Computational analysis of transcriptomic, phylotranscriptomic, and metabolomic diversity

Dissertation

zur Erlangung des akademischen Grades

Doktor der Naturwissenschaften (Dr. rer. nat.)

der Naturwissenschaftlichen Fakultät III
Agrar- und Ernährungswissenschaften,
Geowissenschaften und Informatik
der Martin–Luther–Universität Halle–Wittenberg,

vorgelegt von

Herrn Alexander Gabel
Geb. am 11. Juni 1988 in Schönebeck (Elbe)

Gutachter:
1. Prof. Dr. Ivo Große
2. Prof. Dr. Peter Stadler

Datum der Verteidigung: 15.07.2021

Acknowledgements

I am very grateful to Ivo Grosse for introducing me to the world of Bioinformatics and for actively involving me in research, even as an undergraduate student. Together with Marcel Quint, he gave my colleague Hajk-Georg Drost and me the chance to study the transcriptomic hourglass pattern in plant embryogenesis. For this initial trust and the fruitful collaboration, I would like to thank Ivo Grosse and Marcel Quint. In particular, I would like to extend my gratitude to Ivo Grosse for his support and constant motivation. I will never forget the long productive nights in which, after the usual 12 hours at the institute, he found the time in which we discussed and developed new ideas until the early hours of the next day.

Additionally, I am very grateful to Elliot M. Meyerowitz for allowing me to experience the scientific landscape of the Sainsbury Laboratory at the Cambridge University and to work with him on diverse topics of developmental biology. I am also thankful to Christoph Schuster, with whom I worked on the transcriptome of flowering plants. I very much appreciate the motivating and productive discussions with him, and I am thankful for the inspiring time.

I would like to thank Kevin Ernst and Benjamin Schmidt for helping to develop and implement the PhyloWeb server. I would like to extend my sincere thanks to all my collaborators who offered me the opportunity to contribute to their diverse research projects. I would like to thank Rita Groß-Hardt, Pedro Madrigal, Charles W. Melnyk, Alexandra Schutkowski, Gabrielle I. Stangl, Sophie C. Steinbach, Yanbo Mao, Reinhard Neubert, Frank Wellmer, and Johannes Wohlrab.

Furthermore, I am deeply grateful for the members of the Grosse lab, the best colleagues I could not even imagine. It has been a pleasure to work and to discuss with them in our extended coffee breaks. Especially, I would like to thank Ioana Lemnian, Claus Weinholdt, Martin Porsch, and Jördis-Ann Schüler. What has started with subject-specific discussions during our weekly lab meetings has turned into friendships. I also would like to thank Yvonne Pöschl-Grau, Birgit Möller, and Jan Grau for always having an open door if I had a “small question.”

Finally, I would like to thank my family and friends. They have provided me a world outside my scientific bubble, and I am grateful for such an unforgettable time throughout my PhD studies and beyond. I would also like to express my deepest gratitude to my parents and my brother for their incredible support and belief in my abilities. For having a fantastic time during the past years, I would like to thank the “Schwimmladies” - Annemarie, Cora, and Elisa -, Gabi, Johannes, Sascha, and Kristin, who made Halle feel like home. I especially would like to thank Maria, who motivates and inspires me in unprecedented ways every day.

Abstract

In this thesis, we combine concepts and methods from machine learning and statistics and apply them to analyze diverse biological data in phylotranscriptomics, transcriptomics, and metabolomics.

Starting with phylotranscriptomics, we study the transcriptomic hourglass pattern in animals and plants. It is based on the theory of the developmental hourglass that describes the morphological convergence of animal embryos during mid embryogenesis. Although there is no evidence for a morphological hourglass pattern during embryogenesis in plants, we show a transcriptomic hourglass pattern during the embryogenesis of the model plant *Arabidopsis thaliana*. To study developmental processes in the light of evolution, we introduce the phylotranscriptomic approach to combine gene expression data with evolutionary gene ages. We show that the transcriptomic hourglass pattern is actively maintained in extant species. Furthermore, our results suggest that the transcriptomic hourglass pattern is decoupled from organogenesis and may function as a switch to enable developmental transitions. We further develop a novel phylotranscriptomic measure based on the Shannon entropy showing highly significant transcriptomic hourglass patterns that might be primary to the traditional transcriptomic hourglass patterns.

In order to analyze transcriptomic changes of developmental processes in more detail, we investigate spatio- temporal gene expression data to understand the transcriptome dynamics during grafting. Grafting is a unique feature of plants that allows them to form chimeric organisms through joining previously cut tissues. We detect differentially expressed protein-coding genes showing spatio- and temporal specific expression patterns.

Understanding developmental processes like grafting more precisely, it is essential to study the interaction of protein-coding genes and non-coding transcripts. In plant research, the knowledge of non-coding transcripts is limited. We attempt to overcome these limitations by annotating novel protein-coding splice variants, long non-coding RNAs, and circular RNAs in seven different flowering plants based on organ-specific RNA-Seq data. We develop a reproducible automated pipeline to process the data and annotate novel coding and non-coding transcripts.

Besides this diverse transcriptome landscape, we develop machine learning approaches and apply statistical methods to investigate the metabolomic diversity. Therefore, we quantify and characterize the effects of aging and diabetes on the concentrations of free fatty acids in the outer barrier of the human skin. We also take a closer look at the amino acid concentrations of serum metabolite profiles to unravel the contributions of different dietary protein sources on the metabolism.

In summary, modern life sciences have evolved to an unprecedented level of diversity coupled with immense amounts of data from various sources. Robust methods are required for quantifying diverse observations and for stating their statistical significance. We establish various methods and workflows to analyze and interpret these diverse data.

Contents

1 Introduction 1

2 Transcriptomic hourglass patterns in plant embryogenesis 5

2.1 Introduction 5
2.2 Materials and Methods 7
2.2.1 Phylostratigraphic analysis 7
2.2.2 The transcriptome age index 9
2.2.3 Calculation of sequence divergence levels 10
2.2.4 The transcriptome divergence index 10
2.2.5 Statistical significance of TAI and TDI profiles 11
2.2.6 Relative expression levels for phylostrata 11
2.2.7 Relative expression levels for K_a/K_s quantiles 12
2.3 Results and Discussion 13
2.3.1 Measuring the transcriptome evolution 13
2.3.2 Comparison of TAI and TDI 14
2.3.3 Transcriptome indices uncover transcriptomic hourglass patterns 15
2.4 Conclusions and Outlook 18

3 Evidence for active maintenance of the phylotranscriptomic hourglass 21

3.1 Introduction 21
3.2 Materials and Methods 22
3.2.1 Phylostratigraphy - Construction of phylostratigraphic maps 23
3.2.2 Divergence stratigraphy - Constructing sequence divergence maps 24
3.2.3 Processing of expression data 25
3.2.4 Reductive hourglass test 26
3.2.5 Reductive early conservation test 27
3.3 Results 28
3.3.1 TAI Profiles of *Da. rerio*, *D. melanogaster*, and *A. thaliana* 28
3.3.2 Dependence of PS and DS 31
3.3.3 TDI Profiles of *Da. rerio*, *D. melanogaster*, and *A. thaliana* 32
3.3.4 Statistical testing for potential hourglass patterns 34
3.4 Discussion 36
3.5 Conclusions and Outlook 39

4 Post-embryonic hourglass patterns mark ontogenic transitions in plant development 41

4.1 Introduction 42

4.2 Materials and Methods . . . 43
4.2.1 Germination experiment . . . 43
4.2.2 Floral transition experiment . . . 44
4.2.3 Expression data of flower development . . . 45
4.2.4 Phylotranscriptomic analyses . . . 45
4.3 Results and Discussion . . . 45
4.3.1 Germination - Transition to vegetative phase . . . 45
4.3.2 Floral transition - Vegetative-to-reproductive phase . . . 46
4.3.3 Flower development - Formation of floral organs . . . 49
4.4 Conclusions and Outlook . . . 50

5 Entropic hourglass patterns of animal and plant development 53
5.1 Introduction . . . 53
5.2 Materials and Methods . . . 54
5.2.1 A probabilistic perspective on the transcriptome age index . 55
5.2.2 Entropic TAI . . . 55
5.2.3 Phylostratigraphic analysis pipeline . . . 56
5.2.4 PhyloWeb - Retrieving phylostratigraphic maps . . . 56
5.3 Results and Discussion . . . 58
5.3.1 Hourglass patterns of the entropic TAI . . . 58
5.3.2 Entropic hourglass patterns are highly significant . . . 60
5.4 Conclusions and Outlook . . . 61

6 Deciphering the origin of the hourglass pattern 63
6.1 Introduction . . . 63
6.2 Materials and Methods . . . 64
6.2.1 Sampling of PS distributions and regression analysis . . . 65
6.2.2 Prediction of transcriptomic age index profiles based on a mean PS distribution . . . 66
6.2.3 Normalized transcriptomic age index profiles . . . 67
6.3 Results and Discussion . . . 68
6.3.1 Dependencies between transcriptome indices . . . 69
6.3.2 Reproducing the hourglass pattern . . . 70
6.4 Conclusions and Outlook . . . 72

7 Transcriptome dynamics at Arabidopsis graft junctions 75
7.1 Introduction . . . 76
7.2 Materials and Methods . . . 78
7.2.1 RNA-Seq sample and library preparation . . . 78
7.2.2 Quantification of gene expression . . . 78
7.2.3 Detection of differentially expressed genes . . . 79
7.2.4 Comparison of up- and down- regulated genes with published datasets . . . 80
7.2.5 Fisher's exact test of ratios . . . 80

7.2.6 Dealing with probe ids from microarray datasets 81
7.2.7 Gene sets involved in graft formation 82
7.2.8 GO enrichment analysis 82
7.3 Results 83
7.3.1 Grafting activates vascular formation and cell division genes 83
7.3.2 Genes are asymmetrically expressed around the graft 86
7.3.3 Sugar response correlates with asymmetric gene expression . 88
7.3.4 Auxin response is symmetric at the graft 89
7.3.5 Tissue fusion imparts a unique physiological response that differs from tissue separation 92
7.4 Discussion 95
7.5 Conclusions and Outlook 98

8 Annotation of splice-variants, lncRNAs, and circRNAs in flowering plants 101
8.1 Introduction 101
8.2 Materials and Methods 103
8.2.1 Data 103
8.2.2 Annotation workflow 104
8.2.3 Read mapping and Transcript Assembly 105
8.2.4 Calculation of Coding Potentials 108
8.2.5 Filtering novel putative protein-coding loci and isoforms . . 110
8.2.6 Prediction of long non-coding RNAs 111
8.2.7 Prediction of circular RNAs 113
8.2.8 Implementation 115
8.3 Results 115
8.3.1 Annotation workflow 116
8.3.2 Protein-coding genes and isoforms 118
8.3.3 Long non-coding RNAs 120
8.3.4 Circular RNAs 123
8.3.5 GC content of coding and non-coding RNAs 125
8.4 Discussion 127
8.5 Conclusions and Outlook 130

9 Age- and Diabetes-related Changes in the FFA Composition 131
9.1 Introduction 132
9.2 Materials and Methods 135
9.2.1 Subjects 135
9.2.2 Extraction of SC lipids 135
9.2.3 Gas chromatography – Flame ionization detection analysis . 136
9.2.4 Statistical analysis 137
9.3 Results 137
9.3.1 Quantification and identification of FFA concentrations . . . 137
9.3.2 Comparison of elderly/healthy vs. young/healthy 138

9.3.3 Comparison of young/healthy vs. young/diabetic 139
9.4 Discussion 140
9.5 Conclusions and Outlook 143

10 Metabolic footprint in response to dietry proteins in a pig model 145
10.1 Introduction 146
10.2 Materials and Methods 147
10.2.1 Animals and study design 148
10.2.2 Characterization of the experimental proteins 148
10.2.3 Sample collection 149
10.2.4 Statistical analysis 149
10.2.5 Linear discriminant analysis 150
10.3 Results 152
10.3.1 Amino acid composition of the dietary protein and serum levels of amino acids and metabolites 153
10.3.2 SAM and SAH concentrations in liver and serum homocysteine levels 157
10.3.3 Hepatic methylation 158
10.3.4 Impacts of dietary proteins on health-relevant factors 158
10.4 Discussion 160
10.5 Conclusions and Outlook 164

11 Conclusions and Outlook 165

12 Appendix 195

List of Figures

2.1 Phases of *A. thaliana* embryogenesis 6
2.2 Phylogeny of *A. thaliana* reconstructed from extant species with completely sequenced genomes 8
2.3 Evolutionary age and sequence divergence of *A. thaliana* genes . . . 13
2.4 Transcriptome indices across *A. thaliana* embryogenesis 16
2.5 Relative expression levels over embryo stages 17
2.6 Convergent evolution of a molecular hourglass in animal and plant embryogenesis 19

3.1 NCBI taxonomy tree representing the major groups of species/genomes used for BLAST database 23
3.2 Reductive hourglass test statistic 27
3.3 Phylostratigraphic maps for *Danio rerio*, *Drosophila melanogaster*, and *Arabidopsis thaliana* 29
3.4 TAI profiles across animal and plant embryogenesis 30
3.5 Correlation between phylostratum (PS) and divergence stratum (DS) 31
3.6 TDI profiles across animal and plant embryogenesis 33
3.7 Evaluation of transcriptome index profiles by the reductive hourglass test 34

4.1 The developmental hourglass model in the context of differences in plant and animal development 42
4.2 TAI analysis for germination in *A. thaliana* 47
4.3 TAI analysis for the transition from vegetative to reproductive growth in *Arabidopsis thaliana* 48
4.4 TAI analysis of flower development in *Arabidopsis thaliana* 50

5.1 Screenshot of the PhyloWeb server 57
5.2 Entropic TAI profiles across animal and plant development 59

6.1 Scatterplot of sampled PS distributions 69
6.2 Normalized mean and entropic transcriptome age indices 71

7.1 Transcriptional dynamics of genes associated with provasculature, phloem, and xylem development and cell division 84
7.2 Transcriptional overlap between previously published vascular datasets and the grafting datasets 85
7.3 Asymmetric changes in accumulation of sugar-responsive RNAs and of starch occur at the graft junction 87
7.4 Auxin response is symmetric at the graft junction 90
7.5 $HCA2$ contributes to graft junction formation 91

7.6 Clustering the transcriptome at each time point, based on likelihoods of all possible patterns of differential expression (DE) in grafted, separated, and intact tissues 94
7.7 A subset of genes is differentially expressed only during graft formation compared with intact or separated tissues 95

8.1 Phylogeny of flowering plant species 103
8.2 Overview of the annotation workflow 105
8.3 Mapping statistics 106
8.4 Annotation of protein-coding and non-coding transcripts 108
8.5 Classification of lncRNAs 112
8.6 Annotation of circular RNAs in *A. thaliana* 114
8.7 Amount of coding and non-coding genes and transcripts in flowering plants 116
8.8 Comparison of splicing events 119
8.9 Transcripts per gene and ORF length of lncRNAs and protein-coding mRNA transcripts 121
8.10 Transcript and exon lengths distributions and mean number of exons of RNA species 122
8.11 Distance of lincRNA to nearest protein-coding locus 123
8.12 Location and backsplicing of predicted circRNAs 124
8.13 GC content of exons and introns within different RNA species . . . 125
8.14 GC content vs. transcript lengths 126

9.1 FFA synthesis in the cytosol of a keratinocyte done by multi-enzyme complex and subsequent storage in the lamellar bodies 133
9.2 Description of a chromatogram example of GC with fatty acid standards 136
9.3 Box plot comparing FFA concentrations in µg/mL of elderly/healthy and young/healthy subgroups 138
9.4 Box plot comparing FFA concentrations in µg/mL of young/healthy and young/diabetic subgroups 139

10.1 Composition and concentrations of amino acids in dietary proteins and serum of pigs 153
10.2 Concentrations of characteristic serum metabolites of pigs after ingestion of different dietary proteins 154
10.3 Results of the LDA based on logarithmically transformed serum concentrations [$\log_{10}$(serum concentration in µM+1)] 157
10.4 Methylation analysis and one carbon cycle 159

12.1 5' and 3' UTR lengths of protein-coding transcripts. 195
12.2 Overview of linear splice sites among RNA species 196

List of Tables

5.1 P values of the reductive hourglass test 60

7.1 Contingency table for Fisher's exact test 81

8.1 BUSCO evaluation of annotation completeness 117
8.2 Novel putative protein-coding loci 118
8.3 Statistics of protein-coding isoforms 119

9.1 Relative concentrations of FFAs in the different subgroups (%) ± SD 140
9.2 Bonferroni corrected P values derived from the Mann-Whitney U test 141

10.1 Serum concentrations of amino acid metabolites, health risk parameters, relative mRNA concentrations, mineral concentrations and plasma glucose concentrations 156

12.1 Statistics UTR lengths 195
12.2 Overview sequenced species and organs 201
12.3 Sources of reference annotations 202
12.4 GC content and quintiles of transcript lengths in coding and non-coding RNA species 205

1

Introduction

Bioinformatics has become an essential and indispensable discipline in modern data-driven branches of the life sciences such as biology, biochemistry, pharmacy, agricultural science, nutritional science, and medicine. It marks the turning point from intuitive and qualitative thinking to quantitative reasoning. Its origin can be dated back to the early 1960s when Margaret O. Dayhoff implemented the first *de novo* sequence assembler to decipher large protein sequences. Her efforts, and those of her co-author Robert S. Ledley, to integrate computational science into biology gave rise to a discipline that should 25 years later be known as Bioinformatics.

Starting at sequence analysis of proteins, the deciphering of the genetic code led to the computational analysis of DNA and RNA as sources of biological information. With an increase in cost-efficient DNA sequencing, the need for bioinformatic analyses first peaked during the mid-1990s due to the demand for specialized software to handle the unprecedented amount of data generated by the Human Genome Project [1, 2].

Ever since, Bioinformatics has been established as an interdisciplinary science bridging experimental, computational, and statistical research in the natural sciences. Besides pure sequence analyses of DNA, RNA, and proteins, the applications of bioinformatics research today are intertwined with all modern branches of biology including genomics, transcriptomics, and metabolomics.

The goal of my PhD research work was to contribute to the development of this broad spectrum of bioinformatics applications in natural sciences by combining concepts and methods from machine learning and statistics and by applying them to analyze diverse biological data in phylotranscriptomics, transcriptomics, and metabolomics.

This thesis represents the different research areas in which my coworkers and I developed software to analyze the biological data and interpreted the obtained findings. Because each chapter deals with a different topic, each of them starts with an introduction providing elementary information to understand the biological problem.

In chapter 2, we will start with the developmental hourglass phenomenon; a still-controversially discussed concept in evolutionary developmental biology. We will learn about the morphological hourglass patterns and transcriptomic hourglass patterns only observed in animal embryogenesis. Additionally, we will introduce the phylotranscriptomics approach, which is the combination of gene expression data and evolutionary gene ages to study developmental processes in the light of evolution. Based on this approach, we will be able to detect the transcriptomic hourglass pattern in plant embryogenesis. Our findings suggest, that transcriptomic hourglass patterns of embryogenesis have emerged independently in animals and plants. We will conclude with the question if the observed patterns are non-functional remnants or still functional.

In chapter 3, we will turn to this question by investigating the active maintenance of the transcriptomic hourglass pattern. We will describe the developed statistical approaches to systematically quantify the observed transcriptomic hourglass patterns. Based on our findings, we will conclude that the observed patterns are actively maintained and thus might be of functional relevance.

In chapter 4, based on the active maintenance of the transcriptomic hourglass pattern, we will attempt to find out a functional explanation for the pattern. As proposed in the literature, the hourglass phenomenon in animal embryogenesis is coupled with the establishment of the body plan, and thus coupled with the process of organogenesis [3]. In contrast to animals, organ formation in plants occurs largely postembryonically. In order to determine if the transcriptomic hourglass in plants is connected to developmental transitions or to organogenesis, we will perform phylotranscriptomic analyses of transcriptome data from germination, floral transition, and flower development. We will conclude that the transcriptomic hourglass pattern may function as a switch to enable the transition from one functional program to the next.

In chapter 5, we will redefine the transcriptomic measures of the previous chapters from a probabilistic perspective and develop a novel phylotranscriptomic approach based on the Shannon entropy, which measures the homogeneity of the age distributions as a function of time. We will also learn about the PhyloWeb server enabling the automated calculation of gene ages and allowing researchers to access our calculated gene age assignments. With the updated gene age calculations and the Shannon entropy based phylotranscriptomic approach, we will find highly significant transcriptomic hourglass patterns. We will conclude this chapter with the hypothesis that these novel transcriptomic hourglass patterns could be primary to the traditional transcriptomic hourglass patterns.

In chapter 6, we will continue our attempt to decipher the origin and thus eventually the function of the transcriptomic hourglass pattern. We will start by the investigation of the dependencies between the two proposed phylotranscriptomic measures. Afterwards, we will develop a gradient-based approach to reproduce the entropic hourglass patterns based on the traditional measure or to reproduce the traditional transcriptomic hourglass pattern based on the entropic measure. Based on our findings, we will partially answer the question of the origin, and thus we will get closer to understand the functional relevance of the developmental hourglass pattern. The phylotranscriptomic analyses of the chapters 2 - 6 will be based on temporal gene expression data of a developing organism. In contrast, a more detailed analyses of transcriptomic changes of developmental processes can be performed by adding a spatial dimension.

In chapter 7, we will study the transcriptome dynamics during grafting based on spatio-temporal gene expression data. Grafting is an agriculturally relevant and unique feature of plants enabling increased productivity and yield. The grafting process allows the formation of chimeric organisms through joining previously cut tissue. We will develop and perform statistical analyses to yield information about differential expressed genes that play an essential role in grafting. We will detect sets of genes showing spatio- and temporal specific expression patterns. These genes and their functional relevance will help us to deepen our understanding of the intertissue recognition and wound healing mechanisms. In order to get a more precise picture of the grafting processes, the analysis of non-coding transcripts, such as long non-coding RNAs and circular RNAs, and their interaction with protein-coding genes will be essential.

In chapter 8, we will investigate the transcriptome landscape of flowering plants and present a workflow for the annotation of non-coding transcripts and splice variants of protein-coding genes in seven flowering plants. We will describe the sequencing experiments and the developed annotation workflow to create comprehensive transcriptome annotations of long non-coding RNAs, circular RNAs, and novel splice variants of protein-coding genes based on organ-specific RNA-Seq data. Additionally, we will compare genomic features of non-coding and protein-coding transcripts within and between the different species. Besides the diversity within the transcriptome, a living cell depends on additional groups of compounds such as amino acids, proteins, carbohydrates and lipids. The analyses of these compounds are summarized under the umbrella of metabolomics.

In chapter 9, we will turn our attention to the field of lipidomics, a subdomain of metabolomics. We will analyse the concentrations of free fatty acids of the Stratum corneum, the outer barrier of the human skin. We will learn about the effects of aging and diabetes on the barrier function of this outer layer of the human skin. To quantify and characterize these effects, we will perform univariate and multivariate statistical analyses. By the identification of changes in the free fatty acid composition, we will identify free fatty acids that may have the potential to improve skin recovery and protection.

In chapter 10, we will continue in the field of metabolomics and we will take a closer look at the amino acid concentrations of serum metabolite profiles among others. We will describe statistical analyses to unravel the contributions and effects of three different dietary protein sources on the metabolism. Based on a linear discriminant analysis coupled with leave-one-out cross-validations, we will investigate the multivariate distribution of amino acids in the serum metabolite profiles and attempt to uncover sets of biomarkers helping to discriminate between the intake of the different dietary protein sources.

Chapters 2 - 10 represent a subset of research projects my colleagues and I have worked on during the past years. In addition, I was very fortunate to be allowed to work on additional research projects during my PhD and to contribute to publications in the fields of transcriptomics and metabolomics. In transcriptomics I had the opportunity to work on the gene expression of flower development [4], the transcriptome of polyspermy-derived triparental plants [5], and RNA-Seq data from several space-omics projects [6], topics related to chapters 7 and 8. In the field of lipidomics, I was able to contribute to the work of our colleagues from medicine and pharmacy on the analysis of ceramide lipids in the Stratum corneum to support the repairing of the skin barrier [7], a topic related to chapter 9 of this thesis.

In chapter 11, we will draw conclusions beyond those that could be drawn at the ends of each of the nine chapters and present an outlook for future studies. We will connect the different topics and demonstrate the immense diversity of bioinformatics in various applications. This diversity is reflected by our work with different organisms from the plant and the animal kingdom. We investigated different developmental processes such as embryogenesis, grafting, aging, and others. We analyzed diverse data sets from different sources, such as sequencing data from RNA-Seq or metabolomics data from gas chromatography. Finally, this diversity is reflected in our participation in research projects covering diverse branches of the life sciences such as evolutionary developmental biology, developmental biology, pharmacy, medicine, nutritional science, and space science.

2

Transcriptomic hourglass patterns in plant embryogenesis

The developmental hourglass pattern was first observed based on morphological similarities during vertebrates' embryogenesis. Embryos from different taxa appeared dissimilar in early stages, appear highly similar during mid-embryogenesis, and again appear dissimilar in later stages. Analyses in *Danio rerio* and several subspecies of *Drosophila* embryos could show a similar pattern on the molecular level. Besides, an embryonic hourglass has not been reported in plants. In this chapter, we will investigate the transcriptome of *Arabidopsis thaliana* during embryogenesis and we will provide evidence for a transcriptomic hourglass based on two complementary approaches.

In section 2.1, we will introduce the developmental hourglass model and we will give insights into the embryogenesis of plants. In section 2.2, we will learn about analyses to quantify transcriptomic hourglass patterns that my colleagues and I developed based on evolutionary approaches introduced by [8, 9]. In section 2.3, we will present the transcriptomic hourglass pattern of plant embryogenesis in the model plant *A. thaliana*, and we will discuss possibilities of its origination. In section 2.4, we will speculate about a possible explanation of the transcriptomic hourglass in animals and plants.

The following sections are extracted from Quint et al. 2012 "*A transcriptomic hourglass in plant embryogenesis*" [10].

2.1 Introduction

Animal and plant development starts with a constituting phase called embryogenesis, which evolved independently in both lineages [11]. Comparative anatomy of vertebrate development - based on the Meckel-Serrès law [12] and von Baer's laws of embryology [13] from the early nineteenth century - shows that embryos from various taxa appear different in early stages, converge to a similar form during mid - embryogenesis, and again diverge in later stages. This morphogenetic series is known as the embryonic “hourglass” [14, 15], and its bottleneck of high conservation in mid-embryogenesis is referred to as the phylotypic stage [3].

Recent analyses in *Danio rerio* and *Drosophila* embryos provided convincing molecular support for the hourglass model, because during the phylotypic stage the transcriptome was dominated by ancient genes [9] and global gene expression profiles were reported to be most conserved [16].

In flowering plants, embryogenesis can be separated into three major phases (Fig. 2.1. The early phase is characterized by asymmetric cell divisions to establish apical–basal polarity. In the intermediate phase, major organs and primordia are initiated, which expand in the late phase to the mature embryo [17, 18].

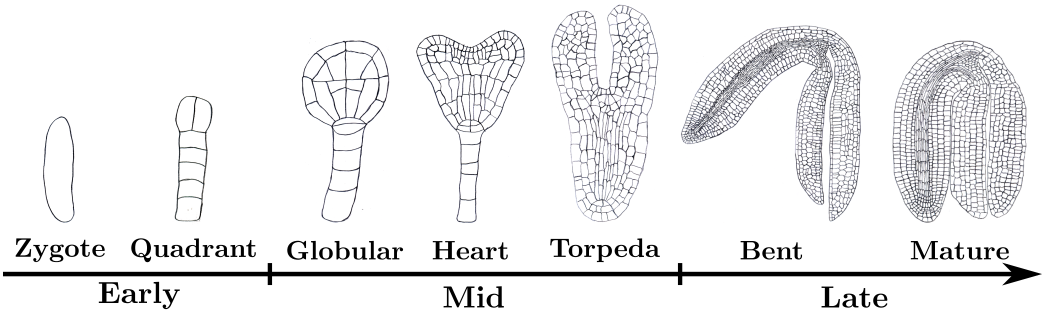


Figure 2.1 | Phases of *A. thaliana* embryogenesis. The early phase of embryogenesis is represented by the developmental stages of zygote and quadrant. During this phase the embryo establishes the apical-basal axis. The zygote divides asymmetrically to form the apical and basal cells of the quadrant. In the mid or intermediate phase represented by the globular, heart, and torpedo stage, the radial axis gets established, major organs and primordia are initiated. In the late phase the mature embryo is formed.

One notable difference between embryogenesis in animals and plants concerns the establishment of morphological variation between taxa. For example, vertebrates develop morphological variation in late embryogenesis, whereas differences between flowering plant taxa are only established during post-embryonic development. Inspired by the historical relevance of the embryonic hourglass model in animals, by recent transcriptional support from studies in *Da. rerio* [9] and *Drosophila* [9, 16], and by the absence of any reported anatomical evidence for such a pattern during plant embryogenesis, we assess the possible existence of a transcriptional hourglass during embryogenesis of the plant reference species *A. thaliana*.

2.2 Materials and Methods

The main findings in this chapter are based on two different transcriptome indices. Both combine gene expression data with evolutionary information. In this section, we will explain the calculation of these two measures and statistical approaches to quantify their time course pattern. In subsection 2.2.1, we will introduce phylostratigraphy to determine the evolutionary gene age, phylostratum. In subsection 2.2.2, we will define the transcriptome age index as a combination of phylostrata and gene expression data. In subsection 2.2.3, we will describe the calculation of sequence divergence levels as K_a/K_s ratios. In subsection 2.2.4, we will combine the K_a/K_s ratios with gene expression data to define the transcriptome divergence index. In subsection 2.2.5, we will explain the test statistics to quantify the existence of a transcriptomic hourglass pattern. In subsection 2.2.6, we will learn about the calculation of relative expression levels of each phylostratum. Finally, in subsection 2.2.7, we will get to know the calculation of relative expression levels of each K_a/K_s quantile.

2.2.1 Phylostratigraphic analysis

Macroevolutionary trends have been generally studied by fossil analysis or morphological comparisons. Due to the increasing number of sequenced genomes and established algorithms for efficient and precise sequence comparisons, phylostratigraphy defines an alternative method, which tries to reveal the origins of each protein-coding gene in a species of interest by identifying homologous sequences in a sequence database representing the tree of life. The phylostratigraphic approach was introduced by Domazet-Lošo et al. in 2007 [8] and can be explained with three essential steps.

In the first step, the target species' phylogeny gets reconstructed based on sequence information of extant species. This reconstruction involves retrieving all available protein sequences from sequenced organisms and building a comprehensive sequence database. The sequences in this database are hierarchically ordered based on the target species' phylogeny, whereas each node in the phylogeny represents an evolutionary age category, called phylostratum (PS).

Following this procedure, we downloaded the amino acid sequences of 1,459 species with completely sequenced genomes (see Supplementary Table 1 of [10]). The species, resp. their amino acid sequences, were then sorted into 13 phylostrata (Fig. 2.2) representing the phylogeny of *A. thaliana*.

In the second step, each protein-coding gene of the target species is blasted against the comprehensive sequence database to identify homologous sequences. Due to their robustness against mutation events, which occur during evolution, we compare only the translated amino acid sequences against the comprehensive sequence database, also containing only amino acid sequences.

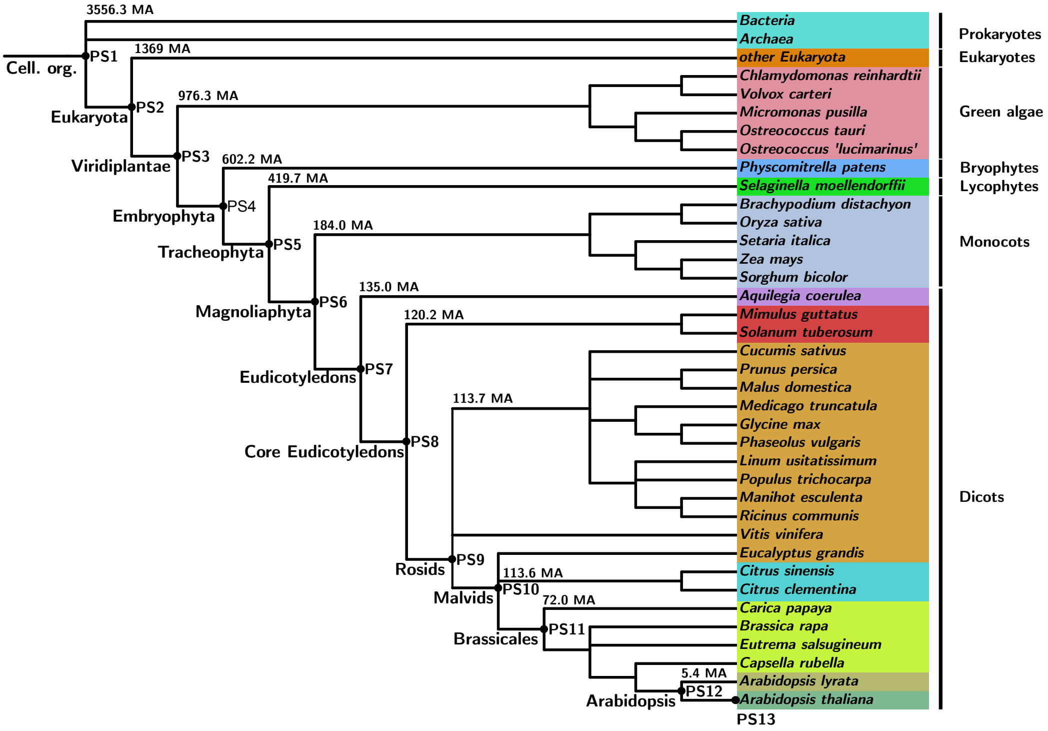


Figure 2.2 | Phylogeny of *A. thaliana* reconstructed from extant species with completely sequenced genomes. Phyla have been assigned according to the NCBI taxonomy database and sorted into the phylostrata (PS) shown at the respective nodes. The phylogenetic tree was created with iTOL [19] (`http://itol.embl.de`). All divergence times were obtained from `http://www.timetree.org` [20]. Cell. org., cellular organisms described by PS1. Adapted Supplementary Figure 1 from [10].

In the third step, according to the phylogenetically most distant species in which BLAST identified homologous sequences, each gene is assigned into a PS. A BLAST hit is homologous if its E-value is below 10^{-5}. If no homologous sequence of a gene from the target species shows similarities to sequences in the sequence database, the gene gets assigned to the youngest PS. The assignment of all genes from a target species into their corresponding PS is called a phylostratigraphic map.

Due to the nature of the BLASTP approach, genes are assigned to PS according to any detectable homology between query and target protein sequences. Thus multi-domain proteins, for example, are mapped to the PS of the oldest domain irrespective if another functional domain has evolved more recently.

For illustration consider the example of a protein of 1,000 amino acids length that consists of a short domain of 50 amino acids conserved among all eukaryotes, whereas the remaining sequence of 950 amino acids is *A. thaliana* specific without any detectable homology to other species. Although in this example the vast majority of the sequence has evolved only recently after divergence of the two sister species *A. thaliana* and *Arabidopsis lyrata*, this protein would be sorted into the ancient PS2. Since roughly 30% of plant proteins have been designated as multi-domain proteins (Supplementary Fig. 4 in [21]), this phenomenon may affect a significant portion of the plant proteins.

Hence, the phylostratigraphic approach identifies founder genes or domains that have emerged at a certain time (or PS) in evolution. Based on this founder gene or domain, additional family members have arisen subsequently by mechanisms like gene duplication or structural rearrangements, resulting in the incorporation of this founder gene or domain in genes of multi-domain proteins. Importantly, phylostratigraphy does not distinguish between orthologs and paralogs. Furthermore, phylostratigraphy considers evolutionary time since the emergence of cellular life roughly four billion years ago until today.

2.2.2 The transcriptome age index

The transcriptome age index (TAI) was initially introduced by Domazet-Lošo in 2010 [9]. It is based on the previously described gene age inference approach of phylostratigraphy [8]. The resulting phylostratigraphic map is an assignment of each protein-coding gene to a discrete age category (PS). To construct the TAI measure, phylostratigraphy based gene age inference is performed for all protein-coding genes of a reference organism of interest (here *A.thaliana*). The phylostratigraphic map is combined with expression levels covering the biological process of interest.

The TAI of developmental stage $s \in \{\text{zygote, quadrant, globular, heart, torpedo, bent cotyledon, mature}\}$ is defined as the weighted mean of the phylostratum ps_i of gene i weighted by the expression level e_{is} of gene i at developmental stage s

$$TAI_s = \frac{\sum_{i=1}^{I} ps_i \; e_{is}}{\sum_{i=1}^{I} e_{is}} \tag{2.1}$$

where I is the total number of genes analysed. Low PS values correspond to evolutionarily old genes, so low TAI values correspond to evolutionarily old transcriptomes. Likewise, high PS values correspond to evolutionarily young genes, so high TAI values correspond to evolutionarily young transcriptomes.

2.2.3 Calculation of sequence divergence levels

The levels of sequence divergence, represented by the K_a/K_s ratio, is an indicator of selective pressure within protein-coding regions. It reflects natural selection, one of the major forces driving evolution.

The basis of the calculation of sequence divergence levels is a global alignment of protein sequences that are orthologous between species. Here, orthologous gene pairs of *A. thaliana* and *A. lyrata*, *A. thaliana* and *Eutrema salsugineum* (formerly known as *Thellungiella halophila*), *A. thaliana* and *Capsella rubella*, or *A. thaliana* and *Brassica rapa* were determined by taking the best hit of the corresponding BLASTP searches.

Then, each orthologous gene pair's amino acid sequences get globally aligned with MAFFT [22] (L-INS-i option) because the K_a/K_s ratio represents the full-length of the sequence and not only a subsequence. The resulting alignment is converted into a codon alignment with PAL2NAL [23] to subsequently compute the K_a/K_s ratio with GESTIMATOR [24].

The K_a/K_s ratio of a gene is the number of nonsynonymous substitutions per nonsynonymous site (K_a) divided by the number of synonymous substitutions per synonymous site (K_s). Gene pairs with $K_a < 0.5$, $K_s < 5$, and ratios $K_a/K_s < 2$ were retained.

2.2.4 The transcriptome divergence index

Analogous to the TAI, we introduce the transcriptome divergence index TDI_s of developmental stage $s \in \{$zygote, quadrant, globular, heart, torpedo, bent cotyledon, mature$\}$ by replacing ps_i in Eq. 2.1 by the K_a/K_s ratio of gene i,

$$TDI_s = \frac{\sum_{i=1}^{I} \frac{K_{ai}}{K_{si}}\, e_{is}}{\sum_{i=1}^{I} e_{is}} \tag{2.2}$$

Hence, low K_a/K_s ratios correspond to conserved genes, so low TDI values correspond to conserved transcriptomes. In contrast, high K_a/K_s ratios correspond to divergent genes, so high TDI values correspond to divergent transcriptomes. The same procedure was repeated for the second independent dataset covering the embryo propers of *A. thaliana* embryo stages pre-globular, globular, heart, linear cotyledon/torpedo, and mature [25, 26] (GEO accession number GSE12404). We normalized this dataset using the GCRMA package (version 2.0) from the Bioconductor project with default parameter settings [27]. For each probe set we computed the stage-wise arithmetic mean of the replicates to get representative expression values for each stage. Here, 20,031 genes represented on the microarrays were included in the analyses

2.2.5 Statistical significance of TAI and TDI profiles

To determine the statistical significance of the TAI and TDI profiles, the following permutation test was performed. The variance V_{TAI} of the seven values of TAI_s was computed as test statistic. For determining the null distribution of V_{TAI}, all PS values within each developmental stage s were randomly permuted, seven surrogate values of TAI_s were computed from this permuted dataset, and a surrogate value of V_{TAI} was computed from these seven surrogate values of TAI_s.

This procedure was repeated 1,000 times, yielding a histogram of 1,000 values of V_{TAI}, which can be approximated by a gamma distribution. The two parameters of the gamma distribution were estimated by the method of moments, the fitted gamma distribution was considered the null distribution of V_{TAI}, and the P value of the observed value of V_{TAI} was computed from this null distribution. The same procedure was repeated for the seven values of TDI_s, yielding a P value of the TDI profile. Likewise, the second dataset [25, 26] was analysed accordingly.

In the following chapters, we will refer to this test statistic as the flat line test, because it statistically quantifies the difference of the observed profile from a flat horizontal line.

2.2.6 Relative expression levels for phylostrata

Relative expression levels were computed as described in [9]. The mean expression level $f_s(ps)$ of phylostratum ps and developmental stage s was computed for each ps and s as the arithmetic mean of expression levels e_{is} of all genes i belonging to phylostratum ps. The mean expression levels $\overline{f}_s(ps)$ were linearly transformed to the interval $[0, 1]$ according to Eq. 2.3 where $\overline{f}_{min}(ps)/\overline{f}_{max}(ps)$ is the minimum/maximum mean expression level of phylostratum ps over the seven developmental stages s.

$$f_s(ps) = \frac{\overline{f}_s(ps) - \overline{f}_{min}(ps)}{\overline{f}_{max}(ps) - \overline{f}_{min}(ps)} \tag{2.3}$$

This linear transformation corresponds to a shift by $\overline{f}_{min}(ps)$ and a subsequent shrinkage by $\overline{f}_{max}(ps) - \overline{f}_{min}(ps)$. As a result, the relative expression level $f_s(ps)$ of developmental stage s with minimum $\overline{f}_s(ps)$ is 0, the relative expression level $f_s(ps)$ of the developmental stage s with maximum $f_s(ps)$ is 1, and the relative expression levels $f_s(ps)$ of all other stages s range between 0 and 1, accordingly.

Next, relative expression levels were grouped into two PS classes, where the first PS class consists of relative expression levels of genes belonging to the three oldest phylostrata PS1–PS3, and where the second PS class consists of relative expression levels of genes belonging to the younger phylostrata PS4–PS13. This grouping was chosen to distinguish phylostrata of plants that pass through embryogenesis (PS4–PS13) from the remaining phylostrata (PS1–PS3), in which the vast majority of species did not evolve embryogenesis. For each developmental stage s and each PS class, the mean value and standard error of the relative expression levels were computed. In addition, the ratio (fold-change) of the two relative expression levels was computed for each developmental stage s, and Welch's two-sample t-test was performed.

2.2.7 Relative expression levels for K_a/K_s quantiles

In contrast to PS values, which are discrete, K_a/K_s ratios are continuous. For computing relative expression levels of genes belonging to different K_a/K_s groups, continuous K_a/K_s ratios were grouped into deciles (10% quantiles). Relative expression levels of these ten K_a/K_s groups were computed in analogy to the computation of relative expression levels of the 13 phylostrata. Likewise, relative expression levels were grouped into two K_a/K_s classes, where the first K_a/K_s class consists of relative expression levels of genes belonging to the first five K_a/K_s groups (K_a/K_s ratios below median, conserved genes), and where the second K_a/K_s class consists of relative expression levels of genes belonging to the remaining five K_a/K_s groups (K_a/K_s ratios above median, divergent genes).

This grouping was chosen because the median is a natural choice, making both K_a/K_s classes equally large, and because the grouping of genes into different PS classes also resulted in two PS classes of roughly similar sizes (first dataset: 10,695 genes in PS1–PS3 and 14,463 genes in PS4–PS13; second dataset: 9,028 and 11,003 genes, respectively). The computation of mean values, standard errors, fold-changes and P values of Welch's two-sample t-test were performed as described in the previous subsection. To investigate the dependence of the results on the grouping into two K_a/K_s classes, the entire analysis was repeated for the following six pairs of K_a/K_s classes: two deciles/eight deciles, three deciles/seven deciles, ..., seven deciles/three deciles. The six resulting plots of means, standard errors, fold-changes and P values are presented in Supplementary Figs 6–9, 11–14 [10].

2.3 Results and Discussion

We obtained genome-wide expression profiles of a complete developmental series from the zygote to the mature embryo in *A. thaliana* from [28]. To investigate the presence of an embryonic transcriptomic hourglass pattern in plants, we combine this transcriptome information with two different measures of evolutionary distance: evolutionary age and sequence divergence. In subsection 2.3.1, we will introduce this procedure, followed by a comparison of the two transcriptome measures TAI and TDI in subsection 2.3.2. Finally, in subsection 2.3.3, we will present the resulting TAI and TDI profiles calculated across the embryogenesis of *A. thaliana*.

2.3.1 Measuring the transcriptome evolution

We compute two different transcriptome indices for each gene, the transcriptome age index (TAI) [9] based on evolutionary age, and the transcriptome divergence index (TDI) based on sequence divergence. We investigate the profiles of these two transcriptome indices across the seven sampled embryo stages, and ask if and to what degree they show an hourglass pattern similar to that found for *Da. rerio* [9] or *Drosophila melanogaster* [9].

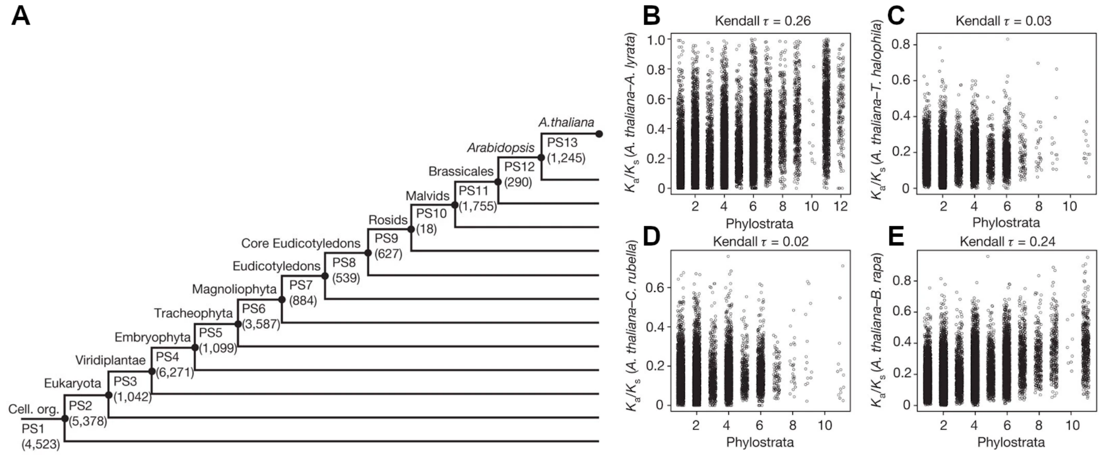


Figure 2.3 | Evolutionary age and sequence divergence of *A. thaliana* genes. (A), Phylostratigraphic map of *A. thaliana*. Numbers in parenthesis denote the number of genes per phylostratum (PS1–PS13). Cell. org., cellular organisms described by PS1.(B)–(E), Scatter plots of phylostratum versus K_a/K_s ratios over all genes. K_a/K_s ratios are derived from orthologous genes between *A. thaliana* and (B), *A. lyrata*, (C), *E. salsugineum*, (D), *C. rubella* and (E), *B. rapa*. Kendall τ values denote the Kendall rank correlation coefficients measuring the association between both parameters. Adapted Figure 1 from [10].

For calculating the TAI, we assign an evolutionary age to each gene in the *A.*

thaliana genome by sorting each gene into its phylostratum, defined as the most distant phylogenetic node containing at least one species with a detectable homologue (Fig. 2.2, Supplementary Tables 1 and 2 in [10]). The resulting phylostratigraphic map [8] contains 13 phylostrata, PS1–PS13 (Fig. 2.3A). PS1 includes the evolutionarily oldest genes with homologous sequences in prokaryotes, and PS13 includes the evolutionarily youngest genes with no homologue in any other species.

For calculating the TDI, we determine the sequence divergence between *A. thaliana* and its sister species *Arabidopsis lyrata* or any one of the closely related Brassicaceaes, *Brassica rapa*, *Capsella rubella* and *Thellungiella halophila*, by computing the K_a/K_s ratio (Supplementary Table 3 of [10]). Here K_a is the number of non-synonymous substitutions per non-synonymous site and K_s is the number of synonymous substitutions per synonymous site for each orthologous gene pair. The K_a/K_s ratio is an indicator of selective pressure within protein coding regions and, thus, reflects natural selection, one of the major forces driving molecular evolution.

Interestingly, evolutionary age and sequence divergence as quantified above show only weak correlations (Kendall's rank correlation coefficient ranging from 0.02 to 0.26; Fig. 2.3B-E), indicating that both measures of evolutionary distance can be regarded as complementary. In combination with transcript information, the TAI quantifies the mean evolutionary age of a transcriptome, where the evolutionary age (phylostratum) of each gene is weighted by its expression level [9]. Analogously, we define the TDI as the mean sequence divergence of a transcriptome, where the sequence divergence (K_a/K_s) of each gene is weighted by its expression level.

2.3.2 Comparison of TAI and TDI

As shown in Fig. 2.3B, the two parameters, gene age and sequence divergence, can be described as largely independent or complementary evolutionary measures. Consequently, the major differences between the TAI and the TDI can be summarized as follows:

1. While the TAI reflects long-term evolutionary changes covering 4 billion years since the origin of life, the TDI reflects short-term evolutionary changes covering only 5-16 million years since the divergence of *A. thaliana* and the other four Brassicaceaes.
2. While the TAI sorts genes into ranked age categories based on founder gene emergence, the TDI addresses selective constraints as detected by ratios of non-synonymous to synonymous substitutions.
3. While the TAI does not distinguish between orthologs and paralogs, the TDI only considers orthologous sequences.

4. While the TAI defines proteins as homologous based on sometimes very short partial sequences, the TDI defines homology based on full-length protein sequences.

Together, TAI and TDI can be seen as complementary transcriptome indices. While the TAI covers deep and long-term evolutionary changes, the TDI detects rather short-term evolutionary changes. With regard to the phylotypic stage both indices are valuable because in many groups of species embryo development occurs rather typical for a closely related group of species than for distantly related species within a phylum [29]. In such cases the phylotypic stage may preferentially reflect recent evolutionary history that may be better detected with the TDI in contrast to the deep evolutionary changes that are primarily addressed with the TAI.

2.3.3 Transcriptome indices uncover transcriptomic hourglass patterns

Figure 2.4 (and Supplementary Fig. 2 of [10]) shows the TAI and TDI profiles across the seven sampled embryo stages of *A. thaliana*. We find that transcriptomes of early plant embryonic stages such as zygote and quadrant are evolutionarily young (high TAI), transcriptomes of the mid-embryogenic phase ranging from the globular to the torpedo stage are older (low TAI), and transcriptomes of later stages of embryogenesis are younger again (Fig. 2.4A). Qualitatively, this TAI profile strikingly resembles the molecular hourglass pattern discovered for *Da. rerio* and *Drosophila* [9].

Likewise, we find that transcriptomes of early stages are divergent (high TDI), transcriptomes of the mid-embryogenic phase are more conserved (low TDI), and transcriptomes of later stages of embryogenesis are more divergent again (Fig. 2.4B). Remarkably, the TDI profile qualitatively resembles the molecular hourglass pattern of the gene expression divergence profile discovered for *Drosophila* [16] and recently also *Caenorhabditis* [30].

Comparing both profiles, we make two observations. First, each of the profiles shows an hourglass pattern, where the TAI reflects long-term evolutionary changes covering 4 billion years since the origin of life, and the TDI reflects short-term evolutionary changes covering roughly 5–16 million years since the divergence of *A. thaliana* and the other four Brassicaceaes [31–34] (Supplementary Note [10]). Second, both profiles point to the torpedo stage as the predicted phylotypic stage, representing simultaneously the stage with the oldest as well as the most conserved/least divergent transcriptome.

An independent, but comparable transcriptome dataset [25, 26] from *A. thaliana* (Supplementary Fig. 3 [10]), which likewise covers embryogenesis from early phases

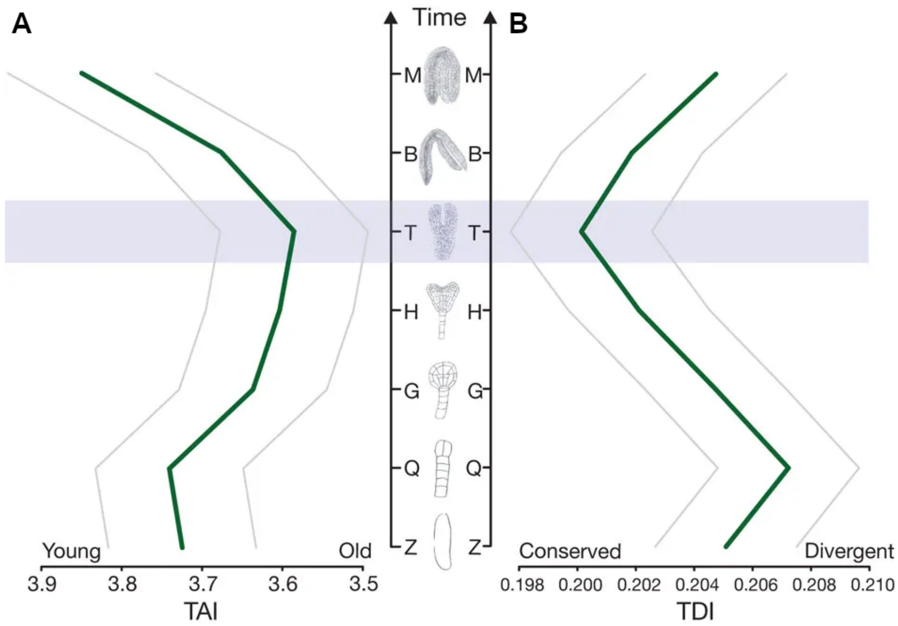


Figure 2.4 | Transcriptome indices across *A. thaliana* embryogenesis. (A), The transcriptome age index (TAI) profile. (B), The transcriptome divergence index (TDI) profile. Embryo stages: Z, zygote; Q, quadrant; G, globular; H, heart; T, torpedo; B, bent cotyledon; M, mature. Representative drawings (not on the same scale) are given for each sampled embryo stage. The blue shaded area marks the predicted phylotypic stage. The grey lines represent the standard error estimated by bootstrap analysis. The overall patterns of the TAI and TDI profiles are highly significant, as measured by permutation tests ($P_{\mathrm{TAI}} = 6 \times 10^{-13}$; $P_{\mathrm{TDI}} = 2 \times 10^{-05}$). Reprinted Figure 2 from [10].

to the mature embryo, confirms the hourglass pattern for both indices (Supplementary Fig. 4 [10]). Together, these observations suggest the possibility of convergent evolution of a molecular embryonic hourglass in animals and *A. thaliana*, and make it tempting to conjecture its universal presence across animal and plant kingdoms.

Given that developmental processes during plant and animal embryogenesis can be very different from the zygote stage on [35], and that the embryonic hourglass must have evolved independently in plants and animals, we wish to understand how the torpedo stage as the *bona fide* phylotypic stage in *A. thaliana* relates to animal phylotypic stages.

Across different animal taxa, the phylotypic stage was defined as the stage at which all major body parts are represented at their final positions as undifferentiated cell condensations [36]. In relation to this ontogenic progression in animals, the mid-embryogenic transition from the globular to the heart stage may conceptually serve as the corresponding stage in flowering plants. Here, polar axes are established and shoot and root apical meristems are initiated [37]. Hence, the ensuing torpedo stage at the transition from mid- to late-embryogenesis marks an ontogenic progression that seems more advanced than the phylotypic stage known from animals. Considering that morphological diversity and many important organs in flowering plants develop post-embryogenically, it is possible that the phylotypic stage maybe shifted towards the transition from mid- to late-embryogenesis compared to animals.

Furthermore, the torpedo stage roughly marks the transition from morphogenesis to the maturation phase. Morphogenesis involves the establishment of the embryo's body plan, whereas maturation involves cell expansion and accumulation of storage macromolecules to prepare for desiccation, germination and early seedling growth [38]. All land plants/embryophytes (all species from PS4 on) including lower land plants pass through a morphogenesis phase, but only the embryogenesis of higher land plants concludes with a maturation phase. Completely different signalling cascades are involved in both phases. One set is switched off and the other one is initiated. Because torpedo stage embryos are in the transition between these different developmental programs, it is conceivable that transcriptional programs are likewise reduced to conserved and evolutionary ancient processes that are reflected by the neck of the hourglass (Fig. 2.4).

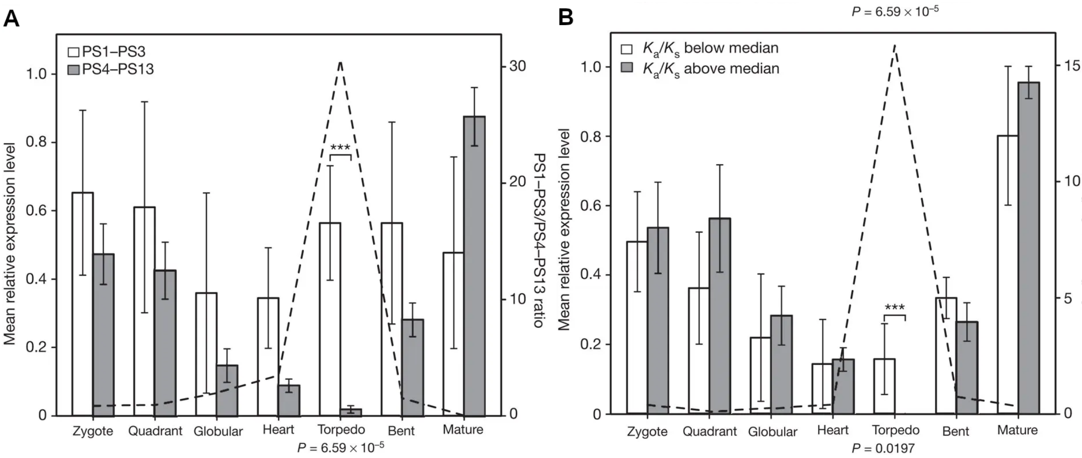


Figure 2.5 | Relative expression levels over embryo stages. (A), Left axis, mean relative expression levels of genes in PS1–PS3 (open bars) and PS4–PS13 (shaded bars); relative expression levels range from 0 to 1. Right axis, ratio of mean relative expression levels between PS1–PS3 and PS4–PS13, data points connected by dashed line. (B), Analogously to A, genes were divided along the median of the K_a/K_s ratios over all genes. Open and shaded bars show K_a/K_s values respectively below and above the median; data points connected by dashed line show the ratio of low to high K_a/K_s values. Error bars, standard error. Asterisks denote significant differences between PS1–PS3 and PS4–PS13 values (A) and conserved (below median) versus divergent (above median) genes at the torpedo stage (B); $^{*}P = 0.05$; $^{***}P = 0.0005$. Reprinted Figure 3 from [10].

Encouraged by these findings, we seek to understand how the molecular hourglass pattern of the TAI profile is determined. Two simple scenarios that would result in a decrease of TAI values include up-regulation of old genes, or down-regulation of young genes during mid-embryogenesis. To distinguish between both scenarios, we compute the relative expression levels of genes from phylostrata containing pre-embryogenesis species (PS1–PS3) versus post-embryogenesis phylostrata (land plants/embryophytes from PS4–PS13, representing plant species that pass through embryogenesis). Whereas expression levels of old genes vary only marginally across embryo stages, young genes are down-regulated towards the torpedo stage, and the ratio of the relative expression levels of old and young genes is maximized in the torpedo stage (Fig. 2.5A, Supplementary Fig. 5 of [10]).

Next, we divide the genes along the median of the K_a/K_s ratios over all genes and perform an analogous analysis for conserved (below median) versus divergent (above median) genes. Interestingly, we find a similar pattern, with divergent genes being more down-regulated towards the torpedo stage than conserved genes (Fig. 2.5B; Supplementary Figs 6–9 [10]). These results are confirmed by the independent dataset [25, 26] (Supplementary Fig. 10–14 of [10]).

Hence, the embryonic hourglass in *A. thaliana* seems to be coordinated by the quantitative down-regulation of young and divergent genes or, qualitatively, by the expression of fewer young and divergent genes towards the torpedo stage. This is in notable agreement with observations from the animal kingdom [9, 16, 39, 40]. As only a fraction of these down-regulated young genes in *A. thaliana* display an hourglass shaped expression profile across the sampled embryo stages themselves, the hourglass pattern is most probably caused by different sets of young genes. One set is involved in morphogenesis (up-regulated before the torpedo stage and down-regulated thereafter) and one set is involved in maturation (down-regulated before the torpedo stage and up-regulated thereafter; Supplementary Figs 15 and 16 of [10]).

In addition, we find that significantly enriched gene ontology terms among young plant genes down-regulated in the torpedo stage compared to early- or late embryogenesis describe signalling processes, such as responses to endogenous stimuli and hormones (Supplementary Tables 5 and 6 of [10]). This indicates that signalling processes controlling transcription of relatively recently evolved genes are down-regulated during the predicted phylotypic stage of *A. thaliana* embryogenesis.

2.4 Conclusions and Outlook

Using a phylotranscriptomic approach based on two complementary measures of evolutionary distance and two independent datasets, we have observed a molecular embryonic hourglass in plants, which seems to be predominantly caused by down-regulation of young and divergent genes towards the torpedo stage (Fig. 2.6). This observation is surprising for two reasons.

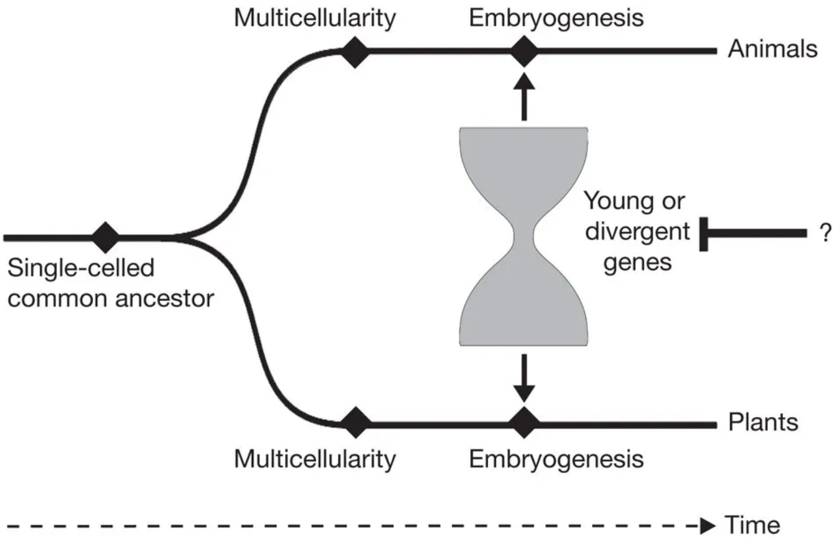


Figure 2.6 | Convergent evolution of a molecular hourglass in animal and plant embryogenesis. Originating from a single-celled common ancestor, animal and plant lineages evolved both multicellularity and embryogenesis independently. For the coordinated progression of the organisms through embryogenesis, the transcriptomes have to follow an hourglass pattern with maximally ancient and conserved transcriptomes during the phylotypic stage. Reprinted Figure 4 from [10].

First, morphological diversity during embryogenesis of flowering plants is negligible, so the increase of both transcriptome indices in late embryogenesis precedes the morphological differences established only during post-embryonic development.

Second, convergent evolution of a molecular hourglass pattern in animals and plants suggests operation of a fundamental developmental profile controlling the expression of evolutionarily young or rapidly evolving genes across kingdoms. We speculate that such a mechanism may be required for enabling spatio-temporal organization and differentiation of complex multicellular life.

If the developmental hourglass pattern is of such importance to multicellularity and thus evolutionary ancient, one question that emerges is whether the transcriptomic hourglass pattern in animals and plants is still functional and actively maintained or whether it is a nonfunctional remnant of a process that was once functional. This question cannot be answered directly. In chapter 3, we will turn to this question.

3

Evidence for active maintenance of the phylotranscriptomic hourglass

In chapter 2, we have introduced the developmental hourglass pattern of embryogenesis. We have found evidence for a transcriptomic hourglass pattern in animals and plants by combining transcriptomic and evolutionary information. However, its biological function could not be identified. Hence, it remains an open question whether the observed pattern is still functional or represents a non-functional evolutionary remnant.

In this chapter, we will address this question by calculating TAI and TDI profiles across the embryogenesis of *Da. rerio*, *D. melanogaster*, and *A. thaliana*. In section 3.1, we will explain the reasons for studying the active maintenance of the transcriptomic hourglass pattern. In section 3.2, we will present an updated phylostratigraphic and divergence stratigraphic approach, and we will get to know advanced statistical tests to assess transcriptomic patterns. In section 3.3, we will show the different transcriptomic patterns and we will determine the dependence between phylostrata and divergence strata. In section 3.4, we will attempt to determine if the hourglass pattern is actively maintained and if its existence may be associated with embryogenesis in extant species. In section 3.5, we will conclude that there is evidence for an actively maintained transcriptomic hourglass pattern and thus there is the potential to uncover this pattern's functionality in the long term.

The following sections are extracted from Drost et al. 2015 "*Evidence for Active Maintenance of Phylotranscriptomic Hourglass Patterns in Animal and Plant Embryogenesis*" [41].

3.1 Introduction

Irrespective of the phylotranscriptomic evidence recently obtained, the developmental hourglass model is controversially discussed to this day. Its biological function is rather poorly understood and hardly goes beyond hypotheses [15, 42].

Although convergent evolution within the animal lineage cannot be excluded, the existence of phylotranscriptomic and morphological hourglass patterns in numerous animal phyla suggests that it might have evolved early in the animal lineage. The developmental hourglass pattern could, therefore, be regarded as evolutionarily ancient. However, it is unclear whether this pattern is being actively maintained and still functional in extant species, or whether it represents a nonfunctional rudiment of a process that was once functional but has since then degenerated.

To be able to one day decipher the function of developmental hourglass patterns, we need to investigate this phenomenon in an experimental manner. Naturally, experiments are restricted to extant species. If actively maintained, such experiments could potentially reveal the molecular function of the developmental hourglass pattern. If, however, the developmental hourglass pattern were an evolutionary relic not functional in extant species, experimental approaches would be largely obsolete. The objective of this study is to investigate whether or not the developmental hourglass pattern is actively maintained in extant species and thus potentially allows to investigate its molecular function by experimental approaches.

To address this question, we study gene ages and TAI profiles as well as sequence divergences and TDI profiles of the vertebrate *D. rerio*, the invertebrate *D. melanogaster*, and the flowering plant *A. thaliana*. TAI profiles are based on both evolutionarily ancient and recent signals all along the tree of life. Hence, the TAI does not convey information about a possible active maintenance of the hourglass pattern. TDI profiles, however, with their distinctive feature of capturing only recent evolutionary signals are potentially able to address this question. To avoid subjective evaluation of the resulting profiles, we introduce three permutation tests, the flat line test, the reductive hourglass test, and the reductive early conservation test, to quantify the statistical significance of the corresponding phylotranscriptomic patterns. In addition, our study will provide support for either the hourglass model or possibly also other models that are currently being discussed.

