantly enriched proteins identified were DMSO-treated CPC^{GFP} samples, where no significant hits were found at all (see Figure 4.23a), and TRY^{TID} samples treated with MG132, which yielded ∼1,300 enriched proteins.

Of the known components of the trichome initiation complex described in section 4.1.1, only TTG1 showed a significant enrichment in the TRY^{TID} biotin pulldown, when proteasome activity was inhibited. Most protein hits were common between the samples treated and untreated with MG132. The GENERAL REGULATORY FACTOR 1 (GRF1), GRF2 and GRF9 proteins are part of the 14-3-3 protein family, which are involved in regulation of "client" proteins. 14-3-3 proteins bind phosphoproteins to effect changes the client functionality. The binding effect can range from the regulation of enzymatic activity to protection from degradation. Some 14-3-3 proteins have been shown to localise to trichome nuclei [Paul et al. 2005]. Multiple GRF proteins were biotin labeled via TRY and CPC and GRF3 is one of the few proteins that also co-immunoprecipitate with CPC after MG132 treatment.

A large number of WD40 proteins, such as TTG1, was biotin labeled in

Table 4.2: Summary of the putative protein interactors described in this section

	This work		STRINGR	
Protein Name	TRY	CPC	TRY	CPC
TRY	Yes	No	Yes	No
CPC	No	Yes	Yes	Yes
GL3	No	No	Yes	Yes
EGL3	No	No	Yes	Yes
ATMYC1	Yes	Yes	Yes	Yes
MYB66	No	No	No	Yes
TT8	No	No	No	Yes
TTG1	Yes	No	No	No
GRF1-3-9	Yes	Yes	No	No
BCHC2	Yes	No	No	No
SPIRRIG	Yes	Yes	No	No
SCAR1-2	Yes	Yes	No	No
WAVE2-5	Yes	Yes	No	No
SCL1	Yes	Yes	No	No
SCL5-14	No	Yes	No	No

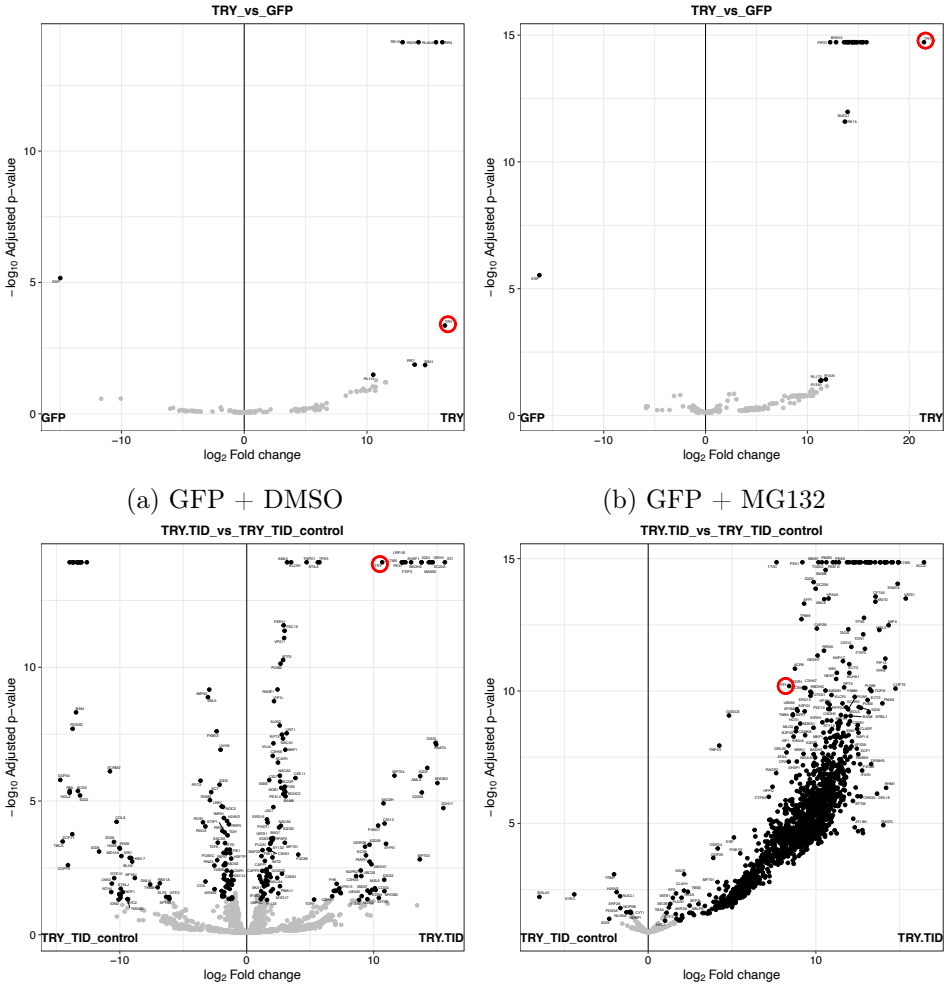
Note:
The table lists the putative protein-protein interactions described in this section together with experimentally validated physical interactions listed in the STRINGR database.

the MG132-treated samples possibly through interaction with TRY or CPC. WD40 repeats are often found following a BEACH domain, which is common in proteins involved in membrane trafficking. Two highly homologous BEACH/WD40 proteins, BEACH-DOMAIN HOMOLOG C 2 (BCHC2) and SPIRRIG (SPI), were both detected in the biotin pulldowns as putative TRY interactors. SPI has previously been shown to be involved in trichome and root hair development. Interestingly, *spi* mutants have distorted trichomes, less lobing of epidermal pavement cells, disconnected epidermal cells on various organs, and shorter root hairs [Steffens et al. 2015; Chin et al. 2021; Saedler et al. 2009].

There were a number of proteins that were retrieved only in biotin pull downs on samples treated with MG132. The inhibition of the proteasome

allowed the detection of interactors belonging to the SCAR/WAVE protein family. Specifically, labeled proteins like SCAR1, SCAR2, WAVE2 and WAVE5 were enriched in both TRY and CPC samples. The SCAR/WAVE protein complex mediates actin nucleation in a localised polar manner, driving directional growth, and cell shape changes. It appears possible that proteins like SPI could interact with the SCAR/WAVE complex and other BEACH/WD40 proteins to regulate trichome development [Chin et al. 2021].

Other interesting protein groups, which were only enriched in MG132 treated samples, included SCARECROW-LIKE (SCL) proteins, members of the JUMONJI (JMJ) family of chromatin remodellers and a large number of proteins involved in the ubiquitin pathway and protein degradation. SCARECROW (SCR) is a well known regulator of root hair patterning and radial organisation of the root. SCR restricts the movement of SHORTROOT (SHR), which is required for the asymmetric cell division responsible for formation of ground tissue, to the endodermis [Koizumi et al. 2012]. Among the members of the JMJ family that were enriched in the biotin pull down, JMJ29, was previously associated with trichome development regulation. JMJ29 regulates trichome development by demethylating lysine 9 on histone H3 on the GL3 locus, with a consequent increase in GL3 expression, increasing trichome density [Hung et al. 2020].



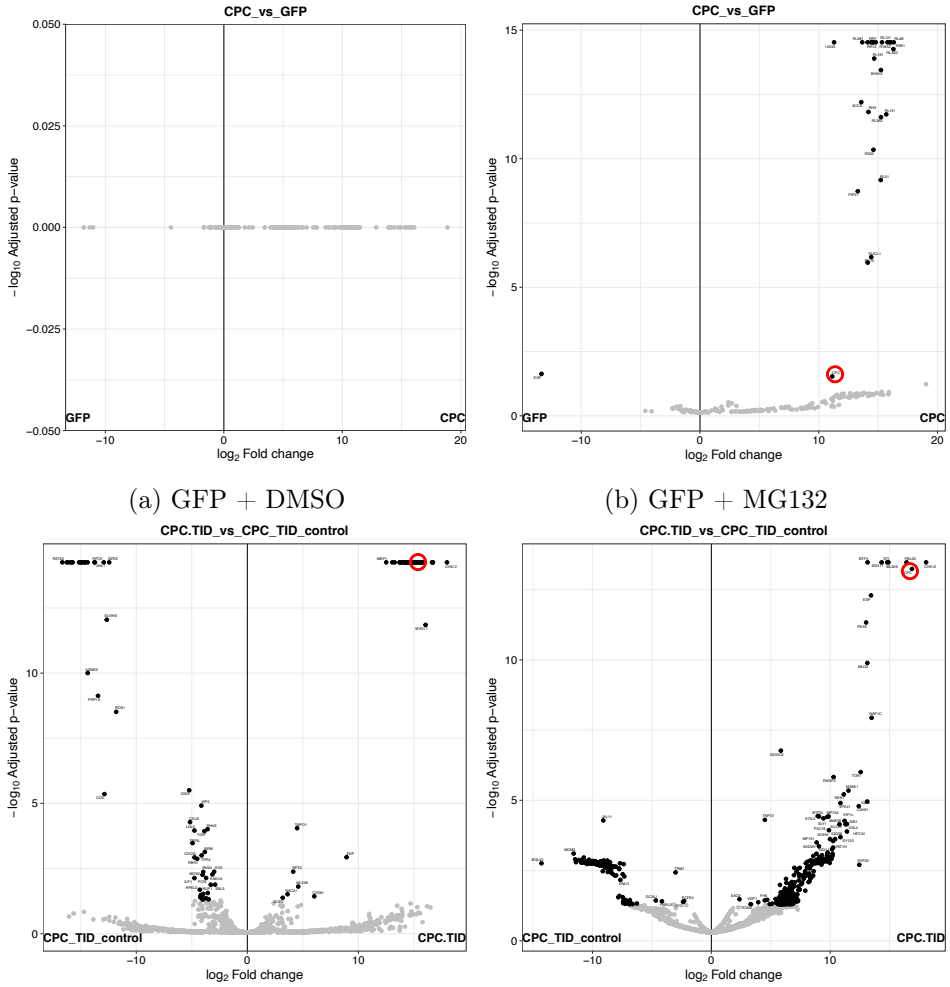
(a) GFP + DMSO

(b) GFP + MG132

(c) Biotin + DMSO

(d) Biotin + MG132

Figure 4.22: Volcano plots summarising the results of the proteomic analyses for TRY. In all figures, the y-axis represents the $-log_{10}$-transformed adjusted p-value, while the x-axis shows the log_2-trasformed fold changes. The y-axis is truncated at 15 to help visualisation. The fold change is measured by calculating the ratio between the LFQ value in the datasets obtained from TRY lines and the LFQ values of the respective controls. Putative TRY interactors have a positive fold change and are found in the right half of the volcano plot. The two top panels represent the results of the co-immunoprecipitation experiments with GFP-tagged lines while the two bottom panels show the results from the biotin labeling experiments with TurboID lines. The results from DMSO-treated plants are shown in the two left panels, while the two panels on the right show the results from MG132-treated plants. The bait proteins are circled in red (TRY or CPC).



(a) GFP + DMSO

(b) GFP + MG132

(c) Biotin + DMSO

(d) Biotin + MG132

Figure 4.23: Volcano plots summarising the results of the proteomic analyses for CPC. In all figures, the y-axis represents the $-log_{10}$-transformed adjusted p-value, while the x-axis shows the log_2-trasformed fold changes. The y-axis is truncated at 15 to help visualisation. The fold change is measured by calculating the ratio between the LFQ value in the datasets obtained from CPC lines and the LFQ values of the respective controls. Putative CPC interactors have a positive fold change and are found in the right half of the volcano plot. The two top panels represent the results of the co-immunoprecipitation experiments with GFP-tagged lines while the two bottom panels show the results of the results from the biotin labeling experiments with TurboID lines. The results from DMSO-treated plants are shown in the two left panels, while the two panels on the right show the results from MG132-treated plants.

4.4 Discussion

Trichome development is controlled by a regulatory complex that is composed of the proteins GL1, TTG1 and GL3/EGL3. It has been shown that the pattern of trichome initiation on leaves is dependent on the concentration of TTG1 and on the activities of the repressor proteins TRY and CPC. An important part of the regulatory mechanism for trichome iniation appears to be the movement of these proteins from cell to cell, resulting in differences in their concentration across developing leaves. How this intracellular movement of regulatory proteins is controlled is currently unknown. Furthermore, the different phenotypes of *try* and *cpc* mutants (as outlined in section 4.1.2) suggest that these closely related R3-MYB proteins have partially distinct functions in the control of trichome iniation. However, the molecular basis of these different activities is not well understood. To address these open questions, I adopted a proteomics approach to characterize the interactomes of TRY and CPC and to identify candidates for proteins involved in the regulation of TRY/CPC activity. To this end, I employed a conventional IP-MS approach as well as proximity labeling using the TurboID biotinylase [Mair et al. 2019]. For the latter work, I first had to establish the technique in the laboratory and design a set of plasmids that can be used for the construction of TurboID vectors before embarking on the proteomics analysis. This vector set has already been made available to other researchers interested in the TurboID method. As part of this project, I also analysed the R3-MYB family of trichome repressors with different software tools in silico, to unravel structural differences and the phylogenesis of its members.

4.4.1 Structural differences of TRY and CPC

TRY and CPC are small proteins of less than 120 amino acids. They share a conserved core region, where the MYB domain resides, but exhibit marked differences at both their N- and C-terminal ends (see Figure 4.5 and 4.8).

CPC contains a short domain in its N-terminus which has previously been connected with cell-to-cell movement of the protein in roots where it is involved in controlling the iniation of root hairs [Kurata et al. 2005]. As I showed by aligning the amino acid sequences of the members of the R3-MYB family, this domain is missing not only from TRY, but also from all the other trichome repressors. On the other hand, TRY has a C-terminal tail which

is rich in both prolines and in polar amino acid residues and is thus predicted to be highly disordered in its 3D structure. Interestingly, in addition to being assigned high scores for structural disorder, the C-terminal tail of TRY received high ANCHOR scores, suggesting that it could be stabilised by the interaction with a globular partner. What could be the function of such an interaction? While the C-terminal tail of TRY is not conserved in sequence among the R3-MYB family, ETC2 and TCL2 have a similar enrichment of prolines and in polar amino acid residues in their C-termini (Figure 4.10 and 4.9), suggesting that TRY shares an unordered domain with other trichome repressors. The substitution of amino acid residues at putative phosphorylation sites in the C-termini of TRY and ETC2 have been shown to prolong their half-lifes and increase their accumulation in roots [Yamada et al. 2018]. Thus, these domains may primarily be involved in controlling the stability of the proteins. In such a scenario, an unordered domain would be recognized by the protein degradation machinery, while stabilization of the tail through interactions with another protein would protect it from degradation. In agreement with this idea, I observed during the biochemical work I conducted on TRY that is it highly unstable (see below). It will be interesting to see whether any of the candidates for interactors identified by the proteomics experiments I did can stabilise TRY protein when co-expressed.

4.4.2 Design and validation of constructs for proteomics analyses

The intrinsic disorder and instability of the TRY protein posed many difficulties for the proteomics analyses I conducted. To circumvent these problems, I used TurboID biotin proximity labeling coupled with proteasome inhibition with good results. For the constructs I generated, I attached to GFP or TurboID to the C-terminiof TRY and CPC. This choice was informed by published data on the R3-MYBs proteins describing the successful of C-terminal tags [Yamada et al. 2018; Tominaga-Wada and Wada 2017; Kurata et al. 2005; Weisburd et al.]. I also reasoned that placing a tag at the N-terminus may affect the functionality especially in the case of CPC proteins due to its proximity to a domain likely involved in mediating protein movement (see preceding section).

Working with a state-of-the-art technique such as TurboID had its chal-

lenges. One of the main problems was protein detection, particularly in the case of the fusion proteins containing TurboID. TurboID was fused with a C-terminal tag to facilitate detection by Western blotting. In hindsight, the choice of a single cMyc tag wasn't optimal and a tag composed of multiple copies of cMyc would have probably led to better detectability. Alternatively, the V5 tag described in Chapter 3.3.4 could have been used as it is small in size and can detected with high sensitivity by Western blotting. The high turnover of TRY but also of CPC and their localised expression in leaf trichomes further exacerbated the protein detection problem (see Figure 4.14). To circumvent the detection problem, I either concentrated the proteins before Western blotting by first isolating nuclei, where TRY and CPC predominantly localise to, or by using seedlings to take advantage of the broader expression of TRY and CPC in young roots and leaves [Kurata et al. 2005].

4.4.3 Proteasome inhibition and biotin labeling enhances IP-MS sensitivity

It has been shown previously that GL3 protein turnover is regulated via proteasomal degradation [Patra et al. 2013]. Similar results have been obtained for members of the R3-MYB family, including for TRY and ETC2, where it has been found that their concentration increases in the presence of proteasome inhibitors [Tominaga-Wada and Wada 2017]. As discussed in section 4.3.5, the instability of TRY and ETC2 is likely due to their unordered C-terminal tails. In fact, removing these regions has been shown to increase their accumulation. In contrast, CPC appears less prone to degradation, which allows it to accumulate to relatively high levels in roots.

Because the finding discussed above suggested that several trichome regulators are subject to degradation and have a rapid turnover, I conducted proteomics experiments with and without the proteasome inhibitor MG132. Use of this inhibitor should promote the accumulation of TRY (and possibly also of CPC) in the plant and thus, increase the concentration of the bait proteins in the samples. This would then lead to higher sensitivity of the mass spectrometry measurements due to the greater protein abundance and to the recovery of more interacting proteins. As shown in section 4.3.12, MG132 treatment resulted in a stark difference in the number of interactors retrieved, both in the GFP pulldowns and the TurboID experiments, when

compared to the experiments where protein degradation was not impaired. Perhaps unsurprisingly, many proteins involved in the ubiquitination system for protein degradation were identified for TRY in the presence of MG132. In this case, TRY – instead of being rapidly degraded by the proteasone – may have been in contact with the blocked protein degradation for long enough to efficiently biotinylate these proteins.

4.4.4 A comparison of two approaches used for the identification of TRY/CPC interactors

I implemented the new TurboID biotin labeling system to use it alongside standard co-immunoprecipitation techniques. This would not only allow me to increase the confidence in candidate interactors recoverd by both methods but also to directly compare the two approaches. As discussed above, biotin labeling by TurboID greatly increased the retrieval of putative interactors, with both TRY and CPC baits. In contrast, the results from the co-immunoprecipitation experiments appeared suboptimal. The IP conditions had already a low stringency, which should help with the sensitivity, but the mass spectrometry hits were few and dominated by ribosomal proteins. To optimise the protocol further, the peptide trypsinisation step was performed on the beads, as it was for the biotin labeled samples. On bead digestions are a delicate process that can lead to the loss of interactors and suboptimal results. The biotin pulldown protocol is more permissive, since the labeled proteins are directly and solidly bound to streptavidin beads. In the case of the GFP protocol, the interactors are bound to the beads indirectly via the protein of interest and the GFP, possibly by weak interactions. This increases the chances of technical errors.

In proximity labeling experiment, which have high recall but generally a lower precision, it's important to discriminate between bona fide and non-specific interactors. I controlled for non-specific interactions by comparing the hits in samples with TRY and CPC baits, with the hits in samples a TurboID bait. The samples were analysed in triplicate to quantify the reproducibility of the observations.

4.4.5 Future perspectives

To date, no datasets from IP-MS experiments have been published for R3-MYBs proteins or for other components of the trichome initiation complex. Thus, while the work described here is entirely novel, I could not compare my datasets against the results from other studies on related proteins. Such comparisons could have been helpful in narrowing down the lists to proteins to those that are likely to be associated with the bait proteins. At the same time, as discussed in section 4.3.12, these lists do contain promising candidates, e.g. proteins that could be involved in the degradation of TRY, which is thought to play an important role in the regulation of its activity. In order to experimentally validate putative interactors that were found by mass spectrometry, protein pairs will be studied in the future using approaches such as yeast two-hybrid (Y2H) screens and bimolecular fluorescence complementation (BiFC). A third, more modern approach would be Fluorescence Resonance Energy Transfer (FRET) but this method requires specialised hardware and more involved computational approaches. I plan to use Y2H and BiFC to validate the interactions, which enough evidence to proceed with further experiments and publication. One advantage is that all the necessary entry constructs were already cloned and tested in the lab, and they are ready for use, which will save valuable time.

