A bioinformatic study on transcriptome conservation patterns in animal and plant development

Dissertation
zur Erlangung des
Doktorgrades der Naturwissenschaften (Dr. rer. nat.)

der

Naturwissenschaftlichen Fakultät III
Agrar- und Ernährungswissenschaften,
Geowissenschaften und Informatik

der Martin-Luther-Universität Halle-Wittenberg

vorgelegt von

M. Sc. **Hajk-Georg Drost**

Geb. am 04.12.1986 in Halle (Saale)

Gutachter:
Prof. Dr. Ivo Große
Prof. Dr. Marcel Quint
Prof. Dr. Günter Theißen

Datum der Verteidigung: 27.09.2016

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.

28.09.2016

Datum / Date

Unterschrift des Antragstellers / *Signature of the applicant*

Abstract

The developmental hourglass concept aims to model a historic phenomenon in biological research. It depicts the morphological observation that animal embryos converge to a common form during mid embryogenesis. This period of morphological conservation between animal embryos was named *phylotypic period* to describe the phenomenon that in these stages, animals of different species appear to be similar morphologically. Recent studies could demonstrate that the transcriptome conservation of animal embryos also follows an hourglass pattern, mirroring the observed morphological pattern. Although plants do not exhibit a morphological hourglass pattern during embryogenesis, we recently reported the existence of a transcriptomic hourglass pattern for the model plant *Arabidopsis thaliana*. To investigate the commonalities between the transcriptomic hourglass patterns in animals and plants in this thesis, I first designed a statistical framework to assess the significance of transcriptome conservation patterns. In a next step, I implemented software tools to answer the question whether or not the currently favoured hypothesis in animals postulating that organogenesis and body plan formation are the major processes that generate the developmental hourglass pattern in animal embryogenesis is sufficient enough to explain the phenomenon observed in plants. Finally, I addressed the question whether or not this organogenesis centred hypothesis is broad enough to explain the independent emergence of the hourglass pattern in both the animal and plant kingdom. For this purpose, I investigated whether or not transcriptomic hourglass patterns are actively maintained in extant species and whether plant hourglass patterns are also present postembryonically. As a result, I found that indeed transcriptomic hourglass patterns are actively maintained in extant species and that the two main phase changes during the life cycle of *Arabidopsis*, from embryonic to vegetative and from vegetative to reproductive development, are also associated with transcriptomic hourglass patterns. In contrast, a process dominated by organ formation, flower development, is not. These results suggest that transcriptomic hourglass patterns in plants are decoupled from organogenesis and body plan establishment and mark general transitions during development. Together, the findings presented in this thesis challenge the previous causal explanation that links the emergence of developmental hourglass patterns to organogenesis and body plan establishment. My co-authors and I argue, that a more fundamental process might shape developmental hourglass patterns and hypothesize that these fundamental processes explain both: the independent emergence of hourglass patterns in animals and plants and their active maintenance in extant species. We refer to these fundamental marks as *organizational checkpoints* and argue that these checkpoints are present in many biological processes and across kingdoms of life.

Acknowledgments

I am very grateful to my advisors Marcel Quint and Ivo Grosse, who guided me through my Bachelor, my Master, and finally through my PhD studies. On this long path, I value most that they always gave me the freedom and support to perform basic research by applying informatic and bioinformatic methods to address fundamental biological questions. Our discussions were often very passionate indicating that we indeed performed cutting edge research. I am grateful for this time and I learned a lot from them.

I would also like to thank Jerzy Paszkowski for giving me the freedom, support, and trust to finish my thesis in Cambridge while already working as Research Associate in his inspiring team at the Sainsbury Laboratory, University of Cambridge.

I would like to thank Bas Dekkers, Leonie Bentsink, Henk Hilhorst, Wilco Ligterink, Pat Ryan, Diarmuid Ó'Maoiléidigh, and Frank Wellmer for the great collaboration on germination and flower development that lead to the publications Dekkers *et al.* (2013), Ryan *et al.* (2015), and Drost *et al.* (2016) and Alexander Gabel for contributing to the publication Drost *et al.* (2015).

Most of my research was mainly conducted at the computer science department of the Martin Luther University Halle–Wittenberg. In addition, I spent three months as visiting researcher in the group of Elliot Meyerowitz at the Sainsbury Laboratory to collaborate with Christoph Schuster on lncRNA and splice variant expression in different tissues of different plant species. I would like to thank both institutions the MLU and SLCU for providing me a scientifically and intellectually stimulating environment and I am grateful to Elliot and Christoph for the opportunity to work on this ambitious and fascinating project. I would like to thank the *SKW Stickstoffwerke Piesteritz GmbH* for awarding me with the *SKWP research award* in 2013 which partially funded my research and research collaborations.

Finally, I would like to express the deepest gratitude to my family who provided me with the intellectual and emotional foundation to start an academic career. I would especially like to thank Claudia, who inspires me every day. Without her my life would not be a good life.

Peer-reviewed publications

This thesis is a cumulative thesis, indicating that it accumulates research papers that have previously been published in peer-reviewed international journals and combines them to a thesis. The following list summarizes these publications:

- **HG Drost**, J Bellstaedt, DS Ó'Maoiléidigh, AT Silva, A Gabel, C Weinholdt, PT Ryan, BJ Dekkers, L Bentsink, HW Hilhorst, W Ligterink, F Wellmer, I Grosse, M Quint. 2016. Post-embryonic hourglass patterns mark ontogenetic transitions in plant development. *Mol. Biol. Evol.* 33 (5): 1158-1163. *doi:10.1093/molbev/msw039* (journal cover)
- **HG Drost**, A Gabel, I Grosse, M Quint. 2015. Evidence for Active Maintenance of Phylotranscriptomic Hourglass Patterns in Animal and Plant Embryogenesis. *Mol. Biol. Evol.* 32 (5): 1221-1231. *doi:10.1093/molbev/msv012*
- PT Ryan, DS Ó'Maoiléidigh, **HG Drost**, K Kwasniewska, A Gabel, I Grosse, E Graciet, M Quint, F Wellmer . 2015. Patterns of gene expression during *Arabidopsis* flower development from the time of initiation to maturation. *BMC Genomics* 16:488. *doi:10.1186/s12864-015-1699-6*
- BJW Dekkers, S Pearce, RP van Bolderen-Veldkamp, A Marshall, P Widera, J Gilbert, **HG Drost**, GW Bassel, K Müller, JR King, AT Wood, I Grosse, M Quint, N Krasnogor, G Leubner-Metzger, MJ Holdsworth, L Bentsink. 2013. Transcriptional dynamics of two seed compartments with opposing roles in *Arabidopsis* seed germination. *Plant Physiology* 163 (1): 205-215. *doi:10.1104/pp.113.223511*

I hereby declare that the copyright of the content of the papers Drost *et al.* 2015 and Drost *et al.* 2016 is by Oxford University Press. These papers are available at:

- `http://mbe.oxfordjournals.org/content/32/5/1221`
- `http://mbe.oxfordjournals.org/content/33/5/1158`

I hereby declare that the copyright of the content of the paper Ryan *et al.* 2015 is by BioMed Central and is available at:

- `http://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-015-1699-6`

I hereby declare that the copyright of the content of the paper Dekkers *et al.* 2013 is by American Society of Plant Biologists and is available at:

- `http://www.plantphysiol.org/content/163/1/205`

The following paper entitled *Capturing Evolutionary Signatures in Transcriptomes with myTAI* is currently submitted to the journal *Molecular Biology and Evolution* and published on bioRxiv but has not been peer-reviewed yet. However, its content summarizes the main applications and functionalities of *myTAI* and *orthologr*. Due to this fact, I decided to include this paper as appendix to this thesis.

HG Drost, A Gabel, T Domazet-Lošo, M Quint, I Grosse. 2016. Capturing Evolutionary Signatures in Transcriptomes with myTAI. (submitted to *Mol. Biol. Evol.* and published on *bioRxiv*)
This paper is available at:

- `http://biorxiv.org/content/early/2016/05/03/051565.abstract`

Contents

1 Introduction 8

2 Objectives and Outline of this thesis 10

3 Methods 11
- 3.1 Gene Age Inference 13
- 3.2 Dollo Parsimony - Phylostratigraphy 15
- 3.3 Divergence Stratigraphy - dNdS Estimation 17
- 3.4 Transcriptome Indices 18
- 3.5 Software Tools to Perform Phylotranscriptomic Analyses 26
  - 3.5.1 R package myTAI 27
  - 3.5.2 R package orthologr 29

4 Results and Discussion 31
- 4.1 The Developmental Hourglass Model 32
- 4.2 From Morphology to Genetics: The Phylotypic Period on the Molecular Level 38
- 4.3 Evidence for Active Maintenance of Phylotranscriptomic Hourglass Patterns 42
- 4.4 Post-embryonic Hourglass Patterns Mark Ontogenetic Transitions in Plant Development 45

5 Conclusions and Outlook 49

6 Paper 1: Evidence for Active Maintenance of Phylotranscriptomic Hourglass Patterns in Animal and Plant Embryogenesis 61

7 Paper 2: Post-embryonic Hourglass Patterns Mark Ontogenetic Transitions in Plant Development 73

8 Paper 3: Patterns of gene expression during Arabidopsis flower development from the time of initiation to maturation 80

9 Paper 4: Transcriptional Dynamics of Two Seed Compartments with Opposing Roles in Arabidopsis Seed Germination 93

10 Appendix 105
- 10.1 Submitted Manuscript: Capturing Evolutionary Signatures in Transcriptomes with myTAI 105
- 10.2 A transcriptomic hourglass in plant embryogenesis 145
- 10.3 Abstract 145
- 10.4 Documentation: Introduction to myTAI 145
- 10.5 Documentation: Introduction to orthologr 229
- 10.6 Suppl. Material: Evidence for Active Maintenance of Phylotranscriptomic Hourglass Patterns in Animal and Plant Embryogenesis 271

10.7 Suppl. Material: Post-embryonic Hourglass Patterns Mark Ontogenetic Transitions in Plant Development 278
10.8 Suppl. Material: Transcriptional Dynamics of Two Seed Compartments with Opposing Roles in Arabidopsis Seed Germination 289

1 Introduction

Understanding the genesis, evolution, and variability of complex organismal forms is among the most fundamental objectives of biological research. Key questions about the origination, maintenance, and evolution of complex life on earth are now approachable through the recent advancements in molecular biology and their intersection with information sciences.

Today, it is well studied that embryo development (= embryogenesis) is the key process to establish complex multicellular life by transitioning from a single celled zygote to a mature multicellular organism (= embryo). Hence, embryogenesis provides the developmental framework to establish the organismal organization (= body plan) of a multicellular organism by passing through a defined series of developmental stages that are governed by genetic programs of gene expression [1]. The concept of characterizing common traits among phylogenetically related species to classify the organizational form of multicellular organisms during comparable stages of embryo development is today referred to as body plan concept. The aim of this concept is to provide a scientific framework to study the origin, evolution, and variability of organismal forms by comparing the function of homologous traits of extant species [2–4].

Due to the vast diversity and variability of organismal forms on earth a scientific comparability can only be achieved by reducing all extent body forms to a common **basic** body plan. This reductionist view of a basic body plan allows us to quantify the variability and diversification of traits shared among phylogenetically related species by comparing the physical properties such as weight, size, and location of basic body plan features [5, 6]. Recent studies suggest that these commonalities are the result of developmental constraints (= a limitation on phenotypic variability caused by the structure, character, composition, or dynamics of the developmental system [7]) limiting the potential combinatorial variability of phenotypes [6–9]. These developmental constraints are proposed to channel the evolutionary conservation of specific body plans resulting in the limited diversity of extant forms in comparison with the combinatorial diversity of potential body plans when assuming an absence of such developmental constraints (limited diversification) [6].

Together, the basic body plan defines the common anatomical features such as head, arms, legs, and other major organ systems shared by organisms belonging to the same species, phyla, or kingdom [5]. These anatomical features allow us to study the evolutionary history of developmental processes and therefore, contributes to the understanding of how complex organismal forms evolve and diversify [5, 6, 10].

The central scientific question arising from the body plan concept however, is *why* and *to what extent* the basic body plan is conserved within and between phyla [6]. Historically, the body plan concept arose from animal studies performed more than 200 years ago [11] describing a fascinating morphological phenomenon observed during mid embryogenesis [10, 12]. In particular, it has been observed that during the organogenic period of mid embryogenesis animal embryos of different species within the same phylum converge to a form of high morphological resemblance when compared with early and late embryogenesis. Due to the high morphological resemblance of anatomical features shared between different vertebrate taxa, this developmental window has been termed *phylotypic stage* [13] or *phylotypic period* [14, 15] and the morphological pattern of dissimilarity - similarity - dissimilarity between animal embryos has been termed *developmental hourglass phenomenon* [10, 12, 16].

Although the existence of a phylotypic stage or period has been controversially debated, and therefore, the existence of a developmental hourglass phenomenon has recently been questioned [14, 15, 17–20] the concept of the developmental hourglass has largely been confirmed on the molecular level [6]. Several studies demonstrated that the degree of sequence conservation, the phylogenetic age of transcriptomes, gene regulatory system conservation or the similarity of gene expression profiles maximize during the phylotypic period [21–44], which is in agreement with a potentially causative association between the phylotypic period and body plan establishment in animals [4].

In 2010, the first transcriptome wide study was performed to confirm the morphological pattern of dissimilarity - similarity - dissimilarity on the molecular level. The data provided support previous studies suggesting the existence of a correlation between phylogeny and ontogeny [26]. In this study, Domazet-Lošo and Tautz [26] concluded that the phylotypic period can be defined as the *ontogenetic progression during which the oldest gene set is expressed, either because this is the phase with the lowest opportunity for lineage-specific adaptations, or because it is internally so constrained that newly evolved genes cannot become integrated* [26].

In this regard, the evolutionary age of genes reflects the phylogenetic component when investigating the evolutionary constraints acting on the transcriptomes of animal embryos. This clear association between gene age and gene expression enables us to capture evolutionary signatures in developmental transcriptomes and furthermore allows us to quantify the sets of genes that are more likely to be negatively selected for constraining organismal diversification.

In 2012, I applied this powerful method to plant embryogenesis by quantifying transcriptome conservation throughout *Arabidopsis thaliana* embryo development and were able to observe an analogous phenomenon of transcriptome conservation (dissimilar - similar - dissimilar) as previously reported in animals [29, 45]. This finding was particularly surprising, because the morphological diversity during angiosperm embryogenesis is negligible due to the establishment of meristems instead of a precise plant body plan [30]. Hence, morphological differences in plants are only established during postembryonic development. The fact however, that both plant and animal embryogenesis follow a molecular hourglass pattern of transcriptome dissimilarity - similarity - dissimilarity raises important questions about the association between transcriptome conservation and body plan establishment (organogenesis) in general.

In our 2012 study, we concluded that the absence of a hourglass pattern based on morphological features in plants suggests that both morphological and molecular patterns might be uncoupled and that the presence of a developmental hourglass phenomenon in animals and plants indicates convergent evolution of the molecular hourglass and a conserved logic of embryogenesis across kingdoms [45].

Our hypothesis, postulating that the morphological and molecular hourglass patterns might be uncoupled, was supported by the study of Cheng *et al.* in 2015 who reported a molecular hourglass pattern in fungi development [46]. This dissociation between the morphological and molecular pattern raises fundamental questions about the findings reported in the animals kingdom aiming to correlate phylogeny and ontogeny via body plan establishment during the phylotypic period. Cheng *et al.* conclude in their study that the presence of a universal molecular hourglass pattern across kingdoms (animals, plants, and fungi) might reflect a mutual strategy for eukaryotes to incorporate evolutionary innovations [46].

Motivated by these findings, the main objectives of my thesis are to first develop a solid statistical framework to assess the statistical significance of observed molecular hourglass patterns (enabling the comparability across kingdoms and studies), second to test whether or not the molecular hourglass patterns in animals and plants are actively maintained and therefore, experimentally assessable in extant species, and third to investigate whether or not postembryonic developmental processes in plants also follow molecular hourglass patterns. For this purpose I developed open source software tools that will allow me and a broad range of researchers to automate and reproduce the quantification of transcriptome conservation.

The following chapters will give the reader a detailed introduction to the biological questions that I aim to address in this thesis by developing and applying methods from computer science and bioinformatics.

2 Objectives and Outline of this thesis

The introduced studies on the molecular hourglass phenomenon suggest that developmental transcriptomes contain evolutionary information which can be captured and quantified using transcriptome indices.

The scientific question I aimed to answer was whether or not there are commonalities between the transcriptomic hourglass patterns in animals and plants.

In order to address this question, my first objective was to build statistical tests to quantify the significance of differential transcriptome conservation between stages by developing a customized statistical framework to quantify and assess the significance of any transcriptome conservation pattern of interest (Paper 1).

A limitation was that no expert group designed and implemented software tools for computing transcriptome indices to apply them in a virtuous and reproducible manner. Hence, although extremely powerful, the application of phylotranscriptomics as a methodology for non-experts was limited by the lack of available and user-friendly software tools.

Hence, the second objective of my thesis was to fill this gap by implementing the R packages *myTAI* and *orthologr* (Paper 1 and Appendix) which allowed me to compute transcriptome indices in a virtuous and reproducible manner. The aim of this part of the thesis was therefore to provide the ability to perform phylotranscriptomic analyses primarily to biologists who are not bioinformatics experts.

The third objective was to apply the implemented software tools to biological data sets that had been generated to a) allow a comparison of the developmental hourglass model between animals and plants (Paper 1), and to b) answer general questions regarding transferrability of the hourglass concept to the plant kingdom (Paper 1 - 4).

Taken together, in this thesis I aim to answer fundamental questions regarding one of the historical concepts of developmental biology by developing and applying bioinformatic tools that allow to address the developmental hourglass concept on a transcriptomic level.

3 Methods

Phylotranscriptomics denotes the methodology of quantifying gene age and gene conservation to then combine this information with the expression of these genes for the computation of the average transcriptome conservation in biological processes by applying transcriptome indices.

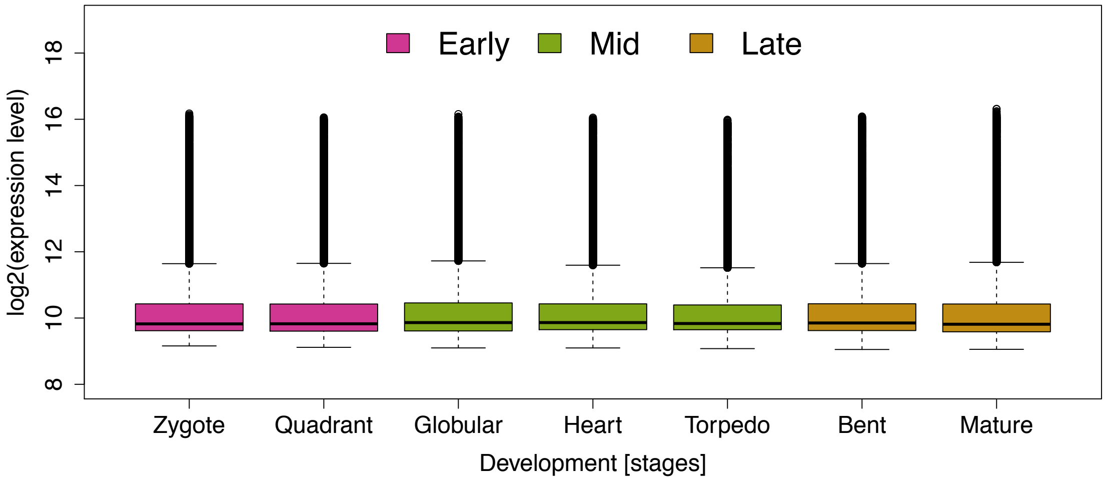


Figure 3.0.1: Gene expression distributions (= developmental transcriptome) throughout seven stages of *A. thaliana* embryo development. Embryo development is devided into three phases: early embryogenesis (purple), mid embryogenesis (green), and late embryogenesis (brown). This boxplot illustrates that the overall distributions of log2 expression levels (y-axis) hardly differ between developmental stages (x-axis) although the difference on the global scale is statistically significant (Kruskal-Wallis Rank Sum Test: $p < 2e\text{-}16$). Hence, a clear visual pattern of gene expression differences between early, mid, and late embryogenesis on the global scale can not be inferred.

The rational for performing the phylotranscriptomic method is to classify a transcriptome (Fig. 3.0.1) into different categories of genes sharing similar evolutionary origins (detectable homologs) or genes being under similar selective pressures and to study the overall expression patterns of these classified genes throughout the biological process of interest (Fig. 3.0.2).

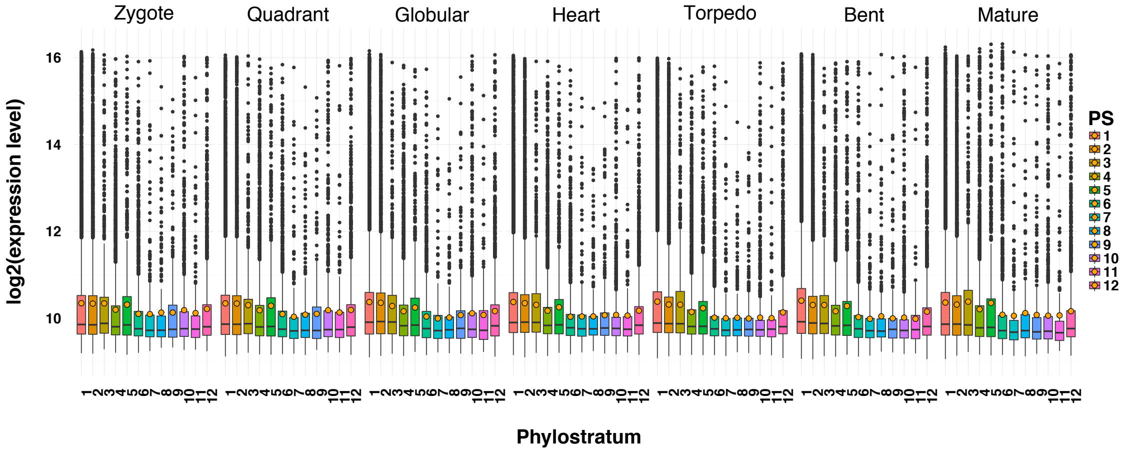


Figure 3.0.2: Gene expression distributions (= developmental transcriptome) throughout seven stages of *A. thaliana* embryo development classified into phylostrata. Each box represents the developmental stage during *A. thaliana* embryogenesis, the y-axis denotes the log2 expression levels of genes that fall into the corresponding phylostratum (age category) shown on the x-axis. Hence, each boxplot represents the gene expression distribution of genes that are classified into the corresponding phylostratum (PS) during a specific developmental stage. The gene age distribution of *A. thaliana* genes ranges from PS1 to PS12 where PS1 represents the evolutionarily most distant age category (cellular org.) and PS12 the evolutionary most recent age category (*A. thaliana* specific; see section Dollo Parsimony - Phylostratigraphy and Fig 3.2.4). Yellow dots in the boxplots denote the mean expression level of the corresponding expression distribution. This visualization illustrates that although the global gene expression distributions do not change visually between developmental stages (Fig. 3.0.1), the global gene expression distributions of PS differ between stages of *A. thaliana* embryo development, and thus, allow to study the effect of transcriptome evolution and conservation on embryo development.

Hence, phylotranscriptomics combines four methods: (1) gene age inference and protein substitution rate quantification, (2) gene expression analysis, (3) transcriptome conservation quantification (transcriptome indices) and evaluation, and (4) relative expression level analysis.

The following sections will introduce these methods in detail and will point out the current status and limitations of this methodology.

3.1 Gene Age Inference

Gene Age Inference is a methodology to trace the evolutionary origin and diversification of protein coding genes in the context of detectable homology [26, 47]. This comparative genomics approach provides a powerful method to study the evolution and diversification of morphological and molecular traits and allows researchers to classify protein coding genes into inter-species or intra-species specific categories. Most inter-species proteins for example can be associated with a highly conserved metabolic function (housekeeping) or for phylum specific developmental processes (e.g. Hox genes). Hence, this approach allows us to quantify the conservation of biological process or trait specific origination events. For practical applications however, the lack of a clear and consistent definition of gene age as discussed by Capra *et al.* led to a rise of different tools and concepts for practical gene age inference [47].

Three major approaches have been proposed to quantify the timing of events (which in most cases is equated with gene age determination) and sequence evolution:

- Gain-loss approaches
- Phylogenetic Reconciliation
- Sequence Divergence Models

Figure 3.1.3 summarizes published methods for gene age inference. The most widely used and established methods are based on two major approaches: *Gain-loss approaches* and *Phylogenetic Reconciliation*.

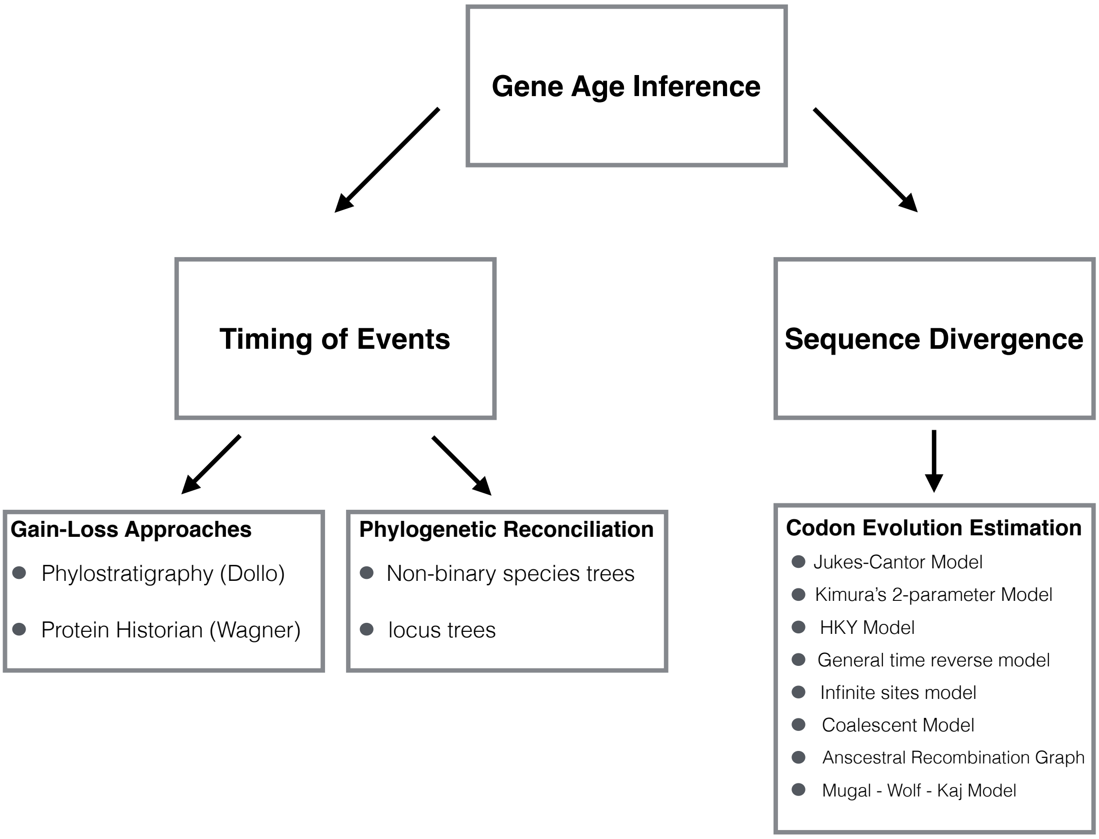


Figure 3.1.3: Most common and established methods aiming to perform gene age inference. The diagram shows two conceptual methods of quantifying gene age: Through the timing of gene origination (detectable homologs) and computing the sequence substitution rate to estimate the divergence and therefore, the age of homolgous sequences [47, 48].

Gain-loss approaches are based on Dollo's law (Phylostratigraphy) and *Wagner Parsimony* (Protein Historian), whereas Phylogenetic Reconciliation approaches include *non-binary species trees* and *locus trees* [26, 47].

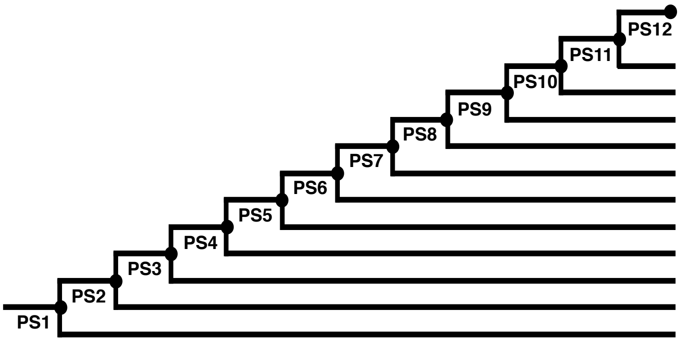
3.2 Dollo Parsimony - Phylostratigraphy

Phylostratigraphy is a computational method to determine the evolutionary origin of protein coding genes based on BLAST homology searches. This sequence homology based method for gene age inference was introduced by Domazet-Lošo *et al.* in 2007 [49].

The process of performing phylostratigraphy can be summarized by the following algorithm:

- Select a taxonomy for a query organism of interest
- Classify annotated genomes into corresponding taxonomic groups (phylogenetic internodes = phylostrata)
- Perform a BLASTp homology search of each protein coding gene of the subject organism against the classified database
- Assign the oldest BLAST hit (in terms of phylogenetic distance) fulfilling the homology detection criteria to the corresponding query gene
- If no homolog can be detected, assign the corresponding gene as species specific

This procedure generates a table storing the gene age assignment in the first column and the corresponding *gene id* of the protein coding gene of the query organism in the second column. The output table is termed *phylostratigraphic map* [45, 49] (see Fig. 3.2.4) and is subsequently joined with the expression data covering the biological process of interest.



Taxonomy	taxid	Phylostratum	#Genes
A. thaliana	3702	PS12	2534
Arabidopsis	3701	PS11	319
Brassicales	3699	PS10	1407
Rosids	71275	PS9	469
Core Eudicotyledons	91827	PS8	322
Eudicotyledons	71240	PS7	332
Magnoliophyta	3398	PS6	1678
Tracheophyta	58023	PS5	909
Embryophyta	3193	PS4	2717
Viridiplantae	33090	PS3	813
Eukaryota	2759	PS2	6611
Cellular Organisms	131567	PS1	9305

Figure 3.2.4: Table summarizing the taxonomy and the corresponding phylostrata of *A. thaliana*. The left side of this figure shows the taxonomic classification of *A. thaliana* from *cellular organisms* up to *A. thaliana*. Each phylogenetic internode is labelled as phylostratum (PS) increasing from PS1 (evolutionary most distant common ancestors; cellular organisms) up to PS12 (*A. thaliana*; species specific). Furthermore, the NCBI taxonomy id and the number of protein coding genes of *A. thaliana* predicted to share detectable homologs with the corresponding PS are shown. The column *#Genes* summarizes the number of genes that share homologs within the same phylostratum.

The interpretation of gene age for each gene can be inferred from the exact definitions of *genomic phylostratigraphy*, *founder gene formation*, *punctuated protein family evolution* and *phylostratum* provided by Domazet-Lošo *et al.* in 2007 [49].

- **Genomic phylostratigraphy:** *a statistical approach for reconstruction of macroevolutionary trends based on the principle of founder gene formation and punctuated emergence of protein families* [49].
- **Founder gene formation:** *first emergence of a gene forming the basis of a new gene lineage or gene family; the origination of founder genes might correlate with functional novelty.* [49]
- **Punctuated protein family evolution:** *a model of genome evolution that assumes that protein families were initiated by founder genes in a scattered manner through evolutionary time.* [49]
- **Phylostratum:** *a set of genes from an organism that coalesce to founder genes having common phylogentic origin.* [49]

Hence, the phylostratigraphic method is based on the assumption that lineage specific genes emerge in a punctuated manner through a process of *de novo* gene birth and quickly evolve to retain an association with a particular biological pathway [49]. This assumption was used by Domazet-Lošo *et al.* to introduce *phylostratigraphy* as a general approach to trace evolutionary innovation using data from genome sequencing projects [49].

On a more abstract level, the gain-loss assumptions implicit to phylostratigraphy match the concept of *Dollo Parsimony* [47]. Capra *et al.* define *Dollo Parsimony* as *a common gain–loss phylogenetic analysis method based on parsimony and the* *assumption that a biological character can only be gained once, although it may experience multiple losses in different lineages* [47].

It is not within the scope of this thesis to discuss all methods in detail, but I encourage the reader to consult the referenced literature to understand the advantages and disadvantages of these gene age inference methods. The advantage of phylostratigraphy is that it allows researchers to quantify gene age in the context of detectable homology and aims to define gene age by its detectable origin. Due to these advantages all gene age values used in this thesis were computed using phylostratigraphy [50].

3.3 Divergence Stratigraphy - dNdS Estimation

In our 2012 study [45], we performed an additional gene age inference method based on sequence divergence estimation (Fig. 3.1.3) to provide a second independent method to phylostratigraphy to verify the transcriptome conservation we observed in plant embryogenesis (= *Divergence Stratigraphy*). In detail, Divergence Stratigraphy is a computational method to determine the degree of selection pressure acting on each protein coding gene of a query organism against a reference organism [40] and hence, quantifies gene age in the context of protein substitution rates. This method differs from phylostratigraphy in that it aims to detect patterns of conservation in closely related species, whereas phylostratigraphy covers homology detection along the tree of life. Divergence Stratigraphy can be summarized by performing the following algorithm:

- Perform orthology inference to determine a set of orthologous genes between closely related species
- Perform a global pairwise alignment of the amino acid sequences of orthologous genes
- Perform a codon alignment of corresponding orthologous genes
- Perform dNdS estimation for the corresponding set of orthologous genes

In the first step orthology inference was performed using the method of *best hit* or *best reciprocal hit* (blastp). Pairwise alignments were performed using MAFFT (L-INS-i option). Codon alignments were computed using *PAL2NAL*, and GESTIMATOR was used for dNdS estimation [45]. This procedure generates a table storing the gene divergence assignment in the first column (dNdS values) and the corresponding *gene id* of the protein coding gene of the query organism in the second column. This table is termed *sequence divergence map* (short divergence map) [45] and is subsequently joined with the expression data covering the biological process of interest.

In general, phylostratigraphy and divergence stratigraphy differ fundamentally due to their underlying biological assumptions. Whereas phylostratigraphy is based on

Dollo's law and therefore aims to detect sequence homologs along the tree of life (detectable homology), divergence stratigraphy is based on the orthology-function-conjecture postulating that orthologs carry out biologically equivalent functions in different organisms [51]. This difference shows that phylostratigraphy aims to quantify the first emergence of a gene independent of its function within the extent species of interest, whereas divergence stratigraphy aims to detect potential functional homologs between closely related species and therefore allows to infer the functional conservation between these species.

In combination, phylostratigraphy and divergence stratigraphy allow researchers to investigate the evolutionary origin of protein coding genes (gene age) as well as their functional conservation between closely related species (gene divergence).

3.4 Transcriptome Indices

Phylostratigraphy and Divergence Stratigraphy generate phylostratigraphic maps and divergence maps for a particular organism of interest. These maps are then joined with the expression data set covering the biological process of interest. To quantify the transcriptome conservation of these joined tables two transcriptome indices have been introduced. These transcriptome measures quantify the transcriptome age and transcriptome divergence within and between biological processes and enable to detect stages or periods of transcriptome conservation in terms of deep evolutionary conservation (Transcriptome Age Index = TAI) [26] or conservation between closely related species (Transcriptome Divergence Index = TDI) [45].

Figure 3.4.5 illustrates that both gene assignments PS and DS are only weakly correlated and therefore, can be used to study transcriptome conservation in deep (TAI) and recent (TDI) evolutionary time scales.

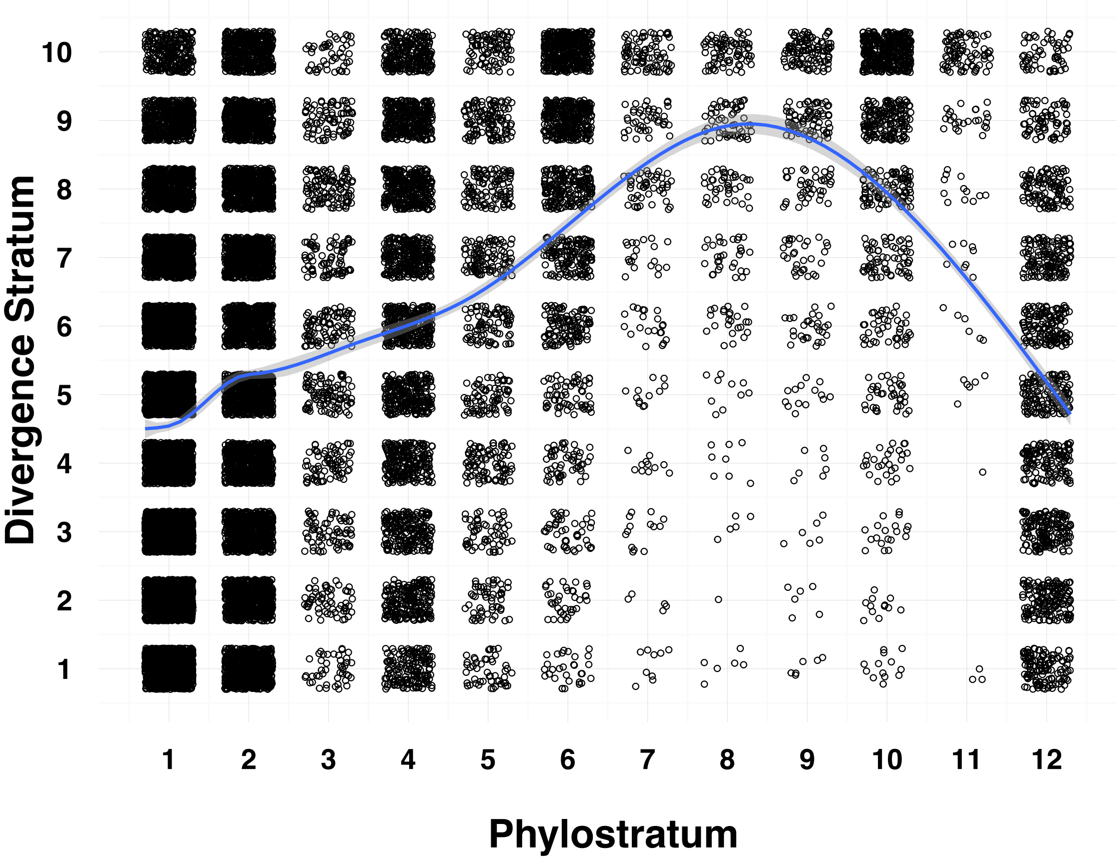


Figure 3.4.5: Linear correlation analysis between phylostrata and divergence strata. This analysis aims to demonstrate that both methods (PS and DS) are linearly independent and therefore, both measures TAI and TDI can be used as independent methods to quantify transcriptome conservation [45]. The x-axis denotes the PS from 1 to 12 and the y-axis denotes the DS ranging from 1 to 10 (deciles of dN/dS values). The result illustrates that PS and DS are weakly correlated (Pearson = 0.2) suggesting a linear independence between both gene age measures [45].

The Transcriptome Age Index is defined as follows:

$$TAI_s = \frac{\sum_{i=1}^{N} e_{is} \cdot ps_i}{\sum_{i=1}^{N} e_{is}} \tag{1}$$

where e_{is} denotes the gene expression value of gene i, in stage s and $ps_i = 1, ..., P$ denotes an integer value representing the phylostratum of gene $i = 1, ...N$, with P denoting the youngest phylostratum and N denoting the total number of protein coding genes. A small ps_i value denotes an old phylostratum and a high ps_i value a younger phylostratum [26].

A higher value of TAI represents the mean expression of a younger transcriptome and a lower value of TAI represents the mean expression of an older transcriptome.

Additionally, TAI values range from 1 to P. As a result, this measure allows us to determine the average evolutionary age of a transcriptome within stages of development or within biological processes in general. Together, the TAI measure quantifies stages or periods of transcriptome conservation in biological processes.

Figure 3.4.6 shows an example visualization of TAI values across seven stages of *A. thaliana* embryo development. The resulting TAI pattern follows an high - low - high pattern of average transcriptome age and illustrates the potential advantage of using TAI to detect stages of transcriptome conservation in biological processes. In this figure, the scope of TAI values is between 2.9 and 3.5 which can be interpreted as transcriptome expression of genes originating (on average) between PS 2.9 - 3.5. Therefore, an evolutionary old set of genes is highly expressed (on average) throughout *A. thaliana* embryogenesis.

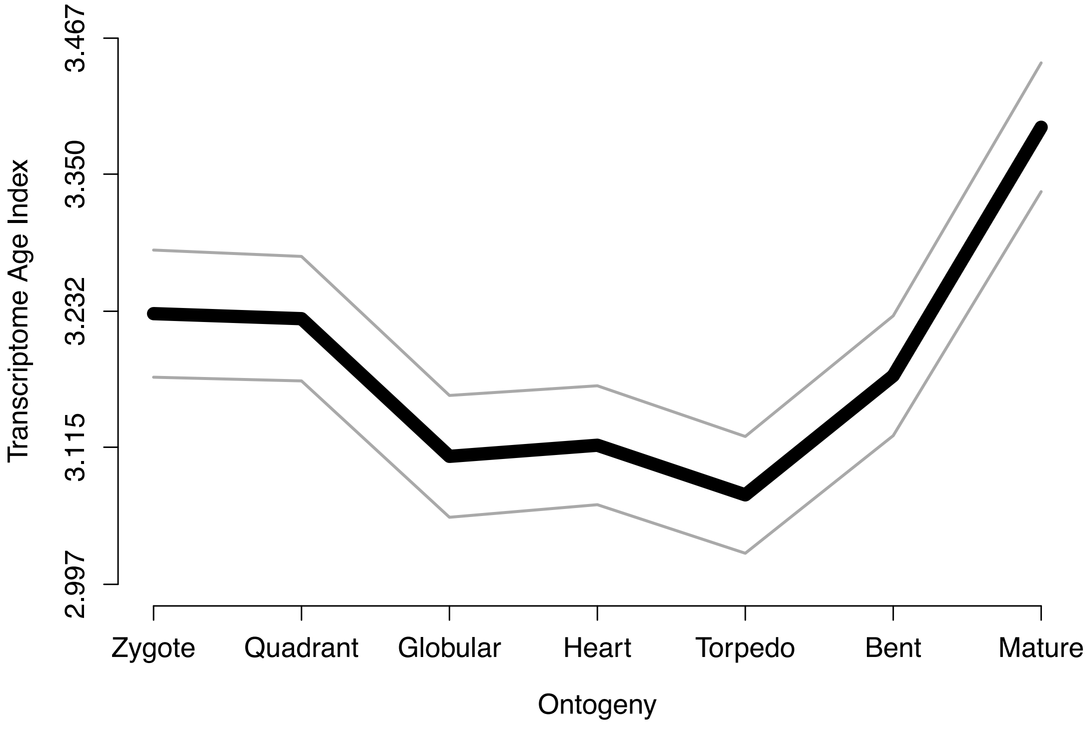


Figure 3.4.6: Transcriptome Age Indices computed for seven stages of *A. thaliana* embryogenesis. The black line connects the transcriptome age indices across seven stages of *A. thaliana* embryogenesis and grey lines represent the standard deviation generated by a permutation test. This example pattern of Transcriptome Age Indices follows a high - low - high pattern (hourglass pattern) of average transcriptome age throughout embryo development [45]. The gray lines represent the standard deviation estimated by permutation analysis.

To assess the statistical significance of observed transcriptome conservation patterns my co-authors and I developed three permutation tests: the *flat line test*, the *reductive hourglass test*, and the *reductive early conservation test* (Paper 1).

The flat line test is defined as follows:

The flat line test [45] is a permutation test based on the variance V of the TAI values of a given TAI profile as test statistic. For any pattern different from a flat horizontal line, V should be high. In order to determine the statistical significance of an observed variance V, we perform the following permutation test. We randomly permute the PS values of the original data set 10,000 times, compute the variance V from each of the 10,000 permuted data set s, approximate the histogram of the 10,000 variances V by a Gamma distribution, and report the probability of exceeding the observed variance V as P-value of the flat line test [40]. The distribution is chosen by applying Cullen and Frey graphs as probabilistic exposure assessment techniques [52] to fit the most reasonable distribution to the corresponding permutation matrices. The parameters of the Gamma distribution are estimated using moment matching estimation [53]. The flat line test can be applied to TDI profiles in exactly the same manner (Paper 1).

The reductive hourglass test is defined as follows:

The reductive hourglass test is a permutation test based on the following test statistic. First, the set of developmental stages is partitioned into three modules, early, mid, and late based on prior biological knowledge. Second, the mean TAI value is computed for each of the three modules, and are denoted by T_{early}, T_{mid}, and T_{late}. Third, we compute the two differences D1 = $T_{early} - T_{mid}$ and D2 = T_{late} - T_{mid}. Fourth, the minimum D_{min} of D1 and D2 are computed as final test statistic of the reductive hourglass test. For a typical hourglass pattern, T_{early} should be high, T_{mid} should be low, and T_{late} should be high, so both differences D1 and D2 should be positive, so the minimum difference D_{min} should be positive, too [40].

In order to determine the statistical significance of an observed minimum difference D_{min}, the following permutation test was performed. We randomly permute the PS values of the original data set 10,000 times, compute the minimum difference D_{min} from each of the 10,000 permuted data sets, approximate the histogram of the 10,000 minimum differences D_{min} by a Gaussian distribution, and report the probability of exceeding the observed minimum difference D_{min} as P-value of the reductive hourglass test [40]. The distribution is chosen by applying Cullen and Frey graphs as probabilistic exposure assessment techniques [52] to fit the most reasonable distribution to the corresponding permutation matrices. The parameters of the Gaussian distribution are estimated using moment matching estimation [53] and the goodness-of-fit is quantified by applying a Lilliefors (Kolmogorov-Smirnov) test [54]. The flat line test can be applied to TDI profiles in exactly the same manner (Paper 1).

The reductive early conservation test is defined as follows:

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 the two differences D1 $= T_{mid} - T_{early}$ and D2 $= T_{late} - T_{early}$. For a typical early conservation pattern, T_{early} should be low, and T_{mid} and T_{late} should be high, so both differences D1 and D2 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 [40]. The parameters of the Gaussian distribution are estimated using moment matching estimation [53] and the goodness-of-fit is quantified by applying a Lilliefors (Kolmogorov-Smirnov) test [54]. The flat line test can be applied to TDI profiles in exactly the same manner (Paper 1).

However, the following disadvantages of using the TAI have been noted [20, 29]:

- TAI is only defined for absolute gene expression levels
- TAI patterns are not always robust against data transformation such as *log* or *sqrt*
- TAI is biased by outlier age assignments or outlier weights (gene expression values)
- TAI only captures the top 2 - 10 % of highly expressed genes

These disadvantages illustrate that the outcome of any TAI analysis must be carefully interpreted before drawing general conclusions on the conservation of non-highly expressed genes [20] and were systematically investigated by me earlier [29].

The Transcriptome Divergence Index (TDI) is defined as follows:

$$TDI_s = \frac{\sum_{i=1}^{N} e_{is} \cdot \frac{dN_i}{dS_i}}{\sum_{i=1}^{N} e_{is}} \tag{2}$$

where e_{is} denotes the gene expression value of gene i, in stage s and $\frac{dN_i}{dS_i} = 0, ..., D$ denotes a continuous value representing the dNdS value of gene $i = 1, ...N$, with D denoting the highest dNdS value and N denoting the total number of protein coding genes. A small $\frac{dN_i}{dS_i}$ value represents a conserved genes and a high $\frac{dN_i}{dS_i}$ value represents a divergent gene [45].

In contrast, the TDI aims to quantify the transcriptome divergence and, therefore, allows to infer potential functional conservation of genes between closely related species [45]. Figure 3.4.7 shows an example visualization of TDI values across seven stages of *A. thaliana* embryo development. The resulting TDI pattern follows a high

- low - high pattern of average transcriptome divergence and illustrates the potential of using TDI to detect stages of functional conservation in biological processes. In this figure, the scope of TDI values is between 0.19 and 0.22 which can be interpreted as transcriptome expression of genes being under strong negative selection quantified in dN/dS (on average) between PS 0.19 - 0.22. Therefore, a highly negatively selected set of genes is highly expressed (on average) throughout *A. thaliana* embryogenesis in comparison with *A. lyrata*.

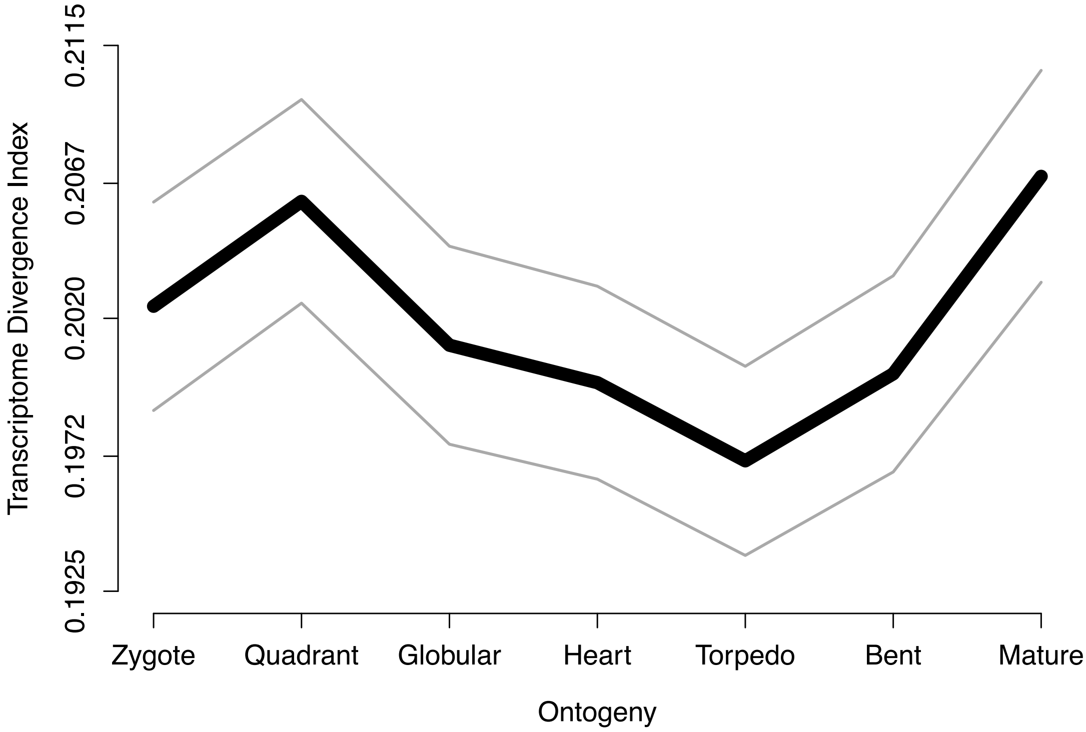


Figure 3.4.7: Transcriptome Divergence Indices computed for seven stages of *A. thaliana* embryogenesis in comparison with *A. lyrata*. The black line connects the transcriptome divergence indices across seven stages of *A. thaliana* embryogenesis and grey lines represent the standard deviation generated by a permutation test. This example pattern of Transcriptome Divergence Indices follows a high - low - high pattern (hourglass pattern) of average transcriptome divergence throughout embryo development [45]. The gray lines represent the standard deviation estimated by permutation analysis.

The TDI measure has similar disadvantages as proposed for the TAI [20, 29]:

- TDI is only defined for absolute gene expression levels
- TDI patterns are not always robust against data transformation such as *log* or *sqrt*
- TDI is biased by outlier age assignments or outlier weights (gene expression values)

- TDI only captures the top 2 - 10 % of highly expressed genes

In summary, transcriptome indices are global measures quantifying the average transcriptome conservation in biological processes. Two independent measures, TAI and TDI aim to determine the deep evolutionary (TAI) and recent evolutionary (TDI) conservation of transcriptomes. Although both measures have disadvantages, they provide a first indication for the potential existence of stages or periods of transcriptome conservation within biological processes of interest.

The TAI and TDI measures capture the global trend of average transcriptome conservation throughout development or a specific biological process and represent an average profile of the transcript contribution of all PS or DS classes. To scrutinize the average expression trend of each individual age or divergence category, relative expression levels were introduced to quantify the global expression trend of each age or divergence category separately ([26,45]).

Relative expression levels are defined as follows:

$$r_{is} = \frac{\bar{f_{is}} - \bar{f}_{\mathrm{imin}}}{\bar{f}_{\mathrm{i}max} - \bar{f}_{\mathrm{i}min}} \tag{3}$$

where $\bar{f_{is}}$ denotes the mean expression level for a specific phylostratum i, $i = 1, ..., N$ and developmental stage s, $s = 1, ..., S$, whereas $\bar{f}_{\mathrm{i}min} = min\{\bar{f_{i1}}, ...\bar{f_{iS}}\}$ denotes the minimum over all absolute mean expression levels for a specific phylostratum i and $\bar{f}_{\mathrm{i}max} = max\{\bar{f_{i1}}, ...\bar{f_{iS}}\}$ denotes the maximum over all absolute mean expression levels for a specific phylostratum i [26].

The *relative expression* value is a linear transformation of the absolute mean expression levels into the interval $[0, 1]$ and allows comparisons between the mean expression patterns of different phylostrata having different absolute mean expression levels. A *relative expression* value of 1 is defined to have the highest expression in relation to the global minimum and *relative expression* value of 0 is defined to represent the lowest mean expression level in relation to the global mean expression pattern.

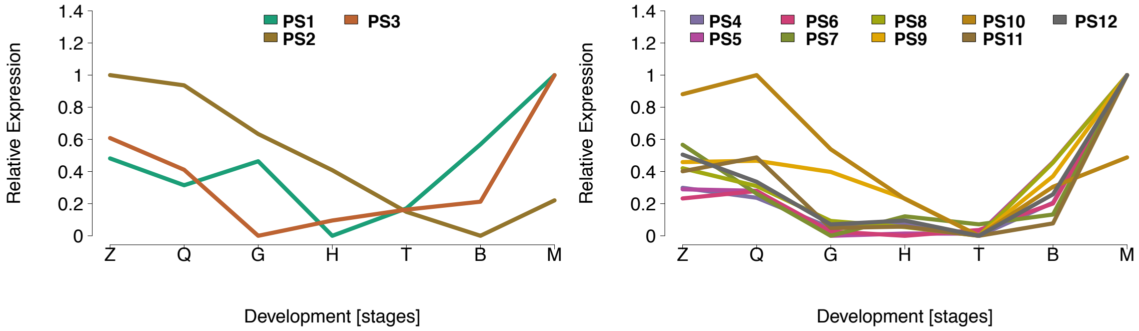


Figure 3.4.8: Relative expression levels (RE) computed for seven stages of *A. thaliana* embryogenesis. **a** RE profiles for 12 phylostrata across seven stages of *A. thaliana* embryogenesis. Each line represents the transformed average expression trend of genes that have been classified into the same phylostratum (age category). The stage with the highest mean expression levels of the genes within a PS (in comparison to all other stages) was set to RE = 1, the stage with the lowest mean expression levels of the genes within a PS (in comparison to all other stages) was set to RE = 0, the remaining stages were adjusted according to the linear transformation. Phylostrata are classified into two groups: group *evolutionarily old* contains PS that categorize genes that originated before complex/multicellular plants evolved (PS1–3) and group *evolutionarily young* contains PS that categorize genes that originated after complex plants evolved (PS4–12). In this example, *evolutionarily young* PS are down-regulated towards the *phylotypic stage* in plants (Torpedo) and up-regulated afterwards [45].

REs allow to investigate whether or not certain PS or DS categories show a common co-expression trend and thus, are contributing to the global TAI or TDI profile. As an example: to detect stages marking an ontogenetic transition in embryo development, the mean relative expression levels of *evolutionarily old* versus *evolutionarily young* PS can be visualized and the significant differences between both classes statistically quantified (Fig. 3.4.9). Developmental stages of significant differences between *evolutionarily old* and *evolutionarily young* RE values provide evidence for the existence of an underlying ontogenetic transition.

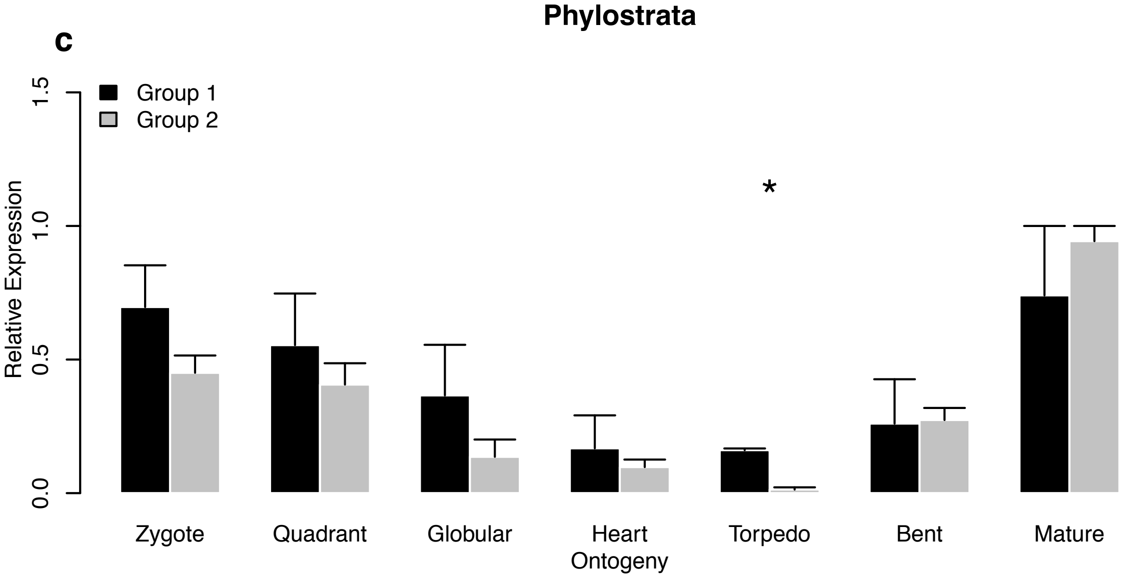


Figure 3.4.9: Mean relative expression levels of *evolutionarily old* versus *evolutionarily young* phylostrata. Group 1 includes the RE values of *evolutionarily old* PS (PS1-3; black bar) and group 2 includes the RE values of *evolutionarily young* PS (PS4-12; gray bar). Asterisks denote significant differences between groups 1 and 2 during the torpedo stage, marking an ontogenetic transition in *A. thaliana* embryo development [45]. Statistical significance was quantified using a Kruskal-Wallis Rank Sum Test.

In summary, transcriptome indices can be computed for each stage of a biological process of interest. These indices quantify transcriptome conservation in two independent ways: the first measure, TAI, quantifies deep evolutionary conservation of the transcriptome whereas the second measure, TDI, quantifies recent evolutionary conservation of the investigated transcriptome (including potential functional conservation). The difference in transcriptome conservation between stages of this biological process is quantified by statistical tests [45] and specific patterns of differential transcriptome conservation between stages can mark conserved intervals within this process. Finally, relative expression level analyses allow researchers to investigate whether or not certain PS or DS categories show a common co-expression trend that could potentially explain the observed global pattern of transcriptome conservation.

3.5 Software Tools to Perform Phylotranscriptomic Analyses

The R programming language is widely used and highly appreciated in scientific research. The advantage of R is that it provides a broad statistical framework which consists of functions that are implemented in *Fortran* and *C/C++* and use the concept of *vectorization* to speed up computations [55, 56]. Computationally costly procedures can be written in *C/C++* and integrated via the Rcpp interface [57]. Based on these facts and a broad community of users and contributors to the R language, I chose to implement *myTAI* and *orthologr* in R to provide researchers easy to use frameworks for performing phylotranscriptomics studies.

3.5.1 R package myTAI

So far, the *myTAI* package consists of 41 functions and was downloaded 5000 times from CRAN. These 41 functions allow users to perform phylotranscriptomic analyses, gene age enrichment quantification (based on Fisher's exact test), differential gene expression analyses, and automated taxonomic information retrieval. The following list shows the detailed functionality of each function.

Phylotranscriptomics Measures:

- **TAI()** : Function to compute TAI
- **TDI()** : Function to compute TDI
- **REMatrix()** : Function to compute the RE profiles of all PS or DS
- **RE()** : Function to transform mean expression levels to relative expression levels
- **pTAI()** : Compute the PS contribution to the global TAI
- **pTDI()** : Compute the DS contribution to the global TDI
- **pMatrix()** : Compute partial TAI or TDI values
- **pStrata()** : Compute partial strata values

Visualization and Analytics Tools:

- **PlotPattern()** : Main function to plot the TAI or TDI profiles and perform statistical tests
- **PlotCorrelation()** : Function to plot the correlation between PS and DS
- **PlotRE()** : Function to plot RE profiles
- **PlotBarRE()** : Function to plot the REs of PS or DS as barplot
- **PlotMeans()** : Function to plot the mean expression profiles of PS or DS
- **PlotDistribution()** : Function to plot the frequency distribution of genes within the corresponding phylostratigraphic map or divergence map
- **PlotContribution()** : Plot the PS or DS contribution to the global TAI or TDI pattern
- **PlotEnrichment()** : Plot the PS or DS enrichment of a given gene set
- **PlotGeneSet()** : Plot the expression profiles of a gene set

- **PlotCategoryExpr()** : Plot the expression levels of each age or divergence category as barplot or violinplot
- **PlotGroupDiffs()** : Plot the significant differences between gene expression distributions of PS or DS groups
- **PlotSelectedAgeDistr()** : Plot the PS or DS distribution of a selected set of genes

A Statistical Framework and Test Statistics

- **FlatLineTest()** : Function to perform the Flat Line Test
- **ReductiveHourglassTest()** : Function to perform the Reductive Hourglass Test
- **EarlyConservationTest()** : Function to perform the Reductive Early Conservation Test
- **EnrichmentTest()** : PS or DS enrichment of a given gene set based on Fisher's exact test
- **bootMatrix()** : Compute a permutation matrix for building custom test statistics

Differential Gene Expression Analysis

- **DiffGenes()** : Implements popular methods for differential gene expression analysis
- **CollapseReplicates()** : Combine replicates in an ExpressionSet
- **CombinatorialSignificance()** : Compute the statistical significance of each replicate combination
- **Expressed()** : Filter expression levels in gene expression matrices (define expressed genes)
- **SelectGeneSet()** : Select a subset of genes in an ExpressionSet
- **PlotReplicateQuality()** : Plot the quality of biological replicates
- **GroupDiffs()** : Quantify the significant differences between gene expression distributions of PS or DS groups

Taxonomic Information Retrieval

- **taxonomy()** : Automatic retrieval of taxonomic information for any organism of interest

Additional Analyses

- **MatchMap()** : Match a PS Map or DS Map with an ExpressionMatrix object

- **tf()** : Transform gene expression levels (e.g. log2 or sqrt)
- **age.apply()** : Age category specific *apply* function
- **ecScore()** : Compute the hourglass score for the Early Conservation Test
- **geom.mean()** : Optimized geometric mean computation
- **harm.mean()** : Optimized harmonic mean computation
- **omitMatrix()** : Compute TAI or TDI profiles omitting a given gene
- **rhScore()** : Compute the hourglass score for the Reductive Hourglass Test

For each function, I designed and implemented automated unit tests to ensure that they compute correct values and to provide a high software quality standard.

In addition, the functions **TAI()**, **TDI()**, **REMatrix()**, **pTAI()**, **pTDI()**, **pMatrix()**, **pStrata()**, **PlotPattern()**, **FlatLineTest()**, **ReductiveHourglassTest()**, **EarlyConservationTest()**, **bootMatrix()**, **ecScore()**, **rhScore()**, **geom.mean()**, **harm.mean()** , and **omitMatrix()** are implemented in C++ and are integrated into R via the Rcpp framework [57].

The functions **PlotPattern()**, **FlatLineTest()**, **ReductiveHourglassTest()**, **EarlyConservationTest()**, **bootMatrix()** are parallelized for multicore processing on a server or HPC cluster using the doParallel framework [58].

These optimizations reduce the computing speed of the corresponding functions by approx. 100 - 1000 fold when compared with the same functionality implemented in R.

Readers will find a detailed documentation of each function as well as detailed information covering the functionality of *myTAI* in the Appendix of this thesis.

3.5.2 R package orthologr

The R package, *orthologr* consists of 21 functions and was downloaded 2000 times from Github. These 21 functions allow users to perform BLAST searches, genome wide orthology inference methods, multiple sequence alignments, codon alignments, genome wide dNdS estimation, and genome wide divergence stratigraphy with R. The following list shows the detailed functionality of each function.

Perform Divergence Stratigraphy

- **divergence_stratigraphy()**: Perform the Divergence Stratigraphy algorithm
- **DivergenceMap()**: Sort dN/dS values into DS

Perform BLAST searches

- **advanced_blast()**: Perform an advanced BLAST+ search (wrapper function for BLAST command line tool)
- **advanced_makedb()**: Create a BLASTable database with makeblastdb (wrapper function for command line tool *makeblastdb*)
- **blast()**: Run BLAST+ search and BLAST output parser
- **blast.nr()**: Run BLASTP search against local *NCBI nr* database and parse the output
- **blast_best()**: Perform a BLAST+ best hit search between two genomes
- **blast_rec()**: Perform a BLAST+ best reciprocal hit (BRH) search between two genomes
- **delta.blast()**: Perform a DELTA-BLAST Search between either two genomes or against local *NCBI nr* database

Perform Pairwise and Multiple Sequence Alignements

- **multi_aln()**: Compute multiple sequence alignments (MSA) based on the *clustalw*, *t_coffee*, *muscle*, *clustalo*, and *mafft* programs (wrapper function for the corresponding MSA command line tools)
- **pairwise_aln()**: Compute pairwise alignments (either Needleman-Wunsch algorithm or Smith-Waterman algorithm)
- **codon_aln()**: Compute a codon alignment based on the PAL2NAL program

Perform Orthology Inference

- **orthologs()**: Main genome wide orthology inference function
- **ProteinOrtho()**: Orthology inference with ProteinOrtho (wrapper function for ProteinOrtho command line tool)

Perform Population Genomics

- **compute_dnds()**: Compute dN/dS values for a given pairwise alignment
- **dNdS()**: Compute dN/dS values for all orthologous genes between two genomes
- **substitutionrate()**: Internal function for dNdS computations

Read and Write CDS, Genomes, and Proteomes

- **read.cds()**: Read the coding sequences of a given organism (genome)
- **read.genome()**: Read the entire genome sequence of a given organism
- **read.proteome()**: Read the entire proteome sequence of a given organism
- **write.proteome()**: Save a proteome in fasta format

The major function of *orthologr*, **divergence_stratigraphy()** has been optimized and parallalized for multicore processing on a server or HPC cluster using the doParallel framework [58]. This function allows users to run the Divergence Stratigraphy algorithm for comparing 2 genomes on 32 cores in 30 minutes (2.5 Ghz per core). An older implementation (Perl Script) using the same settings [59] terminated after 48h and did not provide the same functionality, reproducibility and usability.

Readers will find a detailed documentation of each function as well as detailed information covering the functionality of *orthologr* and all optional algorithms implemented in each function in the Appendix of this thesis.

4 Results and Discussion

In this thesis I aimed to answer fundamental questions regarding one of the historical concepts of developmental biology by developing and applying bioinformatic tools and statistical tests that allow to address the developmental hourglass concept on a transcriptomic level and made these tools accessible to non-bioinformatics experts.

The implementation of the software tools *orthologr* and *myTAI* allowed me and my colleagues to systematically investigate the active maintenance and potential functional conservation of the developmental hourglass phenomenon in extant animal and plant species. As a result of applying *orthologr* and *myTAI* to transcriptome datasets covering animal and plant development, we found that the transcriptomic hourglass patterns in animals and plants are not a rudiment of a process that was once active but has progressively degraded since then, but rather reflects an ongoing process that is still detectable between closely related species. This finding indicates that the molecular function of the hourglass might still be conserved and under selection in extant species allowing future experimental studies to investigate their molecular functions.

When applying *myTAI* to the two most important ontogenetic transitions in the postembryonic development of *A. thaliana*: The transition from the embryonic to the vegetative phase, and the transition from the vegetative to the reproductive phase, we found that in plants these transitions also follow hourglass patterns. This suggests that not a process specific for embryo development, but an even more fundamental process present in all three developmental transitions might generate this pattern. Furthermore, in practically all animal studies, the phylotypic stage has always been associated with the onset of organogenesis. To directly address this aspect in plant development, we performed a control experiment across flower development. As flower development is a process that is dominated by the genesis of the various floral organs, a causal connection between organogenesis and the phylotypic stage would have predicted especially conserved transcriptomes during organogenic stages of flower development. However, flower development displayed no pattern at all, indicating that in plants (and possibly also animals) the mechanism underlying the hourglass pattern is more fundamental than the currently favoured explanation for animal systems. Taken together, the organogenesis-centered explanation is not able to explain the two post-embryonic hourglass patterns we found in plants.

In this chapter, I will first discuss future directions of *orthologr* and *myTAI*. Second, I will elaborate on the observation that the transcriptomic hourglass pattern in animals and plants is actively maintained as well as its implication for future experimental studies aiming to investigate its molecular function. Third, I propose a potential scenario that might explain the observed discrepancy between organogenesis and the hourglass patterns found in postembryonic plant development. Lastly, I will try to provide a synthesis of the presented findings and postulate a model that proposes that transcriptomic hourglass patterns could be a common pattern in many biological processes.

In order to discuss and unify my findings, some of the arguments in the following discussion will be redundant to the arguments raised in the discussions of the individual papers. Unfortunately, this redundancy is impossible to omit, but it will be beneficial for discussing a broader context that leads to a novel hypothesis that might be able to explain the phenomenon of transcriptomic hourglass patterns in biological processes and across kingdoms of life.

4.1 The Developmental Hourglass Model

> One puzzling feature of the debate in this field is that while many authors have written of a conserved embryonic stage, no one has cited any comparative data in support of the idea. It is almost as though the phylotypic stage is regarded as a biological concept for which no proof is needed.
>
> - Richardson et al. (1997)

The developmental hourglass phenomenon has a longstanding and fascinating scientific history. The first scientific reports postulating this embryological phenomenon that animal species pass through stages of morphological resemblance during mid embryogenesis can already be found in the 18th century [11].

Karl Ernst von Baer (1828) was one of the pioneering embryologists who sought for laws of developmental transformation that had the potential to explain the process of animal embryogenesis and its conservation between species [60] [61][pp. 70-91], [30][pp. 14-15]. It was noticed by the leading embryologists at his time (e.g. Johann Meckel, Etienne Serres, and Louis Agassiz) that developmental processes of distinct species resemble each other [11].
In the literature, however, Karl Ernst von Baer is credited to be the first person who scientifically described the phenomenon that embryos of different amniotes often show a striking morphological resemblance [60][p. 221], [15, 62]. Based on his detailed observations he postulated his laws of embryology that correlate morphological complexity with embryonic development [3, 60, 63, 64]:

> *(i) the more general characters of a large group of animals appear earlier in their embryos than the more special characters*
>
> *(ii) from the most general forms the less general are developed, and so on, until finally the most special arise*
>
> *(iii) every embryo of a given animal form instead of passing through the other forms, becomes separate from them*
>
> *(iv) fundamentally, therefore, the embryo of a higher form never resembles any other form, but only its embryo*
>
> - Originally published in german [60][p. 224], translation taken from [65][p. 5], [64][p. 713], also reviewed in [61].

Based on von Baers observations the term developmental hourglass model was coined more than 150 years later by Denis Duboule and Rudolf Raff to illustrate the morphological series of embryo resemblance during embryogenesis [10,12] (Figure 4.1.1). This metaphor of the developmental trajectory of animal embryos substitutes sand by development that flows from the top to the bottom [10][pp. 208 - 209] and aims to categorize embryogenesis into three phases: pre-phylotypic developmental trajectories, phylotypic stage or period, and von Baerian developmental trajectories. Today, this classification into three periods of embryogenesis is referred to as early embryogenesis, mid embryogenesis, and late embryogenesis and the conceptual division of embryogenesis is performed by referring to morphological characteristics. At this stage I would like to point out that von Baers himself was never aware of any evidence that early embryogenesis might be dissimilar between different animal species.

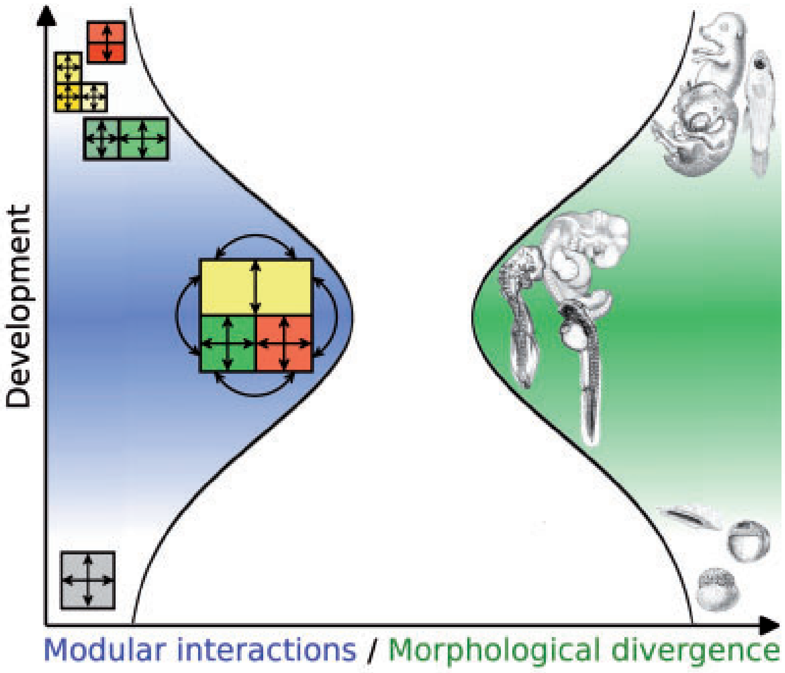


Figure 4.1.1: The developmental hourglass model according to Rudolf Raff (1996) [10]. A web of complex interactions among developmental modules results in selective constraints during midembryogenesis. In the phylotypic period modular interactions maximize and morphological divergence minimizes resulting in the bottleneck of the developmental hourglass model. Rudolf Raff states: *The volume of the hourglass represents probability space, with the width of the hourglass at any level representing the probability that a change can be successfully incorporated into a developmental pathway at that level.* [10] (illustration adapted from [27] and taken from [66]).

Based on this morphological division of embryo development it is known that early embryos of different phyla exhibit a vast morphological diversity [61, 67–69]. It has been suggested that the main reason for the diversity of early embryos is the variability of reproductive lifestyles of animals due to ecological adaptations [69]. The main argument for this comes from Jonathan Slack (2003):

> *[...] organisms come to occupy different niches in which they produce either a lot of small eggs with a poor chance of survival or a few large eggs with a good chance of survival. The presence of more yolk drives various changes in early development, including the disposition of cleavages (meroblastic rather than holoblastic) and the nature of gastrulation movements [...]. Viviparity imposes even more drastic changes on early embryonic life and is necessarily accompanied by the early formation of a variety of extra-embryonic membranes and supporting structures from the zygote as well as the mother. So early development is necessarily diverse because reproductive behaviour is diverse.* [69][pp. 311-312].

Slack's argument is strongly supported by molecular studies comparing the genetic regulation of early embryogenesis (reviewed in [16]). Seminal studies on the genetic regulation of early development provide evidence that in both vertebrates and invertebrates earliest stages of development are most divergent both in terms of gene expression and protein sequence evolution [16][p. 389 - 391]. Kalinka and Tomancak (2012) [16] point out that the proposed measures of gene divergence are well suited for assessing the ease with which early development can be altered and constraints active in early development remain explainable due to the inter-connected nature of gene regulation, that is changes in gene expression or sequence similarity do not necessarily completely ablate the function of individual genes [16][p.389]. They furthermore argue that knockout studies provide only partial measures of selective constraints (= selection on genomic loci that limit the occurrence of non-effecting or beneficial mutations) acting on genes contributing to early development and that the topology of regulatory networks shown to be substantially diverse among insect species will provide a more complete picture on the molecular nature of conserved genes and processes in early development.

I agree with the arguments raised by Slack (2003) [69] and the conclusions presented by Kalinka and Tomancak (2012) [16] and would like to add several aspects of conservation and variability of early development. Brian K. Hall (1999) [61] provides a detailed review on the conserved stages and conserved processes of early vertebrate development [61][pp. 69-71, p. 129, pp. 123 - 196]. Hall points out that the major stages through which all embryonic vertebrates pass are remarkably similar [61][p. 129] :

Zygote → Blastula → Gastrula → Neurula

whereas conserved processes include:

Fertilization → Cleavage → Gastrulation → Neurulation.

Hence, the correct establishment of these stages and processes are common to all vertebrate embryos. In most cases, knockout experiments that aim to prevent early stage embryos from transitioning through these stages or inhibit the functionality of the above mentioned processes will find a conservation across vertebrate embryos due to the conservation of developmental transitions rather due to the total conservation of all genes contributing to early development. In other words, it is the developmental transition itself that might be conserved and not particular genes that are most active during these transitioning stages. I argue that studies investigating the conservation of early development on the molecular level need to be aware of the environmental effects on molecular changes in development (reviewed in [70][pp. 3 - 78]), the correlation between gene regulatory network topology and morphology (reviewed in [71–80]), and the context of epigenetic effects on development (reviewed in [81][pp. 424 - 438]).

In my opinion, it is not surprising to find conserved genes that are lethal to the embryo or show correlations of developmental or morphological effects between species if they can be linked to the generation of the above mentioned embryonic stages that are universal to vertebrates (as I was able to demonstrate earlier [30]). It would be interesting and necessary to design experiments that are able to map conserved genes that have been classified by knockout experiments to the overall gene regulatory apparatus that governs the morphology and transition of vertebrate embryos (a first theoretical approach for this was introduced by Akhshabi *et al.* 2014 [34]).

However, as noted by Hall (1999) [61]: *embryos of closely related species with profoundly different patterns of cleavage and/or gastrulation, produce adults with similar morphologies* and *modified patterns of development are adaptations for embryonic or larval stages of the life cycle; embryos, larvae and adults can evolve independently* [61][p. 136] (see also [82–84]). This empirical observation indicates that the nature of the conservation and divergence of genes contributing to early development is very complex. Differing adaptive strategies (resulting from environmental factors) can still produce taxon specific morphologies that are most likely generated by conserved developmental processes and therefore, are controlled by conserved genes. This complexity illustrates that the definition of gene expression conservation or sequence conservation that is in use today is not sufficient to quantify conservation on the phylum (clade) level in order to correlate genetic constraints with morphological divergence. I believe, however, that the emerging field of gene regulatory network evolution will - one day - be able to elucidate the true nature of early embryogenesis.

Focusing on late animal embryogenesis on the other hand shows that the adaptational strategy causing morphological diversity of late embryos is different from early embryos. Slack (2003) [69] argues that by late development, the animal embryo is becoming quite similar to the postembryonic organism. Hence, free-living organisms are subject to selection and must acquire distinct niches for their survival to the reproductive stage, causing late stage embryos to be diverse. Arguably, also because the organisms to which they give rise are diverse [69].

A more interesting aspect of early, mid, and late embryogenesis from the environmental perspective is the fact that whereas early and late embryo development are accessible by environmental factors, mid embryos are surrounded by egg cases, jelly layers, or a uterus and therefore, are only partially exposed to environmental factors [69].

Slack (2003) [69] furthermore argues that because early-middle stage embryos minimally interact with the environment it might be possible that there is no specific cause for this morphological conservation during mid embryogenesis [69]. Slack states:

> *[...] it is just the stage in the middle at which the selective pressure for change are minimized. As a result it is the stage most likely to retain the features of the common ancestor* [69][p. 312].

Although I agree that common sense based on scientific observations allows to speculate that selective pressure is minimized during mid embryogenesis in vertebrates due to reduced interactions with environmental factors, the exact set of genes that also cause the morphological conservation however, cannot be predicted from this assumption.

Slack (2003) furthermore points out that molecular biology will solve the puzzle concerning ontogeny and phylogeny:

> *The rise of molecular developmental biology has enabled the identification of a number of long-range molecular homologies in animal design that are truely compelling. The key to understanding the strength of this evidence is the arbitrariness of the molecular components used to build multicellular organisms. Some molecules are not at all arbitrary; they have to have certain features to exert their functions* [69][p. 313].

I fully agree with Slack's statement. Only a careful study of the relationships between molecular homologies in distinct species across the animal kingdom will shed light on the origins of the animal body plan and the conservation of its establishment during mid embryogenesis. It will also enable the quantification of the genetic processes allowing ontogeny to create phylogeny. Furthermore, only molecular evidence will support or dismiss the potentially causal relationships between environmental factors and the evolution of selective constraints that limit the diversity of animal forms.

In his article Slack (2003) concludes:

> *Although there are no actual body parts conserved across all animals, the functional domains of key developmental control genes are conserved. We pointed out that the stage of expression and action of these genes in different phyla corresponded to the previously defined pphylotypic stages"within each phylum. This is not surprising if the reason for the existence of phylotypic stages is considered to be the relative lack of selective pressure for change of anatomy at the early-middle developmental stage.*

I decided to include this long introduction to the developmental hourglass model to point out the importance of correlating morphological observations with environmental factors and phenomena observable on the molecular level. In my opinion, new insights on the origin of animal body plan conservation and establishment can only be obtained when all three aspects: environment, morphology, and molecular constraints are fully included into the modelling (hypothesis generation) process, that I try to achieve in this thesis.

In summary, the developmental hourglass model is a metaphor to categorize animal embryogenesis into three morphological states: early, mid, and late embryogenesis. This categorization enables us to quantify the selection pressures acting on each state of embryo development separately. It furthermore, allows us to detect correlations between the morphological conservation of animal embryos across species, environmental factors, and the genetic machinery that controls the corresponding developmental processes. Hence, the correlation between phenotype and genotype during the phylotypic period must be investigated from both perspectives: from the ecological and from the molecular perspective. This can be motivated by the consensus of evolutionary developmental biology research of the past 30 years concluding that ontogeny creates phylogeny and therefore only genetic changes resulting in different developmental outcomes can be exposed to the environment and therefore exposed to selection processes [85].

For this reason, in the following section I will summarize the scientific knowledge covering ecological, developmental, and molecular evidence supporting the correlation with and a possible cause for the morphological and molecular conservation of animal embryos during the phylotypic period. Furthermore, the following chapter will point out the studies (including my own findings) postulating a conserved phylotypic period in plant embryogenesis and fungi development on the molecular level and I will correlate these observations with the findings and conclusions derived from the animal field.

4.2 From Morphology to Genetics: The Phylotypic Period on the Molecular Level

> Homology is indeed morphology's central conception.
>
> - Julian S. Huxley

The presumption that the basic body plan of the phylum is laid down at the stage of maximal morphological resemblance was first proposed by Friedrich Seidel (1960) [86] who defined this stage of maximal morphological resemblance in insects as *Körpergrundgestalt* [86], [61][p. 228]. This term was later adapted to objectively define *basic body patterns* in animals [9, 13, 87] and was picked up by Jack Cohen in 1977 who denoted this body pattern as *phyletic* [88]. Klaus Sander (1983) [13] proposed to rephrase the *phyletic* stage to *phylotypic* stage due to the misconception that the term *phyletic* refers to phylogenesis rather than to characters typical of individual phyla, which is still in use today [61], [13][p. 140]. To overcome this misconception, Sander defined the *phylotypic stage* as *the first stage that reveals the general characters shared by all members of the phylum* [13][p. 140] [13][p. 140].

In my opinion Sander (1983) [13] provides the best morphological explanation of the *phylotypic* stage that can be summarized as follows:

> *Incidentally it is the stage separating 'primitive development' from 'definitive development' in the terminology of the classical embryologist (e.g. Schleip, 1929). Different members of a phylum embark on ontogenesis from very different starting conditions [...]. Generelly speaking, the phylotypic stage is the stage of greatest similarity between forms which, during evolution, have differently specialized both in their modes of adult life and with respect to the earliest stages of ontogenesis which are strongly influenced by special modes of reproduction (e.g. ovipary v. vivipary). It is also the earliest stage which on a general scale permits establishing homologies (Spemann, 1915)* [13][p. 140], [89], [90].

Hence, Sander argues that the transition from *primitive development* to *definitive development* marks a crucial phase during embryogenesis which is reflected by the similarity between organismal forms. This statement is particularly interesting, because although characterized by classical embryological features, this developmental transition is a crucial step for any developing embryo. However, Sander does not present a clear correlation between this developmental transition and the underlying gene regulatory processes that are exposed to selective pressures due to their coordinating role in establishing this important transition.

It is tempting to assume that only this transition from *primitive development* to *definitive development*, hence the transition from cell differentiation to cell growth, which is shared among all multicellular organisms, is the only cause for the morphological resemblance of embryos. Even if this would be the case, this simple explanation is not capable of predicting the exact set of genes that are highly exposed to negative selection. It therefore, lacks the potential to predict the genetic mechanisms that cause the morphological resemblance between species. This illustrates that so far the phylotypic stage was defined and characterized by the homology of morphological features. Finally, technologies in molecular biology were not sufficient enough yet to quantify the evolutionary conservation on the molecular level in that particular stage of development.

Apart from terming and defining phylotypic stages, many different embryonic stages have been proposed to fulfill the definition of being phylotypic in the past 150 years [13,67,91,92] (reviewed in [17]). Ernst Haeckel (1874) argued that the tailbud stage fulfills the criteria for being phylotypic [93], William Ballard (1981) argued for the pharyngula stage [91], Lewis Wolpert (1991) argued for the early stage of somite segregation just after neurulation [94], Slack, Holland, and Graham (1993) also argue for the tailbud stage, because the zootype is most clearly expressed during the tailbud stage [95], Denis Duboule (1994) argues for the stage during gastrulation between the headfold stage and tailbud stage [12], and Galis and Metz (2001) argue for the onset of neurulation until the formation of the last somites [96].

This inconsistency was extensively reviewed and discussed by Michael Richardson (1995) who reexamined the morphological data relating to developmental timing in somite-stage embryos based on the published literature [14]. In his 1995 study, Richardson found convincing patterns of heterochrony (= evolutionary changes in the timing of developmental events of a descendent ontogeny relative to the state in an ancestral ontogeny [97][p. 193]) during vertebrate evolution that strongly affected the shifts of phylotypic stages in different vertebrates [14]. From these patterns he concluded that the phylotypic stage is poorly conserved and is more appropriately described as a phylotypic period [14]. To provide further evidence for his conclusions, Richardson *et al.* (1997) published a study presenting the first review of the external morphology of tailbud embryos since Haeckel, illustrated with original specimens from a variety of vertebrate groups [15]. They found that embryos at the tailbud stage show variations in somite number, embryo sizes and other forms due to allometry, heterochrony, and differences in body plan [15]. From this seminal study Richardson *et al.* concluded:

> *Contrary to recent claims that all vertebrate embryos pass through a stage when they are the same size, we find a greater than 10-fold variation in greatest length at the tailbud stage. Our survey seriously undermines the credibility of Haeckel's drawings [..]. In fact, the taxonomic level of greatest resemblance among vertebrate embryos is below the subphylum. The wide variation in morphology among vertebrate embryos is difficult to reconcile with the idea of a phylogenetically-conserved tailbud stage, and suggests that at least some developmental mechanisms are not highly constrained by the zootype* [15].

In my opinion, the most important conclusion presented by Richardson *et al.* was drawn with regard to molecular studies. They pointed out: *the dangers of drawing general conclusions about vertebrate development from studies of gene expression in a small number of laboratory species* [15].

This striking evidence provided by Richardson *et al.* illustrates the difficulty to quantify the biological variation of embryo development in empirical studies. This lack of quantification on the morphological level makes it difficult to reduce mid embryogenesis to a common morphological stage of intra-phylum resemblance. It, furthermore, supports arguments raised by Adam Sedgewick (1894), Frank Lillie (1919), Gavin deBeer (1940), and Rudolf Raff (1996) who criticized the reduction of embryogenesis to the morphological resemblance and a unifying stage [98], [65][p. 5], [10, 15, 62]. Instead, it motivates comprehensive studies on the molecular level that aim to quantify the exact genetic determinants for inter-species conservation of gene function and gene expression that might be causal for this morphological phenomenon.

The first studies examining this morphological phenomenon using molecular experiments were already performed in the 1980s (reviewed in [30]), but as mentioned above the first transcriptome study to systematically investigate the genetic conservation throughout embryogenesis was performed by Domazet-Lošo and Tautz in 2010 [26]. To map the morphological phenomenon of animal embryo development to the genetic level they first presented a method to quantify the evolutionary age of protein coding genes termed *phylostratigraphy* [49]. Second, they combined this approach with developmental transcriptome across the zebrafish life cycle (spanning development from embryogenesis to senescence). Third, they developed a transcriptome index to quantify the transcriptome conservation throughout the life cycle of zebrafish.

In general, phylostratigraphy aims to map gene origination events to a class of sequenced genomes which is consistent with the evolutionary history of the tree of life (= detectable homology). In particular, phylostratigraphy aims to detect putative functional genomic sequences that are conserved (homolog) between the organism of interest and the most distant common ancestor and classifies matching sequences to the taxonomic category of the most distant hit. In this regard, gene age is defined by the taxonomic category of the most distant detectable homolog and is purely based on sequence homology. These conserved sequences allow us to trace back the putative origination of this particular genomic sequence along the tree of life, answering the question *when* (in which kingdom/phyla/domain/species etc.) this sequence might have emerged. The resulting table generated by the phylostratigraphy algorithm stores the gene age assignment of each protein-coding gene of the query genome and is called *phylostratigraphic map* (see Fig. 3.2.4) [26, 49].

A procedure for associating a phylostratigraphic map with expression data of the developmental process of interest is based on transcriptome indices. The goal of these indices is to obtain the average evolutionary age of the transcriptome for each stage of the biological process of interest. In brief, this procedure works as follows: Compute for each stage the weighted mean of gene age, where the weights are the stage-specific expression levels. This weighted mean is called transcriptome age index (TAI; [26]), and the profile of stage-specific TAI values across all stages of the biological process is called TAI profile. Stages with high TAI values are stages where evolutionarily conserved genes are more lowly expressed and evolutionarily less conserved genes are more highly expressed than in other stages.

By applying this phylotranscriptomics approach, Domazet-Lošo and Tautz found that the phylotypic stage does indeed express the evolutionarily oldest transcriptome and that evolutionarily younger genes are expressed during early and late development, faithfully mirroring the developmental hourglass model on the morphological level [26]. So far, several evolutionary transcriptomics studies were able to provide supporting evidence for the gene expression pattern conservation during the phylotypic period in animals [25, 27, 28, 31, 32, 35–39, 41, 42]. Two main approaches exist to quantify transcriptome conservation. The first approach, phylotranscriptomics, is based on gene age inference combined with gene expression information [26]. The second approach, comparative transcriptomics [25, 27, 32, 35], is based on gene orthology inference and quantification of gene expression pattern conservation. This thesis focuses on the phylotranscriptomic method, but a detailed review on comparative transcriptomics can be found in Roux *et al.* (2015) [99]. The phylotranscriptomic results suggest that the evolutionary transcriptomics approach proposed by Domazet-Lošo and Tautz is capable of mapping this morphological phenomenon to the molecular level (see Methods).

Motivated by the success of the phylotranscriptomic method, in 2012 we asked the question whether or not a similar pattern of transcriptome conservation could be found in the second kingdom of life that evolved embryogenesis as a developmental program to establish multicellularity, that is plants. Surprisingly, we were able to provide evidence for the existence of a similar pattern of transcriptome conservation (dissimilar - similar - dissimilar) for *Arabidopsis thaliana* embryogenesis [29, 45]. This finding was surprising due to the lack of morphological evidence that would suggest the existence of a developmental hourglass phenomenon in plants (there is only some evidence that mid-stage embryos of dicots are morphologically conserved; see [100]). Whereas animal embryos seem to follow a morphological hourglass pattern, plant embryos do not follow a clear morphological pattern of dissimilarity - similarity - dissimilarity between related plant species [30, 45] (e.g. monocot and dicot middle stage embryos differ dramatically on the morphological level [100]). However, the existence of a transcriptomic hourglass pattern in plants suggests that morphological and molecular patterns might be uncoupled, and thus raises fundamental questions about the actual correlation between the morphological hourglass phenomenon and body plan establishment in animals.

In other words, our findings indicate that body plan establishment and organogenesis are not the most fundamental processes that are causal for the emergence of a phylotypic period in animals and that this correlation is not sufficient enough to explain the observation of a molecular hourglass pattern in plants. We therefore speculated that 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 and that such a mechanism may be required for enabling spatio-temporal organization and differentiation of complex multicellular life in general [45].

This *uncoupling* hypothesis is supported by Cheng *et al.* who reported a molecular hourglass pattern in fungi development [46], the third kingdom of life that established multicellularity. In this study, Cheng *et al.* performed the phylotranscriptomic method for the mushroom-forming fungus *Coprinopsis cinerea* and found that the young fruiting body is the stage that expresses the evolutionarily oldest transcriptome, whereas the primordium and mature fruit body express an evolutionarily younger transcriptome. The fungus, thus follows a pattern of transcriptome dissimilarity - similarity - dissimilarity as observed in animals and plants [46].

In summary, the phylotypic period has been correlated with the transition from primitive development to definitive development marking a crucial phase during embryogenesis which is reflected by the similarity between organismal forms [13]. This crucial transition was then interpreted as partial explanation for the causal connection between inter-phylum resemblance and body plan establishment (organogenesis). Recent findings proposing molecular hourglass patterns in animals and later observations postulating the existence of molecular hourglass patterns in plant embryogenesis and fungi development, however, raise the important question whether or not the transition from primitive development to definitive development reflects the only causal connection between body plan formation, organogenesis, and morphological resemblance or whether a more fundamental molecular mechanism is causing both morphological and transcriptome conservation during the establishment of multicellular organisms across kingdoms.

4.3 Evidence for Active Maintenance of Phylotranscriptomic Hourglass Patterns

The controversy about the developmental hourglass model and especially about the hourglass versus early conservation model experiences a new wave of arguments due to the recent finding of transcriptome conservation in animal and plant embryogenesis. For decades it was debated whether the earliest developmental stages of animal embryos are considered foundational and any apparent conservation in later stages is the delayed realization of the conservation of genes and proteins acting early (early conservation model [16]), or if conservation is greatest in mid-embryogenesis and is the result of the need for coordination between growth and patterning when the body plan is being built (developmental hourglass model [16]).

These and other models have traditionally relied on subjective anatomical comparisons, and a lack of measurable quantitative approaches has fed controversial discussions over decades [15,18,19,85,101,102]. Although some of these recent studies favored the early conservation model [18,19], the majority of them supported the developmental hourglass model [21–44].

Later, several studies demonstrated that whole transcriptomes of fly, worm, several vertebrates, and cress followed an hourglass pattern [25–28,32,35,43,45]. This was supported by several additional studies. Among them was for example a study proposing that the conservation of miRNA expression displays an hourglass pattern similar to that observed for protein-coding genes [103], a study in mammals showing that distinct patterns of sequence evolution apply to enhancers with transient *in vivo* activities in mammalian development and, therefore, are in favour of the hourglass model [104], and an epigenetic study reporting widespread DNA demethylation of enhancers during the phylotypic period in zebrafish, *Xenopus tropicalis*, and mouse indicating that DNA methylation might be an upstream regulator of phylotypic enhancer function [44].

In this thesis, I systematically analyzed embryonic transcriptomes of two animal and one plant species. My co-authors and I found that the developmental transcriptomes of fly, fish, and cress follow a transcriptomic hourglass pattern and are actively maintained in extant species. Because the evaluation of transcriptomic patterns in past studies were subjective or relied on statistical tests with different limitations, we developed a statistical framework that extends the Flat Line Test [45], the Reductive Hourglass Test [105], and introduced the Reductive Early Conservation Test to quantify the possibility that a observed transcriptomic pattern might favour the early conservation model. These tests allow researchers to objectively assess transcriptome conservation profiles for any pattern significantly deviating from a flat line, for the significance of high–low–high or low-high-high patterns. In the both latter 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 [40].

The central question arising from these results is whether or not the transcriptomic hourglass pattern might still be associated with a biological function in extant species. If so, the transcriptomic hourglass pattern might either be causal for a downstream biological function or be the result of such a possible function. Alternatively, the transcriptomic 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 level. Hence, answering this question is important for understanding the still unknown 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 transcriptomic hourglass pattern by performing experiments on extant species [40].

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 signatures 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 *D. rerio* greater than 150 Ma, this relatively long time span does 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 ago. The statistical evaluations presented in this thesis show a significant hourglass-like pattern with the minimum during the presumptive phylotypic period, consistent with the developmental hourglass model. Hence, the TDI data shown propose a scenario in which, across kingdoms, the transcriptomic hourglass pattern is actively maintained through stabilizing selection [40].

Interestingly, while vertebrate and invertebrate embryogenesis also follows an hourglass pattern on the morphological level, morphological hourglass patterns are absent from plant embryogenesis. In contrast, comparative embryology in flowering plants, for example, suggests that the complete process of embryogenesis is morphologically highly conserved [106]. 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 [40].

In view of the lack of a morphological hourglass pattern in plants, one could conjecture that although the transcriptomic hourglass pattern might be actively maintained in extant species across kingdoms, transcriptomic 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.

In general, the TAI approach allows researchers to detect patterns of transcriptome conservation in biological processes. Stages of maximal transcriptome conservation might point to functional programs that are evolutionarily more conserved than other functional programs. The TDI approach will then clarify whether or not these stages or periods of maximal transcriptome conservation are still conserved between closely related species. In case these conserved stages are actively maintained, researchers can design experiments to alter or inhibit these functional programs and study the effects on the genetic and morphological level [40]. I believe, that these stages of maximal transcriptome conservation are ideal periods to study the correlation between genotypic changes and their effects on the phenotype. In the context of embryogenesis, phylotypic periods in animals and plants will be ideal stages to study environmental, epigenetic, and gene regulatory effects on development. Such studies might elucidate how developmental programs can change over evolutionary time scales and how this change constrains or promotes phenotypic diversification.

Together, the results presented in this theses allow me to conclude that the hourglass patterns in animal and plant embryogenesis are 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. My co-authors and I argue that due to its active maintenance the commonality and differences between the hourglass phenomena in animals and plants can be systematically studied and have the potential to significantly advance our understanding of the developmental hourglass phenomenon and its correlation with body plan establishment.

4.4 Post-embryonic Hourglass Patterns Mark Ontogenetic Transitions in Plant Development

The most recent studies on the last common ancestor (LCA) of animals and plants suggests that the animal and plant lineages split approx. 1.6 - 1.8 billion years ago [107]. Based on this molecular dating and the absence of paleontological evidence it is believed that this LCA of animals and plants was a unicellular organism and that multicellularity and embryogenesis must have evolved independently in both kingdoms [45, 107].

Hence, the well studied developmental hourglass concept in animals did not consider the plant kingdom. Our finding postulating that plant embryos also follow a transcriptomic hourglass pattern that is comparable to the animal transcriptomic hourglass pattern however, raises fundamental questions about the origination, commonality, and maintenance of these molecular hourglass patterns in both kingdoms [45].