3.2 Materials and Methods

In the following subsections, we will present the novel phylotranscriptomic approaches and the updated data resources for studying the active maintenance of phylotranscriptomic hourglass patterns. In subsection 3.2.1, we will learn about the sequence database for phylostratigraphy and in subsection 3.2.2, we will get to know the novel method to assign protein-coding genes into divergence strata (DS) for later calculations of the transcriptome divergence index. In subsection 3.2.3, we will find out about the gene expression data sets. In subsection 3.2.4, we will introduce the reductive hourglass test to quantify phylotranscriptomic hourglass patterns, and in subsection3.2.5, we will introduce the reductive early conservation test to quantify phylotranscriptomic early conservation patterns.

3.2.1 Phylostratigraphy - Construction of phylostratigraphic maps

The procedure for constructing phylostratigraphic maps follows the method section 2.2.1 of the previous chapter. Instead of only creating a phylostratigraphic map for *A. thaliana*, we computed maps of *Da. rerio* and *D. melanogaster*.

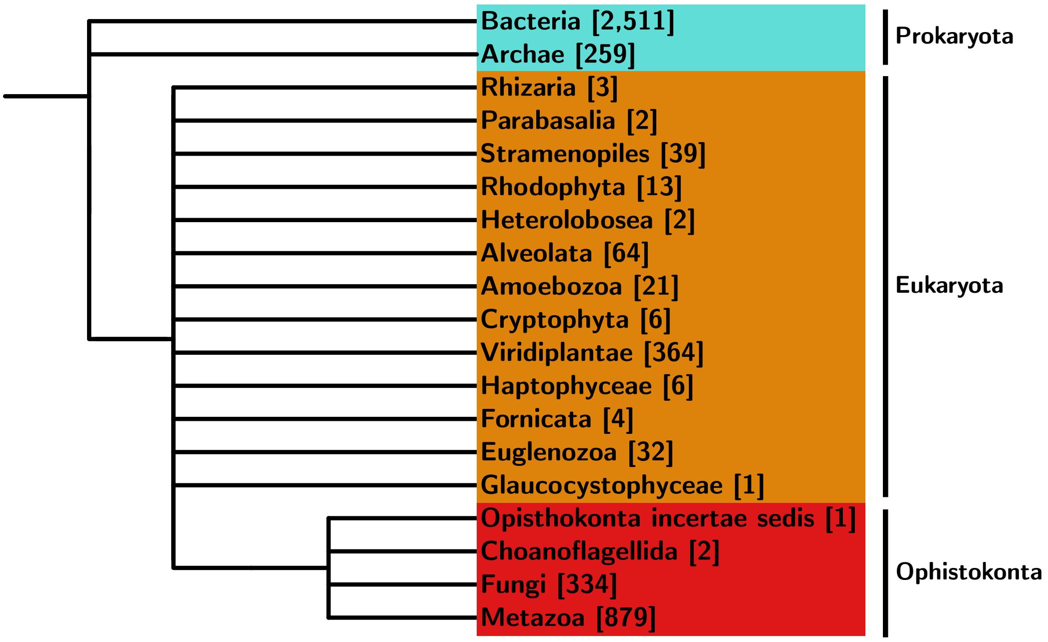


Figure 3.1 | NCBI taxonomy tree representing the major groups of species/genomes used for BLAST database. The clade of Prokaryoto (light blue) represents the biggest group and represents in all three species the oldest PS. The clade of Eukaryota (orange) represents the common taxonomic node of plants (Viridiplantae) and animals (Metazoa). The red clade of Ophistokonta is third major group and contains the clade of Metazoa which is specific for the species *Da. rerio* and *D. melanogaster* in this study. Reprinted Supplementary Figure S1 from [41].

To cover all three species' phylogenies, we generated a sequence database containing 17,582,624 amino acid sequences of 4,557 species from the NCBI, ENSEMBL [43], Flybase [44], and Phytozome [45] databases. We combine the sequences and create one BLASTable target database which is publicly accessible for reproducibility and subsequent phylostratigraphic studies (http://msbi.ipb-halle.de/download/phyloBlastDB_Drost_Gabel_Grosse_Quint.fa.tbz, last accessed August 2, 2015). This database is several times larger than the databases used in previous studies (e.g., [10]) and contains genome information from 2,770 prokaryotes (2,511 bacteria and 259 archea) and 1,787 eukaryotes (883 animals, 364 plants, 344 fungi, and 193 other eukaryotes) (Fig. 3.1)

We performed for each amino acid sequence of *A. thaliana* (TAIR10; 35,386), *Da. rerio* (ENSEMBL release 54; 24,147) and *D. melanogaster* (Flybase release 5.53; 29,357) with a minimum length of 30 amino acids a similarity seatch against this target database using BLASTP (BLAST version 2.2.21). If BLASTP does not identify a hit with an E-value below 10^{-5}, we assign the gene to the youngest PS. Otherwise, we assign it to the oldest PS containing at least one species with at least one BLAST hit with an E-value below 10^{-5}. The reproducible pipeline to perform the phylostratigraphic analysis is available at `https://github.com/AlexGa/Phylostratigraphy` [46].

3.2.2 Divergence stratigraphy - Constructing sequence divergence maps

In chapter 2, we have introduced the TDI based on sequence divergence levels calculated on the global alignment of orthologous genes as a K_a/K_s ratio. In this section, we will introduce an updated version of the sequence divergence calculation previously presented, which shows more robust results due to an improved ortholog detection, and is more comparable to PS values. In section 2.2.3, the divergence levels are continuous values compared to the discrete age categories of PS. We adjust this difference by transforming the continuous ratios into discrete values, called divergence strata (DS). The subsequent divergence map is the assignment of each gene from a particular species to a DS category. We construct the divergence maps of *Da. rerio*, *D. melanogaster*, and *A. thaliana* by the following procedure.

First, we identify orthologous gene pairs of *Da. rerio* and *As. mexicanus* (NCBI annotation release 77; 23.698), *D. melanogaster* and *D. simulans* (Flybase Release 1.4; 15,415), and *A. thaliana* and *A. lyrata* (Phytozome v.9.0; 32,670) by choosing the best reciprocal hit using BLASTP (BLAST version 2.2.29). If the best reciprocal hit has an E-value below 10^{-5}, the gene pair is considered orthologous; otherwise, it is discarded. Second, we construct codon alignments of the orthologous gene pairs using PAL2NAL [23]. Third, we compute K_a/K_s values of the codon alignments using GESTIMATOR [24] and Comeron's substitution model, which combines Li's, Pamillo's, and Bianchi's method with Kimura's method for obtaining robust K_a/K_s estimates [47]. Fourth, we discard all genes with a K_a/K_s value greater than 2 and sort the remaining K_a/K_s values into discrete deciles, which we call DS. DS values for the genomes of all three species are provided in supplementary table S4 in [41].

The same procedure is applied to generate sequence divergence maps for all other pairwise species comparisons (supplementary table S4 [41]). The construction of sequence divergence maps is explained in detail in the advanced vignette of the myTAI R package [48]. It can be applied to any chosen species pair with available coding sequence genomes and can be computed using the orthologr package (`https://github.com/HajkD/orthologr`).

3.2.3 Processing of expression data

For *Da. rerio* we use the microarray expression data set by Domazet-Lošo and Tautz (2010) [9] covering 40 stages corresponding to embryo development. The 16,188 probes of this data set correspond to 12,892 genes according to ENSEMBL predictions [9], and we compute the expression level of each gene as arithmetic mean of the expression levels of the corresponding probes [49]. Intersecting these 12,892 genes with genes in the phylostratigraphic map and the sequence divergence map of *Da. rerio* and *As. mexicanus* yields 12,892 genes and 7,740 genes, respectively. Intersecting sequence divergence maps of *Da. rerio* and *T. rubripes*, *Da. rerio* and *X. maculatus*, and *Da. rerio* and *G. morhua* yields 6,807, 6,997, and 4,734 genes, respectively.

For *D. melanogaster* we use the RNA-seq expression data set by Graveley et al. (2011) [50] covering 12 stages corresponding to embryo development. Intersecting the 15,139 genes of this data set with genes in the phylostratigraphic and the sequence divergence maps of *D. melanogaster* and *D. simulans* yields 12,043 genes and 6,230 genes, respectively. Intersecting sequence divergence maps of *D. melanogaster* and *D. yakuba*, *D. melanogaster* and *D. persimilis*, and *D. melanogaster* and *D. virilis* yielded 6,961, 5,872, and 5,732 genes, respectively.

For *A. thaliana* we use the microarray expression data set by Xiang et al. (2011) [28], which we introduced in the previous chapter (Sec. 2.3). It covers seven stages of embryo development. Intersecting the 26,173 genes of this data set with genes in the phylostratigraphic and sequence divergence maps of *A. thaliana* and *A. lyrata* yields 25,260 genes and 18,240 genes, respectively. Intersecting sequence divergence maps of *A. thaliana* and *C. rubella*, *A. thaliana* and *B. rapa*, and *A. thaliana* and *Car. papaya* yields 17,765, 16,122, and 9,427 genes, respectively.

3.2.4 Reductive hourglass test

The reductive hourglass test statistic quantifies if the observed TAI profile follows an hourglass pattern like shape. First, we partition the set of developmental stages $\{1, \ldots, N\}$ into the three modules - early (S_{early}), mid (S_{mid}), and late (S_{late}) - based on prior biological knowledge. We define the modules as

$$\begin{aligned} S_{early} &= \{1, \ldots, e\} && (3.1) \\ S_{mid} &= \{e+1, \ldots, m\} && (3.2) \\ S_{late} &= \{m+1, \ldots, N\} && (3.3) \end{aligned}$$

, with e denoting the last stage of the early module, m denoting the last stage of the mid module, and N denoting the last stage of the process of interest. Second, we compute the mean TAI value for each of the three modules (Fig. 3.7A left)

$$T_t = \frac{1}{|S_t|} \sum_{s \in S_t} TAI_s \quad ,\forall t \in \{early, mid, late\}. \tag{3.4}$$

Third, we compute the differences between T_{early} and T_{mid} and the difference between T_{late} and T_{mid} as

$$\begin{aligned} D_1 &= T_{early} - T_{mid} && (3.5) \\ D_2 &= T_{late} - T_{mid}. && (3.6) \end{aligned}$$

Fourth, we compute the minimum D_{min} of D_1 and D_2 (Fig. 3.7A right), which defines the final test statistic of the reductive hourglass test

$$D_{min} = min(D_1, D_2). \tag{3.7}$$

Under the assumption of an hourglass shaped profile T_{early} and T_{late} show higher TAI (or TDI) values than T_{mid}. Hence, the two differences should be positive and thus should the minimum difference D_{min}, too.

To determine the statistical significance of D_{min} we perform the same permutation of the dataset as described in the flat line test (subsection 2.2.5). Instead of calculating the variance for each randomly generated profile, we calculate D_{min}. Following this procedure we generate 10,000 values of D_{min} which we use to approximate a Gaussian distribution by estimating the mean and variance of the 10,000 D_{min} values. The significance, the P value, of the observed hourglass test statistic D_{min} is calculated as the probability of exceeding D_{min} (Fig. 3.2). The reductive hourglass test can be applied to TDI profiles in exactly the same manner.

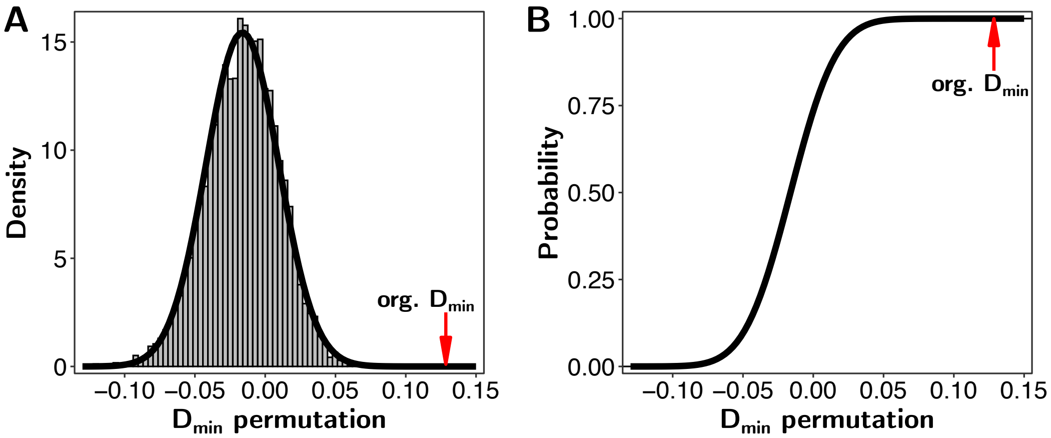


Figure 3.2 | Reductive hourglass test statistic. (A) The histogramm of 10,000 D_{min} values based on permuted PS assignment and recalculating the TAI profile of *A. thaliana* embryogenesis. The estimated Gaussian distribution fits the bell shape of the underlying histogram. (B) Based on the estimated Gaussian distribution, we determined the probability function to calculate the P value of the test statistic. The red arrow shows D_{min} of the original TAI profile within the estimated Gaussian distribution resp. probability function.

3.2.5 Reductive early conservation test

The reductive early conservation test is a permutation test conceptually identical to the reductive hourglass test. Specifically, steps one, two, and four are identical, and in step three we compute the two differences $D_1 = T_{mid} - T_{early}$ and $D_2 = T_{late} - T_{early}$. For a typical early conservation pattern, T_{early} should be low, and T_{mid} and T_{late} should be high, so both differences D_1 and D_2 should be positive, so the minimum difference D_{min} should be positive, too. In order to determine the statistical significance of an observed minimum difference D_{min}, we perform the same permutation test as in the reductive hourglass test, yielding the probability of exceeding the observed minimum difference D_{min} as P value of the reductive early conservation test. Instructions on the application of the flat line test, the reductive hourglass test, and the early conservation test are described in the introductory vignette of the myTAI R package [48].

3.3 Results

In this section, we will describe learn about the dependence of PS and DS and the existence of TAI and TDI hourglass patterns in the embryogenesis of animals and plants. In subsection 3.3.1, we will present the TAI profiles of *Da. rerio* , *D. melanogaster*, and *A. thaliana*. In subsection 3.3.2, we will compare the underlying evolutionary information of TAI and TDI by performing correlation analysis between PS and DS. In subsection 3.3.3, we will present the TDI profiles of *Da. rerio*, *D. melanogaster*, and *A. thaliana*. In subsection 3.3.4, we will learn about novel statistical methods to quantify if the TAI and TDI profiles show potential hourglass patterns.

3.3.1 TAI Profiles of *Da. rerio*, *D. melanogaster*, and *A. thaliana*

As described in subsection 3.2.1, we first set up a common database of 4,557 completely and partially sequenced genomes for the generation of updated phylostratigraphic maps of the three species of interest. This database is several times larger than the databases used in previous studies (e.g., [10]) and contains genome information from 2,770 prokaryotes (2,511 bacteria and 259 archea) and 1,787 eukaryotes (883 animals, 364 plants, 344 fungi, and 193 other eukaryotes) (Fig. 3.1 and Table S1 of [41]. Based on this database, we construct phylostratigraphic maps of *Da. rerio*, *D. melanogaster*, and *A. thaliana* using a customized pipeline. The three resulting phylostratigraphic maps are displayed in Figure 3.3A–C.

We next compute the TAI for each of the three species and each of the developmental stages. The resulting TAI profiles across embryogenesis for all three species are shown in Figure 3.4 (expression values provided in Supplementary Table S3 of [41]). If the mean evolutionary ages of the transcriptomes were the same at different developmental stages, the TAI profile would be a horizontal line.

To objectively test the statistical significance of the observed variations of the TAI at different developmental stages, we apply a permutation test that we refer to as the flat line test (see section 2.2.5 and [10]). When applying this flat line test to the three TAI profiles, we find that the TAI patterns of all three species deviate significantly from a horizontal line ($P < 0.05$).

Visually, the TAI profiles of *Da. rerio* and *A. thaliana* show an hourglass pattern. Although still within the standard deviation of the phylotypic period, the absolute minimum of the *D. melanogaster* TAI profile can be found at the 0–2 h time point in early embryogenesis (Fig. 3.4). This is unexpected and in contrast to comparative transcriptomic approaches, which consistently identified highly divergent transcriptomes in early *Drosophila* embryogenesis [16, 51]. However, we hesitate to over interpret this observation because the overall profile still resembles an hourglass pattern.

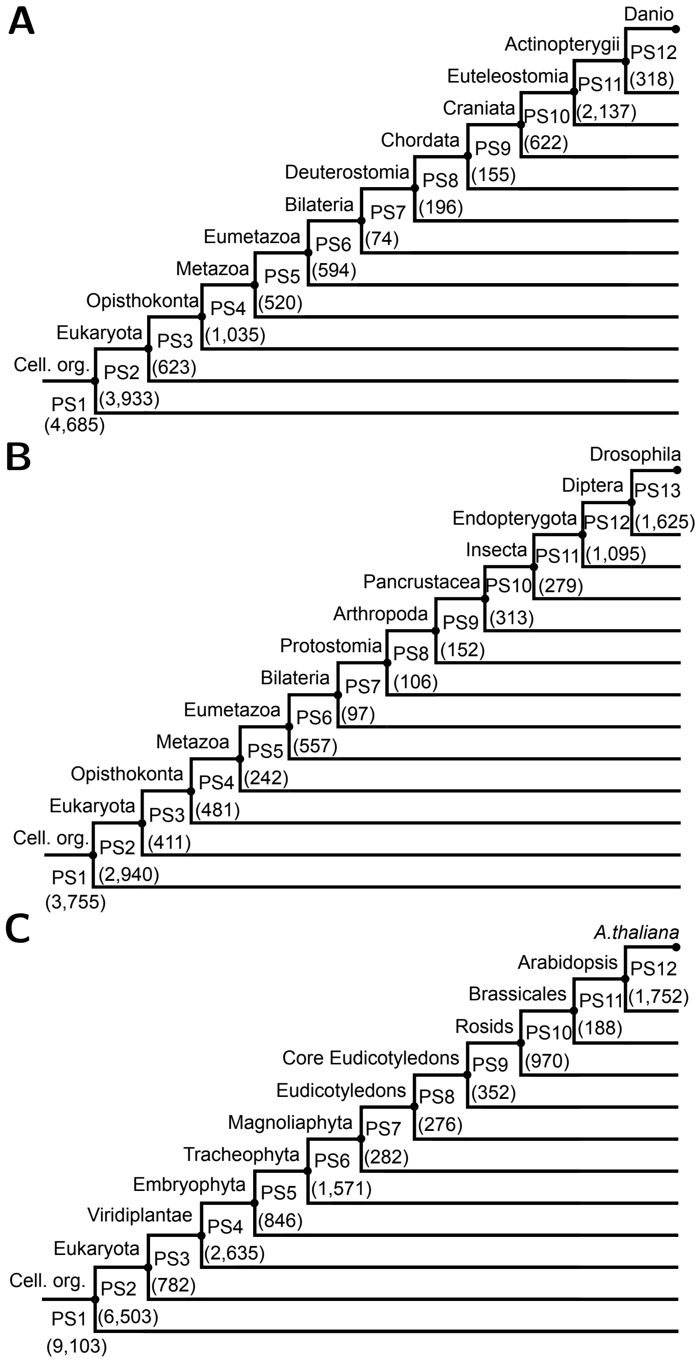


Figure 3.3 | Phylostratigraphic maps for *Danio rerio*, *Drosophila melanogaster*, and *Arabidopsis thaliana*. (A) *Danio rerio.*(B) *Drosophila melanogaster.* (C) *Arabidopsis thaliana.* Numbers in parenthesis denote the number of genes per phylostratum (PS1–PS12/13). Cell. org., cellular organisms described by PS1. Reprinted Figure 1 from [41].

Given that the TAI does not focus on recent evolution and that the majority of genes in all three species map to "old" PS (Fig. 3.3), these results indicate that the phylotranscriptomic hourglass pattern is not a recent innovation. Although TAI patterns alone do not allow this conclusion, the existence of phylotranscriptomic hourglass patterns across kingdoms and the existence of morphological hourglass patterns across animals suggest that these patterns emerged alongside with embryogenesis in early evolution. This suggestion is in accordance with previous findings showing that genes, transcriptomes, and molecular processes are most conserved during the phylotypic period [9, 10, 16, 30, 40, 49, 52–59].

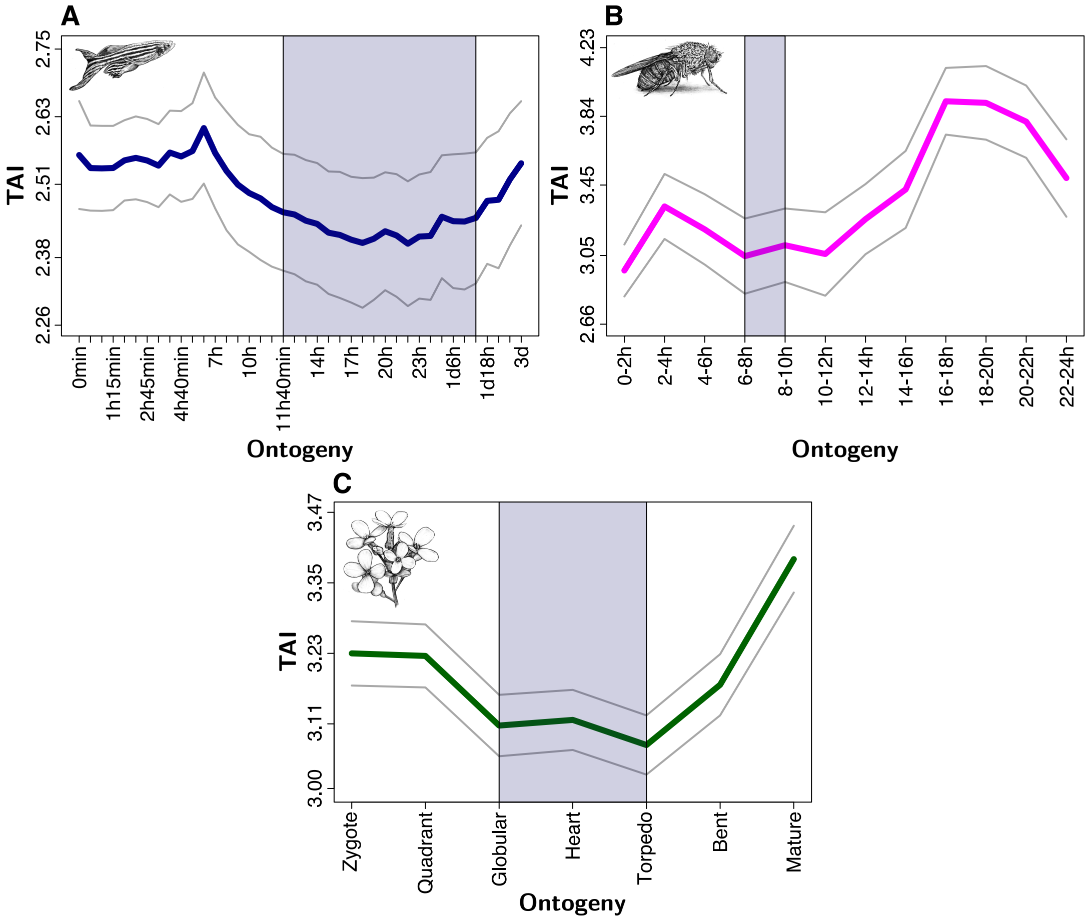


Figure 3.4 | TAI profiles across animal and plant embryogenesis. (A) *Danio rerio*.(B) *Drosophila melanogaster*. (C) *Arabidopsis thaliana*. The blue shaded area marks the predicted phylotypic period. The gray lines represent the standard deviation estimated by permutation analysis. Reprinted Figure 2 from [41].

3.3.2 Dependence of PS and DS

Before turning to the central question of whether or not the observed hourglass patterns might be actively maintained, we test in this section whether PS and DS are sufficiently independent of each other. This independence - or an only weak dependence - of PS and DS is important to assure that TAI and TDI profiles are not dependent on each other. Only in this case, the TDI can provide additional information and conclusions that cannot be drawn based on TAI profiles alone.

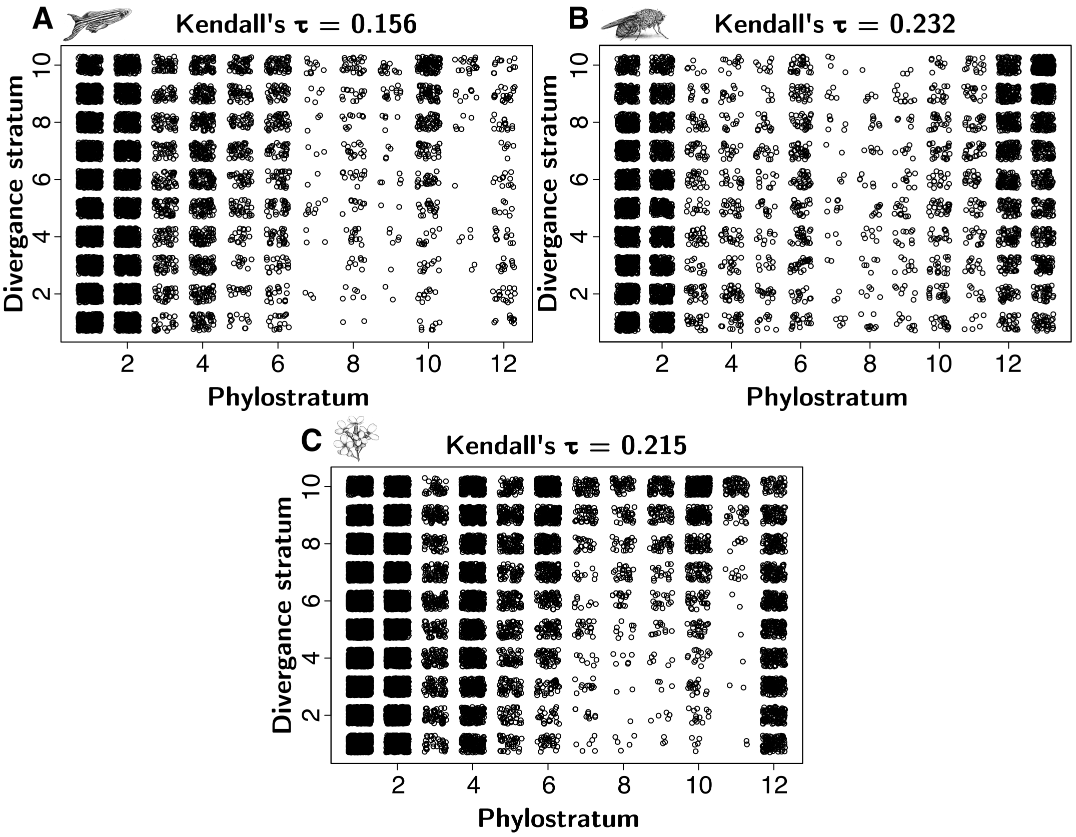


Figure 3.5 | Correlation between phylostratum (PS) and divergence stratum (DS). Scatter plots of phylostratum versus divergence stratum over all genes. (A) *Danio rerio.* (B) *Drosophila melanogaster.* (C) *Arabidopsis thaliana.* K_a/K_s ratios for divergence stratum assignment are derived from orthologous genes between *Da. rerio* and *Astyanax mexicanus* (A), *D. melanogaster* and *D. simulans* (B) and *A. thaliana* and *A. lyrata* (C). Kendall τ values denote the Kendall rank correlation coefficients quantifying the degree of linear dependence between PS and DS in a nonparametric manner. All Kendall τ values are significant (P < 2.2e-16) using Kendall's τ test of no correlation. Reprinted Figure 3 from [41].

For computing DS in analogy to PS, we generate orthologous gene sets for the computation of sequence divergences (K_a/K_s) by pairwise comparisons of the coding sequences of a target species to a related species with a completely sequenced and annotated genome. To lend more support to the TDI profiles to be generated, we compute the sequence divergence for three additional related species for each of the three target species (supplementary Figs. S2–S4 of [41]).

For *Da. rerio* closely related fish genomes are not yet available. Here, we use *Astyanax mexicanus* (divergence time ∼153 Ma [60]), *Takifugu rubripes*, *Xiphophorus maculatus*, and *Gadus morhua* (divergence time for all three species ∼265 Ma [60]). For the assignment of K_a/K_s values of *D. melanogaster* genes, we compare its coding genome to *D. simulans* (divergence time ∼3 Ma [60]), *D. yakuba* (divergence time ∼7 Ma [60]) *D. persimillis* (divergence time ∼34 Ma [60]), and *D. virilis* (divergence time ∼47 Ma [60]). For *A. thaliana* we use the Brassicas *A. lyrata* (divergence time ∼5–10 Ma [61]), *Capsella rubella* (divergence time ∼10–14 Ma [33]), *Brassica rapa* (divergence time ∼16 Ma [60]), and *Carica papaya* (divergence time ∼72 Ma [60]). For each pairwise comparison we sort the continuous K_a/K_s values into deciles and obtain a discrete DS for each gene and each of the four reference species with a detectable ortholog (provided in Supplementary Table S4 and Figs. S5–S7 of [41]).

To study to which degree gene age and sequence divergence are correlated for *Da. rerio*, *D. melanogaster*, and *A. thaliana*, we compute Kendall's rank correlation coefficient of PS and DS, which quantifies the degree of linear dependence between PS and DS per species in a nonparametric manner. In Figure 3.5 we display correlation plots of the three target species to their closest related species. We consistently find that correlations of PS and DS are significant but only weak (Kendall's rank correlation coefficient < 0.25; Fig. 3.5A–C; Supplementary Tables S2 and S4 and Figs. S5–S7 of [41]), stating that TAI and TDI have the potential of capturing independent evolutionary signals for all three species.

3.3.3 TDI Profiles of *Da. rerio*, *D. melanogaster*, and *A. thaliana*

Next, we investigate whether or not the evolutionary selection pressure that has shaped the hourglass pattern might still be active. To address this question, we compute the TDI profiles for all three species, which might potentially identify evidence for or against active maintenance, and thus functionality, of the hourglass pattern in extant species.

If the developmental hourglass pattern were not maintained and therefore under no selective pressure, the TDI profile would resemble a horizontal line. In contrast, if the developmental hourglass pattern were actively maintained in extant species, possibly because it still served an important biological function, the TDI profile should deviate from a horizontal line and take an hourglass-like shape.

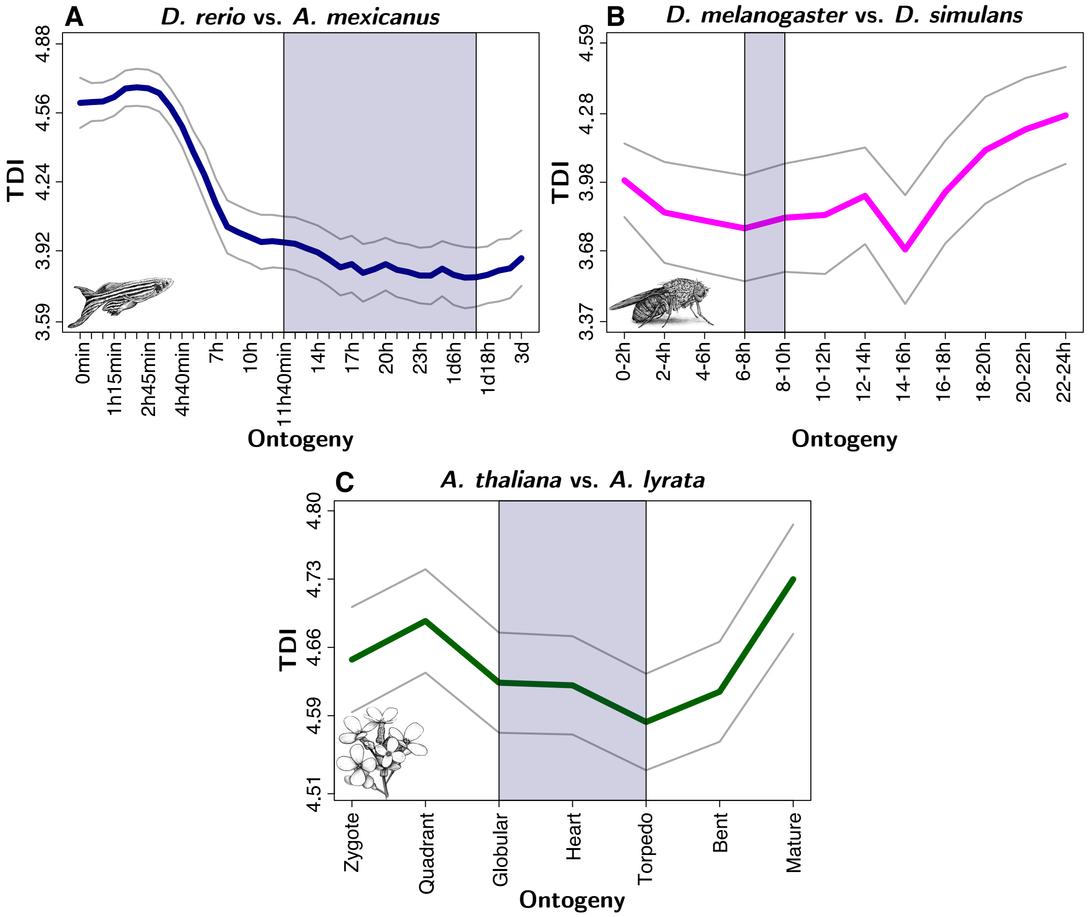


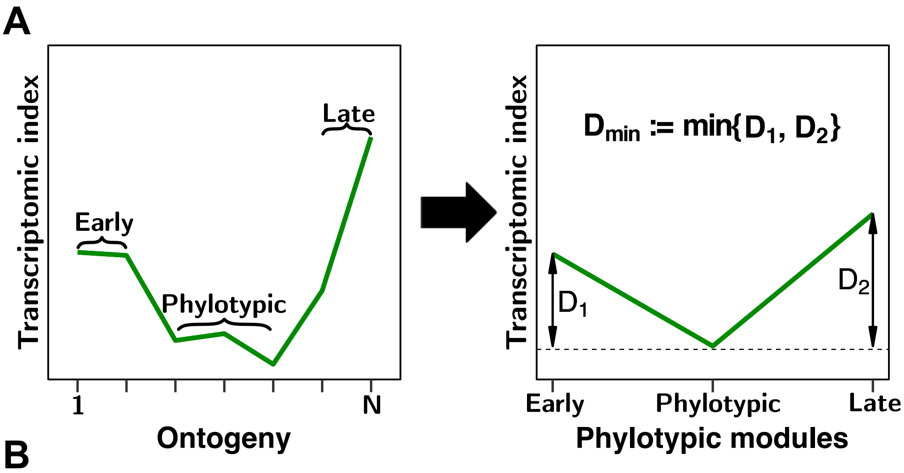
Figure 3.6 | TDI profiles across animal and plant embryogenesis. (A) *Danio rerio*. (B) *Drosophila melanogaster*. (C) *Arabidopsis thaliana*. The blue shaded area marks the predicted phylotypic period. The gray lines represent the standard deviation estimated by permutation analysis. Reprinted Figure 4 from [41].

Figure 3.6 shows the TDI profiles across embryogenesis for all three species based on DS values obtained from ortholog assignment to the closest related species. Applying the flat line test, we find that the TDI patterns of all three species deviate significantly from a horizontal line ($P < 0.05$), demonstrating that selective pressure is acting on embryonic transcriptomes across kingdoms. Visually, the TDI profiles of *D. melanogaster* and *A. thaliana* show an hourglass pattern, whereas the TDI profile of *Da. rerio* shows only the first two-thirds of an hourglass pattern with an increase of TDI values in late embryogenesis being barely noticeable. The TDI profiles for all other pairwise comparisons largely yield similar results (Supplementary Figs. S2–S4 and Table S5 of[41]).

These findings indicate that the phylotranscriptomic hourglass pattern is not a rudiment of a process that was once active but has progressively degraded since then. On the contrary, its evolutionary signal can still be detected even when evolutionary measures are consulted that account only for the last few million years.

3.3.4 Statistical testing for potential hourglass patterns

The studies presented above and all other studies published to date based on distance-based transcriptome comparisons or transcriptome indices have either relied on subjective visual profile interpretation [49, 57], have tested whether the observed profile deviated from a horizontal line [9, 10, 40, 59] (figs. 3.4 and 3.6), or have tested whether the observed profile could be fitted by a parabolic function [16, 30, 53].



B

TI	*Da.rerio*	*D.melanogaster*	*A.thaliana*
TAI	2.9e-02	3.8e-02	1.9e-08
TDI	2.9e-01	4.5e-02	3.2e-03

C

TI	*Da.rerio*	*D.melanogaster*	*A.thaliana*
TAI	1.0	7.1e-01	1.0
TDI	1.0	6.7e-01	9.1e-01

Figure 3.7 | Evaluation of transcriptome index profiles by the reductive hourglass test. (A) Schematic representation of module assignment and derivation of the test statistic. (B) P values derived by application of the reductive hourglass test to the TAI and TDI profiles in all three species. (C) P values derived by application of the reductive early conservation test to the TAI and TDI profiles in all three species. Reprinted Figure 5 from [41].

Naturally, subjective pattern evaluation should be avoided. In addition, the above described statistical approaches have severe limitations: 1) Testing whether the observed profile deviates from a horizontal line does not indicate the existence of an hourglass pattern, because the observed pattern could be anything different from a horizontal line that might even be in agreement with "competing" models such as the early conservation model and 2) testing whether the observed profile could be fitted by a parabolic function indicates the existence of an hourglass pattern, but the strict mathematical form of the pattern (parabola) makes this test highly specific and insensitive to other (nonparabolic) high–low–high patterns. Furthermore, none of these tests provides information about the significance of the localization of the most conserved stages, which is central to the evaluation of potential hourglass patterns.

Here, we propose a statistical test for a general high–low–high hourglass pattern not restricted to a parabolic function where the lowest phase must coincide with the presumptive phylotypic period. We divide embryogenesis in an early module, the phylotypic module, and a late module based on a priori morphological information about the known phylotypic period in animals (Fig. 3.7A). As, in contrast to animals, morphological evidence for a phylotypic period is still lacking in plants, it is impossible to define the phylotypic module for plant embryogenesis in analogy to animal systems. Hence, other biological processes that are likely associated with the phylotypic period had to be taken into account to legitimate a meaningful designation of the *A. thaliana* phylotypic module. Here, the mid-embryonic globular–heart–torpedo stages comprise embryonic morphogenesis and body plan establishment including the initiation and activation of the two apical stem cell niches, that give rise to the vast majority of organs throughout plant life. In addition, essential genes that cause embryo-defective phenotypes are likewise highly expressed during this period, indicating associated selective constraints (Supplementary Fig. S8 of [41]).

Based on these observations, we regard the developmental period encompassing globular, heart, and torpedo embryos as the most reasonable choice for designating the phylotypic period in *A. thaliana*. Next, we compute the differences between the mean values of the transcriptome indices of the early and the phylotypic module and of the late and the phylotypic module. The minimum of these two differences (early vs. phylotypic and late vs. phylotypic) serves as test statistic for a high–low–high pattern. Hence, this test recognizes patterns as hourglass patterns when the most ancient or most conserved transcriptomes occur in the phylotypic module (Fig. 3.7A). As this test reduces the ontogenetic stages to three developmental modules, we refer to this test as the reductive hourglass test.

Applying the reductive hourglass test to the TAI and TDI profiles of the three species reveals significant P values for both patterns of *D. melanogaster* and *A. thaliana* (Fig. 3.7B). For *Da. rerio*, only the TAI hourglass pattern is significant. For the TDI, the evolutionary signal in late embryogenesis seems to be diluted by the comparatively large evolutionary distance between *Da. rerio* and the other fish species (>150 Ma), and the increase of transcriptome divergence in *Da. rerio* development seems to be shifted from late embryogenesis to hatching and postembryonic development (Supplementary Fig. S9 of [41]).

Together, with exception of the *Da. rerio* TDI profile we find that both TAI and TDI values in early and late periods of embryogenesis are significantly higher than in the phylotypic periods in both animals and plants, demonstrating that phylotypic transcriptomes are evolutionarily ancient and highly conserved across kingdoms.

We finally adapt the reductive hourglass test to the early conservation model, call it reductive early conservation test, and apply it to the TAI and TDI profiles of all three species. We find that a low-high-high pattern is rejected in all six cases (Fig. 3.7C), stating that the described TAI and TDI profiles from three model species from two different kingdoms are inconsistent with the early conservation model, but largely consistent with the hourglass model.

3.4 Discussion

The controversy about the developmental hourglass model and especially about the hourglass versus early conservation models is as vibrant as it ever was. These and other models have traditionally relied on subjective anatomical comparisons, and a lack of measurable quantitative approaches has fed controversial discussions over decades [62–67]. However, technological progress recently facilitated quantitative measurements of expression profiles. Although some of these recent studies favored the early conservation model [66, 67], the majority of them supported the developmental hourglass model. Initially, a number of studies demonstrated hourglass-like patterns for limited sets of genes and a variety of genetic parameters [39, 53, 68–70]. Later, several studies demonstrated that whole transcriptomes of fly, worm, several vertebrates, and cress followed an hourglass pattern [9, 10, 16, 30, 40, 59]. For *Drosophila* ssp. it was recently shown that even the conservation of miRNA expression displays an hourglass pattern similar to that observed for protein-coding genes [71].

The later phylotranscriptomic studies have been performed by distance-based transcriptome comparisons [16, 30, 40, 59] or by studies of transcriptome indices [9, 10]; the latter combining evolutionary and transcriptomic information. As of now, there are two flavors of transcriptome indices. The TAI applies the phylogenetic age of a gene as an evolutionary measure [9] and thereby practically covers the complete evolutionary depth of the tree of life. The TDI, on the other hand, is based on sequence divergence of orthologous genes [10] and thereby captures exclusively recent evolutionary signals.

In our study, we systematically analyzed embryonic transcriptomes of two animal and one plant species. The resulting phylotranscriptomic patterns could have followed no profile at all or a variety of different profiles. Because the evaluation of phylotranscriptomic patterns in past studies (including our own) were subjective or relied on statistical tests with different limitations, we developed two more adequate statistical tests, the reductive hourglass test and the reductive early conservation test. These tests allow to objectively assess phylotranscriptomic profiles for the significance of a high–low–high pattern or a low-high-high pattern, respectively. In both cases, a prerequisite is a meaningful division of the set of developmental stages into three modules based on a priori biological knowledge.

Across the three species investigated, TAI analyses showed that early and late embryonic transcriptomes were consistently young (high TAI) and that the oldest transcriptomes were always observed during the presumptive mid-embryonic phylotypic period of each species (low TAI), which represents one of the hallmarks of the developmental hourglass model. For all three species we found that the reductive hourglass test and the reductive early conservation test supported the hourglass model and rejected the early conservation model, providing objective support for the developmental hourglass model.

Confidence in the validity of the developmental hourglass model allowed us posing the central question of this work of whether or not the phylotranscriptomic hourglass pattern might still be associated with a biological function in extant species. If so, the phylotranscriptomic hourglass pattern might either be causal for a downstream biological function or be the result of such a function. Alternatively, the phylotranscriptomic hourglass pattern might simply represent an evolutionary relic of a once important process that continues to exist in a rudimental status.

Only if this pattern were actively maintained, it would be possible to transform the currently predominantly descriptive approaches to a functional, that is, experimental, level. Hence, answering this question is important for understanding the still enigmatic function of the hourglass pattern in the long term and for deciding if it is in principle possible to uncover the molecular function of the phylotranscriptomic hourglass pattern by performing experiments on extant species.

Neither distance-based approaches nor studies of transcriptome indices can address the evolutionary time of emergence of the hourglass pattern in a satisfactory manner. Likewise, its active maintenance in extant species cannot be addressed by distance-based transcriptome comparisons or studies of TAI profiles. However, studies of TDI profiles that consult evolutionary signals from only recent evolution are arguably best suited for investigating the "active maintenance issue".

To date, TDI profiles of animal species had not yet been reported. As the closest related fish species with a completely sequenced genome diverged from *Da. rerio* greater than 150 Ma, this relatively long time span does certainly not qualify to make assumptions on very recent evolutionary trends. Hence, interpretation of these results is less meaningful than those of *D. melanogaster* and *A. thaliana*, whose closest relatives diverged only approximately 3 and 5–10 Ma, respectively. Here, statistical evaluations show a significant hourglass-like pattern with the minimum during the presumptive phylotypic period, consistent with the developmental hourglass model.

This result is supportive of Kalinka et al. (2010) [16], who suggested that the conservation of genes between closely related species that are active during mid-development is the result of natural selection acting to maintain expression levels and their temporal relationships to enable the correct establishment of the body plan. The results provided by [16] and the results from TDI computations reported here propose a scenario in which, across kingdoms, the phylotranscriptomic hourglass pattern is actively maintained through stabilizing selection.

Interestingly, while vertebrate and invertebrate embryogenesis also follows an hourglass pattern on the morphological level, morphological hourglass patterns are apparently absent from plant embryogenesis; at least they have never been reported. In contrast, comparative embryology in flowering plants, for example, suggests that the complete process of embryogenesis is morphologically highly conserved [72]. Mature plant embryos are anatomically much less complex than mature animal embryos. In a simplified manner, animals (such as mammals and many other vertebrates) initiate genesis of the vast majority of organs largely simultaneously in the phylotypic period during embryogenesis.

In contrast, during embryogenesis many plant species including *A. thaliana* establish only a limited set of major organs, consisting of hypocotyl, petioles, cotyledons, the embryonic root, and two stem cell niches (meristems). All other organs are initiated in these two apical meristems or in secondary meristems and are formed only during postembryonic development, where also morphological differences between species are being established. Possibly, plant embryogenesis is not complex enough to generate morphological differences between species, without which a morphological hourglass pattern is obsolete. Alternatively, any trace of a previously existing morphological pattern might have been wiped out and is undetectable by comparing extant species.

Although the TAI profile of *A. thaliana* suggests that the phylotranscriptomic hourglass did not emerge recently, its TDI profile suggests that some functional property of the phylotranscriptomic hourglass is actively maintained in extant plant species. In view of the lack of a morphological hourglass pattern in plants, one could conjecture that although the phylotranscriptomic hourglass pattern might be actively maintained in extant species across kingdoms, phylotranscriptomic and morphological hourglass patterns do not necessitate each other. They might even be uncoupled, which in turn would cast doubt on a possible causal relationship between them.

3.5 Conclusions and Outlook

The existence of hourglass patterns in TAI profiles of animal and plant embryogenesis demonstrates that this pattern is not a recent innovation. Darwin (1859) [73] said "it would be impossible to name one of the higher animals in which some part or other is not in a rudimentary condition." Although we admit that it might not be entirely accurate to directly compare a molecular pattern such as the phylotranscriptomic hourglass with morphological structures, the phylotranscriptomic hourglass pattern might in fact become a molecular addition to the long list of vestigial characters such as the leg bones of whales or the wings of ostriches and other flightless birds, for example.

However, the existence of hourglass patterns in TDI profiles of animal and plant embryogenesis suggests that this pattern is actively maintained in extant species. As evident for most evolutionary questions, experimental studies of processes that were functional in extinct species but have become nonfunctional in the course of evolution are incomparably more difficult to study than processes still functional in extant organisms. Provided that active maintenance of the phylotranscriptomic hourglass pattern would make little sense without it being functional, we hypothesize that this pattern is still functional in extant species and does not represent a nonfunctional relic. Despite this weak evidence for functionality of the phylotranscriptomic hourglass pattern, these data suggest that it might be possible to identify the molecular function(s) of this pattern in the long term. In any case, much remains to be learned, and we believe that a systematic comparative approach between plants and animals has the potential to significantly advance our understanding of the developmental hourglass phenomenon.

In order to accomplish this goal, we have to overcome the limitations of the presented bioinformatics approach. The TAI and TDI profiles provide information about a summarized transcriptomic profile by calculating the weighted arithmetic mean of the evolutionary age (TAI) or sequence divergence (TDI) at each stage in embryogenesis. To systematically uncover the molecular function of the transcriptomic hourglass patterns, we need independent measures to summarize the evolutionary age or sequence divergence distributions which are capable of reflecting the whole underlying distribution.

Alternatively, we can attempt to uncover the function of the phylotranscriptomic hourglass pattern by investigating other developmental or transitional processes that are similar to embryogenesis. In contrast to animals, plants development is not finished after embryogenesis. In plants post-embryonic development is characterized by organ development such as flower development, and by further developmental transitions such as germination and floral transition. Analyzing these processes could help to understand the functional relevance of the transcriptomic hourglass patterns, and we will turn to this investigation in chapter 4.

4

Post-embryonic hourglass patterns mark ontogenic transitions in plant development

In chapter 3, we have discovered that the phylotranscriptomic hourglass pattern may be still actively maintained and of functional relevance during embryogenesis of animals and plants. We have also learned that in animals the phylotypic stage is strongly connected to a window of maximum morphological conservation at the beginning of organogenesis. This morphological connection is not known for plants. In contrast to animals, organogenesis in plants occurs primarily postembryonically. Hence, in the following sections we will attempt to answer the question if in plants the phylotranscriptomic hourglass pattern is connected to organogenesis. To this end, we will investigate the processes of germination, floral transition, and flower development of *A. thaliana*.

In section 4.1, we will compare the embryogenesis of animals and plants and the importance of ontogenetic transitions in postembryonically development of plants. In section 4.2, we will introduce the gene expression data sets of *A. thaliana* covering germination, floral transition, and flower development. In section 4.3, we will present the phylotranscriptomic hourglass patterns during germination and floral transition, implicating a connection of the observed phylotranscriptomic patterns to developmental transitions. In section 4.4, we will conclude our findings and hypothesize that the transcriptomic hourglass pattern may be a feature of developmental processes allowing to switch between two subsequent functional programs.

The following sections are extracted from Drost et al. 2016 "*Post-embryonic Hourglass Patterns Mark Ontogenetic Transitions in Plant Development*" [74].

4.1 Introduction

The window of maximum morphological conservation in mid embryogenesis, i.e., the phylotypic stage [3] or phylotypic period [16, 75], coincides with the onset of organogenesis during body plan establishment. It has been suggested that a likely cause for this conservation is a web of complex interactions among developmental modules (e.g., organ primordia) during body plan establishment, which results in selective constraints that minimize morphological divergence [15](Fig. 4.1A). Although controversially debated for decades, in recent years the concept of the developmental hourglass has been largely confirmed at the transcriptomic level. Several studies showed that the degree of sequence conservation, the phylogenetic age of transcriptomes, or the similarity of gene expression profiles maximize during the phylotypic period [9, 16, 30, 39, 40, 53, 59, 70, 76–78], which is in agreement with a potentially causative association with body plan establishment.

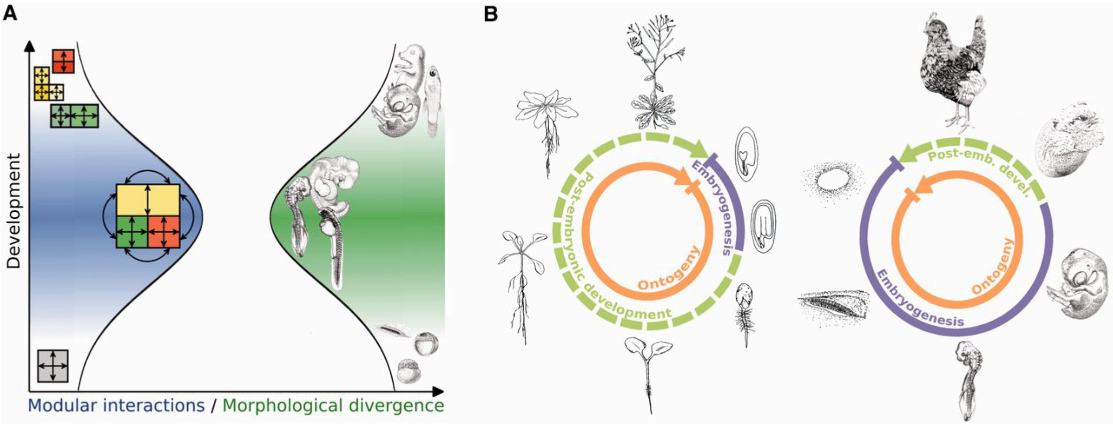

A
B
Development
Modular interactions / Morphological divergence
Post-embryonic development
Embryogenesis
Ontogeny
Post-emb. devel.
Embryogenesis
Ontogeny

Figure 4.1 | The developmental hourglass model in the context of differences in plant and animal development. (A) According to Raff (1996) [15], a web of complex interactions among developmental modules results in selective constraints during midembryogenesis. In the phylotypic period modular interactions maximize and morphological divergence minimizes resulting in the bottleneck of the developmental hourglass model (illustration adapted from Irie and Kuratani 2011 [40]). (B) The part of the ontogenetic life cycle that is covered by embryogenesis varies dramatically between plants and animals. Mature plant embryos have a limited number of organs and little complexity. Most organs develop postembryonically. In contrast to animals, the plant body plan is not fixed. It constantly changes in response to the environment. Animal development is largely embryonic. Mature animal embryos often reach a level of complexity that is comparable with adult individuals. Reprinted Figure 1 from [74].

In contrast to animals with their almost exclusively embryogenic development, organ formation in plants occurs largely postembryonically (Fig. 4.1B). Hence, a web of comparably complex modular interactions between developing organ primordia, which might underly the selective constraints during the phylotypic period in animals, is possibly never achieved during plant embryogenesis. However, a transcriptomic hourglass pattern has nonetheless been observed for plant embryogenesis [10, 41] (as well as for fungal development; [79]), indicating that it may not be causally connected to organogenesis, as suggested by the animal model.

We therefore wondered whether in plants these patterns might instead be associated with developmental transitions. Embryogenesis can be viewed as such a transition, namely from a single-celled zygote to a complex, multicellular embryo. To test this hypothesis, we generated transcriptomic data sets that cover the two most important ontogenetic transitions in postembryonic development in *Arabidopsis thaliana*: The transition from the embryonic to the vegetative phase, and the transition from the vegetative to the reproductive phase. As a control, we also analyzed a transcriptomic time series for flower development, a process that is dominated by organogenesis. We then performed phylotranscriptomic analyses [9, 10, 41], which assess the phylogenetic age of transcriptomes expressed over sequential developmental stages (Supplementary Fig. S1 of [74]), and tested the resulting profiles for the characteristic hourglass shape. If indeed, postembryonic developmental processes would be governed by hourglass patterns, this would suggest that hourglass patterns are not restricted to embryogenesis and possibly a wide-spread phenomenon that governs multiple processes. Furthermore, the potentially causative relationship among organogenesis, body plan establishment, and hourglass patterns would need to be re-evaluated.

4.2 Materials and Methods

This section builds on the methods of phylostratigraphy, divergence stratigraphy, TAI calculation, and the statistical testing presented in chapters 2 and 3. In subsection 4.2.1, we will describe the germination experiment. In subsection 4.2.2, we will explain the synchronization experiment and subsequent RNA-Seq library preparation to gain the floral transition expression data. In subsection 4.2.3, we will briefly introduce the flower development data set. In the subsection 4.2.4, we will explain the phylotranscriptomic analyses.

4.2.1 Germination experiment

Seeds of *A. thaliana*, accession Columbia (Col-0), were cold-stratified at 4°C in the dark for 72 h in Petri dishes on two layers of moistened blue filter paper (Anchor paper Co., U.S.A.). After stratification the seeds were incubated in a growth chamber at 22°C under constant white light. Seeds were collected at different developmental stages: mature dry seeds, six-hours imbibed seeds, seeds at testa rupture, radicle protrusion, appearance of the first root hairs, the onset of photosynthesis defined by appearance of greening cotyledons, and fully opened cotyledons.

Total RNA was extracted according to a modified hot borate method modified [80], as described previously [81]. RNA quality and concentration were assessed by agarose gel electrophoresis (0.1g mL^{-1}) and NanoDrop® measurements.

4.2.2 Floral transition experiment

To achieve synchronization of flowering times, we adapted a previously published cultivation regime [82]. In brief,*A. thaliana* Col-0 seeds were surface sterilized and stratified for 4 days at 4°C in water in the dark. They were then germinated for 7 days on vertical agar plates at 21°C under short day photoperiods (8 h light/16 h dark), before they were vernalized for 6 weeks at 4°C. Although floral transition in Col-0 does not require vernalization, this step significantly increased flowering time synchrony. Subsequently, seedlings were transferred to soil and grown for another 7 days at 21°C under short day photoperiods, before flowering was induced by shifting the plants to long day conditions (16 h light/8 h dark). For RNA-seq analysis we dissected shoot apices beginning 1 day after the shift to long day conditions. Subsequently, shoot apex material was sampled every day for another 8 days resulting in nine time points total. Sampling was performed every day at the same time 8 h after light on.

RNA extraction was performed with the RNeasy Plant Mini Kit (QIAGEN) including the on-column DNase digestion step according to the manufacturer's protocols. Integrity of the RNA was verified by agarose gel electrophoresis.

Library preparation and Illumina RNA-seq was performed by LGC Genomics. Reads were mapped onto the Arabidopsis genome (TAIR10) using TopHat 2 (v2.0.14) [83]. Uniquely mapped reads were counted using the featureCounts (v1.4.6) [84] with the annotation file from TAIR10. The normalized RPKM values were calculated by the function `rpkm()` from the Bioconductor package edgeR [85] using the effective gene length. Finally, the resulting expression set was matched with the phylostratigraphic map of *A. thaliana* and genes having RPKM values < 1 in at least one stage were removed from the dataset. This procedure yielded 16,899 expressed genes. Raw expression data can be downloaded from `http://www.ncbi.nlm.nih.gov/bioproject/311774` (PRJNA311774). Normalized expression data are included in Supplementary Dataset 1 of [74].

4.2.3 Expression data of flower development

Plant material, growth conditions, generation of expression data and data analysis are described in detail in [4]. Expression data can be downloaded from the NCBI GEO database (accession number GSE64581). Normalized expression data are included in Supplementary Dataset 1 of [74].

4.2.4 Phylotranscriptomic analyses

Phylotranscriptomic analyses are performed in the same manner as described in chapters 2 and 3. The phylostratigraphic map of *A. thaliana* is constructed as presented in subsection 3.2.1. The calculation of TAI profiles and relative expression values were performed as presented in subsections 2.2.2 and 2.2.6. The statistical significances of the resulting patterns from the developmental processes of germination, floral transition, and flower development were performed with the flat line test presented in subsection 2.2.5 and reductive hourglass test presented in subsection 3.2.4. All TAI, relative expression level, and statistical test computations are performed using the R package myTAI [48]. The scripts for reproducing the phylotranscriptomic analyses are available at `https://github.com/HajkD/post-embryo`.

4.3 Results and Discussion

In this section, we will investigate the phylotranscriptomic patterns of *A. thaliana* postembryonic developmental processes. We will find out which one of these processes shows an hourglass-shaped TAI profile, and we will discuss the findings in the light of functional relevance of the transcriptomic hourglass patterns. In subsection 4.3.1, we will investigate germination as the transition from embryogenesis to the vegetative phase. In subsection 4.3.2, we will analyse the transcriptomic pattern during floral transition. In subsection 4.3.3, we will find out about the transcriptomic pattern during flower development.

4.3.1 Germination - Transition to vegetative phase

To study the transition from embryogenesis to the vegetative phase, we generated transcriptomic information for seven sequential ontogenetic stages during seed germination [86]. The stages sampled included mature dry seeds, 6-h imbibed seeds, seeds at testa rupture, radicle protrusion, root hair (collet hair) appearance, the appearance of greening cotyledons, and established seedlings with fully opened cotyledons (Fig. 4.2A and Supplementary Fig. S2 of [74]). We then combined the transcriptomic information with previously generated gene age information [41]. Based on an age-assignment approach called phylostratigraphy [8] (Supplementary Fig. S1 of [74]), genes can be sorted into discrete age categories named phylostrata (PS) [8]). For *A. thaliana*, we defined 12 age classes ranging from old (PS1) to young (PS12). Next, we computed the transcriptome age index (TAI) [9] for each developmental stage, which is defined as the weighted mean of gene ages using the stage-specific expression levels as weights. The TAI therefore describes the phylogenetic age of a transcriptome.

As shown in Figure 4.2B, the TAI profile for the embryonic-to-vegetative phase transition displays an hourglass pattern with high TAI values at early and late stages and low TAI values at intermediate stages. We confirmed this observation through statistical tests (flat line test [41]: $P = 8.92 \times 10^{-20}$; reductive hourglass test [41]: $P = 3.08 \times 10^{-16}$; Supplementary Fig. S3a of [74]). The waist of the hourglass corresponded to the phylogenetically oldest transcriptomes stemming from the “testa rupture” to “radicle protrusion” stages. These stages mark the emergence of the seedling from the seed, likely the transition period of this process, at which germination becomes irreversible (Fig. 4.2B). We finally also studied the relative expression levels of genes of different PS and found that the hourglass pattern is caused by a largely antagonistic behavior of old and young genes (Fig. 4.2C), similar to what has been previously reported for embryogenesis [10, 41].

4.3.2 Floral transition - Vegetative-to-reproductive phase

We next tested whether a transcriptomic hourglass pattern also underlies the vegetative-to-reproductive phase transition. During this so-called floral transition, the leaf-producing shoot apical meristem is converted into an inflorescence meristem, which forms flowers [87]. Morphologically, completion of the floral transition can be observed by the bolting inflorescence. However, as the actual transition occurs several days before bolting, we also assessed the expression of floral homeotic genes and other marker genes to better map the time of transition to the reproductive state (Supplementary Fig. S4 of [74]). Based on this information, we synchronized flowering time in the sampling population (Supplementary Fig. S5 of [74]) and generated transcriptome data from the shoot apex before, during, and after floral transition.

Figure 4.3A shows the results from the TAI analysis for nine samples covering the floral transition. We identified a robust hourglass pattern (reductive hourglass test [41]: $P = 2.99 \times 10^{-5}$; Fig. 4.3A and Supplementary Fig. S3b [74]) that significantly deviated from a flat line (flat line test [41]: $P = 3.03 \times 10^{-14}$). Similar to embryogenesis [10, 41] and seed germination (Fig. 4.2C), analysis of relative expression levels of genes assigned to different age classes revealed a largely antagonistic behavior of old and young genes (Fig. 4.3B).

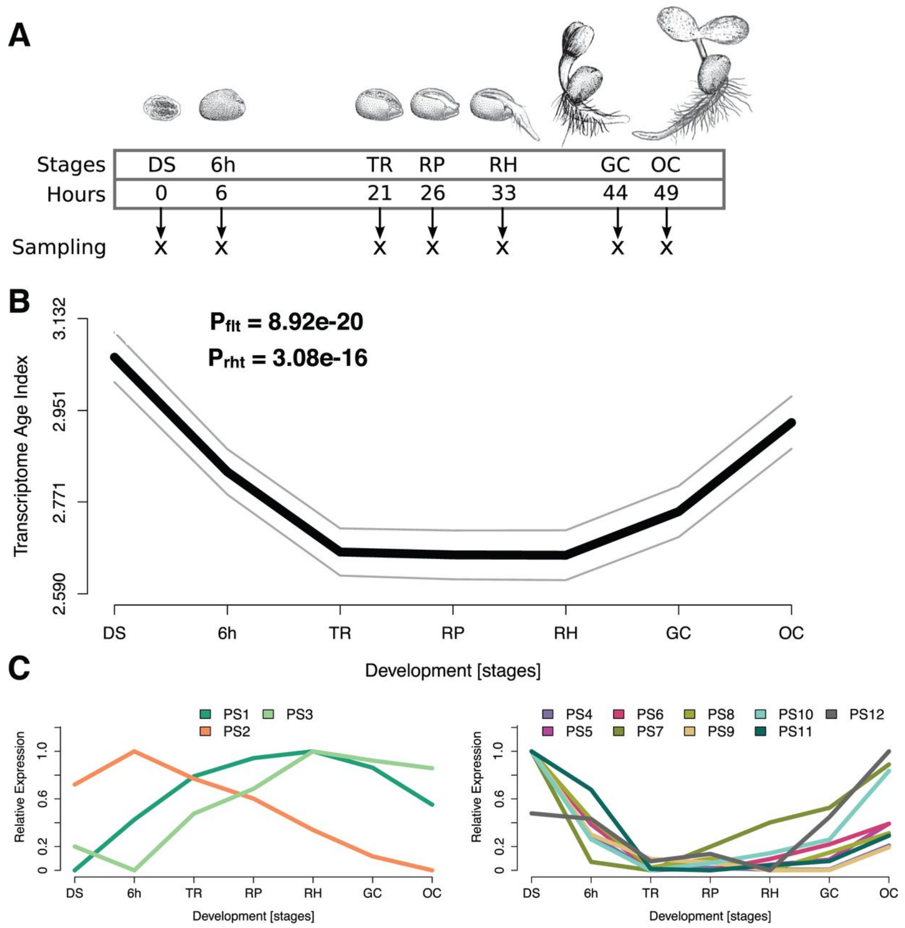


Figure 4.2 | TAI analysis for germination in *A. thaliana*. (A) Illustration of the developmental stages for which transcriptome data were generated. (B) The TAI profile across germination follows an hourglass-like pattern. The gray lines represent the standard deviation estimated by permutation analysis. P values were derived by application of the flat line test [41] (P_{flt}) and the reductive hourglass test [41] (P_{rht}). (C) Relative expression levels for each phylostratum (PS) separately. The stage with the highest mean expression levels of the genes within a PS was set to relative expression level = 1, the stage with the lowest mean expression levels of the genes within a PS was set to relative expression level = 0, the remaining stages were adjusted accordingly. PS was classified into two groups: Group “old” contains PS that categorize genes that originated before complex/multicellular plants evolved (PS1–3) and group “young” contains PS that categorize genes that originated after complex plants evolved (PS4–12). DS, mature dry seeds; 6h, 6-h imbibed seeds; TR, seeds at testa rupture; RP, radicle protrusion; RH, appearance of the first root hairs; GC, appearance of greening cotyledons; OC, fully opened cotyledons. Reprinted Figure 2 from [74].

