Development of a statistical tool for comparative analysis of gene expression dynamics and its application using Brassica and Arabidopsis transcriptomic data

Ruth Kristianingsih

(100317409)

A thesis presented for the degree of
Doctor of Philosophy

John Innes Centre, UK
University of East Anglia, UK
May 2025

Abstract

Comparing gene expression patterns can help identify correspondences of developmental stages within and between species, highlight differences in the timing of key developmental events, and elucidate transcriptional responses to treatments. However, such comparisons are often complicated by variations in timing and the differing timescales of these events. To overcome this challenge, we developed a method based on curve registration, which optimally aligns gene expression dynamics by inferring temporal shifts and stretches. Statistical evaluation of these parameters allows us to compare the fit of a non-registered model (in which expression profiles are considered different) against a registered model (in which differences are resolved through alignment). To make this approach widely accessible, we implemented it as an R package, *greatR*. This tool has been validated using various datasets, both simulated datasets and real biological data. *greatR* has been successfully applied to multiple comparisons, including the floral transition in Arabidopsis, *B. rapa*, and *B. oleracea*, as well as across these species. Additionally, we employed *greatR* to compare expression profiles of two Arabidopsis genotypes during bract formation, offering new insights into the genetic and transcriptional mechanisms underlying this trait. Beyond plant systems, *greatR* can be extended to compare expression responses in other organisms, making it a valuable tool for cross-species analysis. *greatR* has proven to be able to detect pairs of genes with expression profiles which can be superimposed and, therefore, have similar dynamics. This approach enables the exploration of dynamic differences in gene expression within and across species, providing an important foundation for understanding the regulatory networks that govern various biological processes. By comparing these dynamics, it can help uncover both conserved and species-specific regulatory mechanisms. This approach facilitates the transfer of knowledge from well-studied model organisms to less-explored species, the identification of co-regulated gene modules, and the discovery of temporally differentially expressed genes linked to specific conditions or traits.

To my mother,
I hope I've made you proud, Ibuk, and somewhere, you are smiling.

Acknowledgements

One of the mottos that has guided me through life is, "You can if you think you can". Yet, I've come to realise that I couldn't have made this statement true for myself without the immense support, love, and resources I received during my PhD journey.

I would like to express my gratitude to the John Innes Foundation for providing me with the funding that made the rotation PhD programme possible.

I am deeply thankful to my supervisor, Richard Morris, without whom this adventure would not have been possible. Thank you, Richard, for welcoming me into the More Ice group, for your guidance, and belief in me, and for always having your door open when I needed help. Thank you for training me to be better at writing and formulating ideas. Also thank you for teaching me on how to be a better presenter and storyteller. Thank you for the countless fun jokes and delicious food (especially the amazing flammkuchen). Thank you for giving me all the opportunities. Your support and trust helped me navigate the roller coasters of my PhD journey, and for that, I am eternally grateful.

I am also very thankful to my supervisory team: Rachel Wells, Hugh Woolfenden, and Smita Kurup for all their guidance in weekly meetings and for always being there to support me. Thank you for your feedback on my thesis.

Rachel, thank you for your enthusiasm, especially when it came to answering my endless Brassica-related questions. Your excitement and encouragement have been a constant source of motivation. Thank you for always being there for me. Thank you for the biscuit supply, thank you for introducing us to broken biscuits, it is a real gem. Thank you for believing in me, for checking in on me regularly, and for your constant support. I can't thank you enough.

Thank you Hugh for your guidance throughout my PhD and thank you for helping me with my thesis chapters and going through each of your corrections one by one. Thank you for all of the advice you have given me. Thank you for all the inspiring stories about your running journey. You taught me to be more discipline even in such simpler thing such as keeping the streak in Duolingo. Thank you for checking on me and for your kindness. Thank you so much.

I would like to thank all the More Ice members, the Postdocs, and the students, you have been a great team to work with and have made group meetings and socials enjoyable. I would also like to give a special thank you to the present member Aileen Magilin–girl, you've been my rock! Thank you so much for the last couple of years we spent together in the lab. Also thank you to the past lab member Shannon Woodhouse who patiently guided me as I entered the fun Brassica world. Thank you for your mentoring during my rotation.

I would like to sincerely thank George Lomonossoff for the opportunity to spend my rotation weeks in your lab, despite my limited experimental skills. I am incredibly grateful to have been part of your team and for the opportunity to contribute to one of your projects as a co-author.

A special thank you to all the members of George's lab, especially Yulia, for her patience in teaching me wet lab techniques. Yulia, your guidance and support have been invaluable, and I am truly thankful for the experience.

I would like to thank our collaborators in France: Sana Dieudonné and Fabrice Besnard. Thank you for the opportunity and for welcoming me to Lyon and in the lab. It was a great experience to be able to visit France. Thank you Fabrice for always being so enthusiastic with my analysis in every catch-up meeting. I am deeply grateful for the wonderful collaboration and the opportunity to work alongside such dedicated and supportive colleagues.

I would like to thank all of my friends who gave me endless support and help during my PhD. To all of my friends in Norwich, thank you so much for the great time we had during my PhD: for all the dinners, board games, chats, cooking together, walks, and many many more great time spent together. To Julia, thank you for always checking in on me, chica. Your constant care, belief in me, and your willingness to listen to all my stories mean a lot to me. I can't thank you enough. To Sam, I'm so grateful for your advice and always being there whenever I needed it. Your positive affirmations and belief in me kept me going. Jiawen, thank you for reminding me to find joy in the little things and for always being so positive. To my barbiologist gang: Julia, Aileen, Thomasito, Tom, and Pablo, you guys are just amazing people sent for me. Thank you to you guys wonderful people for being there always during my PhD. To Diana, girl thank you so much for your support and always checking on me, thank you for the chocolates and sweets that kept me going during my writing. To Adeline and Pai, thank you for the lunches we have together. To all my friends in Norwich and the UK, Neftaly, Isabel, Emad, Ilina, and Andrei: thank you so much guys for all of the support during my PhD and writing. To my writing buddy, Caroline. Caroline, thank you for the check-ins, the lunches, the coffee breaks, the delicious banana cake, and the dark chocolates that kept me going. I am so grateful for your friendship.

Thank you to all my family in Indonesia, Chile, and Spain. You have been a great motivation for me to keep going. I would like to give my special thank you to my husband, Alfredo. Carinyo, thank you so much for everything, and literally everything. For those nights you spent accompanying me writing, for all the meals you cook when I was busy working and writing, for all of the walks and great discussions. Thank you for helping me when I was stuck with my codes, thank you for all of your help when I had questions about technical details. Thank you for never having any doubt in me. Thank you so much for your endless love and I would not have done this without you.

Ruth Kristianingsih
Norwich, United Kingdom

September 2024

List of Publications

This thesis includes material from the following work.

S. Dieudonné, R. Kristianingsih, S. Laine, B. Jesson, V. Vidal, R. Wells, R. Morris and F. Besnard. Natural variation suggests new mechanisms for bract development in Arabidopsis, desynchronising bract suppression from the floral transition, 08/2024. DOI: 10.1101/2024.08.12.607587

R. Kristianingsih, A. Calderwood, G. S. Sidhu, S. Woodhouse, H. Woolfenden, S. Kurup, R. Wells and R. J. Morris. Comparing gene expression dynamics over different developmental timescales with the R package *greatR*. In preparation, 2024

R. Kristianingsih. *greatR: Gene Registration from Expression and Time-Courses in R*. R package version 2.1.0. 2024

List of GitHub Repositories

All code and data associated with this thesis are available in the following repositories.

greatR package GitHub repository `https://github.com/ruthkr/greatR`

greatR package documentation website `https://ruthkr.github.io/greatR`

Thesis manuscript GitHub repository `https://github.com/ruthkr/greatR-manuscript`

Contents

List of Figures

List of Tables

List of Algorithms and Program Code

1 Introduction

This thesis describes the development of an approach for comparing time series data and explores various applications of this method, such as for analysing gene expression dynamics.

This chapter introduces general concepts in the biology of gene expression and outlines techniques for measuring gene expression. It also highlights the benefits of conducting time-series experiments to capture gene expression over time and reviews previous studies utilising this type of data. Subsequent chapters contain different analyses that can be performed with gene expression time-series data, including pair-wise comparisons and a discussion of currently available methods. Finally, an overview of this thesis will be presented.

1.1 Gene expression and transcriptomics technologies

1.1.1 Gene expression

Gene expression is the process by which genes are transcribed and/or translated into functional gene products, such as proteins or functional RNAs [4]. At a high level, gene expression consists of two main steps: transcription and translation (see fig. 1.1). During transcription, a gene's sequence in the genomic DNA is transcribed into a complementary RNA molecule. For protein-coding genes, this RNA is called messenger RNA (mRNA), which serves as a template for protein synthesis during the translation step. The transcriptome, the collection of all transcripts in a cell or a collection of cells, includes not only mRNAs but also non-coding RNAs (ncRNA). These ncRNAs encompass ribosomal RNAs, transfer RNAs (tRNA), microRNAs (miRNA), and various types of regulatory RNAs [5, 6].

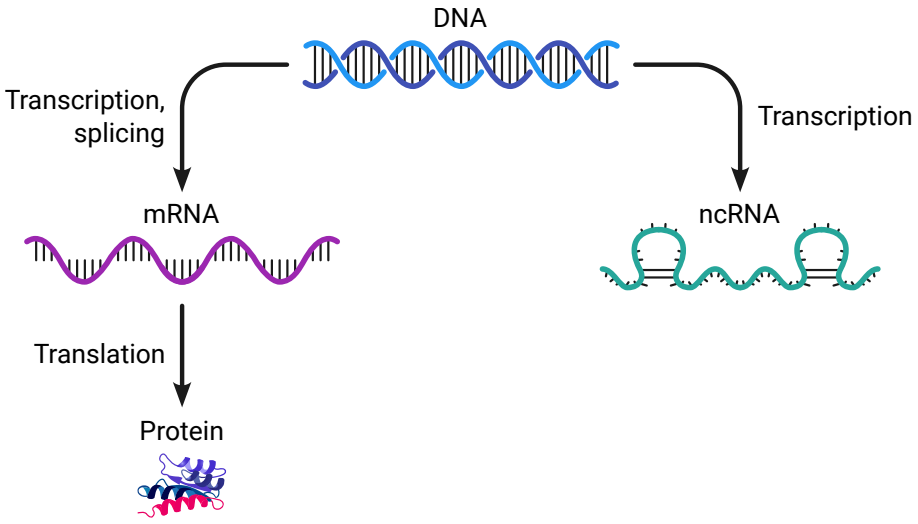


Figure 1.1: Overview of the processes involved in gene expression. Modified from [5].

The process of gene expression can be regulated and modulated at various levels, including transcription initiation, splicing, alternative splicing, mRNA stability, post-transcriptional regulation, and eventually translational and post-translational regulation mechanisms [7]. Gene expression is key to numerous biological processes, such as cell differentiation, morphogenesis, organ development, and disease progression. It also enables cells to adapt to different conditions. By controlling the timing, location, and expression levels, gene functions can be coordinated [7].

1.1.2 Measuring gene expression

Gene expression profiling, or transcriptomics, measures the expression level of mRNAs (transcripts) within a cell population at a specific time [8]. Measuring the expression of an organism's genes across different tissues, conditions, or time points provides valuable information on gene regulation and helps elucidate observed phenotypes [9]. King and Wilson [10] demonstrated that small changes in gene regulatory mechanisms associated with alterations in gene expression can explain significant phenotypic differences between organisms. These phenotypes, including an organism's appearance, behaviour, development, disease history, and temperament, represent the physical properties or characteristics of the organism [11]. In plants, phenotypes can range from visible traits, such as leaf shape, flower colour, and seed size, to less obvious traits like root architecture or drought resistance [12].

Gene expression measurement can also identify genes that have not been previously annotated through transcript assembly and potentially help to infer the functions of their isoforms by comparing their expression dynamics [13]. Transcriptomic analysis has facilitated the study of how gene expression varies within and between organisms, and has been instrumental in understanding biological processes such as disease and development [14].

The first attempts to study the whole transcriptome, which provides a snapshot of all the transcripts present in a cell at a given moment, began in the early 1990s [15]. Since then, technological advancements have made transcriptomics a widely practised discipline. The most commonly used modern techniques include microarrays, which measure a predetermined set of sequences, and RNA sequencing (RNA-Seq), which utilises high-throughput sequencing to capture all sequences [15].

The microarray technique relies on nucleic acid hybridisation to measure RNA concentrations. A labelled DNA or RNA sample, extracted from a biological source such as a specific tissue or cell culture, hybridises to a DNA probe fixed at a defined position (spot) on a solid surface, known as the microarray or slide [4]. The hybridisation process anchors the labelled sample to the probe. Typically, the label is a fluorescent dye, which allows for the estimation of the concentration of various DNA features in the sample. Fluorescence intensity is assumed to correlate with the concentration of the labelled sample, resulting in a relative, dimensionless measurement of gene expression [4]. This measurement usually requires quality control and normalisation to be interpretable. Microarrays can accommodate either one or two samples, using different

fluorescent dyes for each, enabling two hybridisations and thus two measurements on a single array.

RNA-Seq combines high-throughput sequencing techniques with computational methods to capture and quantify transcripts present in an RNA extract [15]. Generally, an RNA population is isolated from a cell or a population of cells. The isolated RNA population is broken down into fragments and reverse-transcribed into copy or complementary DNA (cDNA) fragments library with adaptors attached to both ends [16, 17]. The sequence library is fed to a next-generation sequence machine where the cDNA fragments are sampled uniformly at random. Short substrings, called reads, are read off the ends of these sampled fragments [16]. The reads are around 100 bp in length but can range from 30 bp to over 400 bp, depending on the platform used for sequencing [15, 17]. Following this step, the reads will be either aligned using computer software (e.g. HISAT [18], STAR [19], and Tophat [20]) to a reference genome or reference transcripts, or to each other (*de novo* assembly). A *de novo* assembly approach is used to produce a genome-scale transcription map that consists of the transcriptional structure and/or the level of expression for each gene [17]. This alignment method is also employed to discover transcripts that are missing or incomplete in the reference genome [21]. One of RNA-Seq's key advantages over microarray transcriptomes is its typical dynamic range of five orders of magnitude [17]. Additionally, RNA-Seq requires much lower input RNA amounts (nanograms quantity) compared to microarrays (micrograms quantity) [15, 17]. This allows finer examination of transcriptional activity, down to the single-cell level when combined with linear amplification of cDNA. Theoretically, RNA-Seq has no upper limit of quantification, and the error rate is low for 100 bp reads in nonrepetitive regions [15, 22]. RNA-Seq can be utilised to identify genes within a genome or to determine which genes are active at a specific time point, with read counts enabling accurate modelling of relative gene expression levels. Over time, RNA-Seq methodology has continuously improved, primarily due to advancements in DNA sequencing technologies that enhance throughput, accuracy, and read length [15].

Comparing the two technologies, RNA-Seq stands out by offering a broader dynamic range and easier detection of rare and low-abundance transcripts [17]. Additionally, RNA-Seq has the capability to detect novel transcripts [23]. Since its first description in 2006, RNA-Seq has been rapidly adopted and has overtaken microarrays as the primary transcriptomics technique. In this thesis, the transcriptomics data analysed are mainly RNA-Seq data.

1.2 The importance of time course experiments

Most biological processes are dynamic, and conducting time-series experiments is essential for us to comprehend and model these processes effectively [24]. One of the most abundant and commonly used time-series data is gene expression data, although several types of omics data can also be measured over time. Gene expression time series offers a wide range of insights into biological processes. These include characterising the functions of specific genes, understanding

the relationships among these genes, deciphering their regulation and coordination, and exploring the implications of differential dynamics [24, 25].

Monitoring changes in expression over time also provides a fundamentally different type of information. Instead of focusing only at two different outcomes or responses at specific time points (e.g. phenotype A versus phenotype B, wild type versus mutant, etc), we can observe how coordinated responses emerge from many interacting components over time [25, 26]. Observing and tracking how an organism responds to an external stimulus over time allows us to understand and reverse-engineer the mechanisms that regulate these responses. This understanding will form the rational foundation for generating testable hypotheses. The challenge, then, is how to effectively use time-series data to gain valuable insights to address a number of critical questions [25, 26]. These important questions, which could potentially be addressed, broadly fall under the following categories:

1. Biological systems analysis: understanding the driving dynamics of specific systems by monitoring them over time. One of the most studied systems is the cell cycle system [26]. This system has an important role in disease, development, and many other biological processes. It has been studied in yeast [27] and human [28]. Another example of biological systems which has been studied with time series experiments is the circadian clocks of human [29], mouse [30] and plant [31].

2. Response dynamics: controlled or uncontrolled perturbations (e.g. environmental stress, treatments, and drugs) are applied to systems, and the comprehensive gene expression response is monitored over time. Examples include a study on how *Rhododendron anthopogon* responds and adapts to the extreme conditions across the year in high-altitude alpine regions [32] and how two anticancer drugs affect the transcriptional response of human cancer cells [33].

3. Development: understanding the development of organisms which involves complex sequences of cell proliferation and differentiation. Numerous model organisms have been used over the years to study developmental processes. Examples include a study of seed development in *Brassica napus* [34] and human retina development [35].

4. Disease progression: identifying genes that act in response to a certain pathogen. This identification will be important knowledge in developing strategies to fight and control these diseases [26]. Examples include studies of gene expression time series in human cells that were infected by four different pathogens [36] and a susceptible cultivar of tree tomato *Solanum betaceum* during the infection of pathogen *Phytophthora betacei* [37].

Not surprisingly, generating time-series expression data has become a fundamental method for studying numerous biological processes across various fields (e.g. medicine, agriculture, evolutionary biology, and environmental science) [24]. The advancements in gene expression measurement techniques, such as high-throughput RNA-Seq, make time-series expression studies more

feasible and relevant [25]. Additionally, there has been an exponential increase in the number of time-series datasets deposited in major expression databases in recent years [24]. This makes time-series expression data attractive as a valuable resource for understanding dynamic systems [38].

1.3 Analysing gene expression data

The analysis of gene expression provides insights into cellular processes, as well as gene regulation and function [9]. This analysis involves multiple steps, with a variety of methods available for each. For example, to determine if gene expression levels between two samples at specific times significantly differ, either limma [39], DESeq2 [40], or *edgeR* [41] can be utilised. Unfortunately, there is no universally accepted solution or method suitable for all circumstances. The choice of methods depends on the technology used, the experimental design, and the specific biological questions being addressed. In this section, common methods, such as differentially expressed gene analysis and clustering, will be discussed.

1.3.1 Differentially expressed gene (DEG) analysis

A common biological analysis is to compare molecular expression between samples under different conditions. This type of analysis can, for example, explore the effects of pathogen infection by comparing infected versus uninfected samples, assess drug activity by comparing before and after treatments, or reveal specific molecular functions by comparing species with a knockout gene to the wild type. Typically, this comparison uses static expression data or a single snapshot of expression, measured under steady-state conditions. Several software packages have been developed for this task, including *edgeR* [41], DESeq2 [40], and baySeq [42], all of which use negative binomial models to assess differential gene expression from RNA-Seq data.

However, static expression data cannot determine whether changes are temporary or lasting. This limitation can be addressed by using expression time series data, which captures system dynamics missed by single time point measurements. Early analyses (e.g., as discussed in [24]) often applied differential expression methods originally designed for static experiments. While these methods can successfully identify genes that appear differentially expressed at individual time points, they tend to overlook the inherent temporal correlations in the data. Consequently, the significance calculations derived from such independent analyses may not accurately capture the underlying dynamics and potentially underestimating or overestimating the true statistical significance and missing important temporal patterns in gene regulation.

Some methods have been developed to specifically take the entire trajectory into account [24, 38, 43]. These methods often rely on analysing a more continuous version of the experimental results for each gene, utilising more time points to identify differentially expressed genes. Aryee *et al.* [44] developed BETR, a Bayesian-based method, to estimate the probability of differential

expression for each molecule based on the data. Two other popular methods implemented in R include limma which is based on linear modeling [39] and linear mixed models (LMM) which use coefficients to test for varying trends across different groups [45].

1.3.2 Clustering

Clustering is a widely used technique in time-series expression analysis, aimed at identifying groups of genes that exhibit similar expression patterns over time, referred to as clusters. Identifying these clusters helps structure the data into smaller, more manageable units, making handling and visualisation easier. Additionally, it is hypothesised that there may be biologically relevant relationships between co-expressed genes that are potentially co-regulated [46]. Most clustering methods measure the similarity of expression profiles using various distance measures, with the most common being Euclidean distance, Pearson correlation distance, and Manhattan metric [4]. These distance metrics were used in different methods of clustering.

For static data, popular clustering methods include hierarchical clustering and partitioning clustering [24]. The application of hierarchical clustering methods to microarray data was first popularised by Eisen *et al.* in 1998 [47]. They employed average linkage clustering on expression profiles from a study on *Saccharomyces cerevisiae*. Since then, various other linkage methods and techniques derived from phylogenetics, such as neighbour joining, have been used [4]. A heat map with the resulting clustering tree is normally used for visualisation. Partitioning clustering techniques enable the division of genes into a specified number of clusters, which often need to be pre-determined. An example of this clustering technique is the k-means method [48] which allows various distance measures to facilitate relatively rapid clustering into k groups. Another widely used technique is self-organising maps (SOM), which offer a more meaningful and structured topology of the cluster centres [49].

Although the clustering methods above are developed mainly for static gene expression data, they have been used widely for expression time series data [24, 38]. One of the limitations of these methods is the assumption that expression levels at consecutive time points are independent, which is not valid for transcriptomic time series data [50, 51]. Several methods have been specifically developed for time-series data to mitigate this limitation. Cluster Analysis of Gene Expression Dynamics (CAGED) [52] uses regression analysis to group genes based on their trajectories, Graphical Query Language (GQL) [53] employs hidden Markov models (HMMs)) to group genes based on their transcriptional trends, Short Time-series Expression Miner (STEM) [54] attempts to assign genes to one of several previously defined temporal trajectories, thus allowing users to determine significance levels for the different clusters. More recently, McDowell *et al.* [51] developed a clustering method based on a non-parametric model which models data clusters with a Dirichlet process and temporal dependencies with Gaussian processes. An improvement of this approach by taking into account the variability of experimental replicates was recently published [50].

1.3.3 Post-hoc analysis

With the results from the previous analysis, there are several common analyses that can be performed which include external information.

Gene Ontology analysis. Using clustering or DEG analysis, common genes of interest exhibiting a shared pattern can be identified and grouped together. Subsequently, one is likely to be interested in identifying the biological functions associated with these gene lists. These lists can be, for example, genes that are differentially expressed in *Brassica rapa* plants with different vernalisation periods [55] or a list of seed developmental genes that are upregulated between yellow- and black-seeded *B. napus* [56].

To make sense out of these genes, it requires biological knowledge of the involved genes and their functions. The Gene Ontology (GO) project (http://geneontology.org/ [57]) is a pioneering initiative that developed a database resource to make biological knowledge more accessible, especially given the massive increase in published data that makes it challenging to stay current with all the expanding information and computational methods. This project maintains a controlled hierarchical vocabulary of terms with logical definitions to describe molecular functions, biological processes, and cellular components. This controlled vocabulary is used by several model organism databases to capture experimental and computational findings on the roles specific genes play. This knowledge can be applied to a given list of genes or gene sets to explore the GO terms annotating the genes and to categorise them into functional groups through 'annotation' analysis [58]. Another common analysis is to focus only on terms significantly over-represented in a list of genes submitted, called enrichment analysis. This approach is a particular case of GSEA (gene set enrichment analysis) applied to Gene Ontology annotations [59]. Such analysis can be carried out from the GO project website [60]. This analysis can be also done using developed web applications (e.g. GOrilla [61], DAVID [62], and AmiGO [63]), and packages developed in different languages such as Python (e.g. GOATools [64] and orsum [65]) and R (e.g. *GOstats* [66] and *clusterProfiler* [67]).

Classification analysis. This analysis involves building a model for automatically distinguishing between two or more classes of samples [4]. It helps identify dynamic differences in gene expression profiles, providing insights into varying conditions or responses in different organisms [24]. Golub *et al.* [68] developed a classification model to distinguish between two forms of leukaemia. A variety of classification models are also employed in different microarray studies [4]. This analysis typically involves selecting features to identify genes that distinguish between classes, followed by training and evaluating classification models.

Gene regulatory networks. Gene regulatory networks (GRNs), which control gene expression, are crucial for understanding the system of dynamic biological processes. By studying GRNs over time, we can uncover how they work, how they respond to signals, and how different regulators interact with each other. Although time-series expression data provide important information about the activity of GRNs and many studies to construct GRNs are solely based on this data (such as *ppcor* which is based on the partial correlation [69], *GENIE3* [70] and

OutPredict [71] which are random forest-based methods), time-series expression only provides one possible viewpoint of the system. Specifically, it can be challenging to correctly infer the set of gene interactions when only using gene expression data, infer the specific activity of transcription factors (TFs) since many are post-transcriptionally regulated, infer how changes in epigenetics relate to dynamic responses, and more [72]. Therefore, the integration of several different types of multi-omics time-series, such as ChIP-Seq data which identify bound regions for TFs and time-series epigenetic data ATAC-Seq, are also required to construct accurate GRN models [72].

Pair-wise comparison. A comparison of gene expression dynamics is a widely applied routine to identifying variation in gene expression. This variation or changes in expression can be used to identify genes that might be associated with a specific environmental stimulus or developmental process and to construct or validate gene regulatory networks [24, 25]. The biological processes that give rise to the changes in expression may occur at different rates and timings in different species, strains, individuals, tissues, or conditions. For example, different individuals affected by a similar treatment may progress at different speeds [73], for instance during leaf development [74]. More details of this analysis, such as the importance of this analysis and the currently available methods and techniques, will be discussed in the next sections.

1.4 Comparing expression data and why it is important

The previous section covered several common methods for analysing gene expression data. In this section, a comparison of gene expression dynamics will be discussed in more detail. The common usage and the importance of this procedure will also be highlighted.

1.4.1 Intraspecies comparisons

Same species under different conditions

Within-species comparisons are commonly performed analyses to investigate gene expression variation among individuals of the same species between sample groups under different conditions [75]. These conditions can include the following:

- **Different treatments**
  Comparing expression time series can be used to understand the responses of specific genes under various treatments. Li *et al.* [76] examined the expression time-series of 19 *KNOX* genes in *Dendrobium huoshanense* under different stress treatments, including hormonal applications and drought conditions. *KNOX* genes play important roles in governing plant growth, development, and responses to different abiotic and biotic stresses [76]. In this study, they observed divergent expression patterns among the *KNOX* genes, highlighting their potential roles in stress adaptation. The same approach is also used in other plant

species to understand the response to stress such as in barley (*Hordeum vulgare L.*) under shock-dehydration and slow drought treatments [77], as well as in Arabidopsis gene expression in response to hypobaric stress [78] and eight different abiotic stresses (osmotic, salt, drought, genotoxic, wound, cold, heat, UV-b) [79]. Guo *et al.* [80] also utilised the same approach to investigate the response of *Arabidopsis thaliana* to two *Fusarium oxysporum* strains, an endophyte and a pathogen strain. In this study, they showed the *A. thaliana* and *F. oxysporum* interaction displays both transcriptome conservation and plasticity in the early stages of infection. They found that ∼80% of the Arabidopsis genomes have shared expression patterns in response to the two fungal infections [80]. They highlighted the distinct responding genes which indicate transcriptional plasticity, as the pathogenic interaction activates plant stress responses and suppresses functions related to plant growth and development. In contrast, the endophytic interaction attenuates host immunity but activates plant nitrogen assimilation. This study provides insights into how plants adjust their gene regulation to respond differently to fungal endophytes and pathogens.

- **Different environmental conditions**
  Observing the differential expression patterns of the same species under different conditions helps to understand the cellular response to environmental signals or external stimuli at the transcriptomic level [81]. Robinson *et al.* [82] compared the pair-wise gene expression throughout the developmental stages of grapevines (*Vitis vinifera L.*) grown in two different South Australian vineyards, Willunga and Clare. They discovered that the development rate of grape berries varied between the vineyards, likely due to differences in soil conditions, viticultural management, and climate [82]. Additionally, they identified a set of genes with consistent expression patterns across both sites, indicating that these genes may be developmentally regulated, while other genes showed variation, potentially reflecting vineyard-specific environmental responses.

Same species with phenotypic variation

In addition to the previously mentioned comparison cases, intraspecies comparisons are frequently conducted to study phenotypic variation within species. These comparative studies of expression patterns within species can help identify gene regulatory changes which contribute to complex phenotypic variations within the species [83]. Numerous studies employing this approach have helped to understand within-species variation in phenotypes. For example, Bailon-Zambrano *et al.* [84] discovered that variations in the expression of *mef2ca* paralogues, a gene associated with craniofacial development, influence the severity and variability of craniofacial phenotypes in zebrafish. Calderwood *et al.* [85] identified regulatory differences in the ageing pathway between *B. rapa* varieties *R-o-18* and *Sarisha-14* , which are linked to phenotypic differences in the timing of the floral transition.

1.4.2 Interspecies comparisons

Interspecies comparison is a powerful approach to understand the conservation and divergence of biological processes across species. This knowledge will further enable us to answer many important research questions, such as identifying correspondences between developmental stages [86], differences in the timing of key events during development [86], and transcriptional responses to biological, chemical or physical perturbations [81].

Before performing this comparison, corresponding pairs of genes for each organism have to be defined [87]. These gene pairs are needed to connect expression patterns between species. To define gene pairs, candidate orthologous gene pairs must be identified. Orthologous genes are those genes found in different species that originated from a single gene in their last common ancestor [87, 88]. There are numerous different methods developed for detecting orthologues gene pairs, either based on pairwise sequence comparison between genomes (e.g. InParanoid [89], OrthoMCL [90], and COG [91]) or phylogenetic analysis (e.g. TreeFam [92], HOGENOM [93], and PhylomeDB [94]).

As mentioned above, interspecies comparisons can be used for various biological purposes. These purposes can generally be categorised as follows:

Knowledge transfer from model organisms

Transferring knowledge from model species to less studied species is a keystone for many areas of biological research. This is due to the fact that most functional studies are carried out on model species, such as mice, yeasts, fruit flies, and thale cress (*A. thaliana*). In plants, this comparison is crucial for the translation of knowledge from model to economically important crops [95]. *Arabidopsis thaliana* was established as a universal model plant in the 1980s due to its small genome size of ∼114.5 Mb, short lifecycle, and high seed yield [96]. *Arabidopsis thaliana* is also easy to grow, as well as to be crossed and mutated [96].

Arabidopsis thaliana, as the first model plant, has been extensively used in numerous studies involving gene expression comparison, particularly in Brassica crops due to their close relatedness. Both Arabidopsis and the genus Brassica belong to the Brassicaceae family [97]. Considering this relationship, the orthologues of Arabidopsis genes are likely to have similar roles in Brassicas [85]. Comparisons of expression patterns between *A. thaliana* and Brassica for genes involved in various processes have been conducted previously. These studies include investigations into flowering time genes [55, 85] and circadian clock genes [98]. Dai *et al.* [55] investigated different vernalisation periods (i.e. the prolonged exposure to cold that promotes flowering in annual winter plants) in *B. rapa* in comparison to Arabidopsis. They found that the vernalisation process in *B. rapa* is associated with significant changes in gene expression, particularly in pathways related to plant hormone signal transduction, starch and sucrose metabolism, the photoperiod and circadian clock, and vernalisation, with distinct expression patterns of key genes such as

TPS, *UGP*, *CDF*, *VIN1*, and seven hormone pathway genes observed between two *B. rapa* accessions. Calderwood *et al.* [85] also used the model species Arabidopsis to study the timing difference in floral transition between two different cultivars of *B. rapa Sarisha-14* and *R-o-18*. They discovered that the difference in floral transition phenotypes between the two accessions is attributed to variations in the expression of key components of the ageing pathway. These components interact with *FT* (*FLOWERING LOCUS T*) signals from the leaf to regulate the floral transition [85]. Wang *et al.* [98] explored the circadian transcriptome profiles of Arabidopsis and soybean *Glycine max*. They identified differences in phase, period, and amplitude in the expression patterns of core circadian clock genes between Arabidopsis and soybean. Using this knowledge, they found that translation activities in Arabidopsis and photosynthesis activities in soybeans were more likely to be regulated by the circadian clock [98].

However, the knowledge from *A. thaliana* was more difficult to transfer to more distant species [99]. As a result, plant scientists adopted what they called second-generation plant models in the late 1990s. These include *Brachypodium distachyon* to represent grasses (monocots), *Physcomitrella patens* was chosen as a moss model, *Medicago truncatula* to study the legumes and *Populus trichocarpa* to cover trees [99]. More recently, a few third-generation plant models were also proposed to specifically study specific research areas [99]. Some of these plants are *Setaria viridis* as a model for C4 photosynthesis, *Marchantia polymorpha* to study land plant evolution, *Eutrema salsugineum* for salt tolerance, and many more [99]. The growing collection of model plants has made comparative transcriptomics a more popular method for applying knowledge from model organisms to other plants and crops.

In other organisms, such as humans, comparing gene expression is also a common method for understanding developmental processes [86], disease progression [100], and more. Various model organisms, which differ in complexity and use, are employed in these studies. For instance, small and simple organisms like yeast are often used to investigate gene mutations related to human cancers. The fruit fly (*Drosophila melanogaster*) and zebrafish (*Danio rerio*) are typically used to study genetics and disease development [101]. The nematode (*Caenorhabditis elegans*) is an ideal model organism for understanding the ageing process and the development of simple nervous systems [101]. Additionally, house mouse (*Mus musculus*) models are extensively utilised in biomedical research to investigate disease progression and develop new drugs [102].

Evolutionary studies

Gene regulatory variation causes significant phenotypic changes in the development of organisms [75]. The ability to identify specific sets of genes that are likely to harbour regulatory variations and cause large phenotypic effects relative to other genes is the ultimate goal in the evolutionary-developmental (evo-devo) biology field [75]. To achieve this, the comparison of gene expression patterns is one of the methods for regulatory variation analysis. These comparisons were made among model and non-model species or distantly related species [103]. In plants, studies of this were conducted in different settings, depending on the biological questions.

Leiboff and Hake [104] investigated transcriptional similarity between maize and its closely related species sorghum which have distinct morphology during inflorescence development. They identified that the expression shifts in the key regulators contributed to the morphological differences between the species [104]. Lemmon *et al.* [105] investigated the evolution of diverse floral branching systems, or inflorescences, in tomatoes and related nightshades. They chose five different ones with a maximum range of inflorescence architecture diversity. The selected species included: two with single-flowered inflorescences (*Capsicum annuum*, cultivated pepper; *Nicotiana benthamiana*, model tobacco), two with linear, multiflowered inflorescences (*Solanum lycopersicum*, cultivated tomato; *Solanum prinophyllum*, forest nightshade), and one wild tomato species (*Solanum peruvianum)* with branched inflorescences. This study highlighted that heterochronic shifts in key regulators during a critical transitional window of meristem maturation contribute to the evolutionary diversity of inflorescence complexity [105].

Several studies were attempted to investigate the conservation of gene expression dynamics in specific plant organs [106, 107]. A highly conserved expression pattern of key transcription factors in tip-growing cells between *Physcomitrella patens* and *A. thaliana* were identified by Ortiz-Ramírez *et al.* [106]. Additionally, through the pair-wise expression comparison, they also identified modifications in the expression dynamics of these genes that potentially account for developmental differences between *P. patens* tip-growing cells and *A. thaliana* pollen tubes and root hairs [106]. The conservation between Arabidopsis and maize during leaf development was also previously investigated by Vercruysse *et al.* [107]. They observed significant conservation in transcriptional regulation between the two species.

In other organisms, such as fungi, Guan *et al.* [108] compared the expression patterns of orthologues between *Saccharomyces bayanus* and *Saccharomyces cerevisiae* over different environmental perturbations. They found that most of the expression patterns are conserved. However, when analysing matched perturbations like diauxic shift (i.e. a change in metabolism in yeasts where glucose consumption initially fuels glycolytic fermentation, and then shifts to respiration using ethanol once glucose is depleted) and cell cycle synchrony, roughly 25% of orthologues had different expression patterns between the species. This indicates specific gene expression changes between the two species under certain conditions, suggesting regulatory differences or environmental adaptations.

Albert *et al.* [109] investigated whether similarities between different domestication events exist at the molecular level between domesticated and wild animals. In this study, they compared expression patterns in the brain frontal cortex in three pairs of domesticated and wild species (dogs and wolves, pigs and wild boars, and domesticated and wild rabbits). They also investigated expression differences between domesticated guinea pigs and a distant wild relative *Cavia aperea*, as well as between two lines of rats selected for tameness or aggression towards humans. Albert *et al.* [109] successfully identified expression differences that may correlate with behavioural differences in each domesticated species. These expression differences are unique to each species, suggesting that domestication has followed different genetic pathways in different species [109].

1.5 The challenge and current approaches in comparing expression time series

This section outlines the challenge in comparing expression time series. Subsequently, it describes common and available techniques for comparing pairs of gene expression profiles, along with their limitations.

1.5.1 The challenge in comparing expression time-series

In general, comparing two different time series (such as pairs of gene expression dynamics) requires a way of quantifying how similar they are. For two time-courses, (t_i, y_i) and (t_i, z_i), that consist of measurements, y_i and z_i, at the same time points, t_i, we can define the distance between two datasets at corresponding points, i, using the standard Euclidean distance,

$$d(y, z) = \left[\sum_{i=1}^{N}(y_i - z_i)^2\right]^{1/2}, \tag{1.1}$$

where N is the number of time points or observations, and $y = y_i$ and $z = z_i$ are the values of the two datasets at $i = 1, \ldots, N$. Alternative measures can also be used and these include the sum of absolute differences (Manhattan distance), mean-absolute differences, and others [110, 111, 112]. With a defined distance we can measure how far apart two datasets are and introduce a threshold value below which the datasets are considered to be the same. The threshold will typically depend on the number of data points and the expected errors so that we say the data are the same if they lie within the expected variation. If the expected errors (in the form of the estimated standard deviations $\sigma_{y,i}$ and $\sigma_{z,i}$) are known for each data point then this leads to an expected variance of $\sigma^2_{y,i} + \sigma^2_{z,i}$.

It is important to note that since Euclidean distance is sensitive to scale, in the instance of calculating gene expression distance, using raw gene expression values may cause larger values to dominate the distance calculation. To mitigate this, appropriate normalisation methods should be applied to adjust raw read counts for sequencing depth and compositional biases.

The metrics mentioned above rely on the datasets having corresponding time points to calculate the difference at equivalent time points. This can occur, for instance, when gene expression data is measured at the same time across two different conditions [113]. However, this is often not the case due to different experimental sampling times or variations in the rates and timings of biological processes between samples. Additionally, biological processes such as cell growth, metabolic rates, or disease progression may occur at different speeds in different samples or species. For example, Barry *et al.* [114] found that many genes in mouse epiblast stem (EpiS) cells exhibited faster dynamic changes during neural differentiation compared to human embryonic stem (ES) cells. Similarly, Calderwood *et al.* [85] observed that the developmental progression

of gene expression during the floral transition occurred at different speeds between *B. rapa* and its closely related ancestor, Arabidopsis.

Consider two time series (t_i, y_i) and (τ_j, z_j). These time series consist of measurements y_i at the time points t_i with $i \in [1, N]$ and measurements z_j at time points τ_j with $j \in [1, M]$, where N, M are the numbers of time points of timepoints or observations. If $N = M$ and $t_i = \tau_j$, this is the case when Equation 1.1 can be used because both time series have equivalent sampling time points (see Scenario 1 in Table 1.1). However, if $t_i \neq \tau_j$ for all i, j (see Scenario 2 in Table 1.1) or even if $t_i \neq \tau_j$ for some i, j (see Scenario 3 in Table 1.1), Equation 1.1 cannot be used, as not all time points have corresponding time points between the two time series. Aligning a pair of time series without equivalent time points in such cases is a complex task because direct comparisons cannot be performed. Consequently, the distance metric in Equation 1.1 is no longer applicable. This issue is a primary challenge in comparing expression time series.

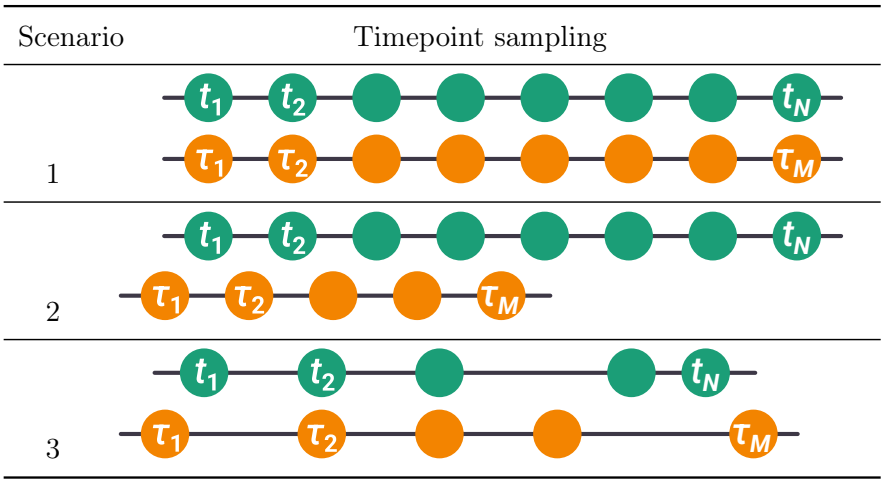


Table 1.1: Different sampling scenarios resulting in two time series, (t_i, y_i) represented by green circles and (τ_j, z_j) represented by orange circles. In Scenario 1, a direct comparison is possible as the time points align. However, in Scenarios 2 and 3, the time points between the datasets are not equivalent, making direct comparisons inapplicable.

1.5.2 Current approaches for comparing time series data and their limitations

Several methods have been developed to compare time series data. Commonly used approaches include dynamic time warping (DTW) [73, 24, 115]. Although DTW was first developed for speech recognition [116], it is also used in many different applications including bioinformatics, medicine, and engineering [116, 113]. This method aligns the time points, t_i and t_j, of two trajectories (t_i, y_i) and (t_j, z_j), that consist of measurements, y_i and z_j, by minimising the distance between them, e.g. using an absolute distance $d(y, z) = |y_i - z_j|$. The smaller the distance the more similar the dynamics. However, DTW does not answer the question of whether a pair of time series can be considered the same within experimental error, which is often of

interest when comparing the expression dynamics of two genes. DTW matches the start and the end indices of the first trajectory with the corresponding indices of the other trajectory, which can be problematic for pairs of gene expression profiles which often have dissimilar patterns and often exhibit differential progression over time which may arise from either time shifts or differences in speed. In addition to this, DTW will match every index from one trajectory to one or more indices from the other trajectory, and vice versa. This will make the optimal alignment results biologically challenging to interpret [82]. Furthermore, the optimal distances between pairs of time series given by DTW do not capture directly the pair's temporal pattern associations. These shortcomings contribute to DTW not being the most ideal method for comparing pairs of gene expression dynamics.

An alternative approach based on DTW, DTW4Omics [117], performs permutations on the order of time points for each pair of time series to estimate the alignment significance. However, this method does not solve DTW's challenges, such as result interpretability and the inability to statistically confirm similar temporal patterns. Additionally, DTW4Omics is computationally expensive [118].

TimeMeter [113] is an algorithm designed to identify similar temporal patterns using DTW results. It addresses one of DTW's limitations by providing measurements of differential progression between time series pairs. However, before these measurements can be evaluated, multiple metrics with predefined thresholds must be computed to assess pattern similarity. This reliance on predefined thresholds introduces bias into the similarity results, and the measurement scores remain difficult to interpret. Additionally, TimeMeter requires DTW results to be obtained separately before analysis, lacking an integrated workflow. This extra preprocessing step adds complexity and may limit its usability in automated pipelines. Furthermore, its dependence on DTW remains a drawback, as DTW struggles with uneven sampling and biological replicates.

Several alternative approaches based on hidden Markov models (HMM) [24, 119, 82] have been proposed. However, these HMM-based methods primarily address time shifts, which can be challenging when comparing pairs of time series with different timescales. Other methods specifically developed for comparing pairs of expression data are also available [120, 121, 122]. Like the HMM-based approaches, these methods focus on identifying time shifts between expression patterns. Additionally, HMM-based approaches are limited to detecting positively correlated dynamics and require substantial computational resources.

Another method, DynOmics [118], based on fast Fourier transform, can identify both time shifts and positive/negative correlations between profiles. However, its computational demands increase significantly when applied to large datasets, such as around 10,000 pairs of gene profiles. Moreover, the parameter results of this approach are difficult to interpret intuitively.

1.6 Project objectives and thesis structure

Following the previous section, we identified a need to develop an approach which is able to identify similarities between two datasets with potentially different rates of change and time ranges, while also providing biologically interpretable results. Additionally, this approach needs to be robust when handling thousands of pairs of gene expression time series and take into account biological replicates. Lastly, there is a need for a user-friendly tool that enables users to perform analyses easily, while limiting preparatory tasks and which provides ways of summarising the results. This thesis describes the development and application of such a method. The approaches of the project can be divided into (1) the development of the method, (2) testing the method on the controlled data, and (3) applying the method to different biological datasets.

Chapter 2 describes the development of the approach which employs curve registration and the Bayesian Information Criterion. This chapter will begin by formulating the problems of comparing two time series over similar ranges for which two possible hypotheses were developed. This will be followed by the discussion of statistical methods to estimate the evidence of these possible hypotheses which later on will be used to determine the similarity between two curves. The approach when the time ranges are not equivalent will also be discussed.

Chapter 3 will focus on the development of the R package *greatR* (Gene Registration from Expression and Time-courses in R) as a wrapper of the method explained in Chapter 2. This will be followed by the method evaluation using simulated datasets to ensure our method can accurately align and perform as intended. This chapter begins with the process of generating the simulated data, followed by a discussion of the results of the approach. Potential limitations of *greatR* will also be discussed.

Chapter 4 will demonstrate another example of possible applications of the approach using different datasets. Rather than comparing expression data from different species, this chapter will focus on comparing the Arabidopsis genotypes *Tsu-0* and *Col-0*. By analysing RNA-Seq time series data across developmental stages, we identified widespread transcriptional desynchronisation during the floral transition between the two genotypes, with key molecular changes occurring during the late vegetative-to-transition stage in the bract-forming accession. While key floral transition genes showed conserved expression, differences in gene timing rather than gene presence suggest that bract development may be driven by temporal shifts in gene expression across accessions.

Chapter 5 applies the method to compare Brassicas and Arabidopsis. This analysis begins by comparing *B. oleracea* with *A. thaliana* to investigate the similarity between the two species during floral transitions. This analysis is then broadened to include *B. rapa*, allowing for comparisons between all three species. Our findings reveal that differences in gene expression timing, rather than expression profiles, largely explain the variation in flowering time gene dynamics across these species. While key floral transition genes exhibited conserved expression, some

paralogues showed species-specific divergence, suggesting potential functional differences in gene regulation. These findings highlight the role of copy number variation and gene regulation in the evolution and adaptation of flowering time in Brassica species.

Chapter 6 is the discussion chapter of the thesis. It will summarise the findings and conclusions of each chapter. Additionally, it will provide an overall conclusion and highlight potential improvements for future research.

2 Formulation of a statistical approach for comparing gene expression patterns

This chapter presents the formulation of our approach for comparing gene expression time series across different conditions or species. The central aim is to accurately identify temporal shifts and rate changes in expression patterns, which requires modelling each time series flexibly and aligning them in a meaningful way. To achieve this, we begin by outlining the hypotheses underlying our comparison framework and proceed to formulate a log-likelihood function that quantifies the alignment quality. We then introduce B-splines to model the expression curves, offering the flexibility needed to capture complex dynamics. Model selection is guided by the Bayesian Information Criterion (BIC), which helps prevent overfitting while balancing fit and complexity. Finally, we describe the optimisation techniques used to estimate the key parameters governing the alignment, specifically stretch and shift, using methods tailored for both small- and large-scale applications. These components, though distinct in method, collectively support a unified goal: to perform robust, scalable curve registration for comparative transcriptomic analysis. This chapter lays the methodological foundation for the applications and evaluations presented in Chapter 3.

2.1 Theoretical foundations

2.1.1 Bayes's theorem for parameter estimation

Bayes' theorem, named after Thomas Bayes, provides a mathematical framework for reversing conditional probabilities, enabling us to determine the probability of a cause given its effect. A key application of this theorem is in Bayesian inference, a statistical approach that uses Bayes' theorem to reverse the probability of observing data given a particular model (the likelihood) to compute the probability of the model itself given the observed data (the posterior probability).

While Bayes' theorem provides a probabilistic approach to parameter estimation by updating prior beliefs, an alternative is Least Squares Estimation (LSE), which estimates parameters by minimising the sum of squared differences between observed and predicted values. Unlike LSE, which provides a single best-fit estimate, Bayesian inference produces a probability distribution over possible parameter values, allowing for a more detailed quantification of uncertainty.

Given a model M with parameters θ, parameter estimation addresses the question of which values of θ are good estimates, given some data $\mathbf{x}$. To calculate the probability of each hypothesis given

the data $\mathbf{x}$, we can use Bayes' theorem. This is expressed as

$$P(\theta|\mathbf{x}) = \frac{P(\mathbf{x}|\theta) \cdot P(\theta)}{P(\mathbf{x})}, \tag{2.1}$$

where:

- $P(\theta)$ is the prior probability about how likely each value of θ might be,
- $P(\mathbf{x}|\theta)$ is the likelihood or evidence, which tells us how well the θ explains the data,
- $P(\mathbf{x})$ is the probability of the observed data, acting as a normalising factor.

To compute the evidence $P(\mathbf{x}|\theta)$, we compute the likelihood function $\Lambda(\theta)$. This function quantifies how likely it is to observe the data $\mathbf{x}$ given a particular set of parameters θ. We will discuss this in more detail in the next subsection.

Likelihood

In statistical analysis, likelihood plays a critical role in estimating the parameters of a model based on observed data. A parameter is a number that defines the characteristics of a distribution. For example, in a Bernoulli distribution, the parameter p represents the probability of success, while in a Uniform distribution, the parameters a and b define the min and the max value [123]. In many real-world situations, we often do not have this information readily available. Instead, we have data generated from an unknown distribution and need to estimate the parameters underlying that distribution. This is where likelihood and maximum likelihood estimation (MLE) come into play. Likelihood assesses how well a set of parameters θ explains the observed data, while MLE helps us find the parameter values that maximise the likelihood function. In practice, likelihood is critical for parameter estimation and model comparison, though it requires careful attention to assumptions such as independence of observations and underlying distributions.

Suppose the data that we are going to use to estimate the parameters have n independent and identically distributed samples. This implies that they share either the same probability mass function (in the case of discrete data) or the same probability density function (for continuous data) [123]. Given set of data $\mathbf{x} = x_1, x_2, \ldots, x_n$ and set of parameters θ, we can define the likelihood as

$$\Lambda(\theta) = \prod_{i=1}^{n} f(x_i|\theta), \tag{2.2}$$

where $f(x_i|\theta)$ is a probability density function. Since we assumed that each data point is independent, then the likelihood of our data is the product of the likelihood for each data point.

Maximum likelihood estimation

The goal of maximum likelihood estimation (MLE) is to find the parameter values (θ) that maximise the likelihood function, hence we need to find:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\ \Lambda(\theta), \tag{2.3}$$

where $\hat{\theta}$ represents the best choice of values for the parameters, and the term argmax stands for "Arguments of the Maxima" and refers to the value(s) in the domain of a function at which the function reaches its maximum.

An important property of the argmax is that, since the logarithmic function is monotonic, the argmax of a function is equivalent to the argmax of the logarithm of that function. This is particularly useful in simplifying mathematical calculations, as logarithms can make complex expressions easier to handle. In the case of likelihood functions, which often involve exponentials, taking the logarithm makes differentiation and optimisation more straightforward. Therefore, when applying Maximum Likelihood Estimation (MLE), finding the argmax of the log-likelihood function yields the same result as finding the argmax of the likelihood function itself [123]. Consequently, in MLE, we often begin by expressing the log-likelihood function to facilitate parameter estimation as follows

$$\log \Lambda(\theta) = \log \prod_{i=1}^{n} f(x_i|\theta) = \sum_{i=1}^{n} \log f(x_i|\theta). \tag{2.4}$$

MLE of normal distributed data

Suppose we have n samples data from normal distribution and for all i, $x_i \sim N(\mu = \theta_0, \sigma^2 = \theta_1)$. Notice here, we have two parameters μ and σ to estimate. Therefore, we can define the likelihood as:

$$\Lambda(\theta) = \prod_{i=1}^{n} f(x_i|\theta) \tag{2.5}$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\theta_1}} e^{-\frac{(x_i-\theta_0)^2}{2\theta_1}} \tag{2.6}$$

and the corresponding log-likelihood is

$$\log \Lambda(\theta) = \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi\theta_1}} e^{-\frac{(x_i-\theta_0)^2}{2\theta_1}} \tag{2.7}$$

$$= \sum_{i=1}^{n} \left[-\log\left(\sqrt{2\pi\theta_1}\right) - \frac{1}{2\theta_1}(x_i - \theta_0)^2 \right]. \tag{2.8}$$

To find the best values θ which maximise the log-likelihood function, we can solve the partial derivative of $\log \Lambda(\theta)$ with respect to both θ_0 and θ_1 equal to zero:

$$\frac{\partial \log \Lambda(\theta)}{\partial \theta_0} = 0, \quad \frac{\partial \log \Lambda(\theta)}{\partial \theta_1} = 0. \tag{2.9}$$

In the context of model selection, once the parameter estimates that maximise the log-likelihood have been determined, the corresponding log-likelihood value can be computed. This value provides a measure of how well the model fits the observed data, with a higher log-likelihood indicating a better fit [123]. However, relying solely on the log-likelihood may not be sufficient when comparing models with varying complexities, as it does not account for the number of parameters or potential overfitting.

2.1.2 BIC

Suppose we have a task to determine which model best fits a given set of data from a range of possible models. Model selection is crucial in statistical analysis because choosing the wrong model can lead to incorrect inferences and predictions. To address this challenge, Schwarz [124] proposed the Bayesian Information Criterion (BIC), a widely used criterion for model selection that balances model fit with model complexity.

The BIC is based on the principle of likelihood, which measures how well a model explains the observed data [125]. However, unlike methods that solely maximise the likelihood function, BIC incorporates a penalty term for the number of parameters in the model. This penalty discourages overfitting, a situation where a model becomes too complex and captures not just the underlying signal in the data but also the noise [125]. By penalising the number of parameters, BIC favours models that provide a good fit with fewer, more essential parameters, leading to simpler and more interpretable models.

Mathematically, the BIC for a given model is calculated as

$$\text{BIC} = -2\mathcal{L} + k \log n, \tag{2.10}$$

where $\mathcal{L}$ is the maximum log-likelihood of the model (representing how well the model fits the data), k is the number of parameters in the model, and n is the number of observations in the data. The first term, $-2\mathcal{L}$, assesses the goodness of fit by looking at the likelihood of the model given the data–the larger the likelihood, the better the fit. The second term, $k \log n$, is the penalty for model complexity. As the number of parameters k increases, the penalty grows, which prevents the selection of unnecessarily complex models that could overfit the data.

In practice, the BIC provides a straightforward and interpretable criterion for model selection: among a set of candidate models, the model with the lowest BIC is preferred. This approach is particularly useful in contexts where multiple models might fit the data reasonably well, but we seek to identify the model that achieves the best trade-off between goodness of fit and simplicity.

One of the strengths of the BIC is that it is derived from a Bayesian framework, which integrates prior information with the likelihood of the observed data. This integration allows BIC to take into account the uncertainty in model parameters while favouring models that are more likely to generalise to new data. However, BIC also has limitations. It works best when the number of data points (n) is large. If the number of parameters (k) is high relative to the number of observations (n), the penalty for complexity can be too strong. This may lead to underfitting, especially in cases with short time series or sparse data. In such scenarios, alternative criteria like Akaike Information Criterion (AIC) or cross-validation may work better.

2.1.3 Curve registration

Functions can exhibit variability in both phase and amplitude [126], as illustrated schematically in Figure 2.1. Phase variation refers to differences in the timing or horizontal alignment of features along the curve, as depicted in Figure 2.1 (a). In contrast, amplitude variation, shown in Figure 2.1 (b), describes changes in the vertical scale or magnitude of the curves. The mean curve in the top panel, represented by the dashed line, does not resemble any individual curve; while it shows reduced amplitude variation, its horizontal extent is larger than that of any single curve. This indicates that the mean has effectively "borrowed" from amplitude to account for phase variation [126].

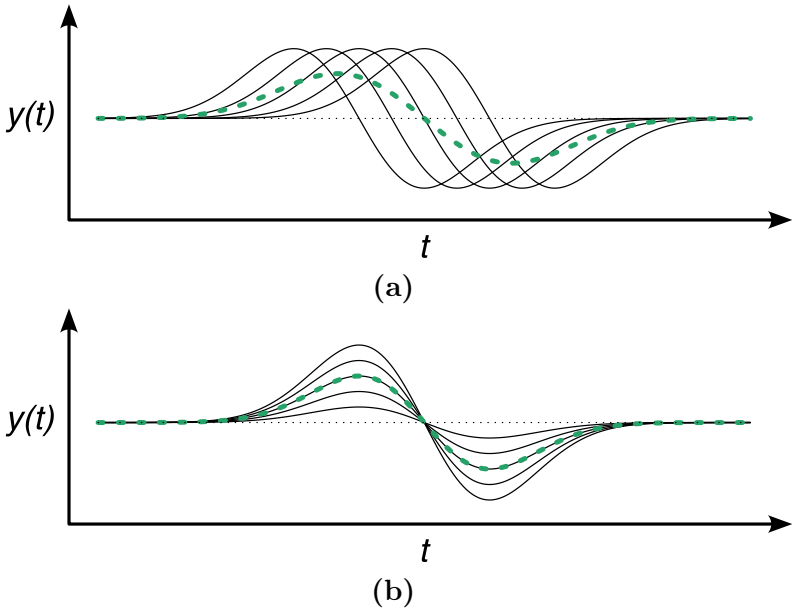


Figure 2.1: Plots show five curves varying (a) in phase and (b) in amplitude. The dashed line in each panel indicates the mean of the five curves. The curves in (b) are superimposed exactly on the central curve. Modified from [126].

In order to address or align curves with phase variation (see Figure 2.1 (a)), we can use a technique known as curve registration [126]. This method systematically adjusts the horizontal

alignment of curves by applying time-warping transformations, allowing for more accurate comparisons between functions by reducing phase variability while preserving amplitude differences. By aligning the phases of the curves, curve registration ensures that the remaining variation reflects amplitude differences (see Figure 2.1 (b)) [126].

Registration in general

In general, a registration method can be defined as a process of aligning features of multiple curves by monotone transformations of their domain [127]. Consider a set of observed functions $y_1(t), y_2(t), \ldots, y_n(t)$, where $y_i(t)$ represents the observed curve for the i-th sample. These functions exhibit both phase and amplitude variability. To achieve this, we introduce a **time-warping function** $h_i(t)$ for each curve [126, 128]. This function adjusts the time axis of the observed curve $y_i(t)$, aligning the key features with those of other curves. The relationship between the observed curve and the aligned curve is given by

$$x_i(t) = y_i(h_i(t)), \tag{2.11}$$

where $x_i(t)$ represents the underlying function after removing the phase variability. The challenge in curve registration often lies in identifying which variations are due to phase and which are due to amplitude. For example, a peak may appear earlier in one curve than another, but distinguishing whether this is a timing (phase) difference or a difference in the magnitude (amplitude) of the peak requires care.

The alignment process involves finding a set of warping functions $h_1(t), h_2(t), \ldots, h_n(t)$ that minimise the phase variability, typically by minimising the difference between the aligned curves. After registration, the remaining differences between the curves should ideally reflect only amplitude variability. Thus, curve registration allows for meaningful comparisons between curves by aligning their key features in time, while leaving amplitude differences intact.

Pairwise registration

In pairwise curve registration, the task is to align two curves: a **reference** or a template curve and a **query** curve, where the reference curve remains fixed, and the query curve is warped to align with the reference. This process ensures that important features, such as peaks or valleys, occur at the same time points on both curves. Let $y_{\text{ref}}(t)$ represent the reference curve and $y_{\text{query}}(t)$ represent the query curve. The goal of the registration process is to find a **time-warping function** $h(t)$, which adjusts the time axis of the query curve, mapping its time points to those of the reference curve. After applying this warping function, the newly aligned version of the query curve is denoted as

$$y_{\text{aligned}}(t) = y_{\text{query}}(h(t)). \tag{2.12}$$

The aim of curve registration is to minimise the difference between the reference curve and the warped query curve. In other words, the objective is to find the warping function $h(t)$ that minimises this difference. This then can be formulated as an optimisation problem where the distance or difference between the two curves across all time points is minimised as follows

$$\min D\left(y_{\text{ref}}(t), y_{\text{query}}(h(t))\right). \tag{2.13}$$

Here, D can represent any suitable distance metric, such as the L_1-norm (absolute differences), L_2-norm (squared differences), or other distance metrics [128].

For the warping function $h(t)$ to be meaningful, it must satisfy certain constraints. First, $h(t)$ must be **monotonic**, meaning it preserves the order of time points (i.e., $t_1 < t_2$ implies $h(t_1) < h(t_2)$) [126, 128]. Additionally, $h(t)$ should be smooth to avoid abrupt changes in the alignment, which could introduce unnatural distortions.

2.1.4 Types of different warping functions to explain phase variations

Depending on the application, Marron *et al.* [128] (see Figure 2.2) classified different possible warping functions to specify phase variations:

1. Uniform scaling: the warping of the time domain simply rescales it by a positive constant $a \in \mathbb{R}_+$ such that $h(t) = at$ for all $t \in \mathbb{R}_+$.
2. Uniform shift: the time axis gets shifted by a constant $c \in \mathbb{R}$, such that $h(t) = c + t$.
3. Linear or affine transform: a combination of the uniform scaling and uniform shift, such that $h(t) = c + at$, $a \in \mathbb{R}_+$ and $c \in \mathbb{R}$.
4. Diffeomorphisms: a more flexible option that can warp the time axis in complex ways, beyond simple scaling and shifting. While diffeomorphisms can be defined on the whole real line, in practice, they are often restricted to a specific interval.
5. Weakly increasing functions: these allow for time-warping functions $h(t)$ that are non-decreasing but not necessarily strictly increasing or smooth. Unlike diffeomorphisms, which are continuously differentiable and invertible, weakly increasing functions may contain flat functions (where $h'(t) = 0$) or discontinuities in the derivative. This flexibility can be useful in applications where parts of the signal are constant or where certain events occur simultaneously across time.

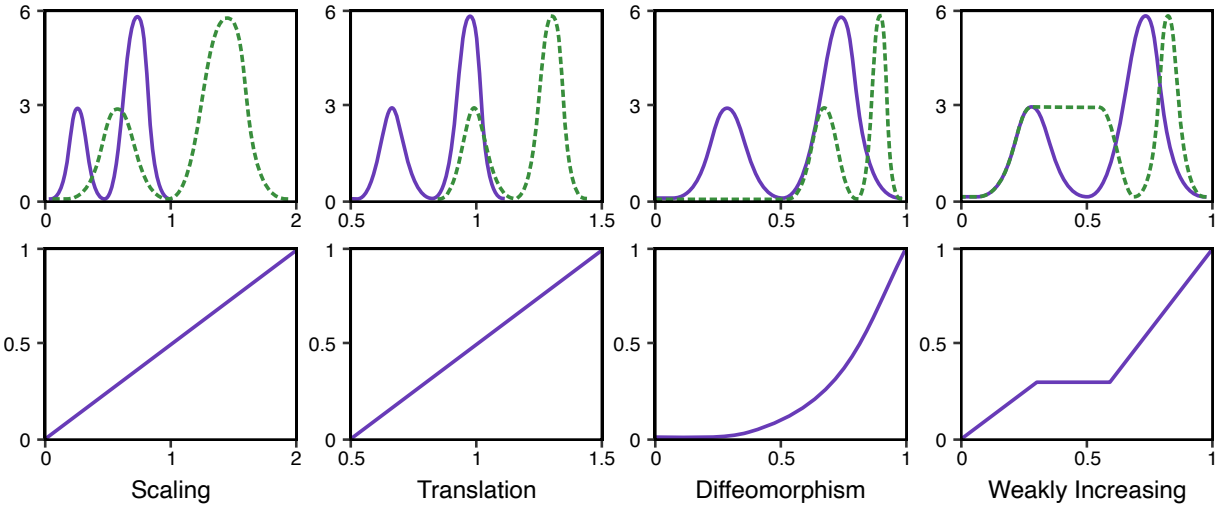


Figure 2.2: Illustrations of various types of warping functions applied to the same function $y(t)$. In the top row, the solid line represents $y(t)$, while the dashed line represents the wrapped function $y(h(t))$. The bottom row shows the corresponding warping functions $h(t)$. Adapted from [128].

2.1.5 Polynomial regression and splines

In this section, we discuss two commonly used techniques for regression: polynomial regression and splines. Polynomial regression provides a foundation for modelling relationships between variables, allowing linear (a polynomial of degree one) and higher-degree polynomials to capture more complex trends. As the degree of the polynomial increases, the model becomes more flexible but can also become prone to overfitting. To address this, splines extend polynomial regression by dividing the data into segments and fitting piecewise polynomials, allowing for more flexibility while maintaining stability. Together, these methods provide a spectrum of approaches for modelling various types of data patterns.

Polynomial regression

Polynomial regression is a technique used to model the relationship between a dependent variable y and an independent variable x as a polynomial of degree n. The general form of a polynomial regression model is given by:

$$y = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n. \tag{2.14}$$

In this formula, $a_0, a_1, \ldots, a_n$ are the coefficients of the polynomial, and n is the degree of the polynomial. The degree of the polynomial determines the flexibility of the model: a higher degree polynomial can capture more complex relationships between x and y. However, increasing the degree also introduces the risk of overfitting, where the model becomes too specific to the dataset and may not generalise well to new data.

To illustrate polynomial regression, let's consider two specific cases: linear and quadratic regression.