It is evident due to the limited complexity of plant embryogenesis (limited extent of organogenesis), that the main hypothesis for the existence of developmental hourglass patterns proposed by animal studies claiming that organogenesis and body plan formation are the main constraints that shape this pattern [6], will be less powerful to predict the observation of a transcriptomic hourglass pattern in plants. However, to test this hypothesis derived from animal studies [6, 10] in plants we designed experiments to capture the transcriptome conservation in the two major postembyronic transitions in the life cycle of *A. thaliana*: the transition from the embryonic to the vegetative phase [108,109] and the transition from the vegetative to the reproductive phase (flowering). As a control, we quantified transcriptome conservation for flower development, a process that is dominated by organogenesis [110] (Paper 3). If indeed, postembryonic developmental processes in plants would be governed by hourglass patterns, this would suggest that transcriptomic hourglass patterns are not restricted to plant embryogenesis and possibly a wide-spread phenomenon that governs multiple processes. Furthermore, the potentially causative relationship among organogenesis, body plan establishment, and hourglass patterns that explains the animal phenomenon to date, would need to be re-evaluated.

My co-authors and I found 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 transcriptomic hourglass pattern which was previously discussed only with regard to embryogenesis, appears to be more widespread, at least in plants. Because no new organs are established during these two postembryonic phase transitions, our results also support the aforementioned conjecture that transcriptomic hourglass patterns are not specifically associated with organogenic processes [45]. This conjecture is crucial, because it challenges the current hypothesis that organogenesis and body plan formation are the major constraints that shape the developmental hourglass in animals.

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 [66]. This hypothesis is supported by the fact that although different sets of genes contribute to the transcriptome in the three major transitions of the plant life cycle: embryogenesis, embryo-to-vegetative, and vegetative-to-reproductive phase transitions, all three developmental processes show a comparable pattern of transcriptome conservation.

Hence, the stage resembling the waist of the hourglass marks the conservation of different developmental processes. If these stages are indeed organizational checkpoints serving as a switch that separates two functional programs we would expect to find crucial transcriptome switches during these stages. Studies investigating these developmental processes indeed provide evidence that this is the case.

For embryogenesis, the most conserved transcriptome can be found at the stage shown to be separating the developmental programs regulating cell differentiation and growth [111, 112]. For the embryo-to-vegetative transition it was shown that two transcriptional phases are separated by testa rupture [108] (Paper 4). The first phase is marked by large transcriptome changes as the seed switches from a dry, quiescent state to a hydrated and active state. The second transcriptional phase indicate a role for mechano-induced signaling at this stage and subsequently highlight the fates of the endosperm and radicle: senescence and growth [108] (Paper 4). The waist of the hourglass of the embryo-to-vegetative transition marks testa rupture as period of maximum transcriptome conservation and thus, supports the organizational checkpoint hypothesis.

During floral transition the leaf-producing shoot apical meristem is converted into an inflorescence meristem, which forms flowers. However, the actual transition occurs several days before bolting [66]. We observed that the waist of the hourglass of floral transition marks stages of maximal expression of floral homeotic genes and other marker genes known to regulate the transition from vegetative-to-reproductive growth [66].

It remains to be shown however, that organizational checkpoints are a feature of multiple developmental processes that also exist in animal development such as metamorphoses etc. The finding that fungi development also follows a transcriptomic hourglass pattern [46] indicates that these features might indeed exist in multiple biological processes and across kingdoms of life. Our hypothesis predicts that the period of maximum transcriptome conservation in fungi development (in the reported case: young fruiting body [46]) marks an organizational checkpoint serving as a switch that separates two functional programs (e.g. for this specific case in fungi: cell differentiation and growth).

As I introduced before, evolutionary developmental biology research of the past 30 years concludes that ontogeny creates phylogeny and does not recapitulate phylogeny [85]. Mutations in the regulatory regions of genes are able to change the expression pattern of the corresponding gene. Changes in expression patterns are then able to change phenotypes and these novel phenotypes are exposed to the environment. Over time, these novel phenotypes are either positively or negatively selected by the environment and only beneficial phenotypes spread through populations [85]. Depending on environmental conditions, these populations in evolutionary time can evolve into new species and thus, over long periods of time ontogenetic changes on the genetic level are able to create new species [85]. On the other side, conservation of gene expression patterns are able to limit phenotypic diversification [85]. In summary, changes in gene expression can generate novel phenotypes and the conservation of gene expression can limit phenotypic diversification.

One interpretation of organizational checkpoints could be that they mark stages in biological processes that show conserved expression patterns. In case of animal embryos, this conservation of expression patterns during mid-embryogenesis might limit morphological diversification and, thus, results in the conserved morphology of animal embryos during this stage (see also discussions in [30, 100, 113]). A plausible explanation for the existence of conserved expression patterns during this stage might be that the previously introduced transition from *primitive development* to *definitive development* marks this mid-embryonic stage in animals. These transitional stages that separate two functional programs of development (e.g. cell differentiation and cell growth) are highly susceptible to environmental factors, because any environmental influence on the developmental program during these stages could result in malformation or lethality of the embryo [10, 12].

In contrast, for plants I provided evidence suggesting that the expression of embryo defective genes (= essential genes for embryogenesis [114]) is maximized during mid-embryogenesis in *A. thaliana* [30, 40]. This finding indicates that the transcriptomic hourglass pattern found in *A. thaliana* not only marks an analogous organizational checkpoint separating *primitive development* and *definitive development* as the one found in animals, but that this checkpoint is also susceptible to environmental factors [100, 106, 113]. This susceptibility is evident due to the increase in lethal effects during mid-embryogenesis in *A. thaliana* when expression of embryo defective genes in this period is inhibited. Future work needs to elucidate how organizational checkpoints evolve in the first place and how the conservation of these checkpoints constrains morphological diversification (see also [30, 100, 113]).

Together, based on the evidence presented in this thesis I speculate that the developmental hourglass pattern found in animal embryogenesis might not be (causally) shaped by organogenesis and body plan establishment, but rather is the result of the conserved organizational checkpoint that separates early embryogenesis and mid embryogenesis which subsequently constrains morphological variability as well. The fact that transcriptomic hourglass patterns evolved independently in both animal and plant embryogenesis and independently mark organizational checkpoints that separate *primitive development* and *definitive development* suggests that organizational checkpoints are common features of developmental processes. These checkpoints reflect the conservation of developmental programs (order of events) and not the conservation of specific homologous genes (e.g. homologs between animals and plants). This hypothesis is supported by the findings of my co-authors and me that show that not only embryo development but also post-embryonic developmental processes in *A. thaliana* follow transcriptomic hourglass patterns which mark ontogenetic transitions. Whether organizational checkpoint not only mark transitions of two functional programs, but also mark stages or periods of environmental susceptibility and therefore limit diversification at this period in general (for any biological process that has a checkpoint) remains to be shown.

5 Conclusions and Outlook

In the past, the developmental hourglass concept has been used to study the relationship between morphological (phenotypic) conservation and body plan establishment. It seemed evident in animal embryogenesis that body plan establishment during embryogenesis can be linked with the morphological hourglass pattern describing the resemblance between animal embryos of different species throughout embryo development.

The fact that in animals transcriptome conservation throughout embryogenesis also follows an hourglass pattern and mirrors the morphological pattern, suggested that genotypic conservation and phenotypic conservation might be causally linked. However, the surprising finding that plant embryos also follow a transcriptomic hourglass pattern in the absence of an analogous morphological pattern followed by the finding that transcriptomic hourglass patterns are present in post-embryonic development of plants in which body plan establishment is absent, raises fundamental questions about the proposed causal relationship between genotypic conservation and phenotypic conservation of animal embryos.

In this thesis, I discussed my observation that transcriptomic hourglass patterns in animal and plant embryogenesis are actively maintained and that post-embryonic hourglass patterns in plants are decoupled from organogenesis and body plan establishment. This finding needs a more fundamental explanation than body plan establishment (which is currently the most supported hypothesis explaining the animal phenomenon). My co-authors and I hypothesized that stages of minimum transcriptome divergence mark organizational checkpoints serving as a switch that separates two functional programs. In animal embryogenesis these checkpoints coincide with phylotypic periods that were defined based on morphological observations. In particular, it is evident that in case of embryo development these two functional programs that are separated by a organizational checkpoint might be cell differentiation (early embryogenesis) and growth (mid/late embryogenesis).

This view is in accordance with the notion of *primitive development* transitioning to *definitive development* during embryogenesis that was introduced by Klaus Sander. The organizational checkpoint hypothesis presented in this thesis provides a testable model for this transition. It predicts that the transition from early embryogenesis to mid-embryogenesis which is shared among all multicellular organisms is causal for the morphological resemblance of embryos and provides a more fundamental explanation than the organogenesis hypothesis.

In post-embryonic development of plants such as germination and flowering these functional programs are most likely to be different from embryogenesis (e.g. during testa rupture in germination marking the emergence of the seedling from the seed, likely the transition period of this process, at which germination becomes irreversible as well as during floral transition).

Evidence discussed in this thesis indicates that organizational checkpoints might be highly susceptible to environmental factors and, thus, might limit diversification in general. I speculate, that the reason why the transition of two functional programs might be present in many biological processes and might be highly susceptible to environmental factors is that the key regulators that govern the corresponding transition are constrained by evolutionary history (Dollo's law).

In particular, these key regulators (e.g. transcription factors in higher hierarchies of gene regulatory networks) evolved to govern a specific sequential order of events (e.g. order of developmental programs). This sequential order of events is constrained by evolutionary history and any drastic change of this order due to mutations or environmental factors will result in malfunction or lethality.

I argue, that organizational checkpoints mark stages in biological processes in which the sequential order of biological events (e.g. developmental programs) is maximally conserved due to the evolutionary constraints acting on their key regulators. The conservation of expression patterns of these regulators is crucial to maintain the sequential order of these events. Hence, environmental factors or other genetic factors that cause the change of expression patterns of these regulators might result in the disruption of the sequential order of events, because changes of expression patterns of key regulators can result in changes of expression patterns of hundreds or thousands of target genes.

In summary, I speculate that changes of gene expression patterns of key regulators cause a disruption of the sequential order of events that guarantee the successful establishment of a specific biological process. In the context of embryogenesis, this sequential order is the correct establishment of organs (animals) or establishment of meristems (plants). This sequential order of events is maximally susceptible to environmental and genetic factors during transitions separating two functional programs (organizational checkpoints). In both, animal and plant embryogenesis these organizational checkpoints mark the transition from cell differentiation (*primitive development*) to cell growth (*definitive development*). Future studies need to determine these key regulators and need to investigate to which degree the sequence of events can be changed in comparison with non-organizational stages.

For example, the developmental sequence common to animal embryogenesis in different species (as introduced earlier): Fertilization → Cleavage → Gastrulation → Neurulation has a particular sequential order that evolved over evolutionary time. Changes in gene expression of the key regulators that govern these transitions might therefore disrupt the correct sequence of these developmental programs and might lead to malfunction or death of the embryo. This example shall illustrate how the sequence of developmental programs itself (once established during evolution) is constraining diversification (Dollo's law). The reason why animals did not evolve any other sequence of developmental programs might be that any change in this sequential order has been disadvantageous for the embryo or was not able to establish complex animal forms. Hence, in the context of embryogenesis developmental transitions are crucial for the correct establishment of complex multicellularity. In the context of biological processes transitions are crucial for executing the correct order of functional programs to establish, maintain, or govern the corresponding biological process.

Together, the organizational checkpoint concept challenges the previous causal explanation that links the emergence of developmental hourglass patterns to organogenesis and body plan establishment and in contrast proposes that in fact it is the organizational checkpoint which is under negative selection and thus, potentially causal for shaping and constraining all downstream processes. Future studies will illuminate whether or not this hypothesis holds true and will investigate whether or not organizational checkpoints are a common mark of many biological processes.

The above discussed analyses could not have been performed without the software tools *myTAI* and *orthologr*. I developed these open source R packages to be able to perform optimized and reproducible phylotranscriptomic analyses. The overall aim however, was to not only apply these tools to my own field of research, but to make them modular, reproducible, optimized, and user-friendly enough so that researchers (including non-bioinformatics experts) are able to apply phylotranscriptomic analyses to any transcriptome dataset and any biological process of interest. These tools will allow future studies to investigate stages of transcriptome conservation in many biological processes.

Future extensions of *myTAI* and *orthologr* could be the following: *myTAI* could be extended to perform gene expression clustering to detect co-expressed genes that can potentially be correlated with the hourglass pattern. These co-expression clusters can then be compared between the different hourglass patterns observed in animals, plants, and fungi and might unravel further commonalities or differences between these phenomena. In addition, a user interface (GUI) could be implemented using Shiny apps (RStudio) to make myTAI's functionality even more accessible. The statistical tests could be extended to test the significance of other patterns of transcriptome conservation than a flat line, a high-low-high, or a low-high-high pattern. Examples could be oscillatory patterns (e.g. circadian patterns), inverse hourglass patterns, or high-low-low patterns.

The R package *orthologr* can be extended by additional orthology inference methods and alternative methods to quantify the selection pressure acting on genes. In particular, these methods could include multiple closely related species instead of pairwise comparisons to quantify the substitution rates of genes. Finally, the divergence stratigraphy algorithm could be extended to perform efficient multiple genome alignments to detect conserved syntenic regions between closely related species.

References

[1] Lewis Wolpert et al. *Principles of Development.* Oxford University Press, 4th edition, 2011.

[2] Brian K Hall. *Homology: The Hierarchial Basis of Comparative Biology.* Academic Press, 1994.

[3] Katherine E Willmore. The body plan concept and its centrality in evo-devo. *Evolution: Education and Outreach*, 5(2):219–230, 2012.

[4] Takayuki Onai, Naoki Irie, and Shigeru Kuratani. The Evolutionary Origin of the Vertebrate Body Plan: The Problem of Head Segmentation. *Annual review of genomics and human genetics*, (May):1–17, 2014.

[5] Wallace Arthur. *The origin of animal body plans: A study in evolutionary developmental biology.* Cambridge University Press, 2000.

[6] Naoki Irie and Shigeru Kuratani. The developmental hourglass model: a predictor of the basic body plan? *Development*, 141(24):4649–55, 2014.

[7] John Maynard Smith, Richard Burian, Stuart Kauffman, Pere Alberch, John Campbell, Brian Goodwin, Russell Lande, David Raup, and Lewis Wolpert. Developmental constraints and evolution: a perspective from the mountain lake conference on development and evolution. *Quarterly Review of Biology*, pages 265–287, 1985.

[8] Rupert Riedl and Richard Peter Spencer Jefferies. *Order in living organisms: a systems analysis of evolution.* Wiley New York, 1978.

[9] Günter P Wagner. *Homology, genes, and evolutionary innovation.* Princeton University Press, 2014.

[10] Rudolf A Raff. *The Shape of Life.* The University of Chicago Press, Chicago, 1996.

[11] Arthur William Meyer. Some historical aspects of the recapitulation idea. *The Quarterly Review of Biology*, 10(4):379–396, 1935.

[12] Denis Duboule. Temporal colinearity and the phylotypic progression: a basis for the stability of avertebrate bauplan and the evolution of morphologies through heterochrony. *Development*, pages 135–142, 1994.

[13] Klaus Sander et al. The evolution of patterning mechanisms: gleanings from insect embryogenesis and spermatogenesis. *Development and evolution*, pages 137–159, 1983.

[14] Michael K Richardson. Heterochrony and the phylotypic period. *Developmental biology*, 172:412 – 421, 1995.

[15] Michael K Richardson et al. There is no highly conserved embryonic stage in the vertebrates: implications for current theories of evolution and development. *Anatomy and embryology*, 196(2):91–106, 1997.

[16] Alex T Kalinka and Pavel Tomancak. The evolution of early animal embryos: conservation or divergence? *Trends in Ecology and Evolution*, 27:385–393, 2012.

[17] Olaf RP Bininda-Emonds et al. Inverting the hourglass: quantitative evidence against the phylotypic stage in vertebrate development. *Proceedings of The Royal Society*, 270:341–346, 2003.

[18] Julien Roux and Marc Robinson-Rechavi. Developmental constraints on vertebrate genome evolution. *PLoS Genetics*, 4(12), 2008.

[19] Aurélie Comte, Julien Roux, and Marc Robinson-Rechavi. Molecular signaling in zebrafish development and the vertebrate phylotypic period. *Evolution and Development*, 12(2):144–156, 2010.

[20] Barbara Piasecka et al. The hourglass and the early conservation models - co-existing evolutionary patterns in vertebrate development. *PLOS Genetics*, 9(4):e1003476, 2013.

[21] Einat Hazkani-Covo 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*, 304:150–158, 2005.

[22] Naoki Irie and Atsuko Sehara-Fujisawa. The vertebrate phylotypic stage and an early bilaterian-related stage in mouse embryogenesis defined by genomic information. *BMC biology*, 5:1, 2007.

[23] Carlo G Artieri, Wilfried Haerty, and Rama S Singh. Ontogeny and phylogeny: molecular signatures of selection, constraint, and temporal pleiotropy in the development of drosophila. *BMC biology*, 7(1):42, 2009.

[24] Tami Cruickshank and Michael J. Wade. Microevolutionary support for a developmental hourglass: Gene expression patterns shape sequence variation and divergence in Drosophila. *Evolution and Development*, 10(5):583–590, 2008.

[25] Alex T Kalinka et al. Gene expression divergence recapitulates the developmental hourglass model. *Nature*, 468:811–814, December 2010.

[26] Tomislav Domazet-Lošo and Diethard Tautz. A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature*, 468:815–818, 2010.

[27] Naoki Irie and Shigeru Kuratani. Comparative transcriptome analysis reveals vertebrate phylotypic period during organogenesis. *Nature Communications*, pages 1–6, 2011.

[28] Michal Levin, Tamar Hashimshony, Florian Wagner, and Itai Yanai. Developmental Milestones Punctuate Gene Expression in the Caenorhabditis Embryo. *Developmental Cell*, 22(5):1101–1108, 2012.

[29] Hajk-Georg Drost. Development of a phylogenetic transcriptome atlas of arabidopsis thaliana. *Bachelor's thesis*, Martin Luther University Halle, 2011.

[30] Hajk-Georg Drost. A bioinformatics approach to study the origin of embryogenesis in animals and plants. Master's thesis, Martin Luther University Halle, 2013.

[31] Alicia N Schep and Boris Adryan. A comparative analysis of transcription factor expression during metazoan embryonic development. *PLOS ONE*, 8(6):1–13, 2013.

[32] Zhuo Wang et al. The draft genomes of softshell turtle and green sea turtle yield insights into the development and evolution of the turtlespecific body plan. *Nat. Genet.*, 45(6):701–708, 2013.

[33] Xiangjun Tian, Joan E. Strassmann, and David C. Queller. Dictyostelium development shows a novel pattern of evolutionary conservation. *Molecular Biology and Evolution*, 30(4):977–984, 2013.

[34] Saamer Akhshabi et al. An explanatory evo-devo model for the developmental hourglass [version 2; referees: 3 approved]. *F1000Research*, 3(156), 2014.

[35] Mark B Gerstein, Joel Rozowsky, Koon-Kiu Yan, Daifeng Wang, Chao Cheng, James B Brown, Carrie A Davis, LaDeana Hillier, Cristina Sisu, Jingyi Jessica Li, et al. Comparative analysis of the transcriptome across distant species. *Nature*, 512(7515):445–448, 2014.

[36] Ewart Kuijk, Niels Geijsen, and Edwin Cuppen. Pluripotency in the light of the developmental hourglass. *Biological Reviews*, pages n/a–n/a, 2014.

[37] Thomas Montavon and Natalia Soshnikova. Hox gene regulation and timing in embryogenesis. In *Seminars in cell & developmental biology*, volume 34, pages 76–84. Elsevier, 2014.

[38] Juan J Tena, Cristina González-Aguilera, Ana Fernández-Miñán, Javier Vázquez-Marín, Helena Parra-Acero, Joe W Cross, Peter WJ Rigby, Jaime J Carvajal, Joachim Wittbrodt, José L Gómez-Skarmeta, et al. Comparative epigenomics in distantly related teleost species identifies conserved cis-regulatory nodes active during the vertebrate phylotypic period. *Genome research*, 24(7):1075–1085, 2014.

[39] Sophie Pantalacci and Marie Sémon. Transcriptomics of developing embryos and organs: a raising tool for evo–devo. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, 324(4):363–371, 2015.

[40] Hajk-Georg Drost, Alexander Gabel, Ivo Grosse, and Marcel Quint. Evidence for active maintenance of phylotranscriptomic hourglass patterns in animal and plant embryogenesis. *Molecular biology and evolution*, 32(5):1221–1231, 2015.

[41] Tamar Friedlander, Avraham E Mayo, Tsvi Tlusty, and Uri Alon. Evolution of bow-tie architectures in biology. *PLoS Comput Biol*, 11(3):e1004055, 2015.

[42] Juan R Martinez-Morales. Toward understanding the evolution of vertebrate gene regulatory networks: comparative genomics and epigenomic approaches. *Briefings in functional genomics*, page elv032, 2015.

[43] Michal Levin, Leon Anavy, Alison G Cole, Eitan Winter, Natalia Mostov, Sally Khair, Naftalie Senderovich, Ekaterina Kovalev, David H Silver, Martin Feder, et al. The mid-developmental transition and the evolution of animal body plans. *Nature*, 2016.

[44] Ozren Bogdanović, Arne H Smits, Elisa de la Calle Mustienes, Juan J Tena, Ethan Ford, Ruth Williams, Upeka Senanayake, Matthew D Schultz, Saartje Hontelez, Ila van Kruijsbergen, et al. Active dna demethylation at enhancers during the vertebrate phylotypic period. *Nature genetics*, 2016.

[45] Marcel Quint, Hajk-Georg Drost, Alexander Gabel, Kristian K Ullrich, Markus Bönn, and Ivo Grosse. A transcriptomic hourglass in plant embryogenesis. *Nature*, 490:98–101, 2012.

[46] Xuanjin Cheng, Jerome Ho Lam Hui, Yung Yung Lee, Patrick Tik Wan Law, and Hoi Shan Kwan. A "developmental hourglass" in fungi. *Molecular biology and evolution*, page msv047, 2015.

[47] John A. Capra, Maureen Stolzer, Dannie Durand, and Katherine S. Pollard. How old is my gene? *Trends in Genetics*, 29(11):659–668, 2013.

[48] Joseph Felsenstein. Theoretical evolutionary genetics. 2013.

[49] Tomislav Domazet-Lošo et al. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *TRENDS in Genetics*, 23(11):533–539, Oktober 2007.

[50] Alexander Gabel. Phylostratigraphy analysis of the arabdopsis thaliana genome. *Bachelor's thesis*, 2011.

[51] Toni Gabaldon and Eugene Koonin. Functional and evolutionary implications of gene orthology. *Nature Reviews Genetics*, 14(5):360–366, 2013.

[52] Alison C Cullen and Christopher Frey. *Probabilistic techniques in exposure assessment: a handbook for dealing with variability and uncertainty in models and inputs.* Springer Science & Business Media, 1999.

[53] Marie Laure Delignette-Muller, Christophe Dutang, Regis Pouillot, Jean-Baptiste Denis, and Maintainer Marie Laure Delignette-Muller. Package 'fitdistrplus'. 2015.

[54] Hubert W Lilliefors. On the kolmogorov-smirnov test for the exponential distribution with mean unknown. *Journal of the American Statistical Association*, 64(325):387–389, 1969.

[55] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016.

[56] John Chambers. *Software for data analysis: programming with R*. Springer Science & Business Media, 2008.

[57] Dirk Eddelbuettel, Romain François, J Allaire, John Chambers, Douglas Bates, and Kevin Ushey. Rcpp: Seamless r and c++ integration. *Journal of Statistical Software*, 40(8):1–18, 2011.

[58] Revolution Analytics and Steve Weston. doparallel: Foreach parallel adaptor for the parallel package. *R package version*, 1(8), 2014.

[59] Sarah Scharfenberg. Pipeline for computing synonymous and nonsynonymous substitutions rates. *Bachelor's thesis*, Martin Luther University Halle, 2012.

[60] Karl Ernst von Baer. *Entwicklungsgeschichte der Thiere: Beobachtung und Reflexion (Vol. 1)*. Königsberg, Bornträger, 1828.

[61] Brian K Hall. *Evolutionary Developmental Biology*. Kluwer Academic Publishers, Dordrecht, 2 edition, 1999.

[62] Gavin De Beer et al. *Embryos and ancestors*. Clarendon Press, Oxford, 1940.

[63] Sabine Brauckmann. Karl ernst von baer (1792-1876) and evolution. *International Journal of Developmental Biology*, 56:653–660, 2012.

[64] Arhat Abzhanov. Von baer's law for the ages: lost and found principles of developmental evolution. *Trends in Genetics*, 29(12):712–722, 2013.

[65] Frank R Lillie. *The development of the chick*. Holt, New York, 1919.

[66] Hajk-Georg Drost, Julia Bellstaedt, Diarmuid S Ó'Maoiléidigh, Anderson T Silva, Alexander Gabel, Claus Weinholdt, Patrick T Ryan, Bas J W Dekkers, Henk W M Hilhorst, Wilco Ligterink, Frank Wellmer, Ivo Grosse, and Marcel Quint. Post-embryonic Hourglass Patterns Mark Ontogenetic Transitions in Plant Development. *Mol. Biol. Evol.*, 33(5):1158–1163, 2016.

[67] Donald T Anderson. *Embryology and Phylogeny in Annelids and Arthropods*. Pergamon Press, Oxford, 1973.

[68] In TJ Horder, JA Witkowski, and CC Wylie, editors, *A History of Embryology*, The Eighth Symposium of The British Society for Developmental Biology, Cambridge, 1986. Cambridge University Press.

[69] Jonathan MW Slack. Phylotype and zootype. *Keywords and Concepts in Evolutionary Developmental Biology*, pages 309–318, 2003.

[70] Scott F Gilbert and David Epel. *Ecological Developmental Biology*. Sinauer Associates, Inc., Sunderland, MA, USA, 2009.

[71] Eric H Davidson. *Genomic Regulatory Systems: Development and Evolution*. Academic Press, San Diego, 2001.

[72] Michael Levine and Eric H Davidson. Gene regulatory networks for development. *PNAS*, 102(14):4936–4942, 2005.

[73] Eric H Davidson and DH Erwin. Gene regulatory networks and the evolution of animal body plans. *Science*, 311:796–800, 2006.

[74] Benjamin Prud'homme et al. Repeated morphological evolution through cis-regulatory changes in a pleiotropic gene. *Nature*, 440:1050–1053, 2006.

[75] Douglas H Erwin and Eric H Davidson. The evolution of hierarchical gene regulatory networks. *Nature Reviews Genetics*, 10:141–148, 2009.

[76] Veronica F Hinman et al. Evolution of gene regulatory network architectures: Examples of subcircuit conservation and plasticity between classes of echinoderms. *Biochimica et Biophysica Acta*, 1789:326–332, 2009.

[77] Eric H Davidson. Emerging properties of animal gene regulatory networks. *Nature*, 468(7326):911–920, 2010.

[78] Isabelle S Peter and Eric H Davidson. Evolution of gene regulatory networks controlling body plan development. *Cell*, 144:970 – 985, 2011.

[79] Colm J Ryan et al. Hierarchical modularity and the evolution of genetic interactomes across species. *Molecular Cell*, 46:691 – 704, 2012.

[80] Eric van Otterloo et al. Gene regulatory evolution and the origin of macroevolutionary novelties: Insights from the neural crest. *Genesis*, 51:457 – 470, 2013.

[81] Benedikt Hallgrímsson and Brian K Hall. *Epigenetics: linking genotype and phenotype in development and evolution*. Univ of California Press, 2011.

[82] Eugenia M del Pino and Richard P Elinson. A novel development pattern for frogs: Gastrulation produces an embryonic disk. *Nature*, 306:589 – 91, 1983.

[83] Wolfgang Dohle and Gerhard Scholtz. Clonal analysis of the crustacean segment: The discordance between genealogical and segmental borders. *Development*, 104:Suppl., 147 – 160, 1988.

[84] Wolfgang Dohle. Differences in cell pattern formation in early embryology and their bearing on evolutionary changes in morphology. *Geobios. Mém. Spécial*, 12:145 – 155, 1989.

[85] Brian K Hall. *Evolutionary developmental biology.* Springer Science & Business Media, 2012.

[86] Friedrich Seidel. Körpergrundgestalt und keimstruktur eine erörterung über die grundlagen der vergleichenden und experimentellen embryologie und deren gültigkeit bei phylogenetischen überlegungen. *Zool Anz*, 164:245–305, 1960.

[87] Klaus Sander. Specification of the basic body pattern in insect embryogenesis. *Adv. Insect Physiol.*, 12:125 – 238, 1976.

[88] Jack Cohen. *Reproduction.* Butterworth Co Publishers Ltd, 1977.

[89] Hans Spemann. Zur geschichte und kritik des begriffs der homologie. In C Chun and W Johannsen, editors, *Allgemeine Biologie*, pages 63 – 86. Teubner, Leipzig, 1915.

[90] Waldemar Schleip. *Die Determination der Primitiventwicklung.* Akademische Verlags-Gesellschaft, Leipzig, 1929.

[91] William W Ballard. Morphogenetic movements and fate maps of vertebrates. *American Zoologist*, 21(2):391–399, 1981.

[92] Charles B Kimmel et al. Stages of embryonic development of the zebrafish. *Developmental dynamics*, 203(3):253–310, 1995.

[93] Ernst HPA Haeckel. *Anthropogenie oder Entwickelungsgeschichte des Menschen: gemeinverständliche wissenschaftliche Vorträge über die Grundzüge der menschlichen Keimes-und Stammes-Geschichte.* Wilhelm Engelmann, 1874.

[94] Lewis Wolpert. *The triumph of the embryo.* Oxford University Press, Oxford, 1991.

[95] Jonathan MW Slack et al. The zootype and the phylotypic stage. *Nature*, 361:490–492, February 1993.

[96] Frietson Galis and Johan AJ Metz. Testing the vulnerability of the phylotypic stage: On modularity and evolutionary conservation. *JOURNAL OF EXPERIMENTAL ZOOLOGY*, 291:195–204, 2001.

[97] Rudolf A Raff et al. Implications of radical evolutionary changes in early development for concepts of developmental constraint. *New perspectives on evolution*, pages 189–207, 1991.

[98] Adam Sedgewick. On the law of development commonly known as von baer's law; and on the significance of ancestral rudiments in embryonic development. *Quarterly Journal of Microscopical Science*, 2(141):35–52, 1894.

[99] Julien Roux, Marta Rosikiewicz, and Marc Robinson-Rechavi. What to compare and how: Comparative transcriptomics for evo-devo. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, 324(4):372–382, 2015.

[100] Andrew G Cridge, Peter K Dearden, and Lynette R Brownfield. Convergent occurrence of the developmental hourglass in plant and animal embryogenesis? *Annals of botany*, 117(5):833–843, 2016.

[101] Michael K Richardson. Vertebrate evolution: the developmental origins of adult variation. *BioEssays*, 21(7):604–613, 1999.

[102] Olaf Bininda-Emonds et al. Inverting the hourglass: quantitative evidence against the phylotypic stage in vertebrate development. *Proceedings of the Royal Society of London B: Biological Sciences*, 270(1513):341–346, 2003.

[103] Maria Ninova, Matthew Ronshaugen, and Sam Griffiths-Jones. Fast-evolving micrornas are highly expressed in the early embryo of drosophila virilis. *RNA*, 20(3):360–372, 2014.

[104] Alex S Nord, Matthew J Blow, Catia Attanasio, Jennifer A Akiyama, Amy Holt, Roya Hosseini, Sengthavy Phouanenavong, Ingrid Plajzer-Frick, Malak Shoukry, Veena Afzal, et al. Rapid and pervasive changes in genome-wide enhancer usage during mammalian development. *Cell*, 155(7):1521–1531, 2013.

[105] Alexander Gabel. Development of a simulated annealing algorithm for uncovering the phylotranscriptomic hourglass pattern. Master's thesis, Martin Luther University Halle, 2013.

[106] Donald R Kaplan and Todd J Cooke. Fundamental concepts in the embryogenesis of dicotyledons: A morphological lnterpretation of embryo mutants. *The Plant Cell*, 9:1903–1919, 1997.

[107] Elliot M Meyerowitz. Plants compared to animals: The broadest comparative study of development. *Science*, 295:1482–1485, February 2002.

[108] Bas JW Dekkers, Simon Pearce, RP van Bolderen-Veldkamp, Alex Marshall, Paweł Widera, James Gilbert, Hajk-Georg Drost, George W Bassel, Kerstin Müller, John R King, et al. Transcriptional dynamics of two seed compartments with opposing roles in arabidopsis seed germination. *Plant physiology*, 163(1):205–215, 2013.

[109] Anderson Tadeu Silva, Pamela A Ribone, Raquel Lia Chan, Wilco Ligterink, and Henk WM Hilhorst. A predictive co-expression network identifies novel genes controlling the seed-to-seedling phase transition in arabidopsis thaliana. *Plant physiology*, pages pp–01704, 2016.

[110] Patrick T Ryan, Diarmuid S Ó'Maoiléidigh, Hajk-Georg Drost, Kamila Kwaśniewska, Alexander Gabel, Ivo Grosse, Emmanuelle Graciet, Marcel Quint, and Frank Wellmer. Patterns of gene expression during arabidopsis flower development from the time of initiation to maturation. *BMC genomics*, 16(1):488, 2015.

[111] Robert B Goldberg, Genaro De Paiva, Ramin Yadegari, et al. Plant embryogenesis: zygote to seed. *SCIENCE-NEW YORK THEN WASHINGTON-*, pages 605–605, 1994.

[112] Colette A. ten Hove, Kuan-Ju Lu, and Dolf Weijers. Building a plant: cell fate specification in the early arabidopsis embryo. *Development*, 142(3):420–430, 2015.

[113] Günter Theißen and Rainer Melzer. Robust views on plasticity and biodiversity. *Annals of Botany*, 117(5):693–697, 2016.

[114] David Meinke et al. Identifying essential genes in arabidopsis thaliana. *Trends in Plant Science*, 13(9):483–491, 2008.

6 Paper 1: Evidence for Active Maintenance of Phylotranscriptomic Hourglass Patterns in Animal and Plant Embryogenesis

HG Drost, A Gabel, I Grosse, M Quint. 2015. Evidence for Active Maintenance of Phylotranscriptomic Hourglass Patterns in Animal and Plant Embryogenesis. *Mol. Biol. Evol.* 32 (5): 1221-1231. *doi:10.1093/molbev/msv012*

Evidence for Active Maintenance of Phylotranscriptomic Hourglass Patterns in Animal and Plant Embryogenesis

Hajk-Georg Drost,[1] Alexander Gabel,[1] Ivo Grosse,*[,1,2] and Marcel Quint*[,3]
[1]Institute of Computer Science, Martin Luther University Halle-Wittenberg, Halle (Saale), Germany
[2]German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany
[3]Department of Molecular Signal Processing, Leibniz Institute of Plant Biochemistry, Halle (Saale), Germany
***Corresponding author:** E-mail: ivo.grosse@informatik.uni-halle.de; mquint@ipb-halle.de.

Associate editor: John True

Abstract

The developmental hourglass model has been used to describe the morphological transitions of related species throughout embryogenesis. Recently, quantifiable approaches combining transcriptomic and evolutionary information provided novel evidence for the presence of a phylotranscriptomic hourglass pattern across kingdoms. As its biological function is unknown it remains speculative whether this pattern is functional or merely represents a nonfunctional evolutionary relic. The latter would seriously hamper future experimental approaches designed to test hypotheses regarding its function. Here, we address this question by generating transcriptome divergence index (TDI) profiles across embryogenesis of *Danio rerio*, *Drosophila melanogaster*, and *Arabidopsis thaliana*. To enable meaningful evaluation of the resulting patterns, we develop a statistical test that specifically assesses potential hourglass patterns. Based on this objective measure we find that two of these profiles follow a statistically significant hourglass pattern with the most conserved transcriptomes in the phylotypic periods. As the TDI considers only recent evolutionary signals, this indicates that the phylotranscriptomic hourglass pattern is not a rudiment but possibly actively maintained, implicating the existence of some linked biological function associated with embryogenesis in extant species.

Key words: **evo-devo, developmental hourglass, embryogenesis, phylotranscriptomics.**

Downloaded from http://mbe.oxfordjournals.org/ by guest on July 4, 2016

Introduction

Embryogenesis coordinates the transformation of a single fertilized egg cell into a differentiated, complex organism. Based on von Baer's third law of embryology (1828), it has been observed that embryos of animal species from the same phylum share a developmental stage with apparent morphological similarities. Animal embryos from the same phylum often appear morphologically different in early embryogenesis, converge to a similar form during mid-embryogenesis, and diverge again in late embryogenesis. This morphological pattern is known as the developmental hourglass pattern (Duboule 1994; Raff 1996), and the stage or period of maximum morphological conservation in mid-embryogenesis is called phylotypic stage (Sander 1983) or phylotypic period (Richardson 1995).

Recently, several groups succeeded in providing a possible explanation for the morphological hourglass pattern in animals by observing an hourglass pattern also at the transcriptome level. Distance-based comparisons of transcriptomes of related species (Kalinka et al. 2010; Irie and Kuratani 2011; Levin et al. 2012) or transcriptome indices based on the combination of evolutionary with transcriptomic information of a single species (Domazet-Lošo and Tautz 2010; Quint et al. 2012) revealed the existence of transcriptomic hourglass patterns in different lineages including even plants. The latter is particularly remarkable because the last common ancestor of animals and plants was most likely unicellular, meaning that both multicellularity and embryogenesis evolved independently in both kingdoms (Meyerowitz 2002). As a consequence, phylotranscriptomic hourglass patterns associated with embryogenesis in animals and plants likely represent an example of convergent evolution.

Distance-based transcriptome comparisons are well established and measure the dissimilarity or distance of expression profiles of orthologous genes among related species. By following this approach, it was found that transcriptomes of *Drosophila* (Kalinka et al. 2010), *Caenorhabditis* (Levin et al. 2012), and turtle (Wang et al. 2013) ssp. are more similar during the morphological phylotypic period in mid-embryogenesis than transcriptomes in early or late embryogenesis. These findings were recently supported by an independent metazoan cross-phyla approach (Gerstein et al. 2014).

While distance-based transcriptome comparisons require transcriptomic information of at least two species or genotypes, transcriptome indices require such information for only a single genotype. Here, evolutionary information of a gene such as phylogenetic age or sequence divergence is combined with its expression level for the computation of transcriptome indices such as the transcriptome age index (TAI, Domazet-Lošo and Tautz 2010) or the transcriptome divergence index (TDI, Quint et al. 2012).

The TAI is based on phylostratigraphy (Domazet-Lošo et al. 2007), which assigns a phylogenetic age to each

© The Author 2015. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Open Access

protein-coding gene in a species of interest by identification of homologous sequences in other species. Following this procedure, genes can be sorted into discrete age categories, named phylostrata (PS), corresponding to hierarchically ordered phylogenetic nodes along the tree of life. The phylogenetic age of each gene quantified by its PS is then weighted by its expression level. The weighted mean of all gene ages yields the TAI (Domazet-Lošo and Tautz 2010), which represents the mean evolutionary age of a transcriptome. As gene age can date back to times before the divergence of prokaryotes and eukaryotes, the TAI incorporates both evolutionarily ancient and recent signals.

The TDI is based on the sequence divergence of protein-coding genes (Ka/Ks ratio) as an indicator of selective pressure (Quint et al. 2012). In analogy to PS, genes can be sorted into discrete sequence divergence categories, named divergence strata (DS), ranging from purifying to positive selection. In analogy to the TAI, the sequence divergence of each gene quantified by its DS is weighted by its expression level, and the weighted mean of all sequence divergences yields the TDI, which represents the mean sequence divergence of a transcriptome. In contrast to the TAI, the TDI focuses on recent evolution among related species. To be more precise, the evolutionary time span investigated by the TDI reaches from “today” to the time when the two selected species split. Depending on the chosen species, this may be as little as a few million years. Hence, distance-based transcriptome comparisons and transcriptome indices such as TAI or TDI quantify different evolutionary properties of one or several transcriptomes.

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 (Raff 1996; Kalinka and Tomancak 2012). 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, we study gene ages and TAI profiles as well as sequence divergences and TDI profiles of the vertebrate *Danio rerio*, the invertebrate *Drosophila melanogaster*, and the flowering plant *Arabidopsis 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.

Results

In the context of the developmental hourglass, morphological and—as we define them—phylotranscriptomic patterns have to be distinguished. This study addresses phylotranscriptomic patterns, which can be divided in distance-based transcriptome comparisons and transcriptome index-based approaches. Transcriptome indices, which are the subject of this work, can again be computed as either TAI or TDI.

Although distance-based transcriptome comparisons show that mid-embryonic stages have a lower gene expression diversity than early and late stages of embryonic development (reviewed in Kalinka and Tomancak 2012), the developmental hourglass model is still controversially discussed. TAI analyses have to date been performed for *Da. rerio*, *D. melanogaster*, *Anopheles gambiae*, and *A. thaliana* (Domazet-Lošo and Tautz 2010; Quint et al. 2012). The results largely confirmed the observations from distance-based transcriptome comparisons in that they identified the most ancient transcriptome during mid-embryogenesis. However, these previous analyses are hardly comparable because they were computed 1) with different genome databases, 2) with different analysis pipelines, and 3) in the case of *D. melanogaster* only for approximately one-quarter of the genes.

To allow for optimal comparability of TAI patterns across species, we here reanalyze TAI profiles of embryonic development of *Da. rerio*, *D. melanogaster*, and *A. thaliana* in a consistent manner based on the same sets of genomes, the same pipeline, and updated phylostratigraphic maps. For *D. melanogaster* we use whole transcriptome expression data (Graveley et al. 2011) instead of the previously used data set that consisted of only 3,550 genes (Arbeitman et al. 2002). Based on the obtained TAI patterns, we will then turn our attention to the TDI and this study’s central question of whether or not the evolutionary signal that shaped the hourglass pattern might be actively maintained.

Downloaded from http://mbe.oxfordjournals.org/ by guest on July 4, 2016

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

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., Quint et al. 2012) 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) (supplementary fig. S1 and table S1, Supplementary Material online, database available for download at http://msbi.ipb-halle.de/download/phyloBlastDB_Drost_Gabel_Grosse_Quint.fa.tbz, last accessed August 2, 2015). 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 1*A*–*C* (supplementary table S2, Supplementary Material online).

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 2 (expression values provided in supplementary table S3, Supplementary Material online). 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* (Quint et al. 2012). 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. 2). This is unexpected and in contrast to comparative transcriptomic approaches, which consistently identified highly divergent transcriptomes in early *Drosophila* embryogenesis (Kalinka et al. 2010; Gerstein et al. 2014). However, we hesitate to overinterpret this observation because the overall profile still resembles an hourglass pattern.

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. 1), 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 (Galis and Metz 2001; Hazkani-Covo et al. 2005; Davidson and Erwin 2009; Domazet-Lošo and Tautz 2010; He and Deem 2010; Kalinka et al. 2010; Irie and Kuratani 2011; Peter and Davidson 2011; Levin et al. 2012; Quint et al. 2012; de Mendoza et al. 2013; Piasecka et al. 2013; Schep and Adryan 2013; Wang et al. 2013).

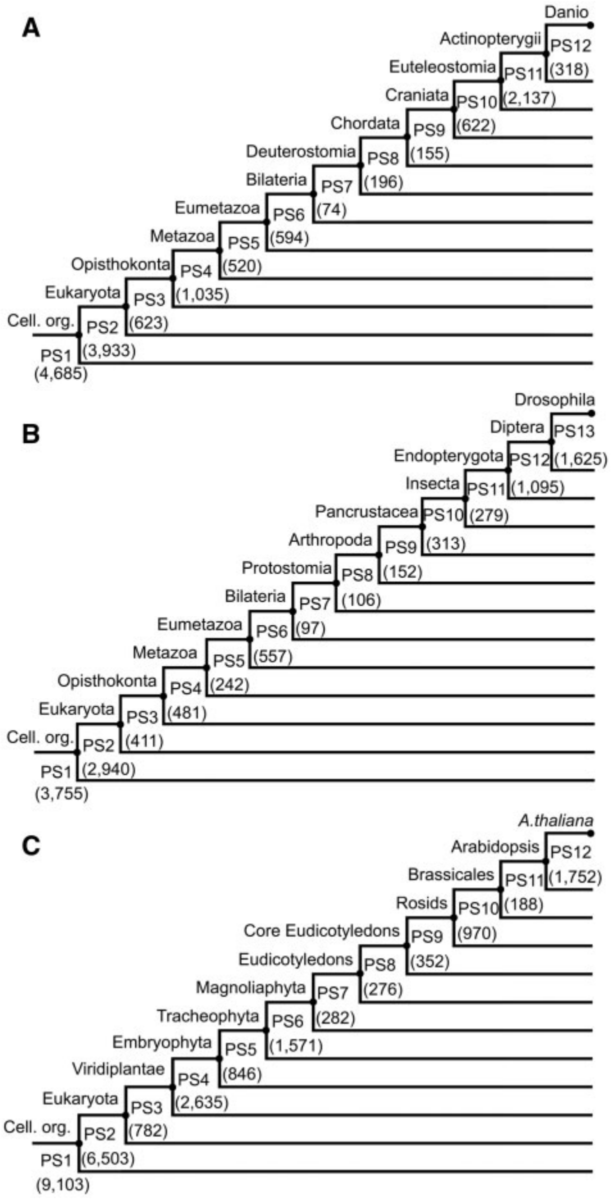


Fig. 1. Phylostratigraphic maps for *Danio rerio*, *Drosophila melanogaster*, and *Arabidopsis thaliana*. (*A*) *Danio rerio*. (*B*) *Drosophila melanogaster*. (*C*) *Arabidopsis thaliana*. Numbers in parenthesis denote the number of genes per phylostratum (PS1–PS12/13). Cell. org., cellular organisms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

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

Downloaded from http://mbe.oxfordjournals.org/ by guest on July 4, 2016

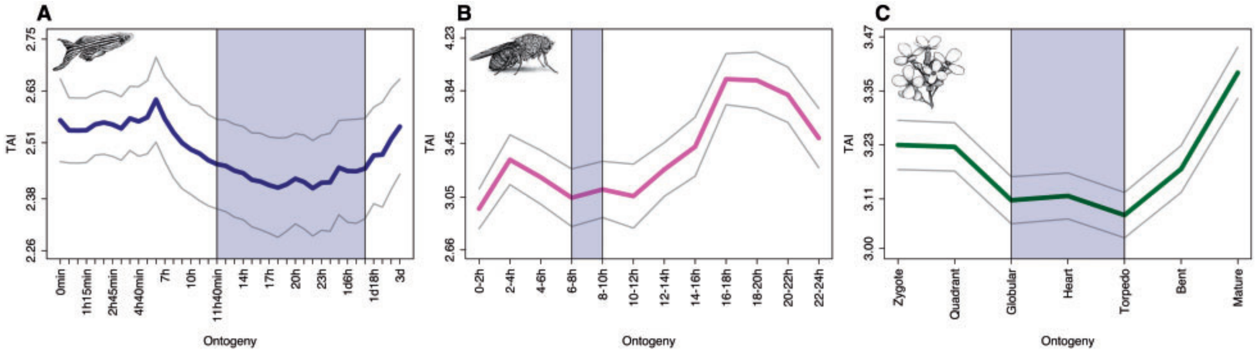


Fig. 2. TAI profiles across animal and plant embryogenesis. (*A*) *Danio rerio*. (*B*) *Drosophila melanogaster*. (*C*) *Arabidopsis thaliana*. The blue shaded area marks the predicted phylotypic period. The gray lines represent the standard deviation estimated by permutation analysis.

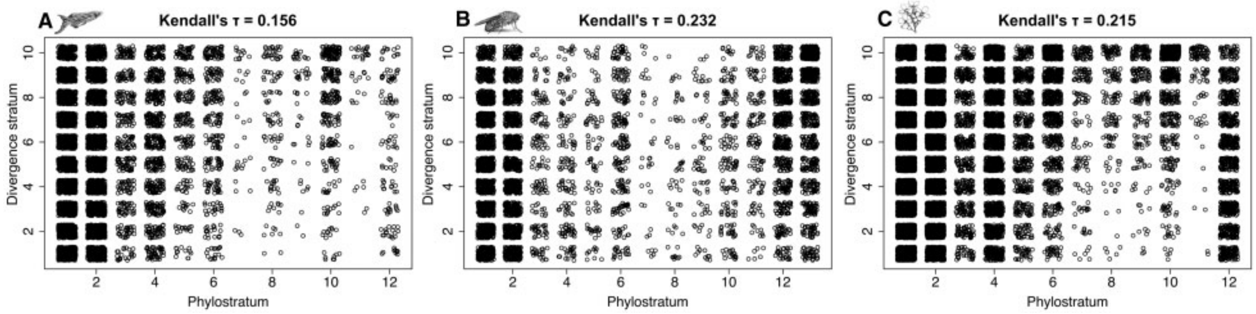


Fig. 3. Correlation between phylostratum (PS) and divergence stratum (DS). Scatter plots of phylostratum versus divergence stratum over all genes. (*A*) *Danio rerio*. (*B*) *Drosophila melanogaster*. (*C*) *Arabidopsis thaliana*. Ka /Ks ratios for divergence stratum assignment are derived from orthologous 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 ($P < 2.2\text{e-}16$) using Kendall's τ test of no correlation.

conclusions that cannot be drawn based on TAI profiles alone.

For computing DS in analogy to PS, we generate orthologous gene sets for the computation of sequence divergences (Ka/Ks) 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, Supplementary Material online).

For *Da. rerio* closely related fish genomes are not yet available. Here, we use *Astyanax mexicanus* (divergence time~153 Ma, Hedges et al. 2006), *Takifugu rubripes*, *Xiphophorus maculatus*, and *Gadus morhua* (divergence time for all three species ~265 Ma, Hedges et al. 2006). For the assignment of Ka/Ks values of *D. melanogaster* genes, we compare its coding genome to *D. simulans* (divergence time ~3 Ma, Hedges et al. 2006), *D. yakuba* (divergence time ~7 Ma, Hedges et al. 2006) *D. persimillis* (divergence time ~34 Ma, Hedges et al. 2006), and *D. virilis* (divergence time ~47 Ma, Hedges et al. 2006). For *A. thaliana* we use the Brassicas *A. lyrata* (divergence time ~5–10 Ma, Hu et al. 2011), *Capsella rubella* (divergence time ~10–14 Ma, Koch and Kiefer 2005), *Brassica rapa* (divergence time ~16 Ma, Hedges et al. 2006), and *Carica papaya* (divergence time ~72 Ma, Hedges et al. 2006). For each pairwise comparison we sort the continuous Ka/Ks 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, Supplementary Material online).

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 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*A*–*C*; supplementary tables S2 and S4 and figs. S5–S7, Supplementary Material online, for the additional species comparisons), stating that TAI and TDI have the potential of capturing independent evolutionary signals for all three species.

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

Next, we finally investigate whether or not the evolutionary selection pressure that has shaped the hourglass pattern

Downloaded from http://mbe.oxfordjournals.org/ by guest on July 4, 2016

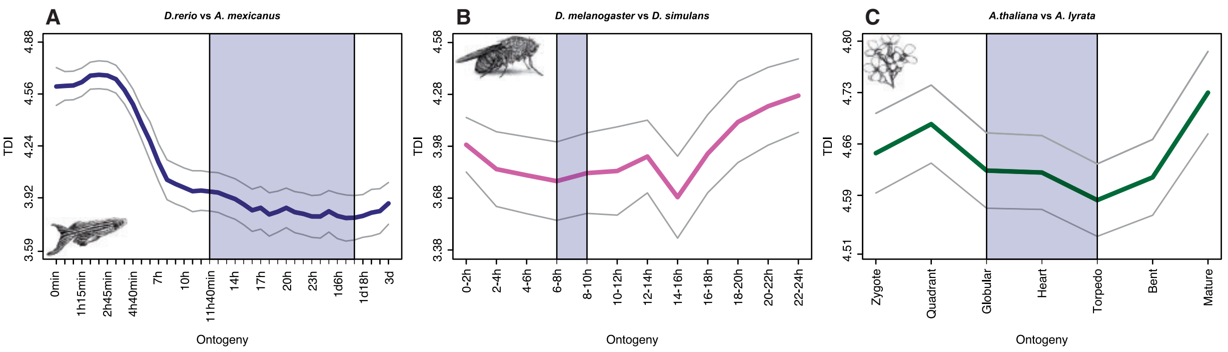


Fig. 4. TDI profiles across animal and plant embryogenesis. (*A*) *Danio rerio*. (*B*) *Drosophila melanogaster*. (*C*) *Arabidopsis thaliana*. The blue shaded area marks the predicted phylotypic period. The gray lines represent the standard deviation estimated by permutation analysis.

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 4 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, Supplementary Material online).

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.

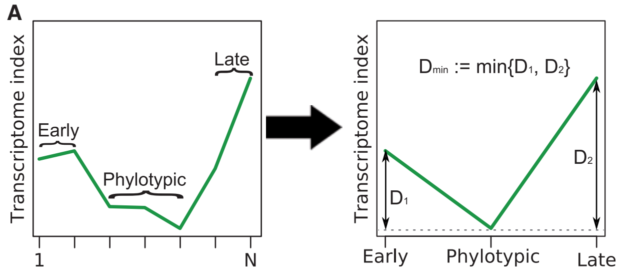
Objective Testing for Potential Hourglass Patterns

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 (de Mendoza et al. 2013; Piasecka et al. 2013), have tested whether the observed profile deviated from a horizontal line (Domazet-Lošo and Tautz 2010; Irie and Kuratani 2011; Quint et al. 2012; Wang et al. 2013, figs. 2 and 4 this study), or have tested whether the observed profile could be fitted by a parabolic function (Hazkani-Covo et al. 2005; Kalinka et al. 2010; Levin et al. 2012).

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. 5*A*). 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, Supplementary Material online). Based on these observations, we regard the developmental period encompassing globular, heart, and torpedo embryos as the most reasonable choice for designating the

Downloaded from http://mbe.oxfordjournals.org/ by guest on July 4, 2016



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-1	1.0
TDI	1.0	6.7e-1	9.1e-1

Fig. 5. 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.

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. 5*A*, see Materials and Methods). 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. 5*B*). 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 My), 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, Supplementary Material online).

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 (see Materials and Methods), 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. 5*C*), 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.

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 (Hall 1997; Richardson et al. 1997; Richardson 1999; Bininda-Emonds et al. 2003; Roux and Robinson-Rechavi 2008; Comte et al. 2010). However, technological progress recently facilitated quantitative measurements of expression profiles. Although some of these recent studies favored the early conservation model (Roux and Robinson-Rechavi 2008; Comte et al. 2010), 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 (Davis et al. 2005; Hazkani-Covo et al. 2005; Demuth et al. 2006; Irie and Sehara-Fujisawa 2007; Cruickshank and Wade 2008). Later, several studies demonstrated that whole transcriptomes of fly, worm, several vertebrates, and cress followed an hourglass pattern (Domazet-Lošo and Tautz 2010; Kalinka et al. 2010; Irie and Kuratani 2011; Levin et al. 2012; Quint et al. 2012; Wang et al. 2013). 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 (Ninova et al. 2014).

The later phylotranscriptomic studies have been performed by distance-based transcriptome comparisons (Kalinka et al. 2010; Irie and Kuratani 2011; Levin et al. 2012; Wang et al. 2013) or by studies of transcriptome indices (Domazet-Lošo and Tautz 2010; Quint et al. 2012); 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 (Domazet-Lošo and Tautz 2010) 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 (Quint et al. 2012) 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

Downloaded from http://mbe.oxfordjournals.org/ by guest on July 4, 2016

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), 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 Kalinka et al. (2010) 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 (Kaplan and Cooke 1997). 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.

Conclusions

The existence of hourglass patterns in TAI profiles of animal and plant embryogenesis demonstrates that this pattern is not a recent innovation. Darwin (1859) 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,

Downloaded from http://mbe.oxfordjournals.org/ by guest on July 4, 2016

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.

Materials and Methods

Scripts for complete reproduction of all data presented in this manuscript including database generation, construction of phylostratigraphic and sequence divergence maps, computation of TAI and TDI patterns, essential gene analysis, and statistical tests are available via the GitHub repository (https://github.com/HajkD/Active-maintenance-of-phylotranscriptomic-hourglasses, last accessed August 2, 2015). Detailed instructions for applications of the same analyses to any expression data set and any species with sufficient genome information can be found in the R packages myTAI (Drost 2014) and orthologr (https://github.com/HajkD/orthologr, last accessed August 2, 2015).

Construction of Phylostratigraphic Maps

Procedures for constructing phylostratigraphic maps have been presented previously (Domazet-Lošo et al. 2007; Domazet-Lošo and Tautz 2010; Quint et al. 2012). Here, we construct phylostratigraphic maps of *Da. rerio*, *D. melanogaster*, and *A. thaliana* based on the same data set and the following procedure. First, we define a set of PS for each of the three species according to the NCBI taxonomy database. Second, we extract all 17,582,624 amino acid sequences of all 4,557 species from the NCBI, ENSEMBL (Flicek et al. 2014), Flybase (St. Pierre et al. 2014), and Phytozome (Goodstein et al. 2012) databases. Third, we generate a target database from these sequences (http://msbi.ipb-halle.de/download/phyloBlastDB_Drost_Gabel_Grosse_Quint.fa.tbz, last accessed August 2, 2015) and BLAST 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 against this target database using BLASTp (BLAST version 2.2.21). Fourth, we assign each gene to its PS by the following rule. If no BLAST hit with an *E*-value below 10^{-5} was identified, 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}. PS for the genomes of all three species are given in supplementary table S2, Supplementary Material online.

Construction of Sequence Divergence Maps

We construct sequence 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 (Suyama et al. 2006). Third, we compute Ka/Ks values of the codon alignments using GESTIMATOR (Thornton 2003) and Comeron's substitution model, which combines Li's, Pamillo's, and Bianchi's method with Kimura's method for obtaining robust Ka/Ks estimates (Comeron 1995). Fourth, we discard all genes with a Ka/Ks value greater than 2 and sort the remaining Ka/Ks values into discrete deciles, which we call DS. DS values for the genomes of all three species are provided in supplementary table S4, Supplementary Material online. The same procedure is applied to generate sequence divergence maps for all other pairwise species comparisons (supplementary table S4, Supplementary Material online). The construction of sequence divergence maps is explained in detail in the advanced vignette of the myTAI R package (Drost 2014). 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, last accessed August 2, 2015).

Processing of Expression Data

For *Da. rerio* we use the microarray expression data set by Domazet-Lošo and Tautz (2010) covering 40 stages corresponding to embryo development. The 16,188 probes of this data set correspond to 12,892 genes according to ENSEMBL predictions (Domazet-Lošo and Tautz 2010), and we compute the expression level of each gene as arithmetic mean of the expression levels of the corresponding probes (Piasecka et al. 2013). 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) 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) covering seven stages of embryo development. Intersecting the 26,173

Downloaded from http://mbe.oxfordjournals.org/ by guest on July 4, 2016

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. Expression values used for TAI and TDI computations are provided in supplementary tables S3 and S5, Supplementary Material online. The introductory vignette of the myTAI R package describes how to define and process expression data sets (Drost 2014).

Transcriptome Age Index

The TAI at stage *s* (TAI_s) has been defined as weighted arithmetic mean over all PS using gene expression intensities e_{is} of gene *i* at developmental stage *s* as weights (Domazet-Lošo and Tautz 2010), that is,

$$TAI_s = \frac{\sum_{i=1}^{n} ps_i e_{is}}{\sum_{i=1}^{n} e_{is}},$$

where ps_i denotes the PS of gene *i*, and *n* denotes the number of genes. A small value of ps_i represents an old PS, and a high value of ps_i a young PS. Hence, a small value of TAI_s represents a high mean evolutionary age of the transcriptome at stage *s*, and a high value of TAI_s a low mean evolutionary age. The standard workflow for TAI analysis is described in detail in the introductory vignette of the myTAI R package (Drost 2014).

Transcriptome Divergence Index

The TDI at stage *s* (TDI_s) has originally been defined as weighted arithmetic mean over all sequence divergence values (Ka/Ks) using gene expression intensities e_{is} of gene *i* at developmental stage *s* as weights (Quint et al. 2012). Here, we slightly modify the definition of the TDI by sorting the continuous Ka/Ks values into deciles yielding ten discrete DS. These discrete DS ranging from 1 to 10 represent the degree of sequence divergence in the same manner in which the discrete PS represent the evolutionary age. We now define the TDI of stage *s* (TDI_s) as weighted arithmetic mean over all DS using gene expression intensities e_{is} as weights, that is,

$$TDI_s = \frac{\sum_{i=1}^{n} ds_i e_{is}}{\sum_{i=1}^{n} e_{is}},$$

where ds_i denotes the DS of gene *i*, and *n* denotes the number of genes. A small value of ds_i represents a conserved DS, and a high value of ds_i a divergent DS. Hence, a small value of TDI_s represents a low mean sequence divergence of the transcriptome at stage *s*, and a high value of TDI_s a high mean sequence divergence. The standard workflow for TDI analysis is described in detail in the introductory vignette of the myTAI R package (Drost 2014).

Essential Genes Expression Level Analysis

Essential genes are defined as genes that are required for normal growth and development which are associated with a loss-of-function phenotype in a standard genetic background (Meinke et al. 2008). For our analysis, we focus on genes causing embryo-defective phenotypes in *A. thaliana*. We took unique essential genes from www.seedgenes.org (Meinke et al. 2008) and only selected genes that were classified as embryo-defective. This procedure yielded 401 unique embryo-defective genes that were used to generate supplementary figure S8, Supplementary Material online. Mean expression levels were plotted for each stage (supplementary fig. S8A, Supplementary Material online) and a Dunn's test of multiple comparisons (Dunn 1964) using Benjamini–Hochberg adjustment (Benjamini and Hochberg 1995) was performed to statistically quantify differences in essential gene expression for pairwise stage comparisons (supplementary fig. S8B, Supplementary Material online). Statistical significance of differences in essential gene expression across all stages was assessed by performing a Kruskal–Wallis rank sum test. Gene IDs, expression values and scripts are included in the accompanying GitHub repository.

Flat Line Test

The flat line test (Quint et al. 2012) is a permutation test based on the variance *V* of the TAI values of a given TAI profile as test statistic. For any pattern different from a flat horizontal line, *V* should be high. In order to determine the statistical significance of an observed variance *V*, we perform the following permutation test. We randomly permute the PS values of the original data set 10,000 times, compute the variance *V* from each of the 10,000 permuted data set *s*, approximate the histogram of the 10,000 variances *V* by a Gamma distribution, and report the probability of exceeding the observed variance *V* as *P*-value of the flat line test.

The flat line test can be applied to TDI profiles in exactly the same manner.

Reductive Hourglass Test

The reductive hourglass test is a permutation test based on the following test statistic. First, we partition the set of developmental stages into three modules—early, mid, and late—based on prior biological knowledge. Second, we compute the mean TAI value for each of the three modules, and we denote these mean TAI values by T_{early}, T_{mid}, and T_{late}. Third, we compute the two differences $D_1 = T_{\text{early}} - T_{\text{mid}}$ and $D_2 = T_{\text{late}} - T_{\text{mid}}$. Fourth, we compute the minimum $D_{\min}$ of D_1 and D_2 as final test statistic of the reductive hourglass test.

For a typical hourglass pattern, T_{early} should be high, T_{mid} should be low, 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

Downloaded from http://mbe.oxfordjournals.org/ by guest on July 4, 2016

following permutation test. We randomly permute the PS values of the original data set 10,000 times, compute the minimum difference D_{min} from each of the 10,000 permuted data sets, approximate the histogram of the 10,000 minimum differences D_{min} by a Gaussian distribution, and report the probability of exceeding the observed minimum difference D_{min} as *P*-value of the reductive hourglass test (fig. 5*A*). Supplementary figure S10, Supplementary Material online, visualizes an example test statistic, the corresponding Gaussian distribution fitting the histogram of the 10,000 minimum differences D_{min}, and the hourglass score of the observed phylotranscriptomic pattern.

The reductive hourglass test can be applied to TDI profiles in exactly the same manner.

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 (Drost 2014). The entire process of building the test statistics for the three tests can be found in its intermediate vignette.

Supplementary Material

Supplementary figures S1–S10 and tables S1–S5 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors thank Sarah Scharfenberg and Kristian Ullrich for valuable discussions and their contributions to the Ka/Ks pipeline, and Julia Bellstädt for graphical support. Furthermore, the authors are grateful to NCBI, ENSEMBL, flybase and Phytozome for providing extensive genome information. This work was supported by the Leibniz Association, the Deutsche Forschungsgemeinschaft (Qu 141/5-1 to M.Q. and Gr 3526/6-1 to I.G.) and the SKW Piesteritz research prize awarded to H.-G.D. and A.G.

References

Arbeitman MN, Furlong EE, Imam F, Johnson E, Null BH, Baker BS, Krasnow MA, Scott MP, Davis RW, White KP. 2002. Gene expression during the life cycle of *Drosophila melanogaster*. *Science* 297:2270–2275.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B Stat Methodol.* 57:289–300.

Bininda-Emonds OR, Jeffery JE, Richardson MK. 2003. Inverting the hourglass: quantitative evidence against the phylotypic stage in vertebrate development. *Proc Biol Sci.* 270:341–346.

Comeron JM. 1995. A method for estimating the numbers of synonymous and nonsynonymous substitutions per site. *J Mol Evol.* 41:1152–1159.

Comte A, Roux J, Robinson-Rechavi M. 2010. Molecular signaling in zebrafish development and the vertebrate phylotypic period. *Evol Dev.* 12:144–156.

Cruickshank T, Wade MJ. 2008. Microevolutionary support for a developmental hourglass: gene expression patterns shape sequence variation and divergence in *Drosophila*. *Evol Dev.* 10:583–590.

Darwin C. 1859. On the origin of species. London: Murray.

Davidson EH, Erwin DH. 2009. An integrated view of precambrian eumetazoan evolution. *Cold Spring Harb Symp Quant Biol.* 74:65–80.

Davis JC, Brandman O, Petrov DA. 2005. Protein evolution in the context of *Drosophila* development. *J Mol Evol.* 60:774–785.

de Mendoza A, Sebé-Pedrós A, Sestak MS, Matejcic M, Torruella G, Domazet-Lošo T, Ruiz-Trillo I. 2013. Transcription factor evolution in eukaryotes and the assembly of the regulatory toolkit in multicellular lineages. *Proc Natl Acad Sci U S A.* 110:E4858–E4866.

Demuth JP, De Bie T, Stajich JE, Cristianini N, Hahn MW. 2006. The evolution of mammalian gene families. *PLoS One* 1:e85.

Domazet-Lošo T, Brajković J, Tautz D. 2007. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends Genet.* 23:533–539.

Domazet-Lošo T, Tautz D. 2010. A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature* 468:815–818.

Drost HG. 2014. A framework to perform phylotranscriptomics analyses for evolutionary developmental biology research. R package version 0.0.1. Available from: http://CRAN.R-project.org/package=myTAI.

Duboule D. 1994. Temporal colinearity and the phylotypic progression: a basis for the stability of a vertebrate Bauplan and the evolution of morphologies through heterochrony. *Dev Suppl.* 135–142.

Dunn OJ. 1964. Multiple comparisons using rank sums. *Technometrics* 6:241–252.

Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, et al. 2014. Ensembl 2014. *Nucleic Acids Res.* 42(Database Issue):D749–D755.

Galis F, Metz JA. 2001. Testing the vulnerability of the phylotypic stage: on modularity and evolutionary conservation. *J Exp Zool.* 291:195–204.

Gerstein MB, Rozowsky J, Yan KK, Wang D, Cheng C, Brown JB, Davis CA, Hillier L, Sisu C, Li JJ, et al. 2014. Comparative analysis of the transcriptome across distant species. *Nature* 512:445–448.

Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, et al. 2012. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* 40:D1178–D1186.

Graveley BR, Brooks AN, Carlson JW, Duff MO, Landolin JM, Yang L, Artieri CG, van Baren MJ, Boley N, Booth BW, et al. 2011. The developmental transcriptome of *Drosophila melanogaster*. *Nature* 470:473–479.

Hall BK. 1997. Phylotypic stage or phantom: is there a highly conserved embryonic stage in vertebrates? *Trends Ecol Evol.* 12:461–463.

Hazkani-Covo E, Wool D, Graur D. 2005. In search of the vertebrate phylotypic stage: a molecular examination of the developmental hourglass model and von Baer's third law. *J Exp Zool B Mol Dev Evol.* 304:150–158.

He J, Deem MW. 2010. Hierarchical evolution of animal body plans. *Dev Biol.* 337:157–161.

Hedges SB, Dudley J, Kumar S. 2006. TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics* 22:2971–2972.

Hu TT, Pattyn P, Bakker EG, Cao J, Cheng JF, Clark RM, Fahlgren N, Fawcett JA, Grimwood J, Gundlach H, et al. 2011. The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nat Genet.* 43:476–481.

Downloaded from http://mbe.oxfordjournals.org/ by guest on July 4, 2016

Irie N, Kuratani S. 2011. Comparative transcriptome analysis reveals vertebrate phylotypic period during organogenesis. *Nat Commun.* 2:248.

Irie N, Sehara-Fujisawa A. 2007. The vertebrate phylotypic stage and an early bilaterian-related stage in mouse embryogenesis defined by genomic information. *BMC Biol.* 5:1.

Kalinka AT, Tomancak P. 2012. The evolution of early animal embryos: conservation or divergence? *Trends Ecol Evol.* 27:385–393.

Kalinka AT, Varga KM, Gerrard DT, Preibisch S, Corcoran DL, Jarrells J, Ohler U, Bergman CM, Tomancak P. 2010. Gene expression divergence recapitulates the developmental hourglass model. *Nature* 468: 811–814.

Kaplan DR, Cooke TJ. 1997. Fundamental concepts in the embryogenesis of dicotyledons: a morphological interpretation of embryo mutants. *Plant Cell* 9:1903–1919.

Koch MA, Kiefer M. 2005. Genome evolution among cruciferous plants: a lecture from the comparison of the genetic maps of three diploid species—*Capsella rubella*, *Arabidopsis lyrata subsppetraea*, and *A. thaliana*. *Am J Bot.* 92:761–767.

Levin M, Hashimshony T, Wagner F, Yanai I. 2012. Developmental milestones punctuate gene expression in the *Caenorhabditis* embryo. *Dev Cell* 22:1101–1108.

Meinke D, Muralla R, Sweeney C, Dickerman A. 2008. Identifying essential genes in *Arabidopsis thaliana*. *Trends Plant Sci.* 13:483–491.

Meyerowitz EM. 2002. Plants compared to animals: the broadest comparative study of development. *Science* 295:1482–1485.

Ninova M, Ronshaugen M, Griffiths-Jones S. 2014. Conserved temporal patterns of microRNA expression in *Drosophila* support a developmental hourglass model. *Genome Biol Evol.* 6:2459–2467.

Peter IS, Davidson EH. 2011. Evolution of gene regulatory networks controlling body plan development. *Cell* 144:970–984.

Piasecka B, Lichocki P, Moretti S, Bergmann S, Robinson-Rechavi M. 2013. The hourglass and the early conservation models—co-existing patterns of developmental constraints in vertebrates. *PLoS Genet.* 9: e1003476.

Quint M, Drost HG, Gabel A, Ullrich KK, Boenn M, Grosse I. 2012. A transcriptomic hourglass in plant embryogenesis. *Nature* 490: 98–101.

Raff RA. 1996. The shape of life: genes, development and the evolution of animal form. Chicago: University Chicago Press.

Richardson MK. 1995. Heterochrony and the phylotypic period. *Dev Biol.* 172:412–421.

Richardson MK. 1999. Vertebrate evolution: the developmental origins of adult variation. *Bioessays* 21:604–613.

Richardson MK, Hanken J, Gooneratne ML, Pieau C, Raynaud A, Selwood L, Wright GM. 1997. There is no highly conserved embryonic stage in the vertebrates: implications for current theories of evolution and development. *Anat Embryol.* 196:91–106.

Roux J, Robinson-Rechavi M. 2008. Developmental constraints on vertebrate genome evolution. *PLoS Genet.* 4:e1000311.

Sander K. 1983. The evolution of patterning mechanisms: gleanings from insect embryogenesis and spermatogenesis. In: Goodwin BC, Holder N, Wylie C, editors. Development and evolution, The Sixth Symposium of the British Society for Developmental Biology. Cambridge: Cambridge University Press. p. 137–160.

Schep AN, Adryan B. 2013. a comparative analysis of transcription factor expression during metazoan embryonic development. *PLoS One* 8: e66826.

St. Pierre SE, Ponting L, Stefancsik R, McQuilton P, FlyBase Consortium. 2014. FlyBase 102—advanced approaches to interrogating FlyBase. *Nucleic Acids Res.* 42:D780–D788.

Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* 34:W609–W612.

Thornton K. 2003. Libsequence: a C11 class library for evolutionary genetic analysis. *Bioinformatics* 19:2325–2327.

von Baer KE 1828. Über Entwicklungsgeschichte der Thiere: Beobachtung und Reflexion. Konigsberg: Gebrüder Bornträger.

Wang Z, Pascual-Anaya J, Zadissa A, Li W, Niimura Y, Huang Z, Li C, White S, Xiong Z, Fang D, et al. 2013. The draft genomes of soft-shell turtle and green sea turtle yield insights into the development and evolution of the turtle-specific body plan. *Nat Genet.* 45:701–706.

Xiang D, Venglat P, Tibiche C, Yang H, Risseeuw E, Cao Y, Babic V, Cloutier M, Keller W, Wang E, et al. 2011. Genome-wide analysis reveals gene expression and metabolic network dynamics during embryo development in *Arabidopsis*. *Plant Physiol.* 156:346–356.

Downloaded from http://mbe.oxfordjournals.org/ by guest on July 4, 2016

7 Paper 2: Post-embryonic Hourglass Patterns Mark Ontogenetic Transitions in Plant Development

HG Drost, J Bellstaedt, DS Ó'Maoiléidigh, AT Silva, A Gabel, C Weinholdt, PT Ryan, BJ Dekkers, L Bentsink, HW Hilhorst, W Ligterink, F Wellmer, I Grosse, M Quint. 2016. Post-embryonic hourglass patterns mark ontogenetic transitions in plant development. *Mol. Biol. Evol.* 33 (5): 1158-1163. *doi:10.1093/molbev/msw039* (journal cover)

Post-embryonic Hourglass Patterns Mark Ontogenetic Transitions in Plant Development

Hajk-Georg Drost,*,†,1 Julia Bellstädt,2 Diarmuid S. Ó'Maoiléidigh,‡,3 Anderson T. Silva,4 Alexander Gabel,1 Claus Weinholdt,1 Patrick T. Ryan,3 Bas J. W. Dekkers,4,5 Leónie Bentsink,4,5 Henk W. M. Hilhorst,4 Wilco Ligterink,4 Frank Wellmer,3 Ivo Grosse,1,6 and Marcel Quint*,2,7

[1]Institute of Computer Science, Martin Luther University Halle-Wittenberg, Halle (Saale), Germany
[2]Department of Molecular Signal Processing, Leibniz Institute of Plant Biochemistry, Halle (Saale), Germany
[3]Smurfit Institute of Genetics, Trinity College Dublin, Dublin 2, Ireland
[4]Wageningen Seed Lab, Laboratory of Plant Physiology, Wageningen University, Wageningen, The Netherlands
[5]Department of Molecular Plant Physiology, Utrecht University, Utrecht, The Netherlands
[6]German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany
[7]Institute of Agricultural and Nutritional Sciences, Martin Luther University Halle-Wittenberg, Halle (Saale), Germany
[†]Present address: Sainsbury Laboratory Cambridge, University of Cambridge, Cambridge, United Kingdom
[‡]Present address: Max Planck Institute for Plant Breeding Research, Cologne, Germany

***Corresponding author:** E-mail: hajk-georg.drost@informatik.uni-halle.de; marcel.quint@landw.uni-halle.de.

Associate editor: Juliette de Meaux

Abstract

The historic developmental hourglass concept depicts the convergence of animal embryos to a common form during the phylotypic period. Recently, it has been shown that a transcriptomic hourglass is associated with this morphological pattern, consistent with the idea of underlying selective constraints due to intense molecular interactions during body plan establishment. Although plants do not exhibit a morphological hourglass during embryogenesis, a transcriptomic hourglass has nevertheless been identified in the model plant *Arabidopsis thaliana*. Here, we investigated whether plant hourglass patterns are also found postembryonically. We found that the two main phase changes during the life cycle of *Arabidopsis*, from embryonic to vegetative and from vegetative to reproductive development, are associated with transcriptomic hourglass patterns. In contrast, flower development, a process dominated by organ formation, is not. This suggests that plant hourglass patterns are decoupled from organogenesis and body plan establishment. Instead, they may reflect general transitions through organizational checkpoints.

Key words: **developmental hourglass, plant development, transcriptomics, germination, floral transition.**

Based on von Baer's third law of embryology (Von Baer 1828), it has been observed that midstage embryos of animal species from the same phylum share morphological similarities. Because these embryos tend to be more divergent at early and late embryogenesis, this morphological pattern has been termed the "developmental hourglass" (Duboule 1994; Raff 1996) (fig. 1A). The window of maximum morphological conservation in midembryogenesis coincides with the onset of organogenesis during body plan establishment and is called phylotypic stage (Sander 1983) or phylotypic period (Richardson 1995, Kalinka et al. 2010). 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 (Raff 1996) (fig. 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 (Hazkani-Covo et al. 2005; Irie and Sehara-Fujisawa 2007; Artieri et al. 2009; Cruickshank and Wade 2008; Kalinka et al. 2010; Domazet-Lošo and Tautz 2010; Yanai et al. 2011; Irie and Kuratani 2011; Levin et al. 2012; Wang et al. 2013; Levin et al. 2016), which is in agreement with a potentially causative association with body plan establishment.

In contrast to animals with their almost exclusively embryogenic development, organ formation in plants occurs largely postembryonically (fig. 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 (Quint et al. 2012; Drost et al. 2015) (as well as for fungal development; Cheng et al. 2015), indicating that it may not be causally connected to organogenesis, as suggested by the animal model. We therefore wondered

© The Author 2016. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Open Access

Downloaded from http://mbe.oxfordjournals.org/ by guest on July 4, 2016

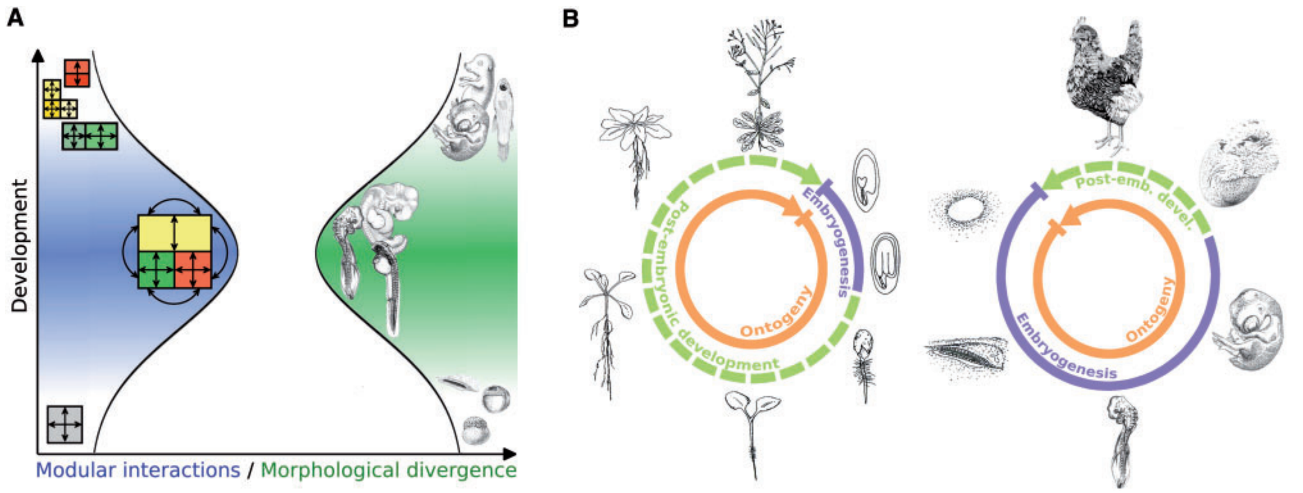

A
Development
Modular interactions / Morphological divergence
B
Embryogenesis
Ontogeny
Post-embryonic development
Post-emb. devel.
Ontogeny
Embryogenesis

Fig. 1. The developmental hourglass model in the context of differences in plant and animal development. (*A*) According to Raff (1996), 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). (*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.

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 (Domazet-Lošo and Tautz 2010; Quint et al. 2012; Drost et al. 2015), which assess the phylogenetic age of transcriptomes expressed over sequential developmental stages (supplementary fig. S1, Supplementary Material online), 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.

Results and Discussion

To study the transition from embryogenesis to the vegetative phase, we generated transcriptomic information for seven sequential ontogenetic stages during seed germination (Silva et al. 2016). 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. 2*A* and supplementary fig. S2, Supplementary Material online). We then combined the transcriptomic information with previously generated gene age information (Drost et al. 2015). Based on an age-assignment approach called phylostratigraphy (Domazet-Lošo et al. 2007) (supplementary fig. S1, Supplementary Material online), genes can be sorted into discrete age categories named phylostrata (PS) (Domazet-Lošo et al. 2007). For *A. thaliana*, we defined 12 age classes ranging from old (PS1) to young (PS12). Next, we computed the transcriptome age index (TAI) (Domazet-Lošo and Tautz 2010) 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 2*B*, 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 [Drost et al. 2015]: $P = 8.92 \times 10^{-20}$; reductive hourglass test (Drost et al. 2015): $P = 3.08 \times 10^{-16}$; supplementary fig. S3*a*, Supplementary Material online). 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. 2*B*). 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. 2*C*), similar to what has been previously reported for embryogenesis (Quint et al. 2012; Drost et al. 2015).

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

Downloaded from http://mbe.oxfordjournals.org/ by guest on July 4, 2016

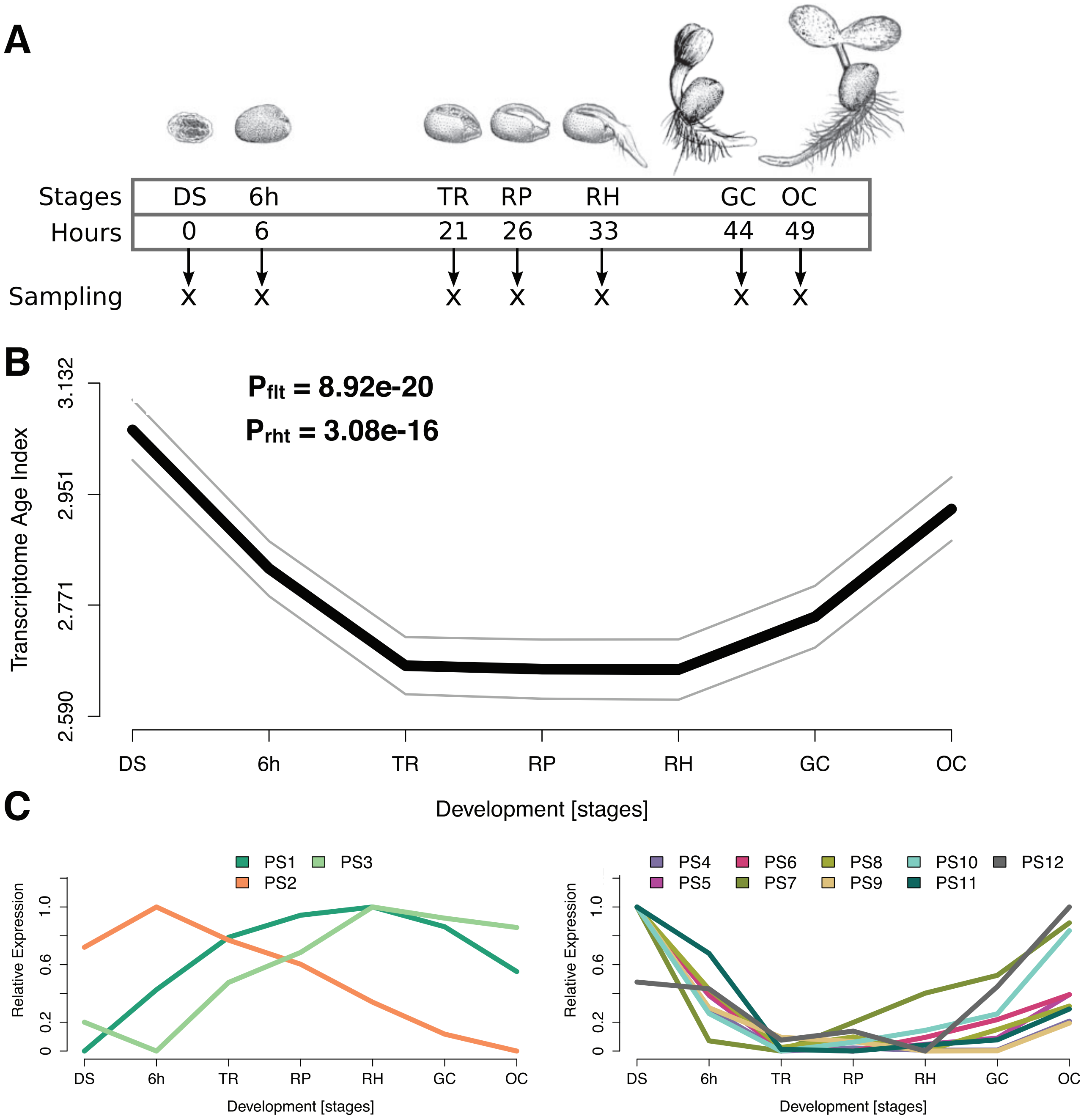


FIG. 2. TAI analysis for germination in *Arabidopsis thaliana*. (*A*) Illustration of the developmental stages for which transcriptome data were generated. (*B*) The TAI profile across germination follows an hourglass-like pattern. The gray lines represent the standard deviation estimated by permutation analysis. *P* values were derived by application of the flat line test (Drost et al. 2015) (P_{flt}) and the reductive hourglass test (Drost et al. 2015) (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.

inflorescence meristem, which forms flowers (Huijser and Schmid 2011). 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, Supplementary Material online). Based on this information, we synchronized flowering time in the sampling population (supplementary fig. S5, Supplementary Material online; see Methods) and generated transcriptome data from the shoot apex before, during, and after floral transition.

Downloaded from http://mbe.oxfordjournals.org/ by guest on July 4, 2016

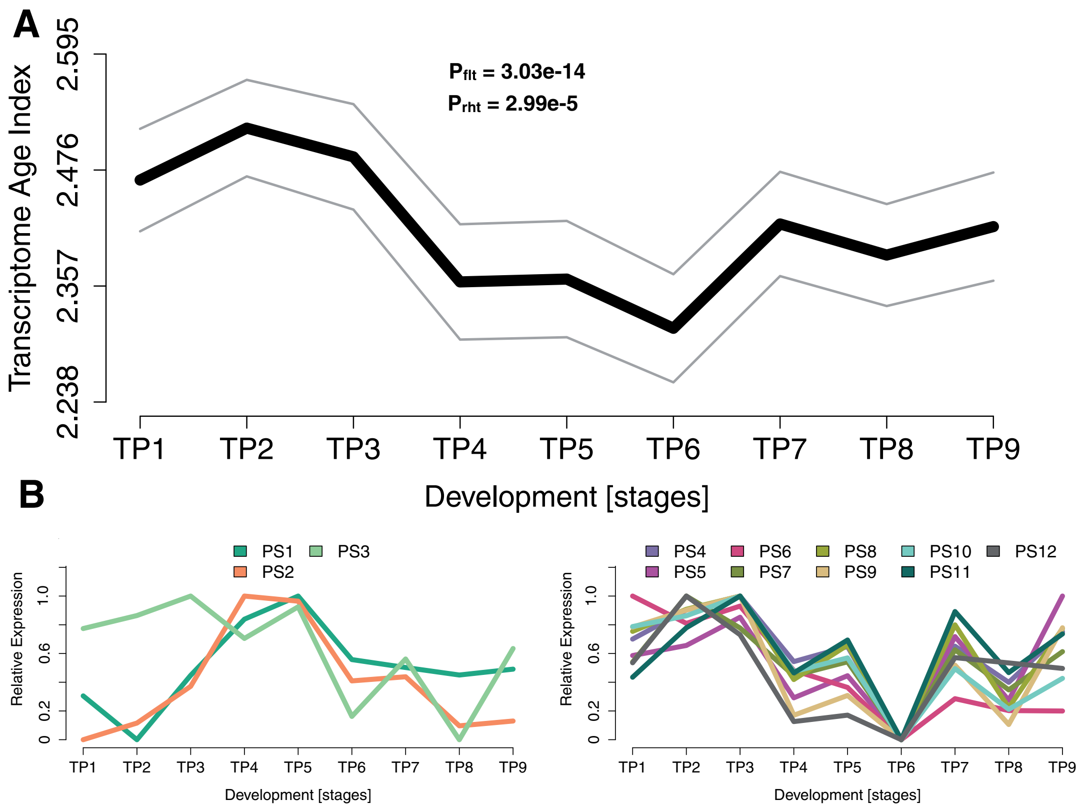


Fig. 3. TAI analysis for the transition from vegetative to reproductive growth in *Arabidopsis thaliana*. (*A*) The TAI profile across the transition to flowering follows an hourglass-like pattern. The gray lines represent the standard deviation estimated by permutation analysis. *P* values were derived by application of the flat line test (Drost et al. 2015) (P_{flt}) and reductive hourglass test (Drost et al. 2015) (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.

Figure 3A shows the results from the TAI analysis for nine samples covering the floral transition. We identified a robust hourglass pattern (reductive hourglass test [Drost et al. 2015]: $P = 2.99 \times 10^{-5}$; fig. 3A and supplementary fig. S3*b*, Supplementary Material online) that significantly deviated from a flat line (flat line test [Drost et al. 2015]: $P = 3.03 \times 10^{-14}$). Similar to embryogenesis (Quint et al. 2012; Drost et al. 2015) and seed germination (fig. 2C), analysis of relative expression levels of genes assigned to different age classes revealed a largely antagonistic behavior of old and young genes (fig. 3B).

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 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. To directly test this, we performed phylotranscriptomic analyses of a flower development data set we previously generated (Ryan et al. 2015). 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 [Drost et al. 2015]: $P = 0.202$; fig. 4*A* and *B*). Likewise, old and young genes did not show a clear antagonistic behavior in their expression (fig. 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 (Raff 1996), 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

Downloaded from http://mbe.oxfordjournals.org/ by guest on July 4, 2016

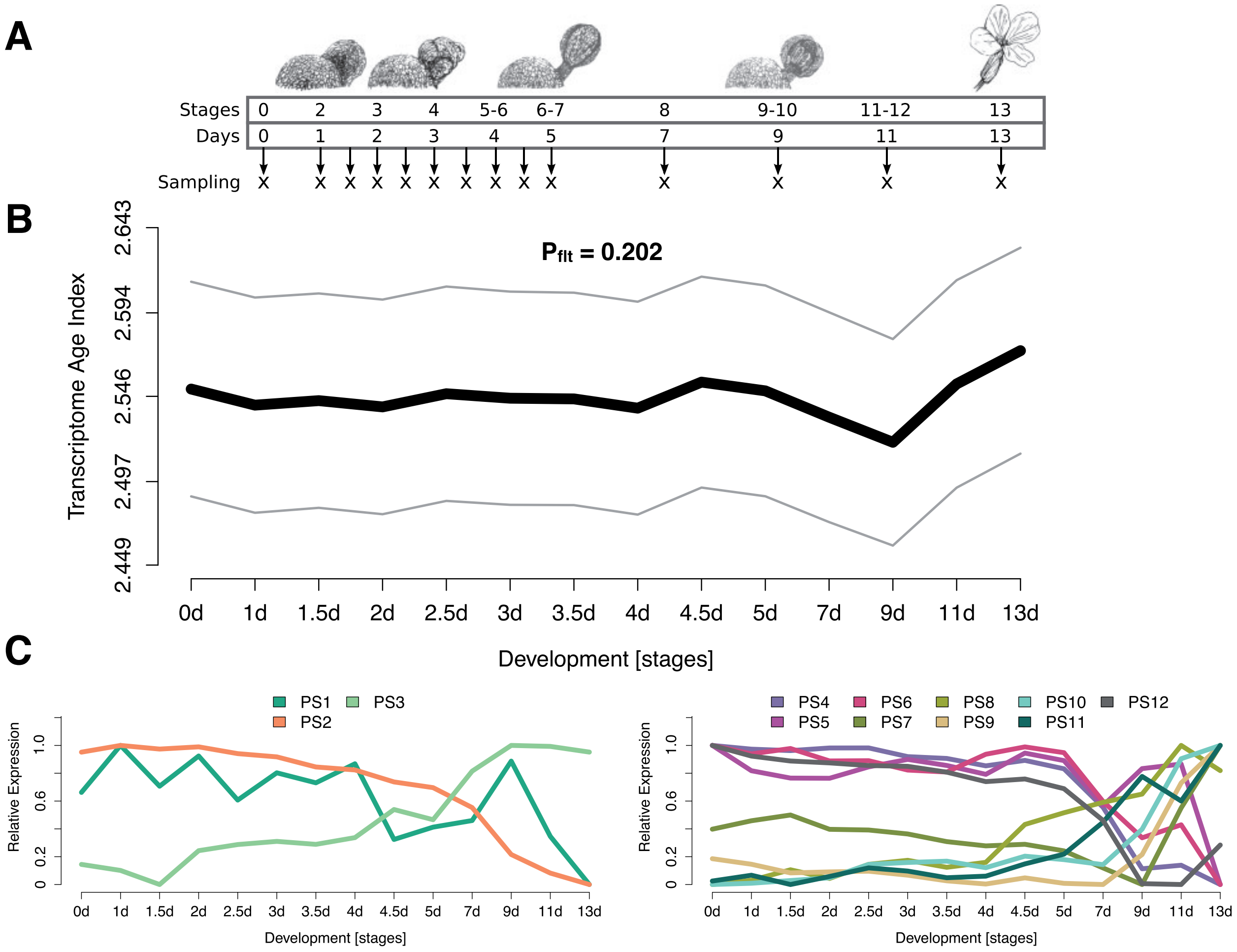


Fig. 4. TAI analysis of flower development in *Arabidopsis thaliana*. (*A*) Illustration of the developmental stages for which transcriptome data were generated; stages according to Ryan et al. 2015. (*B*) The TAI profile across flower development fails to detect evolutionary signal. The gray lines represent the standard deviation estimated by permutation analysis. The *P* value was derived by application of the flat line test (Drost et al. 2015) (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).

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 (Kaplan and Cooke 1997). 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 (Quint et al. 2012). 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 (Irie and Kuratani 2014), it might be worthwhile to directly address this relationship by designing experiments that separate developmental transitions from organogenesis in animals.

In summary, 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, but present in all three major developmental transitions of plant life. 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

Downloaded from http://mbe.oxfordjournals.org/ by guest on July 4, 2016

require passing through an organizational checkpoint serving as a switch that separates two functional programs.

Supplementary Material

Supplementary figures S1–S5, text, and dataset S1 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank the Deutsche Forschungsgemeinschaft (Qu 141/6-1, Qu 141/7-1, GR 3526/7-1, and GR 3526/8-1); the Leibniz Association; the SKWP research foundation; the Science Foundation Ireland; the Netherlands Organization for Scientific Research; the Dutch Technology Foundation, which is the applied science division of the Netherlands Organization for Scientific Research and the Technology Program of the Ministry of Economic Affairs; and the "Coordenação de Aperfeiçoamento de Pessoal de Nível Superior" for financial support.

References

Artieri CG, Haerty W, Singh RS. 2009. Ontogeny and phylogeny: molecular signatures of selection, constraint, and temporal pleiotropy in the development of *Drosophila*. *BMC Biol* 7:42.

Cheng X, Hui JHL, Lee YY, Law PTW, Kwan HS. 2015. A developmental hourglass in fungi. *Mol Biol Evol*. 32:1556–1566.

Cruickshank T, Wade MJ. 2008. Microevolutionary support for a developmental hourglass: gene expression patterns shape sequence variation and divergence in *Drosophila*. *Evol Dev*. 10:583–590.

Domazet-Lošo T, Brajković J, Tautz D. 2007. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends Genet*. 23:533–539.

Domazet-Lošo T, Tautz D. 2010. A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature* 468:815–818.

Drost HG, Gabel A, Grosse I, Quint M. 2015. Evidence for active maintenance of phylotranscriptomic hourglass patterns in animal and plant embryogenesis. *Mol Biol Evol*. 32:1221–1231.

Duboule D. 1994. Temporal colinearity and the phylotypic progression: a basis for the stability of a vertebrate Bauplan and the evolution of morphologies through heterochrony. *Dev Suppl*. 135–142.

Hazkani-Covo E, Wool D, Graur D. 2005. In search of the vertebrate phylotypic stage: a molecular examination of the developmental hourglass model and von Baer's third law. *J Exp Zool B Mol Dev Evol*. 304:150–158.

Huijser P, Schmid M. 2011. The control of developmental phase transitions in plants. *Development* 138:4117–4129.

Irie N, Kuratani S. 2011. Comparative transcriptome analysis reveals vertebrate phylotypic period during organogenesis. *Nat Commun*. 2:248.

Irie N, Kuratani S. 2014. The developmental hourglass model: a predictor of the basic body plan? *Development* 141:4649–4655.

Irie N, Sehara-Fujisawa A. 2007. The vertebrate phylotypic stage and an early bilaterian-related stage in mouse embryogenesis defined by genomic information. *BMC Biol*. 5:1.

Kalinka AT, Varga KM, Gerrard DT, Preibisch S, Corcoran DL, Jarrells J, Ohler U, Bergman CM, Tomancak P. 2010. Gene expression divergence recapitulates the developmental hourglass model. *Nature* 468:811–814.

Kaplan DR, Cooke TJ. 1997. Fundamental concepts in the embryogenesis of dicotyledons: a morphological interpretation of embryo mutants. *Plant Cell*. 9:1903–1919.

Levin M, Anavy L, Cole AG, Winter E, Mostov N, Khair S, Senderovich N, Kovalev E, Silver DH, Feder M, et al. 2016. The middevelopmental transition and the evolution of animal body plans. *Nature* Advance Access publication, doi:10.1038/nature16994.

Levin M, Hashimshony T, Wagner F, Yanai I. 2012. Developmental milestones punctuate gene expression in the *Caenorhabditis* embryo. *Dev Cell*. 22:1101–1108.

Quint M, Drost HG, Gabel A, Ullrich KK, Boenn M, Grosse I. 2012. A transcriptomic hourglass in plant embryogenesis. *Nature* 490:98–101.

Raff RA. 1996. The shape of life: genes, development and the evolution of animal form. Chicago (IL): University of Chicago Press.

Richardson MK. 1995. Heterochrony and the phylotypic period. *Dev Biol*. 172:412–421.

Ryan PT, Ó'Maoiléidigh DS, Drost HG, Kwaśniewska K, Gabel A, Grosse I, Graciet E, Quint M, Wellmer F. 2015. Patterns of gene expression during *Arabidopsis* flower development from the time of initiation to maturation. *BMC Genomics* 16:488.

Sander K. 1983. The evolution of patterning mechanisms: gleanings from insect embryogenesis and spermatogenesis. In: BC Goodwin, N Holder, C Wylie, editors. Development and evolution. The Sixth Symposium of the British Society for Developmental Biology. Cambridge: Cambridge University Press. p. 137–160.

Silva AT, Ribone PA, Chan RL, Ligterink W, Hilhorst HW. 2016. A predictive co-expression network identifies novel genes controlling the seed-to-seedling phase transition in *Arabidopsis thaliana*. *Plant Physiol*. Advance Access publication February 17, 2016; doi: 10.1104/pp.15.01704.

Von Baer KE. 1828. Über Entwicklungsgeschichte der Thiere: Beobachtung und Reflexion. Konigsberg: Gebrüder Bornträger.