Taken together, these observations demonstrate that in plants not only embryogenesis but also the embryo-to-vegetative and vegetative-to-reproductive phase transitions progress through a stage of evolutionary conservation with older transcriptomes being active in mid development. Thus the hourglass pattern, which

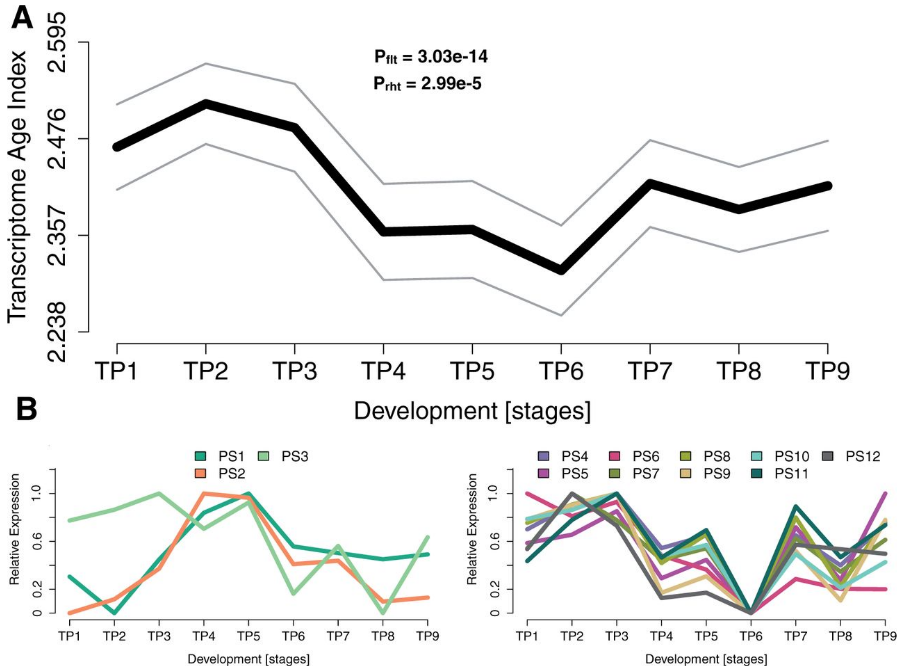


Figure 4.3 | TAI analysis for the transition from vegetative to reproductive growth in *Arabidopsis thaliana.* (A) The TAI profile across the transition to flowering follows an hourglass-like pattern. The gray lines represent the standard deviation estimated by permutation analysis. P values were derived by application of the flat line test [41] (P_{flt}) and reductive hourglass test [41] (P_{rht}). (B) Relative expression levels for each PS separately. The stage with the highest mean expression levels of the genes within a PS was set to relative expression level = 1, the stage with the lowest mean expression levels of the genes within a PS was set to relative expression level = 0, the remaining stages were adjusted accordingly. PS was classified into two groups: Group "old" contains PS that categorize genes that originated before complex/multicellular plants evolved (PS1–3) and group "young" contains PS that categorize genes that originated after complex plants evolved (PS4–12). TP, time point; TP1, 1 day after shift to long day photoperiods (LD); TP2, 2 days after shift to LD; TP3, 3 days after shift to LD; TP4, 4 days after shift to LD; TP5, 5 days after shift to LD; TP6, 6 days after shift to LD; TP7, 7 days after shift to LD; TP8, 8 days after shift to LD; TP9, 9 days after shift to LD. Reprinted Figure 3 from [74].

was previously discussed only with regard to embryogenesis, appears to be more widespread, at least in plants. In fact, the embryonic hourglass is possibly only one of many developmental processes governed by hourglass patterns.

Because no new organs are established during the two postembryonic phase transitions assessed here, our results also support the aforementioned conjecture that transcriptomic hourglass patterns are not specifically associated with organogenic processes.

4.3.3 Flower development - Formation of floral organs

To directly test this, we performed phylotranscriptomic analyses of a flower development data set we previously generated [4]. Flower development follows floral transition and is dominated by the formation of different types of floral organs. In agreement with the idea that hourglass patterns in plants are not tightly associated with organogenesis, the transcriptomic profile across 14 time points from the earliest stages of flower development to mature flowers did not show an hourglass pattern or, in fact, any other pattern at all (flat line test [41]: $P = 0.202$; Fig. 4.4A and B). Likewise, old and young genes did not show a clear antagonistic behavior in their expression (Fig. 4.4C). Together, these data suggest that in plants organogenesis is not the driving factor of hourglass-shaped transcriptome profiles. Hence, the currently favored explanation of animal hourglass patterns, which is based on selective constraints correlated to body plan establishment and organogenesis [15], cannot serve as a plausible explanation for the two postembryonic hourglass patterns reported here.

A simple scenario that might resolve this controversy would be that the transcriptomic hourglass patterns in plants are functionally unrelated to those of animal embryogenesis. They might in fact have evolved to serve a completely different, yet unknown, purpose. This scenario is supported by the lack of reports on morphological hourglass patterns for plant embryogenesis (in contrast to various animal phyla). It seems that morphological similarity among flowering plants is not restricted to a midembryonic period but rather exists throughout embryogenesis [72].

If the biological processes underlying embryonic hourglass patterns in animals and plants are indeed functionally unrelated, we would also have to revoke our earlier hypothesis that the developmental hourglass pattern evolved convergently in both kingdoms [10]. Interestingly, in the three processes we analyzed, it seems that the waist in the hourglass reflects a general transition to a growth or maturation phase. If, however, animal and plant hourglass patterns should serve a similar function, this study would suggest that the underlying cause is not organogenesis or body plan establishment but an even more fundamental process. As also in animal systems a causal relationship between body plan establishment and the phylotypic period remains to be proven [88], it might be worthwhile to directly address this relationship by designing experiments that separate developmental transitions from organogenesis in animals.

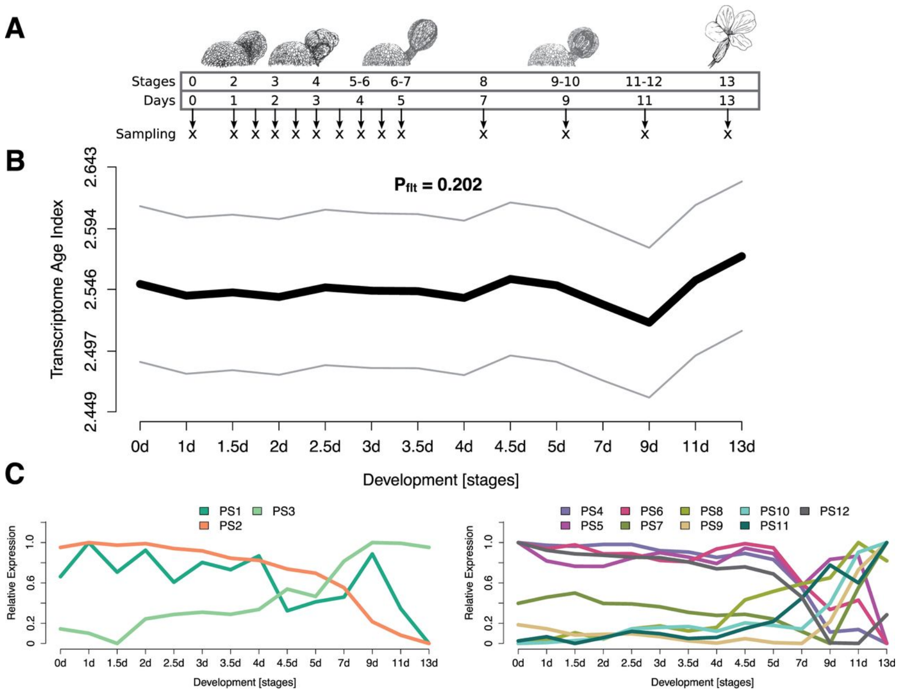


Figure 4.4 | TAI analysis of flower development in *Arabidopsis thaliana.* (A) Illustration of the developmental stages for which transcriptome data were generated; stages according to Ryan et al. 2015 [4].(B) The TAI profile across flower development fails to detect evolutionary signal. The gray lines represent the standard deviation estimated by permutation analysis. The P value was derived by application of the flat line test [41] (P_{flt}). (C) Relative expression levels for each PS separately. The stage with the highest mean expression levels of the genes within a PS was set to relative expression level = 1, the stage with the lowest mean expression levels of the genes within a PS was set to relative expression level = 0, the remaining stages were adjusted accordingly. PS was classified into two groups: Group "old" contains PS that categorize genes that originated before complex/multicellular plants evolved (PS1–3) and group "young" contains PS that categorize genes that originated after complex plants evolved (PS4–12). Reprinted Figure 4 from [74].

4.4 Conclusions and Outlook

The hourglass pattern was historically associated with animal embryogenesis and only recently recognized to govern plant embryogenesis, too. Here, we present evidence that in plants the hourglass pattern is probably even more fundamental and not only characteristic for embryo development. Specifically, it is present in all three major developmental transitions of plant life such as embryogenesis, germination, and floral transition. We could show that due to a missing transcriptomic hourglass pattern in flower development that plant hourglass patterns may not be related to organogenesis and thus body plan establishment, which is a widely assumed cause for animal hourglass patterns.

It will be interesting to test postembryogenic transitions like metamorphoses in animals to see whether this can also be observed for nonplant organisms. We hypothesize that a transcriptomic hourglass pattern is a feature of multiple developmental processes that simply require passing through an organizational checkpoint serving as a switch that separates two functional programs.

In chapter 6, we will turn to this question by developing novel transcriptomic measures based on the Shannon entropy to investigate the origin of the transcriptomic hourglass pattern and thus to attempt to deepen our understanding of its functional relevance.

5

Entropic hourglass patterns of animal and plant development

In chapter 4, we have suggested an explanation about the function of the observed transcriptomic hourglass patterns as a checkpoint for the transition from one developmental program to the next. In this chapter, we will present novel phylotranscriptomic approaches which are based on the Shannon entropy to investigate the whole age distribution of expressed genes during developmental processes like embryogenesis. The application of the developed entropic transcriptome age index could serve as an independent measure to investigate the function and also origin of the transcriptomic hourglass pattern. Additionally, we will present an updated version of the phylostratigraphic pipeline to determine the evolutionary age of genes combined with a web server which will provide publicly available phylostratigraphic maps.

In section 5.1, we will introduce our attempt to a novel entropic transcriptome measure and an updated phylostratigraphic pipeline. In section 5.2, we will redefine the TAI in a probabilistic manner and present the phylostratigraphic pipeline and the PhyloWeb server. In section 5.3, we will present novel hourglass patterns based on the Shannon entropy. Finally, in section 5.4, we will conclude by suggesting an explanation for the detected entropic hourglass patterns and provide an outlook to future methods based on our probabilistic definition of the TAI and its relevance for answering the question of the origin of the transcriptomic hourglass pattern.

5.1 Introduction

The quantification of the transcriptomic hourglass patterns showed that on average evolutionary young genes are expressed at the beginning of embryogenesis, evolutionarily old genes during mid-embryogenesis, and again evolutionarily young genes at the end of embryogenesis.

Focusing on plants, which represent the second major kingdom in the tree of life that evolved embryogenesis, it has been found that phylotranscriptomic hourglass patterns also exist in the two main transitions of post-embryonic plant development, germination and floral transition, suggesting the convergent evolution of phylotranscriptomic hourglass patterns in embryonic and post-embryonic plant development[74].

The origin of the phylotranscriptomic hourglass patterns has remained concealed, but here we find that not only the mean age of expressed genes changes in an hourglass-like manner, but the whole age distribution of expressed genes changes. To improve the calculation of gene ages, we update our phylostratigraphic analyses pipeline and developed a web server which will provide publicly available phylostratigraphic maps and an interface to perform user specific phylostratigraphic analyses.

To study the changes of age distributions, we redefine the traditional transcriptome index in a probabilistic manner and develop the entropic transcriptome age index (eTAI) based on the Shannon entropy [89]. When studying the Shannon entropy of these age distributions as functions of time, we find hourglass patterns that surprisingly show highly significant transcriptomic hourglass patterns. Measuring the whole age distribution and still providing highly significant hourglass patterns might indicate that the phylotranscriptomic hourglass patterns of the entropy [89] could be more fundamental than, and possibly even the mathematical origin of, the traditional transcriptomic hourglass patterns of animal and plant development.

5.2 Materials and Methods

In this section, we will introduce the entropic transcriptome age index and we will learn about the automated web tool for phylostratigraphic analyses, called PhyloWeb, which provides the phylostratigraphic maps in this chapter. In subsection 5.2.1, we will redefine the transcriptome age index in a probabilistic manner and introduce the entropic transcriptome age index in subsection 5.2.2. The calculation of transcriptome indices is based on the evolutionary gene age, provided by phylostratigraphic maps. In subsections 5.2.3 and 5.2.4, we will learn about the updated pipeline for phylostratigraphic analyses and the development of the PhyloWeb server.

5.2.1 A probabilistic perspective on the transcriptome age index

Alternatively to the previous chapters, the TAI can be defined in a probabilistic manner. To introduce this approach, we define a random variable X which is realized over all possible phylostrata $ps \in \{1, \ldots, PS_{max}\}$, whereas PS_{max} defines the youngest phylostratum and 1 the oldest. The probability of observing a PS in a specific developmental stage s can be calculated as

$$P_s(X = ps) = \frac{\sum_{i=1}^{I} e_{is} \delta_{ps,ps_i}}{\sum_{i=1}^{I} e_{is}}. \tag{5.1}$$

Here, the Kronecker delta function δ_{ps,ps_i} is equal to one, if the phylostratum of gene i is equal to its actual ps, otherwise it is zero. Following Eq. 5.1 we get a discrete probability distribution of PS for each stage s of the biological process of interest. Now we are also able to define the TAI, the so far weighted mean of gene ages, as the expectation value of PS in stage s

$$E_s(X) = \sum_{ps=1}^{\mathrm{PS}_{max}} ps \cdot P_s(X = ps) \tag{5.2}$$

which is equal to the traditional transcriptome age index TAI_s (Eq. 2.1). Hence, the TAI captures the mean changes of the age, resp. PS, distribution during a specific stage of the biological process.

In order to avoid naming conflicts, will now denote the traditional TAI as $mTAI$.

5.2.2 Entropic TAI

To study the changes of the age distribution as a function of time, we use the concept of *Shannon Entropy* [89] and introduce the entropic TAI ($eTAI$) for each stage s.

$$eTAI_s(X) = -\sum_{ps=1}^{PS_{max}} P_s(X = ps) \log_2 P_s(X = ps). \tag{5.3}$$

This new metric now allows us to better quantify distribution changes across stages in the biological process of interest. We are able to quantify homogeneities in the age distribution by measuring low $eTAI$ values and heterogeneities in the age distribution by measuring high $eTAI$ values.

5.2.3 Phylostratigraphic analysis pipeline

To automate the calculation of phylostratigraphic maps for each species of interest and to provide this information to a broader audience, we developed the PhyloWeb server application. The calculation of phylostratigraphic maps is computationally demanding. Currently, we are using as the comprehensive sequence database the NCBI non-redundant protein (nr) database [90], containing $\sim 140,000,000$ sequences. A typical eukaryote genome can contain up to $\sim 50,000$ genes and over $\sim 100,000$ proteins. Following these numbers, we would need $\sim 14,000,000,000,000$ comparisons if we had to calculate each sequence comparison.

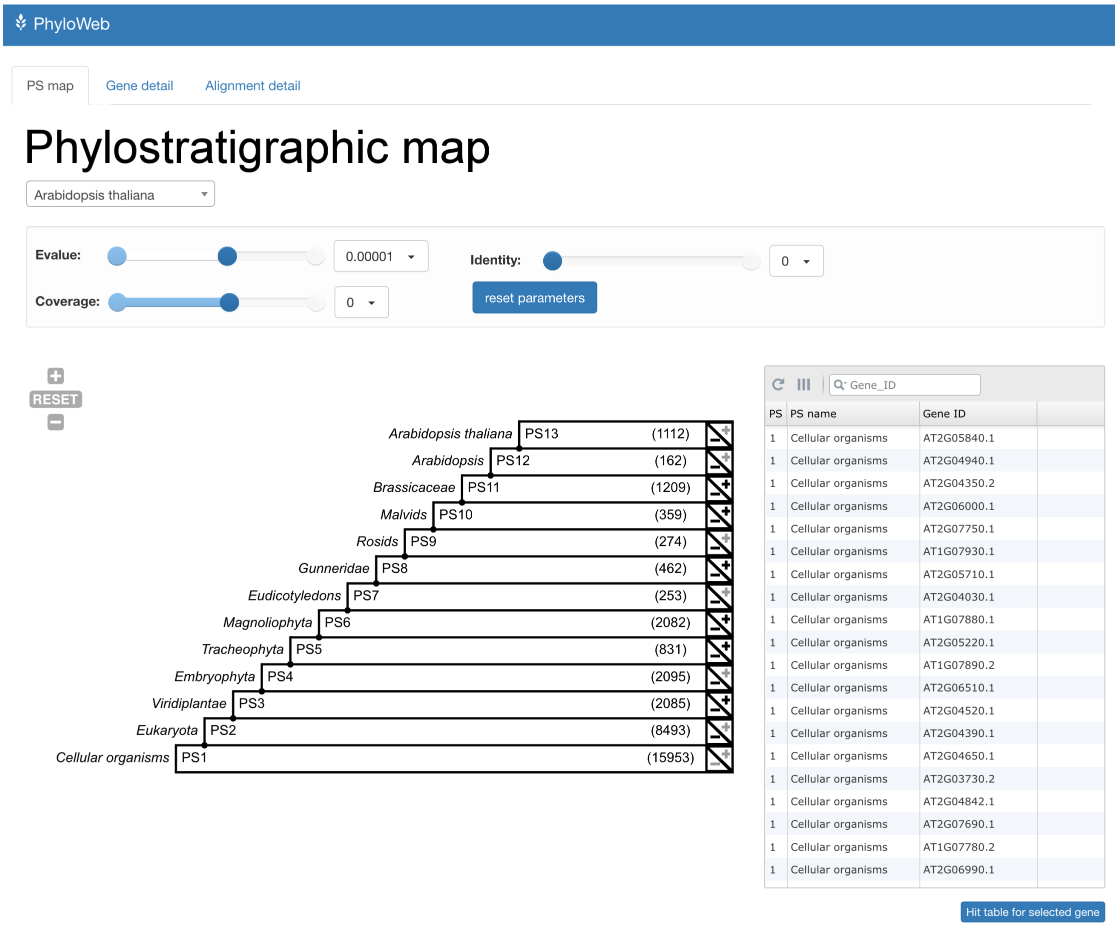
Due to the heuristic nature of the BLAST algorithm [91], this number can be dramatically decreased but the remaining computations to receive a list of all homologous sequences for all genes in a query species is still too high for computing on a personal computer. Hence, we efficiently distributed the similarity searches on the high-performance cluster of the Martin Luther University Halle-Wittenberg. Thus, each node of the cluster computes for a chunk of sequences the similarity searches against the nr database, dramatically reducing the time for similarity searches.

Afterward, the homologous sequences with an E-value < 1 are assigned to their phylogenetic node based on the target species' phylogeny and the taxonomic information from the NCBI taxonomy database [90].

Following this procedure, we store for each homologous sequence its accession id, the species name, the phylogenetic node, the percentage of sequence identity, and the percentage of alignment coverage in a NoSQL MongoDB® database. With the database, we can post hoc filter the BLAST results to study the robustness of the phylostratigraphic map without rerunning the BLAST computations.

5.2.4 PhyloWeb - Retrieving phylostratigraphic maps

To enable researchers the access to our computed phylostratigraphic maps (PS maps) and to adjust the PS assignments based on different filter criteria, we developed a web server front-end. called Phyloweb. The web application uses the Flask micro web framework written in Python and is connected to the MongoDB® database containing the BLAST results. It allows users to interactively manipulate the threshold parameters to modify the phylostratigraphic maps.



Phylostratigraphic map

Arabidopsis thaliana

Evalue: 0.00001

Identity: 0

Coverage: 0

PS	PS name	Gene ID
1	Cellular organisms	AT2G05840.1
1	Cellular organisms	AT2G04940.1
1	Cellular organisms	AT2G04350.2
1	Cellular organisms	AT2G06000.1
1	Cellular organisms	AT2G07750.1
1	Cellular organisms	AT1G07930.1
1	Cellular organisms	AT2G05710.1
1	Cellular organisms	AT2G04030.1
1	Cellular organisms	AT1G07880.1
1	Cellular organisms	AT2G05220.1
1	Cellular organisms	AT1G07890.2
1	Cellular organisms	AT2G06510.1
1	Cellular organisms	AT2G04520.1
1	Cellular organisms	AT2G04390.1
1	Cellular organisms	AT2G04650.1
1	Cellular organisms	AT2G03730.2
1	Cellular organisms	AT2G04842.1
1	Cellular organisms	AT2G07690.1
1	Cellular organisms	AT1G07780.2
1	Cellular organisms	AT2G06990.1

Figure 5.1 | Screenshot of the PhyloWeb server. The upper tabs show three different perspectives on the data. The PS map view presents the user with an interactive phylostratigraphic map of its target species. By applying different thresholds for the E-value, the sequence identity, or the alignment coverage, the user can modify the PS assignments. The tree on the left shows the target species' phylogeny with the number of assigned target genes in brackets. The phylogeny nodes are interactively collapsable or expandable, immediately changing the assignments in the phylostratigraphic map on the right.

It also allows to interactively adjust the phylogeny by expanding or collapsing nodes of the underlying phylogenetic tree. There are three different views available to investigate the phylostratigraphic results (Fig. 5.1). In the 'PS map' view, users can choose between different precalculated PS maps from various species and can modify the similarity thresholds to define homologous sequences in the analysis.

The PS map is summarized in the target species' phylogenetic tree, showing the number of genes assigned to a particular PS. The tree branches can be collapsed or expanded, followed by an immediate change of the PS map. The visualization can be modified by applying different fonts, line widths, or colors and can be downloaded in a publication-ready quality as SVG or pdf.

The table on the right (Fig. 5.1) displays the detailed PS map. The user can search for a particular gene to retrieve more information from the similarity searches in the 'Gene detail' view. Similar to the 'PS map' view, the phylogeny is displayed containing the number of all homologous sequences of the chosen gene assigned to the closest phylogenetic node shared with the query species. Detailed inspections of the alignments between a homologous and its query sequence are possible in the 'Alignment detail' view.

5.3 Results and Discussion

In this section, we will review the entropic transcriptome patterns of *Da. rerio*, *D. melanogaster*, and *A. thaliana* during embryogenesis and also for *A. thaliana* for germination, floral transition, and flower development. In subsection 5.3.1, we will present the entropic TAI profiles and find out if they also show an hourglass pattern. In subsection 5.3.2, we will quantify the significance of the hourglass shaped $eTAI$ patterns and compare them to the traditional $mTAI$ patterns.

5.3.1 Hourglass patterns of the entropic TAI

The traditional $mTAI$ was initially introduced by Domazet-Lošo in 2010 [9]. It is based on a gene age inference approach termed phylostratigraphy [8], which harnesses BLAST searches [91] to assign protein-coding genes a phylogenetic age by identifying homologous sequences in other species on a tree of life scale. Using this approach, genes can be sorted into discrete age categories, referred to as phylostrata (PS), which correspond to taxonomic nodes in the tree of life. To construct the TAI measure, phylostratigraphy based gene age inference is performed for all protein-coding genes of a reference organism of interest. The information about gene ages is then combined with expression levels covering the biological process of interest. Together, the weighted mean of all gene ages and their expression levels is referred to as the TAI.

In previous studies this mean transcriptomic age was capable to uncover the transcriptomic hourglass pattern during embryogenesis and in post-embryonic developmental processes in plants like germination and floral transition. Alternatively to the mean age of expressed genes, the detection of changes in the whole age distribution during the process of interest could provide more insights into the observed phenomena if the phenomena is still detectable. Using the concept entropy [89] we are able to study changes in the age distribution as a function of time.

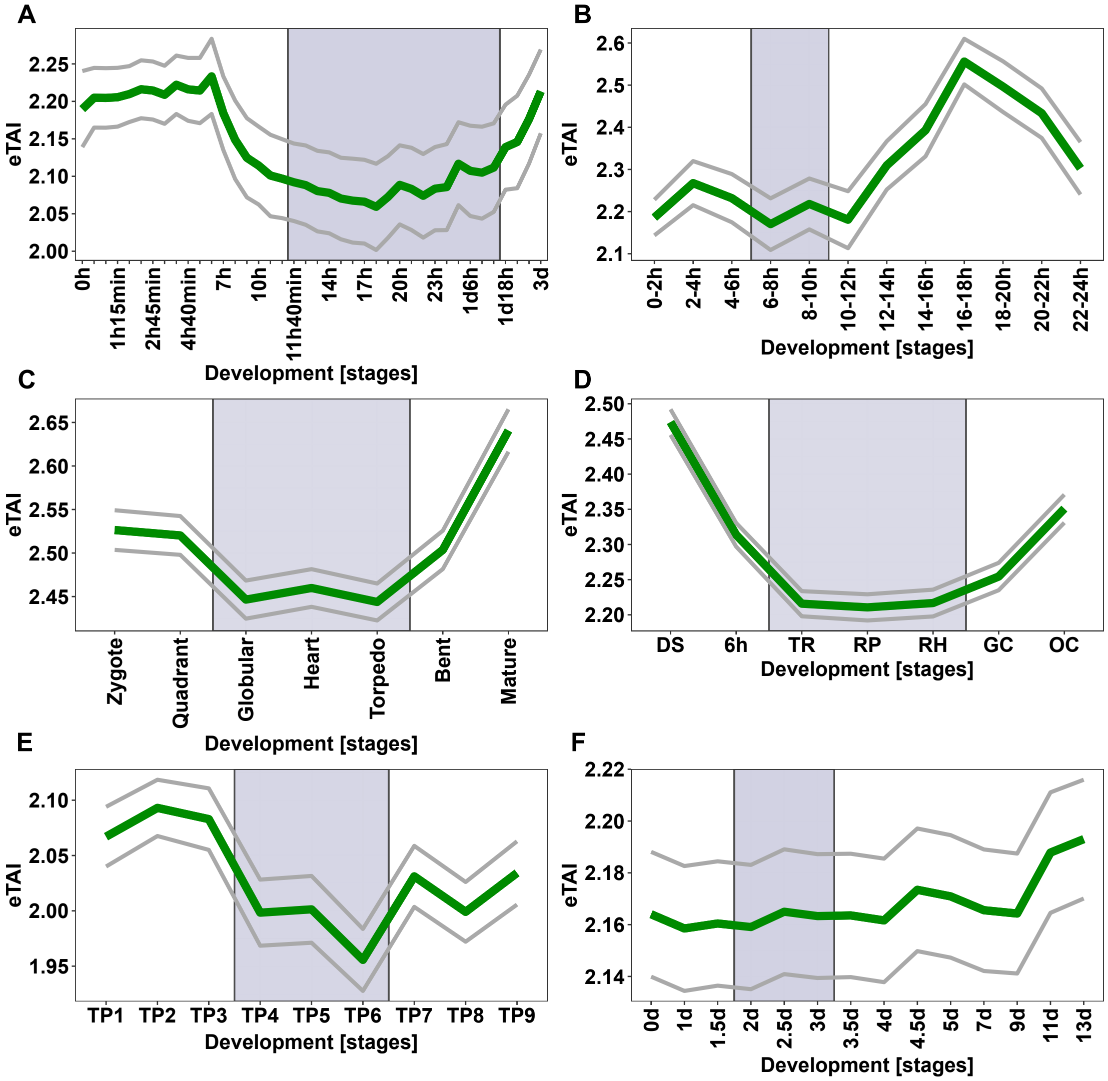


Figure 5.2 | Entropic TAI profiles across animal and plant development. Developmental stages of embryogenesis in (A) *D. melanogaster*, (B) *Da. rerio*, (C) *A. thaliana*. Post-embryonic transitions of *A. thaliana* during (D) germination, (E) floral transition. (F) Flower development of *A. thaliana*. The blue shaded area highlights the predicted phylotypic stage respectively for germination and floral transition the stages with the highest degree of conservation. The gray lines represent the standard deviation estimated by permutation analysis.

In Figure 5.2 we use the entropy on previously published datasets and find the same hourglass shaped patterns during the embryogenesis of *Da. rerio* [9], *D. melanogaster* [92], and *A. thaliana* [28]. Additional to embryogenesis we also detect an hourglass pattern during germination [86] and floral transition [4] in *A. thaliana* indicating a decrease in the age heterogeneity for the predicted conserved stages of the different processes.

The decrease of the age heterogeneity is in accordance to the checkpoint model [93] stating that the highly conserved stages during embryogenesis and post-embryonic processes are a consequence of prohibiting different developmental programs to ensure an ordered transition between developmental programs. This constrain leads to an expression shift of old genes, and thus increasing the homogeneity of the transcriptome age respectively decreasing the heterogeneity.

To confirm the findings of chapter 4, we calculate the $eTAI$ profile of *A. thaliana* flower development (Fig. 5.2 F). Also the $eTAI$ approach is not able to detect a possible hourglass pattern for this process of organogenesis.

5.3.2 Entropic hourglass patterns are highly significant

While Figure 5.2 shows that the concept of entropy is able detect hourglass patterns by plotting the changes of the whole transcriptome age distribution as a function of time, we were interested if the observed pattern are reliable on a quantitative scale.

We perform reductive hourglass tests [41] to verify if the presented profiles are not random observations. As mentioned in section 3.3.4, the test is designed to detect hourglass shaped profiles and determines their significance. As presented in table 5.1 the entropic hourglass patterns of embryogenesis in *Da. rerio*, *D. melanogaster*, and *A. thaliana*, as well as entropic hourglass patterns of germination and floral transition from *A. thaliana* are highly significant. Only the $eTAI$ profile of *A. thaliana* flower development does not show a significant hourglass shaped pattern.

Process	**Species**	**mTAI**	**eTAI**
Embryogenesis	***Da. rerio***	8.22e-03	**1.63e-05**
	D. melanogaster	4.52e-02	**4.50e-06**
	A. thaliana	1.75e-07	**1.36e-33**
Germination	***A. thaliana***	1.87e-16	**6.81e-98**
Floral transition	***A. thaliana***	5.28e-05	**2.85e-14**
Flower development	***A. thaliana***	1.19e-01	**4.41e-1**

Table 5.1 | P values of the reductive hourglass test. We calculate the P values based on the reductive hourglass test for the $mTAI$ and $eTAI$ profiles for embryogenesis in *D. melanogaster*, *Da. rerio*, *A. thaliana*, and Post-embryonic developmental processes of *A. thaliana* such as germination, floral transition, and flower development.

Comparing the calculated P values against the P values of the traditional phylotranscriptomic hourglass patterns of the mean age, the P values of the entropic hourglass patterns are orders of magnitudes lower. Such a strong signal is surprising, and it seems that the observed entropic hourglass is more fundamental than the original mean hourglass patterns. Hence, the question arises if the observed mean hourglass pattern is a consequence based on the changes of the whole age distribution measured by the entropic hourglass. Hence, the entropic hourglass could be considered as the origin of the mean hourglass pattern. The changes in the age distributions observed in the entropic hourglass may provide evidence for our hypothesis that the transcriptomic hourglass pattern may serve as a transition from one transcriptional program to the next.

5.4 Conclusions and Outlook

The quantification of transcriptome conservation based on the traditional transcriptome age index of the mean was and still is a widely used and accepted concept. This method uses the mean of the age distribution to measure the evolutionary conservation during developmental processes. Since it just reflects one parameter of the underlying distribution it is obvious to measure the degree of conservation with a different concept which is able to account the complete age distribution. Such a measure is the entropy which tries to captures the shape of the distribution. Based on this concept we introduced the entropic transcriptome age index to study developmental processes in the light of evolutionary conservation.

In this study we could show that the entropy of age distributions as functions of time is able to detect hourglass patterns during embryogenesis and post-embryonic development in plants and animals based on previously published data sets. Alternative, to the mean transcriptomic hourglass patterns we could detect an decrease in the heterogeneity of the underlying age distribution and thus a shift of the expression towards the set of old and conserved genes. This finding is in concordance with the organisational checkpoint model [93] in which the conserved stages are a consequence of prohibiting different developmental programs to ensure an ordered transition between developmental programs.

The result of this restriction can be detected with the entropy due to the increase of homogeneity in the age distribution. By measuring the significance of the entropic TAI patterns we could show that the P values of the reductive hourglass tests are many orders of magnitudes lower than the P values of the P values of the traditional mean transcriptomic hourglass patterns. Thus, we hypothesize that the entropic hourglass pattern might be the primary pattern, or origin, of the previously observed mean transcriptomic hourglass patterns representing the transition from one transcriptional program to the next by increasing the homogeneity of expressed genes towards the conserved stages.

In chapter 6, we will test this hypothesis by investigating if either the entropic TAI can reproduce the mean TAI hourglass patterns or the mean TAI can reproduce the entropic TAI hourglass patterns.

6

Deciphering the origin of the hourglass pattern

In this chapter, we will continue to attempt to find the origin of the transcriptomic hourglass pattern and to attempt to decipher its function. In chapter 5, we have defined the entropic transcriptome age index, and we have presented transcriptomic hourglass patterns with P values that are orders of magnitudes lower than the traditional transcriptomic hourglass patterns. We hypothesized that the observed $eTAI$ patterns could possibly be primary patterns for the $mTAI$ patterns.

In this chapter, we will attempt to test this hypothesis. After introducing the hypothesis in section 6.1, we will present in section 6.2 the computational approaches to predict the traditional $mTAI$ hourglass patterns based on the observed $eTAI$ hourglass patterns and vice versa. In section 6.3, we will present the reproduced transcriptomic hourglass patterns to find out which transcriptomic measure is more fundamental and might possibly be the primary pattern. In section 6.4, we will conclude that based on our experiments our hypothesis can neither be confirmed nor denied.

6.1 Introduction

Several studies showed that a hourglass pattern is also present at the molecular level during animal and plant embryogenesis [9, 10, 16, 30, 40, 41, 51, 59, 74, 77–79, 94, 95]. This observation was surprising as multicellularity and embryogenesis evolved independently in animals and plants [11] and suggests the convergent evolution of phylotranscriptomic hourglass patterns in animal and plant embryogenesis.

In chapter 4, we found that phylotranscriptomic hourglass patterns also exist in the two main transitions of post-embryonic plant development, germination and floral transition, suggesting the convergent evolution of phylotranscriptomic hourglass patterns in embryonic and post-embryonic plant development [74]. Based on the observed transcriptomic hourglass patterns in embryogenesis and post-embryonic developmental transitions, we postulated the “organisational checkpoint“ hypothesis [74, 93] as an attempt to explain the functional relevance of the transcriptomic hourglass patterns.

The hypothesis states that the waist of the hourglass pattern may serve as a transition by turning down one developmental process and setting up a new developmental process. To investigate this statement, we developed the entropic transcriptome age index, which is able to measure the whole age distribution of expressed genes.

In chapter 5, we thus detected five novel hourglass patterns based on the entropic transcriptome age index ($eTAI$). By determining the significance of these hourglass pattern with the reductive hourglass test, we found that the $eTAI$ hourglass patterns show P values that are orders of magnitudes lower than the P values of the corresponding $mTAI$ hourglass patterns. Representing the whole age distribution and showing such significant hourglass patterns, we wonder if the entropic TAI patterns are the origins of the traditional $mTAI$ hourglass patterns.

In this chapter, we attempt to answer this question by developing computational approaches which connect the traditional and entropic TAI measures. These approaches are designed to reproduce the $mTAI$ profiles based on the $eTAI$ measure and vice versa. Thus, we attempt to find the origin of the transcriptomic hourglass pattern and add evidence or contradict to the “organizational checkpoint“hypothesis.

6.2 Materials and Methods

In this section, we will investigate the dependence of the entropic TAI with the mean TAI. In subsection 6.2.1, we will present a method for sampling PS distributions and calculate the corresponding $mTAI$ and $eTAI$ values. To investigate the dependence of the two transcriptome measures we will present a loess regression analysis. In subsection 6.2.2, we will introduce a gradient-based search algorithm to either reproduce $mTAI$ profiles based on a gradient derived from the $eTAI$ or reproduce $eTAI$ profiles based on a gradient derived from the traditional $mTAI$. In subsection 6.2.3, we will introduce a $zscore$ derived normalization method for the $eTAI$ and $mTAI$ profiles. This normalization will be necessary to qualitatively compare the results of reproducing the $mTAI$ and $eTAI$ profiles.

6.2.1 Sampling of PS distributions and regression analysis

To study the mathematical relationships between $mTAI$ and $eTAI$ we test whether the $eTAI$ pattern might be the underlying origin of the observed $mTAI$ pattern. We therefore simulated mean PS distributions denoted as $\underline{\overline{\theta}}$, by calculating the arithmetic mean for each PS over all stages as

$$\overline{\theta}_{ps} = \frac{1}{S} \sum_{s=1}^{S} P_s(X = ps), \quad \forall ps \in \{1, \ldots, PS_{max}\}. \tag{6.1}$$

Based on $\underline{\overline{\theta}}$ we defined hyperparameters $\underline{\alpha}$ as

$$\alpha_{ps} = c \cdot \overline{\theta}_{ps}, \quad \forall ps \in \{1, \ldots, PS_{max}\} \tag{6.2}$$

and sampled 1,000 data points from this Dirichlet distribution $\mathrm{Dir}(\underline{\overline{\theta}}|\underline{\alpha})$, where the parameter c defines the width of the distribution depending on $\underline{\overline{\theta}}$.

By sampling PS distributions, we are able to calculate the corresponding $mTAI$ and $eTAI$ values and use the sampled data points to estimate a loess regression to study the mathematical relationship between the $mTAI$ and $eTAI$. The calculated the $mTAI$ and $eTAI$ values based on the sampled PS distributions are shown in Fig. 6.1.

When performing the regression analysis, we assumed that the data points are normally distributed, as it is indicated by the histograms of Figure 6.1.

It has to be mentioned that we assume a truncated normal distribution because the $mTAI$ and $eTAI$ values are defined within the closed intervals

$$1 \leq mTAI \leq \mathrm{PS}_{max} \tag{6.3}$$

$$0 \leq eTAI \leq \log_2(\mathrm{PS}_{max}) \tag{6.4}$$

Thus, by using locally estimated scatterplot smoothing regression (loess) we can estimate functions within the sampled data points which allow us to predict artificial $mTAI$ or $eTAI$ values based on the original $eTAI$ and $mTAI$ values.

6.2.2 Prediction of transcriptomic age index profiles based on a mean PS distribution

We perform a gradient search to investigate the dependence between $mTAI$ and $eTAI$. In Figure 6.2 and 6.2 we present the results of predicting the original $mTAI$ values based on an entropic TAI gradient and the prediction of the original $eTAI$ values based on a mean TAI gradient.

Algorithm 1: Gradient Search. This pseudocode outlines the algorithm to reproduce the original $mTAI$ or $eTAI$ value (defined by $xtai_{org}$). It starts at a particular PS distribution $\underline{\theta}^0$ and performs a gradient ascent or descent depending on $xtai_{org}$ and the transcriptome index value of $\underline{\theta}^0$.

```
input : $\underline{\theta}^0 := \overline{\theta}$,
        $xtai \in \{$"mTAI", "eTAI"$\}$,
        $gradXtai \in \{$"mTAI", "eTAI"$\}$,
        $xtai_{org}$ - original eTAI or mTAI value,
        $steps \in \mathbb{R}$ - step size for gradient search,
        $I_{max} \in \mathbb{N}$ - max. number of iterations
output: $xtai_{final}$ - predicted eTAI or mTAI value

// Define direction of the gradient
$d \leftarrow 1$
$xtai_{final} \leftarrow$ calculateXTAI ($\underline{\theta}^0, xtai$)
if $xtai_{org} < xtai_{final}$ then
    $d \leftarrow -1$
for $i \leftarrow 0$ to $I_{max}$ do
    $\underline{\beta} \leftarrow$ log($\underline{\theta}^i$)
    $\underline{\gamma} \leftarrow$ exp($\underline{\beta}$)/exp($\beta_\bullet$)
    $\nabla\theta \leftarrow$ gradient($\underline{\gamma}, gradXtai$)
    $\underline{\nu} \leftarrow \underline{\beta} + d \times step \times \nabla\theta$
    $\underline{\theta}^{(i+1)} \leftarrow$ exp($\underline{\nu}$)/exp($\nu_\bullet$)
    $xtai_{final} \leftarrow$ calculateXTAI($\underline{\theta}^{(i+1)}, xtai$)
    // Loop ends when $xtai_{org}$ is reached or slightly exceeded
    if ($d \times (xtai_{org} - xtai_{final})$) >= 0 then
        break
return ($xtai_{final}$)
```

As described by the pseudocode (Alg. 1), we start at the mean PS distribution ($\underline{\theta}^0 := \overline{\underline{\theta}}$). The variable $xtai$ defines the objective function, i.e., the index we want to predict ($mTAI$ or $eTAI$).

As shown in Fig. 6.1 the gradient has to point into different directions when we start at $\overline{\underline{\theta}}$ (red point in Fig. 6.1) and want to predict all values of the transcriptome age index profile (green points in Fig. 6.1). Thus, depending on the direction we perform either a gradient ascent ($d := 1$ in Alg. 1) or a gradient descent ($d := -1$ in Alg. 1).

Based on the value of the variable $xtai$ and the provided PS distribution $\underline{\theta}$ the function `calculateXTAI` (Line 3 and 12 in Alg. 1) calculates the corresponding transcriptome age index. Additionally, the function `gradient` (Line 9 in Alg. 1) returns the gradient of a requested transcriptome index, defined by $gradXtai$, for a particular PS distribution $\underline{\theta}$. If $gradXtai$ is set to $mTAI$ the gradient is computed as

$$(\nabla\theta)_{ps} := ps, \quad \forall ps \in \{1, ..., PS_{max}\} \tag{6.5}$$

and if $gradXtai$ is set to $eTAI$ the gradient is computed as:

$$(\nabla\theta)_{ps} := \frac{-1 - \texttt{log}(P_s(X = ps))}{\texttt{log}(2)}, \quad \forall ps \in \{1, ..., PS_{max}\} \tag{6.6}$$

Since we investigate the dependence between $mTAI$ and $eTAI$, we try to find the transcriptome index that outperforms the other by finding the PS distribution that fits the original transcriptome index profile the best. Consequently, we either set the variables $xtai := mTAI$ and $gradXtai := eTAI$ or $xtai := eTAI$ and $gradXtai := mTAI$.

The variable $xtai_{org}$ defines the specific time point in the transcriptomic age profile, we want to predict. The search stops if the predicted $xtai_{final}$ value reachs or slightly exceeds $xtai_{org}$. For the experiments shown in section 6.3.2 and Figures 6.2 and 6.2, we use a step size of 10^{-7} and a maximal number of iterations of 10^7.

6.2.3 Normalized transcriptomic age index profiles

In the previous section, we described the procedure to predict the $mTAI$ based on the $eTAI$ function (gradient) and vice versa. To study if the $eTAI$ can predict the traditional $mTAI$ qualitatively better than the $mTAI$ predicts the $eTAI$, we need to compare the differences between the predicted and the corresponding original profiles. As shown in Figure 6.1, the ranges of $mTAI$ (Eq. 6.3) and

$eTAI$ (Eq. 6.4) values are very different. Accordingly, we cannot directly compare the difference between the original and predicted $mTAI$ profiles to the difference between the original and predicted $eTAI$ profiles. We first have to normalize the profiles to compare the differences mentioned above in an objective manner, e.g., by comparing the Manhattan or Euclidean distance between the predicted and original profiles.

To transform the profiles we use the *zscore* approach, in which the values of each profile are normalized to its mean (Eq. 6.7) and its standard deviation (Eq. 6.8).

$$\mu_{xTAI} := \frac{1}{S} \sum_{s=1}^{S} xTAI_s \tag{6.7}$$

$$\sigma_{xTAI} := \sqrt{\frac{1}{S} \sum_{s=1}^{S} (xTAI_s - \mu_{xTAI})^2} \tag{6.8}$$

We therefore define the function `z(`$xTAI_s$`)` as

$$z(xTAI_s) := \frac{xTAI_s - \mu_{xTAI}}{\sigma_{xTAI}} \tag{6.9}$$

,with $xTAI_s$ denoting the $eTAI$ or $mTAI$ at a particular stage s.

6.3 Results and Discussion

In this section, we will investigate the dependence between the $eTAI$ and the $mTAI$. We will also try to decipher if on of the two transcriptome indices is the primary pattern. In subsection 6.3.1, we will attempt to find out to which degree the $eTAI$ and $mTAI$ are related to each other. In subsection 6.3.2, we will find out to which degree is the $eTAI$ able to reproduce the transcriptomic patterns of the $mTAI$ and vice versa.

6.3.1 Dependencies between transcriptome indices

To investigate the relationship between $mTAI$ and $eTAI$, we sample 1,000 PS distributions from a Dirichlet distribution with hyperparameters $\underline{\alpha}$ as defined in Eq. 6.2. For each sampled data point, resp. PS distribution, we calculate the corresponding $mTAI$ and $eTAI$ value. As shown in Fig. 6.1, the $mTAI$ and $eTAI$ values from sampled PS distributions form an ellipse in the scatter plot. The histograms of $mTAI$ and $eTAI$ values seem to follow a Gaussian distribution. Hence, based on the elliptic scatter plot, the presented $mTAI$ and $eTAI$ values could be modeled by a bivariate Gaussian distribution with a symmetric covariance matrix and non-zero off-diagonal elements, resp. non-zero covariances between $mTAI$ and $eTAI$. Thus, based on Fig. 6.1, we assume a particular degree of dependence between the two transcriptome indices.

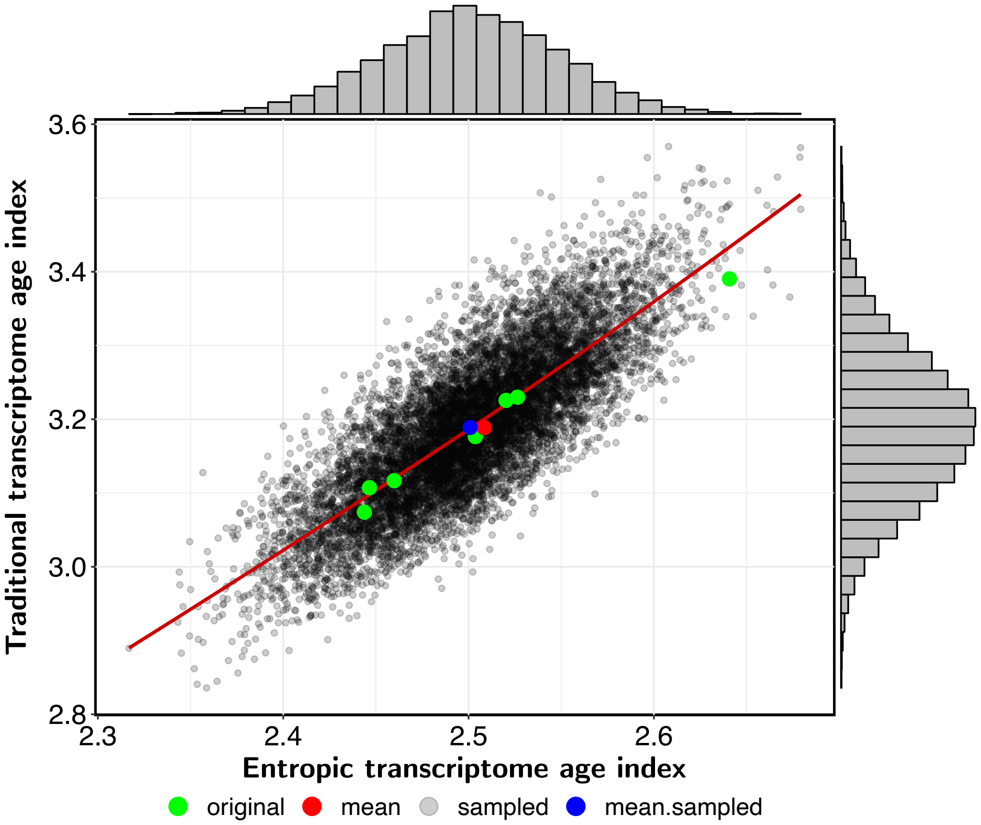


Figure 6.1 | Scatterplot of sampled PS distributions. This plot shows 1,000 sampled PS distributions (gray points) for c = 10^{-3} (Eq. 6.2). We calculate for each PS distribution the entropic TAI ($eTAI$) and traditional TAI ($mTAI$). The x-axis denotes $eTAI$ and the y-axis denotes the $mTAI$ of each PS distribution. The green points represent the PS distributions of the seven stages in *A. thaliana* embryogenesis, the red point corresponds to the mean of the seven stages of *A. thaliana*, and the blue point corresponds to the mean of the Dirichlet distribution, from which we have sampled the PS distributions.

Additionally, we calculate a loess regression within the 1,000 sampled data points to underline the dependency of $mTAI$ and $eTAI$. We see in Fig. 6.1 that the data points which are based on the original data set (green points) are close to the predicted regression line but only the original data points close to the mean PS distribution are also close to the regression line. If a data point is further away from the mean PS distribution the prediction misses the original observed data points.

Taken together, we see a dependence between the $mTAI$ and $eTAI$ value which could mean that one transcriptome measure may be able to predict the other one. As we see by the regression line in Fig. 6.1, that the loess regression may not be capable of precisely predict $mTAI$ values from $eTAI$ values or vice versa because of the deviation we see for PS distributions showing very high $mTAI$-$eTAI$ pairs or very low $mTAI$-$eTAI$ pairs. We need another approach to test if we are able to predict one transcriptome measure with the other.

6.3.2 Reproducing the hourglass pattern

To determine the relationship between $eTAI$ and $mTAI$, we attempt to reproduce the traditional transcriptomic hourglass pattern based on the $eTAI$ and vice versa we attempt to reproduce the entropic transcriptomic hourglass pattern based on the $mTAI$. The results of the study may support our goal to decipher the origin of the transcriptome hourglass patterns and also help us to decipher its function.

In Figure 6.2 we present the reproduced transcriptomic patterns and the original transcriptomic patterns, based on the gradient search approach of Sec. 6.2.2. We find from Figs. 6.2A, E, G, I that the $eTAI$ gradient is capable of producing hourglass shaped patterns which are very similar to the original $mTAI$ patterns. We also find from Figs. 6.2B, D, F, H, J that the $mTAI$ gradient is capable of producing hourglass shaped patterns which are very similar to the original $eTAI$ patterns. To quantify the degree of similarity between the reproduced profiles and the original profiles, we calculate the Euclidean and Manhattan distances between the normalized profiles.

In Fig. 6.2, we find that it is possible to some degree to reproduce $mTAI$ patterns from a $eTAI$ gradient and that it is also possible to a much weaker degree to reproduce $eTAI$ patterns from a $mTAI$ gradient. Only the entropic transcriptomic pattern of *D. melanogaster* embryogenesis (Fig. 6.2C-D) can be reproduced slightly more accurate by the $mTAI$ gradient compared to the entropic gradient.

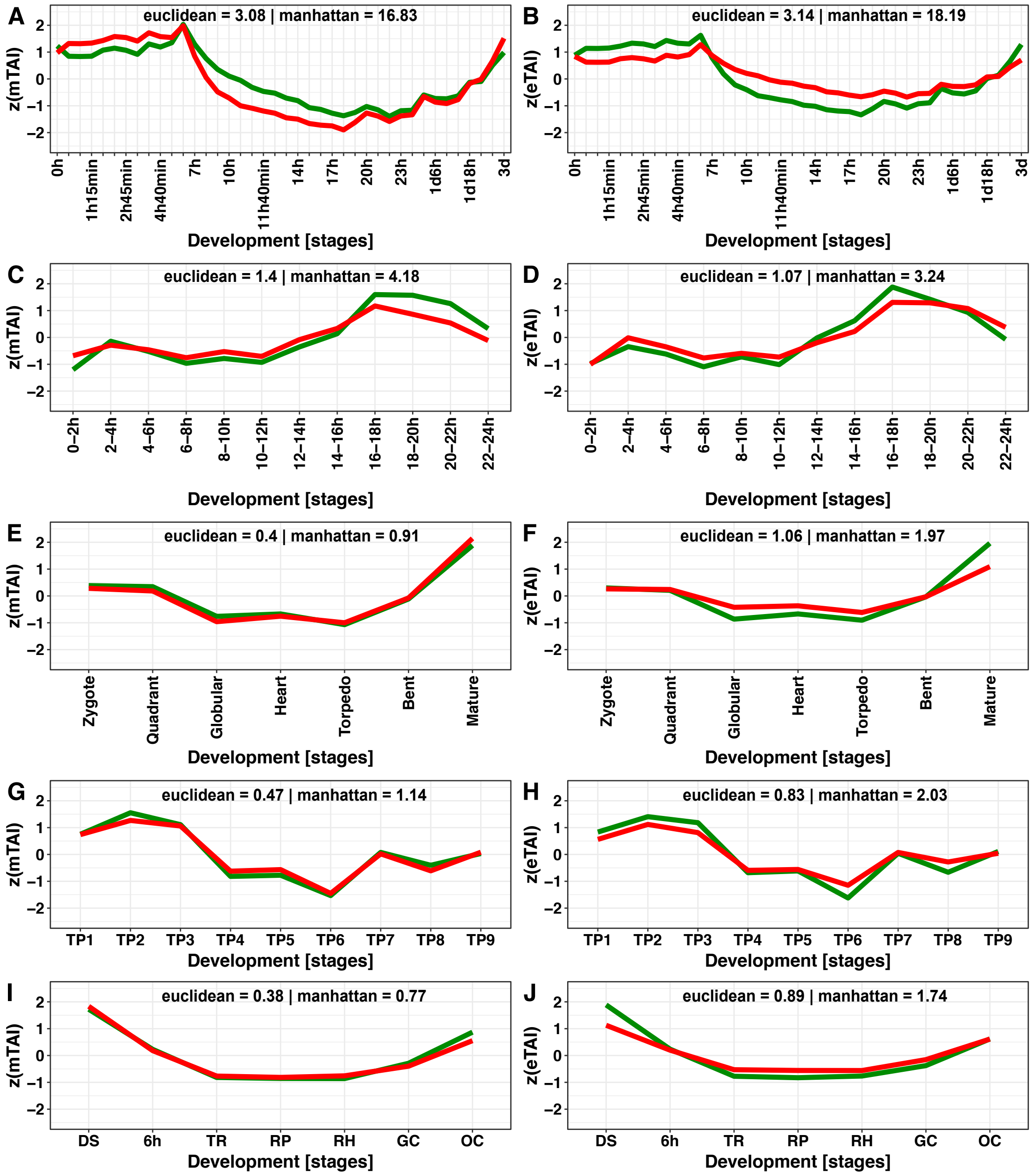


Figure 6.2 | Normalized mean and entropic transcriptome age indices. (A-B) *Da. rerio.*(C-D) *D. melanogaster.* (E-F) *A. thaliana* embryogenesis. (G-H) *A. thaliana* floral transition. (I-J) *A. thaliana* germination. (A, C, E, G, I) Partially reproducing the $mTAI$ hourglass patterns, starting from a mean age distribution and following the gradient of the $eTAI$ function. (B, D, F, H, J) Partially reproducing the traditional $eTAI$ hourglass patterns, starting from a mean age distribution and following the gradient of the $mTAI$ function. The green lines represent the normalized original transcriptomic patterns while the red lines represent the normalized reproduced patterns. The Euclidean and Manhattan distances on the top of each subplot quantify the differences between the original and reproduced normalized patterns.

The differences between the two approaches are very little, hence we cannot confirm nor discard our hypothesis that the entropic hourglass might be the origin of the hourglass pattern of mean. Regarding, the P values of the $eTAI$ patterns, we see a quantitative shift based on the P values which are orders of magnitudes lower than the P values of the corresponding traditional $mTAI$ patterns.

By representing the degree of homogeneity of the whole age distribution as a function of time, the $eTAI$ provides a new perspective to the traditional transcriptomic hourglass patterns. Currently, we cannot confirm if the entropic transcriptomic hourglass pattern is primary to the transcriptomic hourglass pattern of mean, but based on the two transcriptomic approaches we are able to support the "organizational checkpoint" hypothesis as we see not only a decrease of the mean age in the $mTAI$ hourglass patterns but also an increase in the homogeneity of the age distribution in the $eTAI$ hourglass patterns.

We could confirm these patterns for all embryonic and post-embryonic developmental transitions in chapter 5. Thus, we could speculate that at the beginning of the transcriptomic hourglass pattern, we see a heterogeneous and evolutionary young transcriptome. At the waist of the transcriptomic hourglass pattern, we suspect a shut down of the previous developmental process which leads to a homogeneous age distribution and the expression of conserved, evolutionary old, genes. After the shut down of one functional program, the transcriptome can now transition into the next process, resp. start a new functional program. Hence, we see after the waist in the transcriptomic hourglass pattern an increase in heterogeneity and the expression of evolutionary young genes.

6.4 Conclusions and Outlook

The transcriptomic hourglass pattern is a pattern that seems to exist on multiple scales measured by different phylotranscriptomic approaches representing the weighted mean age ($mTAI$) or the degree of age homogeneity as a function of time ($eTAI$). All these hourglass patterns have been confirmed in embryonic development of animals and plants and post-embryonic developmental transitions in plants. In chapter 5, we could show that the $eTAI$ hourglass patterns are highly significant which led us to the speculation that the entropic transcriptomic hourglass patterns could be the primary pattern, rep. origin, of the observed $mTAI$ hourglass patterns and thus could provide evidence to uncover the function of the transcriptomic hourglass patterns.

To test this hypothesis, we have studied the dependency of the $mTAI$ and $eTAI$ approach by developing a method to randomly sample age distributions and compare their corresponding $mTAI$ and $eTAI$ values. By applying a loess regression analysis, we could show that the two measures show to certain degrees a linear dependency. We further investigated this dependency by the attempt to decipher if the $mTAI$ is the primary pattern of the $eTAI$ patterns or vice versa. With a developed gradient search algorithm we attempt to partially reproduce the $mTAI$ patterns based on $eTAI$ gradients and partially reproduce the $eTAI$ patterns based on $mTAI$ gradients.

Based on this experiment we could neither confirm nor discard that one the transcriptomic measures is the origin of the other. But based on the dependency between the two transcriptomic measures we find further evidence that the transcriptomic hourglass pattern may serve as a transcriptional switch. The decrease on the mean age and the increase of age homogeneity could be interpreted as a consequence of shutting down one transcriptional program. Based on the decrease of transcriptional activity only the remnants of previously highly expressed conserved, resp. evolutionary old, transcripts are measured in the waist of the transcriptomic hourglass pattern. After the waist the next transcriptional program gets initialised leading to an increase in transcriptional activity, which leads to an increase of age heterogeneity (high $eTAI$) and thus an increase of the mean evolutionary age (high $mTAI$).

In chapters 2 - 6, we have investigated transcriptomic hourglass patterns, and we have attempted to contribute to the field of evolutionary developmental biology by developing novel phylotranscriptomic approaches. The phylotranscriptomic analysis was based on temporal gene expression data combined with evolutionary information such as the evolutionary gene age. In the next chapter, we will leave the field of evolutionary developmental biology and enter the field of developmental biology by studying the process of grafting, a unique and agriculturally relevant developmental process of plants enabling them to form chimeric organisms and increase yield and productivity. In contrast to the previous chapters, gene expression will be measured on a spatial- and temporal resolution. To analyze this data, we will develop and apply customized bioinformatics analyses and statistical approaches.

7

Transcriptome dynamics at Arabidopsis graft junctions

In this chapter, we will investigate the transcriptome dynamics of grafting, a remarkable property of many plants allowing the formation of a chimeric organism by joining cut tissues. To better understand this process at the molecular level, we will learn about a genome-wide analysis of temporal and spatial gene expression changes in grafted *A. thaliana* hypocotyls. My colleagues and I developed a bioinformatics pipeline to verify and quantify the obtained RNA-Seq data to subsequently analyze the gene expression profiles. The analysis uncovered an intertissue recognition mechanism characterized by an asymmetric gene expression of sugar-associated genes and a symmetric gene expression of auxin-response genes above and below the graft junction. The findings indicate that wound healing is proceeded via different mechanisms depending on the presence or absence of adjoining tissues.

In section 7.1, we will introduce the topic of transcriptome dynamics during grafting and its underlying molecular processes. In section 7.2, we will present the quantification of gene expression data, the detection of differentially expressed genes, followed by the analysis of gene sets connected to grafting their functional analysis. In section 7.3, we will examine the transcriptome dynamics in a spatial-temporal resolution after grafting and we will compare the differentially expressed genes with published datasets. In section 7.4, we will discuss our results with respect to the symmetric and asymmetric gene expression above and below the graft junction. In section 7.5, we will conclude our findings, discuss limitations of the bioinformatics analysis and give an outlook to future work.

The following sections are extracted from Melnyk et al. 2018 "*Transcriptome dynamics at Arabidopsis graft junctions reveal an intertissue recognition mechanism that activates vascular regeneration*" [96].

7.1 Introduction

For millennia people have cut and rejoined plants through grafting. Generating such chimeric organisms combines desirable characteristics from two plants, such as disease resistance, dwarfing and high yields, or can propagate plants and avoid the delays entailed by a juvenile state [97]. Agriculturally, grafting is becoming more relevant as a greater number of plants and species are grafted to increase productivity and yield [98]. However, our mechanistic understanding of grafting and the biological processes involved, including wound healing, tissue fusion and vascular formation, remain limited.

Plants have efficient mechanisms to heal wounds and cuts, in part through the production of wound-induced pluripotent cells termed “callus”. The callus fills the gap or seals the wound, and later, differentiates to form epidermal, mesophyll and vascular tissues [99]. In grafted *Arabidopsis* hypocotyls, tissues adhere 1-2 days after grafting and the phloem, the tissue that transports sugars and nutrients, connects after three days [100, 101]. The xylem, tissue that transports water and minerals, connects after 7 days [100]. Plant hormones are important regulators of vascular formation, and at the graft junction, both auxin and cytokinin responses increase in the vascular tissue [100–102]. Auxin is important for differentiation of vascular tissues whereas cytokinin promotes vascular stem cells, termed “cambium”, to divide and proliferate in a process known as secondary growth [103, 104].

Auxin is produced in the upper parts of a plant and moves towards the roots via cell-to-cell movement. Auxin exporters, including the PIN proteins, transport auxin into the apoplast, whereas auxin importers, such as the AUX and LAX proteins, assist with auxin uptake into adjacent cells [104]. Disrupting this transport, such as by mutating *PIN1*, inhibits healing of a wounded stem [105]. Blocking auxin transport with the auxin transport inhibitor TIBA (2,3,5-triiodobenzoic acid) in the shoot inhibits vascular formation and cell proliferation at the *Arabidopsis* graft junction [102]. In addition to auxin, other compounds, including sugars, contribute to vascular formation. The localised addition of auxin to callus induces phloem and xylem but requires the presence of sugar [106, 107]. In plants, sugars are produced in the leaves and transported through the phloem to the roots [108]. The role of sugars in vascular formation and wound healing is not well established; however, sugars promote cell division and cell expansion [109], processes important for development including vascular formation.

The molecular and cellular mechanisms for wound healing, tissue reunion and graft formation remain largely unknown. One emerging theme is that the top and bottom of the cut do not behave similarly. Such tissue asymmetry occurs in other plant tissues, most notably leaves. Developing leaf primordia have an inherent asymmetry that is established early to specify differences between the top and the bottom of the leaf. External signals promote early leaf polarity changes but how asymmetry is established remains unknown [110]. Auxin depletion in the upper side of the emerging leaf might be important [111] or, alternatively, a meristem-derived lipophilic molecule could activate HD-ZIPIII proteins important for asymmetry [112]. Asymmetry also appears in cut *Arabidopsis* inflorescence stems where the transcription factor *RAP2.6L* expresses exclusively below the cut whereas the transcription factor *ANAC071* expresses exclusively above the cut [105]. Both were important for stem healing and *ANAC071* and a close homologue, *ANAC096*, were important for graft formation [102]. Asymmetry also exists in genetic requirements, since *ALF4* and *AXR1*, two genes involved in auxin perception, are important below but not above the graft junction for phloem connection [100]. However, *ANAC071* is expressed symmetrically around the hypocotyl graft junction 3 days after grafting [102] so the extent of asymmetry and the mechanistic basis for it during wound healing remains largely uncharacterized.

Previous efforts have characterised wound healing and tissue reunion using transcriptomic analyses. Mechanical wounding altered $\sim 8\%$ of the *Arabidopsis* transcriptome and showed a high degree of overlap with transcriptomic changes elicited by pathogen attack and abiotic stress [113]. Stem wounding and wound-induced callus formation altered the expression of hundreds or thousands of genes [105, 114, 115], whereas grafting grape vines, lychee trees and hickory trees induced hundreds or thousands of differentially expressed genes involved in hormone response, wound response, metabolism, cell wall synthesis and signal transduction [101, 116–119]. These grafting studies provide limited information, as tissues from above and below the graft junction were not isolated to test whether these tissues behaved differently, and controls were not performed to distinguish how grafting and tissue fusion might differ from a response associated with cut tissues that remained separated.

Here, we perform an in-depth analysis to describe the spatial and temporal transcriptional dynamics that occur during healing of cut *Arabidopsis* tissues that are joined (grafted) or left unjoined (separated). We find that the majority of genes differentially expressed are initially asymmetrically expressed at the graft junction and that many of these genes are sugar responsive, which correlates with severing of the phloem tissue and the accumulation of starch above the junction. However, genes associated with cell division and vascular formation activate on both sides of the graft and, similarly, auxin responsiveness activates equally on both sides. We propose that the continuous transport of substances, including auxin, independent of functional vascular connections, promoted division and differentiation, while the enhanced auxin response and blockage of sugar transport provided a unique physiological condition to activate genes specific to graft formation that promote wound healing.

7.2 Materials and Methods

In this section, we will describe the data preparation and analysis of the RNA-Seq data. In subsection 7.2.1, we will describe the sampling of the data and the preparation of the RNA-Seq libraries. In subsection 7.2.2, we will explain the quantification of gene expression values. In Subsection 7.2.3, we will learn about the BaySeq analyses performed by Thomas Hardcastle to detect differentially expressed genes. In subsection 7.2.4, we will explain the comparison of differentially expressed genes against published datasets. In subsection 7.2.5, we will describe the statistics to quantify significance of up- and down-regulated gene sets. In subsection 7.2.6, we will present the matching of probe ids to their corresponding gene ids to compare out findings with published microarray data. In subsection 7.2.7, we will describe the one-sided Fisher's exact test to statistically quantify enriched gene sets that are involved in graft formation. Finally in subsection 7.2.8, we will present the GO enrichment analysis for the functional annotation of gene sets.

7.2.1 RNA-Seq sample and library preparation

The grafted wild-type *A. thaliana* accession Col-0 was harvested at the respective time points and care was taken to separate grafts by gently pulling plants apart. Approximately 0.5 mm of tissue was taken above or below each cut site and kept separate. Intact plants had 1 mm of tissue taken from a similar location on the hypocotyl as separated or grafted plants. Grafted, separated, or intact tissues were pooled into groups of ∼24 tissues. Tissues were ground using a microcentrifuge pestle frozen in liquid nitrogen. RNA was extracted using an RNeasy Kit (Qiagen) following the manufacturer's instructions. RNA (90-100 ng) was used to prepare RNAseq libraries using the TruSeq Stranded mRNA LT kit (Illumina) according to the manufacturer's instructions. The final PCR was for 15 cycles, and 11-12 barcoded samples were randomly mixed to make a total of seven mixes for seven flow lanes, one mix per lane. Biological replicates of each sample were sequenced on the HiSeq 4000 platform with paired-end 75-bp transcriptome sequencing (BGI Tech Solutions). RNA-seq data are available from the Gene Expression Omnibus database (GSE107203).