In **linear regression** (a polynomial of degree 1), the relationship between x and y is modeled as a straight line. The formula for a linear model is

$$y = a_0 + a_1 x. \tag{2.15}$$

Here, a_0 is the intercept, representing the value of y when $x = 0$, and a_1 is the slope, representing how much y changes with a unit increase in x. Linear regression is useful when the relationship between the variables is proportional and constant across all values of x.

In **quadratic regression** (a polynomial of degree 2), the relationship between x and y includes a squared term, allowing the model to capture curvature. The formula for a quadratic model is

$$y = a_0 + a_1 x + a_2 x^2. \tag{2.16}$$

In this case, a_2 is the coefficient of the quadratic term x^2, which introduces curvature into the model. This type of regression is commonly used when the data shows a turning point or curvature, such as when growth accelerates and then decelerates.

Splines

In this section, we outline the approach used for modelling the data, with an emphasis on spline-based techniques. Splines are widely recognised for their valuable properties across various fields, including image processing, statistical modelling, and the construction of explanatory models in clinical research [129, 130]. They are particularly effective for generating smooth, flexible curves that can adapt to complex data patterns. By dividing the domain into smaller intervals and fitting piecewise polynomials, splines provide local control of the curve, ensuring smooth transitions between segments [129]. This flexibility is especially useful in avoiding the pitfalls of high-degree polynomial fitting, such as overfitting, numerical instability, and oscillations [129].

One key advantage of splines is their ability to partition the dataset into multiple ranges and fit each range with a separate model. The points where these divisions occur are called knots. A spline function $f(t)$ can be written as a linear combination of basis functions:

$$f(t) = \sum_{i=0}^{d+K} a_i b_i(t), \tag{2.17}$$

where

- $b_i(t)$ are the basis functions,

- a_i are the associated spline coefficients,
- d is the fixed degree of the chosen polynomial, and
- K is the number of knots.

There are several ways to represent cubic splines using different choices of basis function b_i [131, 130]. For example, a degree-d spline with knots at τ_k for $k = 1, \ldots, K$ can be represented by truncated power basis functions as follows

$$f(t) = a_0 + a_1\, b_1(t) + \ldots + a_{K+d}\, b_{K+d}(t), \tag{2.18}$$

where

$$\begin{aligned} b_1(t) &= t^1 \\ &\vdots \\ b_d(t) &= t^d \\ b_{(k+d)}(t) &= (t - \tau_k)_+^d,\ k = 1, \ldots, K \end{aligned} \tag{2.19}$$

and

$$(t - \tau_k)_+^d = \begin{cases} (t - \tau_k)^d, & \text{if } t > \tau_k \\ 0, & \text{otherwise.} \end{cases} \tag{2.20}$$

For example, a linear spline with one knot (τ_1) can be represented as follows (see Figure 2.3)

$$y = \begin{cases} a_0 + a_1 t, & \text{if } t < \tau_1 \\ a_0 + a_1 t + a_2(t - \tau_1), & \text{if } t \geq \tau_1, \end{cases} \tag{2.21}$$

where t is time, τ_1 are the knot, a_0 is the intercept, and a_1, a_2 are the associated spline coefficients (parameters).

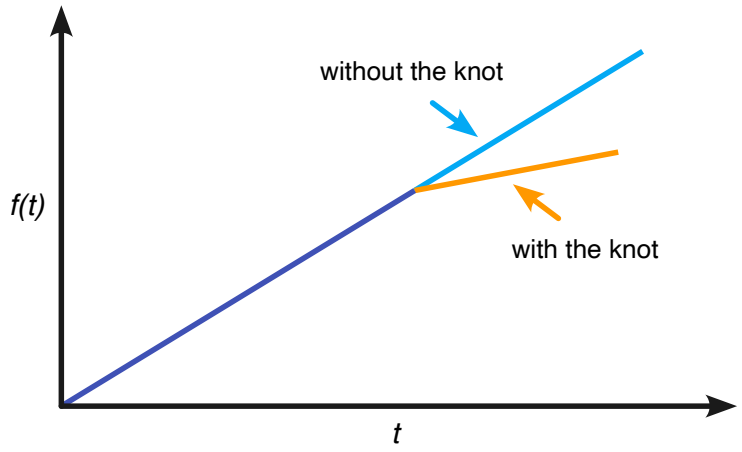


Figure 2.3: An illustration of a linear spline with a single knot. Modified from [132].

A cubic spline with one knot (τ_1) will have five degrees of freedom. By using the representation given in Equation 2.18, the function can be written as:

$$f(t) = a_0 + a_1\, t + a_2\, t^2 + a_3\, t^3 + a_4\, (t - \tau_1)^3, \tag{2.22}$$

where t is time, τ_1 are the knot, a_0 is the intercept, and a_1, a_2, a_3, a_4 are the associated spline coefficients (parameters).

B-splines

While the method described earlier for defining splines is straightforward, it may not be the most flexible or convenient for many applications [129]. An alternative approach is to express a cubic polynomial using a set of normalised basis functions, with the B-spline basis being a popular choice due to its advantageous properties. B-splines are particularly useful for fitting curves to time-series data, such as gene expression profiles, where it is often more convenient to use approximating or smoothing splines rather than interpolating splines [129].

For a partition of a knot sequence (i.e., a non-decreasing sequence $\xi := (\xi_k)$), de Boor [133] defines the B-splines of order one for this knot sequence as the characteristic functions of this partition, i.e., the functions

$$b_k^1(t) = \begin{cases} 1, & \xi_k \leq t < \xi_{k+1} \\ 0, & \text{else.} \end{cases} \tag{2.23}$$

From this first order of B-splines, higher order of B-splines can be obtained by recurrence:

$$\begin{aligned} b_k^n(t) &= \frac{t - \xi_k}{\xi_{k+n-1} - \xi_k} b_k^{n-1}(t) - \frac{\xi_{k+n} - t}{\xi_{k+n} - \xi_{k+1}} b_{k+1}^{n-1}(t), \\ k &= 1, \ldots, K + n, \end{aligned} \tag{2.24}$$

where

- n is the order of the basis polynomials (i.e. for cubic polynomials $n = 4$) and
- ξ_k are the knots, where $k = 1, \ldots, n + K$.

Each spline curve $f(t)$ is constructed based on B-spline basis functions b_k^n and corresponding coefficients $a_k \in \mathbb{R}$, and is defined as follows

$$f(t) = \sum_{i=0}^{K} a_i b_k^n(t). \tag{2.25}$$

B-spline basis functions $b_i(t)$ are defined using a recursive formula [129], and are a commonly used spline basis based on a special parameterisation of a cubic spline [130].

2.1.6 Optimisation methods

L-BFGS-B

Brief description

The L-BFGS-B algorithm is an extension of the L-BFGS algorithm to handle problems with boundary constraints [134, 135]. It is an optimisation algorithm in the family of quasi-Newton methods that approximates the Broyden–Fletcher–Goldfarb–Shanno algorithm (BFGS) using a limited amount of computer memory [136]. Quasi-Newton methods are used to find the minimum (or maximum) of a function by approximating the function's Hessian matrix (which contains information about the curvature of the function). Instead of storing the entire Hessian matrix, L-BFGS only stores a few vectors that summarise the most important parts of the Hessian. This makes it very memory-efficient, which is important for large-scale problems where you have many variables. The "B" in L-BFGS-B stands for box constraints, which simply means that each variable in the problem can be limited to a specific range. These constraints ensure that the algorithm doesn't search for solutions outside the allowable range, which makes it more applicable to real-world problems.

The L-BFGS-B algorithm

L-BFGS-B starts with an initial guess and computes the gradient to determine the direction in which the objective function decreases the most. Instead of using the full Hessian matrix (which describes the curvature of the function), it approximates the Hessian with a limited memory approach to reduce computational cost. The algorithm then finds a search direction, ensuring that any movement stays within the defined bounds for each variable. A line search is performed along this direction to find the point where the function value is minimised, and the parameters are updated. These steps are repeated until one of the stopping criteria is met, such as reaching the maximum number of iterations or a small gradient, indicating that further improvement is unlikely. More details about how L-BFGS-B works are shown in Algorithm 2.1.

Algorithm 2.1 L-BFGS-B algorithm.

Input: Initial guess x_0, objective function $f(x)$, bounds for variables, maximum iterations.
Output: Optimised solution x^*.

Initialise $x \leftarrow x_0$, iteration $k \leftarrow 0$
Set $x^* \leftarrow x$, $f^* \leftarrow f(x)$ ▷ Track the best solution
while stopping criteria not met **do**
    Compute gradient $\nabla f(x_k)$
    Approximate Hessian with limited memory
    Find search direction d_k by solving a quadratic problem
    Perform line search to find step size α_k
    Update $x_{k+1} \leftarrow x_k + \alpha_k d_k$
    Update memory with new gradient and step
    if $f(x) < f^*$ **then**
        Update $x^* \leftarrow x_{k+1}$, $f^* \leftarrow f(x_{k+1})$ ▷ Update the best solution
    end if
    Increment iteration count $k \leftarrow k + 1$
end while
return x^*

Nelder–Mead

The Nelder–Mead (NM) method, also known as the simplex search algorithm, was first introduced by John Nelder and Roger Mead in 1965 [137]. It is one of the best-known algorithms for multidimensional unconstrained optimisation without derivatives. Since it does not require any derivative information, the NM method is particularly suitable for problems involving non-smooth functions.

Brief description

The NM method is designed to solve the classical unconstrained optimisation problems, specifically for minimising a nonlinear function $f : \mathbb{R}^n \to \mathbb{R}$. It belongs to the general class of direct search methods as it relies only on function evaluations at certain points in $\mathbb{R}^n$ without requiring any gradient information [138]. This feature makes NM particularly effective for problems where derivatives are unavailable or difficult to compute. This method is a simplex-based. In this context, a simplex S in $\mathbb{R}^n$ is the convex hull of $n + 1$ vertices $x_0, \ldots, x_n \in \mathbb{R}^n$. For example, a triangle and a tetrahedron are simplex in $\mathbb{R}^2$ and $\mathbb{R}^3$, respectively.

The algorithm begins with an initial simplex formed by $n + 1$ points $x_0, \ldots, x_n \in \mathbb{R}^n$ which are considered as the vertices. The corresponding function values at these points are denoted $f_j := f(x_j)$, for $j := 0, \ldots, n$. This initial simplex S has to be nondegenerate which means that the points $x_0, \ldots, x_n$ must not lie in the same hyperplane. The method then performs a sequence of transformations of the working simplex S, with the goal of decreasing the function values at its vertices. At each iteration, the algorithm tests new points based on various operations, such as reflection, expansion, contraction, or shrinkage of the simplex. The function values at these new

points are computed and compared to the current values at the simplex's vertices. Depending on the results, the simplex is modified to move toward regions of the function where lower values are expected. This process continues until a termination condition is met, such as when the simplex becomes sufficiently small or when the function values at the vertices are sufficiently close if assuming f is continuous.

The NM algorithm

Although the method is relatively simple, its implementations vary based on how the initial simplex is constructed and the criteria used for convergence or termination. The general algorithm can be summarised as follows:

1. Construct the initial simplex S.
2. Repeat the following steps until a termination criterion is met
   - evaluate the termination condition,
   - if not satisfied, transform the working simplex.
3. Return the best vertex of the current simplex and the corresponding function value.

Algorithms 2.2 and 2.3 summarise the initialisation of the simplex and the various transformations applied to the working simplex in more detail. A practical implementation of the Nelder–Mead method requires a robust termination test to ensure that the algorithm halts within a finite amount of time. This test typically comprises three components: (1) Domain convergence, which verifies if the simplex vertices are close enough to indicate convergence in the parameter space; (2) Function-value convergence, which checks whether the function values at the vertices are sufficiently close, signalling that the function is converging; and (3) No-convergence (fail), which triggers when the number of iterations or function evaluations exceeds a predetermined maximum. The algorithm terminates when any of these conditions are met. Different implementations may prioritise one or more of these criteria, with domain convergence being particularly crucial for discontinuous functions to ensure the algorithm identifies a sufficiently accurate solution point.

Algorithm 2.2 Nelder–Mead algorithm: construct initial simplex.

Input: Starting point $x_0 \in \mathbb{R}^n$, step sizes $h_1, \ldots, h_n$.
Output: Simplex S.

for $j \leftarrow 1$ to n **do**
    $x_j \leftarrow x_0 + h_j e_j$ ▷ Generate vertices using step sizes h_j along coordinate axes
end for
return $S \leftarrow \{x_0, x_1, \ldots, x_n\}$

Algorithm 2.3 Nelder–Mead algorithm: simplex transformation.

Input: Current simplex S, function values f_j, parameters α, β, γ, δ.
Output: Updated simplex S.

Identify the worst, second worst, and best vertices: x_h, x_s, x_l
$c \leftarrow \sum_{j \neq h} (x_j/n)$ ▷ Compute centroid c of the simplex opposite x_h

$x_r \leftarrow c + \alpha(c - x_h)$ ▷ Attempt reflection x_r
if $f(x_r) < f(x_h)$ **then**
    $x_h \leftarrow x_r$ and set x_h as the new vertex of the simplex
else
    $x_e \leftarrow c + \gamma(x_r - c)$ ▷ Attempt expansion x_e
    if $f(x_e) < f(x_r)$ **then**
        $x_h \leftarrow x_e$ and set x_h as the new vertex of the simplex
    else
        if $f(x_r) < f(x_s)$ **then**
            $x_c \leftarrow c + \beta(x_r - c)$ ▷ Attempt outer contraction x_c
        else
            $x_c \leftarrow c + \beta(x_h - c)$ ▷ Attempt inner contraction x_c
        end if
        if $f(x_c) < f(x_h)$ **then**
            $x_h \leftarrow x_c$ and set x_h as the new vertex of the simplex.
        else
            $x_j \leftarrow x_l + \delta(x_j - x_l)$ for all vertices x_j in S ▷ Shrink simplex towards x_l
        end if
    end if
end if
return Updated simplex S

Simulated Annealing

We also employ Simulated Annealing (SA), a global optimisation method and one of the most popular metaheuristic techniques for complex, non-linear objective functions [139, 140]. Simulated annealing is inspired by the process of metal annealing, where a material is heated and then slowly cooled to remove defects and reach a stable structure [139]. Unlike gradient-based and deterministic search methods, the key advantage of simulated annealing is its ability to reduce the risk of getting trapped in local optima, which are suboptimal solutions that can stop some algorithms from finding the global optimum.

To explain this, imagine dropping bouncing balls over a landscape. As the balls lose energy and bounce less, they settle in valleys or local minima. If the balls lose energy slowly enough, some will eventually fall into the deepest valleys, representing the global minimum. Simulated annealing mimics this process, moving through potential solutions while gradually reducing the "temperature" (or randomness) of the search. Over time, the system becomes more selective, focusing on areas that offer better solutions while still allowing some exploration of worse ones to escape local minima.

At each step of the search, the algorithm evaluates whether to accept a new solution based on an acceptance probability, even if the new solution is worse than the current one [141]. This probability is calculated as

$$p = \exp\Big[-\frac{\Delta E}{k_B T}\Big], \tag{2.26}$$

where k_B is Boltzmann's constant, T is the temperature for controlling the annealing process and ΔE is the change in energy (or the difference between the current and new objective function values) [141]. As the temperature decreases (i.e., as the algorithm progresses), the system starts behaving more like a hill-climbing method, where only improvements are accepted. Thus, controlling the temperature schedule is key to the efficiency of the algorithm.

There are various ways to manage the cooling, or temperature decrease. Two common cooling schedules are:

1. **Linear cooling schedule.** The temperature decreases linearly over time:

$$T = T_0 - \beta t, \tag{2.27}$$

   where T_0 is the initial temperature, t is the iteration count, and β is the cooling rate.

2. **Geometric cooling schedule.** The temperature decreases by a factor α at each step:

$$T(t) = T_0 \alpha^t, \tag{2.28}$$

   where α is the cooling factor, typically between 0.7 and 0.99, and t is the iteration count. This method is more commonly used because it naturally brings the temperature closer to zero as iterations increase.

For each temperature, the algorithm evaluates the objective function multiple times. If there are too few evaluations, the system might not stabilise, leading to premature convergence. Too many evaluations, on the other hand, can slow down the process. Finding the right balance is crucial for ensuring the system converges to the global optimum in a reasonable amount of time [141]. There are two ways to control the number of iterations:

- fixed iterations with a set number of evaluations are performed at each temperature, or
- variable iterations with more evaluations are performed as the temperature decreases to better explore the local minima.

More details about how Simulated Annealing works can be seen in the Algorithm 2.4.

Algorithm 2.4 Simulated Annealing algorithm.

Input: Initial solution x_0, objective function $f(x)$, initial temperature T_0, cooling schedule parameters, maximum iterations.
Output: Optimised solution x^*.

Initialise $x \leftarrow x_0$, $T \leftarrow T_0$, iteration $t \leftarrow 0$
Set $x^* \leftarrow x$, $f^* \leftarrow f(x)$ ▷ Track the best solution
while stopping criteria not met **do**
    Generate a new candidate solution x' by perturbing x
    Compute the change in objective function $\Delta E \leftarrow f(x) - f(x')$
    if $\Delta E < 0$ **then**
        Accept $x \leftarrow x'$ ▷ Always accept improvements
    else
        Calculate acceptance probability $p \leftarrow \exp\left(-\frac{\Delta E}{k_B T}\right)$
        Accept $x \leftarrow x'$ with probability p
    end if
    if $f(x) < f^*$ **then**
        Update $x^* \leftarrow x$, $f^* \leftarrow f(x)$ ▷ Update the best solution
    end if
    Increment iteration count $t \leftarrow t + 1$
    $T \leftarrow$ UPDATE_TEMPERATURE(T_0, t) ▷ Update temperature
end while
return x^*

function UPDATE_TEMPERATURE(T_0, t)
    Choose a cooling schedule:
    if linear cooling **then**
        $T(t) \leftarrow T_0 - \beta t$
    else if geometric cooling **then**
        $T(t) \leftarrow T_0 \alpha^t$
    end if
    return $T(t)$
end function

2.1.7 Data normalisation and scaling

Data normalisation or scaling is a fundamental technique in data processing, used to adjust the range of independent variables or features to a common scale without distorting differences in the ranges of values. This process is an essential part of data preprocessing, particularly when preparing data for analysis or machine learning tasks. Normalisation ensures that no single feature dominates the model due to its scale, which can be particularly important when algorithms rely on distance measures, such as in clustering or principal component analysis (PCA).

In the context of time series data, normalisation becomes especially critical when comparing datasets that have been recorded under different conditions or units. For example, in gene expression studies, different genes may be expressed at significantly different levels, depending on various

biological factors or experimental conditions. These differences in expression levels can obscure the underlying patterns or dynamics of the data if left unnormalised. If a researcher's primary interest is in comparing the dynamics or trends of the gene expressions rather than their absolute levels, normalisation is necessary to bring the data onto a comparable scale.

By scaling the data before analysis, it is possible to focus on the relative changes and trends over time, facilitating more meaningful comparisons between different time series. This step helps to eliminate biases that may arise from differences in magnitude, ensuring that the analysis reflects the true relationships and patterns within the data. Thus, normalisation is not just a technical step, but a crucial preparatory measure for the subsequent process during registration.

Linear scaling

This scaling method, commonly referred to as min-max scaling or normalisation [142, 143], adjusts the values of each attribute so they fall within a range of 0 to 1. This is achieved by subtracting the minimum value of the attribute and then dividing by the range, which is the difference between the maximum and minimum values, as shown in the following formula

$$\mathbf{x}' = \frac{\mathbf{x} - \mathbf{x}_{\min}}{\mathbf{x}_{\max} - \mathbf{x}_{\min}}. \tag{2.29}$$

Z-score

A Z-score indicates how many standard deviations a particular value deviates from the mean [143]. To perform Z-score scaling, the process begins by subtracting the mean from each value, which centres the data around a mean of zero. Mathematically, it can be formulated as follows

$$\mathbf{x}' = \frac{\mathbf{x} - \mu}{\sigma}, \tag{2.30}$$

where μ is the mean and σ is the standard deviation of the data.

The result is divided by the standard deviation, ensuring that the scaled data has a standard deviation of one. Unlike min-max scaling, Z-score scaling does not confine values to a specific range. However, it has the advantage of being less sensitive to outliers. For instance, consider a scenario where a district's max income is mistakenly recorded as 100 instead of within the typical range of 0 to 15. If min-max scaling were applied to map values to a 0–1 range, this outlier would be scaled to 1, compressing all other values into the narrow range of 0 to 0.15. In contrast, Z-score scaling would remain relatively unaffected by this extreme value, preserving the overall structure of the data [142].

2.2 Results

2.2.1 Two time series over similar ranges can be compared by evaluating how well they can be explained by a joint model

Let's consider a situation where we have two sets of time series data that we want to compare, each collected at specific, but potentially different, time points (see Figure 2.4). We will refer to one time series as the query (q) and the other as the reference (r).

For the reference dataset r, the time points at which the data were collected are labelled as $t_{r,i}$ and the corresponding expression values as $y_{r,i}$. Here, i ranges from 1 to N_r, where N_r represents the total number of time points in the reference dataset. Similarly, for the query dataset q, the time points are labelled as $t_{q,j}$ and the corresponding expression values as $y_{q,j}$. In this case, j ranges from 1 to N_q, with N_q being the total number of time points in the query dataset. The reference and query datasets are denoted by $r = (t_{r,i}, y_{r,i})$ and $q = (t_{q,j}, y_{q,j})$, respectively.

The datasets, denoted as r and q, can represent various types of biological data. For instance, r and q might correspond to measurements or observations related to the same gene but obtained under distinct environmental conditions. This could involve comparing gene expression levels in response to factors such as temperature changes, nutrient availability, or the presence of specific stressors, providing insights into how the gene's behaviour varies in different conditions. Alternatively, r and q could represent data on homologous genes across different species.

If the time points on the x-axis between two datasets do not correspond, a direct comparison of their expression levels (y-axis) is not possible. In such scenarios, it becomes necessary to introduce an underlying model that can account for the differences in time points and provide a framework for interpolation or extrapolation. This underlying model serves as a mathematical function that describes the relationship between the variables over time, allowing for the estimation of new points in one dataset based on the available data in the other. One of the simplest approaches to achieve this is using a piecewise linear curve, which connects adjacent data points with straight lines. This method enables a straightforward interpolation scheme where intermediate values at unmeasured time points in the query dataset can be estimated, thereby facilitating comparison with the reference dataset.

However, the choice of model is not limited to piecewise linear interpolation. More complex models can be employed depending on the complexity of the data and the desired level of accuracy. For example, polynomial models of varying degrees can provide smoother curves that capture the trends in the data more accurately, though they may also introduce the risk of overfitting if the polynomial order is too high. Spline models, which use piecewise polynomials, offer greater flexibility by ensuring smooth transitions between segments and are particularly useful when dealing with non-linear trends.

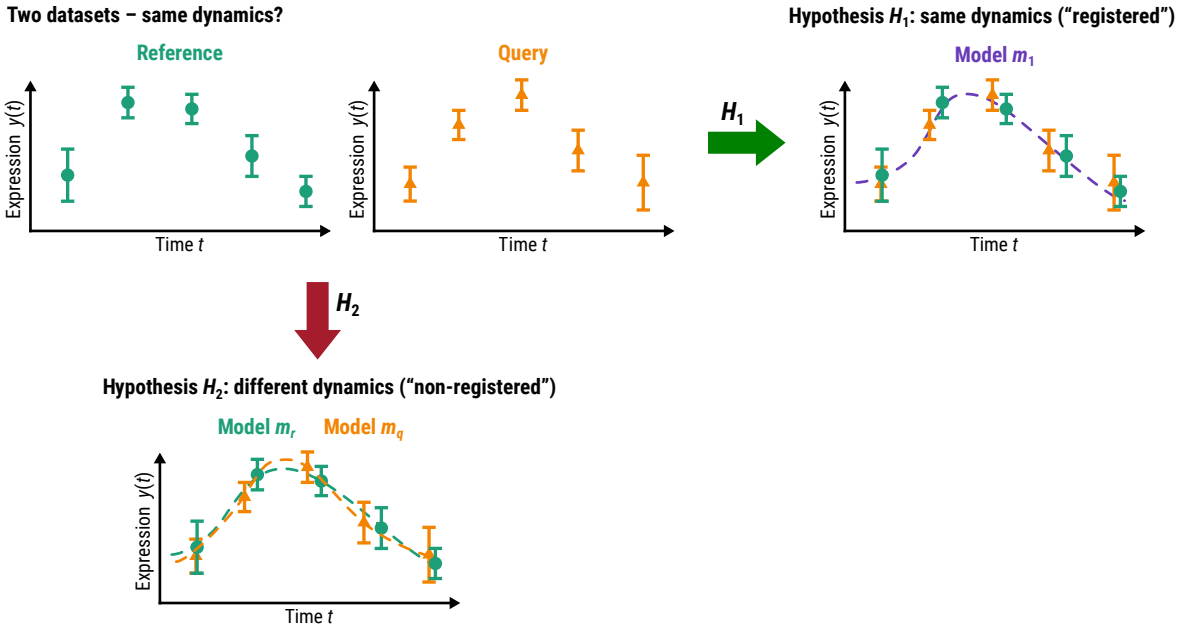


Figure 2.4: Evaluating whether two datasets with similar time points are the same (within experimental error) depends on the choice of model. We depict two datasets (the reference against which we wish to compare the query dataset). We can then ask whether the datasets are best explained by a single model (a function $m_1(\theta_1, t)$ with parameters θ_1), or whether two models, (functions $m_r(\theta_r, t)$ and $m_q(\theta_q, t)$ with parameters θ_r and θ_q that may be different between the functions), provide a statistically better explanation. Hypothesis H_1 = one model (same dynamics of the data), hypothesis H_2 = two models (different dynamics of the data).

The choice of model is important, as it will influence all subsequent inferences drawn from the data. Different models may lead to different interpretations. Therefore, careful consideration must be given to the model selection process. This will be discussed more in Section 2.2.2.

Figure 2.4 provides a conceptual framework for comparing the two datasets. We define the model for the reference dataset by the function $m_r(\theta_r, t)$ and the model for the query dataset by the function $m_q(\theta_q, t)$. Here, θ_r and θ_q denote the parameters associated with the reference and query models, respectively. It is possible that a single model, denoted by $m_1(\theta_1, t)$, with its own set of parameters θ_1, can explain both datasets simultaneously.

To assess whether a single model can sufficiently explain both datasets within experimental error, or whether separate models are required, we can frame the problem within a probabilistic context. This approach allows us to quantify the evidence supporting two hypotheses:

- Hypothesis H_1: the datasets are best explained by *one* common model $m_1(\theta_1, t)$.
- Hypothesis H_2: the datasets are best explained by *two* different models $m_r(\theta_r, t)$ and $m_q(\theta_q, t)$.

Under hypothesis H_1 the assumption is that both datasets can be captured by a shared model, implying that the observed data reflect the same underlying process or mechanism. This could be the case, for instance, when comparing datasets obtained under similar conditions or from related biological processes. In contrast, hypothesis H_2 suggests that the datasets are governed by different dynamics, necessitating separate models to accurately describe each dataset.

In general, the models $m_r(\theta_r, t)$ and $m_q(\theta_q, t)$ do not need to share the same functional form, allowing for a flexible fit to the data. However, for the sake of simplicity in the following analysis, we will focus on cases where the models share the same functional form but may differ in their parameter values. By carefully choosing and comparing these models, we can make informed decisions whether the datasets are likely governed by the same underlying processes or if distinct dynamics are at play. This helps us decide how to analyse and interpret the data going forward.

To calculate the probability of each hypothesis given the data $D = (r, q)$, we can use Bayes' theorem. For hypothesis H_1, this is expressed as

$$P(H_1|D) = P(m_1|r, q) = \frac{P(r, q|m_1) \cdot P(m_1)}{P(r, q)}, \tag{2.31}$$

where:

- m_1 is the joint model that describes both datasets r and q,
- $P(m_1)$ is the prior probability of Hypothesis H_1,
- $P(r, q|m_1)$ is the marginal likelihood or evidence, which tells us how well the model explains the data,
- $P(r, q)$ is the probability of the observed data, acting as a normalising factor.

The model m_1 might have a number, N_{m_1} of parameters denoted by θ. To compute the evidence $P(r, q|m_1)$, we integrate the likelihood function $\Lambda(\theta) = P(\theta|m_1, r, q)$ over the prior distribution of the parameters $P(\theta|m_1)$.

For the likelihood function, we assume that each data point (both from the reference dataset r and the query dataset q) follows a Gaussian distribution. This can be written as

$$\Lambda_{H_1}(\theta_1) \propto \prod_{i=1}^{N_r} \exp\{-[m_1(\theta_1, t_{r,i}) - y_{r,i}]^2/2\sigma_{r,i}^2\} \prod_{j=1}^{N_q} \exp\{-[m_1(\theta_1, t_{q,j}) - y_{q,j}]^2/2\sigma_{q,j}^2\}, \tag{2.32}$$

where $\sigma_{r,i}$ and $\sigma_{q,j}$ are the estimated standard deviations for the data points of the datasets r and q.

In this equation, $\Lambda_{H_1}(\theta_1)$ is the likelihood of hypothesis H_1, and it depends on the parameters θ_1 of the model $m_1(\theta_1, t)$, which aims to describe both the reference and query datasets simultaneously. Essentially, this likelihood tells us how probable the observed data is given a specific set of parameters θ_1 within the joint model m_1.

For Hypothesis H_2, the posterior probability can be expressed as

$$P(H_2|D) = P(m_r, m_q|r, q) = \frac{P(r, q|m_r, m_q) \cdot P(m_r, m_q)}{P(r, q)}, \tag{2.33}$$

where m_r is the model of the reference dataset r and m_q the model of the query dataset q. The likelihood of this hypothesis depends on the parameters θ_r and θ_q of the models m_r and m_q for the respective datasets. This likelihood can be written as

$$\Lambda_{H_2}(\theta_r, \theta_q) \propto \prod_{i=1}^{N_r} \exp\{-[m_r(\theta_r, t_{r,i}) - y_{r,i}]^2/2\sigma_{r,i}^2\} \prod_{j=1}^{N_q} \exp\{-[m_q(\theta_q, t_{q,j}) - y_{q,j}]^2/2\sigma_{q,j}^2\}, \tag{2.34}$$

where $\sigma_{r,i}$ and $\sigma_{q,j}$ are the standard deviations associated with the data points in the reference dataset r and the query dataset q, respectively.

With these likelihood functions, we can compute the marginal likelihoods by integrating over the parameters, θ_1 for H_1 and θ_r, θ_q for H_2,

$$P(r, q|m_1) = \int \Lambda_{H_1}(\theta_1) P(\theta_1|m_1) d\theta_1 \tag{2.35}$$

and

$$P(r, q|m_r, m_q) = \iint \Lambda_{H_2}(\theta_r, \theta_q) P(\theta_r, \theta_q|m_r, m_q) d\theta_r d\theta_q, \tag{2.36}$$

with which the posterior ratio and Bayes factor can be computed,

$$BF_{12} = P(r, q|m_1)/P(r, q|m_r, m_q). \tag{2.37}$$

The advantage of using either the posterior ratios or alternatively Bayes factors is that we have a measure for the confidence in each hypothesis based on established scales. However, a disadvantage of using Bayes factors is that the required integration over the likelihood function is typically computationally intensive. This process often involves techniques like Markov Chain Monte Carlo (MCMC) methods [144], or related approaches like nested sampling [145, 146]. The computational burden is especially pronounced when the analysis needs to be performed across large datasets, such as in whole-transcriptome comparisons where tens of thousands of time series must be analysed.

Given these challenges, we employ the Bayesian Information Criterion (BIC) heuristic [124, 147] as an approximation to Bayes factors. The BIC, also known as the Schwarz Criterion, provides a

statistical measure for evaluating models with different numbers of parameters [124]. It offers a practical and computationally efficient alternative to Bayes factors, especially when dealing with large-scale data.

The BIC is closely related to the Akaike Information Criterion (AIC), another popular model selection criterion. The key difference between them lies in how they penalise the complexity of the model: while both criteria penalise the inclusion of additional parameters to prevent overfitting, BIC applies a stronger penalty as the number of parameters increases. This makes BIC particularly useful when comparing models with varying levels of complexity.

BIC is defined as

$$\mathrm{BIC} = -2\mathcal{L} + k \log N_D, \tag{2.38}$$

where $\mathcal{L} = \max_\theta[\log \Lambda(\theta)]$ is the maximised log-likelihood for the model (m_1 or m_r and m_q with parameters θ or θ_r and θ_q, respectively), k is the total number of parameters (here either k_{m_1} for H_1 or $k_{m_r} + k_{m_q}$ for H_2), and N_D is the sample size ($N_D = N_r + N_q$ unless some of the points overlap or a subset is selected).

The BIC provides a criterion for the model comparison, where lower values indicate a better model. Specifically, if

$$\mathrm{BIC}(H_1) < \mathrm{BIC}(H_2), \tag{2.39}$$

then hypothesis H_1 is considered to provide a better explanation of the data than hypothesis H_2. In this context, it means that the two time series are likely best explained by a single model, implying that the patterns observed in both datasets are similar or consistent.

The difference between BIC values not only tells us which model is better but also provides a measure of how much better one model is compared to another. Specifically, according to Raftery [147]:

- A difference of 0 to 2 between BIC values is regarded as **weak evidence** favouring the model with the lower BIC.
- A difference of 2 to 6 is considered **positive evidence** supporting the lower BIC model.
- A difference greater than 6 is viewed as **strong evidence** that the model with the lower BIC is superior.

One of the significant advantages of using BIC is its computational efficiency. The calculation of BIC requires optimisation, which is typically much faster than the integration required in the full Bayesian framework. Therefore, BIC serves as a practical and computationally efficient alternative to the full Bayesian approach, offering a reasonable approximation for model comparison while being faster to compute. This makes BIC especially useful in large-scale analyses where speed and computational resources are important.

2.2.2 Impact on model selection on the interpretation and inference of data

As mentioned in the previous section, the selection of an appropriate model is an important step in data analysis, as it significantly influences the inferences that can be drawn from the data. Different models may lead to varying interpretations of the same dataset, potentially altering the conclusions of the analysis. Therefore, it is essential to approach the model selection process with careful consideration, balancing the complexity of the model against the need for accurate and reliable inferences. In this section, we will demonstrate the impact of different model choices on the fit to the data and underscore the importance of selecting an appropriate model that balances complexity with the need for accurate, reliable inferences.

Figure 2.5 (a) illustrates two time series generated from the same underlying model. The reference time series (green) and the query time series (orange) were independently sampled from this model, represented by the purple curve, which is defined as follows

$$y(t) = A\sin(\omega t - \phi) + C + \mathcal{N}(\mu, \sigma). \tag{2.40}$$

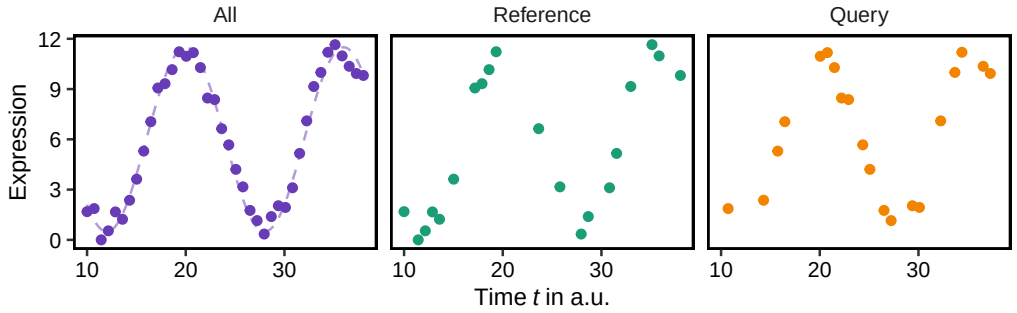
In this equation, the parameters are set as $A = 5.5$, $\omega = \pi/8$, $\phi = 0$, $\mu = 0$, and $\sigma = 0.5$. The term $\mathcal{N}(\mu, \sigma)$ represents a normal distribution with mean μ and standard deviation σ, which describes the noise in the model. The constant C is chosen to ensure that the expression remains non-negative across all time points, specifically $C = \min(|y|)$. Both the reference and query datasets share the same initial and final time points to maintain similar ranges, thereby allowing for direct comparison.

Both the reference and query time series illustrated in Figure 2.5 (a) were subsequently fitted to three different models: linear, cubic B-splines, and sinusoidal (the original model from which the time series were sampled). These fittings were performed both separately and together, following the definitions of hypotheses H_1 and H_2 described in the previous section. The goal was to evaluate whether the two time series are better explained by a single model or by separate models.

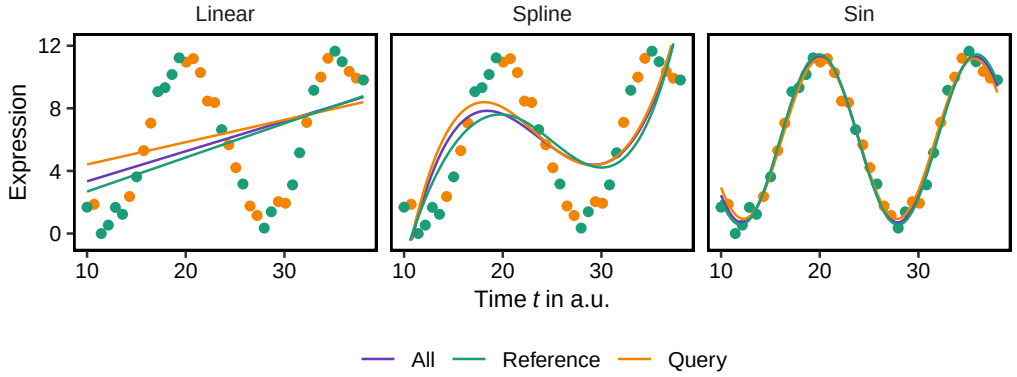
For each model and for both hypotheses H_1 and H_2, the Bayesian Information Criterion (BIC) was calculated to assess the goodness of fit. By comparing the BIC values across the three different models, we can determine not only whether a single model can describe both time series but also which model is the best for these two datasets.

Figure 2.5 (b) illustrates how each model fits the data and Table 2.5 (c) presents the corresponding BIC scores each H_1 and H_2. For the linear model, H_2 is favoured over H_1, indicated by the Bayesian Information Criterion (BIC) values: $\mathrm{BIC}(H_2)_{\mathrm{linear}} < \mathrm{BIC}(H_1)_{\mathrm{linear}}$. This suggests that the query and reference data sets are better explained by two different models in the linear case. In contrast, for both sinusoidal and spline models, H_1 is favoured over H_2, as evidenced by the BIC values: $\mathrm{BIC}(H_1)_{\mathrm{sin}} < \mathrm{BIC}(H_2)_{\mathrm{sin}}$, and $\mathrm{BIC}(H_1)_{\mathrm{spline}} < \mathrm{BIC}(H_2)_{\mathrm{spline}}$. This indicates that a single model best explains both reference and query data sets when sinusoidal and spline models are used.

The different conclusions drawn from these three models underscore the significant impact that the choice of model has on the inferences made from time series data. Furthermore, the BIC(H_2) score for the sinusoidal model is the lowest among all models tested, indicating that this model provides the best fit for the data which is an expected result, given that it is the original model from which the two datasets were sampled.



(a) Reference (green) and query (orange) time series, both sampled from the same underlying model (purple) as defined in Equation 2.40.



(b) Reference, query, and both combined data sets were fitted to three models: linear, cubic B-splines with one knot, sinusoidal following the definitions of H_1 and H_2 as described in the Methods section. Green and orange points indicate reference and query data, respectively. The green, orange, and purple solid lines indicate the models fitted on reference, query, and combined data sets, respectively.

Model	BIC(H_1)	BIC(H_2)	BIC(H_1) – BIC(H_2)	Best explain by a single model
Linear	2070.90	2046.28	24.61	FALSE
Spline	1152.70	1154.43	-1.73	TRUE
Sin	50.07	55.79	-5.72	TRUE

(c) BIC values for H_1 and H_1 across three different models, evaluated on reference and query data sets.

Figure 2.5: The choice of model for fitting time course data significantly impacts subsequent inferences.

2.2.3 Splines offer a flexibility function type to describe gene expression dynamics

In the following section, we explore the use of splines as our chosen model. We begin by considering situations where a mechanistic model, one that explicitly describes the underlying biological processes, is not available. Instead, we rely on more flexible function classes that can capture the observed dynamics without imposing specific assumptions about the mechanisms involved. In this context, we evaluate several polynomial models for their suitability in fitting time series data. These models are chosen for their ability to approximate complex behaviours that are often present in gene expression dynamics. By comparing different polynomial functions, we aim to identify the most effective model for accurately capturing the temporal patterns of gene expression.

We utilised publicly available Arabidopsis [148] and *B. rapa* [85] datasets, each comprising approximately 24,686 genes. These time-series datasets were generated from apex samples taken over developmental time during the vegetative growth phase and the floral transition. The data were fitted to six different polynomial and spline models (see Methods Section 2.1.5 and2.1.5). Figure 2.6 and 2.7 display the BIC values calculated for each model across individual genes.

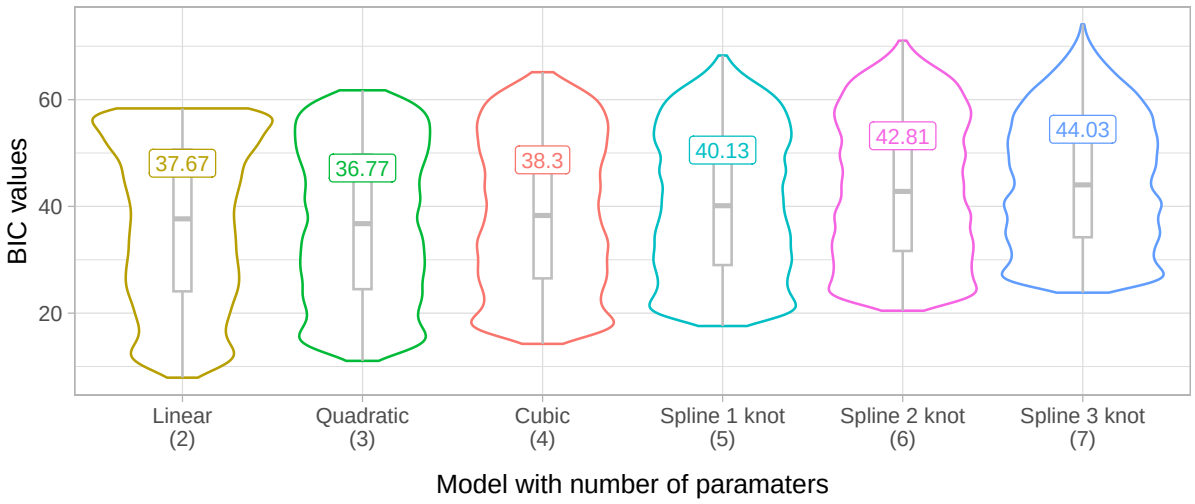


Figure 2.6: Violin plots representing BIC values evaluated on the gene expression of Arabidopsis using six different models. Inside each violin plot is a box plot summarising the distribution range and the individual median (a central line inside the box and a labelled numerical value).

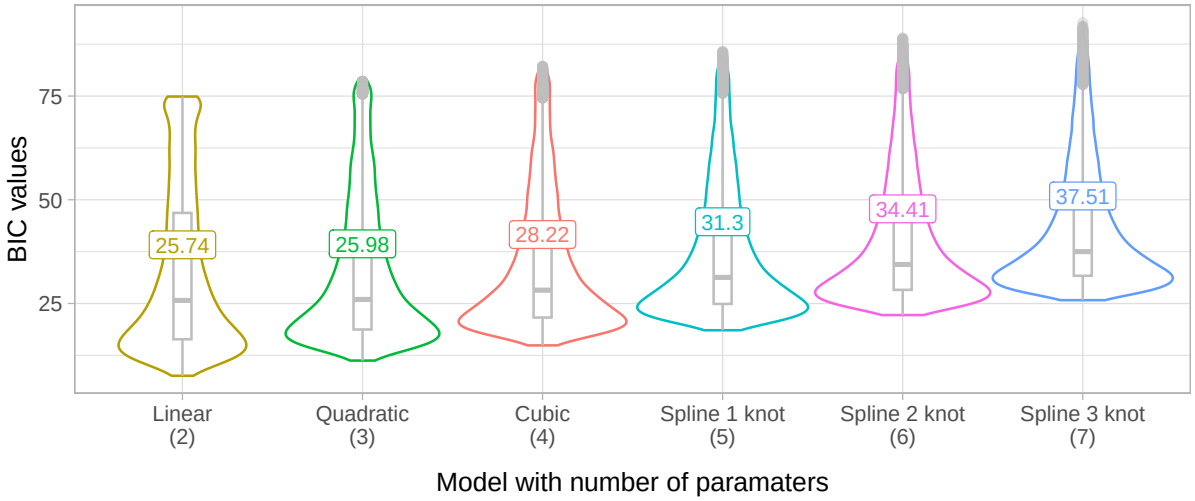


Figure 2.7: Violin plots representing BIC values evaluated on the gene expression of *B. rapa* using six different models. Inside each violin plot is a box plot summarising the distribution range and the individual median (a central line inside the box and a labelled numerical value).

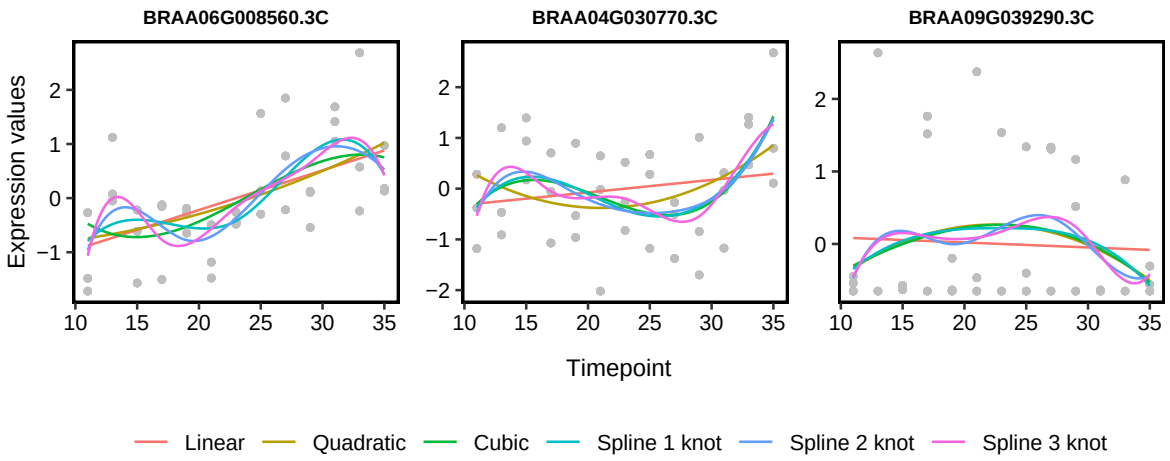
For both *B. rapa* and Arabidopsis, the BIC values for the quadratic and linear models are consistently lower than those for other models, indicating that these models provide the best fit for the data. This is partly because time series that show minimal or no changes across the time course tend to favour simpler models, such as a linear fit. In our analysis, we used a tool we developed, called greatR, which implements a pipeline for time-course gene expression analysis, including pre-processing, model fitting, and curve registration. Within greatR, such low-dynamic time series are filtered out during pre-processing to focus on more informative gene expression patterns. However, even after excluding these time series, the overall ranking of the models remains unchanged, although the margin between the top two models decreases by over 863 genes.

Despite the exclusion of unchanged expression profiles, the linear and quadratic models remain the best fit due to the inherent characteristics of the data. The majority of the expression profiles exhibit high variance and noise (see Figure 2.8, 2.9, and 2.10). Figure 2.8, 2.9, and 2.10 illustrates examples of gene expression data best fitted by a linear model, a quadratic model, and a spline model with one knot, respectively. Although the linear and quadratic models provide the best fit for most gene profiles, they may not adequately capture more complex gene expression dynamics.

The cubic B-spline, on the other hand, offers significant advantages due to its greater flexibility in modelling complex, non-linear relationships between data points which are also characteristic of gene expression data [149]. Unlike quadratic models, which impose a specific form on the data, splines can adapt more naturally to the subtle changes and intricate patterns over time, capturing the true underlying biological processes more effectively. It is regularly used for building explanatory models for biological data, such as expression profiles [129, 150]. Although spline models may result in slightly higher BIC values due to their additional parameters, we consider this a

worthwhile trade-off in return for improved capacity to reflect underlying biological processes more realistically.

Therefore, rather than dismissing the BIC results, we interpret them alongside domain-specific considerations. The cubic B-spline with one knot is chosen as the default model not because it always has the lowest BIC, but because it offers a reasonable balance between flexibility, interpretability, and robustness across diverse datasets. We point out that the best model very much depends on the data and should be evaluated for each case. Some examples of evaluation of different fitting functions are shown in Figure 2.8, 2.9, and 2.10. As is evident, the best model very much depends on the data (see also Figure 2.5) but we find that cubic B-splines with one knot are a reasonable default option.

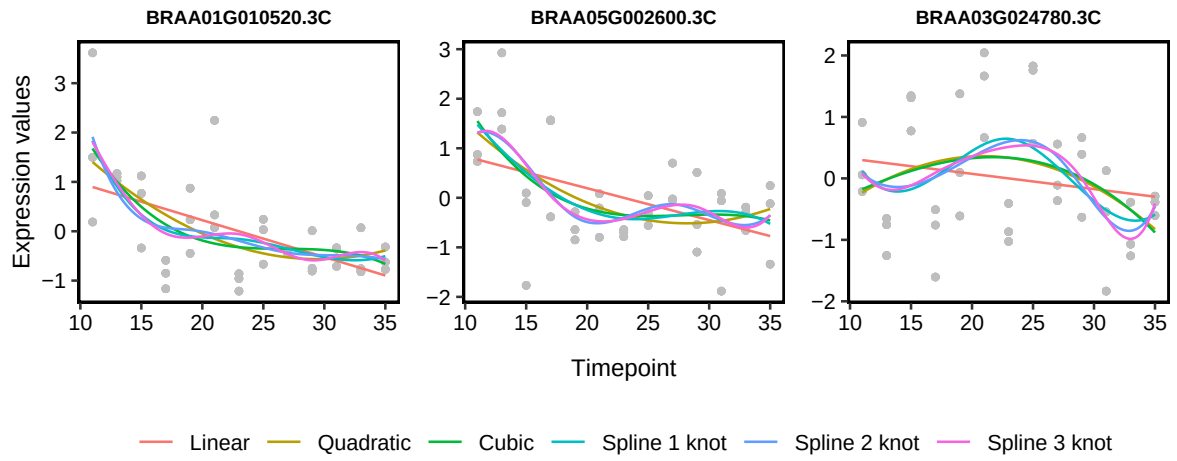


(a)

Gene ID	Linear	Quadratic	Cubic	Spline 1 knot	Spline 2 knot	Spline 3 knot
BRAA06G008560	13.62	17.13	20.74	20.04	23.51	27.03
BRAA04G030770	17.88	20.51	22.79	22.77	26.35	29.76
BRAA09G039290	40.54	41.24	44.82	45.06	47.51	51.40

(b)

Figure 2.8: Examples of three different genes in *B. rapa*, best fitted with a linear model. (a) Data points were plotted with six distinct fitted models. Grey dots represent individual replicates at each time point, while coloured lines correspond to different models used for fitting. (b) Corresponding BIC values for each gene, were evaluated across six models.

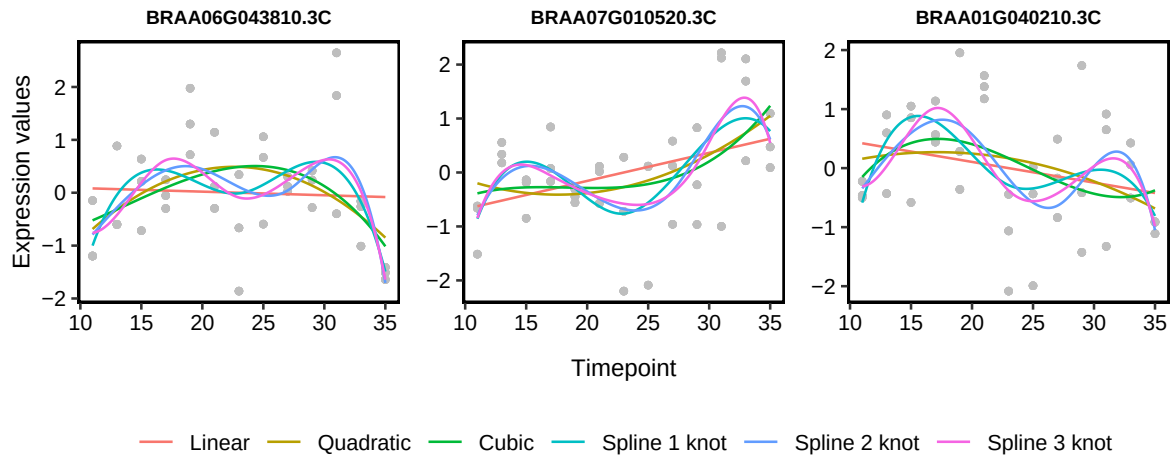


(a)

Gene ID	Linear	Quadratic	Cubic	Spline 1 knot	Spline 2 knot	Spline 3 knot
BRAA01G010520	53.21	51.19	52.99	51.95	55.23	58.82
BRAA05G002600	43.07	41.84	43.89	44.05	46.88	50.59
BRAA03G024780	70.82	68.88	72.47	69.17	71.67	75.00

(b)

Figure 2.9: Examples of three different genes in *B. rapa*, best fitted with a quadratic model. (a) Data points were plotted with six distinct fitted models. Grey dots represent individual replicates at each time point, while coloured lines correspond to different models used for fitting. (b) Corresponding BIC values for each gene, were evaluated across six models.



(a)

Gene ID	Linear	Quadratic	Cubic	Spline 1 knot	Spline 2 knot	Spline 3 knot
BRAA06G043810	68.41	59.57	62.67	54.33	56.17	59.47
BRAA07G010520	53.36	53.91	56.43	49.44	51.92	54.98
BRAA01G040210	60.54	63.41	65.10	56.48	57.52	60.55

(b)

Figure 2.10: Examples of three different genes in *B. rapa*, best fitted with a spline model with one knot. (a) Data points were plotted with six distinct fitted models. Grey dots represent individual replicates at each time point, while coloured lines correspond to different models used for fitting. (b) Corresponding BIC values for each gene, were evaluated across six models.

2.2.4 Two time series over different ranges can be transformed to identify common dynamical features

When comparing gene expression data across different datasets, variations in time ranges or developmental timescales can introduce significant challenges. Gene expression profiles often vary depending on the specific stage of development being studied or the experimental conditions under which the data were collected. These differences can result in datasets that do not align directly, making it difficult to identify common dynamical features that might be shared across different biological contexts. In such cases, it becomes essential to apply transformations to one of the datasets (see Figure 2.11).

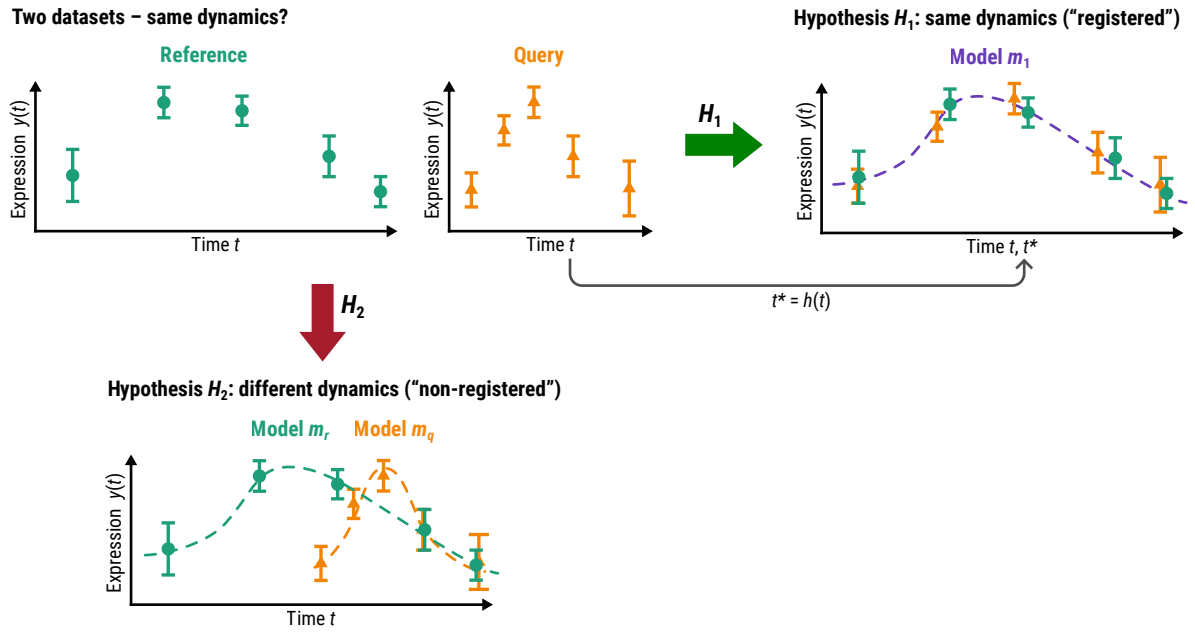


Figure 2.11: Evaluating whether two datasets with different time points are the same (within experimental error) depends on the choice of model. We depict two datasets (the reference against which we wish to compare the query dataset). The green (reference) and orange (query) expression values have different time points associated with them, hindering a simple comparison of their corresponding expression values. We seek a transformation of time, $t^* = h(t)$, that makes the datasets as similar as possible. We can then ask whether the datasets are best explained by a single model (a function $m_1(\theta_1, t)$ with parameters θ_1), or whether two models, (functions $m_r(\theta_r, t)$ and $m_q(\theta_q, t)$ with parameters θ_r and θ_q that may be different between the functions), provide a statistically better explanation. Hypothesis H_1 = one model (same dynamics of the data), hypothesis H_2 = two models (different dynamics of the data).

The goal is to adjust the time values while preserving the expression values, y_i, so that the similarity between the two time series is maximised. Specifically, we aim to identify a transformation function $t^*_{q,j} = h(\beta, t_{q,j})$, where $h(\beta, t_{q,j})$ is a function with parameters β that modifies the time points in the query dataset (see Figure 2.11). By applying this transformation, the query time series is mapped onto a new timescale, producing transformed time values $t^*_{q,j}$. The resulting

transformed query dataset, denoted as $q^* = (t^*_{q,j}, y_{q,j})$, allows for a more direct comparison with the reference dataset.

In general, we are searching for a function, $h(t)$, such that the difference between the query and reference datasets is minimal. The function $h(t)$ is parameterised by a set of parameters β, and we denote the number of these parameters by k_h. To ensure the chronological order of time points remains unchanged, it is crucial that $h(\beta, t)$ is a monotonic function of t, providing order-preserving, one-to-one mapping. This monotonicity guarantees that the transformation does not invert or reorder the time points.

To determine whether the datasets exhibit similar dynamics and can be explained by a single model, we can follow the procedure outlined above, with the additional step of incorporating the time transformation into the hypothesis H_1 (Figure 2.11). Under H_1, the likelihood function now evaluates the model against the transformed query dataset $q^* = (t^*_{q,j}, y_{q,j})$.

The distance function for the likelihood under H_1 between the model and the reference dataset is defined as

$$d(m^*_1, r) = \left[\sum_{i=1}^{N_r} \left(m^*_1(\theta, t_i) - y_{r,i}\right)^2\right]^{1/2}. \tag{2.41}$$

Similarly, the distance function between the joint model and the transformed query dataset is

$$d(m^*_1, q) = \left[\sum_{j=1}^{N_q} \left(m^*_1(\theta, h(t_j)) - y_{q,j}\right)^2\right]^{1/2}. \tag{2.42}$$

The likelihood function $\Lambda_{H_1}(\theta_1, \beta)$ under hypothesis H_1 is then given by

$$\Lambda_{H_1}(\theta_1, \beta) \propto \prod_{i=1}^{N_r} \exp\{-[m_1(\theta_1, t_{r,i}) - y_{r,i}]^2/2\sigma^2_{r,i}\} \prod_{j=1}^{N_q} \exp\{-[m_1(\theta_1, t^*_{q,j}) - y_{q,j}]^2/2\sigma^2_{q,j}\}, \tag{2.43}$$

where $m_1(\theta, t)$ is the joint model to both the reference and transformed query dataset, which uses the transformed time points $t^*_{q,j}$ along with the original expression values $y_{q,j}$.

The additional parameters introduced by the transformation function h can be accounted for in the BIC calculation for H_1. The BIC for H_1 is adjusted as follows

$$\mathrm{BIC}(H_1) = -2\mathcal{L}_{H_1} + (k_{m_1} + k_h) \log N_D, \tag{2.44}$$

where k_h represents the number of parameters associated with the transformation function h.

In contrast, the BIC(H_2) remain unchanged:

$$\mathrm{BIC}(H_2) = -2\mathcal{L}_{H_2} + (k_{m_r} + k_{m_q}) \log N_D, \tag{2.45}$$

since H_2 does not involve the transformation function h, and thus no additional parameters are introduced under this hypothesis. The number of data points included in the distance function can be adjusted depending on whether the goal is to identify global or local similarities in the time series.

As discussed, these comparisons rely on the presence of an underlying model. When dealing with data over different timescales, it is necessary to have both a model for the dynamics and a model for the transformation. The choice of these models is crucial, as all subsequent inferences will depend on how well these models represent the underlying biological processes.

To ensure that the transformation function $h(\beta, t)$ is both meaningful and biologically relevant, it is useful to limit the function space to transformations that reflect common biological phenomena. For instance, a simple but effective transformation could involve parameters that represent a delay or shift in time, β_1, and a change in the timescale, such as a stretch factor, β_2. This can be expressed as

$$h(t) = \beta_1 + \beta_2 t. \tag{2.46}$$

In this case, the number of transformation parameters k_h is equal to 2.

The problem of comparing datasets can thus be reduced to determining the optimal values for the transformation parameters β_1 and β_2 which minimise the distance $d(r, q^*)$ between the reference data r and the transformed query data q. When using a cubic spline function with one knot, this leads to $5+2=7$ parameters for hypothesis H_1 and to $5+5=10$ parameters for hypothesis H_2.

The difference in BIC values between hypothesis H_1 (one model) and H_2 (two models) can then be used to evaluate the similarity between the two time series. If values for the transformation parameters β_1 and β_2 can be identified such that a single joint model provides a better explanation of the data than two separate models–indicated by a negative difference in BIC values:

$$\mathrm{BIC}(H_1) - \mathrm{BIC}(H_2) < 0 \tag{2.47}$$

we can conclude that the expression profiles can be *registered*. In this context, registration means that the two time series share enough similarity in their dynamics to be described by a single model. The larger the negative difference in BIC values, the stronger the statistical evidence supporting this conclusion.

To achieve this, the task of minimising the distance between the time series, or equivalently maximising the difference in BIC values, can be efficiently carried out using well-established optimisation algorithms. These optimisations are performed specifically to find the parameter

values that best align the time series, thereby enabling the identification of common dynamical features across datasets that may initially appear distinct.

2.2.5 Optimisation of BIC can be used to find similarities between two time series

Maximising the likelihood Λ_{H_1} to get the optimal BIC value

To determine if two time-course datasets can be aligned, we compute the difference in the BIC values between two hypotheses: H_1, which allows for a transformation between the datasets, and H_2, which assumes no transformation. This difference is influenced by the chosen model– in this case, cubic B-splines with one knot–and the applied transformation function, here is a linear function. The BIC value for H_2, BIC(H_2), and the likelihood Λ_{H_2} are independent of the transformation parameters, meaning they remain fixed regardless of any transformations applied between the datasets. On the other hand, BIC(H_1), which accounts for the transformation, varies based on the model parameters θ and the transformation parameters β. Since the number of parameters is fixed, maximising the difference between BIC(H_2) and BIC(H_1) is equivalent to minimising BIC(H_1). This requires us to maximise the likelihood Λ_{H_1} by optimising over both the model parameters θ and the transformation parameters β.

Figure 2.12 illustrates the optimisation process for the likelihood Λ_{H_1}. The goal of this optimisation is to find the maximum value of Λ_{H_1} within the parameter space. The parameters that yield the highest Λ_{H_1} will be selected as the candidate transformation parameters. If these parameters result in a negative BIC difference, i.e., $\text{BIC}(H_1) - \text{BIC}(H_2) < 0$, it suggests that the two datasets can be explained by a common underlying structure. In other words, the transformation model H_1 provides a better fit to the data than the no-transformation model H_2. A more negative BIC difference strengthens the evidence for alignment, indicating that the datasets are more likely to conform to a shared model.

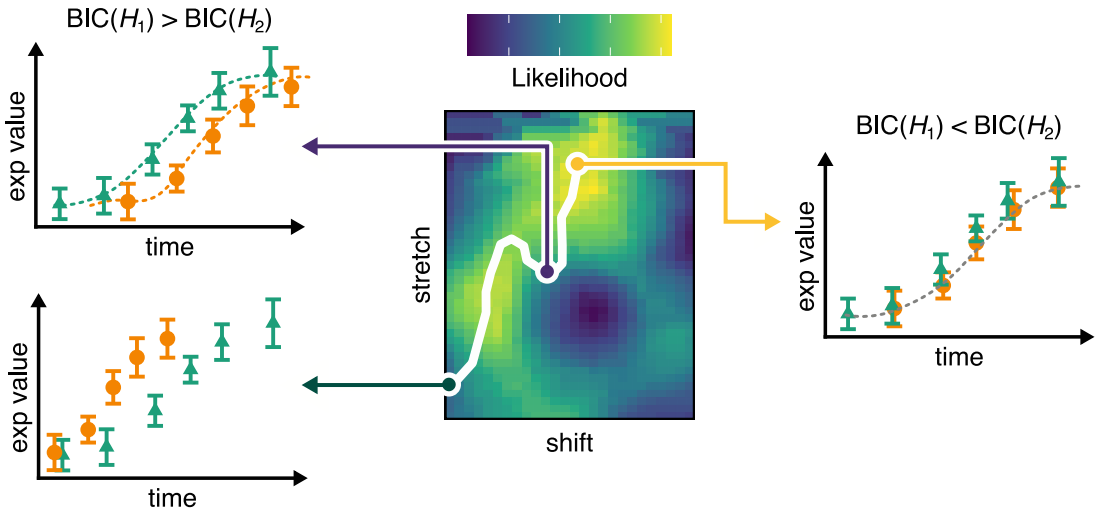


Figure 2.12: Illustration on how likelihood is optimised. The heatmap represents the shift and stretch parameter space. Starting from the initial value of parameters, the optimiser will find the optimal stretch and shift which maximise Λ_{H_1}.