Wang Z, Pascual-Anaya J, Zadissa A, Li W, Niimura Y, Huang Z, Li C, White S, Xiong Z, Fang D, et al. 2013. The draft genomes of soft-shell turtle and green sea turtle yield insights into the development and evolution of the turtle-specific body plan. *Nat Genet*. 45:701–706.

Yanai I, Peshkin L, Jorgensen P, Kirschner MW. 2011. Mapping gene expression in two Xenopus species: evolutionary constraints and developmental flexibility. *Dev Cell*. 20:483–496.

Downloaded from http://mbe.oxfordjournals.org/ by guest on July 4, 2016

8 Paper 3: Patterns of gene expression during Arabidopsis flower development from the time of initiation to maturation

PT Ryan, DS Ó'Maoiléidigh, **HG Drost**, K Kwasniewska, A Gabel, I Grosse, E Graciet, M Quint, F Wellmer . 2015. Patterns of gene expression during *Arabidopsis* flower development from the time of initiation to maturation. *BMC Genomics* 16:488. *doi:10.1186/s12864-015-1699-6*

Ryan *et al. BMC Genomics* (2015) 16:488
DOI 10.1186/s12864-015-1699-6

BMC Genomics

RESEARCH ARTICLE Open Access

Patterns of gene expression during *Arabidopsis* flower development from the time of initiation to maturation

Patrick T. Ryan[1†], Diarmuid S. Ó'Maoiléidigh[1,5†], Hajk-Georg Drost[2], Kamila Kwaśniewska[1], Alexander Gabel[2], Ivo Grosse[2], Emmanuelle Graciet[1,3], Marcel Quint[4*] and Frank Wellmer[1*]

Abstract

Background: The formation of flowers is one of the main model systems to elucidate the molecular mechanisms that control developmental processes in plants. Although several studies have explored gene expression during flower development in the model plant *Arabidopsis thaliana* on a genome-wide scale, a continuous series of expression data from the earliest floral stages until maturation has been lacking. Here, we used a floral induction system to close this information gap and to generate a reference dataset for stage-specific gene expression during flower formation.

Results: Using a floral induction system, we collected floral buds at 14 different stages from the time of initiation until maturation. Using whole-genome microarray analysis, we identified 7,405 genes that exhibit rapid expression changes during flower development. These genes comprise many known floral regulators and we found that the expression profiles for these regulators match their known expression patterns, thus validating the dataset. We analyzed groups of co-expressed genes for over-represented cellular and developmental functions through Gene Ontology analysis and found that they could be assigned specific patterns of activities, which are in agreement with the progression of flower development. Furthermore, by mapping binding sites of floral organ identity factors onto our dataset, we were able to identify gene groups that are likely predominantly under control of these transcriptional regulators. We further found that the distribution of paralogs among groups of co-expressed genes varies considerably, with genes expressed predominantly at early and intermediate stages of flower development showing the highest proportion of such genes.

Conclusions: Our results highlight and describe the dynamic expression changes undergone by a large number of genes during flower development. They further provide a comprehensive reference dataset for temporal gene expression during flower formation and we demonstrate that it can be used to integrate data from other genomics approaches such as genome-wide localization studies of transcription factor binding sites.

Keywords: *Arabidopsis thaliana*, Flower development, Organ specification, Transcriptomics, Temporal gene expression, Paralog, Gene expression atlas

---

* Correspondence: mquint@ipb-halle.de; wellmerf@tcd.ie
† Equal contributors
[4] Leibniz Institute of Plant Biochemistry, Department of Molecular Signal Processing, Halle (Saale), Germany
[1] Smurfit Institute of Genetics, Trinity College Dublin, Dublin, Ireland
Full list of author information is available at the end of the article

© 2015 Ryan et al. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

Background

The formation of flowers is one of the main models for studying the molecular mechanisms underlying the control of plant development. Over the past three decades, a large number of regulatory genes, which control a multitude of different processes during flower morphogenesis, have been identified mainly through a combination of forward and reverse genetics approaches [1–3]. Work in *Arabidopsis thaliana* in particular has led to an understanding of the molecular mechanisms underlying the functions of many of these regulatory genes [4]. Furthermore, it has yielded detailed insights into the regulatory hierarchies among genes that play roles in the control of floral organ formation [5, 6].

With the advent of the genomics era, genetic approaches employed to elucidate the regulation of flower development have been complemented by methods such as global transcript profiling and genome-wide localization studies of transcription factor binding sites. Unfortunately, this work has been hampered in *Arabidopsis* by the fact that flowers of this model plant are small and early-stage floral buds are too minute to be dissected reliably through conventional approaches. Also, *Arabidopsis* flowers are initiated sequentially so that all flowers in an inflorescence are at distinct developmental stages [7]. As a consequence, the collection of sufficient numbers of flowers at particular stages for analysis by genomic technologies is challenging especially for early flower development. To circumvent this problem, a number of approaches have been employed: recently, laser capture microdissection has been used to generate transcriptional profiles of early-stage floral buds [8]. An alternative and largely complementary approach has been the use of floral induction systems, which allow the collection of hundreds of synchronized floral buds from a single plant (see below). These systems have been employed to study both temporal and spatial gene expression during the early stages of flower development [9–14]. Other studies have analyzed gene expression in whole inflorescences of wild-type and mutant plants and in some cases relied on the removal of older (and relatively large) buds that may unduly contribute to RNA preparations from these tissues [15–19]. Moreover, transcript profiling was done with wild-type flowers at individual stages and with distinct floral organ types, but this work has been limited to older flowers, as they can be collected with relative ease [17]. Specific developmental processes such as male-gametophyte/ pollen and female gametophyte/ ovule development have also been studied through transcriptomics experiments, providing detailed information for individual cell and tissue types [20–23].

Although *Arabidopsis* flower development has been studied extensively over the past ten years through the genomics approaches described above, a continuous series of gene expression from the time of initiation to maturation has been lacking. Obtaining this information could be highly informative as it would provide a comprehensive view of stage-specific gene expression activities over the entire course of development and would constitute an important component of a gene expression map. Furthermore, such a dataset could be used in analyses, in which, for example, data from transcript profiling and genome-wide localization studies are integrated to obtain a better understanding of the gene network that controls flower formation.

In this study, we employed a floral induction system to close this knowledge gap and to monitor temporal gene expression during flower development from the time of initiation to maturation. We validated the resulting dataset and used it to obtain novel insights into the processes underlying the formation of flowers on a global scale through computational approaches.

Results and discussion

Temporal gene expression during flower development

To identify patterns of gene expression during flower development from the time of initiation to maturation (stage 13; stages according to [7]), we employed a previously described floral induction system, which allows the collection of hundreds of floral buds from a single plant [9, 13, 24, 25]. This system is based on the expression of the floral meristem identity factor APETALA1 (AP1) fused to the hormone-binding domain of the rat glucocorticoid receptor (GR) from the *AP1* regulatory region ($\mathrm{AP1_{pro}}$) in an *ap1 cauliflower* (*cal*) double-mutant background. *Ap1 cal* plants accumulate inflorescence-like meristems at their shoot apices [26, 27], and activation of the AP1-GR fusion protein in this background through treatment of the plants with the steroid hormone dexamethasone results in the transformation of these meristems into floral primordia, which subsequently develop in a largely synchronized manner. However, at intermediate stages, this synchronization is gradually lost likely due to space constraints [9]. Despite this overall loss of synchronization, we noticed that flowers at the very tip of the inflorescence heads remained fairly synchronized throughout flower development perhaps due to a larger degree of curvature in his area, which may allow floral buds to develop without coming into contact with neighboring flowers. For the gene expression profiling experiments, we therefore collected older floral buds (days 9 to 13 after dexamethasone treatment, corresponding to stages 9-10 to 13, respectively) from this region alone, while younger flowers were harvested more liberally from the inflorescences of $\mathrm{AP1_{pro}}$:AP1-GR *ap1 cal* plants (Fig. 1a-j). To obtain expression data for a large number of distinct floral stages, we collected floral buds at 14 different time-points either immediately before (referred to as 0 d time-point) or from 1 to 13 d after the induction of flower development through treatment with dexamethasone (Fig. 1k). Because

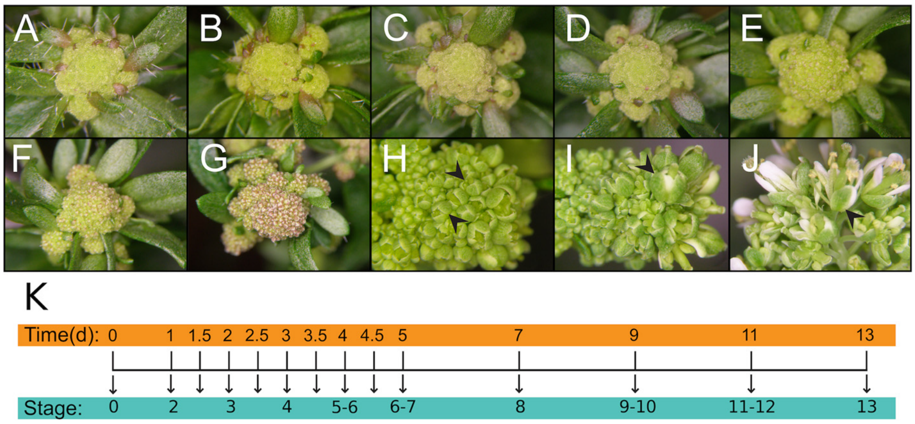

A
B
C
D
E
F
G
H
I
J
K
Time(d): 0 1 1.5 2 2.5 3 3.5 4 4.5 5 7 9 11 13
Stage: 0 2 3 4 5-6 6-7 8 9-10 11-12 13

Fig. 1 Analysis of temporal gene expression during flower development. **a-j** Inflorescences of AP1$_{pro}$:AP1-GR *ap1-1 cal-1* plants **a** before dexamethasone treatment (0 d time-point) , and **b** 1 d, **c** 2 d, **d** 3 d, **e** 4 d, **f** 5 d, **g** 7 d, **h** 9 d, **i** 11 d, and **j** 13 d after treatment with a solution containing 10 μM dexamethasone. The development of flowers on a given inflorescence was largely synchronous until day 7. For later time-points (**h-j**), flowers were harvested from the tip of the inflorescences (arrowheads) after phenotypic assessment. **k** Experimental set-up used for this study. Floral buds were collected from the inflorescences of AP1$_{pro}$:AP1-GR *ap1-1 cal-1* plants at 14 time-points immediately before and after treatment with a dexamethasone ('DEX')-containing solution, which induces flower development by activating the AP1-GR fusion protein. Floral buds from the time of initiation until anthesis (corresponding to stage 13) were sampled

early flower development is characterized by dramatic changes in morphology [7] and involves a large number of transcriptional regulators that control important processes such as floral patterning and floral organ specification [4], we collected most samples at those stages with intervals in-between time-points ranging from 0.5 to 1 d. At later stages of development, the intervals for sample collection were extended to 2 d (Fig. 1k).

For microarray analysis of the tissue samples, we employed a common reference design (e.g., ref. [28]). We then assessed the resulting data for reproducibility and found that the replicates for the individual time-points correlated well (Figure S1 in Additional file 1; see also Fig. 2), implying that the progression of flower development and the tissue collection was highly reproducible over the entire course of the experiment. In order to determine significant expression changes, we applied an F-statistic and searched across the entire dataset for genes with differential expression. We identified ~20,000 genes (i.e., ~75 % of the genes in the *Arabidopsis* genome) that showed differential expression in at least one of fourteen time-points. Because many of these transcriptional changes may be caused by the dramatic alterations in floral size and morphology during the course of development and not by specific gene regulatory events, we next sought to identify genes whose expression changed relatively rapidly. To this end, we compared gene expression between consecutive as well as near-by (within a 2-d time interval) time-points to minimize the effects of morphological alterations and identified 7,405 genes as differentially expressed (Additional file 2). Many of these differentially expressed genes (DEGs) were detected at

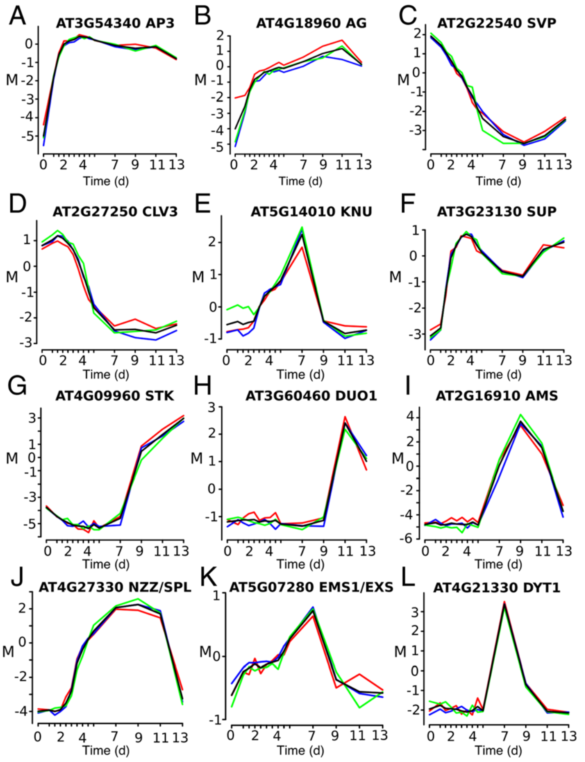


Fig. 2 Expression profiles of known floral regulators. **a-l** *M* values (log$_2$ (expression in sample/expression in common reference)) for selected floral regulators (as indicated) are shown for all time-points. Red, green and blue lines represent data from three biologically independent sets of samples, black lines the mean values of the replicate experiments. Note the high reproducibility of the expression data across all time-points

intermediate (between 5 and 9 d after dexamethasone treatment) and late (between 9 and 13 d) stages of flower development, and overall, a preponderance of gene activation over repression was observed (Table S1 in Additional file 1). Although we found many genes to be repressed immediately after the onset of flower development, this effect was not as pronounced as previously described [9, 29], possibly because of the different floral induction systems and/or different experimental set-ups and data analysis pipelines used.

To validate the results of the microarray experiments, we assessed the expression profiles of genes with known roles in different processes during flower development (Fig. 2 and Figure S2 in Additional file 1) and found that they were in concurrence with their published expression patterns. For example, expression of the floral homeotic genes *APETALA3* (*AP3*) and *AGAMOUS* (*AG*) (Fig. 2a-b) strongly increased in early time-points and then remained high throughout most of flower development in agreement with the activation of these genes at stage 3 and their continued expression in developing floral organs [30, 31]. Down-regulation of the floral repressor *SHORT VEGETATIVE PHASE* (*SVP*) (Fig. 2c) at early floral stages has been described previously and is dependent on AP1 activity [29, 32]. Expression of the stem cell regulator *CLAVATA3* (*CLV3*) was high at early stages and then rapidly decreased in intermediate-stage flowers (Fig. 2d) likely as a consequence of the loss of floral stem cells around stage 6 of development [33]. This termination of floral meristems is at least in part due to the activity of *KNUCKLES* (*KNU*), which we detected to be expressed at intermediate stages (Fig. 2e), in agreement with its known expression pattern at the base of developing carpels and in stamen primordia [34, 35]. Genes with bimodal expression profiles included *SUPERMAN* (*SUP*) (Fig. 2f), which is initially expressed in young floral meristems and at later floral stages during ovule development [36]. Strong up-regulation of the regulator of ovule and seed development *SEEDSTICK* (*STK*) between days 7 and 9 in our experiment (Fig 2g) corresponds to its expression in developing carpels from stage 8 onward [37]. *DUO POLLEN1* (*DUO1*), a regulator of male germline development, was found to be expressed in late flower development (Fig. 2h) in agreement with its specific expression in pollen [38]. *ABORTED MICROSPORES* (*AMS*), which encodes a master regulator of pollen wall formation, was strongly expressed at intermediate stages and reached a maximum around stages 9-10 (9 d after dexamethasone treatment) (Fig. 2i) as previously described [39]. Genes such as *NOZZLE/SPOROCYTELESS* (*NZZ/SPL*) (Fig. 2j), *EXTRA MICROSPOROCYTES1/EXTRA SPOROGENOUS CELLS* (*EMS1/EXS*) (Fig. 2k), and *DYSFUNCTIONAL TAPETUM1* (*DYT1*) (Fig. 2l) were expressed during intermediate stages in agreement with their function in early anther development [40–44]. Activation of *NZZ/SPL* was detected in our experiment around stage 5 and thus earlier than what has been reported previously (i.e. stage 6; [45]). This difference might stem from initially low mRNA levels, which might hamper a reliable detection in *in situ* hybridization or reporter gene essays.

We also compared our dataset to those from several previous studies in which temporal [8–10, 14] and spatial [11, 16] gene expression during flower development had been analyzed either in early or in late-stage flowers using different floral induction systems, laser capture microdissection of wild-type flowers, or through a comparison of the gene expression profiles of inflorescences of floral mutants and of the wild type, respectively. For each pair-wise comparison, we found a significant overlap between the datasets and the one described in this study (Table S2 in Additional file 1 and Additional file 3), further validating the results of our time-course experiment.

Distribution of functional terms among groups of co-expressed genes

Because functionally related genes are often co-expressed during development, we used a *k*-means algorithm to group the DEGs into 15 clusters with distinct gene expression profiles (Fig. 3 and Figure S3 in Additional file 1). Figure 3 shows that the majority of DEGs are predominantly expressed at or after the 9-d time-point. Notable exceptions include genes in clusters 5, 11 and 15, which are up-regulated during early flower development and are repressed at intermediate to late stages. Also, clusters 6 and 7 contain genes that are expressed at the earliest floral stages and are subsequently down-regulated. Genes assigned to clusters 4 and 12 are activated during early flower development when organ primordia are initiated and remain expressed until flowers have reached maturity, suggesting that many of them might play roles during the course of floral organ morphogenesis.

To obtain insights into the functions of the genes assigned to each of the clusters and to further validate the microarray data, we mapped the groups of co-expressed genes onto an *Arabidopsis* gene expression atlas we had generated previously [13] based on published data (Fig. 4a and Additional file 4). We then determined the percentage of genes with maximum (Fig. 4b) and, for comparison, minimum (Fig. 4c) expression in different groups of related tissue samples. For some of the clusters, this analysis allowed predictions of the predominant location of gene expression. For example, a high percentage of genes with maximum expression in pollen was identified in clusters 2, 3, 8-10, and 13-14. Genes assigned to these clusters were predominantly expressed from or after the 9-d time-point and thus at stages when pollen formation occurs [46]. Clusters 6, 7, and 13 contained the highest proportion of genes with maximum expression in meristems, in agreement with

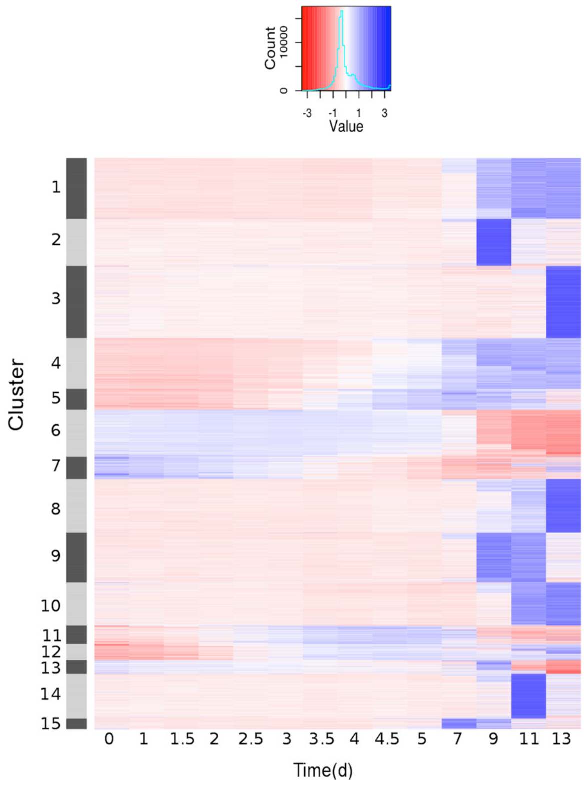


Fig. 3 Genes showing differential expression during flower development. Groups of co-expressed genes were identified among 7,405 differentially expressed genes detected in the time-course experiment. The heat map shows the results of *k*-means clustering ($k = 15$) used to group genes based on the similarity of their *z*-scores (color-coded as per diagram at the top). For a different representation of the individual clusters, see Figure S3

the observation that genes in these clusters are strongly expressed during the earliest floral stages, but are repressed towards more intermediate stages when meristematic activity in flowers ceases. The highest percentage of genes with maximum expression in ovules was found in cluster 15, which contains relatively few genes that are strongly expressed around the 7 and 9-day time-points (corresponding to floral stages 8-10; Fig. 1a) and thus at the time when ovule development commences [47].

We also subjected the groups of co-expressed genes to a Gene Ontology (GO) analysis to identify functionally related genes that are significantly enriched (adjusted $p\text{-value} < 0.05$) in the individual clusters (Figure S4 and Additional file 5). GO terms directly associated with flower formation (e.g., 'Specification of Floral Organ Identity' and terms related to the development of the different floral organ types) and/or floral meristem development (including the terms 'Cell Proliferation' and 'Cell Division') were found to be enriched, in particular, in clusters 6 and 7, as well as in clusters 11 and 12 (Fig. 5a). As described above, these clusters contain genes that are repressed at early to mid-stages (clusters 6 and 7) or are activated during early flower development (clusters 11 and 12) and remain expressed at least until the end of the intermediate phase of flower development. In agreement with the over-representation of flower-related GO terms in these clusters, they contain many of the regulatory genes (which are also typically associated with the GO term 'Regulation of Transcription'; see Fig. 5b) known to be involved in controlling the early phase of flower development (Additional file 2). Genes associated with the term 'Pollen Development' were enriched in clusters 2 and 9, which contain genes with maximal expression around day 9 of the experiment and hence at a time (corresponding to floral stages 9-10; Fig. 1a) when the microspore mother cells appear and meiosis takes place [46]. Genes involved in cell differentiation were enriched in clusters 8 and 10, which contain genes with predominant expression at late stages of flower development (stages 11-13). Many of these genes exhibit maximum expression in pollen (Fig. 4b) and thus, may be involved to a large extent in the differentiation of microspores into pollen grains. Genes involved in the response to different phytohormones such as jasmonic acid, auxin, and abscisic acid were detected as enriched predominantly in cluster 8, in agreement with the known roles of these hormones in various processes during late-stage flower development, which include stamen and pollen formation as well as the maturation of petals [48]. In contrast, genes involved in the response to gibberellin were over-represented in cluster 4, which contains genes that are induced at the end of the early phase of flower development and remain active until floral maturity has been reached. In agreement with this observation, it has been shown that gibberellins are required for proper floral organ growth and elongation [49]. In sum, the results of these analyses allowed us to attribute specific functions to the individual clusters that together account for many of the processes known to occur during flower development.

Distribution of target genes of floral organ identity factors

Floral organ identity factors are necessary and sufficient for the specification and development of the different types of floral organs [5, 6]. They act in a combinatorial manner as predicted by the well-supported (A)BCE model of floral organ identity specification [50–52]. Insights into the functions of these master regulators, which (with the exception of APETALA2) all belong to the family of MADS-domain proteins and are components of higher-order regulatory protein complexes [53], have been obtained in recent years through a combination of genome-wide localization studies and gene

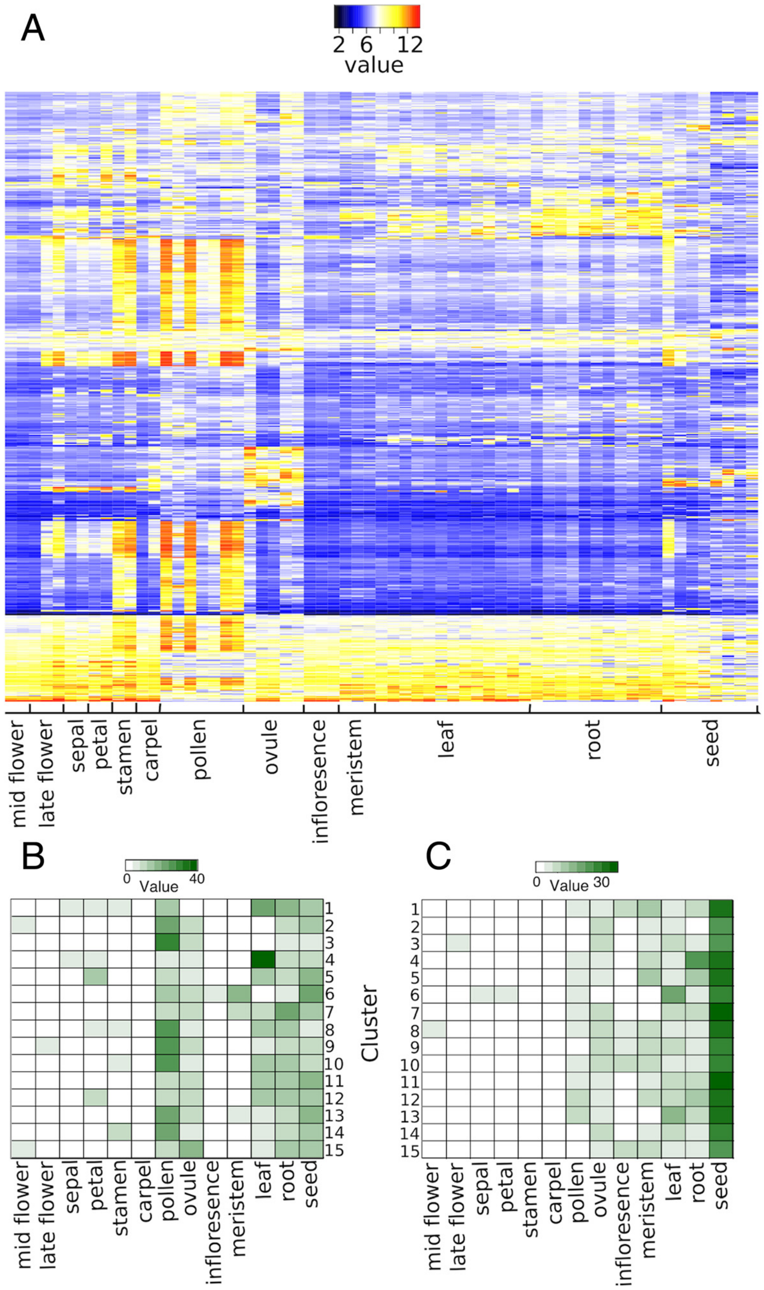


Fig. 4 Mapping groups of co-expressed genes onto an *Arabidopsis* gene expression atlas. **a** Expression data for an *Arabidopsis* gene expression atlas were obtained for genes assigned to each of the 15 *k*-means clusters and hierarchical clustering was performed. Results for cluster 3 are shown as an example. Individual tissue and organ samples of the gene expression atlas (shown in full in Additional file 4) were grouped together as indicated. Note a preponderance of expression in stamen and pollen samples. **b** and **c** The number of genes in each cluster with **b** maximum and **c** minimum expression in each of the tissue samples (as indicated) is shown

perturbation experiments [5, 6]. This work has resulted in the identification of some of their direct target genes and of the cellular and developmental processes they control. Furthermore, it has been shown that the floral organ identity factors bind to many of the same sites in the *Arabidopsis* genome [13] and that their global binding patterns undergo changes as flower development progresses, at least in part as a consequence of stage-specific alterations in chromatin accessibility [14]. Also, the majority of genes bound by these transcription factors at early floral stages do not respond transcriptionally when the activities of the floral homeotic genes are perturbed [12, 13]. While the molecular mechanisms underlying these observations are currently not well

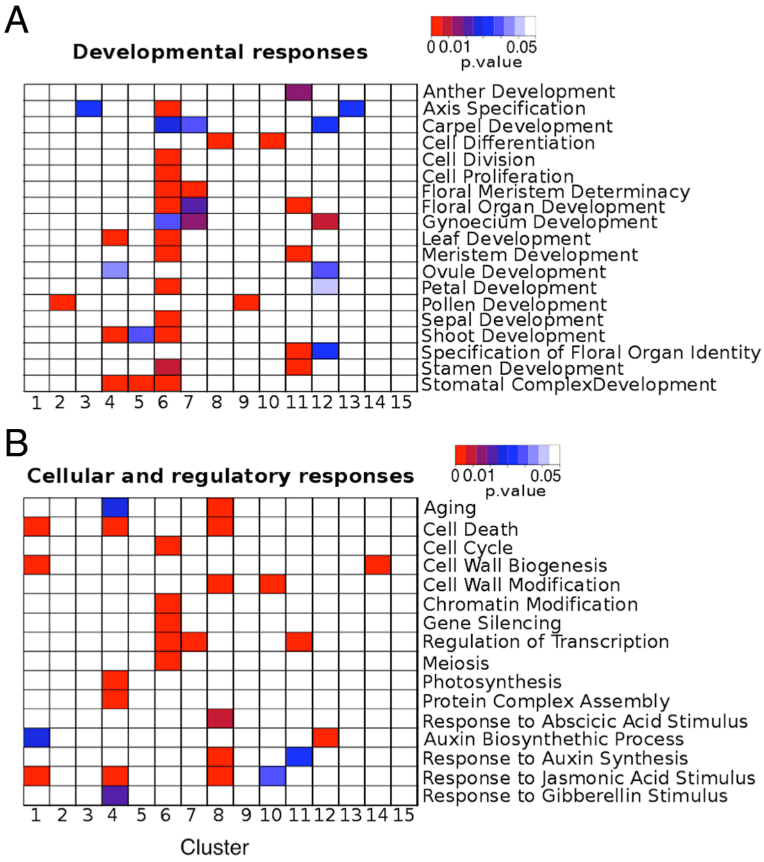


Fig. 5 Gene Ontology terms enriched in the dataset. Adjusted *p*-values for selected GO terms related to **a** developmental functions and **b** cellular and regulatory processes are indicated for each cluster through color-coding (see bars at the top right for colors used). For a full list of GO terms enriched in the dataset, see Additional file 5

understood, it is clear that from binding data alone it is difficult to identify their *bona fide* target genes.

To test whether we could find evidence for the differential expression of genes that are bound by the floral organ identity factors, we projected the global binding patterns of AP1, SEPALLATA3 (SEP3), AP3, PISTILLATA (PI), and AG onto the dataset from the flowering time-course experiment (Additional file 6). Specifically, we identified the percentage of genes in each of the 15 clusters of co-expressed genes that contain binding sites for these transcription factors in their putative regulatory regions (from 3 kb upstream to 1 kb downstream of the transcribed region of a gene). While binding data for AP3, PI and AG are currently available only for ~ stage 4 flowers [12, 13], for AP1 and SEP3, binding data for three distinct stages (2, 5-6, and 7-8) have been generated [14]. Largely independent of the transcription factor under study, we found the highest degree of binding site enrichment in clusters 6, 7, 11, and 12 (Fig. 6). Cluster 5 also showed a significant enrichment for genes with binding sites, but only for SEP3 and AP1, and not at the earliest (stage 2) time-point. The genes assigned to these different clusters have in common that their transcription changes at the time or shortly after the expression of the floral organ identity genes commences around stage 3. Furthermore, they contain many genes associated with the specification of floral organ identity, as well as the regulation of floral organ development and meristem determinacy (Fig. 5) and thus processes that are known to be under control of the floral organ identity factors [5, 6]. Hence, genes in these clusters containing binding sites for the MADS-domain proteins are good candidates for target genes. In fact, they do contain many of the genes known to act directly downstream of these floral regulators (Additional file 6). However, one caveat of this analysis is that the floral organ identity factors appear to have largely distinct sets of target genes despite their overlapping binding patterns [5]. Therefore, while genes that are differentially expressed during early flower development and that contain binding sites for MADS-domain proteins are likely under control of floral organ identity factors, the exact regulatory complex that might be active in the regulation of a given gene cannot be readily deduced without additional data from floral organ identity gene-specific perturbation experiments.

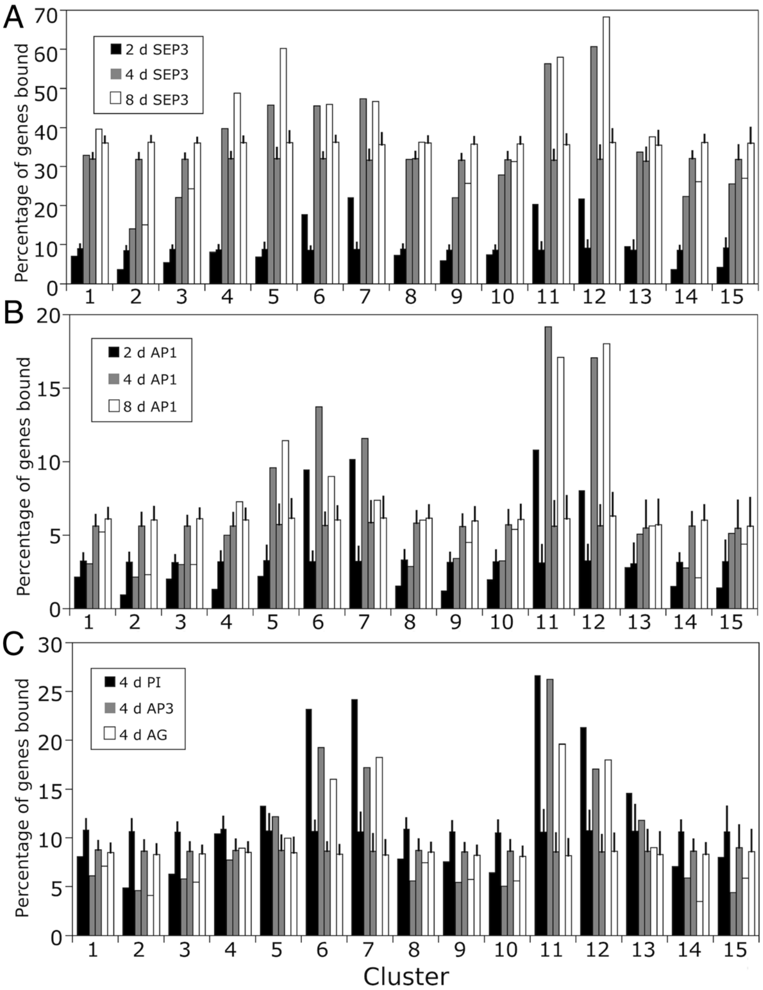


Fig. 6 Distribution of genes with binding sites for floral organ identity factors. The percentage of genes in each cluster bound by **a** SEP3, **b** AP1, and **c** AP3, PI, and AG, respectively, is shown. For **a** and **b**, binding data for SEP3 and AP1, respectively, at three different time points after AP1-GR activation were used for analysis: 2 d (black bars), 4 d (gray bars), and 8 d (white bars). For **c**, binding data for AP3 (black bars), PI (gray bars), and AG (white bars) 4 d after AP1-GR activation were used. In all panels, bars without error bars show the results of the comparisons between binding data for the individual transcription factors and the clusters of co-expressed genes, while bars with error bars show the mean percentage of genes bound by a given floral homeotic transcription factor at the indicated time-point in equally sized groups of genes randomly selected from the dataset of 7,405 DEGs. Error bars indicate one standard deviation calculated based on the results of 100 iterations

In addition to clusters with binding site enrichments, we also found clusters that are significantly depleted for binding sites of the floral organ identity factors. These included especially clusters 2, 3, and 14, which contain genes with predominant expression in the time-course experiment at 9, 13, and 11 d, respectively (Fig. 3). As described above, these clusters comprise in all probability many genes involved in microsporogenesis and pollen development, a process that can progress without the direct involvement of the floral organ identity factors [45]. Taken together, this analysis shows that the results of our transcriptomics study can be used as a reference to integrate different genome-wide datasets and to identify candidates for transcription factor target genes.

Distribution of paralogs within groups of co-expressed genes

In plants, duplicated genes that are retained in a genome are often functionally redundant, although sub- or neo-functionalization may lead to paralogous genes that have only partially overlapping activities or that are employed in entirely different developmental processes, respectively [54]. Shared activities of paralogous genes typically go along with overlapping expression patterns. Therefore, one would expect to find in the clusters of co-expressed genes that paralogs are enriched relative to their genome-wide distribution. In fact, it has been shown previously that paralogous genes are over-represented in some but not all groups of genes with predominant expression at certain stages of early flower development [9]. To test whether this unequal distribution of paralogs extends to intermediate or late stages of flower development, we determined paralogs in each of the 15 clusters described in Fig. 3 (for paralogs identified in the clusters, see Additional file 7). As expected, we found that the percentage of paralogs was significantly (i.e., beyond three standard deviations) increased in all clusters relative to their genome-wide distribution and to a lesser extent (and with the exception of cluster 13) relative to their distribution within the 7,405 DEGs as well (Fig. 7). Notably, the enrichment of paralogs within the clusters varied considerably, with clusters 5, 11-12, and 15 having the highest enrichment values (Table S3 in Additional file 1). In agreement with the idea that genes involved in floral organ development exhibit an increased level of genetic redundancy [9], the genes in these clusters have in common that they are activated during early or intermediate (cluster 15) stages of flower development and many of them have known functions in floral organ morphogenesis and in the control of floral meristem determinacy (Fig. 5). In sum, our results further highlight the varying degree to which paralogous genes contribute to different processes during flower development. Whether such an unequal distribution of paralogs among groups of co-expressed genes extends to other processes during plant development is currently unknown.

Conclusions

The results of our transcriptomics analysis of flower development, which covered most stages from the time of initiation until maturation, shows that the formation of flowers involves the differential expression of at least a quarter of the genes in the *Arabidopsis* genome. While many gene expression changes occur late in development and are likely due to the activation of specific gene sets in developing pollen and - to a lesser extent - ovules, genes with regulatory functions often exhibit intermittent expression during early and late floral stages. Through computational analyses, we have been able to assign functions to groups of co-expressed genes and to provide temporal information on when these processes likely occur during the almost two weeks during which flowers develop from a small number of meristematic cells into a highly complex structure with different organs, tissues and cell types. Using binding data for selected floral organ identity factors, we have further demonstrated that the results of our transcriptomics experiment can help to interpret and mine datasets from genome-wide localization studies. Our data also provide an important component of a gene expression map for flower development. Through the use of techniques such as Translating Ribosome Affinity Purification (TRAP) [11] or Isolation of Nuclei Tagged in specific Cell Types (INTACT) [55], it should be possible to extend this map by introducing detailed spatial information on gene expression for all floral stages.

Methods

Plant material, plant growth, treatment conditions and tissue collection

Plants of genotype AP1$_{pro}$:AP1-GR *ap1-1 cal-1* [13] were grown on a soil:vermiculite:perlite (3:1:1) mixture at 20 °C under constant illumination with cool white fluorescent light. Flower development was induced in ~four week-old plants as described in [9], using a solution containing 10 μM dexamethasone (Sigma-Aldrich), 0.01 % (v/v) ethanol and 0.015 % (v/v) Silwet L-77 (De Sangosse). Floral buds were harvested at different time-points after dexamethasone treatment as described in Fig. 1. Three sets of

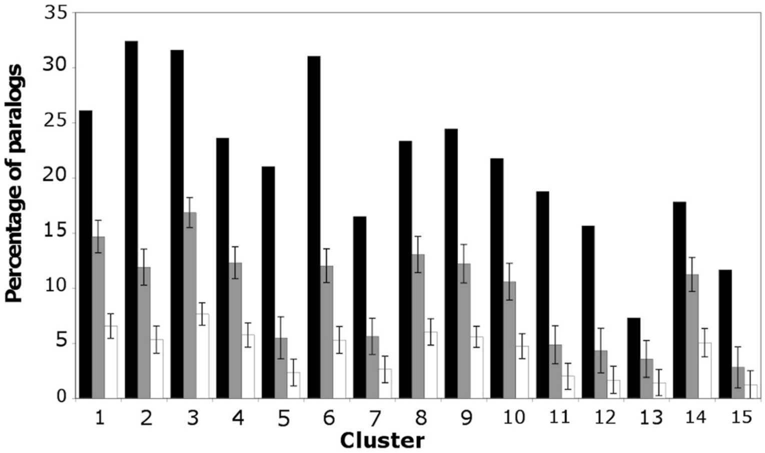


Fig. 7 Distribution of paralogs in groups of co-expressed genes. The percentage of paralogs in each cluster of co-expressed genes (black bars) was determined as described in Methods. To identify clusters with a significant enrichment of paralogous genes, the mean percentage of paralogs was determined in equally sized groups of genes randomly selected from the dataset of 7,405 DEGs (gray bars) and from the *Arabidopsis* genome (white bars), respectively. Error bars indicate one standard deviation calculated based on the results of 100 iterations

biologically independent samples were collected for microarray analysis.

Microarray experiments

Microarray experiments were performed using Agilent whole-genome *Arabidopsis* microarrays. For each microarray hybridization, amplified and dye-labeled RNA samples from a given time-point was co-hybridized with dye-labeled RNA from a common reference sample. This common reference was generated by pooling equal amounts of RNA from the individual time-points from 2 of the 3 sets of independent samples. RNA extractions, amplification and labelling of RNA preparations, microarray hybridizations, as well as washing and scanning of microarrays were done as previously described [12, 13].

Processing of microarray data

Microarray data were analyzed using the software package *limma* (Linear Models for Microarray Data) [56] implemented in *R*. Background correction was done using the *subtract* method and within array normalization was performed with the *loess* method [57]. Between array normalization was done using the *Aquantile* method. Probes within each array were averaged on a gene-level and filtered to remove entries that had expression values below the median value of negative control probes. Linear models were fitted to the data using the *lmscFit* function. Correlograms were generated using the *R* package *corrgram*. Statistics for differential expression were first calculated using the *ebayes* function within *limma*. Genes with a *p*-value (after false discovery rate adjustment using the Benjamini-Hochberg procedure) below 0.01 were considered as differentially expressed. Because this analysis led to a very large number of differentially expressed genes that may not reflect true gene regulatory events (see Results and Discussion), we next compared gene expression between consecutive or nearby time-points using *ebayes*. To this end, we conducted all possible contrasts between time-points that lay within a 2-d interval (see Table S1 in Additional file 1). In order to be called as differentially expressed, genes were required to exhibit a *p*-value below 0.01 after adjustment for false discovery rate across the experiment and a fold-change in expression of 1.7 or greater.

K-means clustering was performed in *R* using scaled $\log_2$-transformed ratios of expression averaged across each replicate across all time-points for each gene, separating differentially expressed genes into 15 clusters on the basis of the similarity of the pattern of their temporal expression. The number of clusters was chosen heuristically based on the elbow method, which aims at maximizing the amount of variance explained while minimizing the number of clusters chosen. To this end, we compared, using the *kmeans* function implemented in *R*, the between-cluster sum-of-squares to the total sum-of-squares for different values for *k* (ranging from 2 to 200). We then plotted the data and selected a value for *k* in the 'elbow' of the plot.

Comparison of expression data with data from an *Arabidopsis* gene expression atlas

Genes assigned to each *k*-means cluster were compared to a previously described [13] *Arabidopsis* gene expression atlas, which is based on published transcriptomics datasets for floral and non-floral tissues, to identify trends in tissue-specific expression within each cluster. This tissue atlas was also used to identify the tissues where genes within a cluster had their highest and lowest expression levels in order to investigate the correlation of changes in temporal expression within developing tissues.

Gene ontology analysis

Gene Ontology analysis was performed using *PlantGSEA* [58]. Statistical significance calculations were performed with a Fisher's exact test using False Discovery Rate adjustment method from Benjamini and Yekutieli [59] with a *p*-value cut off of 0.05.

Identification of paralogs

All known protein sequences from *Arabidopsis* were individually aligned against the sequences from the entire proteome of *Arabidopsis* using *blastp* to select alignments with an E-value cut off of $1\text{x}10^{-20}$ and which covered 80 % of the query sequence [60]. The top 5 non-reciprocal alignments were retained as potential paralogs. Using this information, we determined the percentage of paralogs within each of the 15 clusters of differentially expressed genes described in Fig. 3. To test whether paralogs were significantly enriched in the clusters, we conducted the following background calculation: we first generated, for each cluster, two groups of genes drawn randomly either from the list of 7,405 differentially expressed genes or from genes present on the microarrays used in this study. Both groups contained 100 sets of genes each and the number of genes in a set was identical to the size of a cluster. We then calculated the mean percentage and standard deviations for paralogs in each of the groups and compared them to the percentage of paralogs we had identified in a corresponding cluster. Clusters with percentage values that were beyond three standard deviations from the random gene groups were considered significantly different.

Comparison of expression data with data from genome-wide localization studies

Data from genome-wide localization studies were contrasted with each of the 15 *k*-means clusters to determine the frequency with which genes identified as being bound by the transcription factors AP1 and SEP3 [14], as well as by AP3, PI, and AG [12, 13], occurred in each cluster. This was contrasted against the frequencies with which bound genes occurred in randomized but equally-sized clusters of genes drawn from the 7,405 differentially expressed genes identified in the time-course experiment.

Availability of supporting data

The data sets supporting the results of this article are included within the article (and its additional files). Microarray data have been deposited with the Gene Expression Omnibus (GEO) repository (at http://www.ncbi.nlm.nih.gov/) under GSE64581.

Additional files

Additional file 1: Additional figures, tables and references. Tables and figures in this file are referred to as Table S1-S3 and Figure S1-S4 in the main text.

Additional file 2: Excel spreadsheet containing 7,405 genes identified as differentially expressed in this study. Gene identifiers, aliases, descriptions, and their assignment to one of the fifteen *k*-means clusters are indicated. Also shown are the $\log_2$-transformed expression ratios and adjusted *p*-values for each contrast between time-points ('T').

Additional file 3: Excel spreadsheet listing differentially expressed genes also identified in related studies. Gene identifiers, aliases, descriptions, and their assignment to one of the fifteen *k*-means clusters are indicated. Datasets are described in Table S2 in Additional file 1. For the study by Wellmer et al. (2004) [16], the assignment of genes to one of the four types of floral organs is listed. For the study by Wellmer et al. (2006) [9], the assignment of genes to clusters (A-E) of co-expressed genes is shown. For the study by Gomez-Mena et al. [10], the presence or absence of genes among genes up- or down-regulated after AG-GR activation is indicated. For all other studies, *p*-values for genes from the related studies are shown.

Additional file 4: Mapping groups of co-expressed genes onto an *Arabidopsis* gene expression atlas. Expression data for an *Arabidopsis* gene expression atlas were obtained for genes assigned to each of the 15 *k*-means clusters and hierarchical clustering was performed. Individual tissue and organ samples of the gene expression atlas [12] are indicated.

Additional file 5: Excel spreadsheet containing Gene Ontology terms identified as enriched in the dataset. GO terms and adjusted *p*-values indicating a significant enrichment are shown for all *k*-means clusters.

Additional file 6: Excel spreadsheet containing information on DEGs with binding sites for floral organ identity factors. The gene identifiers, aliases, descriptions, and their assignment to one of the fifteen *k*-means clusters are indicated. Furthermore, the presence of binding sites for floral organ identity factors in the putative promoters of the genes are shown. To this end, data for AP1 and SEP3 from Pajoro et al. [14], for AP3 and PI from Wuest et al. [12], and for AG from O'Maoileidigh et al. [13] have been used. Whether or not a gene has been described previously [12, 13] as a direct target of AP3, PI, and AG is shown as well.

Additional file 7: Microsoft Excel spreadsheet containing information relating to paralogs identified in groups of co-expressed genes. For each gene identified as having paralogs in a given *k*-means cluster (as indicated), the gene identifiers, aliases, and descriptions, as well as the paralogous genes are shown.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

PTR, DSÓM, IG, EG, MQ, and FW designed the study; PTR, H-GD, and AG performed data analysis; DSÓM and KK conducted experiments; and all authors contributed to writing the paper. All authors read and approved the final manuscript.

Acknowledgements

This study was supported by grants from Science Foundation Ireland to F.W. (10/IN.1/B2971) and E.G. (09/SIRG/B1600), and the Deutsche Forschungsgemeinschaft to M.Q. (Qu 141/6-1).

Author details

[1]Smurfit Institute of Genetics, Trinity College Dublin, Dublin, Ireland. [2]Institute of Computer Science, Martin Luther University Halle–Wittenberg, Halle (Saale), Germany. [3]Department of Biology, National University of Ireland Maynooth, Maynooth, Ireland. [4]Leibniz Institute of Plant Biochemistry, Department of Molecular Signal Processing, Halle (Saale), Germany. [5]Present address: Max Planck Institute for Plant Breeding Research, D-50829 Cologne, Germany.

Received: 7 February 2015 Accepted: 15 June 2015
Published online: 01 July 2015

References

1. Hirano HY, Tanaka W, Toriba T. Grass flower development. Methods Mol Biol. 2014;1110:57–84.
2. Causier B, Davies B. Flower development in the asterid lineage. Methods Mol Biol. 2014;1110:35–55.
3. Prunet N, Jack TP. Flower development in Arabidopsis: there is more to it than learning your ABCs. Methods Mol Biol. 2014;1110:3–33.
4. O'Maoileidigh DS, Graciet E, Wellmer F. Gene networks controlling Arabidopsis thaliana flower development. New Phytol. 2014;201(1):16–30.
5. Wellmer F, Graciet E, Riechmann JL. Specification of floral organs in Arabidopsis. J Exp Bot. 2014;65(1):1–9.
6. Sablowski R: Control of patterning, growth, and differentiation by floral organ identity genes. J Exp Bot 2015;66(4):1065–72.
7. Smyth DR, Bowman JL, Meyerowitz EM. Early flower development in Arabidopsis. Plant Cell. 1990;2(8):755–67.
8. Mantegazza O, Gregis V, Chiara M, Selva C, Leo G, Horner DS, et al. Gene coexpression patterns during early development of the native Arabidopsis reproductive meristem: novel candidate developmental regulators and patterns of functional redundancy. Plant J. 2014;79(5):861–77.
9. Wellmer F, Alves-Ferreira M, Dubois A, Riechmann JL, Meyerowitz EM. Genome-wide analysis of gene expression during early Arabidopsis flower development. PLoS Genet. 2006;2(7):e117.
10. Gomez-Mena C, de Folter S, Costa MM, Angenent GC, Sablowski R. Transcriptional program controlled by the floral homeotic gene AGAMOUS during early organogenesis. Development. 2005;132(3):429–38.
11. Jiao Y, Meyerowitz EM. Cell-type specific analysis of translating RNAs in developing flowers reveals new levels of control. Mol Syst Biol. 2010;6:419.
12. Wuest SE, O'Maoileidigh DS, Rae L, Kwasniewska K, Raganelli A, Hanczaryk K, et al. Molecular basis for the specification of floral organs by APETALA3 and PISTILLATA. Proc Natl Acad Sci U S A. 2012;109(33):13452–7.
13. Maoileidigh DS O, Wuest SE, Rae L, Raganelli A, Ryan PT, Kwasniewska K, et al. Control of reproductive floral organ identity specification in Arabidopsis by the C function regulator AGAMOUS. Plant Cell. 2013;25(7):2482–503.
14. Pajoro A, Madrigal P, Muino JM, Matus JT, Jin J, Mecchia MA, et al. Dynamics of chromatin accessibility and gene regulation by MADS-domain transcription factors in flower development. Genome Biol. 2014;15(3):R41.
15. Alves-Ferreira M, Wellmer F, Banhara A, Kumar V, Riechmann JL, Meyerowitz EM. Global expression profiling applied to the analysis of Arabidopsis stamen development. Plant Physiol. 2007;145(3):747–62.
16. Wellmer F, Riechmann JL, Alves-Ferreira M, Meyerowitz EM. Genome-wide analysis of spatial gene expression in Arabidopsis flowers. Plant Cell. 2004;16(5):1314–26.
17. Schmid M, Davison TS, Henz SR, Pape UJ, Demar M, Vingron M, et al. A gene expression map of Arabidopsis thaliana development. Nat Genet. 2005;37(5):501–6.
18. Zhang X, Feng B, Zhang Q, Zhang D, Altman N, Ma H. Genome-wide expression profiling and identification of gene activities during early flower development in Arabidopsis. Plant Mol Biol. 2005;58(3):401–19.

19. Peiffer JA, Kaushik S, Sakai H, Arteaga-Vazquez M, Sanchez-Leon N, Ghazal H, et al. A spatial dissection of the Arabidopsis floral transcriptome by MPSS. BMC Plant Biol. 2008;8:43.
20. Wuest SE, Vijverberg K, Schmidt A, Weiss M, Gheyselinck J, Lohr M, et al. Arabidopsis female gametophyte gene expression map reveals similarities between plant and animal gametes. Curr Biol. 2010;20(6):506–12.
21. Sanchez-Leon N, Arteaga-Vazquez M, Alvarez-Mejia C, Mendiola-Soto J, Duran-Figueroa N, Rodriguez-Leal D, et al. Transcriptional analysis of the Arabidopsis ovule by massively parallel signature sequencing. J Exp Bot. 2012;63(10):3829–42.
22. Pina C, Pinto F, Feijo JA, Becker JD. Gene family analysis of the Arabidopsis pollen transcriptome reveals biological implications for cell growth, division control, and gene expression regulation. Plant Physiol. 2005;138(2):744–56.
23. Honys D, Twell D. Transcriptome analysis of haploid male gametophyte development in Arabidopsis. Genome Biol. 2004;5(11):R85.
24. O'Maoileidigh DS, Wellmer F. A floral induction system for the study of early Arabidopsis flower development. Methods Mol Biol. 2014;1110:307–14.
25. O'Maoileidigh DS, Thomson B, Raganelli A, Wuest SE, Ryan PT, Kwasniewska K, et al. Gene network analysis in Arabidopsis thaliana flower development through dynamic gene perturbations. Plant J. 2015;82.
26. Bowman JL, Alvarez J, Weigel D, Meyerowitz EM, Smyth DR. Control of flower development in Arabidopsis thaliana by APETALA1 and interacting genes. Development. 1993;119:721–43.
27. Ferrandiz C, Gu Q, Martienssen R, Yanofsky MF. Redundant regulation of meristem identity and plant architecture by FRUITFULL, APETALA1 and CAULIFLOWER. Development. 2000;127(4):725–34.
28. Arbeitman MN, Furlong EE, Imam F, Johnson E, Null BH, Baker BS, et al. Gene expression during the life cycle of Drosophila melanogaster. Science. 2002;297(5590):2270–5.
29. Kaufmann K, Wellmer F, Muino JM, Ferrier T, Wuest SE, Kumar V, et al. Orchestration of floral initiation by APETALA1. Science. 2010;328(5974):85–9.
30. Yanofsky MF, Ma H, Bowman JL, Drews GN, Feldmann KA, Meyerowitz EM. The protein encoded by the Arabidopsis homeotic gene agamous resembles transcription factors. Nature. 1990;346(6279):35–9.
31. Jack T, Brockman LL, Meyerowitz EM. The homeotic gene APETALA3 of Arabidopsis thaliana encodes a MADS box and is expressed in petals and stamens. Cell. 1992;68(4):683–97.
32. Liu C, Zhou J, Bracha-Drori K, Yalovsky S, Ito T, Yu H. Specification of Arabidopsis floral meristem identity by repression of flowering time genes. Development. 2007;134(10):1901–10.
33. Fletcher JC, Brand U, Running MP, Simon R, Meyerowitz EM. Signaling of cell fate decisions by CLAVATA3 in Arabidopsis shoot meristems. Science. 1999;283(5409):1911–4.
34. Payne T, Johnson SD, Koltunow AM. KNUCKLES (KNU) encodes a C2H2 zinc-finger protein that regulates development of basal pattern elements of the Arabidopsis gynoecium. Development. 2004;131(15):3737–49.
35. Sun B, Xu YF, Ng KH, Ito T. A timing mechanism for stem cell maintenance and differentiation in the Arabidopsis floral meristem. Gene Dev. 2009;23(15):1791–804.
36. Sakai H, Medrano LJ, Meyerowitz EM. Role of SUPERMAN in maintaining Arabidopsis floral whorl boundaries. Nature. 1995;378(6553):199–203.
37. Pinyopich A, Ditta GS, Savidge B, Liljegren SJ, Baumann E, Wisman E, et al. Assessing the redundancy of MADS-box genes during carpel and ovule development. Nature. 2003;424(6944):85–8.
38. Brownfield L, Hafidh S, Borg M, Sidorova A, Mori T, Twell D. A plant germline-specific integrator of sperm specification and cell cycle progression. PLoS Genet. 2009;5(3):e1000430.
39. Sorensen AM, Krober S, Unte US, Huijser P, Dekker K, Saedler H. The Arabidopsis ABORTED MICROSPORES (AMS) gene encodes a MYC class transcription factor. Plant J. 2003;33(2):413–23.
40. Schiefthaler U, Balasubramanian S, Sieber P, Chevalier D, Wisman E, Schneitz K. Molecular analysis of NOZZLE, a gene involved in pattern formation and early sporogenesis during sex organ development in Arabidopsis thaliana. Proc Natl Acad Sci U S A. 1999;96(20):11664–9.
41. Yang WC, Ye D, Xu J, Sundaresan V. The SPOROCYTELESS gene of Arabidopsis is required for initiation of sporogenesis and encodes a novel nuclear protein. Genes Dev. 1999;13(16):2108–17.
42. Canales C, Bhatt AM, Scott R, Dickinson H. EXS, a putative LRR receptor kinase, regulates male germline cell number and tapetal identity and promotes seed development in Arabidopsis. Curr Biol. 2002;12(20):1718–27.
43. Zhao DZ, Wang GF, Speal B, Ma H. The excess microsporocytes1 gene encodes a putative leucine-rich repeat receptor protein kinase that controls somatic and reproductive cell fates in the Arabidopsis anther. Genes Dev. 2002;16(15):2021–31.
44. Zhang W, Sun Y, Timofejeva L, Chen C, Grossniklaus U, Ma H. Regulation of Arabidopsis tapetum development and function by DYSFUNCTIONAL TAPETUM1 (DYT1) encoding a putative bHLH transcription factor. Development. 2006;133(16):3085–95.
45. Ito T, Wellmer F, Yu H, Das P, Ito N, Alves-Ferreira M, et al. The homeotic protein AGAMOUS controls microsporogenesis by regulation of SPOROCYTELESS. Nature. 2004;430(6997):356–60.
46. Sanders PM, Bui AQ, Weterings K, McIntire KN, Hsu YC, Lee PY, et al. Anther developmental defects in Arabidopsis thaliana male-sterile mutants. Sex Plant Reprod. 1999;11:297–322.
47. Robinson-Beers K, Pruitt RE, Gasser CS. Ovule Development in Wild-Type Arabidopsis and Two Female-Sterile Mutants. Plant Cell. 1992;4(10):1237–49.
48. Chandler JW. The hormonal regulation of flower development. J Plant Growth Regulation. 2011;30(2):242–54.
49. Yu H, Ito T, Zhao Y, Peng J, Kumar P, Meyerowitz EM. Floral homeotic genes are targets of gibberellin signaling in flower development. Proc Natl Acad Sci U S A. 2004;101(20):7827–32.
50. Coen ES, Meyerowitz EM. The war of the whorls: genetic interactions controlling flower development. Nature. 1991;353(6339):31–7.
51. Bowman JL, Smyth DR, Meyerowitz EM. Genes directing flower development in Arabidopsis. Plant Cell. 1989;1(1):37–52.
52. Causier B, Schwarz-Sommer Z, Davies B. Floral organ identity: 20 years of ABCs. Semin Cell Dev Biol. 2010;21(1):73–9.
53. Smaczniak C, Immink RG, Muino JM, Blanvillain R, Busscher M, Busscher-Lange J, et al. Characterization of MADS-domain transcription factor complexes in Arabidopsis flower development. Proc Natl Acad Sci U S A. 2012;109(5):1560–5.
54. Moore RC, Purugganan MD. The evolutionary dynamics of plant duplicate genes. Curr Opin Plant Biol. 2005;8(2):122–8.
55. Deal RB, Henikoff S. A simple method for gene expression and chromatin profiling of individual cell types within a tissue. Dev Cell. 2010;18(6):1030–40.
56. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol. 2004;3:Article3.
57. Smyth GK, Speed T. Normalization of cDNA microarray data. Methods. 2003;31(4):265–73.
58. Yi X, Du Z, Su Z. PlantGSEA: a gene set enrichment analysis toolkit for plant community. Nucleic Acids Res. 2013;41(Web Server issue):W98–103.
59. Banjamini Y, Yekutieli D. False discovery rate–adjusted multiple confidence intervals for selected parameters. J Am Stat Assoc. 2005;100(469):71–81.
60. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421.

Submit your next manuscript to BioMed Central and take full advantage of:

- **Convenient online submission**
- **Thorough peer review**
- **No space constraints or color figure charges**
- **Immediate publication on acceptance**
- **Inclusion in PubMed, CAS, Scopus and Google Scholar**
- **Research which is freely available for redistribution**

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central

9 Paper 4: Transcriptional Dynamics of Two Seed Compartments with Opposing Roles in Arabidopsis Seed Germination

BJW Dekkers, S Pearce, RP van Bolderen-Veldkamp, A Marshall, P Widera, J Gilbert, **HG Drost**, GW Bassel, K Müller, JR King, AT Wood, I Grosse, M Quint, N Krasnogor, G Leubner-Metzger, MJ Holdsworth, L Bentsink. 2013. Transcriptional dynamics of two seed compartments with opposing roles in *Arabidopsis* seed germination. *Plant Physiology* 163 (1): 205-215. *doi:10.1104/pp.113.223511*

Transcriptional Dynamics of Two Seed Compartments with Opposing Roles in Arabidopsis Seed Germination[1][W][OPEN]

Bas J.W. Dekkers[2]*, Simon Pearce[2], R.P. van Bolderen-Veldkamp, Alex Marshall, Paweł Widera, James Gilbert, Hajk-Georg Drost, George W. Bassel, Kerstin Müller, John R. King, Andrew T.A. Wood, Ivo Grosse, Marcel Quint, Natalio Krasnogor, Gerhard Leubner-Metzger[3], Michael J. Holdsworth[3], and Leónie Bentsink[3]

Department of Molecular Plant Physiology, Utrecht University, NL–3584 CH Utrecht, The Netherlands (B.J.W.D., R.P.v.B.-V., L.B.); Wageningen Seed Laboratory, Laboratory of Plant Physiology, Wageningen University and Research Centre, NL–6708 PB Wageningen, The Netherlands (B.J.W.D., R.P.v.B.-V., L.B.); Department of Plant and Crop Sciences (S.P., K.M., M.J.H.) and Centre for Plant Integrative Biology (S.P., J.R.K., A.T.A.W.), School of Biosciences, University of Nottingham, Sutton Bonington Campus, Loughborough, Leicestershire LE12 5RD, United Kingdom; School of Mathematical Sciences, University of Nottingham, Nottingham NG7 2RD, United Kingdom (S.P., J.R.K., A.T.A.W.); The GenePool, Ashworth Laboratories, University of Edinburgh, Edinburgh EH9 3JT, United Kingdom (A.M.); Interdisciplinary Computing and Complex Systems, School of Computer Science, University of Nottingham, Jubilee Campus, Nottingham NG8 1BB, United Kingdom (P.W., J.G., N.K.); Institute of Computer Science, Martin Luther University Halle-Wittenberg, 06120 Halle (Saale), Germany (H.-G.D., I.G.); School of Biosciences, University of Birmingham, Edgbaston, Birmingham B15 2TT, United Kingdom (G.W.B.); Leibniz Institute of Plant Biochemistry, Department of Molecular Signal Processing, 06120 Halle (Saale), Germany (H.-G.D., M.Q.); and School of Biological Sciences, Plant Molecular Science and Centre for Systems and Synthetic Biology, Royal Holloway, University of London, Egham, Surrey TW20 0EX, United Kingdom (G.L.-M.)

Seed germination is a critical stage in the plant life cycle and the first step toward successful plant establishment. Therefore, understanding germination is of important ecological and agronomical relevance. Previous research revealed that different seed compartments (testa, endosperm, and embryo) control germination, but little is known about the underlying spatial and temporal transcriptome changes that lead to seed germination. We analyzed genome-wide expression in germinating Arabidopsis (*Arabidopsis thaliana*) seeds with both temporal and spatial detail and provide Web-accessible visualizations of the data reported (vseed.nottingham.ac.uk). We show the potential of this high-resolution data set for the construction of meaningful coexpression networks, which provide insight into the genetic control of germination. The data set reveals two transcriptional phases during germination that are separated by testa rupture. The first phase is marked by large transcriptome changes as the seed switches from a dry, quiescent state to a hydrated and active state. At the end of this first transcriptional phase, the number of differentially expressed genes between consecutive time points drops. This increases again at testa rupture, the start of the second transcriptional phase. Transcriptome data indicate a role for mechano-induced signaling at this stage and subsequently highlight the fates of the endosperm and radicle: senescence and growth, respectively. Finally, using a phylotranscriptomic approach, we show that expression levels of evolutionarily young genes drop during the first transcriptional phase and increase during the second phase. Evolutionarily old genes show an opposite pattern, suggesting a more conserved transcriptome prior to the completion of germination.

Seeds are important in the plant life cycle, since they represent the link between two successive generations. They are stress-resistant structures that help to bridge unfavorable periods and allow dispersal. Seed formation starts with a double fertilization event, and in Arabidopsis (*Arabidopsis thaliana*), it takes approximately 20 d to form a mature dry seed (Debeaujon et al., 2007; Ohto et al., 2007). At maturity, three major seed compartments can be distinguished (Holdsworth et al., 2008a; Belmonte et al., 2013): the testa (seed coat), a dead tissue that forms a protective outer layer; the endosperm, a single cell layer of tissue positioned directly underneath the testa; and the embryo (enclosed by the testa and endosperm), which emerges to become the future plant (Rajjou et al., 2012; Fig. 1A). A dry seed is a unique structure in the sense that it allows severe dehydration (desiccation tolerance) and enters a phase of

---

[1] This work was supported by the ERA-NET Plant Genomics program (vSEED grant to B.J.W.D., S.P., R.P.v.B.-V., A.M., K.M., J.R.K., A.T.A.W., G.L.-M., and M.J.H.), by the Dutch Technology Foundation of the Netherlands Organization for Scientific Research and the Technology Program of the Ministry of Economic Affairs (to L.B.), by the Leibniz Association (to M.Q.), by the Engineering and Physical Science Council (grant nos. EP/J004111/1, EP/I031642/1, and EP/H000968/1 to N.K.), and by the Royal Society and the Wolfson Foundation (to J.R.K.).

[2] These authors contributed equally to the article.

[3] These authors contributed equally to the article.

* Address correspondence to bas.dekkers@wur.nl.

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantphysiol.org) is: Bas J.W. Dekkers (bas.dekkers@wur.nl).

[W] The online version of this article contains Web-only data.

[OPEN] Articles can be viewed online without a subscription.

www.plantphysiol.org/cgi/doi/10.1104/pp.113.223511

Downloaded from www.plantphysiol.org on July 4, 2016 - Published by www.plantphysiol.org
Copyright © 2013 American Society of Plant Biologists. All rights reserved.

quiescence, bringing processes occurring in "living" organisms to a halt without affecting viability (Farrant and Moore, 2011; Rajjou et al., 2012). Upon imbibition of water, the dry mature seed swells and metabolic activity resumes, marking the start of seed germination and the end of the quiescent state. Arabidopsis germination consists of two visible sequential events (Holdsworth et al., 2008a; Weitbrecht et al., 2011). First, the testa splits (testa rupture [TR]) due to underlying expansion of the endosperm and embryo. Thereafter, the radicle (RAD; embryonic root) protrudes through the endosperm (endosperm rupture [ER]), completing germination sensu stricto (Fig. 1B).

There are two nonexclusive mechanisms proposed to explain seed germination (Nonogaki, 2006; Nonogaki et al., 2007). The first involves the increase in embryo growth potential leading to elongation of the proximal embryonic axis (hypocotyl and RAD) that overcomes the restraint of the covering tissues. The second involves the weakening of these covering layers (including the micropylar endosperm, positioned over the RAD tip; Fig. 1A) to ease the protrusion of the RAD (for review, see Finch-Savage and Leubner-Metzger, 2006). The endosperm has been shown to affect germination even in species with a thin endosperm layer, such as Arabidopsis (Müller et al., 2006; Bethke et al., 2007; Lee et al., 2010). Genome-wide expression studies have been previously applied to gain insight into several aspects of seed biology (Holdsworth et al., 2008a, 2008b; Le et al., 2010), including temporal changes during Arabidopsis germination (Nakabayashi et al., 2005; Preston et al., 2009; Narsai et al., 2011) and in spatial differences between embryo and endosperm (Penfield et al., 2006; Endo et al., 2012). Nevertheless, a detailed knowledge of the temporal changes in gene expression in the different compartments of the Arabidopsis seed is thus far missing, but it is essential to understanding the control of the timing of germination as well as the underlying molecular processes contributed by these different seed compartments. Therefore, we have analyzed the Arabidopsis transcriptome by sampling 11 points along the germination time course,

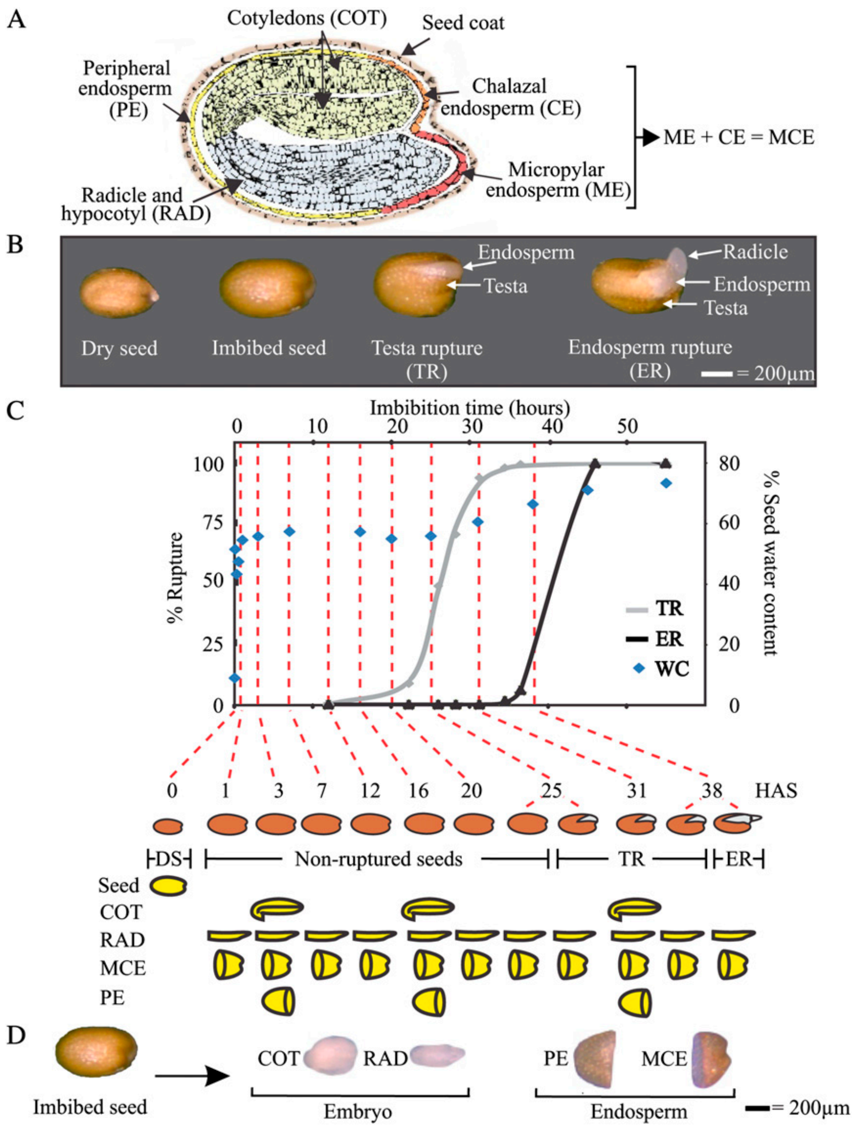


Figure 1. Seed compartments and seed germination kinetics of Arabidopsis seeds. A, A section through an Arabidopsis seed depicting the different seed compartments. B, Different stages during seed germination including TR (which exposes the underlying endosperm layer) and ER (also known as RAD protrusion or germination sensu stricto). C, Arabidopsis seed germination analyzed by measuring TR (gray line), ER (black line), and seed water content (WC; blue diamonds). Below the graph, the time points and physiological stages (dry, NR, TR, and ER) are indicated for each sample. The 29 samples that were analyzed are schematically shown below the germination graph by the yellow pictograms. D, The four seed sections that were used for transcriptome analysis.

Downloaded from www.plantphysiol.org on July 4, 2016 - Published by www.plantphysiol.org
Copyright © 2013 American Society of Plant Biologists. All rights reserved.

including those that allow an analysis of gene expression changes at the key events of germination (TR and ER), with a focus on the micropylar endosperm and the RAD.

RESULTS AND DISCUSSION

Arabidopsis Seed Imbibition, Germination Kinetics, and Transcriptome Analyses

We characterized Arabidopsis seed germination by scoring TR and ER over time. TR started around 20 h after sowing (HAS), and at 31 HAS almost all seeds were fully ruptured. From 31 HAS onward, ER was observed, which was completed in the entire seed population by 45 HAS (Fig. 1C). Microarray experiments were performed using dry seeds and seeds at nine time points along the germination time course until the completion of germination (Fig. 1C). The time points 25 and 38 HAS showed a mixture of nonruptured (NR) and TR seeds and TR and ER seeds, respectively; at these time points, both classes were separated and collected as distinct samples, which enabled us to map the transcriptome changes induced by TR and ER. To capture spatial dynamics, imbibed seeds were dissected into four parts.

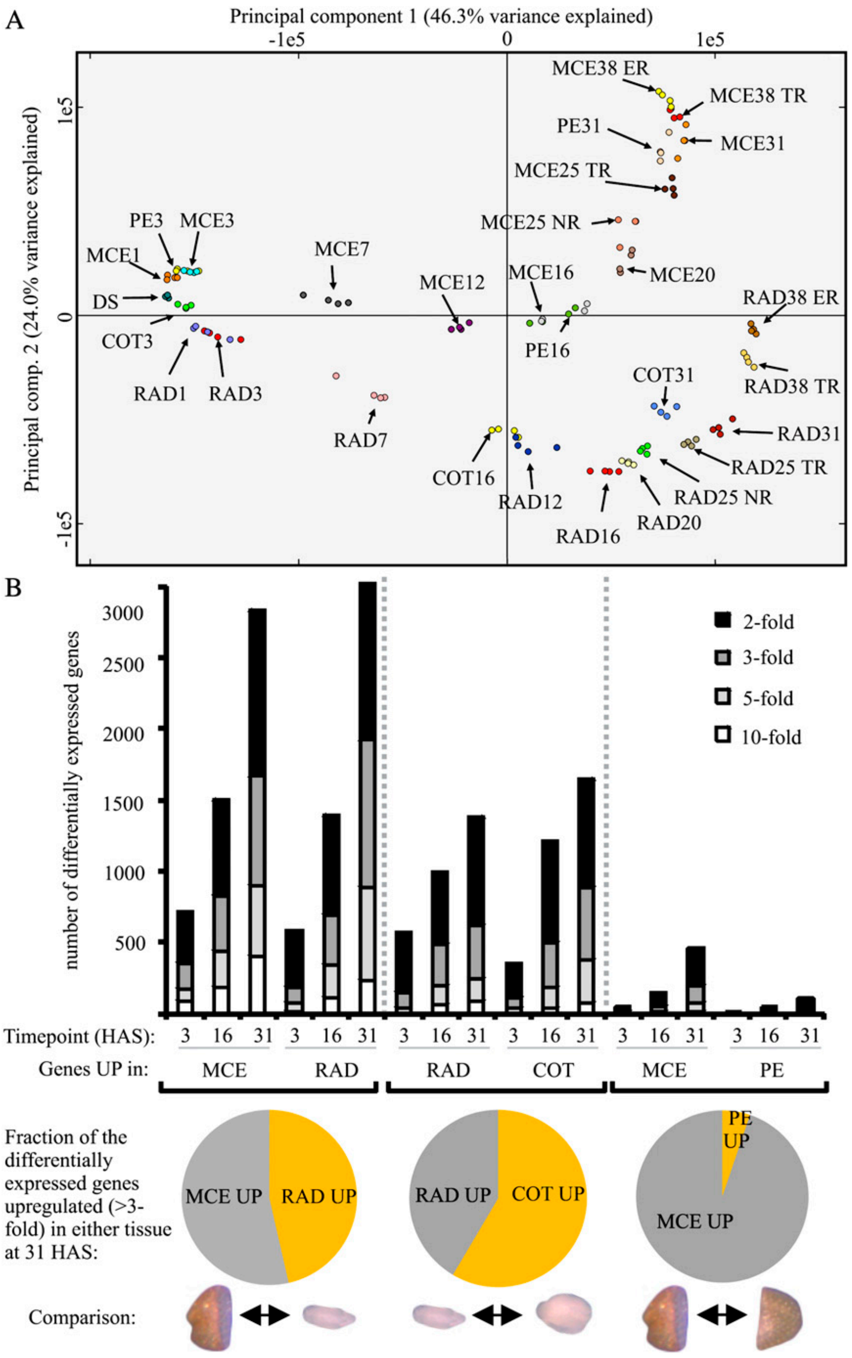
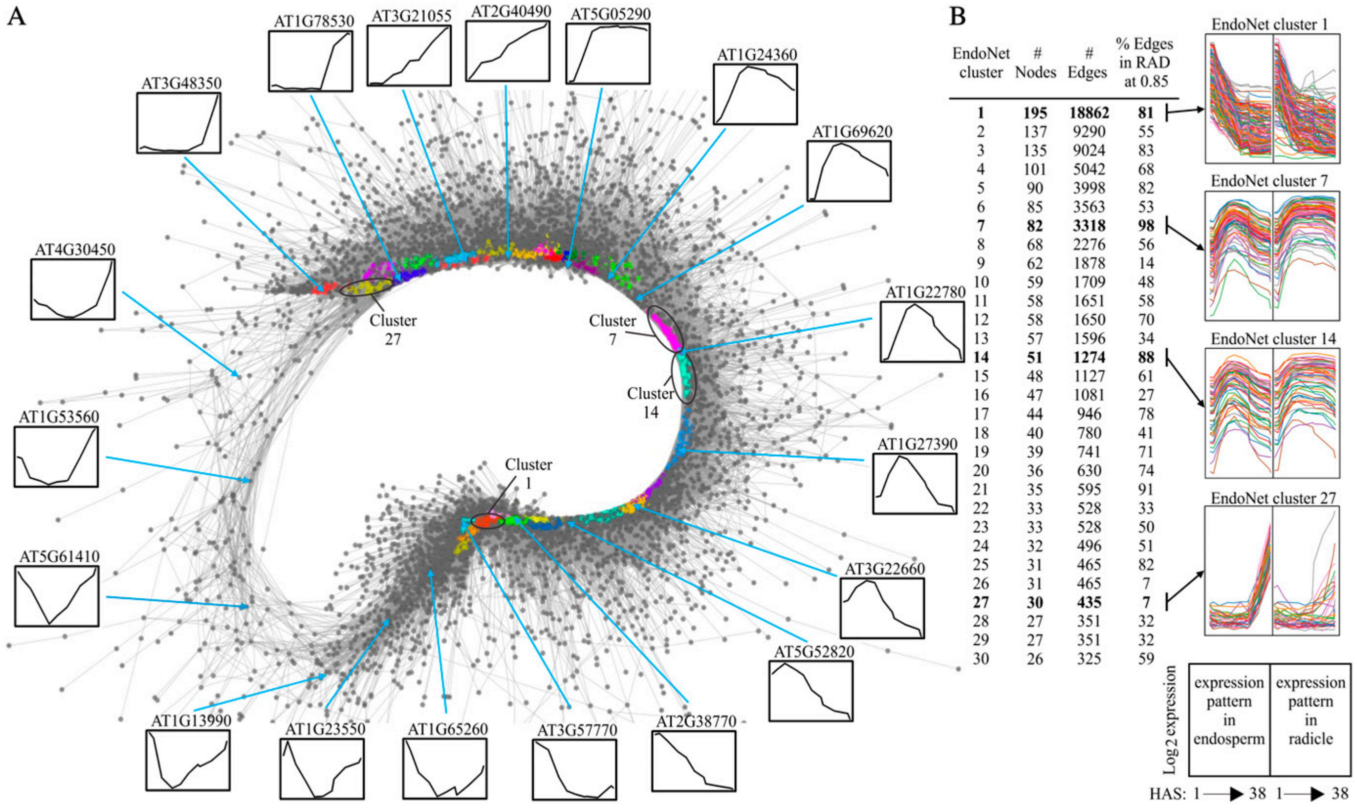


Figure 2. Transcriptional differences between seed compartments. A, PCA of the 116 samples. The four replicates of all 29 samples are indicated by color. B, Tissue differences are represented by the number of differentially expressed genes at three time points during imbibition (3, 16, and 31 HAS, the time points in which all four tissues were sampled). Comparisons were made between endosperm and embryo (MCE versus RAD), between embryo tissues (RAD versus COT), and between both endosperm samples (MCE versus PE). The bars show the number of differentially expressed genes at a 2-, 3-, 5-, and 10-fold cutoff. The pie diagrams below the graph indicate the fraction of the total number of differentially expressed genes (at a 3-fold cutoff level) in either of the two tissues that were compared at 31 HAS.

Downloaded from www.plantphysiol.org on July 4, 2016 - Published by www.plantphysiol.org
Copyright © 2013 American Society of Plant Biologists. All rights reserved.

The key compartments for germination, the RAD (including a large part of the hypocotyl to ensure that it encompasses the region that elongates [Sliwinska et al., 2009]) and the micropylar end of the endosperm (which is a combination of micropylar and chalazal endosperm [MCE]), were sampled at all time points. At three time points (3, 16, and 31 HAS), the cotyledons (COT) and the remainder of the endosperm (peripheral endosperm [PE]) were collected (Fig. 1, A, C, and D; Supplemental Fig. S1). The 29 samples, with four replicates for each sample, were analyzed using Affymetrix ATH1 gene chips. Plotting probe set values in a histogram showed clearly distinguishable peaks for noise and signal and revealed that an appropriate cutoff for considering a gene as potentially expressed was 5 on a $\log_2$ scale (Supplemental Fig. S1). The percentage of genes detected in the different seed compartments was within the same range described for other Arabidopsis seed transcriptome analyses (Nakabayashi et al., 2005; Penfield et al., 2006; Belmonte et al., 2013). In total, 14,317 genes (67.2% of the 21,313 genes on the chip) were found to be expressed at least once in the 29 samples, of which 11,298 (78.8%) were shared between all compartments (Supplemental Fig. S2A).

At the start of the time course, a lower number of genes were found to be expressed, and this number increased during the time course in all tissues, most notably during the first 12 to 16 HAS (Supplemental Fig. S2B). We identified gene sets that were tissue specifically expressed by considering genes as specifically expressed in one tissue when expressed above 6 (on a $\log_2$ scale) in that tissue and expressed below 5 in all the other tissues (which, therefore, is in the noise region). This resulted in 415 genes specific to the endosperm and 546 genes specific to the embryo in our data set (Supplemental Fig. S2; Supplemental Data Set S1), which overlaps with previously published data sets (Penfield et al., 2006; Le et al., 2010; Supplemental Fig. S3). In total, 12,856 genes are expressed above 6 in either tissue, with 10,801 expressed above 6 in both tissues. Thus, according to this definition, 84.01% of the genes are shared between both tissues, while 3.22% are specific to the embryo and 4.24% are specific to the endosperm. The remaining genes (8.53%) are expressed over 6 in one tissue but between 5 and 6 in another tissue and so are not classed as being highly specific to any one tissue. Interrogation using overrepresentation analysis (ORA) revealed that the endosperm gene set was



EndoNet cluster	# Nodes	# Edges	% Edges in RAD at 0.85
1	**195**	**18862**	**81**
2	137	9290	55
3	135	9024	83
4	101	5042	68
5	90	3998	82
6	85	3563	53
7	**82**	**3318**	**98**
8	68	2276	56
9	62	1878	14
10	59	1709	48
11	58	1651	58
12	58	1650	70
13	57	1596	34
14	**51**	**1274**	**88**
15	48	1127	61
16	47	1081	27
17	44	946	78
18	40	780	41
19	39	741	71
20	36	630	74
21	35	595	91
22	33	528	33
23	33	528	50
24	32	496	51
25	31	465	82
26	31	465	7
27	**30**	**435**	**7**
28	27	351	32
29	27	351	32
30	26	325	59

Figure 3. The endosperm coexpression network, EndoNet. A, Sample layout of EndoNet. The nodes (genes) are indicated by gray circles, and edges (gray lines) are drawn between two nodes if their correlation of expression is above 0.932. The 30 largest clusters are indicated by different colors. To visualize the gene expression profiles captured in the network, the expression profiles of exemplar genes are shown around the network. B, Details of the largest 30 clusters are shown, including the number of nodes, edges, and the percentage of edges that are shared with RadNet (at a cutoff of 0.85). The expression profiles of genes in the EndoNet clusters 1, 7, 12, and 27 are shown (the positions of these clusters in EndoNet are shown A). The right side of the graph depicts the expression profiles of the same set of genes in the RAD samples.

Downloaded from www.plantphysiol.org on July 4, 2016 - Published by www.plantphysiol.org
Copyright © 2013 American Society of Plant Biologists. All rights reserved.

overrepresented for genes related to response to abscisic acid, defense response, cell wall macromolecule metabolism/catabolism, and cell death as well as genes associated with the regulation of transcription (Supplemental Fig. S2D), in agreement with recent findings (Endo et al., 2012). In the embryo, the largest class was related to plant development. Other Gene Ontology (GO) classes that were overrepresented included cell division, hormone metabolic process, protein amino acid phosphorylation, signaling, and regulation of transcription (Supplemental Fig. S2E). Thus, different GO classes were found to be overrepresented in each tissue, with regulation of transcription/gene expression appearing in both. Both tissue-specific gene sets are enriched for transcription factors (Supplemental Fig. S2F). In the endosperm, transcription factors of NAC, WRKY, and C3H classes, and in the embryo, transcription factors of bHLH, G2-like, and HB classes, are particularly enriched (Supplemental Fig. S2F). Compartment-specific gene sets containing 106, 47, 21, and two genes were identified for the RAD, COT, MCE, and PE (Supplemental Fig. S2; Supplemental Data Set S1), respectively, and quantitative reverse transcription-PCR confirmed the compartment-specific expression of 20 genes (Supplemental Fig. S4).

In order to globally compare gene expression between the samples, all 116 arrays were plotted using principal component analysis (PCA; Fig. 2A). In general, the largest transcriptome differences were observed between the endosperm and embryo (MCE versus RAD) followed by the comparison between both embryo parts (RAD versus COT). The smallest differences were found between both endosperm (MCE versus PE) parts (Fig. 2). The quality controls (Supplemental Fig. S1), the high correlation between the replicates (Supplemental Table S1), and the confirmation by quantitative reverse transcription-PCR of compartment-specific expression (Supplemental Fig. S4) indicate that this is a robust data set revealing transcriptome changes during seed rehydration and the developmental switch from a quiescent dry seed to germination in both temporal and spatial detail.

Generation of Coexpression Networks and Data Visualization Tools

We generated coexpression networks (Bassel et al., 2011) for the endosperm (EndoNet) and the RAD samples (RadNet). We identified compact clusters of genes in the networks (Supplemental Data Set S1) that were further scrutinized with the network topological analyzer, TopoGSA (http://www.topogsa.net/; Glaab et al., 2010; Supplemental Fig. S5). Interactive visualizations of both networks are available online at http://vseed.nottingham.ac.uk. Compared with our previous visualization tool (Bassel et al., 2011), these visualizations offer improved performance and more advanced gene selection options, such as the highlighting of individual genes or entire clusters, searching for genes by name or descriptive keywords, and visualization of gene expression using our new Electronic Fluorescent Pictograph browser (Winter et al., 2007).

EndoNet shows a ring-like display, a result of the scarcity of genes with constant expression (Fig. 3A). This indicates that the regulation of gene expression is very dynamic in the endosperm during germination. The largest 30 EndoNet clusters are spread around the network and thus represent the major gene expression profiles. ORA revealed cluster-specific overrepresentation of specific biological processes (Supplemental Fig. S6). These clusters consist of 26 to 195 genes and contain at least 99.7% of all possible edges within them (Fig. 3), indicating that genes within such clusters have very similar expression patterns. Genes of some clusters (e.g. EndoNet cluster 1) are also coexpressed in RadNet (81% of the edges in cluster 1 are also found in RadNet at a 0.85 correlation) and show similar expression patterns in both compartments, while other genes (such as EndoNet cluster 27) show an endosperm-specific expression pattern and have few edges in common with RadNet (Fig. 3B). On the other hand, almost all connections in EndoNet clusters 7 and 14 (98% and 88%, respectively) are also present in RadNet (Fig. 3B). Despite strong coexpression between both networks, the expression profiles in these clusters are different between the two compartments, being induced in both but subsequently repressed in the endosperm.

Arabidopsis Seed Germination Is Composed of Two Transcriptional Phases

Analyzing the transcriptional dynamics between consecutive time points of the germination time course

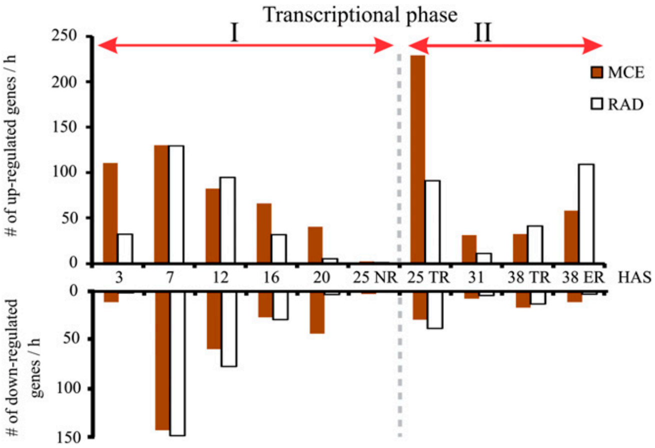


Figure 4. Arabidopsis seed germination is characterized by two transcriptional phases. The number of differentially expressed genes (both up- and down-regulated) between consecutive time points (3 was compared with 1, 7 with 3, 12 with 7, etc.) in the MCE (white bars) and RAD (brown bars) with a reasonable fold change (taking a 3-fold difference as the cutoff) are presented. The two transcriptional phases, phase I from 1 to 25 HAS NR and phase II from 25 HAS NR to 38 HAS ER, are indicated by the red arrows.

Downloaded from www.plantphysiol.org on July 4, 2016 - Published by www.plantphysiol.org
Copyright © 2013 American Society of Plant Biologists. All rights reserved.

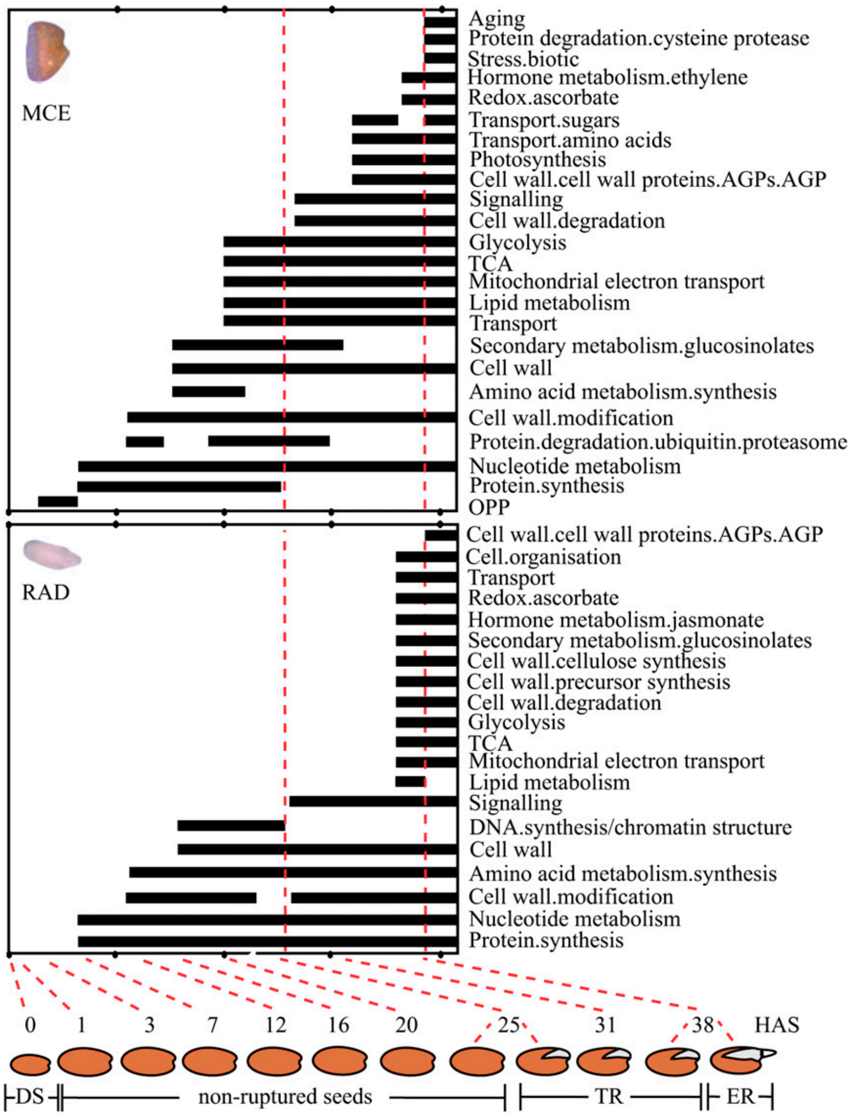
revealed two transcriptional phases (Fig. 4). The first phase runs from 1 to 25 HAS NR and is characterized by large transcriptional changes in both up- and down-regulated genes. At the end of this first phase, the number of differentially expressed genes was reduced (Fig. 4). The second phase, which runs from TR to the completion of germination, was marked by resumption of differential gene expression, most notably at TR. During the second phase, the majority of the differentially expressed genes are induced rather than repressed, in contrast to the first phase.

The First Transcriptional Phase Is Characterized by an Inversion of the Seed Maturation Transcriptional Program

Between 1 and 3 HAS, differential gene expression was observed, particularly in the MCE (Fig. 4). In comparison, the response of the RAD was delayed, which could be due to its slower imbibition kinetics compared with the more outward-positioned MCE (Fig. 4). Large transcriptional changes occurred in the first 16 HAS. ORA of this phase suggests a large overlap in the functional classes that are activated in the MCE and RAD (i.e. genes related to cell wall function, nucleotide metabolism, amino acid metabolism, and protein translation; Fig. 5). A major difference is the activation of classes related to transport and energy metabolism (lipid metabolism, glycolysis, TCA, and mitochondrial electron transport) that are specifically activated in the MCE from 20 HAS, in agreement with findings that storage lipids are more rapidly mobilized in the endosperm compared with the embryo (Penfield et al., 2005).

We compared gene expression during seed germination with gene expression during seed development and identified two gene sets containing 602 and 907 genes (Supplemental Data Set S1) that were strongly up- and down-regulated, respectively, between the embryo COT phase (early seed maturation) and the postmature

Figure 5. Temporal differences between endosperm and embryo using ORA. The overrepresented gene categories of the up-regulated genes of the germination time course (all time points were compared with 1 HAS) were identified in the MCE (top graph) and the RAD (bottom graph) using PageMan (Usadel et al., 2006). Selected categories are summarized in the graphs, and black bars show the time points during germination at which the indicated gene categories are overrepresented. OPP, Oxidative pentose phophate pathway.



Downloaded from www.plantphysiol.org on July 4, 2016 - Published by www.plantphysiol.org
Copyright © 2013 American Society of Plant Biologists. All rights reserved.

green stage (late maturation) from a publicly available data set (Le et al., 2010). The expression of the two gene sets was analyzed during germination, and the majority of the genes of both sets showed inverse expression patterns during seed germination (Fig. 6). The largest overlap (75%) was found between genes that were up-regulated during seed maturation and those down-regulated during germination. Additionally, 67% of the genes from the set that were down-regulated during seed maturation showed an inverse expression pattern (were induced) during germination. The reinduction of these seed maturation down-regulated genes during germination was slower than the removal of the seed maturation-induced genes. Nevertheless, the majority of the seed maturation-repressed genes were reactivated in the first transcriptional phase rather than the second transcriptional phase.

TR Is Marked by High Transcriptional Activity That Overlaps in Part with a Response to Touch-Induced Signaling

TR is characterized by a large number of differentially expressed genes when compared with NR seeds at 25 HAS, mostly genes that are up-regulated in the MCE (Fig. 7A). At TR, 104 genes were over 5-fold up-regulated in the MCE (Supplemental Data Set S1), 30 of which are related to cell wall function. Other classes induced by TR in the MCE include genes related to biotic stress, hormone metabolism, regulation of transcription, signaling (receptor kinases), and transport (Fig. 7B). Possible reasons for these large transcriptional changes between NR and TR seeds include an enhanced access to oxygen, light signaling, and/or a touch (mechano)-sensing response (Fig. 7C). ORA did not reveal a clear indication of the involvement of either oxygen or light. However, the gene set included *TOUCH3* and *TOUCH4* (both more than 8-fold induced), which are known to respond rapidly to touch (Braam, 2005; Fig. 7D). To investigate whether the transcriptional up-regulation at TR resembles touch sensing, we compared our MCE TR up-regulated data set with genes up-regulated upon touch in aerial parts of plants (Lee et al., 2005). We reanalyzed a published touch data set (Lee et al., 2005; Supplemental Materials and Methods S1) and found a 30% overlap with our TR-induced set in the MCE and the touch up-regulated genes, with a lower overlap between the touch data set and the TR-induced genes in the RAD (Fig. 7E). The overlap between the gene sets induced by TR in the MCE and touch was more striking when the gene classes were considered. Touch-induced signaling resulted in a relatively higher abundance of genes related to the GO classes cell wall associated, calcium binding, disease resistance, kinase, and transcription factor (Lee et al., 2005), which match well with the classes identified at TR (Fig. 7B). We also observed that gene expression associated with jasmonate biosynthesis was activated upon TR in the MCE; this plant hormone was recently shown to be a key regulator of plant morphogenesis and enhanced pest resistance upon touch (Chehab et al., 2012). It has been hypothesized that gene expression in the endosperm during germination might be affected by touch/mechano sensing (Martínez-Andújar et al., 2012), and this transcriptome study provides a strong suggestion that touch signaling is indeed, at least in part, responsible for the induction of gene expression in the endosperm.