7.2.2 Quantification of gene expression

The reads acquired through high-throughput sequencing were quality trimmed with sickle [120] to increase the read quality before mapping. Reads were aligned to protein-coding gene sequences acquired from TAIR10 using Bowtie2 [121]. Read assignment was performed using the eXpress tool [122], which also provided effective gene lengths for use in normalisation. Library scaling factors were inferred from the sum of the number of reads assigned to the genes in the lowest seventy-five percentiles of expressed genes for each library [123].

7.2.3 Detection of differentially expressed genes

Analyses of the data were carried out using the R package baySeq [124] and clustering based on the posterior probabilities acquired from this package. For each time point, all possible patterns of differential expression between the graft types were considered, where a 'pattern' defines similarity and difference between different experimental conditions. For example, '$\{Col\_cut\_bottomGenes = Col : Col\_bottomGenes = ungraftedGenes\}, \{Col\_cut\_topGenes = Col : Col\_topGenes\}$' defines a pattern in which gene expression is equivalent in the separated bottoms, the grafted bottoms and the intact plant, but different to the equivalently expressed separated top and grafted top. The total number of possible patterns for five experimental conditions (as in this analysis) is fifty-two.

For a given time point, posterior likelihoods on the likelihood of each pattern of expression are calculated for every gene with greater than ten reads across all experimental conditions. The patterns were then modified to include orderings (denoted by $<$ or $>$), for example, the pattern described would lead to the ordered pattern '$\{Col\_cut\_bottomGenes = Col : Col\_bottomGenes = ungraftedGenes\} > \{Col\_cut\_top - Genes = Col : Col\_topGenes\}$' in which gene expression is equivalent in the separated bottoms, the grafted bottoms and the intact plant and greater than the equivalently expressed separated top and grafted top. In total, 541 ordered patterns exist in this dataset. Posterior likelihoods for an ordered pattern were assigned to that of the unordered pattern for genes in which the (normalised) mean expressions within the equivalently expressed groups conformed to the ordering, and to zero otherwise.

Based on the posterior likelihoods for the ordered patterns, a similarity score s_{ij} was established between two genes i and j as the sum over the products of their likelihoods of each ordered pattern. A single-link agglomerative clustering of genes, in which a gene will join a cluster if it has a greater than 50% similarity to any gene within that cluster was then performed based on these similarity scores. We label each cluster according to the predominant ordered pattern with high likelihood amongst the genes that comprise it. The change in size of these clusters over time is shown for the major clusterings in Fig. 7.6.

We can also find likelihoods on comparisons between pairs of experimental conditions by summing the likelihoods over combinations of patterns. Fig. 7.3A shows the number of genes identified at each time point in a pairwise analysis between the grafted top and grafted bottom samples. The likelihood of symmetric expression (i.e., expression which is equivalent across the graft junction) is calculated as the sum of the likelihoods of all patterns in which the grafted top and grafted bottom samples are equivalent. Conversely, asymmetric expression is calculated as the sum of the likelihoods of all patterns in which the grafted top and grafted bottom samples are not equivalent. Additional sets can be formed by considering the ordering of the grafted top and grafted bottoms samples. Sets of genes are identified at each time point with an FDR of less than 0.05 and a likelihood of symmetric/asymmetric expression greater than 50%. Genes in this analysis were only included if they were differentially expressed relative to intact samples.

7.2.4 Comparison of up- and down- regulated genes with published datasets

To measure if the ratio of up-and down-regulated genes from a previously published dataset is significantly different from the ratio of up-and down-regulated genes in our grafting dataset, we only took into account genes that are differentially expressed at a particular time point. A gene is differentially expressed at a particular time point if the marginal likelihood, calculated by baySeq, is greater than 0.9 and if the absolute log2-foldchange was greater than 1. Hence, we only consider genes that are significantly two-fold up-or down-regulated. We also used this definition of differentially expressed genes to filter the published datasets according to our expression dataset. Hence, some genes were filtered out from the published gene sets because they did not show a significant up-or down-regulation at a particular time point in our expression dataset. The histograms (Figs. 7.2, 7.3, 7.4) show the relative number of up-and down-regulated genes from the published gene sets at a particular time point and a specific condition (separated top, separated bottom, grafted top, grafted bottom) based on the number of genes in the published gene set after filtering. To calculate the significance of the difference of the ratios between the DEGs in the published gene sets and all up-and down-regulated genes, we performed a two-sided Fisher's exact test. To correct for multiple testing, we used the Benjamini-Yekutieli (BY) correction method. Hence, the asterisks in the barplots highlight corrected P values below 0.05. Additionally, to the published Methods section, we describe the test procedure in more detail in the following section.

7.2.5 Fisher's exact test of ratios

We separately performed the test at each time point, assuming temporal independence of the ratios of up- and down-regulated genes. For each time point, we performed the following procedure. First, we extracted all gene ids from our expression dataset that showed a up- or down-regulation in grafted or cut samples compared to the corresponding intact samples, based on the baySeq calculations.

Second, we separately divided the groups of up- and down-regulated genes into foreground and background. As foreground (fg), we defined genes that were taken from a published gene set while the remaining genes were defined as background (bg). Based on this categorisation, we defined the following contingency table (Table 7.1).

regulation **published**	**up**	**down**	$\sum_{\textbf{regulation}}$
yes	$\mathrm{n}_{fg,up}$	$\mathrm{n}_{fg,down}$	n_{fg}
no	$\mathrm{n}_{bg,up}$	$\mathrm{n}_{bg,down}$	n_{bg}
$\sum_{\textbf{published}}$	n_{up}	n_{down}	$\mathrm{n} = \mathrm{n}_{fg} + \mathrm{n}_{bg}$

Table 7.1 | Contingency table for Fisher's exact test. The cells in this table represent the number of genes in a study that are up or down regulated and are already published as differentially expressed (yes) or are determined in out study as differentially expressed (no). The last column summarizes for each row the number of published or unpublished differentially expressed genes as n_{fg} or n_{bg}. The last row summarizes the number of genes that are up (n_{up}) or down regulated (n_{down}) in our study.

We calculated the corresponding P value of a two-sided Fisher exact test as

$$p = 2 \times \frac{\binom{\mathrm{n}_{fg}}{\mathrm{n}_{fg,up}}\binom{\mathrm{n}_{bg}}{\mathrm{n}_{bg,up}}}{\binom{\mathrm{n}}{\mathrm{n}_{up}}} \tag{7.1}$$

$$= 2 \times \frac{\binom{\mathrm{n}_{fg}}{\mathrm{n}_{fg,down}}\binom{\mathrm{n}_{bg}}{\mathrm{n}_{bg,down}}}{\binom{\mathrm{n}}{\mathrm{n}_{down}}}. \tag{7.2}$$

7.2.6 Dealing with probe ids from microarray datasets

Since some published datasets only used probe ids instead of gene ids to represent their differentially expressed genes, we first had to match these probe ids to their corresponding gene ids. This step was done with the R package biomartr [93]. If one probe id matched more than one gene id, we used all the corresponding gene ids and tested afterward if these genes were actually differentially expressed in our dataset. In some cases, one probe id was represented by more than one gene id. Hence, some gene sets contained slightly more gene ids than published probe ids. In contrast, some probe ids did not match to currently existing gene ids. Subsequently, some gene sets contained slightly fewer gene ids than published probe ids.

7.2.7 Gene sets involved in graft formation

Grafting-specific genes (Fig. 7.7, S10, Table S4 of [96]) were identified by taking clusters from the baySeq analysis that were specific to grafting (Table S3) and combining these clusters to generate a list of grafting-specific genes for which further analysis were performed. For calculating the significance of overlapping genes between the baySeq clusters and the published datasets a one-sided Fisher's exact test was applied, to prove if the overlap is greater than expected. The resulting P values were corrected for multiple testing by using the Benjamini-Yekutieli method. This procedure was also applied to generate Table 1 to study the overlaps of symmetrically and asymmetrically expressed genes in the grafting dataset with previously published sugar-responsive genes.

7.2.8 GO enrichment analysis

We performed a GO enrichment analysis on grafting-specific genes. We defined these gene sets based on the baySeq's cluster algorithm for each time point into grafted top, grafted bottom, and grafted top and bottom genes. We extracted the gene ontology annotations from the Bioconductor package org.At.tair.db [125] for each *A. thaliana* gene. To test if a particular GO term is significantly enriched in one of these gene sets, we performed a hypergeometric test using the R package GOstats [126]. We performed the Bonferroni method to correct the resulting P values against multiple testing. A GO term is enriched if the corrected P value is below 0.05.

To reproduce the results of the statistical analysis, the overlap studies, and the gene ontology (GO) enrichment analysis, the required R scripts and expression data are available via GitHub at `https://github.com/AlexGa/GraftingScripts`.

7.3 Results

In this section, we will examine the transcriptome dynamics in a temporal and spatial resolution, i.e., the asymmetrically and symmetrically activated gene expression above and below the graft junction. In subsection 7.3.1, we will introduce the experimental setup to study the grafting process based on RNA-Seq data and we will investigate the transcriptome dynamics at the graft junction based on marker genes and published gene sets. In subsection 7.3.2, we will investigate the asymmetrically expressed genes around the graft junction. In subsection 7.3.3, we will learn about the asymmetrically gene expression of glucose-related genes around the graft junction. In subsection 7.3.4, we will present the symmetrical gene expression of Auxin-induced genes. Finally in subsection 7.3.5, we will present the gene sets of the baySeq clustering approach and their association to published datasets.

7.3.1 Grafting activates vascular formation and cell division genes

To better understand the developmental processes that occur at the graft junction, we generated RNA deep sequencing libraries from *A. thaliana* hypocotyl tissues immediately above and immediately below the graft junction 0, 6, 12, 24, 48, 72, 120, 168 and 240 hours after grafting (HAG) in biological replicates for each tissue at each time point (Fig. 7.1A). Prior to RNA extraction, we separated top and bottom tissues at the graft junction. We found that the strength required to break apart the graft junction increased linearly (Fig. S1 of [96]) similarly to previously reported breaking strength dynamics of grafted *Solanum pennellii* and *Solanum lycopersicum* [127, 128]. When pulling apart grafts to separate top and bottom for sample preparation, grafts broke cleanly with minimal tissue from one half present in the other half (Fig. S1, Movie S1, S2 of [96]). We measured the amount of tissue from tops adherent to bottoms and vice versa (Fig. S1 of [96]) and found less than 4% cross-contamination.

In addition to grafting, we also prepared libraries from ungrafted hypocotyls ("intact" treatment) and cut plants that had not been reattached ("separated" treatment)(Fig. 7.1A). We herein refer to tissues harvested above the graft junction or from the shoot side of separated tissue as "top" and that from below the graft or from the root side of separated tissue as "bottom" (Fig. 7.1A).

To understand which developmental processes occurred at the graft junction, we looked at the expression of markers associated with vascular formation and cell division. Many markers of cambium, phloem and provascular were activated within 6 hours of grafting. Provascular markers typically showed an early peak of expres-

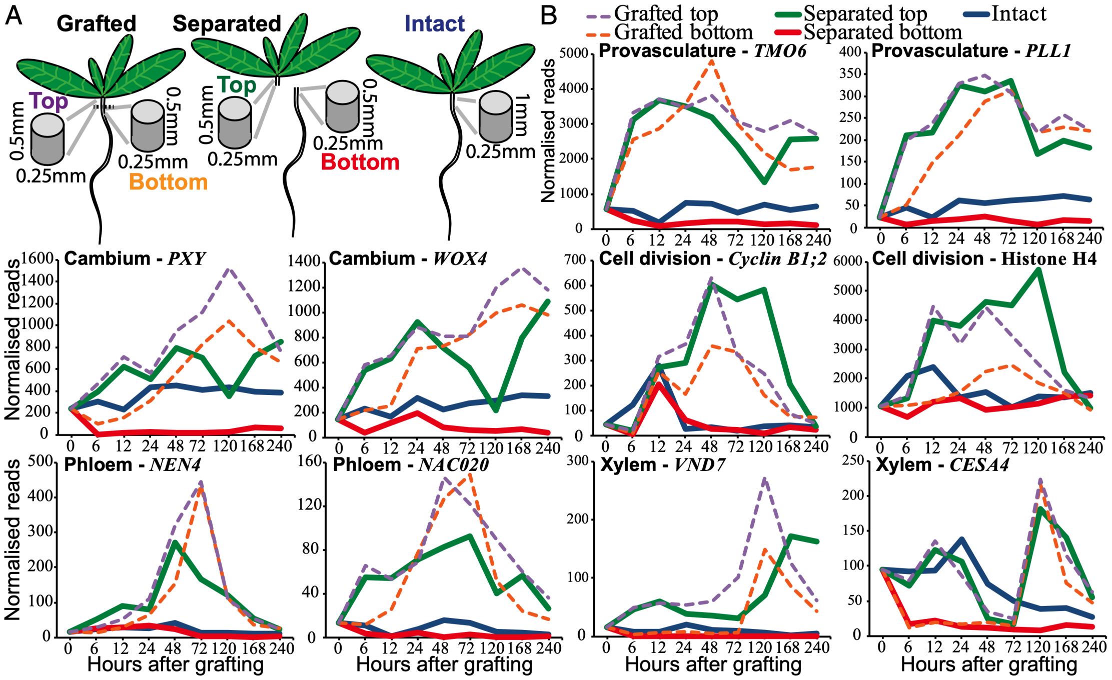


Figure 7.1 | Transcriptional dynamics of genes associated with provasculature, phloem, and xylem development and cell division. (A) Separated and grafted *Arabidopsis* tissues were harvested ~0.5 mm above (top) and ~0.5 mm below (bottom) the cut site. For intact plants, ~1 mm segments were harvested that spanned the region where cuts were made in grafted and separated plants. (B) Expression levels were plotted over time for intact, separated, and grafted samples. Reprinted Figure 1 from [96].

sion followed by a peak of cambial marker expression (Fig. 7.1, S2, S3 of [96]). Expression of phloem markers peaked at 72 hours (Fig. 7.1 and Fig. S2 of [96]), the time when phloem reconnections form in grafted *Arabidopsis* [100, 101]. Notably, the early phloem marker *NAC020* activated before the mid phloem marker *NAC086* which activated before the late-delevopment phloem marker *NEN4*, consistent with the dynamics of phloem transcriptional activation during primary root development and leaf vascular induction (Fig. S2 [96]) [129, 130].

Certain markers associated with xylem formation, such as *VND7* and *BFN1*, activated early in the grafted top. Other xylem markers, such as *IRX3* and *CESA4*, activated late in grafted samples. By 120 hours after grafting, genes activated in xylem development were expressed in top and bottom, consistent with when the first xylem strands differentiate at the graft junction [100]. Genes associated with cell division were activated by 12 hours in the grafted top and by 24 hours in the grafted bottom (Fig. 7.1 and Fig. S2 [96]). On the other hand, control genes, the expression of which does not typically vary between tissues and treatments [131], were not differentially expressed in grafted tops or bottoms (Fig. S2 [96]). The RNAseq expression data appeared to correlate well with transcriptional fluorescent reporters for both activation dynamics and the localization of expression (Fig. S3 [96]).

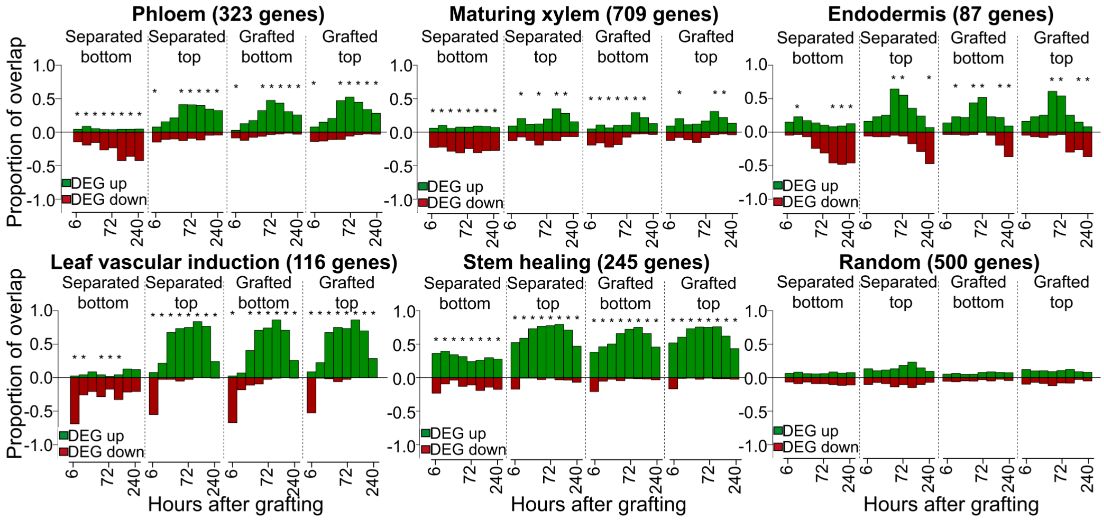


Figure 7.2 | Transcriptional overlap between previously published vascular datasets and the grafting datasets. Genes, the transcripts of which are associated with various cell types or biological processes, were taken from previously published datasets (Dataset S1) and compared with the transcriptomic datasets generated here. The number in parentheses represents the number of cell-type–specific or process-specific genes identified in previous datasets. Overlap is presented as a ratio of 1.0 for differentially expressed genes (DEG) up- or down-regulated in our dataset relative to intact samples compared with up- and down-regulated genes in the previously published transcriptome dataset. An asterisk represents a significant overlap ($P < 0.05$) for a given time point. Reprinted Figure 2 from [96].

The similar activation dynamics of vascular differentiation genes between grafting and leaf vascular formation prompted us to test whether this phenomenon occurred with other known developmental processes. We obtained lists of genes, the expression of which is associated with various biological processes from previous publications (Dataset S1 [96]), and tested how many of the genes differentially expressed in our transcriptomes overlapped with the previously published lists. Differentially expressed genes in grafted samples and separated tops partially overlapped with those the expression of which is associated with phloem, xylem, and procambium formation (Fig. 7.2 and Fig. S4 [96]). There was a high overlap between *Arabidopsis* inflorescence stem healing and grafting, as well as between vascular induction in leaf disk cultures and grafting (Fig. 7.2).

Various genes expressed in a cell-type–specific manner also showed a high transcriptional overlap with graft formation, including phloem, endodermis, and protoxylem (Fig. 7.2 and Fig. S4 [96]). In nearly all cases, the separated top, grafted top, and grafted bottom samples showed similar activation dynamics. The separated bottom samples were exceptional, however, since gene expression associated with vascular development and cell-specific processes was typically down-regulated (Fig. 7.2 and Fig. S4 [96]). We also compared our datasets with RNAs expressed in longitudinal cross-sections of the *Arabidopsis* root [132]. There was little overlap between grafted bottoms and sections from the root meristematic zone, whereas overlap existed between grafted tops and the root meristematic zone at early time points and between grafting and the root maturation zone (Fig. S5 [96]). Our analysis also revealed that two genes expressed in the cambium, *WOX4* and *PXY*, were induced by grafting, but the primary root marker *WOX5* and the lateral root marker *LBD18* were not substantially induced (Fig. 7.1 and Fig. S2 [96]).

7.3.2 Genes are asymmetrically expressed around the graft

Many of the vascular development and cell-division–related genes initially activated in the grafted top whereas, in some instances, activation was delayed in the grafted bottom by up to 24 hours (Fig. 7.1 and Figs. S2 and S3 [96]). Several genes important for tissue reunion or graft formation show an asymmetric pattern of expression above and below the cut [100, 105], suggesting that asymmetry might be a common feature of grafting and tissue reunion.

To investigate the extent of asymmetry at the graft junction, we identified RNAs that were differentially expressed equally in tops and bottoms of grafts (symmetrically expressed) or were more highly expressed in one tissue than in the other (asymmetrically expressed). We identified these genes by performing a pairwise comparison of the protein-coding transcriptome datasets that were differentially expressed as a consequence of grafting relative to intact hypocotyls. Several thousand RNAs were identified that fit either pattern of expression, including the transcript of the cambial markers *TMO6* that was induced symmetrically and *WOX4* that was induced asymmetrically (Figs. 7.1 and 7.3A).

Six to 48 hours after grafting, the number of graft-differentially expressed genes that were asymmetrically expressed was roughly threefold greater than those symmetrically expressed, indicating that tissues above the cut changed their expression dynamics relative to those below the cut. However, at 72 hours the numbers were nearly equal, and by 120 hours, the number of symmetrically differentially expressed genes was threefold greater than those asymmetrically expressed (Fig.7.3A). Some of the observed asymmetry at the graft junction might have been due to a gradient of differential expression along the length of the intact hypocotyl.

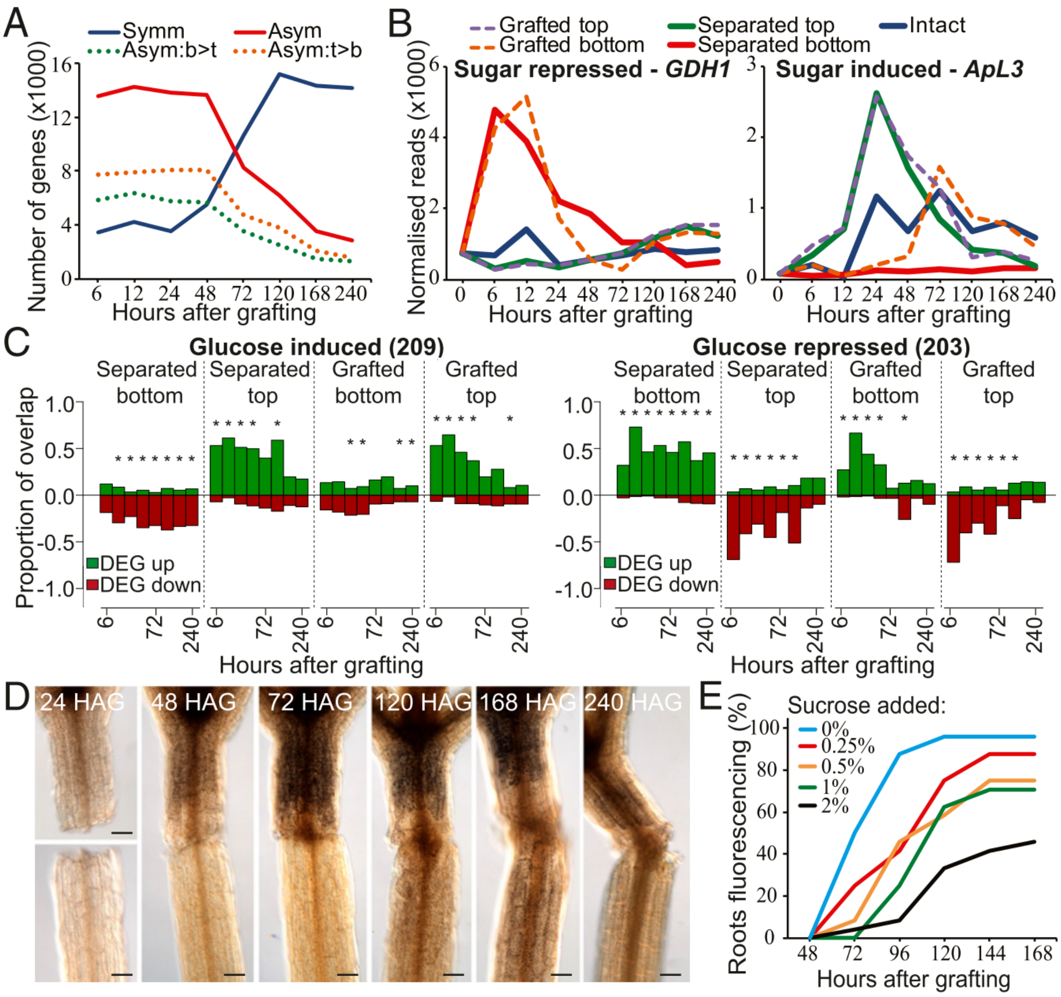


Figure 7.3 | Asymmetric changes in accumulation of sugar-responsive RNAs and of starch occur at the graft junction. A pairwise analysis between the grafted top and grafted bottom identified sets of protein-coding genes symmetrically or asymmetrically expressed at the graft junction [false discovery rate (FDR) < 0.05; likelihood of symmetric/asymmetric expression $> 50\%$]. Asymmetrically expressed genes were further divided into those the RNAs of which were higher in the top (orange dotted line) or higher in the bottom (green dotted line) (FDR < 0.05; likelihood of asymmetric expression $> 50\%$). (B) Expression profiles for transcripts of a sugar-repressed gene (*GDH*1) and a sugar-induced gene (*ApL*3) were plotted for intact, separated, and grafted samples. (C) Transcriptional overlap between previously published glucose-induced or glucose-repressed genes and our dataset. The number in parentheses represents the number of glucose-responsive genes identified in the previous dataset (Dataset S1). Overlap is presented as a ratio of 1.0 for differentially expressed genes (DEG) up- or down-regulated in our dataset relative to intact samples compared with up- and down-regulated genes in the previously published transcriptome dataset. An asterisk represents a significant overlap ($P < 0.05$). (D) Lugol staining of grafted plants at various time points revealed dark brown staining associated with starch accumulation. HAG, hours after grafting. (Scale bars: 100 µm.) (E) *pSUC*2 :: *GFP*-expressing *Arabidopsis* shoots were grafted to Col-0 wild-type roots and GFP movement to the roots was monitored over 7 days for phloem connection in the presence or absence of various concentrations of sucrose. Reprinted Figure 3 from [96].

We reasoned that, if asymmetry was due to inherent asymmetry in intact hypocotyls, then the average expression of a gene above and below the graft junction would be similar to its expression value in intact hypocotyls. We found that, for each time point, between 141 and 1,465 genes had expression values in intact hypocotyls that were similar to the average expression between the grafted top and grafted bottom (Fig. S2C [96]), suggesting that some of these genes may be asymmetrically expressed due to inherent asymmetry in the hypocotyl.

However, these numbers were a small proportion of the 13,000 genes asymmetrically expressed at early time points (Fig. 7.3A). As a second approach, we performed a hierarchical clustering analysis that indicated that the grafted top and grafted bottom were initially dissimilar but by 120 hours had clustered together and had become highly similar (Fig. S1 [96]), consistent with the symmetry analysis (Fig. 7.3A). Thus, graft healing promoted a shift from asymmetry to symmetry.

7.3.3 Sugar response correlates with asymmetric gene expression

The shift from asymmetry to symmetry could be due to phloem reconnection at 72 hours [100] and the resumption of hormone, protein, and sugar transport. We tested a role for sugar by grafting in the presence of exogenous sucrose, which has previously been reported to affect grafting success [133]. Low levels of exogenous sucrose lowered grafting efficiency (Fig. 7.3E), suggesting that differential sugar responses at the graft junction might be important for vascular reconnection. Expression of *ApL3*, a gene the expression of which is induced by sugar [134], was rapidly up-regulated in separated tops and grafted tops, whereas expression of $DIN6$, $GDH1$, and $STP1$, genes the expression of which is repressed by sugar [134–136], was rapidly up-regulated in separated bottoms and grafted bottoms (Fig. 7.3B and Fig. S6 [96]). These observations were consistent with sugar accumulation in the grafted top and sugar depletion in the grafted bottom. The expression of these genes returned to levels similar to intact samples by 120 hours and, with the exception of $ApL3$, the grafted samples normalized expression more rapidly than did the separated tissues.

Genes associated with photosynthesis increase expression in separated bottoms 24 hours after cutting, a common response to starvation [109], but likely too late to affect sugar levels before 24 hours (Fig. S6 [96]). A transcriptional overlap analysis with RNAs from known glucose-responsive genes (Dataset S1 [96]) revealed a substantial overlap with genes differentially expressed by grafting. RNAs from known glucose-induced genes were up-regulated in separated tops and grafted tops, whereas RNAs from known glucose-repressed genes were up-regulated in separated bottoms and grafted bottoms (Fig. 7.3C and Fig. S6 [96]).

This trend was not observed with genes differentially expressed by mannitol treatment (Fig. S6 [96]), suggesting that the effect was specific to metabolically active sugars. To further investigate this effect, we stained grafted, separated, and intact plants with Lugol solution to assay for the presence of starch. Staining above the graft junction increased 48–72 hours after grafting (Fig. 7.3D). By 120 hours, staining was equal on both sides of the graft whereas in separated tops staining became stronger after 72 hours (Fig. 7.3D and Fig.S6 [96]).

We concluded that starch accumulated above the cut, but after 72 hours, this asymmetry disappeared only in grafted plants. To test whether the accumulation of starch and increased sugar responsiveness could explain the observed transition from asymmetry to symmetry, we compared our datasets to previously published genes that are induced by starvation or are induced by sucrose readdition (Dataset S1 [96]). At early time points, 20–31% of asymmetrically expressed genes were known to respond to sugars, whereas only 2–5% of symmetrically expressed genes were known to respond to sugars (Table S1 [96]). However, at 72 hours, the overlap between asymmetrically expressed genes and sugar-responsive genes reduced substantially (Table S1 [96])

7.3.4 Auxin response is symmetric at the graft

The rapid activation of many vascular markers in the grafted bottoms despite the starvation response promoted us to investigate whether other mobile substances such as phytohormones could play a role in gene activation. We compared lists of genes known to respond to cytokinin, ethylene, or methyl jasmonate [137] and found no substantial overlap between these lists and genes differentially expressed by grafting (Fig. S7 and Dataset S1 [96]). Abscisic acid-responsive and brassinosteroid-responsive genes showed overlap with genes differentially expressed in our datasets, but this overlap was of a similar magnitude in both separated and grafted datasets, suggesting that the effect was not specific to grafting (Fig. S7 [96]). Auxin-responsive transcripts were exceptional, however, as they showed a substantial overlap with RNAs differentially expressed by grafting (Fig. 7.4A, B, and Fig. S7 [96]).

Auxin-induced genes were up-regulated in separated tops, grafted bottoms, and grafted tops whereas they were repressed in separated bottoms (Fig. 7.4A and B). Auxin-responsive genes such as $IAA1$ and $IAA2$ [138] were induced to similar levels in grafted tops and grafted bottoms by 24 hours. To further investigate whether the auxin response was uniform between grafted tops and grafted bottoms, we grafted the auxin-responsive fluorescent reporter $p35S : DII - Venus$, the fluorescent protein of which is degraded in the presence of auxin [139]. DII-Venus fluoresced in the separated bottoms but did not fluoresce in grafted bottoms 14 hours after cutting (Fig. 7.4C), indicating that separated bottoms had a low level of auxin response but grafted tops, grafted bottoms, and separated tops had a high level of auxin response.

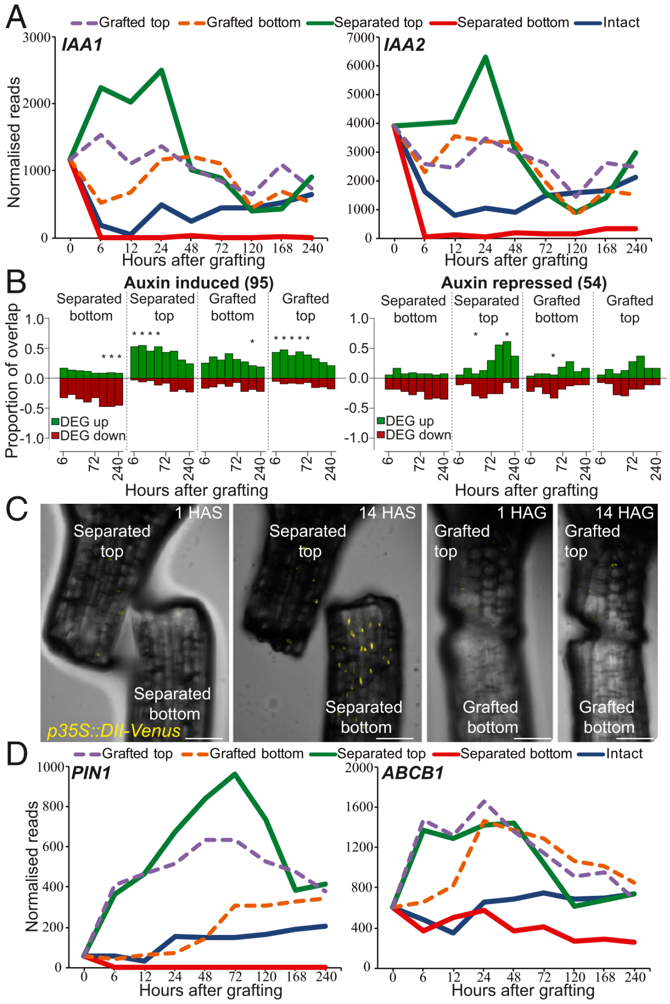


Figure 7.4 | Auxin response is symmetric at the graft junction. (A and D) Expression profiles for various auxin-responsive genes ($IAA1$, $IAA2$) or auxin transporter genes ($PIN1$, $ABCB1$) were plotted for intact, separated, and grafted samples. (B) Overlap between previously published auxin-induced or auxin-repressed RNAs and our dataset. The number in parentheses represents the number of auxin-responsive genes identified in the previous dataset (Dataset S1). Overlap is presented as a ratio of 1.0 for differentially expressed genes (DEG) up- or down-regulated in our dataset relative to intact samples compared with up- and down-regulated genes in the previously published transcriptome dataset. An asterisk represents a significant overlap ($P < 0.05$). (C) Grafted and separated plants expressing the auxin-responsive $p35S :: DII - Venus$ transgene that is degraded in the presence of auxin reveal a reduction of auxin response in cut bottoms, but not in grafted bottoms. HAG, hours after grafting; HAS, hours after separation (Scale bars: 100 µm.). Reprinted Figure 4 from [96].

To test whether auxin contributed to activation of gene expression below the graft junction, we monitored the expression of the symmetrically expressed gene *HIGH CAMBIAL ACTIVITY2* ($HCA2$) (Fig. 7.5A). We generated a transcriptional fluorescent reporter, $pHCA2 :: RFP$, that rapidly activated in grafted bottoms, grafted tops, and separated tops (Fig. 7.5B and C). Separated bottoms did not activate $pHCA2 :: RFP$ expression under grafting conditions or when placed on media containing sucrose or DMSO (Fig. 7.5B and D).

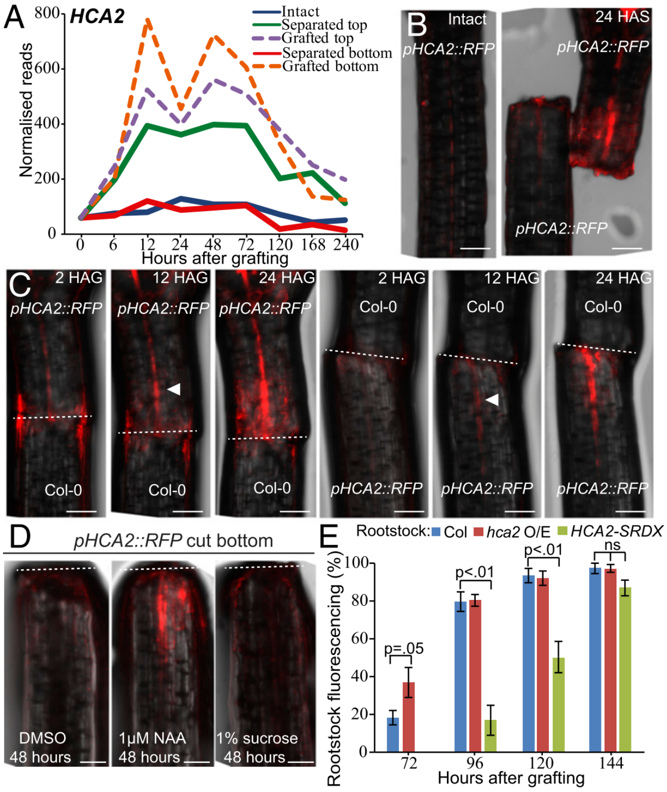


Figure 7.5 | $HCA2$ contributes to graft junction formation. (A) The RNAseq expression profile for $HCA2$ plotted for intact, separated, and grafted samples. (B and C) $HCA2$ transcription is up-regulated above and below the graft junction. $pHCA2 :: RFP$ was grafted to Col-0 roots or Col-0 shoots to avoid ambiguity of signal origin at the junction. $HCA2$ was also up-regulated in separated tops but not in intact samples or in separated bottoms. HAG, hours after grafting; HAS, hours after separation. White arrowhead denotes initial fluorescent signal; dashed lines denote the cut site. (Scale bars: 100 μm.) (D) Separated hypocotyl bottoms activated $pHCA2 :: RFP$ expression upon treatment of the synthetic auxin, NAA, after 48 hours but did not activate $pHCA2 :: RFP$ expression with DMSO or sucrose treatment. (Scale bars: 100 μm.) Dashed lines denote the cut site. (E) $pSUC2 :: GFP$-expressing *Arabidopsis* shoots were grafted to roots of Col-0 wild-type, $hca2$-overexpressing mutants, or plants expressing $p35S :: HCA2 - SRDX$. GFP movement to the roots was monitored 3–7 days after grafting for phloem connection. Reprinted Figure 5 from [96].

However, 26 hours of synthetic auxin [naphthaleneacetic acid (NAA)] treatment was sufficient to activate $pHCA2 :: RFP$ at the cut hypocotyl bottom but was insufficient to activate $pHCA2 :: RFP$ at the primary root tip of intact plants (Fig. 7.5D and Fig. S8 [96]). We also tested whether activation of $HCA2$ below the graft junction was important for grafting. Enhancing $HCA2$ activity ($hca2$ mutant) in grafted roots improved phloem reconnection rates, whereas suppressing $HCA2$ targets ($p35S :: HCA2-SRDX$) delayed phloem reconnection (Fig. 7.5E) [140].

7.3.5 Tissue fusion imparts a unique physiological response that differs from tissue separation

We hypothesized that the symmetric auxin response and asymmetric sugar response at the graft junction could allow a unique transcriptional response since neither separated plants nor intact plants had similar response dynamics to sugars and auxin (Figs. 7.3 and 7.4). To uncover protein-coding genes differentially expressed only by grafting, we segmented the transcriptome into groups of genes that behaved similarly and identified groups that corresponded to genes differentially expressed most highly by grafting (Dataset S2 [96]). We used an empirical Bayesian analysis [124] to define all possible patterns of differential expression between the five tissue types (intact, grafted top, grafted bottom, separated top, and separated bottom) with orderings allowed (< or >) (Fig. 7.6A).

This analysis produced 541 ordered patterns ("clusters") and, for each time point, posterior likelihoods on the likelihood of each pattern of expression were calculated for every gene in every tissue. A gene joined the cluster if it fits best, and a gene could join only one cluster at each time point. Although there were 541 possible clusters, we found that only 113 clusters contained 10 or more genes for at least one time point whereas 28 clusters contained 200 or more genes for at least one time point (Dataset S2 [96]). In the top 113 clusters, ~6,000 genes were differentially expressed in at least one tissue whereas between 1,000 and 4,000 genes were not differentially expressed (Fig. 7.6B).

To simplify the analysis, we considered clusters in which gene expression was grouped into patterns consisting of one comparison between two groups. At early time points, the cluster containing genes with similar differential expression in both grafted tops, grafted bottoms, separated tops, and separated bottoms had high numbers that decreased with time and could represent a general wound response (Fig. 7.6C).

A gene ontology (GO) analysis of the genes in this cluster revealed that they were highly enriched in defense, immune, and wound-responsive genes at 6 hours and that this enrichment decreased as the graft healed (Dataset S3 [96]). Clusters containing genes with similar differential expression in both separated tops and grafted tops had high numbers that decreased with time, similar to the trend observed with clusters containing genes with similar differential expression in both separated bottoms and grafted bottoms. This observation indicated that the grafted top was initially transcriptionally similar to the separated top, whereas the grafted bottom was initially transcriptionally similar to the separated bottom.

After the 48-hours time point, clusters containing genes differentially expressed only in separated tops or differentially expressed only in separated bottoms increased in numbers, suggesting that these tissues gained a unique pattern of gene expression. The clusters containing genes with similar differential expression in both grafted tops, separated tops, and grafted bottoms increase in numbers throughout the healing process (Fig. 7.6C).

We searched for grafting-specific cluster categories with one or more orderings in which genes were most highly differentially expressed by grafting (Dataset S2 [96]). There were very few genes down-regulated by grafting or up-regulated only in the grafted top (Fig. 7.7A). Instead, clusters contained several hundred differentially expressed genes up-regulated either in the grafted bottom only or up-regulated in both grafted bottom and grafted top (Fig. 7.7A). Genes, the expression of which changed only in the grafted bottom sample, were prevalent early during grafting and were most common at 48 hours, whereas genes activated in both top and bottom became prevalent at 48 hours and were most common at 120 hours (Fig. 7.7A and B).

We performed a GO analysis and found that genes differentially expressed most highly in the grafted bottom sample were enriched in the immune response and chitin response biological process categories (Dataset S3 [96]). Previously published chitin-induced RNAs had a high proportion of overlap with differentially expressed graft bottom-specific genes (Fig. 7.7C). A GO analysis also revealed that grafting-specific RNAs expressed in both the grafted top and grafted bottom were enriched in vascular-related biological processes (Dataset S3 [96]).

Previously published phloem-enriched, endodermal-enriched, vascular-induction, and stem-wounding associated RNAs had a high proportion of overlap with these differentially expressed graft-specific genes (Fig. 7.7C and Fig. S9 [96]). Since few genes were grafting-specific and grafted tissues were initially transcriptionally quite similar to separated tissues (Figs. 7.6C and 7.7A), we tested whether tissues separated for short periods could be grafted with similar reconnection dynamics as tissues that had been grafted immediately. Plants were cut and grafted after 0–5 days of separation. Separation did not speed up vascular reconnection, and instead, it always took 3 days from the point of tissue attachment for phloem connections to form (Fig. S9 [96]). Furthermore, the shoot lost competence to graft 2–3 days after separation whereas the root remained competent to graft up to 5 days after separation (Fig. S9 [96]).

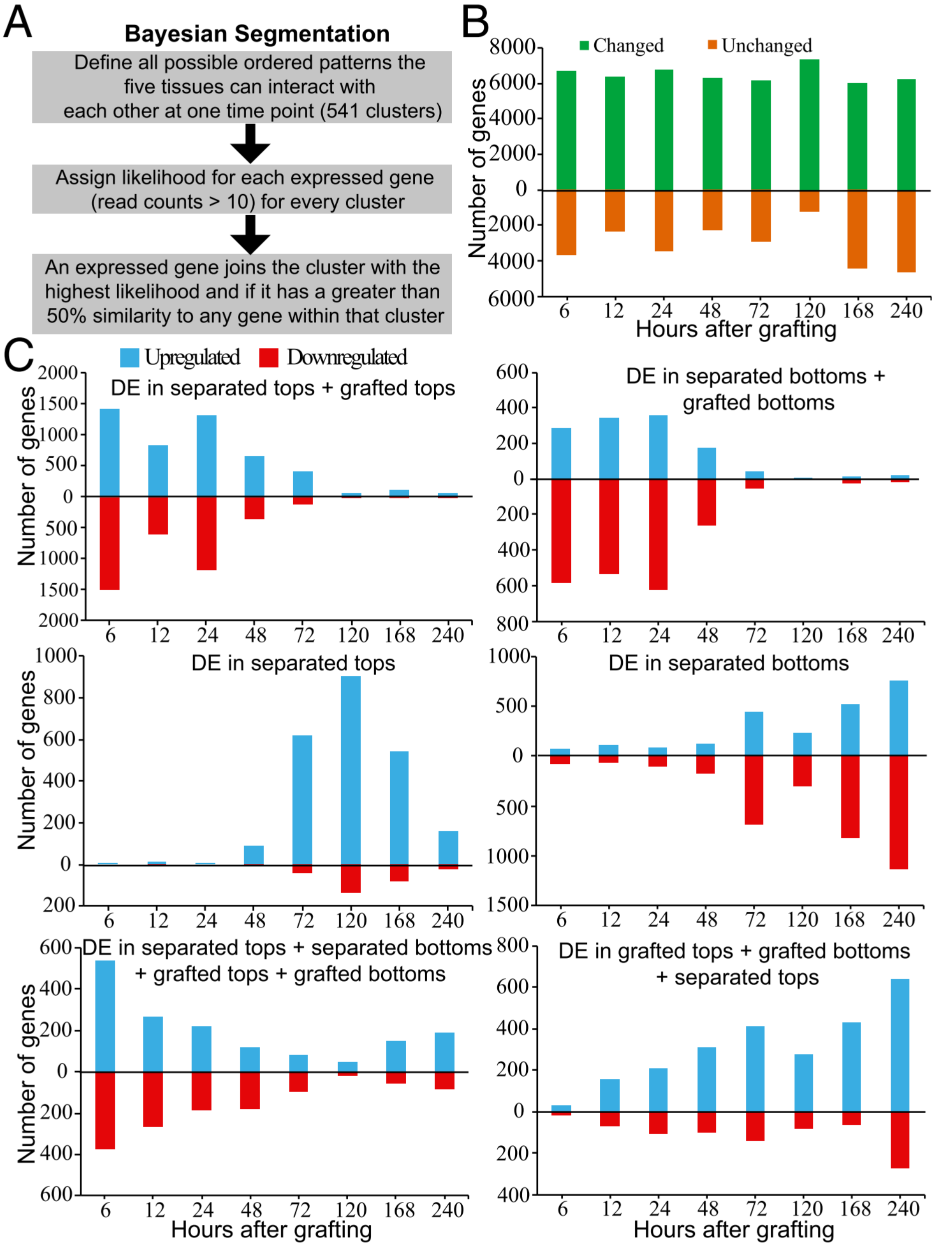


Figure 7.6 | Clustering the transcriptome at each time point, based on likelihoods of all possible patterns of differential expression (DE) in grafted, separated, and intact tissues. (A) Overview depicting the Bayesian segmentation. (B) Analysis of differential behavior produced 113 categories containing at least 10 genes, the expression of which was in a specific differential pattern for at least one time point (Dataset S2). One group is composed of genes the transcript levels of which are not substantially changed in the five tissues (unchanged), whereas the other group is composed of the sum of the other 112 groups (genes the transcript levels of which changed after treatment in at least one tissue) over the time points tested. (C) Major categories in the segmentation revealed RNAs the levels of which changed in all of the treatments listed relative to intact samples. Note that a gene can be represented in only one category for a given time point, that is, the category in which the transcript level changes best fit the category. Reprinted Figure 6 from [96].

Together, it appears that the grafted shoot and root have a unique physiological response that differs from the separated shoot and root and that tissue attachment is required to activate graft formation.

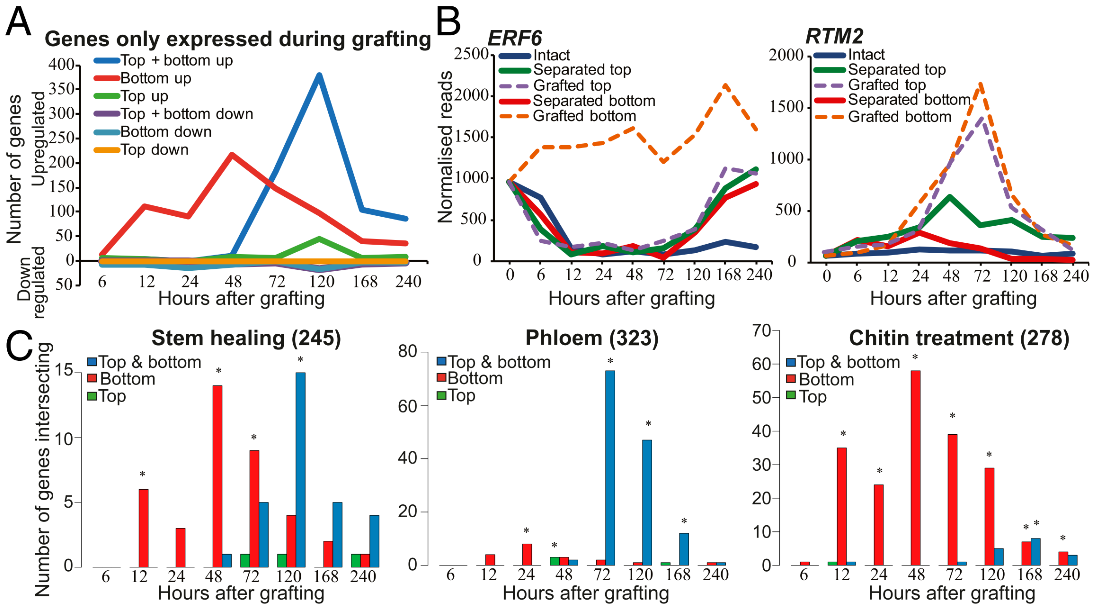


Figure 7.7 | A subset of genes is differentially expressed only during graft formation compared with intact or separated tissues. (A) Certain genes were only differentially expressed in grafted tops, grafted bottoms, or both in grafted tops and grafted bottoms. (B) Expression profiles for a graft bottom-specific ($ERF6$) or a graft top and bottom differentially expressed gene ($RTM2$) were plotted for intact, separated, and grafted samples. (C) Grafting-specific genes are also expressed in other processes such as stem healing, phloem reconnection, and treatment with chitin. Genes, the transcripts of which are associated with these biological processes, were taken from previously published datasets (Dataset S1) and compared with our dataset to assess transcriptional overlap with genes expressed in the grafted top, in the grafted bottom, or in both grafted top and grafted bottom. The number in parentheses represents the number of process-specific genes identified in previous datasets. An asterisk represents a significant overlap ($P < 0.05$). Reprinted Figure 7 from [96].

7.4 Discussion

To better understand how plants graft, we analyzed in depth an RNA deep-sequencing dataset that spatially and temporally distinguished genes activated by cutting followed by tissue attachment or continuous tissue separation. Cutting promoted a similar wound response in both grafted and separated tissues; however, by 72 hours after cutting, the grafted and separated tissues became transcriptionally dissimilar (Fig. 7.6C), indicating that tissue fusion was mechanistically different from healing an unattached cut surface.

During graft formation, tissues had a very high transcriptional overlap with genes differentially expressed by inflorescence stem healing and by vascular induction in leaves (Fig. 7.2 and Fig. S4 [96]) [105, 130], suggesting that grafting is closely related to these processes. Graft formation had little transcriptional overlap with lateral root formation (Fig. S2 [96]) [132] and appeared to follow a pathway similar to secondary root growth since the secondary growth-specific cambium markers $WOX4$ and PXY [141] were activated by grafting (Fig. 7.1 and Fig. S2 [96]).

Grafted tops initially showed a short-lasting and small transcriptional overlap with genes expressed during primary root formation, which may be related to the accumulation of substances activating adventitious root formation, a common response in failed grafts or in cut shoots (Fig. S6C [96]). Thus, we conclude that grafting likely proceeds via a pathway involving secondary growth with radial meristems activating in the mature cambium to heal the wound. Vascular formation genes including those specifying cambium and phloem were activated early, followed by an activation of cell division genes, suggesting that the start of cellular differentiation preceded activation of cell division. Xylem identity genes showed an early and a late activation peak (Fig. 7.1 and Fig. S2 [96]).

There is no visible xylem differentiation at the graft junction during the first peak of expression [100], and this expression could represent programmed cell death that does not lead to xylem differentiation. Alternatively, these genes might be suppressed by phloem differentiation genes that suppress protoxylem formation [142, 143]. The second expression peak of xylem-related genes at 120 hours occurred after the differentiation of functional phloem and coincided with the differentiation of xylem strands at the graft junction [100].

Previous studies highlighted the importance of callus and pericycle cells during regeneration [114, 144], but we see little evidence that genes expressed in the pericycle or during callus formation have high transcriptional overlap with genes differentially expressed by grafting (Fig. S4 [96]). Expression profiles for all protein-coding genes can be found in Datasets S4 and S5 of [96].

A high proportion of genes were initially asymmetrically expressed (Fig. 7.3A), and many had a delay in phloem, cambium, and cell division activation below the graft junction compared with above it (Fig. 7.1 and Fig. S2 [96]). Several genes associated with vascular formation, such as $HCA2$ [140] and $TMO6$ [145], activated equally in both grafted top and grafted bottom at 6 hours after grafting (Figs. 7.1 and 7.5 A–C). These data indicate that, at least transcriptionally, the grafted root rapidly responded to the presence of the grafted shoot and that this response was independent of functional vascular connections. This response was not present in separated roots, indicating that attachment was key for recognition and activation of graft formation (Fig. 7.5 and Fig. S9 [96]).

Sugars are known activators of cell division and cell elongation [109], and, in our datasets, a large proportion of genes asymmetrically expressed are also sugar-responsive (Table S1 [96]). However, sugars are transported in the phloem [108] that is severed upon grafting, and the grafted root exhibited a sugar-starvation response and showed similar sugar-response dynamics as the separated root. Instead, we infer that some other molecular that is transported in the absence of vascular connections activated $HCA2$ and $TMO6$ as well as cell division, phloem-, and cambium-related genes in the grafted bottom.

Given auxin's fundamental role in vascular formation [141], it is a strong candidate for an activating signal. Auxin response was largely symmetric from 12 hours after grafting (Fig. 7.4 and Fig. S7 [96]), consistent with previous findings that the auxin-inducible $DR5$, $IAA5$,and $ANAC071$ genes activate above and below the graft junction within 1–3 days of grafting [100–102, 146]. Furthermore, exogenous auxin application combined with cutting was sufficient to activate $HCA2$ expression in separated root hypocotyls (Fig. 7.5D).

One idea is that grafting caused an interruption in auxin transport, and, where opposing tissues adhered, auxin transport resumed regardless of vascular connections since auxin is transported from cell to cell through the apoplast [104]. The genes encoding the auxin efflux proteins $PIN1$ and $ABCB1$ were transcriptionally activated above the graft junction (Fig. 7.4D), similar to the putative *Pisum sativum* $PIN1$ protein accumulating above a cut stem before vascular reconnection [147], and could reflect a role for these proteins in exporting auxin across the cut. Consistent with these observations, adding an auxin transport inhibitor to grafted *Arabidopsis* shoots prevented the expression of grafting-induced genes below the graft junction [102].

Although auxin response was largely symmetric, our previous work demonstrated that the auxin signaling genes $ALF4$ and $AXR1$ are important for grafting only below the graft junction [100]. Mutating $ALF4$ below the graft junction more strongly reduced auxin response than mutating $ALF4$ above the junction [100]. Thus, proteins such as ALF4 or AXR1 might act by promoting auxin response and vascular regeneration below the graft junction, which could be particularly important when there is incomplete attachment, cellular damage, or inefficient transport. All higher plants transport auxin from shoot to root, yet not all plant species can be successfully grafted [99] so the response to auxin rather than the transport itself may be a determining factor in the ability to graft.

A role for sugars is not completely ruled out, however, since the magnitude of differential expression of vascular-related genes was often lower in the grafted bottom (Fig. S2 [96]). In addition, very low levels of exogenous sugars can improve graft formation under certain conditions [133]. Altogether, endogenous sugars likely enhance cell division and differentiation, perhaps similar to their role in enhancing the rate of pericycle cell divisions in the hypocotyl [148].

7.5 Conclusions and Outlook

The analyses identified two groups of genes, the expression changes of which were unique to graft formation in our experiments (Fig. 7.7). One group activated shortly after grafting below the graft junction and was enriched in immune-responsive and chitin-responsive genes (Fig. 7.7 and Fig. S9 and Dataset S3 of [96]). The breakdown products of cell walls are potent elicitors of defense responses [149], so it is possible that the grafted bottom up-regulates pathways specific to wound damage response. This group was not up-regulated in separated bottoms, however, so the unique physiological state of the grafted root, indicated by the presence of the auxin response but the absence of the sugar response, could have promoted their up-regulation.

The second group activated both above and below the graft junction and became highly expressed later during graft formation (Fig. 7.7). This group was enriched in RNAs associated with vascular development (Fig. 7.7 and Dataset S3 of [96]), and we suggest that the products of these genes are involved in the vascular reconnection processes between the two tissues. Despite many transcriptional similarities between separated tops and grafted tissues, tissues had to be attached for at least 3 days for phloem connections to form, regardless of when cutting occurred (Fig. S9 of [96]). Thus, it appears that RNAs expressed in the separated top or separated bottom are insufficient to drive graft formation. Instead, genes activated uniquely by grafting or genes involved in the recognition response are those that contribute to distinguishing attached from separated plant tissues.

Future work could focus on these genes to identify the pathways required for grafting that could be modified to improve graft formation, wound healing, and vascular regeneration. Likewise, the rapid transcriptional changes below the graft indicate a recognition system that promotes tissue regeneration. Identifying the cues that trigger recognition and understanding how they are perceived could be priorities, as could understanding whether this phenomenon applies more broadly to intertissue communication, tissue regeneration, or tissue fusion events, such as parasitic plant infections [150], epidermal fusions [151, 152], or petal fusions [153].

From a bioinformatics perspective, the detection of graft specific gene sets and their functional association could be improved by also modeling the temporal dependencies between the expressed genes. Additionally, the development of a biological network analysis to uncover the dependencies between the genes within such a gene set could help to deepen our understanding of graft specific gene expression patterns.

In this study, we have investigated only the expression patterns of protein-coding genes and their role in the expression dynamics during grafting. Besides protein-coding transcripts, the transcriptome landscape contains also non-coding transcripts and the interplay between these types of transcripts can provide us with information to better understand the biological process of interest. In chapter 8, we will learn about the variety of coding and non-coding transcripts in flowering plants and bioinformatics approaches to process RNA-Seq data with the attempt to improve the knowledge about protein-coding splice-variants, long non-coding RNAs, and circular RNAs in flowering plants.

8

Annotation of splice-variants, lncRNAs, and circRNAs in flowering plants

In chapters 2 - 7, we have investigated the evolutionary and developmental dynamics of plant transcriptomes based on protein-coding genes. In addition to protein-coding transcripts, there is a big fraction of non-coding transcripts not showing any evidence of a protein-coding potential. These non-coding RNAs have crucial regulatory roles ,e.g., in splicing or post-transcriptional regulation. In this chapter, we will present our workflow to predict and annotate novel protein-coding splice-variants, long non-coding RNAs and circular RNAs from RNA-Seq data of seven related flowering plants represented by samples from eight different organs.

In section 8.1, we will introduce the biological role of long non-coding RNAs and circular RNAs in flowering plants and we will introduce the goals of this chapter. In section 8.2, we will present the experimental setup, the amount of data, and the workflow to process RNA-Seq data and to annotate the different RNA species. In section 8.3, we will get to know the novel annotated RNA species and we will get an insight of their sequence features. In section 8.4, we will discuss the developed workflow and the annotated RNA species. In section 8.5, we will conclude and give an outlook to future work based on the novel annotations.

8.1 Introduction

Extensive deep sequencing studies extended our knowledge of long non-coding and circular RNAs. Long non-coding RNAs (lncRNAs) are a class of RNAs with a sequence length of at least 200 nucleotides and showing, in general, no evidence of protein-coding function. However, some lncRNAs may encode for small proteins [154], and it is speculated that lncRNAs may serve as a source of new proteins [155].

The first lncRNA, called *H19*, was discovered in the early 1990s [156] as an 'unusual gene' with high expression levels during mouse embryo development. In 1993 *Enod40* the first lncRNA in a plant species (soybean) was discovered [157], a highly conserved RNA participating in root symbiotic nodule organogenesis [158]. Since then, the range of regulatory functions mediated by lncRNAs has been studied and expanded, such as mediating gene silencing, mimicking miRNA targets, acting as a guide or scaffold RNA to relocate RNA binding proteins, to name a few [159].

With the rise of deep sequencing, trying to capture the present state of transcripts, extensive data sets have been analyzed trying to annotate and to identify the regulatory role of lncRNAs in different species, on tissue or organ level, especially in the animal kingdom [160, 161]. Besides, advancements in RNA-Seq and the improved analysis of sequencing data led in 2012 to the discovery of the class of circular RNAs [162]. This class of RNA molecules is characterized by a covalent and canonical linkage between the downstream 3' splice site and the 5' upstream splice site of a linear host RNA, i.e., pre-mRNA. Subsequent studies revealed the complex tissue- and stage-specific expression of circRNAs [163] and identified regulatory roles. One of the most prominent examples is the circRNA ciRS-7 (CDR1as) which contains around 70 conserved miRNA target sites. This circRNA acts as a kind of miRNA sponge through binding and suppressing the microRNA miR-7 [164, 165].

Based on the efforts of the last years, several RNA-Seq projects led to a steadily increasing amount of newly discovered circRNAs and lncRNAs in the animal kingdom. In contrast, the discovery and functional identification of long non-coding and circular RNAs in plants are very limited to a few model plants like *A. thaliana*. Initiatives like TAIR [166] and Araport [167] provide comprehensive annotations of the *A. thaliana* genome and provide us with a wealth of valuable information, but for a wide range of other plants, the genome-wide identification of long non-coding and circular RNAs are still at the beginning.

These plant annotations typically rely on one or a few organ types, while several studies in animals [160, 161] and plants [167] show that the expression of especially long non-coding and circular RNAs is strongly organ dependent, so we sequenced eight different organs from seven related plant species based on strand-specific total RNA-Seq experiments. We implement an annotation workflow enabling the prediction of plant-species and organ-specific, protein-coding transcripts, long non-coding, and circular RNAs. Based on the novel annotations, we update and refine the current protein-coding transcriptomes of the seven sequenced flowering plants. The resulting annotations can serve as a resource for protein-coding splice-variants (isoforms), lncRNAs, and circRNAs, providing insights into their genomic structure, their conservation, and potential function in flowering plants.

8.2 Materials and Methods

In this section, we will present the data for the RNA-Seq experiments and the workflow to annotate different RNA species in seven different flowering plants.

In subsection 8.2.1, we will introduce the different plant species and organ samples that were sequenced. In subsections 8.2.2 and 8.2.3, we will present the annotation workflow to create comprehensive assemblies. In subsection 8.2.4 and 8.2.5, we will learn about the calculation of protein-coding potentials and the determination of novel putative protein-coding loci and protein-coding isoforms. In subsections 8.2.6, we will turn our attention to the assembled transcripts without a sufficient protein-coding potential and their classification as long non-coding RNAs. In subsection 8.2.7, we will present the annotation of circRNAs based on backspliced-junctions. Finally, in subsection 8.2.8, we will present details about the implementation of the annotation workflow.

8.2.1 Data

In this study, we chose seven different plants with an evolutionary divergence ranging from 7.1 million years ago (MYA) up to 160 MYA as representatives of the flowering plants phylogeny 8.1 which are *A. thaliana*, *A. lyrata*, *C. rubella*, *E. salsugineum*, *Tarenaya hassleriana*, *Medicago truncatula*, and *Brachypodium distachyon*. All those flowering plants provided complete or partially complete genome sequences and annotations containing at least information about protein-coding genes.

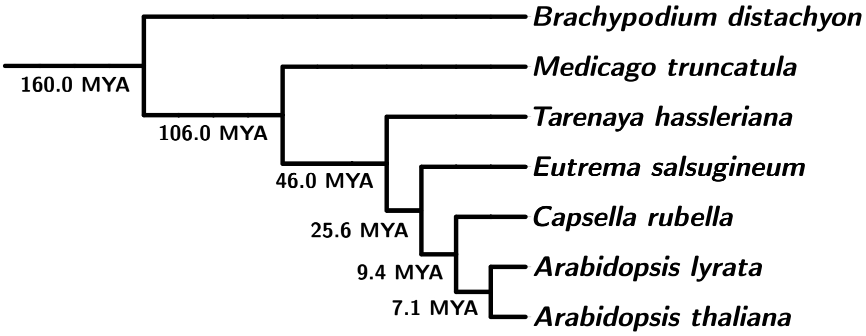


Figure 8.1 | Phylogeny of flowering plant species. The sampled plant species serve as representatives for flowering plants. The group of monocotyledons is represented by *B. distachyon*, while the other sampled plants represent the group of dicotyledons.

From each plant, we sampled and sequenced the organs root, hypocotyl, respectively mesocotyl, leaves, vegetative and inflorescence apex, flower, stamen, carpel, and, consisting of two cell types, the mature pollen. For each organism a similar developmental stage of each organ was sampled, allowing future studies a comparable data set for studying the conservation and evolution of flowering plants in an organ specific manner. Additionally to these organ and pollen samples, for *A. thaliana* several developmental stages of the above mentioned organs have been sequenced, as well as samples from sepals, petals, silliques, and seeds resulting in 43 organ and one pollen sample (Tab. 12.2). Each sample of each organism is represented by three biological replicates, resulting in 294 samples in total.

The sequencing libraries were prepared according to the Illumina Stranded Total RNA Preparation protocol. The resulting sequencing libraries were sequenced with 75bp paired-end reads on an Illumina HiSeq4000, except one lane of re-sequencing on Illumina NextSeq500. Sequencing of total RNA allowed capturing transcripts with poly(A) tails (poly(A^+)) and without poly(A) tails (poly(A^-)). This was essential for annotating novel lncRNAs transcribed by RNA polymerase IV and V [168–171] and for circular RNAs [164, 165, 172] lacking polyadenylated 3' ends. All samples were sequenced with additional ERCC RNA Spike-In control Mix 1 [173] allowing the definition of an objective expression cutoff to differentiate lowly expressed transcripts from transcriptional noise by relating the concentration of ERCC Spike-Ins to their subsequently estimated expression values. Hence, the expression values of ERCC Spike-ins with low concentrations served as a lower bound to define a particular transcript as expressed. The annotation of the seven plant species relied on the current genome sequence releases and the current protein-coding gene annotations providing essential information for the assembly and the prediction of novel coding and non-coding transcripts (Tab. 12).

8.2.2 Annotation workflow

The annotation workflow (Fig. 8.2) starts with the preprocessing of the raw sequenciong data. First, quality controls were performed by *fastqc* [174] to evaluate possible sequencing errors, contaminations during library preparation and/or the subsequent sequencing. Afterwards, remnants of bar coding adapter sequences were removed by cutadapt[175]. Read subsequences showing a poor sequencing quality were removed by a quality trimming performed by the sickle software [120]. The resulting high quality trimmed paired-end reads are now ready to be mapped against the reference genome sequence.