4.5 Conclusion and final remarks

The scope of this work goes beyond the regulation of trichome growth. As I already introduced in section 4.1.2, root hair development shares many of the same developmental regulators with trichome development. Plant roots, and especially root hairs, are key for nutrient uptake and have been one of the main areas of study to improve the yield and sustainability of crops. Interestingly, trichomes have also even been linked to stomatal development in recent years [Simon et al. 2020]. Trichome development is tightly regulated at the protein level, by balancing the concentration of key members of the trichome development complex. In this chapter, I explored the interactome around TRY and CPC, the two main repressors of trichome development, using a combination of co-immunoprecipitation, proximity labeling with high-throughput mass spectrometry. To date, this is the only proteomics experiment of such scale on trichome development factors. Working

with TRY and CPC is challenging, due to their instability, as it was also shown in this work. I was able to overcome this problem by inhibiting the proteasome, which helped to unravel a plethora of new candidate interactors. The exploratory analysis in this work will be the foundation for new projects in the lab, which will shed a new light on the factors that orchestrate the onset of new trichomes and root hairs.

Chapter 5

Identification of genes acting downstream of B and C function regulators at different stages of flower development

5.1 Introduction

As outlined in Chapter 1, flower development is orchestrated by a small number of master regulators, most of which belong to the family of MADS-domain transcription factors. Genome-wide localization studies for most of these regulatory proteins have been conducted in recent years, showing that each of these factors can bind to thousands of sites in the Arabidopsis genome [Pajoro et al. 2014; Wuest et al. 2012; O'Maoileidigh et al. 2013; Kaufmann et al. 2010]. As expected, the bound regions are strongly enriched for CArG-boxes sequences, the canonical binding sites of MADS-domain transcription factors [Aerts et al. 2018]. Surprisingly, in transcriptomics experiments, only a small subset of the genes bound by the floral homeotic factors responded to a perturbation of their activities [Wuest et al. 2012; O'Maoileidigh et al. 2013; Chen et al. 2018]. This limited overlap between the datasets from ChIP-seq and transcriptomics experiments could be due to many of the binding sites being outside the regulatory regions of genes. Alternatively, and perhaps more likely, these sites may be non-functional in the context of floral organ specification but functional when bound by other

MADS-domain transcription factors involved in processes such as the floral transition or fruit development[O'Maoileidigh et al. 2013; Goslin et al. 2017; Chen et al. 2018; Mendes et al. 2013; Smaczniak et al. 2012b]. In this scenario, MADS-domain transcription factors would have similar genome-wide binding patterns with a regulatory output then depending on the "grammar" of motifs in a given promoter and the presence of other transcriptional regulators in a cell that may act cooperatively with the floral homeotic factors in the control of gene expression.

Another possibility is that the transcriptomics experiments mentioned above did not capture all the genes that can be regulated by the floral homeotic factors. In fact, most of these experiments were performed either on whole inflorescences (where older, more mature flowers contribute by far the most to the overall mRNA pool) or with flowers of a distinct developmental stage. Because it is known that the floral homeotic factors are expressed throughout most of floral organ morphogenesis [Ryan et al. 2015a] and carry out diverse functions during flower development, it is possible that some genes bound by them respond only at certain developmental stages or in specific tissues. It therefore seems important to better characterize the gene expression programmes that are controlled by the floral homeotic factors at different stages of flower development and ideally, also with spatial resolution. In fact, there is already some evidence that the floral homeotic factors control different sets of genes at different stages of flower development [Pajoro et al. 2014; Ryan et al. 2015a; Chen et al. 2018; O'Maoileidigh et al. 2013; Wuest et al. 2012]. However, these experiments were often hampered by the fact that a perturbation of floral homeotic factor function early on in flower development leads to a complete loss of certain organ types, precluding an analysis of their functions at more advanced stages. New approaches that allow a stage-specific perturbation of gene activities are therefore needed to analyse their function with stage-specific resolution. The development of lines that mediate the specific expression of artificial microRNAs (amiRNA) targeting the floral homeotic genes provided such a novel approach for transcriptome analysis [Thomson et al. 2017]. These lines are described in detail in section 5.1.1 below.

5.1.1 The role of *AP3* and *AG* in floral organ specification and development

According to the ABC model of floral organ specification (as outlined in section 1.2), the proteins AP3 and PI mediate B function, whereas AG carries out C function. The B function regulators act together with A function proteins to specificy petals and with AG to specify stamens. AG alone is sufficient to specify carpel identity [Coen and Meyerowitz 1991b; Bowman et al. 1989].

AP3 acts as an obligate heterodimer with PI; neither protein alone is able to bind CArG-box motifs by itself. As a consequence, AP3 and PI have largely overlapping sets of target sites in the genome and loss-of-function mutations in either gene results in very similar floral phenotypes, namely the replacement of petals and stamens by extra sepals and carpels, respectively [Riechmann et al. 1996; Goto and Meyerowitz 1994; Jack et al. 1994]. Their expression commences around stage 3 of flower development and continues in petal and stamen primodia for most of their development. During the course of flower development, AP3 and PI seem to regulate the expression of transcription factor-coding genes at the early stages, but their activities at later stages are currently not well understood [Wuest et al. 2012].

AG is a C function gene responsible for both the specification and development of the reproductive floral organs. In addition, it is involved in the formation of ovules and in the control of floral meristem identity where it terminates floral stem cells when all floral organ primordia have been initiated [Bowman et al. 1989; Ito et al. 2004; 2007]. Similar to *AP3/PI*, *AG* expression commenced around stage 3 and continues throughout stamen and carpel primordia for most of their development [Yanofsky et al. 1990]. *Ag* mutants completely lack stamens and carpels but form many extra sepals and petals. Because floral meristem termination is impaired in *ag* mutant plants, a "flower witin a flower" phenotype is observed with a new flower arising from the center of a flower that has already been formed.

As outlined in Chapter 1, the floral homeotic factors act in tetrameric complexes to control the gene expression programmes required for the formation of the different types of flower organs. AP3/PI and AG are part of the complex that regulates stamen development. In addition, AP3/PI are also part of the complex for petal development and AG is a component

of the complex for carpel formation. Given their partially shared activity, B and C function regulators should show some overlap in their genome-wide binding sites and the genes that act downstream of them. In fact, the genome-wide localization and transcriptomics experiments mentioned above identified such an overlap for early flower development. Whether this shared activity is retained at later floral stages is currently unknown.

5.2 Project aims

Although much work has been done in recent years to characterize the activities of the floral homeotic factors, the gene expression programmes they regulate are only partially known. As described above, this is especially true for more advanced stage of flower development that cannot be analysed through the use of floral homeotic mutants. Thus, new methods for stage-specific gene perturbations as well as multi-omics approaches are needed to further our understanding of the regulatory activities of the floral homeotic factor complexes during the course of flower development. To better understand the gene expression programmes acting downstream of the floral homeotic factors, this project aimed at employing the inducible amiRNA lines against *AP3* and *AG* described above in transcriptomics experiments to identify genes that are controlled by the corresponding transcription factors at different stages of flower development. The resulting datasets should then be analysed using bioinformatics methods to identify processes under control of the B and C function regulators during floral organ morphogenesis. Furthermore, the data obtained should be compared to published results from genome-wide localization experiments to distinguish direct targets from secondary response genes.

5.3 Results

To allow a stage-specific perturbation of *AP3* and *AG* activity during different stages of early flower development, I employed transgenic lines that had been developed by my colleague Dr. Bennett Thomson during his time in our laboratory (see his thesis for reference: [Thomson 2018]). These lines employ a dexamethasone (DEX) inducible promoter system to drive the expression of artificial microRNAs (amiRNAs) designed to target either *AG* (line denoted: AG^{amiRNA}) or *AP3* (line denoted: $AP3^{amiRNA}$), respectively.

To allow a stage-specific analysis of flower development, the inducible amiRNA constructs were transformed into a floral induction system background, similar to the one described in section 3.3.1. However, in this case, the induction of flower development was mediated by a fusion protein between AP1 and the ligand-binding domain of the mouse androgen receptor (denoted: AP1-AR), which can be activated by treating plants with the steroid hormone dihydrotestosterone (DHT) (see sections 2.30 and 2.11, respectively for details).

The combination of two activatable systems with different inducers allowed me to first induce flower development and subsequently, to knock down the activities of *AP3* and *AG* at different stages of early flower development. For the remainder of this chapter, when talking about "days after induction" or simply "induction", I am referring to the induction of the flower development by DHT treatment.

5.3.1 Experimental design for the knockdown experiments

To assess the function of B and C function regulators during early flower development, three time-points were chosen for the experiments, namely 3, 5 and 8 days after the induction of flower development by DHT treatment. These time-points corresponded approximately to stages 4, 6-7 and 9, respectively (floral stages according to Smyth et al. [1990]), representing the stage immediately following activation of the floral homeotic factors (stage 4); a stage where all four types of floral organs had been initated and specified and where the floral meristems had just been terminated (stages 6-7); and flowers at a more intermediate stage of development undergoing key events during floral organ maturation (stage 9). Thus, these stages covered the main processes known to be under control of the floral homeotic factors.

For the experiments, plants were grown as described in section 2.3. The inflorescence-like meristems of the plants were then treated with DHT to induce flower development. Subsequently, the plants were treated either with a solution containing DEX to induce amiRNA expression or with a mock solution (i.e., the solution did not contain dexamethasone) as a control. Because previous experiments [O'Maoileidigh et al. 2013] have shown that it takes several hours to efficiently silence the target transcripts after the induction of amiRNA expression and for the transcriptome to react to the silencing, the plants were treated with DEX at 2, 4 and 7 days and tissue was collected 24 hours later to result in the 3, 5 and 8 day time-points specified above. The knockdown experiments were conducted three times, resulting in three biologically independent sets of samples that could be used for analysis.

5.3.2 Determining the knockdown efficiencies in AG^{amiRNA} and $AP3^{amiRNA}$ lines

To confirm that the induction of amiRNA expression led to the intended reduction in target gene expression, I first assessed floral phenotypes by analysing mature flowers of DEX and mock-treated plants. To this end, AG^{amiRNA} and $AP3^{amiRNA}$ plants were treated following the same protocol described above and photos were taken 14 days after induction using a stereo microscope (Figure 5.2). Flowers of mock-treated plants of both AG^{amiRNA} and $AP3^{amiRNA}$ formed flowers that resembled those of the wild type, confirming that flower development had been properly induced by the DHT treatment that was used for the activation of the AP1-AR fusion protein, and that the inducible amiRNA constructs did not lead to an inactivation of their target genes in the absence of the inducer DEX. Flowers of DEX-treated plants showed phenotypic defects similar to those observed in *ap3* and *ag* mutants, respectively. However, they were not comparable in strength to the severe mutant phenotypes of null mutant alleles, where the genes are non-functional throughout flower development. In general, phenotypic defects were strongest in flowers of plants treated at the 3 DAI time-point and got progressively milder the later the treatment was performed. This was expected because in the earliest time-point, the gene knockdowns occurred at a time when floral organs are being specified, whereas in the later time-points, the key structures of the flower had already been established.

Flowers of AG^{amiRNA} plants treated with DEX were generally smaller

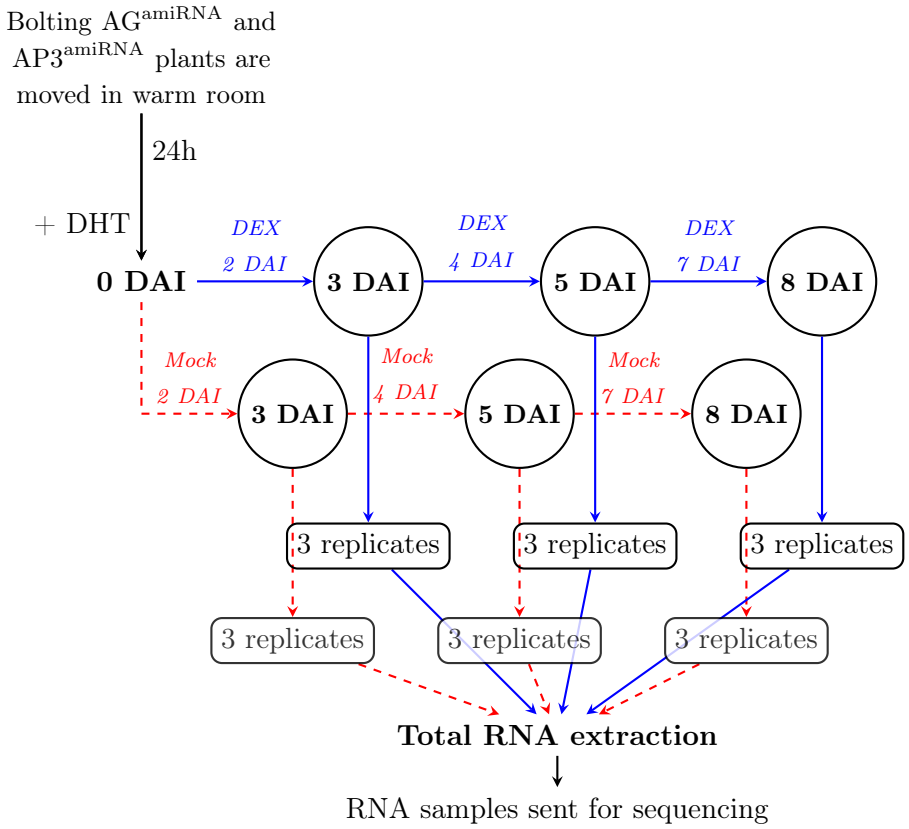


Figure 5.1: Experimental design for the *AG* and *AP3* knockdown experiments. Plants were grown in colder conditions to slow down the differentiation of the meristematic tissue in FIS plants. After bolting they were moved into a warmer room, with standard *Arabidopsis* growth conditions (see section 2.3). The plants were then treated with DHT and subsequently, at different time-points (as indicated) with a DEX containing solution or with a Mock solution. Tissue was then collected and processed for RNA-sequencing. DAI: Days After Induction; FIS: Floral Induction System. The blue, solid arraows refer to the DEX treated samples, while the dashed, red arrows refer to the mock treated samples.

than those of mock-treated control plants and showed a mild *ag* phenotype with supernumerary petals, underdeveloped stamens and underdeveloped gynoecia. In some cases, a weak flower-in-flower phenotype was visible in plants treated with DEX at 3 or 5 DAI. Floral buds treated at 5 DAI with DEX formed sepals and petals correctly but failed to develop proper gynoecia (see Figure 5.2). DEX treatment at 8 DAI had less severe effects on mature flowers: the overall size of the flower was smaller and carpels and stamens

were shorter and underdeveloped.

Most of the flowers from DEX-treated AP3$^{\text{amiRNA}}$ plants resembled those of *ap3* mutants. Flowers of plants treated with DEX at 3 DAI had carpeloid structures in place of petals and stamens were absent. In flowers of plants treated at 5 DAI, petals were severely underdeveloped; small, underdeveloped stamens could be observed in a minority of flowers. Flowers of plants treated at 8 DAI were still quite strongly affected with petals and stamens being much shorter than normal and partially underdeveloped (Figure 5.2).

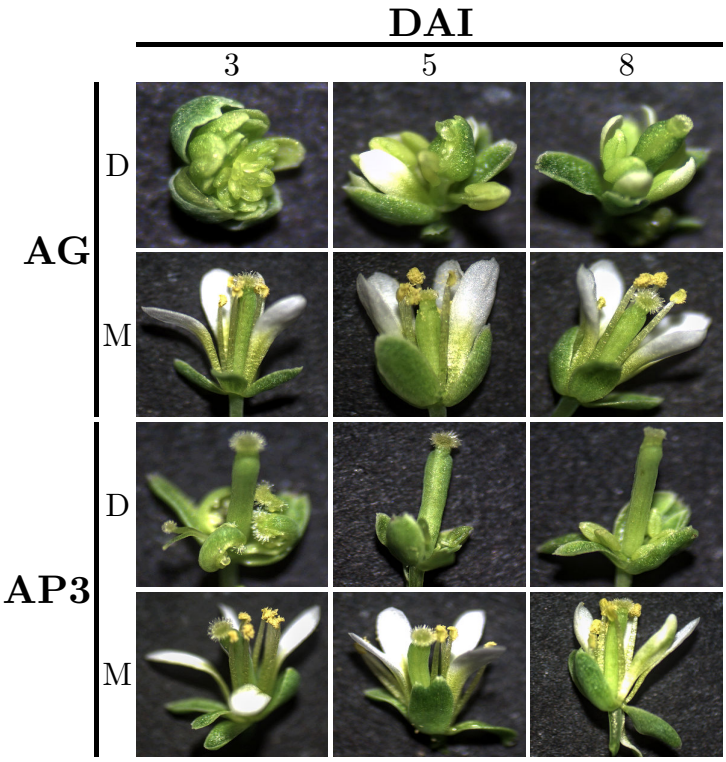


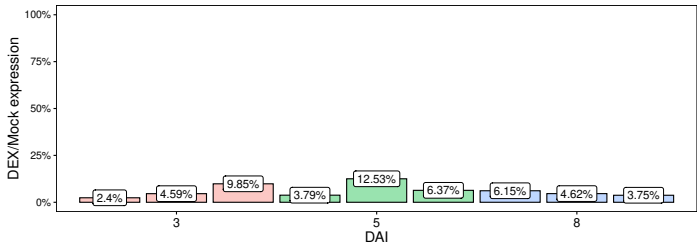
Figure 5.2: Phenotypic effects of gene knockdowns at different stages of flower development. Flowers of AG$^{\text{amiRNA}}$ (AG) and AG$^{\text{amiRNA}}$ (AP3) plants 14 days after the induction of flower development. The time-points for the treatment of plants with either a DEX-containing solution (D) or a mock solution (M) are indicated at the top above each column of images. As outlined in section 5.3.1, the treatments were performed 24 hours before the time-point indicated on top. DAI: Days after induction.

In summary, the phenotypes observed in these experiments suggested that *AG* and *AP3* transcript levels had been knocked-down following the induction of miRNA expression and that the degree of the knockdown was

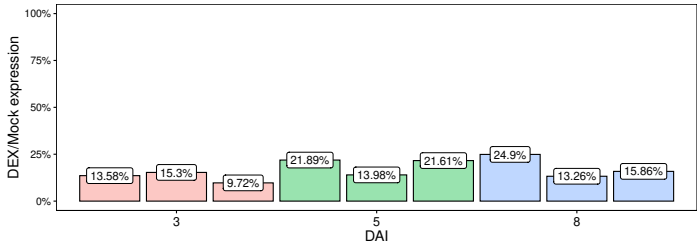
sufficient to disrupt the developmental processes normally controlled by the floral homeotic factors.

To determine the extent of the gene knockdowns, I compared transcript levels for *AP3* and *AG*, respectively, in dexamethasone and mock-treated samples. To this end, total RNA was extracted from the tissue samples and tested for RNA integrity and purity (see section 2.21). Subsequently, cDNA was synthesized from the total RNA preparations and RT-qPCR analysis was conducted using gene-specific primers.

The results of the RT-qPCR experiments showed that the degree of the knockdowns was high across all samples. In general, the knockdowns in the samples from the AP3$^{\text{amiRNA}}$ line were higher than those observed for AG$^{\text{amiRNA}}$, as previously observed [Thomson 2018]. In the former samples, mean expression values ranging from 2.4% and 12.5% were observed in DEX-treated samples relative to the mock-treated controls. This compared to mean expression values ranging from 9.7% to 24.9% in AG$^{\text{amiRNA}}$ samples.