Several well-established optimisation techniques are utilised for maximising the log-likelihood value, each tailored to different use cases. The inclusion of L-BFGS-B, Nelder–Mead (NM), and Simulated Annealing (SA) reflects their complementary strengths, allowing users to choose the method that best fits their specific needs. More details of the implementation of these methods are discussed in the following chapter.

2.2.6 Summary

This chapter presents the theoretical framework and method formulation for comparing two time series. It outlines the key components of our approach, including the development of hypotheses, the construction of the log-likelihood function, and the application of curve registration techniques to align gene expression data. The use of B-splines for curve fitting plays a pivotal role in modelling the complex, non-linear dynamics that are characteristic of gene expression profiles. While simpler models like linear or quadratic regression may offer a good fit based on the BIC score, they often fail to capture the full complexity of the underlying biological processes. In contrast, B-splines provide the necessary flexibility to account for the intricate temporal patterns observed in gene expression data. The BIC is utilised for model selection, balancing model fit and complexity. Optimisation strategies are employed to accurately estimate the stretch and shift parameters for aligning the two time series, ensuring that phase differences are effectively handled. Together, these methods form a cohesive analytical process designed to address the unique challenges of time-course gene expression analysis.

3 | Development of associated R package: *greatR* and method testing using simulated data

3.1 Introduction

To ensure that the methodologies developed in this research are accessible to the broader scientific community, we have developed an R package named *greatR* (Gene Registration from Expression and Time-courses in R). The primary aim of *greatR* is to offer a user-friendly yet robust tool for the analysis of gene expression data, particularly suited to time-course studies and complex developmental processes, where precise alignment and comparison of expression profiles over time are essential.

This package is designed to simplify and automate the process of curve registration, enabling researchers to align gene expression trajectories across different conditions or samples. By providing a streamlined workflow and robust computational algorithms, *greatR* facilitates the exploration of dynamic biological phenomena that are otherwise difficult to quantify due to inherent variability in temporal data.

In this section, we will provide an in-depth description of the package's features, including data requirements, core registration functions, and post-processing utilities. Specifically, we will cover the algorithms that underpin the registration process, offering insights into how the package achieves robust alignment of gene expression curves. The registration process within *greatR* is built around the `register()` function, which implements advanced statistical and computational techniques for aligning gene expression curves. The input data can be provided in multiple formats, including single data frames, lists of data frames, or vectors, offering flexibility for various experimental setups. The algorithm behind the `register()` function applies a combination of spline-based models and time-warping techniques to accurately align expression profiles while accounting for potential noise in the data.

Once the registration is performed, *greatR* provides tools for summarising, visualising, and comparing the aligned data. This includes distance metrics, which allow users to quantify the similarity between samples, and various visualisation functions that offer intuitive ways to explore the results. These tools are designed to enhance the interpretability of the output and support downstream analyses, such as clustering and differential expression analysis.

Furthermore, we will demonstrate the usability of *greatR* through practical examples that illustrate how the package can be applied to real-world data. Figure 3.1 presents an overview of the registration process, showcasing the flow from input data preparation, through registration, to final outputs such as visualisations and statistical summaries. This section will also provide details on how the core algorithms are implemented, offering insights into the internal workings of the package.

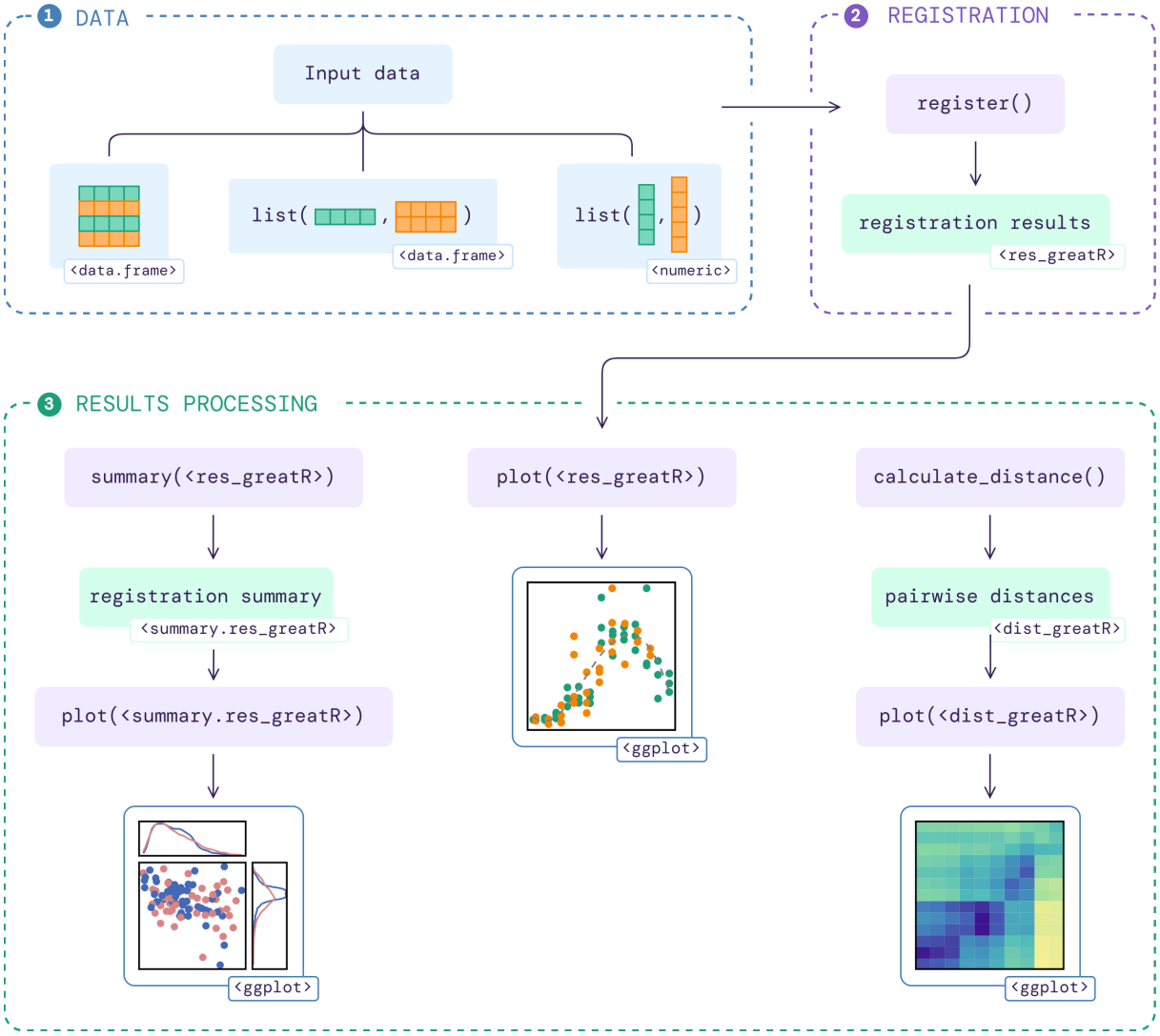


Figure 3.1: Flowchart overview of registering gene expression using *greatR*.

3.2 Implementation details

3.2.1 Object-oriented design: S3

In *greatR*, we utilise the S3 object-oriented system in R to provide a flexible and intuitive interface. The S3 system is particularly suited for packages where simplicity and extensibility are key, as it allows for the easy creation of generic functions that can operate differently depending on the class of the object passed to them. S3 is well-documented and manageable, making it a practical choice for package development. Unlike S4 or R6, which offer more rigid or familiar structures for those with a background in other programming languages, S3 supports the creation of flexible and generic R functions [151]. By choosing S3, we ensure that the package remains intuitive for R users while still providing the necessary extensibility for various data structures and methods.

3.2.2 Package dependencies

The development of *greatR* is supported by several essential R packages, each fulfilling a specific role in data processing, visualisation, statistical computation, and performance optimisation. For data transformation and wrangling, *data.table* [152] is employed due to its efficiency in handling large datasets with minimal memory usage and high speed. Visualisation tasks within the package are managed using *ggplot2* [153], which provides a powerful grammar of graphics for creating complex plots. Additionally, *patchwork* [154] is integrated to facilitate the combination of multiple plots into a single coherent visualisation, while *scales* [155] is used to customise axis scales and legends, enhancing the clarity and impact of the visual outputs.

For statistical calculations, while many are implemented ad-hoc using the base *stats* [156] package, we utilised *optimization* [157] and *neldermead* [158] to find the optimal registration parameters, ensuring precise alignment and accuracy in analyses. To handle large amounts of gene data and significantly reduce computation time, *furrr* [159] and *future* [160] are used for parallel computation, taking full advantage of multi-core processors.

Finally, to make it more user-friendly, particularly in the command-line interface, *cli* [161] is used to construct clear and informative messages and warnings. This ensures that users receive feedback that is both aesthetically pleasing and functionally helpful.

3.2.3 License and package version

Version 2.0.0 of the package is available from the Comprehensive R Archive Network (CRAN) at `https://CRAN.R-project.org/package=greatR` and GitHub at `https://github.com/ruthkr/greatR/` and is distributed under the GPL-3.0 license (GNU General Public License v3.0). Users can easily download and install it, and learn how to use it with articles provided on `https://ruthkr.github.io/greatR/`.

3.2.4 Data requirements

For the *greatR* package, the input data must include both reference (r) and query (q) data, regardless of the format used. If using a single data frame or a data frame within a list, it should contain time-course gene expression data along with all replicates. Figure 3.2 illustrates the required structure of the data frame. This data frame must include both reference and query expression data, organised into the following five columns:

- `gene_id`: The locus name or a unique identifier for each gene.
- `accession`: The accession or name of the reference and query data.
- `timepoint`: The time points corresponding to the gene expression data.

- `expression_value`: The expression values or measures of gene or transcript abundance, such as RPM, RPKM, FPKM, TPM, TMM, or raw read counts.
- `replicate`: The biological replicate ID associated with an expression value at a specific time point.

gene_id	accession	timepoint	expression_value	replicate
gene_1	reference	...	...	1
gene_1	reference	...	...	2
gene_1	query	...	...	1
gene_1	query	...	...	2
...	...	...	...	...
gene_n	reference	...	...	1
gene_n	query	...	...	1
...	...	...	...	...

Figure 3.2: Illustration showing a data frame input format required by *greatR*, containing the gene expression profiles for all replicates of both the reference and query across time points.

If users do not have the reference and query data combined into a single data frame with mapped IDs, they can provide the input as a list of data frames. As illustrated in the diagram below, this list must include separate reference and query data frames, each containing the required columns as specified for the single data frame input (see the previous section). It is important to note that the elements of the list must be named reference and query; however, the order of these elements will not affect the registration process.

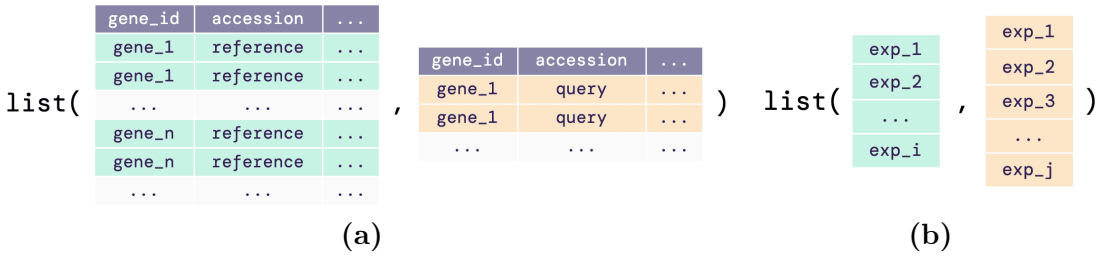


Figure 3.3: Illustration showing (a) a list of reference and query data frames and (b) a list of reference and query vectors, which can be an optional input in *greatR*.

As an alternative to using a list of data frames as input, users can also provide a list of numerical vectors. The illustrative Figure 3.3 (b) demonstrates the required structure for this input format. Since the vectors do not include specific IDs, *greatR* will automatically assign unique IDs to each reference and query pair (for more details, see the documentation register data > Using other inputs). More details about some examples can be accessed through the documentation in https://ruthkr.github.io/greatR/articles/data-requirement.html.

3.2.5 Pre-processing data and registration process

The registration method explained in the previous section is implemented into several functions and wrapped into the main functions called `register()` (see Algorithm 3.1 for more details). The arguments of this function:

Arguments	Description
`input`	An input data containing the gene expression profiles for all replicates of both the reference and query across time points.
`stretches`	A numeric vector of candidate stretch factors to apply to the query data. This argument is used to adjust the time scale of the query relative to the reference and is required only if `use_optimisation = FALSE`.
`shifts`	A numeric vector of candidate shift values to apply to the query data. This argument is used to adjust the starting point of the query relative to the reference and is required only if `use_optimisation = FALSE`.
`reference`	An accession name or identifier for the reference dataset.
`query`	An accession name or identifier for the query dataset. This dataset will be aligned to the reference.
`scaling_method`	A scaling method to apply to the data before the registration process. Options are `"none"` (default, no scaling), `"z-score"`, or `"min-max"` (see Methods 2.1.7).
`overlapping_percent`	A numeric value indicating the minimum percentage of overlapping time points required between the reference and query after applying shifts. Shifts that result in less than this percentage of overlap will be excluded from consideration.
`use_optimisation`	A logical value indicating whether to optimise the registration parameters automatically. If `TRUE` (default), the function will determine the optimal stretch and shift values; if `FALSE`, user-specified `stretches` and `shifts` will be used.
`optimisation method`	A string specifying the optimisation algorithm to use when `use_optimisation = TRUE`. Options include `"lbfgsb"` for the L-BFGS-B algorithm (default), `"nm"` for the Nelder–Mead method, or `"sa"` for Simulated Annealing (see Methods 2.1.6).
`optimisation_config`	An optional list of arguments to override the default configuration of the chosen optimisation method. This allows for customisation of the optimisation process.
`exp_sd`	An optional numeric value representing the experimental standard deviation of the gene expression replicates. This can be used to weight the registration process based on experimental variability.
`num_cores`	An optional integer specifying the number of processor cores to use for parallel processing when registering genes asynchronously. If `NA` (default), the registration will be performed sequentially.

Algorithm 3.1 Algorithm for the registration function.

Input: Data D, scaling method s, experimental standard deviation on the expression replicates σ_{exp} (optional), overlapping percent λ, optimisation method M.
Output: Registered data D^{reg}, model comparison `model_comp`.

function REGISTER(D, λ, M)
    $D \leftarrow$ PREPROCESS_DATA(D, s, σ_{exp})
    $G \leftarrow$ list of unique gene IDs in D

    for $g \in G$ **do** ▷ Iterate over each gene
        $D_g \leftarrow$ filter D for the current gene
        $\mathcal{L}_{H_2} \leftarrow$ CALC_LOGLIK_H2(D_g)
        $\theta \leftarrow$ OPTIMISE(D_g, λ, M) ▷ Optimise parameters
        $\beta_2 \leftarrow \theta_{\beta_2}$
        $\beta_1 \leftarrow \theta_{\beta_1}$
        $D_g^{reg} \leftarrow$ APPLY_REGISTRATION(D_g, β_2, β_1) ▷ Apply registration
        $\mathcal{L}_{H_1} \leftarrow \theta_{\mathcal{L}}$
        `model_comp` $\leftarrow$ COMPARE_H1_H2(D_g^{reg}, β_2, β_1, $\mathcal{L}_{H_1}$, $\mathcal{L}_{H_2}$) ▷ Model comparison
        `reg_res` $\leftarrow \{D_g^{reg}$, `model_comp`$\}$ ▷ Store results
    end for

    Combine `reg_res` results into D^{reg} and `model_comp` from all genes
    $D^{reg} \leftarrow$ merge registered data D^{reg} with original data D
    return `reg_res` $\leftarrow \{D^{reg}$, `model_comp`$\}$
end function

Pre-processing data

Pre-processing the input data is a crucial preliminary step performed before the registration process in *greatR*. This step ensures that the data are clean, consistent, and ready for accurate alignment and analysis. Several key procedures are undertaken during pre-processing:

- Filtering genes: Genes that are present in only one dataset (either the reference or the query) are filtered out to ensure that the registration process operates on genes present in both datasets.
- Filtering low-variance genes: Genes whose expression levels do not exhibit significant variation over time are removed. This step is essential to focus the analysis on genes with dynamic expression patterns.
- Scaling: The gene expression data are scaled according to the method specified by the user via the `scaling_method` argument. This step helps to standardise the range or distribution of expression values, making them comparable across different conditions or datasets.
- Estimating variance: The variance for each time point for every single gene is estimated during this pre-processing.

These pre-processing steps are critical for minimising noise and discrepancies in the data for the subsequent registration process.

Implementation of optimisation methods to maximise the likelihood and parameter settings in *greatR*

We utilised three well-established optimisation techniques: L-BFGS-B, Nelder–Mead (NM), and Simulated Annealing (SA), each offering complementary strengths, allowing users to choose the method that best suits their specific needs. L-BFGS-B is the default optimiser due to its superior speed, making it the fastest among the three. Implemented via the *stats* package [156], it is particularly well-suited for users handling a large number of curves or datasets, typically involving thousands of gene pairs, where computational efficiency is essential.

Nelder–Mead (NM), implemented through the *neldermead* package [158], is also relatively fast and commonly used for large datasets, such as gene expression matrices with thousands of genes measured across multiple time points. NM, being a simplex method, can sometimes fail to converge to the global optimum. To mitigate this, we implemented NM to run in three rounds, which makes it slightly slower than L-BFGS-B. However, the solutions it produces are generally comparable to those of the more robust Simulated Annealing method. NM is a good option for users needing quick, approximate solutions across numerous curve pairs.

Simulated Annealing (SA), available through the *optimization* package [157], is the most robust of the three methods, capable of exploring a wide range of potential solutions. SA is particularly useful when working with a small number of curve pairs, typically in the hundreds or fewer, where a thorough exploration of the solution space is required. Although SA is computationally expensive, its ability to avoid local minima makes it invaluable when optimal accuracy is more important than speed.

In summary, L-BFGS-B is the default optimiser because of its speed and reliability in large-scale scenarios. NM provides an alternative for simpler optimisation tasks, offering a good balance between speed and performance, though it may not always find the optimal solution. SA is best suited for cases where solution robustness is critical, despite its higher computational cost, and is ideal for users working with fewer curve pairs.

The optimisation process in *greatR* is carried out using the method selected by the user via the `optimisation_method` argument. All three methods are constrained optimisation techniques that require minimum and maximum boundary values for each parameter, creating a boundary box for the optimisation. If the stretch and shift values are not specified by the user, initial values are estimated using a default setting that transforms the query data to cover at least 50% of the total range of the reference data. For users who wish to achieve local or global alignment between the reference and query data, the overlapping percentage can be adjusted by setting the `overlapping_percent` parameter to a smaller or bigger value.

When the optimisation process is not utilised (`use_optimisation = FALSE`), the parameters for stretches and shifts must be explicitly defined. The parameter space for the registration process includes these stretches and shifts, with the best parameters being selected based on their log-likelihood values through an iterative search. The BIC score is then calculated for the selected parameter vector to assess whether the reference and query data exhibit similar dynamics (see Algorithm 3.2).

Algorithm 3.2 Algorithm to calculate limits of the search space.

Input: data D, overlap parameters $\lambda_l = 0.5$, $\lambda_u = 1.5$
Output: Search space $S = \{\beta_{2,0}, \beta_{2,l}, \beta_{2,u}, \beta_{1,0}, \beta_{1,l}, \beta_{1,u}\}$

Note: Subscripts l and u refer to **lower** and **upper** bounds, respectively.
function GET_SEARCH_SPACE_LIMITS(D, λ_l, λ_u)
    $S_{\beta_2} \leftarrow$ GET_STRETCH_SEARCH_SPACE_LIMITS(D, λ_l, λ_u)
    $S_{\beta_1} \leftarrow$ GET_SHIFT_SEARCH_SPACE_LIMITS(D, S_{β_2}, λ_l, λ_u)
    return $S \leftarrow S_{\beta_2} \cup S_{\beta_1}$
end function

function GET_STRETCH_SEARCH_SPACE_LIMITS(D, λ_l, λ_u)
    $\beta_{2,0} \leftarrow$ GET_APPROXIMATE_STRETCH(D)
    $\beta_{2,l} \leftarrow \lambda_l \beta_{2,0}$
    $\beta_{2,u} \leftarrow \lambda_u \beta_{2,0}$
    return $S_{\beta_2} = \{\beta_{2,0}, \beta_{2,l}, \beta_{2,u}\}$
end function

function GET_SHIFT_SEARCH_SPACE_LIMITS(D, S_{β_2}, λ_l, λ_u)
    $R_r \leftarrow [R_{r,-} = \min(t_r), R_{r,+} = \max(t_r)]$ ▷ range of timepoints for reference data D_r
    $R_q \leftarrow [R_{q,-} = \min(t_q), R_{q,+} = \max(t_q)]$ ▷ range of timepoints for query data D_q
    $t_\pm \leftarrow R_{r,\pm} \mp \lambda_l * \text{diff}(R_r) \pm \beta_{2,u} * \text{diff}(R_q)$ ▷ max (+) and min (−) possible timepoints

    $\beta_{1,l} \leftarrow t_- - \beta_{2,u} \times R_{r,-}$
    $\beta_{1,u} \leftarrow t_+ - \beta_{2,l} \times R_{r,-}$
    $\beta_{1,0} \leftarrow$ midpoint between $\beta_{1,l}$ and $\beta_{1,u}$
    return $S_{\beta_1} \leftarrow \{\beta_{1,0}, \beta_{1,l}, \beta_{1,u}\}$
end function

The approximation of standard deviation values

For most biological data, the number of replicates are quite limited, for around 3 to five data points per time points. This makes it difficult to calculate the standard deviation as for accurate calculation, a high number of data points are required. Therefore, there are two different ways we implement to approximate the standard deviation value, shown in Algorithm 3.3.

Algorithm 3.3 Algorithm to calculate variance σ^2 for observed expression data.

Input: expression data D with expression values y, experimental standard deviation on the expression replicates σ_{exp} (optional).
Output: expression data D with variance σ^2 for all time points.

function CALC_VARIANCE(D, σ_{exp})
    if σ_{exp} is provided **then**
        $\sigma^2 \leftarrow \sigma^2_{\text{exp}}$ ▷ Fixed σ^2 for all time points
    else
        Group by `gene_id`, `accession`, `timepoint`
        Split data into D_{r} (with replicates) and D_{nr} (with no replicates)
        if D_{r} is not empty **then**
            $\sigma^2_P \leftarrow \max(y)$ ▷ Poisson estimate for expression variance
            $\sigma^2_r \leftarrow (\text{range}(y)/10)^2$ ▷ Global expression variance
            $\sigma^2 \leftarrow \max(\sigma^2_P, \sigma^2_r)$
        end if
        if D_{nr} is not empty **then**
            $\sigma^2 \leftarrow \max(y/10, 0.25)$
        end if
        $D \leftarrow D_{\text{r}} \cup D_{\text{nr}}$ ▷ Combine data with individual σ^2 for all time points
    end if
    return D
end function

Registration results

The function `register()` returns a list with S3 class `res_greatR` containing three different objects:

Return objects	Description
`data`	A data frame containing the expression data and an additional `timepoint_reg` column which is a result of registered time points by applying the registration parameters to the query data.
`model_comparison`	A data frame containing (a) the optimal stretch and shift for each `gene_id` and (b) the difference between Bayesian Information Criterion for the separate model and for the combined model (`BIC_diff`) after applying optimal registration parameters for each gene. If the value of `BIC_diff < 0`, then expression dynamics between reference and query data can be registered (`registered = TRUE`). (Default S3 print).
`fun_args`	A list of arguments used when calling the function (`reference`, `query`, `scaling_method`, ...).

Example of registration on *B. rapa* and Arabidopsis data

greatR provides an example data frame containing two different species, *A. thaliana* and *B. rapa*, with two and three different replicates, respectively. To align this data frame containing gene expression time-course between Arabidopsis *Col-0* and *B. rapa R-o-18*, we can use the function `register()`. When using the default `use_optimisation = TRUE`, *greatR* will find the best stretch and shift parameters through optimisation.

Code 3.4: Running the registration process using `register()` function.

```
# Load the package
library(greatR)

# Load sample data
b_rapa_data <- system.file(
  "extdata/brapa_arabidopsis_data.csv",
  package = "greatR"
) |>
  data.table::fread()

# Run registration
registration_results <- register(
  b_rapa_data,
  reference = "Ro18",
  query = "Col0",
  scaling_method = "z-score"
)

# Registration results
registration_results$model_comparison
```

The results of *greatR* contain the registration information for each gene. These include the ID of the gene, the optimal stretch and shift parameters, the BIC score, and the information on whether the gene is registered or not (see Table 3.1 below as an example).

gene_id	stretch	shift	BIC_diff	registered
BRAA02G018970.3C	4.00	-30.98	2.83	FALSE
BRAA02G043220.3C	2.45	-10.60	-3.57	TRUE
BRAA03G023790.3C	2.25	-4.36	-7.85	TRUE
BRAA03G051930.3C	3.10	-12.56	-8.10	TRUE
BRAA04G005470.3C	3.53	-20.25	-7.54	TRUE
BRAA05G005370.3C	2.28	-5.03	-7.73	TRUE
BRAA06G025360.3C	2.38	-8.02	-6.50	TRUE
BRAA07G030470.3C	4.00	-27.03	-5.45	TRUE
BRAA07G034100.3C	4.00	-27.24	-3.93	TRUE
BRAA09G045310.3C	3.38	-17.91	-7.69	TRUE

Table 3.1: A table showing the results of the registration process performed using the *greatR* package.

From the sample data above, we can see that for nine out of ten genes, `registered = TRUE`, meaning that reference and query data between those nine genes can be aligned or registered. These data frame outputs can further be summarised and visualised; see the documentation on the processing registration results article.

3.2.6 Process results

After running the registration function `register()` as shown in Section 3.2.5, users can summarise and visualise the results as illustrated in Figure 3.1.

Summarising registration results

The total number of registered and non-registered genes can be obtained by running the function `summary()` with `registration_results` object as an input.

The function `summary()` returns a list with S3 class `summary.res_greatR` containing four different objects:

Return objects	Description
`summary`	A data frame containing the summary of the registration results (default S3 print).
`registered_genes`	A vector of gene IDs which are successfully registered.
`non_registered_genes`	A vector of non-registered gene IDs.
`reg_params`	A data frame containing the distribution of registration parameters.

The function `plot()` allows users to plot the bivariate distribution of the registration parameters. Non-registered genes can be ignored by selecting `type = "registered"` instead of the default `type = "all"`. Similarly, the marginal distribution type can be changed from `type_dist = "histogram"` (default) to `type_dist = "density"`.

Plotting registration results

The function `plot()` allows users to plot the registration results of the genes of interest (by default only up to the first 25 genes are shown, for more control over this, use the `genes_list` argument). Notice that the plot includes a label indicating if the particular genes are registered or non-registered, as well as the registration parameters in case the registration is successful. For more details on the other function arguments, go to `plot()`.

Analysing similarity of expression profiles over time before and after registering

After registering the data, users can compare the overall similarity between datasets before and after registering using the function `calculate_distance()`. By default all genes are considered in this calculation, this can be changed by using the `genes_list` argument.

The function `calculate_distance()` returns a list with S3 class `dist_greatR` of two data frames:

Return objects	Description
`result`	distance between scaled reference and query expressions using time points after registration.
`original`	distance between scaled reference and query expressions using original time points before registration.

Each of these data frames above can be visualised using the `plot()` function, by selecting either `type = "result"` (default) or `type = "original"`.

Example of processing registration results on *B. rapa* and Arabidopsis data

Code 3.5: Getting summary of the registration results.

```
# Get registration summary
reg_summary <- summary(registration_results)

reg_summary$summary
```

Result	Value
Total genes	10
Registered genes	9
Non-registered genes	1
Stretch	[2.25, 4]
Shift	[-27.24, -4.36]

Table 3.2: A summary table obtained from function `summary()` in *greatR*.

Code 3.6: Getting the stretch and shift distribution by plotting the registration summary.

```
plot(reg_summary, type = "registered")
```

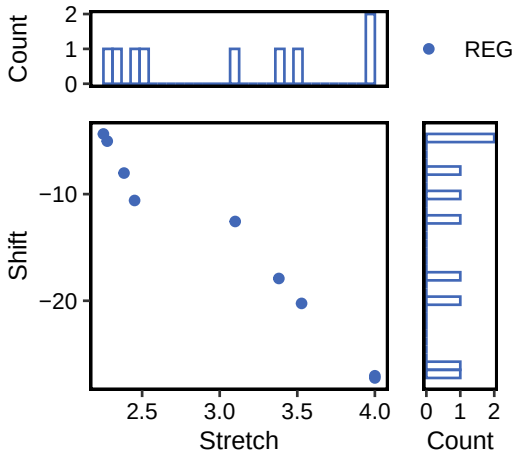


Figure 3.4: Plot of the distribution of stretch and shift parameters.

We also provide a function for users to visualise their registration result for each gene. This can be obtained by executing the function `plot()` on the registration results.

Code 3.7: Getting the plot of registration results for specific gene ID(s).

```
plot(registration_results, genes_list = c("BRAA02G018970.3C",
    "BRAA02G043220.3C"))
```

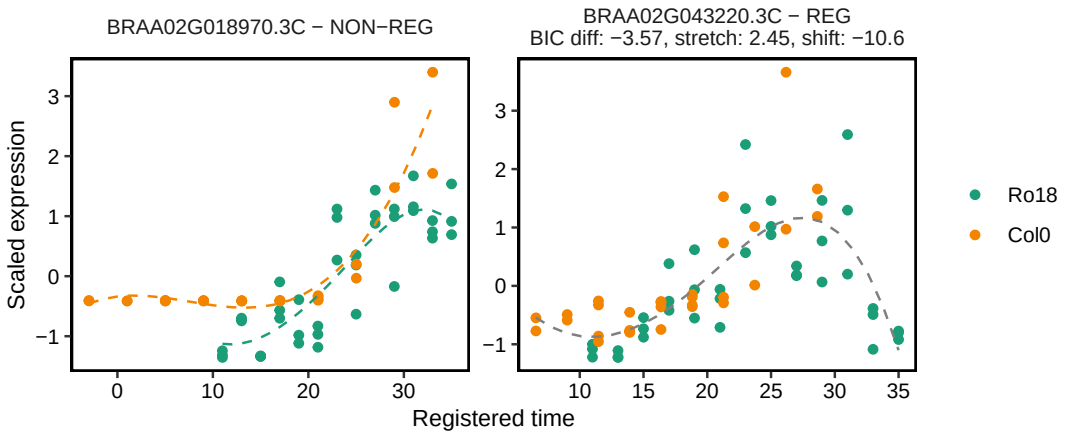


Figure 3.5: Plot showing the registration results of two *B. rapa* genes. Green and orange dots represent *B. rapa* and Arabidopsis expression data, respectively. On the left, the expression time series for each species is fitted to separate models, indicating they are not registered. On the right, the data is registered and fitted to a joint model, shown by the grey dashed line. The plot titles indicate whether the data is registered, and, if so, the optimal stretch and shift parameters are displayed in the title. Time points are shown in days after germination.

Code 3.8: Getting the distance heatmap between samples after registration.

```
# Calculate sample distance
sample_distance <- calculate_distance(registration_results)

# Plot distance heatmap after registration process
plot(sample_distance, type = "result", match_timepoints = TRUE)
```

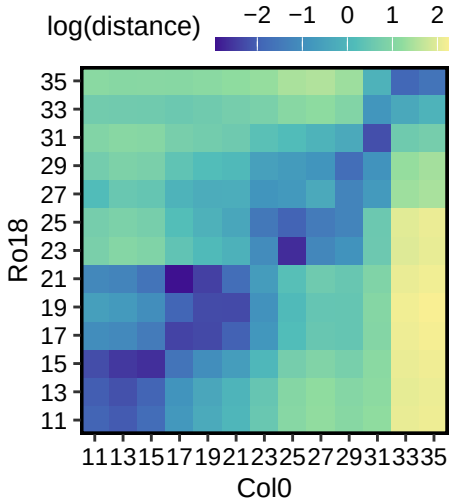


Figure 3.6: Heatmap of mean expression profile distances after the registration process. When `match_timepoints = TRUE` is applied, the heatmap displays the matched query time points to the corresponding reference time points. Time points are given in days after germination.

3.3 Methods tested on the simulated data

3.3.1 Introduction

To validate and evaluate the performance of our method, it was crucial to establish a benchmark or "ground truth" against which we could measure its accuracy. Specifically, we needed to determine whether two time series were indeed the same or different and to have precise information on the shift and stretch values applied during data generation. This baseline allowed us to objectively assess the effectiveness of our approach.

To achieve this, we generated both positive and negative control datasets which will be discussed in the following section (see Methods 3.3.2). The positive control datasets were designed to test whether the method could accurately identify and align expression profiles that originated from the same underlying model with identical parameter values. A successful outcome in this scenario would indicate that our method can recognise matching patterns, even when minor noise or variations are present.

Conversely, the negative control datasets were created to assess the method's ability to distinguish between expression profiles that were intentionally derived from different models, each with distinct parameter values. Here, the goal was to ensure that the method does not mistakenly identify these different profiles as being the same, but rather correctly identifies the divergence between them.

By rigorously testing on both types of control datasets, we evaluated our method's accuracy in aligning similar profiles and its sensitivity in distinguishing distinct ones. This comprehensive evaluation not only confirms the robustness and reliability of our approach across various scenarios but also highlights potential limitations. Recognising these limitations is essential for refining the method and guiding future improvements. In the following sections, we will detail the methods used to generate the simulated datasets, present the registration results obtained from these simulations, and discuss any limitations identified through this testing process.

3.3.2 The generation of simulated data

Positive control data

We generated simulated data by randomly sampling coefficients for a cubic B-spline, as described in the Methods section (Section 2.1.5). The spline coefficients were drawn from a uniform distribution $U(-10, 10)$. For each spline, we selected n time points, $i = 1, \ldots, n$ and calculated the corresponding function values. These values were then rescaled to be positive, ensuring they fell within the range of 0.0 to 10.0.

The minimum and maximum time points for the datasets were randomly sampled to introduce variability in the time ranges used for generating the reference profiles. Specifically, the minimum

time point t_{min} was sampled from a uniform distribution $U(a, b)$, where a and b represent the lower and upper bounds of the range, respectively. Similarly, the maximum time point $tmax$ was sampled from a uniform distribution $U(c, d)$, with c and d representing the corresponding bounds. Note here that $a < b < c < d$, ensures that the maximum time point always follows the minimum time point. This random sampling process ensures that the time intervals over which the data are observed can vary between different simulations, thereby mimicking the natural variability that can occur in real experimental settings. By allowing these time points to vary, we can assess the robustness of the method when applied to datasets with different temporal resolutions.

To generate the query data for the positive control, we used the same set of spline coefficients that were used to generate the reference dataset. The query data were sampled at time points $t_j = 1, \ldots, 10$, ensuring that the underlying dynamics were identical to those in the reference dataset.

The query data were subjected to transformation using time shift and stretch operations. The shift parameter, β_1, was sampled from a uniform distribution $\sim U(0, 5)$, while the stretch factor, β_2, was sampled from a uniform distribution $U(1, 5)$. Mathematically, this transformation can be formulated as follows

$$\mathbf{t_q} = \frac{\mathbf{t_r} - \beta_1}{\beta_2}, \tag{3.1}$$

where $\mathbf{t_q}$ are the transformed time points which are the query time points, and $\mathbf{t_r}$ are the reference time points. The β_1 and β_2 are randomly sampled shift and stretch factors, respectively.

Negative control data

For the negative control, we generated query datasets by vertically reflecting the reference datasets around the horizontal axis. This reflection effectively inverts the original curves, creating dynamics that are fundamentally different from those of the reference data. Mathematically, it can be formulated as follows

$$\mathbf{y_q} = y_{r,max} - \mathbf{y_r}, \tag{3.2}$$

where $\mathbf{y_q}$ are the flipped values (query values), $y_{r,max}$ is the maximum value of the reference data, and $\mathbf{y_r}$ are the reference values.

Following this reflection, we applied the same stretch and shift transformations as were used for the positive control profiles. The stretch factor, β_2, and shift parameter, β_1, were sampled from the same uniform distributions, $U(1, 5)$ and $\beta_1 \sim U(0, 5)$, respectively, to produce the final negative query profiles.

Sampling query time points differently from reference data

To test our method under the conditions where the query and reference data originate from the same underlying model but are sampled at different time points, we generated another set of data specifically for this scenario (see Figure 3.7). The reference dataset was generated using the same methodology as before; however, instead of sampling the query data from the exact same time points as the reference, we sampled from different time points. This setup was designed to challenge the method's ability to align data that, while derived from the same underlying model, are most likely observed at different temporal intervals. This approach allowed us to rigorously test the hypothesis H_1 by examining whether the method could still identify the underlying common model $m_1(\theta_1, t)$ despite the differences in time point sampling. The results from this experiment provide valuable insight into the robustness of our method and its ability to correctly infer a common model even when the data are not perfectly aligned in time.

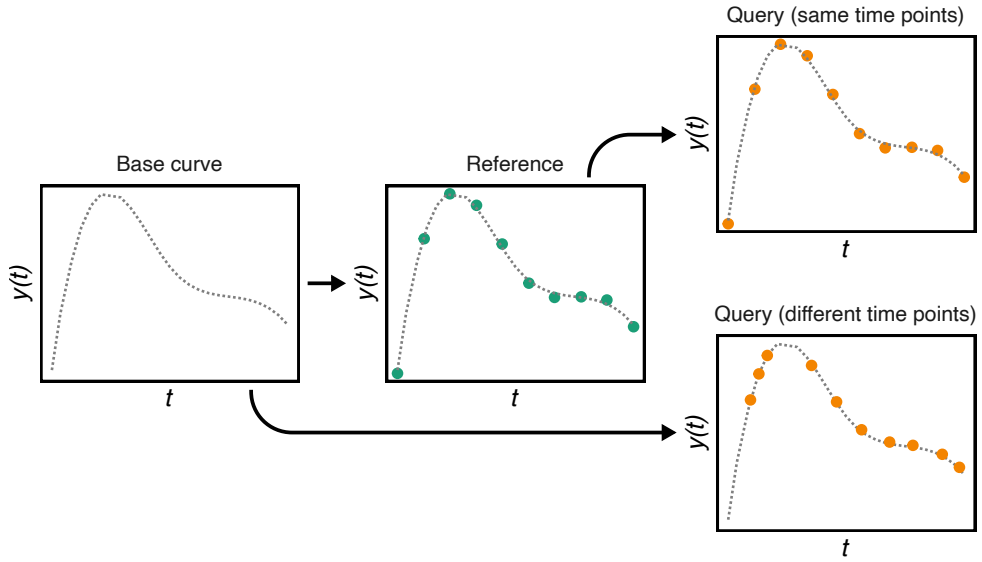


Figure 3.7: Schematic diagram illustrating various scenarios of how query data time points are sampled from the underlying model. These scenarios include query data sampled both at the same and at different time points as the reference data. Green dots represent reference data, while orange dots denote query data.

Time point variability to assess the impact of different temporal resolutions

To assess the method's performance across different temporal resolutions, we constructed additional sets of simulated data for both positive and negative controls. These datasets were designed to include profiles with varying numbers of time points, covering both scenarios with fewer than ten points and those with more. Specifically, we generated additional profiles with five, six, seven, eight, nine, and twenty time points. This approach allows us to evaluate how the method performs under both sparse data conditions, where time points are limited, and dense data conditions, where time points are abundant.

For the scenario where reference and query were sampled from different time points, we generated 1000 reference profiles. Each reference profile was paired with corresponding query data for both positive and negative controls. For the positive dataset, we generated in a total of 14,000 pairs of simulated curves (7-time point configurations $\times$ 2000 profiles for r and q). By evaluating the method across these different time point scenarios, we can rigorously test its robustness, accuracy, and generalisability. This test will evaluate how reliable the method performance is when data are abundant but also when dealing with fewer data points, which is often a common challenge in real experimental conditions.

Noise addition to evaluate the method robustness against experimental variability

To introduce variability and simulate real experimental conditions, we added the same level of noise to both the reference and query data. This noise represents random fluctuations in gene expression that could arise due to biological variability or measurement errors. The noise levels were carefully controlled, ranging from 0% (no added noise) to 200% (high noise), allowing us to assess how the method performs under different levels of data perturbation, which is a common challenge in gene expression studies.

The noise values were sampled from a uniform distribution $U(0, 20)$. This means that for each time point in the query dataset, a random noise value was drawn from this distribution and added to the corresponding gene expression value. The range of the uniform distribution ensures that the noise is evenly spread between 0 and 20, covering a broad spectrum of possible deviations. The upper limit of 20 was chosen to reflect extreme cases where noise could significantly distort the expression data, challenging the method's ability to accurately align and compare the reference and query profiles. In addition to the original 14,000 noise-free curves, we generated 266,000 curves (19 noise-level configurations $\times$ 14,000 profiles) for the positive control dataset and 154,000 curves (11 noise-level configurations ranging from 1 to 10, plus 15 and 20 $\times$ 14,000 profiles) for the negative control dataset.

By systematically varying the noise levels, we could evaluate the robustness of the method under increasingly noisy conditions. This approach helps to determine the threshold at which noise begins to impact the accuracy of the method, providing valuable insights into its reliability in less controlled or more variable experimental settings.

Code availability

The complete code used to generate these control datasets, including all the different settings and configurations for simulating various scenarios, is available on GitHub. This repository contains detailed scripts that allow readers to reproduce the simulated data described in this study, as well as to customise and extend the simulations to suit their own research needs. This can be accessed at the following link: `https://github.com/ruthkr/greatR-manuscript/blob/main/analysis/simulate_control_datasets.R`.

3.3.3 Curve registration with *greatR* can align pairs of time series with similar dynamics (positive control datasets)

We first evaluated our method on the positive control datasets, where both the reference and query data were sampled from the same models and at the exact same time points. This dataset serves as a baseline to verify that our method's formulation and implementation function as intended. As expected, all pairs of curves in this dataset were successfully aligned, with accurately estimated shift and stretch parameters (see Figure 3.8). Figure 3.9 provides an example of a reference and query pair, sampled from identical time points, that were correctly identified as similar by our method.

After demonstrating that our method performs well on this initial positive control dataset, we further evaluated its robustness in a more challenging scenario where the pairs were not sampled at the same time points along the base curve. If the method is implemented correctly, it should identify these pairs as having the same underlying dynamics, regardless of the specific time points at which they were sampled. When evaluated on this positive control dataset, *greatR* successfully registered all 1000 pairs (for each set of data with different time points) with no false negatives in the noise-free scenario (Figure 3.10). This result indicates that the method can reliably align expression profiles that originate from the same underlying model, especially when no external variability is introduced.

Figure 3.10 illustrates how the addition of noise to both datasets can impact the registration results. As the noise level increases–while keeping the σ values fixed in the likelihood (i.e., not accounting for the added noise)–the success rate of registration gradually declines, dropping to almost all non-registered for a noise level of 200% for the data set with five-time points (see Figure 3.10 (a)). This decline is expected, as higher noise levels introduce greater dissimilarity in the dynamics between the datasets, making it more likely for the method to favour different models (hypothesis H_2) over a single model (hypothesis H_1).

Without adjusting for the noise in the likelihood, these elevated noise levels can significantly reduce *greatR*'s ability to accurately align and register the data. However, when the method is adapted to account for the noise by adjusting the σ values, it successfully registers all pairs of curves, even under high noise conditions (see Figure 3.10 (b)). This adjustment highlights the importance of considering noise levels in the likelihood estimation to maintain the accuracy and reliability of the registration process.

As the number of time points in a dataset increases, there is a noticeable improvement in the percentage of successfully registered data, even under a high noise level of 200% (see Figure 3.10 (a)) This improvement is largely due to the better fitting capability of B-splines and the more reliable evaluation provided by the BIC. B-splines, which are used in *greatR*, perform better with a greater number of time points because they can more accurately model the complex dynamics of gene expression profiles. With more data points, the spline has more flexibility to capture the underlying model, leading to more precise alignment during registration. Additionally, the BIC, which balances model fit with complexity, benefits from the increased data density. More time

points provide a more accurate estimate of the model likelihood, and the relative impact of the penalty term for model complexity is reduced, making BIC more effective at distinguishing the correct model.

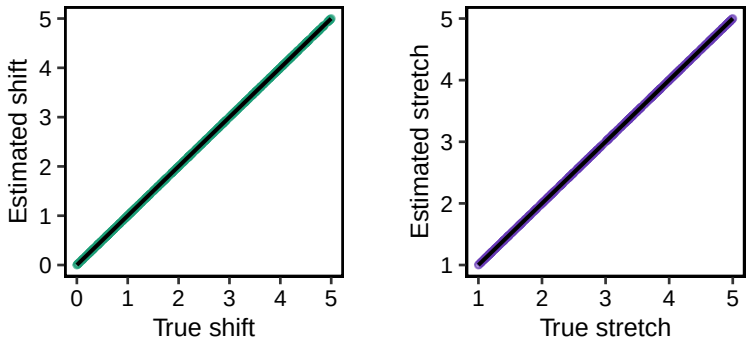


Figure 3.8: Proportion of estimated parameter (a) shift and (b) stretch from the registration results of simulated data (when query and reference data were sampled from the same points) based on cubic B-splines with one knot via *greatR* versus the simulated (true) value of the parameter. Each data point represents an estimation of the parameter for each gene.

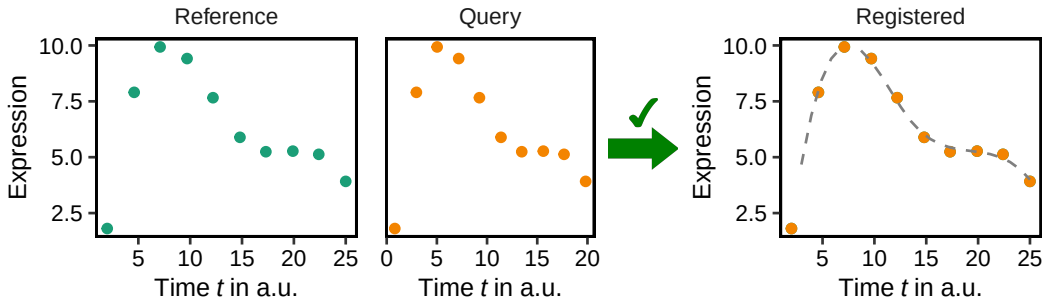


Figure 3.9: A sample from the positive control dataset (with no noise, query and reference were sampled from the same time points). The left-hand side shows the original dynamics of both reference and query curves, and the right-hand side shows the registered curves. Green and orange indicate reference and query data, respectively.

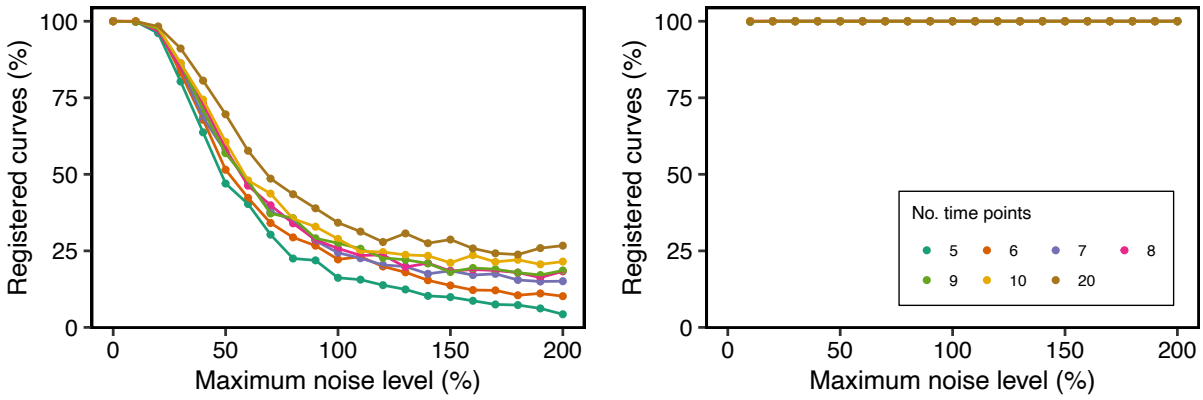


Figure 3.10: Total percentage of registered curves on the positive simulated data with different levels of noises and number of time points without (a) and with adjusted σ (b).

Figure 3.11 and 3.12 provide illustrative examples of the reference and query datasets with ten time points, showcasing the effects of noise on the registration process. In Figure 3.11, the datasets, sampled from different time points but generated from the same model with identical parameters, are shown before and after registration. On the left, you can see the initial misalignment due to the temporal sampling differences, while the right side demonstrates how *greatR* effectively aligns the datasets, in the absence of noise. This successful registration underscores the method's robustness in handling data that share the same underlying dynamics, despite being sampled at different intervals. Conversely, Figure 3.12 highlights the challenges introduced by noise. Here, the query data was subjected to a noise level of 60%, which significantly disrupted the alignment process, leading to a failed registration. This comparison between the noise-free and noisy scenarios emphasises the impact of noise on the registration success and illustrates the method's limitations under more challenging conditions.

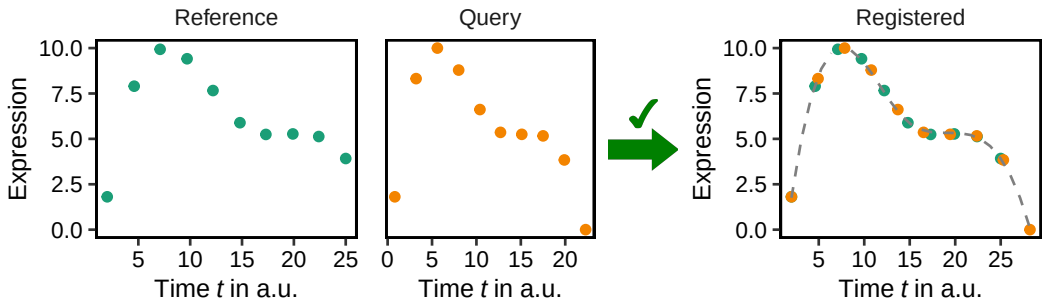


Figure 3.11: A sample from the positive control dataset (with no noise, query and reference were sampled from different time points). The left-hand side shows the original dynamics of both reference and query curves, and the right-hand side shows the registered curves. Green and orange indicate reference and query data, respectively.

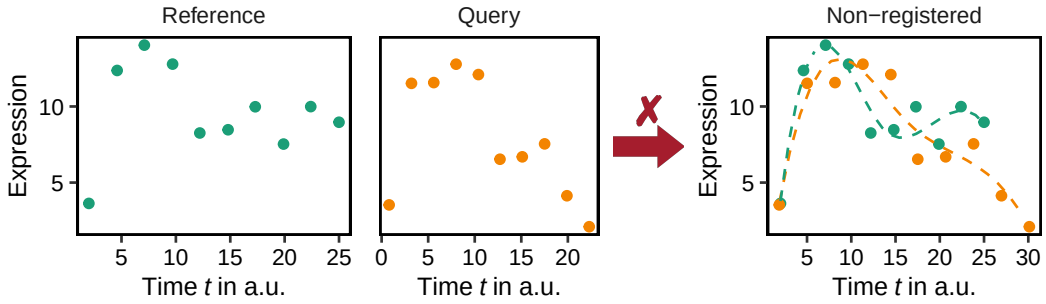
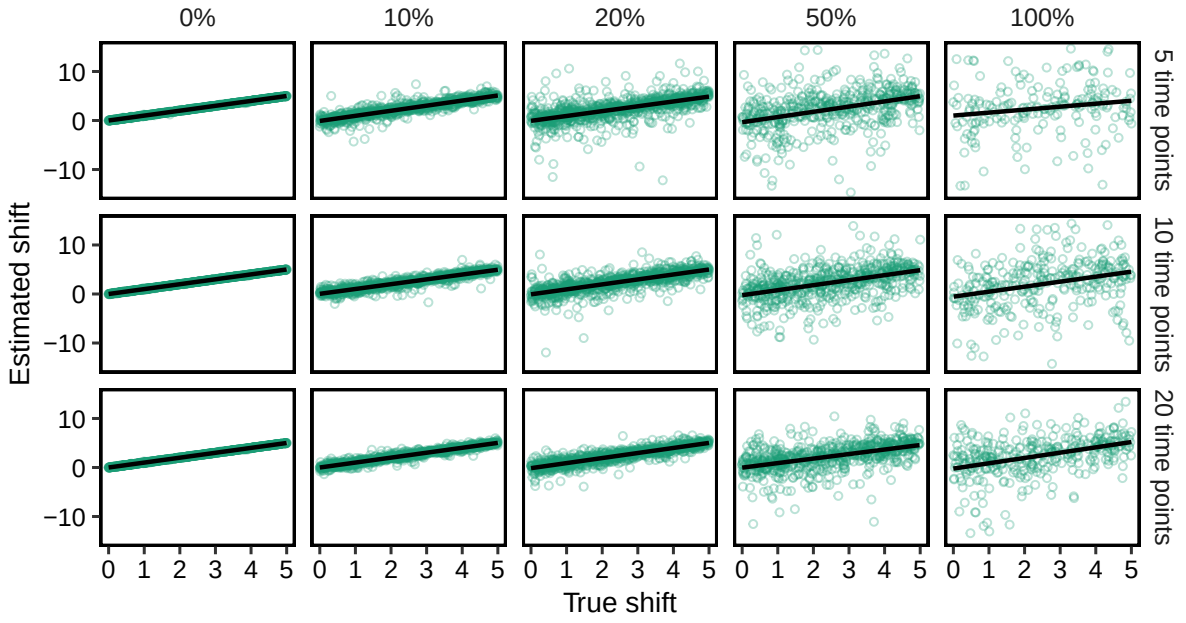


Figure 3.12: A sample from the positive control dataset (noise level is equal to 60%) identified as non-registered curves. The left-hand side shows the original dynamics of both reference and query curves. The right-hand side is the registration results of the corresponding data. Green and orange indicate reference and query data, respectively.

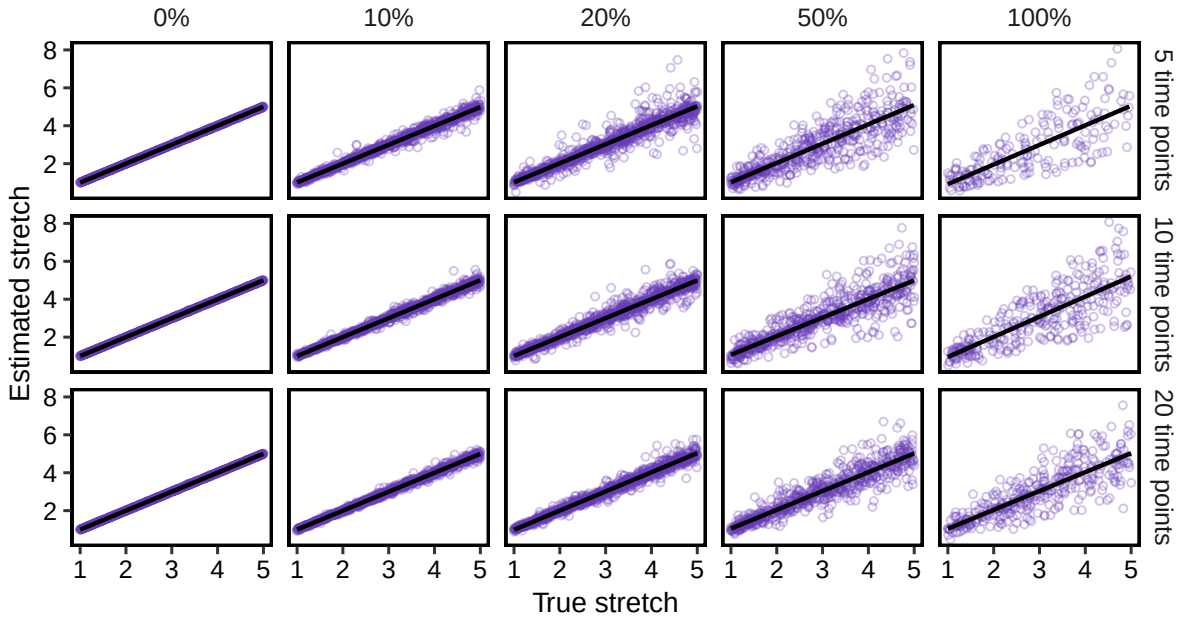
While many pairs were successfully registered even at high noise levels, the estimated stretch and shift parameters deviated from their true values. This deviation becomes particularly evident in Figures 3.13 (a) and 3.13 (b)), which compare the estimated shift and stretch parameters to the true values. Under noise-free conditions, *greatR* performs well, accurately inferring the shift and stretch parameters, closely matching the true underlying dynamics of the data. However, as the noise level increases, the method's ability to accurately estimate these parameters declines. Noise introduces random variability into the data, making it more challenging for the method to identify the true underlying model. As a result, the estimated parameters exhibit greater variability, often deviating from the true values. This is a critical observation, as it indicates that while *greatR* can still align the profiles under noisy conditions, the precision of the alignment (in terms of the exact stretch and shift values) declines.

To further evaluate the consistency and robustness of our method, we conducted an additional experiment using the positive control dataset, where we swapped the roles of the reference and query data. In this setup, the original reference data was treated as the query input, and the original query data was treated as the reference input. The goal was to determine whether our method could consistently register the pairs despite the reversal of input roles. As expected, the method successfully registered all pairs, achieving a 100% registration rate with no false negatives in the noise-free scenario across varying numbers of time points (see Figure 3.14). This outcome confirms the robustness of the method, demonstrating its ability to reliably identify matching curves regardless of how the reference and query inputs are configured. Additionally, the results underscore the method's internal consistency, as it produced correct registrations even when the roles of the datasets were reversed.

We then extended the experiment by introducing varying levels of noise and evaluated both the percentage of successfully registered curves and the accuracy of the estimated stretch and shift parameters under these conditions. Similar registration rates were observed across different noise levels and time points, with results consistent regardless of whether the reference and query data were swapped (see Figure 3.10 and 3.14). Figure 3.15 (a) and (b) demonstrate how the



(a)



(b)

Figure 3.13: Proportion of estimated parameter (a) shift and (b) stretch from the registration results of simulated data based on cubic B-splines with one knot via *greatR* versus the simulated (true) value of the parameter for each different level of σ. Each data point represents an estimation of the parameter for each gene.

accuracy of the estimated shift and stretch parameters are affected by noise. In the noise-free scenario, *greatR* accurately estimates the stretch and shift values, closely matching the true parameters. However, as noise levels increase, the variability in these estimates also increases, leading to a wider distribution around the true values. Figure 3.14 reveals that while many pairs remain successfully registered even at higher noise levels, the precision of the estimated stretch and shift parameters declines (Figure 3.15). Specifically, at 100% noise, the estimates show significant deviation from the true parameters, particularly in cases with fewer time points (Figure 3.15). This effect is illustrated in the scatter plots, where the relationship between true and estimated values becomes more dispersed as noise increases. Note here, the swapping process resulted in different estimated stretch and shift values. Specifically, the stretches were calculated as $1/\beta_1$, and the shifts as β_1/β_2, reflecting the inverse relationship dictated by the curve registration function. These findings are consistent with the theoretical expectation that the transformation function is invertible, meaning that when the roles of the datasets are reversed, the transformation parameters are also inverted. This observation further validates the accuracy and theoretical consistency of our approach.

Overall, these findings confirm that while the method is robust in aligning profiles even under noisy conditions, the accuracy of the estimated parameters can be compromised, particularly with higher noise levels and fewer time points. Nonetheless, the observed trends align with theoretical expectations, further validating the method's reliability.

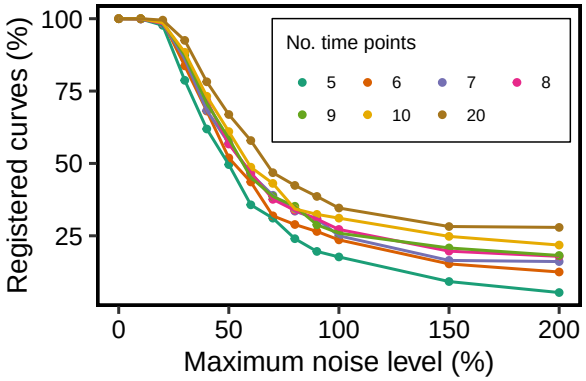
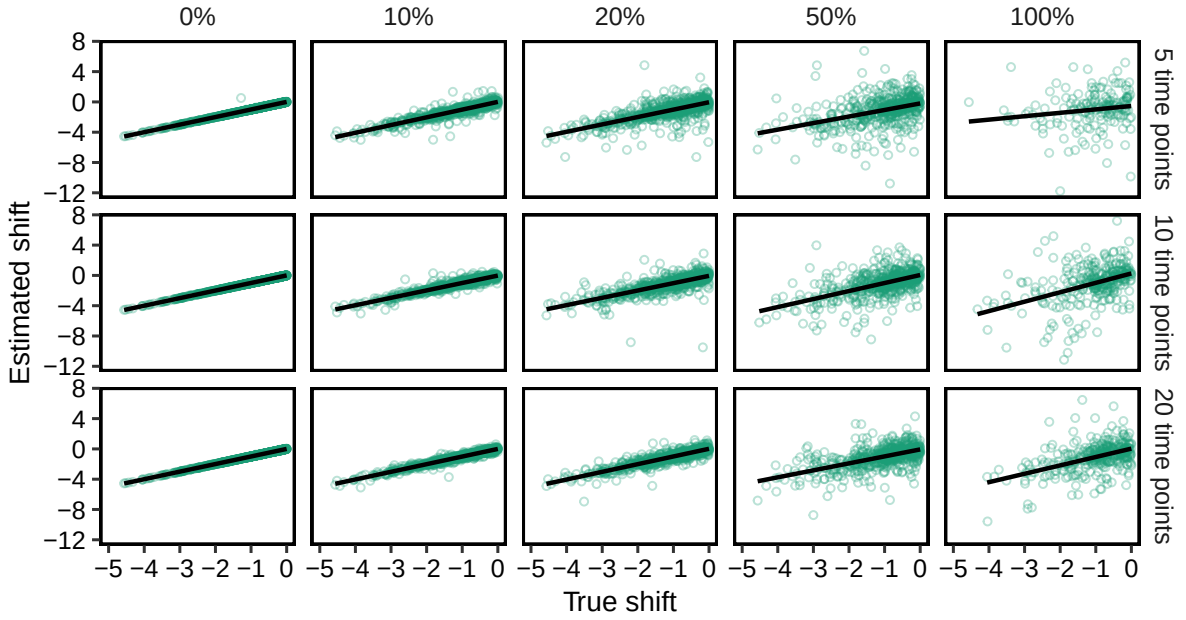
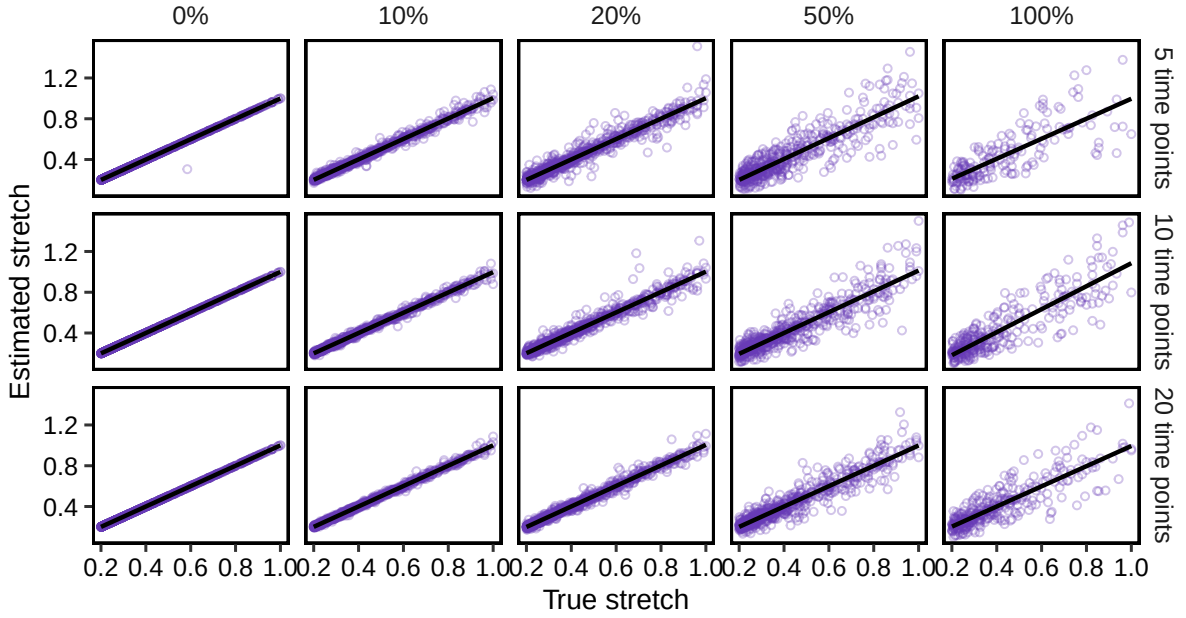


Figure 3.14: Total percentage of registered curves on the positive simulated data with different levels of noises and number of time points. The registration was performed by swapping the reference and query data.