The basis for the subsequent genome-guided transcriptome assembly is the genomic mapping, which was performed by the RNA-Seq aligner STAR [176]. Because the study focuses on the prediction of novel protein-coding isoforms, also known as splice variants, and circRNAs the mapping algorithm aligned split reads. Those reads span an intronic region encapsulated by adjacent exons in a transcript. Hence, the read gets split, and the resulting segments are aligned to the neighboring 5' and 3' exons.

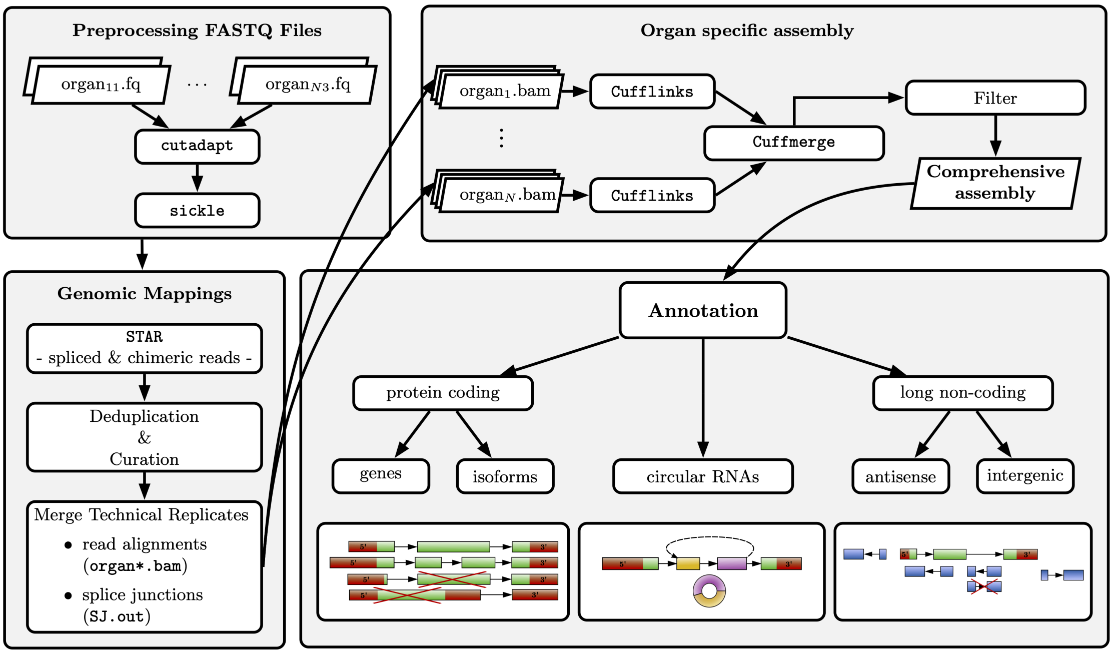


Figure 8.2 | Overview of the annotation workflow. The pipeline starts with the preprocessing, e.g., adapter clipping and quality trimming, of the raw data. Afterward, the mapper `STAR`[176] aligned the processed reads to the genome, followed by read-deduplication and merging of technical replicates. Based on the genomic mappings organ-specific transcriptome assemblies were performed with `Cufflinks` [177]. The organ-specific transcriptomes were merged to achieve a comprehensive assembly, and transcripts were removed due to specific filter criteria. This assembly served as the basis for predicting novel protein-coding loci and isoforms, circular RNAs, and long non-coding RNAs.

Additionally, chimeric reads were aligned for the subsequent prediction of circRNAs. Chimeric reads are a particular class of split reads. These reads align to at least two different locations of the genome, but these locations could be on different chromosomes or could have different strand orientations.

The detection of circRNAs is based on a subclass of chimeric reads that align at different portions in one locus, on the same strand but in reverse order (Fig. 8.6). In the following sections, we refer to these subclass of chimeric reads as back-spliced reads.

8.2.3 Read mapping and Transcript Assembly

Due to the sequencing of total RNA, we performed a duplication analysis after the mapping to detect overrepresented fragments, probably PCR artefacts. Especially, the leaf samples showed a very high rate of duplicated paired-end reads predominantly located in the chloroplast chromosome (Chr. Pt), which is shown in Fig. 8.3A for *A. thaliana*. To eliminate biases rising from the high duplication rates, we performed a deduplication step with *samtools* on each genomic mapping.

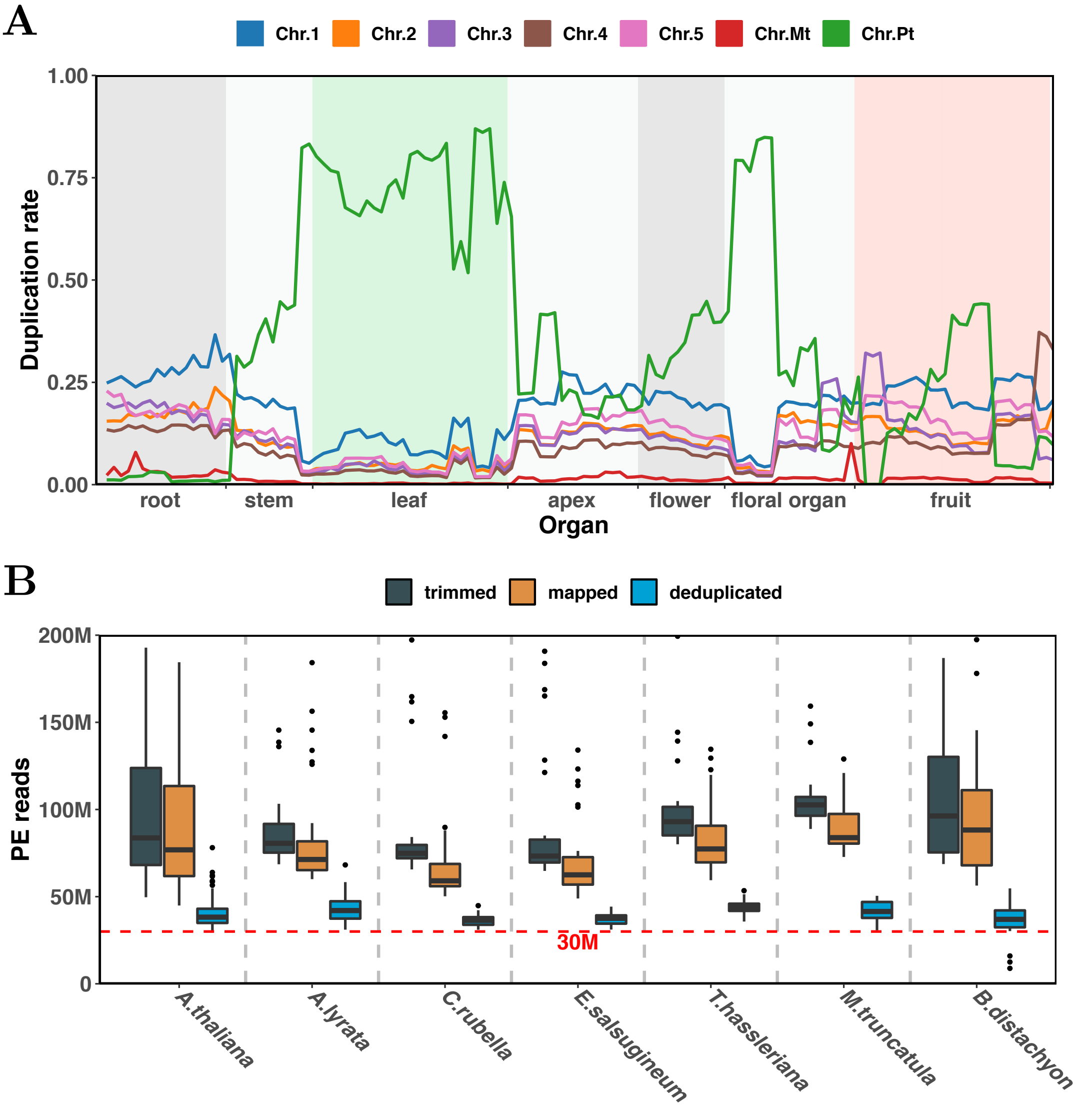


Figure 8.3 | Mapping statistics. (A) The distribution of the duplication rate along the organ samples of *A. thaliana*. The duplication rate was calculated as the fraction of duplicated reads in one chromosome divided by the sum of duplicated reads in all chromosomes. (B) The number of paired-end reads for each plant species after trimming (gray), mapping (yellow), and deduplication (blue). Almost all samples have a sequencing depth of 30 mio. mapped deduplicated paired-end reads, except three pollen samples in *B. distachyon*.

After removing duplicated reads, the majority of samples contained at least 30 million paired end uniquely mapped deduplicated reads. Only three biological replicates from *B. distachyon* pollen (Fig. 8.3B) contained less uniquely mapped deduplicated reads. Subsequently, we curated the mappings in order to fix mating read pairs and to remove soft-clipped subsequences of the reads produced by STAR.

In order to build comprehensive transcriptome annotations of the different plant species, we performed the same annotation workflow for each species and each organ. Cufflinks [177] performed the initial species- and organ-specific transcriptome assembly based on the deduplicated mapped reads for each biological replicates of each sample. Based on Cuffmerge, we summarized the species- and organ-specific assemblies from the three replicates into a species-and organ-specific transcriptome assembly. After this Cufflinks assembly and merge procedure, we produced for each plant species eight organ-specific transcriptome assemblies. We then filtered these species- and organ-specific assemblies by removing transcripts spanning multiple loci and transcripts showing an expression below a defined expression threshold. Transcripts having an expression below thus threshold were called transcriptional noise. We defined the expression threshold based on sampled-specific ERCC Spike-In threshold. We quantified the transcript abundance of each assembled transcript in each biological replicate with salmon [178]. We removed all transcripts which showed a TPM expression below the 5% quantile of expressed ERCC Spike-Ins in more than two out of three biological replicates in at least one sample.

After filtering the species- and organ-specific transcriptome annotations, the current reference annotation was combined with gffcompare [179] into a comprehensive species-specific transcriptome assembly. We filtered these species-specific transcriptome assemblies a second time to remove transcripts showing an insufficient transcript junction coverage. Based on GeMoMa [180], we calculated the splice junction coverages of each assembled transcript. Transcripts having a junction coverage below two split reads in more than two out of three biological replicates in all organ samples of a particular species, we removed this transcripts from the assembly.

Afterwards, we clustered the transcripts with CD-Hit [181] to remove almost identical transcripts. We set the parameter to create a cluster at above 95% identity and we defined that the longest transcript in each cluster is reported as its representative while the other transcripts were removed. Based on BLAST similarity searches [91] against the SILVA ribosomal RNA database [182], we also removed transcripts showing high similarity to known rRNAs. The result of this last filtering step is the comprehensive assembly (Fig. 8.2) which serves as the starting point for the annotation of novel protein-coding genes and isoforms, lncRNAs, and circRNAs. In the next sections, we will explain the classification of the assembled transcripts into these three RNA species in detail.

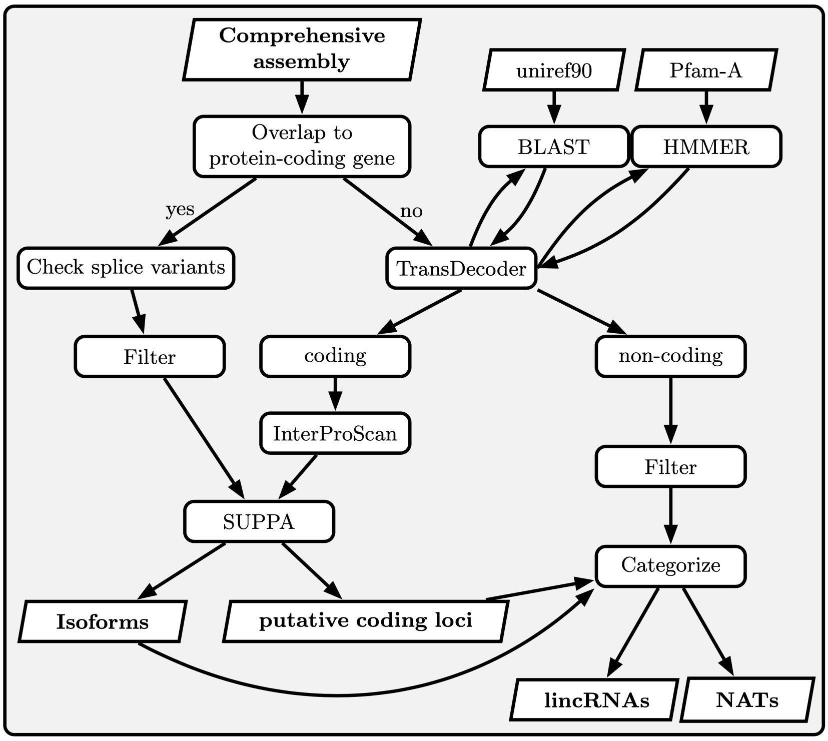


Figure 8.4 | Annotation of protein-coding and non-coding transcripts. Starting from the comprehensive assembly, each transcript is first assigned to known protein-coding transcripts or unknown transcripts. The known transcripts were filtered and used to update the known protein-coding transcriptome with novel isoforms or replace known isoforms. Unknown transcripts were assigned into potentially coding, i.e., probably putative protein-coding loci, and non-coding. The non-coding transcripts were filtered and categorized as intergenic or antisense lncRNAs (NATs).

8.2.4 Calculation of Coding Potentials

The distinction of assembled transcripts into protein-coding and non-coding is a crucial step in the annotation pipeline. First, we assigned the assembled transcripts into three groups. The first group contained all transcripts which overlap with known protein-coding loci. The second group contained all transcripts which showed no overlap to current protein-coding gene models. For these unknown transcripts, we calculated the coding potential to differentiate between candidate transcripts for novel putative protein-coding loci or lncRNAs. We used TransDecoder [183] in combination with BLASTP searches [91] against plant proteins from the Uniref90 database [184] and HMMER searches [185] against the Pfam-A database[186] to calculate the coding potential of each transcript.

TransDecoder algorithm As described by Haas et al. [183], the algorithm started by identifying all open reading frames (ORFs) with a minimum length of 100 amino acids for all provided transcripts. Relying on the strand information of assembled transcripts, TransDecoder searches for ORFs in three instead of six possible reading frames. Each predicted ORF must begin with the start codon (ATG) and must terminate with a stop codon (TAA|TAG|TGA). If TransDecoder predicted more than one ORF, it selected the longest ORF for the subsequent analysis. For all transcripts with ORF lengths above 100 amino acids, TransDecoder calculated a score S_{n}^F as the sum of log likelihood ratios over all positions in the putative coding sequence. The algorithm calculated a single log likelihood ratio $S_{n\ell}^F$ for the reading frame $F \in \{1, 2, 3\}$ in the transcript sequence n at each position ℓ as

$$S_{n\ell}^F = \mathtt{log}\frac{p(x_{n\ell}|x_{n\ell-1}\dots x_{n\ell-5}, \theta_F)}{p(x_{n\ell}|\theta_{bg})} \tag{8.1}$$

,with $\ell \in \{1, \dots, L_n\}$ and L_n denoting the length of the sequence. The numerator $p(x_{n\ell}|x_{n\ell-1}\dots x_{n\ell-5}, \theta_F)$ (8.1) defines the probability of nucleotide $x_{n\ell}$ in a reading frame specific 5^{th}-order Markov model. θ_F defines the parameter set in the reading frame F that was trained on coding sequences from the reference annotation. The denominator $p(x_{n\ell}|\theta_{bg})$ denotes the relative frequency of nucleotide $x_{n\ell}$ which was trained on the set of coding sequences and the set of assembled transcripts. Following equation 8.1, TransDecoder calculated a score for each reading frame. TransDecoder also scored each ORF based on its putative reading frame and with respect to alternative reading frames. If the calculated ORF score based on the putative reading frame of a transcript was positive and greater than the ORF scores on the alternative reading frames than TransDecoder classifies the transcript containing this ORF as coding.

Similarity searches In addition to the calculation of coding potentials based on an Markov model approach, the TransDecoder pipeline defined also transcripts as potentially coding if they showed significant similarities to protein sequences. TransDecoder translated the longest ORF of each assembled transcript into a amino acid sequence and performed a protein BLAST search against a plant specific UniRef90 database and a HMMER search against the Pfam-A. The pipeline classified transcripts as coding, if their amino acid sequence showed significant similarities to known protein sequences or protein domains in the UniRef90 or Pfam-A database.

We considered transcripts having a coding potential and no overlap to known protein-coding loci as transcripts from potentially novel putative coding loci which we furhter analyzed as shown in Fig. 8.4. On the other side, we considered transcripts showing neither in the Markov model approach nor in the similarity searches any potential of being coding as non-coding transcripts and we used additional filters to classify them as long non-coding RNA Fig. 8.4.

8.2.5 Filtering novel putative protein-coding loci and isoforms

To ensure that we did not classify transposable elements as protein-coding locis, we removed novel putative protein-coding transcripts overlapping with de novo annotated transposable elements. Afterwards, we performed for each of the remaining putative coding transcripts an InterProScan [187] against a variety of protein and domain sequence databases, e.g. CDD, Pfam, TIGRFAM, ProDom, to detect functional associations based on sequence comparisons. The InterProScan 5 algorithm used different search algorithms, e.g. BLAST and HMMER. We only considered transcripts showing a significant similarities to sequences in the provided databases to be defined as putative coding loci.

Besides the classification of novel protein-coding loci, we also wanted to identify novel splice-variants (isoforms) of known protein-coding genes. It has been shown that assembling RNA-Seq short reads into full transcripts is a complex task [177, 183, 188]. Since transcripts show a variable read coverage in the genomic mappings and transcript isoforms share exon regions, it is challenging to assign the reads to their origin transcript and provide an unambiguous transcript assembly. As a result of these computational challenges, today there is no algorithm which does not predict falsely assembled transcripts. The generated assemblies have to be filtered by other criteria, e.g. customized with respect to the experimental design such as the library preparation. Because we are interested in the prediction of novel splice variants, i.e., transcript isoforms, of known protein-coding genes, we benefited from the current reference annotations.

For each potential new isoform, we calculated a pairwise global alignment [189] between the assembled transcript and each isoform of the reference annotation at the overlapping gene locus. Subsequently, we predicted ORFs within the assembled transcripts to ensure that the transcripts are not truncated. We considered an assembled transcript as a new protein-coding transcript isoform only if it showed a minimum of 50% sequence identity with at least one reference isoform and if it also shared the same start codon position with its aligned reference isoform. Otherwise, we considered the assembled transcript as incorrectly assembled and removed it from the assembly. The threshold of 50% sequence similarity is arbitrary and a compromise between predicting too many falsely assembled transcripts and removing too many probably functional splice-variants.

Subsequently, we filtered transcript isoforms from novel putative protein-coding loci and novel transcript isoforms based on intron retention events. Since the RNA extraction protocol is intended to capture RNA species regardless of polyadenylation, we probably sequenced unprocessed fragments of immature pre-mRNAs. It has been shown that total RNA sequencing could result in a higher proportion of RNAs originated from intronic sequences [190, 191]. We assumed that the assembled isoforms originated from intron retentions might represent immature pre-mRNAs rather than functional mRNAs. To classify intron retention events, we use SUPPA [192] and removed all assembled transcripts showing retended introns.

8.2.6 Prediction of long non-coding RNAs

LncRNAs are defined as transcripts with a length greater than 200nt, having no capability for coding functional proteins [154, 193–201]. The annotation of lncRNAs mainly based on a length selection, the prediction of a coding potential, and its genomic location with respect to protein-coding genes. With the ongoing effort to study the non-coding transcriptome based on RNA-Seq, several algorithms trying to distinguish between protein-coding and non-coding transcripts were published [183, 202–208].

Their main goal can be summarized in calculating the coding potential of a transcript based on sequence features, such as nucleotide compositions and the length of their ORFs, or evolutionary features, such as codon substitution frequencies. Additionally, most of these algorithms need a set of known non-coding and coding RNAs to train their models for a reliable prediction. Especially in non-model organisms as well as in many plant species, the information about lncRNAs is insufficient or restricted to a few prominent species like *A. thaliana*. In our study, for the majority of species, we had no or insufficient information about lncRNAs.

Hence, the prediction tool that we wanted to use, had to be capable to be trained on species-specific coding transcripts or is already trained based on a generalized approach and thus capable to predict lncRNAs in species with missing information about lncRNAs. As described in the previous section, we could use the non-coding transcript classified by TransDecoder as potential lncRNAs. The advantage of this procedure was that we were capable to train the algorithm just on the known protein-coding transcripts in the reference annotations. To validate our procedure, we compared the TransDecoder algorithm with currently popular lncRNA prediction tools such as FEELnc [206] and CPC2 [205]. FEELnc can be trained only on known species-specific coding transcripts. In contrast, CPC2 [205] is already trained on set of non-coding and protein-coding transcripts. We compared the accuracy of the non-coding prediction between FEELnc, CPC2, and TransDecoder based on a training dataset [205]. In Fig. 8.5B we show, that CPC2 and TranscDecoder had a prediction accuracy of $\sim$ 97% for *A. thaliana* while

FEELnc showed an accuracy of ∼ 94%, confirming our choice of using TransDecoder as a lncRNA prediction tool.

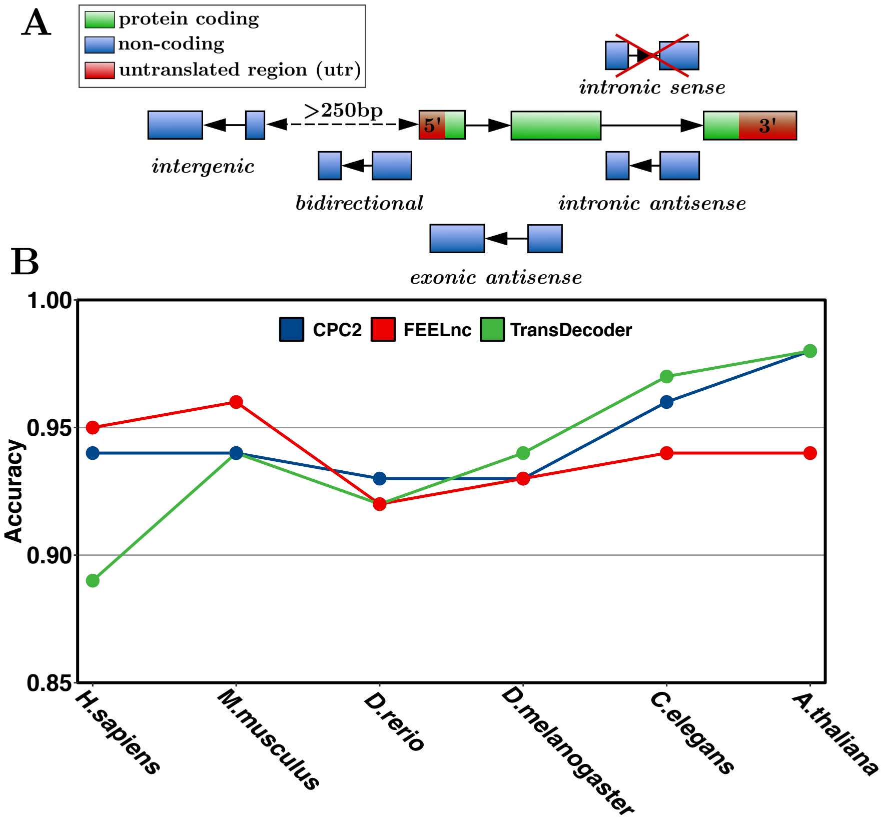


Figure 8.5 | Classification of lncRNAs. (A) Terminology of lncRNAs (blue) based on their genomic location to a protein-coding gene (green and red). Based on the underlying sequencing data we are able to predict intergenic, bidirectional, exonic antisense, and intronic antisense lncRNAs. We were not able to reliable predict intronic sense lncRNAs due to the RNA-Seq library preparation of total RNA leading to a bias of immature spliced transcripts resulting in an increase of probably falsely discovered intronic sense lncRNAs. (B) Comparison of coding potential calculation tools and their accuracy in predicting lncRNAs. Especially for plant lncRNAs TransDecoder and CPC2 showed the highest accuracy of ∼97% outperforming FEELnc.

After we have classified the assembled non-coding RNAs as lncRNAs if they showed a sequence length greater than 200 bp, we categorised the lncRNAs into subgroups based on their genomic location. Relative to their location to protein-coding genes, lncRNAs can be classified into subgroups of bidirectional, intergenic (lincRNA), exonic antisense (NAT), and intronic antisense (intronic NAT) lncRNAs (Fig. 8.5A). The definition of lncRNA subgroups in Fig. 8.5A is mostly based on observations in humans and other vertebrates [209], in which bidirectional promoters activate the transcription of two neighboring loci, facing each other's 5' ends. Distinguishing between intergenic and bidirectional lncRNAs we initially chose an intergenic distance of 250bp. Usually, this distance ranges between 500 bp and up to 5 kbp [160, 201, 210], but the plant species in our study had much smaller and more compressed genomes compared to humans and other vertebrates.

Studies in *A. thaliana* seedlings showed that plants seem to lack significant bidirectional transcription [211]. Despite the possibility of classifying lncRNAs as bidirectional, we assigned lncRNAs in the proximity of protein-coding genes to the group of lincRNAs if they do not overlap with any exon of a protein-coding gene, otherwise we assigned them into the group of NATs. As described in the previous sections, the RNA extraction protocol of total RNA over represents reads in intronic regions. We could not ensure that transcripts assembled and encapsulated in intronic regions are of biological meaning or related to the sequencing experiment. Consequently, we removed transcripts entirely overlapping intronic regions of a protein-coding loci on the same DNA strand. We, therefore, were not capable of annotating intronic sense lncRNAs. Additionally, we removed non-coding transcripts overlapping with annotated transposable elements.

8.2.7 Prediction of circular RNAs

CircRNAs are a class of RNA molecules originated from a circularization event, called backsplicing, showing a covalent and canonical linkage between an upstream 5' splice site and a downstream 3' splice site in a linear immature RNA, like pre-mRNAs [162, 165, 212]. Additionally, circRNAs are distinguished from mRNAs by missing a polyadenylated 3' end. Most circRNAs were found within annotated exon boundaries or at locations having canonical splice signals, the necessary binding motif for the spliceosome [212].

Additionally, circRNAs also appear in lncRNAs or in intergenic regions [209, 213, 214], which we might also detect in the final species-specific annotations. Before we have attempted to detect circRNAs, we updated the reference annotation with the predicted putative coding loci, the novel transcript isoforms, and the newly predicted lncRNAs.

In recent years several circRNA detection algorithms has been published. A detailed list of detection tools and algorithms can be found in the reviews of Szabo et al. [212] and Gao et al. [215]. All these algorithms use the information of back-spliced reads to detect circRNAs (Fig. 8.6A). Their algorithms depend in the first instance on the performance of the mapping tool and its capability of mapping back-spliced reads to the underlying genome [216].

Before integrating a specific circRNA prediction tool into our workflow, we performed an evaluation of the most common circRNA prediction tools such as CIRI2 [217], circExplorer2 [218], and DCC [216]. They could be easily integrated in our workflow and were capable of using the advantage of paired-end read information.

These tools had already been evaluated in various studies, but mainly on human or other animal data sets [216, 219].

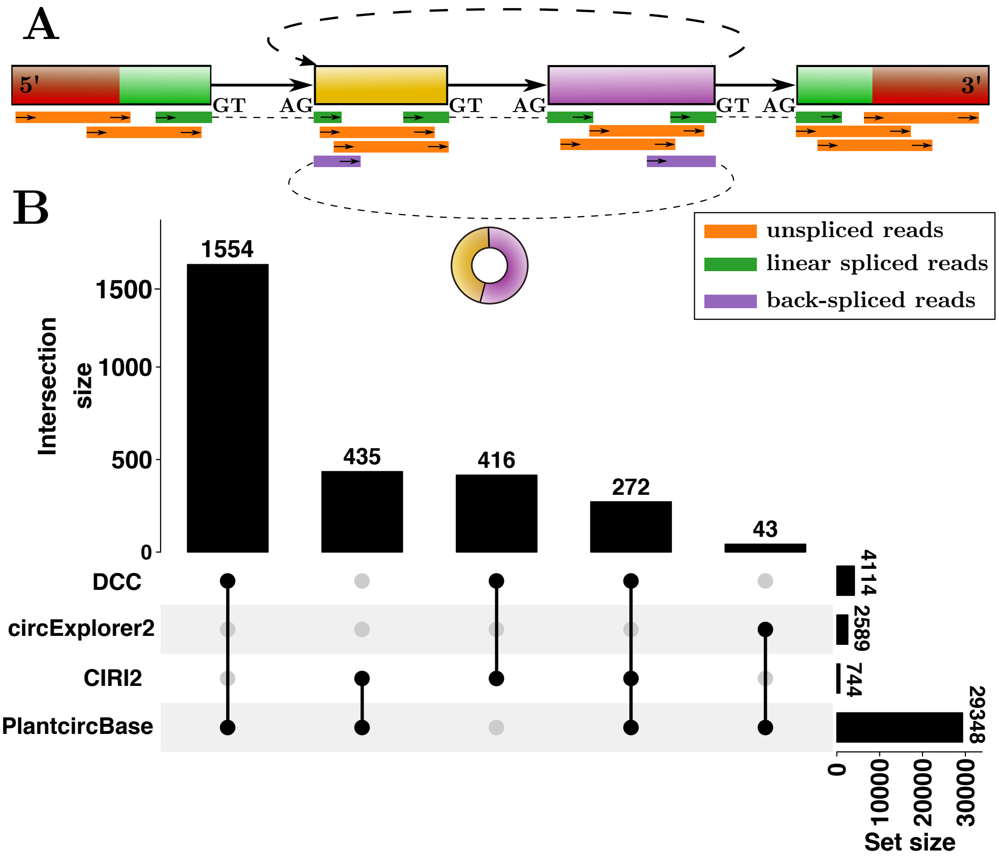


Figure 8.6 | Annotation of circular RNAs in *A. thaliana*. (A) Prediction of circular RNAs within protein-coding genes of *A. thaliana*. Based on back-splice reads (violet) prediction tools are able to annotate circRNAs. In our analysis at least two biological replicates should provide evidence of a circRNA based on back-spliced reads. (B) Comparison of cirCRNA prediction tools. DCC predicts in total not only the most circRNAs but also the most circRNAs which overlap with known circRNAs in the PlantcircBase.

Our benchmark analysis focused on the detection of circRNAs based on our 132 deep sequencing libraries of *A. thaliana* covering ten organs within different developmental stages. We compared our prediction results with already published and annotated circRNAs from the PlantcircBase database [220]. This database currently provides one of the biggest database for predicted circRNAs from different plant species. For *A. thaliana* the database contained 38,938 circRNAs (Version 4, 2019). We filtered these circRNAs with respect of available strand information and used for our benchmark analysis 29,348 published circRNAs. The upset plot in Fig. 8.6B shows that DCC predicted the most circRNAs (4,114) of which 1,554 could also be found in the PlantcircBase. The amount of predicted circRNAs by CIRI2 and circExplorer was much lower and also their overlap with published circRNAs was significant small. Despite the low absolute number of predicted circRNAs by CIRI2, the relative overlap to published circRNAs is the highest ($\sim$ 59%) compared to DCC ($\sim$ 38%) and circExplorer2 ($\sim$ 2%).

Based on the total amount of detected circRNAs and the amount of circRNAs concordant with PlantcircBase, we integrated DCC into the annotation workflow. According to Cheng et al. [216], we performed the circRNA prediction with DCC whereas a circRNA was considered as a candidate only if DCC found evidence in at least two biological replicates supported by at least one back-spliced read.

8.2.8 Implementation

Analyzing the massive amount of sequencing data in a reproducible manner was one of our main goals for implementing the annotation workflow. The results should also be reproducible independent of the underlying computing platform, and it should be possible to perform the analyses on the same software releases. We structured the analyses by using the Snakemake workflow engine [221] in combination with Anaconda [222]. Snakemake allowed us to upscale the calculation to integrate the workflow into our computing cluster environment easily. Additionally, it allows other users with no access to computing clusters or a different computing platform to use our workflow without further changes in the code.

We performed most of our analyses with customized software solutions. We chose the most appropriate programming language for each analysis steps and used Java, C++, R, Python, or Perl for different tasks. The Snakemake workflow and the customized software will be publicly available via GitHub. Based on GitHub, we will be able to forward improvements of the workflow. We did not explicitly tailor the workflow to the plant species in our studies, and it can be easily adapted to annotate other non-model plant species based on RNA-Seq data.

8.3 Results

In this section, we learn about the updated annotations of the seven flowering plants and we will compare several genomic features between protein-coding transcripts, lncRNAs, and circRNAs.

In subsection 8.3.1, we will revisit the annotation workflow and present an overview of the annotated RNA species. In subsections 8.3.2, we will investigate the genomic features of protein-coding transcripts on a loci and isoform level. In subsection 8.3.3, we will present the detected lncRNAs and compare several genomic features with protein-coding transcripts and circRNAs. In subsection 8.3.4, we will investigate the splicing patterns of detected circRNAs. Subsection 8.3.5, will serve as an example for future comparative studies. We will compare the GC contents of the different RNA species between the seven flowering plants.

8.3.1 Annotation workflow

The annotation workflow (Sec. 8.2) enables a species-specific comparison of protein-coding transcripts, long non-coding RNAs, and circular RNAs based on the sequencing experiments from nine different organs of *A. thaliana*, *A. lyrata*, *C. rubella*, *E. salsugineum*, *T. hassleriana*, *M. truncatula*, and *B. distachyon*. Since the workflow utilizes developmental, here organ-specific RNA-Seq data to create comprehensive species-specific annotations, we refer to this workflow as the DevSeq workflow and the corresponding annotations as DevSeq annotations. The workflow reconstructed full-length transcript models of the three RNA species independently for each species. The annotation of each RNA species provides various challenges like the correct reconstruction of complete RNA transcripts based on short-read RNA-Seq data. The main goals were differentiating between coding and non-coding transcripts, predict novel protein-coding loci, predicting circRNAs, and finally create a comprehensive annotation for each plant species.

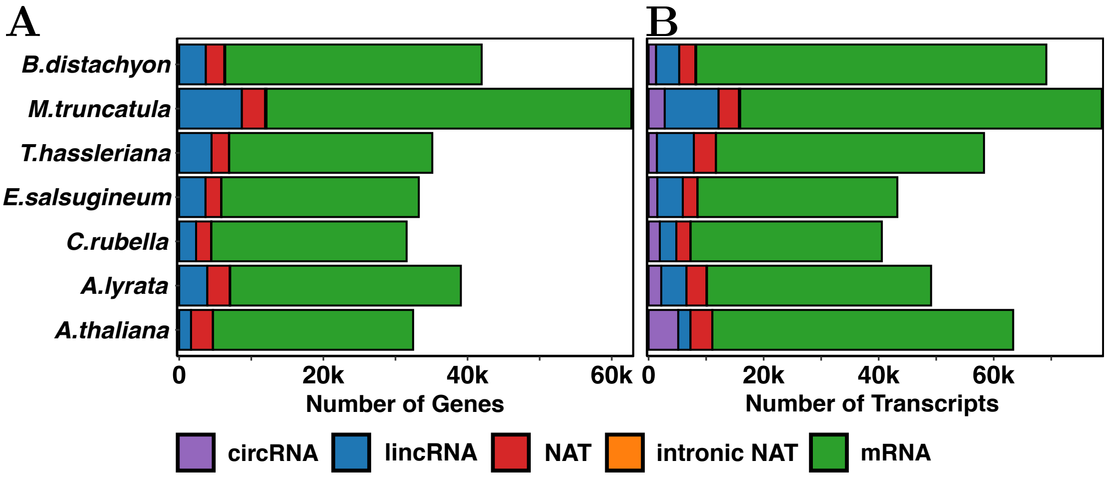


Figure 8.7 | Amount of coding and non-coding genes and transcripts in flowering plants. Absolute amount of annotated coding and non-coding (A) loci (genes) (B) transcript isoforms for circRNAs (purple), lincRNAs (blue), NATs (red), intronic NATs (orange), and protein-coding mRNAs (green). The species on the y-axis are arranged according to the phylogenetic tree. Starting from bottom to top from Brassicacae (dicots) to *B. distachyon* representing grass plants (monocots).

To increase the range of predicting novel non-coding transcripts such as non-coding transcripts without polyA-tails, we sequenced total RNA libraries with depleted rRNAs. After quality checks and the removal of sequencing data showing poor quality, we performed a genomic mapping of the sequenced reads.

Due to the rRNA depletion in the total RNA libraries the amount of reads from chloroplast RNA was up to 87% in green organ samples like leafs or floral organs (Fig. 8.3A). These highly similar reads might originate from PCR artefacts. Those artefacts can skew the mate-pair statistics which several transcriptome assembler use [223] and the assembly algorithm could potentially reconstruct false transcript structures [224]. Sequencing mostly these enriched duplicated fragments of transcripts led to extremely low coverages in less expressed regions. To increase the coverage in these regions, we increased the sequencing depth in samples with high duplication rates and performed a deduplication step after the initial read mapping (Sec. 8.2.2).

Transcript isoforms and lncRNAs are expressed in a organ-specific manner, thus we independently reconstructed the transcripts for each organ. We merged the organ-specific transcriptome assemblies to a comprehensive assembly representing the expressed transcripts found in the sequenced organs of one plant species. Based on that assembly and the genomic mapping we classified the assembled transcripts into protein-coding, long non-coding, and circular RNAs. Figs. 8.7A and B show the final amount of annotated genes and transcripts of the seven species.

To measure the annotation completeness and to quantify possible improvements of the annotations, we performed homologous searches against the Benchmarking Universal Single-Copy Orthologs (BUSCO) [225]. We compared the transcriptomes of each reference annotation and the updated DevSeq annotation against the curated set of embryophyta specific single-copy orthologs, which ideally should be present in all analyzed plant species. The analyses showed that we identified between 96.4% and 99.5% of the curated BUSCO sequences (Tab. 8.1). For the majority of species, except *E. salsugineum* and *M. truncatula*, the DevSeq annotation showed an increase in the percentage of the annotation completeness.

Species	Ref (%)	DevSeq (%)
A. thaliana	1605 (99.4)	1606 (99.5)
A. lyrata	1583 (98.1)	1596 (98.9)
C. rubella	1559 (96.6)	1580 (97.9)
E. salsugineum	1588 (98.4)	1593 (98.7)
T. hassleriana	1600 (99.1)	1600 (99.1)
M. truncatula	1557 (96.5)	1556 (96.4)
B. distachyon	1595 (98.8)	1596 (98.9)

Table 8.1 | BUSCO evaluation of annotation completeness. Absolute number of identified embryophyta specific single copy orthologs in BUSCO. In brackets is shown the percentage of identified orthologs.

8.3.2 Protein-coding genes and isoforms

Based on their overlap with known protein-coding loci and their coding potential (Sec. 8.2.4), we identified novel protein-coding isoforms and novel putative protein-coding loci. The amount of novel putative protein-coding loci ranges between 126 (*A. thaliana*) and 1,256 (*B. distachyon*) genes (Tab. 8.2). These low numbers reflect the high quality of reference annotations used in our study and the efforts of previous studies trying to decipher a complete annotation of protein-coding genes.

Besides these novel loci, we predicted and we updated thousands of known protein-coding transcripts (Tab. 8.3). The deep sequencing libraries of ∼30 mio. paired-end reads and the variety of different sequenced organs, allowed us to detect lowly and organ-specific expressed transcripts. We could verify already annotated isoforms but we also elongated 5' and/or 3' ends of known transcript isoforms (Tab. 8.3). Additionally, the workflow discovered between ∼2,800 (*T. hassleriana*) and over 6,000 (*B. distachyon*) novel transcript isoforms.

Species	**known loci**	**new loci**	**transcripts**
A. thaliana	34,806	126	148
A. lyrata	31,073	915	1,153
C. rubella	26,521	557	711
E. salsugineum	26,351	1,025	1,344
T. hassleriana	27,396	771	846
M. truncatula	50,444	162	172
B. distachyon	34,310	1,256	1,499

Table 8.2 | Novel putative protein-coding loci. Comparison of known protein-coding loci in the reference annotation and novel putative protein-coding loci and their transcript isoforms identified by the DevSeq workflow.

Updating the known set of transcripts led to a shift in the amount of splicing events in protein-coding isoforms. As shown in Fig. 8.8 the fraction of alternative 3' ends slightly increased in most organisms, except for *A. lyrata* and *E. salsugineum*. It is the main splicing event in the reference and also in the DevSeq annotations. The second largest fractions of splicing events are alternative 5' ends and intron retention events. The amount of alternative 5' ends increased in all species, which could be an effect of the random priming during the RNA-Seq library preparation.

It has been shown that random priming leads to more uniformly distributed reads in the genomic mapping but also introducing coverage biases at the 5' ends [226]. Despite, we saw in all annotations a decrease in the fraction of retended introns.

organism	known	alt 3’	alt 5’	alt 3’&5’	new
A. thaliana	34,806	4,852	2,799	5,875	3,818
A. lyrata	18,972	1,635	4,071	8,454	4,680
C. rubella	16,824	908	1,852	8,863	4,080
E. salsugineum	15,580	418	3,035	10,251	4,099
T. hassleriana	27,423	2,280	3,577	9,667	2,805
M. truncatula	46,842	1,364	2,077	7,302	5,112
B. distachyon	42,146	2,856	1,877	6,093	6,405

Table 8.3 | Statistics of protein-coding isoforms. The column “known” and “new” show the numbers of protein-coding isoforms in the reference annotation and predicted by our workflow. Besides the prediction of novel isoforms, we update known isoforms by elongating their 3’ and/or 5’ ends.

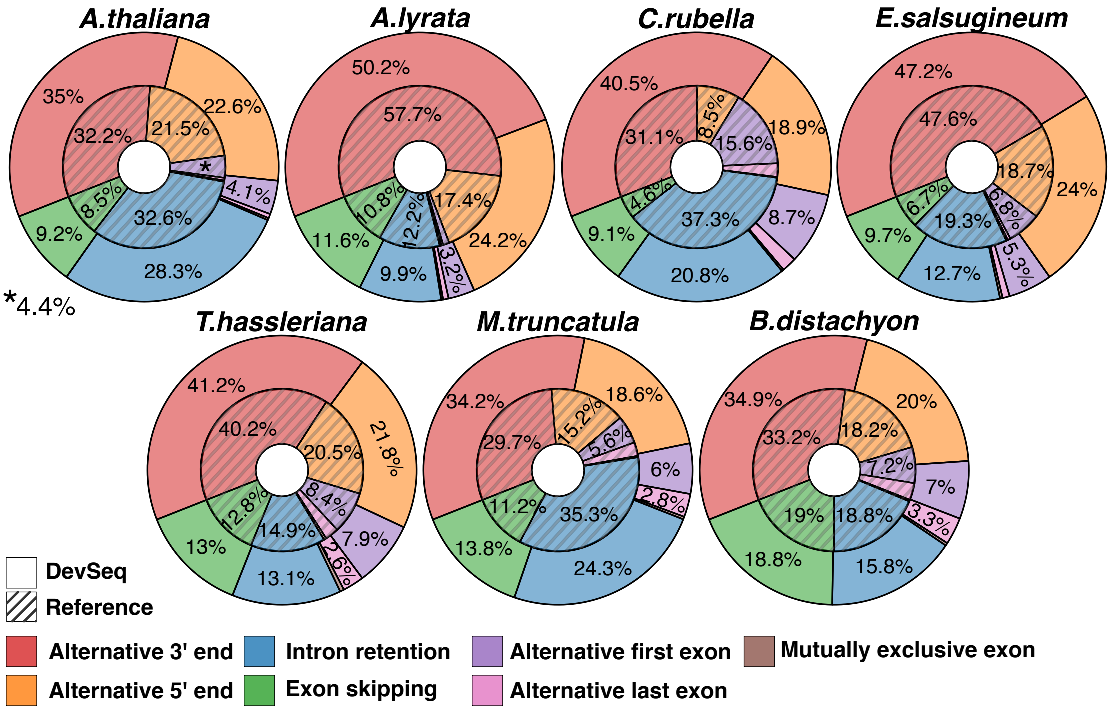


Figure 8.8 | Comparison of splicing events. The outer circles represent the fraction of splicing events in the generated DevSeq annotation whereas the striped inner circles represent the proportions of splicing events observed in the reference annotation. The fractions of alternative 3’ and 5’ ends are the splicing events in the DevSeq annotation while in the reference annotations the alternative 3’ end and the intron retention are the events covering together over 50% of the observed splicing events. The shift between the fractions is mainly a result of the library preparation and also the filtering steps during the prediction of assembled transcripts.

Due to the experimental design of total RNA sequencing, we could not distinguish between retended intron splicing events and read-through artefacts. We, therefore, removed all novel protein-coding isoforms which showed evidence of being originated from retended introns. In contrast, isoforms from the reference annotation showing intron retentions but were already published in the reference annotations are integrated into the DevSeq annotation. As shown in Fig. 8.8, the alternative 3' and 5' end represent largest fractions of novel splice variants. To verify that our annotation has not introduced a bias to the transcriptome annotations, we compared the lengths of the observed 5' and 3' untranslated regions (UTRs). The comparison of UTRs (Fig. 12.1) showed that in all species the 5' UTRs were shorter than the 3' UTRs, which is well known for eucaryotic mRNAs [227]. Comparing the UTR lengths between the species, we saw in the majority of plant species a similar distribution of UTR lengths. Only *M. truncatula* and *B. distachyon* showed an increased length in 3' UTR sequences.

8.3.3 Long non-coding RNAs

The second major group of assembled transcripts are the lncRNAs. As described in the Methods section 8.2.6, we classified the assembled transcripts without any overlap to known protein-coding genes on the same strand as potential candidates for these lncRNAs. After performing several filter steps, we identify thousands of lncRNAs (Fig. 8.7).

LncRNAs as well as protein-coding genes are transcribed as different transcript isoforms. In Fig. 8.9A, we investigated the mean number of isoforms from the three lncRNA classes compared to protein-coding genes (mRNA). Except for *T. hassleriana*, lncRNA loci transcribe on average for 1.00 - 1.25 transcript isoforms. This low amount of transcripts per lncRNA gene was also reported by other studies [228–232]. In contrast to these studies we observed in *A. lyrata*, *C. rubella*, *E. salsugineum*, and *M. truncatula* almost the same low amount of transcript isoforms for protein-coding genes. Only in *A. thaliana* and *B. distachyon*, we saw an increase of transcript isoforms in protein-coding genes compared to lncRNA loci.

The median transcript length was ∼600 bp which is significantly shorter than for protein-coding genes with ∼2 kbp (Fig. 8.10A). Only NATs showed the largest transcript lengths followed by lincRNAs, intronic NATs and circRNAs (Fig. 8.10A).

The pattern of transcript length distributions of the presented RNA species was consistent in all plant species with a small exceptions regarding the transcript length of protein-coding transcripts from *M. truncatula*. The median transcript length of protein-coding transcripts was similar compared to the other plant species but the IQR was much longer and ranged over 10 kbp (Fig. 8.10A).

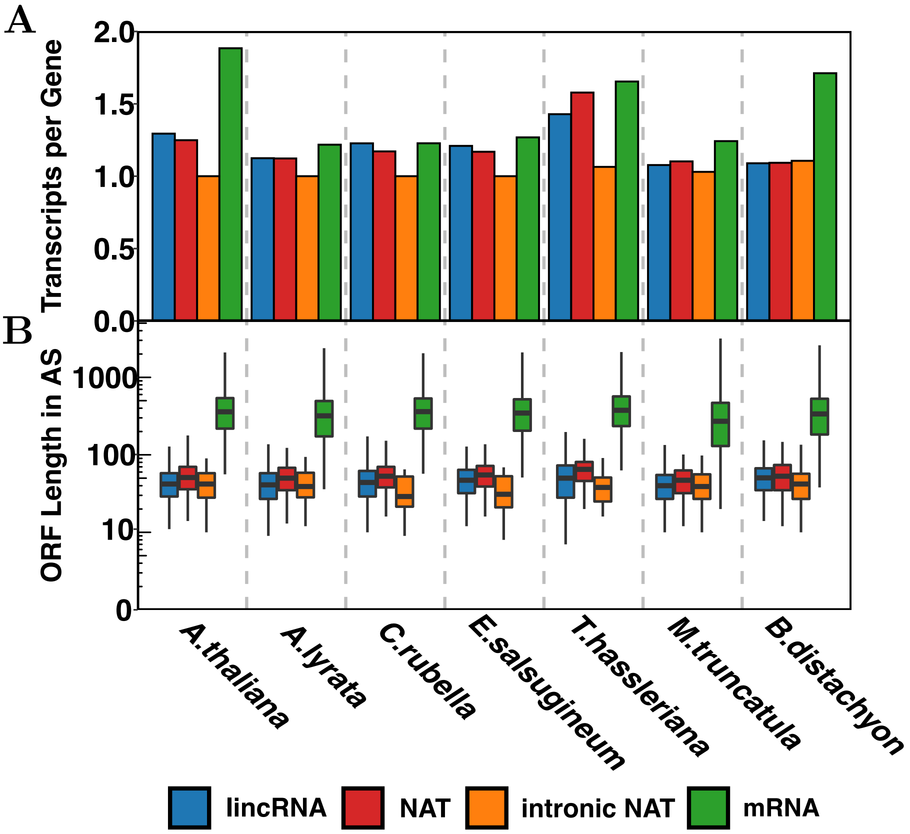


Figure 8.9 | Transcripts per gene and ORF length of lncRNAs and protein-coding mRNA transcripts. (A) shows the mean number of transcripts per lincRNA (blue), NAT (red), intronicNAT (orange), and protein-coding mRNA (green). (B) shows the corresponding ORFs for each RNA species in each organism with the same color code as (A).

However, also the exon length distributions were similar for the majority of species except for *T. hassleriana* showing a very broad length distribution for lincRNA and NAT exon sequences (Fig. 8.10B). Regarding the distribution of exon lengths (Fig. 8.10B) and the mean number of exons per transcript (Fig. 8.10C) over the different RNA species, lncRNAs seemed to have long exons with a median of ∼500 bp, but contained on average only 1.5 exons per transcript. In *C. rubella* and *T. hassleriana* the number of exons per transcript was ∼3 and also the exon length distributions of lncRNAs were similar to protein-coding genes. This similarity might reflect errors in the prediction pipeline, which is in contrast to the ORF lengths shown in Fig. 8.7B. On the other side, this observation could reflect the potential of lncRNAs to serve as sources for novel peptides as proposed by [155].

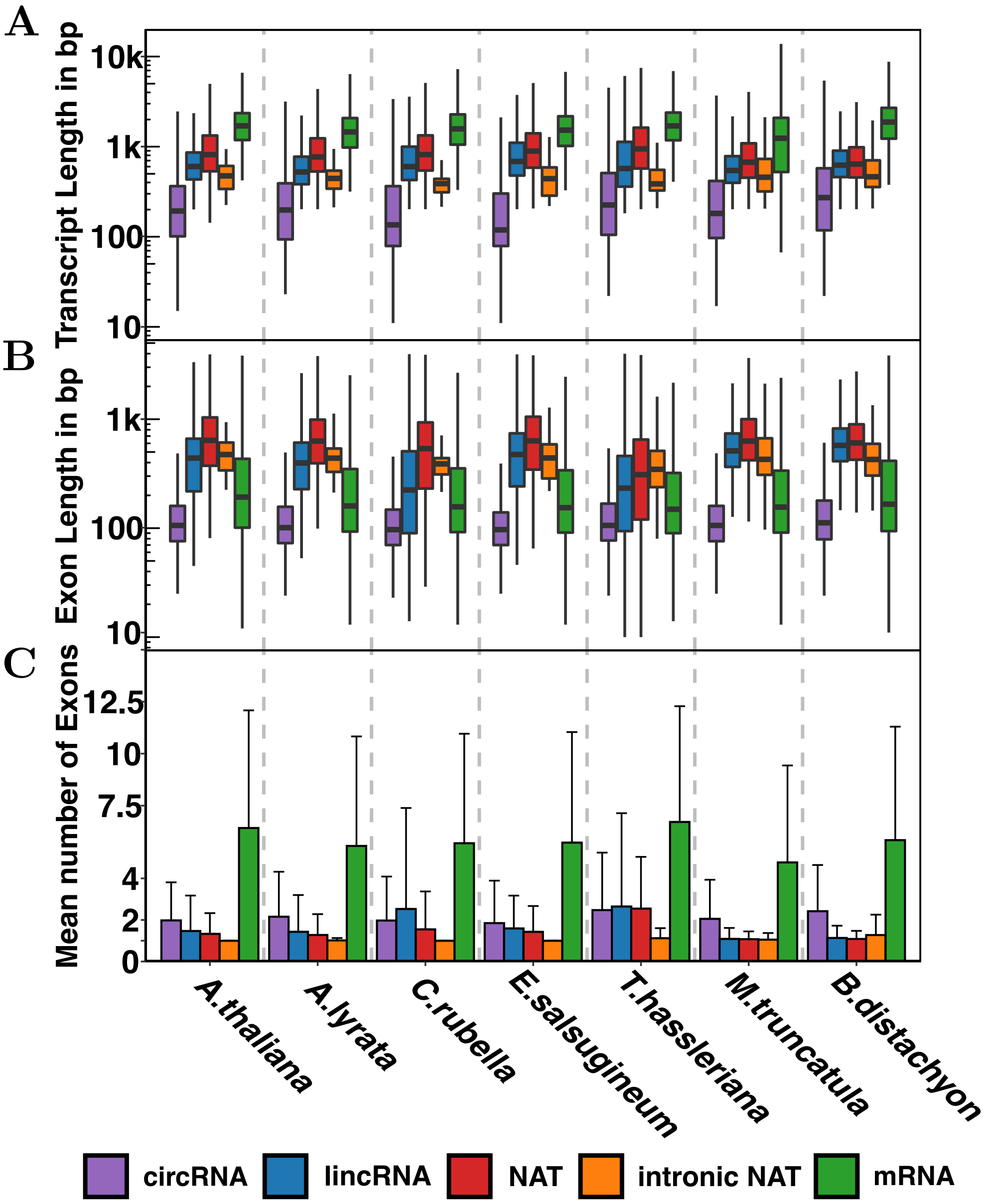


Figure 8.10 | Transcript and exon lengths distributions and mean number of exons of RNA species. (A) Transcript length of transcript isoforms, (B) exon lengths of spliced transcript isoforms, and (C) average number of exons ± SD for each sequenced plant species and each detected circRNA, lincRNA, NAT, intronic NAT, and protein-coding mRNA.

The classification of lncRNAs depended on their genomic location with respect to e.g. protein-coding loci. We saw in all species lincRNAs with distances below 500 bp which could implicate possible clusters of lincRNAs (Fig. 8.11A). Instead, over 75% of lincRNAs were very distant from other lincRNAs by distances >5 kbp. Compared to Fig. 8.11B the median distance of lincRNAs to the closest protein-coding genes was just 1 kbp. The maximal distance to protein-coding genes increased to over 50 kbp, these lincRNA are described as isolated lincRNAs [232].

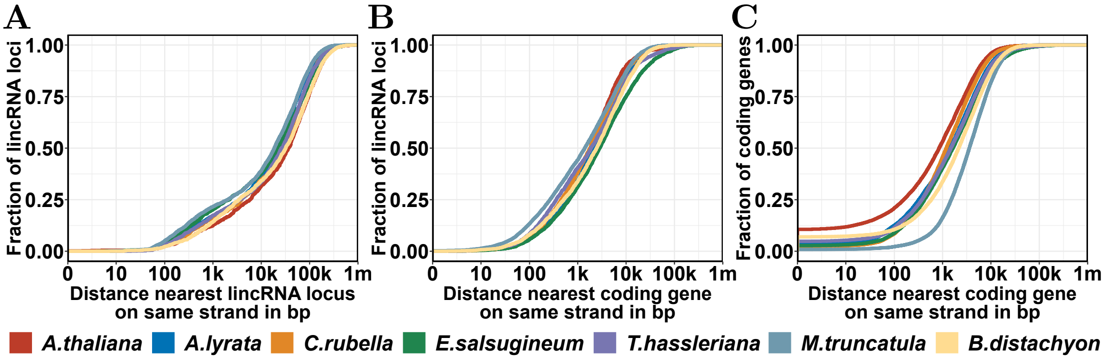


Figure 8.11 | Distance of lincRNA to nearest protein-coding locus. (A) Empirical cumulative density of distances measured from each lincRNA loci to their nearest non-overlapping lincRNA loci located on the same strand. (B) Empirical cumulative density of distances measured from lincRNA loci to their nearest non-overlapping protein-coding loci located on the same strand.

On the other side of the distance spectrum, we saw lincRNAs in close proximity between 0 and <1 kbp. We classified lncRNAs as intergenic if their start or end did not overlap with protein-coding genes [154, 198] or if they were not within a distance between 500 bp and 1 kbp to a protein-coding gene [199, 209, 233, 234]. The later definition of lincRNAs is often combined with the definition bidirectional lncRNAs, which suppose to share the same promotor region with its neighboring protein-coding gene. To our knowledge, there is no comprehensive definition of bidirectional lncRNAs, which could be applied to plants and animals. It depends on the location of the promotor to the closest protein-coding gene to define bidirectional lncRNAs.

8.3.4 Circular RNAs

Besides the prediction of lncRNAs, the DevSeq annotation workflow also detected thousands of novel circRNAs. Predominantly, we found circRNAs in protein-coding loci, i.e., 87-95% of all observed circRNAs were located within protein-coding genes. Besides, we detected ~4-8% circRNAs within intergenic regions and 1-5% within lncRNAs (Fig. 8.12A).

The observed transcript length of circRNAs ranged between 10 bp and 3 kbp (Fig. 8.10A). The median length was ~200 bp and thus much shorter than the median transcript length of lncRNAs or protein-coding transcripts. In contrast, circRNAs seem to contain more exons than lncRNAs (Fig. 8.10C).

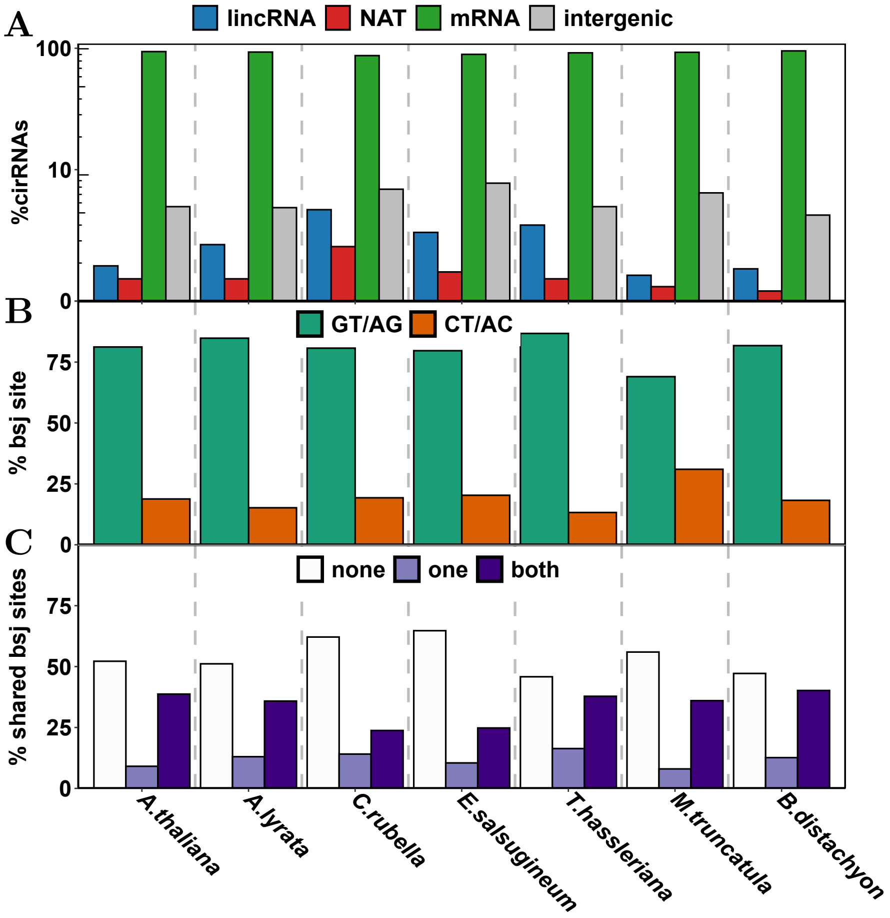


Figure 8.12 | Location and backsplicing of predicted circRNAs. (A) Percentage of detected circRNAs within each species over all lincRNAs, NATs, protein-coding mRNAs, and intergenic regions. (B) Splice donor-acceptor sites of back-splice junctions forming the predicted circRNAs. (C) Fraction of circRNAs within protein-coding mRNAs sharing none, one, or both of their backsplicing sites with exon boundary splice sites of its host transcript.

The majority of back-splice sites forming a circRNA were GT/AG donor-acceptor pairs, which was also the largest fraction in transcript splice sites within the analysed plant species over all RNA species (Fig. 12.2). About 10-25% of all back-splice sites were CT/AC donor-acceptor sites. This pair was almost not present (0.0008% species mean) within linear transcript isoforms (Fig. 12.2). We also found that the back-splice sites which were not intergenic shared no splice site with their host transcript (Fig. 8.12C none) or shared both splice sites with its linear host transcript. Only a minority of back-splice sites shared only a donor or only a acceptor splice site with their linear host transcript.

8.3.5 GC content of coding and non-coding RNAs

In this subsection, we wanted to investigate the GC contents of the different RNA species. The GC content is an characteristic genomic feature to differentiate between coding and non-coding regions. It was shown that in plants that the genomes of grasses contain higher GC contents than other angiosperms [235–238]. These studies focused on the GC abundance and GC variation on the genome level by comparing intergenic and protein-coding regions [236, 238–241].

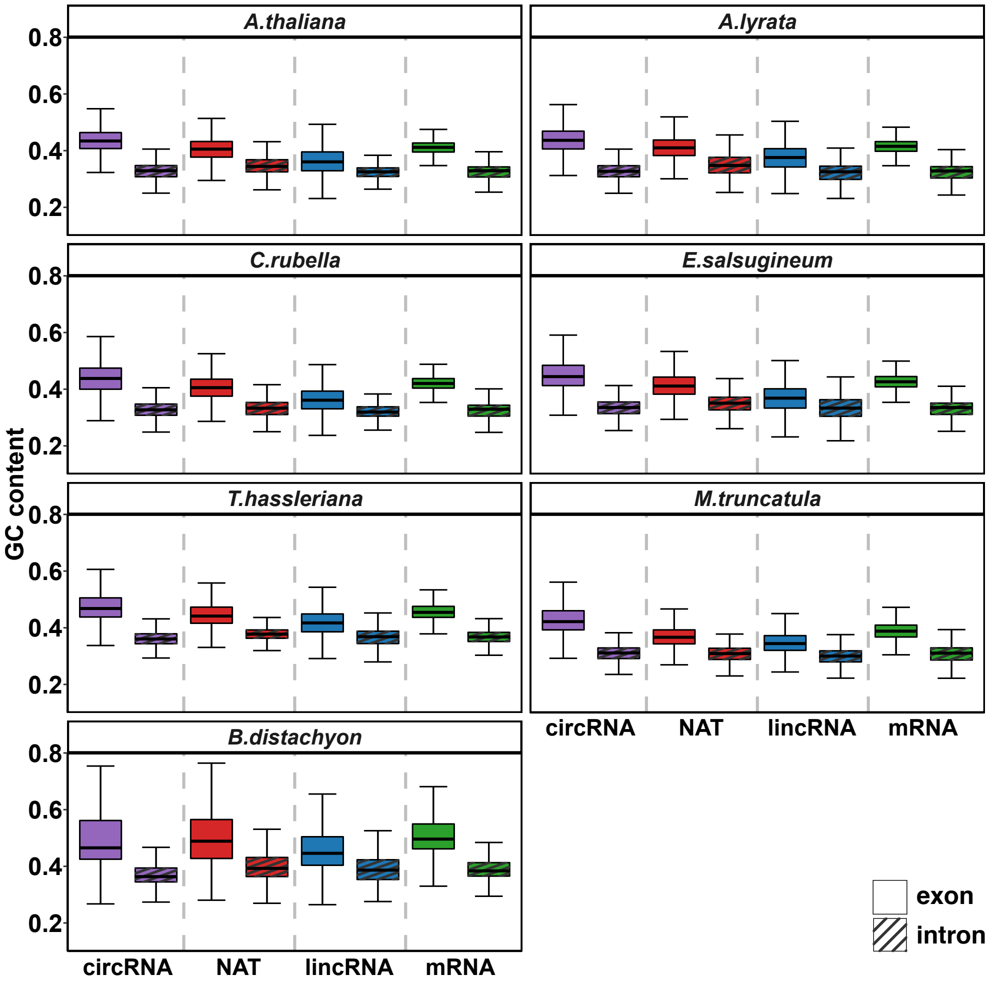


Figure 8.13 | GC content of exons and introns within different RNA species. For each plant species, we calculated the relative GC content for exons and introns within circRNAs, lincRNAs, NATs, and protein-coding mRNAs. The exonic sequences show a higher GC content and a broader IQR compared to the intronic sequences. The range of GC content of the different RNA species is very similar for the majority of plant species, except *B. distachyon*. Only *B. distachyon* shows the largest GC content of all plant species and the widest IQR in all RNA species.

Starting with the comparison of intronic and exonic sequences (Fig. 8.13), we found in all RNA species that the GC content was lower in intronic regions than in exonic regions. This low GC content is in accordance to observations from protein-coding regions in plants and animals [242–244].

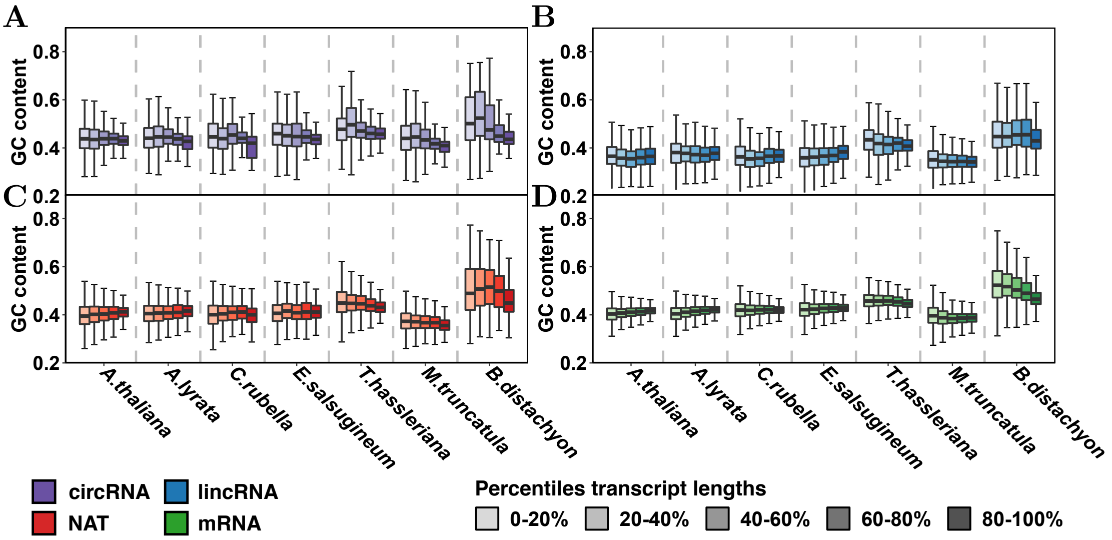


Figure 8.14 | GC content vs. transcript lengths. (A) CircRNAs, (B) lincRNAs, (C) NATs, and (D) protein-coding mRNAs of each plant species are grouped into equally sized quintiles by their transcript lengths. For each group, we plot the relative GC content of each transcript within each quintile. Table 12.4 presents detailed information on transcript lengths and GC content within each quintile.