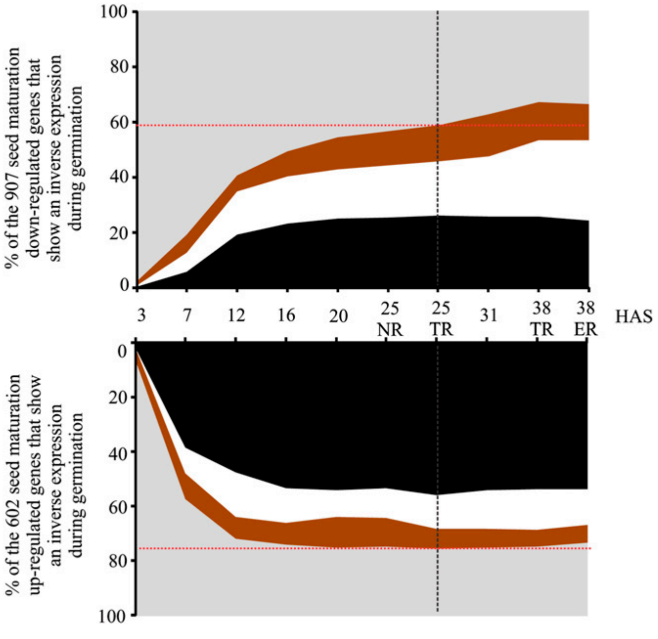
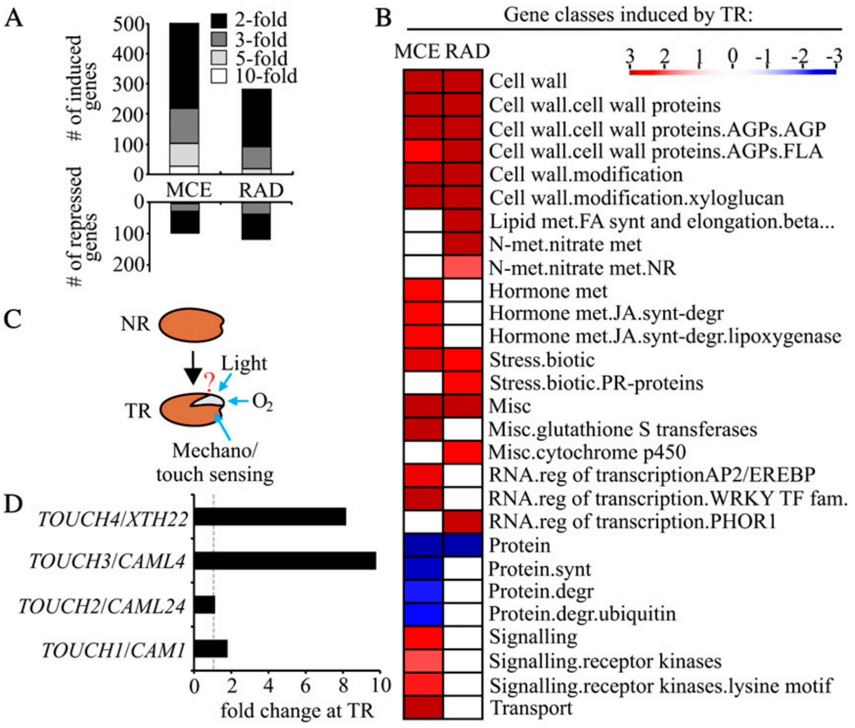


Figure 6. Inverse expression of seed maturation genes during germination in temporal and spatial detail. The top panel shows the percentage of up-regulated genes during germination among a set of 907 genes that are down-regulated during seed maturation. The bottom panel shows the percentage of down-regulated genes during germination among a set of 602 genes that are up-regulated during seed maturation. Genes expressed specifically in the MCE (in brown), in the RAD (in white), and in both (in black) are indicated.

The Second Transcriptional Phase Highlights Distinct Fates for the Embryo and the Endosperm

The second transcriptional phase starts at TR and includes gene expression changes related to the completion of germination. Using ORA, we analyzed the temporal changes in the MCE and the RAD (Fig. 5) as well as gene sets that are more highly expressed within the MCE or RAD along the time course (Supplemental Fig. S7). This revealed that, in the MCE genes related to secondary metabolism, amino acid metabolism and protein synthesis are overrepresented transiently (Fig. 5). Genes more highly expressed in the MCE than the RAD are enriched for protein degradation, transport, and stress-related genes (although the latter are overrepresented in the MCE over the whole time course; Supplemental Fig. S7). The RAD, particularly at the

Downloaded from www.plantphysiol.org on July 4, 2016 - Published by www.plantphysiol.org
Copyright © 2013 American Society of Plant Biologists. All rights reserved.



E Percentage of TR-induced genes that overlap with touch-induced gene set.

	Fold change 2	Fold change 3	Fold change 5
MCE	**23.11%** (116/502)	**29.26%** (67/229)	**30.40%** (31/102)
RAD	**21.99%** (62/282)	**18.68%** (17/91)	**5.56%** (1/18)

Percentage expected by chance considering:

All genes on chip (21,313)	All genes expressed in time-course (14,317)	All genes expressed in MCE (13,085)	All genes expressed in RAD (13,279)
4.11%	**5.38%**	**5.33%**	**5.46%**

Figure 7. Genes induced with respect to TR show a large overlap with touch-induced signaling. A, Number of differentially expressed genes at 25 HAS TR (compared with 25 HAS NR) in the MCE and RAD at different fold change cutoffs. B, Gene classes overrepresented in the TR-induced gene sets in the MCE and RAD. C, Schematic presentation of effectors that could be responsible for the large gene expression changes observed at TR. D, Expression behavior of four *TOUCH* genes at TR in the MCE. E, Table shows the percentage of the TR up-regulated genes in the MCE and the RAD (at 2-, 3-, and 5-fold cutoff) that overlap with the 934 touch up-regulated genes. The percentage expected by chance is indicated using the number of genes present on the chip, genes expressed in the germination time course, genes expressed in the MCE, and genes expressed in the RAD. degr, Degradation; FA, fatty acid; fam, family; FLA, fasciclin-like arabinogalactan; JA, jasmonic acid; met, metabolism; misc, miscellaneous; NR, non-ruptured; PR, pathogenesis-related; reg, regulation; synt, synthesis; TF, transcription factor.

later stage, is enriched for cellular metabolism related to DNA, RNA, and proteins compared with the MCE (Supplemental Fig. S7). ORA suggests that energy metabolism (lipid metabolism, glycolysis, TCA, and mitochondrial electron transport) is activated by 38 HAS. At this stage, genes for cell wall biosynthesis, transport, and secondary metabolism are activated, notably just prior to ER (Fig. 5). In addition, genes related to the cell cycle and lipid and amino acid metabolism are overrepresented within genes more highly expressed in the RAD than the MCE (Supplemental Fig. S7), which are all classes supporting tissue growth. The GO gene class "aging" becomes overrepresented in the latter part of the germination time course in the MCE (Fig. 5; Supplemental Fig. S7). This is in agreement with the down-regulation of key cellular metabolic pathways and the induction of gene classes related to remobilization, reminiscent of the transcriptional changes described for senescence (Lim et al., 2007; Breeze et al., 2011).

The Transition from a Dry Quiescent to a Hydrated and Germinating Seed Coincides with Increased Transcriptional Differences between Seed Compartments

From the PCA of all 116 arrays (Fig. 2A), we conclude that the transcriptome differences between seed compartments are small during early germination and increase with time. This is in agreement with the observation that the number of endosperm- and embryo-specific genes expressed increased along the time course from approximately 40 to 400 (Supplemental Fig. S8A). This may be explained by the fact that the majority of genes induced in seed maturation and that are subsequently removed during germination are shared by the MCE and RAD (72%) and that seed maturation-repressed genes (reactivated during germination) are, in contrast, mostly specific to either the RAD or the MCE (Fig. 6). Presumably, the repression of genes related to development and differentiation is a more general response for an organism passing through a desiccated state, as is shown for the expression of genes involved in stomatal development (in the COT

Downloaded from www.plantphysiol.org on July 4, 2016 - Published by www.plantphysiol.org
Copyright © 2013 American Society of Plant Biologists. All rights reserved.

samples) and root development (in the RAD samples). Many of these genes are induced (sometimes transiently) during germination, with low or no expression initially (Supplemental Fig. S8B).

Differential Gene Expression in the Endosperm Is Concentrated at the Micropylar End

The observation that the transcriptional differences increase with time between seed compartments is also shown, besides the PCA, in the number of differentially expressed genes between the seed compartments. The number of differentially expressed genes was least between both endosperm compartments. At 31 HAS, about 200 genes were differentially expressed (more than 3-fold difference), with the majority of these (95%) being up-regulated in the MCE (Fig. 2B) compared with the PE. Such a skewed division was not observed for other comparisons (Fig. 2B). The micropylar endosperm is hypothesized to possess an inhibitory role in germination, and endosperm changes, in particular of cell wall properties, are suggested to be important for germination control (Nonogaki et al., 2007). Recently, using in situ cell wall epitope detection, Arabidopsis endosperm cell walls were shown to have a different structure compared with the embryo cell wall, and the endosperm walls were shown to contain cellulose, unesterified homogalacturonan, arabinan, and xyloglucan polymers (Lee et al., 2012). However, no spatial or temporal heterogeneity in cell wall polymers was observed prior to germination (Lee et al., 2012). This could indicate that cell wall changes leading to germination are modifications that are not detectable by in situ analysis and/or that occur very locally. We compared both endosperm samples and found many differentially expressed genes between the MCE and PE (Supplemental Data Set S1). The largest differences were found close to the point of germination (31 HAS) in the MCE, and this set was investigated for candidates that are potentially involved in ER.

Several transcription factors were found to be highly expressed in the MCE compared with the PE that may function in gene regulation in this particular compartment. Genes related to cell wall function, including peroxidases, a pectin lyase-like superfamily protein, chitinase family protein, and *ARABINOGALACTAN PROTEIN31* were identified in this set, and these could be potential candidates for affecting cell wall properties to enable seed germination. It is notable that one of the most highly differentially expressed (more than 20-fold) genes in the MCE is *INFLORESCENCE DEFICIENT IN ABSCISSION-LIKE1* (*IDL1*). This encodes a putative ligand that promotes cell separation and floral organ abscission via the interaction with receptor-like kinases (Stenvik et al., 2008). Recently, it has been reported that the *INFLORESCENCE DEFICIENT IN ABSCISSION* (*IDA*) peptide and its receptors *HAESA* (*HAE*) and *HAESA-LIKE2* (*HSL2*) are also important for cell separation during lateral root emergence (Kumpf et al., 2013), suggesting that Arabidopsis seed germination may occur via a cell separation event that is potentially regulated by the *IDA*/*IDL-HAE*/*HSL* signaling module. This detailed data set allowed the identification of transcription factors, cell wall-related genes, and genes related to cell separation, although further research is needed to investigate their potential role in seed germination.

Seed Germination Is Characterized by Coordinated Expression of Evolutionarily Old and Young Genes

Recently, it has been shown that, like animal embryogenesis, plant embryogenesis involves a passage

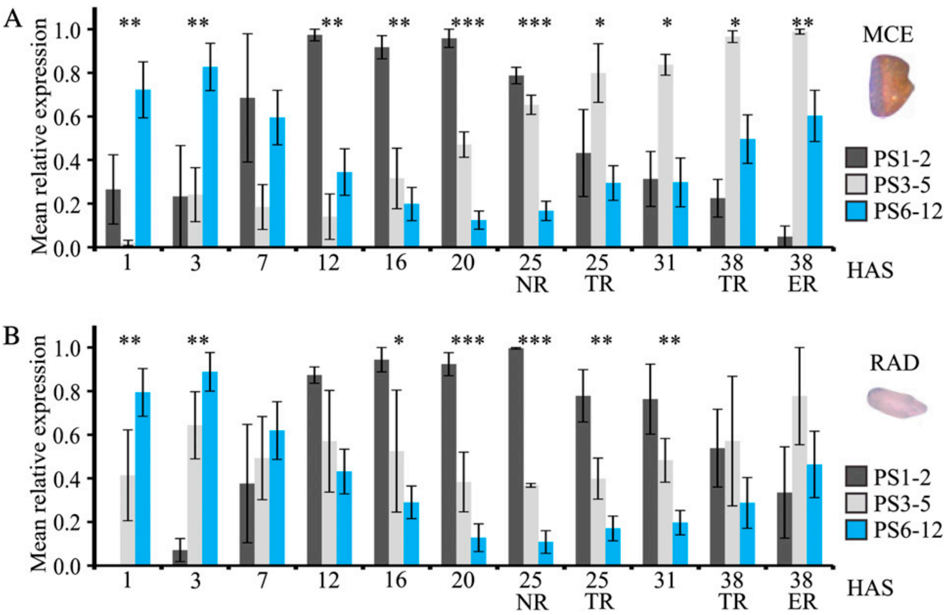


Figure 8. Relative expression of evolutionarily old and young genes across the Arabidopsis germination time course. Plotted are the relative expression levels ± SE of genes of PS1 and PS2, PS3 to PS5, and PS6 to PS12 across the Arabidopsis germination time course in the MCE (A) and RAD (B) compartments. The significance between the relative expression levels between the groups is indicated at each time point by asterisks: *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$. For the phylostratigraphic map and the mean relative expression of individual phylostrata, see Supplemental Figure S9.

Downloaded from www.plantphysiol.org on July 4, 2016 - Published by www.plantphysiol.org
Copyright © 2013 American Society of Plant Biologists. All rights reserved.

through a conserved and evolutionarily old transcriptional stage (Quint et al., 2012). This so-called phylotypic stage is mainly caused by the repression of evolutionarily young genes and is proposed to help the spatiotemporal organization and differentiation of multicellular life (Quint et al., 2012). Since we observed a largely inverse expression pattern during germination of gene sets that are up- and down-regulated during seed development (Fig. 6), we asked whether (1) seed germination is also characterized by the coordinated expression of evolutionarily old and young genes and (2), if so, whether these patterns are linked to the two transcriptomic phases we observed. To answer these questions, we first applied the phylostratigraphic approach (Domazet-Loso et al., 2007; Domazet-Loso and Tautz, 2010; Quint et al., 2012), in which we ordered the Arabidopsis genome into 12 evolutionary age classes (phylostrata; designated PS1–PS12). Each Arabidopsis gene is BLASTed against all genomes underlying the 12 phylostrata and is sorted in its phylostratum, defined as the most distant phylogenetic node containing at least one species with a detectable homolog (Quint et al., 2012). This resulted in the phylostratigraphic map in which PS1 (cellular organisms) contains the evolutionarily oldest genes and PS12 (Arabidopsis) contains the youngest genes that are specific to Arabidopsis, with no homologs detected in any of the other species (Supplemental Fig. S9A).

Next, we interrogated the gene expression data of the MCE and plotted the relative expression values of (1) genes that arose before plant evolution (PS1 and PS2 combined), (2) genes that arose during early plant evolution (algae and non-seed-bearing plants; PS3–PS5), and (3) the evolutionarily youngest genes (which evolved in seed-bearing plants; PS6–PS12). The analysis shows that in the MCE, the relative expression of evolutionarily young genes is high shortly after imbibition but drops during the first transcriptional phase, followed by an increase in the second transcriptional phase (Fig. 8A; Supplemental Fig. S9). Interestingly, the oldest genes (PS1 and PS2) showed an inverse behavior, starting low at the beginning of germination, peaking at the end of the first transcriptional phase, followed by a decrease in the second transcriptional phase. Genes of PS3 to PS5 show a different pattern, starting low and increasing during the course of germination. Comparable results were obtained for the RAD during germination (Fig. 8B; Supplemental Fig. S9). The patterns in both seed parts, particularly the inverse patterns of the evolutionarily old and young genes, suggest that seeds not only pass through an evolutionarily conserved stage during seed development but also during the successive germination phase. Coordinated expression of evolutionarily old and young genes (and, in this way, passage through a conserved transcriptional state) may help to channel large physiological transitions.

CONCLUSION

This study revealed two separate transcriptional phases for seed germination that are separated by TR and provides a strong indication that mechano-induced signaling affects gene expression at TR in the MCE. It also shows that time is an important determinant for spatial expression differences. Surprisingly, we found similar patterns of expression of evolutionarily old and young genes in seed development and seed germination, suggesting that plants passing through a transcriptional old and conserved stage may not be limited to embryogenesis. In addition to the novel biological insight, we are convinced that these detailed transcriptome data, including the tools developed for data visualization and mining, provide a powerful resource to gain further understanding of the roles of different seed compartments in germination, novel regulators, and gene networks underlying seed germination.

MATERIALS AND METHODS

Plant Material, Sampling, and Microarray Analysis

For this experiment, the Arabidopsis (*Arabidopsis thaliana*) accession Columbia-0 (N60000) was used. Seeds were sown on 0.7% water agarose (Eurogentec) and incubated in a germination cabinet at 22°C with continuous light. Germination curves (for both testa and endosperm rupture) were assessed by scoring germination in time. After the indicated HAS, seeds were harvested and dissected using forceps and a scalpel knife. For the isolation of RNA, a commercial kit (Absolutely RNA Nanoprep Kit; Agilent Technologies) was used. In total, 100 ng of RNA was used to synthesize biotin-labeled copy RNA (using the Affymetrix 3′ IVT-Express Labeling Kit), which was hybridized on the Affymetrix GeneChips Arabidopsis ATH1 Genome Array. The raw .cel files were background corrected and normalized using the Robust Microarray Averaging procedure (Irizarry et al., 2003). A detailed version of the methods used is available as Supplemental Materials and Methods S1.

The microarray data used in this article have been deposited in the National Center for Biotechnology Information's Gene Expression Omnibus with accession number GEO 41212.

Supplemental Data

The following materials are available in the online version of this article.

Supplemental Figure S1. ATH1 Genechip quality assessment and reproducibility.

Supplemental Figure S2. General expression numbers.

Supplemental Figure S3. Comparisons with two other seed microarray datasets.

Supplemental Figure S4. RT-qPCR confirms tissue-specific expression found in the microarray dataset.

Supplemental Figure S5. Topological features of the EndoNet and RadNet.

Supplemental Figure S6. Overrepresentation analysis of the 30 largest clusters from the EndoNet co-expression network.

Supplemental Figure S7. ORA using Pageman of genes that are either higher expressed in the MCE or the RAD.

Supplemental Figure S8. Seed tissues differentiate during germination.

Supplemental Figure S9. Expression of evolutionary old and young genes during Arabidopsis seed germination.

Supplemental Figure S10. The node degree distribution for the correlation networks.

Supplemental Table S1. Correlations between the sample replicates.

Supplemental Table S2. Primer information of the genes tested by RT-qPCR.

Downloaded from www.plantphysiol.org on July 4, 2016 - Published by www.plantphysiol.org
Copyright © 2013 American Society of Plant Biologists. All rights reserved.

Supplemental Materials and Methods S1. Supplemental Materials and Methods.

Supplemental Data Set S1. Collection of gene lists.

ACKNOWLEDGMENTS

We thank Kieran Lee (University of Leeds) for providing the image used in Figure 1A, Janet Braam (Rice University) and Dennis Lee for providing the touch transcriptome data set, Nicolas Provart (University of Toronto) and Hardeep Nahal (University of Toronto) for help with developing the Electronic Fluorescent Pictograph browser, and Henk Hilhorst (Wageningen University) and Maarten Koornneef (Max Planck Institute for Plant Breeding Research) for critically reading earlier versions of the manuscript.

Received June 18, 2013; accepted July 8, 2013; published July 16, 2013.

LITERATURE CITED

Bassel GW, Lan H, Glaab E, Gibbs DJ, Gerjets T, Krasnogor N, Bonner AJ, Holdsworth MJ, Provart NJ (2011) Genome-wide network model capturing seed germination reveals coordinated regulation of plant cellular phase transitions. Proc Natl Acad Sci USA **108:** 9709–9714

Belmonte MF, Kirkbride RC, Stone SL, Pelletier JM, Bui AQ, Yeung EC, Hashimoto M, Fei J, Harada CM, Munoz MD, et al (2013) Comprehensive developmental profiles of gene activity in regions and subregions of the Arabidopsis seed. Proc Natl Acad Sci USA **110:** E435–E444

Bethke PC, Libourel IG, Aoyama N, Chung YY, Still DW, Jones RL (2007) The Arabidopsis aleurone layer responds to nitric oxide, gibberellin, and abscisic acid and is sufficient and necessary for seed dormancy. Plant Physiol **143:** 1173–1188

Braam J (2005) In touch: plant responses to mechanical stimuli. New Phytol **165:** 373–389

Breeze E, Harrison E, McHattie S, Hughes L, Hickman R, Hill C, Kiddle S, Kim YS, Penfold CA, Jenkins D, et al (2011) High-resolution temporal profiling of transcripts during *Arabidopsis* leaf senescence reveals a distinct chronology of processes and regulation. Plant Cell **23:** 873–894

Chehab EW, Yao C, Henderson Z, Kim S, Braam J (2012) Arabidopsis touch-induced morphogenesis is jasmonate mediated and protects against pests. Curr Biol **22:** 701–706

Debeaujon I, Lepiniec L, Pourcel L, Routaboul J-M (2007) Seed coat development and dormancy. *In* Annual Plant Reviews, Vol 27: Seed Development, Dormancy and Germination. Blackwell Publishing, Oxford, pp 25–49

Domazet-Loso T, Brajković J, Tautz D (2007) A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. Trends Genet **23:** 533–539

Domazet-Loso T, Tautz D (2010) A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. Nature **468:** 815–818

Endo A, Tatematsu K, Hanada K, Duermeyer L, Okamoto M, Yonekura-Sakakibara K, Saito K, Toyoda T, Kawakami N, Kamiya Y, et al (2012) Tissue-specific transcriptome analysis reveals cell wall metabolism, flavonol biosynthesis, and defense responses are activated in the endosperm of germinating Arabidopsis thaliana seeds. Plant Cell Physiol **53:** 16–27

Farrant JM, Moore JP (2011) Programming desiccation-tolerance: from plants to seeds to resurrection plants. Curr Opin Plant Biol **14:** 340–345

Finch-Savage WE, Leubner-Metzger G (2006) Seed dormancy and the control of germination. New Phytol **171:** 501–523

Glaab E, Baudot A, Krasnogor N, Valencia A (2010) TopoGSA: network topological gene set analysis. Bioinformatics **26:** 1271–1272

Holdsworth MJ, Bentsink L, Soppe WJ (2008a) Molecular networks regulating Arabidopsis seed maturation, after-ripening, dormancy and germination. New Phytol **179:** 33–54

Holdsworth MJ, Finch-Savage WE, Grappin P, Job D (2008b) Post-genomics dissection of seed dormancy and germination. Trends Plant Sci **13:** 7–13

Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics **4:** 249–264

Kumpf RP, Shi CL, Larrieu A, Stø IM, Butenko MA, Péret B, Riiser ES, Bennett MJ, Aalen RB (2013) Floral organ abscission peptide IDA and its HAE/HSL2 receptors control cell separation during lateral root emergence. Proc Natl Acad Sci USA **110:** 5235–5240

Le BH, Cheng C, Bui AQ, Wagmaister JA, Henry KF, Pelletier J, Kwong L, Belmonte M, Kirkbride R, Horvath S, et al (2010) Global analysis of gene activity during Arabidopsis seed development and identification of seed-specific transcription factors. Proc Natl Acad Sci USA **107:** 8063–8070

Lee D, Polisensky DH, Braam J (2005) Genome-wide identification of touch- and darkness-regulated Arabidopsis genes: a focus on calmodulin-like and XTH genes. New Phytol **165:** 429–444

Lee KJD, Dekkers BJW, Steinbrecher T, Walsh CT, Bacic A, Bentsink L, Leubner-Metzger G, Knox JP (2012) Distinct cell wall architectures in seed endosperms in representatives of the Brassicaceae and Solanaceae. Plant Physiol **160:** 1551–1566

Lee KP, Piskurewicz U, Turecková V, Strnad M, Lopez-Molina L (2010) A seed coat bedding assay shows that RGL2-dependent release of abscisic acid by the endosperm controls embryo growth in Arabidopsis dormant seeds. Proc Natl Acad Sci USA **107:** 19108–19113

Lim PO, Kim HJ, Nam HG (2007) Leaf senescence. Annu Rev Plant Biol **58:** 115–136

Martínez-Andújar C, Pluskota WE, Bassel GW, Asahina M, Pupel P, Nguyen TT, Takeda-Kamiya N, Toubiana D, Bai B, Górecki RJ, et al (2012) Mechanisms of hormonal regulation of endosperm cap-specific gene expression in tomato seeds. Plant J **71:** 575–586

Müller K, Tintelnot S, Leubner-Metzger G (2006) Endosperm-limited Brassicaceae seed germination: abscisic acid inhibits embryo-induced endosperm weakening of Lepidium sativum (cress) and endosperm rupture of cress and Arabidopsis thaliana. Plant Cell Physiol **47:** 864–877

Nakabayashi K, Okamoto M, Koshiba T, Kamiya Y, Nambara E (2005) Genome-wide profiling of stored mRNA in Arabidopsis thaliana seed germination: epigenetic and genetic regulation of transcription in seed. Plant J **41:** 697–709

Narsai R, Law SR, Carrie C, Xu L, Whelan J (2011) In-depth temporal transcriptome profiling reveals a crucial developmental switch with roles for RNA processing and organelle metabolism that are essential for germination in Arabidopsis. Plant Physiol **157:** 1342–1362

Nonogaki H (2006) Seed germination: the biochemical and molecular mechanisms. Breed Sci **56:** 93–105

Nonogaki H, Chen F, Bradford KJ (2007) Mechanisms and genes involved in germination sensu stricto. *In* Annual Plant Reviews, Vol 27: Seed Development, Dormancy and Germination. Blackwell Publishing, Oxford, pp 264–304

Ohto M-a, Stone SL, Harada JJ (2007) Genetic control of seed development and seed mass. *In* Annual Plant Reviews, Vol 27: Seed Development, Dormancy and Germination. Blackwell Publishing, Oxford, pp 1–24

Penfield S, Graham S, Graham IA (2005) Storage reserve mobilization in germinating oilseeds: Arabidopsis as a model system. Biochem Soc Trans **33:** 380–383

Penfield S, Li Y, Gilday AD, Graham S, Graham IA (2006) *Arabidopsis* ABA INSENSITIVE4 regulates lipid mobilization in the embryo and reveals repression of seed germination by the endosperm. Plant Cell **18:** 1887–1899

Preston J, Tatematsu K, Kanno Y, Hobo T, Kimura M, Jikumaru Y, Yano R, Kamiya Y, Nambara E (2009) Temporal expression patterns of hormone metabolism genes during imbibition of Arabidopsis thaliana seeds: a comparative study on dormant and non-dormant accessions. Plant Cell Physiol **50:** 1786–1800

Quint M, Drost HG, Gabel A, Ullrich KK, Bönn M, Grosse I (2012) A transcriptomic hourglass in plant embryogenesis. Nature **490:** 98–101

Rajjou L, Duval M, Gallardo K, Catusse J, Bally J, Job C, Job D (2012) Seed germination and vigor. Annu Rev Plant Biol **63:** 507–533

Sliwinska E, Bassel GW, Bewley JD (2009) Germination of Arabidopsis thaliana seeds is not completed as a result of elongation of the radicle but of the adjacent transition zone and lower hypocotyl. J Exp Bot **60:** 3587–3594

Stenvik GE, Tandstad NM, Guo Y, Shi CL, Kristiansen W, Holmgren A, Clark SE, Aalen RB, Butenko MA (2008) The EPIP peptide of INFLORESCENCE DEFICIENT IN ABSCISSION is sufficient to induce abscission in *Arabidopsis* through the receptor-like kinases HAESA and HAESA-LIKE2. Plant Cell **20:** 1805–1817

Usadel B, Nagel A, Steinhauser D, Gibon Y, Bläsing OE, Redestig H, Sreenivasulu N, Krall L, Hannah MA, Poree F, et al (2006) PageMan: an interactive ontology tool to generate, display, and annotate overview graphs for profiling experiments. BMC Bioinformatics **7:** 535

Weitbrecht K, Müller K, Leubner-Metzger G (2011) First off the mark: early seed germination. J Exp Bot **62:** 3289–3309

Winter D, Vinegar B, Nahal H, Ammar R, Wilson GV, Provart NJ (2007) An "Electronic Fluorescent Pictograph" browser for exploring and analyzing large-scale biological data sets. PLoS ONE **2:** e718

Downloaded from www.plantphysiol.org on July 4, 2016 - Published by www.plantphysiol.org
Copyright © 2013 American Society of Plant Biologists. All rights reserved.

10 Appendix

10.1 Submitted Manuscript: Capturing Evolutionary Signatures in Transcriptomes with myTAI

The following paper entitled *Capturing Evolutionary Signatures in Transcriptomes with myTAI* is currently submitted to the journal *Molecular Biology and Evolution* and published on bioRxiv but has not been peer-reviewed yet. However, its content summarizes the main applications and functionalities of *myTAI* and *orthologr*. Due to this fact, I decided to include this paper as appendix to this thesis.

This paper is available at:

- `http://biorxiv.org/content/early/2016/05/03/051565.abstract`

bioRxiv preprint first posted online May. 3, 2016; doi: http://dx.doi.org/10.1101/051565. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

Capturing Evolutionary Signatures in Transcriptomes with myTAI

Hajk-Georg Drost[1,2,7,*], Alexander Gabel[2,7], Tomislav Domazet-Lošo[3,4], Marcel Quint[5,6], Ivo Grosse[2,7]

[1] Sainsbury Laboratory Cambridge, University of Cambridge, Bateman Street, Cambridge CB2 1LR, UK

[2] Martin Luther University Halle-Wittenberg, Institute of Computer Science, Halle (Saale), Germany

[3] Laboratory of Evolutionary Genetics, Division of Molecular Biology, Ruder Boškovic Institute, Zagreb, Croatia

[4] Catholic University of Croatia, Zagreb, Croatia

[5] Department of Molecular Signal Processing, Leibniz Institute of Plant Biochemistry, Halle (Saale), Germany

[6] Martin Luther University Halle-Wittenberg, Institute of Agricultural and Nutritional Sciences, Halle (Saale), Germany

[7] German Center for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany

[*] correspondence to: hgd23@cam.ac.uk

Keywords: Phylostratigraphy, Divergence Stratigraphy, Phylotranscriptomics, Evolution, Development, Evo-Devo

Abstract

Combining transcriptome data of biological processes or response to stimuli with evolutionary information such as the phylogenetic conservation of genes or their sequence divergence rates enables the investigation of evolutionary constraints on these processes or responses. Such *phylotranscriptomic* analyses recently unraveled that mid-developmental transcriptomes of fly, fish, and cress were dominated by evolutionarily conserved genes and genes under negative selection and thus recapitulated the developmental hourglass on the transcriptomic level. Here, we present a protocol for performing phylotranscriptomic analyses on any biological process of interest. When applying this protocol, users are capable of detecting different evolutionary constraints acting on different stages of the biological process of interest in any species. For each

bioRxiv preprint first posted online May. 3, 2016; doi: http://dx.doi.org/10.1101/051565. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

step of the protocol, modular and easy-to-use open-source software tools are provided, which enable a broad range of scientists to apply phylotranscriptomic analyses to a wide spectrum of biological questions.

Introduction

Transcriptomes carry evolutionary information because expressed genes have different evolutionary ages or are exposed to different selective pressures. In major biological processes such as embryogenesis, metamorphosis, fertilization, senescence, etc. or after biological treatments the set of genes expressed at different stages within these processes varies. Analogously, both environmental and endogenous stimuli elicit responses of different sets of genes. In addition to having varying biological functions, these gene sets may vary regarding their evolutionary signatures like phylogenetic conservation of genes (= gene ages) or their sequence divergence rates (= sequence divergence).

To capture and quantify such evolutionary signatures across development, biological processes or response to stimuli, we developed and established *phylotranscriptomic analyses*, which combine information about gene age and gene sequence divergence with transcriptome data of biological processes and response to stimuli. These analyses allowed the molecular confirmation of the developmental hourglass (Domazet-Lošo and Tautz 2010a), one of the historic principles in evolution and developmental biology originally discovered by von Baer (von Baer 1828).

Moreover, phylotranscriptomic analyses unraveled that the hourglass pattern, which was thought to be a hallmark of animal embryogenesis, is restricted neither to animals nor to embryogenesis. Specifically, such analyses identified molecular hourglass patterns in plants (Quint et al. 2012) and fungi (Cheng et al. 2015) as well as in post-embryonic plant development (Drost et al. 2016).

Phylotranscriptomic analyses are not limited to embryogenesis or other post-embryonic developmental processes. Potential applications of these analyses in other disciplines are studies of life cycles of many different animals, plants, fungi, or bacteria, of metabolic or circadian rhythms, of the mitotic and meiotic cell cycle, of tumor progression, or of a

bioRxiv preprint first posted online May. 3, 2016; doi: http://dx.doi.org/10.1101/051565. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

plethora of other fundamental biological processes of life. For example, phylotranscriptomic analyses turned out to be helpful for comparing transcriptomes of three stem cell types in Hydra (Hemmrich et al. 2012) or for reconstructing the evolutionary origin of the neural crest, a *bona fide* innovation of vertebrates (Šestak et al. 2012). In addition, these analyses can be applied to capture evolutionary signatures of temporal responses to different endogenous and exogenous stimuli. Furthermore, phylotranscriptomic analyses can be applied to capture evolutionary signatures on the spatial level in different tissues, organs, cell types, or tumors and allow to study these spatial signatures in development, in other temporal processes, or in response to endogenous and exogenous stimuli.

Despite this great potential, no standardized and reproducible protocol for performing phylotranscriptomic analyses exists to date. To overcome this limitation and to allow a broad range of scientists to capture and quantify evolutionary signatures in transcriptomes by applying such analyses, we developed the open source software packages *createPSmap.pl*, *orthologr*, and *myTAI* and here provide a user-friendly protocol to apply these tools to any organism and biological study of interest.

Methodological Background

Given the wide variety of possible applications of phylotranscriptomic analyses, we focus on development as an example use case of our protocol.

Evolutionary signatures of transcriptomes can be captured by computing transcriptome indices at different measured stages of development, combining these computed values to a transcriptome index profile across the measured stages, and comparing this profile with a flat line. A profile not significantly deviating from a flat line indicates the absence of significant variations of the computed transcriptome index from stage to stage. In contrast, a profile significantly deviating from a flat line indicates the presence of significant variations from stage to stage. We refer to any transcriptome index profile significantly deviating from a flat line as phylotranscriptomic pattern or evolutionary signature.

bioRxiv preprint first posted online May. 3, 2016; doi: http://dx.doi.org/10.1101/051565. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

The computation of the *transcriptome age index* (*TAI*) (Domazet-Lošo and Tautz 2010a) (Supplementary Information) requires the determination of the evolutionary ages of the genes of the studied species. For this purpose, we developed a procedure termed *phylostratigraphy* (Domazet-Lošo et al. 2007), which briefly described works as follows: First, a sequence homology search of the proteins of the studied species against a database of proteins of completely sequenced genomes from species covering all kingdoms of life is performed. Second, these species are sorted into sets named *phylostrata* (PS) corresponding to hierarchically ordered phylogenetic nodes along the tree of life. PS 1 denotes the set of all living species, PS 2 denotes the set of species of the same domain as the query species, PS 3 denotes the set of species of the same kingdom as the query species, etc., and the highest PS denotes the set consisting of only the studied species. Third, each protein of the studied species is assigned to the lowest PS in which at least one homolog with a predefined threshold on the degree of homology was detectable. The resulting assignment of one PS to each protein of the studied genome is called *phylostratigraphic map*, and the PS of a given protein or protein-coding gene is often loosely called *gene age*.

The TAI at a given stage of development is then obtained by joining this phylostratigraphic map with expression data at that stage and by computing the weighted mean of the PS, where the weights are the stage-specific expression levels (Supplementary Information) (Domazet-Lošo and Tautz 2010a). Loosely speaking, the TAI at a given stage of development is the mean evolutionary age of the genes expressed at that stage, and the TAI profile is the profile of these mean ages across different stages of development. Stages with high TAI values are stages where evolutionarily old genes (in low PS) are more lowly expressed - and evolutionarily young genes (in high PS) are more highly expressed - than in other stages.

The computation of the *transcriptome divergence index* (*TDI*) (Quint et al. 2012, Drost et al. 2015) (Supplementary Information) requires the determination of the degree of selection of the genes of the studied species. For this purpose, we developed a procedure termed *divergence stratigraphy* (Quint et al. 2102, Drost et al. 2015), which works analogously to phylostratigraphy as follows: First, a set of orthologs of the proteins of the studied species is obtained in a closely related species. Second, rates of non-synonymous (dN) and synonymous (dS) substitutions as well as the dN/dS ratio are

bioRxiv preprint first posted online May. 3, 2016; doi: http://dx.doi.org/10.1101/051565. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

estimated for each pairwise alignment of the protein-coding genes of the protein of the studied species and its ortholog. Third, the continuous dN/dS ratios are discretized by sorting them into e. g. 10 groups of equal sizes, where group 1 contains the genes with the lowest dN/dS ratios, and group 10 contains the genes with the highest dN/dS ratios. In analogy to PS, these groups of genes of similar dN/dS ratios are called *divergence strata* (*DS*). The resulting assignment of one DS to each protein of the studied species with an ortholog is called *divergence stratigraphic map* (Quint et al. 2012, Drost et al. 2015), and the DS of a given protein-coding gene is often loosely called *sequence divergence* denoting the evolutionary rate of a protein (Zhang and Yang 2015).

The TDI at a given stage of development is then obtained in analogy to obtaining the TAI by joining this divergence stratigraphic map with expression data at that stage and by computing the weighted mean of the DS (Quint et al. 2012, Drost et al. 2015). Loosely speaking, the TDI at a given stage of development is the mean sequence divergence of the genes expressed at that stage, and the TDI profile is the profile of these mean sequence divergences across different stages of development. Stages with high TDI values are stages where genes under strong negative selection (in low DS) are more lowly expressed - and genes under weaker negative selection or even positive selection (in high PS) are more highly expressed - than in other stages.

To assess the significance of deviations of transcriptome index profiles from a flat line, we proposed the *flat-line test* (Quint et al. 2012; Drost et al. 2015). The *flat-line test* is a permutation test that randomly assigns PS or DS to the genes of investigation. This random assignment is used to compute TAI or TDI profiles and produces random patterns of transcriptome conservation. Subsequently, the variance is then used as a measure to quantify the variation of these random transcriptome indices between stages. This procedure is performed independently 10,000 times and the resulting variance values are compared with the actual variance of TAI or TDI patterns. If the p-value of the actual TAI or TDI variance value is less than 0.05 we declare the corresponding pattern as significantly deviating from a flat line and as not deviating from a flat line otherwise (Supplementary Information).

Two phylotranscriptomic patterns of particular interest in plant and animal embryogenesis are the hourglass pattern and the early-conservation pattern. To test the

bioRxiv preprint first posted online May. 3, 2016; doi: http://dx.doi.org/10.1101/051565. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

presence or absence of these patterns we introduced the *reductive hourglass test* and *reductive early-conservation test* (Drost et al. 2015). The reductive hourglass test quantifies the degree of agreement of the transcriptome index profile to an hourglass-like high-low-high pattern, while the reductive early-conservation test quantifies the degree of agreement with an early conservation-like low-low-high pattern. In addition, we designed *myTAI* such that users can easily build customized statistical tests for assessing the significance of any pattern deviating from a flat line (Supplementary Information).

To further scrutinize observed phylotranscriptomic patterns we introduced relative expression profiles per PS and per DS (Domazet-Lošo and Tautz 2010a; Drost et al. 2015) (Supplementary Information). Relative expression levels allow the visualization of the average expression behavior of genes from the same PS or the same DS across the biological process of interest. Specifically, the mean expression profile of each PS and each DS across all stages is linearly transformed to a normalized profile ranging from 0 to 1, and the resulting normalized profile is called *relative expression profile* of the corresponding PS and DS.

Protocol Applications

As introduced before, phylotranscriptomic analyses enabled us to unravel the existence of phylotranscriptomic hourglass patterns in animals (Domazet-Lošo and Tautz 2010a), plants (Quint et al. 2012; Drost et al. 2015; Drost et al. 2016), and fungi (Cheng et al. 2015). These findings suggested that ontogenetic processes occurring during development are related to phylogenetic processes occurring during evolution (Von Baer 1828; Sander 1983; Duboule 1994; Richardson 1995; Raff 1996; Richardson et al. 1997; Richardson 1999; Hazkani-Covo et al. 2005; Irie and Sehara-Fujisawa 2007; Artieri et al. 2009; Cruickshank and Wade 2008; Kalinka et al. 2010; Domazet-Lošo and Tautz 2010a; Yanai et al. 2011; Irie and Kuratani 2011; Levin et al. 2012; Willmore 2012; Svorcová 2012; Tian et al. 2013; Wang et al. 2013; Gerstein et al. 2014; Levin et al. 2016; Gossmann et al. 2016).

Specifically, it has been found that the pattern of morphologically dissimilar-similar-dissimilar embryos between related animal species, the so-called developmental

bioRxiv preprint first posted online May. 3, 2016; doi: http://dx.doi.org/10.1101/051565. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

hourglass pattern (Duboule 1994; Raff 1996), is mirrored by a similar hourglass-like high-low-high pattern of TAI and TDI profiles on the phylotranscriptomic level (Domazet-Lošo and Tautz 2010a; Drost et al. 2015). Moreover, it has been found that the stage or period of maximum transcriptome conservation during mid embryogenesis coincides with the morphological stage of maximum conservation defined as phylotypic stage (Sander 1983) or phylotypic period (Richardson 1995; Raff 1996; Richardson et al. 1997). In this context, phylotranscriptomic analyses have provided a molecular explanation for and thus deepened our understanding of the relation between evolution and development, and we believe that such analyses will advance current and future evo-devo research, too.

However, the applicability of the protocol and the developed software packages is not restricted to the study of development. As specified in section *Methodological Background*, this protocol can be applied to phylotranscriptomic analyses of any transcriptome data set of any biological study of any species. In addition to phylotranscriptomic analyses, individual modules of this protocol can be used for performing different analyses independently of each other.

For example, phylostratigraphy alone has previously been performed to detect orphan genes (Tautz and Domazet-Lošo 2011) or the evolutionary origin of specific classes of genes such as cancer genes (Domazet-Lošo and Tautz 2010b) or transcription factors (De Mendoza et al. 2013). Likewise, divergence stratigraphy can simply be performed for quantifying the rate of synonymous versus non-synonymous substitution rates of protein-coding genes in a genome of choice (Drost et al. 2015).

The modularity of the protocol also allows users to use their own modules for phylostratigraphy or divergence stratigraphy. This modularity for example, enables to study the influence of different phylostratigraphies or divergence stratigraphies on TAI or TDI profiles. In particular, the influence of potential underestimations of gene ages by BLAST approaches (Moyers and Zhang 2015, 2016) can be systematically investigated using this protocol.

In general, this protocol was designed to make it easily applicable for life scientists. For this purpose, we provide software tools and step-by-step instructions for every part of

bioRxiv preprint first posted online May. 3, 2016; doi: http://dx.doi.org/10.1101/051565. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

this protocol. For example, we provide the Perl script *createPSmap.pl* for performing BLAST searches and computing phylostratigraphic maps, the R package *orthologr* for identifying orthologs and computing divergence stratigraphic maps, and the R package *myTAI* for computing TAI and TDI profiles across developmental stages, for performing statistical tests such as the flat-line test, for computing relative expression profiles for all PS and DS, or for producing scientific visualizations of observed phylotranscriptomic patterns in publication quality. All parts of the protocol are demonstrated by using the same example data set covering seven stages of *A. thaliana* embryo development (Quint et al. 2012; Drost et al. 2015).

All software tools are publicly available under an open source license (https://github.com/AlexGa/Phylostratigraphy, https://github.com/HajkD/orthologr, and https://github.com/HajkD/myTAI). An extensive documentation of each function as well as six tutorials covering different phylotranscriptomic analyses are part of the *orthologr* package, the *myTAI* package, and the Supplementary Information.

Furthermore, pre-computed phylostratigraphic maps and divergence stratigraphic maps can be obtained from a public repository (https://github.com/HajkD/published_phylomaps). In this way, users can easily adapt the provided protocol to the organism, biological process, and data sets of their interest with minimal computational effort.

Comparison with other similar techniques

Possible alternatives to phylostratigraphy are based on Wagner-Parsimony or Phylogenetic-Reconciliation (Capra et al. 2013) and are summarized in the software tool ProteinHistorian (Capra et al. 2012; Capra et al. 2013). ProteinHistorian allows performing alternative gene age assignment methods that can then be used by *myTAI* to compute TAI profiles and to perform all other analyses based on phylostratigraphic maps resulting from these alternative methods for estimating gene ages.

Possible alternatives to divergence stratigraphy are based on phylogenetic inference. Here, the metaPhOrs repository (http://orthology.phylomedb.org/) can be accessed to retrieve pre-computed phylogeny-based orthology predictions that can then be used to

bioRxiv preprint first posted online May. 3, 2016; doi: http://dx.doi.org/10.1101/051565. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

estimate sequence substitution rates. Alternative methods for estimating substitution rates are provided by the R package *orthologr*, where the function dNdS() allows users to choose a variety of alternative methods for estimating substitution rates of predicted orthologs (for detailed information see https://github.com/HajkD/orthologr/blob/master/vignettes/dNdS_estimation.Rmd). These alternative divergence stratigraphic maps can then be used by *myTAI* for computing TDI profiles and for performing all other analyses.

A broadly applied alternative approach for associating developmental transcriptomes with evolutionary constraints is comparative transcriptomics, which has been used to study developmental hourglass patterns in several species (Kalinka et al. 2010; Irie and Kuratani 2011; Levin et al. 2012; Romero et al. 2012; Wang et al. 2013; Dunn et al. 2013; Warnefors and Kaessmann 2013; Gerstein et al. 2014; Nesculea and Kaessmann 2014; Levin et al. 2016). Comparative transcriptomics analyzes gene expression diversity between orthologs of two or more species based on the empirical finding that gene expression diversity of orthologs correlates with developmental dissimilarity (Roux et al. 2015).

Limitations of the Protocol

One limitation of the protocol is that it can only be applied to species for which protein sequences are annotated and for which this annotation matches the transcriptome annotation of the corresponding gene expression data set.

A second limitation is that computing a phylostratigraphic map can take several hours, several days, or even several weeks depending on the number of query sequences and the size of the database of proteins of completely sequenced genomes and may thus require a computing cluster.

A third limitation is that users interested in applying a custom taxonomy to *createPSmap.pl* need to perform additional steps to retrieve a customized phylostratigraphic map.

bioRxiv preprint first posted online May. 3, 2016; doi: http://dx.doi.org/10.1101/051565. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

However, as the *myTAI* package is designed to take any custom phylostratigraphic map or divergence stratigraphic map as input, all of the subsequent phylotranscriptomic analyses of this protocol can be performed irrespectively of the origins of the phylostratigraphic or divergence stratigraphic maps (Supplementary Information).

Example Experiment

The following example protocol is divided into four conceptual parts. The first part (steps 1-2) covers the construction of phylostratigraphic and divergence stratigraphic maps. The second part (steps 3 - 10) includes the computation and visualization of TAI and TDI profiles. The third part (steps 11 - 12) covers the application of three statistical tests for quantifying the significance of the observed phylotranscriptomic patterns, and the fourth part (steps 13 - 18) includes the computation and visualization of relative expression profiles (Fig. 1).

In step 1 of the protocol, the Perl script *createPSmap.pl* is used for computing the phylostratigraphic map (Fig. 2). The input files to *createPSmap.pl* are a *fasta* file of protein sequences of the studied species of and a database of proteins of completely sequenced genomes in *fasta* format. Phylogenetic information is provided in the header of *fasta* sequences and can be customized by following the tutorial at https://github.com/AlexGa/Phylostratigraphy. The output of *createPSmap.pl* is the phylostratigraphic map, i.e., a table storing the gene id and the PS of each protein-coding gene of the studied species (https://github.com/AlexGa/Phylostratigraphy).

In step 2, function divergence_stratigraphy() of R package *orthologr* is used for computing the divergence stratigraphic map. In the example data set, the arguments *query_file* and *subject_file* refer to the downloaded CDS files *Athaliana_167_cds.fa* and *Alyrata_107_cds.fa* (see *Prerequisite tools* for details) and denote the input files of function divergence_stratigraphy(). The output of the function divergence_stratigraphy() is the divergence stratigraphic map, i.e. a table storing the gene id and the DS of each protein-coding gene of the studied species that has an ortholog in the other species (https://github.com/HajkD/orthologr/blob/master/vignettes/divergence_stratigraphy.Rmd).

bioRxiv preprint first posted online May. 3, 2016; doi: http://dx.doi.org/10.1101/051565. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

In step 3, the phylostratigraphic map and the divergence stratigraphic map are matched with a transcriptome data set covering the studied biological process. This step is accomplished by function *MatchMap()* of R package *myTAI*, which takes a phylostratigraphic map or divergence stratigraphic map and an expression data set as input and returns a table storing the gene id, its PS or DS, and its expression profile as output. We denote this output data as *PhyloExpressionSet* or *DivergenceExpressionSet*.

Phylostratigraphy and divergence stratigraphy need to be performed only once for each species, and phylostratigraphic maps and divergence stratigraphic maps are available for a variety of species at https://github.com/HajkD/published_phylomaps.

In step 4, R package *myTAI* and pre-formatted data sets required for subsequent analyses of the protocol are loaded into the current R session.

Steps 5 and 6 visualize the phylostratigraphic map (Fig. 2 a) and the divergence stratigraphic map (Fig. 2 b) by plotting histograms of absolute or relative frequencies of genes per PS or per DS, respectively.

TAI and TDI profiles are computed and visualized in steps 7-10 (Fig. 3). Both functions TAI() and TDI() take a *PhyloExpressionSet* or *DivergenceExpressionSet* as input and compute the TAI and TDI values as output. These profiles are then visualized by function PlotPattern().

In steps 11 and 12, *p*-values of the TAI and TDI profiles are computed according to the *flat-line test*, the *reductive hourglass test* (11.A), or the *reductive early-conservation test* (11.B). The flat-line test assesses the deviation of a TAI or TDI profile from a flat line, while the reductive hourglass test and the reductive early-conservation test indicate the presence of a high-low-high pattern or a low-low-high pattern, respectively, based on an *a-priori* definition of early, mid (phylotypic), and late phases of development (Drost et al. 2015) (Supplementary Information).

In steps 13-16, tables of relative expression profiles for each PS and each DS are obtained and visualized by using functions REMatrix() and PlotRE() of R package *myTAI*.

bioRxiv preprint first posted online May. 3, 2016; doi: http://dx.doi.org/10.1101/051565. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

In steps 17-18, a group-specific visualization of relative expression levels is performed by function PlotBarRE() of R package *myTAI*, and potential differences between groups are assessed by a Kruskal-Wallis rank sum test (Fig. 4).

Prerequisite tools

BLAST

The Basic Local Alignment Search Tool (BLAST) (Altschul et al. 1990) is used to determine gene homology relationships in phylostratigraphy and divergence stratigraphy and can be downloaded from ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/. A detailed installation guide can also be found at https://github.com/HajkD/orthologr/blob/master/vignettes/Install.Rmd#install-blast.

Perl and Java Programming Environments

The Perl programming language can be downloaded from https://www.perl.org/ and the Java compiler and interpreter can be obtained from https://www.java.com/de/download/.

The pipeline to perform phylostratigraphy can be downloaded from https://github.com/AlexGa/Phylostratigraphy.

R Programming Environment

The R programming environment (R Core Team 2016) can be downloaded from http://cran.r-project.org/ and installed under Linux, Mac OSX, or Windows.

The R package *orthologr* can be downloaded and installed from

https://github.com/HajkD/orthologr .

The R package *myTAI* can be downloaded and installed from

bioRxiv preprint first posted online May. 3, 2016; doi: http://dx.doi.org/10.1101/051565. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

https://github.com/HajkD/myTAI .

R Package Dependency

myTAI (Drost 2016) uses the following open source packages from CRAN: Rcpp (Eddelbuettel and Francois 2011), nortest (Gross and Ligges 2014), fitdistrplus (Delignette-Muller and Dutang 2015), doParallel (Weston 2014), dplyr (Wickham and Francois 2015), RColorBrewer (Neuwirth 2014), taxis (Chamberlain and Szocs 2013), ggplot2 (Wickham 2009), and edgeR (Robinson et al. 2010).

orthologr (Drost et al. 2015) uses the following open-source packages from CRAN and Bioconductor (Huber et al. 2015): Rcpp (Eddelbuettel and Francois 2011), doParallel (Weston 2014), dplyr (Wickham and Francois 2015), seqinr (Charif and Lobry 2007), data.table (Dowle 2014), Biostrings (Pages et al. 2007), RSQLite (Wickham et al. 2014), stringr (Wichham 2015), IRanges (Lawrence et al. 2013), DBI (R Special Interest Group on Databases 2014), and S4Vectors (Pages et al. 2014).

orthologr and myTAI

After installing R packages *orthologr* and *myTAI*, they can be loaded into the current R session by commands

```
library(orthologr)
library(myTAI)
```

Data

To perform phylostratigraphy, the reference genome database needs to be downloaded from http://msbi.ipb-halle.de/download/phyloBlastDB_Drost_Gabel_Grosse_Quint.fa.tbz. This database currently stores 17,582,624 amino acid sequences covering 4,557 species. After downloading file phyloBlastDB, it can be unpacked by opening a terminal application and typing

bioRxiv preprint first posted online May. 3, 2016; doi: http://dx.doi.org/10.1101/051565. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

```
tar xfvj phyloBlastDB_Drost_Gabel_Grosse_Quint.fa.tbz
```

The header of the FASTA-files of the studied species (e.g. Athaliana_167_protein.fa) must fulfill the following specification

```
>GeneID | [species_name] | [taxonomy]
```

The corresponding taxonomy starts with super kingdoms limited to Eukaryota, Archaea, and Bacteria. For the example data set this yields:

```
>ATCG00500.1|PACid:19637947 | [Arabidopsis thaliana] | [Eukaryota; Viridiplantae;
Streptophyta; Streptophytina; Embryophyta; Tracheophyta; Euphyllophyta;
Spermatophyta; Magnoliophyta; eudicotyledons; core eudicotyledons; rosids;
malvids; Brassicales; Brassicaceae; Camelineae; Arabidopsis]
```

To download the coding sequence (CDS) files of *A. thaliana* and *A. lyrata* from the Phytozome database (Goodstein et al. 2012), the following R command-line tool can be used:

```
# download the CDS file of A. thaliana
download.file( url     = "ftp://ftp.jgi- psf.org/pub/compgen/phytozome/v9.0/
                          Athaliana/annotation/Athaliana_167_cds.fa.gz",
               destfile = "Athaliana_167_cds.fa.gz" )

# download the CDS file of A. lyrata
download.file( url     = "ftp://ftp.jgi-psf.org/pub/compgen/phytozome/v9.0/
                          Alyrata/annotation/Alyrata_107_cds.fa.gz",
               destfile = "Alyrata_107_cds.fa.gz" )
```

Next, the files Athaliana_167_cds.fa.gz and Alyrata_107_cds.fa.gz need to be unpacked. These CDS files are then used by R package *orthologr* for performing divergence stratigraphy.

bioRxiv preprint first posted online May. 3, 2016; doi: http://dx.doi.org/10.1101/051565. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

myTAI includes two example data sets, each containing a gene-expression time course covering seven stages of *A. thaliana* embryogenesis starting from the zygote stage to mature embryos as well as a phylostratigraphic map and divergence stratigraphic map of each protein-coding gene of *A. thaliana* (Xiang et al. 2011; Quint et al. 2012) with an ortholog in *Arabidopsis lyrata* (Drost et al. 2015). These example data sets are loaded into the current R session by commands

```
data(PhyloExpressionSetExample)
data(DivergenceExpressionSetExample)
```

The structure of these example data sets can be displayed by commands

```
head(PhyloExpressionSetExample, 3)
head(DivergenceExpressionSetExample, 3)
```

The structure of the data set resembles a standard format for all functions and procedures implemented in *myTAI*. To distinguish data sets storing phylostratigraphic maps combined with gene expression data and sequence divergence stratigraphic maps combined with gene expression data, the notation *PhyloExpressionSet* and *DivergenceExpressionSet* is used in the documentation of *myTAI*.

Protocol Steps

1. Compute phylostratigraphic map:

```
 perl createPSmap.pl --organism Athaliana_167_protein_with_new_Header.fa
                  --database phyloBlastDB_Drost_Gabel_Grosse_Quint.faphyloBlastDB.fa
                                        --evalue 1e-5 --threads 64 --blastPlus
```

2. Compute divergence stratigraphic map with reciprocal best hit:

```
library(orthologr)
```

```
# compute the divergence stratigraphic map of A. thaliana vs. A. lyrata
```

bioRxiv preprint first posted online May. 3, 2016; doi: http://dx.doi.org/10.1101/051565. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

```
Ath_vs_Aly_DM <- divergence_stratigraphy(
                                query_file     = "Athaliana_167_cds.fa",
                                subject_file    = "Alyrata_107_cds.fa",
                                eval            = "1E-5",
                                ortho_detection = "RBH",
                                comp_cores      = 1 )
```

A. Compute divergence stratigraphic map with best hit:

```
library(orthologr)

# compute the divergence map of A. thaliana vs. A. lyrata using BLAST best hit
Ath_vs_Aly_DM <- divergence_stratigraphy(
                               query_file     = "Athaliana_167_cds.fa",
                               subject_file    = "Alyrata_107_cds.fa",
                               eval            = "1E-5",
                               ortho_detection = "BH",
                               comp_cores      = 1 )
```

3. Match phylostratigraphic map of step 1 or divergence stratigraphic map of step 2 with an expression data set:

```
library(myTAI)

PhyloExpressionSet <- MatchMap(PhyloMap, ExpressionMatrixExample)

DivergenceExpressionSet <- MatchMap(DivergenceMap, ExpressionMatrixExample)
```

4. Load R package *myTAI* and read data of step 3 into the current R session:

```
library(myTAI)
data(PhyloExpressionSetExample)
data(DivergenceExpressionSetExample)
```

bioRxiv preprint first posted online May. 3, 2016; doi: http://dx.doi.org/10.1101/051565. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

A. Load R package *myTAI* and read a custom expression data set from a hard drive:

```
library(myTAI)
myPhyloExpressionSet <- read.csv(“PhyloExpressionSet.csv”, sep = “,”, header = TRUE)
myDivergenceExpressionSet <- read.csv(“DivergenceExpressionSet.csv”, sep = “,”, header = TRUE)
```

5. Visualize the phylostratigraphic map of step 1 (Fig. 2 a):

```
PlotDistribution( PhyloExpressionSet = PhyloExpressionSetExample,
                  xlab               = “Phylostratum”)
```

6. Visualize the divergence stratigraphic map of step 2 (Fig. 2 b):

```
PlotDistribution( PhyloExpressionSet = DivergenceExpressionSetExample,
                  xlab               = “Divergence stratum”)
```

7. Compute the TAI profile from the data set of step 3:

```
TAI(PhyloExpressionSetExample)
```

8. Visualize the TAI profile from step 7 (Fig. 3 a):

```
PlotPattern( ExpressionSet  = PhyloExpressionSetExample,
             type           = “l”,
             lwd            = 6,
             xlab           = “Ontogeny”,
             ylab           = “TAI” )
```

9. Compute the TDI profile from the data set of step 3:

```
TDI(DivergenceExpressionSetExample)
```

10. Visualize the TDI profile from step 9 (Fig. 3 b):

bioRxiv preprint first posted online May. 3, 2016; doi: http://dx.doi.org/10.1101/051565. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

```
PlotPattern( ExpressionSet = DivergenceExpressionSetExample,
             type          = “l”,
             lwd           = 6,
             xlab          = “Ontogeny”,
             ylab          = “TDI” )
```

11. Perform the flat-line test for a PhyloExpressionSet of step 3 and the TAI profile of step 7:

```
FlatLineTest( ExpressionSet = PhyloExpressionSetExample,
              permutations  = 10000,
              plotHistogram = TRUE )
```

A. Perform the reductive hourglass test for a PhyloExpressionSet of step 3 and the TAI profile of step 7:

```
ReductiveHourglassTest( ExpressionSet  = PhyloExpressionSetExample,
                        modules        = list(early = 1:2, mid = 3:5, late = 6:7),
                        permutations   = 10000,
                        plotHistogram = TRUE )
```

B. Perform the reductive early-conservation test for a PhyloExpressionSet of step 3 and the TAI profile of step 7:

```
EarlyConservationTest(  ExpressionSet  = PhyloExpressionSetExample,
                        modules        = list(early = 1:2, mid = 3:5, late = 6:7),
                        permutations   = 10000,
                        plotHistogram = TRUE )
```

12. Perform the flat-line test for a DivergenceExpressionSet of step 3 and the TDI profile of step 9:

```
FlatLineTest( ExpressionSet = DivergenceExpressionSetExample,
```

bioRxiv preprint first posted online May. 3, 2016; doi: http://dx.doi.org/10.1101/051565. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

```
                    permutations  = 10000,
                    plotHistogram = TRUE )
```

A. Perform the reductive hourglass test for a DivergenceExpressionSet of step 3 and the TDI profile of step 9:

```
ReductiveHourglassTest( ExpressionSet  = DivergenceExpressionSetExample,
                        modules        = list(early = 1:2, mid = 3:5, late = 6:7),
                        permutations   = 10000,
                        plotHistogram  = TRUE )
```

B. Perform the reductive early-conservation test for a DivergenceExpressionSet of step 3 and the TDI profile of step 9:

```
EarlyConservationTest(  ExpressionSet  = DivergenceExpressionSetExample,
                        modules        = list(early = 1:2, mid = 3:5, late = 6:7),
                        permutations   = 10000,
                        plotHistogram  = TRUE )
```

13. Compute relative expression profiles for PS based on a PhyloExpressionSet of step 3:

```
REMatrix(PhyloExpressionSetExample)
```

14. Visualize relative expression profiles for PS from the data of step 13:

```
PlotRE( ExpressionSet = PhyloExpressionSetExample,
        Groups        = list(group = 1:12),
        legendName    = “PS”,
        xlab          = “Ontogeny”,
        lty           = 1,
        cex           = 0.7,
        lwd           = 5 )
```

bioRxiv preprint first posted online May. 3, 2016; doi: http://dx.doi.org/10.1101/051565. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

A. Visualize relative expression profiles based on groups of PS from the data of step 13 (Fig. 4 a):

```
PlotRE( ExpressionSet      = PhyloExpressionSetExample,
            Groups         = list(group_1 = 1:3, group_2 = 4:12),
            legendName  = "PS",
            xlab           = "Ontogeny",
            lty            = 1,
            cex            = 0.7,
           lwd             = 5 )
```

15. Compute relative expression profiles for DS based on a DivergenceExpressionSet of step 3:

```
REMatrix(DivergenceExpressionSetExample)
```

16. Visualize relative expression profiles for DS based on the data from step 15:

```
PlotRE( ExpressionSet = DivergenceExpressionSetExample,
         Groups         = list(group = 1:10),
         legendName  = "DS",
         xlab           = "Ontogeny",
         lty            = 1,
         cex            = 0.7,
         lwd            = 5 )
```

A. Visualize relative expression profiles based on groups of DS from the data of step 15 (Fig. 4 b):

```
PlotRE( ExpressionSet      = DivergenceExpressionSetExample,
            Groups         = list(group_1 = 1:5, group_2 = 6:10),
            legendName  = "DS",
            xlab           = "Ontogeny",
            lty            = 1,
```

bioRxiv preprint first posted online May. 3, 2016; doi: http://dx.doi.org/10.1101/051565. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

```
          cex            = 0.7,
          lwd            = 5 )
```

17. Quantify the statistical significance of group differences for PS based on the data of step 14A (Fig. 4 c):

```
PlotBarRE( ExpressionSet = PhyloExpressionSetExample,
            Groups         = list(group_1 = 1:3, group_2 = 4:12),
            xlab           = “Ontogeny”,
           ylab            = “Mean Relative Expression”,
          cex              = 2 )
```

18. Quantify the statistical significance of group differences for DS based on the data of step 16A (Fig. 4 c):

```
PlotBarRE( ExpressionSet = DivergenceExpressionSetExample,
            Groups         = list(group_1 = 1:5, group_2 = 5:10),
            xlab           = “Ontogeny”,
           ylab            = “Mean Relative Expression”,
          cex              = 2 )
```

Protocol Timing

Computing the phylostratigraphic map can take several hours, several days, or even several weeks depending on the number of query sequences and the size of the database of proteins of completely sequenced genomes and may thus require a computing cluster, which can easily run these jobs in parallel, reducing the time to only a few hours.

Computing the divergence stratigraphic map takes 2 - 4 hours on an ordinary PC (step 2). All other steps (3 – 18) take less than 10 minutes on an ordinary PC. Analyses of other data sets might take more or less time depending on the number of samples, genes, and permutations. Computation time and command-input time for most steps in the protocol are only a few seconds. The most time-consuming steps are the computation of phylostratigraphic maps or divergence stratigraphic maps in cases

bioRxiv preprint first posted online May. 3, 2016; doi: http://dx.doi.org/10.1101/051565. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

where they are not yet available, data formatting for obtaining *PhyloExpressionSets* and *DivergenceExpressionSets*, or performing permutation tests if a large number of permutations is chosen.

Protocol Results

Performing phylostratigraphy with Perl script createPSmap.pl and divergence stratigraphy with R package *orthologr* for the example data set (steps 1-2) yields a phylostratigraphic map (**Fig. 2**) and a divergence stratigraphic map for the studied species. Alternatively, users can input their own custom phylostratigraphic maps or divergence stratigraphic maps at this stage of the protocol (Supplementary Information). Matching these maps with gene expression data covering the studied developmental process (step 3) yields a PhyloExpressionSet (**Fig. 2 a**) and a DivergenceExpressionSet (**Fig. 2 b**). Applying steps 7–10 of the protocol to these two data sets yields TAI and TDI profiles for seven stages of *A. thaliana* embryo development (**Fig. 3**).

The *flat-line test* (steps 11-12) yields *p*-values of 1.2e-09 (TAI) and 8.8e-06 (TDI), stating that both profiles deviate significantly from a flat line. The *reductive hourglass test* (steps 11.A and 12.A) yields *p*-values of 8.0e-09 (TAI) and 4.8e-04 (TDI), stating that both profiles are compatible with an hourglass pattern. The *reductive early-conservation test* (steps 11.B and 12.B) yields *p*-values of 0.99 (TAI) and 0.96 (TDI), stating that both profiles deviate significantly from an early-conservation pattern.

Steps 13–16 yield relative expression profiles for all PS (**Fig. 4 a**) and DS (**Fig. 4 b**). For the example data set, we observe that evolutionarily conserved genes (genes of PS 1–3 that emerged before embryogenesis) and evolutionarily young genes (genes of PS 4–12 that emerged after embryogenesis) show qualitatively different relative expression profiles. Figure 4 a shows that high PS have similar relative expression profiles, whereas low PS have relative expression profiles that are dissimilar to each other and dissimilar to those of high PS. Figure 4 b shows that DS 2–10 have similar relative expression profiles, whereas DS 1 shows an antagonistic relative expression profile.

bioRxiv preprint first posted online May. 3, 2016; doi: http://dx.doi.org/10.1101/051565. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

Applying the Kruskal-Wallis rank-sum test to the observed relative expression levels of the two groups of PS (step 17) and the two groups of DS (step 18) yields *p*-values < 0.05 in both cases, stating that the observed differences of the relative expression levels between low and high PS and between low and high DS are statistically significant (**Fig. 4 c**).

Together, this protocol allows users to perform phylotranscriptomic analyses of biological processes of their interest in a standardized manner. The intuitive adaptability of the protocol to any species and any biological process is achieved by the modular structure of the protocol, Perl script createPSmap.pl, and R packages *orthologr* and *myTAI*, which provide an open-source implementation of the protocol. The Perl script and both R packages can be used for performing phylotranscriptomic analyses in a reproducible manner. The documentation of each function as well as the tutorials included in the Supplementary Information provide additional details of the functionality provided by createPSmap.pl, *orthologr*, and *myTAI* so that this protocol can be easily applied by a broad set of users to numerous of phylotranscriptomic studies of various biological processes in the future.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic Local Alignment Search Tool. *J. Mol. Biol.* 215:403-410.

Artieri CG, Haerty W, Singh RS. 2009. Ontogeny and phylogeny: molecular signatures of selection, constraint, and temporal pleiotropy in the development of Drosophila. *BMC Biol.* 7:42.

Von Baer KE. 1828. *Über Entwicklungsgeschichte der Thiere: Beobachtung und Reflexion.* Gebrüder Bornträger.

Chamberlain S, Szocs S. 2013. taxize - taxonomic search and retrieval in R. *F1000Research*. 2:191-219.

bioRxiv preprint first posted online May. 3, 2016; doi: http://dx.doi.org/10.1101/051565. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

Capra JA, Stolzer M, Durand D, Pollard KS. 2013. How old is my gene? *Trends in Genetics* 29 (11):659-668.

Capra JA, Williams AG, Pollard KS. 2012. ProteinHistorian: Tools for the Comparative Analysis of Eukaryote Protein Origin. *PloS Comp Biol*. 8 (6):e1002567.

Charif D, Lobry JR. 2007. Structural approaches to sequence evolution: Molecules, networks, populations. (eds. Bastolla, U. *et al.*) Springer Verlag. 207-232.

Cheng X, Hui JHL, Lee YY, Law PTW, Kwan HS. 2015. A Developmental Hourglass in Fungi. *Mol Biol Evol.* 32:1556-1566.

Cruickshank T, Wade MJ. 2008. Microevolutionary support for a developmental hourglass: gene expression patterns shape sequence variation and divergence in *Drosophila*. *Evol Dev.* 10:583–90.

De Mendoza A, Sebé-Pedrós A, Sestak MS, Matejcic M, Torruella G, Domazet-Lošo T, Ruiz-Trillo I. 2013. Transcription factor evolution in eukaryotes and the assembly of the regulatory toolkit in multicellular lineages. *Proc Natl Acad Sci. USA* 110:E4858–66.

Delignette-Muller ML, Dutang C. 2015. fitdistrplus: An R Package for Fitting Distributions. *Journal of Statistical Software*, 64(4):1-34.

Domazet-Lošo T, Brajković J, Tautz D. 2007. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends Genet.* 23:533-9.

Domazet-Lošo T, Tautz D. 2010a. A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature* 468:815-8.

Domazet-Lošo,T, Tautz, D. 2010b. Phylostratigraphic tracking of cancer genes suggests a link to the emergence of multicellularity in metazoa. *BMC Biology* 8:66.

bioRxiv preprint first posted online May. 3, 2016; doi: http://dx.doi.org/10.1101/051565. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

Dowle M. Srinivasan A, Short T, Lianoglou S, Saporta R, Antonyan E. 2014. data.table: Extension of data.frame. R package version 1.9.4. http://CRAN.R-project.org/package=data.table.

Drost HG, Gabel A, Grosse I, Quint M. 2015. Evidence for active maintenance of phylotranscriptomic hourglass patterns in animal and plant embryogenesis. *Mol Biol Evol.* 32:1221-1231.

Drost HG, Bellstädt J, Ó'Maoiléidigh DS, Silva AT, Gabel A, Weinholdt C, Ryan PT, Dekkers BJW, Bentsink L, Hilhorst H, Ligterink W, Wellmer F, Grosse I, and Quint M. 2016. Post-embryonic hourglass patterns mark ontogenetic transitions in plant development. *Mol. Biol. Evol.* 33(5): 1158-1163.

Drost HG. 2016. Performing Evolutionary Transcriptomics with R. R package version 0.0.4. Available from: http://CRAN.R-project.org/package=myTAI.

Duboule D. 1994. Temporal colinearity and the phylotypic progression: a basis for the stability of a vertebrate Bauplan and the evolution of morphologies through heterochrony. *Dev Suppl.*:135-142.

Dunn CW, Luo X, Wu Z. 2013. Phylogenetic Analysis of Gene Expression. *Integrative and Comparative Biology* 53(5):847–856.

Eddelbuettel D, Francois R. 2011. Rcpp: Seamless R and C++ Integration. *Journal of Statistical Software*. 40(8):1-18.

Gerstein MB, Rozowsky J, Yan KK, Wang D, Cheng C, Brown JB, Davis CA, Hillier L, Sisu C, Li JJ, Pei B, Harmanci AO, Duff MO, Djebali S, Alexander RP, Alver BH, Auerbach R, Bell K, Bickel PJ, Boeck ME, Boley NP, Booth BW, Cherbas L, Cherbas P, Di C, Dobin A, Drenkow J, Ewing B, Fang G, Fastuca M, Feingold EA, Frankish A, Gao G, Good PJ, Guigó R, Hammonds A, Harrow J, Hoskins RA, Howald C, Hu L, Huang H, Hubbard TJ, Huynh C, Jha S, Kasper D, Kato M, Kaufman TC, Kitchen RR, Ladewig E, Lagarde J, Lai E, Leng J, Lu Z, MacCoss M, May G, McWhirter R, Merrihew G, Miller DM, Mortazavi A, Murad R, Oliver B, Olson S, Park PJ, Pazin MJ, Perrimon N, Pervouchine D, Reinke V, Reymond A, Robinson

bioRxiv preprint first posted online May. 3, 2016; doi: http://dx.doi.org/10.1101/051565. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

G, Samsonova A, Saunders GI, Schlesinger F, Sethi A, Slack FJ, Spencer WC, Stoiber MH, Strasbourger P, Tanzer A, Thompson OA, Wan KH, Wang G, Wang H, Watkins KL, Wen J, Wen K, Xue C, Yang L, Yip K, Zaleski C, Zhang Y, Zheng H, Brenner SE, Graveley BR, Celniker SE, Gingeras TR, Waterston R. 2014. Comparative analysis of the transcriptome across distant species. *Nature* 512:445-8.

Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, Rokhsar DS. 2012. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* 40:D1178-86.

Gossmann TI, Saleh D, Schmid MW, Spence MA, Schmid KJ.
Transcriptomes of Plant Gametophytes Havea Higher Proportion of Rapidly Evolving an d Young Genes than Sporophytes. *Mol. Biol. Evol.* doi: 10.1093/molbev/msw044.

Gross J, Ligges U. 2014. nortest: Tests for Normality. R package version 1.0-2. http://CRAN.R-project.org/package=nortest.

Hazkani-Covo E, Wool D, Graur D. 2005. In search of the vertebrate phylotypic stage: a molecular examination of the developmental hourglass model and von Baer's third law. *J Exp Zool B Mol Dev Evol.* 304:150–158.

Hemmrich G, Khalturin K, Boehm AM, Puchert M, Anton-Erxleben F, Wittlieb J, Klostermeier UC, Rosenstiel P, Oberg HH, Domazet-Lošo T, Sugimoto T, Niwa H, Bosch TCG. 2012. Molecular Signatures of the Three Stem Cell Lineages in Hydra and the Emergence of Stem Cell Function at the Base of Multicellularity. *Mol. Biol. Evol.* 29(11):3267-3280.

Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, Bravo HC, Davis S, Gatto L, Girke T, Gottardo R, Hahne F, Hansen KD, Irizarry RA, Lawrence M, Love MI, MacDonald J, Obenchain V, Oleś AK, Pagès H, Reyes A, Shannon P, Smyth GK, Tenenbaum D, Waldron L, Morgan M. 2015. Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods* 12:115.

bioRxiv preprint first posted online May. 3, 2016; doi: http://dx.doi.org/10.1101/051565. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

Irie N, Sehara-Fujisawa A. 2007. The vertebrate phylotypic stage and an early bilaterian-related stage in mouse embryogenesis defined by genomic information. *BMC Biol.* 5:1.

Irie N, Kuratani S. 2011. Comparative transcriptome analysis reveals vertebrate phylotypic period during organogenesis. *Nat Commun.* 2:248.

Irie N, Kuratani S. 2014. The developmental hourglass model: a predictor of the basic body plan? *Development* 141:4649-4655.

Kalinka AT, Varga KM, Gerrard DT, Preibisch S, Corcoran DL, Jarrells J, Ohler U, Bergman CM, Tomancak P. 2010. Gene expression divergence recapitulates the developmental hourglass model. *Nature* 468:811–4.

Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ. 2013. Software for Computing and Annotating Genomic Ranges. *PLoS Comput Biol*. 9(8):e1003118.

Levin M, Hashimshony T, Wagner F, Yanai I. 2012. Developmental milestones punctuate gene expression in the Caenorhabditis embryo. *Dev. Cell* 22:1101-1108.

Levin M, Anavy L, Cole AG, Winter E, Mostov N, Khair S, Senderovich N, Kovalev E, Silver DH, Feder M, Fernandez-Valverde SL, Nakanishi N, Simmons D, Simakov O, Larsson T, Liu S-Y, Jerafi-Vider A, Yaniv K, Ryan JF, Martindale MQ, Rink JC, Arendt D, Degnan SM, Degnan BM, Hashimshony T & Yanai I. 2016. The mid-developmental transition and the evolution of animal body plans. *Nature*. Advance online publication doi:10.1038/nature16994.

Moyers BA, Zhang J. 2015. Phylostratigraphic Bias Creates Spurious Patterns of Genome Evolution. *Mol. Biol. Evol*. 32(1):258-267.

Moyers BA, Zhang J. 2016. Evaluating Phylostratigraphic Evidence for Widespread De Novo Gene Birth in Genome Evolution. *Mol. Biol. Evol*. 33(5):1245-1256.

bioRxiv preprint first posted online May. 3, 2016; doi: http://dx.doi.org/10.1101/051565. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

Neuwirth E. 2014. RColorBrewer: ColorBrewer Palettes. R package version 1.1-2. http://CRAN.R-project.org/package=RColorBrewer.

Nesculea A, Kaessmann H. 2014. Evolutionary dynamics of coding and non-coding transcriptomes. *Nature Reviews Genetics* 15:734-748.

Pages H, Aboyoun P, Gentleman R, DebRoy S. 2007. Biostrings: String objects representing biological sequences, and matching algorithms. R package version 2.36.1.

Pages H, Lawrence M, Aboyoun P. 2014. S4Vectors: S4 implementation of vectors and lists. R package version 0.6.1.

Quint M, Drost HG, Gabel A, Ullrich KK, Boenn M, Grosse I. 2012. A transcriptomic hourglass in plant embryogenesis. *Nature* 490:98-101.

Raff RA. 1996. *The Shape of Life: Genes, Development and the Evolution of Animal Form*. Univ. Chicago Press.

R Core Team. 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

R Special Interest Group on Databases. 2014. DBI: R Database Interface. R package version 0.3.1. http://CRAN.R-project.org/package=DBI.

Richardson MK. 1995. Heterochrony and the phylotypic period. *Dev Biol.* 172:412-21.

Richardson MK, Hanken J, Gooneratne ML, Pieau C, Raynaud A, Selwood L, Wright GM. 1997. There is no highly conserved embryonic stage in the vertebrates: implications for current theories of evolution and development. *Anat Embryol.* 196:91–106.

Richardson MK. 1999. Vertebrate evolution: The developmental origins of adult variation. *Bioessays* 21:604–13.

bioRxiv preprint first posted online May. 3, 2016; doi: http://dx.doi.org/10.1101/051565. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26:139-140.

Romero IG, Ruvinsky I, Gilad Y. 2012. Comparative studies of gene expression and the evolution of gene regulation. *Nature Reviews Genetics* 13:505-516.

Roux J, Rosikiewicz M, Robinson-Rechavi M. 2015. What to Compare and How: Comparative Transcriptomics for Evo-Devo. *J. Exp. Zool.* 324B:372-382.

Sander K. 1983. The evolution of patterning mechanisms: gleanings from insect embryogenesis and spermatogenesis. In: Goodwin BC, Holder N, Wylie C, editors. Development and evolution, The Sixth Symposium of the British Society for Developmental Biology. Cambridge: Cambridge University Press. p. 137–160 (1983).

Šestak MS, Božičević V, Bakarić R, Dunjko V, Domazet-Lošo T. 2013. Phylostratigraphic profiles reveal a deep evolutionary history of the vertebrate head sensory systems. Frontiers in Zoology 10:18.

Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JG, Korf I, Lapp H, Lehväslaiho H, Matsalla C, Mungall CJ, Osborne BI, Pocock MR, Schattner P, Senger M, Stein LD, Stupka E, Wilkinson MD, Birney E. 2002. The Bioperl toolkit: Perl modules for the life sciences. *Genome Res*. 12(10): 1611-1618.

Svorcová J. 2012. The phylotypic stage as a boundary of modular memory: non mechanistic perspective. *Theory Biosci*. 131:31-42.

Tautz D, Domazet-Lošo T. 2011. The evolutionary origin of orphan genes. *Nature Reviews Genetics* 12:692-702.

Tian X, Strassmann JE, Queller DC. 2013. Dictyostelium development shows a novel pattern of evolutionary conservation. *Mol Biol Evol.* 30(4):977-84.

bioRxiv preprint first posted online May. 3, 2016; doi: http://dx.doi.org/10.1101/051565. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

Wang Z, Pascual-Anaya J, Zadissa A, Li W, Niimura Y, Huang Z, Li C, White S, Xiong Z, Fang D, Wang B, Ming Y, Chen Y, Zheng Y, Kuraku S, Pignatelli M, Herrero J, Beal K, Nozawa M, Li Q, Wang J, Zhang H, Yu L, Shigenobu S, Wang J, Liu J, Flicek P, Searle S, Wang J, Kuratani S, Yin Y, Aken B, Zhang G, Irie N. 2013. The draft genomes of soft-shell turtle and green sea turtle yield insights into the development and evolution of the turtle-specific body plan. *Nat Genet.* 45:701–6.

Warnefors M, Kaessmann H. 2013. Evolution of the Correlation between Expression Divergence and Protein Divergence in Mammals. *Genome Biology and Evolution*. 5(7):1324-1335.

Willmore KE. The Body Plan Concept and Its Centrality in Evo-Devo. 2012. *Evo Edu Outreach*. 5:219-230.

Weston S. 2014. doParallel: Foreach parallel adaptor for the parallel package. R package version 1.0.8. http://CRAN.R-project.org/package=doParallel.

Wickham H, Francois R. 2015. dplyr: A Grammar of Data Manipulation. R package version 0.4.1. http://CRAN.R-project.org/package=dplyr.
Wickham H. 2009. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York.

Wickham H. James DA, Falcon S, Healy L. 2014. RSQLite: SQLite Interface for R. R package version 1.0.0 http://CRAN.R-project.org/package=RSQLite.

Wichham H. 2015. stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.0.0. http://CRAN.R-project.org/package=stringr.

Yanai I, Peshkin L, Jorgensen P, Kirschner MW. 2011. Mapping gene expression in two Xenopus species: evolutionary constraints and developmental flexibility. *Dev Cell.* 20:483-96.

Xiang D, Venglat P, Tibiche C, Yang H, Risseeuw E, Cao Y, Babic V, Cloutier M, Keller W, Wang E, Selvaraj G, Datla R. 2011. Genome-wide analysis reveals gene expression and

bioRxiv preprint first posted online May. 3, 2016; doi: http://dx.doi.org/10.1101/051565. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

metabolic network dynamics during embryo development in Arabidopsis. *Plant Physiol.* 156:346-56.

Zhang J and Yang JR. 2015. Determinants of the rate of protein sequence evolution. *Nature Reviews Genetics* 16:409–420.

Author Contributions

HGD conceived the protocol. HGD, AG, TDL, MQ, and IG designed the protocol. HGD designed and implemented *orthologr* and *myTAI*. AG designed and implemented *createPSmap.pl*. HGD, TDL, MQ, and IG wrote the manuscript.

Acknowledgments

We thank Jan Grau, Anne Hoffmann, Philipp Janitza, Kristian Karsten Ullrich, Sarah Scharfenberg, Claus Weinholdt, and Sebastian Wussow for valuable discussions and Deutsche Forschungsgemeinschaft (grants no. FZT 118, GR 3526/2, GR 3526/6, GR 3526/7, GR 3526/8, Qu 141/5, Qu 141/6, and Qu 141/7) and SKW Piesteritz for financial support.

Competing financial interests

The authors declare no competing financial interests.

Figures

bioRxiv preprint first posted online May. 3, 2016; doi: http://dx.doi.org/10.1101/051565. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

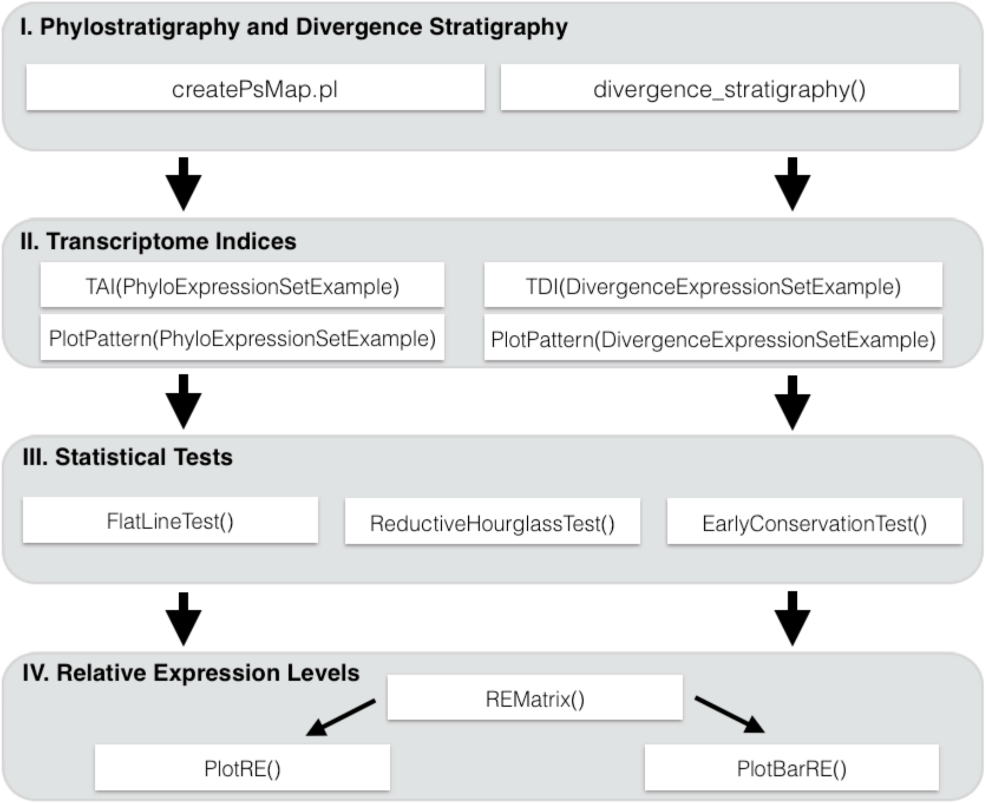

I. Phylostratigraphy and Divergence Stratigraphy
createPsMap.pl
divergence_stratigraphy()
II. Transcriptome Indices
TAI(PhyloExpressionSetExample)
TDI(DivergenceExpressionSetExample)
PlotPattern(PhyloExpressionSetExample)
PlotPattern(DivergenceExpressionSetExample)
III. Statistical Tests
FlatLineTest()
ReductiveHourglassTest()
EarlyConservationTest()
IV. Relative Expression Levels
REMatrix()
PlotRE()
PlotBarRE()

Figure 1. Flow Chart. A step-by-step instruction illustrating the workflow of the protocol.

bioRxiv preprint first posted online May. 3, 2016; doi: http://dx.doi.org/10.1101/051565. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

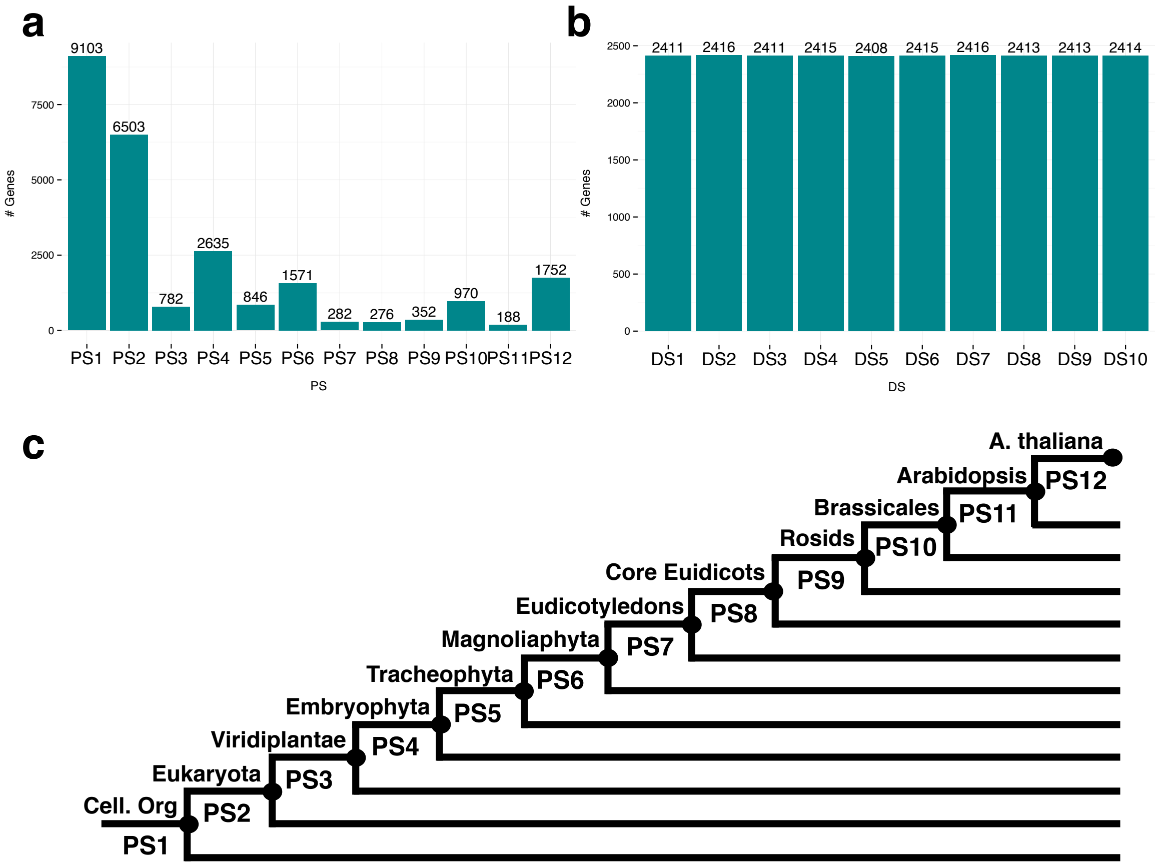


Figure 2. Histograms of phylostrata and divergence strata and phylostratigraphic map. PS histogram (**a**) and DS histogram (**b**) of *A. thaliana* genes. The uniform DS histogram is due to the definition as deciles. **c.** Phylostratigraphic map of *A. thaliana*. The most distant taxonomic category is PS1 (cellular organisms) and the closest category is PS12 (*A. thaliana*).

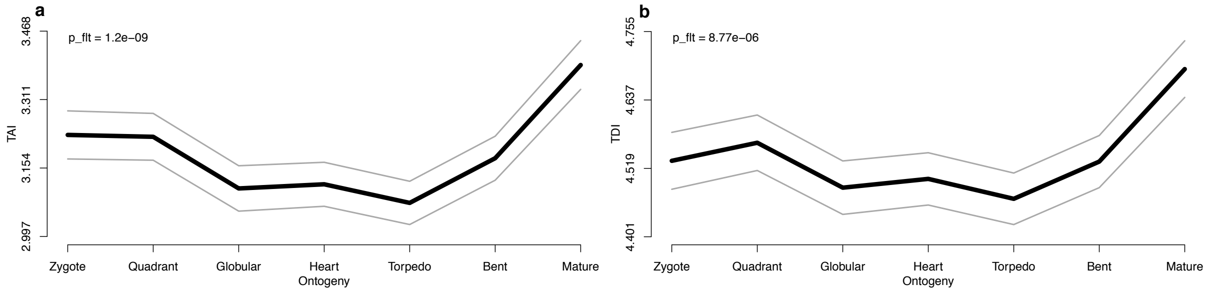


Figure 3. Transcriptome indices of *A. thaliana* embryogenesis. Visualization of (**a**) the TAI profile and (**b**) the TDI profile covering seven stages of *A. thaliana* embryogenesis. Grey lines represent the standard deviation of randomly permuted TAI or TDI profiles. The statistical significance of observed patterns (*p*-values) are computed using the *flat-line test* (Drost et al. 2015).

bioRxiv preprint first posted online May. 3, 2016; doi: http://dx.doi.org/10.1101/051565. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

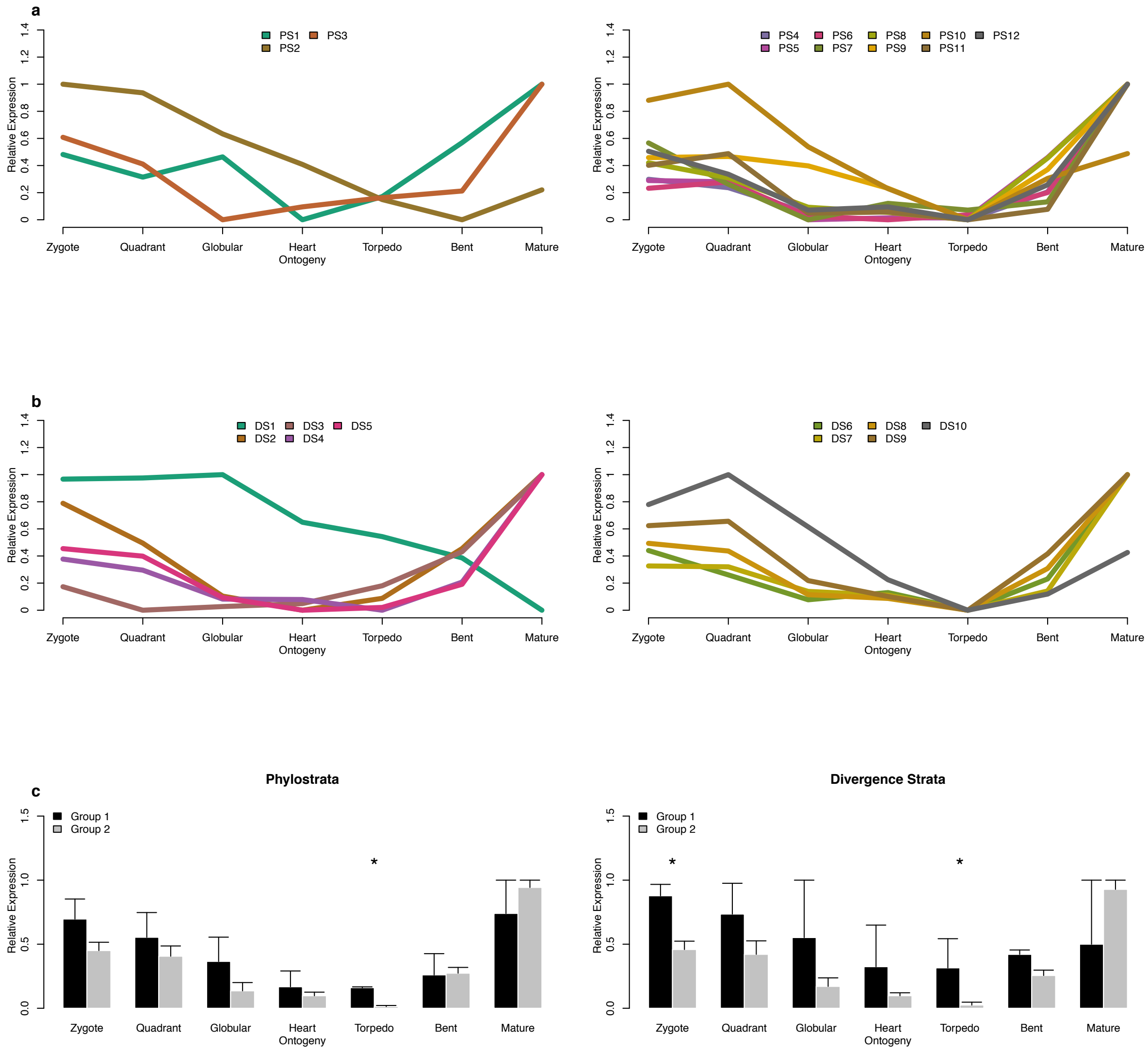
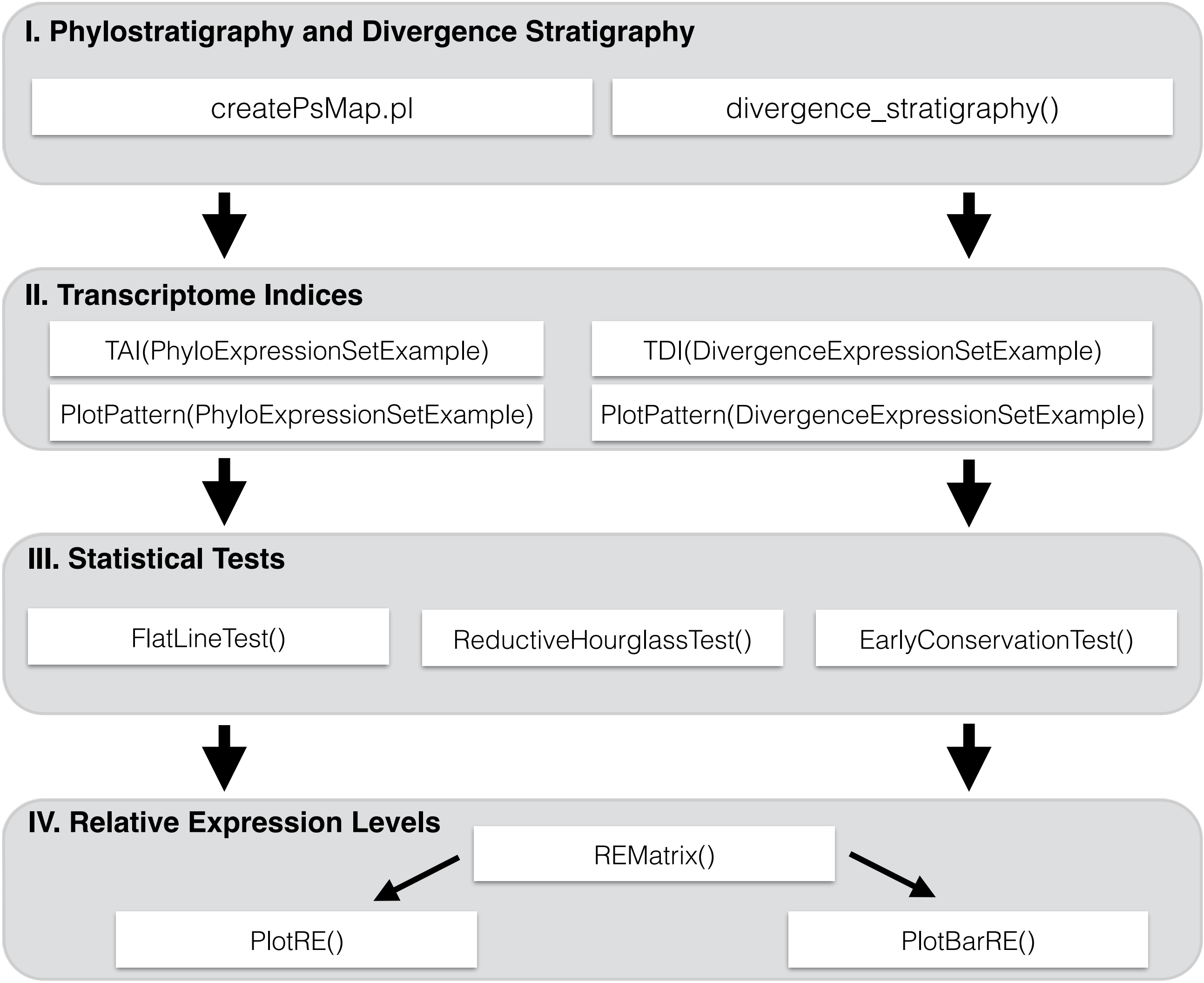
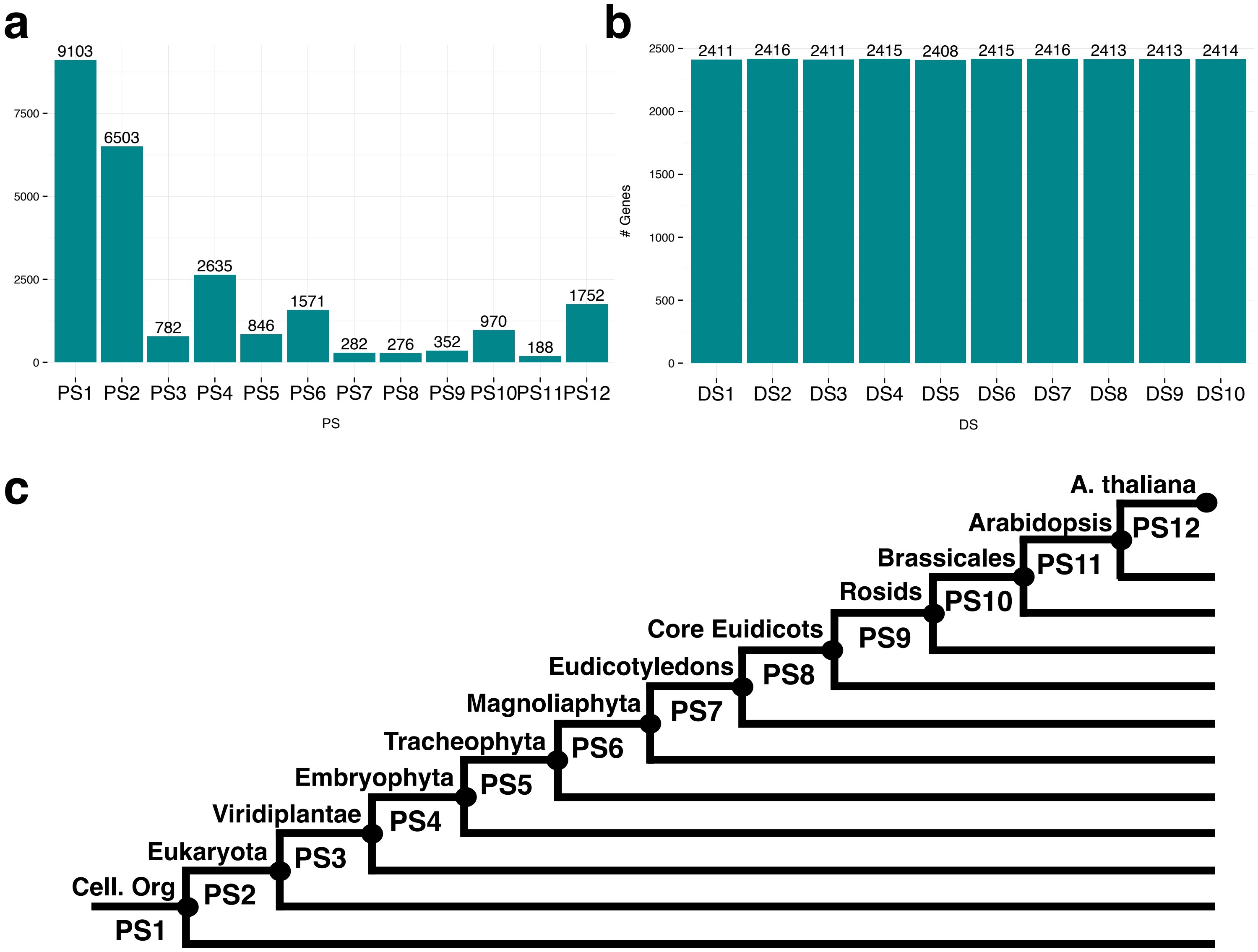


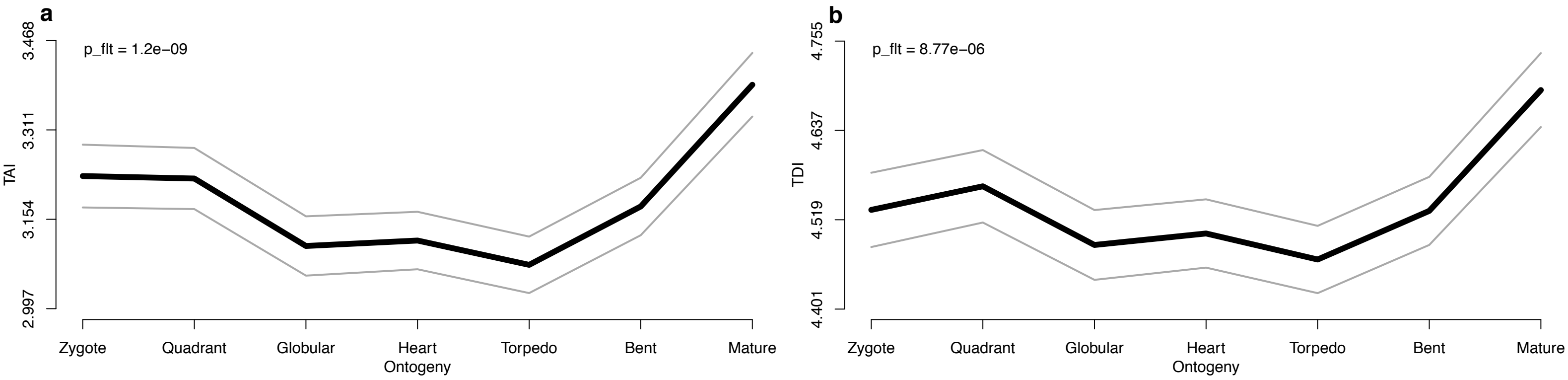
Figure 4. Relative expression profiles of *A. thaliana* embryogenesis. Relative expression profiles of twelve PS (**a**) covering seven stages of *A. thaliana* embryogenesis. PS are divided into two groups to analyze co-expression patterns of PS before (PS 1-3) and after (PS 4-12) the emergence of embryogenesis in plant evolution. Relative expression profiles of ten DS (**b**) of the same stages. DS are divided into two groups (DS 1-2 versus DS 3-10) to analyze co-expression patterns of DS with highly negative (purifying) selection (DS 1-2) and more relaxed or even positive selection (DS 3-10). **c.** Bar plots of mean relative expression levels for PS and DS groups. *p*-values of the difference of mean relative expression levels between PS groups or DS groups are obtained by a Kruskal-Wallis rank-sum test. Developmental stages with significant

bioRxiv preprint first posted online May. 3, 2016; doi: http://dx.doi.org/10.1101/051565. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

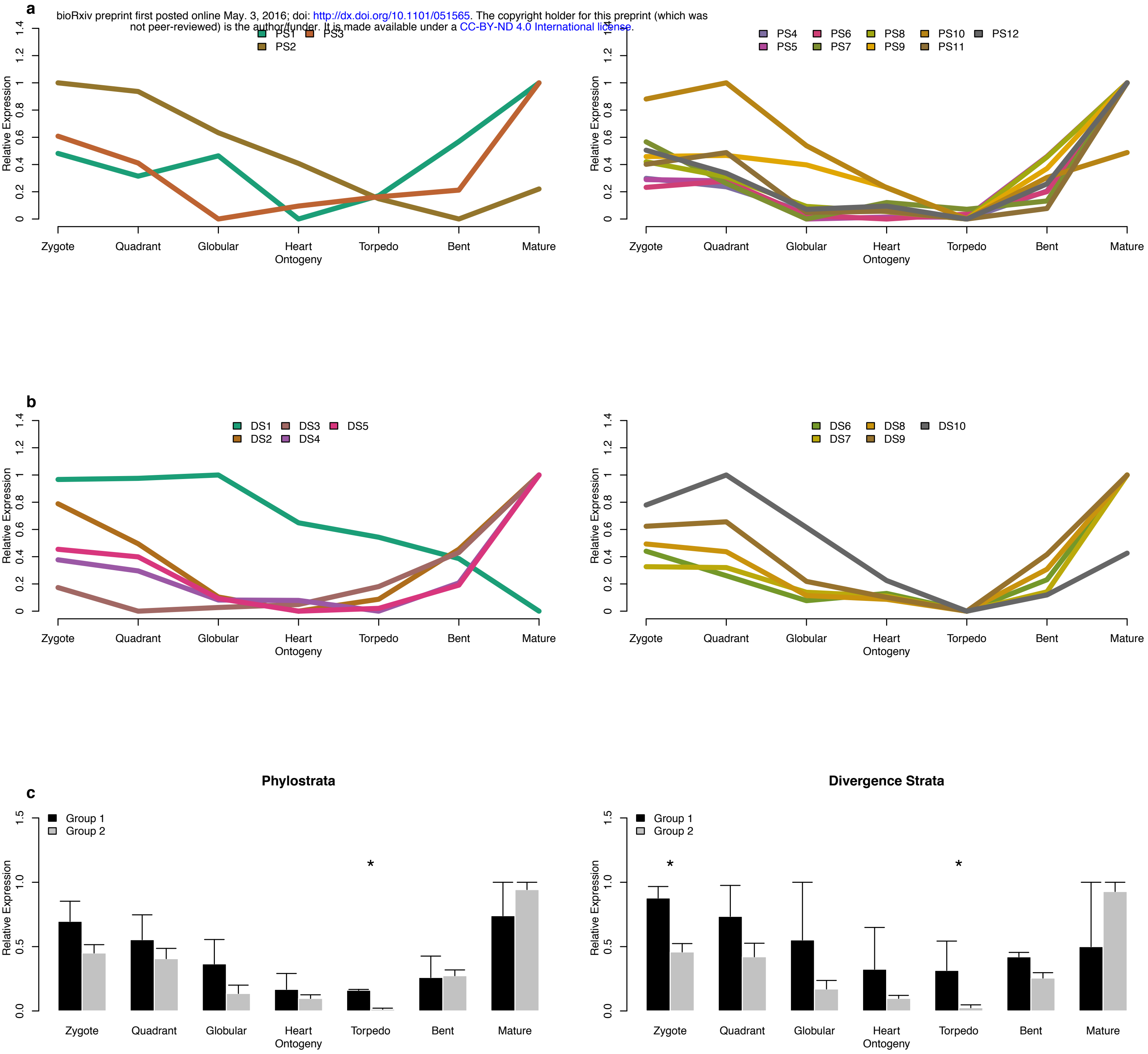
differences of mean relative expression levels are marked by asterisks (Drost et al. 2015).