(a) *AP3* expression in AP3$^{\text{amiRNA}}$ lines after induction.



(b) *AG* expression in AG$^{\text{amiRNA}}$ lines after induction.

Figure 5.3: Knockdown efficiencies for *AG* and *AP3* after amiRNA induction. The efficiency of the knockdowns is represented on the y-axis as the percentage of expression in DEX-treated samples relative to that in mock-treated control samples. Data are shown for three sets of biological independent samples. Results for the different time-points are colour-coded: salmon pink: 3 DAI; green: 5 DAI; blue: 8 DAI. DAI: Days after induction.

5.3.3 RNA-sequencing and a first assessment of the data obtained

The results of the experiments described above indicated that the knockdowns of the floral homeotic genes had been successful in the different samples I generated. To identify genes whose expression depends on AP3 and AG activity at different floral stages, the poly-A fraction of the total RNA preparations were sequenced, generating 100 bp paired-end reads. Sequencing was carried out by the company BGI in China, which employs a technology comparable to Illumina sequencing in terms of performance. BGI technology tethers the DNA fragments to proprietary "nanoballs" instead of a flat flow cell, but the principle is otherwise similar to that employed by Illumina.

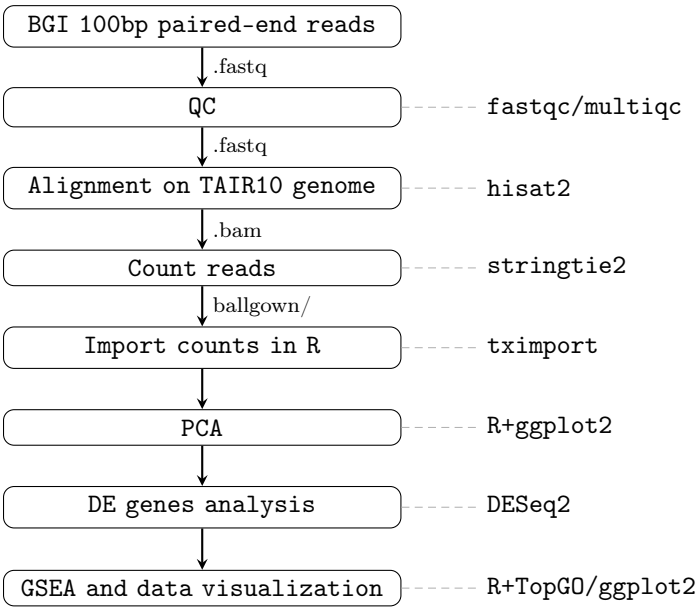


Figure 5.4: Flow chart describing the main components of the bioinformatics pipeline used for the analysis of the RNA-seq results. Individual steps are outlined in the solid boxes. The tools used to perform each step are written on the right and connected to their corresponding step via dashed, grey lines. The intermediate file types are written next to the arrows, which connect the steps as input→output. `Ballgown/` is not an intermediate file type but a standardised directory organisation that `tximport` can use to extract reads counts per gene.

BGI carried out the demultiplexing of the reads and the adapter trimming. To assess the quality of the sequencing reads, I used the software

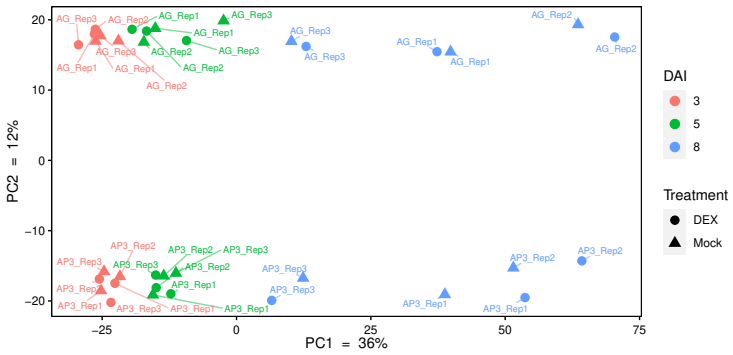
package `fastqc`. As `fastqc` results are difficult to compare across samples, I also employed `multiqc` to easily spot samples that significantly diverge from the mean in any of the metrics. In general, the reads for all samples appeared to be of high quality and no problems with read composition, number, residual adapters, duplication rate, or k-mer bias could be detected.

Next, I aligned the reads to the Arabidopsis genome sequence (version TAIR10) using the software `hisat2` in spliced mode. `Hisat2` is a Burrows-Wheeler transform (BWT) aligner with low hardware requirements that can achieve a good balance between speed and sensitivity, and is well suited for the analysis RNA-Seq reads [Sirén et al. 2014; Kim et al. 2015]. The aligned reads were then counted using `stringtie`. I ran `stringtie` with the `-B` option, which organises the results in a specific directory layout, pioneered by the `R` package `ballgown` [Frazee et al. 2015]. This layout can be easily parsed by `tximport`, an `R` library that standardises the import of counts from a plethora of different bioinformatics tools and pipelines.

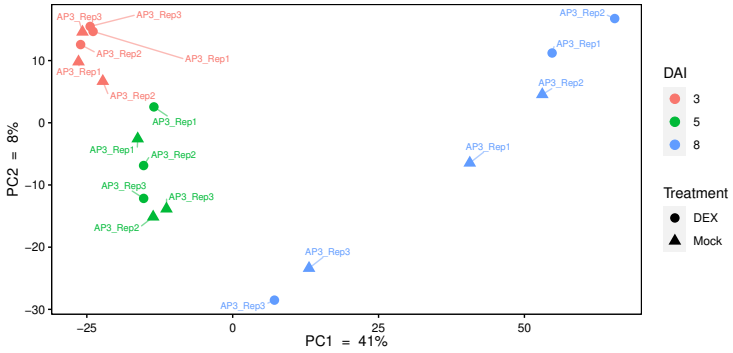
After importing the read counts for each gene, I calculated transcript per million (TPM) values and carried out a Principal Component Analysis (PCA) (see section 3.3.2). The TPM values were computed with the following formula:

$$\mathrm{TPM}_i = \frac{q_i/l_i}{\sum_j q_j/l_j}$$

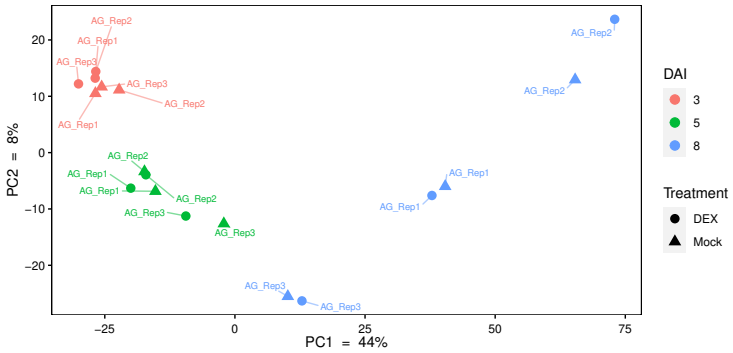
where q_i denotes reads mapped to a transcript i, l_i is the transcript's length, and $\sum_j q_j/l_j$ corresponds to the sum of mapped reads to each transcript i, normalised by the transcript's length.



(a) Full dataset, all samples.



(b) AP3$^{\text{amiRNA}}$ only.



(c) AG$^{\text{amiRNA}}$ only.

Figure 5.5: Results of a PCA analysis of the RNA-seq data. Gene expression values were log-transformed as $\log_{10}(\text{TPM} + 0.5)$ before computing the PCs to reduce the effects of outliers. For all panels, Principal Component 1 (PC1) is shown on the x-axis and PC2 on the y-axis. The percentage of variance explained by each PC is shown after the equal sign. The panels show, from top to bottom, all samples, only the AP3$^{\text{amiRNA}}$ samples, and only AG$^{\text{amiRNA}}$ samples, respectively.

In the PCA plot, the samples tended to cluster primarily in base of the time-point they were collected at. This was expected given the substantial differences in temporal gene expression previously described for flower development [Ryan et al. 2015b].

The PCA on the full set of samples also showed substantial differences between the samples from AG^{amiRNA} and $AP3^{amiRNA}$ plants. Again, this was to some extent expected as the B and C function regulators have distinct functions in flower development in addition to their shared activities. However, I also considered the possibility that these differences are in part the result of the complex backgrounds of these lines. For example, differences in the strength of amiRNA expression may have led to some differences in gene expression. Although undesirable, such a possible biases was kept under control by carrying out paired comparisons between DEX-treated and mock-treated plants when calculating differential expression of genes.

After induction, flowers of FIS plants gradually loose synchrony especially at later stages of development (> stage 8) with flowers in the center of the "cauliflower curd" being typically at more advanced stages than the ones at the edge of the curd [Ryan et al. 2015b]. While great care was taken to sample only flowers in the center of the curd for the experiments, a wider spread of data points was observed at 8 DAI when compared to 3 and 5 DAI (see Figure 5.5). This was probably unavoidable given that at 8 DAI simply more time had passed for the plants of the different treatment populations to get slightly "out of sync".

5.3.4 Identification of differentially expressed genes

To identify genes with differential expression (DE) across the different sample pairs, I analysed the RNA-seq data in `R` using the `DESeq2` package. Each time-point was analysed separately comparing the gene expression in DEX-treated samples to the expression in the mock-treated samples. Genes with an adjusted p-value (or q-value) of less than 0.05 were considered differentially expressed and no fold change cut-off was applied.

When *AG* was perturbed via amiRNA expression, the number of genes that were up and downregulated was almost equal at 3 DAI (130 and 116, respectively; see Table 5.1). At 5 and 8 DAI, there were mainly upregulated genes and the total number of DE genes decreased relative to the 3 DAI time-point. In DEX-treated samples where *AP3* expression was knocked

down, there were 421 significant DE genes at 3 DAI, with the number of upregulated genes more than double than the number of downregulated genes. Similar to *AG*, the number of DE genes identified at later stages decreased and upregulation became the prevalent effect of the *AP3* knockdown (see Table 5.1).

Table 5.1: Total number of up and down regulated genes and the cross section of genes with direct evidence of binding. Genome-wide binding data for *AP3* and *AG* were taken from Ó'Maoiléidigh et al. [2013] and [Wuest et al. 2012]

		q-value < 0.05		With Binding Evidence			
Gene	DAI	up	down	AGup	AGdown	AP3up	AP3down
	3	130	116	20	11	20	12
AG	**5**	69	17	6	6	12	8
	8	84	12	13	5	17	4
	3	294	127	33	18	42	25
AP3	**5**	97	12	12	5	16	6
	8	107	9	8	7	12	7

Note:
The number of genes with q-value < 0.05 that are either up or down-regulated in the knockdown respect to the mock-treated samples. The "AGup" and "AGdown" columns show the number genes that have also direct evidence of AG binding [Ó'Maoiléidigh et al. 2013], while columns "AP3up" and "AP3down" display the ones with direct evidence of AP3 binding [Wuest et al. 2012].

I also compared available list of genes in published ChIP-Seq datasets for AG and AP3 [Wuest et al. 2012; O'Maoileidigh et al. 2013]. Both the ChIP-Seq experiments were performed on stage 5 flowers, or around 4 DAI, from FIS plants expressing GFP-tagged versions of AG or AP3. Surprisingly, the cross-section of DE genes that have evidence of direct AG or AP3 binding is very narrow across the board, ranging from 5% to 15% (see Table 5.1).

5.3.5 Results of a Gene Ontology enrichment analysis

To functionally assess the DE genes, I computed Gene Ontology (GO) enrichments using the Fischer and KS tests (see section 3.3.2). Each set of genes was evaluated separately, and then a union of all the significant terms was clustered using a k-means algorithm (see Figures 5.6 and 5.7).

In the analysis, I found a number of GO terms related to stress responses, particularly those related to abiotic stresses such as *response to wounding*, *response to herbivore*, *cellular response to hypoxia*, *response to osmotic stress* and others. The high number of stress response genes could be due to several reasons, which I will discuss in 5.4. I focused in this work primarily on genes with known or likely functions in the control of developmental processes.

A large number of GO terms showed a significant enrichment at 3 DAI. This group can be divided in three subgroups. In the first two subgroups, there were GO terms that were found only in the *AP3* or *AG* dataset, while the third included terms that were significant in both datasets. GO terms found only in the AG knockdown experiment were *regulation of cell population proliferation*, *response to abscissic acid* and *stamen development*. *AP3* knock down affected terms related to cell growth and development like *unidimensional cell growth*, *regulation of cell size*, and *meristem development*. *AP3* seems to also have an early role in the auxin pathway, specifically auxin movement and compartmentalisation, with terms like *auxin efflux*, *auxin polar transport* and *response to auxin* being significantly enriched. A significant number of genes related to *flower development* was found when knocking down *AG* at 3 DAI and or 5 DAI, and *AP3* at 8 DAI. Gene sets enriched in almost all the time-points belong in broader categories such as *root development*, *cell division*, *leaf development* or *auxin homeostasis*. Notably, *AP3* also regulates a significant number of genes related to pollen development at 5 DAI.

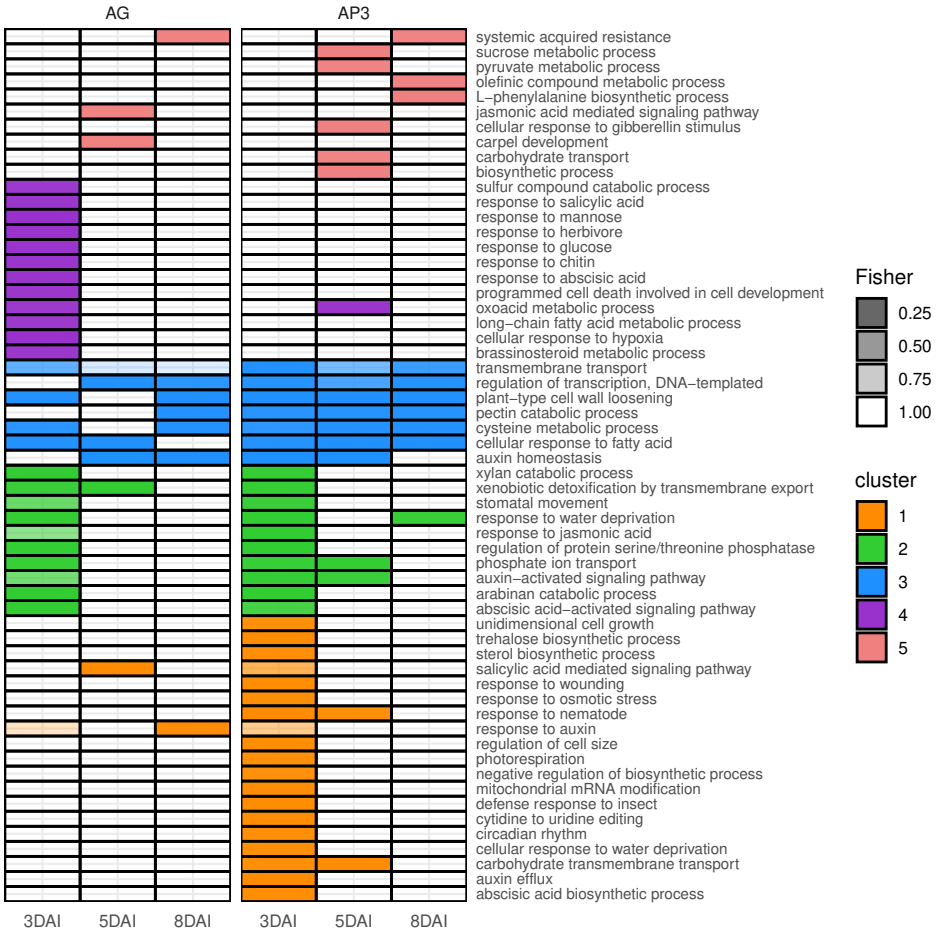


Figure 5.6: GO terms with a significant enrichment in DE genes identified by `topGO` with Fisher's exact test. The cells are colored in base of which cluster they belong to and shaded in base of the p-value they are representing. A lower p-value translates in a more intense color. Each row of cells represents one GO term, which is briefly described on the right side of the plot. Each column of cells refers to one of the three pairwise time-point comparisons performed in `DRIMSeq`.

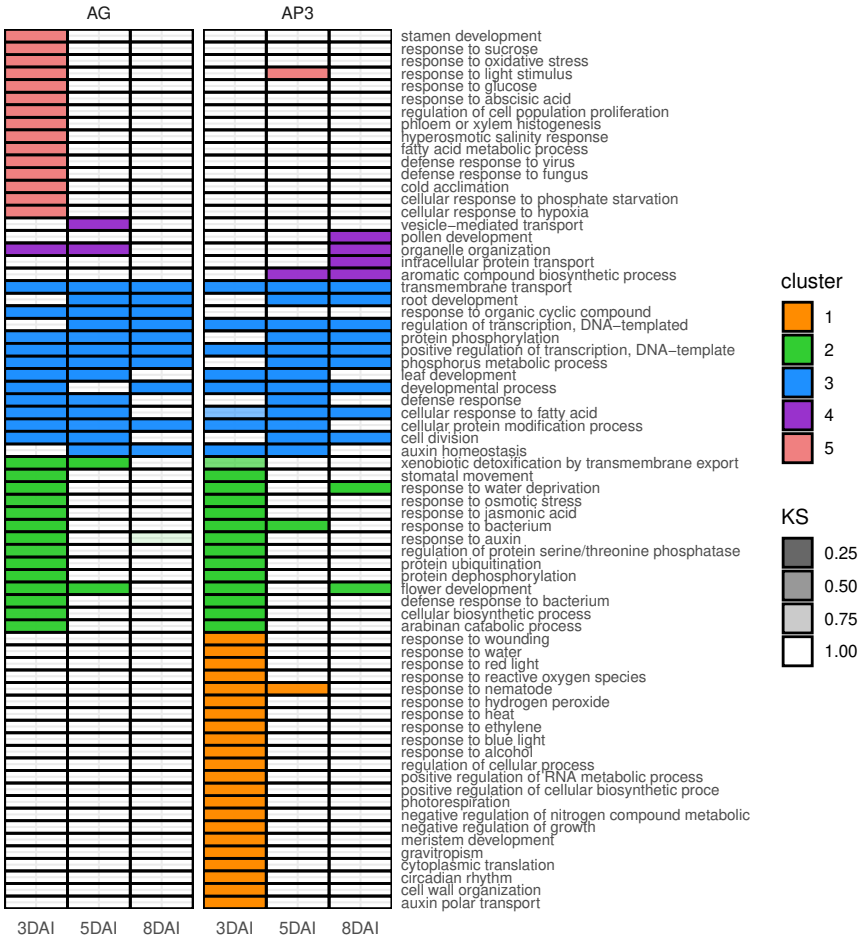
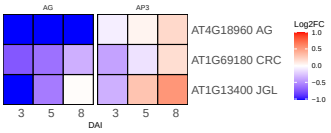
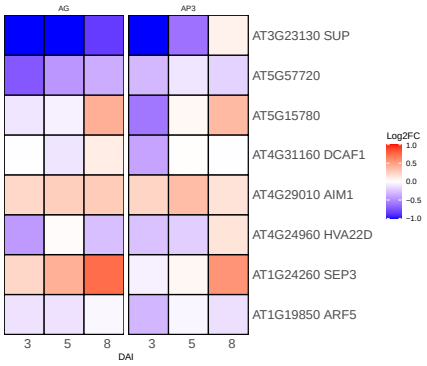


Figure 5.7: GO terms with a significant enrichment in DE genes identified by `topGO` with the Kolgorov-Smirnov (KS) test. The cells are colored in base of which cluster they belong to and shaded in base of the p-value they are representing. A lower p-value translates in a more intense color. Each row of cells represents one GO term, which is briefly described on the right side of the plot. Each column of cells refers to one of the three pairwise time-point comparisons performed in `DRIMSeq`. Due to the large number of significantly enriched GO terms, a lower p-value cutoff of 0.01 was used, to aid visualisation.