(a)



(b)

Figure 3.15: Proportion of estimated parameters (a) shift and (b) stretch from the registration results after swapping the reference and query data. The registration was performed on the simulated data using *greatR*, comparing the estimated parameters to the true (simulated) values across different noise levels (σ). Each data point represents the estimated parameter value for an individual gene, with noise levels increasing from 0% to 100% across columns, and the number of time points varying across rows (5, 10, and 20 time points). The alignment of data points along the diagonal line indicates accurate estimation of the parameters, while deviations reflect the impact of noise and the number of time points on the accuracy of the registration process.

3.3.4 Curve registration with *greatR* can identify pairs of time series with different dynamics using negative control datasets

For the datasets without added noise, *greatR* successfully identified over 98.6% of the negative control datasets as having different dynamics, correctly distinguishing between pairs of time series that originated from distinct models (Table 3.3). The remaining 1.4% of datasets, which were incorrectly registered as similar despite originating from different models, can be attributed to *greatR*'s ability to detect local similarities between the curves. Some examples of these for curves with ten time points are shown in Figure 3.16, where certain pairs exhibit localised matching patterns that lead to their registration, even though their overall dynamics differ. Table 3.3 shows that the number of non-registered curves increases as the number of time points increases. The higher number of time points allows our method to capture more differences between the time series, thus reducing the likelihood of false registrations. This is likely due to the fact that with more time points, the model has additional information to distinguish between slight variations in dynamics that might not be as apparent in shorter time series. Consequently, the ability to detect local similarities becomes less prominent, and the global differences between curves are more clearly recognised, improving the overall accuracy of classification.

When we implemented a stricter criterion requiring that the transformed datasets must overlap completely in time for a successful registration (see Table 3.3 and Figure 3.16), our method demonstrated a better ability to distinguish time series with different dynamics. Under this criterion, *greatR* correctly identified all pairs with different dynamics, eliminating the false positives which were previously observed. This result is consistent for all different numbers of time points in the datasets.

By enforcing full temporal overlap, *greatR* avoided misidentifying local similarities as global alignments, a challenge seen in the initial registration setting. This stricter requirement ensured that only time series with matching global structures were registered as similar, rather than those that shared localised patterns. This adjustment in the registration process significantly reduced the risk of false positives, especially in cases where sparse data might lead to misleading conclusions under more relaxed criteria. This finding underscores the importance of prioritising global alignment when working with time series that exhibit distinct overall dynamics. In particular, it suggests that when there are few data points, ensuring full temporal overlap is crucial for accurate registration. By focusing on global patterns, *greatR* becomes more robust in differentiating between datasets, regardless of the number of time points, which is especially important for applications where local alignment could hide critical differences between curves.

Figure 3.17 provides an example of the original dynamics and the registration results for a pair of curves from the negative control dataset. In this case, the curves exhibit distinct dynamics and were correctly identified as non-registered by *greatR*. This demonstrates the method's effectiveness in recognising and separating time series with differing underlying processes.

No. time points	50% Overlapping		100% Overlapping	
	Non-registered	Registered	Non-registered	Registered
5	986	14	1000	0
6	990	10	1000	0
7	994	6	1000	0
8	992	8	1000	0
9	996	4	1000	0
10	996	4	1000	0
20	999	1	1000	0

Table 3.3: Registration results of negative control datasets with only 50% and full temporal overlap across varying time points.

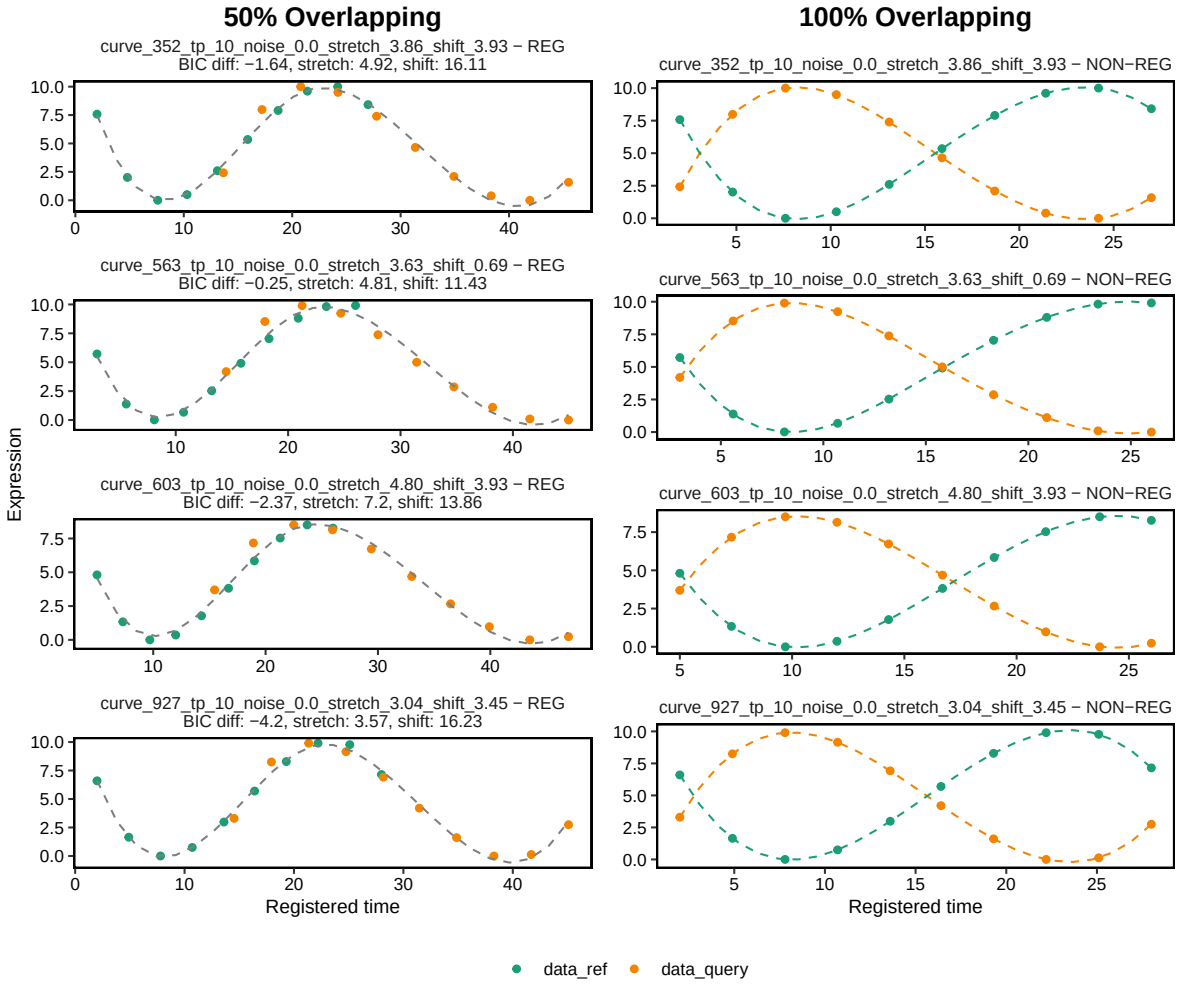


Figure 3.16: Samples from the negative control dataset where curves were initially registered due to local similarities under the 50% temporal overlap criterion. When full temporal overlap was applied, these curves were correctly identified as having different dynamics. Green and orange represent the reference and query data, respectively.

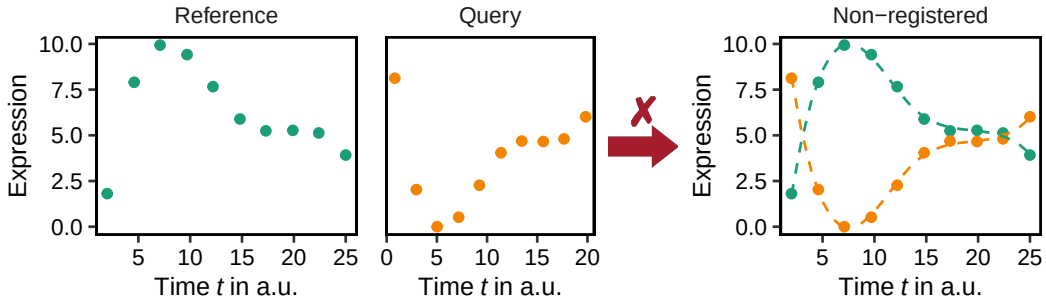


Figure 3.17: A sample from the negative control dataset identified as non-registered curves. The left-hand side shows the original dynamics of both reference and query curves. The right-hand side is the registration results of the corresponding data. Green and orange indicate reference and query data, respectively.

We also examined the impact of adding noise to the negative control datasets. Figure 3.18 shows that the number of non-registered curves decreases as the level of noise increases when full temporal overlap is not enforced. This suggests that as noise increases, *greatR* identifies more curves as having similar dynamics. However, when full temporal overlap was applied, *greatR* consistently identified the curves as having different dynamics across varying numbers of time points, even at high noise levels (up to 100%). At higher noise levels, particularly at 150% and 200%, a small percentage of the curves were incorrectly identified as similar. This can be explained by the fact that as noise levels rise, the added random fluctuations can increasingly dominate the original signal. When the noise perturbs both curves in similar ways, it can mask the underlying differences in their true dynamics. Essentially, the noise introduces enough variability that the original dynamics are harder to distinguish, and the random, noisy fluctuations start to resemble one another. This artificial similarity caused by noise results in *greatR* misclassifying these curves as having similar dynamics, even though they originate from different processes. In extreme cases, the noise distorts the curves to the point where the original differences are no longer distinguishable, leading to false registrations.

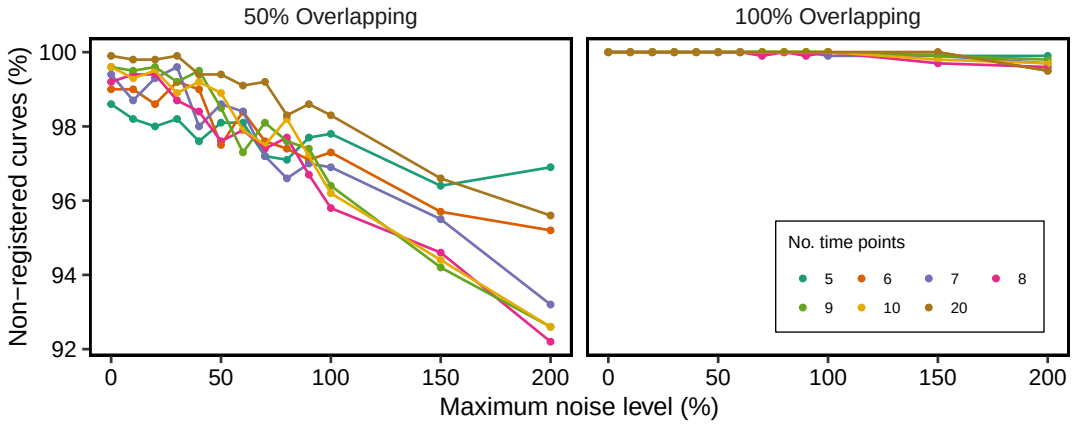


Figure 3.18: Total percentage of non-registered curves on the negative simulated data with different levels of noises and number of time points without and with full temporal overlap.

3.4 Conclusion and future directions

Using both the positive and negative control datasets, we thoroughly evaluated the performance of our method in identifying similar time series and accurately inferring their corresponding shift and stretch factors. The results demonstrate that our method is highly effective in registering pairs of datasets generated from a model with identical parameters, as evidenced by the positive control experiments. Even when noise was introduced to these datasets, the method is still able to identify pairs which are similar, highlighting its robustness in handling real-world data imperfections. However, the degree of this robustness depended on the noise level. As the noise increased, while the pairs were still identified as registered, the estimated stretch and shift parameters began to deviate from their true values. This deviation is expected, as the introduction of noise makes the data less similar to the original, unperturbed profiles, meaning that the registration process adapts to these changes, leading to slight discrepancies in the estimated parameters.

In the case of the negative control datasets, our method also successfully identified pairs of curves generated from different models as non-registered. This ability to differentiate between distinct dynamics is crucial for applications where distinguishing between different biological processes or conditions is necessary. However, the performance was somewhat dependent on the number of time points and noise levels. With fewer time points or higher noise levels, the method's ability to accurately estimate stretch and shift parameters decreased, although it still maintained a reasonable level of accuracy. This suggests that while our method is robust, it benefits from higher data density and lower noise levels, which are common considerations in time series analysis.

Future works

While *greatR* has proven effective in its current form, several potential future developments could improve its usability and expand its functionality. One possible next step is to port the package to other programming languages, such as Python. Python is widely used in the data science and bioinformatics communities, and offering a Python version of *greatR* could make the tool more accessible to a broader audience. This would involve translating the core algorithms and ensuring that the performance remains consistent with the R version. Another useful addition could be the development of a user-friendly interface, such as an R Shiny app. An R Shiny interface would allow users who are less familiar with coding to interact with *greatR* through a graphical user interface. This could include options for uploading datasets, specifying parameters, running analyses, and visualising results in an intuitive and accessible way. Such an interface would make *greatR* more accessible to a wider range of users, including those in clinical and biological research who may not have enough programming experience.

Beyond these usability improvements, there is also potential for improving the underlying models used within the method. Currently, the package uses B-splines to model the dynamics of

time series data, which is effective for capturing a wide range of expression patterns. However, for specific biological processes, such as cell cycle regulation or circadian rhythms, other models like sinusoidal functions could be more appropriate. Sinusoidal models naturally represent periodic dynamics, making them ideal for registering time series data related to cyclic biological processes. Adding such models to our method could increase its applicability in these contexts.

Moreover, given the impact of the number of time points on the method's performance, our method could be further improved by adjusting the complexity of the model based on the data. For datasets with fewer time points (e.g., five or fewer), using simpler models such as cubic polynomials could provide more reliable results. B-splines, while powerful, may require more data points to fit accurately without overfitting. Implementing a mechanism that automatically selects a simpler model when the data are sparse would make our method more versatile and reliable across a wider range of experimental conditions. An alternative, more robust approach could involve fitting each gene pair or curve to different models and selecting the best-suited one for each case. Additionally, machine learning techniques, such as Gaussian processes, could be introduced to replace spline-based methods. Gaussian processes provide greater flexibility, particularly in dealing with missing or sparse time points, as they directly model uncertainties in the data. They are non-parametric models that define a distribution over functions and can capture complex temporal patterns with built-in measures of confidence. Furthermore, implementing nested sampling as an alternative optimisation method could generate robust statistical evidence for selecting optimal registration parameters. Nested sampling is a Bayesian computational technique designed to efficiently explore complex parameter spaces and estimate the marginal likelihood. By computing the marginal likelihood (or Bayesian evidence), nested sampling offers a clear confidence metric for the chosen parameters, ensuring that the selection is based on both optimality and statistical reliability.

In summary, our method has demonstrated strong performance in registering time series data and distinguishing between similar and distinct dynamics. However, as with any computational tool, its performance can be influenced by the quality and quantity of the input data. Future development plans include creating a Python version, designing a user-friendly R Shiny interface, incorporating alternative models such as sinusoidal functions, and enabling adaptive model selection based on the number of time points. Additionally, using machine learning approaches like Gaussian Processes, and implementing nested sampling as an alternative optimisation method, could significantly improve the tool's flexibility, precision, and reliability. These improvements would make our tool a more versatile resource for the broader scientific community, capable of handling diverse time series data with varying characteristics.

4 | Understanding bract formation using comparative transcriptomics

The work presented in this chapter was conducted in collaboration with Sana Dieudonné and Fabrice Besnard from RDP, ENS Lyon. Sana and Fabrice were responsible for the experimental design, data generation, and initial biological interpretation that framed the research question. This includes the comparative investigation of bract formation in Arabidopsis accessions *Tsu-0* and *Col-0*, as described Dieudonné *et al.* [1], where they explored the evolutionary and developmental basis of bract loss, identified QTLs associated with bract formation, and provided key insights into the potential genetic mechanisms involved. They also performed the preprocessing of raw transcriptomic data, including read alignment and generation of count matrices. I conducted all subsequent transcriptomic analyses presented in result section (Section 4.3), including differential expression analysis, gene ontology enrichment, and the comparison of gene expression dynamics using registration. These analyses form the basis of the results and discussion presented in this chapter.

4.1 Bracts and their importance

Plant development is modular, allowing for varied shapes by altering a basic unit called the phytomer [162]. The cells of the plant shoot are produced by a group of stem cells called the shoot meristem, which creates leaf primordia on the edges and the stem tissues, including the vasculature and pith [163]. A new axillary meristem often forms at the junction of the leaf and stem and can grow to repeat this pattern. A typical phytomer includes a node, the internode below it, a leaf growing at the node, and an axillary bud (also known as a lateral bud) situated at the base of the leaf [163, 164] (Figure 4.1). Changes in the growth balance among phytomer components lead to morphological differences seen in life phase transitions and between species. A common change in phytomers during reproductive development is the suppression of leaf growth in the inflorescence, called bracts [163, 165]. The term bract can refer to a leaf associated with a flower or inflorescence, without distinguishing between leaves that are generally associated with an inflorescence and those that specifically support a flower [165, 166]. However, in this chapter, bract refers specifically to the leaves which support a single flower, meaning they are positioned at the junction where the floral peduncle meets the stem [1]. Bracts show a wide range of shapes, sizes, and other morphological features across different species [167] (Figure 4.2).

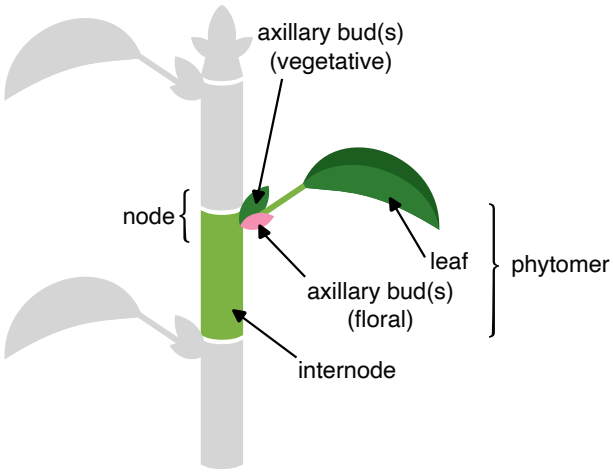


Figure 4.1: Schematic representation of a phytomer. Each phytomer is composed of (at least) an axillary meristem, subtended by a leaf, and an internode. The axillary meristem and the leaf together form the node. Modified from [168].

(a) *Rafflesia arnoldii* [169] **(b)** *Salvia pratensis* [170] **(c)** *Lilum martagon* [171] **(d)** *Euphorbia pulcherrima* [172]

(e) *Arum palaestinum* [173] **(f)** *Passiflora foetida* [174] **(g)** *Cynara cardunculus var. scolymus* [175] **(h)** *Castanea* [176]

Figure 4.2: Bracts have different shapes, sizes, and morphological features across different angiosperm species.

Although bracts can be considered specialised leaves with the capacity to perform photosynthesis [167], their photosynthetic capacity is often lower than that of regular leaves due to their lower mesophyll conductance [177]. In most plant species, showy and colourful bracts, such as bracts in *Rafflesia arnoldii*, *Salvia pratensis*, *Lilum martagon*, and *Euphorbia pulcherrima* (Figure 4.2 (a), (b), (c), and (d), respectively) cannot photosynthesise due to undeveloped chloroplasts [167]. It has been hypothesised that bracts enhance pollination by improving plants' visual

displays, such as *Araceae* (e.g. *Arum palaestinum* shown in Figure 4.2 (e)). However, studies have yielded inconsistent results, some researches show bracts increase pollinator visits and reproduction [178, 179], while another study finds no significant effect [180]. Some studies reported that bracts often protect reproductive organs during their development, such as from herbivores) and from harsh environmental conditions like low temperatures, intense sunlight, strong winds, heavy rain, drought, fire, and mechanical damage [181, 182, 167, 183]. For example, in *Passiflora foetida* (Passifloraceae) (Figure 4.2 (f)), the densely reticulate bracts that envelop the buds and fruits secrete a sticky substance capable of trapping various herbivorous insects [167, 183]. This secretion likely serves to protect the developing buds and fruits from herbivore damage and may also attract predators that contribute to the plant's defence [183]. In *Silybum marianum*, spiny bracts cover and defend the inflorescence during the development, flowering, and seed dispersal stages. In *Cynara cardunculus* var. *scolymus* (artichoke) (Figure 4.2 (g)), the spiny bracts shield the delicate inner part of the bud, which is the flower. Similarly, in *Castanea* species (chestnuts) (Figure 4.2 (h)), the bracts serve to protect the developing fruits [165].

4.1.1 Bract development and suppression differ among different species

Despite their functions, bract development seems to be abandoned by higher plants through evolution [184]. Bract suppression, observed in most angiosperm lineages, involves petals and sepals replacing bract functions. This mechanism was mainly studied in Arabidopsis, rice (*Oryza sativa*) and maize (*Zea mays*). In these plants, bract primordia were visible at a very early stage of apical meristem development but soon stopped developing and the bract primordium was eventually subsumed into the floral meristem [184]. Several mutants in Arabidopsis, such as *lfy* (*LEAFY*), *ap1* (*APETALA1*), *ufo* (*UNUSUAL FLORAL ORGANS*), and *fil* (*FILAMENTOUS FLOWER*), have been reported to develop bracts, highlighting the close relationship between flowering and bract development [184, 1]. A key gene called *NL1* (*NECK LEAF 1*) was identified to regulate bract suppression in rice [1]. The *tsh* mutant is also known to exhibit a similar bract phenotype, *TSH* (*TASSEL SHEATH*) is the homologue of *NL1* in maize. However, the *han* (*HANABA TARANU*) mutant, which is the homologue of *NL1* in Arabidopsis, does not exhibit bract development [185]. This suggests that the mechanisms of bract suppression differ among these species. However, the mechanisms that unlock bract development and suppression in these mutants remain unclear [1].

4.1.2 Arabidopsis natural accessions *Tsu-0* and *Col-0* shows the significant difference in basal bracts production

Although most flowering *Brassicaceae* species lack bracts, some still produce them at the base of the raceme. This suggests that bract loss in *Brassicaceae* is not complete [186]. There is no clear evolutionary pattern of bract presence across different *Brassicaceae* tribes [1].

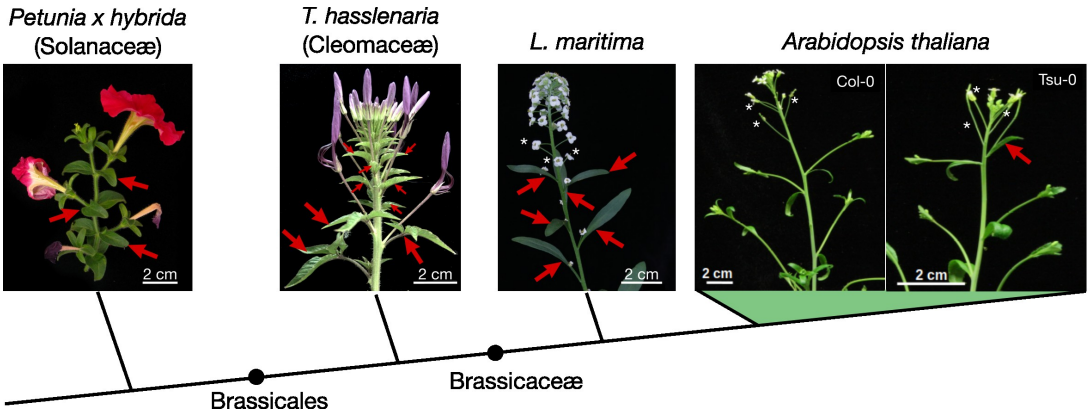


Figure 4.3: Illustrations of bracts (highlighted by red arrows) across different angiosperm species. Although members of the *Brassicaceae* family are generally bractless, some species, like *Lobularia maritima*, retain bracts at the base of inflorescence branches. In Arabidopsis, the presence of basal bracts varies among natural accessions, with some, such as *Tsu-0*, displaying them, while others, like *Col-0*, do not. Taken from [1].

Dieudonné *et al.* [1] highlighted that the natural *Tsu-0* accession often develops bracts on the first one to five flowers of the raceme, unlike the reference *Col-0*, which lacks bracts entirely (Figure 4.3). Through scanning electron microscopy, they found that bract development in *Tsu-0* plants can vary, particularly in the first flowers [1]. These bracts, appearing at the base of a flowering branch, were termed "basal bracts." Other genetically diverse natural accessions also produce basal bracts, though at highly variable rates, with no clear correlation to geographic or genetic origins (see Figure 4.4). They defined a basal bract score by summing all bracts on the main stem and cauline branches, normalised by the total number of branches (see Figure 4.5 (a) for the details). Using this scoring method, they recorded the differences in bract scores within genotypes, with *Col-0* being a low bract producer and *Tsu-0* a high bract producer. These two accessions were then chosen for further study because of their significant difference in bract production.

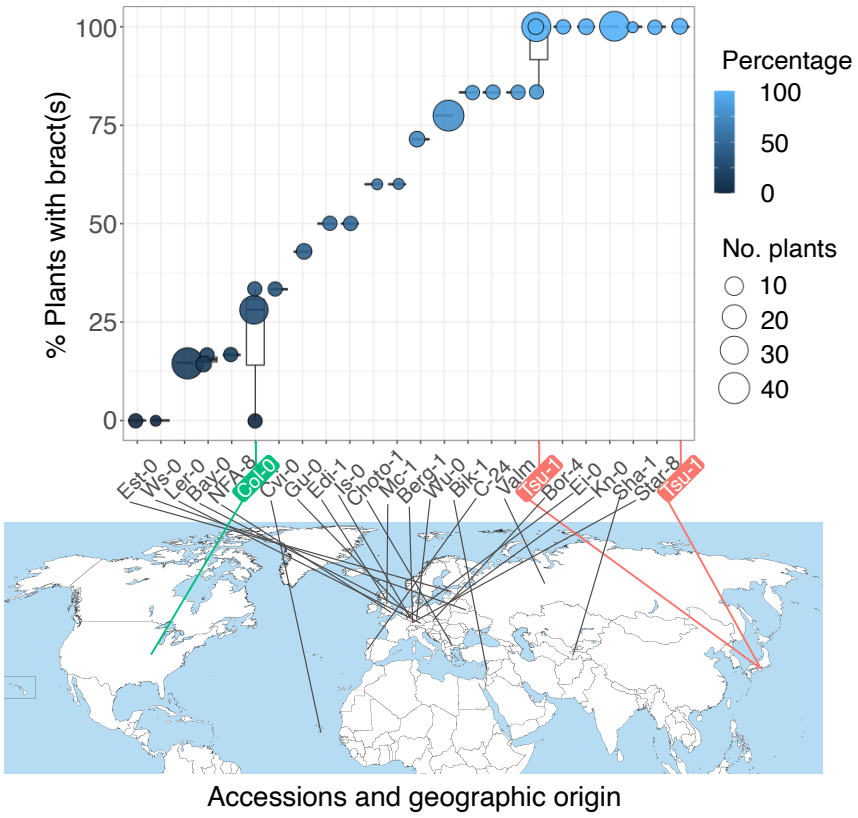


Figure 4.4: Frequency of basal bracts in various Arabidopsis accessions assessed by Dieudonné *et al.* [1]. The frequency is shown as the percentage of plants with at least one basal bract in the inflorescence. Each dot reflects the average value derived from multiple plants (with dot size indicating the number of plants) in a scoring assay. The box plot displays results from several assays per line, with thicker black horizontal lines representing the median value across all scoring assays (range 1-3) within each accession. *Col-0* and *Tsu-0* are highlighted in green and red, respectively. The geographical origins of each accession are indicated on the world map below the frequency plot.

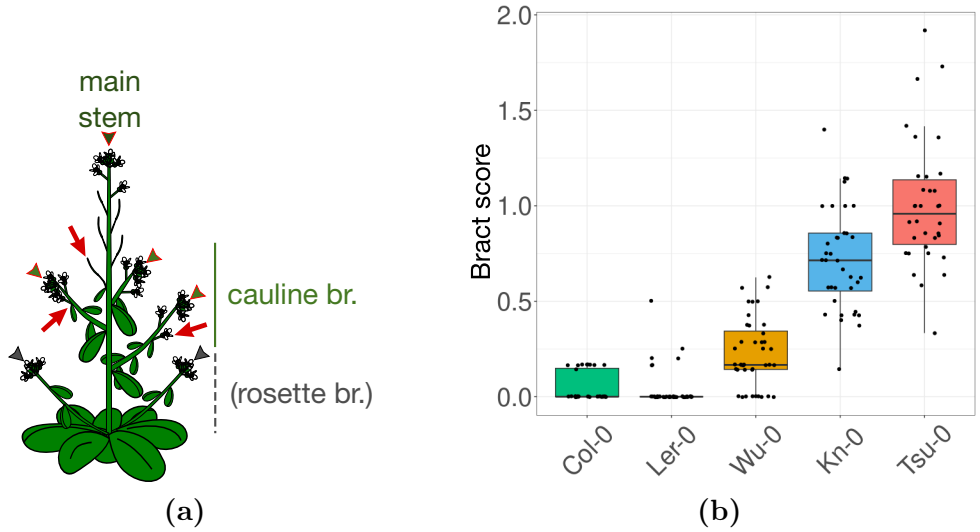


Figure 4.5: *Tsu-0* and *Col-0* were identified as high and low bract producers, respectively using the bract scoring system defined by Dieudonné *et al.* [1]. (a) The description of the plant bracts used for the bract scoring system [1]. (b) Bract scores of different Arabidopsis accessions.

4.1.3 Mutant analysis of basal bract development in Arabidopsis

Dieudonné *et al.* [1] initiated their study of basal bract development by exploring the role of *LFY* through the examination of mutants with altered *LFY* expression, including *lfy*, *ufo*, and *puchi* x *bop1* x *bop2*. In these mutants, they discovered that bracts were typically found at the tips of old branches or more frequently on secondary shoots, contrasting with *Tsu-0*, where basal bracts were limited to the first flowers at the floral transition. This suggests that bract formation in *Tsu-0* is not directly due to a general perturbation of *LFY* function. Furthermore, mutants with reduced *LFY* expression did not show an increase in basal bracts, indicating that bract formation in *Tsu-0* is not simply due to lower *LFY* levels. When examining *tfl1* mutants, which do produce basal bracts, they observed that these bracts are likely cauline leaves transformed into flowers, differing from the true bracts observed in *Tsu-0*.

Collectively, these phenotypic differences suggest that the genetic mechanisms controlling bract development in the mutants differ from those in *Tsu-0*, potentially influenced by the specific conditions of the floral transition [1]. This finding of bracts associated with wild-type flowers in Arabidopsis demonstrates that flower and bract formation are mutually exclusive. This report challenges the traditional view of bract and flower formation, suggesting that the underlying genetic and developmental mechanisms are more complex than previously thought.

4.1.4 Basal bract frequency is independent of photo-induction and plastochron length

Dieudonné *et al.* [1] explored factors that might influence basal bract formation. Hempel and Feldman [187] previously reported that strong photoinduction in certain Arabidopsis accessions

led to the formation of bracts through a process where a young branch meristem (already subtended by a leaf) was converted into a bracteate flower. In these experiments conducted by Hempel and Feldman [187], basal bracts were triggered along with varying forms of chimeric shoot-flowers. However, Dieudonné *et al.* did not observe this in *Tsu-0* plants, nor did they find any link between bract number and shoot-to-flower conversion [1]. Contrary to Hempel and Feldman's experiments which showed rapid flower formation after photoinduction, Dieudonné *et al.* observed that *Tsu-0* plants took longer to transition. This implies that the growth conditions in Dieudonné *et al.* provided weaker photoinductive signals [1].

According to the conversion hypothesis proposed by Hempel and Feldman [187], shorter plastochrons (the intervals between the initiation of two lateral meristems) promote bract formation. Dieudonné *et al.* [1] also examined whether bract production was related to plastochron length but found no consistent difference between *Tsu-0* and *Col-0*. Notably, *Tsu-0* plants flower later than *Col-0*, and they found that late-flowering accessions tend to produce more bracts. This correlation was also observed when they compared five different accessions (*Col-0*, *Kn-0*, *Ler-0*, *Wu-0*, and *Tsu-0*), but not within a single accession [1]. This indicates a complex relationship between flowering time and bract formation. Overall, the results observed by Dieudonné *et al.* suggest that bract formation is not influenced by light or plastochron variations but is more likely linked to the timing of flowering [1].

4.1.5 Basal bract development is controlled by multiple QTLs and correlates with flowering time

To uncover the genetic basis of basal bract formation, Dieudonné *et al.* [1] performed crosses between *Tsu-0* (a high bract-score accession) and *Col-0* (a low bract-score accession). They found that the F_1 plants produced a moderate number of bracts (more than *Col-0* but fewer than *Tsu-0*) and the F_2 generation displayed a broad range of bract numbers spanning the parental extremes. This made it difficult to identify a simple genetic architecture governing the F_2 phenotypic distribution [1]. Using Bulk Segregant Analysis (BSA), they identified four quantitative trait loci (QTLs) on chromosomes 1 and 5. To refine these findings, they used Recombinant Inbred Lines (RILs) and confirmed the presence of these QTLs, focusing on two significant ones (1a and 1b) on chromosome 1. Further analysis revealed that these QTLs exhibit additive effects, meaning that when both *Tsu-0* alleles are present, the bract score closely matches that of the *Tsu-0* parent. They also reported no previously known bract-related genes were found in these QTL regions. However, the study noted correlations between bract scores and the number of cauline branches, suggesting that some genes might influence both traits. Additionally, some unusual phenotypes, such as shifted bract positions and incomplete floral development, were observed in the RILs, indicating that complex genetic interactions might be necessary for proper bract development in *Tsu-0*. This suggests that complex genetic interactions may be required for bract development in *Tsu-0*. Overall, their findings indicate that basal bract formation in *Tsu-0* is controlled by multiple new genetic factors, involving both additive and interacting effects [1].

4.1.6 Hypothesis and aim

The evolution of bract-less or leaf-less inflorescences in multiple plant groups, including the Brassicaceae, serves as a well-known example of evolutionary loss. Previous mutant studies in Arabidopsis suggest that floral meristem identity suppresses bract development, though the exact developmental and evolutionary mechanisms remain elusive. Dieudonné *et al.* [1] revisited this idea by examining the transient bracts that appear at the base of flowering branches in some natural accessions of Arabidopsis. Given the notable differences in bract production between the Arabidopsis accessions *Tsu-0* and *Col-0*, these two accessions were selected for deeper analysis. Preliminary findings by Dieudonné *et al.* [1] indicate that bract production is independent of photoinduction and plastochron length and may instead be linked to later flowering times. The identification of multiple QTLs associated with bract formation, in regions that do not contain any previously identified bract-related genes, suggests the involvement of new genetic mechanisms. Although the mechanisms underlying bract development in mutants remain unclear, earlier studies propose that bract loss in higher plants, like those in the *Poaceae* family, results from changes in developmental timing rather than the gain or loss of specific genes [188].

Building on this, our study explored time series transcriptomics of both accessions. We hypothesised that the differential development of basal bracts in the *Tsu-0* and *Col-0* accessions of Arabidopsis was driven by distinct transcriptional changes during the floral transition. By performing DEG analysis at different stages within accessions and between accessions, we aimed to identify the most critical stage of transcriptional divergence during this transition. We also predicted that key regulatory genes involved in floral transition and bract formation would exhibit unique expression patterns in *Tsu-0* that were absent or significantly different in *Col-0*, leading to the presence of bracts in *Tsu-0*. By analysing RNA-Seq time series data across morphologically matched developmental stages, we expected to identify candidate genes and pathways that were differentially regulated between the two accessions, particularly at the crucial stage during the floral transition, which could provide insights into the genetic mechanisms underlying bract formation.

We also performed an alternative approach, specifically targeting the comparison of bract-less and leaf/bract-producing meristem stages to novel candidate genes involved in bract development. We hypothesised that these candidate genes, potentially linked to bract development, would exhibit distinguishable expression patterns between bract-present and bract-absent stages. By identifying and characterising these genes, we aimed to uncover previously unrecognised pathways and molecular mechanisms that drove the unique bract development in this accession.

Building on previous research by Calderwood *et al.* [85] which demonstrated the transcriptomes of two genotypes (accessions or species) during flowering cannot be aligned to a single developmental timeline (each gene may exhibit varying degrees of resynchronisation), we hypothesised that while most of the genes maintained similar dynamics, they also experienced resynchronisation.

This desynchronisation likely led to heterochrony at the floral transition, contributing to bract derepression in Arabidopsis. We also predicted that in *Tsu-0*, shifts in the timing of gene activation or repression relative to the floral transition stage would differ from those observed in the *Col-0* accession. These timing differences might lead to distinct transcriptional states in *Tsu-0*, which support the transient formation of bracts. By applying a curve-registration approach and analysing the transcriptome-wide timing of gene expression, we aimed to identify key regulatory processes that are desynchronised between *Tsu-0* and *Col-0*. This study built on the findings of Calderwood *et al.* [85], suggested that such transcriptional desynchronisation could reveal novel mechanisms underlying developmental variations, potentially including bract formation.

4.2 Material and methods

4.2.1 Gene expression time series data of *Tsu-0* and *Col-0*

The gene expression data Arabidopsis genotype *Tsu-0* and *Col-0* from Dieudonné *et al.* [1] was used for the analysis of this chapter. Both plant seeds were sown on peaty-clay soil, stratified at 4°C for at least two days, and watered with fertiliser (18-10-18 N-P-K) under LED lighting (sunlight spectrum NS12, 150 µmol.m-2.s-1). Three different day/night regimes were used in the experiments: short-days (SD) with 8h light and 16h dark; long-days (LD) with 16h light and 8h dark and continuous light (CL) with 24h light. Temperature and humidity were controlled as follows: 22°C 60% humidity in CL, and 22°C 60% humidity/18°C 70% humidity day/night in LD and SD conditions. For the RNA-Seq time course, plants were grown for 20 days before switching to SD. *Tsu-0* and *Col-0* meristems were dissected every day in LD conditions, to capture the precise developmental stages.

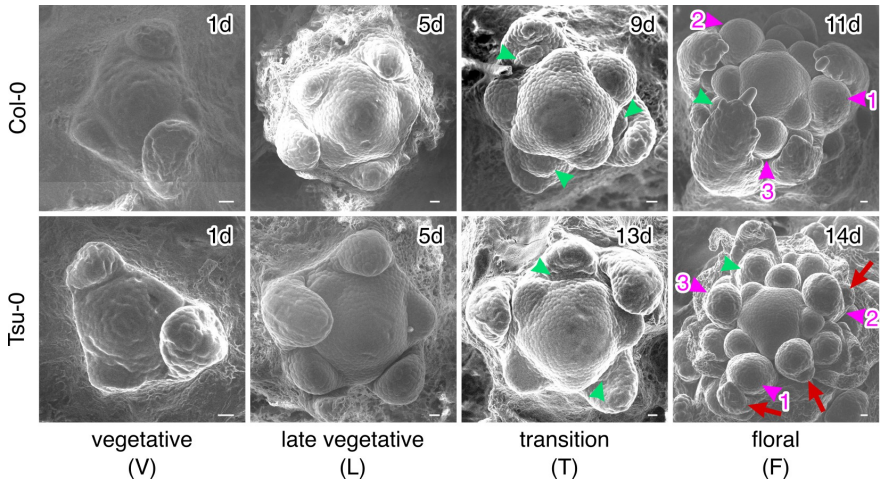


Figure 4.6: Scanning electron microscopy images showing the development of the main meristem in *Col-0* and *Tsu-0* at four stages (V, L, T, F) used for RNA sequencing. Plants were synchronised with 21 days of non-inductive SD light conditions before transitioning to inductive LDs. The number of days after this transition is noted in the top-right corner of each image. Green arrowheads indicate branches with leaves, magenta arrowheads mark the first flowers after floral transition, and in *Tsu-0*, red arrows highlight bracts. Taken from [1].

Figure 4.6 shows scanning electron microscopy images that illustrate the development of the main meristem in both *Col-0* and *Tsu-0* across four distinct stages corresponding to those used in RNA-Seq analysis. These stages are defined as follows: V, the day of transfer from short-day (SD) to long-day (LD) conditions; L, the stage at which *Col-0* and *Tsu-0* exhibit identical meristem shapes; T, the stage at which the first flower emerges, marked by the appearance of a round rather than triangular primordium and the visible initiation of axillary meristem formation at the axils of young leaf primordia; and F, where flowers are distinguishable, with the first whorls differentiated on the initial flower. The timing after the LD transfer is indicated in the top-right corner of each image in Figure 4.6. Green arrowheads highlight the branches bearing leaves, while magenta arrowheads indicate the first flowers produced following the floral transition. In *Tsu-0*, bracts are marked with red arrows. Three independent biological experiments (R1, R2, R3) were performed with 5 to 11 meristems per replicate. Figure 4.7 illustrates the timing of tissue sampling, aligned with the corresponding developmental stages for both *Tsu-0* (represented in red) and *Col-0* (represented in green). Different symbols are used to denote each developmental stage.

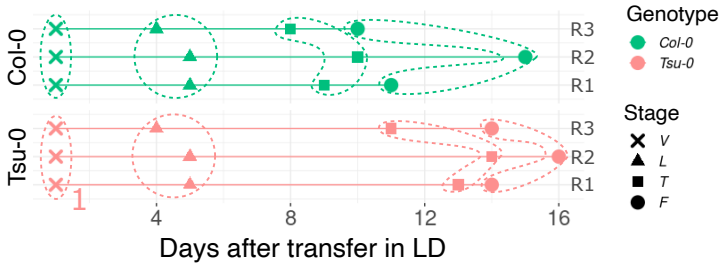


Figure 4.7: Collection times (in days) of the samples used in the RNA-Seq experiment conducted by Dieudonné *et al.* [1] with corresponding developmental stages across different biological replicates.

4.2.2 RNA-Seq analysis, differentially expressed gene analysis, and GO term analysis

Principal component analysis (PCA) was carried out using the `PCA()` function as part of the `FactoMineR` package [189]. Differential analysis was performed using the R Bioconductor package `edgeR` [41]. Reads were first normalised using TMM (Trimmed mean of M-values) to reduce library-specific biases. Normalisation factors were between 0.94 and 1.049. Three types of DEG analysis were considered: DEG at each stage between the different genotypes, DEG between the stage within the same genotype, and DEG across all conditions (stages and genotypes). Multiple DEG analyses were corrected using Benjamin-Hochberg correction, and genes with a p-value < 0.05 were retained. GO term enrichment analyses were performed using function `enrichGo()` from `clusterProfiler` 4.0 [190] with default parameters and the "BH" adjustment method.

4.2.3 Comparative gene expressions analysis on bract-present and -absent developmental stages to identify genes associated with bract development

We categorised the stages into two groups: those with bracts and those without, as illustrated in Figure 4.8. To identify candidate genes potentially associated with bract development, we first assessed the minimum and maximum expression levels across stages with and without bracts. Genes were considered positively associated with bract development if their minimum expression level during bract-present stages was greater than or equal to their maximum expression level during bract-absent stages. Conversely, genes were considered negatively associated if their maximum expression level in the bract-present stages was lower than or equal to their minimum expression level in the bract-absent stages. Genes that did not meet either of these criteria were excluded from further analysis.

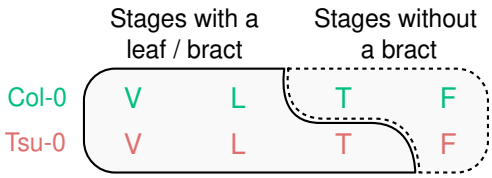


Figure 4.8: Stages characterised by the presence or absence of leaves and/or bracts in both *Tsu-0* and *Col-0* include the vegetative (V), late vegetative (L), transition (T), and floral (F).

4.2.4 Registration

The registration process was carried out with the *greatR* [3] package, without optimisation, 75% overlapping to the reference data (`overlapping_percent = 75`), and z-score scaling (`scaling_method = "z-score"`). The standard deviation for the replicates at each time point was set to 0.01. The shift was specified from [-1, 1], while there was no stretch applied due to the matching developmental stages compared between the two genotypes. The reference and query data used are *Col-0* and *Tsu-0*, respectively. The pairwise registration result was visualised with the `plot()` function from *greatR*.

4.3 Results

4.3.1 Differentially expressed gene analysis between stages and accessions to explore potential regulatory pathways in bract development

To understand the complex regulatory pathways which may drive the bract development, we analysed the RNA-Seq time series data over the floral transition in both accessions [165]. The tissue of both accessions was sampled at the morphologically matching four developmental stages: vegetative (V), late vegetative (L), transition (T), and floral (F). All analyses presented in this chapter were conducted on these developmental stages.

Although the two accessions undergo floral transition at different absolute ages, comparing morphologically matched stages allows us to examine functionally equivalent developmental phases. We hypothesised that key regulatory genes controlling floral transition exhibit similar dynamics and levels between matched stages [165]. Conversely, molecular variations between these stages could reveal why similar developmental phases produce different phenotypes, such as the production or inhibition of bracts, as observed in nightshade meristems within and across species [105].

To test our hypothesis, we compared the gene expression dynamics of key floral transition regulators between the two accessions by plotting their average expression patterns. We selected nine genes known to be crucial for floral transition and flower identity [148, 191]. The expression of most of these genes, such as *FT*, *AP1*, *LFY*, and *FUL*, showed no significant differences between

the two accessions (see Figure 4.9). The identical expression of these genes in both accessions suggests that these genes are not involved in the development of basal bracts in *Tsu-0*. For example, in the case for *LFY*, this observation aligns with findings from p*LFY* transcriptional reporter lines reported by Dieudonné [165]. They observed the same sharp activation of *LFY* transcription in both *Tsu-0* and *Col-0*, despite the differing bract presence phenotypes between the two accessions.

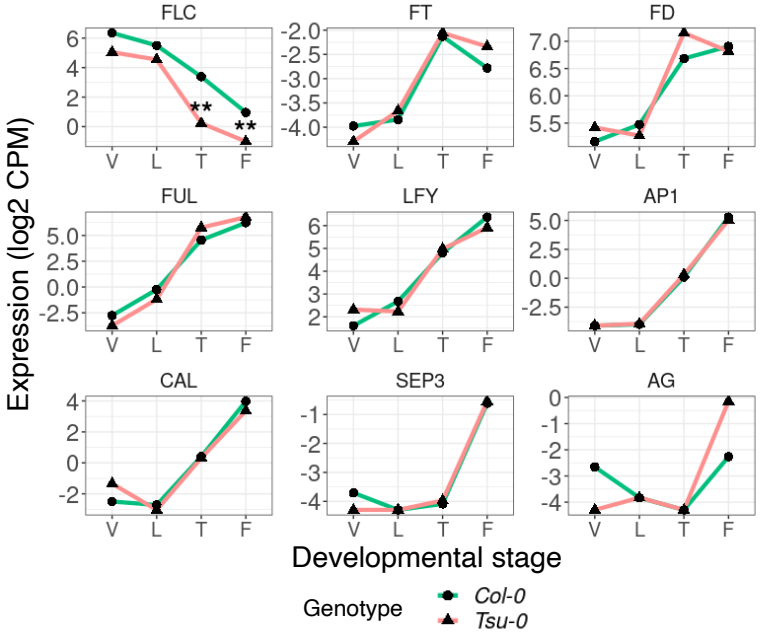
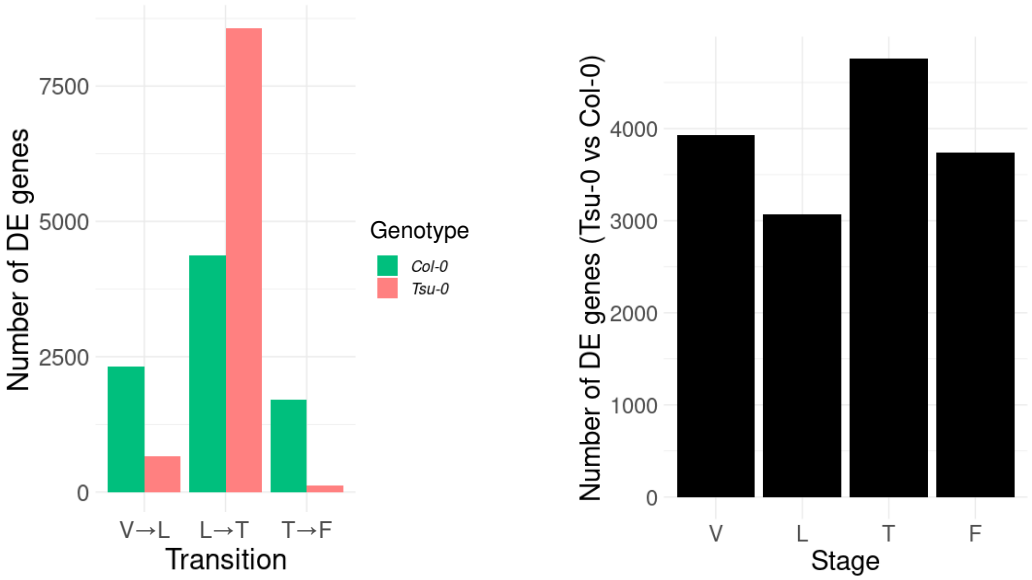


Figure 4.9: Mean of expression dynamics of nine key genes controlling floral transition and identity, in *Col-0* (represented in green) and *Tsu-0* (represented in red). No temporal alignment or curve registration was applied, the profiles represent original expression measurements. Two stars indicate that the fold change of this difference is greater than 1.

For each accession, DEG analysis was carried out through pairwise comparisons between consecutive developmental stages: V vs. L, L vs. T, and T vs. F (Figure 4.10 (a)). The highest number of differentially expressed genes was observed between stages L and T in both accessions, highlighting this transition as the most critical stage. Additionally, we performed DEG analysis between the accessions at each developmental stage (Figure 4.10 (b)). The comparison between *Col-0* and *Tsu-0* revealed 4,759 differentially expressed genes at the T stage, the highest number observed compared to other stages.



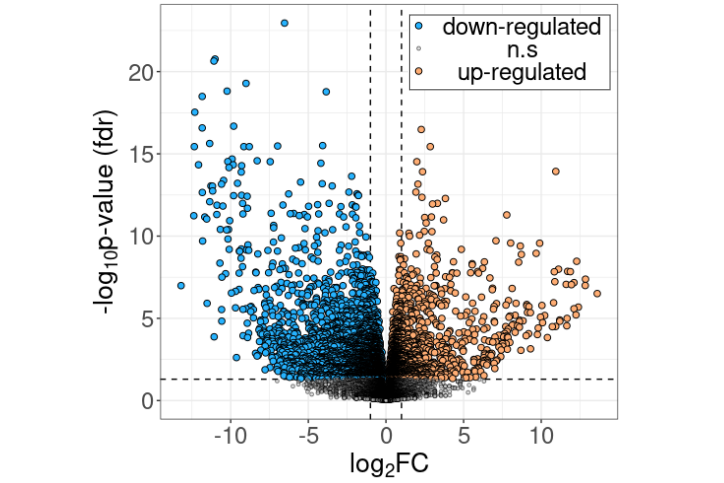
(a) (b)

Figure 4.10: Number of differentially expressed genes (DEGs): (a) between consecutive stages in *Col-0* (green) and *Tsu-0* (red), with the greatest number of changes occurring during the transition from L to T stage, particularly in *Tsu-0*. (b) between *Col-0* and *Tsu-0* at each stage of the time course, with the highest number of DEGs observed at the T stage.

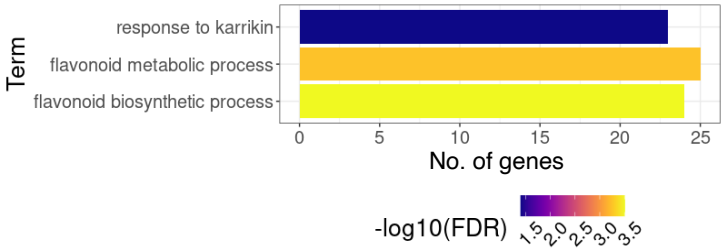
It was previously observed by Dieudonné *et al.* [1] that basal bract formation in *Tsu-0* is observed to be transient since only the 1 to 5 first flowers present a bract. The morphological difference in both *Col-0* and *Tsu-0* during the transitions between T and F are also quite close, with approximately only one to 2 days (see Figure 4.7). Heisler *et al.* [192] demonstrated that bract and flower initiations begin with lateral auxin accumulation before any visible morphological changes occur. Given the previous observation where there is no consistent shorter plastochron in *Tsu-0* compared to *Col-0* [165], it can be inferred that the initiation of the first flowers and their associated bracts likely occurs during stage T. Therefore, we expect the transcriptional changes associated with bract formation in *Tsu-0* should transiently appear at stage T and progressively fade out at F stage.

We performed GO term analysis for significantly up-regulated and down-regulated genes at stage T between the two accessions (Figure 4.11). The enriched GO terms for down-regulated genes include responses to karrikin and flavonoid-related processes. For the up-regulated genes, the enriched GO terms include processes related to single-organism metabolism, glycosinolate biosynthesis, and responses to stimuli and chemicals. However, these GO terms do not appear to be directly related to bract formation and therefore offer limited insights into the molecular mechanisms underlying this specific developmental process.

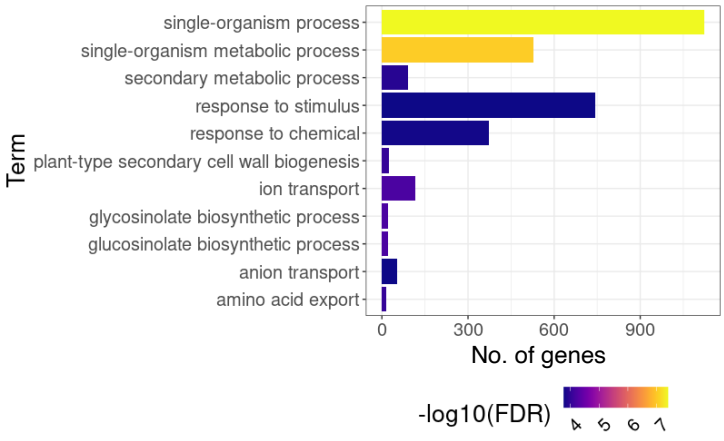
Although Dieudonné *et al.* [1], through mutant studies, previously observed that the genes involved in bract development are unlikely to be the primary drivers of basal bract formation in *Tsu-0*, they may still play a role in the overall bract development process. Figure 4.12 illustrates



(a)



(b)



(c)

Figure 4.11: GO term analysis at the T stage. (a) Volcano plot showing gene expression differences between the two accessions at the T stage. All genes expressed in the shoot apical meristem are plotted as grey dots. Genes above the statistically significant threshold (indicated by the horizontal dashed line) are highlighted in orange for up-regulated and blue for down-regulated genes. Vertical dashed lines indicate an absolute fold change greater than 1. (b) Significantly enriched biological process (BP) GO terms for the up-regulated DEGs identified in (a), along with the corresponding number of genes. (c) The top ten significantly enriched BP GO terms for the down-regulated DEGs identified in (a), along with the corresponding number of genes.

the expression dynamics of bract-related genes in both *Tsu-0* and *Col-0*. Of the 12 genes analysed, *PUCHI*, *SOC1*, and *TFL1* showed significant differences in expression at stage T between *Tsu-0* and *Col-0*. However, these differences remained until stage F, despite bracts disappearing as the flowers matured. This suggests that the differential expression of these genes is not critical for bract formation.

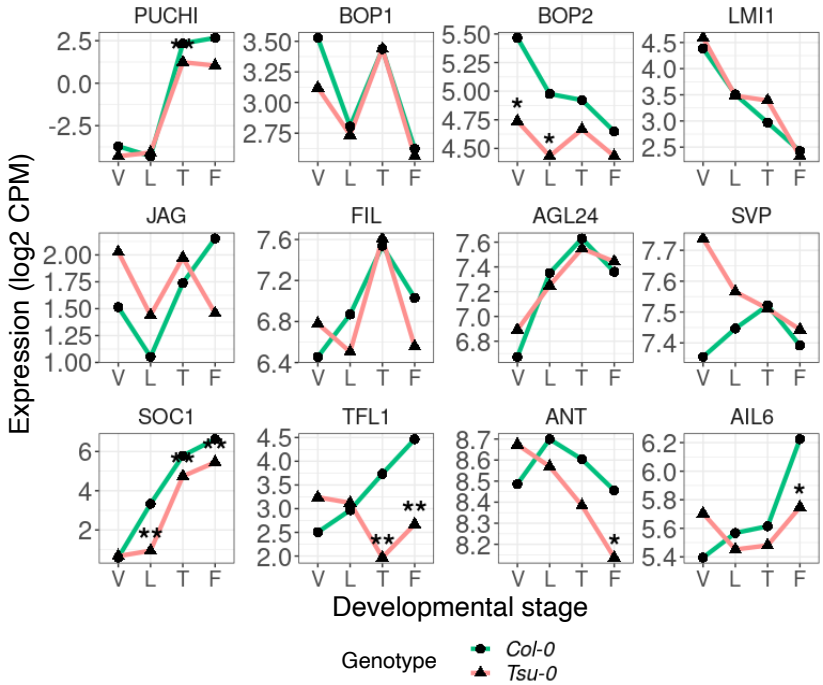


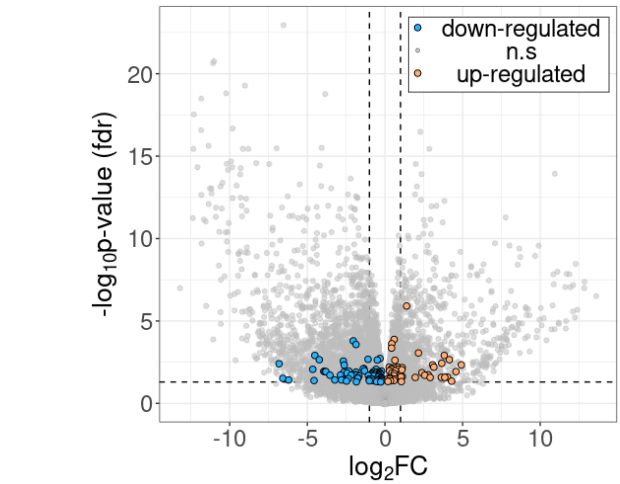
Figure 4.12: Mean of expression dynamics of twelve previously identified bract-controlling genes in *Col-0* (represented in green) and *Tsu-0* (represented in red). One star indicates a significant difference between the two expression levels and two stars indicate that the fold change of this difference is greater than 1.

4.3.2 Comparative analysis between bract-less and bract-producing stages reveals new candidate genes and pathways

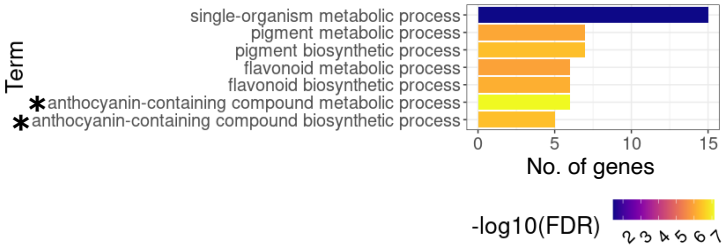
Since no significant regulatory pathways were identified using DEG analysis discussed in the previous section, we employed an alternative approach to identify potential candidate genes involved in bract development. We call these genes potential positive and negative bract regulators. To identify these candidate genes, we first identified two different groups of stages which can be compared to capture potential candidate genes which are responsible for bract developments: bract-less stages (*Col-0* T, F, and *Tsu-0* F) or leaf/bract-producing meristems stages (the other stages, including *Tsu-0* T). We selected genes separating these two groups to be the putative bract regulators (see Method) and found 124 genes in this category. This includes *SOC1* as a putative negative bract regulator.

We performed GO term enrichment analysis for the identified negative and positive bract regulator genes. Among the enriched terms, anthocyanin biosynthesis was associated with up-

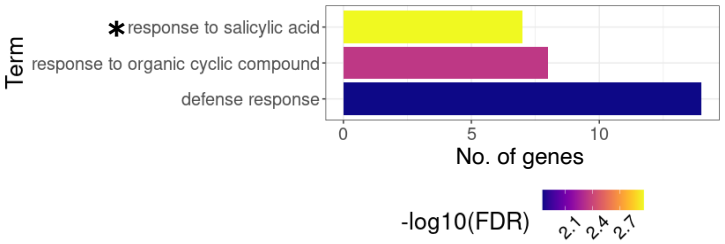
regulated genes, while salicylic acid (SA) was linked to down-regulated genes (Figure 4.13). When these two ontology terms were mapped back to all genes at the T stages, we observed a higher percentage of DE genes in *Tsu-0* associated with these pathways, indicating different activity levels between the two accessions at this critical stage. Dieudonné *et al.* [1] further confirmed the higher anthocyanin production in *Tsu-0* at the L and T stages through purple colouration below the meristems and at the base of the growing stem. In contrast, *Col-0* consistently exhibited pale green tissues (Figure 4.14 (a)). Occasionally, this purple colouration extended to young organs in the meristems (Figure 4.14 (b)). As the stem continued to grow, the pigment persisted at the rosette junction in both *Tsu-0* and *Col-0*, while fading at the apex. This transient meristem colouration supports our results, which identified genes relevant to bract development. Notably, among the 124 putative bract regulators identified based on their expression (see Table S.2), twelve were located within the mapped QTLs previously reported by Dieudonné *et al.* [1].



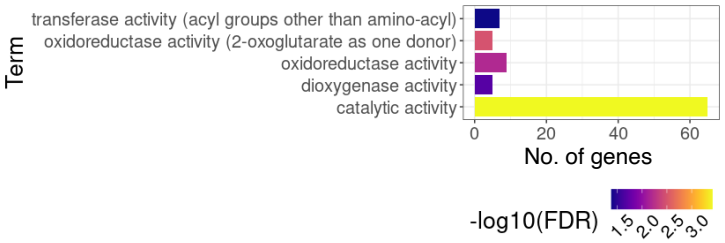
(a)



(b)

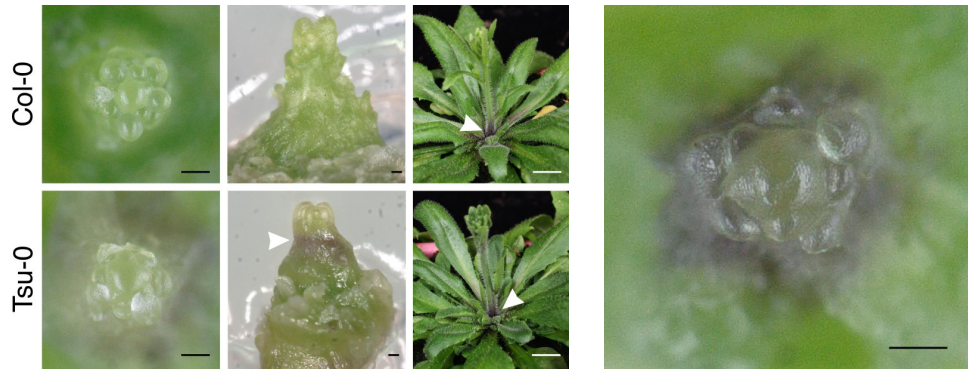


(c)



(d)

Figure 4.13: GO term analysis of bract regulator genes at the T stage. (a) Volcano plot showing gene expression differences between the two accessions at the T stage. All genes expressed in the shoot apical meristem are plotted as grey dots. Genes fulfilling the clustering condition above the statistically significant threshold (indicated by the horizontal dashed line) are highlighted in orange for up-regulated and blue for down-regulated genes. (b) Significantly enriched biological process (BP) GO terms for the up-regulated DEGs identified in (a), along with the corresponding number of genes. (c) Significantly enriched BP GO terms for the down-regulated DEGs identified in (a), along with the corresponding number of genes. (d) Significantly enriched molecular function GO terms of all the genes identified in (a).



(a) (b)

Figure 4.14: (a) Representative images of micro-dissected meristems from *Col-0* (top row) and *Tsu-0* (bottom row) are displayed at different stages: just before or at stage T (left), after stage T (middle), and a close-up of the base of the bolted main stem (right). At stage T, *Tsu-0* meristems exhibit a distinct anthocyanin red colouration just below the meristem, which is absent in *Col-0*. Following bolting, both genotypes show anthocyanin colouration at the base of the stem. Scale bars: 100 µm (left and middle), 1 cm (right). Taken from [1]. (b) A wild-type *Tsu-0* micro-dissected meristem at the T stage showing high anthocyanin colouration at the base of the stem and up to young developing organs. Scale bar: 100 µm. Taken from [1].

Dieudonné *et al.* [1] also performed DEG analysis using data from various mutants to identify genes involved in bract development without prior assumptions. Specifically, mutants like *lfy* and *puchi* x *bop1* x *bop2* that stop producing leaves at the floral transition were compared to *jagged-5d* plants, which consistently produce bracts. As expected, these mutants clustered with *Col-0* at a similar developmental stage. By isolating differentially expressed (DE) genes specific to *jagged-5d*, their analysis revealed an enrichment of genes related to shoot and leaf development, photosynthesis, and metal ion transport. Cross-referencing these DE genes with QTL-mapping data, they identified 33 candidate genes potentially involved in bract formation (Table S.1) [1]. None of these genes are currently known to be linked with bract development or flowering, suggesting the involvement of novel genetic pathways in bract development. However, we identified one anthocyanin biosynthetic enzyme (dihydroflavonol reductase, *DFR*) and five SA-responsive genes. Figure 4.15 illustrates the expression profiles of two of these genes, *DFR* and *FMOGS-OX7*. Further research is needed to determine whether these candidates contribute to basal bract formation in *Tsu-0* and whether the anthocyanin and/or SA pathways play a role in this natural variation. Overall, our analysis suggests that most genes previously linked to bract development in mutants may not be involved, instead highlighting new candidate pathways that could promote bract outgrowth during the unique and transient floral transition stage in the two Arabidopsis accessions.