I. Phylostratigraphy and Divergence Stratigraphy
createPsMap.pl
divergence_stratigraphy()
II. Transcriptome Indices
TAI(PhyloExpressionSetExample)
TDI(DivergenceExpressionSetExample)
PlotPattern(PhyloExpressionSetExample)
PlotPattern(DivergenceExpressionSetExample)
III. Statistical Tests
FlatLineTest()
ReductiveHourglassTest()
EarlyConservationTest()
IV. Relative Expression Levels
REMatrix()
PlotRE()
PlotBarRE()


a
9103
6503
782
2635
846
1571
282
276
352
970
188
1752
0
2500
5000
7500
# Genes
PS1
PS2
PS3
PS4
PS5
PS6
PS7
PS8
PS9
PS10
PS11
PS12
PS
b
2411
2416
2411
2415
2408
2415
2416
2413
2413
2414
0
500
1000
1500
2000
2500
# Genes
DS1
DS2
DS3
DS4
DS5
DS6
DS7
DS8
DS9
DS10
DS
c
A. thaliana
Arabidopsis
PS12
Brassicales
PS11
Rosids
PS10
Core Euidicots
PS9
Eudicotyledons
PS8
Magnoliaphyta
PS7
Tracheophyta
PS6
Embryophyta
PS5
Viridiplantae
PS4
Eukaryota
PS3
Cell. Org
PS2
PS1


a
p_flt = 1.2e−09
TAI
3.468
3.311
3.154
2.997
Zygote
Quadrant
Globular
Heart
Torpedo
Bent
Mature
Ontogeny
b
p_flt = 8.77e−06
TDI
4.755
4.637
4.519
4.401
Zygote
Quadrant
Globular
Heart
Torpedo
Bent
Mature
Ontogeny

bioRxiv preprint first posted online May. 3, 2016; doi: http://dx.doi.org/10.1101/051565. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-ND 4.0 International license.



10.2 A transcriptomic hourglass in plant embryogenesis

The following paper entitled *A transcriptomic hourglass in plant embryogenesis* is not part of the publications which comprise this cumulative dissertation. However, its content summarizes the major findings of my co-authors and me upon which my thesis is built. Due to this fact, I decided to include this paper as appendix to this thesis.

- `http://www.nature.com/nature/journal/v490/n7418/full/nature11394.html`

10.3 Abstract

Animal and plant development starts with a constituting phase called embryogenesis, which evolved independently in both lineages. Comparative anatomy of vertebrate development based on the Meckel-Serre's law and von Baer's laws of embryology 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*, and its bottleneck of high conservation in mid-embryogenesis is referred to as the phylotypic stage. Recent analyses in zebrafish and Drosophila embryos provided convincing molecular support for the hourglass model, because during the phylotypic stage the transcriptome was dominated by ancient genes and global gene expression profiles were reported to be most conserved. Although extensively explored in animals, an embryonic hourglass has not been reported in plants, which represent the second major kingdom in the tree of life that evolved embryogenesis. Here we provide phylotranscriptomic evidence for a molecular embryonic hourglass in Arabidopsis thaliana, using two complementary approaches. This is particularly significant because the possible absence of an hourglass based on morphological features in plants suggests that morphological and molecular patterns might be uncoupled. Together with the reported developmental hourglass patterns in animals, these findings indicate convergent evolution of the molecular hourglass and a conserved logic of embryogenesis across kingdoms.

10.4 Documentation: Introduction to myTAI

Introduction to the myTAI Package

2016-06-25

Installation

Users can download `myTAI` from CRAN :

```
# install myTAI 0.4.0 from CRAN
install.packages("myTAI", dependencies = TRUE)
```

Package Dependencies

```
# to perform differential gene expression analyses with myTAI
# please install the edgeR package
# install edgeR
source("http://bioconductor.org/biocLite.R")
biocLite("edgeR")
```

Overview of the functions implemented in `myTAI`

Phylotranscriptomics Measures:

- `TAI()` : Function to compute the Transcriptome Age Index (TAI)
- `TDI()` : Function to compute the Transcriptome Divergence Index (TDI)
- `REMatrix()` : Function to compute the relative expression profiles of all phylostrata or divergence-strata
- `RE()` : Function to transform mean expression levels to relative expression levels
- `pTAI()` : Compute the Phylostratum Contribution to the global TAI
- `pTDI()` : Compute the Divergence Stratum Contribution to the global TDI
- `pMatrix()` : Compute Partial TAI or TDI Values
- `pStrata()` : Compute Partial Strata Values

Visualization and Analytics Tools:

- `PlotPattern()` : Function to plot the TAI or TDI profiles and perform statistical tests
- `PlotCorrelation()` : Function to plot the correlation between phylostratum values and divergence-stratum values
- `PlotRE()` : Function to plot the relative expression profiles
- `PlotBarRE()` : Function to plot the mean relative expression levels of phylostratum or divergence-stratum classes as barplot
- `PlotMeans()` : Function to plot the mean expression profiles of phylostrata or divergence-strata
- `PlotDistribution()` : Function to plot the frequency distribution of genes within the corresponding phylostratigraphic map or divergence map
- `PlotContribution()` : Plot the Phylostratum or Divergence Stratum Contribution to the Global TAI/TDI Pattern
- `PlotEnrichment()` : Plot the Phylostratum or Divergence Stratum Enrichment of a given Gene Set
- `PlotGeneSet()` : Plot the Expression Profiles of a Gene Set
- `PlotCategoryExpr()` : Plot the Expression Levels of each Age or Divergence Category as Barplot or Violinplot

- `PlotGroupDiffs()` : Plot the significant differences between gene expression distributions of PS or DS groups
- `PlotSelectedAgeDistr()` : Plot the PS or DS distribution of a selected set of genes

A Statistical Framework and Test Statistics:

- `FlatLineTest()` : Function to perform the **Flat Line Test** that quantifies the statistical significance of an observed phylotranscriptomics pattern (significant deviation from a frat line = evolutionary signal)
- `ReductiveHourglassTest()` : Function to perform the **Reductive Hourglass Test** that statistically evaluates the existence of a phylotranscriptomic hourglass pattern (hourglass model)
- `EarlyConservationTest()` : Function to perform the **Reductive Early Conservation Test** that statistically evaluates the existence of a monotonically increasing phylotranscriptomic pattern (early conservation model)
- `EnrichmentTest()` : Phylostratum or Divergence Stratum Enrichment of a given Gene Set based on Fisher's Test
- `bootMatrix()` : Compute a Permutation Matrix for Test Statistics

All functions also include visual analytics tools to quantify the goodness of test statistics.

Differential Gene Expression Analysis

- `DiffGenes()` : Implements Popular Methods for Differential Gene Expression Analysis
- `CollapseReplicates()` : Combine Replicates in an ExpressionSet
- `CombinatorialSignificance()` : Compute the Statistical Significance of Each Replicate Combination
- `Expressed()` : Filter Expression Levels in Gene Expression Matrices (define expressed genes)
- `SelectGeneSet()` : Select a Subset of Genes in an ExpressionSet
- `PlotReplicateQuality()` : Plot the Quality of Biological Replicates
- `GroupDiffs()` : Quantify the significant differences between gene expression distributions of PS or DS groups

Taxonomic Information Retrieval

- `taxonomy()` : Retrieve Taxonomic Information for any Organism of Interest

Minor Functions for Better Usibility and Additional Analyses

- `MatchMap()` : Match a Phylostratigraphic Map or Divergence Map with a ExpressionMatrix
- `tf()` : Transform Gene Expression Levels
- `age.apply()` : Age Category Specific apply Function
- `ecScore()` : Compute the Hourglass Score for the EarlyConservationTest
- `geom.mean()` : Geometric Mean
- `harm.mean()` : Harmonic Mean
- `omitMatrix()` : Compute TAI or TDI Profiles Omitting a Given Gene
- `rhScore()` : Compute the Hourglass Score for the Reductive Hourglass Test

Perform Evolutionary Transcriptomics with R

Gene age inference is the foundation of most evolutionary transcriptomics studies. The concept behind evolutionary transcriptomics is to combine these gene age estimates with gene expression data to quantify the average transcriptome age within a biological process of interest.

In particular, this approach allowed the quantification of transcriptome conservation of animal and plant embryos passing through embryogenesis by first individually estimating the gene ages of specific animal and plant genomes and combining these gene age estimates with transcriptome data covering several stages of embryo development (Domazet-Loso and Tautz, 2010 *Nature* ; Quint, Drost et al., 2012 *Nature* ; Drost et al., 2015 *Mol. Biol. Evol.* ; Drost et al., 2016 *Mol. Biol. Evol.*).

The myTAI package aims to provide a standard tool for evolutionary transcriptomics studies and relies on gene age and sequence conservation estimates as input. This approach allows researchers to study the evolution of biological processes and to detect stages or periods of evolutionary conservation or variability.

Getting Started

As intensely discussed in the past years (Capra et al., 2013; Altenhoff et al., 2016; Liebeskind et al., 2016), gene age inference is not a trivial task and might be biased in some currently existing approaches (Liebeskind et al., 2016).

In particular, Moyers & Zhang argue that genomic phylostratigraphy (a prominent BLAST based gene age inference method) 1) underestimates gene age for a considerable fraction of genes, 2) is biased for rapidly evolving proteins which are short, and/or their most conserved block of sites is small, and 3) these biases create spurious nonuniform distributions of various gene properties among age groups, many of which cannot be predicted a priori (Moyers & Zhang, 2015; Moyers & Zhang, 2016; Liebeskind et al., 2016). However, these arguments are based on simulated data (and the approach has been questioned; pers. comm. with Tomislav Domazet-Lošo) and therefore an objective benchmarking set representing the tree of life is still missing.

Based on this debate a recent study suggested to perform gene age inference by combining thirteen common orthology inference algorithms to create gene age datasets and then characterize the error around each age-call on a per-gene and per-algorithm basis. Using this approach systematic error was found to be a large factor in estimating gene age, suggesting that simple consensus algorithms are not enough to give a reliable point estimate (Liebeskind et al., 2016). However, by generating a consensus gene age and quantifying the possible error in each workflow step, Liebeskind et al., 2016 provide a very useful database of consensus gene ages for a variety of genomes.

Alternatively, Stephen Smith, 2016 argues that *de novo* gene birth/death and gene family expansion/contraction studies should avoid drawing direct inferences of evolutionary relatedness from measures of sequence similarity alone, and should instead, where possible, use more rigorous phylogeny-based methods. For this purpose, I recommend researchers to consult the phylomedb database to retrieve phylogeny-based gene orthology relationships and use these age estimates in combination with `myTAI`.

Evidently, these advancements in gene age research are very recent and gene age inference is a very young and active field of genomic research. Therefore, many more studies need to address the robust and realistic inference of gene age and a community standard is still missing.

Despite the ongoing debate about how to correctly infer gene age, users of `myTAI` can perform any gene age inference method they find most appropriate for their biological question and pass this gene age inference table as input to `myTAI`. To do so, users need to follow the following data format specifications to use their gene age inference table with `myTAI`.

The rational behind gene age inference is to assign each protein coding gene of an organism of iterest with an evolutionary age estimate which aims to quantify its potential origin within the tree of life (detectable sequence homolog; orphan gene (see Tautz & Domazet-Lošo, 2011)). Hence, gene age inference generates a table storing the gene age in the first column and the corresponding gene id of the organism of iterest in the second column. This table is named *phylostratigraphic map*.

Alternatively, we recently proposed an alternative method (Divergence Stratigraphy) to quantify the sequence conservation of protein coding genes between closely related species to study the active maintenance of transcriptome conservation patterns that were captured using a gene age inference approach (Drost et al., 2015 *Mol. Biol. Evol.*). Analogous to gene age inference methods, divergence stratigraphy generates a table storing the sequence conservation estimate in the first column and the corresponding gene id of the organism of iterest in the second column. This table is named *divergence stratigraphic map*.

`myTAI` takes either a *phylostratigraphic map* or a *divergence stratigraphic map* and an expression dataset as input and provides functions to quantify the average transcriptome age or average transcriptome conservation within a biological process of interest.

The three input tables: *phylostratigraphic map*, *divergence stratigraphic map*, and the expression dataset need to fulfill specific data formats when using `myTAI`.

The following code illustrates an example structure of a *phylostratigraphic map* and *divergence stratigraphic map*:

```r
# load myTAI
library(myTAI)

# load example data sets (stored in myTAI)
data(PhyloExpressionSetExample)
data(DivergenceExpressionSetExample)

# show an example phylostratigraphic map of Arabidopsis thaliana
head(PhyloExpressionSetExample[ , c("Phylostratum","GeneID")])
```

```
  Phylostratum      GeneID
1            1 at1g01040.2
2            1 at1g01050.1
3            1 at1g01070.1
4            1 at1g01080.2
5            1 at1g01090.1
6            1 at1g01120.1
```

In detail, a *phylostratigraphic map* stores the gene age assignment generated with phylostratigraphy in the first columns and the corresponding gene id in the second column.

Analogously, a divergence stratigraphic map stores the gene age assignment generated with divergence stratigraphy in the first column and the corresponding gene id in the second column:

```r
# show an example structure of a Divergence Map
head(DivergenceExpressionSetExample[ , c("Divergence.stratum","GeneID")])
```

```
  Divergence.stratum      GeneID
1                  1 at1g01050.1
2                  1 at1g01120.1
3                  1 at1g01140.3
4                  1 at1g01170.1
5                  1 at1g01230.1
6                  1 at1g01540.2
```

Hence, myTAI relies on pre-computed *phylostratigraphic maps* and *divergence stratigraphic maps*. For this purpose, users can consult the following resources to generate or retrieve *phylostratigraphic maps* and *divergence stratigraphic maps* which can then be used with myTAI.

Generate or retrieve phylostratigraphic maps

- generate a phylostratigraphic map with createPSmap.pl (implemented by Alexander Gabel)
- generate a phylostratigraphic map with phylostratigraphy.pl (implemented by Cheng et al. 2015)
- generate a phylostratigraphic map with phylo_pipeline.py (implemented by Shuqing Xu)
- generate a phylostratigraphic map with ORFanFinder
- generate a gene age map with Protein Historian
- download pre-computed and published phylostratigraphic maps
- use a gene age consensus approach to estimating gene ages for model organisms Liebeskind et al., 2016

Generate or retrieve divergence stratigraphic maps

- generate a divergence stratigraphic map with orthologr (implemented by Hajk-Georg Drost)
- generate a divergence stratigraphic map with compute_dNdS.pl (implemented by Cheng et al. 2015)
- retrieve phylogeny-based orthology and paralogy predictions from MetaPhOrs
- download pre-computed and published divergence stratigraphic maps

In general, users can construct their own gene age assignment methods and are not limited to the methods listed above. After formatting the corresponding results to the *phylostratigraphic map* or *divergence stratigraphic map* specification (age assignment in the first column and gene id in the second column), users can use any function in myTAI with their custom gene age assignment table.

Expression dataset specification

The aim of any phylotranscriptomics study is to quantify transcriptome conservation in biological processes. For this purpose, users need to provide the transcriptome dataset of their studied biological process.

In the following examples we will use a `gene expression dataset` covering seven stages of *Arabidopsis thaliana* embryo development. This data format is defined as `ExpressionMatrix` in the `myTAI` data format specification.

```
# gene expression set
```

```
       GeneID     Zygote   Quadrant   Globular      Heart    Torpedo       Bent    Mature
1 at1g01040.2  2173.6352  1911.2001  1152.5553  1291.4224  1000.2529   962.9772 1696.4274
2 at1g01050.1  1501.0141  1817.3086  1665.3089  1564.7612  1496.3207  1114.6435 1071.6555
3 at1g01070.1  1212.7927  1233.0023   939.2000   929.6195   864.2180   877.2060  894.8189
4 at1g01080.2  1016.9203   936.3837  1181.3381  1329.4734  1392.6429  1287.9746  861.2605
5 at1g01090.1 11424.5667 16778.1685 34366.6493 39775.6405 56231.5689 66980.3673 7772.5617
6 at1g01120.1   844.0414   787.5929   859.6267   931.6180   942.8453   870.2625  792.7542
```

The function `MatchMap()` allows users to join a *phylostratigraphic map* with an *ExpressionMatrix* to obtain a joined table referred to as *PhyloExpressionSet*. In some cases, the GeneIDs stored in the `ExpressionMatrix` and in the *phylostratigraphic map* do not match. This is due to GeneID mappings between different databases and annotations. To map non matching GeneIDs between databases and annotations, please consult the Functional Annotation Vignette in the biomartr package. The `biomartr` package allows users to map GeneIDs between database annotations.

After matching a *phylostratigraphic map* with an *ExpressionMatrix* using the `MatchMap()` function, a standard *PhyloExpressionSet* is returned storing the phylostratum assignment of a given gene in the first column, the *gene id* of the corresponding gene in the second column, and the entire gene expression set (time series or treatments) starting with the third column. This format is crucial for all functions that are implemented in the `myTAI` package.

```r
library(myTAI)

# load the example data set
data(PhyloExpressionSetExample)

# construct an example Phylostratigraphic Map
Example.PhylostratigraphicMap <- PhyloExpressionSetExample[ , 1:2]
# construct an example ExpressionMatrix
Example.ExpressionMatrix <- PhyloExpressionSetExample[ , 2:9]

# join a PhylostratigraphicMap with an ExpressionMatrix using MatchMap()
Example.PhyloExpressionSet <- MatchMap(Example.PhylostratigraphicMap,
                                       Example.ExpressionMatrix)

# look at a standard PhyloExpressionSet
str(Example.PhyloExpressionSet)
```

```
#> 'data.frame':    25260 obs. of  9 variables:
#>  $ Phylostratum: int  4 2 3 1 1 2 1 1 1 2 ...
#>  $ GeneID      : chr  "at1g01010.1" "at1g01020.1" "at1g01030.1" "at1g01040.2" ...
#>  $ Zygote      : num  878 1005 819 2174 1501 ...
#>  $ Quadrant    : num  828 1106 772 1911 1817 ...
#>  $ Globular    : num  776 1038 811 1153 1665 ...
#>  $ Heart       : num  754 939 867 1291 1565 ...
#>  $ Torpedo     : num  775 962 774 1000 1496 ...
#>  $ Bent        : num  756 871 748 963 1115 ...
#>  $ Mature      : num  1000 998 786 1696 1072 ...
```

Analogous to a standard *PhyloExpressionSet*, a standard *DivergenceExpressionSet* is a `data.frame` storing the divergence stratum assignment of a given gene in the first column, the gene id of the corresponding gene in the second column, and the entire gene expression set (time series or treatments) starting with the third column.

The following `DivergenceExpressionSet` example illustrates the standard `DivergenceExpressionSet` data set format.

```r
# head of an example standard DivergenceExpressionSet
str(DivergenceExpressionSetExample)
```

```
#> 'data.frame':    24132 obs. of  9 variables:
#>  $ Divergence.stratum: int  1 1 1 1 1 1 1 1 1 1 ...
#>  $ GeneID            : Factor w/ 24132 levels "at1g01010.1",..: 5 12 14 17 24 53 61 88 91 97 ...
#>  $ Zygote            : num  1501 844 1041 1362 894 ...
#>  $ Quadrant          : num  1817 788 908 1042 947 ...
#>  $ Globular          : num  1665 860 1069 1226 933 ...
#>  $ Heart             : num  1565 932 968 1212 965 ...
#>  $ Torpedo           : num  1496 943 1055 1675 871 ...
#>  $ Bent              : num  1115 870 1109 2136 843 ...
#>  $ Mature            : num  1072 793 825 10662 795 ...
```

A *DivergenceExpressionSet* defines the joined table between a *divergence stratigraphic map* and a *Expression Set*. A *DivergenceExpressionSet* can be generated analogous to a *PhyloExpressionSet* by joining a *divergence stratigraphic map* with an *ExpressionMatrix* using the `MatchMap()` function. In some cases, the GeneIDs stored in the *ExpressionMatrix* and in the *divergence stratigraphic map* do not match. This is due to GeneID mappings between different databases and annotations. To map non matching GeneIDs between databases and annotations, please consult the Functional Annotation Vignette in the biomartr package.

Each function implemented in `myTAI` checks internally whether or not the *PhyloExpressionSet* or *DivergenceExpressionSet* standard is fulfilled.

```
# used by all myTAI functions to check the validity of the PhyloExpressionSet standard
is.ExpressionSet(PhyloExpressionSetExample)
```

```
[1] TRUE
```

In case the PhyloExpressionSet standard is violated, the `is.ExpressionSet()` function will return `FALSE` and the corresponding function within the `myTAI` package will return an error message.

```
# used a non standard PhyloExpressionSet
head(PhyloExpressionSetExample[ , 2:5], 2)
```

```
       GeneID   Zygote Quadrant Globular
1 at1g01040.2 2173.635 1911.200 1152.555
2 at1g01050.1 1501.014 1817.309 1665.309
```

```
is.ExpressionSet(PhyloExpressionSetExample[ , 2:5])
```

```
Error in is.ExpressionSet(PhyloExpressionSetExample[, 2:5]) :
  The present input object does not fulfill the ExpressionSet standard.
```

The PhyloExpressionSet and DivergenceExpressionSet formats are crucial for all functions that are implemented in the `myTAI` package.

Keeping these standard data formats in mind will provide users with the most important requirements to get started with the `myTAI` package.

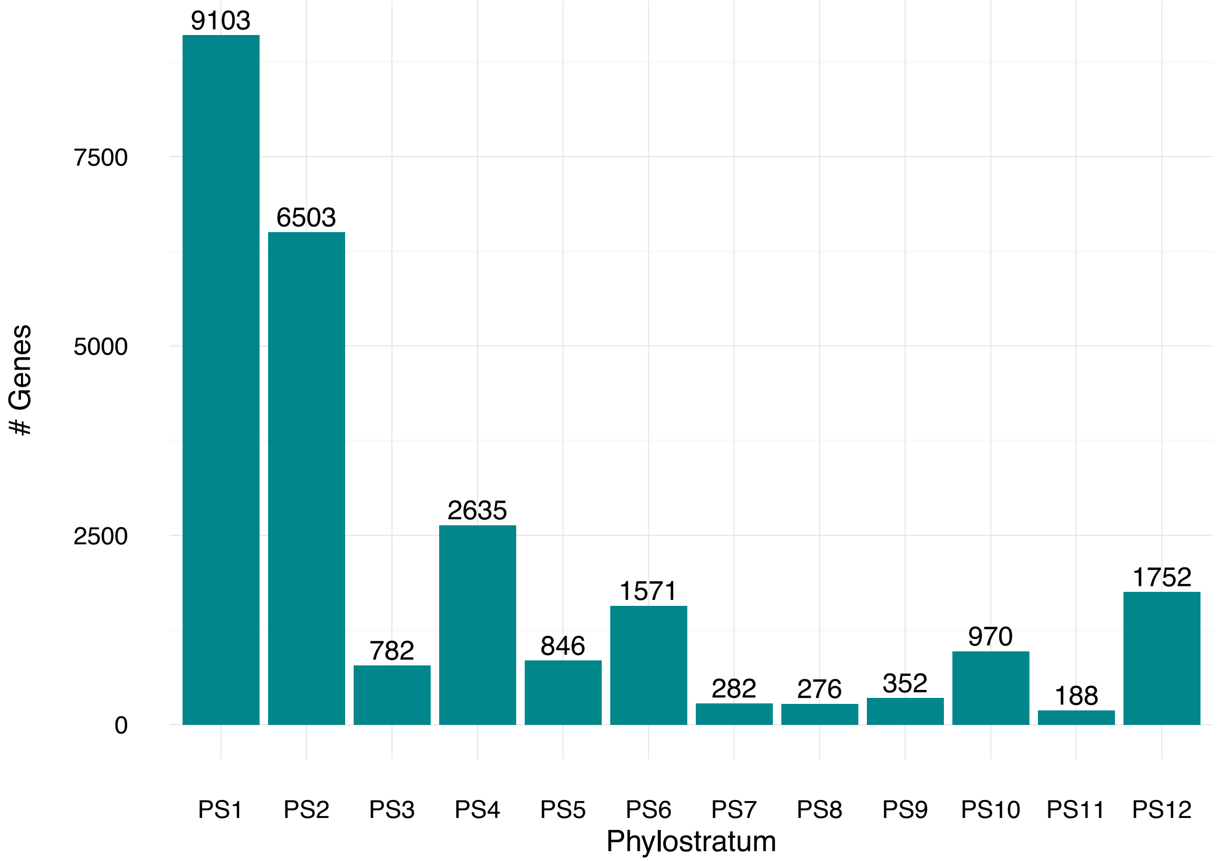
Note, that within the code of each function, the argument `ExpressionSet` always refers to either a PhyloExpressionSet or a DivergenceExpressionSet, whereas in specialized functions some arguments are specified as *PhyloExpressionSet* when they take an PhyloExpressionSet as input data set, or specified as *DivergenceExpressionSet* when they take an *DivergenceExpressionSet* as input data set.

Performing a Standard Workflow for Phylotranscriptomics Analyses

In the beginning of each phylotranscriptomics study users should investigate the distribution of PS or DS within a given PhyloExpressionSet or DivergenceExpressionSet.

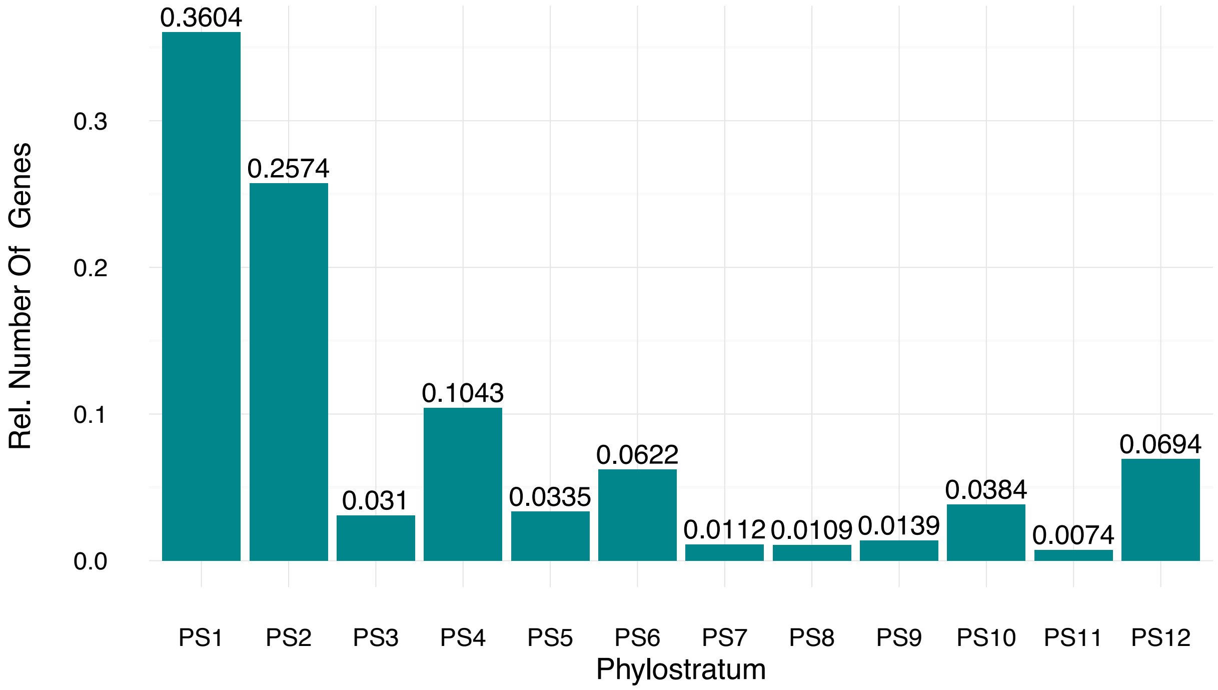
For this purpose, the `PlotDistribution()` function was implemented:

```
# Display the phylostratum distribution (gene frequency distribution)
# of a PhyloExpressionSet as absolute frequency distribution
PlotDistribution( PhyloExpressionSet = PhyloExpressionSetExample,
                  xlab               = "Phylostratum" )
```



or display it as relative frequencies:

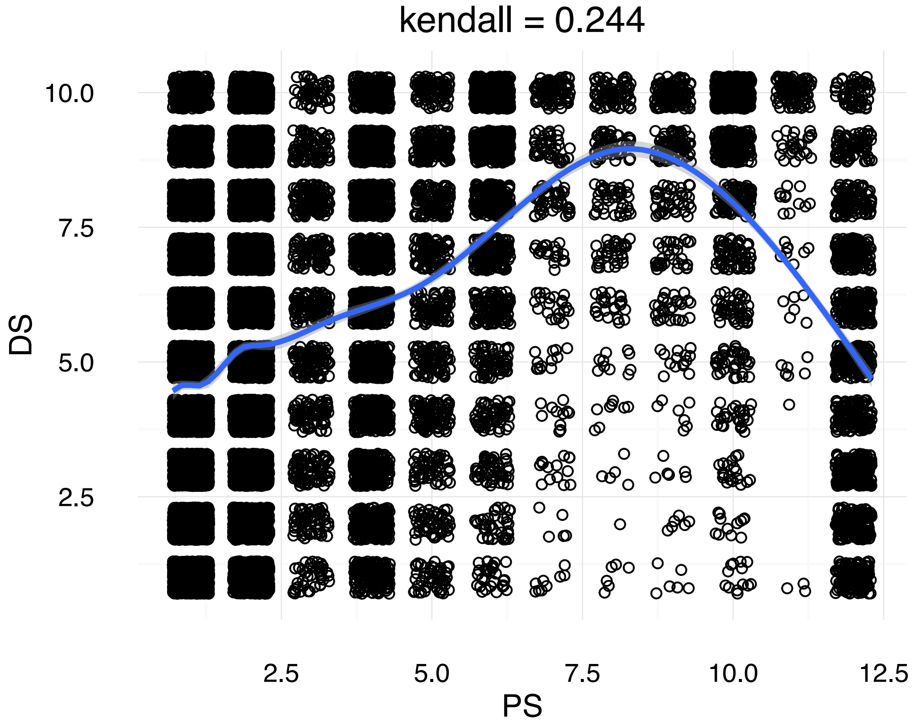
```r
# Plot phylostrata as relative frequency distribution
PlotDistribution( PhyloExpressionSet = PhyloExpressionSetExample,
                  as.ratio           = TRUE,
                  xlab               = "Phylostratum")
```



Another important feature to check is whether the phylostratum assignment and divergence stratum assignment of the genes stored within the PhyloExpressionSet and DivergenceExpressionSet are correlated (linear dependent). This is important to be able to assume the linear independence of **TAI** and **TDI** measures. This step is useful, because the TAI and TDI measures aim to quantify different signatures of evolutionary conservation. Whereas the TAI measure aims to quantify evolutionary signatures based on gene origin along the tree of life, the TDI measure is based on quantifying the selection pressure acting on orthologous genes between closely related species.

For this purpose the `PlotCorrelation()` function was implemented:

```
# Visualizing the correlation between Phylostratum and Divergence-Stratum assignments
# of the intersecting set of genes that are stored within the PhyloExpressionSet
# and DivergenceExpressionSet

PlotCorrelation( PhyloExpressionSet      = PhyloExpressionSetExample,
                 DivergenceExpressionSet = DivergenceExpressionSetExample,
                 method                  = "kendall",
                 linearModel             = TRUE )
```



In this case phylostratum and divergence stratum assignments of the intersecting set of genes that are stored within the PhyloExpressionSet and DivergenceExpressionSet are weakly correlated, but can be assumed to be linear independent.

Note: The `PlotCorrelation()` function always takes a PhyloExpressionSet as **first argument** and a DivergenceExpressionSet as **second argument**.

Visualizing the *Transcriptome Age Index* and the *Transcriptome Divergence Index*

Mathematically, the *Transcriptome Age Index* (TAI) introduced by Domazet-Loso and Tautz, 2010 represents a weighted arithmetic mean of the transcriptome age during a corresponding developmental stage s.

$TAI_s = \sum_{i=1}^{n} \frac{ps_i * e_{is}}{\sum_{i=1}^{n} e_{is}}$

where ps_i denotes the phylostratum assignment of gene i and e_{is} denotes the gene expression level of gene i at developmental time point s. A lower value of TAI describes an older transcriptome age, whereas a higher value of TAI denotes a younger transcriptome age.

Analogous to the TAI measure, the *Transcriptome Divergence Index* (TDI) was introduced by Quint et al., 2012 and Drost et al., 2015 as global measure of average transcriptome selection pressure where s denotes the corresponding developmental stage.

$TDI_s = \sum_{i=1}^{n} \frac{ds_i * e_{is}}{\sum_{i=1}^{n} e_{is}}$

where ds_i denotes the divergence stratum assignment of gene i and e_{is} denotes the gene expression level of gene i at developmental time point s. A lower value of TDI describes an more conserved transcriptome (in terms of sequence dissimilarity), whereas a higher value of TDI denotes a more variable transcriptome.

Transcriptome Age Index Analyses

Evolutionary signatures of transcriptomes can be captured by computing transcriptome indices at different measured stages of development, combining these computed values to a transcriptome index profile across the measured stages, and comparing the resulting profile with a flat line. A profile not significantly deviating from a flat line indicates the absence of significant variations of the computed transcriptome index from stage to stage. In contrast, a profile significantly deviating from a flat line indicates the presence of significant variations from stage to stage. We refer to any transcriptome index profile significantly deviating from a flat line as phylotranscriptomic pattern or evolutionary signature.

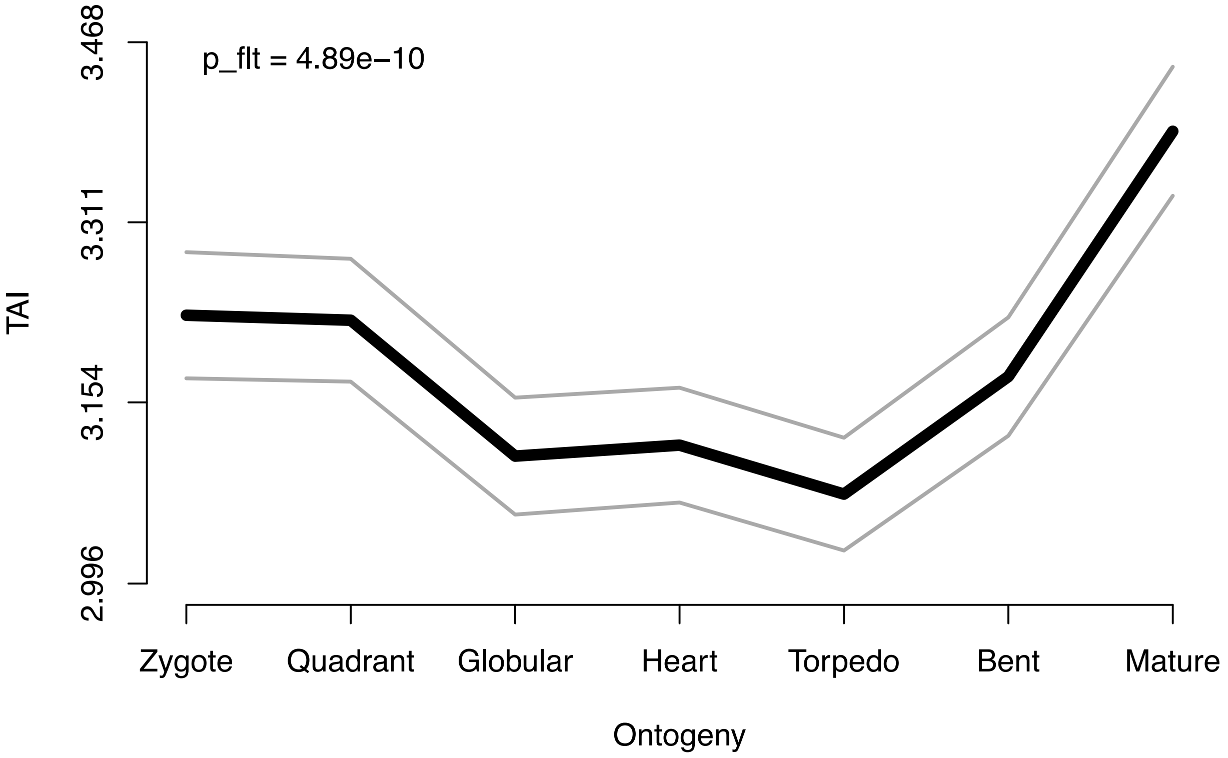
Previously, we introduced three statistical tests to quantify the significance of observed TAI or TDI patterns: `Flat Line Test`, `Reductive Hourglass Test`, and `Reductive Early Conservation Test` (Drost et al., 2015).

The `PlotPattern()` function introduced in the following sections is the main analytics function of myTAI. `PlotPattern()` allows users to visualize TAI or TDI patterns and internally performs the following statistical tests to assess their significance.

Flat Line Test

The `PlotPattern()` function with option `TestStatistic = "FlatLineTest"`, first computes the `TAI` (given a PhyloExpressionSet) or the `TDI` (given a DivergenceExpressionSet) profile as well as their standard deviation, and statistical significance.

```r
# Plot the Transcriptome Age Index of a given PhyloExpressionSet
# Test Statistic : Flat Line Test (default)
PlotPattern( ExpressionSet = PhyloExpressionSetExample,
             TestStatistic = "FlatLineTest",
             type          = "l",
             lwd           = 6,
             xlab          = "Ontogeny",
             ylab          = "TAI" )
```



The p-value (`p_flt`) above the TAI curve is returned by the `FlatLineTest`. As described in the documentation of `PlotPattern()` (`?PlotPattern` or `?FlatLineTest`), the `FlatLineTest` is the default statistical test to quantify the statistical significance of the observed phylotranscriptomic pattern. In detail, the test quantifies any statistically significant deviation of the phylotranscriptomic pattern from a flat line. Here, we define any significant deviation of a phylotranscriptomic pattern from a flat line as evolutionary signature Furthermore, we define corresponding stages of deviation as evolutionary conserved or variable (less conserved) depending on the magnitude of `TAI` and corresponding p-values.

Reductive Hourglass Test

In case the observed phylotranscriptomic pattern not only significantly deviates from a flat line but also visually resembles an *hourglass* shape, one can obtain a p-value quantifying the statistical significance of a visual *hourglass* pattern based on the `ReductiveHourglassTest` (`?ReductiveHourglassTest`).

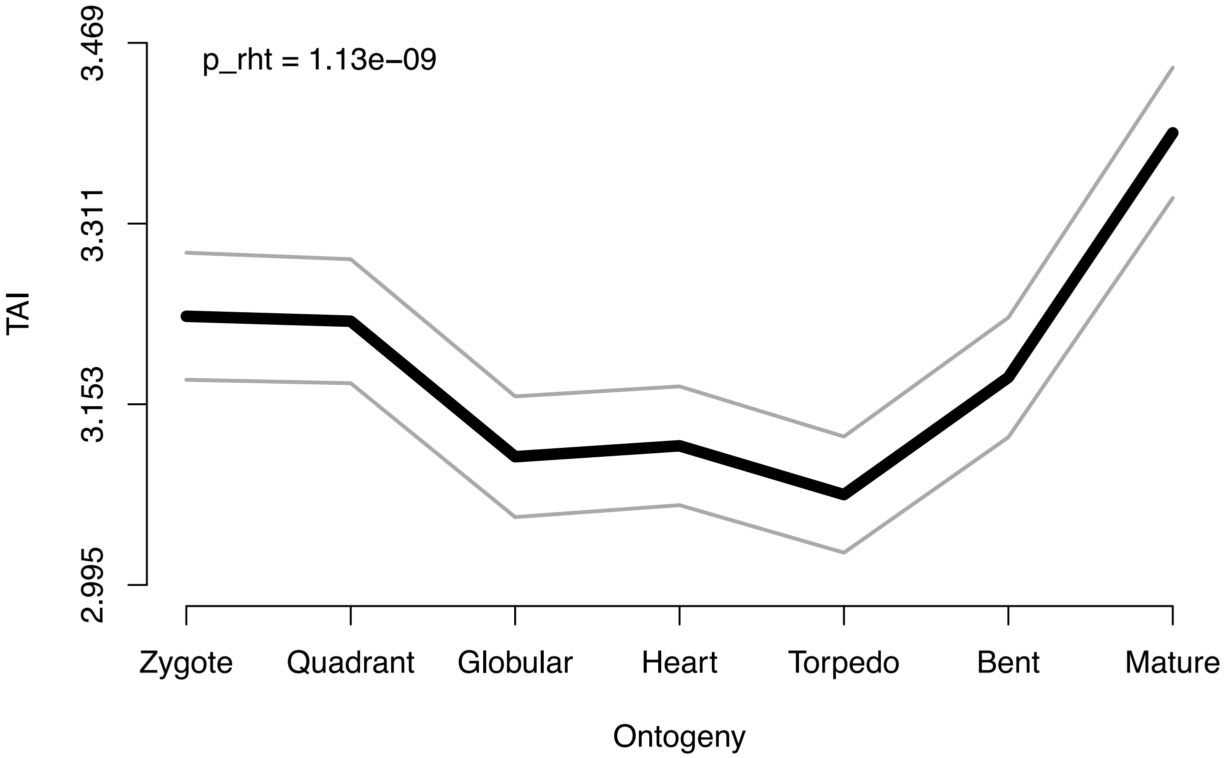
Since the `ReductiveHourglassTest` has been defined for a priori biological knowledge (Drost et al., 2015), the `modules` argument within the `ReductiveHourglassTest()` function needs to be specified.

Three modules need to be specified: an **early-module**, **phylotypic module** (mid), and a **late-module**.

For this example we divide *A. thaliana* embryo development stored within the *PhyloExpressionSetExample* into the following three modules:

- early = stages 1 - 2 (Zygote and Quadrant)
- mid = stages 3 - 5 (Globular, Heart, and Torpedo)
- late = stages 6 - 7 (Bent and Mature)

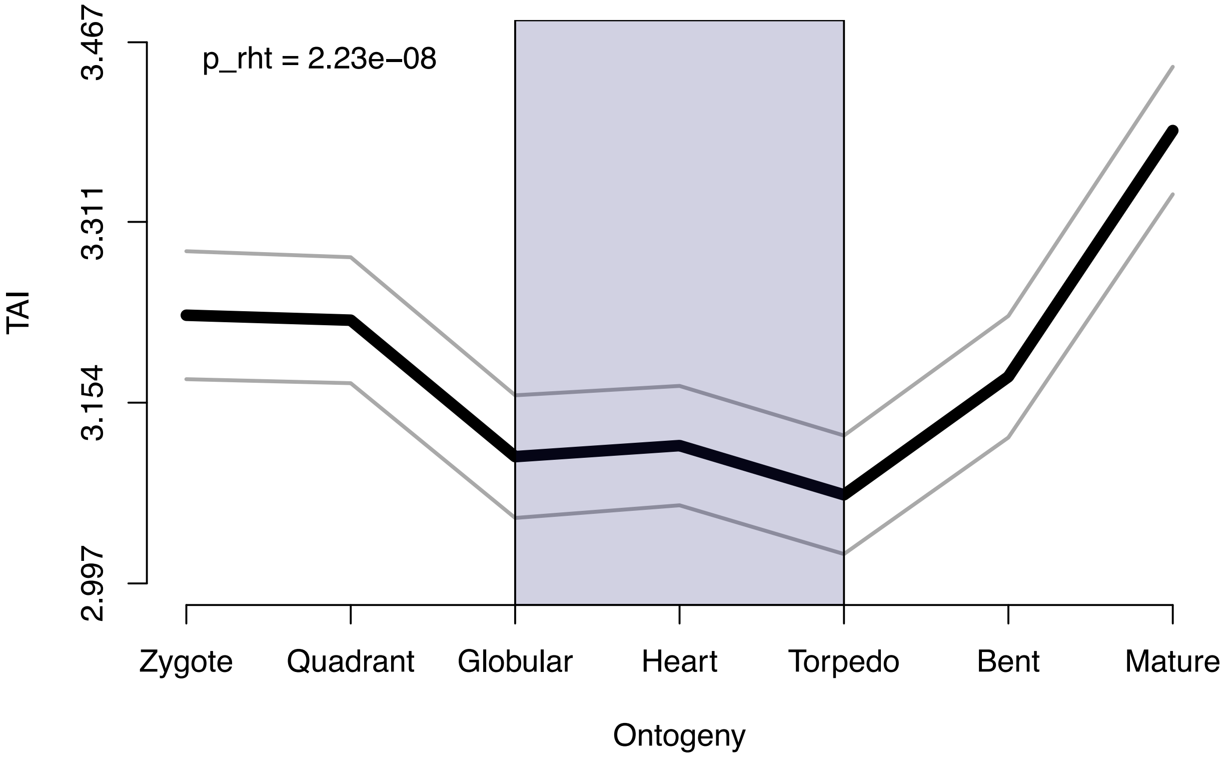
```
# Plot the Transcriptome Age Index of a given PhyloExpressionSet
# Test Statistic : Reductive Hourglass Test
PlotPattern( ExpressionSet = PhyloExpressionSetExample,
             TestStatistic = "ReductiveHourglassTest",
             modules       = list(early = 1:2, mid = 3:5, late = 6:7),
             type          = "l",
             lwd           = 6,
             xlab          = "Ontogeny",
             ylab          = "TAI" )
```



The corresponding p-value `p_rht` now denotes the p-value returned by the `ReductiveHourglassTest` which is different from the p-value returned by the `FlatLineTest` (`p_flt`).

To make sure that correct modules have been selected to perform the `ReductiveHourglassTest`, users can use the `shaded.area` argument to visualize chosen modules:

```
# Visualize the phylotypic period used for the Reductive Hourglass Test
PlotPattern( ExpressionSet = PhyloExpressionSetExample,
             TestStatistic = "ReductiveHourglassTest",
             modules       = list(early = 1:2, mid = 3:5, late = 6:7),
             shaded.area   = TRUE,
             type          = "l",
             lwd           = 6,
             xlab          = "Ontogeny",
             ylab          = "TAI" )
```



Note that for defining a priori knowledge for the `ReductiveHourglassTest` using the `modules` argument, modules need to start at stage 1, . . . , N and do not correspond to the column position in the *PhyloExpressionSet/DivergenceExpressionSet* which in contrast would start at position 3, . . . N + 2.

Reductive Early Conservation Test

The third test statistic that is implemented in the `myTAI` package is the `EarlyConservationTest`.

The `EarlyConservationTest` tests whether an observed phylotranscriptomic pattern follows a low-high-high pattern (monotonically increasing function) supporting the Early Conservation Model of embryogenesis.

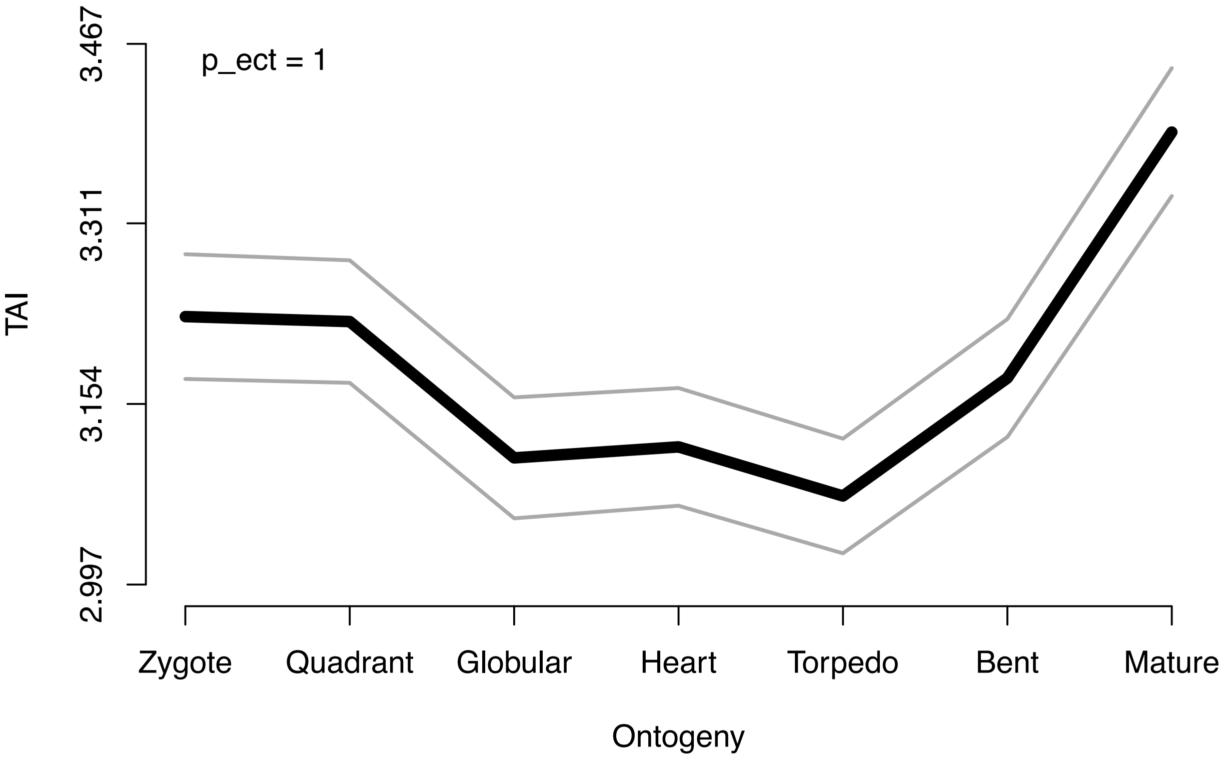
Analogous to the `ReductiveHourglassTest`, the `EarlyConservationTest` needs a priori biological knowledge Drost et al., 2015. So again three `modules` have to be specified for the `EarlyConservationTest()` function.

Three modules need to be specified: an **early-module**, **phylotypic module** (mid), and a **late-module**.

For this example we divide *A. thaliana* embryo development stored within the *PhyloExpressionSetExample* into the following three modules:

- early = stages 1 - 2 (Zygote and Quadrant)
- mid = stages 3 - 5 (Globular, Heart, and Torpedo)
- late = stages 6 - 7 (Bent and Mature)

```
# Plot the Transcriptome Age Index of a given PhyloExpressionSet
# Test Statistic : Reductive Early Conservation Test
PlotPattern( ExpressionSet = PhyloExpressionSetExample,
             TestStatistic = "EarlyConservationTest",
             modules       = list(early = 1:2, mid = 3:5, late = 6:7),
             type          = "l",
             lwd           = 6,
             xlab          = "Ontogeny",
             ylab          = "TAI" )
```



The corresponding p-value `p_ect` now denotes the p-value returned by the `EarlyConservationTest` which is different from the p-value returned by the `FlatLineTest` (`p_flt`) and `ReductiveHourglassTest` (`p_rht`).

Since the present TAI pattern of the *PhyloExpressionSetExample* doesn't support the Early Conservation Hypothesis, the p-value `p_ect` = 1.

Again **note** that for defining a priori knowledge for the `EarlyConservationTest` using the `modules` argument, modules need to start at stage 1, . . . , N and do not correspond to the column position in the *PhyloExpressionSet/DivergenceExpressionSet* which in contrast would start at position 3, . . . N + 2.

To obtain the numerical *TAI* values, the `TAI()` function can be used:

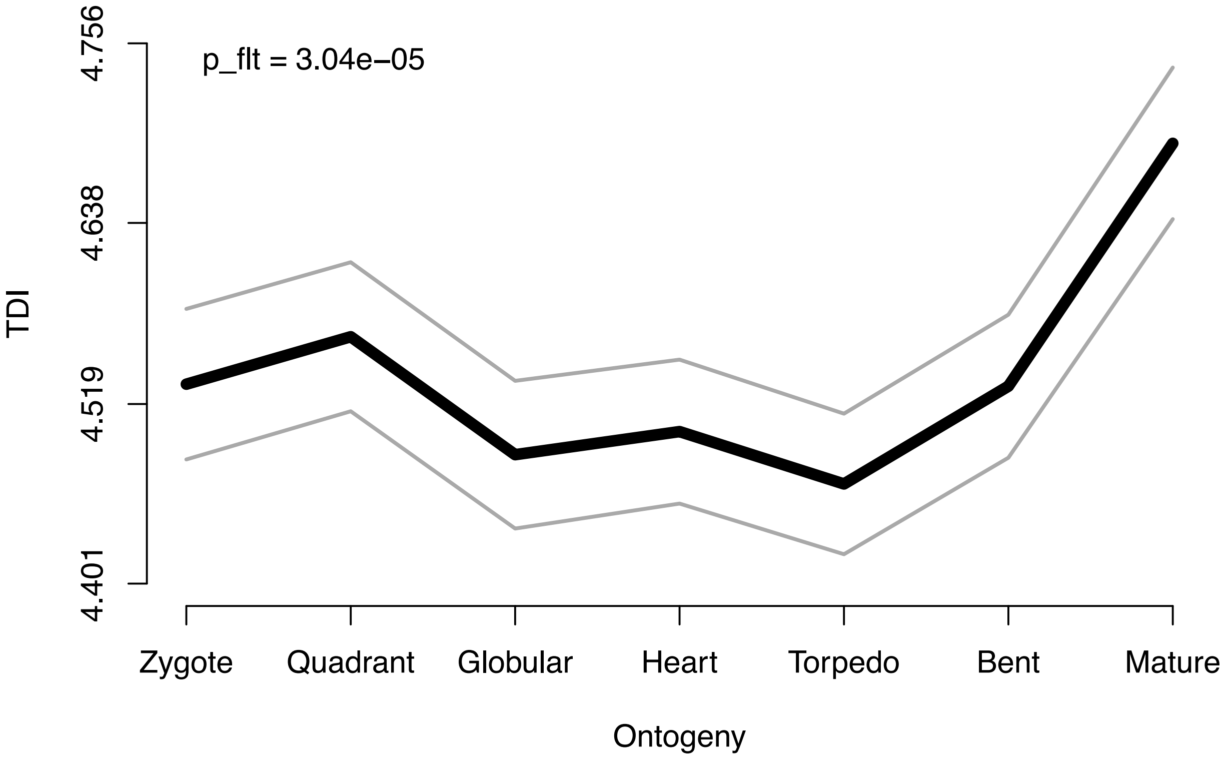
```
# Compute the Transcriptome Age Index values of a given PhyloExpressionSet
TAI(PhyloExpressionSetExample)
```

```
  Zygote Quadrant Globular    Heart  Torpedo     Bent   Mature
3.229942 3.225614 3.107135 3.116693 3.073993 3.176511 3.390334
```

Transcriptome Divergence Index Analyses

Analogous to the *TAI* computations and visualization, the *TDI* computations can be performed in a similar fashion:

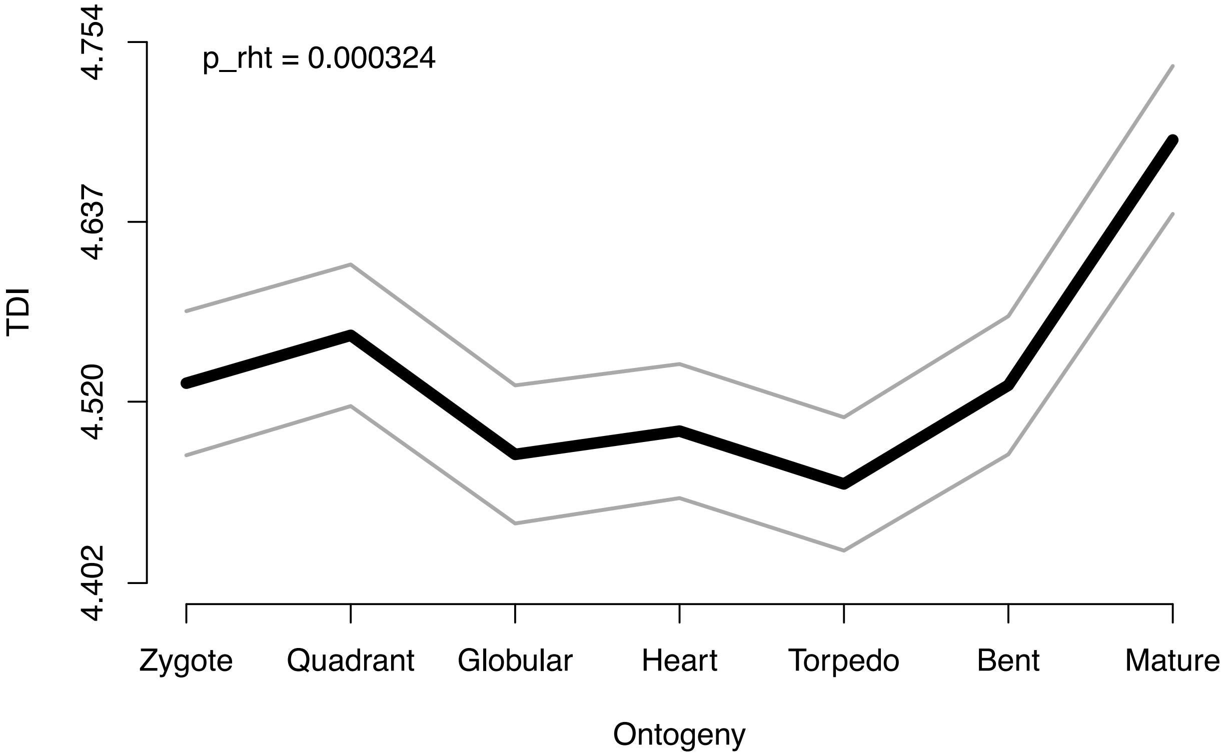
```
# Plot the Transcriptome Divergence Index of a given DivergenceExpressionSet
# Test Statistic : Flat Line Test (default)
PlotPattern( ExpressionSet = DivergenceExpressionSetExample,
             type          = "l",
             lwd           = 6,
             xlab          = "Ontogeny",
             ylab          = "TDI" )
```



Again, for the **ReductiveHourglassTest** we divide *A. thaliana* embryo development into three modules:

- early = stages 1 - 2 (Zygote and Quadrant)
- mid = stages 3 - 5 (Globular, Heart, and Torpedo)
- late = stages 6 - 7 (Bent and Mature)

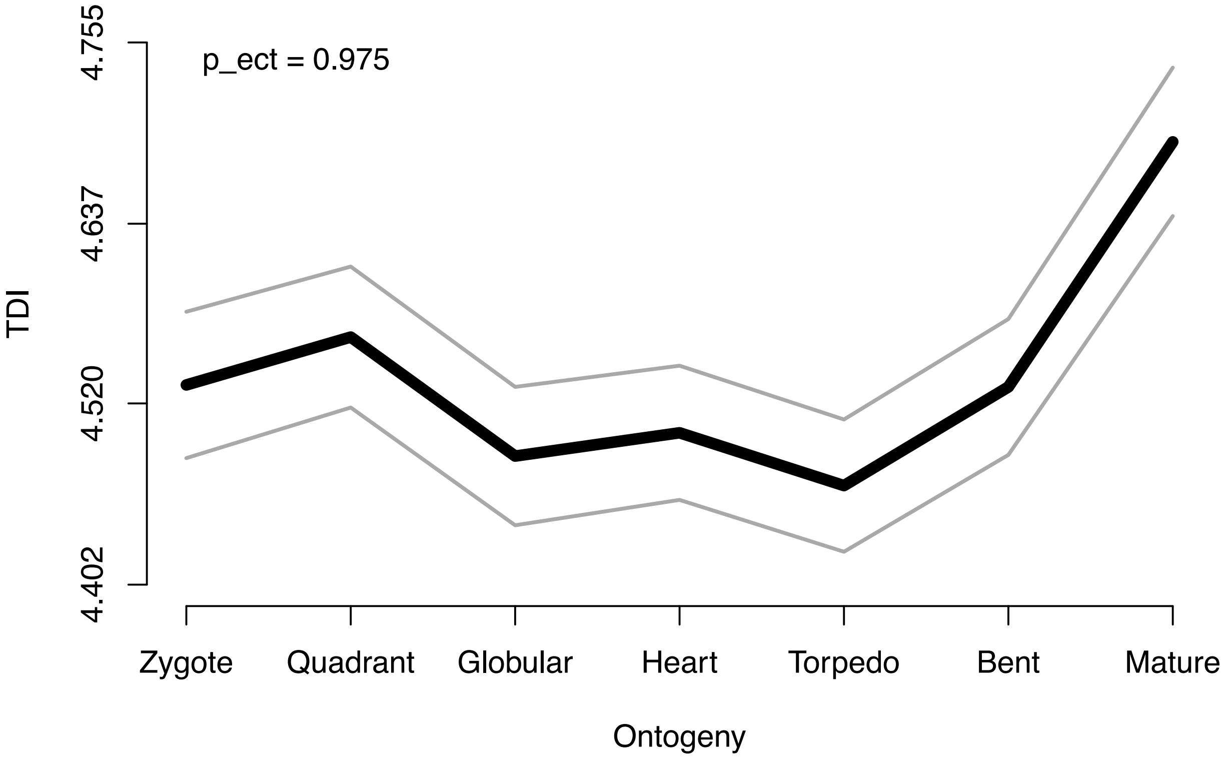
```
# Plot the Transcriptome Divergence Index of a given DivergenceExpressionSet
# Test Statistic : Reductive Hourglass Test
PlotPattern( ExpressionSet = DivergenceExpressionSetExample,
             TestStatistic = "ReductiveHourglassTest",
             modules       = list(early = 1:2, mid = 3:5, late = 6:7),
             type          = "l",
             lwd           = 6,
             xlab          = "Ontogeny",
             ylab          = "TDI" )
```



And for the **EarlyConservationTest** we again divide *A. thaliana* embryo development into three modules:

- early = stages 1 - 2 (Zygote and Quadrant)
- mid = stages 3 - 5 (Globular, Heart, and Torpedo)
- late = stages 6 - 7 (Bent and Mature)

```r
# Plot the Transcriptome Divergence Index of a given DivergenceExpressionSet
# Test Statistic : Reductive Early Conservation Test
PlotPattern( ExpressionSet = DivergenceExpressionSetExample,
             TestStatistic = "EarlyConservationTest",
             modules       = list(early = 1:2, mid = 3:5, late = 6:7),
             type          = "l",
             lwd           = 6,
             xlab          = "Ontogeny",
             ylab          = "TDI" )
```



To obtain the numerical TDI values for a given DivergenceExpressionSet simply run:

```
# Compute the Transcriptome Divergence Index values of a given DivergenceExpressionSet
TDI(DivergenceExpressionSetExample)
```

```
  Zygote Quadrant Globular    Heart  Torpedo     Bent   Mature
4.532029 4.563200 4.485705 4.500868 4.466477 4.530704 4.690292
```

Mean Expression and Relative Expression of Single Phylostrata or Divergence Strata

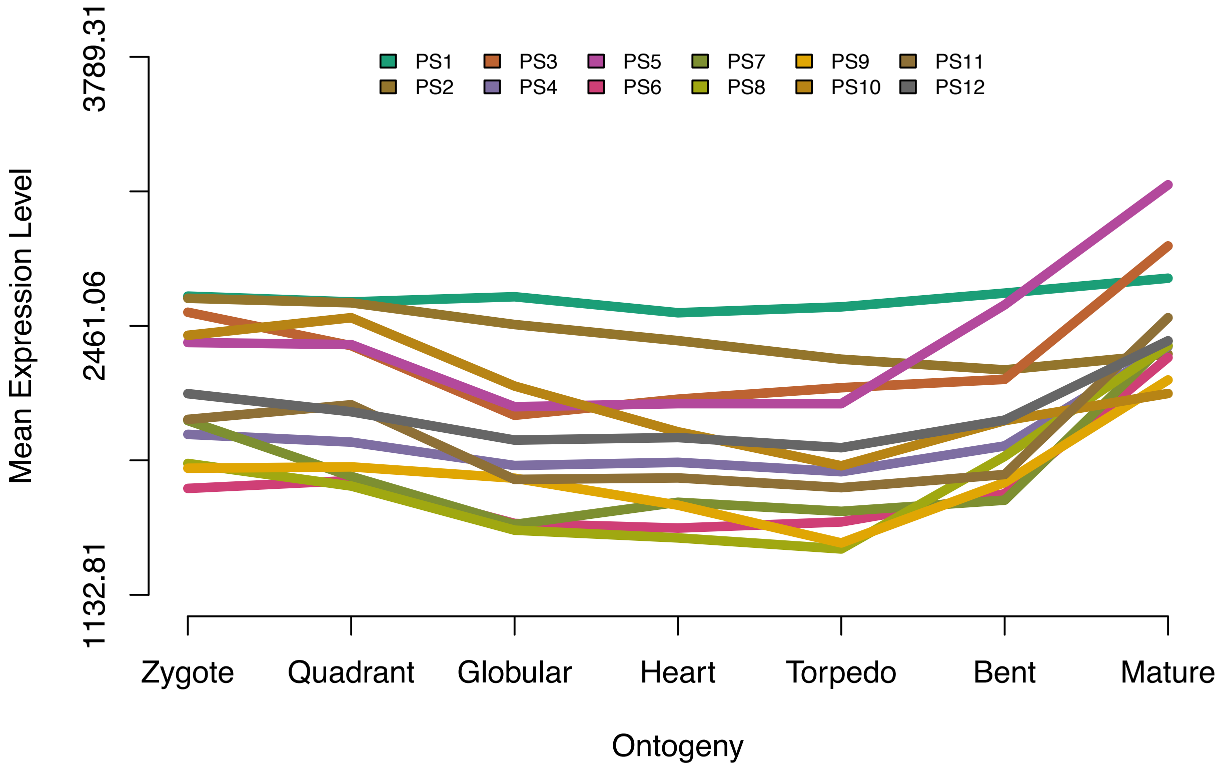
TAI or **TDI** patterns are very useful to gain a first insight into the mean transcriptome age or mean sequence divergence of genes being most active during the corresponding developmental stage or experiment.

To further investigate the origins of the global **TAI** or **TDI** pattern it is useful to visualize the mean gene expression of each Phylostratum or Divergence-Stratum class.

Mean Expression Levels of a PhyloExpressionSet and DivergenceExpressionSet

Visualizing the mean gene expression of genes corresponding to the same Phylostratum or Divergence Stratum class allows users to detect biological process specific groups of Phylostrata or Divergence Strata that are most expressed during the underlying biological process. This might lead to correlating specific groups of Phylostrata or Divergence Strata sharing similar evolutionary origins with common functions or functional contributions to a specific developmental process.

```
# Visualizing the mean gene expression of each Phylostratum class
PlotMeans( ExpressionSet = PhyloExpressionSetExample,
           Groups        = list(1:12),
           legendName    = "PS",
           xlab          = "Ontogeny",
           lty           = 1,
           cex           = 0.7,
           lwd           = 5 )
```



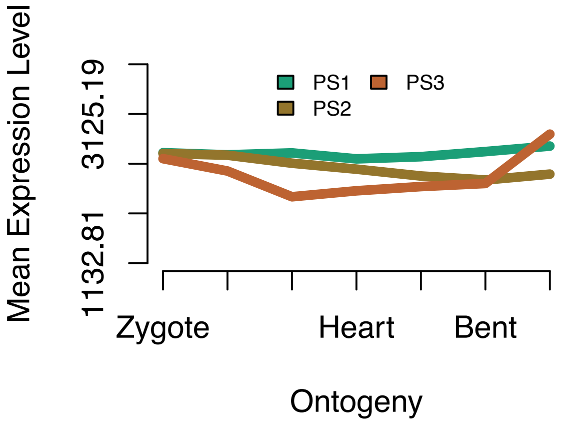
Here we see that the mean gene expression of Phylostratum group: PS1-3 (genes evolved before the establishment of embryogenesis in plants) are more expressed during *A. thaliana* embryogenesis than PS4-12 (genes evolved during or after the establishment of embryogenesis in plants).

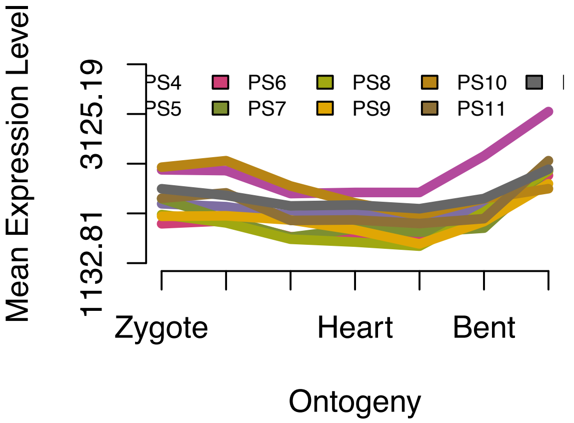
In different biological processes different Phylostratum groups or combination of groups might resemble the majority of expressed genes.

The `PlotMeans()` function takes an PhyloExpressionSet or DivergenceExpressionSet and visualizes for each Phylostratum the mean expression levels of all genes that correspond to this Phylostratum. The `Groups` argument takes a list storing the Phylostrata (classified into the same group) that shall be visualized on the same plot.

For this example we separate groups of Phylostrata into **evolutionary old Phylostrata** (PS1-3) in one plot versus **evolutionary younger Phylostrata** (PS4-12) into another plot:

```
# Visualizing the mean gene expression of each Phylostratum class
# in two separate plots (groups)
PlotMeans( ExpressionSet = PhyloExpressionSetExample,
           Groups        = list(group_1 = 1:3, group_2 = 4:12),
           legendName    = "PS",
           xlab          = "Ontogeny",
           lty           = 1,
           cex           = 0.7,
           lwd           = 5 )
```





To obtain the numerical values (mean expression levels for all Phylostrata) run:

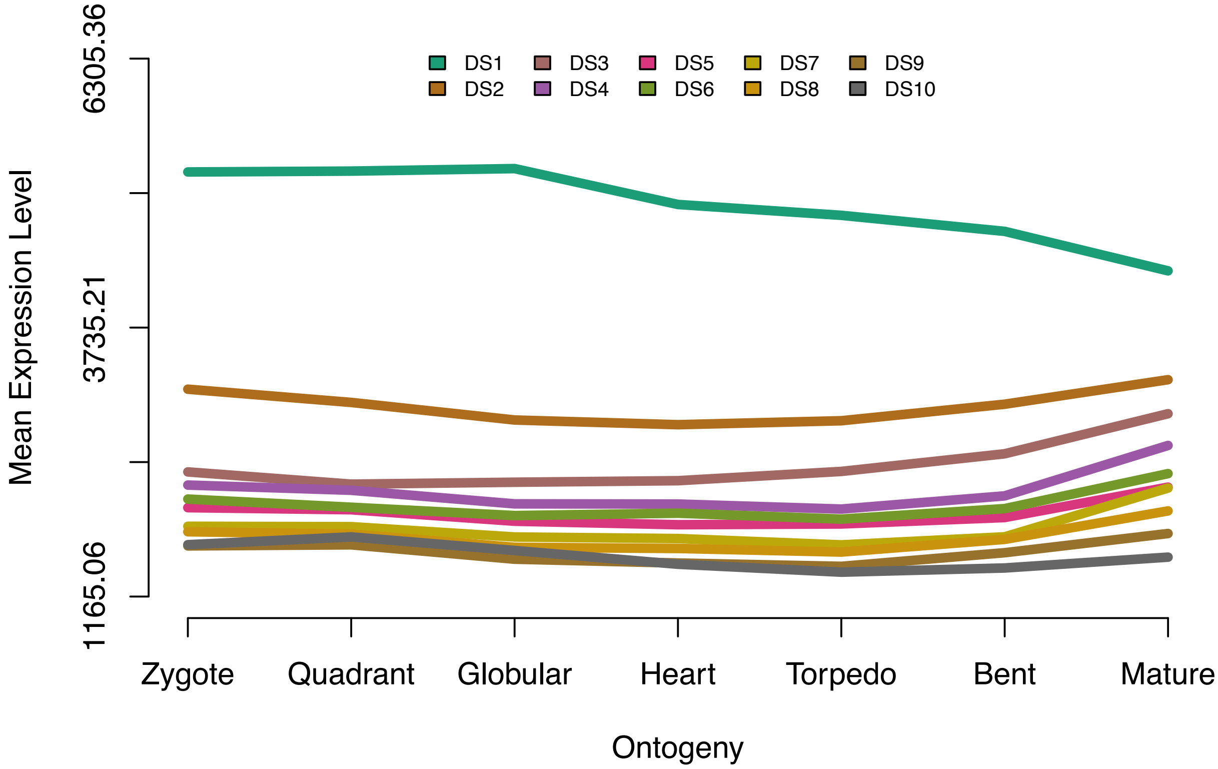
```
# Using the age.apply() function to compute the mean expression levels
# of all Phylostrata
age.apply( ExpressionSet = PhyloExpressionSetExample,
           FUN           = colMeans )
```

```
     Zygote Quadrant Globular    Heart  Torpedo     Bent   Mature
1  2607.882 2579.372 2604.856 2525.704 2554.825 2622.757 2696.331
2  2597.258 2574.745 2467.679 2388.045 2296.410 2243.716 2321.709
3  2528.272 2363.159 2019.436 2099.079 2155.642 2196.875 2855.866
4  1925.320 1887.078 1771.399 1787.175 1740.823 1867.981 2358.893
5  2378.883 2368.593 2061.729 2077.087 2076.693 2564.904 3157.761
6  1658.253 1697.242 1485.401 1462.613 1492.861 1631.741 2304.683
7  1993.321 1717.659 1480.525 1590.009 1545.078 1600.264 2385.409
8  1781.653 1670.106 1452.180 1414.052 1359.376 1816.718 2364.070
9  1758.119 1764.748 1708.815 1575.727 1388.920 1687.314 2193.930
10 2414.456 2501.390 2163.810 1938.060 1770.039 1993.032 2127.015
11 1999.163 2071.456 1702.779 1710.290 1662.099 1726.865 2501.443
12 2126.189 2036.804 1896.964 1909.578 1859.485 1995.732 2387.343
```

Here the `age.apply()` function (`?age.apply`) takes a function as argument that itself receives a `data.frame` as argument (e.g. `colMeans()`).

For a DivergenceExpressionSet run:

```
# Visualizing the mean gene expression of each Divergence-Stratum class
PlotMeans( ExpressionSet = DivergenceExpressionSetExample,
           Groups        = list(1:10),
           legendName    = "DS",
           xlab          = "Ontogeny",
           lty           = 1,
           cex           = 0.7,
           lwd           = 5 )
```



To obtain the numerical values (mean expression levels for all Divergence-Strata) run:

```
# Using the age.apply() function to compute the mean expression levels
# of all Divergence-Strata
age.apply( ExpressionSet = DivergenceExpressionSetExample,
           FUN           = colMeans )
```

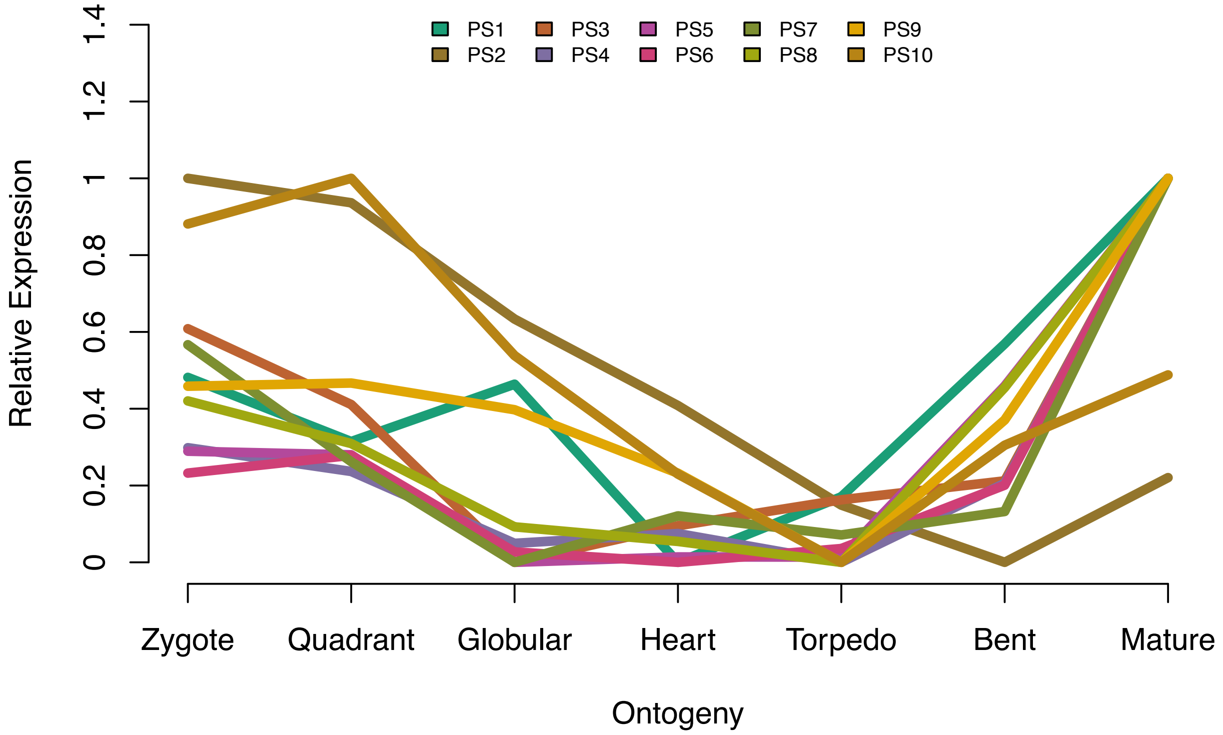
```
    Zygote Quadrant Globular    Heart  Torpedo     Bent   Mature
1  5222.189 5230.547 5254.464 4911.494 4807.936 4654.683 4277.490
2  3146.510 3020.156 2852.072 2807.367 2845.025 3002.967 3237.315
3  2356.008 2239.344 2257.539 2272.270 2360.816 2529.276 2912.164
4  2230.350 2180.706 2050.895 2049.035 2001.043 2127.165 2608.903
5  2014.600 1994.640 1884.899 1851.554 1858.913 1920.185 2210.391
6  2096.593 2018.440 1938.765 1961.828 1905.246 2005.523 2339.767
7  1836.290 1832.815 1734.319 1719.186 1659.044 1736.141 2201.981
8  1784.470 1762.151 1635.529 1624.682 1590.489 1711.439 1983.607
9  1649.254 1659.455 1522.214 1485.560 1453.689 1584.176 1767.276
10 1660.750 1735.086 1605.275 1473.854 1398.067 1438.258 1541.633
```

Relative Expression Levels of a PhyloExpressionSet and DivergenceExpressionSet

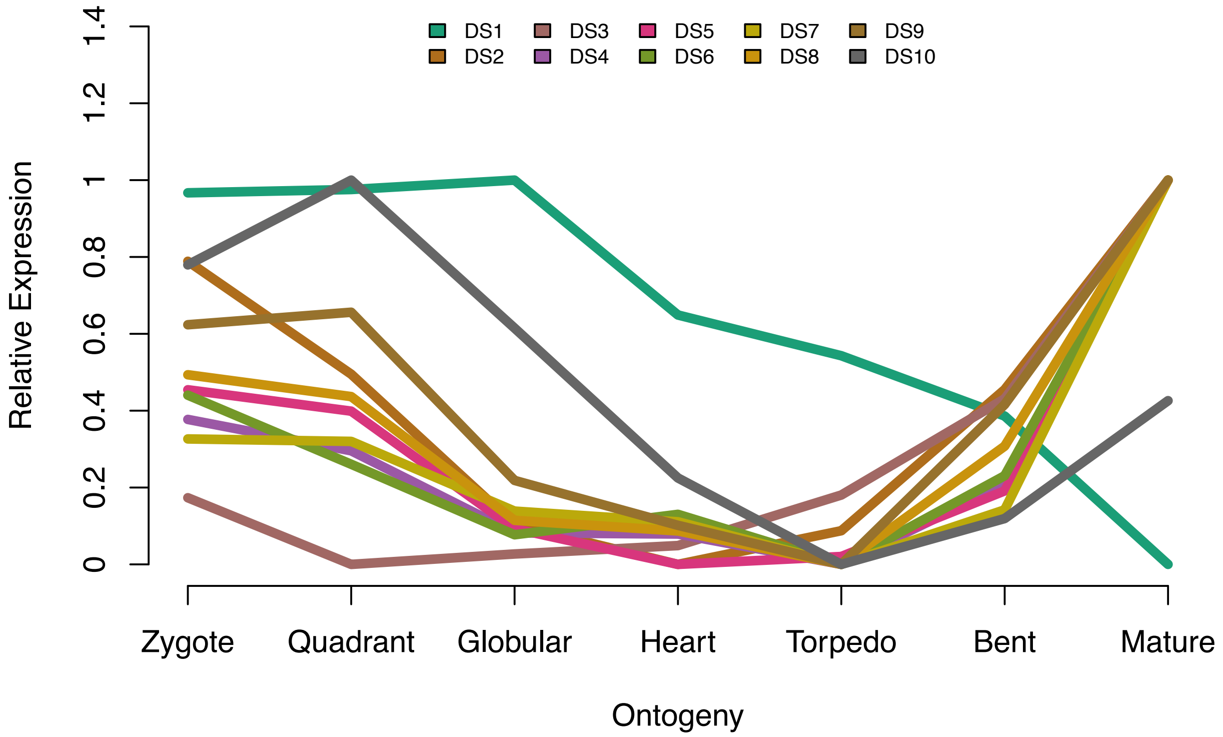
Introduced by Domazet-Loso and Tautz, (2010), relative expression levels are defined as a linear transformation of the mean expression levels (of each Phylostratum or Divergence-Stratum) into the interval $[0, 1]$ (Quint et al., 2012 and Drost et al., 2015). This procedure allows users to compare mean expression patterns between Phylostrata or Divergence Strata independent from their actual magnitude. Hence, relative expression profiles aim to correlate the mean expression profiles of groups of Phylostrata or Divergence Strata due to the assumption that genes or groups of genes sharing a similar expression profile might be regulated by similar gene regulatory mechanisms or contribute to similar biological processes.

The `PlotRE()` function can be used (analogous to the `PlotMeans()` function) to visualize the relative expression levels of a given PhyloExpressionSet and DivergenceExpressionSet:

```
# Visualizing the mean gene expression of each Phylostratum class
PlotRE( ExpressionSet = PhyloExpressionSetExample,
        Groups        = list(1:10),
        legendName    = "PS",
        xlab          = "Ontogeny",
        lty           = 1,
        cex           = 0.7,
        lwd           = 5 )
```

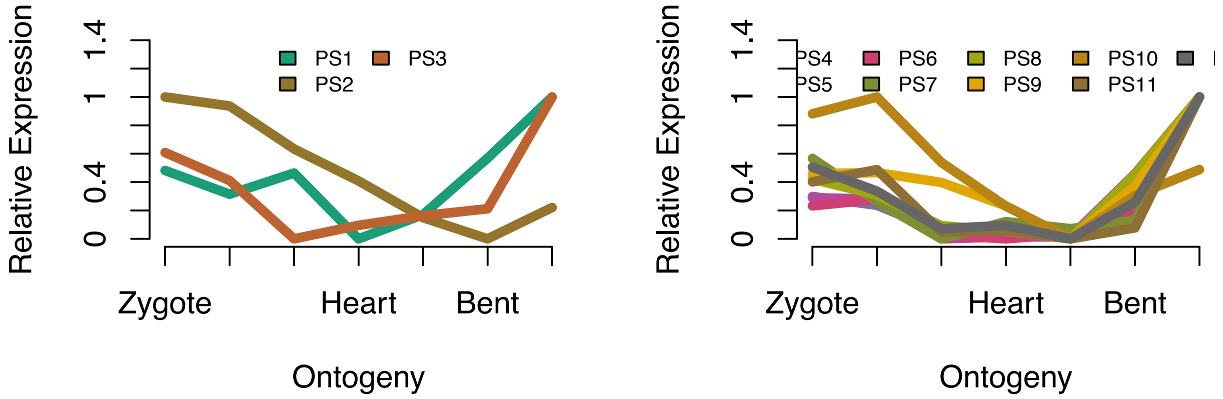


```
# Visualizing the mean gene expression of each Divergence-Stratum class
PlotRE( ExpressionSet = DivergenceExpressionSetExample,
        Groups        = list(1:10),
        legendName    = "DS",
        xlab          = "Ontogeny",
        lty           = 1,
        cex           = 0.7,
        lwd           = 5 )
```



or again by assigning Phylostratum or Divergence-Stratum groups that shall be visualized in different plots:

```r
# Visualizing the mean gene expression of each Phylostratum class
PlotRE( ExpressionSet = PhyloExpressionSetExample,
        Groups        = list(group_1 = 1:3, group_2 = 4:12),
        legendName    = "PS",
        xlab          = "Ontogeny",
        lty           = 1,
        cex           = 0.7,
        lwd           = 5 )
```



The relative expression levels can be obtained using the `REMatrix()` function:

```r
# Getting the relative expression levels for all Phylostrata
REMatrix(PhyloExpressionSetExample)
```

```
      Zygote  Quadrant   Globular      Heart    Torpedo       Bent    Mature
1  0.4816246 0.3145330 0.46389184 0.00000000 0.17067495 0.56880234 1.0000000
2  1.0000000 0.9363209 0.63348381 0.40823711 0.14904726 0.00000000 0.2206063
3  0.6083424 0.4109402 0.00000000 0.09521758 0.16284114 0.21213845 1.0000000
4  0.2985050 0.2366309 0.04946941 0.07499453 0.00000000 0.20573325 1.0000000
5  0.2893657 0.2799777 0.00000000 0.01401191 0.01365328 0.45908792 1.0000000
6  0.2323316 0.2786335 0.02706119 0.00000000 0.03592044 0.20084761 1.0000000
7  0.5666979 0.2620602 0.00000000 0.12099252 0.07133814 0.13232551 1.0000000
8  0.4203039 0.3092784 0.09237036 0.05442042 0.00000000 0.45520558 1.0000000
9  0.4586261 0.4668613 0.39738003 0.23205534 0.00000000 0.37067096 1.0000000
10 0.8811321 1.0000000 0.53841500 0.22974016 0.00000000 0.30490542 0.4881046
11 0.4015809 0.4877111 0.04846721 0.05741594 0.00000000 0.07716367 1.0000000
12 0.5052572 0.3359211 0.07100055 0.09489782 0.00000000 0.25811214 1.0000000
```

```r
# Getting the relative expression levels for all Divergence-Strata
REMatrix(DivergenceExpressionSetExample)
```

```
      Zygote  Quadrant   Globular      Heart    Torpedo      Bent    Mature
1  0.9669643 0.9755188 1.00000000 0.64894653 0.54294759 0.3860827 0.0000000
2  0.7888009 0.4949178 0.10397567 0.00000000 0.08758660 0.4549387 1.0000000
3  0.1733953 0.0000000 0.02704324 0.04893726 0.18054185 0.4309208 1.0000000
4  0.3772372 0.2955661 0.08201140 0.07895260 0.00000000 0.2074848 1.0000000
5  0.4543752 0.3987496 0.09292474 0.00000000 0.02050713 0.1912595 1.0000000
6  0.4403615 0.2605017 0.07713944 0.13021586 0.00000000 0.2307754 1.0000000
7  0.3264585 0.3200581 0.13864386 0.11077270 0.00000000 0.1420009 1.0000000
8  0.4934416 0.4366671 0.11457069 0.08697865 0.00000000 0.3076689 1.0000000
9  0.6236387 0.6561674 0.21851855 0.10163374 0.00000000 0.4161087 1.0000000
10 0.7794318 1.0000000 0.61482564 0.22487531 0.00000000 0.1192539 0.4259882
```

The same result could also be obtained by using the `age.apply()` function in combination with the `RE()` function:

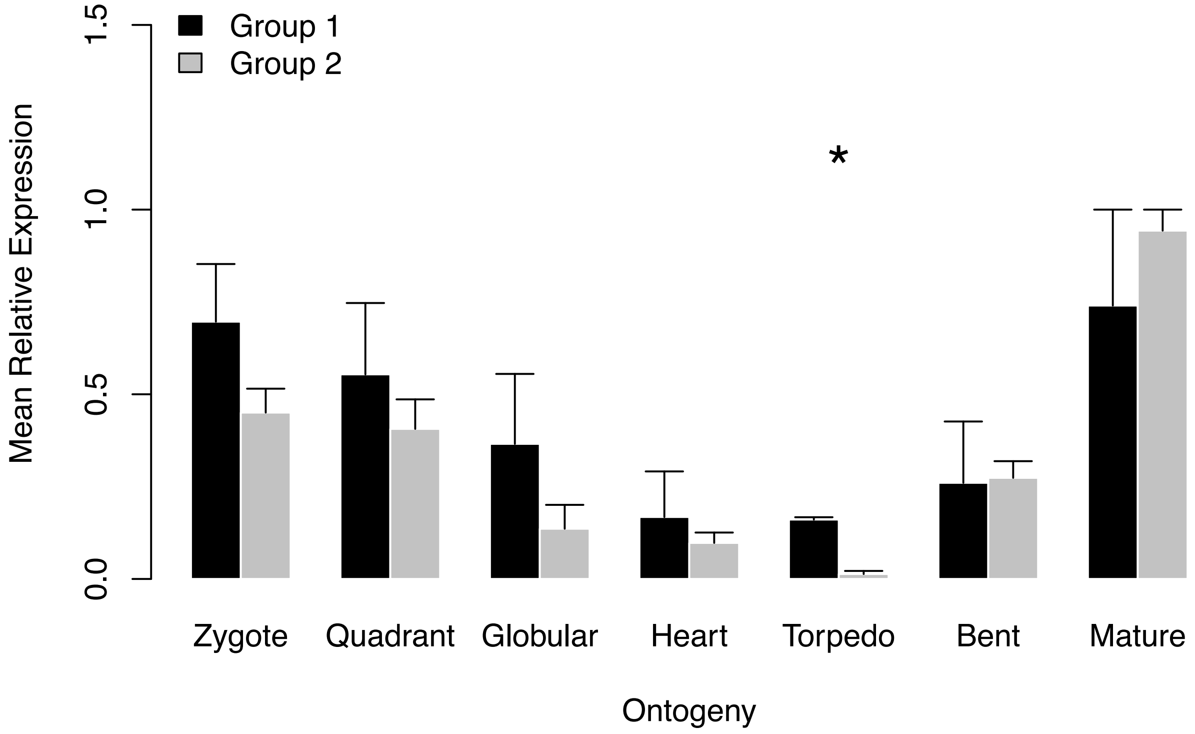
```r
# Getting the relative expression levels for all Phylostrata
age.apply( ExpressionSet = PhyloExpressionSetExample,
           FUN           = RE )
```

```
      Zygote  Quadrant   Globular      Heart    Torpedo       Bent    Mature
1  0.4816246 0.3145330 0.46389184 0.00000000 0.17067495 0.56880234 1.0000000
2  1.0000000 0.9363209 0.63348381 0.40823711 0.14904726 0.00000000 0.2206063
3  0.6083424 0.4109402 0.00000000 0.09521758 0.16284114 0.21213845 1.0000000
4  0.2985050 0.2366309 0.04946941 0.07499453 0.00000000 0.20573325 1.0000000
5  0.2893657 0.2799777 0.00000000 0.01401191 0.01365328 0.45908792 1.0000000
6  0.2323316 0.2786335 0.02706119 0.00000000 0.03592044 0.20084761 1.0000000
7  0.5666979 0.2620602 0.00000000 0.12099252 0.07133814 0.13232551 1.0000000
8  0.4203039 0.3092784 0.09237036 0.05442042 0.00000000 0.45520558 1.0000000
9  0.4586261 0.4668613 0.39738003 0.23205534 0.00000000 0.37067096 1.0000000
10 0.8811321 1.0000000 0.53841500 0.22974016 0.00000000 0.30490542 0.4881046
11 0.4015809 0.4877111 0.04846721 0.05741594 0.00000000 0.07716367 1.0000000
12 0.5052572 0.3359211 0.07100055 0.09489782 0.00000000 0.25811214 1.0000000
```

Quint et al. (2012) introduced an additional way of visualizing the difference of relative expression levels between groups of Phylostrata/Divergence-Strata.