We also saw that *B. distachyon*, as a representative of grass plants and monocotyledons, showed the highest GC contents followed by *T. hassleriana* (Fig. 8.13 and Tab. 12.4). *B. distachyon* also showed the widest range of GC contents in all four RNA species over all plants. The GC content of introns was very similar among the RNA species within each plant species.

Next, we grouped the transcripts of each RNA species within each plant species based on their transcript length into five groups (Fig. 8.14 and Tab. 12.4). In general, we saw in the majority of RNA species a decrease in the IQR and in the standard deviation of GC content with increasing transcript lengths whereas the median GC content was in most plant species constant.

In circRNAs, we found a slight decrease in GC content with increasing transcript lengths, especially in *B. distachyon*. We also detected this decrease in NATs and protein-coding mRNAs of *B. distachyon*. In contrast, the GC contents of lincRNAs over increased transcript lengths were almost constant. Regarding lincRNAs, the GC contents of *T. hassleriana* and *B. distachyon* were the highest compared to the other plant species. *M.truncatula* showed the lowest GC content for lincRNAs, as well as for NATs and for mRNAs.

In contrast to *M. truncatula*, *B. distachyon* showed similar to Fig. 8.13, the highest GC content in all four RNA species. It is the only plant species which showed an decrease in GC content with an increasing transcript length for mRNAs (Fig. 8.14 and Tab. 12.4). A detailed investigation of GC contents of coding and non-coding transcripts in flowering plants was not the focus of this study and will be part of future work.

8.4 Discussion

In chapter 8, we introduced our workflow to predict and annotate protein-coding splice-variants, lncRNAs, and circRNAs in flowering plants based on RNA-Seq experiments performed in an organ-specific manner. The workflow depended on reference genome annotations and the corresponding genome sequences. Many genome annotation workflows have been published with the attempt to be easy to adapt for individual sequencing projects [245–247] but only concentrated on the prediction of protein-coding genes and transcripts.

The DevSeq workflow presented in this study was developed to process RNA-Seq data from total RNA libraries of different plant species represented by different organ samples. Building the workflow, we had to achieve several challenges in the raw data and processed data such as high amounts of duplicated reads, read-through transcripts, and the differentiation between lncRNAs and protein-coding transcripts, and the prediction of circRNAs. After the elimination of read sequences with poor quality, we mapped the reads with the RNA-Seq aligner STAR [248], which is capable of aligning linear split reads in order to predict splice-variants and also aligning chimeric reads in order to detect circularization events, i.e., back-splice junctions, to predict circRNAs.

The subsequent organ- and species-specific transcript assembly determined the results of all subsequent analyses. To achieve a comprehensive transcriptome assembly, we combined the predicted transcripts with their abundances in each organ. Since lncRNA transcripts and certain protein-coding splice-variants show organ-specific gene expression, we removed transcripts with low expression values. We used artificially introduced ERCC Spike-Ins to determine a minimal expression threshold for each species and each organ to remove transcripts that show low expression values and thus might not be of biological relevance. This filtering was a huge advantage compared to other approaches that use arbitrary expression thresholds and thus may remove expressed transcripts with actual biological meaning or keep transcripts which may spoil the transcript assembly.

After we merged the organ-specific transcriptome assemblies to species-specific transcriptome assemblies, we divided the transcriptome into transcripts overlapping protein-coding loci of the reference annotation and novel transcripts without any overlap. By this separation, we could focus on the prediction of novel loci, which we later classified as putative coding or non-coding. To calculate the coding potential to distinguish protein-coding from non-coding transcripts, we performed a benchmark determining an appropriate tool for our classification task. As shown in Fig. 8.5, the accuracy of each tool depended on the species and thus on the training data of the classifier. For our workflow, we chose TransDecoder [183] as it achieved together with CPC2 [205] the highest accuracy, especially for *A. thaliana*. In contrast to CPC2, TransDecoder only relied on annotated protein-coding genes of the species that is investigated. For the majority of our plant species existed no annotation of lncRNAs. We relied on the reference annotation of each plant and tailored the prediction of lncRNAs for each plant species based on their annotated protein sequences.

Current prediction algorithms try to separate as good as possible protein-coding transcripts from non-coding transcripts. LncRNAs, as a subset of non-coding RNAs, are solely defined based on their length (>200 bp) and their genomic location with respect to protein-coding loci. In plants, the number of predicted lncRNAs continually increases, but the biological function of a majority of these transcripts is still unknown. To improve the prediction of lncRNAs, especially in plants, we need curated datasets of lncRNAs analogously to PfamA [186] for protein-coding genes, providing lncRNAs from various plant species with proven biological functions. Current curated datasets mostly rely on *A. thaliana*, but as we have seen in section 8.3.3 the characteristics of lncRNAs are also species-dependent. With plant-specific datasets, we could define lncRNAs more accurately and might increase the prediction rate of biologically relevant lncRNAs.

Based on the classification results of novel transcripts into coding and non-coding, we could increase the number of protein-coding genes and postulate novel splice-variants (Tab. 8.3). Additionally, we classified the non-coding RNAs into different lncRNAs subgroups, such as lincRNAs, NATs, and intronic NATs based on their genomic locations. LncRNAs and protein-coding genes, resp. mRNAs, showed characteristic genomic features.

With the presented findings in flowering plants, we could show that lncRNAs compared to mRNAs seemed to have very short or even no open reading frame (Figs. 8.9), shorter transcripts, less exons per transcript, and longer exons (8.10). These findings are in agreement with lncRNA features found in animals based on short-read sequencing [154]. More recently, it could also be shown by targeted RNA capture and long-read sequencing that in animals the transcript length and the number of exons are similar between lncRNAs and protein-coding transcripts [231]. Taking this into account, our observations could be biased by the sequencing technology. Regarding the uniform pattern of these features throughout the analyzed flowering plants, we might found evidence at least in flowering plants that the transcript length and the number of exons are different between lncRNAs and protein-coding trasncripts.

In contrast, we found differences between the species in the number of transcripts per gene comparing lncRNAs to protein-coding genes. *A. thaliana* and *B. distachyon* showed significant differences between the number of transcripts per gene in protein-coding genes compared to lncRNAs, which could also be observed in animals [154]. In *T. hassleriana* the number of lncRNA transcripts per gene was ∼1.5, which is similar to the protein-coding genes of *T. hassleriana* and in *A. lyrata*, *C. rubella*, *E. salsugineum*, and *M. truncatula* we observed on average between 1-1.2 transcripts per gene for protein-coding genes and lncRNAs. This variation in the ratio of transcripts per gene between the flowering plants could be species-specific or due to a bias in the quality of the provided reference annotations which have an immense impact on the number of protein-coding transcripts because the presented workflow does not discard annotated protein-coding transcripts which we could not detect.

Before the predicting circRNAs, we performed a benchmark of common circRNA prediction tools by comparing detected circRNAs within our *A. thaliana* samples against the PlantcircBase [220]. With DCC [216], we achieved the highest amount of overlap to the PlantcircBase and also to the other circRNA prediction tools. Similar to determining lncRNAs, the choice of the correct tool or cirteria for predicting circRNAs is not straightforward. Besides the lack of validated circRNAs from various plant species, most detected circRNAs in plants were computationally predicted for *A. thaliana*. The decision to use DCC was not solely based on the overlap results presented in the benchmark but also due to its capability of using the information from biological replicates to ensure that the detected backsplice junction were not aligned by chance but were present in at least two out of three biological replicates. The advantage in this study was that we used the same method for all species and all organ samples which further allowed us to compare the prediction results across the different plant species. We found that over ∼90% of all detected circRNAs are hosted by protein-coding genes and the characteristic backsplicing-junctions were only canonical splice-sites GT/AG and CT/AC. This is in accordance with previous studies in *Oryza sativa* and *A. thaliana* [213].

We finally calculated the GC contents of each RNA species for each flowering plant. With this characteristic genomic feature, we could highlight the similarities of circRNAs, lincRNAs, NATs, and mRNAs within the group of Brassicacae, which showed similar GC contents. Additionally, we found high GC contents in grass plants, represented by *B. distachyon.* This monocot showed in all GC comparisons the highest GC content and the highest variation of GC content, which was previously demonstrated only on protein-coding transcripts [235–238].

8.5 Conclusions and Outlook

For predicting and annotating protein-coding splice-variants, lincRNAs, NATs, intronic NATs, and circRNAs, we developed a fully reproducible Snakemake workflow [221]. The workflow produces curated genomic mappings providing information for linear and circular transcript assemblies in an organ-specific manner. Based on the sequencing of ERCC Spike-Ins in each sample, we were capable to introduce expression thresholds for each sequenced organ allowing the elimination of transcriptional noise and reducing the amount of falsely assembled transcripts. Additionally, the expression thresholds enabled the detection of lowly expressed transcripts like lncRNAs. The resulting comprehensive assemblies build the foundation of the three final subworkflows to predict and annotate novel protein-coding transcripts, lncRNAs, and circRNAs.

Due to limited information of non-coding transcripts in non-model plant species, the prediction of lncRNAs and circRNAs is crucial. Hence, we validated the algorithms currently available and suitable for predicting plant non-coding RNAs to choose the best approach for the presented and future sequenced plant species. On average, we found ∼5000 novel lncRNAs and ∼2000 novel circular RNAs for each plant species, identified thousands of novel protein-coding transcript isoforms and updated currently annotated transcripts by elongating their 3’ and 5’ ends. Our publicly available workflow and the seven annotations could serve as a possible starting point of a resource for organ-specific annotations of non-coding RNAs in plants and might potentially become useful for deepening our understanding of the developmental transcriptomes in more complex plants.

In chapters 2 - 8, we have presented bioinformatics software solutions to analyse transcriptomic data in evolutionary developmental biology related to the investigation of the developmental hourglass (chapters 2 - 6), to address scientific problems in developmental biology for the analysis of transcriptome dynamics during grafting (chapter 7), or at the border to genomics to discover and annotate novel protein-coding and non-coding transcripts in flowering plants (chapter 8).

However, the analysis of nucleic acid chains as in transcriptomics, or genomics, is only one way to gain an insight into the biology of a living cell. In addition, metabolomics is able to broaden our view and thus deepen our understanding of biological processes. In an attempt to contribute, we will develop bioinformatics approaches to investigate metabolomic data in order to analyze the free fatty acids composition of the human skin barrier in chapter 9 and to study serum metabolite profiles of pigs in chapter 10.

9

Age- and Diabetes-related Changes in the FFA Composition

In chapters 2 - 8, we have learned about several bioinformatics approaches in different transcriptome analyses. However, the analysis of metabolome data such as the concentration of amino acids, proteins, carbohydrates, or lipids can provide us an additional layer of information to investigate biological processes. In this chapter, we will turn our attention to the analysis of lipidomics data. Lipids are essential for the living cell as they are the building blocks of plasma membranes, serve as energy resources, are involved in cell signaling, and provide a protective layer, to name a few functions. Due to their importance, they play a crucial role in aging and are associated with diseases such as diabetes mellitus.

In the following sections, we will investigate age-related and diabetes-induced changes of the lipid spectrum in the outermost epidermal barrier, the Stratum corneum (SC). For this purpose, we will analyse the free fatty acid (FFA) compositions of the SC, which were qualitatively and quantitatively assessed by gas chromatography-flame ionisation detection. From a bioinformatics perspective, we will present a straightforward statistical analysis to detect and quantify differences in the FFA compositions of even- and odd-numbered FFAs.

In section 9.1, we will describe the influence of aging and diabetes mellitus to the physical barrier function of the SC and we will provide an overview of the FFA synthesis. In section 9.2, we will learn about the experimental design, the quantification of FFA concentrations, and the statistical analysis. In sections 9.3 and 9.4, we will present and discuss the comparisons FFA concentrations in young and old subjects as well as in healthy and diabetic subjects. In section 9.5, we will conclude that the identification of characteristic FFAs could help to understand age- and disease-dependant changes of the epidermal barrier.

The following sections are extracted from Wohlrab et al. 2018 "*Age- and Diabetes-related Changes in the Free Fatty Acid Composition of the Human Stratum Corneum*" [7].

9.1 Introduction

Aging is a complex and multifactorial physiological process, which is not yet fully understood. Nevertheless, various hypotheses have tried to describe the complex interactions involved and to understand the causal correlations. Depending on the perspective, these hypotheses take an evolutionary biological, molecular, cellular or systemic approach [249]. The influence of intrinsic and extrinsic factors on aging is universally acknowledged although the respective interactions and the relation within the historical and scientific context are subjects of controversial debate [250]. The overlapping cascades of different aging processes, which influence each other, have a high relevance for the skin organ and in particular the epidermis which serves as the body's physical barrier [251, 252]. Atrophy in the various layers of the epidermis is mainly caused by a decreasing number of keratinocyte layers, which is thought to be due to a diminished cell division rate. Additionally, reduced proliferative activity and differentiation rate cause corneocytes to increase in size into the stratum corneum (SC). Changes in the physical barrier function are characterised by decreased water-binding capacity of the SC due to a reduction in natural moisturising factors while the lipid spectrum is quantitatively reduced [252, 253].

Hyperglycaemia, which is common to all subtypes of diabetes mellitus, induces a series of biochemical changes in skin. These result in increased oxidative stress and expression of redox-regulated genes and transcription factors, in changes regarding the composition of the extracellular matrix and in functional deficits of proteins [254, 255]. These complex biochemical changes of epidermal micro milieu cause reduction in both keratinocytic proliferation and SC water content [256, 257]. The overall decrease in water content within the epidermis reduces the transcorneal passage of water. These are the main causes for a reduction in the compensation capacity of the barrier, although under normal conditions, none or very minimal change in transepidermal water loss is observed in both aged skin and diabetic xerosis. Functional impairment of the SC often only becomes clinically evident after irritation or occlusion and appears as dry skin or eczema, and the associated pruritus [258].

The physical barrier function of SC is characterised by complex interactions between various physicochemical molecular groups and cellular components [259, 260]. Not only are the quantity and quality of the individual components thought to be vitally important for barrier function but so is their molecular organization, that is, the position of the molecules in relation to each other [261]. Keratinocytic lipid synthesis is mostly an autonomous process and produces not only cholesterol and cholesterol derivates but also free fatty acids (FFAs) with different chain lengths as well as triglycerides [262].

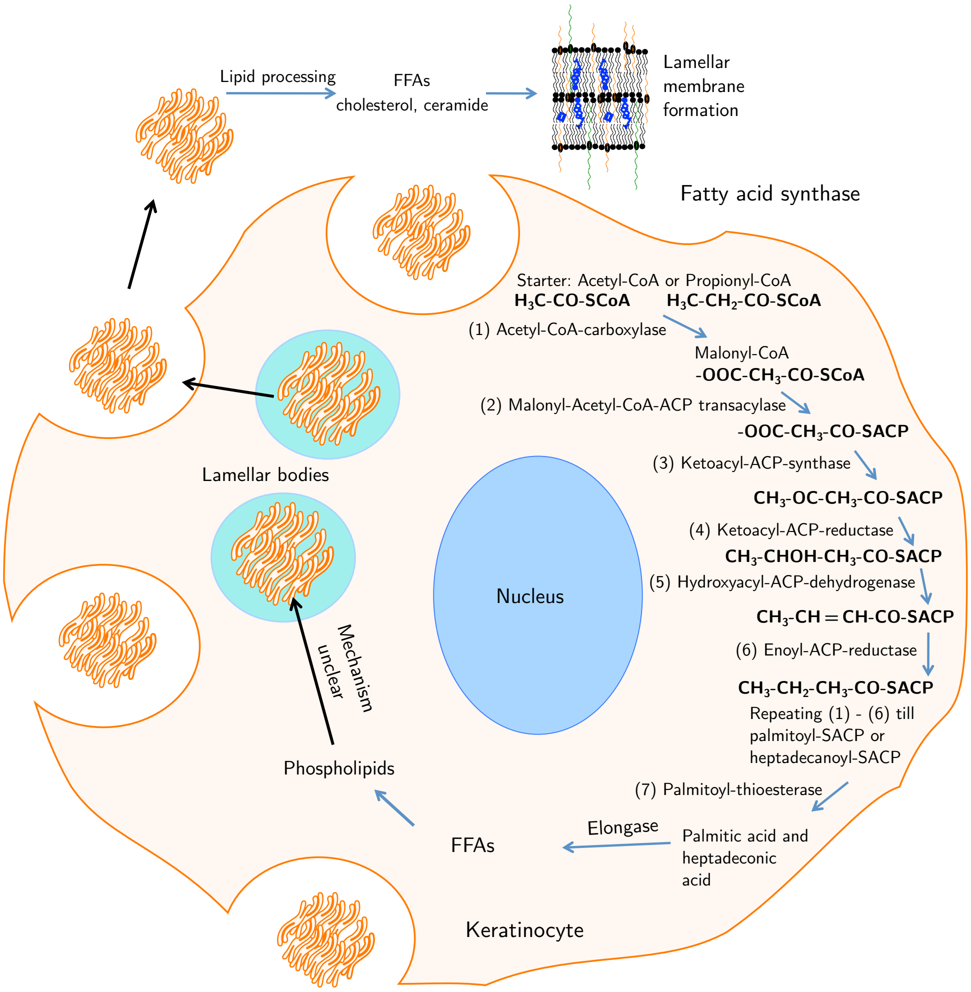


Figure 9.1 | FFA synthesis in the cytosol of a keratinocyte done by multi-enzyme complex and subsequent storage in the lamellar bodies. Finally, the release of the FFA is accomplished as precursors. FFA, free fatty acid. Reprinted Figure 1 from [7].

Additionally, ceramides (CERs) are synthesised in the endoplasmic reticulum of the keratinocytes, and their anisometric molecular structure differs markedly from other lipid classes [263, 264]. Due to load differences within their long-chain molecules, CERs are able to spontaneously form lyotropic mesophases (also known as liquid-crystalline membrane structures) [265, 266]. Unlike phospholipids, functionally important CERs in the SC contain 2 alkyl chains of different lengths. Depending on the hydration level, they show different configurations; various membrane models describe a complex network of membrane sections with polymorphic phase characteristics [259, 267, 268]. In this context, the metabolic and physicochemical interactions of FFAs are of vital importance for the organisation of the SC barrier [269, 270]. Fatty acids, such as intrakeratinocytic substrates, are the basis for the synthesis of CERs as well as phospholipids. After secretion of the lipid matrix from the lamellar bodies, FFAs are formed in the Stratum granulosum by the breakdown of phospholipids [271].

Together with other lipophilic molecules, FFAs are very important for the characteristics of CER membranes, especially for their dense orthorhombic packaging. Previous insufficient attention was placed on the importance of odd-numbered FFAs, which are formed via the phytosphingosine metabolic pathway [272]. CER subspecies have an odd number of carbon atoms in the amide bound fatty acid chain of the molecule [273, 274]. The existence of oddnumbered FFAs in the SC is considered possible, since instead of an acetyl group, a propionyl group may be attached to the fatty acid chain [275]. However, the presence of oddnumbered FFAs and the influence of age on the composition of FFAs have not been extensively studied. Insights into pathogenetic relationships are important for performing a targeted substitution of the epidermal barrier function. For skin lipids, the problem in this case is that qualitative changes, in addition to quantitative changes, could be crucial, in particular for FFAs [276].

It should be mentioned that, aside from the SC, the epidermis is particularly good at synthesising large quantities of both even- and odd-numbered FFAs [277]. Depending on the starting molecule (acetyl-CoA or propionyl-CoA), FFAs are synthesised by fatty acid synthase through the condensation of 2 or 3 carbon units, up to a chain length of C16:0 and C17:0 (Fig. 1) [278]. Chains longer than C16:0 and C17:0 are subsequently synthesised by elongase enzymes [279]. Finally FFAs are released as phospholipids together with lipid hydrolases from the lamellar bodies and accumulate in the intercellular lipid-rich matrix [280]. Aging processes and chronic metabolic changes, such as diabetes mellitus, have an important influence on the lipid metabolism of keratinocytes and cause quantitative as well as qualitative changes of the lipophilic molecules within the membranes of the SC [281]. It is clear that as the skin becomes thinner and drier with age as well as during longconsisting diabetes and this likely influences lipid patterns observed in the SC. Moreover, variances in FFA chain length distribution have already been established as a factor in the pathogenesis of atopic eczema and are currently discussed in psoriasis and ichthyoses [253, 282–285].

In this study, we developed a valid analytical method and evaluated age-related and diabetes-induced changes in the FFA composition of the SC, with the possibility of targeted substitution of the lipid barrier being an antiaging skin or anti-diabetic xerosis treatment approach in the future.

9.2 Materials and Methods

In this section, we will present the quantification and analysis of FFA concentrations. In subsection 9.2.1, we will introduce the participants in this study who provided the skin samples. In subsection 9.2.2, we will describe the extraction of lipids from the SC. In subsection 9.2.3, we will turn to the quantification of FFA concentrations based on gas chromatography. In subsection 9.2.4, we will present the statistical analysis of the FFA concentrations.

9.2.1 Subjects

The study design was approved by the ethics committee of the medical faculty at the Martin Luther University Halle-Wittenberg (Germany) and the study was performed by an experienced dermatologist. Written informed consent was provided by all participants. Inclusion criteria were as follows: non-smoker, Caucasian males or females, exclusion criteria included erosive, ulcerative, or inflamed skin lesions in measurement areas, participation in another clinical trial 4 weeks before study start and the topical application of drugs or skin care products 1 week before study start. Three groups were investigated: healthy subjects aged over 60 years (elderly/healthy), healthy subjects aged 18-40 (young/healthy) and subjects with insulin-dependent diabetes mellitus for at least 5 years aged 18–40 (young/diabetic). Because of the good reachability, clinical involvement and comparability of the specific skin regions lipids were extracted from the following areas: subgroup elderly/healthy = inner forearm, subgroup young/healthy = inner forearm + inner site of foot, subgroup young/diabetic = inner site of foot.

9.2.2 Extraction of SC lipids

For the in vivo extraction of surface lipids, a cylindrical glass ring with an extraction area of 6.15 cm^2 was filled with 5 mL nhexane/ethanol 2:1 (v/v). The open side was pressed tightly to a skin area on the inner forearm. The extraction time was always exactly 5 minutes. The extracts were evaporated at a temperature of 50 °C under a stream of nitrogen. This resulted in a dried residue, which was then stored at –30 °C and subsequently dissolved in 250 µL n-hexane/ethanol 2:1 (v/v) before use [286].

9.2.3 Gas chromatography – Flame ionization detection analysis

GC analyses of FFAs were carried out using the Agilent 7890A GC System (Agilent, Waldbronn, Germany) equipped with an autosampler and a flame ionization detector. The GC column used was Optima FFAPPlus 0.25 µm, 30 × 0.25 mm ID (Macherey-Nagel GmbH & Co. KG, Düren, Germany). Nitrogen was used as the carrier gas. The gradient temperature program was 80 °C for 1 min, 160 °C for 3 min, 250 °C for 6 min and 260 °C for 12.5 min. The injection volume was 1 µL and the injection temperature was 250 °C. TMSH 0.2 M was used as the reagent due to having high volatility. Detection of all FFAs was carried out at 300 °C. A fourfold deuterated FFA (lignocerin-9,9,20,20-d4-acid) was used as an internal standard. Thus, inaccuracies were avoided by calculating the peak area ratio. All runs included 2 recordings.

The calibration curve was constructed using 6 concentrations (0.2; 0.5; 1.0; 1.5; 2.0 and 2.5 µg/mL) of each FFA. The linearity of each plot (concentration versus peak area) was tested using linear regression analysis. Estimation of limit of quantification (LoQ) and limit of detection were calculated from the signal-to-noise ratio. The peak height needed to be 10 times higher than the baseline noise for quantification and 3 times higher for detection. The between-run precision and accuracy of this method were determined by analysing 4 replicates containing 0.2; 0.4; 1.5 and 2.5 µg/mL in hexane/ethanol 2:1 (v/v).

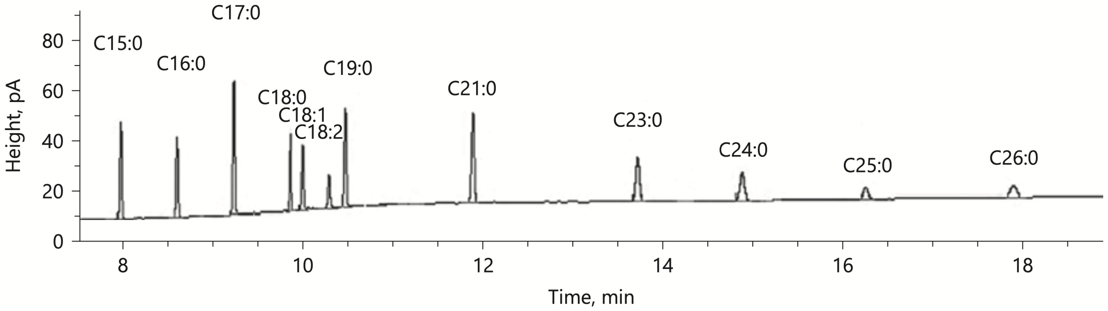


Figure 9.2 | Description of a chromatogram example of GC with fatty acid standards. The x-axis displays the retention time in minutes and the y-axis the peak height in pA. Reprinted Figure 2 from [7].

Five determinants of the same concentrations were conducted over 3 runs on 3 different days. Deviations at the LoQ are allowed in a range of ± 20% and for higher concentrations ±15%. The selectivity of the method was examined through analysis of all FFAs in 1 standard solution and a good peak separation was achieved (Fig. 2). Carryover effects were analysed by observing the occurrence of FFAs in a blank sample of hexane/ethanol 2:1 (v/v) with 2 µL of the derivatisation reagent

TMSH 0.2 M added to 50 µL of sample solution. FFAs that remained as residues in the GC or as contamination could thus be detected. However, in the blank samples, no detectable levels of FFAs were found. Butylhydroxytoluol 0.05% was added to ensure the storage stability of unsaturated FFAs through the prevention of oxidation [287].

9.2.4 Statistical analysis

For each subject, all FFA concentrations were measured twice, and corresponding pairs of FFA concentrations were summarised by using their arithmetic means. These data were used for all the following statistical tests, and the programming language R [288] was used. The obtained FFA data were not normally-distributed, as indicated by the Shapiro-Wilk test [289]. Thus, the non-parametric Mann-Whitney U test was used for assessing statistical significance for each FFA between the two groups. As there were multiple comparisons of each FFA, the resulting P values were adjusted by using the Bonferroni correction method. Differences were considered statistically significant if the adjusted P value was < 0.05. For comparing the complete FFA concentration profiles among the two groups, a multivariate Wilcoxon test was performed [290].

9.3 Results

Overall, 258 subjects were included in this study: 110 subjects in group elderly/healthy, 110 subjects in group young/healthy and 38 subjects in group young/diabetic. In subsection 9.3.1, we will present the FFA concentrations measured in the different groups of subjects. In subsections 9.3.2 and 9.3.3, we will compare the FFA concentrations based on age-related and diabetes-induced changes in the lipid composition of the SC.

9.3.1 Quantification and identification of FFA concentrations

For quantification of FFAs in the SC, a selective gas chromatography – flame ionization detection method was developed and validated. This method demonstrated excellent peak separation and the values of precision and accuracy, according to the criteria of the European Medicines Agency were met [291]. To minimise errors, an internal standard was added to the measurements. The analysis showed no carry-over effect and was shown to be very well suited for measuring even- and odd-numbered FFAs as well as unsaturated FFAs.

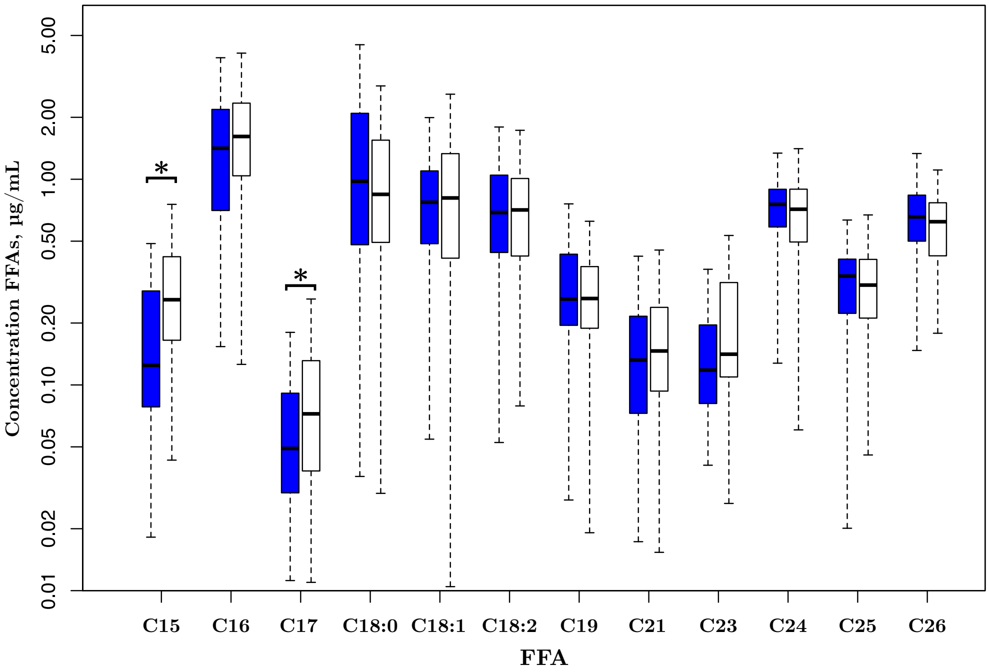


Figure 9.3 | Box plot comparing FFA concentrations in µg/mL of elderly/healthy and young/healthy subgroups. The blue box plots represent subgroup elderly/healthy, while the white box plots represents subgroup young/healthy. Each box plot contains 110 FFA concentrations. * Statistically significant differences between subgroups; $P < 0.05$. FFA, free fatty acid. Reprinted Figure 3 from [7].

All of the FFA standards were recovered in skin extracts from the SC. The most abundant FFAs with a shorter chain length (chain length smaller than 20 carbon atoms) were C18:0, C18:1 and C18:2 and with a longer chain length were C24:0 as well as C26:0. The FFA with the lowest concentration of all investigated FFAs was C17:0 with 0.7% of the total FFA amount. Large interindividual variability in individual FFAs concentrations was observed. In particular, levels of the FFAs C18:1 and C18:2 differed markedly between subjects (Fig. 9.3, 9.4).

9.3.2 Comparison of elderly/healthy vs. young/healthy

Individual FFA levels in the total FFA content of the skin were compared between the 2 groups and the comparisons were analysed for statistical significance (Fig. 9.3). Odd-numbered FFAs comprised 20.7 and 22.6% of the total FFA concentration in the SC on volar forearm in subgroup elderly/healthy and in subgroup young/healthy respectively (Table 9.1). After applying the Mann-Whitney U test, levels of the FFAs C15:0 and C17:0 were found to be significantly lower in subgroup elderly/healthy, compared with subgroup young/healthy (Table 9.2). No statistically significant differences in the complete FFA concentration profiles were observed between these 2 groups.

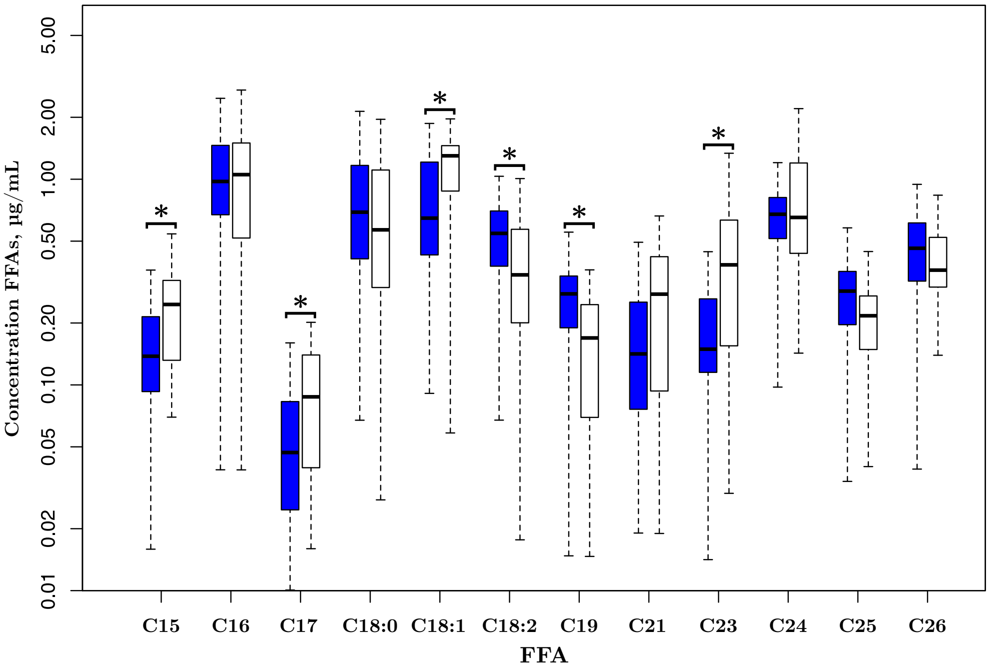


Figure 9.4 | Box plot comparing FFA concentrations in µg/mL of young/healthy and young/diabetic subgroups. The blue box plots represent subgroup young/healthy, while the white box plots represents subgroup young/diabetic. The box plot contains 110 FFA concentrations in subgroup young/healthy and 38 in subgroup young/diabetic. * Statistically significant differences between subgroups; $P < 0.05$. FFA, free fatty acid. Reprinted Figure 4 from [7].

9.3.3 Comparison of young/healthy vs. young/diabetic

Compared with subgroup young/healthy, levels of C18:2 and C19 were significantly decreased in subgroup young/diabetic, ($P < 0.004$, $P < 0.0005$ respectively) and levels of C15, C17, C18:1 and C23 were significantly increased ($P < 0.001$, $P < 0.005$, $P < 0.01$ and $P < 0.01$ respectively; Fig. 9.4). The total contents of odd-numbered FFAs in SC of inner site of foot were 22.8 and 23.60% in subgroup young/healthy and in subgroup young/diabetic respectively (Tables 9.1, 9.2).

	Volar forearm		Inner site of foot	
	elderly/healthy	young/healthy	young/healthy	young/diabetic
C15:0	2.7 ± 2.3	4.5 ± 2.3	2.8 ± 1.7	4.1 ± 1.7
C16:0	19.6 ± 9.9	22 ± 8.1	17.8 ± 7.5	16.6 ± 6.8
C17:0	1.1 ± 2.2	1.3 ± 1.1	0.9 ± 0.8	1.4 ± 0.7
C18:0	13 ±12.8	11.6 ±10.9	13.3 ± 8.6	11 ± 7.1
C18:1	11.6 ± 8.0	12.8 ± 7.0	14.1 ± 6.3	20.4 ± 7.5
C18:2	11.2 ± 4.9	10.7 ± 5.4	10.2 ± 4.3	6.3 ± 3.6
C19:0	5.4 ± 4.1	4.3 ± 2.9	5.1 ± 2.7	2.6 ± 1.8
C21:0	2.7 ± 2.7	2.6 ± 2.1	3.4 ± 3.1	4.3 ± 3.7
C23:0	2.9 ± 2.6	4.2 ± 5.7	4.9 ± 5.4	7.4 ± 5.0
C24:0	12.9 ± 8.2	11 ± 6.0	13 ± 6.1	15.5 ±12.6
C25:0	5.9 ± 4.5	5.7 ± 5.8	5.7 ± 5.0	3.8 ± 1.4
C26:0	11.2 ± 5.8	9.5 ± 4.2	8.7 ± 4.1	6.6 ± 2.2

Table 9.1 | Relative concentrations of FFAs in the different subgroups (%) ± SD. Relative concentrations of each FFA was calculated for each subject in each subgroup. The table displays the mean ± standard deviation of relative FFA concentrations in %. Adapted Table 1 from [7].

9.4 Discussion

In accordance with the aim of this study, we developed as a first step a well-suited and sensitive GC method for analysing FFA from human SC. Excellent peak separation was achieved and we found that the LoQ and Limit of Detection are sufficient to quantify small amounts of FFAs from the SC extracts as a second step. To our knowledge, no such extended SC study with this high number of subjects has been conducted thus far.

The literature regarding FFAs in the human SC is sparse, especially for odd-numbered FFAs, but it has been postulated that FFAs can be processed by stepwise coupling, not only of acetyl groups but also of propionyl groups and therefore it is possible to obtain odd-numbered FFAs [292]. The group of Norlén et al. [292] quantified low concentrations of odd-numbered FFAs in the SC of the inner human forearm. In another study, Norlén et al. [293] determined that these odd-numbered FFAs represent an endogenous component of the SC. However, it is thought that odd-numbered FFAs can also be generated from exogenous triglycerides via enzymatic reactions [294]. Additionally, Nicollier et al. [295] evaluated 14 volunteers with healthy skin and found odd-numbered FFAs with chain lengths ranging from 15 to 27 carbon atoms.

We investigated 12 even- and odd-numbered FFAs and found that odd-numbered FFAs comprised on average 20.7% of the total FFA content in group elderly/healthy,

	A			**B**		
	H_1	H_2	H_3	H_1	H_2	H_3
C15:0	12.000	0.000*	0.000*	11.999	0.001*	0.001*
C16:0	11.262	0.741	1.482	7.618	4.403	8.805
C17:0	11.987	0.013*	0.026*	11.994	0.006	0.012
C18:0	4.226	7.783	8.453	3.846	8.173	7.692
C18:1	10.406	1.600	3.200	11.999	0.001*	0.002*
C18:2	5.105	6.905	10.210	0.004*	11.996	0.008*
C19:0	2.566	9.441	5.132	0.000*	12.000	0.000*
C21:0	10.685	1.319	2.639	11.440	0.565	1.131
C23:0	11.947	0.053	0.107	11.992	0.008*	0.016*
C24:0	1.716	10.290	3.432	8.696	3.322	6.643
C25:0	3.604	8.405	7.207	0.126	11.876	0.252
C26:0	0.402	11.600	0.804	0.379	11.625	0.758

Table 9.2 | Bonferroni corrected P values derived from the Mann-Whitney U test. Columns represent different alternative null hypothesis (H_1–H_3) of Mann-Whitney U test statistic. H_1 and H_2 represent the one-sided test statistic, verifying whether FFA concentrations are greater (H_1) A: in subgroup elderly/healthy compared to subgroup young/ healthy and vice versa (H_2) respectively B: in subgroup young/healthy compared to subgroup young/diabetic and vice versa. H_3 denotes the two-sided Mann-Whitney U test, verifying whether FFA concentrations are different between corresponding subgroups. Asterisks denote statistically significant differences between corresponding subgroups; $^*P < 0.05$. Adapted Table 2 from [7].

about 23% in group young/healthy and 23.6% in group young/diabetic. The even-numbered FFAs C16:0, C18:0, C18:1 and C24:0 were the predominant types in our data pool. All of the FFAs reported here were previously shown to be abundant by Nicollier et al. [295] and the importance of C18:0 as well as C24:0 have been additional shown by Lampe et al. [296]. Large inter-individual variations in individual FFA concentrations were observed in the study.

In particular, the FFAs C18:1 and C18:2 differed markedly between subjects. This is in accordance with other studies using in vivo SC extraction methods [297]. In one of the aforementioned studies by Norlén et al. [292], inter-individual variations in almost all of the studied FFA concentrations in the SC were observed in the study population of 22 female students. This high variability may be caused by different amounts of surface sebum lipids, which vary significantly between individual subjects. These results were supported by a further study by Norlén et al. [293] that compared the FFAs from extracted skin surface samples with FFAs from SC tape stripping samples. Significantly more shorter-chain FFAs were found in the skin surface samples; this was probably due to the different number of sebaceous glands and their activity. Therefore, the authors postulated that only the longer FFAs with chain length of over 20 carbon atoms were of endogenous nature [293].

The obtained data showed large inter-individual differences in the short-chain FFAs too and thus supports this pre-mentioned hypothesis. In the current study, significantly lower levels of C15:0 and C17:0 were observed in elderly healthy subjects compared with younger healthy subjects. However, no significant differences of other FFAs investigated were found between these subgroups. The significant differences between these study groups in C15:0 and C17:0 may be explained by their origin. Recent studies have shown that next to fatty acid synthase, phytosphingosine (PHS) is a major source of odd-numbered FFAs. The authors converted PHS in mammalian tissues via several steps to 2-hydroxypalmitic acid and then to pentadecanoic acid (C15:0) via alpha-oxidation. Finally, elongation of C15:0 to C17:0 occurred via fatty acid synthase [267].

Nevertheless, further studies are necessary to show any correlations between PHS and aged skin. In addition, decreasing levels of C15:0 and C17:0 FFAs could be due to a pH alteration in enzymatic activity (e.g., propionyl-CoA decarboxylase) although potential mechanisms have not yet been identified [298]. However, according to our data the differences between young and aged healthy skin regarding C15:0 and C17:0 FFAs levels may be due to changes in the activity of PHS or fatty acid synthase, and not - like in other skin diseases, such as ichthyoses - due to defects in different elongases or transferases [299]. Whether induction or substitution by topical preparations containing these particular FFAs would be beneficial for aging or dry skin needs to be investigated further. These investigations will show if these FFAs alone or which other substances are essential for maintaining the complex system of young healthy skin [277]. Nonetheless, the results obtained correlate partially with those reported by Rogers et al. [300] who found that individual FFAs did not significantly decrease with age.

Several significant changes in FFAs were observed in the subgroup of the young diabetic subjects, compared with healthy subjects of the same age. These findings were not surprising because FFAs are known regulators of skin surface pH and other studies have found the skin pH to be significantly lower in diabetics [301, 302]. The lower levels of different FFAs in diabetics demonstrate a complex imbalance on SC composition and determine a functional impairment of barrier. Obviously a change in the regulation of keratinocytic proliferation and differentiation is the main reason for an increase of epidermal retention.

With this in mind, the use of oils with an appropriate range of fatty acids as an ingredient of the lipophilic phase in topical cosmetic products is recommended, both for conditioning aged skin but also for care therapy in diabetic patients [303]. However, clinical data to prove practical relevance over a long period of time are still rare [304, 305].

9.5 Conclusions and Outlook

Lipid composition studies of the SC are of great importance for understanding age- and disease-dependent changes in content and in molecular organization of epidermal barrier to develop appropriate cosmetic and medicinal products for recovery and protection.

In this study, we have aimed to identify even- and odd-numbered FFAs within the SC's intercorneocytic lamellar lipid structures and explore age- and diabetes-related changes in FFAs. We investigated the FFA compositions of 258 subjects in total, containing 110 subjects aged over 60 years (elderly/healthy), 110 subjects aged 18–40 (young/healthy), and 38 subjects with diabetes mellitus aged 18–40 (young/diabetic). From a bioinformatics perspective, we performed straightforward statistical analyses to detect and quantify differences in the FFA compositions.

In the future, it might be worthwhile to study if lipid pattern can be influenced by the supplementation of adapted mixtures of FFAs or by induction/normalization of FFAs synthesis, with the aim of strengthening the skin barrier function of the elderly or in subjects with diabetes mellitus [306]. In order to improve the developed bioinformatics approaches, it might be worthwhile to include further information of the participants like gender, medication, or life style aspects such as smoking and non-smoking. Based on these information, association studies could be performed to gain a more detailed insight into the underlying cause of the FFA differences.

In this chapter, we have shown that participants with diabetes mellitus have lower concentrations of different FFAs in the SC, resulting in functional impairments of their skin barrier. The risk of being affected by such chronic diseases like type 2 diabetes, cancer or cardiovascular diseases can depend on various dietary factors, among others.

In chapter 10, we will investigate the effect of different dietary protein sources and their influence on metabolic and functional parameters. In this second metabolomic study, we will analyse the serum metabolites of pigs, which serve as surrogates for humans. To perform structured analyses of the diverse metabolome data, we will present bioinformatics approaches based on machine learning to quantify statistical differences that allow us to interpret the effects of the different dietary protein sources.

10

Metabolic footprint in response to dietry proteins in a pig model

In this chapter, we will investigate the contributions and effects of different dietary proteins on the serum metabolite profile. We will compare the intake of lupin protein, lean beef, and casein, whereas only lupin is a plant based protein whereas lean beef and casein are animal based. Several studies associate the intake of plant proteins with a decreased risk in chronic diseases such as cardiovascular disease or stroke which is in contrast to the animal based proteins. My colleagues and I statistically analyzed the differences in the amino acid composition of the dietary proteins in the pigs' serum metabolites. Additionally, we developed a pipeline based on linear discriminant analysis combined with feature selection to detect sets of biomarkers from the serum metabolites and discriminate between the intake of the three different dietary protein sources.

In section 10.1, we will provide an overview of the impact of dietary protein intake to the risk of being affected by chronic diseases and we will present the goals of this study. In section 10.2, we will introduce the study design and we will present the analysis pipeline and give a detailed overview of the linear discriminant analysis to extract differences in the serum metabolites based on the three different protein intakes. In section 10.3, we will investigate the compositions of the serum metabolites and we will present significant differences in the three dietary groups. In sections 10.4, we will discuss the results from the metabolome analyses and the sets of biomarkers the bioinformatics analyses could detect to distinguish between the intake of the different dietary proteins. Finally, in section 10.5, we will conclude with the findings of the study and we will give an outlook for possible bioinformatics studies and future work.

The following sections are extracted from the publication of Schutkowski et al. 2019 "*Metabolic footprint and intestinal microbial changes in response to dietary proteins in a pig model*" [307].

10.1 Introduction

Dietary factors influence the risk of chronic diseases such as type 2 diabetes (T2D), cancer and cardiovascular diseases (CVDs). Many studies of the contribution of nutrients to the prevention or increased risk of these chronic diseases have focused on the role of fats and carbohydrates, including fibers. However, in the last few years, epidemiological studies have also identified associations between dietary protein sources and disease risks [308, 309].

Data from recently published meta-analyses of cohort studies showed that the intake of protein from red meat and processed red meat is associated with an increased risk of T2D [310, 311]; colorectal, pancreatic, laryngeal and breast cancers [312, 313]; as well as stroke, coronary heart disease, heart failure and hypertension [314, 315]. In contrast, the consumption of plant proteins is often associated with beneficial health effects. Meta-analyses of prospective studies reported an association between a high intake of plant proteins and a decreased risk of obesity, cardiovascular risk factors, CVD and stroke [314–316].

Although data from epidemiological studies suggest that plant proteins are generally healthier than proteins derived from animal sources, protein-rich foods usually consist of several components that might be relevant to health and disease. These factors must be considered when assessing the health benefits or potential risks of proteins. Plant and animal sources of proteins also exhibit substantial differences in the amounts and types of fats and carbohydrates. In addition, plant-based foods contain several phytochemicals, such as polyphenols, to which the beneficial effects of these foods may be attributed. In contrast, processed meat is a source of N-nitroso compounds, mutagenic heterocyclic amines and carcinogenic polycyclic aromatic hydrocarbons that are formed during food processing steps such as frying and grilling [317]. Other issues that limit assumptions on the causal relationships between dietary proteins and health are lifestyle factors associated with the consumption of certain foods. For, example, in contrast to their meat-eating counterparts, vegetarians generally consume healthier diets, are usually nonsmokers, show higher physical activity levels, are less obese, and are more health conscious [318, 319].

Interventional studies that use isolated proteins and control for interfering health-affecting factors are required to obtain data on the actual contributions of dietary proteins to human health. However, the number of interventional studies examining the roles of certain dietary proteins in health is limited. Most human and animal studies have investigated the health potential of soybean protein and observed hypolipidemic effects [320, 321]. Comparable beneficial effects on serum triglyceride and cholesterol levels in humans and animals have also been observed in interventional studies using lupin proteins [322–324]. Strikingly, the majority of the human intervention studies that have investigated the roles of dietary proteins in health are primarily limited to the measurements of classical risk factors, such as blood pressure and blood lipid levels, in patients suffering from various disease conditions, including obesity, T2D, hypertension and hyperlipidemia [324–326]. Intervention studies providing data from comprehensive analyses that have evaluated the changes in metabolites in response to dietary proteins are rare.

The aim of the current study was to compare the effects of three dietary protein sources (lupin protein, lean beef and the milk protein casein) relevant to human nutrition and to describe their abilities to change metabolic and functional parameters. The study focused on changes in amino acids and amino acid derivatives in response to the different dietary proteins, an assessment of the metabolic effects of the dietary proteins and the identification of biomarkers indicative of the type of protein ingested. We further aimed to study the effects of these dietary proteins on modulating cardiovascular risk factors and the composition of the gut microbiome. Pigs were used as an animal model because the morphology and physiology of the gastrointestinal system, the ingesta transit times and the digestive efficiencies are comparable to humans [327], and a closer similarity between the pig and human gut microbiomes than between the mouse and human gut microbiomes has been observed [328]. Pigs and humans are both omnivorous and share greater similarities regarding the eating behavior, the metabolism, and the anatomy and physiology of the gastrointestinal tract compared to rodents [329]. In addition, intervention trials using pigs can be conducted under strictly controlled conditions, such as standardized food intake and physical activity.

10.2 Materials and Methods

In this section, we will describe the experimental design, the quantification of metabolomics data, and the bioinformatics approaches to analyse the diverse data. Further information about additional experimental analysis such as the isolation of RNA, or the profiling of microbial DNA from fecal samples can be obtained from the publication [307]. In subsection 10.2.1, we will introduce the study design and the treatment of the pigs. In subsection 10.2.2, we will present the composition of the protein sources that were fed to the pigs. In subsection 10.2.3, we will learn about the sample collection, especially the collection of blood samples to measure the concentration of serum metabolites. In subsection 10.2.4, we will investigate the bioinformatics approaches for the analysis of variance (ANOVA) to compare concentrations of serum metabolites between the three dietary groups. Additionally to the published study, in subsection 10.2.5, we will present a pipeline based on the linear discriminant analysis combined with feature selection to detect characteristic amino acid combinations distinguishing the pigs' diets.

10.2.1 Animals and study design

The experimental procedure was performed according to the established guidelines for the care and handling of laboratory animals [330]. The study was approved by the local council of Saxony-Anhalt (Landesverwaltungsamt, Halle (Saale), Germany; approval number: H1-4/44G). The pigs were individually housed in pens in an environmentally controlled facility with a temperature of 20°C, relative humidity of 55%–60%, and light from 6:00 a.m. to 6:00 p.m. Water was available ad libitum from a nipple drinking system during the entire study.

Forty-five 13-week-old female crossbred pigs [(German Landrace × Large White) × Pietrain] with an initial body weight of 33.7±2.9 kg (mean ± S.D.) were randomly assigned to 3 groups of 15 pigs each. Group 1 received a lupin protein isolate (Prolupin, Grimmen, Germany) as source of a plant protein, group 2 was fed cooked lean beef (Schirmer & Partner, Seelitz, Germany), and group 3 received casein (Max Walter Handelsvertretungen, Leonberg, Germany) as animal sources of proteins. The experimental diets were calculated to attain comparable contents of crude protein and crude fat. Thus the dietary protein sources were added to a standardized basal diet in amounts of 130-150 g per kg diet. A detailed description of the composition of the basal diet is provided in Supplementary Table 1. The basal diet was supplemented with 4.7 g/kg L-lysine HCl, 2 g/kg DL-methionine, 1.5 g/kg L-threonine and 0.3 g/kg Ltryptophan to meet the amino acid requirements of growing pigs according to the recommendations of the Society of Nutrition Physiology [331]. Minerals and vitamins were added as a commercial premix (Mineral feed, Basu, Bad Sulza, Germany) in amounts of 5 g/kg to meet the recommendations of the National Research Council [332]. The diets were administered for 4 weeks in strictly controlled amounts to prevent differences in feed intake. Pigs were weighed weekly.

10.2.2 Characterization of the experimental proteins

Defatted protein isolated from *Lupinus angustifolius* and casein were not processed. Lean beef was minced and cooked at 100°C for 90 min. Crude protein, fat and ash contents of the dietary proteins were determined using standard methods [333]. Fatty acid analysis was performed using gas chromatography (GC-17V3; Shimadzu Corporation, Kyoto, Japan) equipped with a flame ionization detector and an autosampler (AOC-5000), as described [334]. One GC procedure was required to analyze the fatty acid methyl ester (FAME) distribution of the samples. This method determined the identity and general fatty acid distribution of 4–22 carbon length fatty acids (including straight and branched structures) using a fused-silica capillary column DB-225 ms (30 m, 0.25 mm, i.d. with 0.2 µm film thickness; Jand W, Scientific, USA) and H_2 as carrier gas. Fatty acid concentrations were expressed as percentage of the total area of all FAMEs (% of total

FAME) using GC solution software version 2.3 (Shimadzu). A detailed description of the composition of the experimental proteins is given in Supplementary Table 2 [307].

10.2.3 Sample collection

Blood samples were collected from each pig at the beginning and end of the experiment by puncturing the external jugular vein. Blood was collected in Vacuette Z clot activator tubes (Greiner Bio-One, Kremsmünster, Austria) and centrifuged at 1100×g for 10 min at 4°C to obtain serum for the metabolite analysis. For the determination of plasma glucose levels, blood was collected in NaF-coated tubes. Baseline serum concentrations of all metabolites are summarized in Supplementary Table 4[307]. No significant differences were observed in the baseline levels of any of the metabolites. After 4 weeks, the pigs were anesthetized and euthanized by exsanguination 5 h after their last meal. The liver was harvested, and aliquots were snap-frozen in liquid nitrogen and stored at -80°C until further analysis. A 10-cm section of the duodenum (starting 15 cm behind the pyloric part) was excised, washed several times with a cold NaCl solution (0.9%) and cut lengthwise to collect the duodenal mucosa. The intestinal mucosa was harvested by scraping the surface of the small intestine. Feces were collected from the rectum. Spot urine samples were also collected. The pH of the urine samples was measured using a pH electrode. Feces and urine samples from each pig were frozen at -20°C until further analysis.

10.2.4 Statistical analysis

Statistical analyses were performed using the programming language R [288]. Before performing the statistical tests, data were logarithmically transformed, thus exhibiting a normal distribution. Hence, values are presented as the means±S.D. of $\texttt{log}_{10}$ data. First, Levene's test was used to evaluate the equality of variances, called homoscedasticity, for all parameters measured. The results of the Levene tests and subsequent correction for multiple testing with the Benjamini–Yekutieli method showed that all P values were greater than 0.5, thus suggesting homoscedasticity.

Afterwards, log data were analyzed using one-way analysis of variance (ANOVA). The resulting P values from each ANOVA were also corrected for multiple testing using the Benjamini–Yekutieli method. If the adjusted P values revealed significant effects ($P < 0.05$), means of the three groups were compared using Tukey's multiple-comparison test. Logarithmically transformed means were considered significantly different at $P < 0.05$.

10.2.5 Linear discriminant analysis

An LDA was performed to investigate the differences in amino acid concentrations between the three groups and combined with feature selection to decipher either single amino acids or pairs of amino acids that differentiated between the three groups. The procedures used to select the amino acids and to validate the classification approach are described below.

For the purpose of classification the pig samples were defined as the set $\{(\underline{x}_1, y_1), (\underline{x}_2, y_2), ..., (\underline{x}_N, y_N)\}$, with $\underline{x}_n \in \mathcal{R}^D$ representing the logarithm of concentrations of D selected amino acids, where $n \in \{1, \ldots, N\}$, while N denoting the total number of pigs in the study, and y representing the dietary protein source of each pig, respectively its group. Thus, we defined the set of groups as $C = \{$lupin, beef, casein$\}$ with $y_n \in C$. The number of pigs in each group was defined as N_c with $c \in C$ and $\sum_{c \in C} N_c = N$. Furthermore, it was assumed that the vectors $\underline{x}_n$ follow a multivariate normal distribution. Hence, we were able to define the probability of the class posterior $p(y_n = c|\underline{x}_n, \theta_c)$ for all $c \in C$ and $\underline{x}_n$ $n \in \{1, \ldots, N\}$ with θ_c representing the model parameters for group c.

$$\begin{aligned} p(y_n = c|\underline{x}_n, \theta_c) &= \frac{p(\underline{x}_n|y_n = c, \theta_c) \cdot p(y_n = c|\theta_c)}{\sum_{c \in C} p(\underline{x}_n|y_n = c, \theta_c) \cdot p(y_n = c|\theta_c)} && (10.1) \\ &\propto \mathcal{N}(\underline{x}_n|\underline{\mu}_c, \Sigma_c) \cdot \pi_c && (10.2) \end{aligned}$$

, with $p(y_n = c|\theta_c) = \pi_c$ denoting the class probability. Based on Leven's test we assumed equal variance and thus a shared covariance matrix $\Sigma = \Sigma_c$, $\forall c \in C$. Hence, the posterior probability can be written as

$$\begin{aligned} p(y_n = c|\underline{x}_n, \theta_c) &\propto \mathcal{N}(\underline{x}_n|\underline{\mu}_c, \Sigma) \cdot \pi_c && (10.3) \\ &\propto \pi_c \cdot \texttt{exp}\left(\underline{\mu}_c^T \Sigma^{-1} \underline{x}_n - \frac{1}{2}\underline{x}_n^T \Sigma^{-1} \underline{x}_n - \frac{1}{2}\underline{\mu}_c^T \Sigma^{-1} \underline{\mu}_c\right) && (10.4) \\ &= \texttt{exp}\left(\underline{\mu}_c^T \Sigma^{-1} \underline{x}_n - \frac{1}{2}\underline{\mu}_c^T \Sigma^{-1} \underline{\mu}_c + \texttt{log}\pi_c\right) && (10.5) \\ &\underbrace{\cdot\, \texttt{exp}\left(-\frac{1}{2}\underline{x}_n^T \Sigma^{-1} \underline{x}_n\right)}_{x_n^{const}} && (10.6) \end{aligned}$$

As shown in Eq. 10.6, the term x_n^{const} is independent of group c and thus can be removed.

$$p(y_n = c|\underline{x}_n, \theta_c) \propto \mathtt{exp}\left(\underline{\mu}_c^T \Sigma^{-1} \underline{x}_n - \frac{1}{2}\underline{\mu}_c^T \Sigma^{-1} \underline{\mu}_c + \mathtt{log}\pi_c\right) \tag{10.7}$$

The parameters of the LDA model like the mean vectors $\underline{\mu}_c$ and the covariance matrix Σ were estimated in an unbiased manner as

$$\hat{\underline{\mu}}_c = \frac{1}{N_c}\sum_{n=1}^{N} \underline{x}_n \delta(y_n = c) \tag{10.8}$$

$$\hat{\Sigma}_c = \frac{1}{N_c - 1}\sum_{n=1}^{N} (\underline{x}_n - \hat{\underline{\mu}}_c)(\underline{x}_n - \hat{\underline{\mu}}_c)^T \delta(y_n = c), \tag{10.9}$$

$$\hat{\Sigma} = \frac{\sum_{c \in C}(N_c - 1)\hat{\Sigma}_c}{N - |C|} \tag{10.10}$$

$$\pi_c = \frac{N_c}{N} \tag{10.11}$$

Using the trained model each data point was able to be classified based on its maximal posterior probability. Thus, the model predicted a new group label $\hat{y}_n$ for each data point $\underline{x}_n$ as

$$\hat{y}_n = \underset{c \in C}{\operatorname{argmax}}\; p(y_n = c|\underline{x}_n, \theta_c) \tag{10.12}$$

$$= \underset{c \in C}{\operatorname{argmax}}\; \underline{\beta}_c^T \underline{x}_n + \gamma_c. \tag{10.13}$$

The classifier was validated using a leave one out cross-validation (LOOCV) approach, which is referred to as the outer LOOCV. In each run of this outer LOOCV, 44 data points were used for feature selection, namely, the selection of amino acids, and the left out data point was used to validate the selected trained model. As a model, we refer the results of an LDA based on the selected amino acids. Feature selection, the process used to choose the best combination of amino acids for the LDA, was performed in an internal LOOCV based on the remaining 44 data points. This internal LOOCV used 43 data points in each run to train an

LDA model for each single amino acid and for all pairs of amino acids. The data point that was left out was used to measure the performance of each model. The classification error Eq. 10.14, which is the mean of misclassified data points, was used to compare the models after the internal LOOCV. It is defined as

$$\overline{err} = \frac{1}{N}\sum_{n=1}^{N} \delta(y_n = \hat{y}_n). \tag{10.14}$$

Thus, in each run of the outer LOOCV, models with a minimum classification error based on the internal LOOCV were selected. At the end of each run of the outer LOOCV, each left-out data point was classified by each selected model, followed by calculating each model's classification rate. Models showing the highest classification rate were selected. Based on this procedure, six pairs of amino acids were obtained that could distinguish the three groups of pigs the best.

10.3 Results

In this section, we will present the results of the statistical analysis of the amino acid composition of the dietary proteins and the metabolites measured in the serum. We will also investigate the LDA results to identify pairs of amino acids in the serum which enable a differentiation of the pigs based on their diet.

In subsection 10.3.1, we will describe the differences in the concentrations of amino acids and other metabolites quantified from the pigs' serum and we will describe the biomarkers provided by the LDA analysis. In subsections 10.3.2, we will present the results of the methylation analysis based on SAM and SAH concentrations. In subsection 10.3.3, we will describe the relative mRNA expressions of genes involved in DNA methylation. In subsection 10.3.4, we will investigate the impacts of dietary proteins on health-relevant factors such as serum mineral concentrations or the relative mRNA expression of genes involved in apoptosis and stress responses.

10.3.1 Amino acid composition of the dietary protein and serum levels of amino acids and metabolites

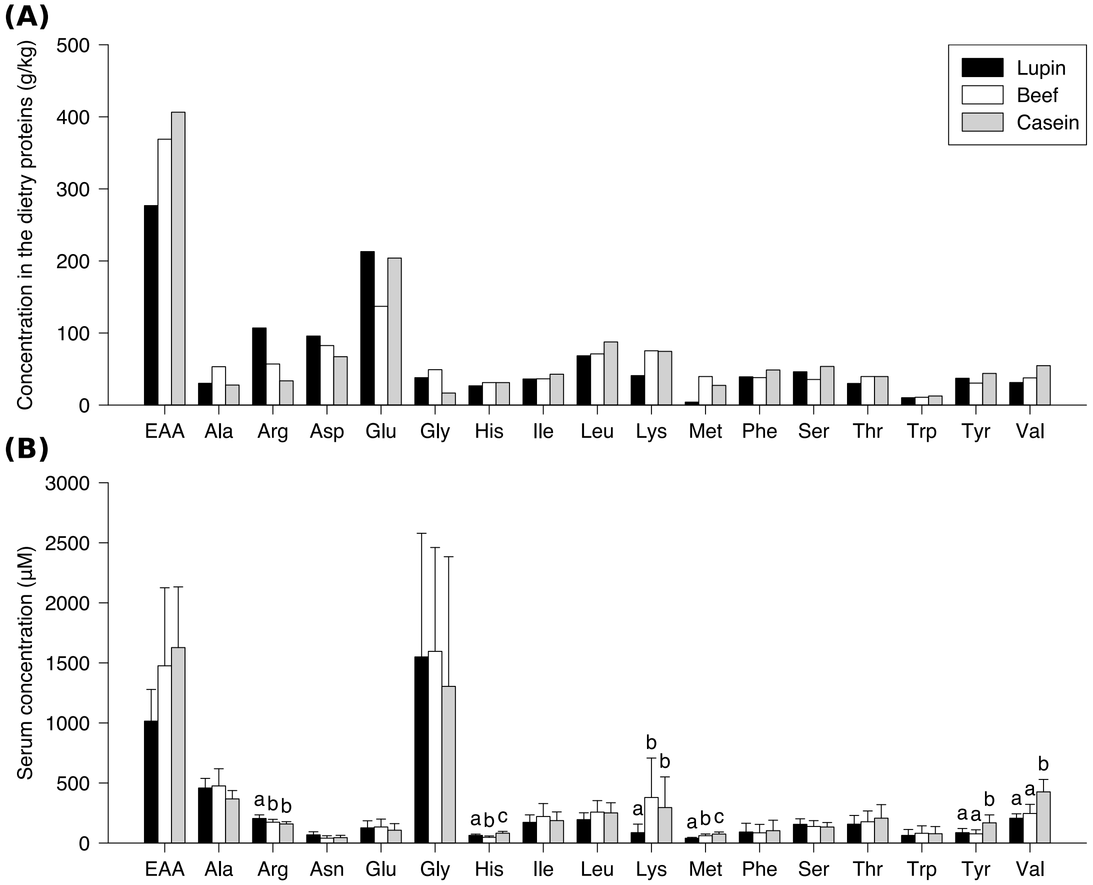


Figure 10.1 | Composition and concentrations of amino acids in dietary proteins and serum of pigs. (A) Amino acid composition of the dietary proteins and (B) serum amino acid concentrations in pigs fed lupin, beef or casein for 4 weeks. The results are presented as means ± S.D., n = 15. Levene's test for homoscedasticity showed no significant heterogeneity of variances. Metabolite concentrations were analyzed using one-way ANOVA followed by Tukey's test. a, b and c: Means without a common letter differed significantly ($P < 0.05$). EAA, indispensable amino acids. Reprinted Figure 1 from [307].

First, we analyzed the amino acid composition of the dietary proteins. Fig. 10.1A shows marked differences in the alanine, arginine, glutamate, lysine, methionine, tyrosine and valine concentrations between the dietary proteins. The highest arginine content was detected in the lupin protein isolate, the highest methionine content was observed in the cooked beef, and casein was characterized by the highest valine content. Both proteins from animal origin, beef protein and casein, were characterized by higher lysine contents than lupin protein. We quantified the free amino acid concentrations in the 5-h fasting serum samples collected from the pigs to assess whether dietary proteins alter the serum concentrations of nonprotein-bound amino acids.