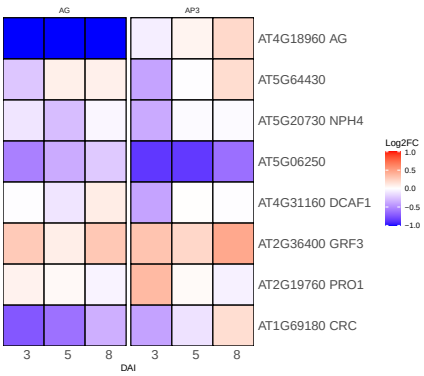
I then plotted the $\log_2$ FC of significant genes belonging to the three GO terms I found to be most relevant from downstream analysis: *carpel development*, *flower development* and *leaf development*. For each GO term, I plotted the genes that were differentially expressed in at least one time-point (see Figure 5.8). The plots show a number of genes being differentially expressed in more the one time-point, including *SUPERMAN (SUP)*, *CRABS CLAW (CRC)*, *SEPALLATA3 (SEP3)*, *ABNORMAL INFLORESCENCE MERISTEM (AIM1)* and *JAGGED-LIKE (JGL)* (discussed in section 5.4).



(a) *Carpel development*



(b) *Flower development*



(c) *Leaf development*

Figure 5.8: The $\log_2$ FC of genes differentially expressed in relevant Gene Ontology terms. The cells are coloured in a scale from blue to red, in base of the $\log_2$ FC between DEX and mock treatments (negative to positive).

5.3.6 Comparing differentially expressed genes with Ryan et al. [2015a]

To gain more insight into the expression patterns of the differentially expressed genes assigned to key Gene Ontology terms (see Figure 5.8), I compared my results with available gene expression data [Ryan et al. 2015a]. In Ryan et al. [2015a], the authors used the floral induction system to col-

lect a large amount of synchronised flowers form 0 to 13 days after flower transition. The RNA extracted from the tissue was then used in microarray experiments for high-throughput quantification of gene expression. In the paper, the gene expression at each time-point was compared with that in the preceding time-point and the authors reported the fold changes of the gene expression and the adjusted p-value of the observation. The fold changes of genes with a significant adjusted p-value in at least one of the comparisons were reported, for a total of 7,045 genes. First, I compare the two gene lists to see which of the genes I found in the *carpel development*, *leaf development* and *flower development* gene ontology categories, were also differentially expressed in Ryan et al. [2015a] (see Table 5.2). Half of the genes were found in both datasets, including *CRC*, *JGL*, *SEP3* and *SUP*. Interestingly, *CRC* expression, which is activated by both AP3 and AG in the first stages of development, shows a positive change in expression between 3 and 5 days after the flower transition, right after the largest increase in *AG* and *AP3* expression. This is in concordance with the results of the knockdown experiments, where *CRC* expression decreased in 3 and 5 days following *AG* and *AP3* knockdown. *JGL* fold changes show a monotonic decrease, turning negative at around 6 days after flower transition. Notably, in the knockdown experiments, *JGL* expression is promoted by AG in the first stages of flower development (3 and 5 days), while being repressed by AP3 at later stages (5 and 8 days). *SUP* expression is promoted by both AG and AP3 which had been previously shown to directly bind to its promoter [Sakai et al. 2000]. Is not surprising to observe that *SUP* expression changes closely resemble *AG* and *AP3*, just out of phase by one or two days. The knockdown experiments in this work showed that *SEP3* expression is repressed by both AG and AP3, particularly by AG at 8 days after flower transition. However, the fold changes in Ryan et al. [2015a] *SEP3* expression closely follow that of *AG*, which suggests that *SEP3* is regulated by other factors, like genes involved in flowering time regulation [Lee et al. 2012; Liu et al. 2009].

Table 5.2: Overlap with genes found in Ryan et al. [2015a]

GO	GeneID	GeneName	Ryan et al. [2015a]
Flower Development	*AT1G69180*	*CRC*	Yes
	AT2G19760	*PRO1*	No
	AT2G36400	*GRF3*	No
	AT4G31160	*DCAF1*	No
	AT5G06250		Yes
	AT5G20730	*NPH4*	No
	AT5G64430		No
Carpel Development	*AT1G13400*	*JGL*	Yes
	AT1G69180	*CRC*	Yes
Leaf Development	*AT1G19850*	*ARF5*	No
	AT1G24260	*SEP3*	Yes
	AT4G24960	*HVA22D*	Yes
	AT4G29010	*AIM1*	No
	AT4G31160	*DCAF1*	No
	AT5G15780		No
	AT5G57720		Yes
	AT3G23130	*SUP*	Yes

Note:
The table shows which of the genes shown in Figure5.8 overlap with the list of genes differentially expressed during flower development, described in Ryan et al. [2015a].

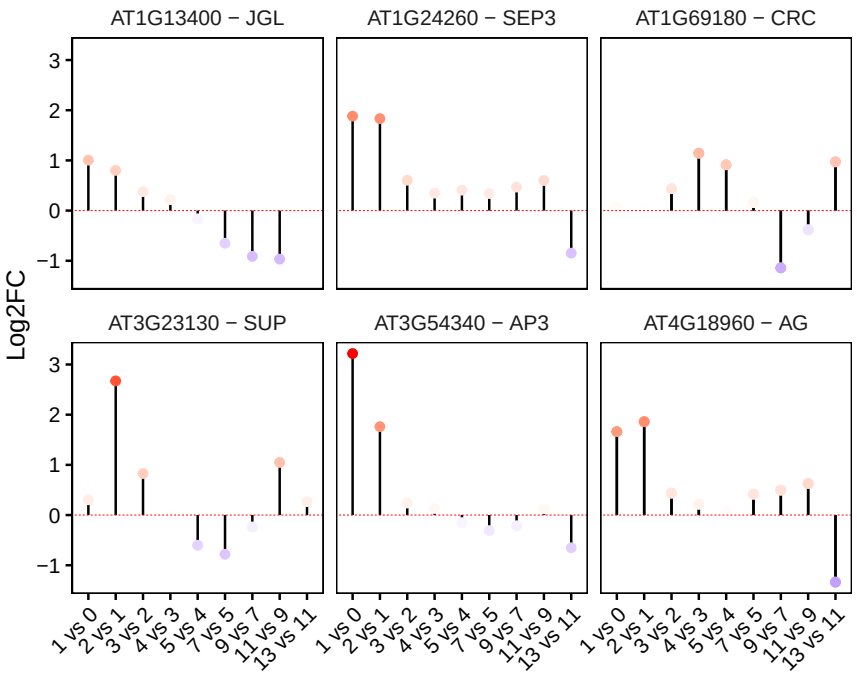


Figure 5.9: Lollipop plot on log_2-transformed fold changes from pairwise comparisons of adjacent time-points in Ryan et al. [2015a]. Each panel shows the time series for a subset of key genes found in this work, that was also found in Ryan et al. [2015a] (see Table 5.2). In all panels, the y-axis represents the log_2-transformed fold changes, while the x-axis shows the comparisons between flower development time-points. A red horizontal line marks a fold change of zero. The data points at the end of the "lollipop" segments are coloured in base of the fold change from blue (negative) to red (positive). *AG* and *AP3* fold changes were added for comparison.

5.3.7 Comparison of the transcriptomics datasets with results from genome-wide localization experiments

As mentioned above, a number of published ChIP-seq dataset are available that describe the genome-wide binding sites of MADS-domain transcription factors involved in flower development [Pajoro et al. 2014; Wuest et al. 2012; O'Maoileidigh et al. 2013]. These include binding data for AG and AP3, which were generated using stage 5 flowers [Wuest et al. 2012; O'Maoileidigh et al. 2013]. A comparison of the genes with known AP3 or AG binding sites and genes I identified as differentially expressed in the RNA-seq experiments showed a limited but significant overlap (see Table 5.1). I also conducted a pre-ranked gene set enrichment analysis (GSEA), using the `R` package `fgsea`. This method can be used to detect the enrichment of an a priori defined set of genes in a dataset of differentially expressed genes. For this analysis, I not only considered genes bound by AP3 and AG but also included binding data for the floral homeotic factors AP1 and SEP3 which had been generated for three floral stages (2, 4 and 8 days after induction of flower development) [Pajoro et al. 2014; Wuest et al. 2012; O'Maoileidigh et al. 2013].

In the GSEA, the sign of the enrichment and its magnitude represent how much each set of genes is enriched for up or down-regulated genes in a transcriptomics dataset. As the name suggests, a pre-ranked GSEA requires the genes to be ranked before GSEA score calculation. I ranked the genes by their fold change and p-value using the formula:

$$GeneRank = sign(FoldChange) * -\log_{10} pvalue$$

This transformation combines the directionality (sign) of the fold change, with the p-value of the gene. The p-value is log_{10} transformed so that smaller, p-values would have a more extreme rank, depending on their sign. Six sets of ranks were calculated on the gene expression datasets at 3, 5 and 8 DAI, for the *AG* and *AP3* knockdowns (see section 5.3.1). The output of `fgsea` is composed by two key metrics: a normalised enrichment score (NES) and an adjusted p-value. The NES is a measure of how enriched the gene set is in a subset of ranks, with respect to drawing the same number of genes from the ranks at random. Because the GSEA was run on signed ranks, the NES will be positive if the enrichment is weighted towards positive fold

changes, and negative for negative fold changes. The adjusted p-value is a measure of the False Discovery Rate (FDR). The adjusted p-value is obtained by adjusting the p-values for multiple testing with the Benjamini-Hochberg correction. The results were summarised in a bubble plot matrix, where the NES is displayed as a blue (negative) to red (positive) color gradient, and the adjusted p-value by the size of the bubble (see Figure 5.10).

The GSEA highlighted interesting synergies between the genes that responded to *AG* and *AP3* knockdowns, and the gene targets of the floral homeotic transcription factors. The first thing that stands out is the strong correlation between the enrichments for AP3 and PI gene sets. That is expected because, as it was noted in Wuest et al. [2012], AP3 and PI share most of the binding sites in the genome at stage 5 of flower development (4 DAI).

In addition to such general correlations, the available lists of direct gene targets offer insights into the possible dynamics between different MADS-domain transcription factors (see Figure 5.10). In particular, the information about the "direction" of the enrichment offers a proxy for cases where AG and AP3 may act in concert in the regulation of downstream genes, possibly as part of a higher order complex. For example, among the genes responding to *AG* perturbation at 3 DAI, those that are also bound by AP3/PI 4 DAI tend to be positively regulated by AG. In contrast, the same gene sets are enriched in genes that are under AG repression at 5 DAI and 8 DAI.

The datasets for AP1 and SEP3 binding datasets span more than one time-point, which allows for a finer grained view of their activities over the course of flower development. Even if the time-points do not exactly coincide with the ones used here, the positive enrichment values in the bottom left quadrant of the AG panel suggest that the genes bound by AP1 and SEP3 at early stages of flower development (at 2 and 4 DAI) are enriched for genes repressed by AG (at 3 and 5 DAI). At the 8 DAI time-point, the SEP3 binding data show a significant overlap with genes that are upregulated by AG, while the AP1 gene set does not show such a significant enrichment. These results are in agreement with the observation that AG interacts with SEP3 but not with AP1. In general, AG seems to work in concert with other MADS-domain transcription factors during the early stages of flower development mostly for gene repression. The exception is genes also bound by AP3/PI bound, which seem to have an overlapping function with AG in

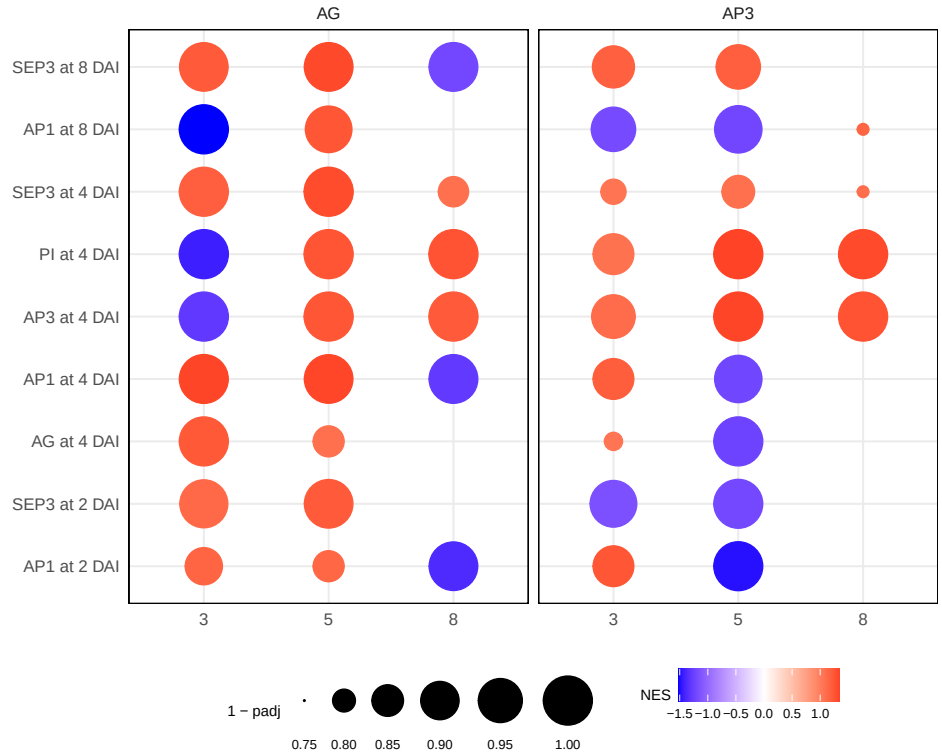


Figure 5.10: Gene set enrichment analysis (GSEA) using genome-wide binding datasets for MADS-domain transcription factors [Wuest et al. 2012; O'Maoileidigh et al. 2013; Pajoro et al. 2014]. The results are organised in two grids, for AG and AP3 respectively. The columns correspond to the ranks calculated from the fold changes and p-values from the pairwise comparisons in the knockdown experiments, while the rows correspond to the gene sets tested for enrichment in the GSEA. Bubbles on the grid are coloured in a gradient from blue (more negative) to red (more positive) enrichment values (GSEA scores). The bubble size reflects the underlying adjusted p-value of the enrichment. The adjusted p-value cutoff is 0.25, enrichments below this cutoff are not shown. The more significant adjusted p-values, the lower their value. To highlight the more significant adjusted p-values the bubbles size is proportional to 1 - adjusted p-value (the wider the bubble area, the closer the adjusted p-value is to zero).

gene activation the early stages of flower development (3-4 DAI).

When the GSEA was run on AP3 ranked data, the most significant enrichment values were detected for early floral stages. In particular, genes bound by AP1, AG and SEP3 between 2 and 4 DAI exhibited negative enrichment values in the 5 DAI expression data. In contrast, AP3/PI bound

genes that are usually repressed by AP3 at the 5 DAI time-point. This hints at a separation of targets, where AP3/PI could function in concert with other MADS-domain transcription to promote gene expression, and as an obligate heterodimer when repressing gene expression.

5.4 Discussion

In this chapter, I investigated the gene expression programmes acting downstream of the B and C function regulators AP3 and AG at different stages of early flower development. To this end, I used transgenic lines developed by Dr. Bennett Thomson [Thomson 2018] that allow the specific knockdown of the gene activities through of specific amiRNAs. As shown in section 5.3.2, I confirmed that the gene knockdowns had been successful. In general, the knockdown efficiencies were higher in the $AP3^{amiRNA}$ than in the $AP3^{amiRNA}$ line, possibly as a consequence of differences in amiRNA expression.

5.4.1 Genes acting downstream of the B and C function regulators during early flower development

The results of the transcriptomics experiments showed for both AP3 and AG that most of the genes that were identified in the knockdown experiments responded at the earliest (3 DAI) time and a GO analysis showed that genes commonly associated with flower development are enriched in the dataset, indicating that the experiment had worked as planned. These included for example, *SUP*, which is known to be involved in defining the boundary between stamen and carpel whorls. *Sup* loss-of-function mutants have supernumerary stamens that develop at the expense of the central carpel. Furthermore, in *sup* mutants, *AP3* expression expands to the fourth whorl, while *AG* expression is unaffected. In the experiments I conducted, *SUP* expression is strongly decreased after both *AG* and *AP3* knockdowns in most time-points taken, in agreement with previous results showing that *SUP* is regulated by B and C function factors [Sakai et al. 2000].

Another example of a known target gene of floral homeotic factors that responded in the experiment is *CRC*, which encodes a transcription factor involved in regulating cell division and expansion during carpel development. CRC is known to be activated by AG [Yamaguchi et al. 2017; Gómez-Mena et al. 2005; O'Maoileidigh et al. 2013; Lee et al. 2005] but appears to be repressed by AP3/PI [Wuest et al. 2012]. In agreement with this model, *CRC* was downregulated in the transcriptomics experiment in response to *AG* perturbation but up-regulated at 8 DAI after a knockdown of *AP3*.

As previously discussed, SEP3 plays an important role in the formation of high-order complexes with other floral homeotics factors [Honma and Goto

2001; Hugouvieux et al. 2019; Yang and Jack 2004; Smaczniak et al. 2012b]. Interestingly, *SEP3* expression appears to be repressed by both AP3 and AG especially at the 8 DAI time-point and thus at more intermediate stages of flower development. The significance of this downregulation event is currently unknown but it may point towards differences in the composition of floral quartets as flower development progresses.

ABNORMAL INFLORESCENCE MERISTEM1 (AIM1) is a gene involved in the beta-oxidation of fatty acids. In *aim1* mutants, floral meristems are disorganised and the affected plants exhibit severely reduced fertility [Richmond and Bleecker 1999]. *AIM* expression seems to be repressed by both AG and AP3.

There is an interesting difference in how AG and AP3 regulate *JAGGED-LIKE (JGL)* expression. AG promotes *JGL* expression, mainly in the early stages of flower development, while *AP3* seem to have only a mild, positive effect. However, at 3 and especially 8 DAI, *JGL* is strongly repressed by *AP3*. *JGL*, which is also known as *NUBBIN (NUB)*, encodes a C2H2 zinc-finger transcription factor. *JGL* functions redundantly with *JAGGED (JAG)* to promote the growth of micro-sporangia in the anthers and the carpel walls of the gynoecium [Dinneny et al. 2006]. Unlike *JAG*, *JGL* expression is confined to the adaxial side of lateral organs, such as the carpel walls of the gynoecium [Dinneny et al. 2006].

In addition to known regulators of flower development, a number of stress-related genes were differentially expressed in the experiments. Aside from special cases, such as *FLOWERING LOCUS C (FLC)*, a gene that encodes a key regulator of flowering [Sheldon et al. 2000] but is also assigned to the *response to cold* GO term, these genes have not been previously associated with flowering. For the experiment, plants were used that were grown side-by-side so that DEX and mock-treated plants experienced the exact same growth conditions. Also, the DEX treatment alone is not thought to lead to any major changes in gene expression [Ó'Maoiléidigh et al. 2015]. In previous experiments conducted in our laboratory, it was shown that inducible promoter systems can lead to the non-specific induction of genes, including many related to stresses. However, this was mainly observed for an ethanol-dependent promoter system, while the DEX-dependent promoter used for this study appeared to be much less affected [Ó'Maoiléidigh et al. 2015]. Another possibility is that the high expression of amiRNAs necessary for the

efficient knockdown of the floral homeotic genes, result the onset of cellular stress responses. Lastly, it is also possible that the stress related genes identified as differentially expressed in the experiments are *bona fide* targets of the AG and AP3 transcription factors. In fact, a large number of stress-related genes have been shown to act downstream of the key floral regulator LFY, which appears to play a role in the response of flowers to pathogens [Winter et al. 2011].