This bar plot comparing the mean relative expression levels of one Phylostratum/Divergence-Stratum group with all other groups can be plotted analogous to the `PlotMeans()` and `PlotRE()` functions:

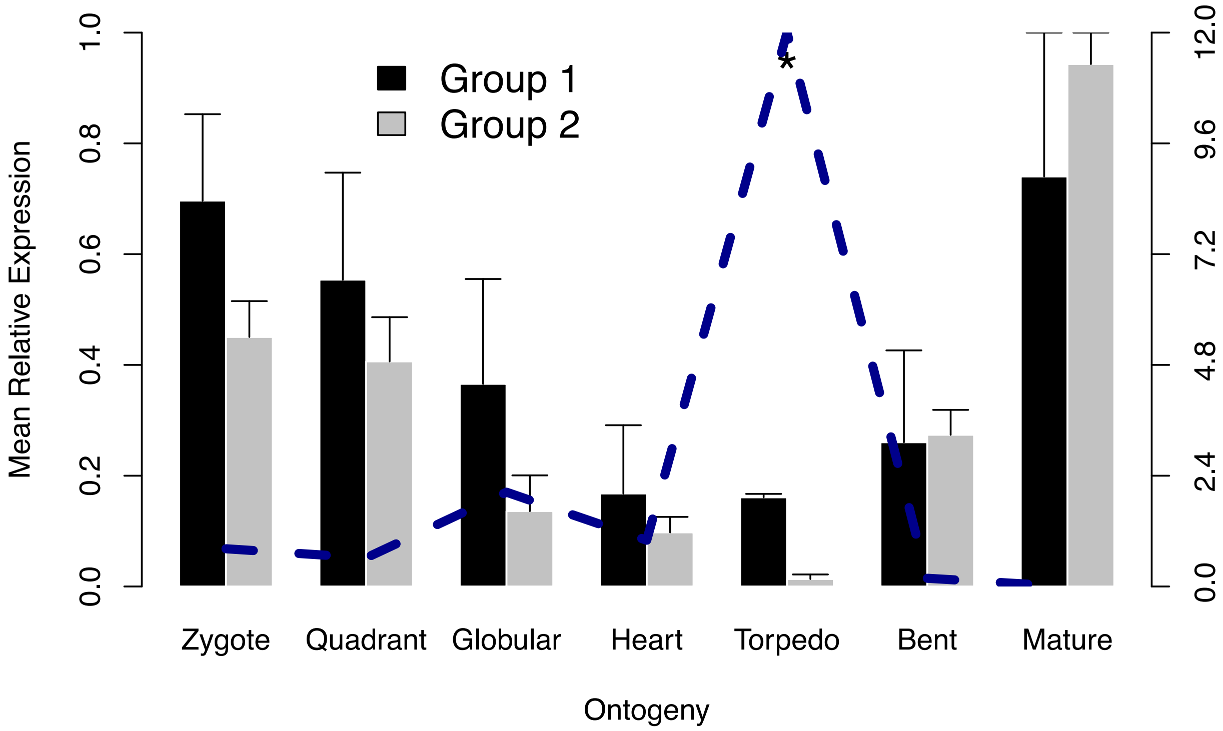
```
# Visualizing the mean relative expression of two Phylostratum groups
PlotBarRE( ExpressionSet = PhyloExpressionSetExample,
           Groups        = list(group_1 = 1:3, group_2 = 4:12),
           xlab          = "Ontogeny",
           ylab          = "Mean Relative Expression",
           cex           = 2)
```



Here the argument `Groups = list(1:3, 4:12)` corresponds to dividing Phylostrata 1-12 into Phylostratum groups defined as *origin before embryogenesis* (group one: PS1-3) and *origin during or after embryogenesis* (group two: PS4-12). A Kruskal-Wallis Rank Sum Test is then performed to test the statistical significance of the different bars that are compared. The '*' corresponds to a statistically significant difference.

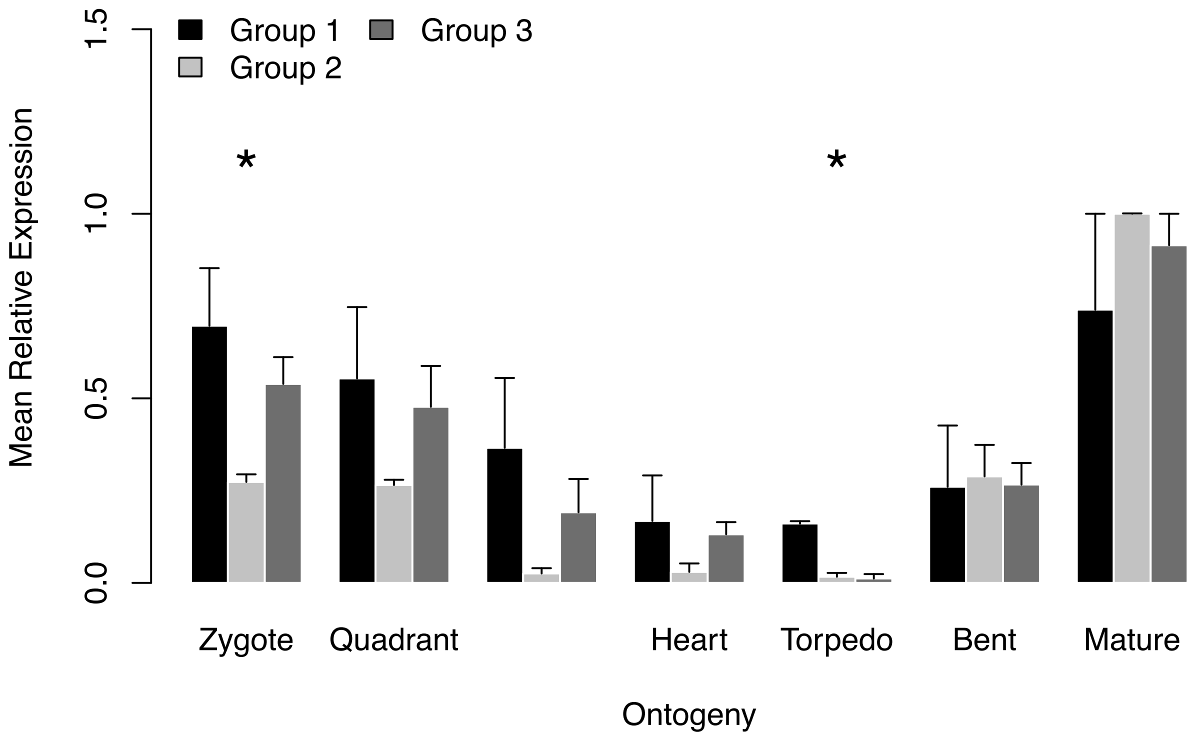
Additionally the ratio between both values represented by the bars to be compared can be visualized as function within the bar plot using the `ratio = TRUE` argument:

```
# Visualizing the mean relative expression of two Phylostratum groups
PlotBarRE( ExpressionSet = PhyloExpressionSetExample,
           Groups        = list(group_1 = 1:3, group_2 = 4:12),
           ratio         = TRUE,
           xlab          = "Ontogeny",
           ylab          = "Mean Relative Expression",
           cex           = 2 )
```



It is also possible to compare more than two groups:

```
# Visualizing the mean relative expression of three Phylostratum groups
PlotBarRE( ExpressionSet = PhyloExpressionSetExample,
           Groups        = list(group_1 = 1:3, group_2 = 4:6, group_3 = 7:12),
           wLength       = 0.05,
           xlab          = "Ontogeny",
           ylab          = "Mean Relative Expression",
           cex           = 2 )
```



For the corresponding statistically significant stages, a *Posthoc* test can be performed to detect the combinations of differing bars that cause the global statistical significance.

Investigating the Statistical Significance of Phylotranscriptomic Patterns

Three methods have been proposed to quantify the statistical significance of the observed phylotranscriptomics patterns (Quint et al., 2012; Drost et al., 2015).

- **Flat Line Test**
- **Reductive Hourglass Test**
- **Reductive Early Conservation Test**

Here, we will build the test statistic of each test step by step so that future modifications or new test statistics can be built upon the existing methods implemented in the `myTAI` package.

Flat Line Test

The **Flat Line Test** is a permutation test quantifying the statistical significance of an observed phylotranscriptomic pattern. The goal is to detect any evolutionary signal within a developmental time course that significantly deviates from a flat line.

To build the test statistic we start with a standard PhyloExpressionSet. The `myTAI` package provides an example PhyloExpressionSet named `PhyloExpressionSetExample`:

```
library(myTAI)

# load an example PhyloExpressionSet stored in the myTAI package
data(PhyloExpressionSetExample)
```

```r
# look at the standardized data set format
head(PhyloExpressionSetExample, 3)
```

```
  Phylostratum      GeneID   Zygote Quadrant Globular     Heart  Torpedo
1            1 at1g01040.2 2173.635 1911.200 1152.555 1291.4224 1000.253
2            1 at1g01050.1 1501.014 1817.309 1665.309 1564.7612 1496.321
3            1 at1g01070.1 1212.793 1233.002  939.200  929.6195  864.218
       Bent    Mature
1  962.9772 1696.4274
2 1114.6435 1071.6555
3  877.2060  894.8189
```

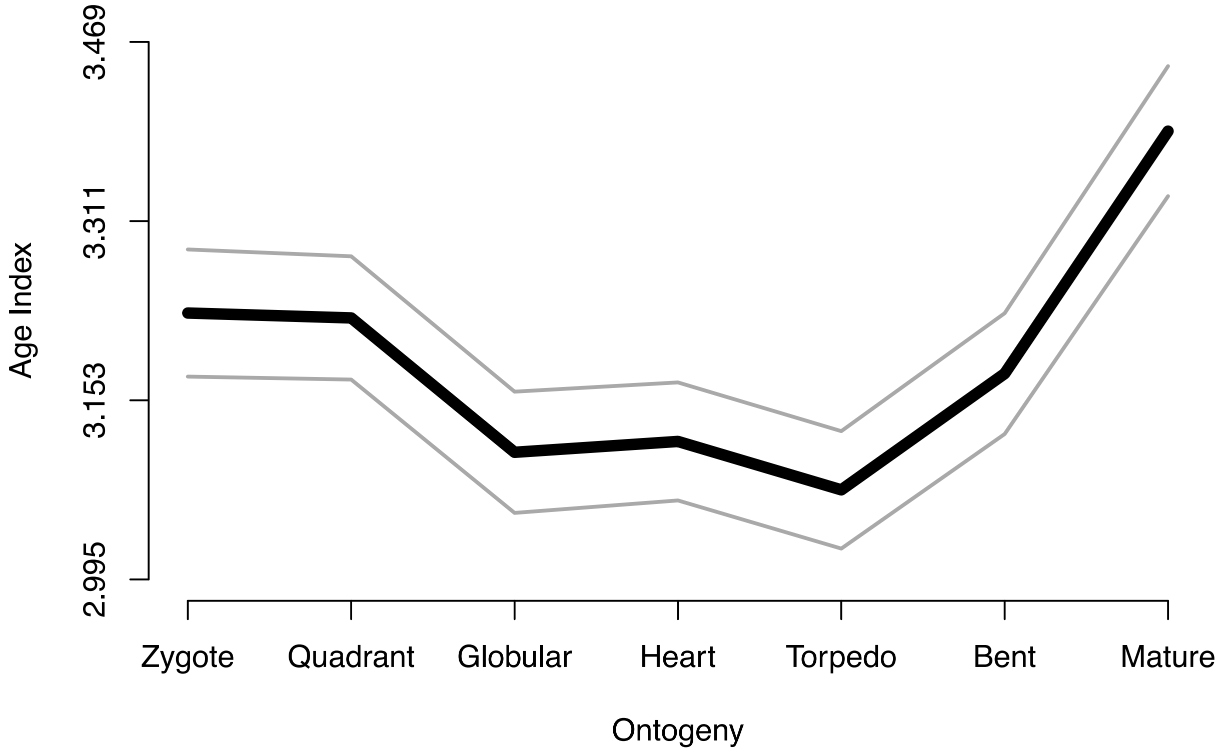
Users will observe that the first column of the `PhyloExpressionSetExample` stores the Phylostratum assignments of the corresponding genes. The permutation test is based on random sampling of the Phylostratum assignment of genes. The underlying assumption is that the `TAI` profile of correctly assigned Phylostrata is significantly deviating from `TAI` profiles based on randomly assigned Phylostrata.

```r
# TAI profile of correctly assigned Phylostrata
TAI(PhyloExpressionSetExample)
```

```
  Zygote Quadrant Globular    Heart  Torpedo     Bent   Mature
3.229942 3.225614 3.107135 3.116693 3.073993 3.176511 3.390334
```

Visualization:

```r
# Visualize the TAI profile of correctly assigned Phylostrata
PlotPattern( ExpressionSet = PhyloExpressionSetExample,
             type          = "l",
             lwd           = 6,
             p.value       = FALSE )
```

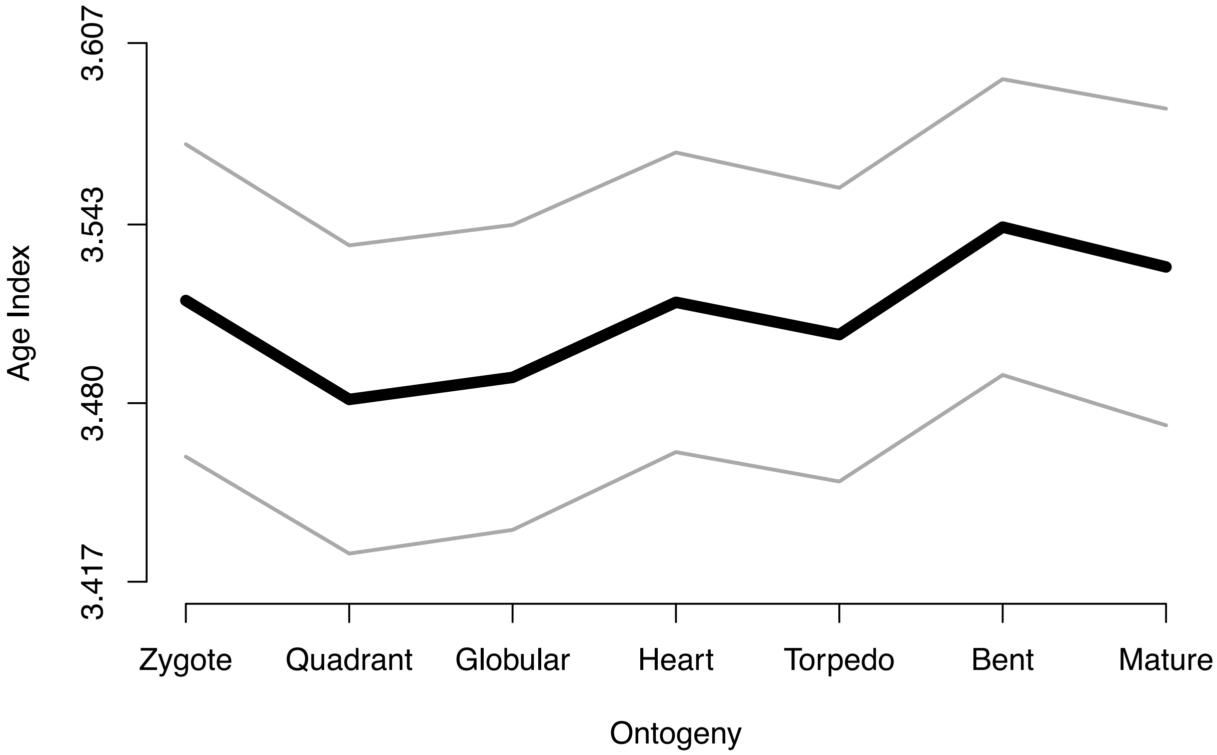


```
#>   Zygote Quadrant Globular    Heart  Torpedo     Bent   Mature
#> 3.516207 3.481292 3.489073 3.515569 3.504153 3.542093 3.528021
```

```
  Zygote Quadrant Globular    Heart  Torpedo     Bent   Mature
3.543700 3.573999 3.554707 3.543376 3.582905 3.562813 3.523409
```

Visualization:

```
# Visualize the TAI profile based on randomly assigned Phylostrata
PlotPattern( ExpressionSet =  randomPhyloExpressionSetExample,
             type          = "l",
             lwd           = 6,
             p.value       = FALSE )
```



Users will observe that the visual pattern of the correctly assigned `TAI` profile and the randomly assigned `TAI` profile differ qualitatively.

Now we investigate the variance of the two observed patterns.

```
# Variance of the TAI profile based on correctly assigned Phylostrata
var(TAI(PhyloExpressionSetExample))
```

```
[1] 0.01147725
```

```
# Variance of the TAI profile based on randomly assigned Phylostrata
var(TAI(randomPhyloExpressionSetExample))
```

```
[1] 0.0004102549
```

We observe that the variance of the randomly assigned TAI profile is much smaller than the variance of the correctly assigned TAI profile. Here we use the variance to quantify the *flatness* of a given TAI profile. In theory the variance of a perfect flat line would be zero. So any TAI profile that is close to zero would resemble a flat line. But how exactly are the variances of randomly assigned TAI profiles distributed? For this purpose the `bootMatrix()` function was implemented.

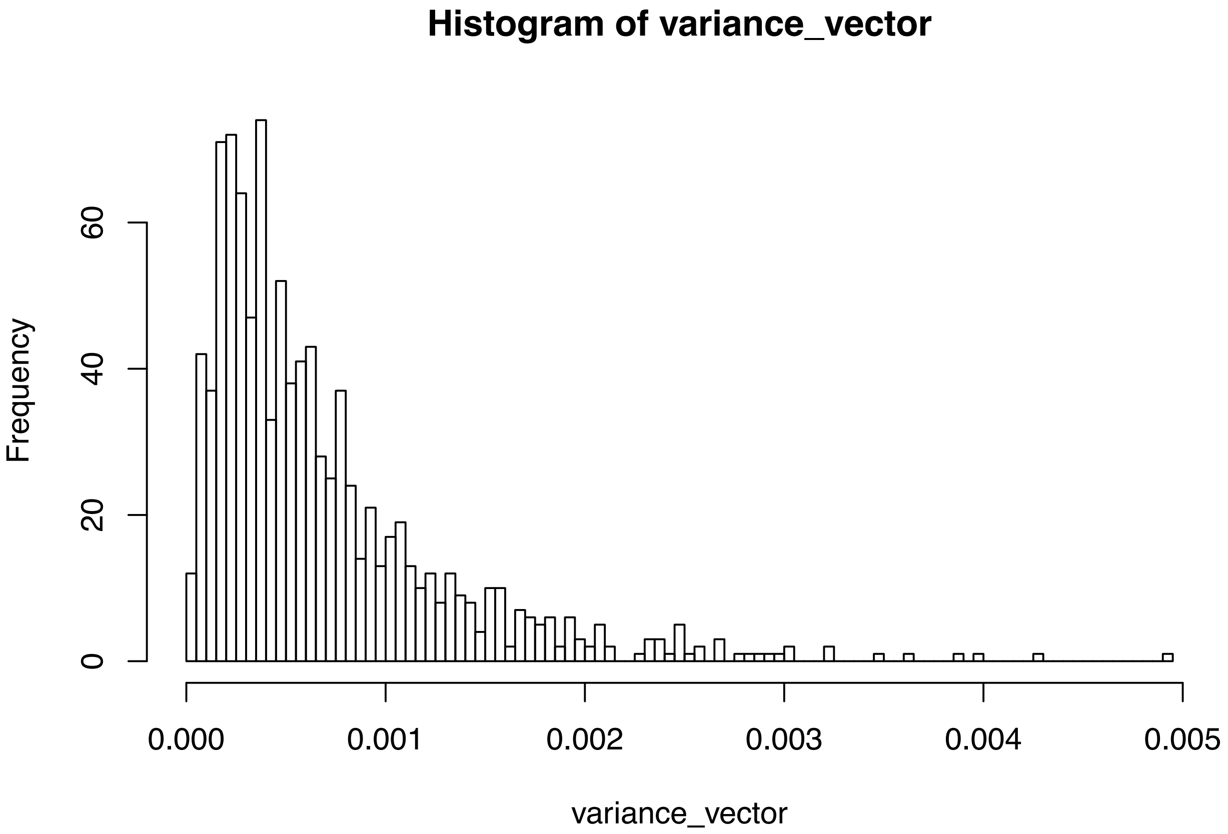
The `bootMatrix()` takes an PhyloExpressionSet or DivergenceExpressionSet as input and computes N `TAI` or `TDI` profiles based on randomly assigned Phylostrata or Divergence-Strata.

```
#>     Zygote Quadrant Globular    Heart  Torpedo     Bent   Mature
#> 1 3.488903 3.494434 3.439236 3.446201 3.449732 3.424719 3.410703
#> 2 3.553589 3.538320 3.537404 3.559358 3.569401 3.549278 3.494447
#> 3 3.472941 3.477267 3.483564 3.507338 3.510584 3.548868 3.552674
#> 4 3.489978 3.493074 3.505159 3.501532 3.528903 3.514788 3.552641
#> 5 3.451095 3.482781 3.482203 3.465931 3.486338 3.469059 3.416965
#> 6 3.510306 3.511145 3.529838 3.531301 3.549431 3.556543 3.508643
```

```
    Zygote Quadrant Globular    Heart  Torpedo     Bent   Mature
1 3.578000 3.577732 3.552230 3.556756 3.555627 3.553580 3.510978
2 3.590339 3.564544 3.589012 3.598754 3.583721 3.597553 3.633530
3 3.615284 3.616286 3.569486 3.566816 3.575761 3.554789 3.601522
4 3.501172 3.506554 3.486261 3.491282 3.531536 3.541634 3.494997
5 3.450857 3.458339 3.447274 3.463692 3.444261 3.468433 3.497926
6 3.539454 3.543827 3.595298 3.588149 3.582898 3.588062 3.565487
```

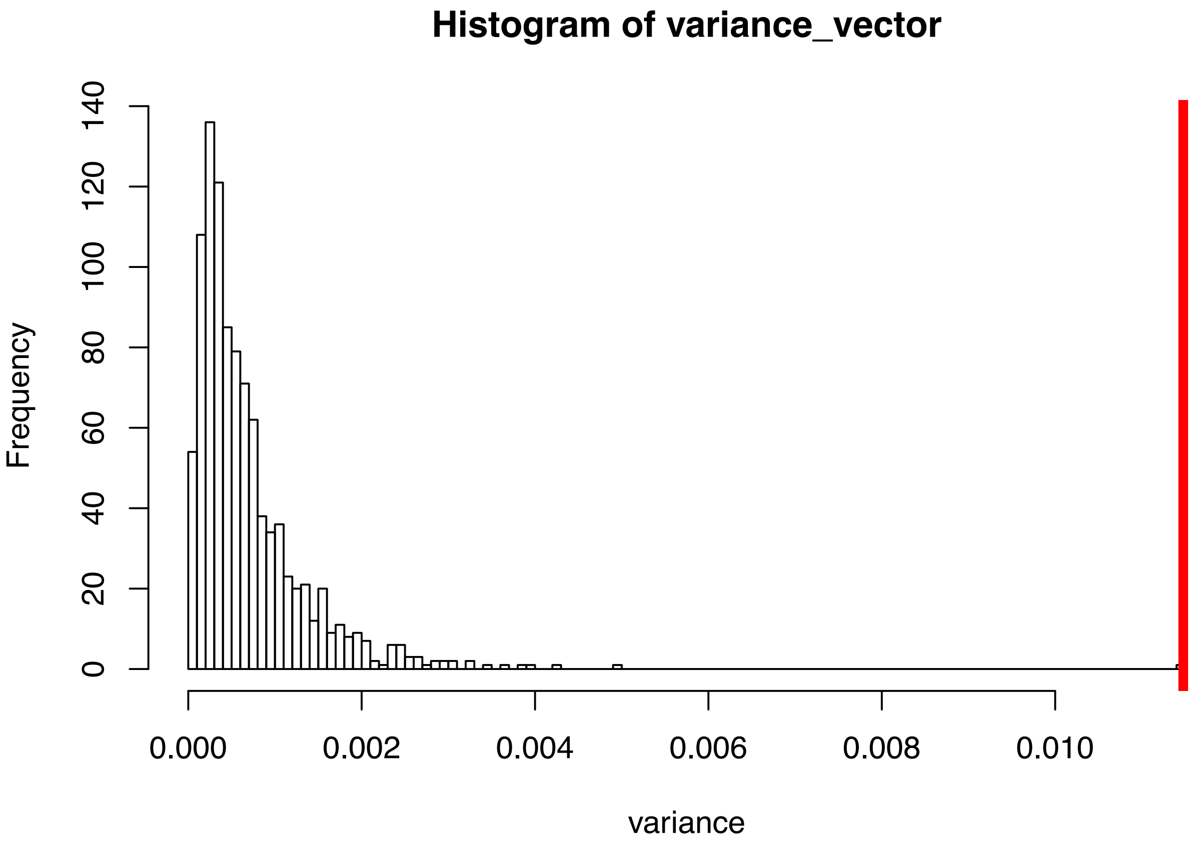
Based on this `booMatrix` we can compute the variance of each random TAI profile.

```
# compute the variance of the random TAI profile for each row
variance_vector <- apply(randomTAIs, 1 , var)

# and visualize the distribution of variances
hist(variance_vector, breaks = 100)
```



Now it is interesting to see where we can find the variance of the correctly assigned TAI.

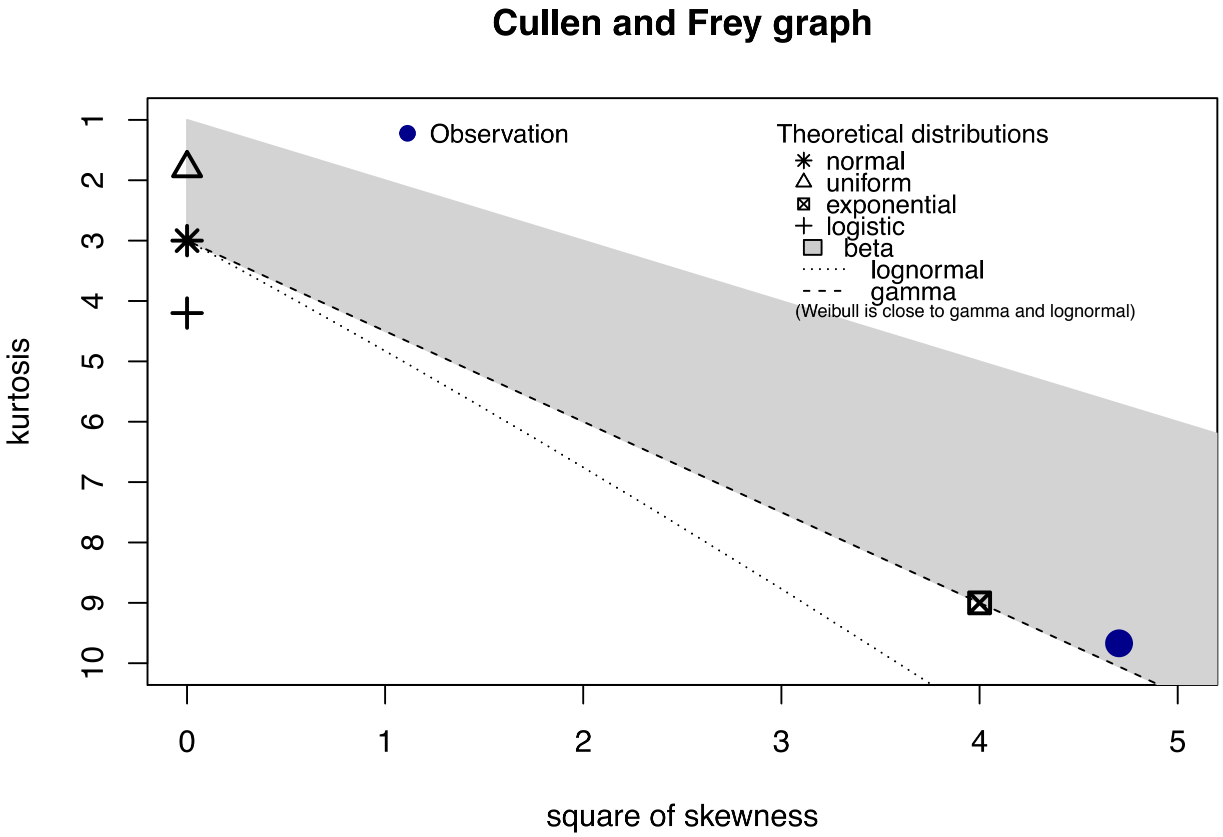
```
# variance of the TAI profile based on correctly assigned Phylostrata
var_real <- var(TAI(PhyloExpressionSetExample))

# visualize the distribution of variances
hist( x      = c(variance_vector,var_real),
       breaks = 100,
       xlab   = "variance",
       main   = "Histogram of variance_vector" )

# and plot a red line at the position where we can find the
# real variance
abline(v = var_real, lwd = 5, col = "red")
```



This plot illustrates that variances based on random `TAI` profiles seem to have a smaller variance than the variance based on the correct `TAI` profile. To obtain a p-value that now quantifies this difference, we need to fit the histogram of `variance_vector` with a specific probability distribution.

Visually it would be possible to choose a gamma distribution to fit the histogram of `variance_vector`. To validate this choice a Cullen and Frey graph provided by the **fitdistrplus** package can be used.

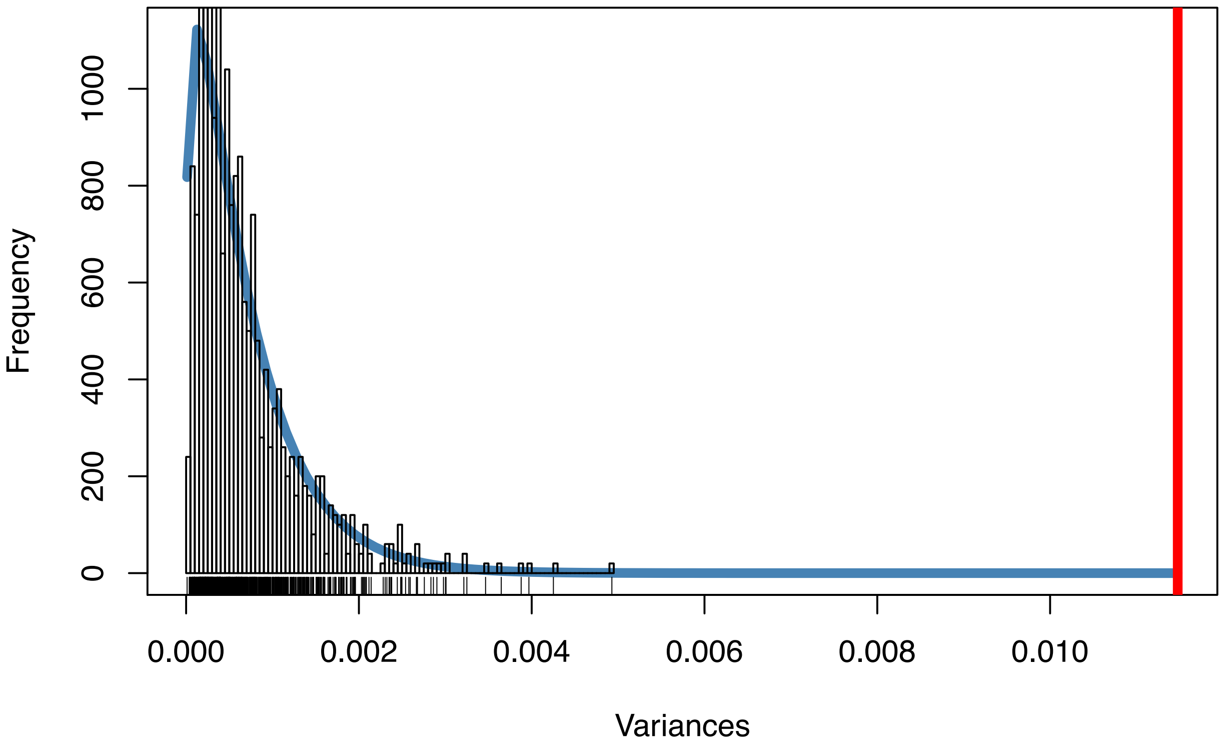
```
# install.packages("fitdistrplus")

# plot a Cullen and Frey graph
fitdistrplus::descdist(variance_vector)
```



```
#> summary statistics
#> ------
#> min:  1.142992e-05   max:  0.00492798
#> median:  0.0004985414
#> mean:  0.0006854571
#> estimated sd:  0.0006217058
#> estimated skewness:  2.168862
#> estimated kurtosis:  9.670302
```

Based on the observation that a gamma distribution is a suitable fit for `variance_vector`, we can now estimate the parameters of the gamma distribution that fits the data.

```
# estimate the parameters: shape and rate using 'moment matching estimation'
gamma_MME <- fitdistrplus::fitdist(variance_vector,distr = "gamma", method = "mme")
# estimate shape:
shape <- gamma_MME$estimate[1]
# estimate the rate:
rate <- gamma_MME$estimate[2]

# define an expression written as function as input for the curve() function
gamma_distr <- function(x){ return(dgamma(x = x, shape = shape, rate = rate)) }

# plot the density function and the histogram of variance_vector
curve( expr = gamma_distr,
       xlim = c(min(variance_vector),max(c(variance_vector,var_real))),
       col  = "steelblue",
       lwd  = 5,
       xlab = "Variances",
       ylab = "Frequency" )

# plot the histogram of variance_vector
histogram <- hist(variance_vector,prob = TRUE,add = TRUE, breaks = 100)
rug(variance_vector)

# plot a red line at the position where we can find the real variance
abline(v = var_real, lwd = 5, col = "red")
```



Using the gamma distribution with estimated parameters the corresponding p-value of `var_real` can be computed.

```
# p-value of var_real
pgamma(var_real, shape = shape,rate = rate, lower.tail = FALSE)
```

```
#> [1] 3.012488e-09
```

Hence, the variance of the correct TAI profile significantly deviates from random TAI profiles and this allows us to assume that the underlying TAI profile captures a real evolutionary signal.

Using the `FlatLineTest()` Function

This entire procedure of computing the p-value having the variance of TAI profiles as test statistic is done by the `FlatLineTest()` function.

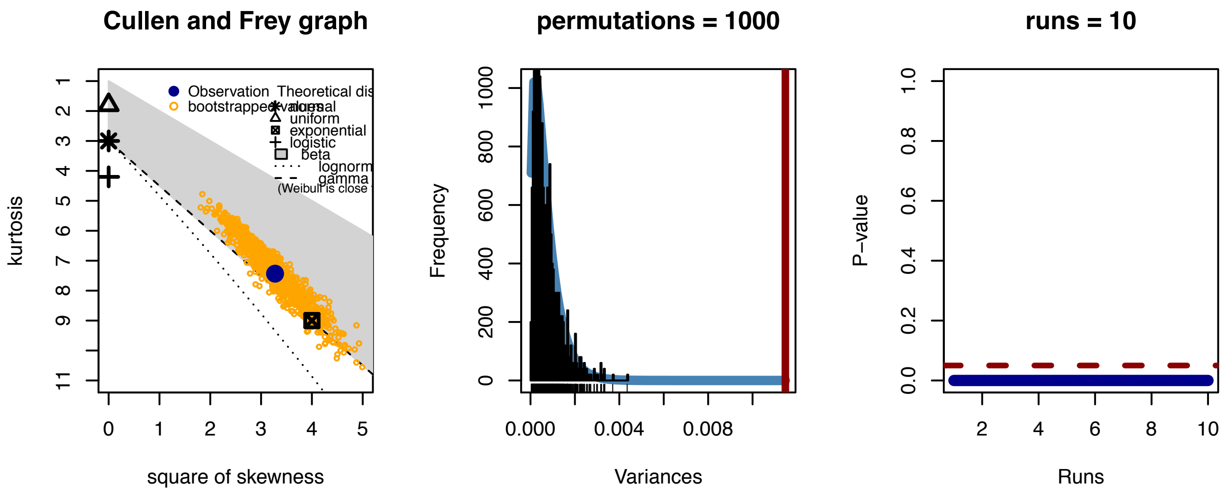
```
# Perform the FlatLineTest
FlatLineTest( ExpressionSet = PhyloExpressionSetExample,
              permutations  = 1000 )
```

```
#> $p.value
#> [1] 1.356011e-08
#>
#> $std.dev
#> [1] 0.05508008 0.05394724 0.05399575 0.05271312 0.05142109 0.05421588 0.05633689
```

This function returns the p-value of the test statistic.

Additionally the `FlatLineTest()` function allows users to investigate the goodness of the test statistic.

```
# perform the FlatLineTest and investigate the goodness of the test statistic
FlatLineTest( ExpressionSet = PhyloExpressionSetExample,
              permutations  = 1000,
              plotHistogram = TRUE )
```



```
#> $p.value
#> [1] 4.437918e-10
#>
#> $std.dev
#> [1] 0.05397621 0.05184950 0.04970483 0.04882796 0.04801541 0.04988094 0.05481457
```

The `plotHistogram` argument specifies whether analytics plots shall be drawn to quantify the goodness of the test statistic returned by the `FlatLineTest`.

The three resulting plots show:

- a Cullen and Frey graph
- a histogram of the test statistic and the corresponding gamma distribution that was fitted to the test statistic
- a plot showing the p-values (`p_flt`) for 10 individual runs. Since the underlying test statistic is generated by a permutation test, the third plot returned by `FlatLineTest()` shows the influence of different permutations to the corresponding p-value

In other words, to test whether the underlying permutation of the permutation test is causing the significance of the p-value, you can specify the `runs` argument within the `FlatLineTest()` function to perform several independent runs. In case there exists a permutation that causes a previous significant p-value to become non-significant, the corresponding phylotranscriptomic pattern shouldn't be considered as statistically significant.

Reductive Hourglass Test

The **Reductive Hourglass Test** has been developed to statistically evaluate the existence of a phylotranscriptomic hourglass pattern based on TAI or TDI computations. The corresponding p-value quantifies the probability that a given TAI or TDI pattern (or any phylotranscriptomics pattern) does not follow an hourglass like shape. A p-value < 0.05 indicates that the corresponding phylotranscriptomics pattern does indeed follow an hourglass (high-low-high) shape.

To build the test statistic again we start with a standard PhyloExpressionSet.

```r
library(myTAI)

# load an example PhyloExpressionSet stored in the myTAI package
data(PhyloExpressionSetExample)

# look at the standardized data set format
head(PhyloExpressionSetExample, 3)
```

```
  Phylostratum      GeneID   Zygote Quadrant Globular     Heart  Torpedo
1            1 at1g01040.2 2173.635 1911.200 1152.555 1291.4224 1000.253
2            1 at1g01050.1 1501.014 1817.309 1665.309 1564.7612 1496.321
3            1 at1g01070.1 1212.793 1233.002  939.200  929.6195  864.218
       Bent    Mature
1  962.9772 1696.4274
2 1114.6435 1071.6555
3  877.2060  894.8189
```

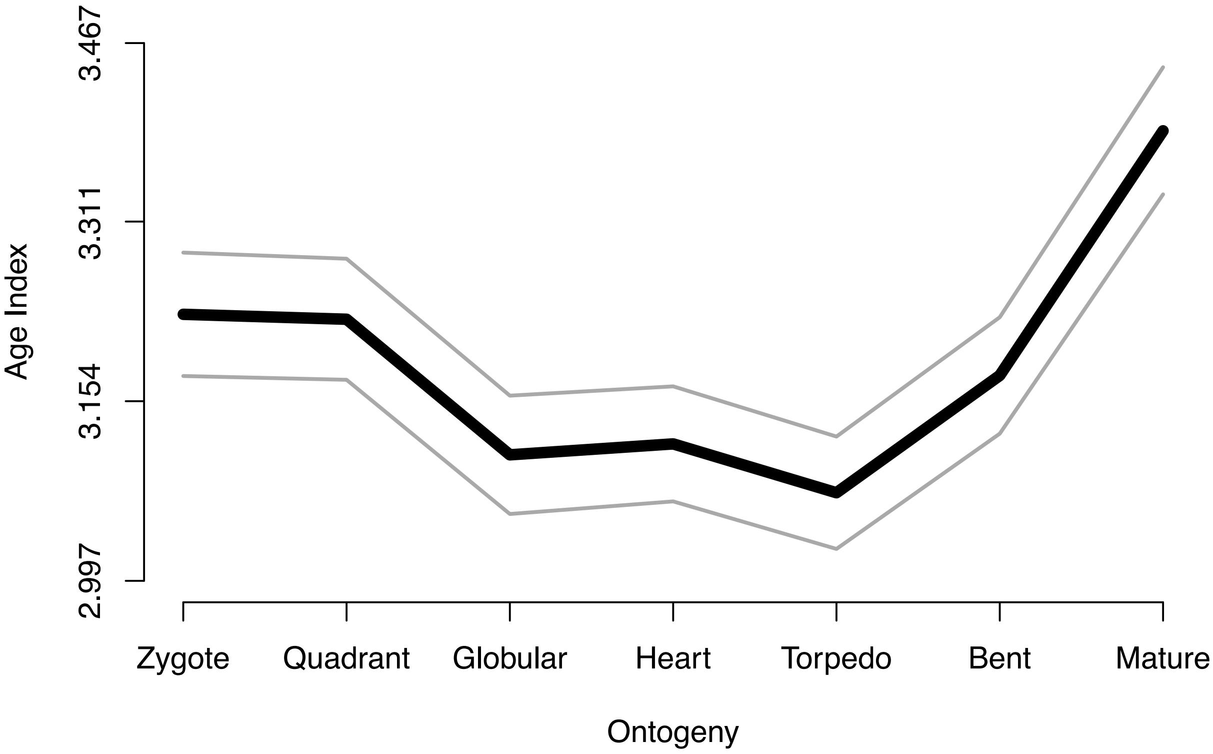
And again compute the `TAI()` profile of the `PhyloExpressionSetExample`.

```r
# TAI profile of correctly assigned Phylostrata
TAI(PhyloExpressionSetExample)
```

```
  Zygote Quadrant Globular    Heart  Torpedo     Bent   Mature
3.229942 3.225614 3.107135 3.116693 3.073993 3.176511 3.390334
```

Visualization:

```r
# visualize the TAI profile of correctly assigned Phylostrata
PlotPattern( ExpressionSet = PhyloExpressionSetExample,
             type          = "l",
             lwd           = 6,
             p.value       = FALSE )
```



The Reductive Hourglass Test is a permutation test based on the following test statistic.

1) A set of developmental stages is partitioned into three modules - `early`, `mid`, and `late` - based on prior biological knowledge (see Drost et al., 2015 for details).
2) The mean TAI or TDI value for each of the three modules T_{early}, T_{mid}, and T_{late} are computed.
3) The two differences D1 = T_{early} - T_{mid} and D2 = T_{late} - T_{mid} are calculated.
4) The minimum D_{min} of D1 and D2 is computed as final test statistic of the **Reductive Hourglass Test**.

In order to determine the statistical significance of an observed minimum difference D_{min} the following permutation test was performed. Based on the `bootMatrix()` D_{min} is calculated from each of the permuted TAI or TDI profiles, approximated by a Gaussian distribution with method of moments estimated parameters returned by `fitdistrplus::fitdist()`, and the corresponding p-value is computed by `pnorm` given the estimated parameters of the Gaussian distribution. The goodness of fit for the random vector D_{min} is statistically quantified by a Lilliefors (Kolmogorov-Smirnov) test for normality.

To perform the **Reductive Hourglass Test** you can use the `ReductiveHourglassTest()` function. Using this function you need to divide the given phylotranscriptomic pattern into three developmental modules:

- early module
- mid module
- late module

This can be done using the `modules` argument: `module = list(early = 1:2, mid = 3:5, late = 6:7)`. In this example (`PhyloExpressionSetExample`) we divide the corresponding developmental process into the three modules:

- early module: Stages 1 - 2 = Zygote and Quadrant
- mid module: Stages 3 - 5 = Globular, Heart, and Torpedo
- late module: Stages 6 - 7 = Bent and Mature

```r
# Perform the Reductive Hourglass Test
ReductiveHourglassTest( ExpressionSet = PhyloExpressionSetExample,
                        modules       = list(early = 1:2, mid = 3:5, late = 6:7),
                        lillie.test   = TRUE )
```

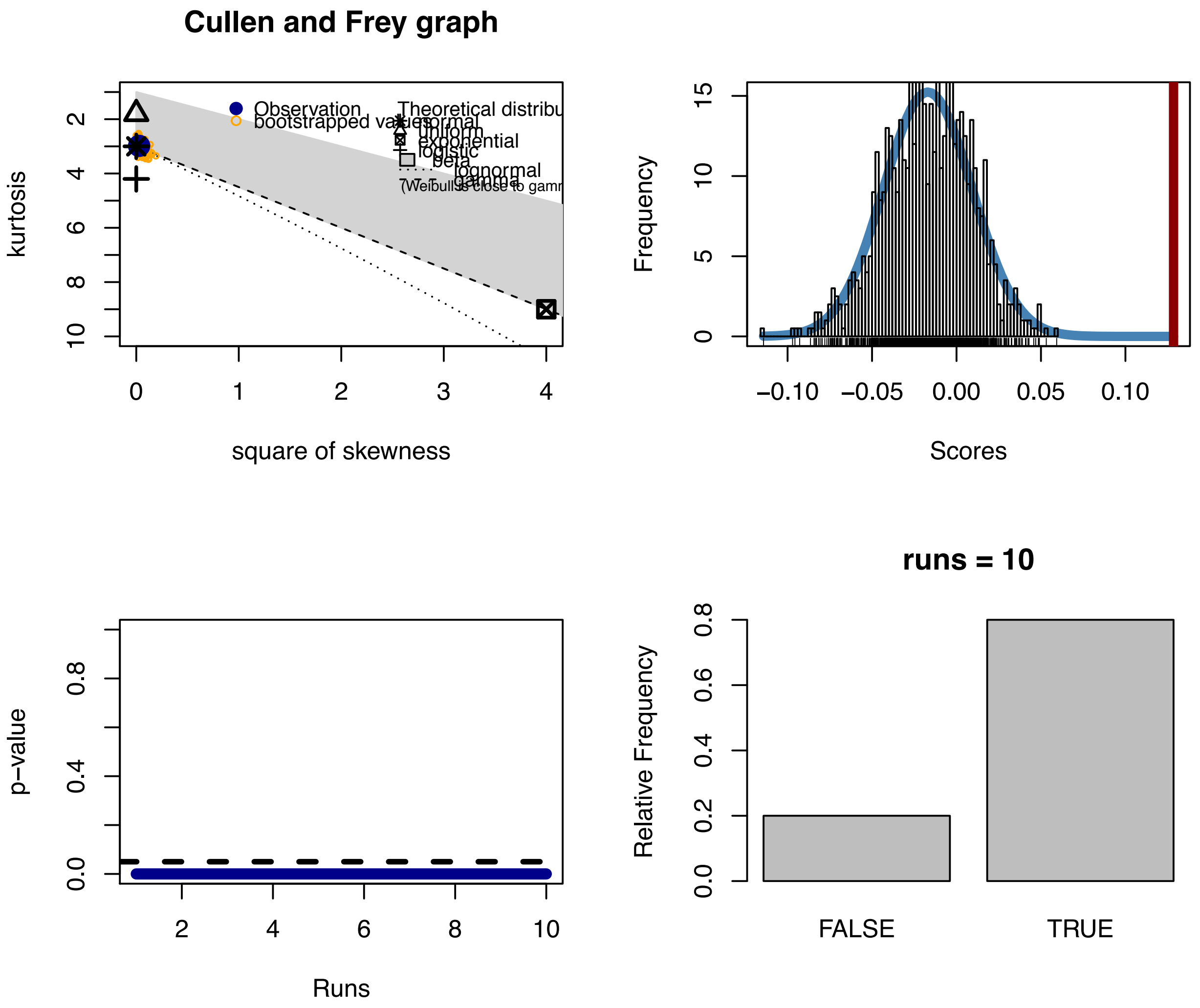
```
#> $p.value
#> [1] 1.240371e-08
#>
#> $std.dev
#> [1] 0.05374032 0.05284646 0.05096514 0.04926126 0.04876508 0.05049721 0.05519132
#>
#> $lillie.test
#> [1] TRUE
```

The corresponding output shows the p-value returned by the **Reductive Hourglass Test**, the standard deviation of randomly permuted TAI profiles returned by `bootMatrix()` (`apply(bootMatrix(PhyloExpressionSetExample) , 2 , sd)`) and in case the argument `lillie.test = TRUE`, a logical value representing the goodness of fit statistic returned by the Lilliefors (Kolmogorov-Smirnov) test for normality. In case `lillie.test` is `TRUE` the corresponding Lilliefors (Kolmogorov-Smirnov) test passed the goodness of fit criterion. In case `lillie.test` is `FALSE` the corresponding goodness of fit by a normal distribution is not statistically significant.

Analogous to the `plotHistogram` argument that is present in the `FlatLineTest()` function, the `ReductiveHourglassTest()` function also takes an argument `plotHistogram`. When `plotHistogram = TRUE`, the `ReductiveHourglassTest()` function returns a multi-plot showing:

- A Cullen and Frey skewness-kurtosis plot. This plot illustrates which distributions seem plausible to fit the resulting permutation vector D_{min}. Here a normal distribution seems most plausible.
- A histogram of D_{min} combined with the density plot is visualized. D_{min} is then fitted by a normal distribution. The corresponding parameters are estimated by moment matching estimation.
- A plot showing the p-values for N independent runs to verify that a specific p-value is biased by a specific permutation order.
- A bar plot showing the number of cases in which the underlying goodness of fit (returned by Lilliefors (Kolmogorov-Smirnov) test for normality) has shown to be significant (`TRUE`) or not significant (`FALSE`). This allows to quantify the permutation bias and their implications on the goodness of fit.

```r
# perform the Reductive Hourglass Test and plot the test statistic
ReductiveHourglassTest( ExpressionSet = PhyloExpressionSetExample,
                        modules       = list(early = 1:2, mid = 3:5, late = 6:7),
                        plotHistogram = TRUE,
                        lillie.test   = TRUE )
```



```
#> $p.value
#> [1] 1.359878e-08
#>
#> $std.dev
#> [1] 0.05434786 0.05421480 0.05283417 0.05163244 0.05089917 0.05125313 0.05577403
#>
#> $lillie.test
#> [1] FALSE
```

The corresponding output shows the `summary statistics` of the fitted normal distribution as well as the p-value, standard deviation, and Lilliefors (Kolmogorov-Smirnov) test result.

This example output nicely illustrates that although the Lilliefors (Kolmogorov-Smirnov) test for normality is violated for some permutations, the Cullen and Frey graph shows that there is no better approximation than a normal distribution (which is also supported visually by investigating the fitted frequency distribution). The corresponding p-value returned by `ReductiveHourglassTest()` is significant and illustrates that the observed phylotranscriptomic pattern of PhyloExpressionSetExample does follow the Hourglass Model assumption.

Reductive Early Conservation Test

The Early Conservation Test has been developed to statistically evaluate the existence of a monotonically increasing phylotranscriptomic pattern based on TAI or TDI computations. The corresponding p-value quantifies the probability that a given TAI or TDI pattern (or any phylotranscriptomic pattern) does not follow an early conservation like pattern. A p-value < 0.05 indicates that the corresponding phylotranscriptomics pattern does indeed follow an early conservation (low-high-high) shape.

To build the test statistic again we start with a standard PhyloExpressionSet.

```
library(myTAI)

# load an example PhyloExpressionSet stored in the myTAI package
data(PhyloExpressionSetExample)

# look at the standardized data set format
head(PhyloExpressionSetExample, 3)
```

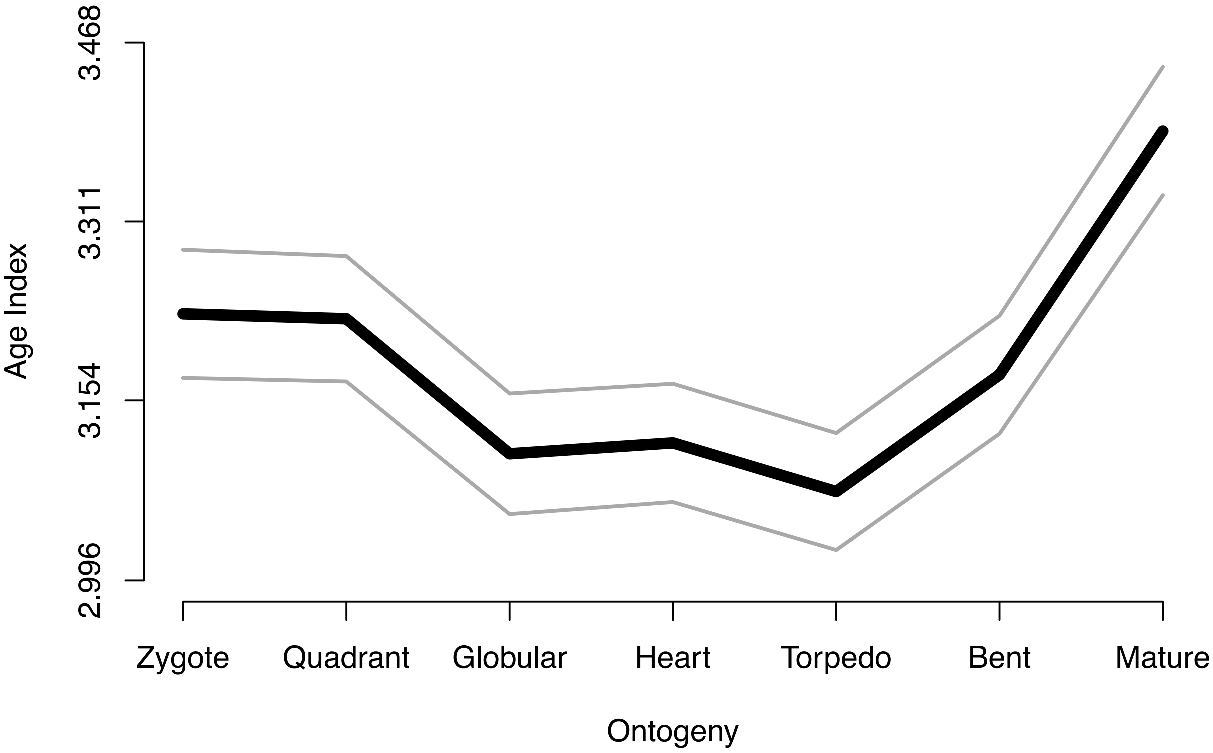
And again compute the `TAI()` profile of the `PhyloExpressionSetExample`.

```
# TAI profile of correctly assigned Phylostrata
TAI(PhyloExpressionSetExample)
```

```
  Zygote Quadrant Globular    Heart  Torpedo     Bent   Mature
3.229942 3.225614 3.107135 3.116693 3.073993 3.176511 3.390334
```

Visualization:

```
# Visualize the TAI profile of correctly assigned Phylostrata
PlotPattern( ExpressionSet = PhyloExpressionSetExample,
             type          = "l",
             lwd           = 6,
             p.value       = FALSE )
```



The reductive early conservation test is a permutation test based on the following test statistic.

1) A set of developmental stages is partitioned into three modules - `early`, `mid`, and `late` - based on prior biological knowledge.
2) The mean TAI or TDI value for each of the three modules T_{early}, T_{mid}, and T_{late} are computed.
3) The two differences D1 = T_{mid} - T_{early} and D2 = T_{late} - T_{early} are calculated.
4) The minimum D_{min} of D1 and D2 is computed as final test statistic of the **Reductive Early Conservation Test**.

In order to determine the statistical significance of an observed minimum difference D_{min} the following permutation test was performed. Based on the `bootMatrix()` D_{min} is calculated from each of the permuted TAI or TDI profiles, approximated by a Gaussian distribution with method of moments estimated parameters returned by `fitdistrplus::fitdist()`, and the corresponding p-value is computed by `pnorm` given the estimated parameters of the Gaussian distribution. The goodness of fit for the random vector D_{min} is statistically quantified by an Lilliefors (Kolmogorov-Smirnov) test for normality.

To perform the **Reductive Early Conservation Test** you can use the `EarlyConservationTest()` function. Using this function you need to divide the given phylotranscriptomics pattern into three developmental modules:

- early module
- mid module
- late module

This can be done using the `modules` argument: `module = list(early = 1:2, mid = 3:5, late = 6:7)`. In this example (`PhyloExpressionSetExample`) we divide the corresponding developmental process into the three modules:

- early module: Stages 1 - 2 = Zygote and Quadrant
- mid module: Stages 3 - 5 = Globular, Heart, and Torpedo
- late module: Stages 6 - 7 = Bent and Mature

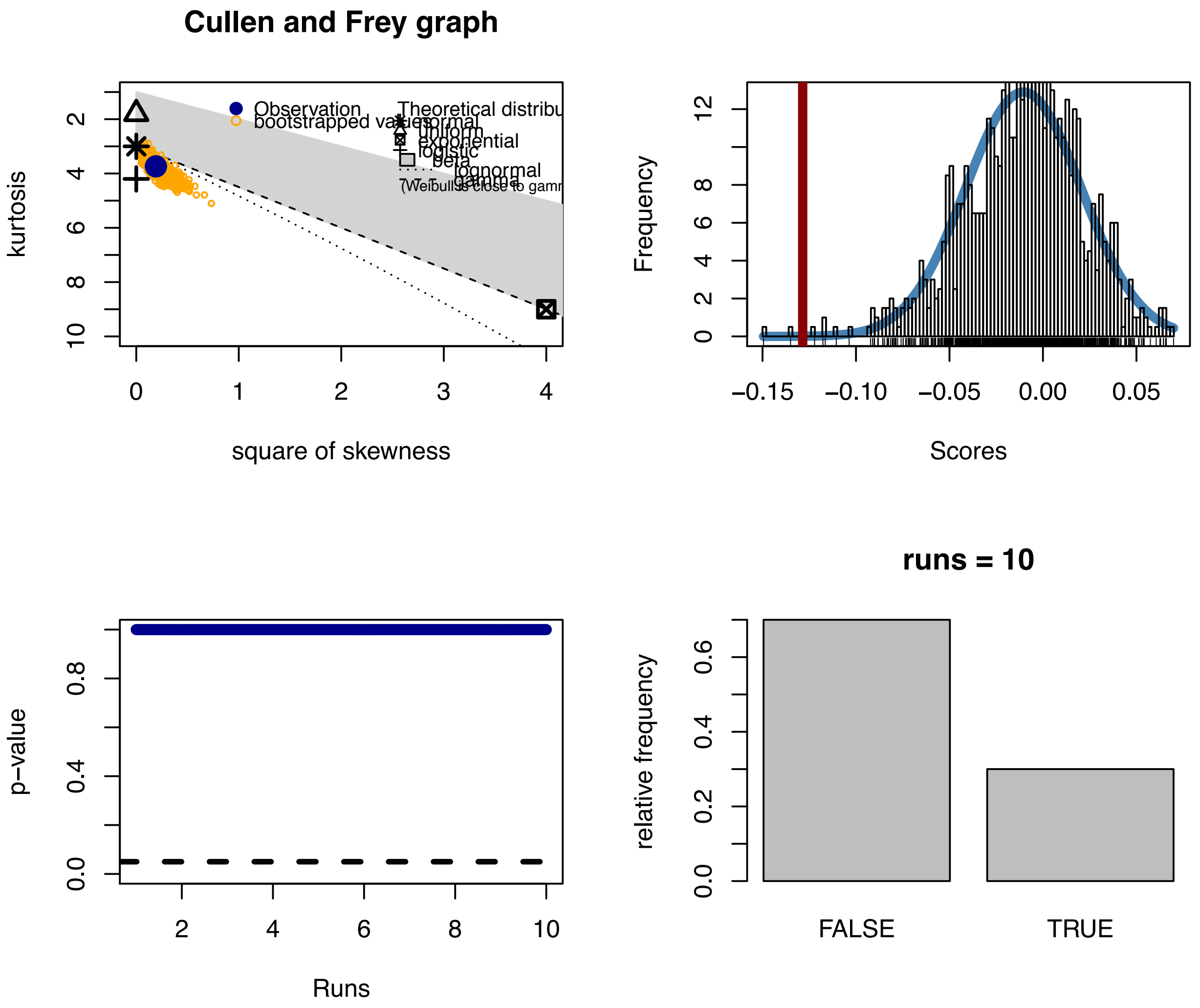
```r
# Perform the Reductive Early Conservation Test
EarlyConservationTest( ExpressionSet = PhyloExpressionSetExample,
                       modules       = list(early = 1:2, mid = 3:5, late = 6:7),
                       lillie.test   = TRUE )
```

```
#> $p.value
#> [1] 0.9999032
#>
#> $std.dev
#> [1] 0.05548875 0.05407686 0.05156524 0.05105336 0.05058059 0.05226052 0.05682250
#>
#> $lillie.test
#> [1] FALSE
```

Analogous to the `plotHistogram` argument that is present in the `FlatLineTest()` and `ReductiveHourglassTest()` function, the `EarlyConservationTest()` function also takes an argument `plotHistogram`. When `plotHistogram = TRUE`, the `EarlyConservationTest()` function returns a multi-plot showing:

- A Cullen and Frey skewness-kurtosis plot. This plot illustrates which distributions seem plausible to fit the resulting permutation vector D_{min}. Again a normal distribution seems most appropriate.
- A histogram of D_{min} combined with the density plot is visualized. D_{min} is then fitted by a normal distribution. The corresponding parameters are estimated by moment matching estimation.
- A plot showing the p-values for N independent runs to verify that a specific p-value is biased by a specific permutation order.
- A bar plot showing the number of cases in which the underlying goodness of fit (returned by Lilliefors (Kolmogorov-Smirnov) test for normality) has shown to be significant (`TRUE`) or not significant (`FALSE`). This allows to quantify the permutation bias and their implications on the goodness of fit.

```r
# perform the Reductive Early Conservation Test and plot the test statistic
EarlyConservationTest( ExpressionSet = PhyloExpressionSetExample,
                       modules       = list(early = 1:2, mid = 3:5, late = 6:7),
                       plotHistogram = TRUE,
                       lillie.test   = TRUE )
```



```
#> $p.value
#> [1] 0.9999331
#>
#> $std.dev
#> [1] 0.05393911 0.05235373 0.05132276 0.05013268 0.04949230 0.05283076 0.05776743
#>
#> $lillie.test
#> [1] FALSE
```

This example output nicely illustrates that although the Lilliefors (Kolmogorov-Smirnov) test for normality is violated, the Cullen and Frey graph shows that there is no better approximation than a normal distribution (which is also supported visually by investigating the fitted frequency distribution). The corresponding p-value returned by the `EarlyConservationTest()` is highly non-significant and illustrates that the observed phylotranscriptomic pattern of `PhyloExpressionSetExample` does not follow the Early Conservation Model assumption.

This example shall illustrate that finding the right test statistic is a multi-step process of investigating different properties of the underlying permutation test. Although single aspects might fit or fit not corresponding criteria, the overall impression (sum of all individual analyses) must be considered to obtain a valid p-value.

Data Transformation

Motivated by the discussion raised by Piasecka et al., 2013, the influence of gene expression transformation on the global phylotranscriptomic patterns does not seem negligible. Hence, different transformations can result in qualitatively different TAI or TDI patterns.

Initially, the TAI and TDI formulas were defined for absolute expression levels. So using the initial TAI and TDI formulas with transformed expression levels can result in qualitatively different patterns when compared with non-transformed expression levels, but might also belong to a different class of models, since different valid expression level transformation functions result in different patterns.

The purpose of the `tf()` function is to allow the user to study the qualitative impact of different transformation functions on the global TAI and TDI pattern, or on any subsequent phylotranscriptomic analysis.

The examples using the `PhyloExpressionSetExample` data set show that using common gene expression transformation functions: log2 (Quackenbush, 2001 and 2002), sqrt (Yeung et al., 2001), boxcox, or inverse hyperbolic sine transformation, each transformation results in qualitatively different patterns. Nevertheless, for each resulting pattern the statistical significance can be tested using either the `FlatLineTest()`, `ReductiveHourglassTest()`, or `EarlyConservationTest()` (Drost et al., 2015) to quantify the significance of observed patterns.

The `tf()` function takes a standard `PhyloExpressionSet` or `DivergenceExpressionSet` and transformation function and returns the corresponding `ExpressionSet` with transformed gene expression levels.

```r
library(myTAI)

data(PhyloExpressionSetExample)

# a simple example is to transform the gene expression levels of a given PhyloExpressionSet
# using a sqrt or log2 transformation

PES.sqrt <- tf(PhyloExpressionSetExample, sqrt)

head(PES.sqrt)
```

```
  Phylostratum      GeneID    Zygote  Quadrant  Globular     Heart
1            1 at1g01040.2  46.62226  43.71728  33.94930  35.93637
2            1 at1g01050.1  38.74292  42.62990  40.80820  39.55706
3            1 at1g01070.1  34.82517  35.11413  30.64637  30.48966
4            1 at1g01080.2  31.88919  30.60039  34.37060  36.46195
5            1 at1g01090.1 106.88576 129.53057 185.38244 199.43831
6            1 at1g01120.1  29.05239  28.06409  29.31939  30.52242
    Torpedo      Bent   Mature
1  31.62678  31.03187 41.18771
2  38.68230  33.38628 32.73615
3  29.39759  29.61766 29.91352
4  37.31813  35.88836 29.34724
5 237.13197 258.80566 88.16213
6  30.70579  29.50021 28.15589
```

```r
PES.log2 <- tf(PhyloExpressionSetExample, log2)

head(PES.log2)
```

```
  Phylostratum      GeneID    Zygote  Quadrant  Globular     Heart
```

```
1            1 at1g01040.2 11.085894 10.900263 10.170620 10.334745
2            1 at1g01050.1 10.551722 10.827588 10.701574 10.611727
3            1 at1g01070.1 10.244117 10.267960  9.875289  9.860497
4            1 at1g01080.2  9.989991  9.870956 10.206206 10.376639
5            1 at1g01090.1 13.479852 14.034298 15.068722 15.279598
6            1 at1g01120.1  9.721170  9.621306  9.747567  9.863595
    Torpedo      Bent    Mature
1  9.966149  9.911358 10.728284
2 10.547204 10.122367 10.065625
3  9.755251  9.776772  9.805452
4 10.443610 10.330888  9.750306
5 15.779093 16.031451 12.924174
6  9.880877  9.765307  9.630730
```

```
# in case a given PhyloExpressionSet already stores gene expression levels
# that are log2 transformed and need to be re-transformed to absolute
# expression levels, to perform subsequent phylotranscriptomics analyses
# (that are defined for absolute expression levels),
# one can re-transform a PhyloExpressionSet like this:

PES.absolute <- tf(PES.log2 , function(x) 2^x)

# which should be the same as  PhyloExpressionSetExample :
head(PhyloExpressionSetExample)
head(PES.absolute)
```

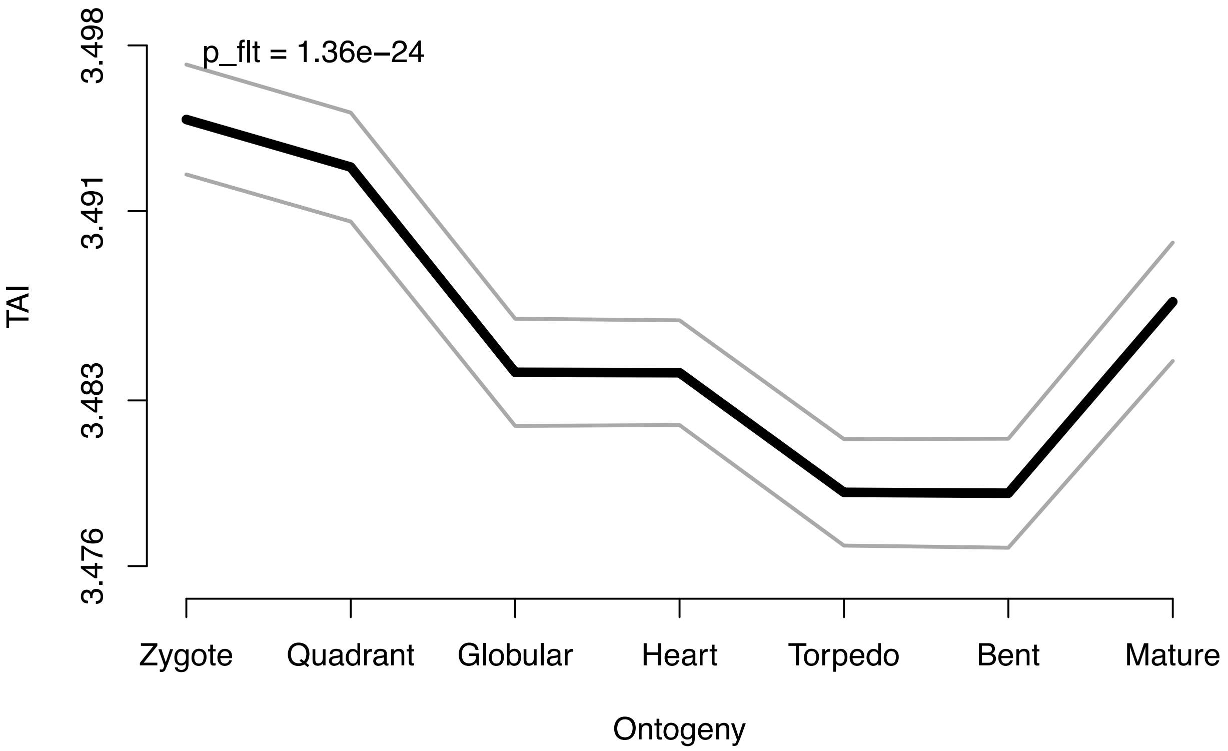
```
> head(PhyloExpressionSetExample)
  Phylostratum      GeneID     Zygote   Quadrant   Globular      Heart
1            1 at1g01040.2  2173.6352  1911.2001  1152.5553  1291.4224
2            1 at1g01050.1  1501.0141  1817.3086  1665.3089  1564.7612
3            1 at1g01070.1  1212.7927  1233.0023   939.2000   929.6195
4            1 at1g01080.2  1016.9203   936.3837  1181.3381  1329.4734
5            1 at1g01090.1 11424.5667 16778.1685 34366.6493 39775.6405
6            1 at1g01120.1   844.0414   787.5929   859.6267   931.6180
     Torpedo       Bent    Mature
1  1000.2529   962.9772 1696.4274
2  1496.3207  1114.6435 1071.6555
3   864.2180   877.2060  894.8189
4  1392.6429  1287.9746  861.2605
5 56231.5689 66980.3673 7772.5617
6   942.8453   870.2625  792.7542

> head(PES.absolute)
  Phylostratum      GeneID     Zygote   Quadrant   Globular      Heart
1            1 at1g01040.2  2173.6352  1911.2001  1152.5553  1291.4224
2            1 at1g01050.1  1501.0141  1817.3086  1665.3089  1564.7612
3            1 at1g01070.1  1212.7927  1233.0023   939.2000   929.6195
4            1 at1g01080.2  1016.9203   936.3837  1181.3381  1329.4734
5            1 at1g01090.1 11424.5667 16778.1685 34366.6493 39775.6405
6            1 at1g01120.1   844.0414   787.5929   859.6267   931.6180
     Torpedo       Bent    Mature
1  1000.2529   962.9772 1696.4274
2  1496.3207  1114.6435 1071.6555
3   864.2180   877.2060  894.8189
```

```
4  1392.6429  1287.9746  861.2605
5 56231.5689 66980.3673 7772.5617
6   942.8453   870.2625  792.7542
```

When transforming the `ExpressionMatrix` of the `PhyloExpressionSetExample` using different transformation functions, the resulting phylotranscriptomic patterns qualitatively differ:

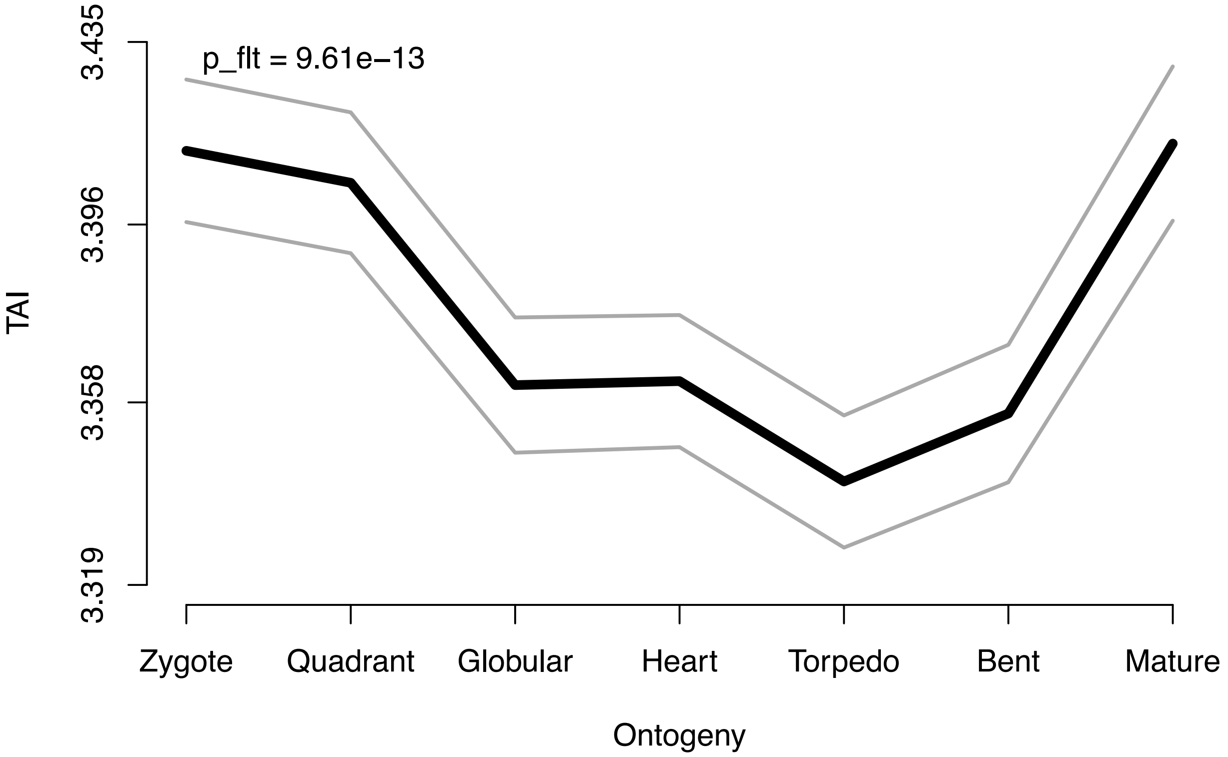
log2 transformation (TAI)

```
data(PhyloExpressionSetExample)

# plotting the TAI using log2 transformed expression levels
# and performing the Flat Line Test to obtain the p-value
PlotPattern( ExpressionSet = tf(PhyloExpressionSetExample, log2),
             type          = "l",
             lwd           = 5,
             TestStatistic = "FlatLineTest",
             xlab          = "Ontogeny",
             ylab          = "TAI" )
```



sqrt transformation (TAI)

```
data(PhyloExpressionSetExample)

# plotting the TAI using sqrt transformed expression levels
# and performing the Flat Line Test to obtain the p-value
PlotPattern( ExpressionSet = tf(PhyloExpressionSetExample, sqrt),
             TestStatistic = "FlatLineTest",
             type          = "l",
             lwd           = 5,
             xlab          = "Ontogeny",
             ylab          = "TAI" )
```



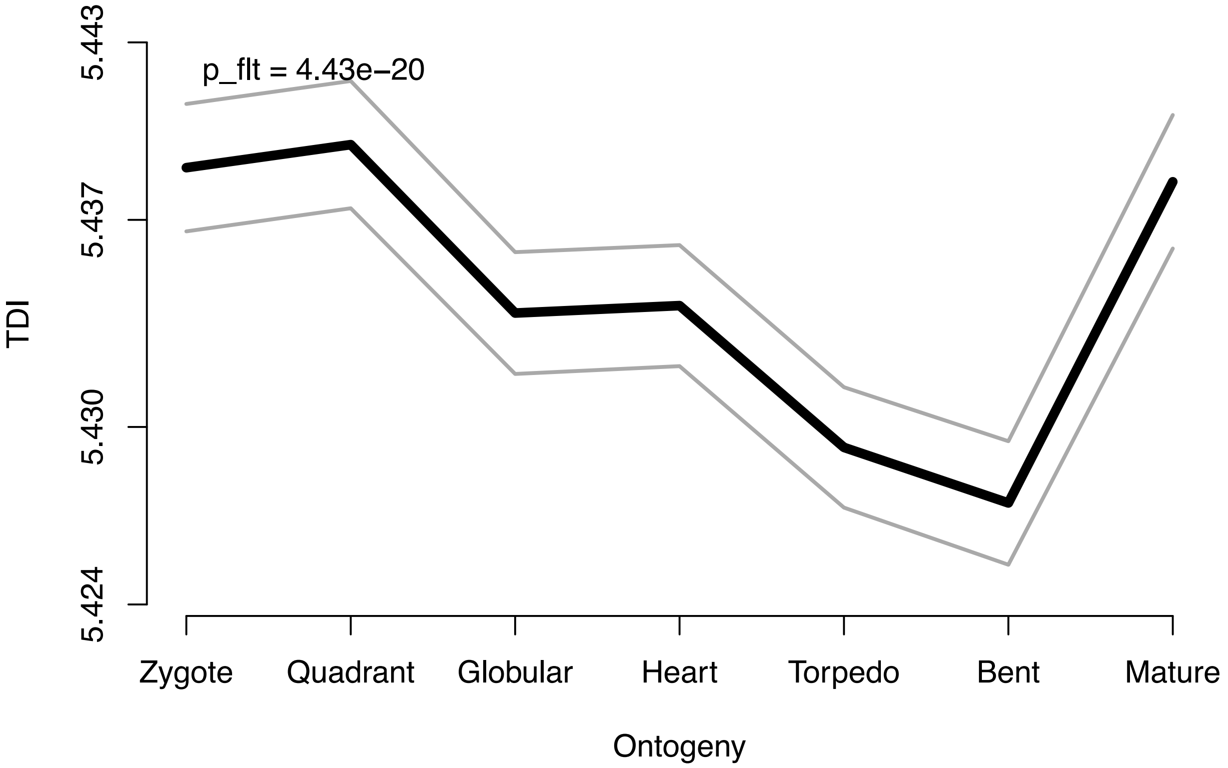
For the `PhyloExpressionSetExample` all transformations result in a significant phylotranscriptomics pattern deviating from a flat line.

Nevertheless, it is not clear which transformation is the most appropriate one since the original TAI and TDI measure were defined for absolute expression levels.

The same accounts for `TDI` profiles:

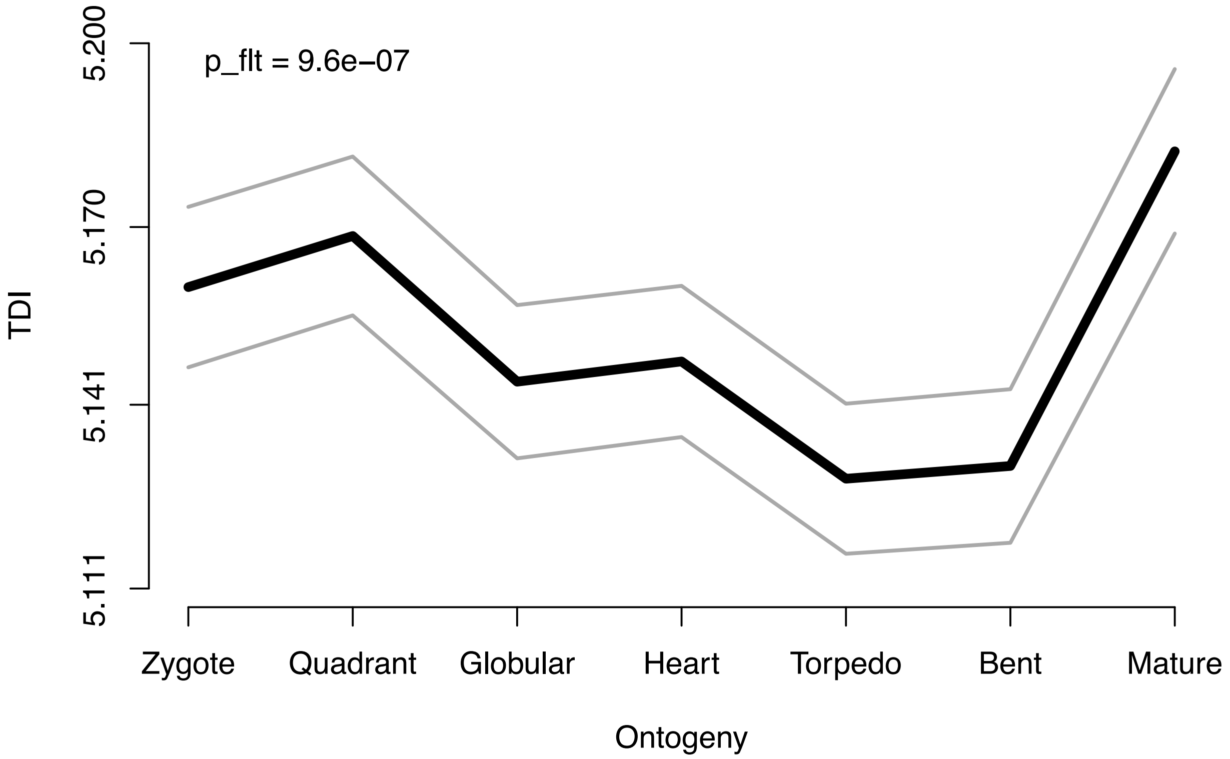
log2 transformation (TDI)

```
data(DivergenceExpressionSetExample)

# plotting the TDI using log2 transformed expression levels
# and performing the Flat Line Test to obtain the p-value
PlotPattern( ExpressionSet = tf(DivergenceExpressionSetExample, log2),
             TestStatistic = "FlatLineTest",
             type          = "l",
             lwd           = 5,
             xlab          = "Ontogeny",
             ylab          = "TDI" )
```



sqrt transformation (TDI)

```
data(DivergenceExpressionSetExample)

# plotting the TDI using sqrt transformed expression levels
# and performing the Flat Line Test to obtain the p-value
PlotPattern( ExpressionSet = tf(DivergenceExpressionSetExample, sqrt),
             TestStatistic = "FlatLineTest",
             type          = "l",
             lwd           = 5,
             xlab          = "Ontogeny",
             ylab          = "TDI" )
```



As a result, observed patterns should always be quantified using statistical tests (for ex. `FlatLineTest()`, `ReductiveHourglassTest()`, and `EarlyConservationTest()`). In case the observed pattern is significant, qualitative differences of the observed patterns based on different data transformations must be investigated in more detail, since most data transformations are known to cause different effects on a measure that isn't robust against data transformations.

Expression Data Analysis with `myTAI`

In the `Introduction` we introduced and discussed how phylotranscriptomics can be applied to capture evolutionary signatures in (developmental) transcriptomes. Furthermore, in the `Enrichment Analyses` section we provide a use case to correlate specific groups or sets of genes with their predicted evolutionary origin. Here, we aim to combine previously introduced techniques with *classic* gene expression analyses to detect possible functional causes for the observed transcriptome conservation.

In other words, phylotranscriptomics allows us to detect stages or periods of evolutionary conservation and is able to predict the evolutionary origin of process or trait specific genes based on enrichment analyses. By combining evolutionary enrichment analyses with the functional annotation of process or trait specific genes (see Functional Annotation and Phylotranscriptomics for details) the detection of evolutionary signals can be correlated with functional processes. Then, performing gene expression analyses on corresponding process or trait specific genes allows users to detect potential causes of stage/period specific evolutionary transcriptome conservation.

The following sections introduce main gene expression data analysis techniques implemented in `myTAI`:

- Detection of Differentially Expressed Genes (DEGs)
- Fold-Change
- Welch t-test

- Wilcoxon Rank Sum Test (Mann-Whitney U test)
- Negative Binomial (Exact Tests)
- Collapsing Replicate Samples
- Filter for Expressed Genes
- Compute the Statistical Significance of Each Replicate Combination

Detection of Differenentially Expressed Genes (DEGs)

A variety of methods have been published to detect differentially expressed genes. Some methods are based on non-statistical quantification of expression differences (e.g. fold-change and log-fold-change), but most methods are based on statistical tests to quantify the significance of differences in gene expression between samples. These statistical methods can furthermore be divided into two methodological categories: parametric tests and non-parametric tests. The `DiffGenes()` function available in `myTAI` implements the most popular and useful methods to detect differentially expressed genes. In the literature, different methods have been introduced and discussed for microarray technologies versus RNA-Seq technologies.

In this section we will introduce all methods implemented in `DiffGenes()` using small examples and will furthermore, discuss published advantages and disadvantages of each method and each mRNA quantification technology.

Note that when using `DiffGenes()` it is assumed that your input dataset has been normalized before passing it to `DiffGenes()`. For RNA-Seq data `DiffGenes()` assumes that the libraries have been normalized to have the same size, i.e., to have the same expected column sum under the null hypothesis (or the lib.size argument in `DiffGenes()` is specified accordingly).

Fold-Changes

A fold change in gene expression is simply the ratio of the gene expression level of one sample against a second sample: $\frac{e_{i1}}{e_{i2}}$, where e_{i1} is the expression level of gene i in sample one and e_{i2} is the expression level of gene i in sample two. In case replicate expression levels are present for each sample the ratio of means of the corresponding replicates is computed: $\frac{\bar{e}_{i1}}{\bar{e}_{i2}}$, where $\bar{e}_{i1}$ is the mean of replicate expression levels of gene i in sample one and $\bar{e}_{i2}$ is the mean of replicate expression levels of gene i in sample two.

- **Advantages:** Given a small number of replicate values the statistical evaluation of differentially expressed genes might be biased (depending on the statistical test chosen) by underlying sample distributions which are not fulfilled or because a small number of replicate values is not sufficient enough to perform non-parametric tests. Here, fold-changes provide a simple way to quantify gene expression differences between samples by n-fold enrichment. In our opinion, although the process of choosing a threshold for defining genes as being differentially expressed or not based on fold-change values is purely subjective and relies on common sense, in some cases this procedure will be more suitable than defining differentially expressed genes based on p-values obtained from a test statistic with violated test assumptions.
- **Disadvantages:** If used appropriately, statistical tests not only systematically quantify the significance of the observed gene-by-gene differences of expression, but furthermore, accounts the variance of replicate expression levels when comparing the mean difference of replicate expression levels between samples. Hence, the gene specific variance between replicates is also quantified by the p-value returned by the sufficient test statistic which is not quantified by a simple fold-change measure.

Example: Fold-Change

For the following example we assume that `PhyloExpressionSetExample[1:5,1:8]` stores 5 genes and 3 developmental stages with 2 replicate expression levels per stage.

```
data("PhyloExpressionSetExample")

# Detection of DEGs using the fold-change measure
DEGs <- DiffGenes(ExpressionSet = PhyloExpressionSetExample[1:5,1:8],
                  nrep          = 2,
                  method        = "foldchange",
                  stage.names   = c("S1","S2","S3"))


head(DEGs)
```

```
  Phylostratum      GeneID    S1->S2    S1->S3    S2->S1    S2->S3    S3->S1    S3->S2
1            1 at1g01040.2 1.6713881 2.0806706 0.5983051 1.2448758 0.4806143 0.8032930
2            1 at1g01050.1 1.0273222 1.2709185 0.9734045 1.2371177 0.7868325 0.8083305
3            1 at1g01070.1 1.3087379 1.4044799 0.7640949 1.0731560 0.7120073 0.9318310
4            1 at1g01080.2 0.7779572 0.7286769 1.2854177 0.9366542 1.3723503 1.0676299
5            1 at1g01090.1 0.3803866 0.2288961 2.6289042 0.6017460 4.3687939 1.6618307
```

The resulting output shows all combinations of fold-changes between samples (developmental stages). Here, `S1->S2` denotes that the fold-change was computed for expression levels of stage `S1` against stage `S2`.

Example: Log-Fold-Change

When selecting `method = "log-foldchange"` it is assumed that the input `ExpressionSet` stores `log2` expression levels. Here, we transform absolute expression levels stored in `PhyloExpressionSetExample` to `log2` expression levels using the `tf()` function before log-fold-changes are computed.

```
data("PhyloExpressionSetExample")

# Detection of DEGs using the logfold-change measure
log.DEGs <- DiffGenes(ExpressionSet = tf(PhyloExpressionSetExample[1:5,1:8],log2),
                      nrep          = 2,
                      method        = "log-foldchange",
                      stage.names   = c("S1","S2","S3"))


head(log.DEGs)
```

```
 Phylostratum      GeneID      S1->S2     S1->S3      S2->S1      S2->S3     S3->S1      S3->S2
1            1 at1g01040.2  0.74104679  1.0570486 -0.74104679  0.31600182 -1.0570486 -0.31600182
2            1 at1g01050.1  0.03888868  0.3458715 -0.03888868  0.30698280 -0.3458715 -0.30698280
3            1 at1g01070.1  0.38817621  0.4900360 -0.38817621  0.10185975 -0.4900360 -0.10185975
4            1 at1g01080.2 -0.36223724 -0.4566488  0.36223724 -0.09441158  0.4566488  0.09441158
5            1 at1g01090.1 -1.39446159 -2.1272350  1.39446159 -0.73277345  2.1272350  0.73277345
```

The resulting output stores all combinations of log fold-changes between samples (developmental stages).

Welch t-test

The `Welch t-test` is a parametric test to statistically quantify the difference of sample means in cases where the assumption of homogeneity of variance (equal variances in the two populations) is violated (Boslaugh, 2013). The `Welch t-test` is a sufficient parameter test for small sample sizes and thus, has been used to detect differentially expressed genes based on p-values returned by the test statistic (Hahne et al., 2008).

In detail, the test statistic is computed as follows:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where $\bar{x}_1$ and $\bar{x}_2$ are sample means, s_1^2 and s_2^2 are the sample variances, and n_1 and n_2 are the sample sizes.

The degrees of freedom for Welch's t-test are then computed as follows:

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_2^4}{n_2^2(n_2-1)}}$$

To perform a sufficient `Welch t-test` the following assumptions about the input data need to be fulfilled to test whether two samples come from populations with equal means:

Assumptions about input data

- independent samples
- continuous data
- (approximate) normality

Nevertheless, although in most cases `log2` expression levels are used to perform the `Welch t-test` assuming that expression levels are log-normal distributed which approximates a normal distribution in infinity, in most cases the small number of replicates is not sufficient enough to fulfill the (approximate) normality assumption of the `Welch t-test`.

Due to this fact, non-parametric, sampling based, or generalized linear model based methods have been proposed to quantify p-values of differential expression. Nevertheless, the `DiffGenes()` function implements the `Welch t-test` for the detection of differentially expressed genes, allowing users to compare the results with more recent DEG detection methods/methodologies also implemented in `DiffGenes()`.

- **Advantages:**
- DEG detection based on statistical quantification
- Parametric test resulting in a strong test statistic
- Can handle small sample sizes
- **Disadvantages:**
- Test assumptions must be fulfilled to return sufficient p-values

- Can hardly assure normality with very sample sizes of $n = 3, 4, 5, ..$ (replicates)
- Pairwise comparisons between different stages or experiments

Example: Welch t-test

Performing `Welch t-test` with `DiffGenes()` can be done by specifying `method = "t.test"`. Internally `DiffGenes()` performs a two sided `Welch t-test`. This means that the `Welch t-test` quantifies only whether or not a gene is significantly differentially expressed, but not the direction of enrichment (over-expressed or under-expressed).

The `PhyloExpressionSetExample` we use in the following example stores absolute expression levels. In case your `ExpressionSet` also stores absolute expression levels (which is likely due to the `ExpressionSet` standard for Phylotranscriptomics analyses), you can use the `tf()` function implemented in `myTAI` to transform absolute expression levels to `log2` expression levels before performing `DiffGenes()` with a `Welch t-test`, e.g. `tf(PhyloExpressionSetExample[1:5,1:8],log2)`. In general, using `log2` transformed expression levels as input `ExpressionSet` of `DiffGenes()` allows us to (at least) assume that samples (replicate expression levels) used to perform the `Welch t-test` are log-normal distributed and therefore, somewhat approximate normal distributed.

Please notice however, that RNA-Seq data can include count values of 0. So when transforming absolute counts to `log2` counts infinity values of `log2(0) = -Inf` will be produced and therefore, p-value computations will not be possible. To avoid this case you could either remove RNA-Seq count values of 0 from the input dataset using the `Expressed()` function (see section *Filter for Expressed Genes*), e.g. pass `tf(Expressed(PhyloExpressionSetExample[1:5,1:8], cut.off = 1),log2)` as `ExpressionSet` argument to `DiffGenes()` or shift all count values by a constant value, e.g. pass `tf(PhyloExpressionSetExample[1:5,1:8], function(x) log2(x + 1))` as `ExpressionSet` argument to `DiffGenes()`.

Internally, `DiffGenes()` will also check for 0 values in input data and will automatically shift all expression levels by `+1` in case 0 values are included.

```r
data("PhyloExpressionSetExample")

# Detection of DEGs using the p-value returned by a Welch t-test
ttest.DEGs <- DiffGenes(ExpressionSet = tf(PhyloExpressionSetExample[1:5,1:8],log2),
                        nrep          = 2,
                        method        = "t.test",
                        stage.names   = c("S1","S2","S3"))

# look at the results
ttest.DEGs
```

```
  Phylostratum      GeneID     S1<->S2    S1<->S3    S2<->S3
1            1 at1g01040.2 0.027832572 0.04020203 0.13481563
2            1 at1g01050.1 0.852379466 0.31471871 0.36326955
3            1 at1g01070.1 0.003200692 0.00113536 0.02236621
4            1 at1g01080.2 0.086426813 0.03092924 0.45999438
5            1 at1g01090.1 0.090387087 0.04638872 0.04978092
```

The resulting `data.frame` stores the p-values of stage-wise comparisons for each gene. To adjust p-values for multiple testing of stage-wise comparisons you can specify the `p.adjust.method` argument with one of the p-value adjustment methods implemented in `DiffGenes()`.

In detail, correcting for multiple testing allows to appropriately choose selection cut-offs for p-values fulfilling the differential expression criteria. Hahne et al., 2008 (p. 87) give a nice example of correcting for multiple testing to determine appropriate selection cut-offs.

Please consult the documentation of `?p.adjust` to see which p-value adjustment methods are implemented in `DiffGenes()`.

Please also consult these reviews (Biostatistics Handbook, Gelman et al., 2008, and Slides) to decide whether or not to apply p-value adjustment to your own dataset.

```r
data("PhyloExpressionSetExample")

# Detection of DEGs using the p-value returned by a Welch t-test
# and furthermore, adjust p-values for multiple comparison
# using the Benjamini & Hochberg (1995) method: method = "BH"
ttest.DEGs.p_adjust <-
        DiffGenes(
        ExpressionSet   = tf(PhyloExpressionSetExample[1:5, 1:8], log2),
        nrep            = 2,
        method          = "t.test",
        p.adjust.method = "BH",
        stage.names     = c("S1", "S2", "S3")
        )


ttest.DEGs.p_adjust
```

```
  Phylostratum      GeneID    S1<->S2   S1<->S3   S2<->S3
1            1 at1g01040.2 0.06958143 0.0579859 0.2246927
2            1 at1g01050.1 0.85237947 0.3147187 0.4540869
3            1 at1g01070.1 0.01600346 0.0056768 0.1118311
4            1 at1g01080.2 0.11298386 0.0579859 0.4599944
5            1 at1g01090.1 0.11298386 0.0579859 0.1244523
```

The resulting p-value adjusted `data.frame` can be used to filter for differentially expressed genes. Here, specifying the arguments: `comparison`, `alpha`, and `filter.method` in `DiffGenes()` allows users to obtain only significant differentially expressed genes.

```r
# Detection of DEGs using the p-value returned by a Welch t-test
# and furthermore, adjust p-values for multiple comparison
# using the Benjamini & Hochberg (1995) method: method = "BH"
# and filter for significantly differentially expressed genes (alpha = 0.05)
ttest.DEGs.p_adjust.filtered <-
        DiffGenes(
        ExpressionSet   = tf(PhyloExpressionSetExample[1:10 , 1:8], log2),
        nrep            = 2,
        method          = "t.test",
        p.adjust.method = "BH",
        stage.names     = c("S1", "S2", "S3"),
        comparison      = "above",
        alpha           = 0.05,
        filter.method   = "n-set",
        n               = 1
        )

# look at the genes fulfilling the filter criteria
ttest.DEGs.p_adjust.filtered
```

```
  Phylostratum    GeneID   S1<->S2  S1<->S3  S2<->S3
3      1 at1g01070.1 0.03200692 0.0113536 0.2192432
```

In this example, only 1 out of 10 genes fulfills the p-value criteria (`alpha = 0.05`) in at least one stage comparison.

Rank top p-values

Finally, users can rank genes in increasing p-value order for each stage comparison by typing:

```
ttest.DEGs.p_adjust <-
        DiffGenes(
        ExpressionSet   = tf(PhyloExpressionSetExample[1:500, 1:8], log2),
        nrep            = 2,
        method          = "t.test",
        p.adjust.method = "BH",
        stage.names     = c("S1", "S2", "S3")
        )


        head(ttest.DEGs.p_adjust[order(ttest.DEGs.p_adjust[, "S1<->S2"], decreasing = FALSE) , 1:3])
```

```
    Phylostratum      GeneID  S1<->S2
54             1 at1g02400.1 0.151388
119            1 at1g03870.1 0.151388
137            1 at1g04380.1 0.151388
289            1 at1g08110.4 0.151388
383            1 at1g10360.1 0.151388
413            1 at1g11040.1 0.151388
```

Here the line `ttest.DEGs.p_adjust[order(ttest.DEGs.p_adjust[ , "S1<->S2"], decreasing = FALSE) , 1:3]` will sort p-values of stage comparison `"S1<->S2"` in increasing order.