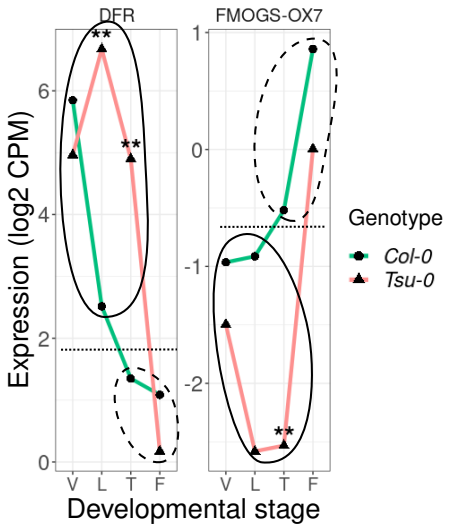


Figure 4.15: Expression profiles of two candidate genes at the T stage, showing up-regulation (left) and down-regulation (right). *DFR* is involved in anthocyanin biosynthesis, while *FMOGS-OX7* responds to salicylic acid. Two stars indicate significant differential expression with a fold change greater than 1. The bract and bract-less clusters are marked by solid and dashed circles, respectively, with the horizontal dotted line showing their separation.

4.3.3 Curve registration analysis shows that bract development happens during a period when many genes are desynchronised

The approach discussed in the previous section (see Method, Figure 4.8) could mostly identify genes where the RNA levels in *Tsu-0* change later than in *Col-0* (see Figure 4.15). This behaviour is an example of transcriptional heterochrony, which means a shift in the timing of when genes are activated or repressed, often due to changes in cis-regulatory gene regions [193]. These timing differences in gene activity can occur at different stages in different genotypes [1]. We propose that bracts represent a classic case of heterochrony because they are a juvenile trait (leaf-like structures) that continue to appear later in development [1]. This fits with the fact that *Tsu-0* flowers later than *Col-0*. Even though a direct comparison of the developmental stages between *Tsu-0* and *Col-0* should minimise most of this difference, as shown in Figure 4.12, some differences in timing between the two accessions could still be observed.

We hypothesise that looking at changes in gene activity timing across the whole transcriptome could reveal important processes happening in *Tsu-0* during the floral transition. To do this, we first performed a PCA analysis on the RNA-Seq data from both accessions. The result is presented in Figure 4.16, which shows the two primary axes that account for the most variance in the data. The second axis, which explains about 24.4% of the variance, separates the two genotypes. The main axis, accounting for approximately 47.9% of the total variance, organises the sampling time points from the earliest to the latest stages in both genotypes. We interpret this main axis as representing the "transcriptomic age."

Interestingly, along this main axis, the T stage in *Tsu-0* does not align with the T stage in *Col-0*. Instead, it clusters more closely with the F stages, indicating that the transcriptome at this stage in *Tsu-0* is more similar to the F stage. This suggests that in *Tsu-0*, the transition from the T stage to the F stage involves minimal changes in gene expression. This pattern implies a possible delay or shift in the timing of gene expression in *Tsu-0* compared to *Col-0*, which could be a key factor in the distinct developmental processes that occur in *Tsu-0* during the floral transition.

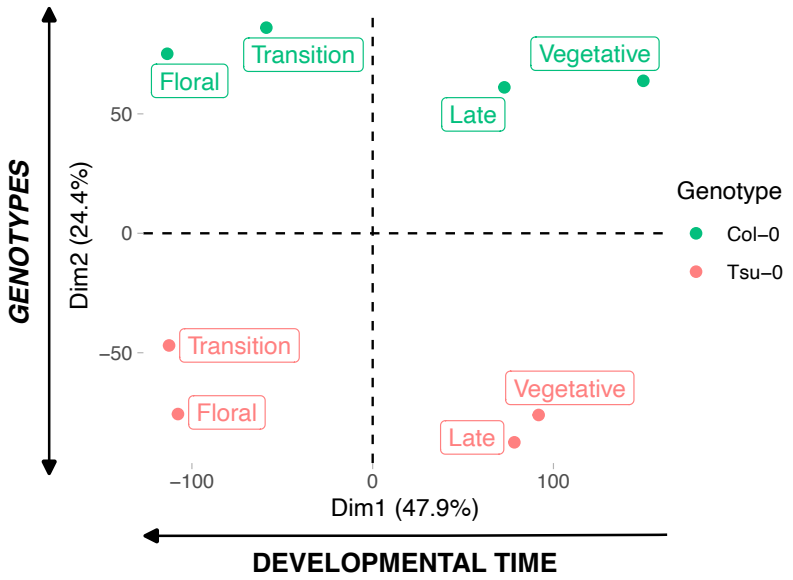


Figure 4.16: PCA analysis results of RNA-Seq time series over four developmental stages (vegetative (V), late vegetative (L), transition (T), and floral (F)). The x-axis can be interpreted as developmental time, and the y-axis as genotypes.

A previous study by Calderwood *et al.* [85] found that during flowering, the transcriptomes of two different accession genotypes (*B. rapa* cv. *R-o-18* and *B. rapa* cv. *Sarisha-14*) and species genotypes (Arabidopsis *Col-0* and *B. rapa* cv. *R-o-18*) cannot be perfectly aligned to a single developmental timeline. Each gene may exhibit different timing, with some genes becoming active earlier or later in one genotype compared to the other. This means that gene expression patterns are not consistently synchronised between the two genotypes during flowering. We hypothesise that this is also the case in *Col-0* and *Tsu-0* where most of the genes between the two accessions exhibit similar dynamics but are differently synchronised. This means that differences in *Col-0* and *Tsu-0* are primarily due to timing variations rather than inherent differences in expression profiles.

To measure gene desynchronisation between *Col-0* and *Tsu-0*, we applied the same curve-registration approach [85, 3]. This method is particularly useful for detecting subtle temporal shifts in gene expression within our dataset. In this analysis, the shifts are measured relative to the developmental stages (V, L, T, F), which is used as the common reference point for both accessions. Therefore, genes that perfectly match between the two accessions (have no shift), such as *AP1* (see Figure 4.17), may still be shifted in absolute time. On the other hand, positive

(e.g. *AG*) and negative shifts (e.g. *DFR*) indicate whether gene expression dynamics occur earlier or later, reflecting a desynchronisation of the floral transition. If a gene is not successfully registered, as in the case of *CYP705A9* in Figure 4.17, it suggests that these genes exhibit different expression dynamics between the two accessions.

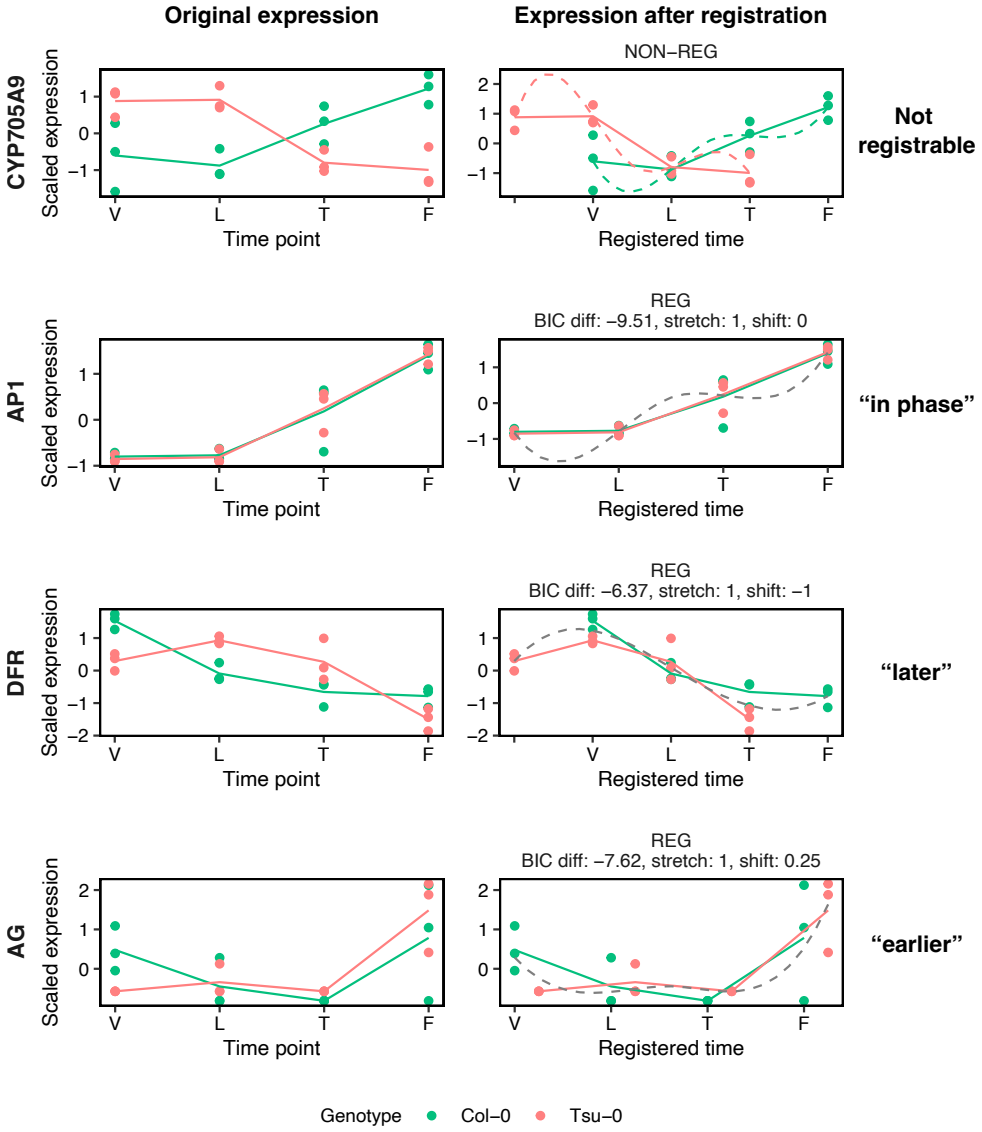


Figure 4.17: Examples of temporal registration of gene expression dynamics (right panels) between *Col-0* (green) and *Tsu-0* (red) based on scaled expression levels (left panel). In the left panel, dots represent the expression levels of independent biological replicates, while lines indicate the mean expression level at each time point. In the right panel, the green (*Col-0*) and red (*Tsu-0*) dotted curves represent the fitted models for each genotype independently, while the grey dotted curve represents the joint model for both *Col-0* and *Tsu-0*. When the green and red dotted curves are used, it indicates that two independent models best explain the time series, suggesting dissimilarity between the genotypes. Conversely, if the grey dotted curve is used, it indicates that a single model best explains both time series, suggesting a similarity between them. The last column provides a biological interpretation of the computed shift: a null shift indicates that the expression dynamics in *Tsu-0* remain 'in phase' with the floral transition, while negative or positive shifts indicate that the expression dynamics in *Tsu-0* are desynchronised and occur later or earlier, respectively, compared to the phenotypic progression of the floral transition.

Since we only wanted to consider the expression dynamics between the two accessions, we used the scaled expression dynamics for this analysis. Out of the total 21,568 genes, only 43 genes were not successfully registered (Table S.3), this includes *CYP705A9* in Figure 4.17. The successful registration of the vast majority of genes suggests that most genes follow very similar temporal dynamics in both accessions.

Based on the registration results, we categorised the genes into three groups according to their shift factors: null, negative, and positive shifts (Figure 4.18). To understand these shifts biologically, consider how a gene's expression dynamics align with the phenotypic progression of the floral transition through the four stages (V, L, T, F) in *Col-0*. A null shift means that the gene's expression dynamics in *Tsu-0* remain "in phase" with the floral transition, such as in the case of *AP1*. On the other hand, a positive or negative shift indicates that the gene's expression dynamics occur earlier or later, respectively than the floral transition. For instance, *AG* exhibits a positive shift (earlier expression), while *DFR* shows a negative shift (later expression). These shifts highlight heterochronies between the transcriptomic and phenotypic levels, revealing differences in timing between gene expression and the observable stages of floral development.

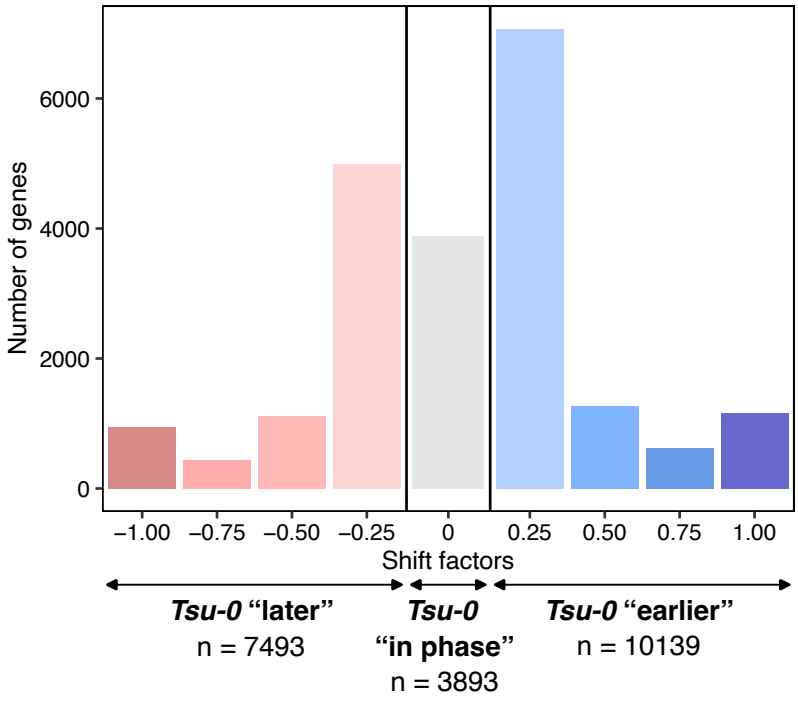


Figure 4.18: Distribution of heterochronic shifts resulting from the registration of the entire transcriptome between *Tsu-0* and *Col-0*. Shift values are colour-coded using a red-to-blue gradient, ranging from -1 to 1.

Across the entire transcriptome, the shifts were widely distributed, showing that gene dynamics are complex and often out of sync between the two accessions (Figure 4.18). Overall, more genes tend to shift earlier in relation to the floral transition. This finding aligns with the PCA results (Figure 4.16) and suggests that, even though the bract can be seen as a "juvenile" trait, it does

not represent the majority of heterochronies observed at the transcriptomic level. This means while the bract may appear to indicate delayed development, most gene expression changes are actually occurring earlier, highlighting a deeper and more widespread desynchronisation in gene activity. Furthermore, in *Tsu-0*, the floral transition is not imposing the clock for the entire meristematic transcriptome since it is only 3879 genes (18% of total genes) stay in phase with this phenotypic event while a majority have either delayed or advanced expression dynamics.

When we looked closely at the major regulators of flowering (Figure 4.19 and 4.21), genes in phase (shift = 0) are mostly related to flowering and developmental phase change. This finding is also supported by GO term enrichment analysis shown in Figure 4.22. In contrast, genes associated with known bract mutants exhibit a wide range of shifts, from very early to very late stages (Figure 4.20 and 4.21). This variability suggests that these genes are not working together in a coordinated manner for bract development in *Tsu-0*. Instead, the diverse timing of these genes indicates that a single, unified genetic program might not govern bract formation.

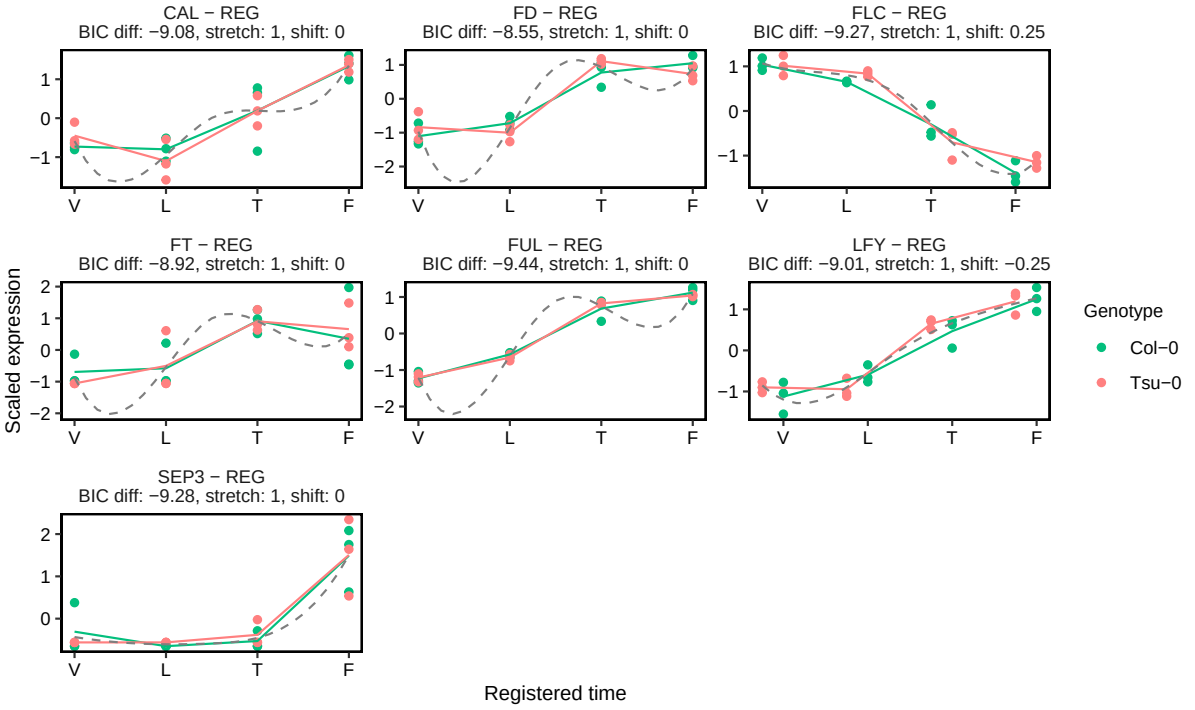


Figure 4.19: Registration results of scaled expression dynamics between *Tsu-0* (red) and *Col-0* (green, used as the reference) during the floral transition for a selected set of key genes controlling floral transition and identity (see Figure 4.9).

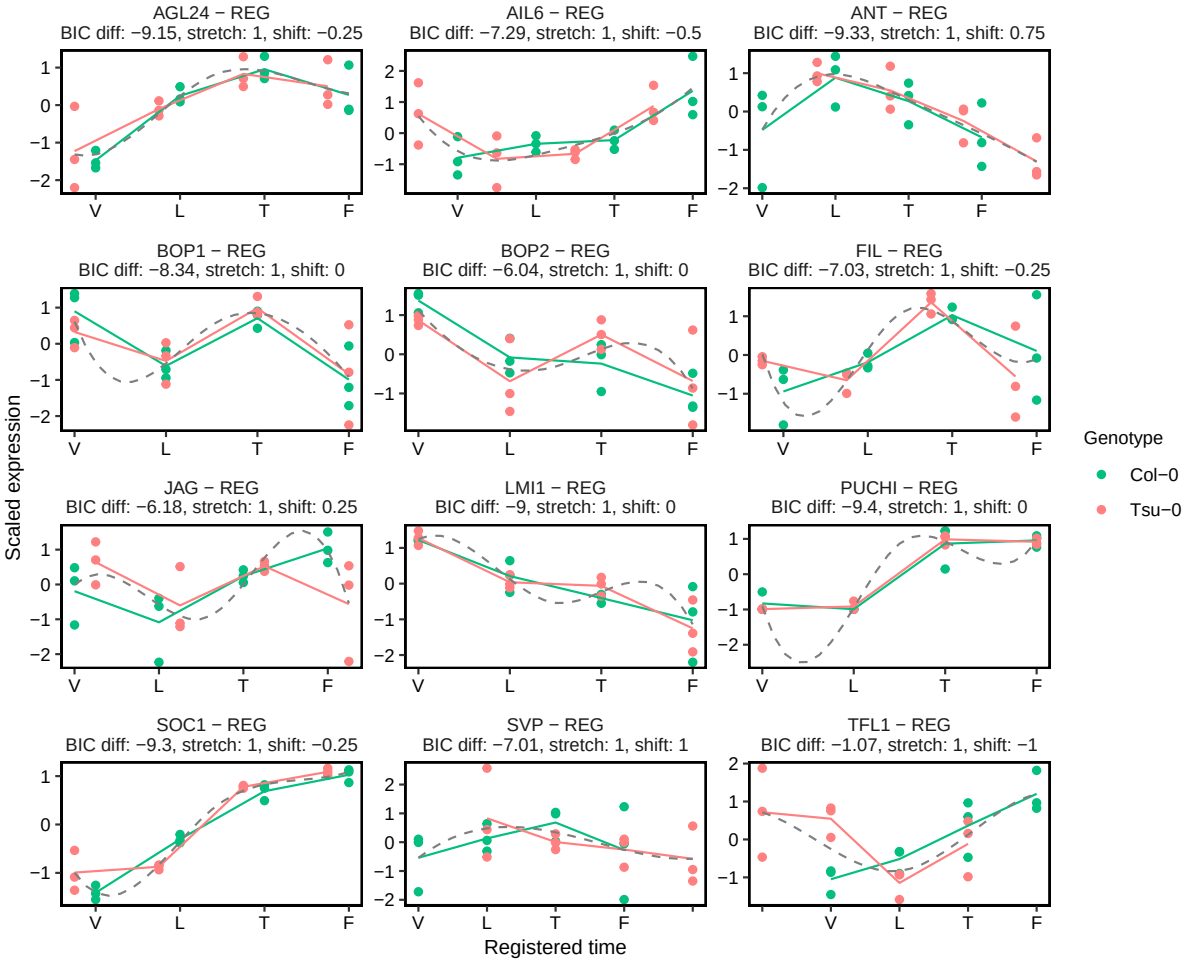


Figure 4.20: Registration results of scaled expression dynamics between *Tsu-0* (red) and *Col-0* (green, used as the reference) during the floral transition for a selected set of previously identified 'bract' genes (see Figure 4.12).

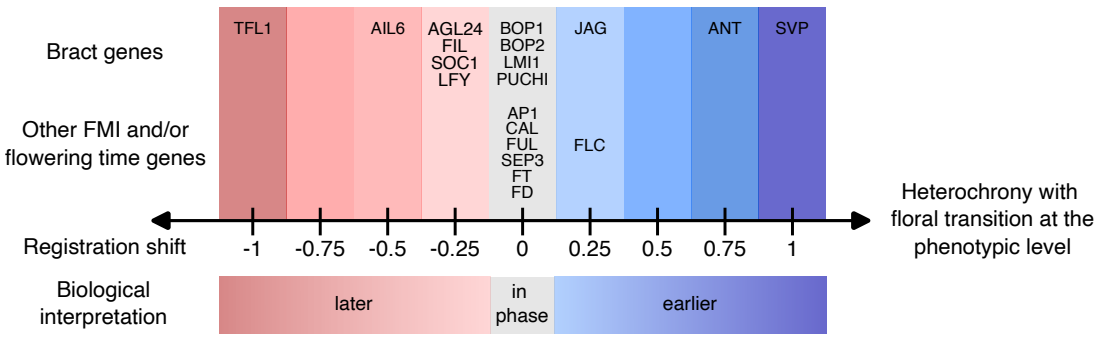


Figure 4.21: List of genes related to bract development and/or floral transition and identity and their corresponding shift factors.

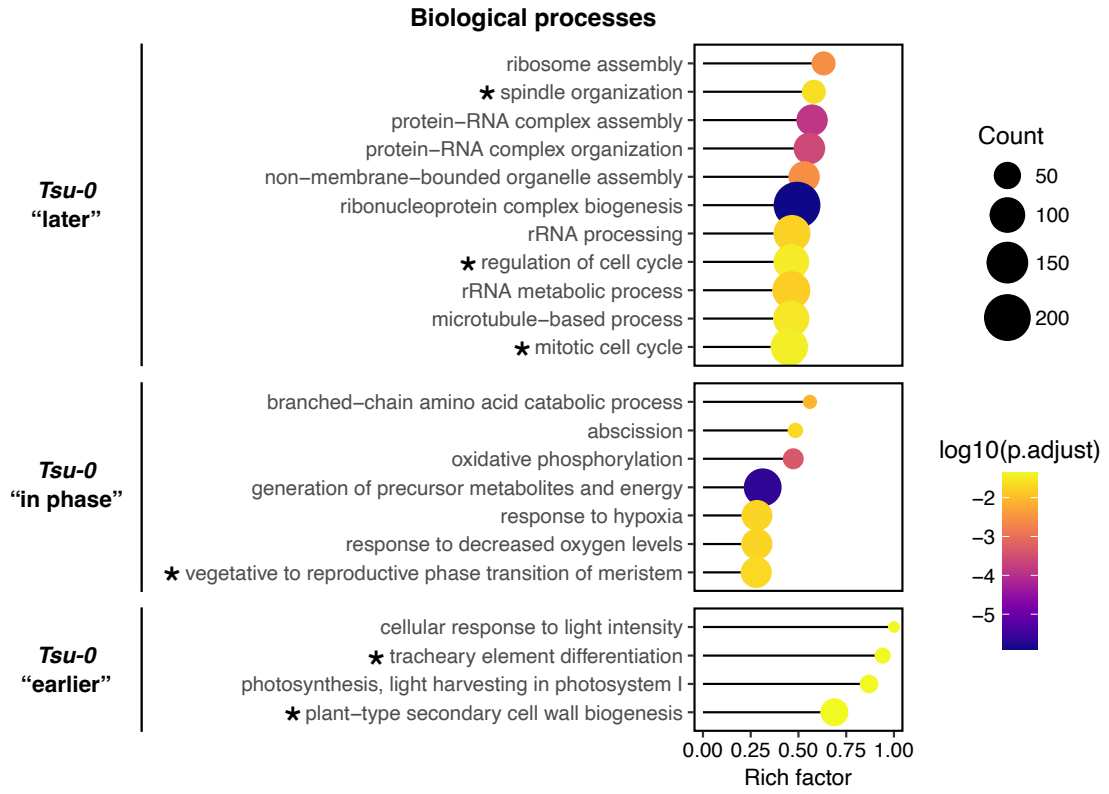


Figure 4.22: GO term enrichment analysis for the three categories of heterochronic shifts. Significant BPGO terms (BH-adjusted p-value < 0.05) were simplified using semantic similarity (cutoff $= 0.7$). The Rich Factor, representing the proportion of genes involved among all genes associated with a specific GO term, was calculated for the remaining terms. The size of the dots indicates the number of genes, while the colour scale indicates the statistical significance (BH-adjusted p-value) of the enrichment within each shift category. Stars denote GO terms related to developmental processes.

GO term enrichment analysis performed on each timing category also revealed which processes are desynchronised from flowering in *Tsu-0* compared to *Col-0* (Figure 4.22). Some vascular differentiation processes (like tracheary element and secondary cell wall formation) happen earlier in *Tsu-0* before flowering (Figure 4.22). This might be due to factors like the plant's age, indicating it's not closely linked to flowering. On the other hand, processes like cell division (spindle, cell cycle, mitosis) and ribosomal biogenesis (terms related to ribosome, rRNA, and protein-RNA complexes) occur later in *Tsu-0* (Figure 4.22), suggesting that key meristematic functions last longer. Further research is needed to explore whether this extended activity is related to bract development.

This analysis focusing on gene desynchronisation helped us understand why gene expression varied the most at stage T (Figure 4.10 (b)). Even though the gene expression states at the beginning and end of this process are similar in both accessions, gene expression changes controlling the floral transition may happen at a time when the rest of the genes are expressed at different levels because they are not synchronised with flowering. Such desynchronisation, especially in fast and gene-specific processes like flowering (Figure 4.23 and 4.24), is likely to occur in

varying gene expression states. Figure 4.23 and 4.24 illustrate the proposed model for the natural formation of basal bracts in Arabidopsis [1]. Figure 4.23 illustrates the transition from the vegetative stage, where the plant produces leaves (V), to the flowering stage, where it produces flowers (F), in two different accessions, *Col-0* (top row) and *Tsu-0* (bottom row). During the transition stage (T), marked by the formation of the first flower (indicated by a purple arrow and labelled '1'), there is a noticeable shift from the development of earlier axillary meristems (indicated by green arrowheads, numbered in reverse order starting from the first flower). In the *Tsu-0* accession, the first flowers are accompanied by the development of a bract (marked by red arrows), even though floral meristem identity genes are actively expressed in the floral meristem (represented in purple). In contrast, in *Col-0*, the bract remains undeveloped, residing within what is referred to as the cryptic bract domain. Interestingly, *Tsu-0* stops producing bracts shortly after the formation of the initial flowers. Additionally, the figure shows that the developmental stages from V to F occur at different absolute times, with *Tsu-0* flowering later than *Col-0*. The progression of development from the vegetative to the flowering stage is depicted using a green-to-red colour gradient for each accession, highlighting the differences in timing and developmental processes between the two.

Figure 4.24 illustrates how desynchronisation of gene expression dynamics in the *Tsu-0* accession leads to a new gene expression state during the floral transition stage. It compares the expression patterns of four hypothetical genes between *Col-0* (top row) and *Tsu-0* (bottom row). Gene A remains synchronised with the flowering transition in both accessions, such as genes like *LFY* or *AP1*. However, in *Tsu-0*, gene B's expression is delayed, while gene C's expression occurs earlier compared to *Col-0*. These shifts in expression timing, indicated by horizontal grey arrows, lead to differences in expression levels for genes B and C at the transition stage (T) in *Tsu-0*, as highlighted by the brown vertical arrows. The small letters a–d label the gene expression curves at each time point. However, these heterochronic shifts do not account for all differences in gene expression between the accessions, as shown by the behaviour of gene D. Ultimately, the new gene expression state that arises during the T stage in *Tsu-0* supports the development of both flowers and bracts, but this state is temporary, as bract inhibition is re-established when the meristem progresses towards the flowering stage (F).

This extensive heterochronic desynchronisation of gene dynamics creates a transcriptional noise which might lead to developmental variations during flowering, such as differences in bract development. In summary, our study provides a comprehensive analysis of transcriptome-wide timing differences (heterochronies) between two Arabidopsis accessions. It shows that the natural development of bracts during the floral transition cannot simply be attributed to a longer vegetative phase. Instead, it is influenced by the complex interplay and timing of gene expression changes.

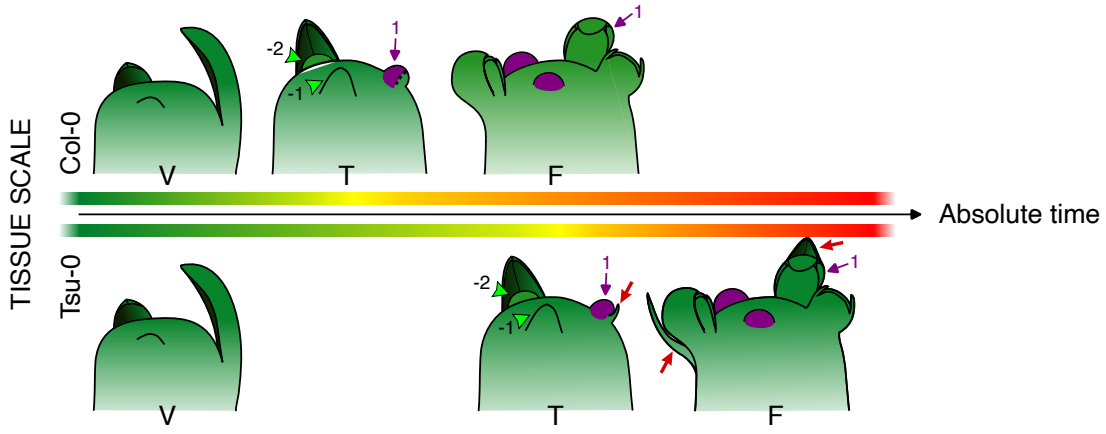


Figure 4.23: The transition from vegetative (V) to flowering (F) stages in *Col-0* (top) and *Tsu-0* (bottom) accessions. During the transition stage (T), marked by the first flower (purple arrow, '1'), *Tsu-0* develops a bract (red arrows) alongside the flower, while *Col-0* does not. *Tsu-0* ceases bract production after the initial flowers. The stages occur at different times, with *Tsu-0* flowering later. The developmental progression is illustrated with a green-to-red gradient.

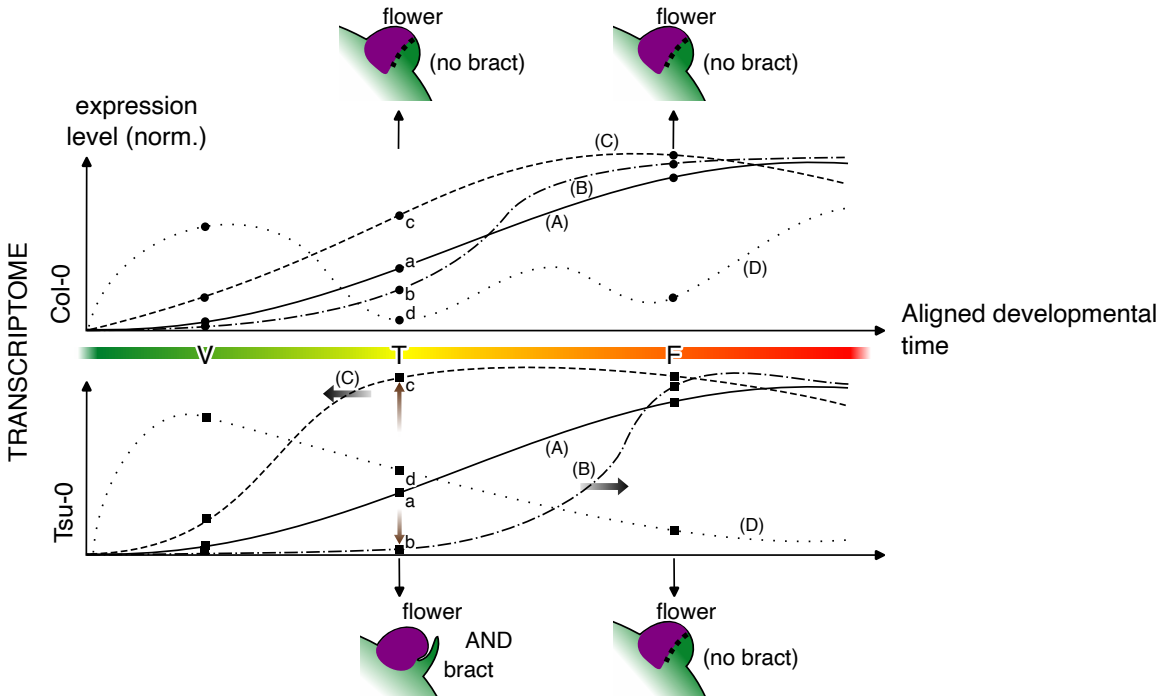


Figure 4.24: Desynchronisation of gene expression dynamics in *Tsu-0* creates a new gene expression state during the floral transition. Gene A is synchronised in both *Col-0* and *Tsu-0*, while genes B and C shift in timing, leading to different expression levels in *Tsu-0* at the transition stage (T). These shifts support flower and bract development in *Tsu-0*, but this state is temporary, with bract inhibition resuming as the meristem advances to the flowering stage (F). Gene D illustrates that not all differences are due to heterochronic shifts.

4.4 Discussion

Overall, this study explored the genetic and developmental processes behind bract formation, using natural variation in the presence of bracts at the base of flowering branches in Arabidopsis. The comprehensive phenotypic characterisations by Dieudonné *et al.* [1] revealed significant differences between these bracts and those found in known mutants, while also highlighting similarities with bracts observed in naturally bracteate species. By combining quantitative genetics, genomics, and transcriptomics in the *Tsu-0* accession, this study proposes new mechanisms controlling bract outgrowth.

My contribution to this research focused on the transcriptomic analysis, which included clustering processes to identify potential bract regulators, differential gene expression analysis across stages and accessions, and the application of the curve-registration method via *greatR* to pinpoint gene desynchronisations during the floral transition between *Tsu-0* and *Col-0*. These findings raise intriguing questions about the impact of these gene expression shifts on development and evolution, particularly regarding their influence on bract formation and loss.

4.4.1 Genetic mechanisms of basal bract development in *Tsu-0*

Dieudonné *et al.* [1] explored the genetic mechanisms underlying basal bract development in *Tsu-0* by employing quantitative genetics approaches, such as BSA and RIL, alongside transcriptomic analysis. They identified four major QTLs, with the two most significant ones located on chromosome 1. These QTLs were found to have additive effects on basal bract formation [1]. Despite identifying these regions, the high density of polymorphisms and the number of differentially expressed genes within these mapped intervals pose challenges in pinpointing specific candidate genes without further fine-mapping or a Genome-Wide Association Study (GWAS) on a larger accession panel.

Other analyses, to which I also contributed, revealed novel pathways potentially involved in bract development, including genetic interactions between *SOC1*, *TFL1*, and *PUCHI* (Figure 4.9), as well as pathways related to chloroplast function, metal ion homeostasis, anthocyanin biosynthesis and response to salicylic acid (Figure 4.13), as well as ribosome biogenesis (Figure 4.22). The study highlighted the role of basic metabolic and cellular function genes in controlling specific developmental processes, as seen in analogous pathways in other plant species [194].

Despite the remaining uncertainty about the causal genes and pathways described above, based on mutant analysis performed by Dieudonné *et al.* [1], their findings suggest the existence of a new bract developmental process distinct from the regained bracts observed in mutants. *Tsu-0* basal bracts display unique characteristics associated with wild-type flowers, wild-type bract shape and position, presence restricted to the base of the raceme, with no modification of plastochron rate, and independence from the light regime. These phenotypes differ from those reported in bract mutants, and none of the known "bract mutant genes" were identified within the QTL intervals.

This indicates that *Tsu-0* bract development may involve unique genetic mechanisms that do not interfere with floral meristem identity.

It is important to note that this particular study, including the identification of QTL intervals and analysis of bract mutants, was conducted by our collaborators (Dieudonné *et al.* [1]). While our work has provided crucial insights into the genetic basis of bract formation, there remains the possibility, as indicated by the transgressive indeterminism observed in some RILs [1], that some bract causal genes may also influence flower development. Genetic interactions could suppress these floral phenotypes while still allowing for the presence of basal bracts. This collaborative research highlights the complexity of these genetic pathways and suggests that further studies are needed to fully understand the mechanisms involved.

4.4.2 Transcriptomic heterochronies during floral transition

In the section of the study I also contributed to, we focused on transcriptomic analysis to understand the transient formation of bracts in *Tsu-0* at the base of each raceme. This trait is common in the *Brassicaceae* family, and it's been observed that some species naturally show variations at the base of the flowering branch, such as the presence of bracts or changes in flower structure. These basal nodes are formed during the floral transition, where the plant shifts from making leaves to producing flowers. This phase appears to be less strictly controlled, leading to more variation in the traits produced, such as bracts. Our data suggest that natural genetic differences between plants can lead to more frequent and varied traits at the base of branches, but no specific explanation has yet been proposed for why this reduced control, or "developmental canalisation," happens during the floral transition.

We propose that differences in the timing of gene expression, referred to as transcriptional heterochronies, could explain this phenomenon. In species without bracts, bract development is typically viewed as an extension of the vegetative phase, where a juvenile trait coexists with an adult trait [188, 195]. When we compared matching developmental stages between *Tsu-0* and *Col-0* for key floral transition and bract-related genes, we did not observe significant differences in the expression dynamics of these genes between the two accessions. We also attempted to identify potential bract-regulating genes by comparing gene expression between stages where bracts are present and absent, but no specific genes responsible for bract formation in *Tsu-0* were identified. However, when we analysed the entire transcriptome using PCA, we found that the floral transition in *Tsu-0* occurs within an older transcriptome, rather than a younger one (Figure 4.16).

Since *Tsu-0* plants flower later in absolute time, the genes related to flowering are delayed. Yet, many other genes fall out of sync with this delayed flowering. Some genes maintain their timing or shift earlier, while others shift later than the flowering time (Figure 4.18). This complex desynchronisation of gene expression during the floral transition was also previously observed in *B. rapa* and Arabidopsis [85], and appears to be a common occurrence across and within species.

Flowering time is a trait that undergoes strong selective pressure in plants [196, 197]. When the timing of flowering-related genes is constantly fine-tuned to adapt to environmental or genetic changes, it can lead to a desynchronisation with other genes that are not directly involved in flowering. This desynchronisation can disrupt the usual coordination of gene expression, potentially leading to the emergence of new or altered patterns of gene activity. Such shifts in gene expression may, in turn, give rise to transient developmental variations, like the formation of basal bracts, as the plant's developmental processes respond to these new genetic signals. This suggests that evolutionary changes in flowering time might unintentionally affect other parts of the plant, leading to unexpected traits like basal bracts.

When genes become desynchronised during development, it leads to a significant divergence in gene expression patterns, particularly during critical phases like the floral transition. This divergence has been observed in the *Solanaceae* family and is associated with the evolution of more complex flower arrangements [105]. This phenomenon parallels the inverse hourglass model proposed in animal development, where the middle stages of embryogenesis show greater variability across species than the early or late stages [105]. In contrast to the classic hourglass model where mid-development is highly conserved, the inverse hourglass model suggests that intermediate developmental stages are especially susceptible to evolutionary change, contributing to increased morphological diversity [105]. In plants, this means that during the floral transition, when gene expression is highly variable, even small shifts in gene timing can result in notable changes in plant structure and form. Consequently, the sensitivity of floral transition to these timing shifts, or transcriptional heterochronies, might play a crucial role in driving phenotypic evolution. This could influence not only the diversity within populations but also the divergence between species.

In conclusion, changes in the timing of gene expression, known as heterochrony, might be important in the evolution of bract formation and loss in plants like *Brassicaceae*. While traditional models for bract loss often focus on specific genes found in mutants, our findings suggest that shifts in gene timing could also play a role in both losing and reactivating bracts. This idea connects bract development in different plant families, such as *Brassicaceae* and *Poaceae*. Furthermore, the concept that lost traits cannot be regained, known as Dollo's law, might be challenged by the role of heterochrony in reactivating dormant developmental programs [198, 1]. Future research, especially involving comparisons between species that retain or have lost bracts, will be essential to further elucidate the genetic and developmental pathways that have shaped bract evolution in plants.

5 Exploring genetic variation in the floral transition between *B. rapa* and *B. oleracea* using comparative transcriptomics

5.1 Introduction

5.1.1 The importance and origins of *B. oleracea* and *B. rapa*

Brassica species are one of the most highly diverse and largest genera of plants, and they are cultivated worldwide as both horticultural and field crops. These plants have many important uses, including as vegetables, sources of oil and medicine, fodder, green manure, biofumigants, and spices [199, 200]. Brassicaceous vegetables include several species, such as *B. oleracea* (cabbage, broccoli, cauliflower, kale, Brussels sprouts, collard greens, and Chinese kale), *B. rapa* (turnip, napa cabbage, and turnip rape), *B. napus* (swede), and *B. juncea* (Indian mustard) [199, 201]. In 2022, the UK produced approximately 63,000 tonnes of broccoli and over 71,000 tonnes of cauliflower, highlighting the significance of these crops [202]. The significance of these Brassica vegetables extends beyond their vitamin and mineral content, as they also provide numerous beneficial plant secondary metabolites that contribute to human health [199].

The Brassica genus belongs to the tribe Brassiceae which is part of the Brassicaceae family [203]. In 1935, U established the foundational relationship between the most important Brassica species, known as the "Triangle of U" [204]. It was proposed here that three allotetraploids *B. napus* (AACC), *B. carinata* (BBCC), and *B. juncea* (AABB) originated from interspecific hybridisation of the diploid genomes of *B. nigra* (BB), *B. oleracea* (CC), and *B. rapa* (AA). In this chapter, we will focus on the diploid species *B. rapa* and *B. oleracea* as not only are they the most important cruciferous vegetable crops [205], but they are also closely related and diverged from the ancestor ∼4 MYA [206]. Despite their close relatedness, it has been reported that the two species' genomes have substantial size differences between their syntenic regions (about 26 Mb of transposable elements (TEs) in *B. rapa* versus about 88 Mb of TEs in *B. oleracea* [206]). Although several studies have compared these two species to understand different biological processes, such as DNA methylation-related genes [207], glucosinolate and phenolic compound content, antioxidant capacity [208], interspecific hybridisation [205], and general genomic comparisons [209], the impact of these structural genomic differences on the conservation of gene function across different developmental stages, such as flowering time, has not been extensively explored.

5.1.2 The significance and current knowledge of flowering time in model species Arabidopsis

During development plants switch from vegetative to reproductive growth. This stage is known as the floral transition [210]. Controlling the timing of this stage is essential for reproductive success, ensuring plants pollinate and seeds develop in favourable conditions. In plants grown as crops, this adaptation helps plants flower synchronously at the correct time to maximise yields [210, 211]. Research on flowering physiology and genetics, with Arabidopsis as a model plant, has revealed that the timing of the switch is influenced by various environmental and internal factors [210]. There are seven pathways known to affect flowering time in Arabidopsis: the photoperiod pathway, the vernalisation pathway, the autonomous pathway, the hormone pathway, the sugar pathway, the ambient temperature pathway, and the ageing pathway [212, 191]. These pathways trigger the expression of genes which are responsible for initiating the floral transition [210] (see Figure 5.1).

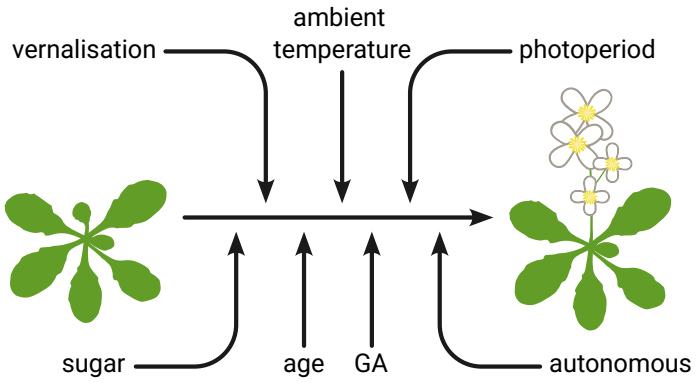


Figure 5.1: Floral transition is induced by environmental signals and endogenous factors. Adapted from [213, 191].

The findings of the molecular mechanisms of flowering extensively researched in the model plant Arabidopsis are compiled in FLOR-ID (Flowering Interactive Database), an interactive resource detailing the networks of flowering-time genes [191]. The Arabidopsis flowering-time networks consist of 306 genes distributed across the seven pathways [191]. Genes that participate in multiple pathways are referred to as "flowering-time integrators" [191]. These integrator genes, such as *FLOWERING LOCUS T* (*FT*) and *SUPPRESSOR OF OVEREXPRESSION OF CONSTANS 1* (*SOC1*), govern flowering time by merging signals from multiple pathways [214]. *FT* expression is inhibited by the transcriptional repressor *FLOWERING LOCUS C* (*FLC*) through the vernalisation pathway and promoted by the transcriptional activator *CONSTANS* (*CO*) through the photoperiod pathway and circadian clock [215]. In addition to the group of flowering-time integrator genes, there is another group of genes which switch the fate of the meristem from vegetative to floral (floral meristem identity genes) [216]. This group includes *LEAFY* (*LFY*), *APETALA1* (*AP1*), *CAULIFLOWER* (*CAL*), *APETALA2* (*AP2*), and *UNUSUAL FLORAL ORGANS* (*UFO*) [216, 217]. Floral meristem identity genes can influence

flowering time. For example, overexpressing *LFY* and *AP1* leads to the early formation of determinate floral meristems [216]. In contrast, mutations in *TERMINAL FLOWER1* (*TFL1*) affect both flowering time and meristem identity, leading to early flowering and transforming the inflorescence meristem into a determinate flower. This significantly reduces the number of flowers and branches produced in the inflorescence [216].

In most plants, including Brassica vegetables, the floral transition is a highly responsive developmental phase, particularly sensitive to environmental and endogenous cues. The regulation of this process is a key focus in plant breeding and adaptation strategies [218, 219].Knowledge about flowering time genes is crucial for improving Brassica crops, particularly vegetable varieties such as *B. oleracea* and *B. rapa* [215]. This effort is essential to reduce uncertainties in harvest time predictability and market availability [220], as flowering influences not only seasonal growth patterns but also many agronomic traits, including crop yield and quality. Variation in flowering time plays a significant role in shaping the diverse morphological forms found in cultivated brassicas, which are economically important [221]. Additionally, optimising flowering regulation can help minimise crop waste by reducing the risks of bolting, reduced produce quality, and post-harvest losses. Better control over flowering timing allows breeders to improve yield consistency, cut down on waste, and maximise resource efficiency throughout the entire production cycle.

5.1.3 Transfering knowledge of the floral transition from Arabidopsis to Brassica and its challenge

While flowering time is well-studied in Arabidopsis, our understanding of this process in Brassica species remains limited. Given the close relationship between Arabidopsis and Brassica species, as both belong to the Brassicaceae family [222, 221], Arabidopsis serves as an ideal model for studying flowering time in Brassica. However, transferring this knowledge between the two species comes with a challenge due to the genome multiplication events in Brassica evolution. As a result of these various duplication events (such as whole genome triplication, tandem duplication, and segmental duplication [223]), the Brassica genome contains multiple gene paralogues. In a polyploid organism like Brassicas, not all paralogues of a gene are equally important for its primary function. The presence of multiple copies reduces selection pressure, allowing mutations to occur with minimal phenotypic effects [224, 201]. Over time, these mutations can accumulate, causing genes to acquire new functions (neofunctionalisation), lose part of their original function (subfunctionalisation), or become completely non-functional [225]. Consequently, a major question to investigate is the extent to which gene copies have diverged and whether they retain the same function as their orthologous genes as it will influence the knowledge transfer from Arabidopsis, such as when considering a gene regulatory network of flowering time. The polyploidy in Brassica results in multiple copies not only in specific transcription factors but also of their regulators and target genes. This significantly expands the number of regulatory connections within the network [201]. The initial crucial step in addressing this challenge is to identify the specific functions performed by each gene copy. This involves determining which

paralogues have become redundant and which have undergone neo- or subfunctionalisation during flowering time, which is essential for simplifying the complexity. Identifying genes that have retained their original function will be an important step in using existing knowledge from Arabidopsis to construct gene regulatory networks in Brassica. Previous comparative studies on flowering time between Brassicas and Arabidopsis have demonstrated that they share similar gene content, as well as similar functions and regulatory networks [85, 215, 226]. However, similar studies focusing on the conservation of gene content and function related to flowering time within the Brassica genus, particularly in *B. rapa* and *B. oleracea*, have not yet been explored. Conducting this study between these two species will improve our understanding of how gene expression dynamics differences, despite their close relatedness, impact the regulation and evolution of flowering time. This insight could have significant implications for breeding programs and agricultural practices, as flowering time is a critical trait for crop yield and adaptation to different environments.

5.1.4 Hypothesis and aims

This chapter aims to investigate the regulation of the floral transition in *B. rapa* and *B. oleracea* by comparing their gene expression dynamics with those of Arabidopsis. This comparison will be done using the approach which we developed for the time series expression data described in Chapter 2. Using a similar method, it was previously observed that in *B. rapa* [85] and *B. oleracea* [201], most of the gene expression dynamics are similar to those gene dynamics in Arabidopsis. However, one of the disadvantages of the previous method is the inability to find optimal parameters. In this section, using the updated technique we developed, we will re-register these datasets. We hypothesise that using the new updated approach, more genes will be identified to be similar due to the ability of the approach to find the optimal sets of parameters. This potentially helps in finding the potential similar genes between two Brassica cultivars and Arabidopsis which have not been identified before.

Due to gene duplication, multiple copies of genes can either accumulate deleterious mutations and become nonfunctional, or acquire beneficial mutations leading to new functions or subfunctionalisation [227]. As two important diploid crops, it is crucial to understand the genetic variation associated with flowering time in *B. rapa* and *B. oleracea*. When compared to Arabidopsis, we hypothesise that *B. oleracea* retains a higher number of flowering time gene copies than *B. rapa*, likely due to the presence of transposons that promote gene retention. In addition to comparing Arabidopsis with *B. rapa* and Arabidopsis with *B. oleracea*, a transcriptomic comparison between *B. rapa* and *B. oleracea* will be explored. We hypothesise that most of the flowering time genes in *B. rapa* and *B. oleracea* exhibit similar dynamics. However, for those genes that do not exhibit the same dynamics, we hypothesise that subfunctionalisation or neofunctionalisation may have occurred, potentially contributing to phenotypic differences in flowering time regulation between the two species.

5.2 Materials and methods

5.2.1 Gene expression time series data

The gene expression data *B. rapa* cv. *R-o-18* from Calderwood *et al.* [85] was used for the analysis of this chapter. Plants were grown in cereal mix (40% medium grade peat, 40% sterilised soil, 20% horticultural grit, 1.3 kg/m3 PG mix 14-16-18 + Te base fertiliser, 1 kg/m3 Osmocote Mini 16-8-11 2 mg + Te 0.02% B, wetting agent, 3 kg/m3 maglime and 300 g/m3 Exemptor). The material was grown in a Conviron MTPS 144 controlled environment room with Valoya NS1 LED lighting (250 μmol m-2 s-1) 18 °C day/15 °C night, 70% relative humidity with a 16-hr day [85]. Sampling of the plant apex was performed 10 hr into the day.

For *B. oleracea*, we used *B. oleracea* cv *DH1012* collected and analysed by Woodhouse [201]. The seeds from *B. oleracea* cv *DH1012* were grown and sampled under the exact same growth (16-hr day) conditions as the *B. rapa* cv *R-o-18* plants [85]. For both *B. rapa* and *B. oleracea*, apex samples were taken over development during the vegetative growth and the floral transition, continuing until floral buds were visible (developmental stage BBCH51) [228]. For *B. rapa* and *B. oleracea*, BBCH51 was reached at 35d and 51d post-sowing, respectively. At each time point, three replicated samples were collected. All preprocessing, including read alignment and transcript quantification, was carried out by the original authors of each dataset and not repeated as part of this study.

For the model species, publicly available gene expression data in Arabidopsis *Col-0* shoot apex from 7 to 16 days after germination (see Figure 5.2) grown under similar 16-hr day conditions were downloaded from NCBI SRA, project ID PRJNA268115 [148]. Gene expression levels were quantified using the previously published approach HISAT v2.0.4 [18] and StringTie v1.2.2 [229]. *Brassica rapa* reads were aligned to Chiifu v3 reference genome [230, 85], *B. oleracea* reads were aligned to both pan transcriptome [231] and the Brasssica pan genome [232], and Arabidopsis reads were aligned to the TAIR10 reference genome [233, 85]. Gene expression level is reported in Counts per Million (CPM). Orthologues of Arabidopsis genes in *B. rapa* were identified by A. Calderwood and H. Woolfenden, while those in *B. oleracea* were identified by H. Woolfenden. These orthologues were determined based on sequence similarity and gene synteny, using either the SynOrths tool [234] or BLAST [235]. To ensure confidence in orthologue assignments, only matches with greater than 95% sequence identity to Arabidopsis were retained. Although sequence similarity between Brassica paralogues was not directly assessed, the high-confidence mapping to Arabidopsis orthologues provides strong support for the reliable identification of Brassica paralogues used in downstream expression analysis.

To simplify the discussion of this chapter, *B. rapa* cv *R-o-18* will be referred to *B. rapa*, *B. oleracea* cv *DH1012* as *B. oleracea*, and Arabidopsis *Col-0* as Arabidopsis hereafter.

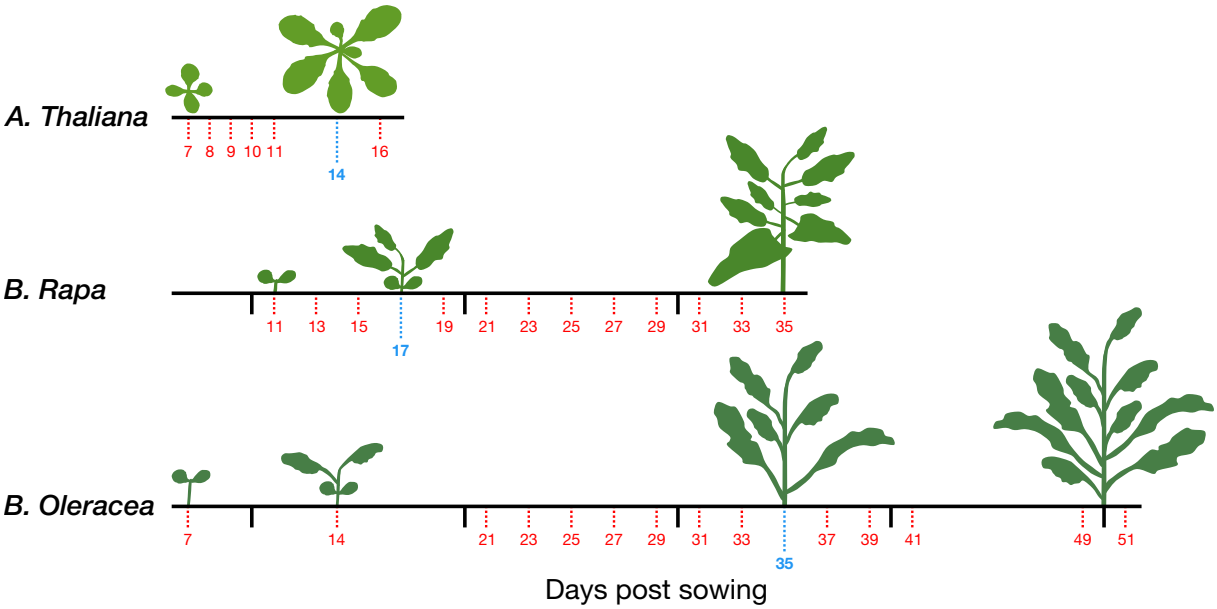


Figure 5.2: Schematic illustration of the developmental stages of Arabidopsis, *B. rapa*, and *B. oleracea*. Red dotted lines and numbers displayed below the bottom axis represent days post-sowing at which apex tissues were sampled, and blue dotted lines and numbers represent the day of floral transition. The representations of the plants indicate the approximate number of leaves on the plants at the indicated time points.

5.2.2 Registration

The registration process was carried out with the *greatR* [3] package, with the default parameter optimisation method LBFGSB (`optimisation_method = "lbfgsb"`), 90% overlapping to the reference data (`overlapping_percent = 90`), and z-score scaling (`scaling_method = "z-score"`). The pairwise registration result was visualised with the `plot()` function from *greatR*. Additionally, the sum of gene expression per time point was calculated for the registration of the total gene expression.

5.2.3 The distance calculation between samples

The dist function which was implemented in *greatR* [3] was used to calculate the Euclidean distance between transcriptomes. Only the flowering time genes for each Arabidopsis, *B. rapa*, and *B. oleracea* were used for the analysis in this chapter.

5.2.4 Gene regulatory networks inference

To validate the regulatory function of each copy of genes which were identified using curve registration via *greatR*, we generated gene regulatory networks for different sets of flowering time genes. To generate the networks, the likelihood of regulatory links between genes was inferred

using CSI v1.0, using a pipeline implemented by Calderwood *et al.* [85]. The gene networks were visualised using R package `ggnetwork` [236] and `ggplot2` [153].

5.3 Results

5.3.1 Copy number variation of flowering time genes between *B. rapa* and *B. oleracea*

According to the FLOR-ID database [191], there are 306 genes involved in flowering time in Arabidopsis. The corresponding orthologues of these genes were identified (see Methods 5.2.1), revealing 487 genes in *B. rapa* and 524 genes in *B. oleracea*. Additionally, 20 Arabidopsis genes are not present in either *B. rapa* or *B. oleracea*. The detailed list of genes and their associated pathways is provided in Table 5.1. Among the genes listed above, some miRNAs are absent in the genome assemblies of *B. rapa* and *B. oleracea*. This absence is likely due to the read alignment protocols, which were not designed to capture miRNAs. Consistent with previous findings [197], no orthologues were detected for several genes related to photoperiodism/light perception and signalling pathways, including *CRYPTOCHROME-INTERACTING BASIC-HELIX-LOOP-HELIX 5* (*CIB5*), *CALCIUM-DEPENDENT PROTEIN KINASE 33* (*CPK33*), *FLAVIN-BINDING-KELCH REPEAT-F BOX 1* (*FKF1*), and *SCHNARCHZAPFEN* (*SNZ*). In the circadian clock pathway, *ZTL* was the only gene found without any orthologue, a finding that Li *et al.* [197] also reported. Additionally, we identified several genes that are absent in both *B. rapa* and *B. oleracea* and have not been previously reported. These include *DAY NEUTRAL FLOWERING* (*DNF*) (photoperiodism/light perception and signalling pathways), *HUA2 LIKE 2* (*HULK2*) (general pathway), as well as *HEXOKINASE 1, GLUCOSE INSENSITIVE 2* (*HXK1*) and *SUCROSE SYNTHASE 4* (*SUS4*) (sugar pathway).

Gene name	Gene details	Pathway
CIB5	AT1G26260	Photoperiodism, light perception and signalling
CPK33	AT1G50700	Photoperiodism, light perception and signalling
DNF	AT3G19140	Photoperiodism, light perception and signalling
FKF1	AT1G68050	Photoperiodism, light perception and signalling
HULK2	AT2G48160	General
HXK1	AT4G29130	Sugar
MIR156A	AT2G25095	Aging
MIR156B	AT4G30972	Aging
MIR156C	AT4G31877	Aging
MIR156D	AT5G10945	Aging
MIR156E	AT5G11977	Aging
MIR156F	AT5G26147	Aging
MIR156G	AT2G19425	Aging
MIR156H	AT5G55835	Aging
MIR172A	AT2G28056	Aging Photoperiodism, light perception and signalling
MIR172B	AT5G04275	Aging Photoperiodism, light perception and signalling
MIR172C	AT3G11435	Aging Photoperiodism, light perception and signalling
SNZ	AT2G39250	Aging Photoperiodism, light perception and signalling
SUS4	AT3G43190	Sugar
ZTL	AT5G57360	Circadian Clock Photoperiodism, light perception and signalling

Table 5.1: List of the genes in Arabidopsis which have no copy in either *B. rapa* or *B. oleracea*.

For the genes that have copies in both *B. rapa* and *B. oleracea*, three genes do not have available data in our gene expression datasets. These genes include *CYCLING DOF FACTOR 4* (*CDF4*), *GA2-oxidase 3* (*GA2ox3*), and *WRKY DNA-BINDING PROTEIN 34* (*WRKY34*). *CDF4* is known to reduce *CONSTANS* (*CO*) expression and is responsible for a photoperiodic flowering response [237]. *GA2ox3* is one of the *GA2-oxidase* genes which regulate the deactivation of bioactive gibberellins, which play multiple roles in plant development and stress response [219]. *WRKY34* is known as a key transcription factor that negatively regulates the cold sensitivity of mature Arabidopsis pollen [238]. However, the absence of these three gene copies in our *B. rapa* and *B. oleracea* datasets may be due to read mapping issues. These issues could arise from incomplete genome coverage, sequence divergence, or the quality of sequencing data, potentially causing reads to fail to align properly to the reference genome.

Out of 306 FLOR-ID flowering time genes, 273 have orthologues in both *B. rapa* and *B. oleracea*, 23 have orthologues only in *B. oleracea*, and 9 only in *B. rapa*. Table 5.2 lists the genes present exclusively in either *B. rapa* or *B. oleracea*, highlighting differences in copy number variations. Most genes with copies only in *B. oleracea* are involved in general flowering time functions, as well as photoperiodism, light perception, and signalling pathways. Specifically, two genes are asso-

ciated with vernalisation (*FRIGIDA LIKE 1* (*FRL1*) and *MADS AFFECTING FLOWERING 3* (*MAF3*)), two with the circadian clock pathway (*LUX ARRHYTHMO* (*LUX*) and *LIGHT-REGULATED WD 2* (*LWD2*)), and one with the sugar pathway (*SUCROSE-PROTON SYMPORTER 9* (*SUC9*)). Fewer genes were found to have copies only in *B. rapa* compared to *B. oleracea*. These genes are mostly involved in the general flowering time pathway, with exceptions such as *GIBBERELLIN 2-OXIDASE 7* (*GA2ox7*) (hormones), *MADS AFFECTING FLOWERING 3* (*MAF3*) (vernalisation), and *AT-STUbl4* and *PHYTOCHROME D* (*PHYD*) (photoperiodism, light perception, and signalling).

Gene name	Gene details	Copy no. in *B. rapa*	Copy no. in *B. oleracea*	Pathway
AHL22	AT2G45430	0	2	General
ASH2R	AT1G51450	0	1	General
ATC	AT2G27550	0	1	Photoperiodism, light perception and signaling
BFT	AT5G62040	0	1	General
ELF4	AT2G40080	0	4	Circadian Clock Photoperiodism, light perception and signaling
ELF7	AT1G79730	0	1	General
FBH2	AT4G09180	0	1	Photoperiodism, light perception and signaling
FPF1	AT5G24860	0	2	Hormones
FRL1	AT5G16320	0	2	Vernalisation
FTIP1	AT5G06850	0	1	Photoperiodism, light perception and signaling
FWA	AT4G25530	0	1	General
LDL1	AT1G62830	0	1	General
LUX	AT3G46640	0	2	Circadian Clock
LWD2	AT3G26640	0	1	Circadian Clock
MAF3	AT5G65060	0	2	Vernalisation
NF-YB3	AT4G14540	0	3	Photoperiodism, light perception and signaling
NF-YC3	AT1G54830	0	1	Photoperiodism, light perception and signaling
RGA1	AT2G01570	0	2	Hormones
RGL1	AT1G66350	0	1	Hormones
SUC9	AT5G06170	0	1	Sugar
VIM2	AT1G66050	0	1	General
VIM3	AT5G39550	0	1	General
VIP5	AT1G61040	0	2	General
AT-STUbl4	AT1G66650	1	0	Photoperiodism, light perception and signaling
ATJ3	AT3G44110	1	0	General
AtBMI1C	AT3G23060	1	0	General
ELF8	AT2G06210	1	0	General
GA2ox7	AT1G50960	2	0	Hormones
JMJ15	AT2G34880	1	0	General
MAF5	AT5G65080	1	0	Vernalisation
MRG1	AT4G37280	1	0	General
PHYD	AT4G16250	1	0	Photoperiodism, light perception and signaling

Table 5.2: List of genes which are only present (in different copy numbers) in *B. rapa* and *B. oleracea*.

A total of 487 genes are present in *B. rapa* and 524 in *B. oleracea*, with the distribution of copy number variations illustrated in Figure 5.3. Most Arabidopsis genes have more than one copy in *B. rapa* and *B. oleracea*, including key floral integrator genes: *SOC1* (3 copies in both *B. rapa* and *B. oleracea*), *AP1* (2 and 3 copies in *B. rapa* and *B. oleracea*, respectively), and *LFY* (2 copies in both *B. rapa* and *B. oleracea*). Approximately 100 genes have only one copy.