5.4.2 Future perspectives

As stated in the project aims, the work I described in this chapter is part of a bigger project involving stage-specific multi-omics datasets. The next step will be to combine the results obtained from this work with data from stage-specific protein-protein interaction analyses and genome-wide localization studies for the B and C function regulators. The datasets will be completed over the next few months and I will be involved in their analysis. In the meantime, I would like to follow up on new AG and AP3 targets that I have identified in this work. In the immediate future, I would like to validate the genes I discussed in section 5.4 by qPCR. I still have cDNA as well as frozen tissue left from the flowers I sampled for sequencing. Those samples can be used for qPCR experiments to have an independent verification of the gene expressions detected by sequencing (similar to Figure 5.3.2).

5.5 Conclusion and final remarks

Differences in genome-wide binding sites accounts only in part for the specificity of MADS-domain transcription factors. In fact, they share most of these binding-sites, which suggest a combined function (see section 1.3). The study of these complex regulatory pathways is slowed down by the lack of orthogonal datasets obtained at multiple stages during flower development. The transcriptomics dataset generated in this work is only a part of a bigger project involving multiple -omics approaches, which will shed a new light on the interplay of MADS-domain transcription factors during flower development.

Chapter 6

Conclusion

The main goal of my PhD work was to further the understanding of the biology of flowers, with a special emphasis on the function of floral trichomes. Trichomes offer a physical barrier for pests, and hence a better comprehension of the gene networks that regulate their formation on fruits may aid in efforts to improve the insect resistance of crops. Advances in this area could further lead to more sustainable approaches for pest control, a potential that is currently unexplored (see section 1). Previous work from our laboratory has shown that floral homeotic transcription factors, and in particular AG, repress the genetic programme for leaf development, which includes that for trichome formation [O'Maoileidigh et al. 2013; Ó'Maoiléidigh et al. 2018]. However, a knockdown of *AG* activity alone leads to the initiation of only a few trichomes on carpel valves. *TRY* and *CPC* are the two main trichome repressors in Arabidopsis. When *AG* expression is knocked-down in a *try cpc* double-mutant background, plants produce carpels covered in trichomes [Ó'Maoiléidigh et al. 2018]. To better our understand this process, I divided my work into two broad areas: a) studying the transcriptomal programme downstream of the floral homeotic transcription factors with stage-specific resolution, and b) gaining novel insights into the regulation of trichome development by exploring the interactomes of the TRY and CPC repressor proteins.

As described in sections 3.1.3 and 3.1.4, alternative splicing and lncRNA expression are particularly elusive phenomena due to the necessity of assembling and quantifying the expression of single transcripts. To explore the differential expression at the level of single transcript isoforms during the

onset of flower formation, I used a cutting-edge, third-generation transcriptomics approach based on the Oxford Nanopore MinION sequencer, which is capable of sequencing full-transcripts. The use of the floral induction system allowed a stage-specific sampling of sufficient floral tissue for Oxford Nanopore sequencing. For this project, I set up an experimental pipeline for Nanopore sequencing in our laboratory, from RNA extraction to data analysis. I was able to identify previously unknown splicing events which could be independently validate by qPCR. However, the qPCR results also revealed biases in the quantification of transcript expression. Oxford Nanopore sequencing is still a novel technology, which is under constant development, and there is still a high base calling error rate and a general lack of good bioinformatics tools for the analysis of the data it produces. Nevertheless, the dataset I generated in this work has the potential to become a valuable resource for the broader scientific community.

In chapter 5, I described the importance of high-order complexes of MADS-domain transcription factors in the regulation of flower development and how new and improved -omics datasets are necessary for further advancements in the field. To this end, I used gene perturbation techniques to knock-down the B and C class genes *AP3* and *AG* coupled with deep sequencing to unravel their downstream targets. The list of gene targets I obtained revealed interesting genes and expression dynamics that I plan to investigate further. The transcriptomics work described here is part of a concerted effort in our laboratory that aims at creating a multiomics dataset, with ChIP-seq and IP-MS experiments on the way, to shed a light on the composition and function of the floral quartets during the first stages of flower development.

The remainder of my PhD work focused on trichome development, exploring the interactome around TRY and CPC via biotin labeling and immunoprecipitation followed by mass spectrometry (see chapter 4). As outlined in section 4.1.2, *try* and *cpc* loss-of-function mutants have very different phenotypes, but the genes are very similar and their expression patterns in the leaf appear identical. Furthermore, *TRY* and *CPC* expression is found in adult trichomes while the proteins they encode are able to travel in neighbouring cells, a process that is thought could regulate trichome density and distribution on the leaf surface. The interactions between trichome-related genes have been studied extensively at the genetic level, but this work is

the first peek at the interactome around the two main repressors of trichome development. I identified a number of novel protein-protein interactions that could further our understanding on how the trichome development program is regulated between cells on the surface of the leaf.

In conclusion, the work described in this thesis offers a view into the actors governing trichome development in flowers using a variety of -omics approaches and molecular techniques. In the near future, I plan to publish the datasets to make them available to the scientific community and to further advance research in the field.

Bibliography

Anaconda software distribution, 2020.

N. Aerts, S. de Bruijn, H. van Mourik, G. C. Angenent, and A. D. van Dijk. Comparative analysis of binding patterns of mads-domain proteins in arabidopsis thaliana. *BMC Plant Biology*, 18:1–16, 6 2018. ISSN 14712229. doi: 10.1186/S12870-018-1348-8/FIGURES/6.

G. P. Alamancos, A. Pagès, J. L. Trincado, N. Bellora, and E. Eyras. Leveraging transcript quantification for fast computation of alternative splicing profiles. *RNA (New York, N.Y.)*, 21:1521–1531, 9 2015. ISSN 1469-9001. doi: 10.1261/RNA.051557.115.

E. R. Alvarez-Buylla, M. Benítez, A. Corvera-Poiré, Álvaro Chaos Cador, S. de Folter, A. G. de Buen, A. Garay-Arroyo, B. García-Ponce, F. Jaimes-Miranda, R. V. Pérez-Ruiz, A. Piñeyro-Nelson, Y. E. Sánchez-Corrales, A. C. Cador, S. de Folter, A. G. de Buen, A. Garay-Arroyo, B. García-Ponce, F. Jaimes-Miranda, R. V. Pérez-Ruiz, A. Piñeyro-Nelson, and Y. E. Sánchez-Corrales. Flower development. *The Arabidopsis Book / American Society of Plant Biologists*, 8:e0127, 1 2010. ISSN 1543-8120. doi: 10.1199/TAB.0127.

M. Alves-Ferreira, F. Wellmer, A. Banhara, V. Kumar, J. L. Riechmann, and E. M. Meyerowitz. Global expression profiling applied to the analysis of arabidopsis stamen development. *Plant physiology*, 145:747–762, 2007. ISSN 0032-0889. doi: 10.1104/PP.107.104422.

T. Ammunét, N. Wang, S. Khan, and L. L. Elo. Deep learning tools are top performers in long non-coding rna prediction. *Briefings in Functional Genomics*, 21:230–241, 5 2022. ISSN 20412657. doi: 10.1093/BFGP/ELAB045.

F. Ariel, T. Jegu, D. Latrasse, N. Romero-Barrios, A. Christ, M. Benhamed, and M. Crespi. Noncoding transcription by alternative rna polymerases dynamically regulates an auxin-driven chromatin loop. *Molecular Cell*, 55: 383–396, 8 2014. ISSN 1097-2765. doi: 10.1016/J.MOLCEL.2014.06.011.

D. Arora, N. B. Abel, C. Liu, P. van Damme, K. Yperman, D. Eeckhout, L. D. Vu, J. Wang, A. Tornkvist, F. Impens, B. Korbei, J. van Leene, A. Goossens, G. de Jaeger, T. Ott, P. N. Moschou, and D. van Damme. Establishment of proximity-dependent biotinylation approaches in different plant model systems. *The Plant Cell*, 32:3388–3407, 11 2020. ISSN 1040-4651. doi: 10.1105/TPC.20.00235.

N. Arteaga, B. Méndez-Vigo, A. Fuster-Pons, M. Savic, A. Murillo-Sánchez, F. X. Picó, and C. Alonso-Blanco. Differential environmental and genomic architectures shape the natural diversity for trichome patterning and morphology in different arabidopsis organs. *Plant, Cell Environment*, 45: 3018–3035, 10 2022. ISSN 1365-3040. doi: 10.1111/PCE.14308.

N. Bache, P. E. Geyer, D. B. Bekker-Jensen, O. Hoerning, L. Falkenby, P. V. Treit, S. Doll, I. Paron, J. B. Müller, F. Meier, J. V. Olsen, O. Vorm, and M. Mann. A novel lc system embeds analytes in pre-formed gradients for rapid, ultra-robust proteomics. *Molecular cellular proteomics : MCP*, 17: 2284–2296, 11 2018. ISSN 1535-9484. doi: 10.1074/MCP.TIR118.000853.

R. Balkunde, D. Bouyer, and M. Hülskamp. Nuclear trapping by gl3 controls intercellular transport and redistribution of ttg1 protein in arabidopsis. *Development*, 138:5039–5048, 11 2011. ISSN 09501991. doi: 10.1242/dev.072454.

F. E. Baralle and J. Giudice. Alternative splicing as a regulator of development and tissue identity. *Nature Reviews Molecular Cell Biology*, 18: 437–451, 7 2017. ISSN 14710080. doi: 10.1038/nrm.2017.27.

S. Behera, A. Voshall, and E. N. Moriyama. Plant transcriptome assembly: Review and benchmarking. *Bioinformatics*, pages 109–130, 3 2021. doi: 10.36255/EXONPUBLICATIONS.BIOINFORMATICS.2021.CH7.

E. Benková, M. Michniewicz, M. Sauer, T. Teichmann, D. Seifertová, G. Jürgens, and J. Friml. Local, efflux-dependent auxin gradients as a common

module for plant organ formation. *Cell*, 115:591–602, 11 2003. ISSN 00928674. doi: 10.1016/S0092-8674(03)00924-3.

R. Benlloch, A. Berbel, A. Serrano-Mislata, and F. Madueño. Floral initiation and inflorescence architecture: A comparative view. *Annals of Botany*, 100:659, 9 2007. ISSN 03057364. doi: 10.1093/AOB/MCM146.

A. Bhate, D. J. Parker, T. W. Bebee, J. Ahn, W. Arif, E. H. Rashan, S. Chorghade, A. Chau, J. H. Lee, S. Anakk, R. P. Carstens, X. Xiao, and A. Kalsotra. Esrp2 controls an adult splicing programme in hepatocytes to support postnatal liver maturation. *Nature Communications*, 6: 8768, 11 2015. ISSN 20411723. doi: 10.1038/ncomms9768.

H. Blum, H. Beier, and H. J. Gross. Improved silver staining of plant proteins, rna and dna in polyacrylamide gels. *ELECTROPHORESIS*, 8:93–99, 1 1987. ISSN 1522-2683. doi: 10.1002/ELPS.1150080203.

M. A. Blázquez, L. N. Soowal, I. Lee, and D. Weigel. Leafy expression and flower initiation in arabidopsis. *Development*, 124:3835–3844, 10 1997. ISSN 09501991.

D. Bouyer, F. Geier, F. Kragler, A. Schnittger, M. Pesch, K. Wester, R. Balkunde, J. Timmer, C. Fleck, and M. Hülskamp. Two-dimensional patterning by a trapping/depletion mechanism: The role of ttg1 and gl3 in arabidopsis trichome formation. *PLOS Biology*, 6:e141, 6 2008. ISSN 1545-7885. doi: 10.1371/JOURNAL.PBIO.0060141.

J. L. Bowman, D. R. Smyth, and E. M. Meyerowitz. Genes directing flower development in arabidopsis. *The Plant Cell*, 1:37–52, 1 1989. ISSN 1040-4651. doi: 10.1105/tpc.1.1.37.

J. L. Bowman, J. Alvarez, D. Weigel, E. M. Meyerowitz, and D. R. Smyth. Control of flower development in arabidopsis thaliana by apetala1 and interacting genes. *Development*, 119, 1993.

J. L. Bowman, D. R. Smyth, and E. M. Meyerowitz. The abc model of flower development: Then and now. *Development (Cambridge)*, 139:4095–4098, 11 2012. ISSN 09501991. doi: 10.1242/DEV.083972.

D. C. Boyes, A. M. Zayed, R. Ascenzi, A. J. McCaskill, N. E. Hoffman, K. R. Davis, and J. Görlach. Growth stage-based phenotypic analysis of

arabidopsis: a model for high throughput functional genomics in plants. *The Plant cell*, 13:1499–1510, 7 2001. ISSN 1040-4651. doi: 10.1105/TPC.010011.

D. Bradley, O. Ratcliffe, C. Vincent, R. Carpenter, and E. Coen. Inflorescence commitment and architecture in arabidopsis. *Science (New York, N.Y.)*, 275:80–3, 1 1997. ISSN 0036-8075.

N. L. Bray, H. Pimentel, P. Melsted, and L. Pachter. Near-optimal probabilistic rna-seq quantification. *Nature Biotechnology 2016 34:5*, 34:525–527, 4 2016. ISSN 1546-1696. doi: 10.1038/nbt.3519.

A. Byrne, A. E. Beaudin, H. E. Olsen, M. Jain, C. Cole, T. Palmer, R. M. DuBois, E. C. Forsberg, M. Akeson, and C. Vollmers. Nanopore long-read rnaseq reveals widespread transcriptional variation among the surface receptors of individual b cells. *Nature Communications*, 8:16027, 7 2017. ISSN 2041-1723. doi: 10.1038/ncomms16027.

L. Cao, Y. Wang, C. Bi, Q. Ye, T. Yin, and N. Ye. Prelnc: An accurate tool for predicting lncrnas based on multiple features. *Genes*, 11:1–20, 9 2020. ISSN 20734425. doi: 10.3390/GENES11090981.

G. Capovilla, E. Symeonidi, R. Wu, and M. Schmid. Contribution of major flm isoforms to temperature-dependent flowering in arabidopsis thaliana. *Journal of Experimental Botany*, 68:5117–5127, 11 2017. ISSN 0022-0957. doi: 10.1093/jxb/erx328.

B. Causier, Z. Schwarz-Sommer, and B. Davies. Floral organ identity: 20 years of abcs. *Seminars in Cell and Developmental Biology*, 21:73–79, 2010. ISSN 10849521. doi: 10.1016/J.SEMCDB.2009.10.005.

D. Chen, W. Yan, L.-Y. Fu, and K. Kaufmann. Architecture of gene regulatory networks controlling flower development in arabidopsis thaliana. *Nature Communications*, 9:4534, 12 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-06772-3.

A. W. Cheng, J. Shi, P. Wong, K. L. Luo, P. Trepman, E. T. Wang, H. Choi, C. B. Burge, and H. F. Lodish. Muscleblind-like 1 (mbnl1) regulates pre-mrna alternative splicing during terminal erythropoiesis. *Blood*, 124:598–610, 7 2014. ISSN 15280020. doi: 10.1182/blood-2013-12-542209.

S. Chin, T. Kwon, B. R. Khan, J. A. Sparks, E. L. Mallery, D. B. Szymanski, and E. B. Blancaflor. Spatial and temporal localization of spirrig and wave/scar reveal roles for these proteins in actin-mediated root hair development. *The Plant cell*, 33:2131–2148, 7 2021. ISSN 1532-298X. doi: 10.1093/PLCELL/KOAB115.

K. F. Cho, T. C. Branon, N. D. Udeshi, S. A. Myers, S. A. Carr, and A. Y. Ting. Proximity labeling in mammalian cells with turboid and split-turboid. *Nature Protocols 2020 15:12*, 15:3971–3999, 11 2020. ISSN 1750-2799. doi: 10.1038/s41596-020-0399-0.

E. S. Coen and E. M. Meyerowitz. The war of the whorls: Genetic interactions controlling flower development. *Nature*, 353:31–37, 1991a. ISSN 00280836. doi: 10.1038/353031A0.

E. S. Coen and E. M. Meyerowitz. The war of the whorls: Genetic interactions controlling flower development. *Nature*, 353:31–37, 1991b. ISSN 00280836. doi: 10.1038/353031a0.

J. Cox, M. Y. Hein, C. A. Luber, I. Paron, N. Nagaraj, and M. Mann. Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed maxlfq. *Molecular Cellular Proteomics : MCP*, 13:2513, 9 2014. ISSN 15359484. doi: 10.1074/MCP.M113.031591.

T. Csorba, J. I. Questa, Q. Sun, and C. Dean. Antisense coolair mediates the coordinated switching of chromatin states at flc during vernalization. *Proceedings of the National Academy of Sciences of the United States of America*, 111:16160–16165, 11 2014. ISSN 10916490. doi: 10.1073/pnas.1419030111.

T. Czechowski, M. Stitt, T. Altmann, M. K. Udvardi, and W. R. Scheible. Genome-wide identification and testing of superior reference genes for transcript normalization in arabidopsis. *Plant physiology*, 139:5–17, 2005. ISSN 0032-0889. doi: 10.1104/PP.105.063743.

T. da Costa Negri, W. A. L. Alves, P. H. Bugatti, P. T. M. Saito, D. S. Domingues, and A. R. Paschoal. Pattern recognition analysis on long noncoding rnas: a tool for prediction in plants. *Briefings in Bioinformatics*, 4 2018. ISSN 1467-5463. doi: 10.1093/bib/bby034.

M. W. Davis and E. M. Jorgensen. Ape, a plasmid editor: A freely available dna manipulation and visualization program. *Frontiers in Bioinformatics*, 2:5, 2 2022. ISSN 2673-7647. doi: 10.3389/FBINF.2022.818619.

S. Digiuni, S. Schellmann, F. Geier, B. Greese, M. Pesch, K. Wester, B. Dartan, V. Mach, B. P. Srinivas, J. Timmer, C. Fleck, and M. Hulskamp. A competitive complex formation mechanism underlies trichome patterning on arabidopsis leaves. *Molecular Systems Biology*, 4:217, 2008. ISSN 17444292. doi: 10.1038/msb.2008.54.

J. R. Dinneny, D. Weigel, and M. F. Yanofsky. Nubbin and jagged define stamen and carpel shape in arabidopsis. *Development*, 133:1645–1655, 5 2006. ISSN 0950-1991. doi: 10.1242/DEV.02335.

G. Ditta, A. Pinyopich, P. Robles, S. Pelaz, and M. F. Yanofsky. The sep4 gene of arabidopsis thaliana functions in floral organ and meristem identity. *Current Biology*, 14:1935–1940, 11 2004. ISSN 09609822. doi: 10.1016/j.cub.2004.10.028.

Z. Dosztányi, B. Mészáros, and I. Simon. Anchor: Web server for predicting protein binding regions in disordered proteins. *Bioinformatics*, 25:2745–2746, 10 2009. ISSN 13674803. doi: 10.1093/bioinformatics/btp518.

K. Edwards, C. Johnstone, and C. Thompson. A simple and rapid method for the preparation of plant genomic dna for pcr analysis. *Nucleic Acids Research*, 19:1349–1349, 3 1991. ISSN 0305-1048. doi: 10.1093/NAR/19.6.1349.

J. M. Engreitz, J. E. Haines, E. M. Perez, G. Munson, J. Chen, M. Kane, P. E. McDonel, M. Guttman, and E. S. Lander. Local regulation of gene expression by lncrna promoters, transcription and splicing. *Nature*, 539:452–455, 11 2016. ISSN 0028-0836. doi: 10.1038/nature20149.