Wilcoxon-Mann-Whitney test (Mann-Whitney U test)

The Wilcoxon-Mann-Whitney test is a *nonparametric* test to quantify the shift in empirical distribution parameters. *Nonparametric* tests are useful when sample populations do not meet the test assumptions of *parametric* tests.

```
data("PhyloExpressionSetExample")

# Detection of DEGs using the p-value returned by a Wilcoxon-Mann-Whitney test
Wilcox.DEGs <- DiffGenes(ExpressionSet = PhyloExpressionSetExample[1:5,1:8],
                        nrep          = 2,
                        method        = "wilcox.test",
                        stage.names   = c("S1","S2","S3"))

# look at the results
Wilcox.DEGs
```

```
  Phylostratum      GeneID   S1<->S2   S1<->S3   S2<->S3
1            1 at1g01040.2 0.3333333 0.3333333 0.3333333
2            1 at1g01050.1 1.0000000 0.3333333 0.3333333
3            1 at1g01070.1 0.3333333 0.3333333 0.3333333
4            1 at1g01080.2 0.3333333 0.3333333 0.6666667
5            1 at1g01090.1 0.3333333 0.3333333 0.3333333
```

Again, users can adjust p-values by specifying the `p.adjust.method` argument.

```
data("PhyloExpressionSetExample")

# Detection of DEGs using the p-value returned by a Wilcoxon-Mann-Whitney test
# and furthermore, adjust p-values for multiple comparison
# using the Benjamini & Hochberg (1995) method: method = "BH"
# and filter for significantly differentially expressed genes (alpha = 0.05)
Wilcox.DEGs.adj <- DiffGenes(ExpressionSet   = PhyloExpressionSetExample[1:5,1:8],
                             nrep            = 2,
                             method          = "wilcox.test",
                             stage.names     = c("S1","S2","S3"),
                             p.adjust.method = "BH")

# look at the results
Wilcox.DEGs.adj
```

```
  Phylostratum      GeneID   S1<->S2   S1<->S3   S2<->S3
1            1 at1g01040.2 0.4166667 0.3333333 0.4166667
2            1 at1g01050.1 1.0000000 0.3333333 0.4166667
3            1 at1g01070.1 0.4166667 0.3333333 0.4166667
4            1 at1g01080.2 0.4166667 0.3333333 0.6666667
5            1 at1g01090.1 0.4166667 0.3333333 0.4166667
```

Negative Binomial (Exact Tests)

Exact Tests for Differences between two groups of negative-binomial counts implemented in `DiffGenes()` are based on the `edgeR` function `exactTest()`. Please consult the edgeR Users Guide for mathematical details.

Install edgeR Package

The detection of DEGs using negative binomial models is based on the powerful implementations provided by the edgeR package. Hence, before using the negative binomial models in `DiffGenes()` users need to install the edgeR package.

```
# install edgeR
source("http://bioconductor.org/biocLite.R")
biocLite("edgeR")
```

Double Tail Method

This method computes two-sided p-values by doubling the smaller tail probability (see `?exactTestByDeviance` for details). To compute p-values for stagewise comparisons based on negative binomial models, the `DiffGenes()` argument `method = "doubletail"`, the number of replicates per stage `nrep`, and `lib.size` quantifying the library size to equalize sample library sizes by quantile-to-quantile normalization need to be specified (see also `?equalizeLibSizes`).

```
data("PhyloExpressionSetExample")

# Detection of DEGs using the p-value returned by the Double Tail Method
DoubleTail.DEGs <- DiffGenes(ExpressionSet = PhyloExpressionSetExample[1:5,1:8],
                        nrep          = 2,
                        method        = "doubletail",
                        lib.size      = 1000,
                        stage.names   = c("S1","S2","S3"))

# look at the results
DoubleTail.DEGs
```

```
  Phylostratum      GeneID    S1<->S2     S1<->S3   S2<->S3
1            1 at1g01040.2 0.26026604 0.110233012 0.6304508
2            1 at1g01050.1 0.95314428 0.598102712 0.6398757
3            1 at1g01070.1 0.55461941 0.456018563 0.8774231
4            1 at1g01080.2 0.58130025 0.487028051 0.8860005
5            1 at1g01090.1 0.03615134 0.001773543 0.2645537
```

Again, users can adjust p-values by specifying the `p.adjust.method` argument.

```
data("PhyloExpressionSetExample")

# Detection of DEGs using the p-value returned by the Double Tail Method
# and furthermore, adjust p-values for multiple comparison
# using the Benjamini & Hochberg (1995) method: method = "BH"
# and filter for significantly differentially expressed genes (alpha = 0.05)
DoubleTail.DEGs.adj <- DiffGenes(ExpressionSet   = PhyloExpressionSetExample[1:5,1:8],
                                  nrep            = 2,
                                  method          = "doubletail",
                                  lib.size        = 1000,
                                  stage.names     = c("S1","S2","S3"),
                                  p.adjust.method = "BH")

# look at the results
DoubleTail.DEGs.adj
```

```
  Phylostratum      GeneID   S1<->S2     S1<->S3   S2<->S3
1            1 at1g01040.2 0.6506651 0.275582530 0.8860005
2            1 at1g01050.1 0.9531443 0.598102712 0.8860005
3            1 at1g01070.1 0.7266253 0.598102712 0.8860005
4            1 at1g01080.2 0.7266253 0.598102712 0.8860005
5            1 at1g01090.1 0.1807567 0.008867715 0.8860005
```

Small-P Method

This method performs the method of small probabilities as proposed by Robinson and Smyth (2008) (see `exactTestBySmallP` for details). To compute p-values for stagewise comparisons based on negative binomial models, the `DiffGenes()` argument `method = "doubletail"`, the number of replicates per stage `nrep`, and `lib.size` quantifying the library size to equalize sample library sizes by quantile-to-quantile normalization need to be specified (see also `?equalizeLibSizes`).

```
data("PhyloExpressionSetExample")

# Detection of DEGs using the p-value returned by the Small-P Method
```

```
SmallP.DEGs <- DiffGenes(ExpressionSet = PhyloExpressionSetExample[1:5,1:8],
                         nrep          = 2,
                         method        = "smallp",
                         lib.size      = 1000,
                         stage.names   = c("S1","S2","S3"))

# look at the results
SmallP.DEGs
```

```
  Phylostratum      GeneID    S1<->S2     S1<->S3   S2<->S3
1            1 at1g01040.2 0.26026604 0.110233012 0.6304508
2            1 at1g01050.1 0.95314428 0.598102712 0.6398757
3            1 at1g01070.1 0.55461941 0.456018563 0.8774231
4            1 at1g01080.2 0.58130025 0.487028051 0.8860005
5            1 at1g01090.1 0.03615134 0.001773543 0.2645537
```

Again, users can adjust p-values by specifying the `p.adjust.method` argument.

```
data("PhyloExpressionSetExample")

# Detection of DEGs using the p-value returned by the Small-P Method
# and furthermore, adjust p-values for multiple comparison
# using the Benjamini & Hochberg (1995) method: method = "BH"
# and filter for significantly differentially expressed genes (alpha = 0.05)
SmallP.DEGs.adj <- DiffGenes(ExpressionSet  = PhyloExpressionSetExample[1:5,1:8],
                             nrep            = 2,
                             method          = "smallp",
                             lib.size        = 1000,
                             stage.names     = c("S1","S2","S3"),
                             p.adjust.method = "BH")

# look at the results
SmallP.DEGs.adj
```

```
  Phylostratum      GeneID   S1<->S2     S1<->S3   S2<->S3
1            1 at1g01040.2 0.6506651 0.275582530 0.8860005
2            1 at1g01050.1 0.9531443 0.598102712 0.8860005
3            1 at1g01070.1 0.7266253 0.598102712 0.8860005
4            1 at1g01080.2 0.7266253 0.598102712 0.8860005
5            1 at1g01090.1 0.1807567 0.008867715 0.8860005
```

Deviance Method

This method uses the deviance goodness of fit statistics to define the rejection region, and is therefore equivalent to a conditional likelihood ratio test (see `exactTestByDeviance` for details). To compute p-values for stagewise comparisons based on negative binomial models, the `DiffGenes()` argument `method = "doubletail"`, the number of replicates per stage `nrep`, and `lib.size` quantifying the library size to equalize sample library sizes by quantile-to-quantile normalization need to be specified (see also `?equalizeLibSizes`).

```
data("PhyloExpressionSetExample")
```

```
# Detection of DEGs using the p-value returned by the Deviance
Deviance.DEGs <- DiffGenes(ExpressionSet = PhyloExpressionSetExample[1:5,1:8],
                       nrep          = 2,
                       method        = "deviance",
                       lib.size      = 1000,
                       stage.names   = c("S1","S2","S3"))

# look at the results
Deviance.DEGs
```

```
  Phylostratum      GeneID    S1<->S2     S1<->S3   S2<->S3
1            1 at1g01040.2 0.26026604 0.110233012 0.6304508
2            1 at1g01050.1 0.95314428 0.598102712 0.6398757
3            1 at1g01070.1 0.55461941 0.456018563 0.8774231
4            1 at1g01080.2 0.58130025 0.487028051 0.8860005
5            1 at1g01090.1 0.03615134 0.001773543 0.2645537
```

Again, users can adjust p-values by specifying the `p.adjust.method` argument.

```
data("PhyloExpressionSetExample")

# Detection of DEGs using the p-value returned by the Deviance Method
# and furthermore, adjust p-values for multiple comparison
# using the Benjamini & Hochberg (1995) method: method = "BH"
# and filter for significantly differentially expressed genes (alpha = 0.05)
Deviance.DEGs.adj <- DiffGenes(ExpressionSet     = PhyloExpressionSetExample[1:5,1:8],
                               nrep              = 2,
                               method            = "deviance",
                               lib.size          = 1000,
                               stage.names       = c("S1","S2","S3"),
                               p.adjust.method = "BH")

# look at the results
Deviance.DEGs.adj
```

```
  Phylostratum      GeneID   S1<->S2     S1<->S3   S2<->S3
1            1 at1g01040.2 0.6506651 0.275582530 0.8860005
2            1 at1g01050.1 0.9531443 0.598102712 0.8860005
3            1 at1g01070.1 0.7266253 0.598102712 0.8860005
4            1 at1g01080.2 0.7266253 0.598102712 0.8860005
5            1 at1g01090.1 0.1807567 0.008867715 0.8860005
```

Replicate Quality Check

Users can also perform replicate quality checks to quantify the variability between replicate expression levels fo each stage separately.

The `PlotReplicateQuality()` is designed to perform customized replicate variablity checks for any `ExpressionSet` object storing replicates.

```
data(PhyloExpressionSetExample)

# visualize the sd() between replicates
PlotReplicateQuality(ExpressionSet = PhyloExpressionSetExample[ , 1:8],
                     nrep          = 2,
                     legend.pos   = "topright",
                     ylim          = c(0,0.2),
                     lwd           = 6)
```

The resulting plot visualizes the kernel density estimates for the variance (log variance) between replicates. Each curve represents the density function for the replicate variation within one stage or experiment. In this case the variance between replicates of `Stage 1` to `Stage 3` (each including 2 replicates) seem to deviate from each other allowing the conclusion that each stage has a different expression level variability between replicates.

The `FUN` argument implemented in `PlotReplicateQuality()` allows users to furthermore, specify customized criteria quantifying replicate varibility. Please notice that the function specified in `FUN` will be performed separately on each gene and stage.

In the following example the median bbsolute deviation function `mad()` is used to quantify replicate variability.

```
data(PhyloExpressionSetExample)

# visualize the mad() between replicates
PlotReplicateQuality(ExpressionSet = PhyloExpressionSetExample[ , 1:8],
                     nrep          = 2,
                     FUN           = mad,
                     legend.pos    = "topright",
                     ylim          = c(0,0.015),
                     lwd           = 6)
```

In general, users are not limited to specific functions implememnted in R. By writing customized functions such as `FUN = function(x) return((x - mean(x))^2)` users can define their own criteria to quantify replicate variability and can then apply this criteria to `PlotReplicateQuality()` by specifying the `FUN` argument.

Collapsing Replicate Samples

After performing differential gene expression analyses, replicate expression levels are collapsed to a single stage specific expression level. For this purpose, `myTAI` implements the `CollapseReplicates()` function, allowing users to combine replicate expression levels stored in a standard `PhyloExpressionSet` or `DivergenceExpressionSet` object to a stage specific expression level using a specified window function.

```
library(myTAI)

# load example data
data(PhyloExpressionSetExample)

# genrate an example PhyloExpressionSet with replicates
ExampleReplicateExpressionSet <- PhyloExpressionSetExample[ ,1:8]

# rename stages
names(ExampleReplicateExpressionSet)[3:8] <- c("Stage_1_Repl_1","Stage_1_Repl_2",
                                               "Stage_2_Repl_1","Stage_2_Repl_2",
                                                "Stage_3_Repl_1","Stage_3_Repl_2")
# have a look at the example dataset
head(ExampleReplicateExpressionSet, 5)
```

```
  Phylostratum      GeneID Stage_1_Repl_1 Stage_1_Repl_2 Stage_2_Repl_1
1            1 at1g01040.2       2173.635      1911.2001       1152.555
2            1 at1g01050.1       1501.014      1817.3086       1665.309
3            1 at1g01070.1       1212.793      1233.0023        939.200
4            1 at1g01080.2       1016.920       936.3837       1181.338
5            1 at1g01090.1      11424.567     16778.1685      34366.649
```

Now, assume that this example `PhyloExpressionSet` stores three developmental stages and 2 biological replicates for each developmental stage. Of course, we could now compute and visualize the TAI profile by typing:

```
# visualize the TAI profile over 3 stages of development
# and 2 replicates per stage
PlotPattern(ExpressionSet = ExampleReplicateExpressionSet,
            type          = "l",
            lwd           = 6)
```

Usually, one would expect that variations in replicate values are smaller than variations between developmental stages. In this example however, we constructed replicate values that vary larger than expression levels between developmental stages. For many applications it might be useful to visualize TAI/TDI values of replicates as well, but normally replicate values are collapsed to one gene and stage specific value after differential gene expression analyses and replicate quality control have been performed.

The following example illustrates how to collapse replicates with `CollapseReplicates()`:

```
# combine the expression levels of the 2 replicates (const) per stage
# using geom.mean as window function and rename new stages to: "S1","S2","S3"
CollapssedPhyloExpressionSet <- CollapseReplicates(
                                        ExpressionSet = ExampleReplicateExpressionSet,
                                        nrep          = 2,
                                        FUN           = geom.mean,
                                        stage.names   = c("S1","S2","S3"))

# have a look at the collpased PhyloExpressionSet
head(CollapssedPhyloExpressionSet)
```

```
   Phylostratum      GeneID         S1         S2         S3
1            1 at1g01040.2  2038.1982  1220.0147   981.4381
2            1 at1g01050.1  1651.6070  1614.2524  1291.4582
3            1 at1g01070.1  1222.8557   934.3975   870.6878
4            1 at1g01080.2   975.8215  1253.2189  1339.2866
5            1 at1g01090.1 13844.9740 36972.3612 61371.0937
6            1 at1g01120.1   815.3288   894.8987   905.8272
```

The `nrep` argument specifies either a constant number of replicates per stage or a numeric vector storing variable numbers of replicates for each developmental stage. In our example, each developmental stage had a constant (equal) number of replicates per developmental stage (`nrep = 2`). In case a variable stage specific number of replicates is present, one could specify `nrep = c(2,3,2)` defining the case that developmental stage 1 stores 2 replicates, stage 2 stores 3 replicates, and stage 3 again, stores 2 replicates.

The argument `FUN` specifies the window function to collapse replicate expression levels to a single stage specific value. In this example, we chose the `geom.mean()` (geometric mean) function implemented in `myTAI`, because our example `PhyloExpressionSet` stores absolute expression levels. Notice that the mathematical equivalent of performing arithmetic mean (`mean()`) computations on `log` expression levels is to perform the geometric mean (`geom.mean()`) on absolute expression levels.

The `stage.names` argument then specifies the new names of collapsed stages.

Filter for Expressed Genes

After differential gene expression analyses and replicate aggregation have been performed, some studies filter gene expression levels in RNA-Seq count tables or microarray expression matrices for non-expressed or outlier genes. For example, in most studies performing RNA-Seq experiments FPKM/RPKM values < 1 are remove from the processed (final) count table.

For this purpose `myTAI` implements the `Expressed()` function to filter (remove) expression levels in RNA-Seq count tables or microarray expression matrices which do not pass a defined expression threshold.

The `Expressed()` function takes a standard `PhyloExpressionSet` or `DivergenceExpressionSet` object storing a RNA-Seq count table (CT) or microarray gene expression matrix and removes genes from this count table or gene expression matrix that have an expression level below a defined `cut.off` value.

`Expressed()` allows users to choose from several gene extraction methods (see `?Expressed` for details):

- `const`: all genes that have at least one stage that undercuts or exceeds the expression `cut.off` will be excluded from the `ExpressionSet`. Hence, for a 7 stage `ExpressionSet` genes passing the expression level `cut.off` in 6 stages will be retained in the `ExpressionSet`.
- `min-set`: genes passing the expression level `cut.off` in `ceiling(n/2)` stages will be retained in the `ExpressionSet`, where `n` is the number of stages in the `ExpressionSet`.
- `n-set`: genes passing the expression level `cut.off` in `n` stages will be retained in the `ExpressionSet`. Here, the argument `n` is defining the number of stages for which the threshold criteria should be fulfilled.

```r
# check number of genes in PhyloExpressionSetExample
nrow(PhyloExpressionSetExample)
#> [1] 25260

# remove genes that have an expression level below 8000
# in at least one developmental stage
FilterConst <- Expressed(ExpressionSet = PhyloExpressionSetExample,
                         cut.off       = 8000,
                         comparison    = "below",
                         method        = "const")

nrow(FilterConst) # check number of retained genes
#> [1] 449
```

Users will observe that only 449 out of 25260 genes in `PhyloExpressionSetExample` have an absolute expression level above `8000` when omitting genes using `method = 'const'`. The argument `comparison` specifies whether genes having expression levels below, above, or below AND above (both) the `cut.off` value should be removed from the dataset.

The following comparison methods can be selected:

- `comparison = "below"`: define genes as not expressed which undercut the `cut-off` threshold.

- `comparison = "above"`: define genes as outliers which exceed the `cut-off` threshold.
- `comparison = "both"`: remove genes fulfilling the `comparison = "below"` **AND** `comparison = "above"` criteria.

```
# again: check number of genes in PhyloExpressionSetExample
nrow(PhyloExpressionSetExample)
#> [1] 25260

# remove genes that have an expression level above 12000
# in at least one developmental stage (outlier removal)
FilterConst.above <- Expressed(ExpressionSet = PhyloExpressionSetExample,
                               cut.off       = 12000,
                               comparison    = "above",
                               method        = "const")

nrow(FilterConst.above) # check number of retained genes
#> [1] 23547
```

For this example 25260 - 23547 = 1713 have been classified as outliers (expression levels above 12000) and were removed from the dataset.

```
# again: check number of genes in PhyloExpressionSetExample
nrow(PhyloExpressionSetExample)
#> [1] 25260

# remove genes that have an expression level below 8000 AND above 12000
# in at least one developmental stage (non-expressed genes AND outlier removal)
FilterConst.both <-  Expressed(ExpressionSet = PhyloExpressionSetExample,
                               cut.off       = c(8000,12000),
                               comparison    = "both",
                               method        = "const")

nrow(FilterConst.both) # check number of retained genes
#> [1] 2
```

When selecting `comparison = 'both'`, the `cut.off` argument receives 2 threshold values: the *below* `cut.off` as first element and the *above* `cut.off` as second element. In this case `cut.off = c(8000,12000)`. Here, only 2 genes fulfill these criteria.

Analogously, users can specify the number of stages that should fulfill the threshold criteria using the `n-set` method.

```
# remove genes that have an expression level below 8000
# in at least 5 developmental stages (in this case: n = 2 stages fulfilling the criteria)
FilterNSet <- Expressed(ExpressionSet = PhyloExpressionSetExample,
                        cut.off       = 8000,
                        method        = "n-set",
                        comparison    = "below",
                        n             = 2)

nrow(FilterMinSet) # check number of retained genes
#> [1] 20
```

Here, 20 genes are fulfilling these criteria.

Compute the Statistical Significance of Each Replicate Combination

In some cases (high variability of replicates) it might be useful to verify that there is no sequence of replicates (for all possible combination of replicates) that results in a non-significant `TAI` or `TDI` pattern, when the initial pattern with combined replicates was shown to be significant.

The `CombinatorialSignificance()` function implemented in `myTAI` allows users to compute the p-values quantifying the statistical significance of the underlying pattern for all combinations of replicates.

A small Example:

Assume a `PhyloExpressionSet` stores 3 developmental stages with 3 replicates measured for each stage. The 9 replicates in total are denoted as: $1.1, 1.2, 1.3, 2.1, 2.2, 2.3, 3.1, 3.2, 3.3$. Now the function computes the statistical significance of each pattern derived by the corresponding combination of replicates, e.g.

- 1.1, 2.1, 3.1 : p-value for combination 1
- 1.1, 2.2, 3.1 : p-value for combination 2
- 1.1, 2.3, 3.1 : p-value for combination 3
- 1.2, 2.1, 3.1 : p-value for combination 4
- 1.2, 2.1, 3.1 : p-value for combination 5
- 1.2, 2.1, 3.1 : p-value for combination 6
- 1.3, 2.1, 3.1 : p-value for combination 7
- 1.3, 2.2, 3.1 : p-value for combination 8
- 1.3, 2.3, 3.1 : p-value for combination 9
- ...

This procedure yields 27 p-values for the 3^3 (n^m) replicate combinations, where n denotes the number of developmental stages and m denotes the number of replicates per stage.

Note that in case users have a large amount of stages/experiments and a large amount of replicates the computation time will increase by n^m. For 11 stages and 4 replicates, $4^{11} = 4194304$ p-values have to be computed. Each p-value computation itself is based on a permutation test running with $1,000, 10,000, ...$ or more permutations. Be aware that this might take some time.

The p-value vector returned by this function can then be used to plot the p-values to see whether an critical value α is exceeded or not (e.g. $\alpha = 0.05$).

```r
# load a standard PhyloExpressionSet
data(PhyloExpressionSetExample)

# we assume that the PhyloExpressionSetExample
# consists of 3 developmental stages
# and 2 replicates for stage 1, 3 replicates for stage 2,
# and 2 replicates for stage 3
# FOR REAL ANALYSES PLEASE USE: permutations = 1000 or 10000
# BUT NOTE THAT THIS TAKES MUCH MORE COMPUTATION TIME
p.vector <- CombinatorialSignificance(ExpressionSet = PhyloExpressionSetExample,
                                      replicates    = c(2,3,2),
                                      TestStatistic = "FlatLineTest",
                                      permutations  = 100,
                                      parallel      = FALSE)
```

```
 [1] 2.436296e-03 2.288593e-02 1.608399e-03 1.185615e-02 1.835306e-06 1.077012e-05
 [7] 2.025515e-07 5.148342e-07 1.654885e-07 6.251145e-06 9.265520e-10 1.047479e-06
```

Users will observe that none of the replicate combinations resulted in p-values > 0.05 and thus we can assume that the phylotranscriptomic pattern computed based on collapsed replicates is not biased by insignificant replicate combinations.

```
any(p.vector > 0.05)
#> FALSE
```

`CombinatorialSignificance()` can perform all significance tests introduced in the `Introduction`.

Furthermore, the `parallel` argument allows users to perform significance computations in parallel on a multicore machine. This will speed up p-value computations for a large number of combinations.

Performing `Phylostratum` and `Divergence Stratum` Enrichment Analyses

`Phylostratum` and `Divergence Stratum` enrichment analyses have been performed by several studies to correlate organ or metabolic pathway evolution with the origin of organ or pathway specific genes (Sestak and Domazet-Loso, 2015).

In detail, `Phylostratum` and `Divergence Stratum` enrichment analyses can be performed analogously to Gene Ontology and Kegg enrichment analyses to study the enrichment of evolutionary age or sequence divergence in a set of selected genes against the entire genome/transcriptome. In case specific age categories are significantly over- or underrepresented in the selected gene set, assumptions or potential correlations between the evolutionary origin of a particular organ or metabolic pathway can be implied.

In this vignette we will use the data set published by Sestak and Domazet-Loso, 2015 to demonstrate how to perform enrichment analyses using `myTAI`.

Enrichment Analyses using `PlotEnrichment()`

The `PlotEnrichment()` function implemented in `myTAI` computes and visualizes the significance of enriched (over- or underrepresented) `Phylostrata` or `Divergence Strata` within an input set of process/tissue specific genes. In detail this function takes the `Phylostratum` or `Divergence Stratum` distribution of all genes stored in the input `ExpressionSet` as background set and the `Phylostratum` or `Divergence Stratum` distribution of the specific gene set and performs a Fisher's exact test for each `Phylostratum` or `Divergence Stratum` to quantify the statistical significance of over- or under-represented `Phylostrata` or `Divergence Strata` within the set of selected genes. In other words, the frequency distribution of `Phylostrata` or `Divergence Strata` in the complete sample is compared with the frequency distribution of `Phylostrata` or `Divergence Strata` in the set of selected genes and over- or under-representation is visualized by log-odds (or odds), where a log-odd of zero means that both frequency distributions are equal (see also Sestak and Domazet-Loso, 2015).

Example Data Set Retrieval

Before using the `PlotEnrichment()` function, we need to download the example data set from Sestak and Domazet-Loso, 2015.

Download the `Phylostratigraphic Map` of *D. rerio*:

```
# download the Phylostratigraphic Map of Danio rerio
# from Sestak and Domazet-Loso, 2015
download.file(url      = "http://mbe.oxfordjournals.org/
              content/suppl/2014/11/17/msu319.DC1/TableS3-2.xlsx",
              destfile = "MBE_2015a_Drerio_PhyloMap.xlsx")
```

Read the `*.xlsx` file storing the `Phylostratigraphic Map` of *D. rerio* and format it for the use with `myTAI`:

```
# install the readxl package
install.packages("readxl")

# load package readxl
library(readxl)

# read the excel file
DrerioPhyloMap.MBEa <-
        read_excel("MBE_2015a_Drerio_PhyloMap.xlsx",
        sheet = 1,
        skip = 4)

# format Phylostratigraphic Map for use with myTAI
Drerio.PhyloMap <- DrerioPhyloMap.MBEa[ , 1:2]

# have a look at the final format
head(Drerio.PhyloMap)
```

```
  Phylostrata            ZFIN_ID
1           1 ZDB-GENE-000208-13
2           1 ZDB-GENE-000208-17
3           1 ZDB-GENE-000208-18
4           1 ZDB-GENE-000208-23
5           1  ZDB-GENE-000209-3
6           1  ZDB-GENE-000209-4
```

Now, `Drerio.PhyloMap` stores the `Phylostratigraphic Map` of *D. rerio* which is used as background set to perform enrichment analyses with `PlotEnrichment()`.

Enrichment Analyses

Now, the `PlotEnrichment()` function visualizes the over- and underrepresented `Phylostrata` of brain specific genes when compared with the total number of genes stored in the `Phylostratigraphic Map` of *D. rerio*.

```
# read expression data (organ specific genes) from Sestak and Domazet-Loso, 2015
Drerio.OrganSpecificExpression <- read_excel("MBE_2015a_Drerio_PhyloMap.xlsx", sheet = 2, skip = 3)

# select only brain specific genes
Drerio.Brain.Genes <- unique(na.omit(Drerio.OrganSpecificExpression[ , "brain"]))

# visualize enriched Phylostrata of genes annotated as brain specific
PlotEnrichment(Drerio.PhyloMap,
               test.set     = Drerio.Brain.Genes,
               measure      = "log-foldchange",
               use.only.map = TRUE,
               legendName   = "PS")
```

Here, the first argument is either a standard `ExpressionSet` object (in case `use.only.map = FALSE`: default) or a `Phylostratigraphic Map` or `Divergence Map` (in case `use.only.map = TRUE`; see Introduction for details). The second argument `test.set` specifies the gene ids also stored in the corresponding `ExpressionSet` or `Phylostratigraphic Map/ Divergence Map` for which enrichment shall be quantified and visualized.

To visualize the odds or log-odds of over- or underrepresented genes within the `test.set` the following procedure is performed:

- N_{ij} denotes the number of genes in group j and deriving from PS i, with $i = 1, .., n$ and where $j = 1$ denotes the background set and $j = 2$ denotes the `test.set`
- $N_{i.}$ denotes the total number of genes within PS i
- $N_{.j}$ denotes the total number of genes within group j
- $N_{..}$ is the total number of genes within all groups j and all PS i
- $f_{ij} = N_{ij}$ / $N_{..}$ and $g_{ij} = f_{ij}$ / $f_{.j}$ denote relative frequencies between groups
- $f_{i.}$ denotes the between group sum of f_{ij}

The result is the **fold-change value** (odds; `measure = "foldchange"`) denoted as $C_2 = g_{i2}/f_{i.}$ which is visualized above and below zero or the **log fold-change** value (log-odds; `measure = "log-foldchange"`), where log_2(C) = $log_2(g_{i2})$ - $log_2(f_{i.})$ which is visualized symmetrically above and below zero by `PlotEnrichment()`. Analogously, $C_1 = g_{i1}/f_{i.}$ but is not visualized by this function.

Internally, `PlotEnrichment()` performs a Fisher's exact test for each `Phylostratum` or `Divergence Stratum` separately, to quantify the significance of over- or under-representation of corresponding `Phylostrata` or `Divergence Strata` within the `test.set` when compared with the entire `ExpressionSet`. `PlotEnrichment()` visualizes significantly enriched (over- or underrepresented) `Phylostrata` or `Divergence Strata` with asterisks '*'.

Notation:

- '*' = P-Value $\leq$ 0.05
- '**' = P-Value $\leq$ 0.005
- '***' = P-Value $\leq$ 0.0005

Users will notice that when performing the `PlotEnrichment()` function, the p-values and the enrichment matrix (storing C_1 and C_2) will be returned.

```
PlotEnrichment(Drerio.PhyloMap,
               test.set     = Drerio.Brain.Genes,
               measure      = "log-foldchange",
               use.only.map = TRUE,
               legendName   = "PS")
```

```
$p.values
         PS1          PS2          PS3          PS4          PS5          PS6 
8.283490e-01 8.362880e-05 6.778981e-02 1.373816e-02 7.946309e-13 6.017041e-01 
         PS7          PS8          PS9         PS10         PS11         PS12 
2.185021e-03 2.281194e-03 8.943147e-01 5.699612e-01 4.717058e-02 9.367759e-01 
```

```
        PS13          PS14
3.929949e-03 1.593834e-05

$enrichment.matrix
           BG_Set     Test_Set
PS1  -0.001132832  0.007668216
PS2   0.023733936 -0.172380714
PS3  -0.040879607  0.250587496
PS4  -0.048920465  0.294399729
PS5  -0.114888949  0.603817643
PS6   0.008678915 -0.060350168
PS7  -0.062948352  0.367240944
PS8   0.115630474 -1.206210187
PS9  -0.007353969  0.048964218
PS10 -0.031971192  0.200141519
PS11  0.039742253 -0.303363314
PS12 -0.002418079  0.016311853
PS13  0.101449988 -0.984621732
PS14  0.098211044 -0.938724783
```

In case users are only interested in the p-values of the Fisher test and the enrichment matrix without illustrating the bar plot, they can specify the `plot.bars = FALSE` argument to only retrieve the numeric results.

```
# specify plot.bars = FALSE to retrieve only numeric results
EnrichmentResult <- PlotEnrichment(Drerio.PhyloMap,
                                   test.set     = Drerio.Brain.Genes,
                                   measure      = "log-foldchange",
                                   use.only.map = TRUE,
                                   legendName   = "PS",
                                   plot.bars    = FALSE)

# access p-values quantifying the enrichment for each Phylostratum
EnrichmentResult$p.values
```

```
         PS1          PS2          PS3          PS4          PS5          PS6
8.283490e-01 8.362880e-05 6.778981e-02 1.373816e-02 7.946309e-13 6.017041e-01
         PS7          PS8          PS9         PS10         PS11         PS12
2.185021e-03 2.281194e-03 8.943147e-01 5.699612e-01 4.717058e-02 9.367759e-01
        PS13         PS14
3.929949e-03 1.593834e-05
```

```
# access enrichment matrix storing C_1 and C_2
EnrichmentResult$enrichment.matrix
```

```
           BG_Set     Test_Set
PS1  -0.001132832  0.007668216
PS2   0.023733936 -0.172380714
PS3  -0.040879607  0.250587496
PS4  -0.048920465  0.294399729
PS5  -0.114888949  0.603817643
PS6   0.008678915 -0.060350168
PS7  -0.062948352  0.367240944
```

```
PS8   0.115630474 -1.206210187
PS9  -0.007353969  0.048964218
PS10 -0.031971192  0.200141519
PS11  0.039742253 -0.303363314
PS12 -0.002418079  0.016311853
PS13  0.101449988 -0.984621732
PS14  0.098211044 -0.938724783
```

Defining the Background Set

The Fisher test which is performed inside `PlotEnrichment()` assumes that all genes stored in the input `ExpressionSet` or `Phylostratigraphic Map/Divergence Map` are used to define the background set for constructing the test statistic. However, since in most cases the `test.set` is an subset of the input `ExpressionSet` or `Phylostratigraphic Map/Divergence Map` one could also specify the `complete.bg` argument to remove all `test.set` genes from the background set when performing the Fisher test and visualization.

The following two examples allow users to compare the results when retaining all genes as background set compared with the option to remove `test.set` genes from the background set.

```
# complete.bg = TRUE (default) -> retain test.set genes in background set
PlotEnrichment(Drerio.PhyloMap,
               test.set     = Drerio.Brain.Genes,
               measure      = "log-foldchange",
               complete.bg  = TRUE,
               use.only.map = TRUE,
               legendName   = "PS")
```

```
# complete.bg = FALSE -> remove test.set genes from background set
PlotEnrichment(Drerio.PhyloMap,
               test.set     = Drerio.Brain.Genes,
               measure      = "log-foldchange",
               complete.bg  = FALSE,
               use.only.map = TRUE,
               legendName   = "PS")
```

Users will notice that although some p-values change, the qualitative result did not change. In border line cases however, the results might influence whether or not some `Phylostrata` or `Divergence Strata` are denoted as significantly enriched or not. So always be aware of the interpretation when retaining or removing the `test.set` from the background set, because both options are valid and have advantages and disadvantages and depend on a valid interpretation.

Interpretation of Enrichment Results

For the *D. rerio* brain genes example you can see that PS4, PS5, and PS7 are significantly over-represented in the set of brain specific genes.

```
PlotEnrichment(Drerio.PhyloMap,
               test.set     = Drerio.Brain.Genes,
               measure      = "foldchange",
               complete.bg  = TRUE,
               use.only.map = TRUE,
               legendName   = "PS")
```

Again, we retrieve the *D. rerio* specific taxonomy represented by PS1-14 using the `taxonomy()` funtion (see Introduction and Taxonomy for details).

```r
# retrieve the taxonomy of D. rerio
taxonomy(organism = "Danio rerio")
```

```
                    id         name    rank
1   cellular organisms      no rank  131567
2            Eukaryota superkingdom    2759
3         Opisthokonta      no rank   33154
4              Metazoa      kingdom   33208
5            Eumetazoa      no rank    6072
6            Bilateria      no rank   33213
7        Deuterostomia      no rank   33511
8             Chordata       phylum    7711
9             Craniata    subphylum   89593
10          Vertebrata      no rank    7742
11       Gnathostomata      no rank    7776
12          Teleostomi      no rank  117570
13        Euteleostomi      no rank  117571
14      Actinopterygii   superclass    7898
15        Actinopteri        class  186623
16        Neopterygii     subclass   41665
17          Teleostei   infraclass   32443
18 Osteoglossocephalai      no rank 1489341
19       Clupeocephala      no rank  186625
20           Otomorpha      no rank  186634
21        Ostariophysi      no rank   32519
22            Otophysa      no rank  186626
23       Cypriniphysae   superorder  186627
24       Cypriniformes        order    7952
25         Cyprinoidea  superfamily   30727
26          Cyprinidae       family    7953
27               Danio        genus    7954
28         Danio rerio      species    7955
```

Sestak and Domazet-Loso, 2015 collapsed these 28 taxonomic nodes into 14 taxonomic nodes (see `Figure 2` in Sestak and Domazet-Loso, 2015) and labelled them as phylostrata 1 to phylostrata 14, where PS1 represents `cellular organisms` and PS14 represents `D. rerio` specific genes. Based on the phylostratum categorization of Sestak and Domazet-Loso, 2015, PS4 represents `Holozoa (= Metazoa + allies)`, PS5 represents `Metazoa`, and PS7 represents `Bilateria`.

Now, the over-representation results of brain specific genes returned by `PlotEnrichment()` provide evidence, that brain specific genes might indeed have originated during the emergence of the nervous system at the metazoan-eumetazoan transition leading to the interpretation that the vertebrate brain has a step wise adaptive history where most of its extant organization was already present in the chordate ancestor as argued by Sestak and Domazet-Loso, 2015.

This example shall illustrate how the `PlotEnrichment()` function can be used to trace the evolutionary origin of tissue or process specific genes by investigating their age enrichment.

In case users have an `ExpressionSet` storing the `Phylostratigraphic Map` of *D. rerio* as well as an expression set, they can furthermore use the `PlotGeneSet()` function implemented in `myTAI` to visualize the expression levels of brain specific genes which have been shown to be significantly enriched in `Metazoa` specific phylostrata.

Example:

```
# the best parameter setting to visualize this plot:
# png("DrerioBrainSpecificGeneExpression.png",700,400)
PlotGeneSet(ExpressionSet = DrerioPhyloExpressionSet,
            gene.set      = Drerio.Brain.Genes,
            plot.legend   = FALSE,
            type          = "l",
            lty           = 1,
            lwd           = 4,
            xlab          = "Ontogeny",
            ylab          = "Expression Level")

# dev.off()
```

Here `DrerioPhyloExpressionSet` denotes a hypothetical `ExpressionSet` of *D. rerio* development.

Additionally, the `SelectGeneSet()` function allows users to obtain the `ExpresisonSet` subset of selected genes (`gene.set`) for subsequent analyses.

```
# select the ExpressionSet subset of Brain specific genes
Brain.PhyloExpressionSet <- SelectGeneSet( ExpressionSet = DrerioPhyloExpressionSet,
                                           gene.set      = Drerio.Brain.Genes )

head(Brain.PhyloExpressionSet)
```

Adjust P-values for Multiple Comparisons

In case a large number of Phylostrata or Divergence Strata is included in the input `ExpressionSet`, p-values returned by `PlotEnrichment()` should be adjusted for multiple comparisons. For this purpose `PlotEnrichment()` includes the argument `p.adjust.method`. Here, all methods implemented in `?p.adjust` can be specified:

```
# adjust p-values for multiple comparisons with Benjamini & Hochberg (1995)
PlotEnrichment(Drerio.PhyloMap,
               test.set        = Drerio.Brain.Genes,
               measure         = "log-foldchange",
               complete.bg     = FALSE,
               use.only.map    = TRUE,
               legendName      = "PS",
               p.adjust.method = "BH")
```

Please consult these reviews (Biostatistics Handbook, Gelman et al., 2008, and Slides) to decide whether or not to apply p-value adjustment to your own dataset.

Combine Functional Annotation with Enrichment Analyses

The greatest advantage of `Phylostratum` and `Divergence Stratum` enrichment analyses can be unfolded when it is based on the functional annotation of the `test.set`. The Functional Annotation and Phylotranscriptomics vignettes of our biomartr package provide detailed tutorials on how to retrieve functional annotation for a subset of genes. Categorizing genes into common functional groups or processes via `biomartr` will then allow users to search for genes deriving from enriched `Phylostrata` or `Divergence Strata`. The Introduction vignette illustrated how `Phylostrata` or `Divergence Strata` can be linked with the origin of evolutionary events. This correlation between the predicted origin of genes (`Phylostratigraphic Map`) or its current state of selection between closely related species (`Divergence Map`) and their functional annotation furthermore allows users to detect signals of potential evolutionary origins of specific biological functions, processes, or tissues and their active maintenance between closely related species.

Investigating Age or Divergence Category Specific Expression Level Distributions

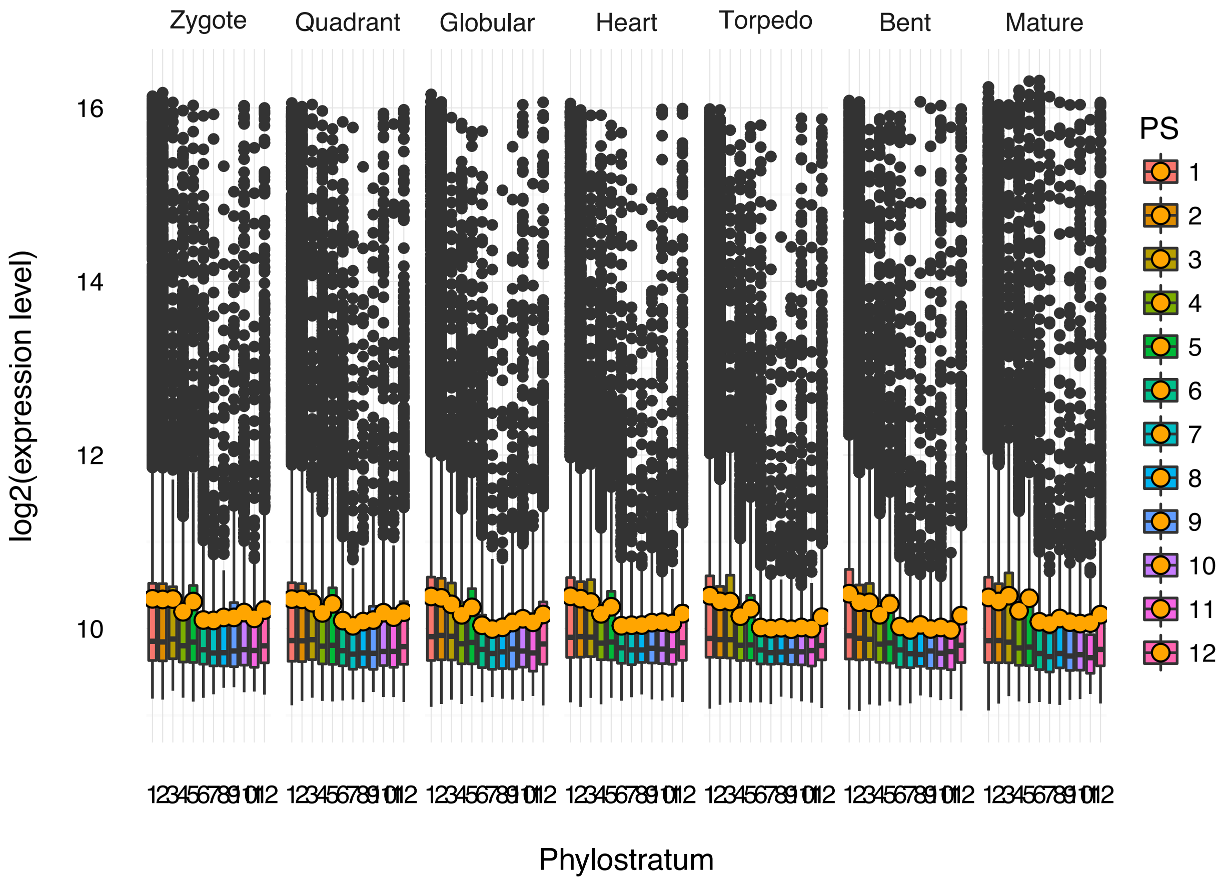
Gene expression levels are a fundamental aspect of phylotranscriptomics studies. In detail, phylotranscriptomic measures aim to quantify the expression intensity of genes deriving from common age or divergence categories to detect stages of evolutionary constraints. Hence, the gene expression distribution of age or divergence categories as well as their differences within and between stages or categories allow us to investigate the age (PS) or divergence (DS) category specific contribution to the corresponding transcriptome.

For this purpose, the `PlotCategoryExpr()` aims to visualize the expression level distribution of each phylostratum during each time point or experiment as barplot, dot plot, or violin plot enabling users to quantify the age (PS) or divergence (DS) category specific contribution to the corresponding transcriptome.

This way of visualizing the gene expression distribution of each age (PS) or divergence (DS) category during all developmental stages or experiments allows users to detect specific age or divergence categories contributing significant levels of gene expression to the underlying biological process (transcriptome).

```r
library(myTAI)

data(PhyloExpressionSetExample)

# category-centered visualization of PS specific expression level distributions (log-scale)
PlotCategoryExpr(ExpressionSet = PhyloExpressionSetExample,
                 legendName    = "PS",
                 test.stat     = TRUE,
                 type          = "category-centered",
                 distr.type    = "boxplot",
                 log.expr      = TRUE)
```

```
#>                   Zygote Quadrant Globular Heart Torpedo Bent  Mature
#> category-centered "***"  "***"    "***"    "***" "***"   "***" "***"
```



```
                  Zygote Quadrant Globular Heart Torpedo Bent  Mature
category-centered "***"  "***"    "***"    "***" "***"   "***" "***"
```

The resulting boxplot illustrates the log expression levels of each phylostratum during each developmental stage. Additionally, a Kruskal-Wallis Rank Sum Test as well as a Benjamini & Hochberg p-value adjustment for multiple comparisons is performed (`test.stat = TRUE`) to statistically quantify the differences between expression levels of different age or divergence categories. This type of analysis allows users to detect stages or experiments that show high diviation between age or divergence category contributions to the overall transcriptome or no significant deviations of age or divergence categories, suggesting equal age or divergence category contributions to the overall transcriptome. The corresponding P-values are printed to the console using the following notation:

- ’*’ = P-Value ≤ 0.05
- ’**’ = P-Value ≤ 0.005
- ’***’ = P-Value ≤ 0.0005
- ‘n.s.’ = not significant = P-Value > 0.05

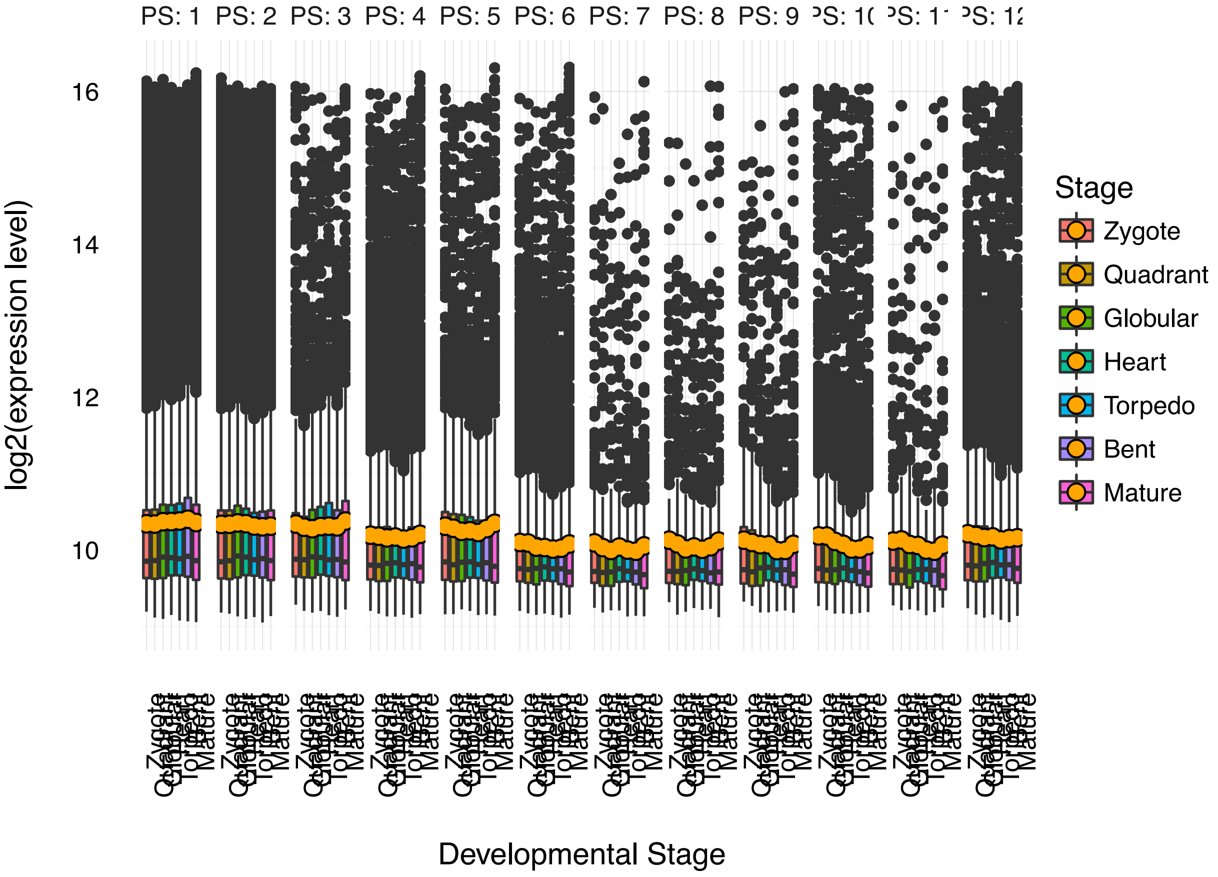
In this case all developmental stages show significant differences in phylostratum specific gene expression.

Please notice that users need to define the `legendName` argument as `PS` or `DS` to specify whether the input `ExpressionSet` is a `PhyloExpressionSet` (`legendName = 'PS'`) or `DivergenceExpressionSet` (`legendName = 'DS'`).

Alternatively, users can investigate the differences of gene expression **between** all stages or experiments for **each** age or divergence category by specifying `type = 'stage-centered'`.

```r
library(myTAI)

data(PhyloExpressionSetExample)

# stage-centered visualization of PS specific expression level distributions (log-scale)
PlotCategoryExpr(ExpressionSet = PhyloExpressionSetExample,
                 legendName    = "PS",
                 test.stat     = TRUE,
                 type          = "stage-centered",
                 distr.type    = "boxplot",
                 log.expr      = TRUE)
```

```
#>                PS1   PS2   PS3    PS4 PS5    PS6 PS7    PS8    PS9    PS10  PS11   PS12
#> stage-centered "***" "***" "n.s." "*" "n.s." "*" "n.s." "n.s." "n.s." "***" "n.s." "***"
```



```
    PS1   PS2   PS3   PS4  PS5    PS6 PS7    PS8    PS9    PS10  PS11   PS12
sc "***" "***" "n.s." "*" "n.s." "*" "n.s." "n.s." "n.s." "***" "n.s." "***"
```

Here, the Kruskal-Wallis Rank Sum Test (with Benjamini & Hochberg p-value adjustment) quantifies whether or not the gene expression distribution of a single age or divergence category significantly changes throughout development or experiments. This type of analysis allows users to detect specific age or divergence categories that significantly change their expression levels throughout development or experiments.

In this case, users will observe that PS3,5,7-9,11 do not show significant differences of gene expression between developmental stages suggesting that their contribution to the overall transcriptome remains constant throughout development.

Finally, users can choose the following plot types to visualize expression distributions:

Argument: `distr.type`

- `distr.type = "boxplot"` This specification allows users to visualize the expression distribution of all PS or DS as boxplot.
- `distr.type = "violin"` This specification allows users to visualize the expression distribution of all PS or DS as violin plot.
- `distr.type = "dotplot"` This specification allows users to visualize the expression distribution of all PS or DS as dot plot.

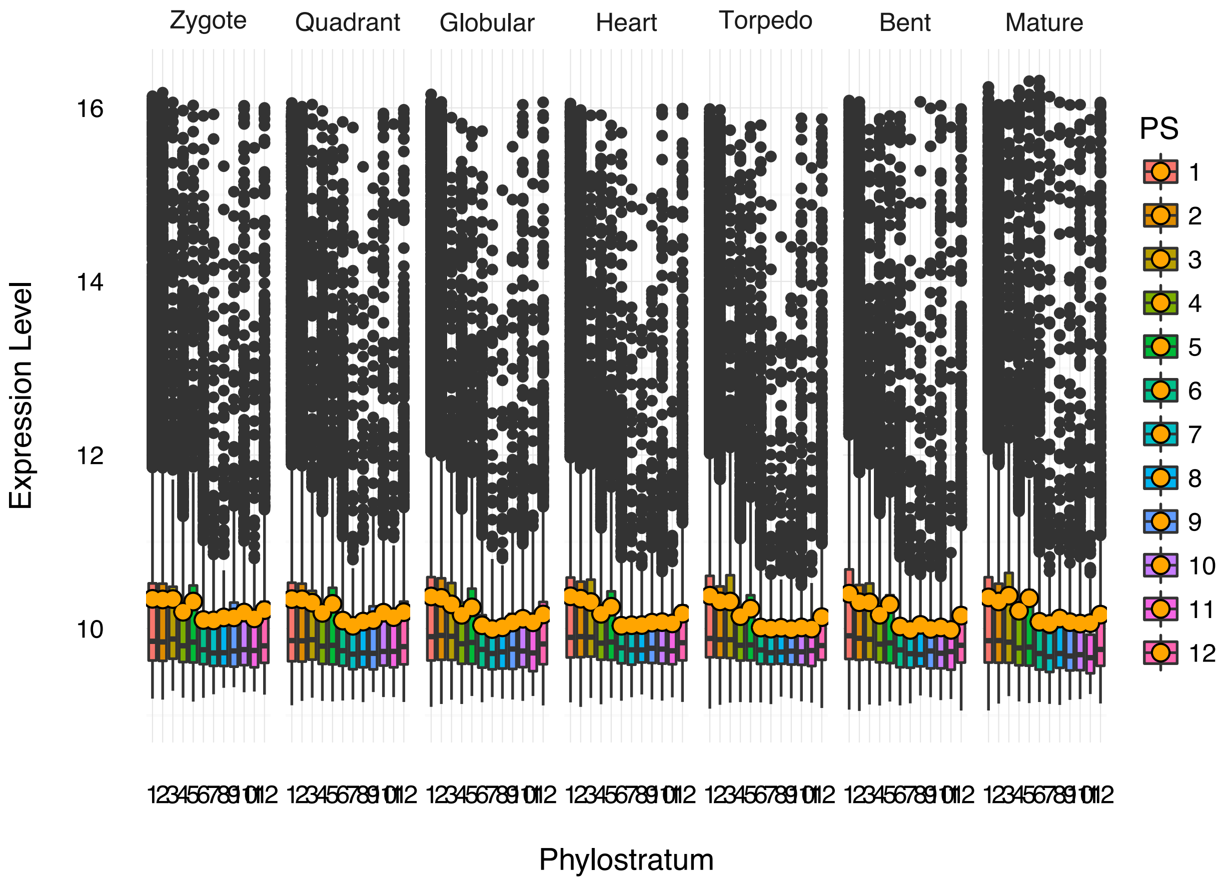
Together, studies perfomed with `PlotCategoryExpr()` allow users to conclude that genes originating in specific PS or DS contribute significantly more to the overall transcriptome than other genes originating from different PS or DS categories. More specialized analyses such as `PlotMeans()`, `PlotRE()`, `PlotBarRE()`, `TAI()`, `TDI()`, etc. will then allow them to study the exact mean expression patterns of these age or divergence categories.

Users will notice that so far all examples shown above specified `log.expr = TRUE` illustrating boxplots based on log2 expression levels. This way of visualization allows better visual comparisons between age or divergence categories. However, when specifying `log.expr = FALSE` absolute expression levels will be visualized in the corresponding boxplot.

Alternatively, instead of specifying `log.expr = TRUE` users can directly pass log2 transformed expression levels to `PlotCategoryExpr()` via `tf(PhyloExpressionSetExample,log2)` (when `log.expr = FALSE`):

```r
data(PhyloExpressionSetExample)

# category-centered visualization of PS specific expression level distributions (log-scale)
PlotCategoryExpr(ExpressionSet = tf(PhyloExpressionSetExample, log2),
                 legendName    = "PS",
                 test.stat     = TRUE,
                 type          = "category-centered",
                 distr.type    = "boxplot",
                 log.expr      = FALSE)
```

```
#>                   Zygote Quadrant Globular Heart Torpedo Bent  Mature
#> category-centered "***"  "***"    "***"    "***" "***"   "***" "***"
```

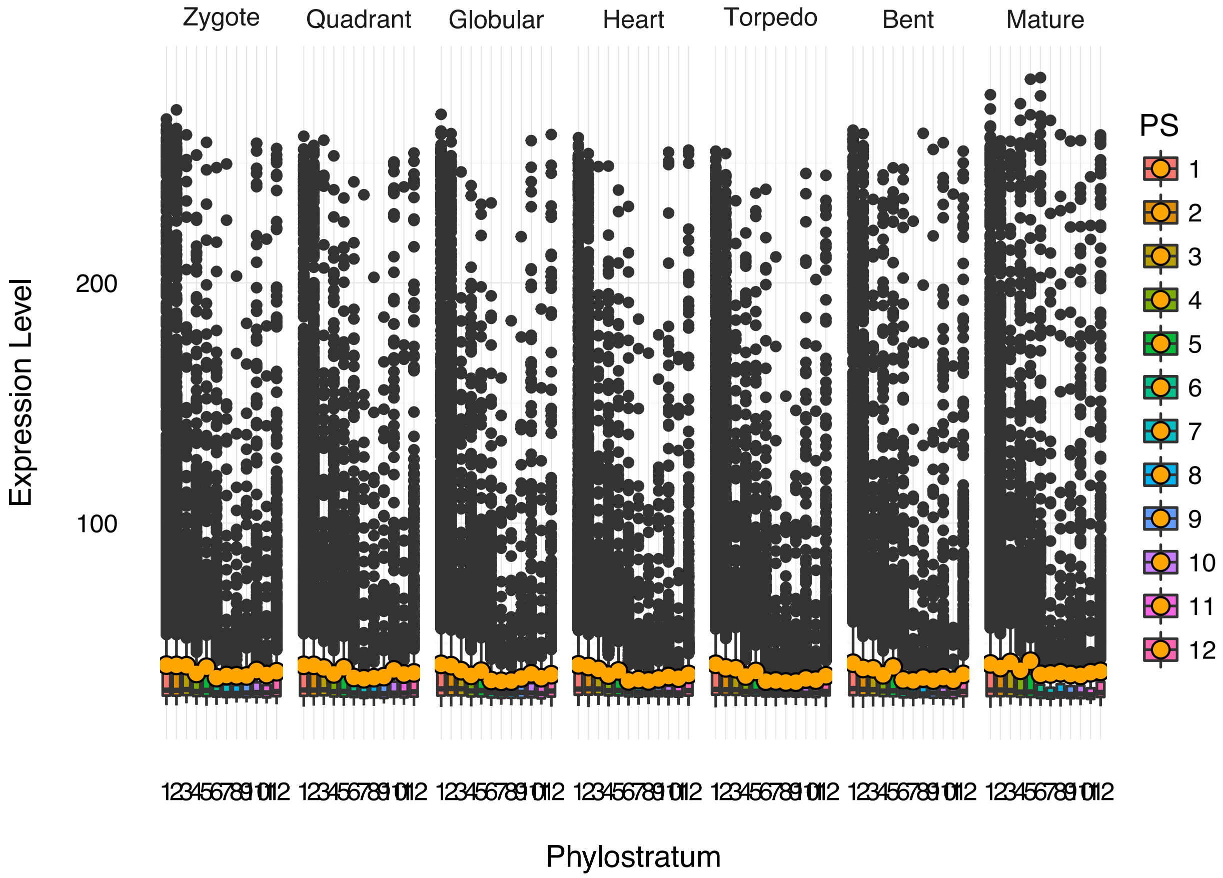


```
                  Zygote Quadrant Globular Heart Torpedo Bent  Mature
category-centered "***"  "***"    "***"    "***" "***"   "***" "***"
```

Or any other expression level transformation, e.g. `sqrt`.

```
data(PhyloExpressionSetExample)

# category-centered visualization of PS specific expression level distributions (sqrt-scale)
PlotCategoryExpr(ExpressionSet = tf(PhyloExpressionSetExample, sqrt),
                 legendName    = "PS",
                 test.stat     = TRUE,
                 type          = "category-centered",
                 distr.type    = "boxplot",
                 log.expr      = FALSE)
```

```
#>                   Zygote Quadrant Globular Heart Torpedo Bent  Mature
#> category-centered "***"  "***"    "***"    "***" "***"   "***" "***"
```



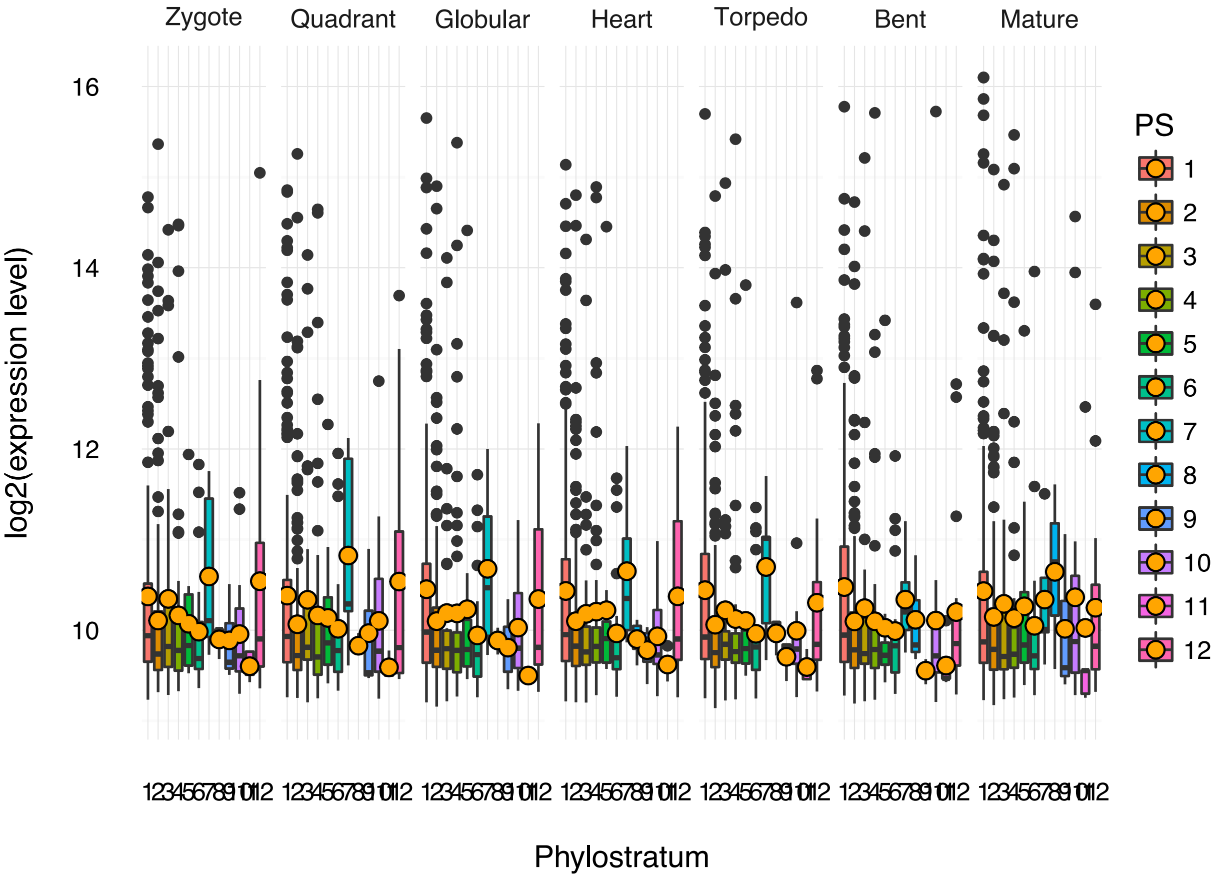
```
                  Zygote Quadrant Globular Heart Torpedo Bent  Mature
category-centered "***"  "***"    "***"    "***" "***"   "***" "***"
```

Gene Subset Age or Divergence Category Specific Expression Level Distributions

In some cases, users wish to visualize the gene expression distributions for a subset of genes in contrast to the entire transcriptome. For this purpose, the `gene.set` argument allows users to specify the gene ids of a subset of genes that shall be matched in the input `ExpressionSet` and for which expression level distributions shall be visualized.

```
library(myTAI)
data(PhyloExpressionSetExample)

# define an example gene subset (500 genes) which
# can be found in the input ExpressionSet
set.seed(234)
example.gene.set <- PhyloExpressionSetExample[sample(1:25260,500) , 2]

# visualize the gene expression distributions for these 500 genes (category-centered)
PlotCategoryExpr(ExpressionSet = PhyloExpressionSetExample,
                 legendName    = "PS",
                 test.stat     = TRUE,
                 type          = "category-centered",
                 distr.type    = "boxplot",
                 log.expr       = TRUE,
                 gene.set       = example.gene.set)
```

```
#>                      Zygote Quadrant Globular Heart Torpedo Bent Mature
#> category-centered "*"    "*"      "*"      "*"   "*"     "*"  "n.s."
#>
#> # Genes:
#>   1   2   3   4   5   6   7   8   9  10  11  12
#> 193 137  30  52  13  22   5   3   3  20   5  17
```



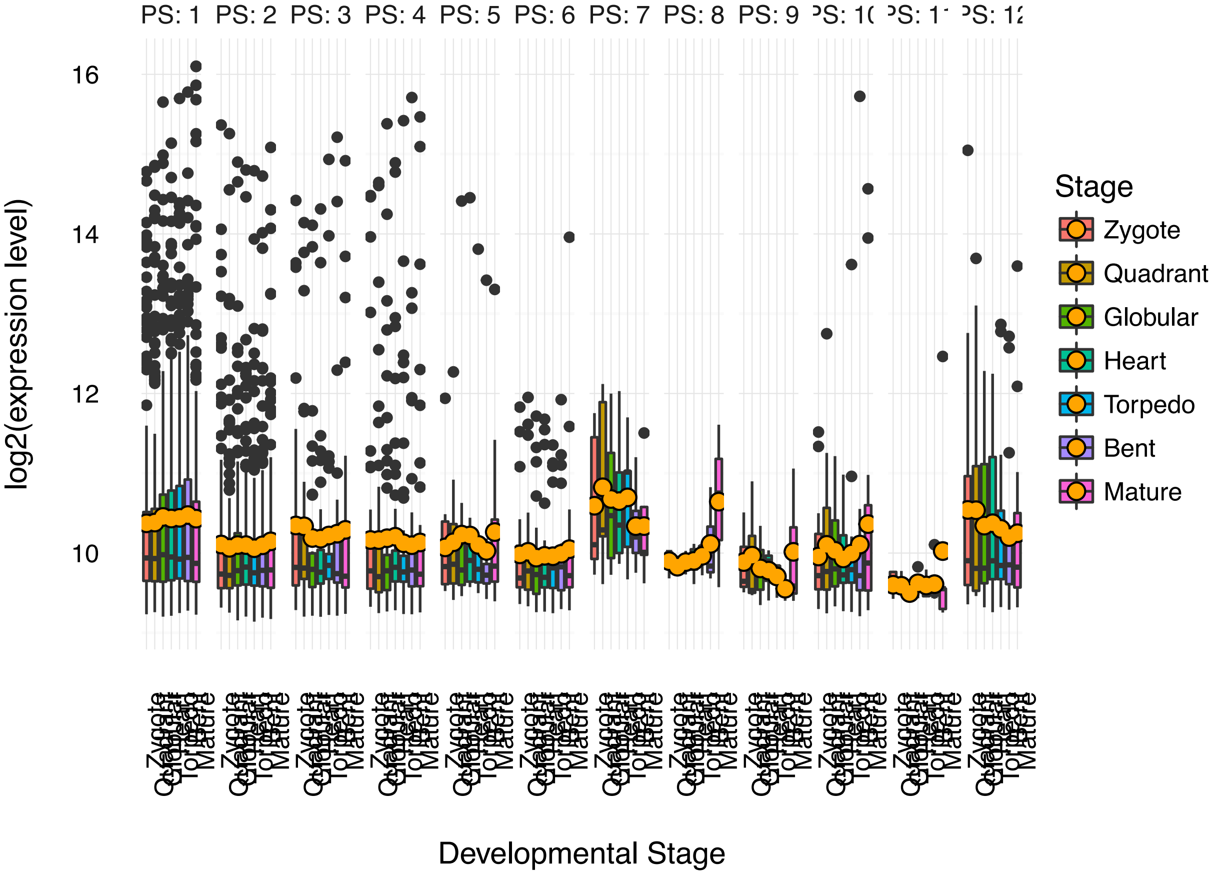
```
                  Zygote Quadrant Globular Heart Torpedo Bent Mature
category-centered "*"    "*"      "*"      "*"   "*"     "*"  "n.s."
```

Or analogously `stage-centered`:

```
library(myTAI)

data(PhyloExpressionSetExample)

# define an example gene subset (500 genes) which
# can be found in the input ExpressionSet
set.seed(234)
example.gene.set <- PhyloExpressionSetExample[sample(1:25260,500) , 2]
```

```r
# visualize the gene expression distributions for these 500 genes (stage-centered)
PlotCategoryExpr(ExpressionSet = PhyloExpressionSetExample,
                 legendName    = "PS",
                 test.stat     = TRUE,
                 type          = "stage-centered",
                 distr.type    = "boxplot",
                 log.expr      = TRUE,
                 gene.set      = example.gene.set)
```

```
#>                 PS1    PS2    PS3    PS4    PS5    PS6    PS7    PS8    PS9    PS10   PS11   PS12
#> stage-centered "n.s." "n.s." "n.s." "n.s." "n.s." "n.s." "n.s." "n.s." "n.s." "n.s." "n.s." "n.s."
#>
#> # Genes:
#>   1   2   3   4   5   6   7   8   9  10  11  12
#> 193 137  30  52  13  22   5   3   3  20   5  17
```



```
               PS1    PS2    PS3    PS4    PS5    PS6    PS7    PS8    PS9    PS10   PS11
stage-centered "n.s." "n.s." "n.s." "n.s." "n.s." "n.s." "n.s." "n.s." "n.s." "n.s." "n.s."
               PS12
stage-centered "n.s."
```

For example, users interested in the enrichment of PS or DS values in *D. rerio* brain genes (see Enrichment Vignette for details) could also visualize their gene expression distributions throughout development with `PlotCategoryExpr()` in cases where expression data is available.

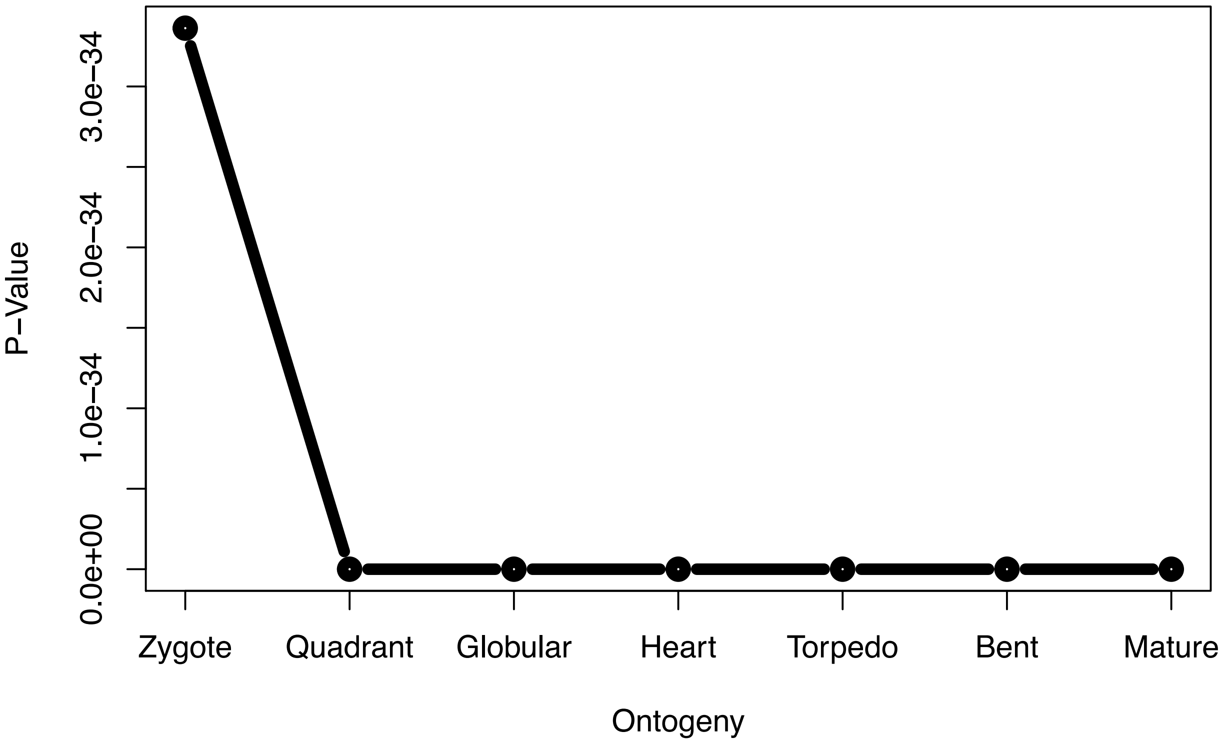
Computing the significant differences between gene expression distributions of PS or DS groups

As proposed by Quint et al., 2012 in some cases users whish to compare the difference of group specific expression levels using a statsitical test.

For this purpose, the `PlotGroupDiffs()` function performs a test to quantify the statistical significance between the global expression level distributions of groups of PS or DS. It therefore allows users to investigate significant groups of PS or DS that significantly differ in their gene expression level distibution within specific developmental stages or experiments.

Analogous to the `PlotRE()` or `PlotMeans()` function (see Introduction for details), users need to pass the `Groups` to `PlotGroupDiffs()` specifying the groups that shall be compared.

```
library(myTAI)

data(PhyloExpressionSetExample)

PlotGroupDiffs(ExpressionSet = PhyloExpressionSetExample,
               Groups        = list(group_1 = 1:3,group_2 = 4:12),
               legendName    = "PS",
               plot.type     = "p-vals",
               type          = "b",
               lwd           = 6,
               xlab          = "Ontogeny")
```



```
#>                                  Zygote    Quadrant     Globular        Heart      Torpedo          Ben
#> p.value ( wilcox.test ) 3.362424e-34 3.017793e-42 1.987181e-71 1.546567e-66 1.174961e-85 9.975191e-9
```

In cases where no plot shall be drawn and only the resulting p-value shall be returned users can specify the `plot.type = NULL` argument to receive only p-values returned by the underlying test statistic.

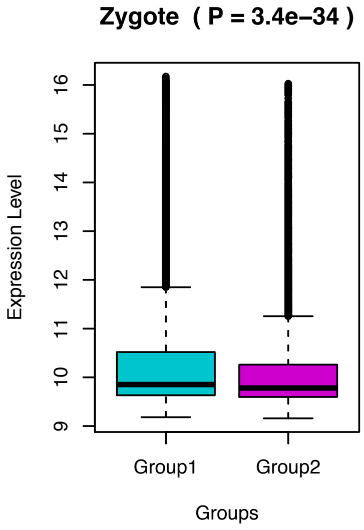
```
library(myTAI)

data(PhyloExpressionSetExample)

# only receive the p-values without the corresponding plot
PlotGroupDiffs(ExpressionSet = PhyloExpressionSetExample,
               Groups        = list(group_1 = 1:3,group_2 = 4:12),
               legendName    = "PS",
               plot.type     = NULL)
```

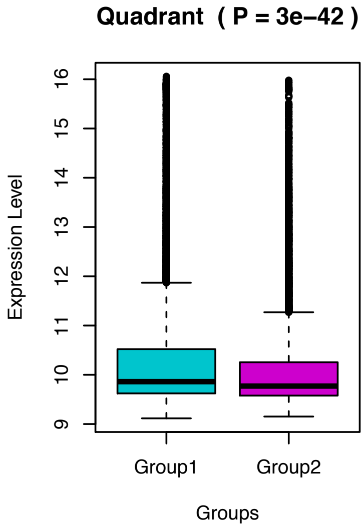
```
#>                               Zygote     Quadrant     Globular        Heart      Torpedo          Ben
#> p.value ( wilcox.test ) 3.362424e-34 3.017793e-42 1.987181e-71 1.546567e-66 1.174961e-85 9.975191e-9
```

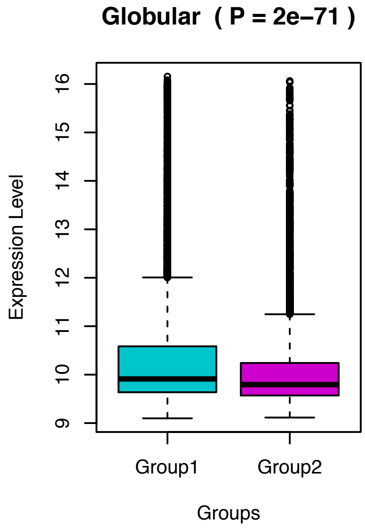
Optionally, users can also visualize the difference in expression level distributions of groups of PS/DS during each developmental stage by specifying the `plot.type = "boxplot"` argument.

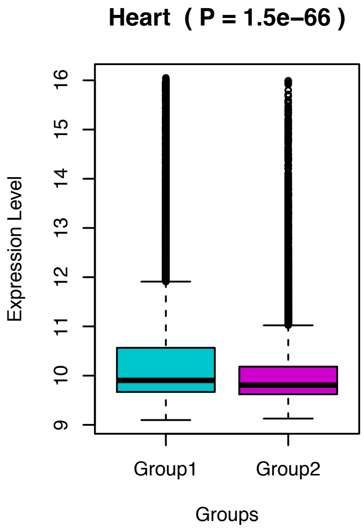
```
library(myTAI)

data(PhyloExpressionSetExample)

# visualize difference as boxplot
PlotGroupDiffs(ExpressionSet = tf(PhyloExpressionSetExample,log2),
               Groups        = list(group_1 = 1:3,group_2 = 4:12),
               legendName    = "PS",
               plot.type     = "boxplot")
```

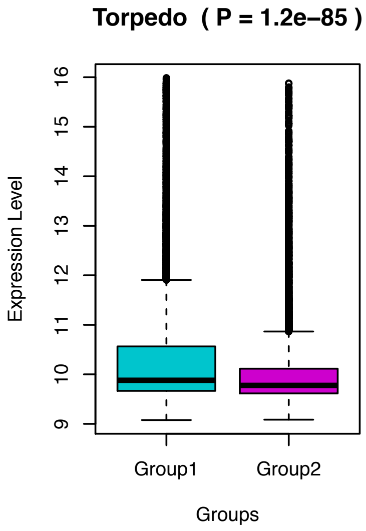
```
#>                               Zygote     Quadrant     Globular        Heart      Torpedo          Ben
#> p.value ( wilcox.test ) 3.362424e-34 3.017793e-42 1.987181e-71 1.546567e-66 1.174961e-85 9.975191e-9
```

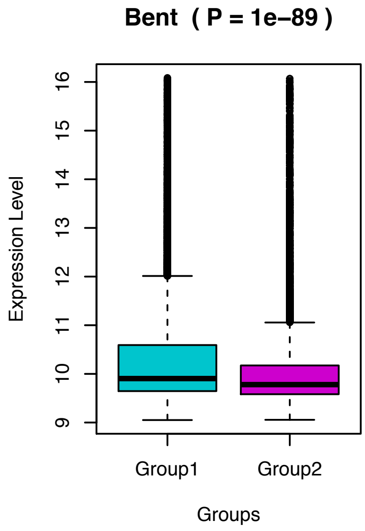

Zygote (P = 3.4e–34)
Expression Level
9
10
11
12
13
14
15
16
Group1
Group2
Groups

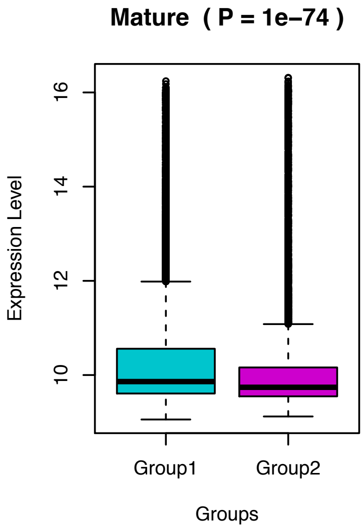

Quadrant (P = 3e–42)
Expression Level
9
10
11
12
13
14
15
16
Group1
Group2
Groups


Globular (P = 2e–71)
Expression Level
9
10
11
12
13
14
15
16
Group1
Group2
Groups


Heart (P = 1.5e–66)
Expression Level
9
10
11
12
13
14
15
16
Group1
Group2
Groups


Torpedo (P = 1.2e–85)
Expression Level
9
10
11
12
13
14
15
16
Group1
Group2
Groups


Bent (P = 1e–89)
Expression Level
9
10
11
12
13
14
15
16
Group1
Group2
Groups


Mature (P = 1e–74)
Expression Level
10
12
14
16
Group1
Group2
Groups

Here, we use log2 transformed expression levels for better visualization (`tf(PhyloExpressionSetExample,log2)`).

Internally, the `PlotGroupDiffs()` function performs a Wilcoxon Rank Sum test to quantify the statistical significance of PS/DS group expression. This quantification allows users to detect developmental stages of significant expression level differences between PS/DS groups. In this example we chose genes originated before the evolution of embryogenesis evolved in plants (Group1 = PS1-3) versus genes originated after the evolution of embryogenesis evolved in plants (Group2 = PS4-12). As a result, we observe that indeed the difference in total gene expression between these groups is significant throughout embryogenesis. In terms of the P-value quantification we observe that the P-value is minimized towards the phylotypic period. Hence, the expression level difference between the studied PS groups is maximized during the phylotypic period.

References

Domazet-Loso T, Tautz D. 2010. **A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns**. *Nature* **468**:815-8.

Drost HG *et al.* 2015. **Evidence for Active Maintenance of Phylotranscriptomic Hourglass Patterns in Animal and Plant Embryogenesis**. *Mol. Biol. Evol.* **32**(5):1221-1231.

Drost HG *et al.* 2016. **Post-embryonic hourglass patterns mark ontogenetic transitions in plant development**. *Mol. Biol. Evol.* 33(5):1158-1163.

Piasecka B. *et al.* 2013. **The hourglass and the early conservation models - co-existing patterns of developmental constraints in vertebrates**. *PLoS Genet.* **9**:e1003476.

Quackenbush J. 2001. Computational Analysis of Microarray Data. *Nature Reviews.* **2**:418-427.

Quackenbush J. Microarray data normalization and transformation. *Nature Genetics.* **32**:496-501.

Quint M, Drost HG *et al.* 2012. **A transcriptomic hourglass in plant embryogenesis**. *Nature* **490**:98-101.

Sestak MS, Domazet-Loso T. 2015. **Phylostratigraphic Profiles in Zebrafish Uncover Chordate Origins of the Vertebrate Brain**. *Mol. Biol. Evol.* 32(2):299-312.

Yeung KY *et al.* 2001. Model-based clustering and data transformations for gene expression data. *Bioinformatics.* **17**:977-987.

Yeung KY *et al.* 2001. Supplement to *Model-based clustering and data transformations for gene expression data – Data Transformations and the Gaussian mixture assumption. Bioinformatics.* **17**:977-987.

10.5 Documentation: Introduction to orthologr

Introduction to the orthologr Package

2016-06-25

Overview

The `orthologr` package allows users to perform BLAST searches, orthology inference methods, multiple sequence alignments, codon alignments, dNdS estimation, and divergence stratigraphy with R. The following tutorial will cover these topics in detail:

- Perform BLAST Searches
- Perform Sequence Alignments
- Perform Orthology Inference
- Perform dNdS Estimation
- Perform Divergence Stratigraphy

Installation Guide

Before you can load and install `orthologr` you need to install the following packages from Bioconductor:

```r
# install all Bioconductor packages orthologr depends on

# install Bioconductor base packages
source("http://bioconductor.org/biocLite.R")
biocLite()

# install package: Biostrings
biocLite("Biostrings")

# install package: S4Vectors
biocLite("S4Vectors")

# install package: XVector
biocLite("XVector")
```

Users might be asked during the installation process of `Biostrings`, `S4Vectors`, and `IRanges` whether or not they would like to update all package dependencies of the corresponding packages. Please type `a` specifying that all package dependencies of the corresponding packages shall be updated. This is important for the sufficient functionality of `orthologr`.

On Unix Based Systems

Now users can use the `devtools` package to install `orthologr` from GitHub.

```r
# install.packages("devtools")

# install the current version of orthologr on your system
library(devtools)
install_github("HajkD/orthologr", build_vignettes = TRUE, dependencies = TRUE)
```

On Windows Systems

In some cases (when working with **WINDOWS** machines), the installation via `devtools` will not work properly. In this case users can try the follwing steps:

```
# On Windows, this won't work - see ?build_github_devtools
install_github("HajkD/orthologr", build_vignettes = TRUE, dependencies = TRUE)

# When working with Windows, first users need to install the
# R package: rtools -> install.packages("rtools")

# Afterwards users can install devtools -> install.packages("devtools")
# and then they can run:

devtools::install_github("HajkD/orthologr", build_vignettes = TRUE, dependencies = TRUE)

# and then call it from the library
library("orthologr", lib.loc = "C:/Program Files/R/R-3.1.1/library")
```

Overview of the functions that are implemented in `orthologr`:

Perform BLAST searches with R

- `advanced_blast()`: Perform an advanced BLAST+ search
- `advanced_makedb()`: Create a BLASTable database with makeblastdb (advanced options)
- `blast()`: Perform a BLAST+ search
- `blast.nr()`: Perform a BLASTP search against NCBI nr
- `blast_best()`: Perform a BLAST+ best hit search
- `blast_rec()`: Perform a BLAST+ best reciprocal hit (BRH) search
- `delta.blast()`: Perform a DELTA-BLAST Search

Perform Pairwise and Multiple Sequence Alignements with R

- `multi_aln()`: Compute Multiple Sequence Alignments based on the `clustalw`, `t_coffee`, `muscle`, `clustalo`, and `mafft` programs.
- `pairwise_aln()`: Compute Pairwise Alignments
- `codon_aln()`: Compute a Codon Alignment

Perform Orthology Inference with R

- `orthologs()`: Main Orthology Inference Function
- `ProteinOrtho()`: Orthology Inference with ProteinOrtho

Perform Population Genomics with R

- `dNdS()`: Compute dNdS values for two organisms
- `substitutionrate()`: Internal function for dNdS computations

Read and Write CDS, Genomes, and Proteomes

- `read.cds()`: Read the CDS of a given organism
- `read.genome()`: Read the genome of a given organism
- `read.proteome()`: Read the proteome of a given organism
- `write.proteome()`: Save a proteome in fasta format

Getting Started

Performing BLAST Searches

The `orthologr` package stores 20 example genes (orthologs) between *Arabidopsis thaliana* and *Arabidopsis lyrata*. The following example BLAST search shall illustrate a simple search with standard parameters provided by the `blast()` function.

When running the subsequent functions please make sure you can call BLAST+ from your console either in the standard `PATH` or in case you have BLAST+ installed in a separate folder, please specify the `path` argument that can be passed to `blast()`.

To check whether BLAST+ can be executed from the default `PATH` (`usr/bin/local` on UNIX systems), you can run:

```
system("blastp -version")
```

This should return something like this:

```
blastp: 2.2.29+
Package: blast 2.2.29, build Dec 10 2013 15:51:59
```

If everything works properly, you can get started with you first BLAST+ search.

The blast() function

The `blast()` function provides the easiest way to perform a BLAST search.

```
library(dplyr)

# performing a BLAST search using blastp (default)
hit_tbl <-
blast(
query_file   = system.file('seqs/ortho_thal_cds.fasta', package = 'orthologr'),
subject_file = system.file('seqs/ortho_lyra_cds.fasta', package = 'orthologr')
)


glimpse(hit_tbl)
```

```
Variables:
$ query_id      (chr) "AT1G01010.1", "AT1G01020.1", "AT1G01030.1",...
$ subject_id    (chr) "333554|PACid:16033839", "470181|PACid:16064...
$ perc_identity (dbl) 73.99, 91.06, 95.54, 91.98, 100.00, 89.51, 9...
$ alig_length   (dbl) 469, 246, 359, 1970, 213, 648, 366, 300, 434...
$ mismatches    (dbl) 80, 22, 12, 85, 0, 58, 14, 22, 8, 34, 4, 6, ...
$ gap_openings  (dbl) 8, 0, 2, 10, 0, 5, 2, 2, 3, 0, 0, 1, 3, 2, 1...
$ q_start       (dbl) 1, 1, 1, 6, 1, 1, 1, 1, 1, 1, 1, 1, 5, 4, 2,...
$ q_end         (dbl) 430, 246, 359, 1910, 213, 646, 366, 294, 429...
$ s_start       (dbl) 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 16, 2, 4...
$ s_end         (dbl) 466, 246, 355, 1963, 213, 640, 362, 299, 433...
$ evalue        (dbl) 0e+00, 7e-166, 0e+00, 0e+00, 2e-160, 0e+00, ...
$ bit_score     (dbl) 627, 454, 698, 3704, 437, 1037, 696, 491, 85...
```

As you can see, the hit table shows the output of the BLAST+ search. The `blast()` function runs `blastp` as default BLAST+ algorithm. Different BLAST+ algorithms can be selected by specifying the `blast_algorithm` argument, e.g. `blast_algorithm = "tblastn"`. See `?blast` for further details. The `blast()` function returns the BLAST arguments: `query_id`, `subject_id`, `perc_identity`, `alig_length`, `mismatches`, `gap_openings`, `q_start`, `q_end`, `s_start`, `s_end`, `evalue`, and `bit_score`.

Since `blast()` stores the hit table returned by BLAST in a data.table object, you can access each column, using the data.table notation.

In case you need to specify the `PATH` to BLAST+ please use the `path` argument:

```
# performing a BLAST search using blastp (default)
hit_tbl <-
blast(
query_file   = system.file('seqs/ortho_thal_cds.fasta', package = 'orthologr'),
subject_file = system.file('seqs/ortho_lyra_cds.fasta', package = 'orthologr'),
path         = "/path/to/blastp"
)


hit_tbl
```

```
# access columns: query_id, subject_id, evalue, and bit_score
hit_tbl[ , list(query_id, subject_id, evalue, bit_score)]
```

```
       query_id             subject_id evalue bit_score
 1: AT1G01010.1 333554|PACid:16033839  0e+00       627
 2: AT1G01020.1 470181|PACid:16064328 7e-166       454
 3: AT1G01030.1 470180|PACid:16054974  0e+00       698
 4: AT1G01040.1 333551|PACid:16057793  0e+00      3704
 5: AT1G01050.1 909874|PACid:16064489 2e-160       437
 6: AT1G01060.3 470177|PACid:16043374  0e+00      1037
 7: AT1G01070.1 918864|PACid:16052578  0e+00       696
 8: AT1G01080.1 909871|PACid:16053217 1e-178       491
 9: AT1G01090.1 470171|PACid:16052860  0e+00       859
10: AT1G01110.2 333544|PACid:16034284  0e+00       972
11: AT1G01120.1 918858|PACid:16049140  0e+00      1092
12: AT1G01140.3 470161|PACid:16036015  0e+00       918
13: AT1G01150.1 918855|PACid:16037307 3e-150       421
14: AT1G01160.1 918854|PACid:16044153  1e-93       268
15: AT1G01170.2 311317|PACid:16052302  3e-54       158
16: AT1G01180.1 909860|PACid:16056125  0e+00       576
17: AT1G01190.1 311315|PACid:16059488  0e+00      1036
```

```
18: AT1G01200.1 470156|PACid:16041002 3e-172       470
19: AT1G01210.1 311313|PACid:16057125  7e-76       215
20: AT1G01220.1 470155|PACid:16047984  0e+00      2106
```

The `blast()` function also allows you to pass additional parameters to the BLAST+ search using the `blast_params` argument. In the following example, a remote BLAST+ search is performed.

```
blast(
        query_file =
        system.file('seqs/ortho_thal_cds.fasta', package = 'orthologr'),
        subject_file =
        system.file('seqs/ortho_lyra_cds.fasta', package = 'orthologr'),
        blast_params = "-max_target_seqs 1"
        )
```

In all cases the default `e-value` BLAST+ searches is `1E-5` and the default `blast_algorithm` is `blastp`.

Since BLAST+ searches can be computationally expensive, it is possible to specify the `comp_cores` argument when working with an multicore machine.

```
# BLAST computations using the comp_cores parameter: here with 2 cores
blast(
query_file   = system.file('seqs/ortho_thal_cds.fasta', package = 'orthologr'),
subject_file = system.file('seqs/ortho_lyra_cds.fasta', package = 'orthologr'),
comp_cores   = 2
)
```

The `query_file` and `subject_file` arguments specify the path to the corresponding fasta files storing the `CDS` files, `amino acid` files, or `genome` files of the query organism and subject organism of interest. Make sure that when using `CDS`files, `amino acid` files, or `genome` files the corresponding argument `seq_type` must be adapted according to the input data format.

Use :

- `CDS` files -> `seq_type = "cds"`
- `amino acid` files -> `seq_type = "protein"`
- `genome` files -> `seq_type = "dna"`

The `format` argument specifies the input file format, e.g. "fasta" or "gbk". The `blast_algorithm` argument specifies the BLAST program (algorithm) that shall be used to perform BLAST searches, e.g. "blastp","blastn","tblastn",etc. Again, the `eval` argument defines the default e-value that shall be chosen as best hit threshold.

Using the split-apply-combine strategy for a BLAST hit table

All `blast` functions implemented in `orthologr` can easily be processed using the split-apply-combine strategy to detect for example `one-to-one`, `one-to-many`, and `many-to-many` gene homology relationships.

Here a simple example:

```
# install.packages(c("plyr","dplyr"), dependencies = TRUE)
library(plyr)
library(dplyr)
```

```
# perform a blastp search of 20 A. thaliana genes against 1000 A. lyrata genes
hit_tbl <-
blast(
query_file   = system.file('seqs/ortho_thal_cds.fasta', package = 'orthologr'),
subject_file = system.file('seqs/ortho_lyra_cds_1000.fasta', package = 'orthologr')
)

# determine 'one-to-many' and 'one-to-one' gene relationships
rel_hit_tbl <-
ddply(.data = hit_tbl,
.variables = "query_id",
.fun = nrow)
colnames(rel_hit_tbl)[2] <- "n_genes"

rel_hit_tbl
```

```
      query_id n_genes
1  AT1G01010.1       4
2  AT1G01020.1       1
3  AT1G01030.1       1
4  AT1G01040.1       1
5  AT1G01050.1       1
6  AT1G01060.3       2
7  AT1G01070.1       3
8  AT1G01080.1       4
9  AT1G01090.1       1
10 AT1G01110.2       1
11 AT1G01120.1       3
12 AT1G01140.3      36
13 AT1G01150.1       1
14 AT1G01160.1       1
15 AT1G01170.2       1
16 AT1G01180.1       1
17 AT1G01190.1       6
18 AT1G01200.1       8
19 AT1G01210.1       1
20 AT1G01220.1       1
```

Now you can sort genes into classes: `one-to-one` and `one-to-many`.

```
# classify into 'one-to-one' relationships
one_to_one <- filter(rel_hit_tbl,n_genes == 1)

# classify into 'one-to-many' relationships
one_to_many <- filter(rel_hit_tbl,n_genes > 1)
```

```
# look at one_to_one
one_to_one
```

```
      query_id n_genes
1  AT1G01020.1       1
2  AT1G01030.1       1
```

```
3  AT1G01040.1       1
4  AT1G01050.1       1
5  AT1G01090.1       1
6  AT1G01110.2       1
7  AT1G01150.1       1
8  AT1G01160.1       1
9  AT1G01170.2       1
10 AT1G01180.1       1
11 AT1G01210.1       1
12 AT1G01220.1       1
```

```r
# look at one_to_many
one_to_many
```

```
     query_id n_genes
1 AT1G01010.1       4
2 AT1G01060.3       2
3 AT1G01070.1       3
4 AT1G01080.1       4
5 AT1G01120.1       3
6 AT1G01140.3      36
7 AT1G01190.1       6
8 AT1G01200.1       8
```

Now we can treat classes: `one_to_one` and `one_to_many` differently:

one-to-one genes

```r
# look at the evalue, perc_identity, and alig_length of one_to_one genes
oo_genes <-
dplyr::filter(hit_tbl, query_id %in% one_to_one[, "query_id"])

oo_genes[, list(query_id, subject_id, evalue, perc_identity, alig_length)]
```

```
      query_id            subject_id evalue perc_identity alig_length
1  AT1G01020.1 470181|PACid:16064328 3e-164         91.06         246
2  AT1G01030.1 470180|PACid:16054974  0e+00         95.54         359
3  AT1G01040.1 333551|PACid:16057793  0e+00         91.98        1970
4  AT1G01050.1 909874|PACid:16064489 1e-158        100.00         213
5  AT1G01090.1 470171|PACid:16052860  0e+00         96.77         434
6  AT1G01110.2 333544|PACid:16034284  0e+00         93.56         528
7  AT1G01150.1 918855|PACid:16037307 1e-148         72.63         285
8  AT1G01160.1 918854|PACid:16044153  5e-92         84.92         179
9  AT1G01170.2 311317|PACid:16052302  1e-52         85.57          97
10 AT1G01180.1 909860|PACid:16056125  0e+00         92.58         310
11 AT1G01210.1 311313|PACid:16057125  3e-74         95.33         107
12 AT1G01220.1 470155|PACid:16047984  0e+00         96.69        1056
```

Now you could filter for additional criteria to define a first set of true orthologs. In this example we define true orthologs as `one_to_one` genes having a minimum alignment length of 300, a perc_identity of > 80 percent and an e-value $< 1E-5$.

```
# look at the evalue, perc_identity, and alig_length of one_to_one genes
oo_genes <-
dplyr::filter(hit_tbl, query_id %in% one_to_one[, "query_id"])

true_orthologs <-
dplyr::filter(oo_genes, evalue < 1e-5, perc_identity > 80, alig_length > 300)

true_orthologs[, list(query_id, subject_id, evalue, perc_identity, alig_length)]
```

```
      query_id            subject_id evalue perc_identity alig_length
1: AT1G01030.1 470180|PACid:16054974      0         95.54         359
2: AT1G01040.1 333551|PACid:16057793      0         91.98        1970
3: AT1G01090.1 470171|PACid:16052860      0         96.77         434
4: AT1G01110.2 333544|PACid:16034284      0         93.56         528
5: AT1G01180.1 909860|PACid:16056125      0         92.58         310
6: AT1G01220.1 470155|PACid:16047984      0         96.69        1056
```

This way we could filter out a high confidence set of orthologous genes from the `one_to_one` class of genes.

In reality most orthology inference programs and methods perform way more complicated and sophisticated analyses to distinguish true orthologs from true paralogs (in-paralogs, out-paralogs, etc.). These subsequent analyses can also be performed using the above introduced split-apply-combine strategy.

Note, that you can perform self-BLAST searches `blast(query,query)` and `blast(subject,subject)` to distinguish between orthologous and paralogous genes.

Now we continue with the `one_to_many` class of genes.

one-to-many genes

Here we want to address the question how to deal with multiple hits returned by BLAST+ .

Again we investigate all `one_to_many` genes:

```
one_to_many
```

```
     query_id n_genes
1 AT1G01010.1       4
2 AT1G01060.3       2
3 AT1G01070.1       3
4 AT1G01080.1       4
5 AT1G01120.1       3
6 AT1G01140.3      36
7 AT1G01190.1       6
8 AT1G01200.1       8
```

When looking at gene_id `AT1G01200.1` we see that it was found 8 times in the corresponding subject set of *A. lyrata*.

```
hit_tbl["AT1G01200.1", list(query_id, subject_id, evalue, perc_identity, alig_length)]
```

```
      query_id            subject_id evalue perc_identity alig_length
1: AT1G01200.1 470156|PACid:16041002 2e-170         95.80         238
```

```
2: AT1G01200.1 909905|PACid:16035105  7e-06         21.64         171
3: AT1G01200.1 910431|PACid:16035207  8e-74         52.97         219
4: AT1G01200.1 918732|PACid:16054958  2e-50         44.56         193
5: AT1G01200.1 919287|PACid:16060536  1e-68         58.10         179
6: AT1G01200.1 919355|PACid:16050170  7e-72         53.30         212
7: AT1G01200.1 919721|PACid:16036935  9e-80         59.31         204
8: AT1G01200.1 919852|PACid:16055066  4e-07         24.03         154
```

Now we have to decide which hit shall be considered as potential *ortholog*. In this example `subject_id` `470156|PACid:16041002` has the highest `perc_identity` as well as the lowest e-value `2e-170`. So a straightforward approach would be to choose subject gene `470156|PACid:16041002` as potential ortholog of query gene `AT1G01200.1`.