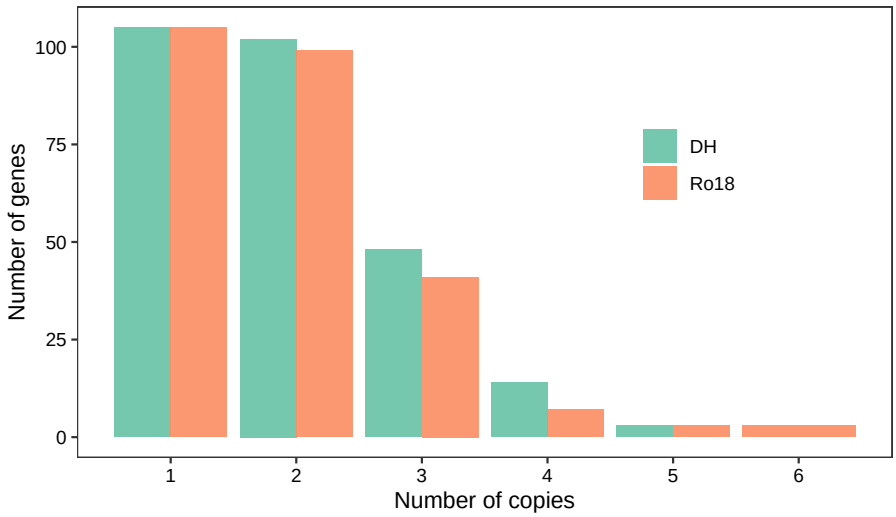


Figure 5.3: Copy number variation of Arabidopsis flowering time genes in *B. rapa* and *B. oleracea* (in green and orange colour, respectively).

5.3.2 Registration results show that the developmental progression to flowering in two Brassicas and Arabidopsis is similar

The illustrated diagram in Figure 5.2 shows the difference in developmental stages between Arabidopsis, *B. rapa*, and *B. oleracea*. Although these species were sampled at the matching morphological developmental stage (according to the BBCH system), the timing of their progression is desynchronised. To investigate whether desynchronisation of gene expression profiles explains the differences in transcriptomic gene expression between Arabidopsis and both *B. rapa* and *B. oleracea*, curve registration via *greatR* was employed. Table 5.3 shows a summary of the registration results between Arabidopsis and both *B. rapa* and *B. oleracea*. Of the 30,612 *B. oleracea* genes analysed using curve registration, approximately 80% were found to have similar dynamics to their ortologues in Arabidopsis. This is an improvement over the previously reported result of around 60% [201] registered genes. Furthermore, among the *B. oleracea* genes identified to have ortologues in Arabidopsis, approximately 84% of the 524 flowering time-related genes were successfully registered.

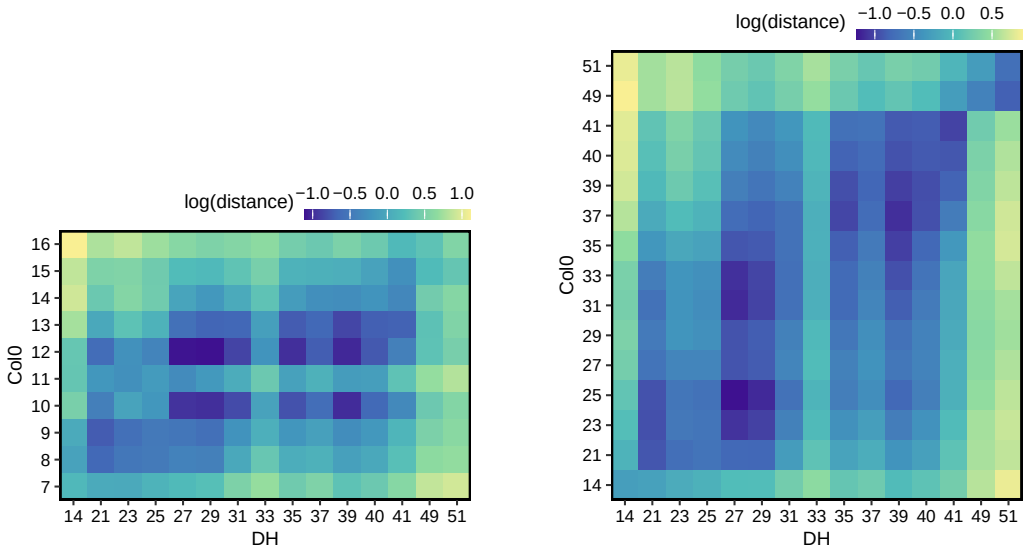
These results support our hypothesis and align with the previous finding [201], indicating that the differences in gene expression between *B. oleracea* and Arabidopsis are due to timing discrepancies rather than differences in the expression profiles. This is further visualised in the heatmap

in Figure 5.3. After registration (Figure 5.3 (b)), we observed closer distances between nearby time points between Arabidopsis and *B. oleracea*, highlighting a common progression from early to late gene expression states. This suggests that most gene expressions are more similar between Arabidopsis and *B. oleracea* after registration. Such similarity would not be evident through a naive comparison (Figure 5.3 (a)).

Similarly, we also explored whether the differences in gene expression profiles between Arabidopsis and *B. rapa* are due to desynchronisation. From Table 5.3, among the 27856 *B. rapa* genes analysed, approximately 87% exhibited similar dynamics to their paralogues in Arabidopsis. This finding also exceeds the previously reported number of registered genes of around 62% [85]. In addition to this, about 82% of the total 524 flowering time-related genes were successfully registered. Similar to the results observed in *B. oleracea* and Arabidopsis, this result also indicates that the differences in *B. rapa* and Arabidopsis are primarily due to timing variations rather than inherent differences in expression profiles. The heatmap in Figure 5.5 (b) shows that post-registration there are reduced distances between corresponding time points of Arabidopsis and *B. rapa*. This also implies that gene expression patterns between Arabidopsis and *B. rapa* become more aligned after registration, which would not be apparent if the distances were measured without any registration, only using scaled expressions (Figure 5.5 (a)).

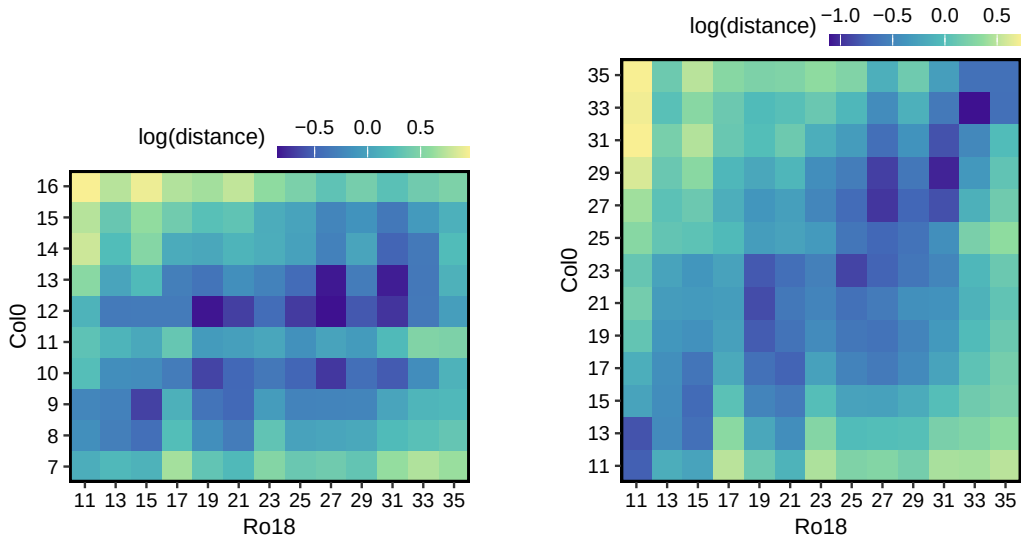
Comparison	Genes	Total genes	Registered genes
Arabidopsis vs *B. oleracea*	all	30 612	24 779 (80.9%)
Arabidopsis vs *B. rapa*	all	27 856	24 259 (87.1%)
Arabidopsis vs *B. oleracea*	flowering	524	439 (83.8%)
Arabidopsis vs *B. rapa*	flowering	487	390 (80.1%)

Table 5.3: Registration results of Arabidopsis vs *B. rapa* and Arabidopsis vs *B. oleracea* for both all genes and flowering time related genes.



(a) **(b)**

Figure 5.4: Heatmaps showing the gene expression distance of samples taken from Arabidopsis and *B. oleracea* over time since germination. The gene expression distance is measured using the average squared difference between homologous gene pairs. (a) The heatmap displays the distance measured from the scaled expressions. While similarities between developmental time points are observed, there is no indication of a correlation between species samples at closer time points. (b) The heatmap shows the distance measured from the expressions after registration. A darker purple diagonal appears in the heatmap, indicating that curve registration effectively resolves differences in gene expression, suggesting that these differences are more likely due to desynchronisation rather than distinct expression patterns.



(a) **(b)**

Figure 5.5: Heatmaps showing the gene expression distance of samples taken from Arabidopsis and *B. rapa* over time since germination. The gene expression distance is measured using the average squared difference between homologous gene pairs. (a) The heatmap displays the distance measured from the scaled expressions. Similar to what was observed in *B. oleracea*, while the similarities between developmental time points are observed, there is no indication of correlation between species samples at closer time points. (b) The heatmap shows the distance measured from the expressions after registration. A darker purple diagonal appears in the heatmap, indicating that curve registration effectively resolves differences in gene expression, suggesting that these differences are more likely due to desynchronisation rather than distinct expression patterns.

5.3.3 Comparison between Arabidopsis to the two Brassica species shows species-specific expression

We investigated the extent of conservation of both *B. rapa* and *B. oleracea* to Arabidopsis by analysing the similarity between gene expression ortologues in both Brassicas and Arabidopsis. This analysis helps to understand the evolutionary relationships in gene expression, identify potential conserved genetic functions, and provide insights into the genetic basis of flowering time shared among these species. Figure 5.6 provides a schematic diagram illustrating the comparison methodology. Each species was compared to the others, resulting in several defined cases based on the similarity of their dynamics.

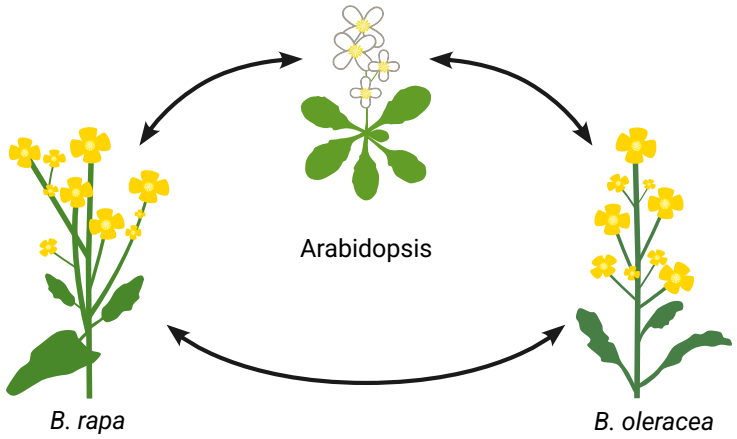


Figure 5.6: Schematic diagram showing the comparison between two Brassica cultivars *B. rapa* and *B. oleracea* and the model species Arabidopsis. The arrows represent cross-species comparisons of gene expression dynamics for selected orthologous genes. These comparisons were performed using curve registration, which aligns expression profiles across species to account for differences in developmental timing.

We hypothesised that the dynamics of both *B. rapa* and *B. oleracea* are similar to Arabidopsis, and they are also similar to each other within Brassica, which indicates evolutionary retention. However, if *B. rapa* and *B. oleracea* exhibit similar dynamics to each other but not to Arabidopsis, this suggests a Brassica-specific case. Finally, if either *B. rapa* or *B. oleracea* displays dissimilar dynamics to both Arabidopsis and the other Brassica species, this points to a potential Brassica species-specific scenario. For further details, refer to Table 5.4 below.

Arabidopsis vs *B. rapa*	Arabidopsis vs *B. oleracea*	*B. rapa* vs *B. oleracea*	Case
✓	✓	✓	Conserved
✗	✗	✓	Brassica specific
✗	✓	✗	*B. rapa* specific
✓	✗	✗	*B. oleracea* specific

Table 5.4: Potential cases based on gene expression similarity when comparing each orthologous pair among Arabidopsis, *B. rapa*, and *B. oleracea*.

Investigating the expression of flowering time genes in *B. rapa* and *B. oleracea* which have a single orthologue of those in Arabidopsis

From all orthologues of Arabidopsis identified in *B. rapa* and *B. oleracea*, there is a high percentage of genes which exclusively exist as singletons. From Paterson *et al.* [239], single-copy genes are coined as "duplication-resistant" genes and could be important to the long-term survival of polyploid lineage. De Smet *et al.* [240] further investigated the possibility that such a pattern could be explained by random gene loss only and therefore propose that there is selection pressure to preserve such genes as singletons. Regardless of the mechanisms that maintain these genes as single copies, they likely play significant roles, as the phenotypic effects of mutations in one copy of a duplicated pair are notably smaller compared to those observed in singleton genes [241].

Table 5.5 lists various cases based on the registration results of single-copy genes among the three species: Arabidopsis, *B. rapa*, and *B. oleracea*. Our results reveal more complexity than initially hypothesised, uncovering additional cases. One notable case is lineage-specific divergence, where gene expression profiles are registered between *B. rapa* versus Arabidopsis and *B. oleracea* versus Arabidopsis, but not between *B. rapa* versus *B. oleracea*. Another observed case is the Brassica-predominant divergence, where curve registration shows that gene expression profiles are the same between pairs within Brassica species but differ from one of those Brassica when compared to Arabidopsis.

Approximately 71% of the 72 single-copy genes show conserved gene expression between Arabidopsis and Brassicas, as curve registration revealed that the expression profiles for each gene pair across species were identical. Among these conserved genes are *TRITHORAX-LIKE PROTEIN 2* (*ATX2*), *GA REQUIRING 2* (*GA2*), and *CIRCADIAN CLOCK ASSOCIATED 1* (*CCA1*). *ATX2* is a histone methylation enzyme that regulates *FLC* and *FT* [242]. *GA2* (or *ATKS1*) is involved in gibberellin biosynthesis, and a single mutation in this gene results in late flowering under both short and long days [243]. *CCA1* plays a key role in the circadian clock, and its overexpression leads to circadian rhythm disruption, extended hypocotyl growth, and delayed flowering [244].

Case	Count	Percentage (%)
Conserved	51	70.83
Lineage-specific divergence	5	6.94
Brassica specific	2	2.78
B. rapa specific	2	2.78
Else (Brassica pre-dominant)	12	16.67

Table 5.5: Summary of the analysis results on the conservation of single-copy gene expression through registration between three species Arabidopsis, *B. rapa*, and *B. oleracea*.

Approximately five genes are best categorised as "lineage-specific divergence," as their expression patterns are similar between either Brassica species and Arabidopsis but differ within the Brassica species themselves. These genes show divergence specifically within the Brassica lineage while maintaining some ancestral expression characteristics seen in Arabidopsis. This suggests that although these genes retain certain traits from their common ancestor (as observed in Arabidopsis), they potentially have undergone changes within the Brassica genus that make their expression patterns unique among Brassicas. However, it is also possible that conserved expression patterns exist between the Brassica species but are not readily detectable due to limitations in expression resolution, variability across replicates, or subtle differences falling below statistical thresholds.

The genes identified in this category include *NODULIN HOMEOBOX* (*AtNDX*), *DICER-LIKE 3* (*DCL3*), *FLOWERING LOCUS D* (*FLD*), *GIBBERELLIN 2-OXIDASE 4* (*GA2ox4*), and *GA INSENSITIVE DWARF 1C* (*GID1C*). Figure 5.8 shows the registration results of *AtNDX*, *FLD*, and *GA2ox4*.

AtNDX is known to regulate *COOLAIR*, an antisense *FLC* transcript [245]. *FLD*, part of the autonomous pathway, mediates histone H3 Lys-4 demethylation at the *FLC* locus and acts in partial redundancy with *LSD1-LIKE 1* (*LDL1*) and *LSD1-LIKE 2* (*LDL2*) genes to repress *FLC* expression [191]. *GA2ox4* is reported to be the main GA2 oxidase involved in controlling flowering time [246]. *GID1C* functions as a receptor for gibberellins (*GA*s), essential hormones that regulate growth and development in plants [191]. Notably, these "lineage-specific divergence" genes are primarily associated with the vernalisation pathway and gibberellins.

Two genes are classified as Brassica-specific because they exhibit similar expression patterns within Brassica species but differ from those in Arabidopsis. These genes are *HOMOLOGUE OF TRITHORAX 1* (*ATX1*) and *PSEUDO-RESPONSE REGULATOR 3* (*PRR3*). *ATX1* (also known as *SDG27*) is a histone methylation enzyme that regulates *FLC*, *FT*, and *AG*, similar to *ATX2* [191, 242]. In rice and maize, the ortologues of *ATX1* are known to perform similar functions [247, 248]. Although the specific function of *ATX1* in Brassica remains unclear [249, 250], its expression patterns suggest it likely serves a comparable role in both *B. rapa* and *B. oleracea*, as evidenced by their similar expression dynamics (Figure 5.9 *ATX1*).

In Arabidopsis, *PRR3* is a member of the *PSEUDO-RESPONSE REGULATOR* (*PRR*) family and plays a role in regulating the photoperiodic flowering response [191]. In *B. rapa*, *PRR3* is predicted to be nonfunctional due to genetic rearrangements and partial deletions [251, 252]. This likely accounts for the observed differences in expression patterns between *B. rapa* and Arabidopsis, as the orthologous genes are not registered. Given that, *PRR3* expression is registered between *B. rapa* and *B. oleracea* (see Figure 5.9 *PRR3*), it is potentially that *PRR3* is also nonfunctional in *B. oleracea*. However, a comparison of the protein sequences is required to confirm this.

We found that the genes *HISTONE DEACETYLASE 9* (*HDA9*) and *HOMOLOG OF YEAST YAF9 A* (*YAF9A*) fall into the *B. rapa* specific category, where their *B. rapa* dynamics do not show similarity to either *B. oleracea* or Arabidopsis, while their *B. oleracea* expression exhibits a similar dynamic to Arabidopsis, as shown in Figure 5.10. In Arabidopsis, *HDA9* plays a role in inhibiting *FT* expression under short-day (SD) conditions [191]. Although much is still unknown about *HDA9*'s role in other plants [253], a study by Wang *et al.* [254] found that *HDA9* in *B. oleracea* does not interact directly with *AGL24* and *SOC1*, as it does in *B. juncea* [247]. Instead, a study in a Chinese cabbage cultivar of *B. rapa* by Ma *et al.* [255] speculated that *HDA9* may have a similar mechanism to that in *B. juncea*. These findings support our results (shown in Figure 5.10) as *HDA9* is identified as a *B. rapa*-specific gene that potentially interacts with *AGL24*, as speculated in Chinese cabbage and known in *B. juncea*, but differs from *B. oleracea*. Since *B. oleracea* has a similar expression pattern to Arabidopsis, it is likely that *HDA9* in Arabidopsis also does not interact with *AGL24* and *SOC1*. This is further supported by findings that *HDA9* in Arabidopsis only interacts with *AGL19*, a known transcriptional activator of *FT* [191, 256].

Another gene in the *B. rapa*-specific category is *YAF9*. In Arabidopsis, *YAF9* binds to *FLC* chromatin and regulates *FLC* expression by modulating the acetylation levels of *H2A.Z* and *H4* [191, 257]. Although the role of this gene in Brassica is not yet fully understood, our observations of expression similarity (see Figure 5.10 for *YAF9*) suggest that *YAF9* in *B. oleracea* may function similarly to its role in Arabidopsis but differs from the copy in *B. rapa*.

For genes that show similar expression patterns within Brassica species but exhibit variability compared to Arabidopsis, we defined this case as "Brassica-predominant". This classification applies to gene pairs that maintain similar expression profiles across Brassica species, indicating a predominant expression pattern unique to this genus. At the same time, there may also be variability or divergence in expression when compared to Arabidopsis. Among the genes identified in this category are *CLEAVAGE STIMULATING FACTOR 77* (*CSTF77*) and *TARGET OF EARLY ACTIVATION TAGGED 3* (*TOE3*) (Supplementary Table S.4). *CSTF77* is known to be essential for the 3-end processing of *FLC* antisense transcripts in Arabidopsis [258]. *TOE3*, along with *AP*, plays a significant role in repressing *AG* expression in Arabidopsis [259]. However, the specific functions of these genes within Brassicas remain unexplored. According to our analysis, the expression of *CSTF77* and *TOE3* genes is consistent between *B. rapa* and *B. oleracea*

(Figure 5.11), suggesting that their roles might be similar across these Brassica species. However, their functions could potentially differ from those observed in Arabidopsis.

It is important to note that the observed lineage-specific divergence and Brassica-predominant expression patterns could be influenced by noise or errors during sequencing or sampling. Despite these potential sources of error, the consistent similarity in expression patterns between gene pairs in Arabidopsis and Brassicas suggests that some genes may indeed exhibit true lineage-specific divergence and Brassica-predominant expression.

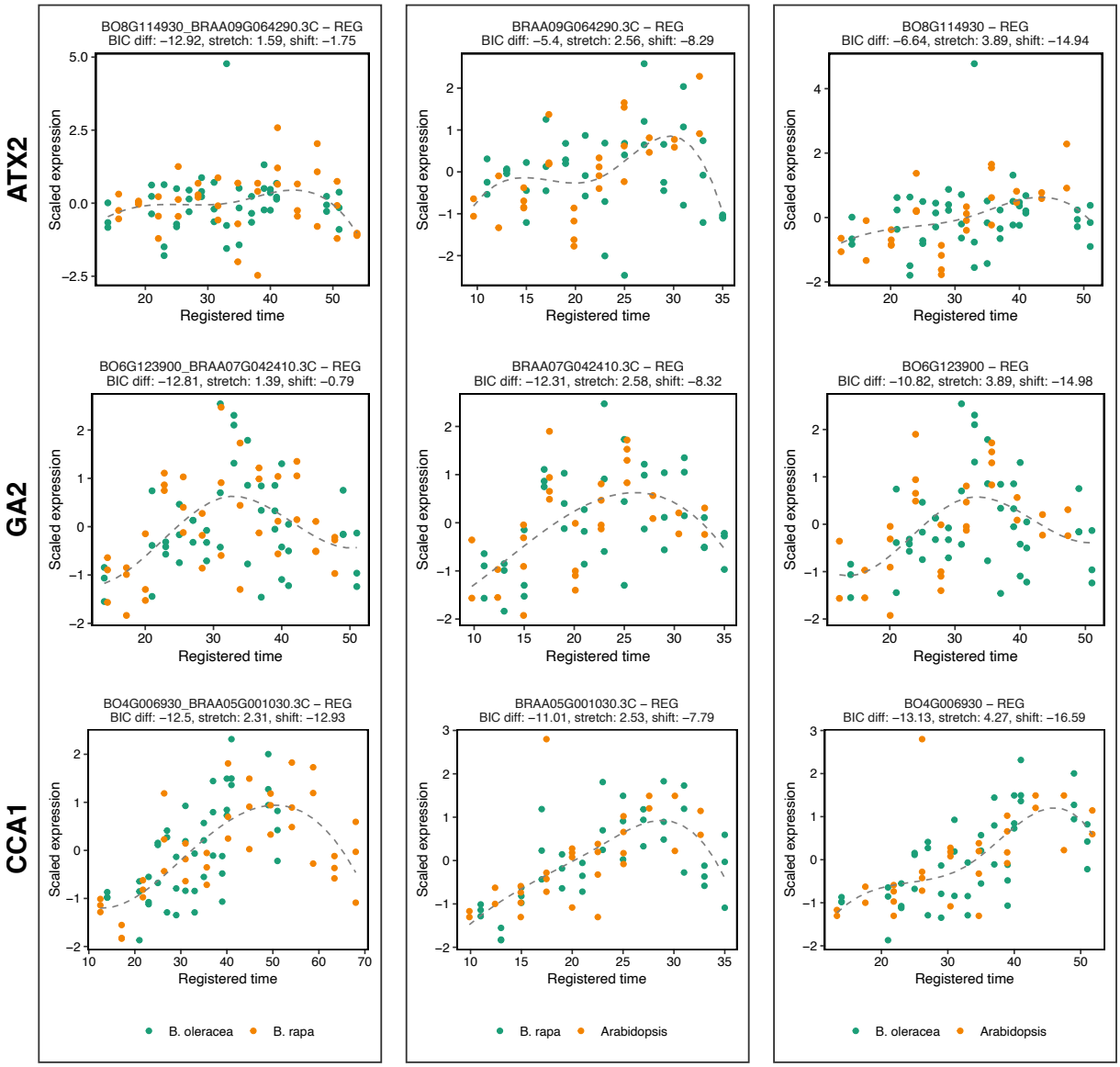


Figure 5.7: *ATX2*, *GA2*, and *CCA1* are among the genes with conserved expression across all three species. Each panel shows the pairwise comparison between species, from left to right: *B. rapa* vs. *B. oleracea*, *B. rapa* vs. Arabidopsis, and *B. oleracea* vs. Arabidopsis. Each dot represents a gene expression replicate at each time point. Green and orange colours denote gene expression for the species indicated in the legend below each plot. A grey dashed line indicates that the two profiles are registered as a single model fitted to both gene expression profiles. The subtitle in each plot provides the gene ID and the optimal stretch and shift parameters for each pair of profiles.

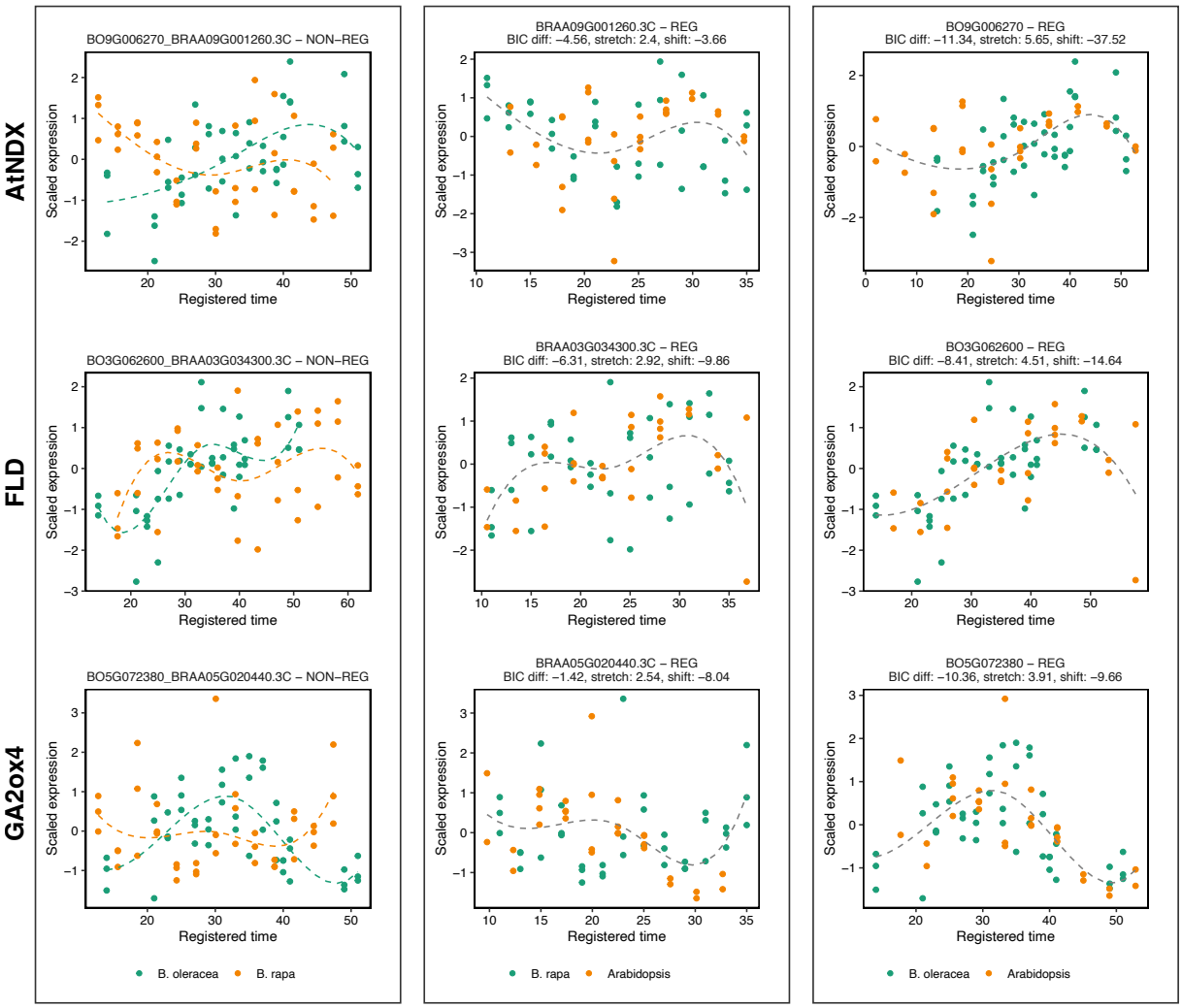


Figure 5.8: *AtNDX*, *FLD*, and *GA2ox4* are categorised as "lineage-specific divergence" genes. Each panel shows the pairwise comparison between species, from left to right: *B. rapa* vs. *B. oleracea*, *B. rapa* vs. Arabidopsis, and *B. oleracea* vs. Arabidopsis. Each dot represents a gene expression replicate at each time point. Green and orange colours denote gene expression for the species indicated in the legend below each plot. A grey dashed line indicates that the two profiles are registered as a single model fitted to both gene expression profiles. If no grey dashed line is present, the pair is not registered, with each profile fitted separately (green and orange dashed lines corresponding to the species indicated in the legend). The subtitle in each plot provides the gene ID and the optimal stretch and shift parameters for each pair of profiles.

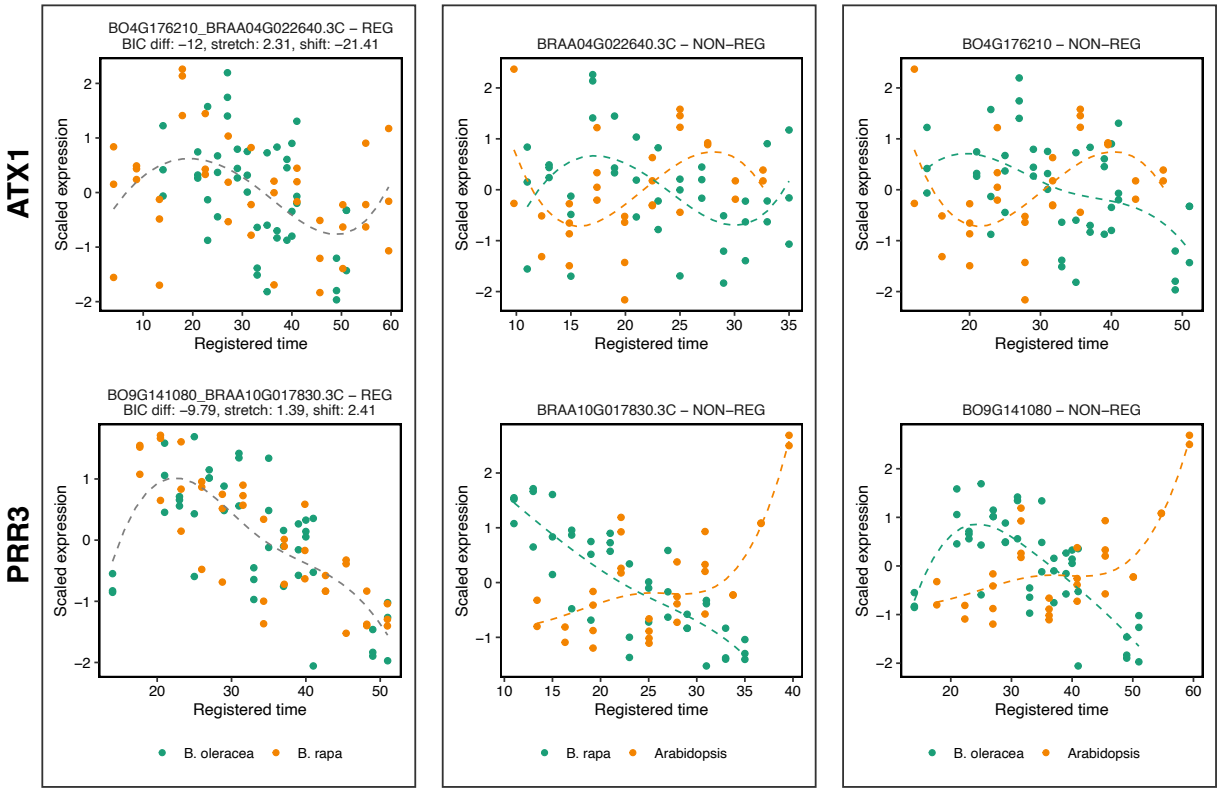


Figure 5.9: *ATX1* and *PRR3* are categorised as Brassica-specific genes. Each panel shows the pairwise comparison between species, from left to right: *B. rapa* vs. *B. oleracea*, *B. rapa* vs. Arabidopsis, and *B. oleracea* vs. Arabidopsis. Each dot represents a gene expression replicate at each time point. Green and orange colours denote gene expression for the species indicated in the legend below each plot. A grey dashed line indicates that the two profiles are registered as a single model fitted to both gene expression profiles. If no grey dashed line is present, the pair is not registered, with each profile fitted separately (green and orange dashed lines corresponding to the species indicated in the legend). The subtitle in each plot provides the gene ID and the optimal stretch and shift parameters for each pair of profiles.

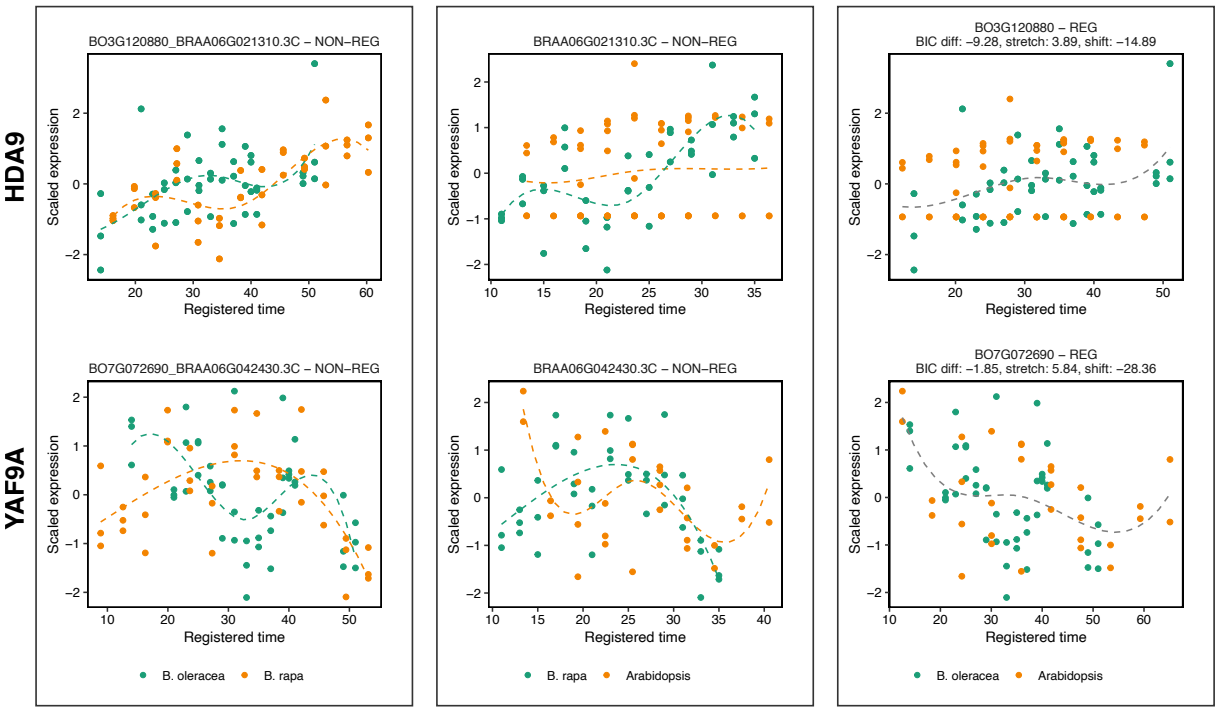


Figure 5.10: *HDA9* and *YAF9A* are categorised as *B. rapa*-specific genes. Each panel shows the pairwise comparison between species, from left to right: *B. rapa* vs. *B. oleracea*, *B. rapa* vs. Arabidopsis, and *B. oleracea* vs. Arabidopsis. Each dot represents a gene expression replicate at each time point. Green and orange colours denote gene expression for the species indicated in the legend below each plot. A grey dashed line indicates that the two profiles are registered as a single model fitted to both gene expression profiles. If no grey dashed line is present, the pair is not registered, with each profile fitted separately (green and orange dashed lines corresponding to the species indicated in the legend). The subtitle in each plot provides the gene ID and the optimal stretch and shift parameters for each pair of profiles.

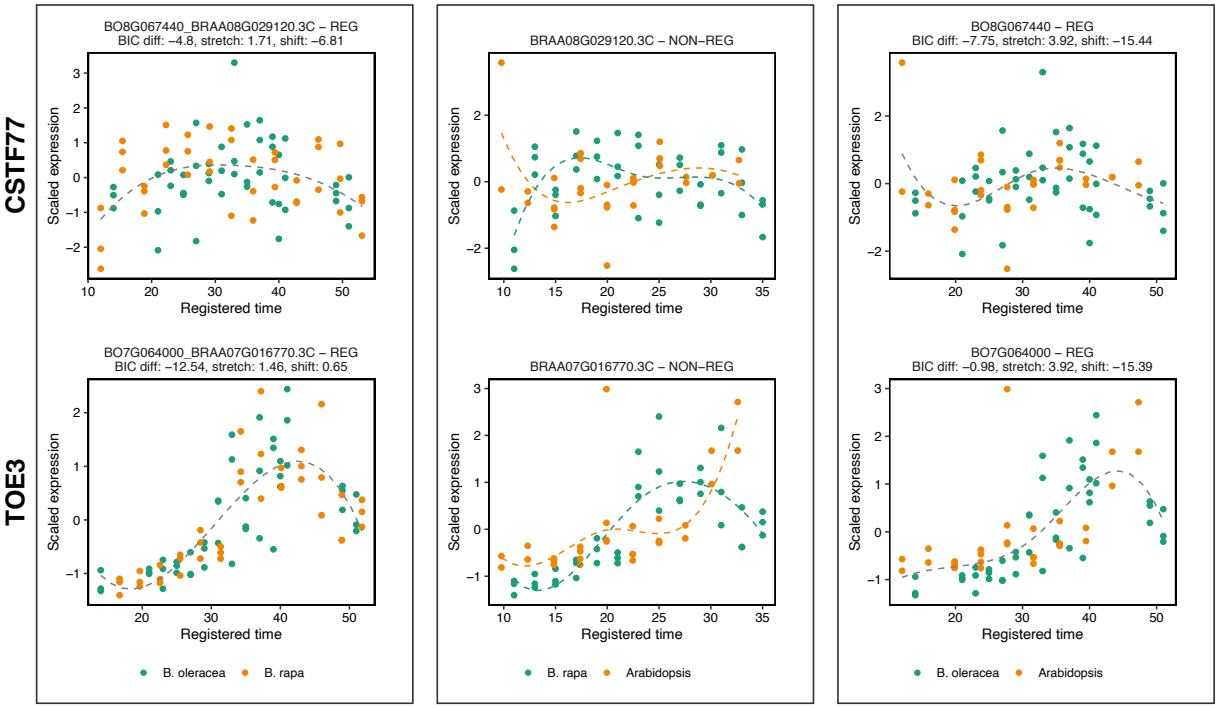


Figure 5.11: *CSTF77* and *TOE3* are categorised as Brassica-predominant genes. Each panel shows the pairwise comparison between species, from left to right: *B. rapa* vs. *B. oleracea*, *B. rapa* vs. Arabidopsis, and *B. oleracea* vs. Arabidopsis. Each dot represents a gene expression replicate at each time point. Green and orange colours denote gene expression for the species indicated in the legend below each plot. A grey dashed line indicates that the two profiles are registered as a single model fitted to both gene expression profiles. If no grey dashed line is present, the pair is not registered, with each profile fitted separately (green and orange dashed lines corresponding to the species indicated in the legend). The subtitle in each plot provides the gene ID and the optimal stretch and shift parameters for each pair of profiles.

5.3.4 Investigating the expression of key floral genes in Arabidopsis which have multiple copies in *B. rapa* and *B. oleracea*

Following on the previous work conducted in *B. rapa* by Calderwood *et al.* [85], we also investigated the expression profiles of five key floral transition genes in *B. rapa* and *B. oleracea*. The selected genes are *SOC1*, *AGL24*, *AP1*, *FLY*, and *TFL1* (see Figure 5.12). We chose these genes because previous studies reported that their expression patterns are diagnostic for different developmental stages in Arabidopsis [215]. In addition to these genes, we also chose *FLC* as is shown in Figure 5.12, *FLC* is a direct repressor of *SOC1* as a response to the vernalisation pathway.

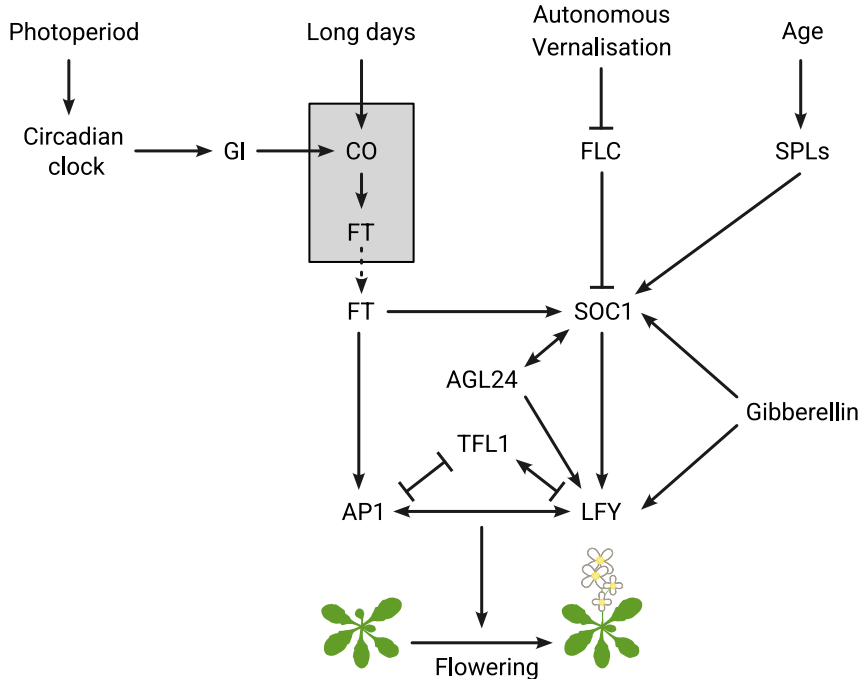


Figure 5.12: Schematic representation of the regulatory interactions between key floral integrators in Arabidopsis influenced by five different pathways. Adapted from [260, 261, 262]

Briefly, in Arabidopsis, when *SOC1* expression is activated in the apex together with *AGL24* it directly induces expression of *LFY*, one of the major floral meristem identity genes. Another gene which is also necessary to establish and maintain flower meristem identity, *AP1* is activated mainly by *FT* (Figure 5.12). When *LFY* and *AP1* are induced, flower development occurs at the SAM according to the ABC model, through the activation of *AP3* which is classified in B-class function [263]. *TFL1* acts antagonistically to *FT* and has a function in regulating the length of the inflorescence phase between induction of *FT* expression and conversion of the SAM into a floral meristem [264]. *TFL1* is reported to be important to flexibly counterbalance incoming *FT* signals. Its expression increases in proportion to the strength of the floral inductive signal [265].

Figure 5.13 shows the absolute expression dynamics of all copies of key regulatory genes across three species: Arabidopsis, *B. rapa*, and *B. oleracea*. The timings of gene expression relative

to the floral transition time (marked by the vertical black line) highlight a contrast between the Brassica species and Arabidopsis. In Arabidopsis, the expression of *LFY* and *AP1* genes begins to increase rapidly at the onset of the floral transition. In contrast, these genes in *B. rapa* and *B. oleracea* show a sharp rise in expression shortly after the floral transition. This indicates a delay in the activation of these genes in the Brassica species compared to Arabidopsis. Moreover, the expression of *TFL1* remains relatively low throughout the developmental stages in both Arabidopsis and *B. rapa*. However, in *B. oleracea*, *TFL1* shows a significantly higher expression level, even from the earliest time points observed. This suggests a potential distinct regulatory mechanism for *TFL1* in *B. oleracea* compared to the other two species. Interestingly, the expression levels of *TFL1* among biological replicates in *B. oleracea* vary widely. Some replicates show consistently high expression, while others are much lower. This variation could be due to biological differences, such as distinct regulatory states, or it might result from technical noise. It also raises the possibility that there are two expression patterns, or clusters, within *B. oleracea*.

Despite their differences in their timing of activation, most genes in *B. rapa* and *B. oleracea* have similar dynamics with those in Arabidopsis. For instance, in Arabidopsis *SOC1* expression starts to increase before *LFY* before the floral transition, the same dynamics can be observed in *B. oleracea*. In *B. rapa*, however, *SOC1* and *LFY* both increase at approximately the same time. Additionally, while the expression of *AP1* in both *B. rapa* and *B. oleracea* rises rapidly after the floral transition, which is considered relatively late compared to Arabidopsis, the overall expression dynamics of *AP1* in these Brassica species are similar to those in Arabidopsis. This suggests that despite some differences in timing, the fundamental regulatory mechanisms governing these key floral integrator genes are potentially conserved across these species.

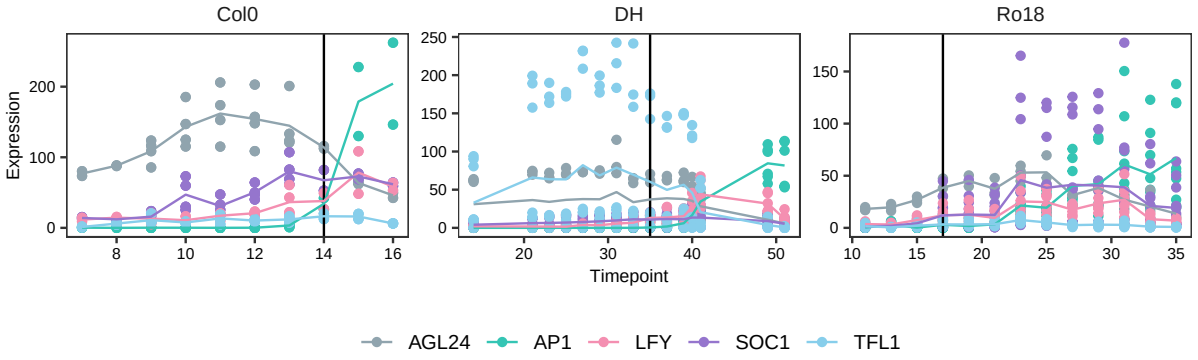


Figure 5.13: The expression profiles of five selected floral genes were investigated in Arabidopsis (Col0), *B. oleracea* (DH), and *B. rapa* (Ro18). Each gene is represented by a distinct colour. Each dot represents a gene expression replicate at each time point. The morphologically identified floral transition is marked by a vertical black line: at 14d in Arabidopsis, 35d in *B. oleracea*, and 17d in *B. rapa*. The timing of gene expression changes relative to other genes differs among the three species.

We performed registration to investigate the dynamics of these five homologous genes in Brassica

relative to Arabidopsis. This process was applied to each gene copy in Brassica. Additionally, we examined the similarity of the total expression of these gene copies across species. Calderwood *et al.* [266] previously reported that in *B. napus*, which has nine copies of *FLC*, the total expression of all *FLC* copies explains the differences in vernalisation requirements among *B. napus* types. This suggests that all *FLC* paralogues in *B. napus* are important in determining the cold requirement and response of each crop type. Building on this, we hypothesised that all paralogues of some of these genes are essential for their regulatory functions.

For *AP1*, *LFY*, and *AGL24* (Figure 5.14– 5.17), the curve registration effectively superimposed the expression patterns between Arabidopsis and *B. oleracea*, Arabidopsis and *B. rapa* homologues, as well as the total paralogues across the three species. Although Calderwood *et al.* [85] previously reported that these genes are similar between *B. rapa* and Arabidopsis, their analysis used fixed and equally sampled ranges for the shift and stretch parameters, which may have affected the parameter optimality between paralogues. Using mean data for registration may also introduce issues, as the variation between replicates is not accounted for.

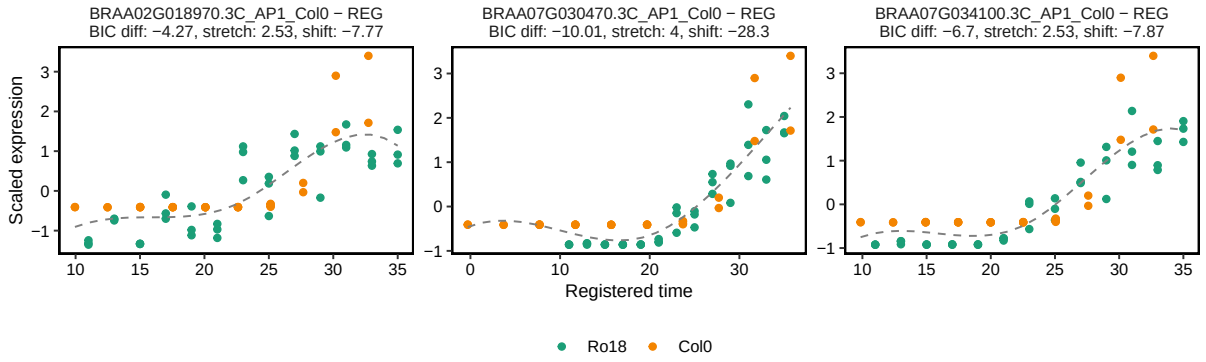
In contrast, the copy of *AP1* (BO6G095760) and *AGL24* (BO1G039080) in *B. oleracea* were not successfully registered to their Arabidopsis homologues in previous study by Woodhouse [201], although their dynamic similarities were evident without registration. This failure likely resulted from the sampling method for parameter selection, where the true optimal parameters might not have been included in the sampled set. Although all *LFY* paralogues in *B. oleracea* were successfully registered by Woodhouse [201], the best alignment achieved was local, with approximately 50% or fewer overlapping time points. Such a small percentage of overlapping should not be considered, especially since the sampling across the three species was performed to match morphological phenotypes.

The successful registration of these three genes across the three species indicates that they are highly conserved, particularly with the ancestor Arabidopsis, as well as within the Brassica genus. This high level of conservation suggests that their functions have likely been maintained across these species.

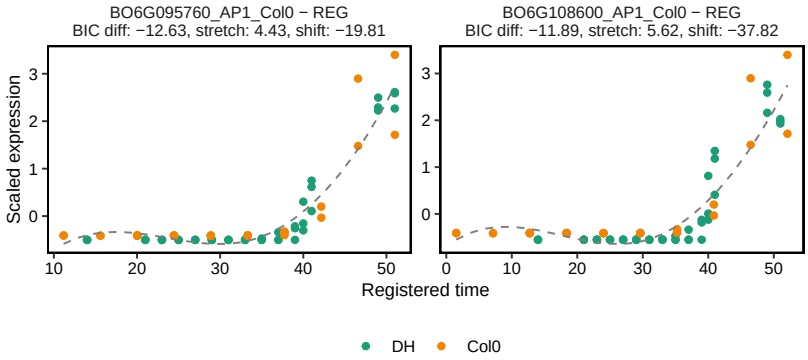
Three paralogues of *SOC1* have been identified in *B. oleracea*, with one located on C03 and two on C04 [215]. The functions of these paralogues have not yet been explored. In Woodhouse's study [201], the *SOC1* paralogue on C04, BO4G195720, was excluded due to its low expression level. However, we included this paralogue for its dynamic interest. Woodhouse's study successfully registered the two remaining *SOC1* copies with the best alignment found to be a local alignment with low overlapping timepoint coverage. Our analysis, with a higher percentage of overlapping alignment, revealed that one of the C03 paralogues, BO3G038880, did not register (Figure 5.15), suggesting a potential functional divergence from the Arabidopsis homologue. The other two paralogues on C04 were successfully registered but have different optimal parameters. This indicates that while these paralogues have similar expression profiles, they are desynchronised from each other and the Arabidopsis homologue. Additionally, all paralogues of *SOC1* in *B. rapa* are observed to be highly conserved with the dynamics being registered for all paralogues. Despite the dissimilarity of one *SOC1* paralogue in *B. oleracea*, the overall

comparison of all *SOC1* paralogues in *B. rapa* and *B. oleracea* shows a high degree of similarity among the three species. This indicates that the divergent dynamics of the *SOC1* paralogue on C03 in *B. oleracea* do not necessarily imply a potential difference in function. Instead, the total expression still maintains dosage balance and interacts with other genes to perform the SOC1 function, though a protein sequence comparison is needed to confirm this.

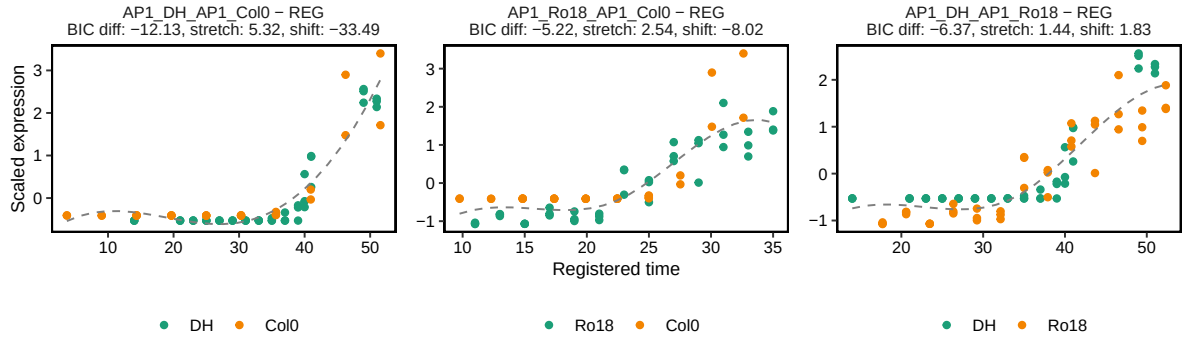
While *AP1*, *LFY*, *SOC1*, and *AGL24* show high similarity between their homologues in *B. rapa* and Arabidopsis, as well as in *B. oleracea* and Arabidopsis, and also in the total number of their paralogues, we observed different behaviour in *TFL1* (see Figure 5.18). All copies of *TFL1* in *B. rapa* exhibit similar dynamics to their homologue in Arabidopsis. However, this similarity is not observed in *B. oleracea*, as two of the three *TFL1* copies do not display dynamics similar to their homologues in Arabidopsis. When comparing the sums of the paralogues for each species, we observed dissimilarities between *B. rapa* and *B. oleracea*, but not between each species and Arabidopsis. This suggests a potential functional divergence in *TFL1* within Brassica while retaining a consistent function to some degree with Arabidopsis.



(a)

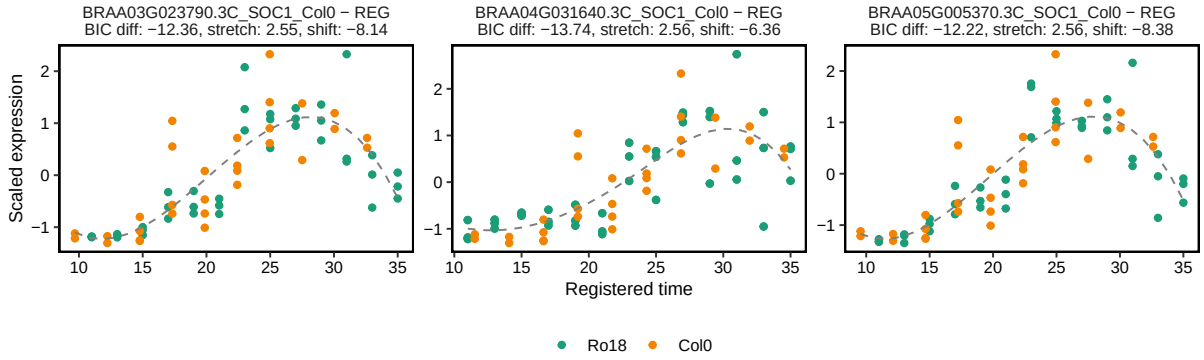


(b)

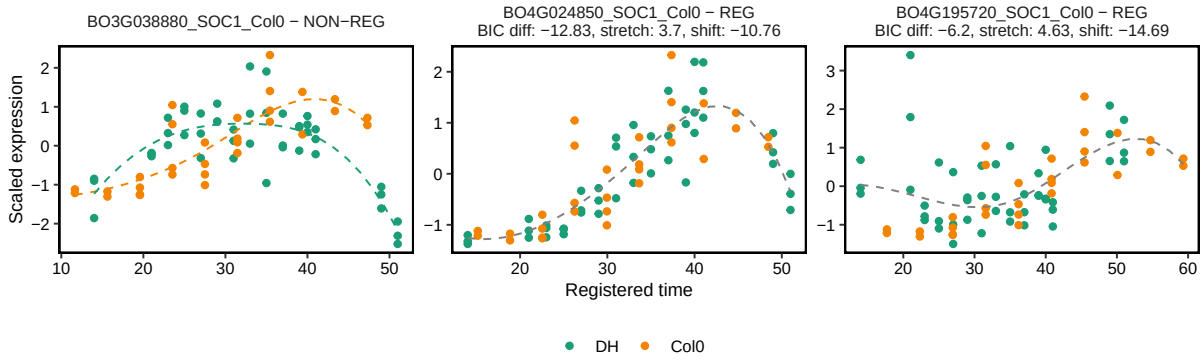


(c)

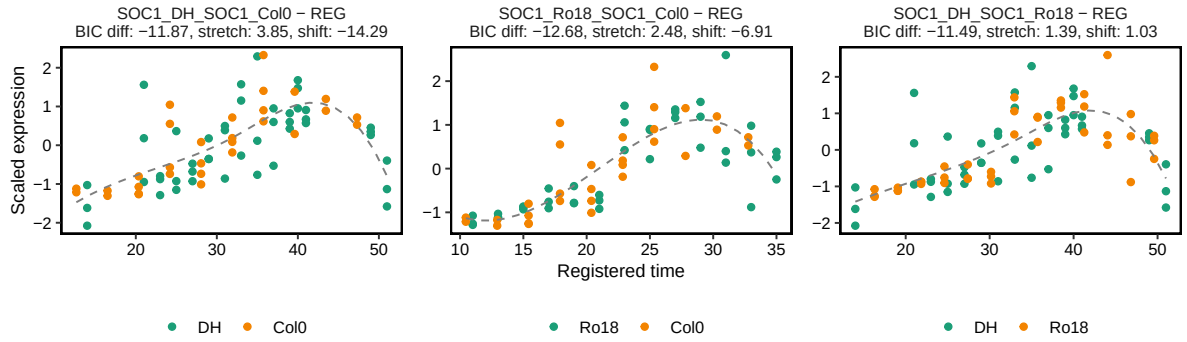
Figure 5.14: Registration results of (a) *AP1* paralogues in *B. rapa* and Arabidopsis, (b) in *B. oleracea* and Arabidopsis, and (c) total *AP1* copies among *B. rapa*, *B. oleracea*, and Arabidopsis. Each dot represents a gene expression replicate at each time point. Green and orange indicate gene expression for the species indicated in the legend below each plot. A grey dashed line indicates that the two profiles are registered as a single model fitted to both gene expression profiles. If no grey dashed line is present, the pair is not registered, with each profile fitted separately (green and orange dashed lines corresponding to the species indicated in the legend). The subtitle in each plot provides the gene ID and the optimal stretch and shift parameters for each pair of profiles.



(a)

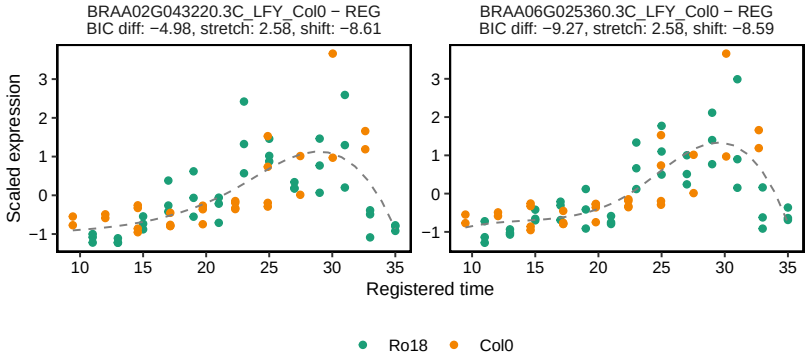


(b)

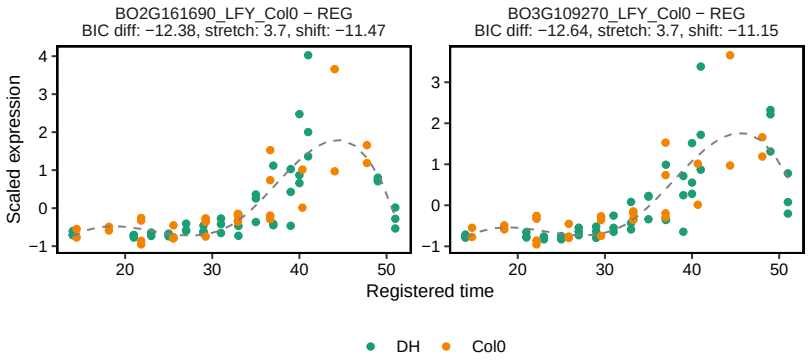


(c)

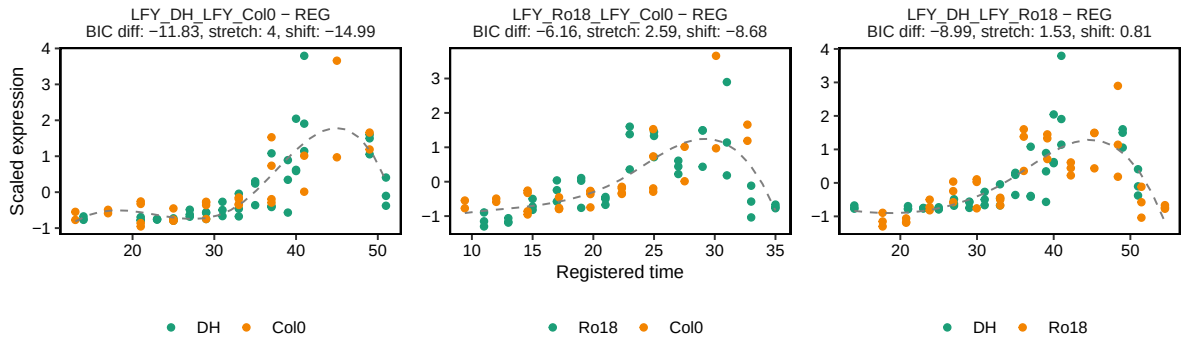
Figure 5.15: Registration results of (a) *SOC1* paralogues in *B. rapa* and Arabidopsis, (b) in *B. oleracea* and Arabidopsis, and (c) total *SOC1* copies among *B. rapa*, *B. oleracea*, and Arabidopsis. Each dot represents a gene expression replicate at each time point. Green and orange indicate gene expression for the species indicated in the legend below each plot. A grey dashed line indicates that the two profiles are registered as a single model fitted to both gene expression profiles. If no grey dashed line is present, the pair is not registered, with each profile fitted separately (green and orange dashed lines corresponding to the species indicated in the legend). The subtitle in each plot provides the gene ID and the optimal stretch and shift parameters for each pair of profiles.



(a)

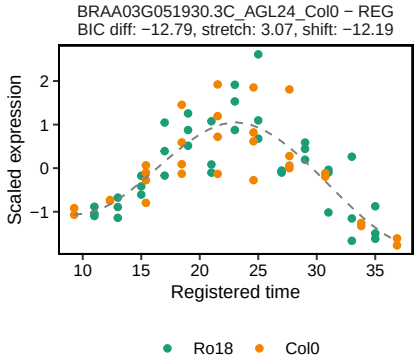


(b)

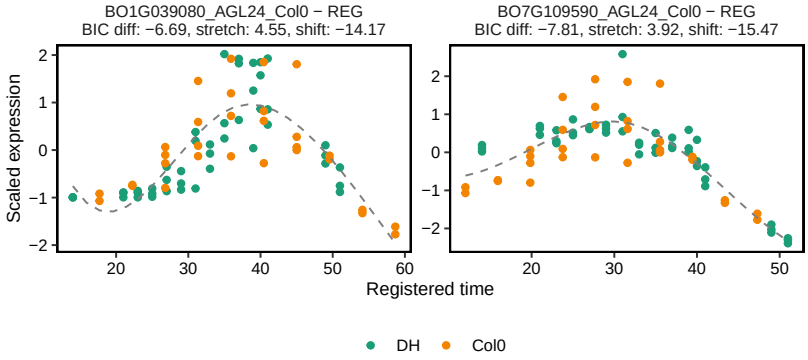


(c)

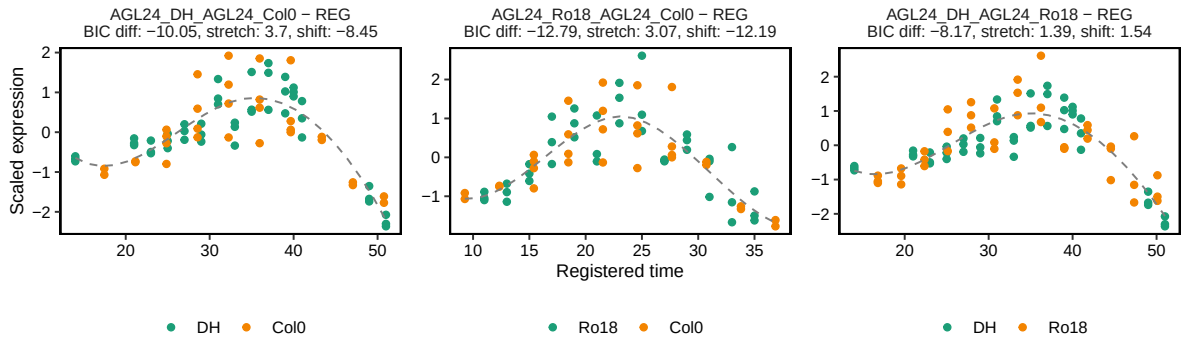
Figure 5.16: Registration results of (a) *LFY* paralogues in *B. rapa* and Arabidopsis, (b) in *B. oleracea* and Arabidopsis, and (c) total *LFY* copies among *B. rapa*, *B. oleracea*, and Arabidopsis. Each dot represents a gene expression replicate at each time point. Green and orange indicate gene expression for the species indicated in the legend below each plot. A grey dashed line indicates that the two profiles are registered as a single model fitted to both gene expression profiles. If no grey dashed line is present, the pair is not registered, with each profile fitted separately (green and orange dashed lines corresponding to the species indicated in the legend). The subtitle in each plot provides the gene ID and the optimal stretch and shift parameters for each pair of profiles.