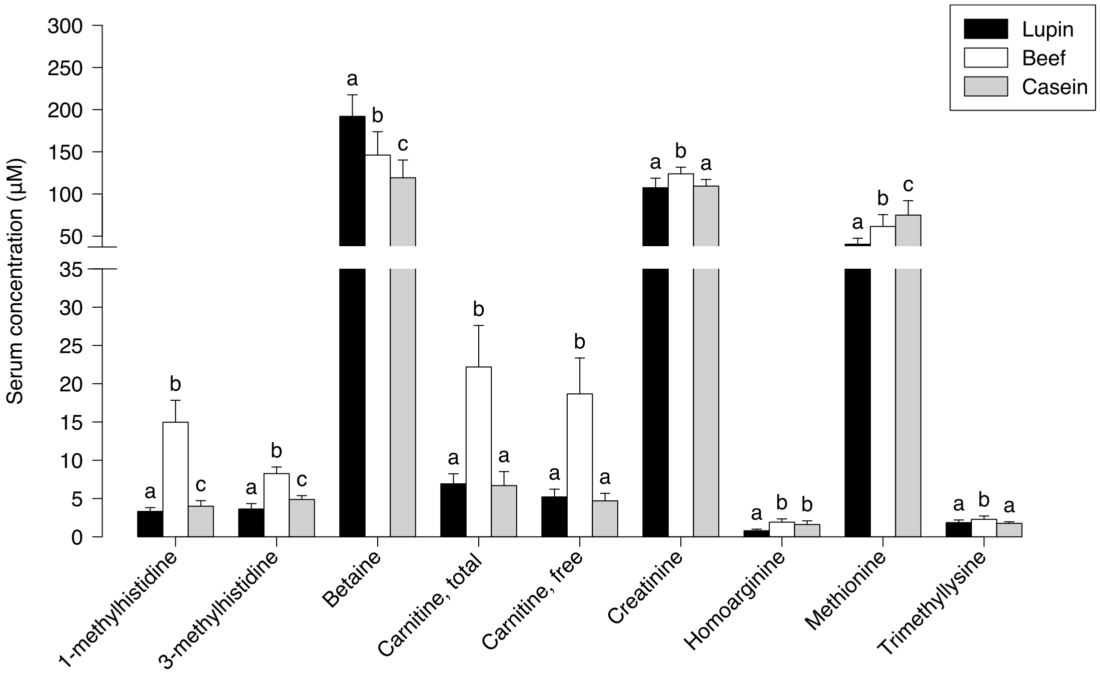


Figure 10.2 | Concentrations of characteristic serum metabolites of pigs after ingestion of different dietary proteins. Serum concentrations of 1- and 3-methylhistidine, betaine, total carnitine (sum of free and acetyl- and acyl carnitines) and free carnitine, homoarginine, methionine and trimethyllysine in pigs fed lupin, beef or casein for 4 weeks. The results are presented as means ± S.D., n = 15. Levene's test for homoscedasticity showed no significant heterogeneity of variances. Logarithmically transformed data were analyzed using one-way ANOVA followed by Tukey's test. a, b and c: Means without a common letter differed significantly ($P < 0.05$). Reprinted Figure 2 from [307].

The serum concentrations of indispensable amino acids partially reflected the contents of indispensable amino acids in the dietary proteins (Fig. 10.1A and B). Significant differences in serum amino acid concentrations between the three groups of pigs were observed for arginine, histidine, lysine, methionine, tyrosine and valine, where the lupin protein group exhibited the lowest concentrations of lysine and methionine (Fig. 10.1B). The serum concentrations of the other amino acids were similar (Fig. 10.1B).

A set of additional amino acid derivatives was quantified using HPLC and GC–MS/MS to identify serum metabolites that were characteristic of the type of protein ingested. Significantly higher serum betaine concentrations were observed in the lupin protein group than in the beef protein and casein groups (Fig. 10.2).

Serum concentrations of six amino acid metabolites were higher in the group fed beef protein than in the other two groups: 1-methylhistidine and 3-methylhistidine, total and free carnitines, creatinine and trimethyllysine (Fig. 10.2). The serum homoarginine concentrations were comparable between the groups fed beef protein and casein but higher than those in the lupin protein group. Serum levels of choline, citrulline, creatine, asymmetric and symmetric dimethylarginine, dimethylglycine, methionine sulfoxide, ornithine, taurine and TMAO did not differ between the treatment groups (Table 10.1).

	Lupin	**Beef**	**Casein**	***P*-value**
Amino acid derivatives (μM)				
Choline	16.85±1.67	15.37±1.76	15±2.78	ns
Citrulline	93.34±18.03	79.33±26.75	92.76±22.39	ns
Creatine	237.53±112.49	243.07±98.11	269.81±110.52	ns
Dimethylarginine asym.	1.49±0.18	1.64±0.26	1.50±0.23	ns
Dimethylarginine sym.	0.89±0.16	0.91±0.12	0.93±0.14	ns
Dimethylglycine sym.	7.76±2.99	7.77±2.71	6.37±1.42	ns
Methionine sulfoxide	2.13±0.45	2.32±0.60	2.58±1.01	ns
Ornithine	124.02±68.41	96.04±61.30	86.21±54.60	ns
Taurine	161.68±53.48	186.12±82.22	171.69±66.37	ns
TMAO	2.85±2.17	4.02±2.98	2.62±1.94	ns
Glucose metabolism				
Glucose (mM)	4.80±0.47	4.85±0.42	4.86±0.32	ns
Insulin (mg/L)	0.04±0.02	0.05±0.03	0.04±0.03	ns
HOMA index	0.01±0.01	0.01±0.01	0.01±0.01	ns
Lipid metabolism				
Cholesterol (mM)	2.94±0.30	2.78±0.26	2.87±0.24	ns
Triglycerides (mM)	0.46^a±0.09	0.34^b±0.09	0.32^b±0.09	< 0.05
LDL (mM)	1.32±0.16	1.31±0.16	1.37±0.19	ns
HDL (mM)	1.82±0.23	1.58±0.16	1.61±0.17	ns
HDL/LDL Ratio	0.73±0.07	0.83±0.07	0.86±0.15	ns
Inflammation markers				
CRP (mg/L)	23.02±6.23	21.52±7.01	22.26±6.53	ns
Rel. mRNA expr. of DNA methyltransferase and histone demethylase genes				
Dnmt1	1.00±0.31	1.01±0.28	1.02±0.38	ns
Dnmt3a	1.00±0.43	0.87±0.22	1.10±0.33	ns
Kdm3a	1.00±0.15	0.99±0.34	1.34±0.41	ns
Rel. mRNA expr. of genes related to intestinal health in the intestinal mucosa				
Birc5	1.08±0.37	1.08±0.57	1.10±0.47	ns
Eif4ebp1	1.04±0.20	1.12±0.31	1.16±0.17	ns
Hmox1	1.18±0.98	1.26±1.21	0.87±0.66	ns
Sod1	1.04±0.28	1.13±0.36	1.22±0.26	ns
Tp53	1.07±0.47	0.83±0.32	1.24±0.57	ns
Growth factors				
IGF-1 (mg/L)	217.71±45.52	245.40±46.39	236.31±82.08	ns

Continued on next page

	Lupin	**Beef**	**Casein**	***P*-value**
Serum mineral concentrations				
P (mM)	2.82±0.13	2.74±0.23	2.88±0.14	ns
Ca (mg/L)	108.70±6.09	107.89±8.04	111.18±9.76	ns
Mg (mg/L)	18.84±2.06	19.97±2.74	20.71±2.07	ns
Fe (mg/L)	3.50±0.88	3.70±0.91	4.07±0.99	ns
Zn (mg/L)	0.63±0.18	0.70±0.18	0.82±0.50	ns
Se (µg/L)	144.47±15.63	143.48±14.95	159.50±19.74	ns

Table 10.1 | Serum concentrations of amino acid metabolites, health risk parameters, relative mRNA concentrations, mineral concentrations and plasma glucose concentrations. Serum concentrations of several metabolites and minerals in pigs fed lupin, beef or casein for 4 weeks. C-reactive protein (CRP), IGF-1 and TMAO levels were quantified. The results are presented as means ± S.D., n = 15. Levene's test for homoscedasticity showed no significant heterogeneity of variances. Metabolite concentrations were analyzed using one-way ANOVA followed by Tukey's test. a, b and c: Means not sharing a common letter differed significantly ($P < 0.05$). Adapted Table 1 from [307].

An LDA was conducted to identify a subset of biomarkers that was useful for discriminating between the intake of the three dietary protein sources. The LDA analyzes relative differences between the three groups and is a widely used method for pattern classification.

As depicted in Fig. 10.3, the three groups showed well-defined clusters for six combinations of two biomarkers each. Combination 1 comprises betaine and 3-methylhistidine; combination 2, homoarginine and 3-methylhistidine; combination 3, methionine and 1-methylhistidine; combination 4, methionine and 3-methylhistidine; combination 5, methionine and free carnitine; and combination 6, homoarginine and free carnitine. Thus, these combinations are potential biomarkers to classify samples from animals fed the corresponding dietary protein source.

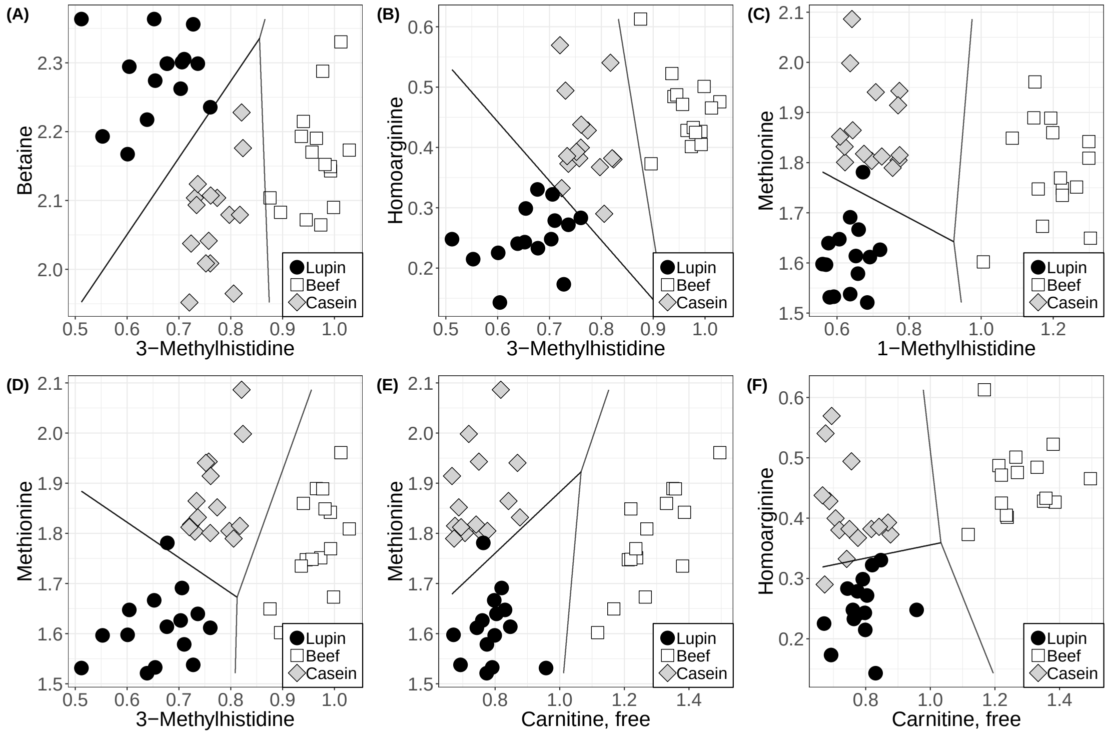


Figure 10.3 | Results of the LDA based on logarithmically transformed serum concentrations [$\log_{10}$(serum concentration in µM+1)]. Straight lines show the classification borders defined by LDA to distinguish between the three groups of pigs that were fed lupin protein (black points), beef (white rectangles) and casein (gray diamonds). (A–F) Classification results based on log concentrations of pairs of amino acids. The x- and y-axes are labeled with the selected pairs of amino acids defined by the feature selection within the LOOCV analysis. All three groups are nearly perfectly distinguishable based on six different pairs of amino acid derivatives. Reprinted Figure 3 from [307].

10.3.2 SAM and SAH concentrations in liver and serum homocysteine levels

Fig. 10.4C illustrates the relationship between the methyl donators methionine and betaine and the synthesis of dimethylglycine, SAM, SAH and homocysteine. Because the dietary proteins specifically differed in their impact on serum levels of methionine and betaine, molecules that are involved in one-carbon metabolism, we analyzed the SAM:SAH ratio in the liver and serum homocysteine concentrations as marker of the methylation capacity. The SAM:SAH ratio in liver was higher in the group fed the lupin protein than in groups fed beef or casein (Fig. 10.4A), although this difference did not reach statistical significance. The lowest serum homocysteine concentration was observed in the group fed the lupin protein, followed by the groups fed the beef and casein (Fig. 10.4B).

10.3.3 Hepatic methylation

Methylation processes depend on the availability of methyl groups. As the diets of the three groups of pigs differed in their concentrations of methyl group donators methionine and betaine, we analyzed mRNA levels of genes involved in DNA methylation. Here, the expression of mRNAs encoding the hepatic DNA methyltransferases Dnmt1 and Dnmt3a did not differ between the three treatment groups (Table 10.1). We additionally analyzed the relative mRNA expression of the histone demethylase Kdm3a and did not detect significant differences between the three groups of pigs (Table 10.1).

We analyzed the global liver methylation status of histone H4 using MALDI-TOF mass spectrometry to assess whether differences in the methyl group donators methionine and betaine induce changes in the histone methylation pattern. However, these analyses did not reveal significant differences in the global histone H4 methylation pattern between the three groups (Fig. 10.4D). The findings were confirmed by a site-specific Western blot analysis, which did not show differences in the monomethylation of lysine 4 in histone H3 (H3K4me1), the dimethylation of lysine 4 in histone H3 (H3K4me2) and the trimethylation of lysine 4 in histone H3 (H3K4me3) between the treatment groups (Fig. 10.4E).

10.3.4 Impacts of dietary proteins on health-relevant factors

We analyzed fasting glucose and insulin levels and calculated the homeostatic model assessment (HOMA) index in response to the consumption of lupin protein, beef protein and casein to elucidate the effects of dietary proteins on glucose metabolism but did not observe differences between the three diets (Table 10.1). Furthermore, serum concentrations of total, LDL and HDL cholesterol were comparable between the three groups, whereas the concentration of triglycerides was significantly higher in pigs fed lupin protein compared to those fed beef and casein (Table 10.1). Other factors associated with diseases, such as the inflammatory marker C-reactive protein (CRP) and the proliferative factor IGF-1, also did not differ between the groups (Table 10.1). However, the urine of pigs that were fed beef had a significantly lower pH value than the urine of pigs fed lupin protein or casein (pH values of urine: lupin protein: 6.12±0.18a; beef: 5.82±0.27b; casein: 6.20±0.25a; $P < 0.05$).

The serum concentrations of calcium, iron, magnesium, phosphorus, selenium and zinc were quantified to elucidate whether the consumption of the different dietary proteins was associated with changes in mineral concentrations. The analyses failed to show any differences in serum mineral concentrations between the groups (Table 10.1). Because the intake of red meat is associated with an increased colon cancer risk, we analyzed the mRNA expression of intestinal genes involved in apoptosis and stress responses. However, neither the expression of genes involved in regulating apoptosis, such as survivin (Birc5) and the tumor suppressor Tp53, nor that of stress responsive genes, including heme oxygenase 1 (Hmox1), Eif4eb1 (4ebp1) and superoxide dismutase 1 (Sod1), was differently regulated by the intake of lupin protein, beef and casein (Table 10.1). As shown in Table 10.1, the expression of these genes was not influenced by the dietary treatment.

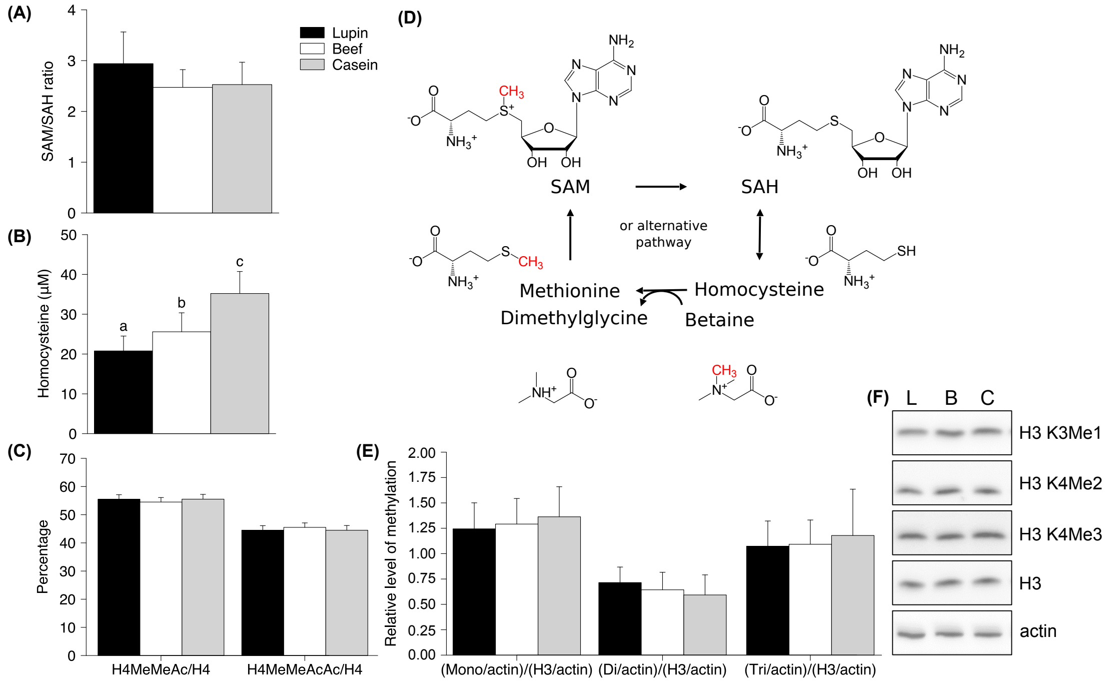


Figure 10.4 | Methylation analysis and one carbon cycle . (A) SAM/SAH ratio in the liver; (B) serum concentrations of homocysteine; (C) schematic representation of the one-carbon cycle; (D) percentage of global methylation of histone H4 and (E) Western blot analysis of histone H3K4me1, H3K4me2 and H3K4me3 levels in the liver tissues of pigs fed lupin protein, beefor casein for 4 weeks. The results are presented as changes in methylation relative to lupin-fed pigs. (F) Representative Western blot images. All results are presented as means ± S.D., $n = 15$. Levene's test for homoscedasticity showed no significant heterogeneity of variances. Data were analyzed using one-way ANOVA followed by Tukey's test. a, b and c: Means without a common letter differ significantly ($P < 0.05$). Reprinted Figure 4 from [307].

10.4 Discussion

Dietary proteins play a crucial role in providing indispensable amino acids and nitrogen to synthesize structural and functional proteins in the organism. In addition to providing amino acids for protein synthesis, dietary proteins and their degradation products may modulate the risk of chronic diseases. As chronic diseases usually develop because of metabolic changes or possibly alterations in the gut microbiome, we investigated the profiles of amino acid metabolites in serum and the fecal microbiota in response to three different dietary protein sources that may exert beneficial or detrimental effects on health.

Here, the dietary protein sources changed the serum concentrations of free amino acids and their derivatives. Metabolites that differed significantly between the groups were the nonprotein-bound amino acids arginine, histidine, lysine, methionine, tyrosine and valine and the amino acid derivatives betaine, creatinine, homoarginine, trimethyllysine, carnitines and methylhistidines. A classification analysis was conducted to elucidate whether the dietary proteins induce a distinguishable metabolic footprint. Data obtained from the LDA analysis identified six sets of two serum biomarkers each that discriminated between the intake of lupin protein, beef or casein. The sets of metabolites whose combination was useful for distinguishing between the dietary proteins included 1- and 3-methylhistidine, betaine, methionine, free carnitine and homoarginine.

1-Methylhistidine is part of the histidine dipeptide anserine, which is a natural antioxidant and is mainly present in skeletal muscle [335]. As the diet is the main source of histidine dipeptides in humans [335], 1-methylhistidine can be used as a marker for the intake of meat. 3-Methylhistidine is produced by the posttranslational methylation of histidine in muscle-derived actin and myosin. In contrast to 1-methylhistidine, 3-methylhistidine is either derived from the diet and or produced endogenously during muscle protein turnover and degradation [336].

Under our experimental conditions, we did not expect differences in the endogenous muscle protein turnover; thus, we propose that the differences in serum 3-methylhistidine concentrations were attributed to differences in the dietary protein intake. This hypothesis is supported by our findings that the intake of beef, which mainly consists of skeletal muscle proteins, results in a significant increase in the serum 1- and 3-methylhistidine concentrations.

As the diet is the only source of 1-methylhistidine for humans, we propose that the quantification of 1-methylhistidine or the combination of both methylhistidines will provide a more precise indication of the true meat intake compared to 3-methylhistidine. Researchers previously proposed that 24-h urinary measurements of 1- and 3-methyhistidine are necessary for the verification of meat intake in humans [337, 338]. Here, these metabolites were also present in high concentrations in serum after beef intake. Thus, we propose that 24-h urine sampling may not be required to quantify methylhistidine concentrations as marker of meat intake.

In the current study, we detected increased levels of carnitine and its precursor trimethyllysine in the serum of pigs fed the beef. Although carnitine can be synthesized endogenously, meat intake significantly increases circulating carnitine levels [339]. This finding has also been confirmed in a human cohort study that reported lower serum carnitine levels in vegans than in individuals consuming food of animal origin [340]. The carnitine content of meat normally ranges from 300 to 500 µmol/100 g, depending on the cut and cooking method [341, 342].

Data from that association study confirmed that carnitine can be used to distinguish between protein sources, although carnitine is produced endogenously. Another biomarker is trimethyllysine, which is closely linked to carnitine and is important in animals as the metabolic precursor of carnitine [343]. To our knowledge, trimethyllysine has not currently been used as a marker of meat intake.

Homoarginine is a nonproteinogenic amino acid, and several studies have shown that, due to its structural similarity to arginine, it can serve as an alternative substrate for NO synthase and might thus be involved in the mechanism regulating blood pressure [344]. Recent studies have suggested a role for homoarginine in vascular function and identified a possible association between low homoarginine concentrations and an increased CVD risk [345, 346]. In the present study, significantly lower serum homoarginine concentrations were observed in lupin-protein-fed pigs compared to the other two groups, which, together with the higher serum homocysteine content in beef and casein fed animals, suggest a higher CVD risk.

As the lupin-protein-fed pigs had lower concentrations of homoarginine and homocysteine, we postulate that lupin protein may positively impact CVD risk, in contrast to beef and casein. Currently, lupin protein is used in only small quantities in human nutrition. Results from our study can contribute to force the use of lupin proteins as an alternative for soy bean protein or proteins from animal sources.

In addition to the amino acid metabolites, parameters such as urine pH have been used to corroborate the finding that high meat intake leads to renal acid load and an increase in renal protein excretion [347, 348].

Although 1- and 3-methylhistidine, homoarginine, carnitine and trimethyllysine are markers of meat intake, the serum betaine level is a biomarker of the intake of plant proteins, in particular lupin protein. Betaine is an essential osmolyte that accumulates in most plant tissues to regulate cell volume, and it also supplies methyl groups to convert homocysteine to methionine [349]. Betaine is present at high concentrations in wheat products, pulses, potatoes, spinach, broccoli, beet and cabbage, but it is not a marker for vegetable intake in general, as most vegetables, particularly fruits, contain very low amounts of betaine [350, 351].

In addition, lupin protein intake was associated with low serum methionine concentrations and a high arginine concentration. Here, we were not able to identify biomarkers of dairy protein intake. For dairy intake, the two odd-chain fatty acids pentadecanoic acid (15:0) and heptadecanoic acid (17:0) are widely used as intake markers [352]. Nevertheless, markers of dairy protein intake are lacking.

Interestingly, among the metabolites that are potential biomarkers of dietary meat and plant protein intake, methionine and betaine are physiological methyl donors that play important roles in the one-carbon cycle. Differences in the concentrations of metabolites from the one-carbon cycle might alter the methylation capacity and subsequently result in DNA or histone hypo- or hypermethylation [353]. Based on accumulating evidence, dietary components that influence the supply of methyl groups modulate DNA or histone methylation patterns and in turn gene transcription [354].

SAM, which is generated in the one-carbon cycle, serves as an essential cofactor for almost all DNA and posttranslational protein methylation reactions. Diets rich in methyl group donors, including methionine, choline, betaine, folic acid or vitamin B12, alter global and gene-specific promotor DNA or histone methylation by affecting the methylation capacity of the cell and thus the activity of DNA or histone methyltransferases [355]. Epigenetic changes have increasingly been correlated with metabolic disorders, including obesity, T2D and CVD [356]. In addition, differences in DNA methylation in patients with cancer are related to methyl donor availability.

Betaine serves as a methyl donor in the one-carbon cycle that converts homocysteine to methionine and has been used to lower serum homocysteine concentrations [357]. Betaine (also named trimethylglycine) is mainly eliminated by transmethylation to dimethylglycine [358]. This observation is consistent with the finding that pigs fed lupin protein display the highest serum betaine concentrations and the lowest serum homocysteine concentrations. We speculated that all crops rich in proteins contain considerable amounts of betaine.

The finding that the SAM:SAH ratio was not significantly different between the groups did not support the hypothesis that differences in the serum concentrations of metabolites from the one-carbon cycle may impact epigenetic regulation. Furthermore, no differences in the global or site-specific histone methylation or in the mRNA expression of the DNA methyltransferases or histone demethylase Kdm3a were detected. Thus, dietary proteins are unlikely to affect the disease risk by influencing epigenetic processes following a short-term intervention.

However, it has to be considered that the experiment lasted only 4 weeks and all diets contained methyl donors in amounts to meet the requirements of growing pigs. The effect of dietary protein sources on epigenetics might have been different in cases of long-term intervention studies or low basal levels of methyl donors.

Still, data from a recent meta-analysis of human studies revealed that even dietary interventions lasting only 5 days can impact DNA methylation [358].

Quantification of classical risk factors such as glucose, lipids and CRP shows that the intake of proteins from animal or plant origin does not exert substantial health effects within 4 weeks of treatment. Here, the most important findings were that lupin protein intake, in contrast to beef and casein intake, was capable of lowering homocysteine concentrations. High serum homocysteine concentrations are associated with an increased CVD risk. Here, no significant differences in serum cholesterol levels were observed. This finding is consistent with animal and human intervention studies that only observed beneficial effects on serum total cholesterol or HDL cholesterol under hypercholesterolemic conditions [324, 359]. Furthermore, recent clinical trials questioned the association between an increase in HDL cholesterol levels and the improvement of cardiovascular outcomes [360].

Surprisingly, serum triglycerides were even higher in pigs fed lupin protein than in those fed beef or casein. No further significant differences were observed in the serum concentrations of TMAO, CRP or IGF-1; in parameters of glucose metabolism; or in the serum mineral concentrations. Interestingly, recent cohort and prospective studies reported positive correlations between the concentrations of trimethylamine-containing dietary nutrients, including choline, betaine, carnitine and trimethyllysine, and CVD risks [361, 362]. According to these studies, these metabolites are converted to the TMAO precursor trimethylamine by gut microorganisms. The causality of this association and the mechanisms responsible for this association are not clear [361]. In our study, the diets failed to affect serum TMAO levels.

Next, serum concentrations of asymmetric and symmetric dimethylarginine, which are independent risk factors for all-cause mortality and CVD [363], were comparable between the three groups of pigs. Furthermore, no differences were detected in the serum mineral concentrations, although plant proteins, such as lupin proteins, are characterized by higher concentrations of the mineral binding phytic acid. Meat intake has repeatedly been shown to exert adverse effects on colon health by increasing the risk of colon cancer [364, 365]. Factors associated with cellular stress and proliferation control were suggested to be responsible for the adverse effects of meat on the gut [366, 367].

In the current study, we did not detect any difference in the relative mRNA expression of genes associated with cellular stress and proliferation between the three groups of pigs. Based on our findings, we did not find adverse effects of beef intake on classical cardiovascular risk factors.

10.5 Conclusions and Outlook

Although combinations of two metabolites each enable 100% discrimination between the intake of lupin protein, beef or casein in our study, further controlled human intervention studies are necessary to determine whether our biomarkers are able to categorize individuals according to their regular protein intake, regardless of other components of the diet.

In addition, studies comparing the metabolic profile after the consumption of different plant protein sources are needed, as we do not know whether our potential biomarkers are suitable to identify lupin intake or if they reflect plant protein intake in general. As beef contained higher amounts of lipids than lupin protein or casein, we cannot exclude that some differences observed between the three groups were caused by the lipids associated with the beef intake. However, we propose that the impact of fat derived from the dietary protein source on the parameters measured was small because 90% of the dietary fat came from the basal dietary compounds.

Despite its limitations, this study had several strengths. The data were obtained from a randomized, highly controlled feeding study. The study was conducted with an animal model that shares great similarities with humans regarding the eating behavior, metabolism, and the anatomy and physiology. The administration of isolated proteins provided a good indication of the metabolic changes in response to a certain protein source.

In conclusion, dietary proteins induce distinct metabolic fingerprints in serum, which served as biomarkers for the type of dietary protein consumed. Based on the obtained data, the dietary proteins differed in their impact on serum homocysteine and homoarginine concentrations, which are risk factors of cardiovascular diseases. As the lupin protein group was characterized by lower levels of these risk factors, we postulate that lupin protein may positively impact cardiovascular risk compared to beef or casein. The data also did not indicate any epigenetic effect induced by the dietary protein source.

From a bioinformatics perspective, the analysis of serum concentrations based on LDA to detect amino acid biomarkers had its limitations due to the small sample size of 15 data points for each of the three dietary protein sources. In order to avoid overparameterization, resp. overfitting, we could not increase the parameter space. Hence, we were not able to model dependencies between the amino acid concentrations. If we had sufficient data, we could model dependencies by performing a quadratic discriminant analysis (QDA) which might lead to an increased classification rate or different sets of biomarkers.

11

Conclusions and Outlook

During my PhD studies, it was my goal to work on a broad spectrum of bioinformatic applications, and I am grateful that I had the opportunity to work with my colleagues on various topics of biology in various branches of the life sciences.

In this thesis, I attempted to show this diversity. In chapters 2 - 6, we studied the transcriptomic hourglass pattern in evolutionary developmental biology based on phylotranscriptomic analyses. In chapter 7, we entered the field of transcriptomics and learned about the transcriptome dynamics during grafting in developmental biology. In chapter 8, we developed a workflow for the annotation of protein-coding, long non-coding, and circular transcripts of flowering plants based on RNA-Seq data, covering the fields of transcriptomics and genomics.

To broaden the spectrum further, in chapters 9 and 10, we analyzed metabolomics data. In chapter 9, we entered the field of metabolomics and specifically the field of lipidomics by comparing the lipid composition of the human skin related to age-and disease-induced changes. This topic is directly related to applications in the fields of medicine and pharmacy. In chapter 10, we analyzed metabolic and transcript data from the metabolite serum of pigs in the field of agricultural and nutritional sciences.

The common objective of all of these applications was to uncover patterns hidden in the given data, such as the transcriptomic hourglass pattern, the gene expression patterns during grafting, differences of coding and non-coding transcripts, differences in lipid compositions, or differences in metabolite concentrations. To detect and to quantify these patterns, we combined concepts and methods from machine learning and statistics with subject-specific biological knowledge.

In chapter 2, we introduced the developmental hourglass pattern during embryogenesis in animals, which was also supported on the transcriptomic level for *Da. rerio* and *Drosophila* [9, 16]. Despite the absence of morphological evidence in plant species, my colleagues and I could discover a transcriptional hourglass pattern during embryogenesis for *A. thaliana.* Based on the previously published transcriptome age index (TAI) [9], which captures deep and long-term evolutionary changes, we developed the transcriptome diversity index (TDI), detecting rather short-term evolutionary changes. Demonstrating the existence of a transcriptomic hourglass pattern in animals and plants that probably evolved independently in both kingdoms by convergent evolution might suggest a necessary mechanism which allows a living organism to progress through embryogenesis coordinately.

In chapter 3, we continued investigating the transcriptomic hourglass by studying its functional relevance. As we had seen in chapter 2, the transcriptomic hourglass pattern could be regarded as evolutionary ancient due to its independent evolution in animals and plants. By calculating the TAI and TDI profiles for *Da. rerio*, *D. melanogaster*, and *A. thaliana* embryogenesis and by systematically evaluating the resulting patterns, we could quantify the statistical significance of transcriptomic hourglass patterns in all species with both measures. Especially the TDI profiles provided evidence for an actively maintained developmental transcriptomic hourglass pattern during embryogenesis which may imply functional relevance, suggesting that it might be possible to identify the molecular function of this pattern in the long term.

In chapter 4, we hypothesized that the transcriptomic hourglass pattern might be associated with other developmental transitions such as embryogenesis. In contrast to animals, the development of plants is not completed after embryogenesis, and organ formation occurs largely postembryonically. To test this hypothesis, we performed phylotranscriptomic analyses on postembryonic developmental transitions of *A. thaliana* such as germination and floral transition, and we detected in both transitions significant transcriptomic hourglass patterns. This implies that hourglass patterns are not restricted to embryogenesis, but that they may be present in several developmental processes.

In chapter 5, we tried to shed light on a possible functional explanation of the transcriptomic hourglass patterns presented in the previous chapters. We redefined the known phylotranscriptomic measures in a probabilistic manner, and we developed an entropic transcriptome age index. Applying the entropic TAI, we detected transcriptomic hourglass patterns with P values that were orders of magnitudes smaller than those of the transcriptomic hourglass patterns of the traditional TAI. This led us to the question if the transcriptomic hourglass patterns of the entropic TAI could possibly be the origins of the transcriptomic hourglass patterns of the traditional TAI.

In chapter 5, we tested the hypothesis that the entropic TAI might be the origin of the traditional TAI. We developed an approach to reproduce either the entropic TAI based on the traditional TAI or to reproduce the traditional TAI based on the entropic TAI. We found that the entropic TAI is capable of reproducing the traditional TAI patterns more accurately than the traditional TAI is capable of reproducing the entropic TAI patterns.

The phylotranscriptomic analyses of the chapters 2 - 6 were based on temporal gene expression data in combination with evolutionary information such as gene age. All of these studies had in common that the spatial resolution of gene expression was neglected. In contrast, our next goal was to analyze transcriptome dynamics of developmental processes at spatial and temporal resolution.

In chapter 7, we developed approaches to study transcriptomic dynamics in the developmental process of grafting. We analyzed the spatio-temporal gene expression above and below the graft junction. We found that different mechanisms are responsible for wound healing depending on the presence or absence of adjoining tissues. We found that an intertissue recognition mechanism is characterized by an asymmetric gene expression of sugar-associated genes and a symmetric gene expression of auxin-response genes above and below the graft junction.

Chapters 2 - 7 only considered the expression of protein-coding genes. Additionally, the investigation of expressed non-coding transcripts such as long non-coding RNAs, and their interaction with protein-coding genes could uncover an even more detailed view into developmental processes.

In chapter 8, we developed a workflow for annotating protein-coding splice variants, lncRNAs and circRNAs in seven flowering plants. From each plant, we sequenced eight organs and the mature pollen based on total RNA-Seq. We developed an annotation workflow to generate a comprehensive annotation for each plant species. We updated the current genomic annotations of the seven plants by thousands of novel protein-coding splice variants, lncRNAs, and circRNAs. The resulting annotations provided novel insights into the genomic structure of theses RNA species.

Understanding the complexity of the transcriptome in an organ-specific or a cell-specific manner on an evolutionary or developmental scale is the basis for understanding the complex biology of a living cell or a living organism. To approach the complex system of a cell, we need to investigate not only its expressed transcripts, but also compounds such as amino acids, proteins, carbohydrates, and lipids, which are analyzed in metabolomics.

In chapter 9, we entered lipidomics, a subdomain of metabolomics, by investigating age-related and diabetes-related changes in the free fatty acid composition of the Stratum corneum. Our straightforward statistical analysis uncovered a significant decrease of free fatty acid concentrations predominantly in young diabetic subjects compared to healthy subjects and in elderly subjects compared to young healthy subjects.

Interestingly, the risk of being affected by such chronic diseases like diabetes or cardiovascular diseases can depend on dietary factors such as different protein sources.

In chapter 10, we investigated in a second metabolomics study the contributions and effects of dietary proteins on the serum metabolite profile and thus their association to the risk of being affected by chronic diseases. Pigs, which serve as an animal model in the nutritional sciences, were fed with lupin protein, lean beef, and casein, and we analyzed their metabolite serums. We developed a linear discriminant analysis coupled with feature selection, and we uncovered combinations of metabolites from the serum that discriminate between the intake of the three different dietary protein sources.

At the end of our attempts to answer the scientific questions of chapters 2 - 10, several new questions have raised that open the way to future work. In the phylotranscriptomics studies of chapters 2 - 6, the development of novel approaches to study the origins of the transcriptomic hourglass patterns could help to extend our perspective to the developmental processes that seem to maintain the conserved transcriptomic patterns. It might be worthwhile to incorporate spatial data as we saw in the analysis of transcriptome dynamics during grafting or expression data from non-coding transcripts.

As we saw in the analyses of grafting in chapter 7, it might also be appropriate to consider a more detailed perspective into the process. Future work could focus on genes activated uniquely by grafting or genes involved in the recognition response to distinguish attached from separated plant tissues. These genes might help to identify the pathways required for grafting, wound healing, and vascular regeneration. Additionally, the inclusion of non-coding transcripts such as long non-coding RNAs or circular RNAs might help to identify and to extend pathways that are involved in grafting.

The annotation of novel protein-coding splice variants, long non-coding transcripts, and circular transcripts, as we saw in chapter 8, has the potential to open new ways of understanding the transcriptome. Based on comparative analyses, we might be capable of studying the conservation and potential functions of the different protein-coding and non-coding RNA species. Additionally, comparative transcriptome analyses of the different organs from the various flowering plants could help to deepen the understanding of the transcriptome evolution in an organ-specific manner.

The increased complexity in the transcriptome raised new questions for future work in transcriptomics and evolutionary biology. This rise of novel questions was also shown in the metabolomics studies. Based on the detected differences in free fatty acids in chapter 9, future work might be capable of finding evidence that the lipid pattern can be influenced, which might lead to the development of cosmetic and medical products to strengthen the skin barrier function of elderly patients or patients suffering from diabetes mellitus. In chapter 10, we developed a bioinformatics approach to obtained biomarkers to discriminate between the intake of different protein sources. Based on the findings, further controlled human intervention studies could determine whether the obtained biomarkers might categorize individuals according to their regular protein intake regardless of other diet components. Since the sample size in this study was small, it would be interesting, if advanced machine learning techniques would be more appropriate than the proposed linear discriminant analysis.

The presented thesis did not cover all bioinformatics studies my colleagues and I performed in the past years, but it is rather a subset of published and yet unpublished work. The following four transcriptomics and lipidomics studies were not presented in this thesis: In transcriptomics, we additionally worked on patterns of gene expression during *A. thaliana* flower development [4], analyzed the transcriptome of polyspermy-derived triparental plants to investigate the bypassing of the postzygotic polyploidization barrier, known as the triploid block [5], and supported RNA-Seq studies to analyze space-omics data [6]. In metabolomics, we worked on the targeted delivery of ceramide lipids into the Stratum corneum to support the repairing of the skin barrier [7].

My goal when composing and writing this thesis was to show that modern bioinformatics research is intertwined with various fields of the life sciences. The massive generation of huge biological data sets and the diverse topics reported in this thesis require efficient data processing coupled with subject-orientated analysis. In chapters 2 - 6, we generated large amino acid sequence databases for performing phylostratigraphy, and in chapters 3 – 8, we analyzed RNA-Seq data from various tissues and species. However, in all studies presented, we developed or applied computational and statistical approaches with the common goal of extracting hidden patterns from the data.

Analyzing the given data, we need to understand the biological subject to model the question mathematically and statistically. As we saw in all chapters, the goal of our colleagues was to extract and to interpret hidden patterns from their data that are of biological relevance and not observed by chance. To this end, robust methods are needed for quantifying these diverse observations and for stating their statistical significance, or establishing methods and workflows that enable other researchers to comprehend and to reproduce the published results.

The landscape of research in the natural sciences has evolved to an unprecedented level of diversity coupled with immense amounts of data from all sources. Data management, statistical analysis, and the development and application of machine learning techniques have become essential parts of modern data-driven research. With this thesis I wanted to demonstrate that these techniques coupled with biological knowledge are the core of bioinformatics research, enabling insightful collaborations with specialists of various natural sciences to answer fundamental questions and to expand our understanding of nature.

References

1. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
2. Craig Venter, J. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
3. Sander, K. In (eds Goodwin, C. *et al.*) 137–160 (Cambridge University Press, Cambridge, 1983).
4. Ryan, P. T. *et al.* Patterns of gene expression during Arabidopsis flower development from the time of initiation to maturation. *BMC Genomics* **16**, 488 (2015).
5. Mao, Y. *et al.* Selective egg cell polyspermy bypasses the triploid block. *eLife* **9** (2020).
6. Madrigal, P. *et al.* Revamping Space-omics in Europe. *Cell Systems,* 10–11 (2020).
7. Steinbach, S. C. *et al.* Retarder action of isosorbide in a microemulsion for a targeted delivery of ceramide NP into the stratum corneum. *Die Pharmazie* **72**, 440–446 (2017).
8. Domazet-Lošo, T. *et al.* A phylostratigraphic approach to uncover the genomic history of major adaptions in metazoan lineages. *Trends in Genetics* **23**, 533–539 (2007).
9. Domazet-Lošo, T. & Tautz, D. A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature* **468**, 815–818 (2010).
10. Quint, M. *et al.* A transcriptomic hourglass in plant embryogenesis. *Nature* **490**, 98–101 (2012).
11. Meyerowitz, E. M. Plants compared to animals: the broadest comparative study of development. *Science (New York, N.Y.)* **295**, 1482–1485 (2002).
12. Meckel, J. F. *Beyträge zur vergleichenden Anatomie* (Reclam, Leipzig, 1811).
13. Von Baer, K. *Über Entwicklungsgeschichte der Thiere : Beobachtungen und Reflexion* (Gebrüder Bornträger, Königsberg, 1828).
14. Duboule, D. Temporal colinearity and the phylotypic progression: a basis for the stability of a vertebrate Bauplan abd the evolution of morphologies through heterochrony. *Dev. Suppl.* 135–142 (1994).
15. Raff, R. A. *The Shape of Life: Genes, Development and Evolution of Animal Form* (University Chicago Press, 1996).
16. Kalinka, A. T. *et al.* Gene expression divergence recapitulates the developmental hourglass model. *Nature* **468**, 811–814 (2010).

17. Smet, I. D. *et al.* Embryogenesis – the humble beginnings of plant life. *The Plant Journal* **61**, 959–970 (2010).
18. Peris, C. I. L. *et al.* Green Beginnings — Pattern Formation in the Early Plant Embryo. *Current Topics in Developmental Biology* **91**, 1–27 (2010).
19. Letunic, I. & Bork, P. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Research,* 1–4 (2011).
20. Hedges, S. B. & Kumar, S. *The timetree of life* (OUP Oxford, 2009).
21. Kersting, A. R. *et al.* Dynamics and Adaptive Benefits of Protein Domain Emergence and Arrangements during Plant Genome Evolution. *Genome Biology and Evolution* **4**, 316–329 (2012).
22. Katoh, K. *et al.* MAFFT version 5: Improvement in accuracy of multiple sequence alignment. *Nucleic Acids Research* **33**, 511–518 (2005).
23. Suyama, M. *et al.* PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research* **34**, 609–612 (2006).
24. Thornton, K. libsequence: A C++ class library for evolutionary genetic analysis. *Bioinformatics* **19**, 2325–2327 (2003).
25. Zuber, H. *et al.* The Seed Composition of *Arabidopsis* Mutants for the Group 3 Sulfate Transporters Indicates a Role in Sulfate Translocation within Developing Seeds. *Plant Physiology* **154**, 913–926 (2010).
26. Le, B. H. *et al.* Global analysis of gene activity during *Arabidopsis* seed development and identification of seed-specific transcription factors. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 8063–8070 (2010).
27. Wu, Z. *et al.* A model-based background adjustment for oligonucleotide expression arrays. *Journal of the American Statistical Association* **99**, 909–917 (2004).
28. Xiang, D. *et al.* Genome-wide analysis reveals gene expression and metabolic network dynamics during embryo development in Arabidopsis. *Plant Physiology* **156**, 346–356 (2011).
29. Müller, W. A. & Hassel, M. *Entwicklungsbiologie und Reproduktionsbiologie von Mensch und Tieren* (Springer-Verlag, 2006).
30. Levin, M. *et al.* Developmental Milestones Punctuate Gene Expression in the *Caenorhabditis* Embryo. *Developmental Cell* **22**, 1101–1108 (2012).
31. Koch, M. A. *et al.* Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in *Arabidopsis*, *Arabis*, and related genera (Brassicaceae). *Molecular Biology and Evolution* **17**, 1483–1498 (2000).

32. Arakaki, M. *et al.* Contemporaneous and recent radiations of the world's major succulent plant lineages. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 8379–8384 (2011).
33. Koch, M. A. & Kiefer, M. Genome evolution among cruciferous plants: A lecture from the comparison of the genetic maps of three diploid species - *Capsella rubella*, *Arabidopsis lyrata subsp. petraea*, and *A. thaliana*. *American Journal of Botany* **92**, 761–767 (2005).
34. Oh, D. H. *et al.* Genome Structures and Halophyte-Specific Gene Expression of the Extremophile *Thellungiella parvula* in Comparison with *Thellungiella salsuginea* (*Thellungiella halophila*) and *Arabidopsis*. *Plant Physiology* **154**, 1040–1052 (2010).
35. Nodine, M. D. & Bartel, D. P. Maternal and paternal genomes contribute equally to the transcriptome of early plant embryos. *Nature* **482**, 94–97 (2012).
36. Slack, J. M. *et al.* The zootype and the phylotypic stage. *Nature* **361**, 490–492 (1993).
37. Lau, S. *et al.* Early Embryogenesis in Flowering Plants: Setting Up the Basic Body Pattern. *Annual Review of Plant Biology* **63**, 483–506 (2012).
38. Park, S. & Harada, J. J. *Arabidopsis* Embryogenesis. *Methods in Molecular Biology* **427**, 3–16 (2008).
39. Irie, N. & Sehara-Fujisawa, A. The vertebrate phylotypic stage and an early bilaterian-related stage in mouse embryogenesis defined by genomic information. *BMC Biology* **5** (2007).
40. Irie, N. & Kuratani, S. Comparative transcriptome analysis reveals vertebrate phylotypic period during organogenesis. *Nature communications* **2** (2011).
41. Drost, H.-G. *et al.* Evidence for Active Maintenance of Phylotranscriptomic Hourglass Patterns in Animal and Plant Embryogenesis. *Molecular Biology and Evolution* **32**, 1221–1231 (2015).
42. Kalinka, A. T. & Tomancak, P. The evolution of early animal embryos: Conservation or divergence? *Trends in Ecology and Evolution* **27**, 385–393 (2012).
43. Flicek, P. *et al.* Ensembl 2014. *Nucleic Acids Research* **42**, 749–755 (2014).
44. St. Pierre, S. E. *et al.* FlyBase 102 - Advanced approaches to interrogating FlyBase. *Nucleic Acids Research* **42**, 780–788 (2014).
45. Goodstein, D. M. *et al.* Phytozome: A comparative platform for green plant genomics. *Nucleic Acids Research* **40**, 1178–1186 (2012).
46. Gabel, A. *AlexGa/Phylostratigraphy: createPSmap* version v0.0.4. 2019.

47. Comeron, J. M. A method for estimating the numbers of synonymous and nonsynonymous substitutions per site. *Journal of Molecular Evolution* **41**, 1152–1159 (1995).
48. Drost, H. G. *et al.* MyTAI: Evolutionary transcriptomics with R. *Bioinformatics* **34**, 1589–1590 (2018).
49. Piasecka, B. *et al.* The Hourglass and the Early Conservation Models-Co-Existing Patterns of Developmental Constraints in Vertebrates. *PLoS Genetics* **9** (2013).
50. Graveley, B. R. *et al.* The developmental transcriptome of *Drosophila melanogaster. Nature* **471**, 473–9 (2011).
51. Gerstein, M. B. *et al.* Comparative analysis of the transcriptome across distant species. *Nature* **512**, 445–448 (2014).
52. Galis, F. & Metz, J. A. Testing the vulnerability of the phylotypic stage: On modularity and evolutionary conservation. *Journal of Experimental Zoology* **291**, 195–204 (2001).
53. Hazkani-Covo, E. *et al.* In Search of the Vertebrate Phylotypic Stage : A Molecular Examination of the Developmental Hourglass Model and von Baer ' s Third Law. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution* **304B**, 150–158 (2005).
54. Davidson, E. H. & Erwin, D. H. An integrated view of precambrian eumetazoan evolution. *Cold Spring Harbor Symposia on Quantitative Biology* **LXXIV** (2009).
55. He, J. & Deem, M. W. Hierarchical evolution of animal body plans. *Developmental Biology* **337**, 157–161 (2010).
56. Peter, I. S. & Davidson, E. H. Evolution of gene regulatory networks controlling body plan development. *Cell* **144**, 970–985 (2011).
57. De Mendoza, A. *et al.* Transcription factor evolution in eukaryotes and the assembly of the regulatory toolkit in multicellular lineages. *Proceedings of the National Academy of Sciences of the United States of America* **110**, E4858–E4866 (2013).
58. Schep, A. N. & Adryan, B. A Comparative Analysis of Transcription Factor Expression during Metazoan Embryonic Development. *PLoS ONE* **8** (2013).
59. Wang, Z. *et al.* The draft genomes of soft-shell turtle and green sea turtle yield insights into the development and evolution of the turtle-specific body plan. - Supplement. *Nature genetics* **45**, 701–706 (2013).
60. Hedges, S. B. *et al.* TimeTree: A public knowledge-base of divergence times among organisms. *Bioinformatics* **22**, 2971–2972 (2006).
61. Hu, T. T. *et al.* The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nature genetics* **43**, 476–81 (2011).

62. Hall, B. K. Phylotypic stage or phantom: is there a highly conserved embryonic stage in vertebrates? *Trends in Ecology and Evolution* **12**, 461–463 (1997).

63. Richardson, M. K. *et al.* There is no highly conserved embryonic stage in the vertebrates: Implications for current theories of evolution and development. *Anatomy and Embryology* **196**, 91–106 (1997).

64. Richardson, M. K. Vertebrate evolution: The developmental origins of adult variation. *BioEssays* **21**, 604–613 (1999).

65. Bininda-Emonds, O. R. *et al.* Inverting the hourglass: Quantitative evidence against the phylotypic stage in vertebrate development. *Proceedings of the Royal Society B: Biological Sciences* **270**, 341–346 (2003).

66. Roux, J. & Robinson-Rechavi, M. Developmental constraints on vertebrate genome evolution. *PLoS Genetics* **4** (2008).

67. Comte, A. *et al.* Molecular signaling in zebrafish development and the vertebrate phylotypic period. *Evolution and Development* **12**, 144–156 (2010).

68. Davis, J. C. *et al.* Protein evolution in the context of *Drosophila* development. *Journal of Molecular Evolution* **60**, 774–785 (2005).

69. Demuth, J. P. *et al.* The evolution of mammalian gene families. *PLoS ONE* **1** (2006).

70. Cruickshank, T. & Wade, M. J. Microevolutionary support for a developmental hourglass: Gene expression patterns shape sequence variation and divergence in *Drosophila. Evolution and Development* **10**, 583–590 (2008).

71. Ninova, M. *et al.* Conserved temporal patterns of microRNA expression in *Drosophila* support a developmental hourglass model. *Genome Biology and Evolution* **6**, 2459–2467 (2014).

72. Kaplan, D. R. & Cooke, T. J. Fundamental concepts in the embryogenesis of dicotyledons: A morphological interpretation of embryo mutants. *Plant Cell* **9**, 1903–1919 (1997).

73. Darwin, C. *On the origin of species* (London: Murray, 1859).

74. Drost, H.-G. *et al.* Post-embryonic Hourglass Patterns Mark Ontogenetic Transitions in Plant Development. *Molecular Biology and Evolution* **33**, 1158–1163 (2016).

75. Richardson, M. K. Heterochrony and the phylotypic period. *Dev. Biol.* **172**, 412–421 (1995).

76. Artieri, C. G. *et al.* Ontogeny and phylogeny: Molecular signatures of selection, constraint, and temporal pleiotropy in the development of *Drosophila. BMC Biology* **7**, 1–14 (2009).

77. Yanai, I. *et al.* Mapping Gene Expression in Two Xenopus Species: Evolutionary Constraints and Developmental Flexibility. *Developmental Cell* **20**, 483–496 (2011).
78. Levin, M. *et al.* The mid-developmental transition and the evolution of animal body plans. *Nature* **531**, 637–641 (2016).
79. Cheng, X. *et al.* A "developmental hourglass" in fungi. *Molecular Biology and Evolution* **32**, 1556–1566 (2015).
80. Wan, C. & Wilkins, T. A Modified Hot Borate Method Significantly Enhances the Yield of High-Quality RNA from Cotton (*Gossypium hirsutum L.*) *Analytical Biochemistry* **223**, 7–12 (1994).
81. Maia, J. *et al.* The re-establishment of desiccation tolerance in germinated *Arabidopsis thaliana* seeds and its associated transcriptome. *PLoS ONE* **6** (2011).
82. Schmid, M. *et al.* Dissection of floral induction pathways using global expression analysis. *Development* **130**, 6001–6012 (2003).
83. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology* **14**, R36 (2013).
84. Liao, Y. *et al.* FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
85. Zhou, X. *et al.* Robustly detecting differential expression in RNA sequencing data using observation weights. *Nucleic Acids Research* **42** (2014).
86. Silva, A. T. *et al.* A predictive coexpression network identifies novel genes controlling the seed-to-seedling phase transition in *Arabidopsis thaliana. Plant Physiology* **170**, 2218–2231 (2016).
87. Huijser, P. & Schmid, M. The control of developmental phase transitions in plants. *Development* **138**, 4117–4129 (2011).
88. Irie, N. & Kuratani, S. The developmental hourglass model: a predictor of the basic body plan? *Development (Cambridge, England)* **141**, 4649–55 (2014).
89. Shannon, C. E. A Mathematical Theory of Communication. *Bell System Technical Journal* **27**, 379–423 (1948).
90. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* **44**, D7–D19 (2016).
91. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**, 3389–3402 (1997).
92. Arbeitman, M. N. *et al.* Gene expression during the life cycle of *Drosophila melanogaster. Science* **297**, 2270–2275 (2002).

93. Drost, H. G. *et al.* Cross-kingdom comparison of the developmental hourglass. *Current Opinion in Genetics and Development* **45**, 69–75 (2017).
94. Xu, F. *et al.* High expression of new genes in trochophore enlightening the ontogeny and evolution of trochozoans. *Scientific Reports* **6**, 1–10 (2016).
95. Wu, L. *et al.* Gene Expression Does Not Support the Developmental Hourglass Model in Three Animals with Spiralian Development. *Molecular biology and evolution* **36**, 1373–1383 (2019).
96. Melnyk Charles W and**Gabel**, A. *et al.* Transcriptome dynamics at Arabidopsis graft junctions reveal an intertissue recognition mechanism that activates vascular regeneration. *Proceedings of the National Academy of Sciences of the United States of America* **115**, E2447–E2456 (2018).
97. Goldschmidt, E. E. Plant grafting: New mechanisms, evolutionary implications. *Frontiers in Plant Science* **5**, 1–9 (2014).
98. Lee, J. M. *et al.* Current status of vegetable grafting: Diffusion, grafting techniques, automation. *Scientia Horticulturae* **127**, 93–105 (2010).
99. Melnyk, C. W. Plant grafting: insights into tissue regeneration. *Regeneration* **4**, 3–14 (2017).
100. Melnyk, C. W. *et al.* A developmental framework for graft formation and vascular reconnection in arabidopsis thaliana. *Current Biology* **25**, 1306–1318 (2015).
101. Yin, H. *et al.* Graft-union development: a delicate process that involves cell–cell communication between scion and stock for local auxin accumulation. *Journal of Experimental Botany* **63**, 4219–4232 (2012).
102. Matsuoka, K. *et al.* Differential cellular control by cotyledon-derived phytohormones involved in graft reunion of arabidopsis hypocotyls. *Plant and Cell Physiology* **57**, 2620–2631 (2016).
103. Matsumoto-Kitano, M. *et al.* Cytokinins are central regulators of cambial activity. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 20027–20031 (2008).
104. Leyser, O. Auxin, self-organisation, and the colonial nature of plants. *Current Biology* **21**, R331–R337 (2011).
105. Asahina, M. *et al.* Spatially selective hormonal control of RAP2.6L and ANAC071 transcription factors involved in tissue reunion in Arabidopsis. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 16128–16132 (2011).
106. Wetmore, R. H. & Rier, J. P. Experimental Induction of Vascular Tissues in Callus of Angiosperms. *American Journal of Botany* **50**, 418–430 (1963).
107. Aloni, R. Role of auxin and sucrose in the differentiation of sieve and tracheary elements in plant tissue cultures. *Planta* **150**, 255–263 (1980).

108. Lough, T. J. & Lucas, W. J. INTEGRATIVE PLANT BIOLOGY: Role of Phloem Long-Distance Macromolecular Trafficking. *Annual Review of Plant Biology* **57**, 203–232 (2006).
109. Wang, L. & Ruan, Y. L. Regulation of cell division and expansion by sugar and auxin signaling. *Frontiers in Plant Science* **4**, 1–9 (2013).
110. Kuhlemeier, C. & Timmermans, M. C. The Sussex signal: Insights into leaf dorsiventrality. *Development (Cambridge)* **143**, 3230–3237 (2016).
111. Qi, J. *et al.* Auxin depletion from leaf primordia contributes to organ patterning. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 18769–18774 (2014).
112. McConnell, J. R. *et al.* Role of PHABULOSA and PHAVOLUTA in determining radial patterning in shoots. *Nature* **411**, 709–713 (2001).
113. Cheong, Y. H. *et al.* Transcriptional profiling reveals novel interactions between wounding, pathogen, abiotic stress, and hormonal responses in Arabidopsis. *Plant Physiology* **129**, 661–677 (2002).
114. Iwase, A. *et al.* The AP2/ERF transcription factor WIND1 controls cell dedifferentiation in arabidopsis. *Current Biology* **21**, 508–514 (2011).
115. Ikeuchi, M. *et al.* Wounding triggers callus formation via dynamic hormonal and transcriptional changes. *Plant Physiology* **175**, 1158–1174 (2017).
116. Zheng, B. S. *et al.* CDNA-AFLP analysis of gene expression in hickory (Carya cathayensis) during graft process. *Tree Physiology* **30**, 297–303 (2009).
117. Cookson, S. J. *et al.* Graft union formation in grapevine induces transcriptional changes related to cell wall modification, wounding, hormone signalling, and secondary metabolism. *Journal of Experimental Botany* **64**, 2997–3008 (2013).
118. Cookson, S. J. *et al.* Heterografting with nonself rootstocks induces genes involved in stress responses at the graft interface when compared with autografted controls. *Journal of Experimental Botany* **65**, 2473–2481 (2014).
119. Chen, Z. *et al.* Transcriptome changes between compatible and incompatible graft combination of Litchi chinensis by digital gene expression profile. *Scientific Reports* **7**, 1–12 (2017).
120. Joshi, N. & Fass, J. *Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software]* 2011.
121. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357–359 (2012).
122. Roberts, A. & Pachter, L. Streaming fragment assignment for real-time analysis of sequencing experiments. *Nature methods* **10**, 71–3 (2013).
123. Hardcastle, T. J. *et al.* Identifying small interfering RNA loci from high-throughput sequencing data. *Bioinformatics* **28**, 457–463 (2012).

124. Hardcastle, T. J. & Kelly, K. A. BaySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics* **11** (2010).

125. Carlson, M. *org.At.tair.db: Genome wide annotation for Arabidopsis* R package version 3.4.1 (2017).

126. Falcon, S. & Gentleman, R. Using GOstats to test gene lists for GO term association. *Bioinformatics* **23**, 257–258 (2007).

127. Lindsay, D. W. *et al.* An Analysis of the Development of the Graft Union in Lycopersicon esculentum. *Annals of Botany* **38**, 639–646 (1974).

128. Moore, R. Graft formation in Solanum pennellii (Solanaceae). *Plant Cell Reports* **3**, 172–175 (1984).

129. Furuta, K. M. *et al.* Plant development. Arabidopsis NAC45/86 direct sieve element morphogenesis culminating in enucleation. *Science (New York, N.Y.)* **345**, 933–937 (2014).

130. Kondo, Y. *et al.* Vascular cell induction culture system using arabidopsis leaves (VISUAL) reveals the sequential differentiation of sieve element-like cells. *Plant Cell* **28**, 1250–1262 (2016).

131. Czechowski, T. *et al.* Genome-Wide Identification and Testing of Superior Reference Genes for Transcript Normalization in Arabidopsis. *Plant Physiology* **139**, 5–17 (2005).

132. Brady, S. M. *et al.* A high-resolution root spatiotemporal map reveals dominant expression patterns. *Science (New York, N.Y.)* **318**, 801–6 (2007).

133. Marsch-Martínez, N. *et al.* An efficient flat-surface collar-free grafting method for Arabidopsis thaliana seedlings. *Plant Methods* **9**, 1–9 (2013).

134. Villadsen, D. & Smith, S. M. Identification of more than 200 glucose-responsive Arabidopsis genes none of which responds to 3 -O-methylglucose or 6-deoxyglucose. *Plant Molecular Biology* **55**, 467–477 (2004).

135. Thum, K. E. *et al.* Genome-wide investigation of light and carbon signaling interactions in Arabidopsis. *Genome biology* **5**, 1–20 (2004).

136. Cordoba, E. *et al.* Sugar regulation of SUGAR TRANSPORTER PROTEIN 1 (STP1) expression in Arabidopsis thaliana. *Journal of Experimental Botany* **66**, 147–159 (2015).

137. Nemhauser, J. L. *et al.* Different Plant Hormones Regulate Similar Processes through Largely Nonoverlapping Transcriptional Responses. *Cell* **126**, 467–475 (2006).

138. Abel, S. *et al.* ThePS-IAA4/5-like Family of Early Auxin-inducible mRNAs inArabidopsis thaliana. *Journal of Molecular Biology* **251**, 533–549 (1995).

139. Brunoud, G. *et al.* A novel sensor to map auxin response and distribution at high spatio-temporal resolution. *Nature* **482**, 103–106 (2012).

140. Guo, Y. *et al.* Dof5.6/HCA2, a dof transcription factor gene, regulates interfascicular cambium formation and vascular tissue development in Arabidopsis. *Plant Cell* **21**, 3518–3534 (2009).

141. De Rybel, B. *et al.* Plant vascular development: From early specification to differentiation. *Nature Reviews Molecular Cell Biology* **17**, 30–40 (2016).

142. Bonke, M. *et al.* APL regulates vascular tissue identity in Arabidopsis. *Nature* **426**, 181–186 (2003).

143. Ito, Y. & Fukuda, H. Dodeca-CLE Peptides as Suppressors. *Science,* 842–845 (2006).

144. Sugimoto, K. *et al.* Arabidopsis Regeneration from Multiple Tissues Occurs via a Root Development Pathway. *Developmental Cell* **18**, 463–471 (2010).

145. Gardiner, J. *et al.* Expression of DOF genes identifies early stages of vascular development in Arabidopsis leaves. *International Journal of Developmental Biology* **54**, 1389–1396 (2010).

146. Pitaksaringkarn, W. *et al.* ARF6 and ARF8 contribute to tissue reunion in incised Arabidopsis inflorescence stems. *Plant Biotechnology* **31**, 49–53 (2014).

147. Sauer, M. *et al.* Canalization of auxin flow by Aux/IAA-ARF-dependent feedback regulation of PIN polarity. *Genes and Development* **20**, 2902–2911 (2006).

148. Skylar, A. *et al.* Metabolic sugar signal promotes Arabidopsis meristematic proliferation via G2. *Developmental Biology* **351**, 82–89 (2011).

149. Souza, C. d. A. *et al.* Cellulose-Derived Oligomers Act as Damage-Associated Molecular Patterns and Trigger Defense-Like Responses. *Plant Physiology* **173**, 2383–2398 (2017).

150. Musselmann, L. J. The biology of Striga, Orobanche, and other root-parasitic weeds. *Annual Review of Phytopathology* **18**, 463–489 (1980).

151. Becraft, P. W. *et al.* Crinkly4: A TNFR-like receptor kinase involved in maize epidermal differentiation. *Science* **273**, 1406–1409 (1996).

152. Lolle, S. J. *et al.* Genetic analysis of organ fusion in Arabidopsis thaliana. *Genetics* **149**, 607–619 (1998).

153. Zhong, J. & Preston, J. C. Bridging the gaps: Evolution and development of perianth fusion. *New Phytologist* **208**, 330–335 (2015).

154. Derrien, T. *et al.* The GENCODE v7 catalog of human long noncoding RNAs. *Genome Research* **22**, 1775–1789 (2012).

155. Ruiz-Orera, J. *et al.* Long non-coding RNAs as a source of new peptides. *eLife* **3**, 1–24 (2014).

156. Brannan, C. I. *et al.* The product of the H19 gene may function as an RNA. *Molecular and Cellular Biology* **10**, 28–36 (1990).

157. Yang, W.-C. *et al.* Characterization of GmENOD40, a gene showing novel patterns of cell-specific expression during soybean nodule development. *The Plant Journal* **3**, 573–585 (1993).

158. Romero-Barrios, N. *et al.* Splicing regulation by long noncoding RNAs. *Nucleic Acids Research* **46**, 2169–2184 (2018).

159. Kung, J. T. Y. *et al.* Long noncoding RNAs: Past, present, and future. *Genetics* **193**, 651–669 (2013).

160. Necsulea, A. *et al.* The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* **505**, 635–640 (2014).

161. Sarropoulos, I. *et al.* Developmental dynamics of lncRNAs across mammalian organs and species. *Nature* **571**, 510–514 (2019).

162. Salzman, J. *et al.* Circular RNAs are the predominant transcript isoform from hundreds of human genes in diverse cell types. *PLoS ONE* **7** (2012).

163. Rybak-Wolf, A. *et al.* Circular RNAs in the Mammalian Brain Are Highly Abundant, Conserved, and Dynamically Expressed. *Molecular Cell* **58**, 870–885 (2015).