G. A. Erd, Os, M. A. Pajkos, Z. Dosztányi, and D. Dosztányi. Iupred3: prediction of protein disorder enhanced with unambiguous experimental annotation and visualization of evolutionary conservation. *Nucleic Acids Research*, 49:W297–W303, 7 2021. ISSN 0305-1048. doi: 10.1093/NAR/GKAB408.

J. J. Esch, M. Chen, M. Sanders, M. Hillestad, S. Ndkium, B. Idelkope, J. Neizer, and M. D. Marks. A contradictory glabra3 allele helps define gene interactions controlling trichome development in arabidopsis. *Development*, 130:5885–5894, 12 2003. ISSN 09501991. doi: 10.1242/dev.00812.

M. Fambrini and C. Pugliesi. The dynamic genetic-hormonal regulatory network controlling the trichome development in leaves. *Plants*, 8:253, 2019. doi: 10.3390/plants8080253.

S. Filichkin, H. D. Priest, M. Megraw, and T. C. Mockler. Alternative splicing in plants: Directing traffic at the crossroads of adaptation and environmental stress, 4 2015. ISSN 13695266.

S. D. Folter, R. G. Immink, M. Kieffer, L. Pařenicová, S. R. Henz, D. Weigel, M. Busscher, M. Kooiker, L. Colombo, M. M. Kater, B. Davies, and G. C. Angenent. Comprehensive interaction map of the arabidopsis mads box transcription factors. *The Plant cell*, 17:1424–1433, 2005. ISSN 1040-4651. doi: 10.1105/TPC.105.031831.

A. C. Frazee, G. Pertea, A. E. Jaffe, B. Langmead, S. L. Salzberg, and J. T. Leek. Ballgown bridges the gap between transcriptome assembly and expression analysis. *Nature biotechnology*, 33:243–6, 3 2015. ISSN 1546-1696. doi: 10.1038/nbt.3172.

J. Giudice, Z. Xia, E. T. Wang, M. A. Scavuzzo, A. J. Ward, A. Kalsotra, W. Wang, X. H. Wehrens, C. B. Burge, W. Li, and T. A. Cooper. Alternative splicing regulates vesicular trafficking genes in cardiomyocytes during postnatal heart development. *Nature Communications*, 5:3603, 4 2014. ISSN 20411723. doi: 10.1038/ncomms4603.

K. Goslin, B. Zheng, A. Serrano-Mislata, L. Rae, P. T. Ryan, K. Kwaśniewska, B. Thomson, D. S. Ó'Maoiléidigh, F. Madueño, F. Wellmer, and E. Graciet. Transcription factor interplay between leafy and apetala1/cauliflower during floral initiation. *Plant Physiology*, 174: 1097–1109, 6 2017. ISSN 15322548. doi: 10.1104/pp.17.00098.

K. Goslin, A. Finocchio, and F. Wellmer. A golden gate-based plasmid library for the rapid assembly of biotin ligase constructs for proximity labelling, 2021. ISSN 26928205.

K. Goto and E. M. Meyerowitz. Function and regulation of the arabidopsis floral homeotic gene pistillata. *Genes development*, 8:1548–1560, 1994. ISSN 0890-9369. doi: 10.1101/GAD.8.13.1548.

L. Gramzow and G. Theissen. A hitchhiker's guide to the mads world of plants. *Genome Biology*, 11, 6 2010. ISSN 14747596. doi: 10.1186/GB-2010-11-6-214.

I. Gupta, P. G. Collier, B. Haase, A. Mahfouz, A. Joglekar, T. Floyd, F. Koopmans, B. Barres, A. B. Smit, S. A. Sloan, W. Luo, O. Fedrigo, M. E. Ross, and H. U. Tilgner. Single-cell isoform rna sequencing characterizes isoforms in thousands of cerebellar cells. *Nature Biotechnology*, 36:1197–1202, 10 2018. ISSN 1087-0156. doi: 10.1038/nbt.4259.

R. A. Gupta, N. Shah, K. C. Wang, J. Kim, H. M. Horlings, D. J. Wong, M. C. Tsai, T. Hung, P. Argani, J. L. Rinn, Y. Wang, P. Brzoska, B. Kong, R. Li, R. B. West, M. J. V. D. Vijver, S. Sukumar, and H. Y. Chang. Long non-coding rna hotair reprograms chromatin state to promote cancer metastasis. *Nature*, 464:1071–1076, 4 2010. ISSN 00280836. doi: 10.1038/nature08975.

C. Gustafson-Brown, B. Savidge, and M. F. Yanofsky. Regulation of the arabidopsis floral homeotic gene apetala1. *Cell*, 76:131–143, 1 1994. ISSN 00928674. doi: 10.1016/0092-8674(94)90178-3.

A. Gómez-Felipe, D. Kierzkowski, and S. de Folter. The relationship between agamous and cytokinin signaling in the establishment of carpeloid features. *Plants*, 10:827, 5 2021. ISSN 22237747. doi: 10.3390/PLANTS10050827/S1.

C. Gómez-Mena, S. de Folter, M. M. R. Costa, G. C. Angenent, and R. Sablowski. Transcriptional program controlled by the floral homeotic gene agamous during early organogenesis. *Development (Cambridge, England)*, 132:429–438, 2 2005. ISSN 0950-1991. doi: 10.1242/DEV.01600.

K. S. H. T. H Nakamura, KI Taoka. Functional diversification of fd transcription factors in rice, components of florigen activation complexes. *Plant Cell Physiol.*, 54:385–397, 2013. doi: 10.1093/pcp/pct005.

E. J. Hawkes, S. P. Hennelly, I. V. Novikova, J. A. Irwin, C. Dean, and K. Y. Sanbonmatsu. Coolair antisense rnas form evolutionarily conserved elaborate secondary structures. *Cell reports*, 16:3087–3096, 9 2016. ISSN 2211-1247. doi: 10.1016/j.celrep.2016.08.045.

R. Henriques, H. Wang, J. Liu, M. Boix, L. F. Huang, and N. H. Chua. The antiphasic regulatory module comprising cdf5 and its antisense rna flore links the circadian clock to photoperiodic flowering. *The New phytologist*, 216:854–867, 11 2017. ISSN 1469-8137. doi: 10.1111/NPH.14703.

M. A. G. Herz, M. G. Kubaczka, G. Brzyżek, L. Servi, M. Krzyszton, C. Simpson, J. Brown, S. Swiezewski, E. Petrillo, and A. R. Kornblihtt. Light regulates plant alternative splicing through the control of transcriptional elongation. *Molecular Cell*, 73:1066–1074.e3, 3 2019. ISSN 10974164. doi: 10.1016/j.molcel.2018.12.005.

T. Honma and K. Goto. Complexes of mads-box proteins are sufficient to convert leaves into floral organs. *Nature*, 409:525–529, 1 2001. ISSN 00280836. doi: 10.1038/35054083.

E. Huala and I. M. Sussex. Leafy interacts with floral homeotic genes to regulate arabidopsis floral development. *The Plant Cell*, 4:901–913, 8 1992. ISSN 1040-4651. doi: 10.1105/tpc.4.8.901.

N. C. Hubner, A. W. Bird, J. Cox, B. Splettstoesser, P. Bandilla, I. Poser, A. Hyman, and M. Mann. Quantitative proteomics combined with bac transgeneomics reveals in vivo protein interactions. *The Journal of cell biology*, 189:739–754, 5 2010. ISSN 1540-8140. doi: 10.1083/JCB.200911091.

R. Huertas, R. Catalá, J. M. Jiménez-Gómez, M. M. Castellano, P. Crevillén, M. Piñeiro, J. A. Jarillo, and J. Salinas. Arabidopsis sme1 regulates plant development and response to abiotic stress by determining spliceosome activity specificity. *The Plant cell*, 31:537–554, 2 2019. ISSN 1532-298X. doi: 10.1105/tpc.18.00689.

V. Hugouvieux, C. S. Silva, A. Jourdain, A. Stigliani, Q. Charras, V. Conn, S. J. Conn, C. C. Carles, F. Parcy, and C. Zubieta. Tetramerization of mads family transcription factors sepallata3 and agamous is required for floral meristem determinacy in arabidopsis. *Nucleic Acids Research*, 46: 4966–4977, 2019. ISSN 13624962. doi: 10.1093/nar/gky205.

F. Y. Hung, J. H. Chen, Y. R. Feng, Y. C. Lai, S. Yang, and K. Wu. Arabidopsis jmj29 is involved in trichome development by regulating the core trichome initiation gene glabra3. *The Plant Journal*, 103:1735–1743, 8 2020. ISSN 1365-313X. doi: 10.1111/TPJ.14858.

J. Hwang and Y. K. Kim. When a ribosome encounters a premature termination codon. *BMB Reports*, 46:9, 1 2013. ISSN 19766696. doi: 10.5483/BMBREP.2013.46.1.002.

H. Inoue, H. Nojima, and H. Okayama. High efficiency transformation of escherichia coli with plasmids. *Gene*, 96:23–28, 1990. ISSN 0378-1119. doi: 10.1016/0378-1119(90)90336-P.

T. Ito, F. Wellmer, H. Yu, P. Das, H. Ito, M. Alves-Ferreira, J. L. Riechmann, and E. M. Meyerowitz. The homeotic protein agamous controls microsporogenesis by regulation of sporocyteless. *Nature 2004 430:6997*, 430:356–360, 7 2004. ISSN 1476-4687. doi: 10.1038/nature02733.

T. Ito, K. H. Ng, T. S. Lim, H. Yu, and E. M. Meyerowitz. The homeotic protein agamous controls late stamen development by regulating a jasmonate biosynthetic gene in arabidopsis. *Plant Cell*, 19:3516–3529, 2007. ISSN 1532298X. doi: 10.1105/TPC.107.055467.

T. Jack, G. L. Fox, and E. M. Meyerowitz. Arabidopsis homeotic gene apetala3 ectopic expression: transcriptional and posttranscriptional regulation determine floral organ identity. *Cell*, 76:703–716, 2 1994. ISSN 0092-8674. doi: 10.1016/0092-8674(94)90509-6.

A. B. James, N. H. Syed, S. Bordage, J. Marshall, G. A. Nimmo, G. I. Jenkins, P. Herzyk, J. W. Brown, and H. G. Nimmo. Alternative splicing mediates responses of the arabidopsis circadian clock to temperature changes. *Plant Cell*, 24:961–981, 3 2012. ISSN 10404651. doi: 10.1105/tpc.111.093948.

Y. Jiao and E. M. Meyerowitz. Cell-type specific analysis of translating rnas in developing flowers reveals new levels of control. *Molecular Systems Biology*, 6:419, 1 2010. ISSN 17444292. doi: 10.1038/MSB.2010.76.

J. Jin, F. Tian, D. C. Yang, Y. Q. Meng, L. Kong, J. Luo, and G. Gao. Planttfdb 4.0: Toward a central hub for transcription factors and regula-

tory interactions in plants. *Nucleic Acids Research*, 45:D1040–D1045, 1 2017. ISSN 13624962. doi: 10.1093/NAR/GKW982.

J. Jin, P. Lu, Y. Xu, Z. Li, S. Yu, J. Liu, H. Wang, N. H. Chua, and P. Cao. Plncdb v2.0: a comprehensive encyclopedia of plant long noncoding rnas. *Nucleic Acids Research*, 49:D1489–D1495, 1 2021. ISSN 0305-1048. doi: 10.1093/NAR/GKAA910.

J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis. Highly accurate protein structure prediction with alphafold. *Nature 2021 596:7873*, 596:583–589, 7 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03819-2.

Y. J. Kang, D. C. Yang, L. Kong, M. Hou, Y. Q. Meng, L. Wei, and G. Gao. Cpc2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Research*, 45:W12–W16, 7 2017. ISSN 0305-1048. doi: 10.1093/NAR/GKX428.

K. Kaufmann, J. M. Muiño, R. Jauregui, C. A. Airoldi, C. Smaczniak, P. Krajewski, and G. C. Angenent. Target genes of the mads transcription factor sepallata3: Integration of developmental and hormonal pathways in the arabidopsis flower. *PLOS Biology*, 7:e1000090, 2009. ISSN 1545-7885. doi: 10.1371/JOURNAL.PBIO.1000090.

K. Kaufmann, F. Wellmer, J. M. Muiñ, T. Ferner, S. E. Wuest, V. Kumar, A. Serrano-Mislata, F. Madueño, P. Kraiewski, E. M. Meyerowitz, G. C. Angenent, and J. L. Riechmann. Orchestration of floral initiation by apetala1. *Science*, 328:85–89, 4 2010. ISSN 10959203. doi: 10.1126/SCIENCE.1185244.

O. Kelemen, P. Convertini, Z. Zhang, Y. Wen, M. Shen, M. Falaleeva, and S. Stamm. Function of alternative splicing. *Gene*, 514:1, 2 2013. ISSN 03781119. doi: 10.1016/J.GENE.2012.07.083.

L. A. Kelley, S. Mezulis, C. M. Yates, M. N. Wass, and M. J. Sternberg. The phyre2 web portal for protein modelling, prediction and analysis. *Nature protocols*, 10:845, 6 2015. ISSN 17502799. doi: 10.1038/NPROT.2015.053.

M. Khan, J. Y. Youn, A. C. Gingras, R. Subramaniam, and D. Desveaux. In planta proximity dependent biotin identification (bioid). *Scientific Reports*, 8:1–8, 12 2018. ISSN 20452322. doi: 10.1038/s41598-018-27500-3.

N. Khemka, V. K. Singh, R. Garg, and M. Jain. Genome-wide analysis of long intergenic non-coding rnas in chickpea and their potential role in flower development. *Scientific reports*, 6:33297, 2016. ISSN 2045-2322. doi: 10.1038/srep33297.

D. Kim, B. Langmead, and S. L. Salzberg. Hisat: a fast spliced aligner with low memory requirements. *Nature Methods 2015 12:4*, 12:357–360, 3 2015. ISSN 1548-7105. doi: 10.1038/nmeth.3317.

D.-H. Kim, Y. Xi, and S. Sung. Modular function of long noncoding rna, coldair, in the vernalization response. *PLOS Genetics*, 13:e1006939, 7 2017. ISSN 1553-7404. doi: 10.1371/journal.pgen.1006939.

D.-H. H. Kim and S. Sung. Vernalization-triggered intragenic chromatin loop formation by long noncoding rnas. *Developmental Cell*, 40:302–312.e4, 2 2017. ISSN 1534-5807. doi: 10.1016/J.DEVCEL.2016.12.021.

M. S. Kim, S. M. Pinto, D. Getnet, R. S. Nirujogi, S. S. Manda, R. Chaerkady, A. K. Madugundu, D. S. Kelkar, R. Isserlin, S. Jain, J. K. Thomas, B. Muthusamy, P. Leal-Rojas, P. Kumar, N. A. Sahasrabuddhe, L. Balakrishnan, J. Advani, B. George, S. Renuse, L. D. N. Selvan, A. H. Patil, V. Nanjappa, A. Radhakrishnan, S. Prasad, T. Subbannayya, R. Raju, M. Kumar, S. K. Sreenivasamurthy, A. Marimuthu, G. J. Sathe, S. Chavan, K. K. Datta, Y. Subbannayya, A. Sahu, S. D. Yelamanchi, S. Jayaram, P. Rajagopalan, J. Sharma, K. R. Murthy, N. Syed, R. Goel, A. A. Khan, S. Ahmad, G. Dey, K. Mudgal, A. Chatterjee, T. C. Huang, J. Zhong, X. Wu, P. G. Shaw, D. Freed, M. S. Zahari, K. K. Mukherjee, S. Shankar, A. Mahadevan, H. Lam, C. J. Mitchell, S. K. Shankar, P. Satishchandra, J. T. Schroeder, R. Sirdeshmukh, A. Maitra, S. D. Leach, C. G. Drake, M. K. Halushka, T. S. Prasad, R. H. Hruban, C. L. Kerr, G. D. Bader, C. A. Iacobuzio-Donahue, H. Gowda, and A. Pandey. A draft

map of the human proteome. *Nature 2014 509:7502*, 509:575–581, 5 2014. ISSN 1476-4687. doi: 10.1038/nature13302.

T.-W. Kim, C. H. Park, C.-C. Hsu, Y.-W. Kim, Y.-W. Ko, Z. Zhang, J.-Y. Zhu, Y.-C. Hsiao, T. Branon, K. Kaasik, E. Saldivar, K. Li, A. Pasha, N. J. Provart, A. L. Burlingame, S.-L. Xu, A. Y. Ting, and Z.-Y. Wang. Mapping the signaling network of bin2 kinase using turboid-mediated biotin labeling and phosphoproteomics. *The Plant Cell*, 35:975–993, 3 2023. ISSN 1040-4651. doi: 10.1093/PLCELL/KOAD013.

R. J. Kinsella, A. Kähäri, S. Haider, J. Zamora, G. Proctor, G. Spudich, J. Almeida-King, D. Staines, P. Derwent, A. Kerhornou, P. Kersey, and P. Flicek. Ensembl biomarts: a hub for data retrieval across taxonomic space. *Database*, 2011, 1 2011. ISSN 17580463. doi: 10.1093/DATABASE/BAR030.

K. Koizumi, T. Hayashi, and K. L. Gallagher. Scarecrow reinforces short-root signaling and inhibits periclinal cell divisions in the ground tissue by maintaining shr at high levels in the endodermis. *Plant Signaling Behavior*, 7: 1573, 12 2012. ISSN 15592324. doi: 10.4161/PSB.22437.

T. Kurata, T. Ishida, C. Kawabata-Awai, M. Noguchi, S. Hattori, R. Sano, R. Nagasaka, R. Tominaga, Y. Koshino-Kimura, T. Kato, S. Sato, S. Tabata, K. Okada, and T. Wada. Cell-to-cell movement of the caprice protein in arabidopsis root epidermal cell differentiation. *Development*, 132:5387–5398, 12 2005. ISSN 0950-1991. doi: 10.1242/DEV.02139.

S. Käppel, R. Eggeling, F. Rümpler, M. Groth, R. Melzer, and G. Theißen. Dna-binding properties of the mads-domain transcription factor sepallata3 and mutant variants characterized by selex-seq. *Plant Molecular Biology*, 105:543–557, 3 2021. ISSN 15735028. doi: 10.1007/S11103-020-01108-6.

X. Lai, R. Vega-Léon, V. Hugouvieux, R. Blanc-Mathieu, F. van der Wal, J. Lucas, C. S. Silva, A. Jourdain, J. M. Muino, M. H. Nanao, R. Immink, K. Kaufmann, F. Parcy, C. Smaczniak, and C. Zubieta. The intervening domain is required for dna-binding and functional identity of plant mads transcription factors. *Nature Communications*, 12, 12 2021. ISSN 20411723. doi: 10.1038/S41467-021-24978-W.

A. Lampropoulos, Z. Sutikovic, C. Wenzl, I. Maegele, J. U. Lohmann, and J. Forner. Greengate—a novel, versatile, and efficient cloning system for plant transgenesis. *PloS one*, 8:e83043, 2013. ISSN 1932-6203. doi: 10.1371/journal.pone.0083043.

J. H. Lee, J. J. Kim, and J. H. Ahn. Role of sepallata3 (sep3) as a downstream gene of mir156-spl3-ft circuitry in ambient temperature-responsive flowering. *Plant Signaling Behavior*, 7:1151, 9 2012. ISSN 15592316. doi: 10.4161/PSB.21366.

J. Y. Lee, S. F. Baum, J. Alvarez, A. Patel, D. H. Chitwood, and J. L. Bowman. Activation of crabs claw in the nectaries and carpels of arabidopsis. *The Plant Cell*, 17:25, 2005. ISSN 10404651. doi: 10.1105/TPC.104.026666.

M. M. Lee and J. Schiefelbein. Cell pattern in the arabidopsis root epidermis determined by lateral inhibition with feedback. *The Plant cell*, 14:611–8, 3 2002. ISSN 1040-4651.