(a)

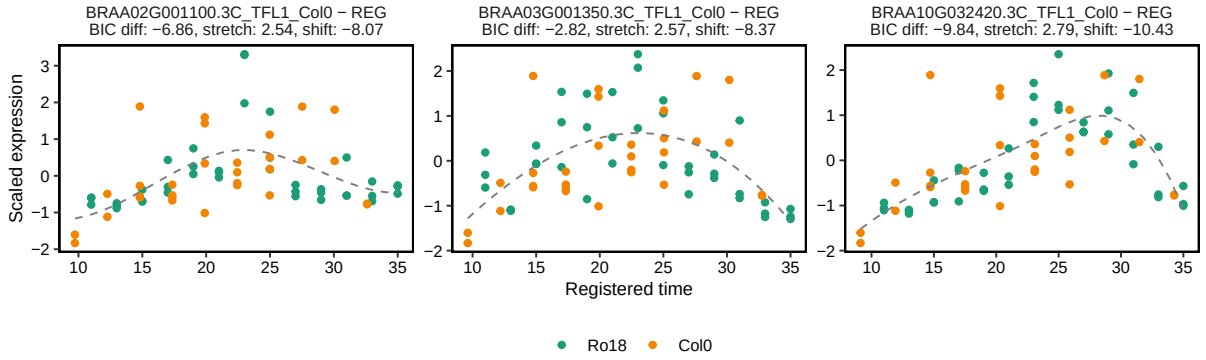


(b)

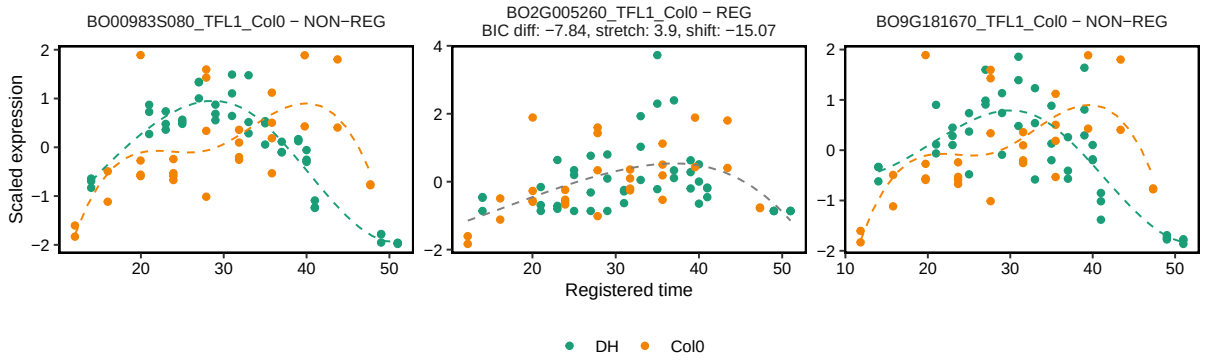


(c)

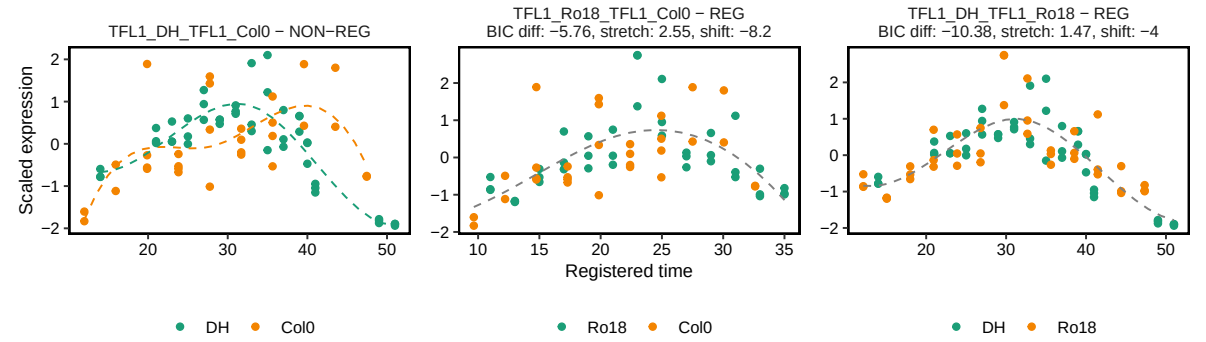
Figure 5.17: Registration results of (a) *AGL24* paralogues in *B. rapa* and Arabidopsis, (b) in *B. oleracea* and Arabidopsis, and (c) total *AGL24* copies among *B. rapa*, *B. oleracea*, and Arabidopsis. Each dot represents a gene expression replicate at each time point. Green and orange indicate gene expression for the species indicated in the legend below each plot. A grey dashed line indicates that the two profiles are registered as a single model fitted to both gene expression profiles. If no grey dashed line is present, the pair is not registered, with each profile fitted separately (green and orange dashed lines corresponding to the species indicated in the legend). The subtitle in each plot provides the gene ID and the optimal stretch and shift parameters for each pair of profiles.



(a)



(b)



(c)

Figure 5.18: Registration results of (a) *TFL1* paralogues in *B. rapa* and Arabidopsis, (b) in *B. oleracea* and Arabidopsis, and (c) total *TFL1* copies among *B. rapa*, *B. oleracea*, and Arabidopsis. Each dot represents a gene expression replicate at each time point. Green and orange indicate gene expression for the species indicated in the legend below each plot. A grey dashed line indicates that the two profiles are registered as a single model fitted to both gene expression profiles. If no grey dashed line is present, the pair is not registered, with each profile fitted separately (green and orange dashed lines corresponding to the species indicated in the legend). The subtitle in each plot provides the gene ID and the optimal stretch and shift parameters for each pair of profiles.

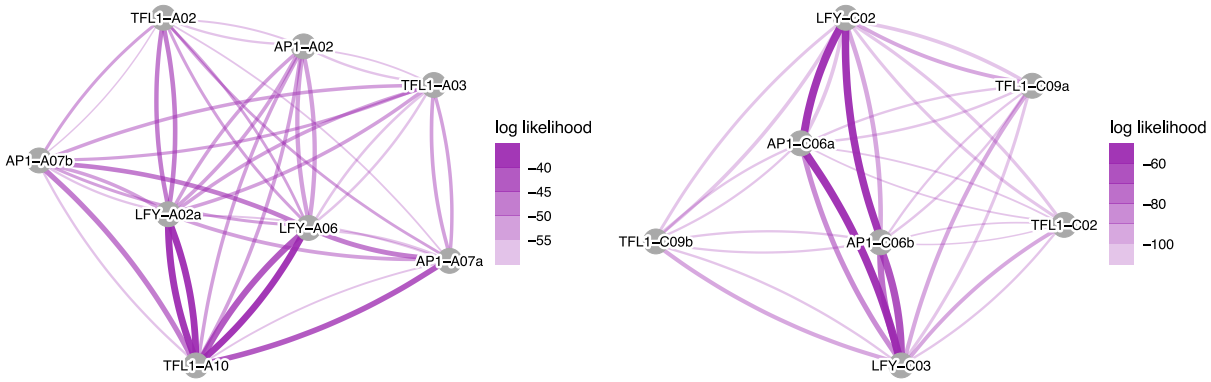
Gene regulatory networks to explore the differences between TFL1 paralogues

The successful registration of all *TFL1* paralogues in *B. rapa*, compared to only one out of three paralogues in *B. oleracea* may suggest that *TFL1* is regulated differently between these two species. To explore the roles of these paralogues, we will use gene regulatory networks (GRNs) inferred with the CSI (Causal Structure Identification) method [267]. CSI is a Gaussian process-based approach for inferring GRNs from multiple time series datasets. It jointly learns a single regulatory network from all datasets, or alternatively, can infer separate but related networks for each context using a hierarchical approach, enabling the detection of both shared and context-specific regulatory relationships.

The successful registration of all *AP1* copies in both *B. rapa* and *B. oleracea* suggests that they may perform roles similar to that of Arabidopsis *AP1*, as indicated by their similar expression profiles identified through curve registration (Figure 5.18). Conversely, the inability to register two AP1 copies in C09 (BO00983S080 and BO9G181670) suggests that these copies may be non-functional or have different functions compared to the Arabidopsis homologue.

Figure 5.19 presents the generated networks for each *TFL1* paralogue in both *B. rapa* and *B. oleracea*. We focused on two genes known to interact with Arabidopsis *TFL1*: *LFY* and *AP1* [268, 201, 269]. These three genes form a feedback loop, with *AP1* and *LFY* acting as floral meristem identity genes that regulate each other synergistically [261]. *AP1* and *LFY* have antagonistic roles in regulating *TFL1* expression: *LFY* promotes *TFL1* expression, while *AP1* suppresses it [268, 201].

Due to the identified similarities between all *TFL1* paralogues in *B. rapa*, we hypothesised that a similar network could be generated, showing a strong association between the copies of the three genes. Interestingly, only one *TFL1* copy on A10 in *B. rapa* (BRAA10G032420.3C) is strongly associated with both *LFY* paralogues and two *AP1* copies, whereas the other *TFL1* copies are not. Notably, the BIC value for the *TFL1* on A10 is the smallest compared to the other copies in A02 (BRAA02G001100.3C) and A03 (BRAA03G001350.3C), suggesting that the expression of this copy is the closest to the Arabidopsis homologue. However, none of the *TFL1* copies in *B. oleracea* have a strong association with either *AP1* or *LFY* copies, even though one of the copies on C02 was registered. This discrepancy may explain the differences observed when comparing the total *TFL1* copies among the three species: the total in *B. oleracea* differs from Arabidopsis *TFL1* but still shows some similarity when compared to *B. rapa*. This suggests a potential divergence of *TFL1* copies in *B. oleracea* from Arabidopsis, while still retaining some overlapping functions with *TFL1* copies in *B. rapa*.

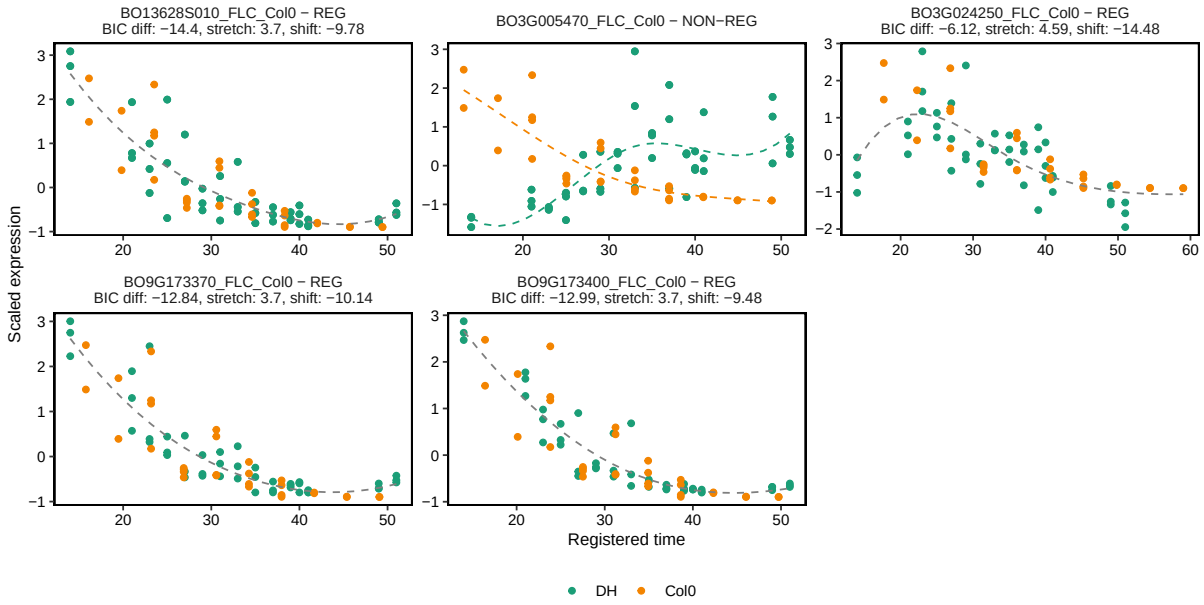


(a) **(b)**

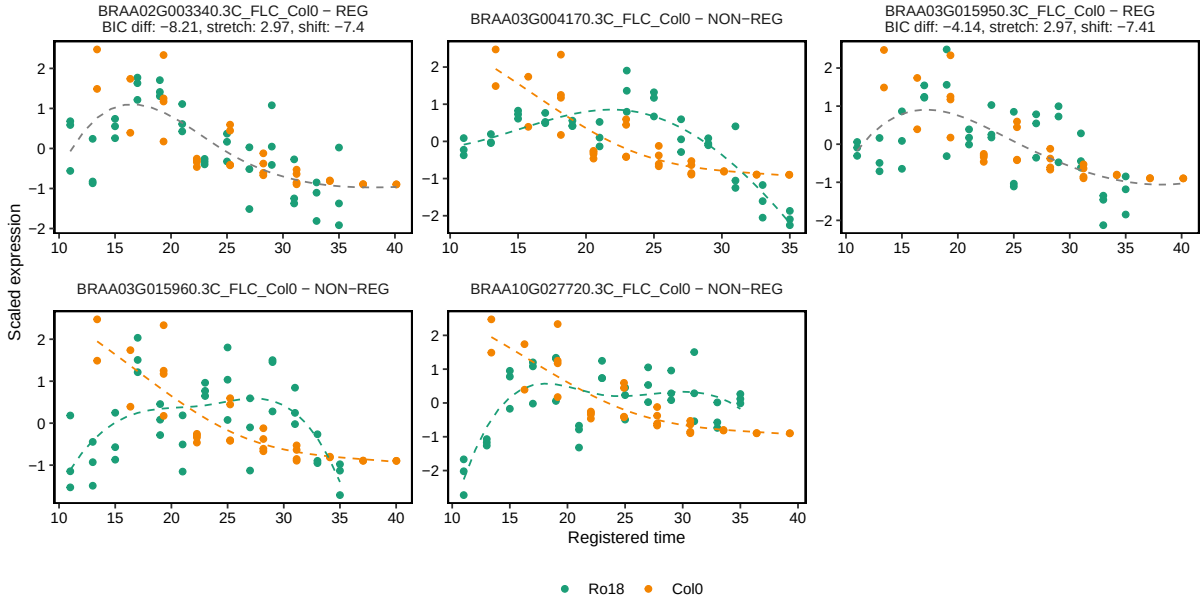
Figure 5.19: CSI inference [267] showing interactions between *AP1*, *LFY*, and *TFL1* in (a) *B. rapa* and (b) *B. oleracea*. The analysis reveals differential regulation of *TFL1* between *B. rapa* and *B. oleracea*, with most *TFL1* paralogues in *B. oleracea* showing no strong associations with either *AP1* and *LFY* paralogues, whereas in *B. rapa*: one paralogue in A010 (BRAA10G032420.3C) strongly associate with both *LFY* copies and one *AP1* copy in A07 (BRAA07G030470.3C).

5.3.5 Exploring the role of *FLC* in *B. rapa* and *B. oleracea*

FLC is a well-studied floral repressor known for its key role in the vernalisation pathway and its dosage-dependent control of flowering time. The number of *FLC* gene copies present appears to influence flowering time [201]. Many plant species, including Brassica, require prolonged cold exposure, typically encountered during winter, before they can flower and set seed [218]. Without this exposure, known as vernalisation, flowering is prevented [218]. Vernalisation is an evolutionary adaptation to temperate climates, preventing premature flowering before winter and ensuring flowering occurs under the favourable conditions of spring. Although *B. oleracea DH1012* and *B. rapa R-o-18* are rapid cycling varieties that do not require vernalisation, their *FLC* copies were expressed in the apex (Figure 5.20). Therefore, we investigated whether the *FLC* paralogues in *B. rapa* and *B. oleracea* are potentially functional by comparing their gene expressions to their homologue Arabidopsis. We identified five *FLC* paralogues in both *B. rapa* and *B. oleracea*. Table 5.6 provides a detailed list of these paralogues along with their chromosome locations.



(a) Registration results of *FLC* paralogues in *B. oleracea* and Arabidopsis. Green and orange dots represent expression replicates for each time point in *B. oleracea* and Arabidopsis, respectively. BO3G005470 *FLC* is the only gene that did not register to its Arabidopsis homologue.



(b) Registration results of *FLC* paralogues in *B. rapa* and Arabidopsis. Out of the five identified copies of *FLC* in *B. rapa*, only two paralogues, BRAA02G003340 *FLC* and BRAA03G015950 *FLC*, were successfully registered to their Arabidopsis homologue.

Figure 5.20: Registration results of each individual copy of *FLC* in (1) *B. rapa* and (2) *B. oleracea*.

Species	Gene ID	Chromosome	Alias
B. oleracea	BO13628S010	C01	C01
B. oleracea	BO3G005470	C03	C03a
B. oleracea	BO3G024250	C03	C03b
B. oleracea	BO9G173370	C09	C09a
B. oleracea	BO9G173400	C09	C09b
B. rapa	BRAA02G003340.3C	A02	A02
B. rapa	BRAA03G004170.3C	A03	A03a
B. rapa	BRAA03G015950.3C	A03	A03b
B. rapa	BRAA03G015960.3C	A03	A03c
B. rapa	BRAA10G027720.3C	A10	A10

Table 5.6: A list of *FLC* paralogues along with their respective chromosome locations in *B. rapa* and *B. oleracea*.

In *B. oleracea*, the expression of three out of five *FLC* copies (BO13628S010, BO9G173370, and BO9G173400) continuously decreases from the first sampled time point, a trend that begins before the meristem transition from vegetative to inflorescence at 35d. Although one copy, BO3G024250, shows a slight increase before the second time point, its expression dynamics align with the other three copies at later time points. Additionally, curve registration identified the expression profiles of these four copies closely match their Arabidopsis homologue (Figure 5.20 (a)), suggesting they may have similar functions. However, one copy, BO3G005470, exhibits different dynamics compared to its Arabidopsis homologue. The expression of this copy increases until the floral transition and stabilises afterwards, contrasting with the other four paralogues. Consequently, curve registration identified this copy as having different dynamics from its Arabidopsis homologue. These findings align with Woodhouse's study [201], although we found different optimal parameters for the registrations.

Four *FLC* copies were identified on *B. rapa* [270]: one paralogue on A02, two on A03, and one on A10. Using curve registration, we observed that two of the paralogues, one on A03 and the one on A10, did not match their Arabidopsis homologue. The other two copies, BRAA02G003340.3C on A02 and BRAA03G015950.3C on A03, were successfully registered, although their expression patterns slightly upregulated before the floral transition at 17d, followed by a continuous decrease afterwards. This behaviour contrasts with the dynamics in Arabidopsis, where expression continuously decreases prior to the floral transition. These suggest that the *FLC* paralogues in *B. rapa* may have different roles from *FLC* in Arabidopsis.

Each *FLC* homologue between *B. rapa* and *B. oleracea* was compared to one another. Figure 5.21 summarises the results of these comparisons. As shown in Figure 5.20 (a) and (b) (indicated by green checkmarks), more *FLC* copies were found to be similar to their homologues in Arabidopsis compared to those in *B. rapa*. When pairwise comparisons of homologues between *B. rapa*

and *B. oleracea* were performed, all copies in *B. rapa* were registered to the *FLC* paralogue on Chromosome C03 (BO3G024250 or C03b). However, no other significant similarities were identified, except for the *B. rapa* copies on Chromosomes A03 (BRAA03G015960.3C or A03c) and A10 (BRAA10G027720.3C), which showed similarity to one *FLC* copy in *B. oleracea* on Chromosome C03 (BO3G005470 or C03a). Supplementary Figure S.1 provides detailed registration results for each pair of homologues.

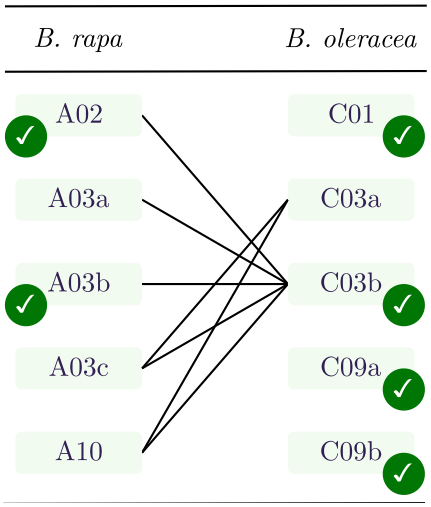


Figure 5.21: Registration results of *FLC* homologues in Arabidopsis on different chromosomes in *B. rapa* and *B. oleracea* are shown. A black line indicates registered homologue pairs, while the absence of a line signifies non-registered pairs. Green checkmarks denote that each copy is registered to the Arabidopsis homologue.

When comparing the total *FLC* expression in *B. oleracea* and *B. rapa* to each other and to Arabidopsis, we observed that the *FLC*s in *B. rapa* exhibit a distinct pattern compared to *B. oleracea* and Arabidopsis (Figure 5.22). This finding again aligns with the results reported by Calderwood *et al.* [85]. Although one *FLC* copy in *B. oleracea* was not registered, the overall *FLC* dynamics in *B. oleracea* still match those of the homologues in Arabidopsis. This consistency may be due to the dosage-dependent manner in which *FLC* copies function, where the expression of the remaining paralogues compensates for the unregistered copy.

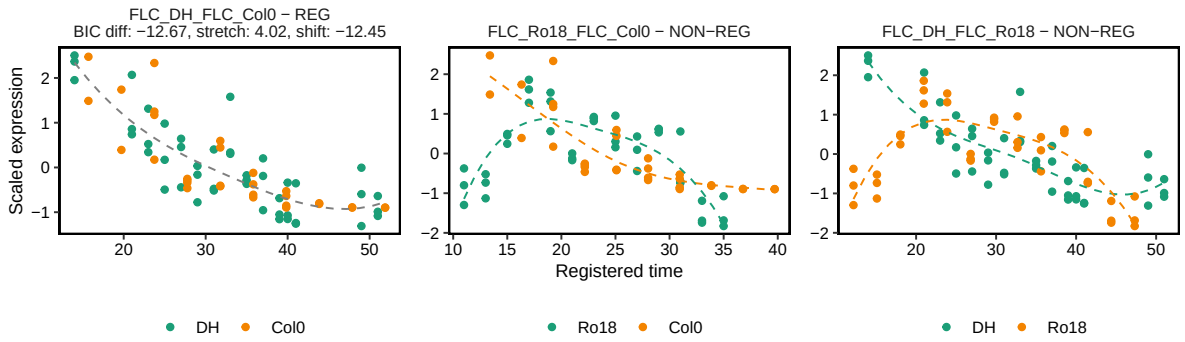


Figure 5.22: Registration results of total *FLC* copies among *B. oleracea*, *B. rapa*, and Arabidopsis. The total *FLC* copies in *B. rapa* was not successfully registered to either the total *FLC* copies in *B. oleracea* or its homologue in Arabidopsis. However, the total FLC in *B. oleracea* shows similar dynamics to the *FLC* in Arabidopsis.

Additionally, we identified another *FLC* copy in *B. rapa* located in A03, BRAA03G015960.3C, which exhibits different dynamics compared to its homologue in Arabidopsis. In the following section, we will further investigate the role of *FLC* by exploring the gene regulatory networks involving this gene in Arabidopsis, *B. rapa*, and *B. oleracea*.

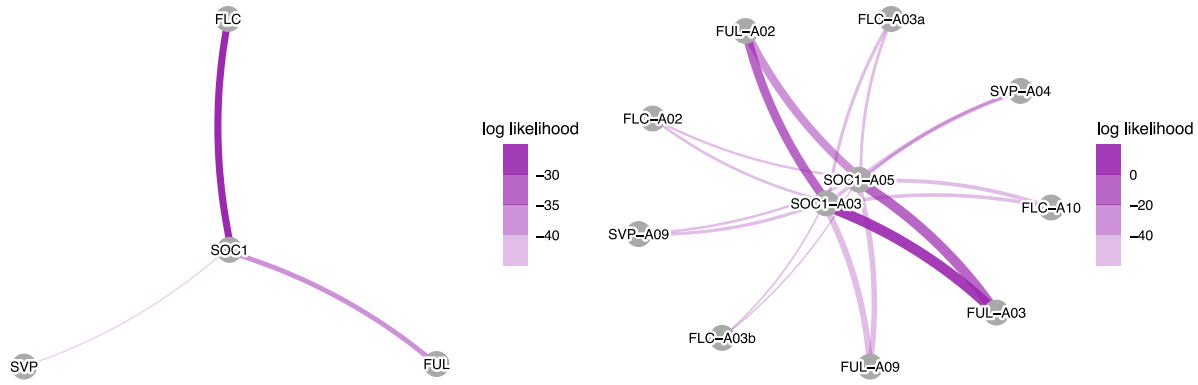
Gene regulatory networks to explore the role of *FLC* in *B. rapa* and *B. oleracea*

In a previous study, Calderwood *et al.* [85] inferred gene regulatory networks involving *SHORT VEGETATIVE PHASE* (*SVP*), *FLC*, *FRUITFULL* (*FUL*), and *SOC1* in both Arabidopsis and *B. rapa*. Consistent with earlier findings, their inferred network demonstrated that in Arabidopsis, *SOC1* expression is regulated by *FLC* and activated by *FUL*. However, in *B. rapa*, none of the *FLC* paralogues show a strong association with *SOC1*. Instead, *SOC1* expression is primarily linked to the expression of two *FUL* paralogues located on Chromosomes A02 and A03. We generated gene regulatory networks mirroring those by Calderwood *et al.* (Figure 5.23) but included the additional *FLC* copy BRAA03G015960.3C (shown in Figure 5.24 (b)). From this network, we observed that this *FLC* copy does not appear to associate with *SOC1*, as indicated by curve registration, similar to the other *FLC* copies. On the other hand, consistent with Calderwood *et al.* [85], *FUL* paralogues in A03 and A02 are highly associated with the two copies of *SOC1* in *B. rapa*.

We also examined the role of *FLC* copies in *B. oleracea* by constructing a gene regulatory network based on known interactions with *FLC* in Arabidopsis. As shown in Figure 5.24 (b), unlike *B. rapa*, most *FLC* paralogues in *B. oleracea* exhibit a strong association with *SOC1* copies. The *FLC* paralogues with similar dynamics identified through curve registration (Figure 5.20) show a significant association with *SOC1* copies. However, the two copies on C03: BO3G024250, which is registered but exhibits slightly different dynamics initially, and BO3G005470, which is not registered, demonstrate a much weaker association with *SOC1* copies.

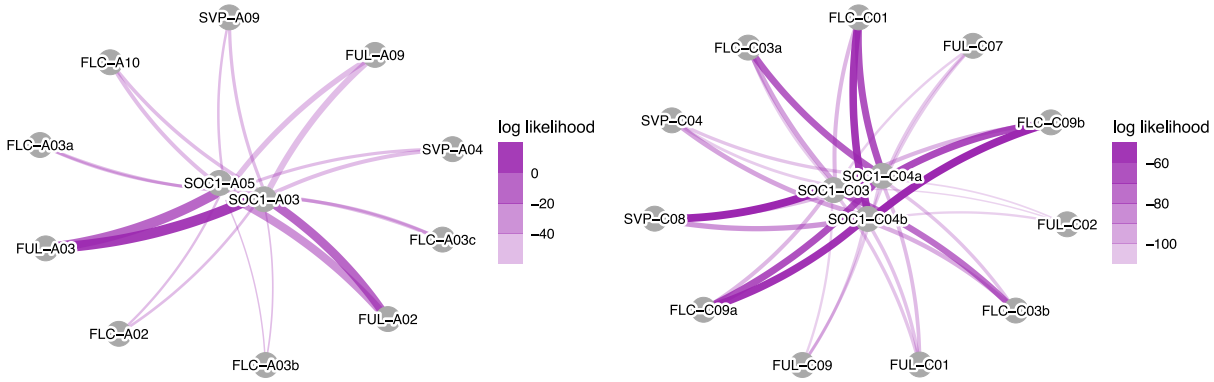
By comparing the relationships between *FLC* and *SOC1* copies in the regulatory networks of *B. rapa* and *B. oleracea*, we found that, despite the *B. oleracea* cultivar being a rapid-cycling variety that does not require vernalisation, its *FLC* copies may still be functional. This observation indicates that *FLC* genes still play a significant role in the regulatory network even in rapid-cycling cultivars. This insight could be highly valuable for breeding strategies. Specifically, targeting and knocking out *FLC* copies that show the highest association with *SOC1* could lead to the development of even faster flowering rapid-cycling cultivars.

Additionally, our analysis highlights the utility of curve registration as a powerful tool for dissecting and understanding gene functionality. By providing detailed insights into the dynamics and interactions of specific genes, curve registration helps unravel complex genetic relationships and functions.



(a) **(b)**

Figure 5.23: Replication of CSI inference [267] showing the interactions between *SOC1*, *FLC*, *FUL*, and *SVP* in (a) Arabidopsis and (b) *B. rapa*. These GRNs, based on Calderwood *et al.* [85], reveal differential regulation of *SOC1* between Arabidopsis and *B. rapa*.



(a) **(b)**

Figure 5.24: CSI inference [267] showing interactions between *SOC1*, *FLC*, *FUL*, and *SVP* in (a) *B. rapa*, including an additional *FLC* copy in A03 (BRAA03G015960.3C), and (b) *B. oleracea*. The analysis reveals differential regulation of *SOC1* between *B. rapa* and *B. oleracea*, with most *FLC* paralogues in *B. oleracea* showing strong associations with *SOC1* paralogues, a pattern not observed in *B. rapa*.

5.3.6 Discussion

B. rapa and *B. oleracea* are among the most important vegetable crops [207], both members of the *Brassicaceae* family, along with the model species Arabidopsis. Despite their close relatedness and many similarities, the expression profiles of various developmental stages, including flowering time genes, differ among these three species. However, through curve registration, we observed that most of these differences can be resolved by desynchronising gene expression between them. These findings are consistent with previous studies comparing *B. rapa* with Arabidopsis [85] and *B. oleracea* with Arabidopsis [201].

While this work builds on the initial concept of applying curve registration to gene expression data, originally introduced by Calderwood *et al.* [85], the methodological implementation presented here represents a substantial advance. The original pipeline lacked integration, included redundant steps, and relied on sampling rather than formal parameter optimisation. In this thesis, I redesigned and streamlined the approach, reformulated the registration framework, introduced parameter optimisation, and implemented the method in a flexible R package, *greatR*. All analyses presented in this chapter, and throughout the thesis, were carried out by me. These innovations extend the original concept into a generalisable and practical tool for analysing biological time-series data across species.

In this study, we expanded on Woodhouse's analysis [201], examining not only key floral genes, which are often present in multiple copies but also genes present as single copies. Our comparisons extended beyond pairwise analyses between each Brassica crop and the model species to include comparisons among the Brassica crops themselves. This analysis aims to understand the molecular differences between the three crops and the extent of their conservation.

Before comparing the transcriptomic profiles, we analysed the copy number variation of flowering time genes in *B. rapa* and *B. oleracea* orthologues to Arabidopsis flowering time genes recorded in the FLOR-ID database [191]. In polyploid species, copy number variation is important as it affects gene and protein expression levels, ultimately influencing phenotype and evolutionary adaptation [271]. Our analysis revealed that more than half of the flowering time genes in *B. rapa* and *B. oleracea* have multiple copies. While the distribution of lower copy numbers is similar between the two crops, the total number of gene copies is higher in *B. oleracea* than in *B. rapa*. Li *et al.* [272] previously reported that *B. oleracea* has a larger number of transposable elements (TEs) compared to *B. rapa*, resulting in a larger genome. This genomic characteristic likely contributes to the higher number of flowering time gene copies in *B. oleracea*. Understanding these differences in copy number variation is crucial for analysing how each gene, as well as all copies collectively, contribute to the regulatory mechanisms and evolutionary pressures that shape these crops. This knowledge provides a foundation for exploring gene function and interaction, ultimately leading to insights into the adaptation and phenotypic diversity of *B. rapa* and *B. oleracea*.

For all identified flowering time genes in both *B. rapa* and *B. oleracea*, curve registration demonstrated that the differences between Arabidopsis and each of the two Brassicas can be mostly explained by delays or shifts in the timing of gene expression, rather than differences in the expression patterns themselves. We further investigated the extent of gene paralogue expression conservation in each Brassica relative to Arabidopsis and whether the registration results were consistent between the two Brassicas. We began by examining single-copy genes present in both Brassicas and then extended our analysis to key floral genes, which predominantly have multiple copies.

A potential limitation of this analysis is the assumption that paralogue-specific expression can be reliably estimated. Although sequence similarity between Brassica paralogues was not directly assessed, the high-confidence mapping to Arabidopsis orthologues with a minimum 95% sequence identity threshold provides strong support for the reliable identification of Brassica paralogues used in downstream expression analysis [235]. Furthermore, the expression data were generated with sufficient read depth across replicates, which supports the robustness of gene-level quantification. While alternative splicing was not the focus of this study, the use of CPM values helps reduce ambiguity caused by isoform variation.

Pairwise curve registration performed between the three species demonstrated that the majority of single-copy genes exhibit similar expression dynamics, indicating strong evolutionary retention and conservation. This suggests that these genes have maintained their functional roles across species through evolutionary time. Our registration results also revealed the instances where certain genes displayed divergent dynamics only between Brassicas, which we identified as lineage-specific divergence. These genes potentially have evolved differently in each lineage, adapting to species-specific needs or environmental pressures. Furthermore, we observed a subset of genes that exhibited Brassica-predominant dynamics, suggesting that this group of

genes have similar roles within Brassicas but may slightly differ from the ancestor Arabidopsis.

Two genes, *ATX1* and *PRR3*, were categorised as Brassica-specific because they differ from Arabidopsis but exhibit similar dynamics within the Brassicas. These genes potentially contribute to traits unique to Brassicas, distinguishing them from Arabidopsis. Additionally, two other genes, *HDA9* and *YAF9A*, were categorised as *B. rapa*-specific, as their expression in *B. rapa* is different from both *B. oleracea* and Arabidopsis. Although these genes have not been extensively explored, identifying them as potential candidates may help elucidate the genetic basis of *B. rapa*-specific traits related to flowering time. The observed transcriptomic differences between the two Brassica diploids are thought to result from their polyploid history during species formation [273]. Given that single-copy genes are often crucial and can have significant phenotypic effects when mutated [239, 240, 241], identifying these genes is essential. This knowledge will facilitate the exploration of genes related to species-specific characteristics and accelerate Brassica breeding programs.

The five key floral transition genes *AGL24*, *AP1*, *LFY*, *SOC1*, and *TFL1* were closely examined, building on previous studies [85, 201]. We found that all paralogues of *AP1*, *LFY*, and *AGL24* in both Brassicas are highly conserved compared to their homologues in Arabidopsis. In *B. rapa*, this result aligns with Calderwood *et al.* [85], who also observed dynamic similarities between *B. rapa* and Arabidopsis. However, in *B. oleracea*, the results differed; not all paralogues of *AP1* and *AGL24* were registered in Woodhouse's study [201]. Although *SOC1* paralogues were identified and registered in the previous study, each copy aligned only locally with their Arabidopsis homologue covering less than 50% of the entire *SOC1* copies at the beginning of their time ranges. We identified that these discrepancies are due to the methods used for curve registration, including parameter sampling and the log-likelihood functions employed. Despite these differences, the conservation of these genes is evident in their overall expression patterns across the three species.

SOC1 plays a crucial role in integrating signals from different pathways [262]. We observed that the three identified paralogues of this gene in *B. rapa* exhibit dynamics similar to the Arabidopsis homologue. However, one out of the three paralogues in *B. oleracea* displays different dynamics compared to *SOC1* in Arabidopsis. Despite this variation, the paralogues of genes that directly interact with *SOC1* in Arabidopsis, such as *AGL24* and *LFY*, do not show any dissimilarity to their homologues in Arabidopsis. This suggests that the other *SOC1* copies may have greater importance in fulfilling *SOC1* functions in *B. oleracea*, as evidenced by the overall conservation of total *SOC1* expression patterns across the three species. This observation highlights the potential of functional redundancy and specialisation that can occur within gene families.

The expression dynamics of *TFL1* paralogues in both *B. oleracea* and *B. rapa* compared to *TFL1* in Arabidopsis had not yet been explored. In Arabidopsis, *TFL1* plays a crucial role in actively suppressing the expression of *AP1* and *LFY* [268, 201]. In our investigation, we found that all three paralogues of *TFL1* in *B. rapa* exhibit the same dynamics as the *TFL1* homologue in

Arabidopsis, indicating a potential functional similarity between these two species. Conversely, in *B. oleracea*, two out of three *TFL1* paralogues show different dynamics compared to *TFL1* in Arabidopsis. When examining the total number of *TFL1* paralogues, we found that the total expressions of *TFL1* copies in *B. oleracea* differ from those in Arabidopsis, while maintaining similarity to *B. rapa*. To understand these roles further, we generated GRNs to capture the interactions between *TFL1* and its direct interactors, *AP1* and *LFY*. In *B. oleracea*, there is no strong association between *TFL1* paralogues and either *AP1* or *LFY* paralogues, contrary to what is observed in Arabidopsis [268, 201, 269]. This explains the registration results for both total and individual *TFL1* paralogues in Arabidopsis. For the GRN generated for *B. rapa*, we found one copy of *TFL1* on chromosome A10 that has a strong association with some *LFY* and *SOC1* paralogues. This suggests that this particular *TFL1* copy may have a greater role in performing the *TFL1* function, as it showed to be the most similar to *TFL1* in Arabidopsis. Overall, our analysis suggests that *TFL1* may not be functional in *B. oleracea*. In *B. rapa*, there is potential for non-functional gene paralogues and redundancy, but there is also evidence pointing to specific *TFL1* paralogues that potentially retain functionality. Functional investigations into these *TFL1* paralogues could clarify the roles of these specific gene copies.

Despite their rapid-cycling phenotype that does not require vernalisation, the *FLC* genes were expressed in both Brassica species. We investigated the floral repressor *FLC* in both species to understand the expressions of their copies and how these expressions are mediated in rapid-cycling lines. Our observations revealed that four out of five copies of *FLC* in *B. oleracea* exhibit the same dynamics as their homologue in Arabidopsis, suggesting that these genes may remain functional in the *B. oleracea* rapid-cycling line *DH1012*. In contrast, only two out of five copies of *FLC* were successfully registered between *B. rapa* and Arabidopsis. Building on the previous study by Calderwood *et al.* [85], we generated regulatory networks for *FLC* and its downstream genes in Arabidopsis, such as *SOC1* and other *SOC1*-interacting genes: *SVP* and *FUL*. Consistent with previous findings, we observed that none of the *FLC* copies, including a newly identified copy in *B. rapa*, were associated with *SOC1*. Instead, both copies of *FUL* showed strong associations with *SOC1* in *B. rapa*. Conversely, in *B. oleracea*, most *FLC* copies demonstrated strong associations with *SOC1*, particularly those located on chromosomes C01 and C09. This difference can be further observed when comparing the total *FLC* copies among the three species, with *B. rapa* showing a distinct pattern compared to Arabidopsis and *B. oleracea*. Further functional investigations into the roles of *FLC* paralogues could clarify their relative importance, offering valuable insights for crop improvement strategies. For example, knocking out the *FLC* paralogue with the strongest association in the GRN could potentially lead to faster flowering time in *B. oleracea DH1012*, demonstrating a practical application of our findings in breeding programs targeting specific desirable traits.

6 General Discussions

6.1 Chapter summaries

6.1.1 Chapter 2: Formulation of a statistical approach for comparing gene expression patterns

In Chapter 2, we presented the formulation of a novel statistical approach specifically developed to compare gene expression patterns. This method addresses the challenge of comparing two sets of time series data, each collected at potentially different time points. One dataset is referred to as the query, while the other is the reference. These datasets could represent gene expression measurements under distinct environmental conditions or data from homologous genes across different species. Since the time points at which data are collected may not align between the two datasets, a direct comparison of expression levels is not possible without employing an interpolation or extrapolation model that allows for the estimation of values at unmeasured time points, thereby facilitating the comparison of the datasets. Simple models, such as linear models, can provide an initial approach, but more complex models like polynomials or splines offer greater accuracy when dealing with non-linear trends. The core of the approach lies in determining whether both datasets can be described by a single model, implying shared dynamics, or if separate models are required, reflecting distinct underlying processes. This comparison is assessed by the statistical criterion BIC, which helps in choosing the model assumptions that best explain the data while balancing complexity and computational efficiency.

One effective option is to use spline models, particularly cubic B-splines with one knot, which offer the flexibility to capture non-linear trends while maintaining smooth transitions between different time segments. In this chapter, we also introduced curve registration, an approach we utilised when comparing gene expression data across different datasets with varying time ranges or developmental timescales. In such cases, it becomes necessary to apply transformations to align the datasets for meaningful comparison. Gene expression profiles often vary depending on the specific stage of development or experimental conditions, which can result in non-aligned datasets. To address this issue, a curve registration or a time transformation is applied to the query datasets, adjusting the time values while preserving the expression values. This transformation allows for a more direct comparison, making it possible to identify shared dynamical features across different biological contexts. The time transformation function, which adjusts the time axis of the query dataset, ensures that the differences between the reference and query datasets are minimised. This transformation can involve parameters such as shifts and stretch factors, which modify the time points in a biologically meaningful way. For example, a linear transformation with parameters for time shifts and stretches allows for simple adjustments, while a more complex model can account for non-linearities in the data.

To find the optimal values for the shift and stretch parameters that best align the datasets, an optimisation process is employed. This process involves maximising the likelihood of the model, we used techniques such as L-BFGS-B, Nelder–Mead, or Simulated Annealing. These optimisation algorithms search the parameter space to identify the values that minimise the difference between the transformed query dataset and the reference dataset. Once the optimal parameters are found, the model selection criterion BIC, is used to evaluate whether a single joint model can explain the datasets or if two separate models are required. A negative BIC difference would indicate that the transformation successfully aligns the datasets, supporting the hypothesis that the time series share common dynamics.

6.1.2 Chapter 3: Development of associated R package: *greatR* and methods testing using simulated data

In Chapter 3, we presented the development of the *greatR* package, an R package designed to facilitate the comparison of gene expression patterns, especially in time-course and developmental studies. The package implements the statistical approach described in Chapter 2, enabling researchers to align and compare gene expression dynamics in a flexible and efficient manner. By providing a user-friendly interface, *greatR* allows for the application of advanced statistical methods while also addressing common challenges such as time shifts, unequal sampling densities, and biological variability in gene expression data.

The design of *greatR* was guided by the need for a tool that could handle the complexity of gene expression data while being accessible to users with varying levels of expertise in statistical programming. The core functionality of *greatR* is centred on the alignment and comparison of gene expression time series. The package includes functions to process, such as scaling gene expression data and filtering low expression data, as well as performing pairwise comparisons of time series. Additionally, the package integrates methods for visualising the results of time-series alignments, providing clear and interpretable output that aids in the biological interpretation of the data. To optimise performance, the package utilised existing R packages for efficient computation.

To validate the functionality of *greatR*, we conducted extensive testing using simulated gene expression data. The simulated datasets were designed to mimic real-world biological scenarios, including time shifts, unequal sampling densities, and varying levels of noise. These simulations allowed for the evaluation of the package's performance in different contexts, ensuring that it can handle the challenges commonly encountered in comparative transcriptomics. The results of these tests demonstrated that *greatR* is capable of accurately aligning and comparing gene expression profiles under a wide range of conditions. The package successfully detected subtle differences in gene expression dynamics, even when the datasets had different sampling times or exhibited shifts in timing. The statistical tests implemented in the package were able to quantify the significance of these differences, providing users with robust tools for analysing time-course gene expression data.

6.1.3 Chapter 4: Understanding bract formation using comparative transcriptomics

Chapter 4 explored the regulatory pathways involved in bract development using RNA-Seq time series data across different developmental stages in two Arabidopsis accessions *Tsu-0* and *Col-0*. Our analysis compared gene expression at four morphologically matched stages, vegetative (V), late vegetative (L), transition (T), and floral (F), between the two accessions, even though the transition to flowering occurred at different absolute times. We found that key floral transition genes, such as *FT*, *AP1*, *LFY*, and *FUL*, exhibited similar expression patterns across accessions, suggesting they are not directly involved in the development of basal bracts. However, differentially expressed gene (DEG) analysis between stages and accessions revealed that the greatest molecular changes occurred during the transition from late vegetative to transition stages, particularly in the bract-forming accession, indicating a critical period for bract formation.

Despite identifying several candidate genes and pathways, including potential regulators like *SOC1*, the molecular mechanisms driving bract development remain unclear. A subsequent GO analysis highlighted processes like glycosinolate biosynthesis, although these do not appear to be directly linked to bract formation. Using comparative analysis, we identified 124 genes as potential bract regulators, which exhibited distinct expression patterns between bract-producing and bract-less stages. We also applied our approach to examine temporal shifts in gene expression dynamics between these two accessions. This revealed widespread transcriptional desynchronisation during the floral transition, with some genes expressed earlier or later in one accession compared to the other. These heterochronies suggest that bract development may result from timing differences in gene expression rather than inherent differences in gene profiles. Our findings suggest that bract formation is influenced by complex, temporally shifted gene activity, particularly during the transition stage, leading to variations in development across accessions.

6.1.4 Chapter 5: Exploring genetic variation in the floral transition between *B. rapa* and *B. oleracea* using comparative transcriptomics

In Chapter 5, we investigated the copy number variation of flowering time genes between two Brassica species, *B. rapa* and *B. oleracea*, and compared their gene expression dynamics to Arabidopsis. We identified orthologues of 306 flowering time-related genes in Arabidopsis, resulting in 487 and 524 genes in *B. rapa* and *B. oleracea*, respectively. Additionally, we observed that certain genes related to key pathways, such as photoperiodism and sugar signalling, are not present in both *B. rapa* and *B. oleracea*. This analysis showed that both *B. rapa* and *B. oleracea* have multiple copies of most Arabidopsis genes, with higher gene copy numbers in *B. oleracea*, possibly due to the presence of more transposable elements.

Our registration analysis of the genes involved in flowering time demonstrated that the differences in gene expression between Arabidopsis and the two Brassicas can largely be explained by desynchronisation in gene expression timing rather than differences in gene expression profiles. This result suggests a high degree of conservation in gene dynamics across the three species. We explored the conservation of gene paralogues in *B. rapa* and *B. oleracea*, identifying that most single-copy genes exhibit similar dynamics across the species, with a few genes showing species- or lineage-specific divergence. For key floral transition genes, such as *SOC1*, *AP1*, and *LFY*, all paralogues in *B. rapa* exhibited conserved expression dynamics with Arabidopsis, while in *B. oleracea*, one *SOC1* paralogue and two *TFL1* paralogues showed divergent dynamics, suggesting potential functional differences. Gene regulatory network analysis revealed distinct regulatory interactions between these genes in *B. rapa* and *B. oleracea*, further highlighting the possible divergence in gene regulation within these species.

For the floral repressor *FLC*, which plays a key role in the vernalisation pathway, we found that the majority of paralogues in *B. oleracea* showed conserved expression with Arabidopsis, while *B. rapa* exhibited divergent expression in some paralogues. These findings suggest that *FLC* paralogues may contribute differently to flowering regulation in these Brassicas, with implications for breeding strategies focused on manipulating flowering time. Overall, this chapter highlights the potential role of copy number variation and gene regulation in the evolution and adaptation of flowering time in Brassica species.

6.2 Outlook and limitations

The development of the method and the accompanying R package, *greatR*, has shown strong potential in accurately registering time series data and differentiating between distinct biological dynamics. To broaden the tool's accessibility and functionality, future enhancements could include making *greatR* available in other programming languages, such as Python, and developing a user-friendly interface, such as an R Shiny app, for researchers with limited programming experience. Additionally, expanding the underlying models to incorporate sinusoidal functions for cyclical biological processes, like circadian rhythms, would improve its applicability across a wider range of biological systems.

Despite its promising performance, *greatR* has certain limitations, particularly when dealing with a small number of time points. The current cubic B-spline fitting method performs well with five or more time points but can lead to overfitting when only fewer data points are available. When time points are limited, the model becomes too complex for the available information, leading to less reliable results. One potential solution would be to adapt the method by automatically selecting simpler models, such as quadratic polynomials, when data is limited. Alternatively, a more robust approach could involve fitting each pair of genes or curves to different models and selecting the most appropriate one for each case. Machine learning techniques, like Gaussian processes, could also be incorporated to replace spline-based methods. Gaussian processes offer

greater flexibility, particularly in handling missing or sparse time points, by modelling uncertainties directly in the data.

Another useful direction for future development would be to add gene clustering to the modelling process. Right now, each gene is modelled separately, but many genes probably follow similar expression patterns, especially during coordinated developmental processes. Grouping genes with similar time-course profiles and estimating parameters within each group could allow the model to share information across genes, leading to more stable and accurate fits. This idea is similar to what is done in differential expression tools like DESeq2 [40] and edgeR [41], where variance estimates are improved by sharing information across genes. Adding a similar approach to the time-series framework in *greatR* might improve the reliability of the curve fitting, especially for noisy data or when there are few replicates, and could help reveal biologically meaningful groups of co-expressed genes.

Additionally, implementing nested sampling as an alternative optimisation method could generate robust statistical evidence for selecting optimal registration parameters. While the current focus is on gene expression dynamics in plants, expanding *greatR* to handle a broader range of input data, such as general curves representing various processes across organisms, would significantly enhance its utility and make it a valuable resource for diverse scientific disciplines.

An important consideration in this approach is related to the estimation of standard deviation. Gene expression time series typically contain only a small number of replicates (often as few as 2 or 3 in many experimental setups) per time point, making it difficult to accurately calculate variability. As a result, we rely on an estimated standard deviation to account for this limitation, as calculating it directly from so few replicates can compromise the precision of downstream analyses. While our approximation performs well, refining this aspect would help better address uncertainty in cases with limited replicates. In *greatR*, users have the flexibility to specify the standard deviation if it is known. Another consideration involves the scaling method applied to the data. Scaling can significantly influence the inferred results as scaling can alter the distance between pairs of curves (query and reference). While this is a normal effect of scaling, it may impact how users interpret the alignment accuracy, and they should remain mindful of this when analysing their results. Finally, the percentage of overlapping time points specified by users plays an important role in shaping the alignment. Lower overlap may result in local alignments rather than global ones. Although the tool was designed with this flexibility in mind, users should consider how this may affect the overall interpretation of the alignment in their specific context.

After formulating and developing our approach, we applied it in a collaborative study aimed at understanding the genetic and developmental mechanisms underlying bract formation. This research integrated quantitative genetics, genomics, and transcriptomics, which involved significant contributions from both our group and collaborators. The identification of QTLs related to bract development in *Tsu-0* was achieved through the combined efforts of the Dieudonné *et al.* [1] team, and this provides a strong foundation for future research aimed at fine-mapping

these regions or conducting Genome-Wide Association Studies (GWAS) to pinpoint specific causal genes. Additionally, our contribution focused on the discovery of transcriptional heterochronies during the floral transition, which opens new avenues for studying how shifts in gene expression timing influence the evolution of plant traits, such as bract formation. These findings suggest that heterochrony may play a key role in shaping phenotypic diversity across species, potentially challenging traditional models of trait loss, such as Dollo's law. Looking forward, further exploration of the role of transcriptional heterochrony across different plant families, such as *Brassicaceae* and *Poaceae*, could yield valuable insights into the reactivation or loss of developmental traits. Expanding the comparative analysis between species that retain or have lost bracts will also be critical for understanding the evolutionary pathways that govern bract formation.

Several limitations need to be acknowledged, one of them arises from the reliance on transcriptomic data. While this study, with significant contributions from our team, identified potential transcriptional heterochronies and gene desynchronisations, confirming whether these shifts correspond to functional changes will require additional layers of analysis. Protein sequence comparisons, proteomic data, and experimental validation are essential to verify whether the observed gene expression shifts translate into functional changes in bract development. Furthermore, while this study suggests that heterochrony plays a role in bract formation, understanding the precise genetic mechanisms involved will require a deeper investigation into the regulatory networks at play, including potential epigenetic modifications that could influence gene expression timing. Additionally, the study's focus on *Tsu-0* and *Col-0* limits the generalisability of the findings across other species. Further research comparing different accessions or species that exhibit variation in bract formation will be essential to validate the broader applicability of these results. Addressing these limitations through continued collaboration will provide a more complete understanding of the genetic and developmental pathways shaping bract evolution and offer new opportunities for advancing our knowledge of plant development.

Another application of our method was to compare the gene expression dynamics across the related species Arabidopsis, *B. rapa*, and *B. oleracea*. By focusing on flowering time genes and extending the analysis to include genes with both single and multiple copies, we have advanced our understanding of the molecular differences between these crops. Future studies could build on this work by expanding beyond flowering time genes to explore other developmental and stress-related pathways, which may further clarify the evolutionary adaptations of these species. Additionally, the potential for lineage-specific divergence and Brassica-predominant dynamics, identified through gene expression analysis, presents an exciting opportunity for exploring the functional evolution of these genes in response to environmental pressures. To confirm the functionality of genes where similar expression dynamics were observed, comparative analysis of protein sequences will be necessary. This will help determine whether the conserved gene expression is also reflected in protein structure and function, further validating the inferred roles of these genes in the regulatory networks of these species. Further development of gene regulatory networks (GRNs), as demonstrated with *SOC1*, *FLC*, and *TFL1* in this study, offers a valuable approach for identifying gene copies that may retain similar functions across species.

This information can be highly beneficial for crop improvement efforts. For instance, the identification of *B. rapa* specific genes with altered expression dynamics highlights potential targets for selective breeding programs focused on enhancing flowering time and other important traits. The observed functional redundancy and specialisation among paralogues emphasise the need to further explore the roles of gene copies, particularly in polyploid crops, where gene duplication plays a significant role in driving adaptation and influencing phenotypic traits. Looking into the protein families of paralogues that are switched on at different times may help reveal whether these timing differences reflect functional changes, especially in families known to be structurally flexible and able to perform multiple functions.

Several limitations must be considered when interpreting the findings of this study. While we successfully analysed copy number variation and gene expression in *B. rapa* and *B. oleracea*, the presence of multiple gene copies complicates the interpretation of individual gene functions. The current study focused primarily on the transcriptomic level, which does not fully capture the functional consequences of gene copy variation. Future work should integrate additional layers of data, such as protein activity or epigenetic modifications, to provide a clearer understanding of how these gene copies contribute to phenotypic traits. Lastly, while the regulatory networks generated in this study provide valuable insights, they rely on correlative data and would benefit from experimental validation. Functional experiments, such as gene knockouts, are essential to confirm the roles of specific genes, particularly those identified as Brassica- or *B. rapa*-specific. These experiments would strengthen the causal links between gene expression patterns and phenotypic outcomes, helping to refine the understanding of gene function in these important crop species.

6.3 Concluding remarks

The primary goal of this project was to develop a statistical tool for comparing pairs of gene expression dynamics. This objective has been successfully accomplished through the formulation of the method, its implementation into a practical and user-friendly R package, and subsequent validation using both simulated datasets and real biological data.

In this study, we present a comprehensive statistical approach to analysing gene expression dynamics using curve registration, building on the work of Calderwood *et al.* [85]. This methodology has been implemented as an R package, making it accessible to a broader audience. We validated the approach using both positive and negative control datasets with varying time points and noise levels. As expected, increasing noise levels made it more challenging to detect similarities between gene expression dynamics, as the dynamics deteriorated. We used this approach to analyse transcriptomic data from two Arabidopsis natural genotypes, *Tsu-0* and *Col-0*, to investigate bract formation in Brassicaceae. Bract formation, a trait present in some species, appears during the floral transition when gene expression becomes less tightly regulated. Our findings indicate that natural genetic variations and transcriptional heterochronies, shifts in the timing of gene expression, may explain this phenomenon. While no specific bract-regulating genes

were identified, the desynchronisation of gene expression during flowering likely contributes to the development of traits like bracts, offering a potential mechanism for evolutionary divergence in plant morphology. We also applied the method to compare Arabidopsis with two Brassica species, *B. rapa* and *B. oleracea*, focusing on paralogues involved in flowering time. Our findings confirmed that single-copy paralogues in *B. rapa* and *B. oleracea* are generally conserved with Arabidopsis, though we identified some genes with species-specific dynamics. In contrast, for key floral regulators with multiple paralogues, we observed differences in gene dynamics, suggesting the evolution of potential novel functions across species.

This thesis describes a novel statistical approach that has proven effective and user-friendly, while also providing insights into the dynamic differences between *B. rapa*, *B. oleracea*, and Arabidopsis. These findings represent a preliminary step toward understanding the broader gene regulatory networks in *B. rapa* and *B. oleracea* compared to Arabidopsis, as well as the relationships between these species for genes involved in other pathways. While our tool and analysis have provided valuable insights, further research is needed to confirm the function by examining not only gene expression dynamics but also protein sequences and activity across these species. Additionally, our work on *Tsu-0* and *Col-0* offers valuable insights into how transcriptional heterochrony, or shifts in gene expression timing, may explain the transient formation of bracts in *Tsu-0* and related species. Although specific genes responsible for bract formation have not yet been identified, more comprehensive comparative studies between other members of the *Brassicaceae* could help uncover the genetic basis of bract development.

S | Supplementary

Gene ID	Symbol
AT1G10340	AT1G10340
AT1G10600	AMSH2
AT1G10800	AT1G10800
AT1G12070	AT1G12070
AT1G12090	ELP
AT1G12860	SCRM2
AT1G14000	VIK
AT1G63650	EGL3
AT1G63840	AT1G63840
AT1G63860	AT1G63860
AT1G63940	MDAR6
AT1G64940	CYP89A6
AT1G65050	AT1G65050
AT1G65130	AT1G65130
AT1G65260	PTAC4
AT1G65370	AT1G65370
AT5G25140	CYP71B13
AT5G27690	AT5G27690
AT5G41670	AT5G41670
AT5G41950	AT5G41950
AT5G42146	AT5G42146
AT5G43210	AT5G43210
AT5G43380	TOPP6
AT5G43910	AT5G43910
AT5G44070	CAD1
AT5G44568	AT5G44568
AT5G45060	AT5G45060
AT5G45280	AT5G45280
AT5G45470	AT5G45470
AT5G45510	AT5G45510
AT5G46370	KCO2
AT5G47130	AT5G47130
AT5G47470	UMAMIT7

Table S.1: List of 33 candidate genes potentially involved in bract formation, identified by cross-referencing differentially expressed genes with QTL-mapping data [1].

Locus	TAIR Alias	Cluster	In mapped QTL	GO terms	Overlap with DE in *jagged-5d*
AT1G03495		bract positive			
AT1G03940		bract positive			
AT1G08050		bract negative			
AT1G12160	FMOGS-OX7	bract negative	1a*	SA response	
AT1G14720	ATXTH28, EXGT-A2, XTH28, XTR2	bract positive			
AT1G17330	CN-PDE1, PDE1	bract negative			
AT1G21100	IGMT1	bract positive			
AT1G21250	AtWAK1, PRO25, WAK1	bract negative		SA response	yes
AT1G22480		bract negative			yes
AT1G23480	ATCSLA03, ATCSLA3, CSLA03, CSLA3	bract positive			
AT1G23560		bract positive			
AT1G27340	LCR	bract positive			yes
AT1G29820		bract positive			
AT1G30320		bract negative			yes
AT1G30540	GNK	bract positive			
AT1G53540	HSP17.6C	bract positive			yes
AT1G60920	AGL55	bract positive			
AT1G62660	VI1	bract positive			
AT1G62960	ACS10	bract negative			
AT1G64780	AMT1;2, ATAMT1;2	bract positive	1b*		
AT1G64840	AtFDA3	bract negative	1b*		
AT1G65470	FAS1, FUGU2, NFB2	bract negative	1b*		
AT1G70950	WDL7	bract negative			yes
AT1G72810	TSY	bract positive			
AT1G73060	LPA3	bract positive			yes
AT1G76900	AtTLP1, TLP1	bract positive			
AT1G78210		bract negative			
AT1G79450	ALIS5	bract negative			
AT1G80280		bract negative			
AT2G02061		bract negative			
AT2G04090		bract positive			
AT2G04495		bract negative			
AT2G04515		bract negative			
AT2G14560	LURP1	bract negative		SA response	
AT2G17470	ALMT6, AtALMT6	bract positive			
AT2G19800	MIOX2	bract negative			
AT2G22770	NAI1	bract positive			yes
AT2G25580	MEF8	bract negative			
AT2G30010	TBL45	bract positive			
AT2G30740	CARK8	bract positive			yes
AT2G30750	CYP71A12	bract negative			
AT2G32680	AtRLP23, RLP23	bract negative			
AT2G33080	AtRLP28, RLP28	bract negative			
AT2G39750		bract positive			
AT2G42990		bract negative			
AT2G43570	CHI	bract negative			
AT2G44180	MAP2A	bract negative			
AT2G45660	AGL20, ATSOC1, SOC1	bract negative			
AT2G46930	PAE3	bract positive			
AT3G09540		bract positive			
AT3G09780	ATCRR1, CCR1	bract positive			
AT3G12040		bract negative			
AT3G12220	scpl16	bract negative			
AT3G18550	AtBRC1, ATTCP18, BRC1, TCP18	bract negative			
AT3G19010		bract negative			
AT3G21950		bract negative			
AT3G25050	AtXTH3, XTH3	bract positive			
AT3G27880		bract positive			
AT3G28880		bract negative			
AT3G29575	AFP3	bract negative			
AT3G29590	AT5MAT	bract positive		anthocyanin	
AT3G44610	AGC1-12	bract negative			

Table S.2: List of 124 potential bract regulators identified through comparative gene expression analysis on bract-present and -absent developmental stages (see Methods 4.2.3).

Locus	TAIR Alias	Cluster	In mapped QTL	GO terms	Overlap with DE in *jagged-5d*
AT3G44830		bract positive			
AT3G48080		bract negative			yes
AT3G51860	ATCAX3, ATHCX1, CAX1-LIKE, CAX3	bract negative			
AT3G51890	CLC3	bract negative			
AT3G57240	AtBG3, BG3, GNS3	bract negative			yes
AT3G61280		bract negative			
AT4G00200	AHL7	bract negative			
AT4G01370	ATMPK4, MAPK4, MPK4	bract negative		SA response	yes
AT4G02820	RTP7	bract negative			
AT4G04750	IF1	bract positive			
AT4G11630		bract negative			
AT4G14090		bract positive		anthocyanin	
AT4G14510	ATCFM3B, CFM3B, SPRT2	bract negative			
AT4G14660	NRPE7	bract negative			
AT4G14980		bract negative			yes
AT4G15110	CYP97B3	bract positive			yes
AT4G17615	ATCBL1, CBL1, SCABP5	bract negative			
AT4G18810		bract positive			yes
AT4G19590		bract negative			
AT4G21540	SPHK1	bract negative			
AT4G22880	ANS, LDOX, TDS4, TT18	bract positive		anthocyanin	
AT4G25110	AtMC2, AtMCP1c, MC2, MCP1c	bract negative			
AT4G33150	LKR, LKR/SDH, SDH	bract positive			yes
AT4G34900	ATXDH2, XDH2	bract positive			
AT4G37320	CYP81D5	bract positive			
AT5G01370	ACI1, TRM29	bract negative			yes
AT5G02370		bract negative			
AT5G06450	RICE2	bract negative			
AT5G08275		bract negative			
AT5G09290		bract negative			
AT5G09390	CD2b	bract negative			
AT5G09850	MED26C	bract positive			
AT5G10760	AED1	bract negative			yes
AT5G11010	GRC3	bract negative			
AT5G11280		bract negative			
AT5G16010		bract positive			
AT5G17220	ATGSTF12, GST26, GSTF12, TT19	bract positive		anthocyanin	
AT5G19040	ATIPT5, IPT5	bract positive			
AT5G20400		bract negative			
AT5G20430		bract negative			yes
AT5G20550		bract negative			
AT5G24210	PRLIP1	bract negative			
AT5G24530	AtDMR6, DMR6	bract negative		SA response	
AT5G24850	CRY3	bract negative			
AT5G25250	FLOT1	bract negative	5a		
AT5G32460		bract negative			
AT5G35450		bract positive			
AT5G38840		bract negative			
AT5G41130	ELT5	bract positive	5b		
AT5G41160	ATPUP12, PUP12	bract negative	5b		
AT5G41750		bract negative	5b		
AT5G42800	DFR, M318, TT3	bract positive	5b	anthocyanin	
AT5G44560	VPS2.2	bract negative	5b		
AT5G45000		bract negative	5b		
AT5G46050	AtNPF5.2, ATPTR3, NPF5.2, PTR3	bract negative	5b	SA response	
AT5G48657		bract negative			
AT5G49760	CARD1, HPCA1	bract negative			
AT5G51380		bract negative			
AT5G52540		bract positive			
AT5G53390	FOP1, WSD11	bract negative			
AT5G54060	UF3GT, UGT79B1	bract positive		anthocyanin	
AT5G54610	ANK, BDA1	bract negative		SA response	

Table S.2: List of 124 potential bract regulators identified through comparative gene expression analysis on bract-present and -absent developmental stages (see Methods 4.2.3).