164. Hansen, T. B. *et al.* Natural RNA circles function as efficient microRNA sponges. *Nature* **495**, 384–388 (2013).

165. Memczak, S. *et al.* Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* **495**, 333–338 (2013).

166. Berardini, T. Z. *et al.* The *Arabidopsis* information resource: Making and mining the "gold standard" annotated reference plant genome. *Genesis* **53**, 474–485 (2015).

167. Cheng, C.-Y. *et al.* Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome. *The Plant Journal* **89**, 789–804 (2017).

168. Wierzbicki, A. T. *et al.* Noncoding Transcription by RNA Polymerase Pol IVb/Pol V Mediates Transcriptional Silencing of Overlapping and Adjacent Genes. *Cell* **135**, 635–648 (2008).

169. Wierzbicki, A. T. *et al.* RNA polymerase v transcription guides ARGONAUTE4 to chromatin. *Nature Genetics* **41**, 630–634 (2009).

170. Liu, T. T. *et al.* A global identification and analysis of small nucleolar RNAs and possible intermediate-sized non-coding RNAs in *Oryza sativa*. *Molecular Plant* **6**, 830–846 (2013).

171. Wang, Y. *et al.* Genomic Features and Regulatory Roles of Intermediate-Sized Non-Coding RNAs in *Arabidopsis*. *Molecular Plant* **7**, 514–527 (2014).

172. Burd, C. E. *et al.* Expression of linear and novel circular forms of an INK4/ARF-associated non-coding RNA correlates with atherosclerosis risk. *PLoS Genetics* **6**, 1–15 (2010).

173. Corporation, L. T. *ERCC RNA Spike-In Control Mixes User Guide* 4456740 (2012), 1–28.

174. Andrews, S. *FastQC: a quality control tool for high throughput sequence data* 2010.

175. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10 (2011).

176. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

177. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology* **28**, 511–515 (2010).

178. Patro, R. *et al.* Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods* **14**, 417–419 (2017).

179. Pertea, G. & Pertea, M. GFF Utilities: GffRead and GffCompare. *F1000 Research* **9**, 1–20 (2020).

180. Keilwagen, J. *et al.* Using intron position conservation for homology-based gene prediction. *Nucleic Acids Research* **44** (2016).

181. Li, W. *et al.* Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics* **17**, 282–283 (2001).

182. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Research* **41**, 590–596 (2013).

183. Haas, B. J. *et al.* De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols* **8**, 1494–1512 (2013).

184. The UniProt Consortium. UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Research* **47**, D506–D515 (2018).

185. Johnson, L. S. *et al.* Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics* **11**, 431 (2010).

186. Finn, R. D. *et al.* The Pfam protein families database: Towards a more sustainable future. *Nucleic Acids Research* **44**, D279–D285 (2016).

187. Jones, P. *et al.* InterProScan 5: Genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).

188. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology* **33**, 290–295 (2015).

189. Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. eng. *Journal of molecular biology* **48**, 443–453 (1970).

190. Ameur, A. *et al.* Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. *Nature Structural and Molecular Biology* **18**, 1435–1440 (2011).

191. Zhao, S. *et al.* Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: PolyA+ selection versus rRNA depletion. *Scientific Reports* **8**, 1–12 (2018).

192. Alamancos, G. P. *et al.* Leveraging transcript quantification for fast computation of alternative splicing profiles. *RNA (New York, N.Y.)* **21**, 1521–1531 (2015).

193. Cabili, M. *et al.* Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes and Development* **25**, 1915–1927 (2011).

194. Belgard, T. G. *et al.* A transcriptomic atlas of mouse neocortical layers. *Neuron* **71**, 605–616 (2011).

195. Pauli, A. *et al.* Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. *Genome Research* **22**, 577–591 (2012).

196. Young, R. S. *et al.* Identification and properties of 1,119 candidate LincRNA loci in the *Drosophila melanogaster* genome. *Genome Biology and Evolution* **4**, 427–442 (2012).

197. Li, T. *et al.* Identification of long non-protein coding RNAs in chicken skeletal muscle using next generation sequencing. *Genomics* **99**, 292–298 (2012).

198. Liu, J. *et al.* Genome-wide analysis uncovers regulation of long intergenic noncoding RNAs in *Arabidopsis. The Plant cell* **24**, 4333–45 (2012).

199. Li, S. *et al.* High resolution expression map of the *Arabidopsis* root reveals alternative splicing and lincRNA regulation. *Developmental Cell* **39**, 508–522 (2016).

200. Nelson, A. D. *et al.* Evolinc: A tool for the identification and evolutionary comparison of long intergenic non-coding RNAs. *Frontiers in Genetics* **8**, 1–12 (2017).

201. Darbellay, F. & Necsulea, A. Comparative Transcriptomics Analyses across Species, Organs, and Developmental Stages Reveal Functionally Constrained lncRNAs. *Molecular Biology and Evolution* **37**, 240–259 (2020).

202. Lin, M. F. *et al.* PhyloCSF: A comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27**, 275–282 (2011).

203. Wang, L. *et al.* CPAT: Coding-potential assessment tool using an alignment-free logistic regression model. *Nucleic Acids Research* **41**, 1–7 (2013).

204. Sun, L. *et al.* Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. *Nucleic Acids Research* **41** (2013).

205. Kang, Y. J. *et al.* CPC2: A fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Research* **45**, W12–W16 (2017).

206. Wucher, V. *et al.* FEELnc: A tool for Long non-coding RNAs annotation and its application to the dog transcriptome. *Nucleic Acids Research,* 1–12 (2016).

207. Han, S. *et al.* Long Noncoding RNA Identification: Comparing Machine Learning Based Tools for Long Noncoding Transcripts Discrimination. *BioMed Research International* **2016**, 1–14 (2016).

208. Negri, T. D. C. *et al.* Pattern recognition analysis on long noncoding RNAs: A tool for prediction in plants. *Briefings in Bioinformatics* **20**, 682–689 (2019).

209. Ziegler, C. & Kretz, M. The More the Merrier — Complexity in Long non-Coding RnA Loci. *Frontiers in Endocrinology* **8**, 1–6 (2017).

210. Jin, J. *et al.* PLncDB: Plant long non-coding RNA database. *Bioinformatics* **29**, 1068–1071 (2013).

211. Hetzel, J. *et al.* Nascent RNA sequencing reveals distinct features in plant transcription. *Proceedings of the National Academy of Sciences of the United States of America* **113**, 12316–12321 (2016).

212. Szabo, L. & Salzman, J. Detecting circular RNAs: bioinformatic and experimental challenges. *Nature reviews. Genetics* **17**, 679–692 (2016).

213. Ye, C. Y. *et al.* Widespread noncoding circular RNAs in plants. *New Phytologist* **208**, 88–95 (2015).

214. Quinn, J. J. & Chang, H. Y. Unique features of long non-coding RNA biogenesis and function. *Nature Publishing Group* **17** (2016).

215. Gao, Y. & Zhao, F. Computational Strategies for Exploring Circular RNAs. *Trends in Genetics* **34**, 389–400 (2018).

216. Cheng, J. *et al.* Specific identification and quantification of circular RNAs from sequencing data. *Bioinformatics* **32**, 1094–1096 (2016).

217. Gao, Y. *et al.* Circular RNA identification based on multiple seed matching. *Briefings in bioinformatics* **19**, 803–810 (2018).

218. Zhang, X. O. *et al.* Diverse alternative back-splicing and alternative splicing landscape of circular RNAs. *Genome Research* **26**, 1277–1287 (2016).

219. Zeng, X. *et al.* A comprehensive overview and evaluation of circular RNA detection tools. *PLoS Computational Biology* **13**, 1–21 (2017).

220. Chu, Q. *et al.* PlantcircBase: A Database for Plant Circular RNAs. *Molecular Plant* **10**, 1126–1128 (2017).

221. Köster, J. & Rahmann, S. Snakemake – a scalable bioinformatics workflow engine. *Bioinformatics* **28**, 2520–2522 (2012).

222. *Anaconda Software Distribution* version Vers. 2-2.4.0. 2020.

223. Martin, J. A. & Wang, Z. Next-generation transcriptome assembly. *Nature Reviews Genetics* **12**, 671–682 (2011).

224. Kozarewa, I. *et al.* Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nature Methods* **6**, 291–295 (2009).

225. Simão, F. A. *et al.* BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).

226. Hansen, K. D. *et al.* Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Research* **38**, 1–7 (2010).

227. Pesole, G. *et al.* Structural and functional features of eukaryotic mRNA untranslated regions. *Gene* **276**, 73–81 (2001).

228. Tilgner, H. *et al.* Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Research* **22**, 1616–1625 (2012).

229. Melé, M. *et al.* Chromatin environment, transcriptional regulation, and splicing distinguish lincRNAs and mRNAs. *Genome Research* **27**, 27–37 (2017).

230. Mukherjee, N. *et al.* Integrative classification of human coding and noncoding genes through RNA metabolism profiles. *Nature Structural and Molecular Biology* **24**, 86–96 (2017).

231. Lagarde, J. *et al.* High-throughput annotation of full-length long noncoding RNAs with capture long-read sequencing. *Nature Genetics* **49**, 1731–1740 (2017).

232. Ransohoff, J. D. *et al.* The functions and unique features of long intergenic non-coding RNA. *Nature Reviews Molecular Cell Biology* **19**, 143–157 (2018).

233. Hezroni, H. *et al.* Principles of Long Noncoding RNA Evolution Derived from Direct Comparison of Transcriptomes in 17 Resource Principles of Long Noncoding RNA Evolution Derived from Direct Comparison of Transcriptomes in 17 Species. *CellReports* **11**, 1110–1122 (2015).

234. Ariel, F. *et al.* Battles and hijacks: Noncoding transcription in plants. *Trends in Plant Science* **20**, 362–371 (2015).

235. King, G. J. Through a genome, darkly: Comparative analysis of plant chromosomal DNA. *Plant Molecular Biology* **48**, 5–20 (2002).

236. Barow, M. & Meister, A. Lack of correlation between AT frequency and genome size in higher plants and the effect of nonrandomness of base sequences on dye binding. *Cytometry* **47**, 1–7 (2002).

237. Šmarda, P. & Bureš, P. In (eds Wendel, J. F. *et al.*) 209–235 (Springer Vienna, Vienna, 2012).

238. Singh, R. *et al.* Comparative Analysis of GC Content Variations in Plant Genomes. *Tropical Plant Biology* **9**, 136–149 (2016).

239. King, G. J. & Ingrouille, M. J. DNA base composition heterogeneity in the grass genus *Briza L. Genome* **29**, 621–626 (1987).

240. Salinas, J. *et al.* Compositional compartmentalization and compositional patterns in the nuclear genomes of plants. *Nucleic Acids Research* **16**, 4269–4285 (1988).

241. Wang, H. C. *et al.* Mutational Bias Affects Protein Evolution in Flowering Plants. *Molecular Biology and Evolution* **21**, 90–96 (2004).

242. Carels, N. *et al.* Compositional properties of homologous coding sequences from plants. *Journal of Molecular Evolution* **46**, 45–53 (1998).

243. Tatarinova, T. V. *et al.* GC3biology in corn, rice, sorghum and other grasses. *BMC Genomics* **11** (2010).

244. Muyle, A. *et al.* GC-biased gene conversion and selection affect GC content in the *Oryza* genus (rice). *Molecular Biology and Evolution* **28**, 2695–2706 (2011).

245. Haas, B. J. *et al.* Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research,* 5654–5666.

246. Campbell, M. S. *et al.* MAKER-P : A Tool Kit for the Rapid Creation, Management, and Quality Control of Plant Genome Annotations. *Plant Physiology* **164**, 513–524 (2014).

247. Hoff, K. J. *et al.* BRAKER1: Unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics* **32**, 767–769 (2015).

248. Star, B. *et al.* The genome sequence of Atlantic cod reveals a unique immune system. *Nature* **477**, 207–10 (2011).

249. Weinert, B. T. & Timiras, P. S. Invited Review: Theories of aging. *Journal of Applied Physiology* **95**, 1706–1716 (2003).

250. Ouwehand, C. *et al.* A review of successful aging models: Proposing proactive coping as an important additional strategy. *Clinical Psychology Review* **27**, 873–884 (2007).

251. Jafferany, M. *et al.* Geriatric dermatoses: a clinical review of skin diseases in an aging population. eng. *International journal of dermatology* **51**, 509–522 (2012).

252. Ghadially, R. *et al.* The aged epidermal permeability barrier. Structural, functional, and lipid biochemical abnormalities in humans and a senescent murine model. *The Journal of Clinical Investigation* **95**, 2281–2290 (1995).

253. Sahle, F. F. *et al.* Skin Diseases Associated with the Depletion of Stratum Corneum Lipids and Stratum Corneum Lipid Substitution Therapy. *Skin Pharmacology and Physiology* **28**, 42–55 (2015).
254. Wohlrab, J. *et al.* Skin diseases in diabetes mellitus. *JDDG: Journal of the German Society of Dermatology* **5**, 37–53 (2007).
255. Yokota, M. & Tokudome, Y. The Effect of Glycation on Epidermal Lipid Content, Its Metabolism and Change in Barrier Function. *Skin Pharmacology and Physiology* **29**, 231–242 (2016).
256. Sakai, S. *et al.* Characteristics of the Epidermis and Stratum Corneum of Hairless Mice with Experimentally Induced Diabetes Mellitus. *Journal of Investigative Dermatology* **120**, 79–85 (2003).
257. Behm, B. *et al.* Impact of a Glycolic Acid-Containing pH 4 Water-in-Oil Emulsion on Skin pH. *Skin Pharmacology and Physiology* **28**, 290–295 (2015).
258. Jackson, S. M. *et al.* Pathobiology of the stratum corneum. eng. *The Western journal of medicine* **158**, 279–285 (1993).
259. Elias, P. M. Stratum corneum architecture, metabolic activity and interactivity with subjacent cell layers. eng. *Experimental dermatology* **5**, 191–201 (1996).
260. Danby, S. G. *et al.* The Effect of an Emollient Containing Urea, Ceramide NP, and Lactate on Skin Barrier Structure and Function in Older People with Dry Skin. *Skin Pharmacology and Physiology* **29**, 135–147 (2016).
261. Elias, P. M. Structure and Function of the Stratum Corneum Extracellular Matrix. *Journal of Investigative Dermatology* **132**, 2131–2133 (2012).
262. Ponec, M. *et al.* Regulation of lipid synthesis in relation to keratinocyte differentiation capacity. *Biochimica et Biophysica Acta (BBA) - Lipids and Lipid Metabolism* **921**, 512–521 (1987).
263. Farwanah, H. *et al.* Profiling of human stratum corneum ceramides by means of normal phase LC/APCI–MS. *Analytical and Bioanalytical Chemistry* **383**, 632–637 (2005).
264. Tessema, E. N. *et al.* Potential Applications of Phyto-Derived Ceramides in Improving Epidermal Barrier Function. *Skin Pharmacology and Physiology* **30**, 115–138 (2017).
265. Kessner, D. *et al.* Properties of Ceramides and Their Impact on the Stratum Corneum Structure: *Skin Pharmacology and Physiology* **21**, 58–74 (2008).
266. Gupta, R. *et al.* Molecular Dynamics Simulation of Skin Lipids: Effect of Ceramide Chain Lengths on Bilayer Properties. *The Journal of Physical Chemistry B* **120**, 12536–12546 (2016).
267. Norlén, L. *et al.* Stratum corneum lipid organization as observed by atomic force, confocal and two-photon excitation fluorescence microscopy. eng. *International journal of cosmetic science* **30**, 391–411 (2008).

268. Schröter, A. *et al.* Basic nanostructure of stratum corneum lipid matrices based on ceramides [EOS] and [AP]: a neutron diffraction study. *Biophysical journal* **97**, 1104–1114 (2009).

269. Katsuta, Y. *et al.* Unsaturated Fatty Acids Induce Calcium Influx into Keratinocytes and Cause Abnormal Differentiation of Epidermis. *Journal of Investigative Dermatology* **124**, 1008–1013 (2005).

270. Nikolakopoulou, Z. *et al.* Omega-3 polyunsaturated fatty acids selectively inhibit growth in neoplastic oral keratinocytes by differentially activating ERK1/2. *Carcinogenesis* **34**, 2716–2725 (2013).

271. Feingold, K. R. Thematic review series: Skin Lipids. The role of epidermal lipids in cutaneous permeability barrier homeostasis. *Journal of Lipid Research* **48**, 2531–2546 (2007).

272. Kondo, N. *et al.* Identification of the phytosphingosine metabolic pathway leading to odd-numbered fatty acids. *Nature Communications* **5**, 5338 (2014).

273. Hinder, A. *et al.* Investigation of the molecular structure of the human stratum corneum ceramides [NP] and [EOS] by mass spectrometry. eng. *Skin pharmacology and physiology* **24**, 127–135 (2011).

274. Latulippe, M. E. *et al.* ILSI Brazil International Workshop on Functional Foods: a narrative review of the scientific evidence in the area of carbohydrates, microbiome, and health. *Food & Nutrition Research* **0** (2013).

275. Gurr, M. I. & Harwood, J. L. In *Lipid Biochemistry* 23–118 (Springer, 1991).

276. Ananthapadmanabhan, K. P. *et al.* Stratum corneum fatty acids: their critical role in preserving barrier integrity during cleansing. eng. *International journal of cosmetic science* **35**, 337–345 (2013).

277. Grubauer, G. *et al.* Relationship of epidermal lipogenesis to cutaneous barrier function. *Journal of Lipid Research* **28**, 746–752 (1987).

278. Laposata, M. Fatty Acids: Biochemistry to Clinical Significance. *American Journal of Clinical Pathology* **104**, 172–179 (1995).

279. Jakobsson, A. *et al.* Fatty acid elongases in mammals: Their regulation and roles in metabolism. *Progress in Lipid Research* **45**, 237–249 (2006).

280. Feingold, K. R. Lamellar Bodies: The Key to Cutaneous Barrier Function. *Journal of Investigative Dermatology* **132**, 1951–1953 (2012).

281. Jia, Y. *et al.* The mechanism of skin lipids influencing skin status. *Journal of Dermatological Science* **89**, 112–119 (2018).

282. Van Smeden, J. *et al.* The importance of free fatty acid chain length for the skin barrier function in atopic eczema patients. eng. *Experimental dermatology* **23**, 45–52 (2014).

283. Li, S. *et al.* Lipidomic analysis of epidermal lipids: a tool to predict progression of inflammatory skin disease in humans. *Expert Review of Proteomics* **13**, 451–456 (2016).

284. Van Smeden, J. & Bouwstra, J. A. In *Current Problems in Dermatology* 8–26 (2016).

285. Joo, K.-M. *et al.* Relationship of ceramide–, and free fatty acid–cholesterol ratios in the stratum corneum with skin barrier function of normal, atopic dermatitis lesional and non-lesional skins. *Journal of dermatological science* **77**, 71–74 (2015).

286. Smesny, S. *et al.* Skin ceramide alterations in first-episode schizophrenia indicate abnormal sphingolipid metabolism. eng. *Schizophrenia bulletin* **39**, 933–941 (2013).

287. Goebel, A. S. B. *et al.* Dermal targeting using colloidal carrier systems with linoleic acid. *European Journal of Pharmaceutics and Biopharmaceutics* **75**, 162–172 (2010).

288. R Development Core Team. *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing (Vienna, Austria, 2018).

289. Shapiro, S. S. & Wilk, M. B. An Analysis of Variance Test for Normality (Complete Samples). *Biometrika* **52**, 591–611 (1965).

290. Liu, R. Y. & Singh, K. A Quality Index Based on Data Depth and Multivariate Rank Tests. *Journal of the American Statistical Association* **88**, 252–260 (1993).

291. Agency, E. M. Guideline on validation of bioanalytical methods (2011).

292. Norlén, L. *et al.* Inter-and intra-individual differences in human stratum corneum lipid content related to physical parameters of skin barrier function in vivo. *Journal of investigative dermatology* **112**, 72–77 (1999).

293. Norlén, L. *et al.* A new HPLC-based method for the quantitative analysis of inner stratum corneum lipids with special reference to the free fatty acid fraction. eng. *Archives of dermatological research* **290**, 508–516 (1998).

294. Linseisen, J. & Wolfram, G. Odd-Numbered Medium-Chain Triglycerides (Trinonanoin) in Total Parenteral Nutrition: Parameters of Carbohydrate and Protein Metabolism. *Annals of Nutrition and Metabolism* **37**, 320–327 (1993).

295. Nicollier, M. *et al.* Free fatty acids and fatty acids of triacylglycerols in normal and hyperkeratotic human stratum corneum. eng. *The Journal of investigative dermatology* **87**, 68–71 (1986).

296. Lampe, M. A. *et al.* Human stratum corneum lipids: characterization and regional variations. *Journal of Lipid Research* **24**, 120–130 (1983).

297. Lavrijsen, A. P. M. *et al.* Validation of an in vivo extraction method for human stratum corneum ceramides. *Archives of Dermatological Research* **286**, 495–503 (1994).

298. Kalousek, F. *et al.* Isolation and characterization of propionyl-CoA carboxylase from normal human liver. Evidence for a protomeric tetramer of nonidentical subunits. *Journal of Biological Chemistry* **255**, 60–65 (1980).

299. Elias, P. M. *et al.* Thematic review series: Skin Lipids. Pathogenesis of permeability barrier abnormalities in the ichthyoses: inherited disorders of lipid metabolism. *Journal of Lipid Research* **49**, 697–714 (2008).

300. Rogers, J. *et al.* Stratum corneum lipids: the effect of ageing and the seasons. *Archives of Dermatological Research* **288**, 765–770 (1996).

301. Yosipovitch, G. *et al.* Skin surface pH in intertriginous areas in NIDDM patients. Possible correlation to candidal intertrigo. eng. *Diabetes care* **16**, 560–563 (1993).

302. Mackiewicz-Wysocka, M. *et al.* Skin pH Is Lower in Type 1 Diabetes Subjects and Is Related to Glycemic Control of the Disease. *Diabetes Technology & Therapeutics* **17**, 16–20 (2014).

303. Neubert, R. H. H. *et al.* Controlled Penetration of a Novel Dimeric Ceramide into and across the Stratum Corneum Using Microemulsions and Various Types of Semisolid Formulations. eng. *Skin pharmacology and physiology* **29**, 130–134 (2016).

304. Buldo, P. *et al.* Multivariate data analysis for finding the relevant fatty acids contributing to the melting fractions of cream. *Journal of the Science of Food and Agriculture* **93**, 1620–1625 (2013).

305. Shapiro, M. (& Rosen, G. H. Topical Oil Applications in Essential Fatty Acid Deficiency. *Nutrition in Clinical Practice* **4**, 140–144 (1989).

306. Daehnhardt, D. *et al.* The Influence of Two Different Foam Creams on Skin Barrier Repair of Foot Xerosis: A Prospective, Double-Blind, Randomised, Placebo-Controlled Intra-Individual Study. *Skin Pharmacology and Physiology* **29**, 266–272 (2016).

307. Schutkowski, A. *et al.* Metabolic footprint and intestinal microbial changes in response to dietary proteins in a pig model. *The Journal of Nutritional Biochemistry* **67**, 149–160 (2019).

308. Chalvon-Demersay, T. *et al.* A systematic review of the effects of plant compared with animal protein sources on features of metabolic syndrome1-3. *Journal of Nutrition* **147**, 281–292 (2017).

309. Grosso, G. *et al.* Health risk factors associated with meat, fruit and vegetable consumption in cohort studies: A comprehensive meta-analysis. *PLoS ONE* **12**, 1–21 (2017).

310. Tian, S. *et al.* Dietary protein consumption and the risk of type 2 diabetes: A systematic review and meta-analysis of cohort studies. *Nutrients* **9**, 1–17 (2017).

311. Aune, D. *et al.* Meat consumption and the risk of type 2 diabetes: A systematic review and meta-analysis of cohort studies. *Diabetologia* **52**, 2277–2287 (2009).

312. Larsson, S. C. *et al.* Red meat consumption and risk of cancers of the proximal colon, distal colon and rectum: The Swedish Mammography Cohort. *International Journal of Cancer* **113**, 829–834 (2005).

313. Wu, J. *et al.* Dietary protein sources and incidence of breast cancer: A dose-response meta-analysis of prospective studies. *Nutrients* **8** (2016).

314. Bechthold, A. *et al.* Food groups and risk of coronary heart disease, stroke and heart failure: A systematic review and dose-response meta-analysis of prospective studies. *Critical Reviews in Food Science and Nutrition* **59**, 1071–1090 (2017).

315. Schwingshackl, L. *et al.* Food groups and risk of hypertension: A systematic review and dose-response meta-analysis of prospective studies. *Advances in Nutrition* **8**, 793–803 (2017).

316. Lin, Y. *et al.* Dietary animal and plant protein intakes and their associations with obesity and cardio-metabolic indicators in European adolescents: The HELENA cross-sectional study. *Nutrition Journal* **14**, 1–11 (2015).

317. Sinha, R. *et al.* Meat, meat cooking methods and preservation, and risk for colorectal adenoma. *Cancer Research* **65**, 8034–8041 (2005).

318. Davey, G. K. *et al.* EPIC–Oxford:lifestyle characteristics and nutrient intakes in a cohort of 33 883 meat-eaters and 31 546 non meat-eaters in the UK. *Public Health Nutrition* **6**, 259–268 (2003).

319. Walter, P. *et al.* *Gesundheitliche Vor- und Nachteile einer vegetarischen Ernährung. Expertenbericht der Eidgenössischen Ernährungskommission* tech. rep. (Bern, 2006).

320. Sirtori, C. R. *et al.* Effects of Dietary Proteins on the Regulation of Liver Lipoprotein Receptors in Rats. *The Journal of Nutrition* **114**, 1493–1500 (1984).

321. Maki, K. C. *et al.* Effects of soy protein on lipoprotein lipids and fecal bile acid excretion in men and women with moderate hypercholesterolemia. *Journal of Clinical Lipidology* **4**, 531–542 (2010).

322. Schutkowski, A. *et al.* Additive effects of lupin protein and phytic acid on aortic calcification in ApoE deficient mice. *Journal of Clinical and Translational Endocrinology* **2**, 6–13 (2015).

323. Sirtori, C. R. *et al.* Proteins of White Lupin Seed, a Naturally Isoflavone-Poor Legume, Reduce Cholesterolemia in Rats and Increase LDL Receptor Activity in HepG2 Cells. *Journal of Nutrition* **134**, 18–23 (2004).

324. Weisse, K. *et al.* Lupin protein compared to casein lowers the LDL cholesterol: HDL cholesterol-ratio of hypercholesterolemic adults. *European Journal of Nutrition* **49**, 65–71 (2010).

325. Sucher, S. *et al.* Comparison of the effects of diets high in animal or plant protein on metabolic and cardiovascular markers in type 2 diabetes: A randomized clinical trial. eng. *Diabetes, obesity & metabolism* **19**, 944–952 (2017).

326. Wofford, M. R. *et al.* Effect of soy and milk protein supplementation on serum lipid levels: A randomized controlled trial. *European Journal of Clinical Nutrition* **66**, 419–425 (2012).

327. Miller, E. R. & Ullrey, D. E. The Pig as a Model for Human Nutrition. *Annual Review of Nutrition* **7**, 361–382 (1987).

328. Xiao, L. *et al.* A reference gene catalogue of the pig gut microbiome. *Nature Microbiology* **1**, 1–6 (2016).

329. Roura, E. *et al.* Critical review evaluating the pig as a model for human nutritional physiology. *Nutrition Research Reviews* **29**, 60–90 (2016).

330. National Research Council. *Guide for the care and use of laboratory animals* 8th editio (National Academies Press, Washington D.C., 2010).

331. Flachowsky, G. *et al.* *Empfehlungen zur Energie- und Nährstoffversorgung von Schweinen* (DLG, 2006).

332. National Research Council. *Nutrient requirements of swine* 11th revis (National Academies Press, 2012).

333. Bassler, R. *Die chemische Untersuchung von Futtermitteln* (VDLUFA-Verlag, 1976).

334. Dawczynski, C. *et al.* Amino acids, fatty acids, and dietary fibre in edible seaweed products. *Food Chemistry* **103**, 891–899 (2007).

335. Gil-Agustí, M. *et al.* Anserine and carnosine determination in meat samples by pure micellar liquid chromatography. *Journal of Chromatography A* **1189**, 444–450 (2008).

336. Young, V. R. & Munro, H. N. Ntau-methylhistidine (3-methylhistidine) and muscle protein turnover: an overview. eng. *Federation proceedings* **37**, 2291–2300 (1978).

337. Altorf-van der Kuil, W. *et al.* Identification of biomarkers for intake of protein from meat, dairy products and grains: a controlled dietary intervention study. *British Journal of Nutrition* **110**, 810–822 (2013).

338. Cross, A. J. *et al.* Urinary biomarkers of meat consumption. *Cancer Epidemiology Biomarkers and Prevention* **20**, 1107–1111 (2011).

339. Mitchell, M. E. Carnitine metabolism in human subjects I. Normal metabolism. *The American Journal of Clinical Nutrition* **31**, 293–306 (1978).

340. Krajcovicová-Kudlácková, M. *et al.* Correlation of carnitine levels to methionine and lysine intake. eng. *Physiological research* **49**, 399–402 (2000).

341. Demarquoy, J. *et al.* Radioisotopic determination of l-carnitine content in foods commonly eaten in Western countries. *Food Chemistry* **86**, 137–142 (2004).

342. Rigault, C. *et al.* Changes in l-carnitine content of fish and meat during domestic cooking. *Meat Science* **78**, 331–335 (2008).

343. Servillo, L. *et al.* Where Does Nϵ-Trimethyllysine for the Carnitine Biosynthesis in Mammals Come from? *PLOS ONE* **9**, e84589 (2014).

344. Tsikas, D. & Wu, G. Homoarginine, arginine, and relatives: analysis, metabolism, transport, physiology, and pathology. *Amino Acids* **47**, 1697–1702 (2015).

345. Atzler, D. *et al.* L-Homoarginine and cardiovascular disease. *Current Opinion in Clinical Nutrition & Metabolic Care* **18** (2015).

346. Pilz, S. *et al.* Low homoarginine concentration is a novel risk factor for heart disease. *Heart* **97**, 1222 LP –1227 (2011).

347. Welch, A. A. *et al.* Urine pH is an indicator of dietary acid–base load, fruit and vegetables and meat intakes: results from the European Prospective Investigation into Cancer and Nutrition (EPIC)-Norfolk population study. *British Journal of Nutrition* **99**, 1335–1343 (2008).

348. Moe, S. M. *et al.* Vegetarian Compared with Meat Dietary Protein Source and Phosphorus Homeostasis in Chronic Kidney Disease. *Clinical Journal of the American Society of Nephrology* **6**, 257 LP –264 (2011).

349. Craig, S. A. S. Betaine in human nutrition. *The American Journal of Clinical Nutrition* **80**, 539–549 (2004).

350. Sakamoto, A. *et al.* Betaine and homocysteine concentrations in infant formulae and breast milk. eng. *Pediatrics international : official journal of the Japan Pediatric Society* **43**, 637–640 (2001).

351. Zeisel, S. H. *et al.* Concentrations of Choline-Containing Compounds and Betaine in Common Foods. *The Journal of Nutrition* **133**, 1302–1307 (2003).

352. Wolk, A. *et al.* Evaluation of a biological marker of dairy fat intake. *The American Journal of Clinical Nutrition* **68**, 291–295 (1998).

353. Keating, S. T. & El-Osta, A. Epigenetics and Metabolism. *Circulation Research* **116**, 715–736 (2015).

354. Obeid, R. The metabolic burden of methyl donor deficiency with focus on the betaine homocysteine methyltransferase pathway. *Nutrients* **5**, 3481–3495 (2013).

355. Tammen, S. A. *et al.* Epigenetics: The link between nature and nurture. *Molecular Aspects of Medicine* **34**, 753–764 (2013).

356. De Rosa, S. *et al. Type 2 Diabetes Mellitus and Cardiovascular Disease: Genetic and Epigenetic Links* 2018.

357. Yap, S. Classical homocystinuria: Vascular risk and its prevention. *Journal of Inherited Metabolic Disease* **26**, 259–265 (2003).

358. ElGendy, K. *et al.* Effects of dietary interventions on DNA methylation in adult humans: systematic review and meta-analysis. *British Journal of Nutrition* **120**, 961–976 (2018).

359. Radtke, J. *et al.* Lupin protein isolate versus casein modifies cholesterol excretion and mRNA expression of intestinal sterol transporters in a pig model. *Nutrition & Metabolism* **11**, 9 (2014).

360. Woudberg, N. J. *et al. Pharmacological Intervention to Modulate HDL: What Do We Target?* 2018.

361. Li, X. S. *et al.* Untargeted metabolomics identifies trimethyllysine, a TMAO-producing nutrient precursor, as a predictor of incident cardiovascular disease risk. *JCI insight* **3** (2018).

362. Wang, Z. *et al.* Prognostic value of choline and betaine depends on intestinal microbiota-generated metabolite trimethylamine-N-oxide. *European Heart Journal* **35**, 904–910 (2014).

363. Schlesinger, S. *et al.* Asymmetric and Symmetric Dimethylarginine as Risk Markers for Total Mortality and Cardiovascular Outcomes: A Systematic Review and Meta-Analysis of Prospective Studies. *PLOS ONE* **11**, e0165811 (2016).

364. Zhao, Z. *et al.* Red and processed meat consumption and esophageal cancer risk: a systematic review and meta-analysis. *Oncotarget* **8**, 83306–14 (2017).

365. English, D. R. *et al.* Red Meat, Chicken, and Fish Consumption and Risk of Colorectal Cancer. *Cancer Epidemiology Biomarkers &amp; Prevention* **13**, 1509 LP –1514 (2004).

366. Guéraud, F. *et al.* Dietary polyunsaturated fatty acids and heme iron induce oxidative stress biomarkers and a cancer promoting environment in the colon of rats. *Free Radical Biology and Medicine* **83**, 192–200 (2015).

367. Ijssennagger, N. *et al.* Dietary heme induces acute oxidative stress, but delayed cytotoxicity and compensatory hyperproliferation in mouse colon. *Carcinogenesis* **34**, 1628–1635 (2013).

12

Appendix

This chapter contains additional figures and tables. Further supplementary figures and tables of published articles can be found on the publishers' websites.

Species	5' UTR Mean ± SD	3' UTR Mean ± SD
A. thaliana	277.40 ± 292.17	415.31 ± 421.95
A. lyrata	221.62 ± 263.67	313.39 ± 348.94
C. rubella	253.22 ± 290.19	484.07 ± 552.56
E. salsugineum	273.81 ± 390.82	416.27 ± 480.33
T. hassleriana	264.68 ± 319.48	348.05 ± 470.19
M. truncatula	271.11 ± 320.60	570.94 ± 613.95
B. distachyon	382.98 ± 546.25	651.26 ± 748.50

Table 12.1 | Statistics UTR lengths. Arithmetic mean ± standard deviation of 5' and 3' UTR sequences from all protein-coding transcripts having annotated UTRs.

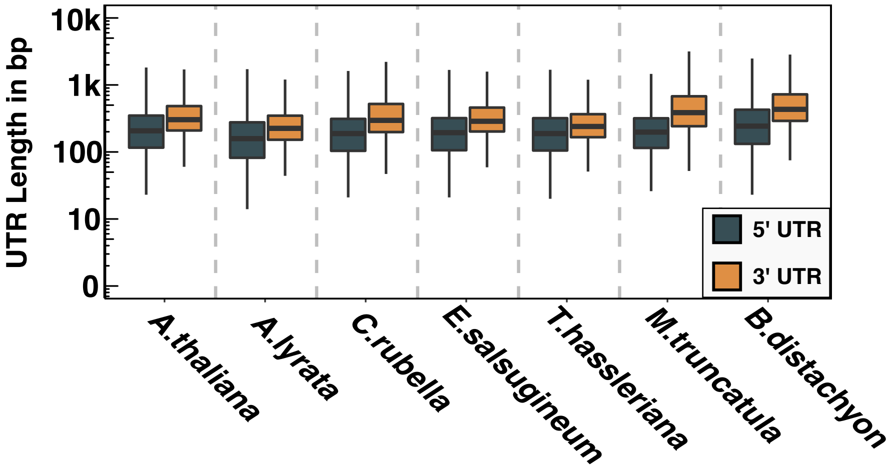


Figure 12.1 | 5' and 3' UTR lengths of protein-coding transcripts.. In dark gray the 5' UTR lengths are between 220 and 380 bp long with a variation between 300 and 500 bp. In orange the 3' UTR lengths are between 300 and 650 bp long with a standard variation between 340 and 740 bp. Table 12.1 presents a detailed list of observed UTR lengths in all 7 species.

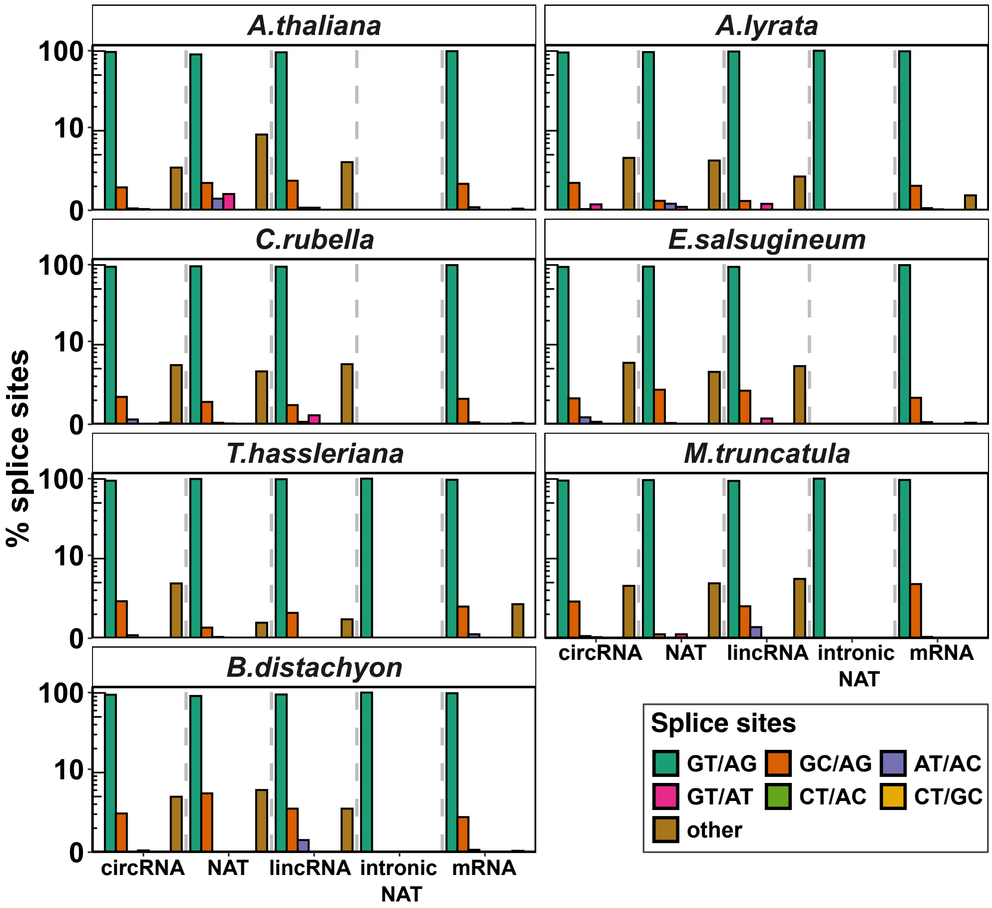


Figure 12.2 | Overview of linear splice sites among RNA species. Percentage of splice sites detected within each transcript isoform of circRNAs, NATs, lincRNAs, intronic NATs, and protein-coding mRNAs of each plant species. Displayed are the canonical donor/acceptor splice sites. Non-canonical splice sites are summarized as “other”.

ID	Species	Organ sample	Age
1	*A. thaliana*	root, root tip (top 0.5-1mm)	5d
2	*A. thaliana*	root, maturation zone (2-3mm piece above root tip)	5d
3	*A. thaliana*	root, whole root (3mm piece including root tip)	5d
4	*A. thaliana*	root, whole root	7d
5	*A. thaliana*	root, whole root	14d
6	*A. thaliana*	root, whole root	21d
7	*A. thaliana*	hypocotyl	10d
8	*A. thaliana*	3rd internode (0.5cm piece)	24d
9	*A. thaliana*	2nd internode (1cm piece)	24d
10	*A. thaliana*	1st internode (1cm piece)	28d+
11	*A. thaliana*	cotyledons	7d
12	*A. thaliana*	leaves 1+2	7d
13	*A. thaliana*	leaves 1+2	10d
14	*A. thaliana*	leaves 1+2 - petiole	10d
15	*A. thaliana*	leaves 1+2 - leaf tip	10d
16	*A. thaliana*	leaves 5+6	17d
17	*A. thaliana*	leaves 9+10	27d
18	*A. thaliana*	leaves senescing	35d
19	*A. thaliana*	cauline leaves	24d
20	*A. thaliana*	apex, vegetative	7d
21	*A. thaliana*	apex, vegetative	10d
22	*A. thaliana*	apex, vegetative, before bolting	14d
23	*A. thaliana*	apex, inflorescence (dissected up to stage 4/5 flower)	21d
24	*A. thaliana*	apex, inflorescence, clv1-8 (dissected as above)	21d+
25	*A. thaliana*	apex, inflorescence (dissected as above)	28d
26	*A. thaliana*	flower stage 9	21d+

Continued on next page

ID	Species	Organ sample	Age
27	*A. thaliana*	flower stage 10/11	21d+
28	*A. thaliana*	flower stage 12	21d+
29	*A. thaliana*	flower stage 15	21d+
30	*A. thaliana*	sepals stage 12	21d+
31	*A. thaliana*	sepals stage 15	21d+
32	*A. thaliana*	petals stage 12	21d+
33	*A. thaliana*	petals stage 15	21d+
34	*A. thaliana*	stamens stage 12	21d+
35	*A. thaliana*	stamens stage 15	21d+
36	*A. thaliana*	mature pollen	28d
37	*A. thaliana*	carpels stage 12 (early mid-stage)	21d+
38	*A. thaliana*	carpels stage 12 (late-stage)	21d+
39	*A. thaliana*	carpels stage 15 (with seeds - pre-globular stage embryo)	21d+
40	*A. thaliana*	fruit stage 16 - siliques with seeds (pre-globular to globular stage embryo)	28d+
41	*A. thaliana*	fruit stage 17a - siliques with seeds (heart stage embryo)	28d+
42	*A. thaliana*	seeds - fruit stage 16 (pre-globular and globular stage embryo)	28d+
43	*A. thaliana*	seeds - fruit stage 17a (heart to torpedo stage embryo)	28d+
44	*A. thaliana*	seeds - fruit stage 18 (mature green stage embryo)	28d+
45	*A. lyrata*	root, whole root (3mm piece including root tip)	6d
46	*A. lyrata*	hypocotyl	12d
47	*A. lyrata*	leaves 1+2	10d
48	*A. lyrata*	apex, vegetative	11d
49	*A. lyrata*	apex, inflorescence (dissected up to stage 4/5 flower)	8w/10w/13d
50	*A. lyrata*	flower stage 12 (mid-stage)	8w/10w/15d
51	*A. lyrata*	mature pollen	8w/10w/25d
52	*A. lyrata*	carpels stage 12 (early mid-stage)	8w/10w/17d
53	*A. lyrata*	stamen stage 11	8w/10w/25d

Continued on next page

ID	Species	Organ sample	Age
54	*A. lyrata*	stamen stage 12 (early-stage)	8w/10w/23d
55	*A. lyrata*	stamen stage 12 (mid-stage)	8w/10w/24d
56	*A. lyrata*	stamen stage 12 (late-stage)	8w/10w/21d
57	*C. rubella*	root, whole root (3mm piece including root tip)	4d
58	*C. rubella*	hypocotyl	9d
59	*C. rubella*	leaves 1+2	7d
60	*C. rubella*	apex, vegetative	7d
61	*C. rubella*	apex, inflorescence (dissected up to stage 4/5 flower)	6w/7w/17d
62	*C. rubella*	flower stage 12 (mid-stage)	6w/7w/20d
63	*C. rubella*	mature pollen	6w/7w/22d
64	*C. rubella*	carpels stage 12 (early mid-stage)	6w/7w/21d
65	*C. rubella*	stamen stage 12 (mid-stage)	6w/7w/20d
66	*E. salsugineum*	root, whole root (3mm piece including root tip)	6d
67	*E. salsugineum*	hypocotyl	12d
68	*E. salsugineum*	leaves 1+2	9d
69	*E. salsugineum*	apex, vegetative	9d
70	*E. salsugineum*	apex, inflorescence (dissected up to stage 4/5 flower)	7w/8w/12d
71	*E. salsugineum*	flower stage 12 (mid-stage)	7w/8w/15d
72	*E. salsugineum*	mature pollen	7w/8w/17d
73	*E. salsugineum*	carpels stage 12 (early mid-stage)	7w/8w/18d
74	*E. salsugineum*	stamen stage 12 (mid-stage)	7w/8w/17d
75	*T. hassleriana*	root, whole root (3mm piece including root tip)	5d
76	*T. hassleriana*	hypocotyl	11d
77	*T. hassleriana*	leaves 1+2	11d
78	*T. hassleriana*	apex, vegetative	11d
79	*T. hassleriana*	apex, inflorescence	11w

Continued on next page

ID	Species	Organ sample	Age
80	*T. hassleriana*	flower stage 12 equivalent	11w
81	*T. hassleriana*	mature pollen	11w
82	*T. hassleriana*	carpels stage 12 equivalent	11w
83	*T. hassleriana*	stamen stage 12 equivalent	11w
84	*M. truncatula*	root, whole root (3mm piece including root tip)	4d
85	*M. truncatula*	hypocotyl	8d
86	*M. truncatula*	leaf 2	7d
87	*M. truncatula*	apex, vegetative	6d
88	*M. truncatula*	inflorescence meristem including I1, I2 and F, Br, L and Stp primordia	7w
89	*M. truncatula*	Medicago flower stage 8 (before anther dehiscence)	7w
90	*M. truncatula*	mature pollen	7w
91	*M. truncatula*	carpels Medicago flower stage 8 (inflected carpel)	7w
92	*M. truncatula*	stamen Medicago flower stage 8 (before anther dehiscence)	7w
93	*B. distachyon*	root, whole root (3mm piece including root tip)	4d
94	*B. distachyon*	mesocotyl	7d
95	*B. distachyon*	leaf 1	4d
96	*B. distachyon*	apex, vegetative	5d
97	*B. distachyon*	lateral spiklet meristem ('naked-stage' to 'awn initiation' stage meristem)	18-20d
98	*B. distachyon*	floret with anthers immediatly before dehiscence from basal lateral spiklet	32d
99	*B. distachyon*	mature pollen from basal lateral spiklet	32d
100	*B. distachyon*	carpels 'stage 12 equivalent' from basal lateral spiklet)	32d
101	*B. distachyon*	stamen 'stage 12 equivalent' from basal lateral spiklet)	32d

Continued on next page

ID	Species	Organ sample	Age

Table 12.2 | Overview sequenced species and organs. We sequenced from the seven flowering plants the organs root, hyphocotyle (mesocotyle for *B. distachyon*), leaf, vegetative and inflorescence apex (lateral spikelet meristem for *B. distachyon*), carpels, and stamens. Additionally, we sequenced for each species the mature pollen consisting of two cell types. In contrast to all other sequenced species in this study, we chose for *A. thaliana* additional developmental stages of organs, resulting in 44 biological samples for this species. All biological samples are represented by three independent biological replicates, which sums up to a total of 303 sequencing libraries.

Species	**Release**	**Annotation data**	**Genome sequence**
A. thaliana	Ensembl 34	`Arabidopsis_thaliana.TAIR10.34`	`Arabidopsis_thaliana.TAIR10.dna.toplevel`
A. lyrata	Phytozome v.12	`Alyrata_384_v2.1.gene_exons`	`Alyrata_384_v1`
C. rubella	Phytozome v.12	`Crubella_183_v1.0.gene_exons`	`Crubella_183_v1`
E. salsugineum	Phytozome v.12	`Esalsugineum_173_v1.0.gene_exons`	`Esalsugineum_173_v1`
T. hassleriana	RefSeq	`GCF_000463585.1_ASM46358v1_genomic`	`GCF_000463585.1_ASM46358v1_genomic`
M. truncatula	Phytozome v.12	`Mtruncatula_285_Mt4.0v1.0.gene_exons`	`Mtruncatula_285_Mt4.0`
B. distachyon	Phytozome v.12	`Bdistachyon_314_v3.1.gene_exons`	`Bdistachyon_314_v3.0`

Table 12.3 | Sources of reference annotations. List of reference annotations and their corresponding genome sequences updated by the DevSeq workflow.

Species	RNA species	Percentile	Transcript length in bp	GC
A. thaliana	circRNA	0-20%	(0, 90]	0.44 ± 0.07
	circRNA	20-40%	(90, 138]	0.44 ± 0.06
	circRNA	40-60%	(138, 249]	0.44 ± 0.05
	circRNA	60-80%	(249, 421]	0.44 ± 0.04
	circRNA	80-100%	(421, 5.6e+03]	0.43 ± 0.04
	lincRNA	0-20%	(0, 399]	0.37 ± 0.06
	lincRNA	20-40%	(399, 533]	0.36 ± 0.05
	lincRNA	40-60%	(533, 682]	0.36 ± 0.05
	lincRNA	60-80%	(682, 957]	0.36 ± 0.05
	lincRNA	80-100%	(957, 6.28e+03]	0.37 ± 0.05
	NAT	0-20%	(0, 480]	0.40 ± 0.05
	NAT	20-40%	(480, 687]	0.40 ± 0.05
	NAT	40-60%	(687, 974]	0.40 ± 0.04
	NAT	60-80%	(974, 1.48e+03]	0.41 ± 0.04
	NAT	80-100%	(1.48e+03, 7.28e+03]	0.41 ± 0.03
	mRNA	0-20%	(0, 1.07e+03]	0.40 ± 0.04
	mRNA	20-40%	(1.07e+03, 1.49e+03]	0.41 ± 0.03
	mRNA	40-60%	(1.49e+03, 1.92e+03]	0.41 ± 0.02
	mRNA	60-80%	(1.92e+03, 2.57e+03]	0.41 ± 0.02
	mRNA	80-100%	(2.57e+03, 3.09e+04]	0.42 ± 0.02
A. lyrata	circRNA	0-20%	(0, 82]	0.44 ± 0.07
	circRNA	20-40%	(82, 132]	0.45 ± 0.07
	circRNA	40-60%	(132, 261]	0.45 ± 0.06
	circRNA	60-80%	(261, 460]	0.44 ± 0.04
	circRNA	80-100%	(460, 2.54e+04]	0.42 ± 0.04
	lincRNA	0-20%	(0, 357]	0.38 ± 0.06
	lincRNA	20-40%	(357, 459]	0.38 ± 0.05
	lincRNA	40-60%	(459, 610]	0.38 ± 0.05
	lincRNA	60-80%	(610, 872]	0.37 ± 0.05
	lincRNA	80-100%	(872, 6.93e+03]	0.38 ± 0.05
	NAT	0-20%	(0, 485]	0.41 ± 0.05
	NAT	20-40%	(485, 659]	0.41 ± 0.04
	NAT	40-60%	(659, 915]	0.41 ± 0.04
	NAT	60-80%	(915, 1.39e+03]	0.41 ± 0.04
	NAT	80-100%	(1.39e+03, 5.81e+03]	0.41 ± 0.03
	mRNA	0-20%	(0, 875]	0.41 ± 0.04
	mRNA	20-40%	(875, 1.27e+03]	0.41 ± 0.03
	mRNA	40-60%	(1.27e+03, 1.68e+03]	0.42 ± 0.03
	mRNA	60-80%	(1.68e+03, 2.26e+03]	0.42 ± 0.02
	mRNA	80-100%	(2.26e+03, 1.66e+04]	0.42 ± 0.02
C. rubella	circRNA	0-20%	(0, 72]	0.45 ± 0.08
	circRNA	20-40%	(72, 104]	0.44 ± 0.07
	circRNA	40-60%	(104, 198]	0.45 ± 0.06
	circRNA	60-80%	(198, 457]	0.44 ± 0.04
	circRNA	80-100%	(457, 5.48e+03]	0.41 ± 0.05
	lincRNA	0-20%	(0, 386]	0.37 ± 0.06
	lincRNA	20-40%	(386, 528]	0.36 ± 0.05
	lincRNA	40-60%	(528, 713]	0.35 ± 0.05
	lincRNA	60-80%	(713, 1.14e+03]	0.37 ± 0.05
	lincRNA	80-100%	(1.14e+03, 8.18e+03]	0.37 ± 0.04

Continued on next page

Species	RNA species	Percentile	Transcript length in bp	GC
	NAT	0-20%	(0, 495]	0.40 ± 0.05
	NAT	20-40%	(495, 693]	0.41 ± 0.05
	NAT	40-60%	(693, 998]	0.41 ± 0.04
	NAT	60-80%	(998, 1.5e+03]	0.41 ± 0.04
	NAT	80-100%	(1.5e+03, 7.22e+03]	0.40 ± 0.04
	mRNA	0-20%	(0, 937]	0.42 ± 0.040
	mRNA	20-40%	(937, 1.36e+03]	0.42 ± 0.03
	mRNA	40-60%	(1.36e+03, 1.81e+03]	0.42 ± 0.03
	mRNA	60-80%	(1.81e+03, 2.5e+03]	0.42 ± 0.02
	mRNA	80-100%	(2.5e+03, 1.68e+04]	0.42 ± 0.02
E. salsugineum	circRNA	0-20%	(0, 72]	0.46 ± 0.08
	circRNA	20-40%	(72, 98]	0.45 ± 0.07
	circRNA	40-60%	(98, 152]	0.45 ± 0.08
	circRNA	60-80%	(152, 375]	0.45 ± 0.05
	circRNA	80-100%	(375, 3.84e+03]	0.43 ± 0.04
	lincRNA	0-20%	(0, 437]	0.36 ± 0.06
	lincRNA	20-40%	(437, 595]	0.36 ± 0.06
	lincRNA	40-60%	(595, 801]	0.37 ± 0.05
	lincRNA	60-80%	(801, 1.25e+03]	0.37 ± 0.05
	lincRNA	80-100%	(1.25e+03, 1.01e+04]	0.38 ± 0.05
	NAT	0-20%	(0, 533]	0.41 ± 0.05
	NAT	20-40%	(533, 753]	0.42 ± 0.05
	NAT	40-60%	(753, 1.05e+03]	0.41 ± 0.05
	NAT	60-80%	(1.05e+03, 1.57e+03]	0.42 ± 0.05
	NAT	80-100%	(1.57e+03, 6.69e+03]	0.41 ± 0.04
	mRNA	0-20%	(0, 909]	0.42 ± 0.04
	mRNA	20-40%	(909, 1.32e+03]	0.42 ± 0.03
	mRNA	40-60%	(1.32e+03, 1.75e+03]	0.43 ± 0.03
	mRNA	60-80%	(1.75e+03, 2.37e+03]	0.43 ± 0.03
	mRNA	80-100%	(2.37e+03, 1.87e+04]	0.43 ± 0.02
T. hassleriana	circRNA	0-20%	(0, 95]	0.48 ± 0.07
	circRNA	20-40%	(95, 157]	0.50 ± 0.08
	circRNA	40-60%	(157, 295]	0.48 ± 0.05
	circRNA	60-80%	(295, 602]	0.46 ± 0.04
	circRNA	80-100%	(602, 1.97e+04]	0.46 ± 0.04
	lincRNA	0-20%	(0, 323]	0.43 ± 0.07
	lincRNA	20-40%	(323, 474]	0.41 ± 0.07
	lincRNA	40-60%	(474, 723]	0.41 ± 0.06
	lincRNA	60-80%	(723, 1.34e+03]	0.41 ± 0.05
	lincRNA	80-100%	(1.34e+03, 1.15e+04]	0.41 ± 0.03
	NAT	0-20%	(0, 515]	0.45 ± 0.06
	NAT	20-40%	(515, 764]	0.45 ± 0.05
	NAT	40-60%	(764, 1.17e+03]	0.45 ± 0.05
	NAT	60-80%	(1.17e+03, 1.85e+03]	0.44 ± 0.04
	NAT	80-100%	(1.85e+03, 8.93e+03]	0.43 ± 0.03
	mRNA	0-20%	(0, 1.07e+03]	0.46 ± 0.04
	mRNA	20-40%	(1.07e+03, 1.49e+03]	0.46 ± 0.03
	mRNA	40-60%	(1.49e+03, 1.92e+03]	0.46 ± 0.03
	mRNA	60-80%	(1.92e+03, 2.60e+03]	0.46 ± 0.03
	mRNA	80-100%	(2.60e+03, 1.66e+04]	0.45 ± 0.03

Continued on next page

Species	**RNA species**	**Percentile**	**Transcript length in bp**	**GC**
M. truncatula	circRNA	0-20%	(0, 87]	0.44 ± 0.08
	circRNA	20-40%	(87, 135]	0.45 ± 0.07
	circRNA	40-60%	(135, 264]	0.44 ± 0.06
	circRNA	60-80%	(264, 474]	0.42 ± 0.03
	circRNA	80-100%	(474, 1.41e+04]	0.40 ± 0.04
	lincRNA	0-20%	(0, 368]	0.36 ± 0.05
	lincRNA	20-40%	(368, 481]	0.35 ± 0.04
	lincRNA	40-60%	(481, 618]	0.35 ± 0.04
	lincRNA	60-80%	(618, 881]	0.35 ± 0.04
	lincRNA	80-100%	(881, 1.01e+04]	0.34 ± 0.03
	NAT	0-20%	(0, 415]	0.38 ± 0.05
	NAT	20-40%	(415, 576]	0.37 ± 0.04
	NAT	40-60%	(576, 804]	0.37 ± 0.04
	NAT	60-80%	(804, 1.25e+03]	0.37 ± 0.04
	NAT	80-100%	(1.25e+03, 1.46e+04]	0.36 ± 0.03
	mRNA	0-20%	(0, 413]	0.40 ± 0.05
	mRNA	20-40%	(413, 962]	0.39 ± 0.05
	mRNA	40-60%	(962, 1.54e+03]	0.39 ± 0.04
	mRNA	60-80%	(1.54e+03, 2.32e+03]	0.39 ± 0.03
	mRNA	80-100%	(2.32e+03, 1.71e+04]	0.39 ± 0.03
B. distachyon	circRNA	0-20%	(0, 105]	0.52 ± 0.11
	circRNA	20-40%	(105, 185]	0.54 ± 0.11
	circRNA	40-60%	(185, 367]	0.51 ± 0.10
	circRNA	60-80%	(367, 676]	0.47 ± 0.08
	circRNA	80-100%	(676, 1.22e+04]	0.45 ± 0.06
	lincRNA	0-20%	(0, 425]	0.46 ± 0.09
	lincRNA	20-40%	(425, 554]	0.46 ± 0.08
	lincRNA	40-60%	(554, 709]	0.46 ± 0.08
	lincRNA	60-80%	(709, 1.02e+03]	0.46 ± 0.08
	lincRNA	80-100%	(1.02e+03, 5.89e+03]	0.44 ± 0.06
	NAT	0-20%	(0, 416]	0.51 ± 0.11
	NAT	20-40%	(416, 558]	0.51 ± 0.10
	NAT	40-60%	(558, 729]	0.52 ± 0.09
	NAT	60-80%	(729, 1.11e+03]	0.50 ± 0.08
	NAT	80-100%	(1.11e+03, 1.04e+04]	0.46 ± 0.07
	mRNA	0-20%	(0, 1.08e+03]	0.53 ± 0.09
	mRNA	20-40%	(1.08e+03, 1.62e+03]	0.53 ± 0.06
	mRNA	40-60%	(1.62e+03, 2.15e+03]	0.51 ± 0.06
	mRNA	60-80%	(2.15e+03, 2.96e+03]	0.50 ± 0.06
	mRNA	80-100%	(2.96e+03, 2.12e+04]	0.47 ± 0.05

Table 12.4 | GC content and quintiles of transcript lengths in coding and non-coding RNA species. For each species and each RNA species (except intronic NATs), we group the corresponding transcripts into quintiles based on their transcript length in bp and calculate the relative mean GC content ± standard deviation. The transcript length for each quanitle is given as an interval representing the minimal and maximal transcript length.

Eidesstattliche Erklärung / Declaration under Oath

Ich erkläre an Eides statt, dass ich die Arbeit selbstständig und ohne fremde Hilfe verfasst, keine anderen als die von mir angegebenen Quellen und Hilfsmittel benutzt und die den benutzten Werken wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

I declare under penalty of perjury that this thesis is my own work entirely and has been written without any help from other people. I used only the sources mentioned and included all the citations correctly both in word or content.

Halle (Saale), 28.7.2021 Alexander Gabel

Curriculum Vitae

Personal information

Name: Alexander Gabel
Date of birth: 11.06.1988
Nationality: German

Education

09/2013	Master of Science in Bioinformatics Martin Luther University Halle-Wittenberg “Development of a simulated annealing algorithm for uncovering the phylotranscriptomic hourglass pattern” Advisors: Prof. Dr. Ivo Grosse & Prof. Dr. Marcel Quint
10/2011	Bachelor of Science in Bioinformatics Martin Luther University Halle-Wittenberg “A phylostratigraphic analysis of the *Arabidopsis thaliana* genome” Advisors: Prof. Dr. Ivo Grosse & Prof. Dr. Marcel Quint
07/2007	Abitur Dr. Carl-Hermann-Gymnasium, Schönebeck (Elbe)

Experience

since 2013	Research And Teaching Assistant Bioinformatics group of Prof. Dr. Ivo Große Martin Luther University Halle-Wittenberg, Germany
2015 – 2018	Visiting scientist Research group of Prof. Dr. Elliot M. Meyerowitz Sainsbury Laboratory, Cambridge University, UK: Analysis of high-throughput RNA sequencing data with special focus on the prediction and evolution of protein-coding genes, splice variants, long non-coding RNAs and circular RNAs of developmental transcriptomes in flowering plants
2012 – 2013	Research Intern Medicinal Chemistry group of Prof. Dr. Wolfgang Sippl Martin Luther University Halle-Wittenberg, Germany Implementation of an automated pipeline for benchmarking protein ligand docking algorithms
2011 – 2013	Research Intern Research group of Prof. Dr. Marcel Quint Department of Molecular Signal Processing Leibniz Institute of Plant Biochemistry, Halle (Saale), Germany Development of information theoretical and phylogenomic approaches to study developmental processes
2011	Research Intern Molecular Ecology group of Dr. Walter Durka Helmholtz Centre for Environmental Research, Halle (Saale), Germany Implementation of an R package to analyse AFLP data sets for population genetics

Awards

2014	FERCHAU Engineering GmbH sponsorship award
2013	SKWP Research Award for for excellent scientific achievements by young scientists and scholars
2012	Georg-Cantor award for exceptional research in Mathematics and Computer Science

Publications

2020	Madrigal, P., **Gabel**, A., Villacampa, A., Manzano, A., Deane, C. S., Bezdan, D., Carnero-Diaz, E., Medina, F. J., Hardiman, G., Grosse, I., Szewczyk, N., Weging, S., Giacomello, S., Harridge, S. D., Morris-Paterson, T., Cahill, T., da Silveira, W. A. & Herranz, R. Revamping Space-omics in Europe. *Cell Systems,* 10–11 (2020).
	Mao, Y., **Gabel**, A., Nakel, T., Viehöver, P., Baum, T., Tekleyohans, D. G., Vo, D., Grosse, I. & Groß-Hardt, R. Selective egg cell polyspermy bypasses the triploid block. *eLife* **9** (2020).
2019	Schutkowski, A., König, B., Kluge, H., Hirche, F., Henze, A., Schwerdtle, T., Lorkowski, S., Dawczynski, C., **Gabel**, A., Große, I. & Stangl, G. I. Metabolic footprint and intestinal microbial changes in response to dietary proteins in a pig model. *The Journal of Nutritional Biochemistry* **67**, 149–160 (2019).
2018	Melnyk, C. W., **Gabel**, A., Hardcastle, T. J., Robinson, S., Miyashima, S., Grosse, I. & Meyerowitz, E. M. Transcriptome dynamics at Arabidopsis graft junctions reveal an intertissue recognition mechanism that activates vascular regeneration. *Proceedings of the National Academy of Sciences of the United States of America* **115**, E2447–E2456 (2018).
	Drost, H. G., **Gabel**, A., Liu, J., Quint, M. & Grosse, I. MyTAI: Evolutionary transcriptomics with R. *Bioinformatics* **34**, 1589–1590 (2018).

	Wohlrab, J., **Gabel**, A., Wolfram, M., Grosse, I., Neubert, R. H. & Steinbach, S. C. Age-and diabetes-related changes in the free fatty acid composition of the human stratum corneum. *Skin Pharmacology and Physiology* **31**, 283–291 (2018).
2017	Steinbach, S. C., Triani, R., Bennedsen, L., **Gabel**, A., Haeusler, O., Wohlrab, J. & Neubert, R. H. H. Retarder action of isosorbide in a microemulsion for a targeted delivery of ceramide NP into the stratum corneum. *Die Pharmazie* **72**, 440–446 (2017).
2016	Drost, H.-G., Bellstädt, J., Ó'Maoiléidigh, D. S., Silva, A. T., **Gabel**, A., Weinholdt, C., Ryan, P. T., Dekkers, B. J., Bentsink, L., Hilhorst, H. W. M., Ligterink, W., Wellmer, F., Grosse, I. & Quint, M. Post-embryonic Hourglass Patterns Mark Ontogenetic Transitions in Plant Development. *Molecular Biology and Evolution* **33**, 1158–1163 (2016).
2015	Drost, H.-G., **Gabel**, A., Grosse, I. & Quint, M. Evidence for Active Maintenance of Phylotranscriptomic Hourglass Patterns in Animal and Plant Embryogenesis. *Molecular Biology and Evolution* **32**, 1221–1231 (2015).
	Ryan, P. T., Ó'Maoiléidigh, D. S., Drost, H.-G., Kwaśniewska, K., **Gabel**, A., Grosse, I., Graciet, E., Quint, M. & Wellmer, F. Patterns of gene expression during Arabidopsis flower development from the time of initiation to maturation. *BMC Genomics* **16**, 488 (2015).
2012	Quint, M., Drost, H.-G., **Gabel**, A., Ullrich, K. K., Bönn, M. & Grosse, I. A transcriptomic hourglass in plant embryogenesis. *Nature* **490**, 98–101 (2012).
	Gabel, A., Quint, M. & Grosse, I. A phylostratigraphic analysis of the *Arabidopsis thaliana* genome. In *Lecture Notes in Informatics (LNI) - Seminars* **11** (2012), 179–182.

Halle (Saale), 28.7.2021