D. A. Levin. The role of trichomes in plant defense. *https://doi.org/10.1086/407484*, 48:3–15, 3 1973. ISSN 0033-5770. doi: 10.1086/407484.

P. Li, Y. Meng, L. Wang, and L. J. Di. Bioid: A proximity-dependent labeling approach in proteomics study, 2019. ISSN 10643745.

S. Li, M. Yamada, X. Han, U. Ohler, and P. N. Benfey. High-resolution expression map of the arabidopsis root reveals alternative splicing and lincrna regulation. *Developmental cell*, 39:508–522, 11 2016. ISSN 1878-1551. doi: 10.1016/J.DEVCEL.2016.10.012.

Q. Lin, Z. Zhou, W. Luo, M. Fang, M. Li, and H. Li. Screening of proximal and interacting proteins in rice protoplasts by proximity-dependent biotinylation. *Front. Plant Sci.*, 8:1–10, 5 2017. ISSN 1664462X. doi: 10.3389/fpls.2017.00749.

C. Liu, W. Xi, L. Shen, C. Tan, and H. Yu. Regulation of floral patterning by flowering time genes. *Developmental Cell*, 16:711–722, 5 2009. ISSN 15345807. doi: 10.1016/j.devcel.2009.03.011.

X. Liu, Y. J. Kim, R. Müller, R. E. Yumul, C. Liu, Y. Pan, X. Cao, J. Goodrich, and X. Chen. Agamous terminates floral stem cell maintenance in arabidopsis by directly repressing wuschel through recruitment of polycomb group proteins. *The Plant cell*, 23:3654–3670, 2011. ISSN 1532-298X. doi: 10.1105/TPC.111.091538.

A. Mair, S. L. Xu, T. C. Branon, A. Y. Ting, and D. C. Bergmann. Proximity labeling of protein complexes and cell type specific organellar proteomes in arabidopsis enabled by turboid. *eLife*, 8, 9 2019. ISSN 2050084X. doi: 10.7554/eLife.47864.

M. A. Mandel and M. F. Yanofsky. A gene triggering flower formation in arabidopsis. *Nature*, 377:522–524, 10 1995. ISSN 0028-0836. doi: 10.1038/377522a0.

M. A. Mandel, C. Gustafson-Brown, B. Savidge, and M. F. Yanofsky. Molecular characterization of the arabidopsis floral homeotic gene apetala1. *Nature*, 360:273–277, 11 1992. ISSN 0028-0836. doi: 10.1038/360273a0.

O. Mantegazza, V. Gregis, M. Chiara, C. Selva, G. Leo, D. S. Horner, and M. M. Kater. Gene coexpression patterns during early development of the native arabidopsis reproductive meristem: novel candidate developmental regulators and patterns of functional redundancy. *The Plant journal : for cell and molecular biology*, 79:861–877, 2014. ISSN 1365-313X. doi: 10.1111/TPJ.12585.

M. D. Marsico, A. P. Gallart, W. Sanseverino, and R. A. Cigliano. Greenc 2.0: a comprehensive database of plant long non-coding rnas. *Nucleic Acids Research*, 50:D1442–D1447, 1 2022. ISSN 0305-1048. doi: 10.1093/NAR/GKAB1014.

K. E. McBride and K. R. Summerfelt. Improved binary vectors for agrobacterium-mediated plant transformation. *Plant molecular biology*, 14:269–276, 2 1990. ISSN 0167-4412. doi: 10.1007/BF00018567.

F. Meier, A. D. Brunner, S. Koch, H. Koch, M. Lubeck, M. Krause, N. Goedecke, J. Decker, T. Kosinski, M. A. Park, N. Bache, O. Hoerning, J. Cox, O. Räther, and M. Mann. Online parallel accumulation-serial

fragmentation (pasef) with a novel trapped ion mobility mass spectrometer. *Molecular cellular proteomics : MCP*, 17:2534–2545, 12 2018. ISSN 1535-9484. doi: 10.1074/MCP.TIR118.000900.

M. A. Mendes, R. F. Guerra, M. C. Berns, C. Manzo, S. Masiero, L. Finzi, M. M. Kater, and L. Colombo. Mads domain transcription factors mediate short-range dna looping that is essential for target gene expression in arabidopsis. *Plant Cell*, 25:2560–2572, 2013. ISSN 1532298X. doi: 10.1105/TPC.112.108688.

K. Morohashi and E. Grotewold. A systems approach reveals regulatory circuitry for arabidopsis trichome initiation by the gl3 and gl1 selectors. *PLoS Genetics*, 5, 2 2009. ISSN 15537390. doi: 10.1371/journal.pgen.1000396.

K. Morohashi, M. Zhao, M. Yang, B. Read, A. Lloyd, R. Lamb, and E. Grotewold. Participation of the arabidopsis bhlh factor gl3 in trichome initiation regulatory events. *Plant Physiology*, 145:736–746, 11 2007. ISSN 00320889. doi: 10.1104/pp.107.104521.

B. Mészáros, G. Erdös, and Z. Dosztányi. Iupred2a: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Research*, 46:W329–W337, 7 2018. ISSN 0305-1048. doi: 10.1093/NAR/GKY384.

A. Nickless, J. M. Bailis, and Z. You. Control of gene expression through the nonsense-mediated rna decay pathway. *Cell Bioscience 2017 7:1*, 7:1–12, 5 2017. ISSN 2045-3701. doi: 10.1186/S13578-017-0153-7.

Y. Ohashi, A. Oka, I. Ruberti, G. Morelli, and T. Aoyama. Entopically additive expression of glabra2 alters the frequency and spacing of trichome initiation. *The Plant Journal*, 29:359–369, 2 2002. ISSN 09607412. doi: 10.1046/j.0960-7412.2001.01214.x.

D. S. O'Maoileidigh, S. E. Wuest, L. Rae, A. Raganelli, P. T. Ryan, K. Kwasniewska, P. Das, A. J. Lohan, B. Loftus, E. Graciet, F. Wellmer, D. S. Ó'Maoiléidigh, S. E. Wuest, L. Rae, A. Raganelli, P. T. Ryan, K. Kwaśniewska, P. Das, A. J. Lohan, B. Loftus, E. Graciet, and F. Wellmer. Control of reproductive floral organ identity specification in

arabidopsis by the c function regulator agamous. *Plant Cell*, 25:2482–2503, 7 2013. ISSN 10404651. doi: 10.1105/tpc.113.113209.

D. S. O'Maoileidigh, E. Graciet, and F. Wellmer. Genetic control of arabidopsis flower development. *Advances in Botanical Research*, 72:159–190, 1 2014. ISSN 0065-2296. doi: 10.1016/B978-0-12-417162-6.00006-7.

K. J. Oparka, A. G. Roberts, P. Boevink, S. S. Cruz, I. Roberts, K. S. Pradel, A. Imlau, G. Kotlizky, N. Sauer, and B. Epel. Simple, but not branched, plasmodesmata allow the nonspecific trafficking of proteins in developing tobacco leaves. *Cell*, 97:743–754, 6 1999. ISSN 00928674. doi: 10.1016/S0092-8674(00)80786-2.

A. Pajoro, P. Madrigal, J. M. Muiño, J. T. Matus, J. Jin, M. A. Mecchia, J. M. Debernardi, J. F. Palatnik, S. Balazadeh, M. Arif, D. S. Ó'Maoiléidigh, F. Wellmer, P. Krajewski, J. L. Riechmann, G. C. Angenent, and K. Kaufmann. Dynamics of chromatin accessibility and gene regulation by mads-domain transcription factors in flower development. *Genome Biology*, 15, 3 2014. ISSN 1474760X. doi: 10.1186/GB-2014-15-3-R41.

Q. Pan, O. Shai, L. J. Lee, B. J. Frey, and B. J. Blencowe. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics*, 40:1413–1415, 12 2008. ISSN 10614036. doi: 10.1038/ng.259.

B. Patra, S. Pattanaik, and L. Yuan. Ubiquitin protein ligase 3 mediates the proteasomal degradation of glabrous 3 and enhancer of glabrous 3, regulators of trichome development and flavonoid biosynthesis in arabidopsis. *The Plant journal : for cell and molecular biology*, 74:435–447, 5 2013. ISSN 1365-313X. doi: 10.1111/TPJ.12132.

R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, and C. Kingsford. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods 2017 14:4*, 14:417–419, 3 2017. ISSN 1548-7105. doi: 10.1038/nmeth.4197.

A. L. Paul, P. C. Sehnke, and R. J. Ferl. Isoform-specific subcellular localization among 14-3-3 proteins in arabidopsis seems to be driven by client

interactions. *Molecular Biology of the Cell*, 16:1735–1743, 4 2005. ISSN 10591524. doi: 10.1091/MBC.E04-09-0839/ASSET/IMAGES/LARGE/ZMK0040530910004.JPEG.

C. T. Payne, F. Zhang, and A. M. Lloyd. Gl3 encodes a bhlh protein that regulates trichome development in arabidopsis through interaction with gl1 and ttg1. *Genetics*, 156:1349–1362, 2000. ISSN 00166731.

S. Pelaz, G. S. Ditta, E. Baumann, E. Wisman, and M. F. Yanofsky. B and c floral organ identity functions require sepallata mads-box genes. *Nature*, 405:200–203, 2000. ISSN 00280836. doi: 10.1038/35012103.

G. Pertea and M. Pertea. Gff utilities: Gffread and gffcompare. *F1000Research*, 9, 2020. ISSN 20461402. doi: 10.12688/F1000RESEARCH.23297.2/DOI.

M. Pesch and M. Hülskamp. Role of triptychon in trichome patterning in arabidopsis. *BMC Plant Biology*, 11:130, 9 2011. ISSN 1471-2229. doi: 10.1186/1471-2229-11-130.

M. Pesch, I. Schultheiß, S. Digiuni, J. F. Uhrig, and M. Hülskamp. Mutual control of intracellular localisation of the patterning proteins atmyc1, gl1 and try/cpc in arabidopsis. *Development (Cambridge, England)*, 140:3456–67, 8 2013. ISSN 09501991. doi: 10.1242/dev.094698.

M. Pesch, I. Schultheiß, K. Klopffleisch, J. F. Uhrig, M. Koegl, C. S. Clemen, R. Simon, S. Weidtkamp-Peters, and M. Hülskamp. Transparent testa glabra1 and glabra1 compete for binding to glabra3 in arabidopsis. *Plant Physiology*, 168:584–597, 2015. ISSN 15322548. doi: 10.1104/pp.15.00328.

H. Pimentel, N. L. Bray, S. Puente, P. Melsted, and L. Pachter. Differential analysis of rna-seq incorporating quantification uncertainty. *Nature Methods*, 14:687–690, 7 2017. ISSN 1548-7091. doi: 10.1038/nmeth.4324.

M. I. Rai, M. Alam, D. A. Lightfoot, P. Gurha, and A. J. Afzal. Classification and experimental identification of plant long non-coding rnas. *Genomics*, 4 2018. ISSN 0888-7543. doi: 10.1016/J.YGENO.2018.04.014.

N. Rani, T. J. Nowakowski, H. Zhou, S. E. Godshalk, V. Lisi, A. R. Kriegstein, and K. S. Kosik. A primate lncrna mediates notch signaling during

neuronal development by sequestering mirna. *Neuron*, 90:1174–1188, 6 2016. ISSN 10974199. doi: 10.1016/j.neuron.2016.05.005.

O. Ratcliffe, D. Bradley, E. Coen, G. Ditta, P. Zambryski, L. J. Feldman, and M. F. Yanofsky. Separation of shoot and floral identity in arabidopsis. *Development*, 126:1109–1120, 2 2000. ISSN 0950-1991.

W. G. Rerie, K. A. Feldmann, and M. D. Marks. The glabra2 gene encodes a homeo domain protein required for normal trichome development in arabidopsis. *Genes and Development*, 8:1388–1399, 6 1994. ISSN 08909369. doi: 10.1101/gad.8.12.1388.

P. Rice, L. Longden, and A. Bleasby. Emboss: the european molecular biology open software suite. *Trends in genetics : TIG*, 16:276–277, 6 2000. ISSN 0168-9525. doi: 10.1016/S0168-9525(00)02024-2.

T. A. Richmond and A. B. Bleecker. A defect in b-oxidation causes abnormal inflorescence development in arabidopsis. *The Plant Cell*, 11:1911, 10 1999. ISSN 10404651. doi: 10.2307/3871086.

J. L. Riechmann and E. M. Meyerowitz. Determination of floral organ identity by arabidopsis mads domain homeotic proteins ap1, ap3, pi, and ag is independent of their dna-binding specificity. *Molecular Biology of the Cell*, 8:1243–1259, 1997. ISSN 10591524. doi: 10.1091/MBC.8.7.1243.

J. L. Riechmann, B. A. Krizek, and E. M. Meyerowitz. Dimerization specificity of arabidopsis mads domain homeotic proteins apetala1, apetala3, pistillata, and agamous. *Proceedings of the National Academy of Sciences of the United States of America*, 93:4793–4798, 5 1996. ISSN 0027-8424. doi: 10.1073/PNAS.93.10.4793.

J. L. Rinn, M. Kertesz, J. K. Wang, S. L. Squazzo, X. Xu, S. A. Brugmann, L. H. Goodnough, J. A. Helms, P. J. Farnham, E. Segal, and H. Y. Chang. Functional demarcation of active and silent chromatin domains in human hox loci by noncoding rnas. *Cell*, 129:1311–1323, 6 2007. ISSN 00928674. doi: 10.1016/j.cell.2007.05.022.

J. T. Robinson, H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. S. Lander, G. Getz, and J. P. Mesirov. Integrative genomics viewer. *Nature Biotechnology 2011 29:1*, 29:24–26, 1 2011. ISSN 1546-1696. doi: 10.1038/nbt.1754.

M. D. Robinson and M. Nowicka. Drimseq: a dirichlet-multinomial framework for multivariate count outcomes in genomics. *F1000Research*, 5, 2016. ISSN 1759796X. doi: 10.12688/F1000RESEARCH.8900.2.

K. J. Roux, D. I. Kim, M. Raida, and B. Burke. A promiscuous biotin ligase fusion protein identifies proximal and interacting proteins in mammalian cells. *Journal of Cell Biology*, 196:801–810, 3 2012. ISSN 00219525. doi: 10.1083/jcb.201112098.

P. T. Ryan, D. S. Ó'Maoiléidigh, H. G. Drost, K. Kwaśniewska, A. Gabel, I. Grosse, E. Graciet, M. Quint, and F. Wellmer. Patterns of gene expression during arabidopsis flower development from the time of initiation to maturation. *BMC Genomics*, 16:1–12, 12 2015a. ISSN 14712164. doi: 10.1186/S12864-015-1699-6/FIGURES/7.

P. T. Ryan, D. S. Ó'Maoiléidigh, H.-G. Drost, K. Kwaśniewska, A. Gabel, I. Grosse, E. Graciet, M. Quint, and F. Wellmer. Patterns of gene expression during arabidopsis flower development from the time of initiation to maturation. *BMC Genomics*, 16:488, 12 2015b. ISSN 1471-2164. doi: 10.1186/s12864-015-1699-6.

R. Saedler, M. Jakoby, B. Marin, E. Galiana-Jaime, and M. Hülskamp. The cell morphogenesis gene spirrig in arabidopsis encodes a wd/beach domain protein. *The Plant journal : for cell and molecular biology*, 59:612–621, 8 2009. ISSN 1365-313X. doi: 10.1111/J.1365-313X.2009.03900.X.

N. Saitou and M. Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4: 406–425, 7 1987. ISSN 0737-4038. doi: 10.1093/OXFORDJOURNALS.MOLBEV.A040454.

H. Sakai, B. A. Krizek, S. E. Jacobsen, and E. M. Meyerowitz. Regulation of sup expression identifies multiple regulators involved in arabidopsis floral meristem development. *The Plant Cell*, 12:1607, 2000. ISSN 10404651. doi: 10.1105/TPC.12.9.1607.

I. Sarropoulos, R. Marin, M. Cardoso-Moreira, and H. Kaessmann. Developmental dynamics of lncrnas across mammalian organs and species. *Nature*, 571:510–514, 7 2019. ISSN 14764687. doi: 10.1038/s41586-019-1341-x.

S. Schellmann, A. Schnittger, V. Kirik, T. Wada, K. Okada, A. Beermann, J. Thumfahrt, G. Jürgens, M. Hülskamp, A. Beerman, J. Thumfahrt, G. Jürgens, and M. Hülskamp. Triptychon and caprice mediate lateral inhibition during trichome and root hair patterning in arabidopsis. *The EMBO Journal*, 21:5036–5046, 10 2002. ISSN 0261-4189. doi: 10.1093/emboj/cdf524.

J. Schiefelbein. Cell-fate specification in the epidermis: a common patterning mechanism in the root and shoot. *Current Opinion in Plant Biology*, 6: 74–78, 2 2003. ISSN 1369-5266. doi: 10.1016/S136952660200002X.

H. Schoof, M. Lenhard, A. Haecker, K. F. Mayer, G. Jürgens, and T. Laux. The stem cell population of arabidopsis shoot meristems is maintained by a regulatory loop between the clavata and wuschel genes. *Cell*, 100: 635–644, 3 2000. ISSN 00928674. doi: 10.1016/S0092-8674(00)80700-X.

E. A. Schultz and G. W. Haughn. Leafy, a homeotic gene that regulates inflorescence development in arabidopsis. *The Plant Cell*, 3:771–781, 8 1991. ISSN 1040-4651. doi: 10.1105/tpc.3.8.771.

D. Sehnal, S. Bittrich, M. Deshpande, R. Svobodová, K. Berka, V. Bazgier, S. Velankar, S. K. Burley, J. Koča, and A. S. Rose. Mol* viewer: modern web app for 3d visualization and analysis of large biomolecular structures. *Nucleic Acids Research*, 49:W431–W437, 7 2021. ISSN 0305-1048. doi: 10.1093/NAR/GKAB314.

M. Seki, E. Katsumata, A. Suzuki, S. Sereewattanawoot, Y. Sakamoto, J. Mizushima-Sugano, S. Sugano, T. Kohno, M. C. Frith, K. Tsuchihara, and Y. Suzuki. Evaluation and application of rna-seq by minion. *DNA Research*, 26:55–65, 2 2019. ISSN 1340-2838. doi: 10.1093/dnares/dsy038.

J. S. Seo, H. X. Sun, B. S. Park, C. H. Huang, S. D. Yeh, C. Jung, and N. H. Chua. Elf18-induced long-noncoding rna associates with mediator to enhance expression of innate immune response genes in arabidopsis. *The Plant cell*, 29:1024–1038, 5 2017. ISSN 1532-298X. doi: 10.1105/TPC.16.00886.

S. Shannon and D. R. Meeks-Wagner. A mutation in the arabidopsis tfl1 gene affects inflorescence meristem development. *THE PLANT CELL ONLINE*, 3:877–892, 9 1991. ISSN 10404651. doi: 10.1105/tpc.3.9.877.

C. C. Sheldon, D. T. Rouse, E. J. Finnegan, W. J. Peacock, and E. S. Dennis. The molecular basis of vernalization: The central role of flowering locus c (flc). *Proceedings of the National Academy of Sciences of the United States of America*, 97:3753–3758, 3 2000. ISSN 00278424. doi: 10.1073/pnas.97.7.3753.

F. Sievers, A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, J. D. Thompson, and D. G. Higgins. Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Molecular Systems Biology*, 7:539, 1 2011. ISSN 1744-4292. doi: 10.1038/MSB.2011.75.

N. M. Simon, J. Sugisaka, M. N. Honjo, S. A. Tunstad, G. Tunna, H. Kudoh, and A. N. Dodd. Altered stomatal patterning accompanies a trichome dimorphism in a natural population of arabidopsis. *Plant Direct*, 4:e00262, 9 2020. ISSN 2475-4455. doi: 10.1002/PLD3.262.