Gene ID	Symbol
AT1G02670	
AT1G03170	FAF2, FTM5
AT1G09440	
AT1G25370	
AT1G33600	
AT1G34260	FAB1D
AT1G64940	CYP89A6
AT1G80000	BTZ1
AT1G80350	AAA1, ATKTN1
AT2G05760	NAT1
AT2G20850	SRF1
AT2G27010	CYP705A9
AT2G36290	
AT2G40380	PRA1.B2
AT2G40570	
AT2G40650	PRP38
AT3G11860	
AT3G21560	BRT1, UGT84A2
AT3G46370	
AT3G52710	
AT3G55515	DVL8, RTFL7
AT3G62460	
AT4G03460	
AT4G08770	Prx37
AT4G13630	MyoB13
AT4G17770	ATTPS5, TPS5
AT4G20010	OSB2, PTAC9
AT4G22390	
AT4G36850	
AT4G37820	
AT5G02040	PRA1.A1
AT5G07870	
AT5G35180	
AT5G39330	
AT5G40382	
AT5G44560	VPS2.2
AT5G44570	CWM2
AT5G45275	
AT5G47960	ATRABA4C, RABA4C
AT5G51210	OLEO3
AT5G59305	
AT5G60270	LecRK-I.7
AT5G64430	

Table S.3: List of non-registered genes between two Arabidopsis accessions *Col-0* and *Tsu-0*.

Genes	Case	Genes	Case
HDA9	*B. rapa* specific	GIS5	Conserved
YAF9A	*B. rapa* specific	HAM2	Conserved
CSTF77	Brassica pre-dominant	HLP1	Conserved
DCL4	Brassica pre-dominant	HUB2	Conserved
DET1	Brassica pre-dominant	JMJ14	Conserved
HTA8	Brassica pre-dominant	JMJ32	Conserved
JMJ30	Brassica pre-dominant	LD	Conserved
LDL2	Brassica pre-dominant	LWD1	Conserved
RFI2	Brassica pre-dominant	MED16	Conserved
RGL2	Brassica pre-dominant	MED18	Conserved
RRP6L2	Brassica pre-dominant	MRG2	Conserved
SPA4	Brassica pre-dominant	MYR1	Conserved
TOE3	Brassica pre-dominant	NF-YA4	Conserved
XAL2	Brassica pre-dominant	NF-YB2	Conserved
ATX1	Brassica specific	OTS1	Conserved
PRR3	Brassica specific	OTS2	Conserved
ATX2	Conserved	PHP	Conserved
ATXR7	Conserved	PHYC	Conserved
CBP20	Conserved	PUB12	Conserved
CCA1	Conserved	REF6	Conserved
CDF3	Conserved	RRP6L1	Conserved
CLF	Conserved	RUP2	Conserved
CRY1	Conserved	RVE2	Conserved
EFS	Conserved	SDG26	Conserved
ELF6	Conserved	SDG7	Conserved
ESD7	Conserved	SHL1	Conserved
FBH3	Conserved	SUVR5	Conserved
FCA	Conserved	TEM2	Conserved
FIO1	Conserved	VIL1	Conserved
FLX	Conserved	VIL3	Conserved
FLX4	Conserved	VRN2	Conserved
FRL2	Conserved	AtNDX	Lineage-specific divergence
GA2	Conserved	DCL3	Lineage-specific divergence
GASA5	Conserved	FLD	Lineage-specific divergence
GI	Conserved	GA2ox4	Lineage-specific divergence
GID1A	Conserved	GID1C	Lineage-specific divergence

Table S.4: List of 72 flowering time genes present in single copies in both *B. rapa* and *B. oleracea*.

Species	Gene name	TAIR symbol	Brassica ID	Chromosome
B. oleracea	AGL24	AT4G24540	BO1G039080	C01
B. oleracea	AGL24	AT4G24540	BO7G109590	C07
B. rapa	AGL24	AT4G24540	BRAA03G051930.3C	A03
B. oleracea	AP1	AT1G69120	BO6G095760	C06
B. oleracea	AP1	AT1G69120	BO6G108600	C06
B. rapa	AP1	AT1G69120	BRAA02G018970.3C	A02
B. rapa	AP1	AT1G69120	BRAA07G030470.3C	A07
B. rapa	AP1	AT1G69120	BRAA07G034100.3C	A07
B. oleracea	LFY	AT5G61850	BO2G161690	C02
B. oleracea	LFY	AT5G61850	BO3G109270	C03
B. rapa	LFY	AT5G61850	BRAA02G043220.3C	A02
B. rapa	LFY	AT5G61850	BRAA06G025360.3C	A06
B. oleracea	SOC1	AT2G45660	BO3G038880	CO3
B. oleracea	SOC1	AT2G45660	BO4G024850	C04
B. oleracea	SOC1	AT2G45660	BO4G195720	C04
B. rapa	SOC1	AT2G45660	BRAA03G023790.3C	A03
B. rapa	SOC1	AT2G45660	BRAA05G005370.3C	A05
B. rapa	SOC1	AT2G45660	BRAA04G031640.3C	A04
B. oleracea	TFL1	AT5G03840	BO00983S080	C09
B. oleracea	TFL1	AT5G03840	BO9G181670	C09
B. oleracea	TFL1	AT5G03840	BO2G005260	C02
B. rapa	TFL1	AT5G03840	BRAA02G001100.3C	AO2
B. rapa	TFL1	AT5G03840	BRAA03G001350.3C	A03
B. rapa	TFL1	AT5G03840	BRAA10G032420.3C	A10

Table S.5: Orthologues table of *B. rapa* and *B. oleracea* mapped to Arabidopsis FLOR-ID genes *AGL24*, *AP1*, *LFY*, *SOC1*, and *TFL1*.

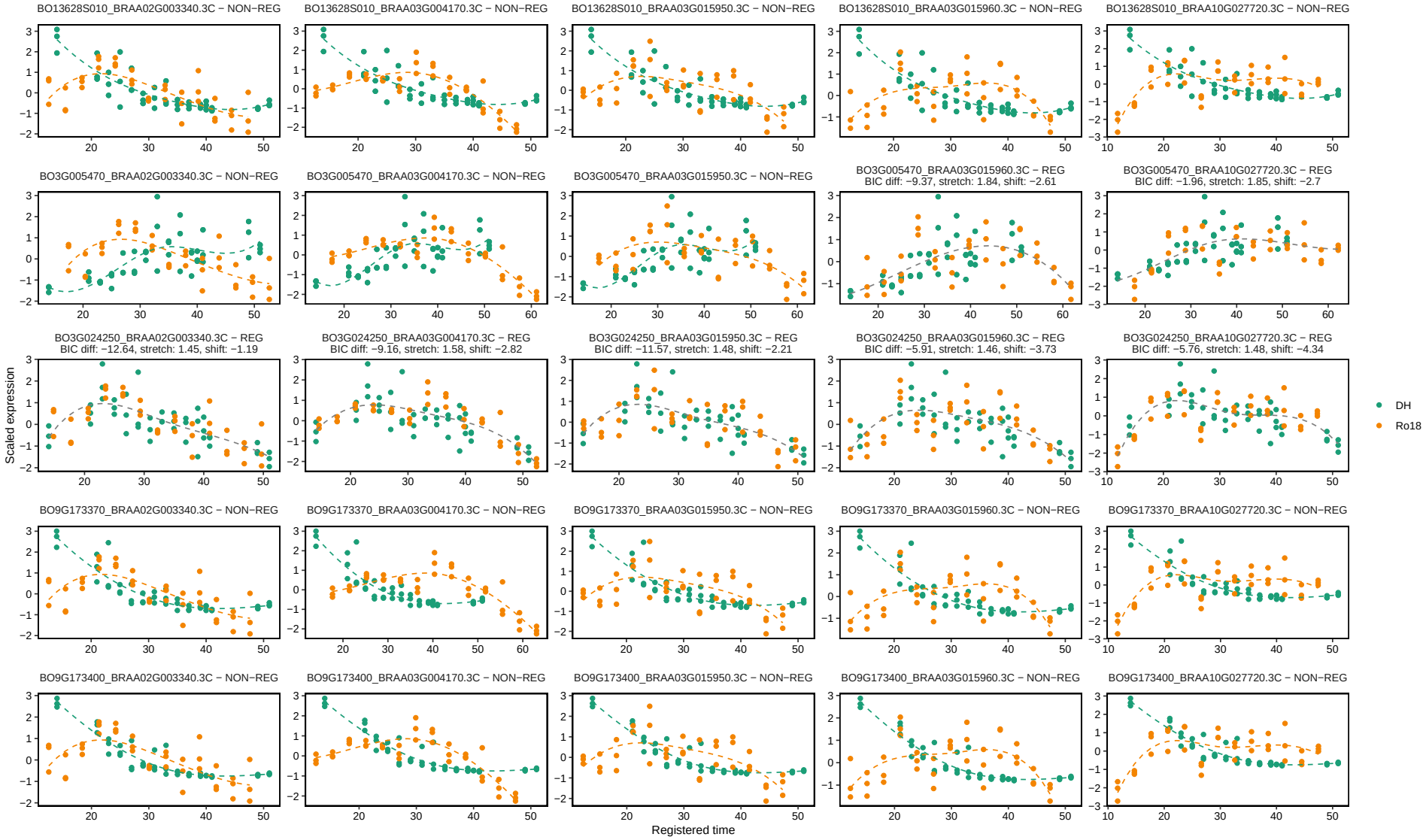


Figure S.1: Pairwise registration results of *FLC* paralogues between *B. rapa* and *B. oleracea*.

Bibliography

[1] S. Dieudonné *et al.* Natural variation suggests new mechanisms for bract development in Arabidopsis, desynchronising bract suppression from the floral transition, 08/2024. DOI: 10.1101/2024.08.12.607587.

[2] R. Kristianingsih *et al.* Comparing gene expression dynamics over different developmental timescales with the R package *greatR*. In preparation, 2024.

[3] R. Kristianingsih. *greatR: Gene Registration from Expression and Time-Courses in R*. R package version 2.1.0. 2024.

[4] S. H. G. Symons. *Analysis and Visualization of Gene Expression Data*. Thesis, 2011.

[5] J. E. Krebs, E. S. Goldstein and S. T. Kilpatrick. *Lewin's genes XII*. eng. Jones & Bartlett Learning, Burlington, MA, 2018.

[6] C. Wahlestedt. Targeting long non-coding RNA to therapeutically upregulate gene expression. *Nature Reviews Drug Discovery*, 12(6):433–446, 05/2013. DOI: 10.1038/nrd4018.

[7] D.-W. Doug Chung and K. Le Roch. *Genome-wide analysis of gene expression*. In *Encyclopedia of Biological Chemistry*. Elsevier, 2013, pages 369–374. DOI: 10.1016/b978-0-12-378630-2.00634-4.

[8] P. Muller and R. ten Cate. *Oligoarticular and polyarticular juvenile idiopathic arthritis*. In *Pediatrics in Systemic Autoimmune Diseases*. Elsevier, 2016, pages 1–30. DOI: 10.1016/b978-0-444-63596-9.00001-3.

[9] K. P. Singh *et al.* Mechanisms and measurement of changes in gene expression. *Biological Research For Nursing*, 20(4):369–382, 04/2018. DOI: 10.1177/1099800418772161.

[10] M.-C. King and A. C. Wilson. Evolution at two levels in humans and chimpanzees: their macromolecules are so alike that regulatory mutations may account for their biological differences. *Science*, 188(4184):107–116, 04/1975. DOI: 10.1126/science.1090005.

[11] Y.-C. Chen. *Introductory chapter: gene expression and phenotypic traits*. In *Gene Expression and Phenotypic Traits*. IntechOpen, 04/2020. DOI: 10.5772/intechopen.89863.

[12] M. Rippy *et al.* Plant functional traits and viewer characteristics co-regulate cultural services provisioning by stormwater bioretention. *Ecological Engineering*, 168:106284, 10/2021. DOI: 10.1016/j.ecoleng.2021.106284.

[13] C. Trapnell *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28(5):511–515, 05/2010. DOI: 10.1038/nbt.1621.

[14] Z. Alpay Savasan *et al. Advances in cerebral palsy biomarkers*. In *Advances in Clinical Chemistry*. Elsevier, 2021, pages 139–169. DOI: 10.1016/bs.acc.2020.04.006.

[15] R. Lowe *et al.* Transcriptomics technologies. *PLOS Computational Biology*, 13(5):e1005457, 05/2017. DOI: 10.1371/journal.pcbi.1005457.

[16] M. N. Bernstein. *Computational methods for transcriptome-based cellular phenotyping.* Thesis, 2019.

[17] Z. Wang, M. Gerstein and M. Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, 01/2009. DOI: 10.1038/nrg2484.

[18] D. Kim, B. Langmead and S. L. Salzberg. HISAT: a fast spliced aligner with low memory requirements. *Nature Methods*, 12(4):357–360, 03/2015. DOI: 10.1038/nmeth.3317.

[19] A. Dobin *et al.* STAR: ultrafast universal RNA-Seq aligner. *Bioinformatics*, 29(1):15–21, 10/2012. DOI: 10.1093/bioinformatics/bts635.

[20] C. Trapnell, L. Pachter and S. L. Salzberg. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9):1105–1111, 03/2009. DOI: 10.1093/bioinformatics/btp120.

[21] B. J. Haas and M. C. Zody. Advancing RNA-Seq analysis. *Nature Biotechnology*, 28(5):421–423, 05/2010. DOI: 10.1038/nbt0510-421.

[22] F. Ozsolak and P. M. Milos. RNA sequencing: advances, challenges and opportunities. *Nature Reviews Genetics*, 12(2):87–98, 12/2010. DOI: 10.1038/nrg2934.

[23] Illumina. RNA-Seq vs microarrays: compare technologies. <https://emea.illumina.com/science/technology/next-generation-sequencing/beginners/advantages/rna-seq-vs-arrays.html> (visited on 04/06/2024).

[24] Z. Bar-Joseph, A. Gitter and I. Simon. Studying and modelling dynamic biological processes using time-series gene expression data. *Nature Reviews Genetics*, 13(8):552–564, 07/2012. DOI: 10.1038/nrg3244.

[25] I. Androulakis, E. Yang and R. Almon. Analysis of time-series gene expression data: methods, challenges, and opportunities. *Annual Review of Biomedical Engineering*, 9(1):205–228, 08/2007. DOI: 10.1146/annurev.bioeng.9.060906.151904.

[26] Z. Bar-Joseph. Analyzing time series gene expression data. *Bioinformatics*, 20(16):2493–2503, 05/2004. DOI: 10.1093/bioinformatics/bth283.

[27] I. Simon *et al.* Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell*, 106(6):697–708, 09/2001. DOI: 10.1016/s0092-8674(01)00494-9.

[28] M. L. Whitfield *et al.* Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Molecular Biology of the Cell*, 13(6):1977–2000, 06/2002. M. J. Solomon, editor. DOI: 10.1091/mbc.02-02-0030.

[29] K.-F. Storch *et al.* Extensive and divergent circadian gene expression in liver and heart. *Nature*, 417(6884):78–83, 04/2002. DOI: 10.1038/nature744.

[30] S. Panda *et al.* Coordinated transcription of key pathways in the mouse by the circadian clock. *Cell*, 109(3):307–320, 05/2002. DOI: 10.1016/s0092-8674(02)00722-5.

[31] F. Yi *et al.* Time-series transcriptomics reveals a drought-responsive temporal network and crosstalk between drought stress and the circadian clock in foxtail millet. *The Plant Journal*, 110(4):1213–1228, 04/2022. DOI: 10.1111/tpj.15725.

[32] N. Rathore *et al.* Time-series RNA-Seq transcriptome profiling reveals novel insights about cold acclimation and de-acclimation processes in an evergreen shrub of high altitude. *Scientific Reports*, 12(1), 09/2022. DOI: 10.1038/s41598-022-19834-w.

[33] W. Berkofsky-Fessler *et al.* Preclinical biomarkers for a cyclin-dependent kinase inhibitor translate to candidate pharmacodynamic biomarkers in phase i patients. *Molecular Cancer Therapeutics*, 8(9):2517–2525, 09/2009. DOI: 10.1158/1535-7163.mct-09-0083.

[34] L. Siles, P. Eastmond and S. Kurup. Big data from small tissues: extraction of high-quality RNA for RNA-sequencing from different oilseed Brassica seed tissues during seed development. *Plant Methods*, 16(1), 06/2020. DOI: 10.1186/s13007-020-00626-0.

[35] A. Hoshino *et al.* Molecular anatomy of the developing human retina. *Developmental Cell*, 43(6):763–779.e4, 12/2017. DOI: 10.1016/j.devcel.2017.10.029.

[36] G. J. Nau *et al.* Human macrophage activation programs induced by bacterial pathogens. *Proceedings of the National Academy of Sciences*, 99(3):1503–1508, 01/2002. DOI: 10.1073/pnas.022649799.

[37] D. Bautista *et al.* Comprehensive time-series analysis of the gene expression profile in a susceptible cultivar of tree tomato (solanum betaceum) during the infection of phytophthora betacei. *Frontiers in Plant Science*, 12, 10/2021. DOI: 10.3389/fpls.2021.730251.

[38] Z. Yida and L. Shu. A general review on time-series gene expression data. *ACM Transcactions on Applied Perception*, 2(3), 05/2010.

[39] G. K. Smyth. *Limma: linear models for microarray data.* In *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. R. Gentleman *et al.*, editors. Springer New York, New York, NY, 2005, pages 397–420. DOI: 10.1007/0-387-29362-0_23.

[40] M. I. Love, W. Huber and S. Anders. Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *Genome Biology*, 15(12), 12/2014. DOI: 10.1186/s13059-014-0550-8.

[41] M. D. Robinson, D. J. McCarthy and G. K. Smyth. *edgeR*: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–40, 2010. DOI: 10.1093/bioinformatics/btp616.

[42] T. J. Hardcastle and K. A. Kelly. baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*, 11(1), 08/2010. DOI: 10.1186/1471-2105-11-422.

[43] J. M. Straube. *Development of statistical tools for integrating time course 'omics' data.* Thesis, 2016.

[44] M. J. Aryee *et al.* An improved empirical bayes approach to estimating differential gene expression in microarray time-course data: BETR (Bayesian Estimation of Temporal Regulation). *BMC Bioinformatics*, 10(1), 12/2009. DOI: 10.1186/1471-2105-10-409.

[45] J. C. Pinheiro and D. M. Bates. *Mixed-Effects Models in S and S-PLUS*. Springer, New York, 2000. DOI: 10.1007/b98882.

[46] K. Y. Yeung, M. Medvedovic and R. E. Bumgarner. From co-expression to co-regulation: how many microarray experiments do we need? *Genome Biology*, 5(7), 06/2004. DOI: 10.1186/gb-2004-5-7-r48.

[47] M. B. Eisen *et al.* Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25):14863–14868, 12/1998. DOI: 10.1073/pnas.95.25.14863.

[48] J. A. Hartigan. *Clustering algorithms*. John Wiley & Sons, Inc., 1975.

[49] J. Vesanto and E. Alhoniemi. Clustering of the self-organizing map. *IEEE Transactions on Neural Networks*, 11(3):586–600, 05/2000. DOI: 10.1109/72.846731.

[50] E. Nushi *et al.* Bayesian model-based method for clustering gene expression time series with multiple replicates, 05/2024. DOI: 10.1101/2024.05.23.595463.

[51] I. C. McDowell *et al.* Clustering gene expression time series data using an infinite gaussian process mixture model. *PLOS Computational Biology*, 14(1):e1005896, 01/2018. Q. Nie, editor. DOI: 10.1371/journal.pcbi.1005896.

[52] M. F. Ramoni, P. Sebastiani and I. S. Kohane. Cluster analysis of gene expression dynamics. *Proceedings of the National Academy of Sciences*, 99(14):9121–9126, 06/2002. DOI: 10.1073/pnas.132656399.

[53] A. Schliep, C. Steinhoff and A. Schönhuth. Robust inference of groups in gene expression time-courses using mixtures of HMMs. *Bioinformatics*, 20(suppl_1):i283–i289, 08/2004. DOI: 10.1093/bioinformatics/bth937.

[54] J. Ernst and Z. Bar-Joseph. STEM: a tool for the analysis of short time series gene expression data. *BMC Bioinformatics*, 7(1), 04/2006. DOI: 10.1186/1471-2105-7-191.

[55] Y. Dai *et al.* Comparative transcriptome analysis of gene expression and regulatory characteristics associated with different vernalization periods in *Brassica rapa*. *Genes*, 11(4):392, 04/2020. DOI: 10.3390/genes11040392.

[56] J. Jiang *et al.* Transcriptomic comparison between developing seeds of yellow- and black-seeded *Brassica napus* reveals that genes influence seed quality. *BMC Plant Biology*, 19(1), 05/2019. DOI: 10.1186/s12870-019-1821-z.

[57] Expansion of the gene ontology knowledgebase and resources. *Nucleic Acids Research*, 45(D1):D331–D338, 11/2016. DOI: 10.1093/nar/gkw1108.

[58] M. Pomaznoy, B. Ha and B. Peters. GOnet: a tool for interactive Gene Ontology analysis. *BMC Bioinformatics*, 19(1), 12/2018. DOI: 10.1186/s12859-018-2533-3.

[59] A. Subramanian *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 09/2005. DOI: 10.1073/pnas.0506580102.

[60] H. Mi *et al.* PANTHER version 11: expanded annotation data from gene ontology and reactome pathways, and data analysis tool enhancements. *Nucleic Acids Research*, 45(D1):D183–D189, 11/2016. DOI: 10.1093/nar/gkw1138.

[61] E. Eden *et al.* GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*, 10(1), 02/2009. DOI: 10.1186/1471-2105-10-48.

[62] D. W. Huang, B. T. Sherman and R. A. Lempicki. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, 4(1):44–57, 12/2008. DOI: 10.1038/nprot.2008.211.

[63] S. Carbon *et al.* AmiGO: online access to ontology and annotation data. *Bioinformatics*, 25(2):288–289, 11/2008. DOI: 10.1093/bioinformatics/btn615.

[64] D. V. Klopfenstein *et al.* GOATOOLS: a Python library for gene ontology analyses. *Scientific Reports*, 8(1), 07/2018. DOI: 10.1038/s41598-018-28948-z.

[65] O. Ozisik, M. Térézol and A. Baudot. Orsum: a python package for filtering and comparing enrichment analyses using a simple principle. *BMC Bioinformatics*, 23(1), 07/2022. DOI: 10.1186/s12859-022-04828-2.

[66] S. Falcon and R. Gentleman. Using GOstats to test gene lists for GO term association. *Bioinformatics*, 23(2):257–258, 11/2006. DOI: 10.1093/bioinformatics/btl567.

[67] G. Yu *et al.* *clusterProfiler*: an R package for comparing biological themes among gene clusters. *OMICS: A Journal of Integrative Biology*, 16(5):284–287, 05/2012. DOI: 10.1089/omi.2011.0118.

[68] T. R. Golub *et al.* Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, 10/1999. DOI: 10.1126/science.286.5439.531.

[69] S. Kim. Ppcor: an r package for a fast calculation to semi-partial correlation coefficients. *Communications for Statistical Applications and Methods*, 22(6):665–674, 11/2015. DOI: 10.5351/csam.2015.22.6.665.

[70] V. A. Huynh-Thu *et al.* Inferring regulatory networks from expression data using tree-based methods. *PLoS ONE*, 5(9):e12776, 09/2010. M. Isalan, editor. DOI: 10.1371/journal.pone.0012776.

[71] J. Cirrone *et al.* OutPredict: multiple datasets can improve prediction of expression and inference of causality. *Scientific Reports*, 10(1), 04/2020. DOI: 10.1038/s41598-020-63347-3.

[72] J. Ding and Z. Bar-Joseph. Analysis of time-series regulatory networks. *Current Opinion in Systems Biology*, 21:16–24, 06/2020. DOI: 10.1016/j.coisb.2020.07.005.

[73] J. Aach and G. M. Church. Aligning gene expression time series with time warping algorithms. *Bioinformatics*, 17(6):495–508, 06/2001. DOI: 10.1093/bioinformatics/17.6.495.

[74] H. Vanhaeren, N. Gonzalez and D. Inzé. A journey through a leaf: phenomics analysis of leaf growth in *Arabidopsis thaliana*. *The Arabidopsis Book*, 13:e0181, 01/2015. DOI: 10.1199/tab.0181.

[75] W. Zheng *et al.* Regulatory variation within and between species. *Annual Review of Genomics and Human Genetics*, 12(1):327–346, 09/2011. DOI: 10.1146/annurev-genom-082908-150139.

[76] G. Li *et al.* Comparative analysis of *KNOX* genes and their expression patterns under various treatments in *Dendrobium huoshanense*. *Frontiers in Plant Science*, 14, 10/2023. DOI: 10.3389/fpls.2023.1258533.

[77] F. Gürel *et al.* Comparison of expression patterns of selected drought-responsive genes in barley (hordeum vulgare l.) under shock-dehydration and slow drought treatments. *Plant Growth Regulation*, 80(2):183–193, 03/2016. DOI: 10.1007/s10725-016-0156-0.

[78] A.-L. Paul *et al.* Patterns of Arabidopsis gene expression in the face of hypobaric stress. *AoB PLANTS*, 9(4), 07/2017. DOI: 10.1093/aobpla/plx030.

[79] B. S. Yadav *et al.* Multidimensional patterns of metabolic response in abiotic stress-induced growth of *Arabidopsis thaliana*. *Plant Molecular Biology*, 92(6):689–699, 09/2016. DOI: 10.1007/s11103-016-0539-7.

[80] L. Guo *et al.* Metatranscriptomic comparison of endophytic and pathogenic Fusarium–Arabidopsis interactions reveals plant transcriptional plasticity. *Molecular Plant-Microbe Interactions*, 34(9):1071–1083, 09/2021. DOI: 10.1094/mpmi-03-21-0063-r.

[81] Y. Gao *et al.* Time series transcriptome analysis in *Medicago truncatula* shoot and root tissue during early nodulation. *Frontiers in Plant Science*, 13, 04/2022. DOI: 10.3389/fpls.2022.861639.

[82] S. Robinson *et al.* Alignment of time course gene expression data and the classification of developmentally driven genes with hidden markov models. *BMC Bioinformatics*, 16(1), 06/2015. DOI: 10.1186/s12859-015-0634-9.

[83] A. A. Pai, J. K. Pritchard and Y. Gilad. The genetic and mechanistic basis for variation in gene regulation. *PLoS Genetics*, 11(1):e1004857, 01/2015. T. Lappalainen, editor. DOI: 10.1371/journal.pgen.1004857.

[84] R. Bailon-Zambrano *et al.* Variable paralog expression underlies phenotype variation. *eLife*, 11, 09/2022. DOI: 10.7554/elife.79247.

[85] A. Calderwood *et al.* Comparative transcriptomics reveals desynchronisation of gene expression during the floral transition between Arabidopsis and Brassica rapacultivars. *Quantitative Plant Biology*, 2, 2021. DOI: 10.1017/qpb.2021.6.

[86] M. Cardoso-Moreira *et al.* Gene expression across mammalian organ development. *Nature*, 571(7766):505–509, 06/2019. DOI: 10.1038/s41586-019-1338-5.

[87] G. Lelandais and S. Le Crom. *Cross-species comparison using expression data.* In *Introduction to Systems Biology.* Humana Press, 2007, pages 147–159. DOI: 10.1007/978-1-59745-531-2_8.

[88] R. S. Datta *et al.* Berkeley PHOG: PhyloFacts orthology group prediction web server. *Nucleic Acids Research*, 37(suppl_2):W84–W89, 05/2009. DOI: 10.1093/nar/gkp373.

[89] M. Remm, C. E. Storm and E. L. Sonnhammer. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *Journal of Molecular Biology*, 314(5):1041–1052, 12/2001. DOI: 10.1006/jmbi.2000.5197.

[90] F. Chen *et al.* OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Research*, 34(90001):D363–D368, 01/2006. DOI: 10.1093/nar/gkj123.

[91] R. L. Tatusov *et al.* The *COG* database: an updated version includes eukaryotes. *BMC Bioinformatics*, 4(1), 09/2003. DOI: 10.1186/1471-2105-4-41.

[92] J. Ruan *et al.* TreeFam: 2008 update. *Nucleic Acids Research*, 36(Database):D735–D740, 12/2007. DOI: 10.1093/nar/gkm1005.

[93] J.-F. Dufayard *et al.* Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases. *Bioinformatics*, 21(11):2596–2603, 02/2005. DOI: 10.1093/bioinformatics/bti325.

[94] J. Huerta-Cepas *et al.* PhylomeDB: a database for genome-wide collections of gene phylogenies. *Nucleic Acids Research*, 36(Database):D491–D496, 12/2007. DOI: 10.1093/nar/gkm899.

[95] A. S. Kasianov *et al.* Interspecific comparison of gene expression profiles using machine learning. *PLOS Computational Biology*, 19(1):e1010743, 01/2023. A. Kahles, editor. DOI: 10.1371/journal.pcbi.1010743.

[96] M. Somssich. A short history of *Arabidopsis thaliana (L.) Heynh. Columbia-0*, 05/2018. DOI: 10.7287/peerj.preprints.26931v1.

[97] Y.-W. Yang *et al.* Molecular phylogenetic studies of Brassica, Rorippa, Arabidopsis and allied genera based on the internal transcribed spacer region of 18S–25S rDNA. *Molecular Phylogenetics and Evolution*, 13(3):455–462, 12/1999. DOI: 10.1006/mpev.1999.0648.

[98] X. Wang, Y. Hu and W. Wang. Comparative analysis of circadian transcriptomes reveals circadian characteristics between Arabidopsis and soybean. *Plants*, 12(19):3344, 09/2023. DOI: 10.3390/plants12193344.

[99] I. Cesarino *et al.* Plant science's next top models. *Annals of Botany*, 126(1):1–23, 04/2020. DOI: 10.1093/aob/mcaa063.

[100] J. A. Miller, S. Horvath and D. H. Geschwind. Divergence of human and mouse brain transcriptome highlights alzheimer disease pathways. *Proceedings of the National Academy of Sciences*, 107(28):12698–12703, 06/2010. DOI: 10.1073/pnas.0914257107.

[101] D. Simmons. The use of animal models in studying genetic disease: transgenesis and induced mutation. *Nature Education*, 1(1):70, 2008.

[102] R. L. Perlman. Mouse models of human disease: an evolutionary perspective. *Evolution, Medicine, and Public Health*:eow014, 04/2016. DOI: 10.1093/emph/eow014.

[103] S. Li, H. Nakayama and N. R. Sinha. How to utilize comparative transcriptomics to dissect morphological diversity in plants. *Current Opinion in Plant Biology*, 76:102474, 12/2023. DOI: 10.1016/j.pbi.2023.102474.

[104] S. Leiboff and S. Hake. Reconstructing the transcriptional ontogeny of maize and sorghum supports an inverse hourglass model of inflorescence development. *Current Biology*, 29(20):3410–3419.e3, 10/2019. DOI: 10.1016/j.cub.2019.08.044.

[105] Z. H. Lemmon *et al.* The evolution of inflorescence diversity in the nightshades and heterochrony during meristem maturation. *Genome Research*, 26(12):1676–1686, 11/2016. DOI: 10.1101/gr.207837.116.

[106] C. Ortiz-Ramírez *et al.* A transcriptome atlas of physcomitrella patens provides insights into the evolution and development of land plants. *Molecular Plant*, 9(2):205–220, 02/2016. DOI: 10.1016/j.molp.2015.12.002.

[107] J. Vercruysse *et al.* Comparative transcriptomics enables the identification of functional orthologous genes involved in early leaf growth. *Plant Biotechnology Journal*, 18(2):553–567, 08/2019. DOI: 10.1111/pbi.13223.

[108] Y. Guan *et al.* Comparative gene expression between two yeast species. *BMC Genomics*, 14(1), 01/2013. DOI: 10.1186/1471-2164-14-33.

[109] F. W. Albert *et al.* A comparison of brain gene expression levels in domesticated and wild animals. *PLoS Genetics*, 8(9):e1002962, 09/2012. J. M. Akey, editor. DOI: 10.1371/journal.pgen.1002962.

[110] D. R. Bull. *Coding moving pictures: motion prediction.* In *Communicating Pictures*. Elsevier, 2014, pages 255–289. DOI: 10.1016/b978-0-12-405906-1.00008-8.

[111] B. Ait-Amir, P. Pougnet and A. El Hami. *Meta-model development.* In *Embedded Mechatronic Systems 2*. Elsevier, 2015, pages 151–179. DOI: 10.1016/b978-1-78548-014-0.50006-2.

[112] J. Han, M. Kamber and J. Pei. *Getting to know your data.* In *Data Mining*. Elsevier, 2012, pages 39–82. DOI: 10.1016/b978-0-12-381479-1.00002-2.

[113] P. Jiang *et al.* TimeMeter assesses temporal gene expression similarity and identifies differentially progressing genes. *Nucleic Acids Research*, 48(9):e51–e51, 03/2020. DOI: 10.1093/nar/gkaa142.

[114] C. Barry *et al.* Species-specific developmental timing is maintained by pluripotent stem cells ex utero. *Developmental Biology*, 423(2):101–110, 03/2017. DOI: 10.1016/j.ydbio.2017.02.002.

[115] T. Giorgino. Computing and visualizing dynamic time warping alignments in R: the dtw package. *Journal of Statistical Software*, 31(7), 2009. DOI: 10.18637/jss.v031.i07.

[116] M. Herrmann and G. I. Webb. Amercing: an intuitive and effective constraint for dynamic time warping. *Pattern Recognition*, 137:109333, 05/2023. DOI: 10.1016/j.patcog.2023.109333.

[117] R. Cavill, J. Kleinjans and J.-J. Briedé. DTW4Omics: comparing patterns in biological time series. *PLoS ONE*, 8(8):e71823, 08/2013. P. Csermely, editor. DOI: 10.1371/journal.pone.0071823.

[118] J. Straube, B. E. Huang and K.-A. L. Cao. DynOmics to identify delays and co-expression patterns across time course experiments. *Scientific Reports*, 7(1), 01/2017. DOI: 10.1038/srep40131.

[119] T.-h. Lin, N. Kaminski and Z. Bar-Joseph. Alignment and classification of time series gene expression in clinical studies. *Bioinformatics*, 24(13):i147–i155, 07/2008. DOI: 10.1093/bioinformatics/btn152.

[120] Y. X. R. Wang *et al.* Generalized correlation measure using count statistics for gene expression data with ordered samples. *Bioinformatics*, 34(4):617–624, 10/2017. O. Stegle, editor. DOI: 10.1093/bioinformatics/btx641.

[121] X. Leng and H.-G. Müller. Time ordering of gene coexpression. *Biostatistics*, 7(4):569–584, 02/2006. DOI: 10.1093/biostatistics/kxj026.

[122] Y. Yuan *et al.* Development and application of a modified dynamic time warping algorithm (*DTW-S*) to analyses of primate brain expression time series. *BMC Bioinformatics*, 12(1), 08/2011. DOI: 10.1186/1471-2105-12-347.

[123] P. Chris. Probability for computer scientists. 2023. <https://chrispiech.github.io/probabilityForComputerScientists/>.

[124] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[125] S. I. Vrieze. Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological Methods*, 17(2):228–243, 2012. DOI: 10.1037/a0027127.

[126] J. Ramsay, G. Hooker and S. Graves. *Registration: aligning features for samples of curves.* In *Functional Data Analysis with R and MATLAB*. Springer New York, 2009, pages 117–130. DOI: 10.1007/978-0-387-98185-7_8.

[127] A. Kneip and J. O. Ramsay. Combining registration and fitting for functional models. *Journal of the American Statistical Association*, 103(483):1155–1165, 09/2008. DOI: 10.1198/016214508000000517.

[128] J. S. Marron *et al.* Functional data analysis of amplitude and phase variation. *Statistical Science*, 30(4), 11/2015. DOI: 10.1214/15-sts524.

[129] Z. Bar-Joseph *et al.* Continuous representations of time-series gene expression data. *Journal of Computational Biology*, 10(3–4):341–356, 06/2003. DOI: 10.1089/10665270360688057.

[130] A. Perperoglou *et al.* A review of spline function procedures in R. *BMC Medical Research Methodology*, 19(1), 03/2019. DOI: 10.1186/s12874-019-0666-3.

[131] G. James *et al. An Introduction to Statistical Learning: with Applications in R*. Springer, 2021.

[132] J. Samuel. Machine learning. 2024. <%5Curl%7Bhttps://bookdown.org/ssjackson300/Machine-Learning-Lecture-Notes/splines.html#spline-basis-representation/%7D>.

[133] C. de Boor. B(asic)-Spline Basics. In 1986.

[134] C. Zhu *et al.* Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software*, 23(4):550–560, 12/1997. DOI: 10.1145/279232.279236.

[135] G. Antonie. Bound constrained optimization: application to the dip estimation problem. 2004. <https://sepwww.stanford.edu/data/media/public/docs/sep117/antoine1/paper_html/>.

[136] D. C. Liu and J. Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45(1–3):503–528, 08/1989. DOI: 10.1007/bf01589116.

[137] J. Nelder and R. Mead. A simplex method for function minimization. *The Computer Journal*, 8(1):27–27, 04/1965. DOI: 10.1093/comjnl/8.1.27.

[138] S. Singer and J. Nelder. Nelder–Mead algorithm. *Scholarpedia*, 4(7):2928, 2009. DOI: 10.4249/scholarpedia.2928.

[139] S. Kirkpatrick, C. D. Gelatt and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 05/1983. DOI: 10.1126/science.220.4598.671.

[140] P. García Nieto *et al.* Forecast of the higher heating value in biomass torrefaction by means of machine learning techniques. *Journal of Computational and Applied Mathematics*, 357:284–301, 09/2019. DOI: 10.1016/j.cam.2019.03.009.

[141] X.-S. Yang. Metaheuristic optimization. *Scholarpedia*, 6(8):11472, 2011. DOI: 10.4249/scholarpedia.11472.

[142] A. Geron. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, Inc., 2nd edition, 2019.

[143] G. Developers. Normalization: machine learning crash course. 2024. <https://developers.google.com/machine-learning/crash-course/numerical-data/normalization>.

[144] C. J. Geyer and E. A. Thompson. Constrained monte carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 54(3):657–683, 07/1992. DOI: 10.1111/j.2517-6161.1992.tb01443.x.

[145] J. Skilling. Nested sampling for general bayesian computation. *Bayesian Analysis*, 1(4), 12/2006. DOI: 10.1214/06-ba127.

[146] N. Pullen and R. J. Morris. Bayesian model comparison and parameter inference in systems biology using nested sampling. *PLoS ONE*, 9(2):e88419, 02/2014. S. Rogers, editor. DOI: 10.1371/journal.pone.0088419.

[147] A. E. Raftery. Bayesian model selection in social research. *Sociological Methodology*, 25:111, 1995. DOI: 10.2307/271063.

[148] A. V. Klepikova *et al.* RNA-Seq analysis of an apical meristem time series reveals a critical point in *Arabidopsis thaliana* flower initiation. *BMC Genomics*, 16(1), 06/2015. DOI: 10.1186/s12864-015-1688-9.

[149] Y. Luan and H. Li. Clustering of time-course gene expression data using a mixed-effects model with b-splines. *Bioinformatics*, 19(4):474–482, 03/2003. DOI: 10.1093/bioinformatics/btg014.

[150] S. Déjean *et al.* Clustering time-series gene expression data using smoothing spline derivatives. *EURASIP Journal on Bioinformatics and Systems Biology*, 2007:1–10, 2007. DOI: 10.1155/2007/70561.

[151] H. Wickham. *Advanced R*. Chapman and Hall/CRC, 05/2019. DOI: 10.1201/9781351201315.

[152] M. Dowle and A. Srinivasan. *data.table: Extension of 'data.frame'*. https://r-datatable.com. 2023.

[153] H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

[154] T. L. Pedersen. *patchwork: The Composer of Plots*. R package version 1.2.0. 2024.

[155] H. Wickham and D. Seidel. *scales: Scale Functions for Visualization*. https://scales.r-lib.org. 2022.

[156] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2023.

[157] K. Husmann, A. Lange and E. Spiegel. *The R Package optimization: Flexible Global Optimization with Simulated-Annealing*. 2017.

[158] S. Bihorel. *neldermead: R Port of the Scilab Neldermead Module*. R package version 1.0-12. 2022.

[159] D. Vaughan and M. Dancho. *furrr: Apply Mapping Functions in Parallel using Futures*. R package version 0.3.1. 2022.

[160] H. Bengtsson. A unifying framework for parallel and distributed processing in R using futures. *The R Journal*, 13(2):208–227, 2021. DOI: 10.32614/RJ-2021-048.

[161] G. Csárdi. *cli: Helpers for Developing Command Line Interfaces*. https://cli.r-lib.org. 2023.

[162] A. Gray. *Structural botany*, volume 1. American Book Company, 1879.

[163] C. J. Whipple *et al.* A conserved mechanism of bract suppression in the grass family. *The Plant Cell*, 22(3):565–578, 03/2010. DOI: 10.1105/tpc.109.073536.

[164] V. Brukhin and N. Morozova. Plant growth and development - basic knowledge and current views. *Mathematical Modelling of Natural Phenomena*, 6(2):1–53, 10/2010. DOI: 10.1051/mmnp/20116201.

[165] S. Dieudonné. *Bract inhibition during floral transition in Arabidopsis thaliana*. PhD thesis, L'Ecole Normale Supérieure de Lyon, 12/2021.

[166] J. R. Dinneny *et al.* The role of *JAGGED* in shaping lateral organs. *Development*, 131(5):1101–1110, 03/2004. DOI: 10.1242/dev.00949.

[167] B. Song *et al.* Multifunctionality of angiosperm floral bracts: a review. *Biological Reviews*, 99(3):1100–1120, 01/2024. DOI: 10.1111/brv.13060.

[168] Bailey lab UC Davis. The Phytomer: Basic Unit of a plant shoot. 2022. <https://baileylab.ucdavis.edu/software/helios/_plant_architecture_doc.html>.

[169] Wikimedia Commons. *Rafflesia arnoldii* flowers in Bengkulu, Indonesia. 2015. <https://commons.wikimedia.org/wiki/File:Rafflesia_arnoldii_Bengkulu_01.jpg>.

[170] Wikimedia Commons. *Salvia pratensis*, Lamiaceae, Meadow Clary, Meadow Sage, flower. Karlsruhe, Germany. 2009. <https://commons.wikimedia.org/wiki/File:Salvia_pratensis_005.JPG>.

[171] Wikimedia Commons. Turk's cap lily (*Lilium martagon*) found on the root path to the 1,456m high Staufenberg near Dornbirn. 2009. <https://commons.wikimedia.org/wiki/File:T%C3%BCrkenbund_Lilie,_Lilium_martagon.JPG>.

[172] Wikimedia Commons. Poinsettia *Euphorbia pulcherrima*. 2006. <https://commons.wikimedia.org/wiki/File:E_pulcherrima_ies.jpg>.

[173] Wikimedia Commons. *Arum palaestinum* located near Modi'in, Israel. 2007. <https://commons.wikimedia.org/wiki/File:PikiWiki_Israel_3833_arum_palestinum.jpg>.

[174] Wikimedia Commons. Passionfruit cream. 2013. <https://commons.wikimedia.org/wiki/File:Passionfruit_cream.jpg>.

[175] Wikimedia Commons. Artichoke, in a allotment in Tourcoing (Nord), France. 2012. <https://commons.wikimedia.org/wiki/File:Artichoke_J1.jpg>.

[176] Wikimedia Commons. Cracked fruit of a chestnut tree (edible chestnut). 2005. <https://commons.wikimedia.org/wiki/File:Chestnuts.jpg>.

[177] Y.-Y. Hu *et al.* Important photosynthetic contribution from the non-foliar green organs in cotton at the late growth stage. *Planta*, 235(2):325–336, 09/2011. DOI: 10.1007/s00425-011-1511-z.

[178] B. Song *et al.* The bracts of the alpine 'glasshouse' plant *Rheum alexandrae* (Polygonaceae) enhance reproductive fitness of its pollinating seed-consuming mutualist: Rheum bracts enhance pollinator fitness. *Botanical Journal of the Linnean Society*, 179(2):349–359, 08/2015. DOI: 10.1111/boj.12312.

[179] T. Keasar *et al.* Honesty of signaling and pollinator attraction: the case of flag-like bracts. *Israel Journal of Plant Sciences*, 54(2):119–128, 04/2006. DOI: 10.1560/ijps_54_2_119.

[180] L. Zhang *et al.* Phylogeny and evolution of bracts and bracteoles in tacca (dioscoreaceae). *Journal of Integrative Plant Biology*, 53(11):901–911, 11/2011. DOI: 10.1111/j.1744-7909.2011.01076.x.

[181] M. von Balthazar and P. K. Endress. Floral bract function, flowering process and breeding systems of Sarcandra and Chloranthus (Chloranthaceae). *Plant Systematics and Evolution*, 218(3/4):161–178, 1999.

[182] H. Sun *et al.* Survival and reproduction of plant species in the qinghai–tibet plateau. *Journal of Systematics and Evolution*, 52(3):378–396, 04/2014. DOI: 10.1111/jse.12092.

[183] T. R. Radhamani, L. Sudarshana and R. Krishnan. Defense and carnivory: dual role of bracts in *Passiflora foetida*. *Journal of Biosciences*, 20(5):657–664, 12/1995. DOI: 10.1007/bf02703305.

[184] X. Wu *et al.* Inducing bract-like leaves in Arabidopsis through ectopically expressing an *ASR* gene from the dove tree. *Industrial Crops and Products*, 180:114796, 06/2022. DOI: 10.1016/j.indcrop.2022.114796.

[185] Y. Zhao *et al.* *HANABA TARANU* is a *GATA* transcription factor that regulates shoot apical meristem and flower development in Arabidopsis. *The Plant Cell*, 16(10):2586–2600, 10/2004. DOI: 10.1105/tpc.104.024869.

[186] D. A. German *et al.* An updated classification of the Brassicaceae (Cruciferae). *PhytoKeys*, 220:127–144, 03/2023. DOI: 10.3897/phytokeys.220.97724.

[187] F. D. Hempel and L. J. Feldman. Specification of chimeric flowering shoots in wild-type Arabidopsis. *The Plant Journal*, 8(5):725–731, 11/1995. DOI: 10.1046/j.1365-313x.1995.08050725.x.

[188] P. Alberch *et al.* Size and shape in ontogeny and phylogeny. *Paleobiology*, 5(3):296–317, 1979. DOI: 10.1017/s0094837300006588.

[189] S. Le, J. Josse and F. Husson. *FactoMineR*: a package for multivariate analysis. *Journal of Statistical Software*, 25(1):1–18, 2008. DOI: 10.18637/jss.v025.i01.

[190] T. Wu *et al.* *clusterProfiler* 4.0: a universal enrichment tool for interpreting omics data. *The Innovation*, 2(3):100141, 2021. DOI: 10.1016/j.xinn.2021.100141.

[191] F. Bouché *et al.* FLOR-ID: an interactive database of flowering-time gene networks in *Arabidopsis thaliana*. *Nucleic Acids Research*, 44(D1):D1167–D1171, 10/2015. DOI: 10.1093/nar/gkv1054.

[192] M. G. Heisler *et al.* Patterns of auxin transport and gene expression during primordium development revealed by live imaging of the Arabidopsis inflorescence meristem. *Current Biology*, 15(21):1899–1911, 11/2005. DOI: 10.1016/j.cub.2005.09.052.

[193] T. Pham *et al.* The evolutionary origination of a novel expression pattern through an extreme heterochronic shift. *Evolution & Development*, 19(2):43–55, 01/2017. DOI: 10.1111/ede.12215.

[194] H. Tsukaya *et al.* How do 'housekeeping' genes control organogenesis?–unexpected new findings on the role of housekeeping genes in cell and organ differentiation. *Journal of Plant Research*, 126(1):3–15, 08/2012. DOI: 10.1007/s10265-012-0518-2.

[195] M. Buendía-Monreal and C. S. Gillmor. The times they are a-changin': heterochrony in plant development and evolution. *Frontiers in Plant Science*, 9, 09/2018. DOI: 10.3389/fpls.2018.01349.

[196] R. H. Bloomer and C. Dean. Fine-tuning timing: natural variation informs the mechanistic basis of the switch to flowering in *Arabidopsis thaliana*. *Journal of Experimental Botany*, 68(20):5439–5452, 08/2017. DOI: 10.1093/jxb/erx270.

[197] H. Li *et al.* Genome-wide identification of flowering-time genes in Brassica species and reveals a correlation between selective pressure and expression patterns of vernalization-pathway genes in *Brassica napus*. *International Journal of Molecular Sciences*, 19(11):3632, 11/2018. DOI: 10.3390/ijms19113632.

[198] S. J. Gould. Dollo on dollo's law: irreversibility and the status of evolutionary laws. *Journal of the History of Biology*, 3(2):189–212, 1970. DOI: 10.1007/bf00137351.

[199] K. Witzel, A. B. Kurina and A. M. Artemyeva. Opening the treasure chest: the current status of research on *Brassica oleracea* and *B. rapa* vegetables from *ex situ* germplasm collections. *Frontiers in Plant Science*, 12, 05/2021. DOI: 10.3389/fpls.2021.643047.

[200] P. Subramanian, S.-H. Kim and B.-S. Hahn. Brassica biodiversity conservation: prevailing constraints and future avenues for sustainable distribution of plant genetic resources. *Frontiers in Plant Science*, 14, 07/2023. DOI: 10.3389/fpls.2023.1220134.

[201] S. M. Woodhouse. *Unravelling the floral transition in Brassica oleracea using transcriptomics*. PhD thesis, University of East Anglia, 09/2021.

[202] L. Brown. Horticulture statistics–2023. 2024. <https://www.gov.uk/government/statistics/latest-horticulture-statistics/horticulture-statistics-2023#section-1--vegetables> (visited on 20/06/2024).

[203] F. Cheng, J. Wu and X. Wang. Genome triplication drove the diversification of Brassica plants. *Horticulture Research*, 1(1), 05/2014. DOI: 10.1038/hortres.2014.24.

[204] U. Nagaharu and N. Nagaharu. Genome analysis in Brassica with special reference to the experimental formation of *B. napus* and peculiar mode of fertilization. *Jpn J Bot*, 7(7):389–452, 1935.

[205] X. Zhang *et al.* Interspecific hybridization, polyploidization and backcross of *Brassica oleracea var. alboglabra* with *B. rapa var. purpurea* morphologically recapitulate the evolution of Brassica vegetables. *Scientific Reports*, 6(1), 01/2016. DOI: 10.1038/srep18618.

[206] M. Zhao *et al.* Shifts in the evolutionary rate and intensity of purifying selection between two Brassica genomes revealed by analyses of orthologous transposons and relics of a whole genome triplication. *The Plant Journal*, 76(2):211–222, 08/2013. DOI: 10.1111/tpj.12291.

[207] A. Feng *et al.* Genome-wide identification, evolutionary selection, and genetic variation of DNA methylation-related genes in *Brassica rapa* and *Brassica oleracea*. *Journal of Integrative Agriculture*, 21(6):1620–1632, 06/2022. DOI: 10.1016/s2095-3119(21)63827-3.

[208] P. Soengas *et al.* New vegetable Brassica foods: a promising source of bioactive compounds. *Foods*, 10(12):2911, 11/2021. DOI: 10.3390/foods10122911.

[209] S. Liu *et al.* The *Brassica oleracea* genome reveals the asymmetrical evolution of polyploid genomes. *Nature Communications*, 5(1), 05/2014. DOI: 10.1038/ncomms4930.

[210] P. K. Boss *et al.* Multiple pathways in the decision to flower: enabling, promoting, and resetting. *THE PLANT CELL ONLINE*, 16(suppl_1):S18–S31, 03/2004. DOI: 10.1105/tpc.015958.

[211] F. Andrés and G. Coupland. The genetic basis of flowering responses to seasonal cues. *Nature Reviews Genetics*, 13(9):627–639, 08/2012. DOI: 10.1038/nrg3291.

[212] A. Srikanth and M. Schmid. Regulation of flowering time: all roads lead to rome. *Cellular and Molecular Life Sciences*, 68(12):2013–2037, 04/2011. DOI: 10.1007/s00018-011-0673-y.

[213] H. Hõrak. Remodeling flowering: *CHROMATIN REMODELING4* promotes the floral transition. *The Plant Cell*, 32(5):1346–1347, 03/2020. DOI: 10.1105/tpc.20.00196.

[214] F. Fornara, A. de Montaigu and G. Coupland. SnapShot: control of flowering in Arabidopsis. *Cell*, 141(3):550–550.e2, 04/2010. DOI: 10.1016/j.cell.2010.04.024.

[215] S. V. Schiessl *et al.* Flowering time gene variation in Brassica species shows evolutionary principles. *Frontiers in Plant Science*, 8, 10/2017. DOI: 10.3389/fpls.2017.01742.

[216] Y. Y. Levy and C. Dean. The transition to flowering. *The Plant Cell*, 10(12):1973–1989, 12/1998. DOI: 10.1105/tpc.10.12.1973.

[217] F. Wellmer and J. L. Riechmann. Gene networks controlling the initiation of flower development. *Trends in Genetics*, 26(12):519–527, 12/2010. DOI: 10.1016/j.tig.2010.09.001.

[218] S. Takada *et al.* The role of *FRIGIDA* and *FLOWERING LOCUS C* genes in flowering time of *Brassica rapa* leafy vegetables. *Scientific Reports*, 9(1), 09/2019. DOI: 10.1038/s41598-019-50122-2.

[219] C. Li *et al.* Comprehensive expression analysis of Arabidopsis *GA2*-oxidase genes and their functional insights. *Plant Science*, 285:1–13, 08/2019. DOI: 10.1016/j.plantsci.2019.04.023.

[220] R. Uptmoor *et al.* Prediction of flowering time in *Brassica oleracea* using a quantitative trait loci-based phenology model. *Plant Biology*, 14(1):179–189, 06/2011. DOI: 10.1111/j.1438-8677.2011.00478.x.

[221] J. A. Irwin *et al.* Functional alleles of the flowering time regulator *FRIGIDA* in the *Brassica oleracea* genome. *BMC Plant Biology*, 12(1), 02/2012. DOI: 10.1186/1471-2229-12-21.

[222] A. Akter *et al.* Genome triplication leads to transcriptional divergence of *FLOWERING LOCUS C* genes during vernalization in the genus Brassica. *Frontiers in Plant Science*, 11, 02/2021. DOI: 10.3389/fpls.2020.619417.

[223] S. Das Laha *et al.* Gene duplication and stress genomics in Brassicas: current understanding and future prospects. *Journal of Plant Physiology*, 255:153293, 12/2020. DOI: 10.1016/j.jplph.2020.153293.

[224] J. Gu *et al.* The story of a decade: genomics, functional genomics, and molecular breeding in *Brassica napus*. *Plant Communications*, 5(4):100884, 04/2024. DOI: 10.1016/j.xplc.2024.100884.

[225] G. C. Conant and K. H. Wolfe. Turning a hobby into a job: how duplicated genes find new functions. *Nature Reviews Genetics*, 9(12):938–950, 12/2008. DOI: 10.1038/nrg2482.

[226] D. M. Jones *et al.* Spatio-temporal expression dynamics differ between homologues of flowering time genes in the allopolyploid *Brassica napus*. *The Plant Journal*, 96(1):103–118, 08/2018. DOI: 10.1111/tpj.14020.

[227] S. D. Iohannes and D. Jackson. Tackling redundancy: genetic mechanisms underlying paralog compensation in plants. *New Phytologist*, 240(4):1381–1389, 09/2023. DOI: 10.1111/nph.19267.

[228] U. Meier *et al.* The BBCH system to coding the phenological growth stages of plants-history and publications. *Journal für Kulturpflanzen*, 61(2):41–52, 2009. DOI: 10.5073/JFK.2009.02.01.

[229] M. Pertea *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-Seq reads. *Nature Biotechnology*, 33(3):290–295, 02/2015. DOI: 10.1038/nbt.3122.

[230] L. Zhang *et al.* Improved *Brassica rapa* reference genome by single-molecule sequencing and chromosome conformation capture technologies. *Horticulture Research*, 5(1), 08/2018. DOI: 10.1038/s41438-018-0071-9.

[231] Z. He *et al.* Construction of Brassica A and C genome-based ordered pan-transcriptomes for use in rapeseed genomic research. *Data in Brief*, 4:357–362, 09/2015. DOI: 10.1016/j.dib.2015.06.016.

[232] A. A. Golicz *et al.* The pangenome of an agronomically important crop plant *Brassica oleracea*. *Nature Communications*, 7(1), 11/2016. DOI: 10.1038/ncomms13390.

[233] T. Z. Berardini *et al.* The Arabidopsis information resource: making and mining the "gold standard" annotated reference plant genome. *genesis*, 53(8):474–485, 08/2015. DOI: 10.1002/dvg.22877.

[234] F. Cheng *et al.* Syntenic gene analysis between *Brassica rapa* and other Brassicaceae species. *Frontiers in Plant Science*, 3, 2012. DOI: 10.3389/fpls.2012.00198.

[235] S. F. Altschul *et al.* Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 10/1990. DOI: 10.1016/s0022-2836(05)80360-2.

[236] F. Briatte. *ggnetwork: Geometries to Plot Networks with 'ggplot2'*. R package version 0.5.13. 2024.

[237] F. Fornara *et al.* Arabidopsis DOF transcription factors act redundantly to reduce *CONSTANS* expression and are essential for a photoperiodic flowering response. *Developmental Cell*, 17(1):75–86, 07/2009. DOI: 10.1016/j.devcel.2009.06.015.

[238] C. Zou, W. Jiang and D. Yu. Male gametophyte-specific *WRKY34* transcription factor mediates cold sensitivity of mature pollen in Arabidopsis. *Journal of Experimental Botany*, 61(14):3901–3914, 07/2010. DOI: 10.1093/jxb/erq204.

[239] A. Paterson *et al.* Many gene and domain families have convergent fates following independent whole-genome duplication events in Arabidopsis, Oryza, Saccharomyces and Tetraodon. *Trends in Genetics*, 22(11):597–602, 11/2006. DOI: 10.1016/j.tig.2006.09.003.

[240] R. De Smet *et al.* Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. *Proceedings of the National Academy of Sciences*, 110(8):2898–2903, 02/2013. DOI: 10.1073/pnas.1300127110.

[241] K. Hanada *et al.* Evolutionary persistence of functional compensation by duplicate genes in Arabidopsis. *Genome Biology and Evolution*, 1:409–414, 01/2009. DOI: 10.1093/gbe/evp043.

[242] S. Shafiq, A. Berr and W.-H. Shen. Combinatorial functions of diverse histone methylations in *Arabidopsis thaliana* flowering time regulation. *New Phytologist*, 201(1):312–322, 09/2013. DOI: 10.1111/nph.12493.

[243] M. Martin-Trillo *et al.* *EARLY IN SHORT DAYS 1* (*ESD1*) encodes *ACTIN-RELATED PROTEIN 6* (*AtARP6*), a putative component of chromatin remodelling complexes that positively regulates *FLC* accumulation in Arabidopsis. *Development*, 133(7):1241–1252, 04/2006. DOI: 10.1242/dev.02301.

[244] S. Fujiwara *et al.* Circadian clock proteins *LHY* and *CCA1* regulate *SVP* protein accumulation to control flowering in Arabidopsis. *The Plant Cell*, 20(11):2960–2971, 11/2008. DOI: 10.1105/tpc.108.061531.

[245] Q. Sun *et al.* R-loop stabilization represses antisense transcription at the Arabidopsis *FLC* locus. *Science*, 340(6132):619–621, 05/2013. DOI: 10.1126/science.1234848.

[246] I. Rieu *et al.* Genetic analysis reveals that *C19-GA* 2-oxidation is a major gibberellin inactivation pathway in Arabidopsis. *The Plant Cell*, 20(9):2420–2436, 09/2008. DOI: 10.1105/tpc.108.058818.

[247] P. Jiang *et al.* *SIP1* participates in regulation of flowering time in rice by recruiting *OsTrx1* to *Ehd1*. *New Phytologist*, 219(1):422–435, 04/2018. DOI: 10.1111/nph.15122.

[248] S. Wang *et al.* *SDG128* is involved in maize leaf inclination. *The Plant Journal*, 108(6):1597–1608, 10/2021. DOI: 10.1111/tpj.15527.

[249] H. Dong *et al.* Diversification and evolution of the SDG gene family in *Brassica rapa* after the whole genome triplication. *Scientific Reports*, 5(1), 11/2015. DOI: 10.1038/srep16851.

[250] S. Sehrish *et al.* Genome-wide identification and characterization of SET domain family genes in *Brassica napus* l. *International Journal of Molecular Sciences*, 23(4):1936, 02/2022. DOI: 10.3390/ijms23041936.

[251] C. R. McClung. A modern circadian clock in the common angiosperm ancestor of monocots and eudicots. *BMC Biology*, 8(1), 05/2010. DOI: 10.1186/1741-7007-8-55.

[252] P. Lou *et al.* Preferential retention of circadian clock genes during diploidization following whole genome triplication in *Brassica rapa*. *The Plant Cell*, 24(6):2415–2426, 06/2012. DOI: 10.1105/tpc.112.099499.

[253] P. G. H. de Rooij *et al.* The diverse and unanticipated roles of histone deacetylase 9 in coordinating plant development and environmental acclimation. *Journal of Experimental Botany*, 71(20):6211–6225, 07/2020. D. Gibbs, editor. DOI: 10.1093/jxb/eraa335.

[254] Y. Wang *et al.* Protein interactions of flowering inhibitors *AGL18* and *HDA9* with integrator factors in *Brassica oleracea var. italica*. *Acta Horticulturae Sinica*, 45(12):2383–2394, 2383, 2018. DOI: 10.16420/j.issn.0513-353x.2018-0335.

[255] G. Ma *et al.* Expression patterns of *HDA9*, *HDA6* and *FLD* in chinese cabbage (brassica rapa l. ssp. pekinensis) under different photoperiods and their protein interactions. *Brazilian Journal of Botany*, 46(3):549–561, 07/2023. DOI: 10.1007/s40415-023-00897-6.

[256] M.-J. Kang *et al.* Repression of flowering under a noninductive photoperiod by the *HDA9-AGL19-FT* module in Arabidopsis. *New Phytologist*, 206(1):281–294, 11/2014. DOI: 10.1111/nph.13161.

[257] T. Bieluszewski *et al.* *AtEAF1* is a potential platform protein for Arabidopsis *NuA4* acetyltransferase complex. *BMC Plant Biology*, 15(1), 03/2015. DOI: 10.1186/s12870-015-0461-1.

[258] F. Liu *et al.* Targeted 3' processing of antisense transcripts triggers Arabidopsis *FLC* chromatin silencing. *Science*, 327(5961):94–97, 01/2010. DOI: 10.1126/science.1180278.

[259] J.-H. Jung *et al.* The *miR172* target *TOE3* represses *AGAMOUS* expression during Arabidopsis floral patterning. *Plant Science*, 215–216:29–38, 02/2014. DOI: 10.1016/j.plantsci.2013.10.010.

[260] I. Castro Marín *et al.* Nitrate regulates floral induction in Arabidopsis, acting independently of light, gibberellin and autonomous pathways. *Planta*, 233(3):539–552, 11/2010. DOI: 10.1007/s00425-010-1316-5.

[261] L. Corbesier and G. Coupland. The quest for florigen: a review of recent progress. *Journal of Experimental Botany*, 57(13):3395–3403, 09/2006. DOI: 10.1093/jxb/erl095.

[262] J. Lee and I. Lee. Regulation and function of *SOC1*, a flowering pathway integrator. *Journal of Experimental Botany*, 61(9):2247–2254, 04/2010. DOI: 10.1093/jxb/erq098.

[263] P. Huijser and M. Schmid. The control of developmental phase transitions in plants. *Development*, 138(19):4117–4129, 10/2011. DOI: 10.1242/dev.063511.

[264] C. Jung and A. E. Müller. Flowering time control and applications in plant breeding. *Trends in Plant Science*, 14(10):563–573, 10/2009. DOI: 10.1016/j.tplants.2009.07.005.

[265] K. E. Jaeger *et al.* Interlocking feedback loops govern the dynamic behavior of the floral transition in Arabidopsis. *The Plant Cell*, 25(3):820–833, 03/2013. DOI: 10.1105/tpc.113.109355.

[266] A. Calderwood *et al.* Total *FLC* transcript dynamics from divergent paralogue expression explains flowering diversity in *Brassica napus*. *New Phytologist*, 229(6):3534–3548, 12/2020. DOI: 10.1111/nph.17131.

[267] C. A. Penfold *et al.* Nonparametric bayesian inference for perturbed and orthologous gene regulatory networks. *Bioinformatics*, 28(12):i233–i241, 06/2012. DOI: 10.1093/bioinformatics/bts222.

[268] A. Serrano-Mislata *et al.* DELLA genes restrict inflorescence meristem function independently of plant height. *Nature Plants*, 3(9):749–754, 08/2017. DOI: 10.1038/s41477-017-0003-y.

[269] D. M. Jones. *Effects of gene multiplication on flowering time regulation in spring and winter varieties of Brassica napus.* PhD thesis, University of East Anglia, 09/2017.

[270] A. Akter *et al.* Characterization of *FLOWERING LOCUS C 5* in *Brassica rapa L. Molecular Breeding*, 43(8), 07/2023. DOI: 10.1007/s11032-023-01405-0.

[271] A. Dolatabadian *et al.* Copy number variation among resistance genes analogues in *Brassica napus*. *Genes*, 13(11):2037, 11/2022. DOI: 10.3390/genes13112037.

[272] X. Li *et al.* Large-scale gene expression alterations introduced by structural variation drive morphotype diversification in *Brassica oleracea*. *Nature Genetics*, 56(3):517–529, 02/2024. DOI: 10.1038/s41588-024-01655-4.

[273] J. Jiang *et al.* Digital gene expression analysis of gene expression differences within Brassica diploids and allopolyploids. *BMC Plant Biology*, 15(1), 01/2015. DOI: 10.1186/s12870-015-0417-5.