C. M. A. Simopoulos, E. A. Weretilnyk, and G. B. Golding. Prediction of plant lncrna by ensemble machine learning classifiers. *BMC Genomics*, 19:316, 12 2018. ISSN 1471-2164. doi: 10.1186/s12864-018-4665-2.

J. Sirén, N. Välimäki, and V. Mäkinen. Indexing graphs for path queries with applications in genome research. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 11:375–388, 3 2014. ISSN 15455963. doi: 10.1109/TCBB.2013.2297101.

P. J. Skene and S. Henikoff. An efficient targeted nuclease strategy for high-resolution mapping of dna binding sites. *eLife*, 6, 1 2017. ISSN 2050084X. doi: 10.7554/ELIFE.21856.

C. Smaczniak, R. G. Immink, J. M. Muiño, R. Blanvillain, M. Busscher, J. Busscher-Lange, Q. D. Dinh, S. Liu, A. H. Westphal, S. Boeren, F. Parcy, L. Xu, C. C. Carles, G. C. Angenent, and K. Kaufmann. Characterization of mads-domain transcription factor complexes in arabidopsis flower development. *Proceedings of the National Academy of Sciences of the United States of America*, 109:1560–1565, 1 2012a. ISSN 00278424. doi: 10.1073/PNAS.1112871109.

C. Smaczniak, R. G. H. Immink, J. M. Muiño, R. Blanvillain, M. Busscher, J. Busscher-Lange, S. Liu, A. H. Westphal, S. Boeren, F. Parcy, L. Xu,

C. C. Carles, G. C. Angenent, and K. Kaufmann. Characterization of mads-domain transcription factor complexes in arabidopsis flower development. 109, 2012b. doi: 10.1073/pnas.1112871109.

C. Smaczniak, J. M. Muiño, D. Chen, G. C. Angenent, and K. Kaufmann. Differences in dna binding specificity of floral homeotic protein complexes predict organ-specific target genes. *Plant Cell*, 29:1822–1835, 2017. ISSN 1532298X. doi: 10.1105/TPC.17.00145.

D. R. Smyth, J. L. Bowman, and E. M. Meyerowitz. Early flower development in arabídopsis, 1990.

C. Soneson, Y. Yao, A. Bratus-Neuenschwander, A. Patrignani, M. D. Robinson, and S. Hussain. A comprehensive examination of nanopore native rna sequencing for characterization of complex transcriptomes. *Nature Communications*, 10:3359, 12 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-11272-z.

S. K. Song, K. H. Ryu, Y. H. Kang, J. H. Song, Y. H. Cho, S. D. Yoo, J. Schiefelbein, and M. M. Lee. Cell fate in the arabidopsis root epidermis is determined by competition between werewolf and caprice. *Plant Physiology*, 157:1196–1208, 11 2011. ISSN 00320889. doi: 10.1104/pp.111.185785.

A. Steffens, A. Bräutigam, M. Jakoby, and M. Hülskamp. The beach domain protein spirrig is essential for arabidopsis salt stress tolerance and functions as a regulator of transcript stabilization and localization. *PLOS Biology*, 13:e1002188, 2015. ISSN 1545-7885. doi: 10.1371/JOURNAL.PBIO.1002188.

C. H. Su, D. Dhananjaya, and W. Y. Tarn. Alternative splicing in neurogenesis and brain development, 2 2018. ISSN 2296889X.

S. Sureshkumar, C. Dent, A. Seleznev, C. Tasset, and S. Balasubramanian. Nonsense-mediated mrna decay modulates flm-dependent thermosensory flowering response in arabidopsis. *Nature Plants*, 2:1–7, 5 2016. ISSN 20550278. doi: 10.1038/NPLANTS.2016.55.

D. B. Szymanski and M. D. Marks. Glabrous1 overexpression and triptychon alter the cell cycle and trichome cell fate in arabidopsis. *The Plant cell*, 10:2047–62, 12 1998. ISSN 1040-4651. doi: 10.1105/TPC.10.12.2047.

A. D. Tang, C. M. Soulette, M. J. van Baren, K. Hart, E. Hrabeta-Robinson, C. J. Wu, and A. N. Brooks. Full-length transcript characterization of sf3b1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *Nature Communications 2020 11:1*, 11:1–12, 3 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-15171-6.

A. Telfer, K. M. Bollman, R. S. Poethig, S. Ray, P. C. Zambryski, L. J. Feldman, and M. F. Yanofsky. Phase change and the regulation of trichome distribution in arabidopsis thaliana. *Development (Cambridge, England)*, 124:645–54, 2 1997. ISSN 0950-1991.

G. Theißen. Development of floral organ identity: stories from the mads house. *Current opinion in plant biology*, 4:75–85, 2001. ISSN 1369-5266. doi: 10.1016/S1369-5266(00)00139-4.

G. Theißen, R. Melzer, and F. Ruümpler. Mads-domain transcription factors and the floral quartet model of flower development: linking plant development and evolution. *Development (Cambridge, England)*, 143:3259–3271, 9 2016. ISSN 1477-9129. doi: 10.1242/DEV.134080.

B. Thomson. Analysis of stage-specific gene perturbations and characterisation of two novel f-box genes during flower development in arabidopsis thaliana. 2018.

B. Thomson, E. Graciet, and F. Wellmer. Inducible promoter systems for gene perturbation experiments in arabidopsis. *Methods in molecular biology (Clifton, N.J.)*, 1629:15–25, 2017. ISSN 1940-6029. doi: 10.1007/978-1-4939-7125-1_2.

C. Tian, Y. Wang, H. Yu, J. He, J. Wang, B. Shi, Q. Du, N. J. Provart, E. M. Meyerowitz, and Y. Jiao. A gene expression map of shoot domains reveals regulatory mechanisms. *Nature Communications*, 10:141, 12 2019. ISSN 2041-1723. doi: 10.1038/s41467-018-08083-z.

R. Tominaga, M. Iwata, K. Okada, and T. Wada. Functional analysis of the epidermal-specific myb genes caprice and werewolf in arabidopsis. *The Plant cell*, 19:2264–77, 7 2007. ISSN 1040-4651. doi: 10.1105/tpc.106.045732.

R. Tominaga-Wada and T. Wada. Analysis of ttg1 and cpc-like myb genes during arabidopsis epidermal cell differentiation. *Plant Biotechnology*, 33: 1–6, 2016. doi: 10.5511/plantbiotechnology.16.0629a.

R. Tominaga-Wada and T. Wada. Extended c termini of cpc-like myb proteins confer functional diversity in arabidopsis epidermal cell differentiation. *Development (Cambridge, England)*, 144:2375–2380, 7 2017. ISSN 1477-9129. doi: 10.1242/dev.149542.

L. S. Toni, A. M. Garcia, D. A. Jeffrey, X. Jiang, B. L. Stauffer, S. D. Miyamoto, and C. C. Sucharov. Optimization of phenol-chloroform rna extraction. *MethodsX*, 5:599–608, 1 2018. ISSN 2215-0161. doi: 10.1016/J.MEX.2018.05.011.

J. L. Trincado, J. C. Entizne, G. Hysenaj, B. Singh, M. Skalic, D. J. Elliott, and E. Eyras. Suppa2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome biology*, 19, 3 2018. ISSN 1474-760X. doi: 10.1186/S13059-018-1417-1.

L. Trinkle-Mulcahy. Recent advances in proximity-based labeling methods for interactome mapping [version 1; referees: 2 approved], 2019. ISSN 1759796X.

B. Uszczynska-Ratajczak, J. Lagarde, A. Frankish, R. Guigó, and R. Johnson. Towards a complete map of the human long non-coding rna transcriptome. *Nature reviews. Genetics*, 19:535, 9 2018. ISSN 14710064. doi: 10.1038/S41576-018-0017-Y.

J. P. Verta and A. Jacobs. The role of alternative splicing in adaptation and evolution. *Trends in Ecology and Evolution*, 37:299–308, 4 2022. ISSN 01695347. doi: 10.1016/j.tree.2021.11.010.

H. Wang, J. Wang, J. Jiang, S. Chen, Z. Guan, Y. Liao, and F. Chen. Reference genes for normalizing transcription in diploid and tetraploid arabidopsis. *Scientific Reports 2014 4:1*, 4:1–8, 10 2014. ISSN 2045-2322. doi: 10.1038/srep06781.

S. Wang, L. Hubbard, Y. Chang, J. Guo, J. Schiefelbein, and J.-G. Chen. Comprehensive analysis of single-repeat r3 myb proteins in epidermal cell

patterning and their transcriptional regulation in arabidopsis. *BMC plant biology*, 8:81, 7 2008. ISSN 1471-2229. doi: 10.1186/1471-2229-8-81.

X. Wang, C. Shen, P. Meng, G. Tan, and L. Lv. Analysis and review of trichomes in plants. *BMC Plant Biology 2021 21:1*, 21:1–11, 2 2021. ISSN 1471-2229. doi: 10.1186/S12870-021-02840-X.

Y. Wang, X. Luo, F. Sun, J. Hu, X. Zha, W. Su, and J. Yang. Overexpressing lncrna lair increases grain yield and regulates neighbouring gene cluster expression in rice. *Nature Communications*, 9:3516, 12 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-05829-7.

D. Weigel, J. Alvarez, D. R. Smyth, M. F. Yanofsky, and E. M. Meyerowitz. Leafy controls floral meristem identity in arabidopsis. *Cell*, 69:843–59, 5 1992. ISSN 0092-8674. doi: 10.1016/0092-8674(92)90295-N.

J. L. Weirather, M. de Cesare, Y. Wang, P. Piazza, V. Sebastiano, X.-J. Wang, D. Buck, and K. F. Au. Comprehensive comparison of pacific biosciences and oxford nanopore technologies and their applications to transcriptome analysis. *F1000Research*, 6:100, 6 2017. ISSN 2046-1402. doi: 10.12688/f1000research.10571.2.

B. Weisburd, G. Tiao, and H. L. Rehm. Insights from a genome-wide truth set of tandem repeat variation. doi: 10.1101/2023.05.05.539588.

F. Wellmer, J. L. Riechmann, M. Alves-Ferreira, and E. M. Meyerowitz. Genome-wide analysis of spatial gene expression in arabidopsis flowers. *Plant Cell*, 16:1314–1326, 2004. ISSN 10404651. doi: 10.1105/TPC.021741.

F. Wellmer, E. Graciet, and J. L. Riechmann. Specification of floral organs in arabidopsis. *Journal of Experimental Botany*, 65:1–9, 1 2014. ISSN 14602431. doi: 10.1093/JXB/ERT385.

K. Wester, S. Digiuni, F. Geier, J. Timmer, C. Fleck, and M. Hülskamp. Functional diversity of r3 single-repeat genes in trichome development. *Development (Cambridge, England)*, 136:1487–1496, 5 2009. ISSN 0950-1991. doi: 10.1242/DEV.021733.

S. M. Weyn-Vanhentenryck, H. Feng, D. Ustianenko, R. Duffié, Q. Yan, M. Jacko, J. C. Martinez, M. Goodwin, X. Zhang, U. Hengst, S. Lomvardas, M. S. Swanson, and C. Zhang. Precise temporal regulation of alternative splicing during neural development. *Nature Communications*, 9:1–17, 12 2018. ISSN 20411723. doi: 10.1038/s41467-018-04559-0.

H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4.

C. M. Winter, R. S. Austin, S. Blanvillain-Baufumé, M. A. Reback, M. Monniaux, M. F. Wu, Y. Sang, A. Yamaguchi, N. Yamaguchi, J. E. Parker, F. Parcy, S. T. Jensen, H. Li, and D. Wagner. Leafy target genes reveal floral regulatory logic, cis motifs, and a link to biotic stimulus response. *Developmental Cell*, 20:430–443, 4 2011. ISSN 1534-5807. doi: 10.1016/J.DEVCEL.2011.03.019.

S. E. Wuest, D. S. O'Maoileidigh, L. Rae, K. Kwasniewska, A. Raganelli, K. Hanczaryk, A. J. Lohan, B. Loftus, E. Graciet, and F. Wellmer. Molecular basis for the specification of floral organs by apetala3 and pistillata. *Proceedings of the National Academy of Sciences of the United States of America*, 109:13452–13457, 8 2012. ISSN 00278424. doi: 10.1073/PNAS.1207075109.

A. Wutz and R. Jaenisch. A shift from reversible to irreversible x inactivation is triggered during es cell differentiation. *Molecular Cell*, 5:695–705, 2000. ISSN 10972765. doi: 10.1016/S1097-2765(00)80248-8.

K. Yamada, M. Sasabe, Y. Fujikawa, T. Wada, and R. Tominaga-Wada. Amino acid substitutions in cpc-like myb reveal residues important for protein stability in arabidopsis roots. *PLOS ONE*, 13:e0205522, 10 2018. ISSN 1932-6203. doi: 10.1371/JOURNAL.PONE.0205522.

N. Yamaguchi, J. Huang, Y. Xu, K. Tanoi, and T. Ito. Fine-tuning of auxin homeostasis governs the transition from floral stem cell maintenance to gynoecium formation. *Nature Communications*, 8:1–15, 12 2017. ISSN 20411723. doi: 10.1038/s41467-017-01252-6.

W. Yan, D. Chen, and K. Kaufmann. Molecular mechanisms of floral organ specification by mads domain proteins. *Current Opinion in Plant Biology*, 29:154–162, 2 2016. ISSN 13695266. doi: 10.1016/J.PBI.2015.12.004.

X. Yang, J. Coulombe-Huntington, S. Kang, G. M. Sheynkman, T. Hao, A. Richardson, S. Sun, F. Yang, Y. A. Shen, R. R. Murray, K. Spirohn, B. E. Begg, M. Duran-Frigola, A. MacWilliams, S. J. Pevzner, Q. Zhong, S. A. Trigg, S. Tam, L. Ghamsari, N. Sahni, S. Yi, M. D. Rodriguez, D. Balcha, G. Tan, M. Costanzo, B. Andrews, C. Boone, X. J. Zhou, K. Salehi-Ashtiani, B. Charloteaux, A. A. Chen, M. A. Calderwood, P. Aloy, F. P. Roth, D. E. Hill, L. M. Iakoucheva, Y. Xia, and M. Vidal. Widespread expansion of protein interaction capabilities by alternative splicing. *Cell*, 164:805–817, 2 2016. ISSN 10974172. doi: 10.1016/j.cell.2016.01.029.

Y. Yang and T. Jack. Defining subdomains of the k domain important for protein-protein interactions of plant mads proteins. *Plant Molecular Biology*, 55:45–59, 5 2004. ISSN 01674412. doi: 10.1007/s11103-004-0416-7.

Y. Yang, L. Fanning, and T. Jack. The k domain mediates heterodimerization of the arabidopsis floral organ identity proteins, apetala3 and pistillata. *The Plant journal : for cell and molecular biology*, 33:47–59, 1 2003a. ISSN 0960-7412. doi: 10.1046/J.0960-7412.2003.01473.X.

Y. Yang, H. Xiang, and T. Jack. pistillata-5, an arabidopsis b class mutant with strong defects in petal but not in stamen development. *Plant Journal*, 33:177–188, 1 2003b. ISSN 09607412. doi: 10.1046/j.1365-313X.2003.01603.x.

M. F. Yanofsky, H. Ma, J. L. Bowman, G. N. Drews, K. A. Feldmann, and E. M. Meyerowitz. The protein encoded by the arabidopsis homeotic gene agamous resembles transcription factors. *Nature*, 346:35–39, 1990. ISSN 00280836. doi: 10.1038/346035A0.

F. Zhang. A network of redundant bhlh proteins functions in all ttg1-dependent pathways of arabidopsis. *Development*, 130:4859–4869, 10 2003. ISSN 0950-1991. doi: 10.1242/dev.00681.

L. Zhang, L. Wang, Y. Yang, J. Cui, F. Chang, Y. Wang, and H. Ma. Analysis of arabidopsis floral transcriptome: Detection of new florally expressed genes and expansion of brassicaceae-specific gene families. *Frontiers in Plant Science*, 5:1–9, 1 2015. ISSN 1664462X. doi: 10.3389/FPLS.2014.00802/ABSTRACT.

R. Zhang, C. P. G. Calixto, Y. Marquez, P. Venhuizen, N. A. Tzioutziou, W. Guo, M. Spensley, J. C. Entizne, D. Lewandowska, S. T. Have, N. F. D. Frey, H. Hirt, A. B. James, H. G. Nimmo, A. Barta, M. Kalyna, and J. W. S. Brown. A high quality arabidopsis transcriptome for accurate transcript-level analysis of alternative splicing. *Nucleic acids research*, 45: 5061–5073, 5 2017. ISSN 1362-4962. doi: 10.1093/nar/gkx267.

X. Zhang, R. Henriques, S. S. Lin, Q. W. Niu, and N. H. Chua. Agrobacterium-mediated transformation of arabidopsis thaliana using the floral dip method. *Nature Protocols 2006 1:2*, 1:641–646, 6 2006. ISSN 1750-2799. doi: 10.1038/nprot.2006.97.

Y. Zhang, G. Song, N. K. Lal, U. Nagalakshmi, Y. Li, W. Zheng, P. jui Huang, T. C. Branon, A. Y. Ting, J. W. Walley, and S. P. Dinesh-Kumar. Turboid-based proximity labeling reveals that ubr7 is a regulator of n nlr immune receptor-mediated immunity. *Nature Communications 2019 10:1*, 10:1–17, 7 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-11202-z.

Y. Zhang, Y. Li, X. Yang, Z. Wen, U. Nagalakshmi, and S. P. Dinesh-Kumar. Turboid-based proximity labeling for in planta identification of protein-protein interaction networks. *Journal of visualized experiments : JoVE*, 2020, 5 2020. ISSN 1940-087X. doi: 10.3791/60728.

L. Zhao, H. Zhang, M. V. Kohnen, K. V. S. K. Prasad, L. Gu, and A. S. N. Reddy. Analysis of transcriptome and epitranscriptome in plants using pacbio iso-seq and nanopore-based direct rna sequencing. *Frontiers in Genetics*, 10:253, 3 2019. ISSN 1664-8021. doi: 10.3389/fgene.2019.00253.

H. Zheng, M. Hernaez, K. Brennan, and O. Gevaert. Abstract 2472: Benchmark of lncrna quantification in rna-seq of cancer samples. *Cancer Research*, 78:2472–2472, 7 2018. ISSN 0008-5472. doi: 10.1158/1538-7445.AM2018-2472.

D. S. Ó'Maoiléidigh, S. E. Wuest, L. Rae, A. Raganelli, P. T. Ryan, K. Kwaśniewska, P. Das, A. J. Lohan, B. Loftus, E. Graciet, and F. Wellmer. Control of reproductive floral organ identity specification in arabidopsis by the c function regulator agamous. *The Plant Cell*, 25: 2482, 2013. ISSN 1532298X. doi: 10.1105/TPC.113.113209.

D. S. Ó'Maoiléidigh, B. Thomson, A. Raganelli, S. E. Wuest, P. T. Ryan, K. Kwasniewska, C. C. Carles, E. Graciet, and F. Wellmer. Gene network analysis of arabidopsis thaliana flower development through dynamic gene perturbations. *The Plant Journal*, 83:344–358, 7 2015. ISSN 1365-313X. doi: 10.1111/TPJ.12878.

D. S. Ó'Maoiléidigh, D. Stewart, B. Zheng, G. Coupland, F. Wellmer, D. S. Ó'Maoiléidigh, D. Stewart, B. Zheng, G. Coupland, and F. Wellmer. Floral homeotic proteins modulate the genetic program for leaf development to suppress trichome formation in flowers. *Development*, 145:dev.157784, 2 2018. ISSN 0950-1991. doi: 10.1242/dev.157784.