We can validate this approach by running a reciprocal best hit search with `blast_rec()`and compare the output of gene `AT1G01200.1` with our choice `470156|PACid:16041002`.

```
rbh_hit_tbl <-
        blast_rec(
        query_file   = system.file('seqs/ortho_thal_cds.fasta', package = 'orthologr'),
        subject_file = system.file('seqs/ortho_lyra_cds_1000.fasta', package = 'orthologr')
        )

rbh_hit_tbl
```

```
      query_id            subject_id evalue
1  AT1G01010.1 333554|PACid:16033839  0e+00
2  AT1G01020.1 470181|PACid:16064328 3e-164
3  AT1G01030.1 470180|PACid:16054974  0e+00
4  AT1G01040.1 333551|PACid:16057793  0e+00
5  AT1G01050.1 909874|PACid:16064489 1e-158
6  AT1G01060.3 470177|PACid:16043374  0e+00
7  AT1G01070.1 918864|PACid:16052578  0e+00
8  AT1G01080.1 909871|PACid:16053217 5e-177
9  AT1G01090.1 470171|PACid:16052860  0e+00
10 AT1G01110.2 333544|PACid:16034284  0e+00
11 AT1G01120.1 918858|PACid:16049140  0e+00
12 AT1G01140.3 470161|PACid:16036015  0e+00
13 AT1G01150.1 918855|PACid:16037307 1e-148
14 AT1G01160.1 918854|PACid:16044153  5e-92
15 AT1G01170.2 311317|PACid:16052302  1e-52
16 AT1G01180.1 909860|PACid:16056125  0e+00
17 AT1G01190.1 311315|PACid:16059488  0e+00
18 AT1G01200.1 470156|PACid:16041002 2e-170
19 AT1G01210.1 311313|PACid:16057125  3e-74
20 AT1G01220.1 470155|PACid:16047984  0e+00
```

When we now look at gene `AT1G01200.1` we find that indeed subject gene `470156|PACid:16041002` has been detected as potential ortholog having the same evalue `2e-170`.

```
rbh_hit_tbl["AT1G01200.1" , ]
```

```
     query_id            subject_id evalue
1 AT1G01200.1 470156|PACid:16041002 2e-170
```

Nevertheless there might be cases in which it is hard to decide for or against the best hit compared with all other hits.

For example we can investigate gene `AT1G01070.1` :

```
hit_tbl["AT1G01070.1", list(query_id,subject_id,evalue,perc_identity,alig_length)]
```

```
      query_id            subject_id evalue perc_identity alig_length
1: AT1G01070.1 918864|PACid:16052578  0e+00         95.08         366
2: AT1G01070.1 919693|PACid:16048878  2e-67         32.87         356
3: AT1G01070.1 919961|PACid:16062329  0e+00         79.29         338
```

Here both e-values for subject genes `918864|PACid:16052578` and `919961|PACid:16062329` are the same and only `perc_identity` and `alig_length` differ. The *reciprocal best hit* approach chose gene `918864|PACid:16052578` which also had the highest `perc_identity`.

```
rbh_hit_tbl["AT1G01070.1" , ]
```

```
     query_id            subject_id evalue
1 AT1G01070.1 918864|PACid:16052578      0
```

But since the `blast_rec()` function was implemented to choose the bidirectional best hit based on the e-value, in border line cases a different gene as expected could be chosen.

An alternative analysis that can be performed with these three candidate subject genes is the following:

```
# read CDS sequences of the 20 example query genes of A. thaliana
Ath.cds <-
read.cds(
file   = system.file('seqs/ortho_thal_cds.fasta', package = 'orthologr'),
format = "fasta"
)

# read CDS sequences of the 1000 example subject genes of A. lyrata
Aly.cds <-
read.cds(
file   = system.file('seqs/ortho_lyra_cds_1000.fasta', package = 'orthologr'),
format = "fasta"
)


# show the sequence of gene AT1G01070.1
Ath.cds["AT1G01070.1" , seqs]
```

```
[1] "atggctggagatatgcaaggagtgagagtagtagaaaaatattcaccggtcatagtgatggtgatgtcaaatgta
gcgatgggttcggtgaatgcacttgtgaagaaagctcttgatgttggtgtgaaccatatggtcattggtgcttatcgaat
ggctatttccgctttaattttggttccctttgcctatgtcttggaaaggaaaacaagaccacaaataacgtttaggctaa
tggtcgatcatttcgtcagtggccttctcggggcgagtttgatgcagtttttctttttgcttggtctgtcgtacacgtca
gcaactgtttcgtgtgctttggtaagcatgttgcctgcaatcaccttcgctttggcccttattttcaggactgaaaatgt
gaagattctaaagaccaaagcaggaatgttgaaggtgattggaactttgatctgtataagtggagctttgttcttaacat
tttacaaaggcccacaaatatcaaactctcactctcactctcacggtggggcttcccacaacaacaacgatcaagacaag
gccaataattggcttcttggatgtctttatttaaccataggaacagtgttgctatctctatggatgttgtttcaaggac
tttaagtattaagtacccttgcaaatactcgagcacttgtcttatgtcaattttcgcggcatttcaatgtgctctcttga
```

```
gcctttacaagagcagagacgttaatgattggatcatagatgatagattcgttatcaccgtcatcatatacgctggagtg
gtaggacaagcaatgacgacggttgcaacaacatggggattaaaaaattaggagctgtgttcgcatcggcgtttttccc
acttactctcatttcggctactctatttgatttcctcattttacacactcctttataccttggaagtgtgattggatcac
tagtgaccataacgggtctctacatgttcttgtgggggaagaacaaagaaacggaatcatcaactgcattgtcttcagga
atggataacgaagctcaatatactactcctaataaggataacgactctaagtcgcccgtttaa"
```

Now you can perform a global alignment between the CDS sequences of `AT1G01070.1` and the three subject genes as follows:

```r
library(Biostrings)

# perform 3 global alignments between:  AT1G01070.1 and 918864|PACid:16052578,
# 919693|PACid:16048878, 919961|PACid:16062329
sapply(Aly.cds[hit_tbl["AT1G01070.1", subject_id], seqs],
pairwiseAlignment,
pattern = Ath.cds["AT1G01070.1" , seqs],
type    = "global")
```

```
$...
Global PairwiseAlignmentsSingleSubject (1 of 1)
pattern: [1] atggctggagatatgcaaggagtgagagta...aaggataacgactctaagtcgcccgtttaa
subject: [1] atgggtgaaggtatgattggagtgagagta...aaggataacgactctaagtcgcccgtttaa
score: 1768.965

$...
Global PairwiseAlignmentsSingleSubject (1 of 1)
pattern: [1] atggctgga---gatatgcaaggagtgaga...----cgac----tctaagtcgcccgtttaa
subject: [1] atggctaaatcagatatgc------tg---...ggttccacaaggtctatatcgcc---ttaa
score: -2318.726

$...
Global PairwiseAlignmentsSingleSubject (1 of 1)
pattern: [1] atggctggagatatgcaaggagtgagagta...aaggataacgactctaagtcgcccgtttaa
subject: [1] atgagtgaggatatgggaggagtgaaagta...----------------------------aa
score: 486.462
```

Note: To obtain the score value, you need to specify the `scoreOnly = TRUE` in the `pairwiseAlignment` function.

As you can see, subject gene `918864|PACid:16052578` also has the highest global alignment score `1768.965` based on the Needleman-Wunsch algorithm. This strategy might help you to differentiate between border line cases.

The examples shown above shall demonstrate the use cases that can be performed using the `blast` functions implemented in `orthologr`.

Another useful analysis can be to take the length of the initial query genes into account using the `nchar()` function:

```r
# show the length distribution of all genes
# stored in "Ath.cds"
Ath.cds[ , nchar(seqs)]
```

```
[1] 1290  738 1077 5730  639 1938 1098  882 1287 1584 1587 1356
1038  588  252 1437 1608  714  321 3168
```

Or the length of a specific gene:

```
Ath.cds["AT1G01070.1" , nchar(seqs)]
```

```
[1] 1098
```

This way you can easily visualize the length distribution of genes stored in your query organism file.

```
Ath.cds <-
        read.cds(system.file('seqs/ortho_thal_cds_1000.fasta', package = 'orthologr'),
        format = "fasta")

hist(Ath.cds[, nchar(seqs)], breaks = 100)
```

The blast_best() function

For some analyses it is sufficient to perform BLAST+ best hit searches. The `blast_best()` function is optimized to perform BLAST+ best hit searches (only based on the minimum e-value) and returns the best hit when performing a BLAST+ search of a query organisms (or set of query genes) against a subject organism (or set of subject genes).

```
# performing gene orthology inference using the best hit (BH) method
blast_best(
        query_file    = system.file('seqs/ortho_thal_cds.fasta', package = 'orthologr'),
        subject_file  = system.file('seqs/ortho_lyra_cds.fasta', package = 'orthologr'),
        clean_folders = TRUE
        )
```

```
       query_id            subject_id evalue
 1: AT1G01010.1 333554|PACid:16033839  0e+00
 2: AT1G01020.1 470181|PACid:16064328 7e-166
 3: AT1G01030.1 470180|PACid:16054974  0e+00
 4: AT1G01040.1 333551|PACid:16057793  0e+00
 5: AT1G01050.1 909874|PACid:16064489 2e-160
 6: AT1G01060.3 470177|PACid:16043374  0e+00
 7: AT1G01070.1 918864|PACid:16052578  0e+00
 8: AT1G01080.1 909871|PACid:16053217 1e-178
 9: AT1G01090.1 470171|PACid:16052860  0e+00
10: AT1G01110.2 333544|PACid:16034284  0e+00
11: AT1G01120.1 918858|PACid:16049140  0e+00
12: AT1G01140.3 470161|PACid:16036015  0e+00
13: AT1G01150.1 918855|PACid:16037307 3e-150
14: AT1G01160.1 918854|PACid:16044153  1e-93
15: AT1G01170.2 311317|PACid:16052302  3e-54
16: AT1G01180.1 909860|PACid:16056125  0e+00
17: AT1G01190.1 311315|PACid:16059488  0e+00
18: AT1G01200.1 470156|PACid:16041002 3e-172
19: AT1G01210.1 311313|PACid:16057125  7e-76
20: AT1G01220.1 470155|PACid:16047984  0e+00
```

The `blast_best()` function returns: `query_id`, `subject_id`, and `eval`.

In case you need more parameters returned by a BLAST+ best hit search, you can specify the `detailed_output` argument (`detailed_output = TRUE`).

```
# BLAST+ best hit search
best_hit_tbl <-
blast_best(
query_file      = system.file('seqs/ortho_thal_cds.fasta', package = 'orthologr'),
subject_file    = system.file('seqs/ortho_lyra_cds.fasta', package = 'orthologr'),
detailed_output = TRUE
)

dplyr::glimpse(best_hit_tbl)
```

```
Variables:
$ query_id      (chr) "AT1G01010.1", "AT1G01020.1", "AT1G01030.1",...
$ subject_id    (chr) "333554|PACid:16033839", "470181|PACid:16064...
$ perc_identity (dbl) 73.99, 91.06, 95.54, 91.98, 100.00, 89.51, 9...
$ alig_length   (dbl) 469, 246, 359, 1970, 213, 648, 366, 300, 434...
$ mismatches    (dbl) 80, 22, 12, 85, 0, 58, 14, 22, 8, 34, 4, 6, ...
$ gap_openings  (dbl) 8, 0, 2, 10, 0, 5, 2, 2, 3, 0, 0, 1, 3, 2, 1...
$ q_start       (dbl) 1, 1, 1, 6, 1, 1, 1, 1, 1, 1, 1, 1, 5, 4, 2,...
$ q_end         (dbl) 430, 246, 359, 1910, 213, 646, 366, 294, 429...
$ s_start       (dbl) 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 16, 2, 4...
$ s_end         (dbl) 466, 246, 355, 1963, 213, 640, 362, 299, 433...
$ evalue        (dbl) 0e+00, 7e-166, 0e+00, 0e+00, 2e-160, 0e+00, ...
$ bit_score     (dbl) 627, 454, 698, 3704, 437, 1037, 696, 491, 85...
```

The blast_rec() function

The `blast_rec()` function was implemented to optimize BLAST+ reciprocal best hit searches (only based on the minimum e-value). BLAST+ reciprocal best hit searches are used to perform orthology inference.

Running `blast_rec()` using default parameter settings:

```
# performing gene orthology inference using the reciprocal best hit (RBH) method
blast_rec(
query_file   = system.file('seqs/ortho_thal_cds.fasta', package = 'orthologr'),
subject_file = system.file('seqs/ortho_lyra_cds.fasta', package = 'orthologr')
)
```

```
      query_id            subject_id evalue
1  AT1G01010.1 333554|PACid:16033839  0e+00
2  AT1G01020.1 470181|PACid:16064328 7e-166
3  AT1G01030.1 470180|PACid:16054974  0e+00
4  AT1G01040.1 333551|PACid:16057793  0e+00
5  AT1G01050.1 909874|PACid:16064489 2e-160
6  AT1G01060.3 470177|PACid:16043374  0e+00
7  AT1G01070.1 918864|PACid:16052578  0e+00
8  AT1G01080.1 909871|PACid:16053217 1e-178
9  AT1G01090.1 470171|PACid:16052860  0e+00
10 AT1G01110.2 333544|PACid:16034284  0e+00
11 AT1G01120.1 918858|PACid:16049140  0e+00
12 AT1G01140.3 470161|PACid:16036015  0e+00
```

```
13 AT1G01150.1 918855|PACid:16037307 3e-150
14 AT1G01160.1 918854|PACid:16044153  1e-93
15 AT1G01170.2 311317|PACid:16052302  3e-54
16 AT1G01180.1 909860|PACid:16056125  0e+00
17 AT1G01190.1 311315|PACid:16059488  0e+00
18 AT1G01200.1 470156|PACid:16041002 3e-172
19 AT1G01210.1 311313|PACid:16057125  7e-76
20 AT1G01220.1 470155|PACid:16047984  0e+00
```

Again you can specify the `detailed_output` argument to get more parameters returned by `blast_rec()`.

```
# running blast_rec() using detailed_output = TRUE
rbh <-
blast_rec(
query_file      = system.file('seqs/ortho_thal_cds.fasta', package = 'orthologr'),
subject_file    = system.file('seqs/ortho_lyra_cds.fasta', package = 'orthologr'),
detailed_output = TRUE
)

dplyr::glimpse(rbh)
```

```
Variables:
$ query_id      (chr) "AT1G01010.1", "AT1G01020.1", "AT1G01030.1",...
$ subject_id    (chr) "333554|PACid:16033839", "470181|PACid:16064...
$ perc_identity (dbl) 73.99, 91.06, 95.54, 91.98, 100.00, 89.51, 9...
$ alig_length   (dbl) 469, 246, 359, 1970, 213, 648, 366, 300, 434...
$ mismatches    (dbl) 80, 22, 12, 85, 0, 58, 14, 22, 8, 34, 4, 6, ...
$ gap_openings  (dbl) 8, 0, 2, 10, 0, 5, 2, 2, 3, 0, 0, 1, 3, 2, 1...
$ q_start       (dbl) 1, 1, 1, 6, 1, 1, 1, 1, 1, 1, 1, 1, 5, 4, 2,...
$ q_end         (dbl) 430, 246, 359, 1910, 213, 646, 366, 294, 429...
$ s_start       (dbl) 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 16, 2, 4...
$ s_end         (dbl) 466, 246, 355, 1963, 213, 640, 362, 299, 433...
$ evalue        (dbl) 0e+00, 7e-166, 0e+00, 0e+00, 2e-160, 0e+00, ...
$ bit_score     (dbl) 627, 454, 698, 3704, 437, 1037, 696, 491, 85...
```

The advanced_blast() function

The `advanced_blast()` function was implemented to allow you to perform BLAST+ searches in a flexible environment. All parameters that shall be passed to the corresponding BLAST+ search need to be specified using the `blast_params` argument. In case you work with very large hit tables, that do not fit into memory, you can specify the `sql_database` argument to store the corresponding hit table in a SQLite database and access it via the `dplyr database notation`.

```
ab <-
        advanced_blast(
        query_file  = system.file('seqs/ortho_thal_cds.fasta', package = 'orthologr'),
        subject_file    = system.file('seqs/ortho_lyra_cds.fasta', package = 'orthologr'),
        blast_algorithm = "blastp",
        blast_params    = "-evalue 1E-5 -num_threads 1"
        )

dplyr::glimpse(ab)
```

```
Variables:
$ query_id                 (chr) "AT1G01010.1", "AT1G01020.1", "AT1...
$ subject_id               (chr) "333554|PACid:16033839", "470181|P...
$ perc_identity            (dbl) 73.99, 91.06, 95.54, 91.98, 100.00...
$ num_ident_matches        (dbl) 347, 224, 343, 1812, 213, 580, 348...
$ alig_length              (dbl) 469, 246, 359, 1970, 213, 648, 366...
$ mismatches               (dbl) 80, 22, 12, 85, 0, 58, 14, 22, 8, ...
$ gap_openings             (dbl) 8, 0, 2, 10, 0, 5, 2, 2, 3, 0, 0, ...
$ q_start                  (dbl) 1, 1, 1, 6, 1, 1, 1, 1, 1, 1, 1, 1...
$ q_end                    (dbl) 430, 246, 359, 1910, 213, 646, 366...
$ s_start                  (dbl) 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1...
$ s_end                    (dbl) 466, 246, 355, 1963, 213, 640, 362...
$ evalue                   (dbl) 0e+00, 7e-166, 0e+00, 0e+00, 2e-16...
$ bit_score                (dbl) 627, 454, 698, 3704, 437, 1037, 69...
$ score_raw                (dbl) 1617, 1169, 1801, 9604, 1125, 2681...
$ query_coverage_per_subj  (dbl) 100, 100, 100, 99, 100, 100, 100, ...
```

As you can see, the `advanced_blast()` function returns more parameters than all other `blast()` functions.

Selecting the best hit using advanced_blast()

The flexibility of `advanced_blast()` enables you to perform all previously introduced analyses (`blast()`, `blast_best()`, `blast_rec()`, etc.) as well as a variety of addtional analyses. The following example shall illustrate how to perform BLAST best hit searches using `advanced_blast()`

```r
# when performing an advanced BLAST search, you can easily select the best hit
library(dplyr)

advB <- advanced_blast(
        query_file      = system.file('seqs/ortho_thal_cds.fasta', package = 'orthologr'),
        subject_file    = system.file('seqs/ortho_lyra_cds_1000.fasta', package = 'orthologr'),
        seq_type        = "cds",
        blast_algorithm = "blastp",
        blast_params    = "-evalue 1E-5 -num_threads 1"
        )

best_hit <- advB %>% group_by(query_id) %>% do(min(evalue))

best_hit
```

```
      query_id min(evalue)
1  AT1G01010.1       0e+00
2  AT1G01020.1      3e-164
3  AT1G01030.1       0e+00
4  AT1G01040.1       0e+00
5  AT1G01050.1      1e-158
6  AT1G01060.3       0e+00
7  AT1G01070.1       0e+00
8  AT1G01080.1      5e-177
9  AT1G01090.1       0e+00
10 AT1G01110.2       0e+00
11 AT1G01120.1       0e+00
```

```
12 AT1G01140.3        0e+00
13 AT1G01150.1       1e-148
14 AT1G01160.1        5e-92
15 AT1G01170.2        1e-52
16 AT1G01180.1        0e+00
17 AT1G01190.1        0e+00
18 AT1G01200.1       2e-170
19 AT1G01210.1        3e-74
20 AT1G01220.1        0e+00
```

The `blast_params` argument allows you to specify all parameters that shall be passed to the corresponding `blast_algorithm` based on the NCBI blast stand-alone notation.

Using DELTA-BLAST in advanced_blast()

Domain enhanced lookup time accelerated BLAST (DELTA-BLAST) is a new BLAST algorithm provided by NCBI. It was introduced as a useful program for the detection of remote protein homologs.

You can download the Conserved Domain Database (CDD) file `cdd.tar.gz` from NCBI and store all `cdd_deltablast.* files` in a folder: `path/to/cdd_database/folder`.

More information on how to install and use DELTA-BLAST can be found here. Make sure that when using DELTA-BLAST in `advanced_blast()` the `db_path` argument denoting the path to the folder storing the corresponding `cdd_deltablast.* files` need to be specified.

```r
# DELTA-BLAST
#
# you can also use deltablast to perform BLAST searches
# make sure you have specified the the db_path argument to the cdd_deltablast.* files
advanced_blast(
query_file      = system.file('seqs/ortho_thal_cds.fasta', package = 'orthologr'),
subject_file    = system.file('seqs/ortho_lyra_cds.fasta', package = 'orthologr'),
blast_algorithm = "deltablast",
db_path         = "path/to/cdd_files",
blast_params    = "-evalue 1E-5 -num_threads 2"
)
```

A more intuitive way (but less flexible) to perform delta-blast searches is to use the `delta.blast()` function. The `delta.blast()` function works in two ways.

1) Use the `cdd_delta` database as additional information by performing `deltablast` searches between `query organism A` and `subject organism B`.
2) Perform a `deltablast` search between `query organism A` and the `cdd_delta` database

The following example illustrates the first option (A vs B via cdd_delta).

Here the `cdd.path` argument specifies the path to the `cdd_deltablast.* files`:

```r
# perform a delta-blast serach between A. thaliana and A. lyrata genes
delta.blast(
query_file   = system.file('seqs/ortho_thal_cds.fasta', package = 'orthologr'),
subject_file = system.file('seqs/ortho_lyra_cds.fasta', package = 'orthologr'),
cdd.path     = "path/to/cdd/database/folder",
comp_cores   = 1
)
```

The following example illustrates the second option (A vs cdd_delta). The difference is the specification of argument `subject_file = "cdd_delta"` instead of a second subject organism:

```
# perform a delta-blast serach between A. thaliana and the cdd database
delta.blast(
query_file   = system.file('seqs/ortho_thal_cds.fasta', package = 'orthologr'),
subject_file = "cdd_delta",
cdd.path     = "path/to/cdd/database/folder",
comp_cores   = 1
)
```

Performing BLASTP searches against the NCBI nr database with `blast.nr()`

The `blast.nr()` function implemented in `orthologr` allows users to perform local `blastp` searches against the NCBI nr database.

For this purpose, the first step is to download the NCBI nr database to a local machine using the `biomartr` package.

Please notice that the NCBI nr database contains about 30 files storing approx. 500 MB data per file. So when downloading this large database via `biomartr` users need to make sure that they are constantly connected to the internet.

```
# install the 'biomartr' package from CRAN
install.packages("biomartr")

# load biomartr
library("biomartr")

# download the NCBI nr database to your local machine to a folder
# names 'DB'
sapply(listDatabases("nr"),download_database)
```

After downloading all NCBI nr files to the `DB` folder users need to go to the `DB` folder and unpack all `env_nr*` files and store them in the same `DB` folder. In case users whish to specify a different folder for downloading NCBI nr they can specify the `path` argument of the `?download_database` function implemented in `biomartr`.

Now users can use the `blast.nr()` function to perform BLASTP searches against NCBI nr.

```
nr.res <-
        blast.nr(
        query_file      = system.file('seqs/aa_seqs.fasta', package = 'orthologr'),
        nr.path         = "DB",
        seq_type        = "protein",
        max.target.seqs = 20,
        comp_cores      = 2
        )

nr.res
```

Here, the `nr.path` argument specifies the folder path to the NCBI nr database. In case users unpacked all NCBI nr files to the `DB` folder created by `download_database()` the `nr.path` argument can be specified as `nr.path = "DB"`.

The corresponding `gi id` can then be clipped by running:

```
nr.res[, "subject_id"] <-
        unlist(lapply(stringr::str_split(nr.res[, subject_id], "[|]"), function(x)
        x[2]))

nr.res
```

```
                query_id subject_id perc_identity num_ident_matches alig_length mismatches
1: 333554|PACid:16033839  139291046         40.00                24          60         35
2: 333554|PACid:16033839  140103111         46.03                29          63         34
   gap_openings n_gaps pos_match ppos q_start q_end q_len qcov qcovhsp
1:            1      1        39 65.0       8    66   246   24      24
2:            0      0        47 74.6       8    70   246   26      26
                                                          query_seq s_start s_end s_len
1:    CVGC-GFRVKSLFIQYSPGNIRLMKCGNCKEVADEYIECERMIIFIDLILHRPKVYRHVL       9    68    71
2: CVGCGFRVKSLFIQYSPGNIRLMKCGNCKEVADEYIECERMIIFIDLILHRPKVYRHVLYNAI      48   110   114
                                                        subject_seq evalue bit_score score_raw
1:    CVECMCDENESIYKKYSEGNLRLTRCIRCSDFVDRYVEYDNVLIVLDLVLHKNPAYRHLL  1e-07      54.3       129
2: CVECGKSVQQVFKMFGKGNIRLSRCKSCHSISDKYVEYEFVLIFIDLLLHKTQVYRHLIFNRL  8e-14      72.0       175
```

By performing this `gi` clipping the initial `gi id` returned by blasting against the `NCBI nr` database `gi|139291046|gb|ECE46929.1|` is clipped to the `gi id 139291046` for which corresponding `taxonomy ids` can be extracted from NCBI Taxonomy.

Storing very large hit tables in a SQLite database with advanced_blast()

The `advanced_blast()` function can also store the BLAST output CSV file in a SQLite database. This works only with `dplyr` version >= 0.3 . To store the BLAST output in an SQLite database and to receive an SQLite connection as specified by `tbl()` in dplyr please use the `sql_database = TRUE` argument.

```
# using a SQLite database to store the BLAST output and
# select some example data

library(dplyr)

sqlE <- advanced_blast(
        query_file      = system.file('seqs/ortho_thal_cds.fasta', package = 'orthologr'),
        subject_file    = system.file('seqs/ortho_lyra_cds_1000.fasta', package = 'orthologr'),
        seq_type        = "cds",
        blast_algorithm = "blastp",
        blast_params    = "-evalue 1E-5 -num_threads 1",
        sql_database    = TRUE
        )

head(sqlE)
```

```
# look at the structure of sqlE
glimpse(sqlE)
```

```
Variables:
$ V1  (chr) "AT1G01010.1", "AT1G01010.1", "AT1G01010.1", "AT1G01010.1", "AT1G01020.1...
$ V2  (chr) "333554|PACid:16033839", "909883|PACid:16051052", "311334|PACid:16035114...
$ V3  (dbl) 73.99, 34.46, 43.64, 33.75, 91.06, 95.54, 91.98, 100.00, 89.51, 64.52, 9...
$ V4  (int) 347, 122, 72, 54, 224, 343, 1812, 213, 580, 40, 348, 268, 117, 271, 29, ...
$ V5  (int) 469, 354, 165, 160, 246, 359, 1970, 213, 648, 62, 366, 338, 356, 300, 95...
$ V6  (int) 80, 162, 81, 80, 22, 12, 85, 0, 58, 22, 14, 68, 232, 22, 66, 91, 47, 8, ...
$ V7  (int) 8, 12, 4, 6, 0, 2, 10, 0, 5, 0, 2, 1, 3, 2, 0, 3, 0, 3, 0, 0, 5, 5, 1, 6...
$ V8  (int) 1, 1, 1, 6, 1, 1, 6, 1, 1, 13, 1, 1, 16, 1, 105, 109, 110, 1, 1, 1, 32, ...
$ V9  (int) 430, 343, 162, 153, 246, 359, 1910, 213, 646, 74, 366, 338, 365, 294, 19...
$ V10 (int) 1, 1, 1, 10, 1, 1, 2, 1, 1, 1048, 1, 1, 8, 1, 336, 13, 76, 1, 1, 1, 20, ...
$ V11 (int) 466, 295, 156, 155, 246, 355, 1963, 213, 640, 1109, 362, 336, 362, 299, ...
$ V12 (dbl) 0e+00, 7e-43, 1e-32, 8e-21, 3e-164, 0e+00, 0e+00, 1e-158, 0e+00, 2e-22, ...
$ V13 (dbl) 627.0, 152.0, 123.0, 87.8, 454.0, 698.0, 3704.0, 437.0, 1037.0, 98.2, 69...
$ V14 (int) 1617, 384, 309, 216, 1169, 1801, 9604, 1125, 2681, 243, 1797, 1409, 551,...
$ V15 (int) 100, 80, 38, 34, 100, 100, 99, 100, 100, 10, 100, 92, 96, 100, 48, 48, 2...
```

This way you can perfom filtering steps on very large tables.

```r
# select all rows that have an evalue of zero
filter(sqlE, V12 > 1e-15)
```

```r
# select the best hit using the evalue criterion
sqlE %>% group_by(V1) %>% do(best_hit_eval = min(V12))
```

```
Source: sqlite 3.8.6 [blast_sql_db.sqlite3]
From: <derived table> [?? x 2]

            V1 best_hit_eval
1  AT1G01010.1        0e+00
2  AT1G01020.1       3e-164
3  AT1G01030.1        0e+00
4  AT1G01040.1        0e+00
5  AT1G01050.1       1e-158
6  AT1G01060.3        0e+00
7  AT1G01070.1        0e+00
8  AT1G01080.1       5e-177
9  AT1G01090.1        0e+00
10 AT1G01110.2        0e+00
..         ...          ...
```

Since the BLAST output does not return a CSV header, the corresponding SQLite table the BLAST output in columns named V1 - V15. In SQLite, renaming the columns can only be done by creating a new table which is very time consuming. The dplyr::rename() function does not support to rename multiple columns and therefore this “not so elegent way” is implemented in the current version of `orthologr`. Future versions will tackle this issue.

The set_blast() function

The `set_blast()`function reads a file storing a specific sequence type, such as “cds”, “protein”, or “dna” in a standard sequence file format such as “fasta”, etc. and depending of the makedb parameter either creates a blast-able database, or returns the corresponding protein sequences as data.table object for further BLAST searches.

```
head(set_blast(file =
 system.file('seqs/ortho_thal_cds.fasta',
            package = 'orthologr'))[[1]] , 2)
```

The advanced_makedb() function

The `advanced_makedb` function provides a simple, but powerful interface between the R language and `makeblastdb`. You can specify the `params` argument to pass all parameters defined for `makeblastdb` to the corresponding `makeblastdb` call.

```
# make the A. thaliana genome to a blast-able database
advanced_makedb(
database_file = system.file('seqs/ortho_thal_aa.fasta', package = 'orthologr'),
params        = "-input_type fasta -dbtype prot -hash_index"
)
```

```
Building a new DB, current time: 11/10/2014 16:56:58
New DB name:   _blast_db/ortho_thal_aa.fasta
New DB title:  _blast_db/ortho_thal_aa.fasta
Sequence type: Protein
Keep Linkouts: T
Keep MBits: T
Maximum file size: 1000000000B
Adding sequences from FASTA; added 20 sequences in 0.0583858 seconds.
```

Performing Sequence Alignements

The `orthologr` package provides multiple functions to perform pairwise and multiple sequence alignments. The following functions are implemented in `orthologr`:

- `multi_aln()` : Perform Multiple Sequence Alignments
- `pairwise_aln()` : Perform Pairwise Sequence Alignments
- `codon_aln()` : Perform Codon Alignments

Getting Started

Prior to be able to use all sequence alignment functions implemented in `orthologr` you need to install corresponding alignment tools of interest. The above mentioned functions provide interfaces to the following alignment programs:

The `multi_aln()` function

- **ClustalW** : Advanced multiple alignment tool of nucleic acid and protein sequences
- **T_Coffee** : A collection of tools for processing multiple sequence alignments of nucleic acids and proteins as well as their 3D structure
- **MUSCLE** : Fast and accurate multiple alignment tool of nucleic acid and protein sequences
- **ClustalO** : Fast and scalable multiple alignment tool for nucleic acid and protein sequences that is also capable of performing HMM alignments

- **MAFFT** : A tool for multiple sequence alignment and phylogeny

The easiest way to use the `multi_aln()` function is to store the corresponding multiple sequence alignment tools in the default execution `PATH` of you system (e.g. `/usr/local/bin` on UNIX machines).

You can test whether the corresponding multiple sequence alignment tool can be executed from the default `PATH` by running:

ClustalW

- MacOS: `system("clustalw2 -help")`
- Linux: `system("clustalw -help")`
- Windows: `system("clustalw2.exe -help")`

In case everything is installed appropriately, you should see:

```
CLUSTAL 2.1 Multiple Sequence Alignments


                DATA (sequences)

-INFILE=file.ext                             :input sequences.
-PROFILE1=file.ext  and  -PROFILE2=file.ext  :profiles (old alignment).


                VERBS (do things)

-OPTIONS            :list the command line parameters
-HELP  or -CHECK    :outline the command line params.
-FULLHELP           :output full help content.
-ALIGN              :do full multiple alignment.
```

Perform A Multiple Alignment Using ClustalW

The `multi_aln()` function takes a fasta file storing the genes (proteins) that shall be aligned. The `tool` argument specifies the alignment tool that shall be used to perform a multiple sequence alignment (in this case `tool = clustalw`). The `get_aln` argument specifies whether or not the alignment shall be printed out to the console. In case `get_aln = FALSE`, the corresponding alignment file is stored in the `file.path(tempdir(),_alignment,multi_aln)` directory.

```
# in case the default execution path of clustalw runs properly on your system
multi_aln(
file    = system.file('seqs/aa_seqs.fasta', package = 'orthologr'),
tool    = "clustalw",
get_aln = TRUE
)
```

It is also possible to pass additional parameters to the ClustalW call:

```r
# running clustalw using additional parameters
# details: system("clustalw2 -help")
multi_aln(
file     = system.file('seqs/aa_seqs.fasta', package = 'orthologr'),
tool     = "clustalw",
get_aln = TRUE,
params  = "-PWMATRIX=BLOSUM -TYPE=PROTEIN"
)
```

T_COFFEE

```r
system("t_coffee -version")
```

In case everything is installed appropriately, you should see:

```
PROGRAM: T-COFFEE Version_11.00.8cbe486 (2014-08-12 21:55:14 - Revision 8cbe486 - Build 470)
```

MUSCLE

```r
system("muscle -help")
```

In case everything is installed appropriately, you should see:

```
MUSCLE v3.8.31 by Robert C. Edgar

http://www.drive5.com/muscle
This software is donated to the public domain.
Please cite: Edgar, R.C. Nucleic Acids Res 32(5), 1792-97.


Basic usage

    muscle -in <inputfile> -out <outputfile>

Common options (for a complete list please see the User Guide):

    -in <inputfile>    Input file in FASTA format (default stdin)
    -out <outputfile>  Output alignment in FASTA format (default stdout)
    -diags             Find diagonals (faster for similar sequences)
    -maxiters <n>      Maximum number of iterations (integer, default 16)
    -maxhours <h>      Maximum time to iterate in hours (default no limit)
    -html              Write output in HTML format (default FASTA)
    -msf               Write output in GCG MSF format (default FASTA)
    -clw               Write output in CLUSTALW format (default FASTA)
    -clwstrict         As -clw, with 'CLUSTAL W (1.81)' header
    -log[a] <logfile>  Log to file (append if -loga, overwrite if -log)
    -quiet             Do not write progress messages to stderr
    -version           Display version information and exit
```

```
Without refinement (very fast, avg accuracy similar to T-Coffee): -maxiters 2
Fastest possible (amino acids): -maxiters 1 -diags -sv -distance1 kbit20_3
Fastest possible (nucleotides): -maxiters 1 -diags
```

```r
# in case the default execution path of muscle runs properly on your system
multi_aln(
file    = system.file('seqs/aa_seqs.fasta', package = 'orthologr'),
tool    = "muscle",
get_aln = TRUE
)
```

ClustalO

```r
system("clustalo --help")
```

MAFFT

```r
system("mafft -help")
```

In case everything is installed appropriately, you should see:

```
------------------------------------------------------------------------------
  MAFFT v7.187 (2014/10/02)
  http://mafft.cbrc.jp/alignment/software/
  MBE 30:772-780 (2013), NAR 30:3059-3066 (2002)
------------------------------------------------------------------------------
High speed:
  % mafft in > out
  % mafft --retree 1 in > out (fast)

High accuracy (for <~200 sequences x <~2,000 aa/nt):
  % mafft --maxiterate 1000 --localpair  in > out (% linsi in > out is also ok)
  % mafft --maxiterate 1000 --genafpair  in > out (% einsi in > out)
  % mafft --maxiterate 1000 --globalpair in > out (% ginsi in > out)

If unsure which option to use:
  % mafft --auto in > out

--op # :         Gap opening penalty, default: 1.53
--ep # :         Offset (works like gap extension penalty), default: 0.0
--maxiterate # : Maximum number of iterative refinement, default: 0
--clustalout :   Output: clustal format, default: fasta
--reorder :      Outorder: aligned, default: input order
--quiet :        Do not report progress
--thread # :     Number of threads (if unsure, --thread -1)
```

The `multi_aln()` function

The `multi_aln()` function is an interface function between R and common multiple sequence alignment tools. When working with this function a new folder named `_alignment` is being created and stores the multiple alignment returned by the corresponding alignment tool. The argument `get_aln = TRUE` allows to work with the multiple alignment generated by the corresponding alignment tool within the current R session.

This small pairwise alignment example shall illustrate how the `multi_aln()` output can be used:

```r
multi_aln(
        system.file('seqs/aa_seqs.fasta', package = 'orthologr'),
        tool = "clustalw",
        get_aln = TRUE
        )
```

```
$nb
[1] 2

$nam
[1] "AT1G01010.1"           "333554|PACid_16033839"

$seq
$seq[[1]]
[1] "medqvgfgfrpndeelvghylrnkiegntsrdvevaisevnicsydpwnlrfqskyksrdamwyffsrrennk
gnrqsrttvsgkwkltgesvevkdqwgfcsegfrgkighkrvlvfldgrypdktksdwvihefhydllpehqrtyvic
rleykgddadilsayaidptpafvpnmtssagsvvnqsrqrnsgsyntyseydsanhgqqfnensnimqqqplqgsfn
plleydfanhggqwlsdyidlqqqvpylapyenesemiwkhvieenfeflvdertsmqqhysdhrpkkpvsgvlpdds
sdtetgsmifedtssstdsvgssdepghtriddipslniieplhnykaqeqpkqqskekvissqksecewkmaedsik
ippstntvkqswivlenaqwnylknmiigvllfisviswiilvg"

$seq[[2]]
[1] "--------maasehrcvgcgfr---------------vkslfiqyspgnirlmk-------------------
---------------cgnckevadey----------iecermiifid---------------------lilhrpkvyr
hvlynainpetvniqhllwklvfvyllldsyrslllrrtdeess----------------fshssvlisikvligvls
anaafifs----------------------------------------faiaakgllnevs---rgreimlgicissy
fkifllamlvwefp----------------msvifivdilvltsnsmalkvmtestmtrciavcliahlvrfsvgqif
ep-------tifltqfgslmqylsylfrtv--------------"


$com
[1] NA

attr(,"class")
[1] "alignment"
```

Furthermore, multiple alignments are returned as follows:

```r
multi_aln(
        system.file('seqs/multi_aln_example.fasta', package = 'orthologr'),
        tool = "clustalw",
        get_aln = TRUE
        )
```

```
 CLUSTAL 2.1 Multiple Sequence Alignments
```

```
$nb
[1] 4

$nam
[1] "AT1G01010.1|PACid_19656964"      "Thhalv10006531m|PACid_20187082"
[3] "Bra032623|PACid_22715924"        "311315|PACid_16059488"

$seq
$seq[[1]]
[1] "-----------------------------------------------------------------
----------------------------------------------------------------------
----------------------------------------------------------------------
----------------------------------------------------------------------
----------------------------------------------------------------------
----------------------------------------------------------------------
----------------------------------------------------------------------
----------------------------------------------------------------------
----------------------------------------------------------------------
----------------------------------------------------------------------
----------------------------------------------------------------------
----------------------------------------------------------------------
----------------------------------------------------------------------
----------------------------------------------------------------------
----------------------------------------------------------------------
----------------------------------------------------------------------
-----------------------------------------medqvgfg------------------frp
ndeelvghylrnkiegntsr----------dvevaisevnicsydp---------wnlrfqskyksrd--
-------amwyffsrre------------nnkgnrqsrttvsgkwkltgesvevkdqwgfcseg------
--frgkig-------------------------------------------hkrvlvfldgrypdktksd
wvihefhydllpehqr-----------------tyvicr--leykgddadilsayaidptpafvpnmtss
agsvvnqsrqrnsgsyntyseydsanhgqqfnensn-imqqqplqgsfnplleydfanhg-------gqw
l----------------------------sdyidlqqqvpylapye------------------------
-----------------------------------nesemiwkhvieenfeflvdertsmqqhysdhrpk
kpvsgvlp----------------ddssdtetgsmifedtssstds------------------------
---------------------------------vgssdepghtriddipslniieplhnykaqeqpkqqs
kekvissqksecewkmaeds---------ikippstntvkqswivlenaqwnylknmiigvllfisvisw
iilvg-------------------------------"

$seq[[2]]
[1] "mvmederrgdikppsywldacediscdliddlvsdfdpssvavaesvdengvnndffggidhild
sikngggglpnrahingvsetnsqringnsevseaaqliagettvsvkgnvlqkcggkrdevskeegeknr
krarvcsyqrersnlsgrgqansregdrfmnrkrtrnwdeaghnkrrdgynyrrdgrdreargywerdkv
gsnelvyrsgtweadherdlkkesgrnresdekaeenkskpeehkekvveeqarryqldvleqakaknti
afletgagktliailliksihkdltsqnrkmlsvflvpkvplvyqqaevirnqtcfqvghycgemgqdfw
darrwqrefeskqvlvmtaqillnilrhsiirmeainllildechhavkkhpyslvmsefyhttpkdkrp
aifgmtaspvnlkgvssqvdcaikirnletkldstvctikdrkelekhvpmpseivveydkaatmwslhe
kikqmiaaveeaaqassrkskwqfmgardagakdelrqvygvsertesdgaanlihklrainytlaelgq
wcaykvaqsfltalqsdervnfqvdvkfqesylsevvsllqcellegaaaekavaelskpengnandeie
egelpddhvvsggehvdkvigaavadgkvtpkvqsliklllkyqhtadfraivfvervvaalvlpkvfae
lpslgfircasmighnnsqemkssqmqdtiskfrdgqvtllvatsvaeegldirqcnvvmrfdlaktvla
yiqsrgrarkpgsdyilmverenvshaaflrnarnseetlrkeaiertdlshlkdssrlisidavpgtvy
kveatgamvslnsavglihfycsqlpgdryailrpefsmvkhekpgghteyscrlqlpcnapfeilegpv
cssmrlaqqavclagckklhemgaftdmllpdkgsgqdaekadqddegepipgtarhrefypegvadvlk
gewilsgkeicessklfhlymysvrcvdsgvskdpfltevsefavlfgneldaevlsmsmdlyvaramit
```

```
kaslafrgslditesqlssikkfhvrlmsivldvdvepsttpwdpakaylfvpvadnssaepikginwel
vekitkttvwdnplqrarpdvylgtnertlggdrreygfgklrhnigfgqkshptygirgavasfdvvra
sgllpvrdalekevegdlsqgklmmadgcmvaenllgkivtaahsgkrfyvdsicydmsaetsfprkegy
lgpleyntyadyykqkygvdlsckqqplikgrgvsycknllsprfeqsgesetildktyyvflppelcvv
hplsgslvrgaqrlpsimrrvesmllavqlknlisypiptskilealtaascqetfcyeraellgdaylk
wivsrflflkypqkhegqltrmrqqmvsnlvlyqyalvkglqsyiqadrfapsrwsapgvppvydedtkd
ggssffdeeekpegnkdvfedgemedgelegdlssyrvlssktladvvealigvyyveggktaanhlmkw
igihveddpeetegsvkpvynvpesvlksidfvgleralkyeftekgllveaithasrpssgvscyqrle
fvgdavldhlitrhlfftytslppgrltdlraaavnnenfarvavkhklhlylrhgssalekqirdfvke
vltesskpgfnsfglgdckapkvlgdivesiagaifldsgkdttaawkvfqpllqpmvtpetlpmhpvre
lqercqqqaegleykasrsgntatvevfidgvqvgaaqnpqkkmaqklaarnalaalkekeaeeskkkqa
ngnaagenqddnengnkkngnqtftrqtlndiclrknwpmpsyrcvkeggpahakrftfgvrvntsdrgw
tdecigepmpsvkkakdsaailllellnktys----"

$seq[[3]]
[1] "-----------------------------------------------------------------
----------------------------------------------------------------------
----------------------------------------------------------------------
----------------------------------------------------------------------
----------------------------------------------------------------------
----------------------------------------------------------------------
----------------------------------------------------------------------
----------------------------------------------------------------------
----------------------------------------------------------------------
----------------------------------------------------------------------
----------------------------------------------------------------------
----------------------------------------------------------------------
----------------------------------------------------------------------
----------------------------------------------------------------------
----------------------------------------------------------------------
----------------------------------------------------------------------
----------------------------------------------------------------------
----------------------------------------------------------------------
----------------------------------------------------------------------
----------------------------------------------------------------------
-----------------------------------------------mqqppqmfpmapsmpptnitteq
iqkyleenkklimaimenqnlgklaecaqyqallqknlmylaaiadaqpppstagatppp----------
---------------------------------amasqmgaphpg-------------------------
---------------------------------------------mqppsyfmqhp------qasgmaqq
appagifp----------------------prgplqfgsphqlqdp------------------------
-----------------------------------qqqhmhqqamqghmgmrpmginnnngmqhqmqqqp
etslggsaanvgirggkqdg-------------------------adgqgkddgk---------------
------------------------------------"

$seq[[4]]
[1] "-----------------------------------------------------------------
----------------------------------------------------------------------
----------------------------------------------------------------------
----------------------------------------------------------------------
----------------------------------------------------------------------
----------------------------------------------------------------------
----------------------------------------------------------------------
----------------------------------------------------------------------
----------------------------------------------------------------------
```

```
----------------------------------------------------------------------
----------------------------------------------------------------------
----------------------------------------------------------------------
----------------------------------------------------------------------
----------------------------------------------------------------------
--------------------------------------------------------mlslnmrteienlw
vfalaskfnifmqehfaslllaiaitwctltivfwstpggpawg--------------------------
--------------------------kyfftrrfsslghnrksknlipgprgfplvgsmslrsshvahqr
iasvaemsnakrlmafslgdtkvvvtchpdvakeilnssvfadrpvdetayg---lmfnramgfapngty
wrtlrrlssnhlfnpkqikrseeqrrviatqmvnafarnaksafavrdllktaslcnmmglvfg------
-----------------------------------------------------reyelesnnnveseclk
glveegydllg-----------------------tlnwtdhlpwlagldfqqirfrcsqlvpkvnlllsr
iihehyatgnfldvllslqrseklsdsdivavlwemifrgtdtvavliewvlarialhpk----------
----------------------------------vqstvhdeldr-------------------------
---------------------------------------------vvgrsrtvdesdlpsltyltamike
vlr--lhp-----------------------pgpllswarlsitdt------------------------
------------------------------------tvdgyhvpagttamvnmwaiardphvwedplefk
perfvakdgeaefsvfgsdlr---------------lapfgsgkrvcpgknlglttvsfwvatllhefew
lpsveanppdlsevlrlscemacplivnvsprrksv"


$com
[1] NA

attr(,"class")
[1] "alignment"
```

Performing Divergence Stratigraphy

The `orthologr` package allows users to perform **Divergence Stratigraphy** for any query and subject organisms of interest.

Divergence Stratigraphy is the process of quantifying the selection pressure (in terms of protein evolutionary rate) acting on orthologous genes between closely related species. The resulting sequence divergence map (short divergence map), stores the divergence stratum in the first column and the `query_id` of inferred orthologous genes in the second column (Quint et al., 2012 *Nature*; Drost et al., 2015 *Mol. Biol. Evol.*; Drost et al., 2016 *Mol. Biol. Evol.*; Introduction to myTAI).

The following Algorithm implemented in `divergence_stratigraphy()` defines **Divergence Stratigraphy** as method (see Drost et al., 2015):

1) Orthology Inference using BLAST best reciprocal hit ("RBH") based on blastp
2) Pairwise global amino acid alignments of orthologous genes using the Needleman-Wunsch algorithm
3) Codon alignments of orthologous genes using PAL2NAL
4) dNdS estimation using Comeron's method (1995)
5) Categorize estimated dNdS values into `divergence strata` (= deciles of all dNdS values)

In `orthologr` the Needleman-Wunsch algorithm, PAL2NAL and Comeron's method (1995) are already included in the `orthologr` package and do not have to be installed separately. Nevertheless, users need to make sure **they have BLAST installed on their machine before using the `divergence_stratigraphy()` function**.

Note: The following examples assume that the **BLAST** program is installed and stored in the default execution path `usr/local/bin`. In case users do not have **BLAST** installed yet or the following command in R produces a different output, please consult the Installation Vignette to corretly set up the **BLAST** program to perform **Divergence Stratigraphy**.

```
system("blastp -version")
```

```
blastp: 2.2.30+
Package: blast 2.2.30, build Oct 27 2014 17:10:51
```

Divergence Map Computations

In Drost et al., 2015 *Mol. Biol. Evol.* we define a `Divergence Map` as table storing the degree of selection pressure (= `divergence strata`) for each protein coding gene of a given query organism. In this case selection pressure was quantified by dNdS estimation (ratio of synonymous versus non-synonymous `codon -> amino acid` sequence substitution rates). The resulting dNdS values for all protein coding genes of the query organism are then categorized into deciles (10%-quantiles) allowing users to compare the results obtained from `Phylostratigraphy` with results obtained form `Divergence Stratigraphy`.

To perform **Divergence Stratigraphy** using `orthologr` users need to retrieve the following input files:

- a CDS file covering all protein coding genes of the query organism of interest
- a CDS file covering all protein coding genes of the subject organism of interest

Sequence Data Retrieval

In the following example, we will use *Arabidopsis thaliana* as query organism and *Arabidopsis lyrata* as subject organism.

First, we need to download the CDS sequences for all protein coding genes of *A. thaliana* and *A. lyrata*.

Option 1:

The CDS retrieval can be done using a `Terminal` or by manual downloading the files

- `Arabidopsis_thaliana.TAIR10.23.cds.all.fa.gz`
- `Arabidopsis_lyrata.v.1.0.23.cds.all.fa.gz`

```
# download CDS file of A. thaliana
curl ftp://ftp.ensemblgenomes.org/pub/
plants/release-23/fasta/arabidopsis_thaliana/
cds/Arabidopsis_thaliana.TAIR10.23.cds.all.fa.gz
-o Arabidopsis_thaliana.TAIR10.23.cds.all.fa.gz

# unzip the fasta file
gunzip -d Arabidopsis_thaliana.TAIR10.23.cds.all.fa.gz

# download CDS file of A. lyrata

curl ftp://ftp.ensemblgenomes.org/pub/plants/
release-23/fasta/arabidopsis_lyrata/cds/
Arabidopsis_lyrata.v.1.0.23.cds.all.fa.gz
```

```
-o Arabidopsis_lyrata.v.1.0.23.cds.all.fa.gz

# unzip the fasta file
gunzip -d Arabidopsis_lyrata.v.1.0.23.cds.all.fa.gz
```

When the download is finished you need to unzip the files.

Option 2:

We implemented the biomartr package to automate the process of performing biological data retrieval. The Sequence Retrieval Vignette stored in `biomartr` provides detailed use cases for the automation of biological sequence retrieval.

Note: Users need to make sure they have biomartr installed before running any `biomartr` functions.

Computation Time

Please note that performing **Divergence Stratigraphy** with two large genomes can take (even on a multicore machine) some time -> **up to several hours**. On a 4 core machine with 3.4 GHz i7 processors the computation time of generating a divergence map between *A. thaliana* and *A. lyrata* was **2.5-3 hours**.

The `comp_cores` argument implemented in the `divergence_stratigraphy()` function allows users to specify the number of cores they would like to use on their machine. The default value is `comp_cores = 1` which might take **10-12h** to execute. So users need to make sure that they use all cores available on their machine to speed up the computation time.

Running `divergence_stratigraphy()`

As mentioned earlier the `divergence_stratigraphy()` function is the main function to perform the **Divergence Stratigraphy** algorithm.

In `divergence_stratigraphy()` the `query_file` and `subject_file` arguments take an character string storing the path to the corresponding **fasta** files containing the CDS sequences of these organisms. Here the previously downloaded CDS sequence files of *A. thaliana* (= `query_file`) and *A. lyrata* (= `subject_file`) need to be specified. The `eval` is set to `1E-5` (default ; see Quint et al., 2012 *Nature*) and `BLAST best reciprocal hit` is used for orthology inference (see Drost et al., 2015). In case '`orthologr` is running on a multicore machine, users can set the `comp_cores` argument to any number of cores supported by their machine. The `clean_folders` argument indicates whether or not the internal folder structure should be deleted (cleaned) after processing is finished. In this case all output files generated by `divergence_stratigraphy` (stored in `tempdir()`) will be removed after the `Divergence Map` was returned. The `quiet` argument indicates whether or not a successful interface call should be printed out to the console (`quiet = FALSE`) or not (`quiet = TRUE`).

```r
library(orthologr)

# compute the divergence map of A. thaliana
Athaliana_DM <- divergence_stratigraphy(
        query_file      = "path/to/Arabidopsis_thaliana.TAIR10.23.cds.all.fa",
        subject_file    = "path/to/Arabidopsis_lyrata.v.1.0.23.cds.all.fa",
        eval            = "1E-5",
        ortho_detection = "RBH",
        comp_cores      = 1,
        quiet           = TRUE,
        clean_folders   = TRUE
        )
```

Before running `divergence_stratigraphy()` with two complete genomes, users can first run a test **Divergence Stratigraphy** with 20 example genes that are stored in the `orthologr` package:

```
library(orthologr)

# performing standard divergence stratigraphy
divergence_stratigraphy(
query_file      = system.file('seqs/ortho_thal_cds.fasta', package = 'orthologr'),
subject_file    = system.file('seqs/ortho_lyra_cds.fasta', package = 'orthologr'),
eval            = "1E-5",
ortho_detection = "RBH",
dnds.threshold  = 2,
comp_cores      = 1,
quiet           = TRUE,
clean_folders   = TRUE
)
```

```
   divergence_strata    query_id
1                 10 AT1G01010.1
2                  9 AT1G01020.1
3                  5 AT1G01030.1
4                  4 AT1G01040.1
5                  1 AT1G01050.1
6                  9 AT1G01060.3
7                  6 AT1G01070.1
8                  8 AT1G01080.1
9                  2 AT1G01090.1
10                 7 AT1G01110.2
11                 2 AT1G01120.1
12                 3 AT1G01140.3
13                10 AT1G01150.1
14                 8 AT1G01160.1
15                 1 AT1G01170.2
16                 6 AT1G01180.1
17                 7 AT1G01190.1
18                 4 AT1G01200.1
19                 5 AT1G01210.1
20                 3 AT1G01220.1
```

The resulting output is a `Divergence Map` of the 20 example genes.

To save corresponding `Divergenec Maps` to a hard drive users can pass the resulting `divergence_stratigraphy()` output to a variable and then use the `write.table()` function implemented in R to store the `Divergenec Map` as `*.csv` file.

```
Athaliana_DM <- divergence_stratigraphy(...)

write.table(
x         = Athaliana_DM,
file      = "Ath_Aly_DivergenceMap.csv",
sep       = ";",
col.names = TRUE,
row.names = FALSE,
quote     = FALSE
)
```

This way `write.table()` will store the `Divergence Map` to the users current working directory (= `getwd()`).

Specifying the arguments in `divergence_stratigraphy()`

Several argument combinations can be specified in `divergence_stratigraphy()` (see `Arguments` in `?divergence_stratigraphy`). This section introduces additional output options of `divergence_stratigraphy()`.

Example: `blast_path`

Sometimes the machine users are working on does not allow them to install **BLAST** in the default execution path `usr/local/bin`. For this purpose the `blast_path` argument is implemented in `divergence_stratigraphy()`. This argument takes an character string specifying the `PATH` to the user's `blastp` execution file that is stored in a different place than `usr/local/bin`.

The following example shows a possible specification of `blast_path`.

```
library(orthologr)

# performing standard divergence stratigraphy
divergence_stratigraphy(
query_file      = system.file('seqs/ortho_thal_cds.fasta', package = 'orthologr'),
subject_file    = system.file('seqs/ortho_lyra_cds.fasta', package = 'orthologr'),
blast_path      = "here/the/path/to/blastp",
eval            = "1E-5",
ortho_detection = "RBH",
dnds.threshold  = 2,
comp_cores      = 1,
quiet           = TRUE,
clean_folders   = TRUE
)
```

```
   divergence_strata    query_id
1                 10 AT1G01010.1
2                  9 AT1G01020.1
3                  5 AT1G01030.1
4                  4 AT1G01040.1
5                  1 AT1G01050.1
6                  9 AT1G01060.3
7                  6 AT1G01070.1
8                  8 AT1G01080.1
9                  2 AT1G01090.1
10                 7 AT1G01110.2
11                 2 AT1G01120.1
12                 3 AT1G01140.3
13                10 AT1G01150.1
14                 8 AT1G01160.1
15                 1 AT1G01170.2
```

```
16                  6 AT1G01180.1
17                  7 AT1G01190.1
18                  4 AT1G01200.1
19                  5 AT1G01210.1
20                  3 AT1G01220.1
```

Example: `ds.values`

As defined earlier, a `Divergence Map` stores the `divergence strata` for protein coding genes of a `query` organism. However, `divergence strata` are based on `dNdS` values that were categorized into deciles. For this reason it is not possible to map a `divergence stratum` value to the exact initial `dNdS` value. So in case users are interested in the the exact `dNdS` value of protein coding genes, they can specify the `ds.values` argument in `divergence_stratigraphy()` allowing them to retrieve a `dNdS Map` instead of a `Divergence Map`. For this purpose users need to set `ds.values = FALSE`.

```
library(orthologr)

# performing standard divergence stratigraphy
# but receive a dNdS Map instead of a Divergence Map
divergence_stratigraphy(
query_file      = system.file('seqs/ortho_thal_cds.fasta', package = 'orthologr'),
subject_file    = system.file('seqs/ortho_lyra_cds.fasta', package = 'orthologr'),
eval            = "1E-5",
ortho_detection = "RBH",
ds.values       = FALSE,
dnds.threshold  = 2,
comp_cores      = 1,
quiet           = TRUE,
clean_folders   = TRUE
)
```

```
      dNdS    query_id
1  0.41950 AT1G01010.1
2  0.38790 AT1G01020.1
3  0.11850 AT1G01030.1
4  0.11560 AT1G01040.1
5  0.00000 AT1G01050.1
6  0.39670 AT1G01060.3
7  0.17280 AT1G01070.1
8  0.32170 AT1G01080.1
9  0.04174 AT1G01090.1
10 0.26620 AT1G01110.2
11 0.02317 AT1G01120.1
12 0.04324 AT1G01140.3
13 0.64120 AT1G01150.1
14 0.37310 AT1G01160.1
15 0.00000 AT1G01170.2
16 0.16830 AT1G01180.1
17 0.17730 AT1G01190.1
18 0.11370 AT1G01200.1
19 0.13420 AT1G01210.1
20 0.10230 AT1G01220.1
```

The corresponding output now stores `dNdS` values instead of `DS` values in the first column.

Example: `subject.id`

Although the `Divergence Map` standard is specified as storing `DS` values in the first column and `GeneIDs` in the second column, in some cases it is important to store the GeneIDs of orthologous genes in the `subject` organism. The `subject.id` argument implemented in `divergence_stratigraphy()` allows users to retrieve the GeneIDs of the orthologous genes of the `subject` organism. For this purpose users need to specify `subject.id = TRUE`.

```
# receive a Divergence Map with DS | query GeneID | orthologous subject GeneID
divergence_stratigraphy(
query_file      = system.file('seqs/ortho_thal_cds.fasta', package = 'orthologr'),
subject_file    = system.file('seqs/ortho_lyra_cds.fasta', package = 'orthologr'),
eval            = "1E-5",
ortho_detection = "RBH",
comp_cores      = 1,
quiet           = TRUE,
clean_folders   = TRUE,
subject.id      = TRUE
)
```

```
   DS    query_id            subject_id
1  10 AT1G01010.1 333554|PACid:16033839
2   9 AT1G01020.1 470181|PACid:16064328
3   5 AT1G01030.1 470180|PACid:16054974
4   4 AT1G01040.1 333551|PACid:16057793
5   1 AT1G01050.1 909874|PACid:16064489
6   9 AT1G01060.3 470177|PACid:16043374
7   6 AT1G01070.1 918864|PACid:16052578
8   8 AT1G01080.1 909871|PACid:16053217
9   2 AT1G01090.1 470171|PACid:16052860
10  7 AT1G01110.2 333544|PACid:16034284
11  2 AT1G01120.1 918858|PACid:16049140
12  3 AT1G01140.3 470161|PACid:16036015
13 10 AT1G01150.1 918855|PACid:16037307
14  8 AT1G01160.1 918854|PACid:16044153
15  1 AT1G01170.2 311317|PACid:16052302
16  6 AT1G01180.1 909860|PACid:16056125
17  7 AT1G01190.1 311315|PACid:16059488
18  4 AT1G01200.1 470156|PACid:16041002
19  5 AT1G01210.1 311313|PACid:16057125
20  3 AT1G01220.1 470155|PACid:16047984
```

The resulting output now shows DS values, query GeneIDs, and orthologous subject GeneIDs.

A similar output can be generated for `dNdS` values instead of `DS` values by specifying `ds.values = FALSE` and `subject.id = TRUE`.

```
# receive a dNdS Map with dNdS | query GeneID | orthologous subject GeneID
divergence_stratigraphy(
query_file      = system.file('seqs/ortho_thal_cds.fasta', package = 'orthologr'),
subject_file    = system.file('seqs/ortho_lyra_cds.fasta', package = 'orthologr'),
eval            = "1E-5",
ortho_detection = "RBH",
comp_cores      = 1,
ds.values        = FALSE,
quiet            = TRUE,
clean_folders    = TRUE,
subject.id       = TRUE
)
```

```
      dNdS    query_id            subject_id
1  0.41950 AT1G01010.1 333554|PACid:16033839
2  0.38790 AT1G01020.1 470181|PACid:16064328
3  0.11850 AT1G01030.1 470180|PACid:16054974
4  0.11560 AT1G01040.1 333551|PACid:16057793
5  0.00000 AT1G01050.1 909874|PACid:16064489
6  0.39670 AT1G01060.3 470177|PACid:16043374
7  0.17280 AT1G01070.1 918864|PACid:16052578
8  0.32170 AT1G01080.1 909871|PACid:16053217
9  0.04174 AT1G01090.1 470171|PACid:16052860
10 0.26620 AT1G01110.2 333544|PACid:16034284
11 0.02317 AT1G01120.1 918858|PACid:16049140
12 0.04324 AT1G01140.3 470161|PACid:16036015
13 0.64120 AT1G01150.1 918855|PACid:16037307
14 0.37310 AT1G01160.1 918854|PACid:16044153
15 0.00000 AT1G01170.2 311317|PACid:16052302
16 0.16830 AT1G01180.1 909860|PACid:16056125
17 0.17730 AT1G01190.1 311315|PACid:16059488
18 0.11370 AT1G01200.1 470156|PACid:16041002
19 0.13420 AT1G01210.1 311313|PACid:16057125
20 0.10230 AT1G01220.1 470155|PACid:16047984
```

Example: `dnds.threshold`

`Divergence Strata` are obtained by categorizing dNdS values into deciles. For decilation the range of dNdS values is important. The `dnds.threshold` defines the upper level cut off of dNdS values. Since dNdS values are in the range [0, +Inf] a upper threshold needs to be specified. The default value for `dnds.threshold` in `divergence_stratigraphy()` is `dnds.threshold = 2` due to the interpretation of dNdS values for predicting sequence evolution (dNdS < 1 -> negative selection; dNdS $= 1$ -> neutral selection; dNdS > 1 -> positive selection). Hence, all dNdS values `> 1` predict positive selection. In my experience of computing dNdS values between hundreds of pairwise species comparisons covering all evolutionary distances, dNdS values of orthologous genes rarely take values > 2. Nevertheless, in case you wish to extend or reduce the upper threshold for dNdS values, you can specify the `dnds.threshold` in `divergence_stratigraphy()`.

```
# upper threshold for dNdS: dnds.threshold = 5
divergence_stratigraphy(
query_file       = system.file('seqs/ortho_thal_cds.fasta', package = 'orthologr'),
subject_file     = system.file('seqs/ortho_lyra_cds.fasta', package = 'orthologr'),
eval             = "1E-5",
ortho_detection = "RBH",
ds.values        = FALSE,
dnds.threshold   = 5,
comp_cores       = 1,
quiet            = TRUE,
clean_folders    = TRUE
)
```

Example: `ortho_detection`

According to Drost et al., 2015 *Mol. Biol. Evol.* the **Divergence Stratigraphy** algorithm performs **BLAST** best reciprocal hit (RBH) as orthology inference method. Despite this convention, the `ortho_detection` argument allows users to perform orthology inference within the **Divergence Stratigraphy** algorithm that is based on any orthology inference method implemented in `orthologr` (see `?orthologs` or Orthology Inference Vignette for details). For example in Quint et al., 2012 *Nature* instead of using **BLAST** best reciprocal hit, the method **BLAST** best hit (BH) was used to perform orthology inference within the **Divergence Stratigraphy** algorithm.

```
# orthology inference method: ortho_detection = "BH"
divergence_stratigraphy(
query_file      = system.file('seqs/ortho_thal_cds.fasta', package = 'orthologr'),
subject_file    = system.file('seqs/ortho_lyra_cds.fasta', package = 'orthologr'),
eval            = "1E-5",
ortho_detection = "BH",
ds.values       = TRUE,
dnds.threshold  = 2,
comp_cores      = 1,
quiet           = TRUE,
clean_folders   = TRUE
)
```

```
   DS    query_id
1  10 AT1G01010.1
2   9 AT1G01020.1
3   5 AT1G01030.1
4   4 AT1G01040.1
5   1 AT1G01050.1
6   9 AT1G01060.3
7   6 AT1G01070.1
8   8 AT1G01080.1
9   2 AT1G01090.1
10  7 AT1G01110.2
11  2 AT1G01120.1
12  3 AT1G01140.3
13 10 AT1G01150.1
14  8 AT1G01160.1
15  1 AT1G01170.2
16  6 AT1G01180.1
17  7 AT1G01190.1
18  4 AT1G01200.1
19  5 AT1G01210.1
20  3 AT1G01220.1
```

Skip `Divergence Stratigraphy` and Download Already Published `Divergence Maps`

Users can find a detailed list of published Phylostratigraphic Maps and Divergence Maps by following the link. This way the computation time of 3-4 h on a local machine for 2 genome comparisions can be skipped.

Combine a Divergence Map with Gene Expression Data

`Divergence Maps` can be used for a wide range of analyses. One example is to combine `Divergence Maps` with gene expression data to capture evolutionary signals in developmental transcriptomes (= `Phylotranscriptomics`; see Drost et al., 2015 *Mol. Biol. Evol.*). Performing phylotranscriptomic analyses based on an existing `Divergence Map` can easily be done by using the `myTAI` package. You can consult the Introduction to the myTAI package Vignette for more details.

Performing dN/dS Estimation

The dN/dS ratio quantifies the mode and strength of selection acting on a pair of orthologous genes. This selection pressure can be quantified by comparing synonymous substitution rates (dS) that are assumed to be neutral with nonsynonymous substitution rates (dN), which are exposed to selection as they change the amino acid composition of a protein (Mugal et al., 2013).

The `orthologr` package provides a function named `dNdS()` to perform dNdS estimation on pairs of orthologous genes. The `dNdS()` function takes the CDS files of two organisms of interest (`query_file` and `subject_file`) and computes the dNdS estimation values for orthologous gene pairs between these organisms.

Note: the following dNdS estimation methods are based on KaKs_Calculator:

- "NG": Nei, M. and Gojobori, T. (1986)
- "LWL": Li, W.H., et al. (1985)
- "LPB": Li, W.H. (1993) and Pamilo, P. and Bianchi, N.O. (1993)
- "MLWL" (Modified LWL), MLPB (Modified LPB): Tzeng, Y.H., et al. (2004)
- "YN": Yang, Z. and Nielsen, R. (2000)
- "MYN" (Modified YN): Zhang, Z., et al. (2006)

It is assumed that when you choose one of these dNdS estimation methods you have KaKs_Calculator installed on your machine and it can be executed from the default execution `PATH`.

The following pipeline resembles an example dNdS estimation procedure:

1) Orthology Inference: e.g. BLAST reciprocal best hit (RBH)
2) Pairwise sequence alignment: e.g. clustalw for pairwise amino acid sequence alignments
3) Codon Alignment: e.g. pal2nal program
4) dNdS estimation: e.g. Yang, Z. and Nielsen, R. (2000) (YN)

Note: it is assumed that when using `dNdS()` all corresponding multiple sequence alignment programs you want to use are already installed on your machine and are executable via either the default execution PATH or you specifically define the location of the executable program via the `aa_aln_path` or `blast_path` argument that can be passed to `dNdS()`. See the Sequence Alignments vignette for details.

The following example shall illustrate a dNdS estimation process.

```r
library(orthologr)

# get a dNdS table using:
# 1) reciprocal best hit for orthology inference (RBH)
# 2) clustalw for pairwise amino acid alignments
# 3) pal2nal for codon alignments
# 4) Yang, Z. and Nielsen, R. (2000) (YN) for dNdS estimation
# 5) single core processing 'comp_cores = 1'
dNdS(
query_file      = system.file('seqs/ortho_thal_cds.fasta', package = 'orthologr'),
subject_file    = system.file('seqs/ortho_lyra_cds.fasta', package = 'orthologr'),
ortho_detection = "RBH",
aa_aln_type     = "multiple",
aa_aln_tool     = "clustalw",
codon_aln_tool  = "pal2nal",
dnds_est.method = "YN",
comp_cores      = 1,
clean_folders   = TRUE,
quiet           = TRUE
)
```

```
       query_id            subject_id         dN        dS      dNdS method
 1: AT1G01010.1 333554|PACid_16033839 0.10581700 0.2844350 0.3720250     YN
 2: AT1G01020.1 470181|PACid_16064328 0.04164150 0.0951677 0.4375590     YN
 3: AT1G01030.1 470180|PACid_16054974 0.01664670 0.1163900 0.1430260     YN
 4: AT1G01040.1 333551|PACid_16057793 0.01421700 0.1314360 0.1081670     YN
 5: AT1G01050.1 909874|PACid_16064489         NA 0.2092450 0.0000000     YN
 6: AT1G01060.3 470177|PACid_16043374 0.04387800 0.1131710 0.3877130     YN
 7: AT1G01070.1 918864|PACid_16052578 0.02028020 0.0960773 0.2110820     YN
 8: AT1G01080.1 909871|PACid_16053217 0.03930610 0.0995795 0.3947210     YN
 9: AT1G01090.1 470171|PACid_16052860 0.00992436 0.2496940 0.0397461     YN
10: AT1G01110.2 333544|PACid_16034284 0.03292970 0.1293160 0.2546450     YN
11: AT1G01120.1 918858|PACid_16049140 0.00356132 0.1225410 0.0290623     YN
12: AT1G01140.3 470161|PACid_16036015 0.00582238 0.1354990 0.0429699     YN
13: AT1G01150.1 918855|PACid_16037307 0.13565500 0.1962460 0.6912480     YN
14: AT1G01160.1 918854|PACid_16044153 0.11558300 0.1929560 0.5990120     YN
15: AT1G01170.2 311317|PACid_16052302 0.00557175 0.2903370 0.0191906     YN
16: AT1G01180.1 909860|PACid_16056125 0.04065370 0.1557400 0.2610360     YN
17: AT1G01190.1 311315|PACid_16059488 0.02849220 0.1538610 0.1851810     YN
18: AT1G01200.1 470156|PACid_16041002 0.01983450 0.1512510 0.1311360     YN
19: AT1G01210.1 311313|PACid_16057125 0.02106910 0.1433630 0.1469630     YN
20: AT1G01220.1 470155|PACid_16047984 0.01530070 0.1446480 0.1057780     YN
```

The output includes `NA` values. To filter for `NA` values or a specific `dnds.threshold`, you can use the `filter_dNdS()` function. The `filter_dNdS()` function takes the output data.table returned by `dNdS()` and filters the output by the following criteria:

1) all dN values having an NA value are omitted
2) all dS values having an NA value are omitted
3) all dNdS values >= the specified `dnds.threshold` are omitted

```
library(orthologr)

# get dNdS estimated for orthologous genes between A. thaliana and A. lyrata
Ath_Aly_dnds <-
dNdS(
query_file      = system.file('seqs/ortho_thal_cds.fasta', package = 'orthologr'),
subject_file    = system.file('seqs/ortho_lyra_cds.fasta', package = 'orthologr'),
dnds_est.method = "YN",
comp_cores      = 1,
clean_folders   = TRUE,
quiet           = TRUE
)

# filter for:
# 1) all dN values having an NA value are omitted
# 2) all dS values having an NA value are omitted
# 3) all dNdS values >= 2 are omitted
filter_dNdS(Ath_Aly_dnds, dnds.threshold = 2)
```

```
      query_id            subject_id         dN        dS      dNdS method
1  AT1G01010.1 333554|PACid_16033839 0.10581700 0.2844350 0.3720250     YN
2  AT1G01020.1 470181|PACid_16064328 0.04164150 0.0951677 0.4375590     YN
3  AT1G01030.1 470180|PACid_16054974 0.01664670 0.1163900 0.1430260     YN
4  AT1G01040.1 333551|PACid_16057793 0.01421700 0.1314360 0.1081670     YN
5  AT1G01060.3 470177|PACid_16043374 0.04387800 0.1131710 0.3877130     YN
6  AT1G01070.1 918864|PACid_16052578 0.02028020 0.0960773 0.2110820     YN
7  AT1G01080.1 909871|PACid_16053217 0.03930610 0.0995795 0.3947210     YN
8  AT1G01090.1 470171|PACid_16052860 0.00992436 0.2496940 0.0397461     YN
9  AT1G01110.2 333544|PACid_16034284 0.03292970 0.1293160 0.2546450     YN
10 AT1G01120.1 918858|PACid_16049140 0.00356132 0.1225410 0.0290623     YN
11 AT1G01140.3 470161|PACid_16036015 0.00582238 0.1354990 0.0429699     YN
12 AT1G01150.1 918855|PACid_16037307 0.13565500 0.1962460 0.6912480     YN
13 AT1G01160.1 918854|PACid_16044153 0.11558300 0.1929560 0.5990120     YN
14 AT1G01170.2 311317|PACid_16052302 0.00557175 0.2903370 0.0191906     YN
15 AT1G01180.1 909860|PACid_16056125 0.04065370 0.1557400 0.2610360     YN
16 AT1G01190.1 311315|PACid_16059488 0.02849220 0.1538610 0.1851810     YN
17 AT1G01200.1 470156|PACid_16041002 0.01983450 0.1512510 0.1311360     YN
18 AT1G01210.1 311313|PACid_16057125 0.02106910 0.1433630 0.1469630     YN
19 AT1G01220.1 470155|PACid_16047984 0.01530070 0.1446480 0.1057780     YN
```

Instead of using a multiple alignment tool for pairwise alignments you can also choose a global pairwise alignment of orthologous genes based on the Needleman-Wunsch algorithm. For this purpose the argument `aa_aln_type` must be set to `aa_aln_type = "pairwise"` and `aa_aln_tool = "NW"` for Needleman-Wunsch.

```
library(orthologr)

# get a dNdS table using:
# 1) reciprocal best hit for orthology inference (RBH)
# 2) pairwise amino acid alignments using Needleman-Wunsch
# 3) pal2nal for codon alignments
# 4) Comeron (1995) for dNdS estimation
# 5) single core processing 'comp_cores = 1'
dNdS(
        query_file      = system.file('seqs/ortho_thal_cds.fasta', package = 'orthologr'),
        subject_file    = system.file('seqs/ortho_lyra_cds.fasta', package = 'orthologr'),
        ortho_detection = "RBH",
        aa_aln_type     = "pairwise",
        aa_aln_tool     = "NW",
        codon_aln_tool  = "pal2nal",
        dnds_est.method = "Comeron",
        comp_cores      = 1,
        clean_folders   = TRUE,
        quiet           = TRUE
        )
```

```
       query_id             subject_id       dN     dS    dNdS
 1: AT1G01010.1 333554|PACid:16033839 0.106400 0.2537 0.41950
 2: AT1G01020.1 470181|PACid:16064328 0.040230 0.1037 0.38790
 3: AT1G01030.1 470180|PACid:16054974 0.014990 0.1265 0.11850
 4: AT1G01040.1 333551|PACid:16057793 0.013470 0.1165 0.11560
 5: AT1G01050.1 909874|PACid:16064489 0.000000 0.1750 0.00000
 6: AT1G01060.3 470177|PACid:16043374 0.044950 0.1133 0.39670
 7: AT1G01070.1 918864|PACid:16052578 0.018300 0.1059 0.17280
 8: AT1G01080.1 909871|PACid:16053217 0.033980 0.1056 0.32170
 9: AT1G01090.1 470171|PACid:16052860 0.009104 0.2181 0.04174
10: AT1G01110.2 333544|PACid:16034284 0.032480 0.1220 0.26620
11: AT1G01120.1 918858|PACid:16049140 0.003072 0.1326 0.02317
12: AT1G01140.3 470161|PACid:16036015 0.005672 0.1312 0.04324
13: AT1G01150.1 918855|PACid:16037307 0.130000 0.2028 0.64120
14: AT1G01160.1 918854|PACid:16044153 0.104600 0.2804 0.37310
15: AT1G01170.2 311317|PACid:16052302 0.000000 0.3064 0.00000
16: AT1G01180.1 909860|PACid:16056125 0.029680 0.1763 0.16830
17: AT1G01190.1 311315|PACid:16059488 0.028690 0.1618 0.17730
18: AT1G01200.1 470156|PACid:16041002 0.019050 0.1675 0.11370
19: AT1G01210.1 311313|PACid:16057125 0.020670 0.1540 0.13420
20: AT1G01220.1 470155|PACid:16047984 0.015690 0.1533 0.10230
```

The `dNdS()` function can be used choosing the following options:

- `ortho_detection` : `RBH` (BLAST best reciprocal hit), `BH` (BLAST best reciprocal hit), `PO` (ProteinOrtho), and `OrthoMCL` (OrthoMCL)
- `aa_aln_type` : `multiple` or `pairwise`
- `aa_aln_tool` : `clustalw`, `t_coffee`, `muscle`, `clustalo`, `mafft`, and `NW` (in case `aa_aln_type = "pairwise"`)
- `codon_aln_tool` : `pal2nal`
- `dnds_est.method` : `Li`, `Comeron`, `NG`, `LWL`, `LPB`, `MLWL`, `YN`, and `MYN`

Please see `?dNdS` for details.

In case your BLAST program, or multiple alignment program can not be executed from the default execution `PATH` you can specify the `aa_aln_path` or `blast_path` arguments.

```
library(orthologr)

# using the `aa_aln_path` or `blast_path` arguments
dNdS(
query_file      = system.file('seqs/ortho_thal_cds.fasta', package = 'orthologr'),
subject_file    = system.file('seqs/ortho_lyra_cds.fasta', package = 'orthologr'),
ortho_detection = "RBH",
blast_path      = "here/path/to/blastp",
aa_aln_type     = "multiple",
aa_aln_tool     = "clustalw",
aa_aln_path     = "here/path/to/clustalw",
codon_aln_tool  = "pal2nal",
dnds_est.method = "Comeron",
comp_cores      = 1,
clean_folders   = TRUE,
quiet           = TRUE
)
```

Advanced options

Additional arguments can be passed to `dNdS()`. This allows you to use more advanced options of several interface programs.

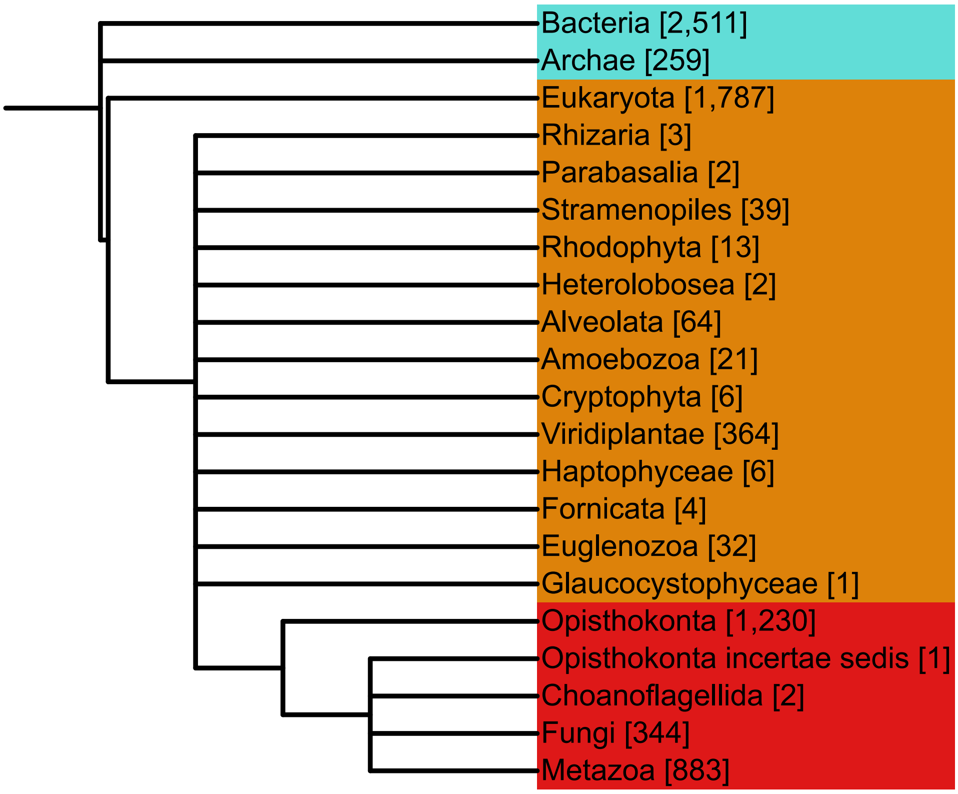
To pass additional parameters to the interface programs, you can use the `blast_params` and `aa_aln_params` arguments. The `aa_aln_params` argument assumes that when you chose e.g. `aa_aln_tool = "mafft"` you will pass the corresponding additional parameters in MAFFT notation.

```
library(orthologr)

# get dNdS estimated for orthologous genes between A. thaliana and A. lyrata
# using additional parameters:

# get a dNdS table using:
# 1) reciprocal best hit for orthology inference (RBH)
# 2) multiple amino acid alignments using MAFFT
# 3) pal2nal for codon alignments
# 4) Comeron (1995) for dNdS estimation
# 5) single core processing 'comp_cores = 1'
Ath_Aly_dnds <-
        dNdS(
        query_file      = system.file('seqs/ortho_thal_cds.fasta', package = 'orthologr'),
        subject_file    = system.file('seqs/ortho_lyra_cds.fasta', package = 'orthologr'),
        ortho_detection = "RBH",
        blast_params    = "-matrix BLOSUM80",
        aa_aln_tool     = "mafft",
        aa_aln_params   = "--maxiterate 1 --clustalout",
        dnds_est.method = "Comeron",
        comp_cores      = 1,
        clean_folders   = TRUE,
        quiet           = TRUE
        )

# filter for:
# 1) all dN values having an NA value are omitted
# 2) all dS values having an NA value are omitted
# 3) all dNdS values >= 0.1 are omitted
filter_dNdS(Ath_Aly_dnds, dnds.threshold = 0.1)
```

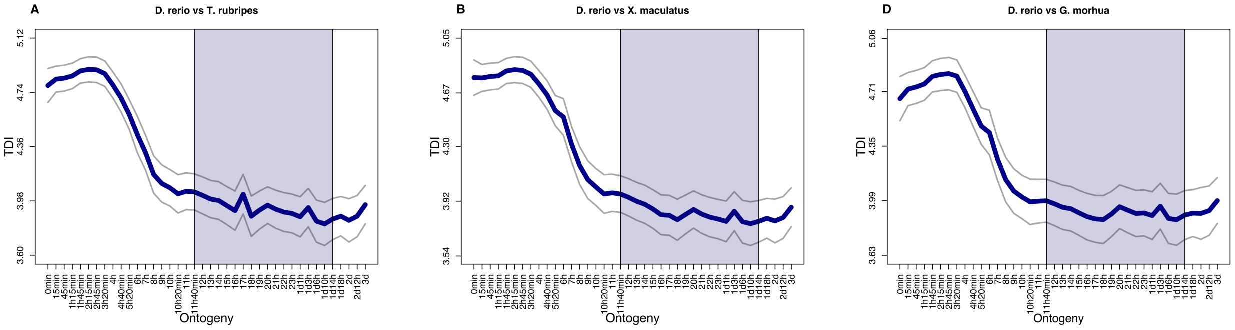
```
     query_id      subject_id       dN     dS    dNdS
1 AT1G01050.1 909874|PACid:16 0.000000 0.1750 0.00000
2 AT1G01090.1 470171|PACid:16 0.009843 0.2150 0.04579
3 AT1G01120.1 918858|PACid:16 0.003072 0.1326 0.02317
4 AT1G01140.3 470161|PACid:16 0.005672 0.1312 0.04324
5 AT1G01170.2 311317|PACid:16 0.008750 0.2827 0.03095
6 AT1G01220.1 470155|PACid:16 0.015210 0.1533 0.09919
```

Here `blast_params` and `aa_aln_params` take an character string specifying the parameters that shall be passed to BLAST and MAFFT. The notation of these parameters must follow the command line call of the stand alone versions of BLAST and MAFFT: e.g. `blast_params = "blast_params = -matrix BLOSUM80"` and `aa_aln_params = "--maxiterate 1 --clustalout"`.

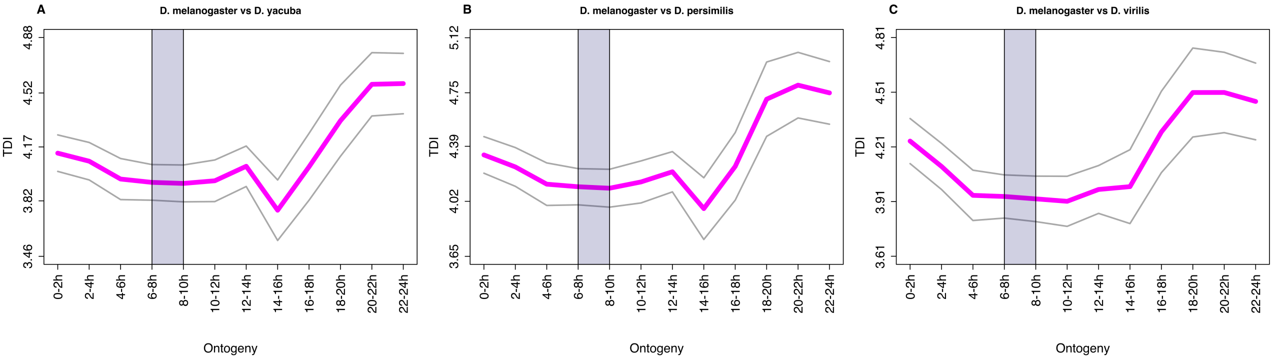
10.6 Suppl. Material: Evidence for Active Maintenance of Phylotranscriptomic Hourglass Patterns in Animal and Plant Embryogenesis



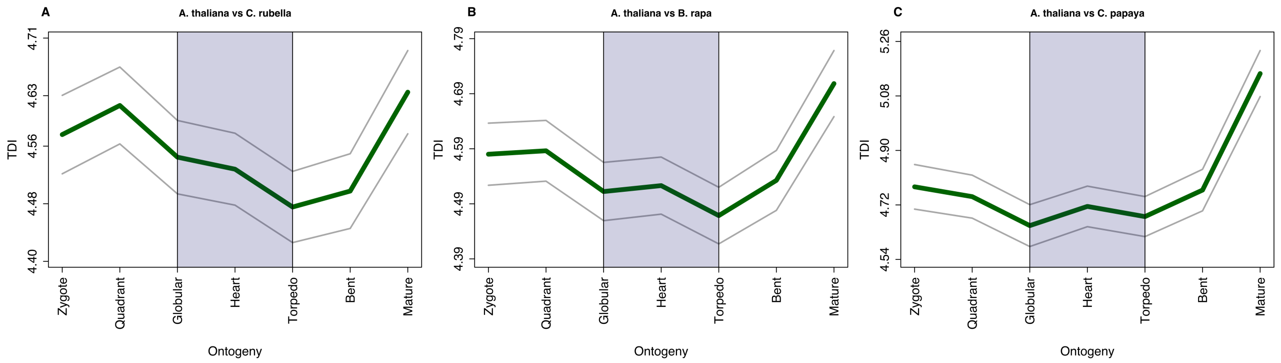
Supplementary figure S1: NCBI taxonomy tree representing the major groups of species/genomes used for PS map data base.



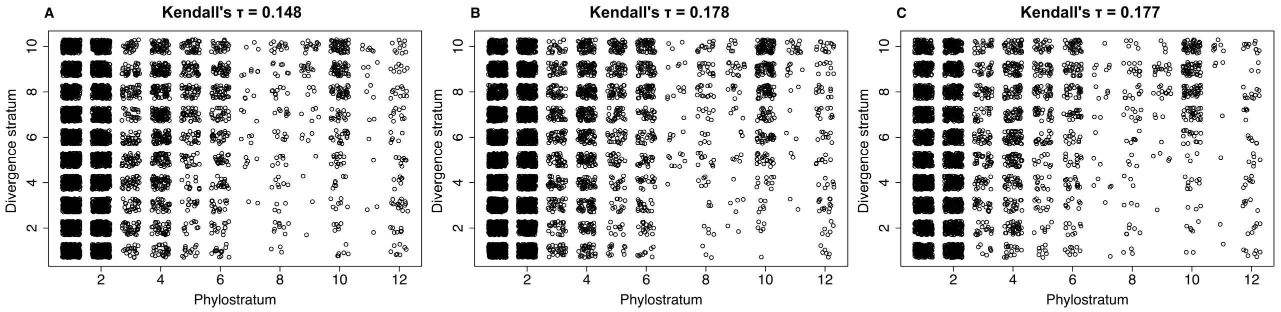
Supplementary figure S2: Transcriptome divergence index profiles across *D. rerio* embryogenesis. **A**, *D. rerio* vs *T. rubripes* (P-value red. hourgl. test = 0.504). **B**, *D. rerio* vs *X. maculatus* (P-value red. hourgl. test = 0.36). **C**, *D. rerio* vs *G. morhua* (P-value red. hourgl. test = 0.138). The blue shaded area marks the predicted phylotypic period. The grey lines represent the standard deviation estimated by permutation analysis.



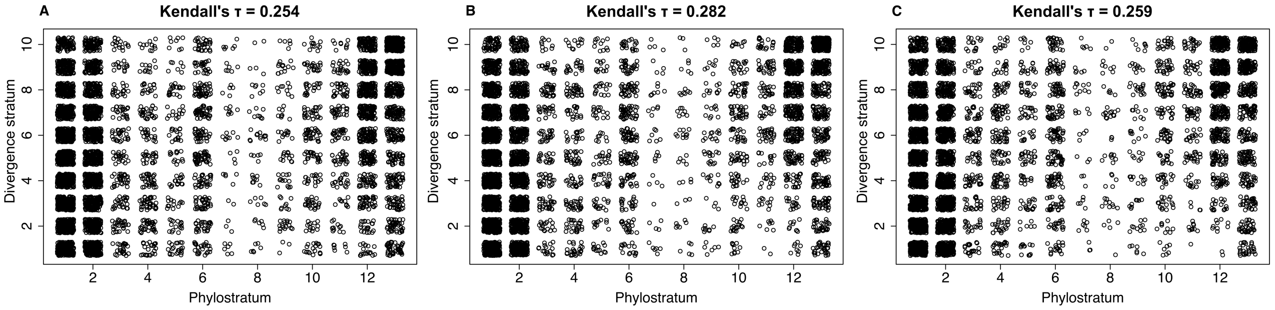
Supplementary figure S3: Transcriptome divergence index profiles across *D. melanogaster* embryogenesis. **A**, *D. melanogaster* vs *D. yakuba* (P-value red. hourgl. test = 0.021). **B**, *D. melanogaster* vs *D. persimilis* (P-value red. hourgl. test = 0.0215). **C**, *D. melanogaster* vs *D. virilis* (P-value red. hourgl. test = 0.00713). The blue shaded area marks the predicted phylotypic period. The grey lines represent the standard deviation estimated by permutation analysis.



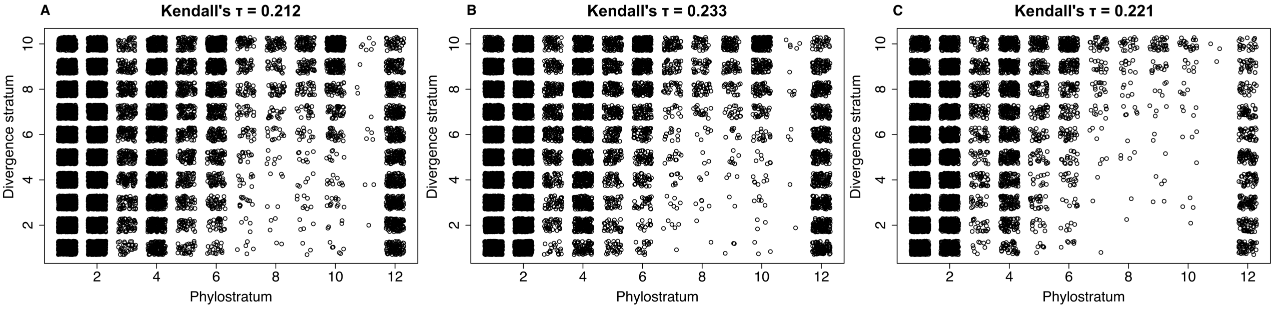
Supplementary figure S4: Transcriptome divergence index profiles across *A.thaliana* embryogenesis. **A**, *A.thaliana* vs *C. rubella* (P-value red. hourgl. test = 0.00745). **B**, *A.thaliana* vs *B. rapa* (P-value red. hourgl. test = 0.000249). **C**, *A.thaliana* vs *C. papaya* (P-value red. hourgl. test = 0.00239). The blue shaded area marks the predicted phylotypic period. The grey lines represent the standard deviation estimated by permutation analysis.



Supplementary figure S5: Correlation between phylostrata and divergence strata. Scatter plots of phylostratum vs. divergence stratum over all genes of *D. rerio*. Ka /Ks ratios for divergence stratum assignment are derived from orthologous genes between **A**, *D. rerio* vs *T. rubripes*. **B**, *D. rerio* vs *X. maculatus*. **C**, *D. rerio* vs *G. morhua*. Kendall τ values denote the Kendall rank correlation coefficients measuring the association between both parameters.

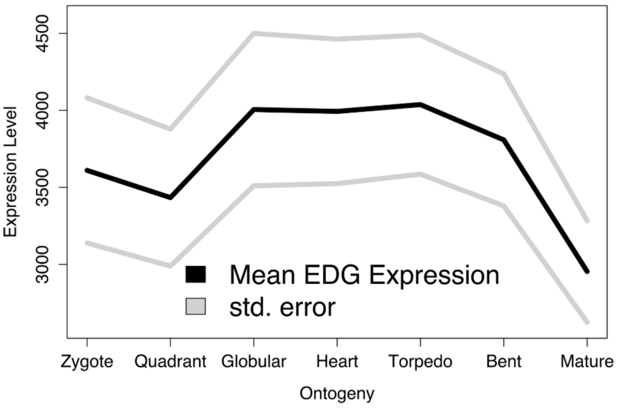


Supplementary figure S6: Correlation between phylostrata and divergence strata. Scatter plots of phylostratum vs. divergence stratum over all genes of *D. melanogaster*. Ka /Ks ratios for divergence stratum assignment are derived from orthologous genes between **A**, *D. melanogaster* vs *D. yakuba.* **B**, *D. melanogaster* vs *D. persimilis.* **C**, *D. melanogaster* vs *D. virilis.* Kendall τ values denote the Kendall rank correlation coefficients measuring the association between both parameters.



Supplementary figure S7: Correlation between phylostrata and divergence strata. Scatter plots of phylostratum vs. divergence stratum over all genes of *A.thaliana*. Ka /Ks ratios for divergence stratum assignment are derived from orthologous genes between **A**, *A.thaliana* vs *C. rubella.* **B**, *A.thaliana* vs *B. rapa.* **C**, *A.thaliana* vs *C. papaya.* Kendall τ values denote the Kendall rank correlation coefficients measuring the association between both parameters.

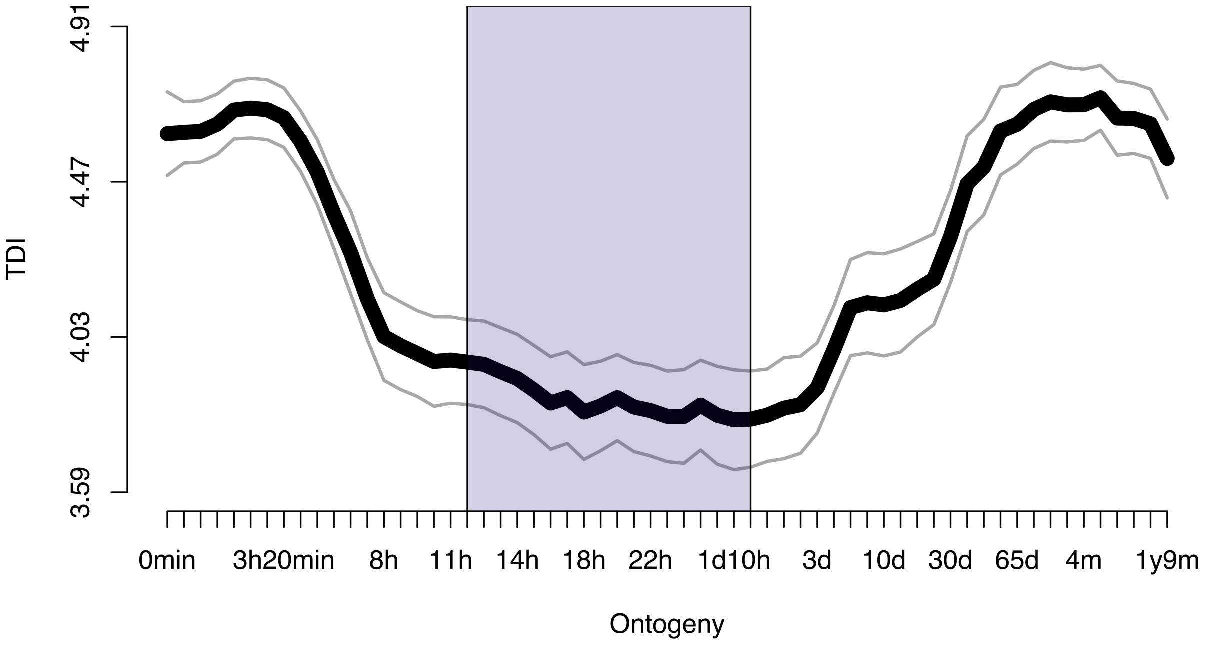
A



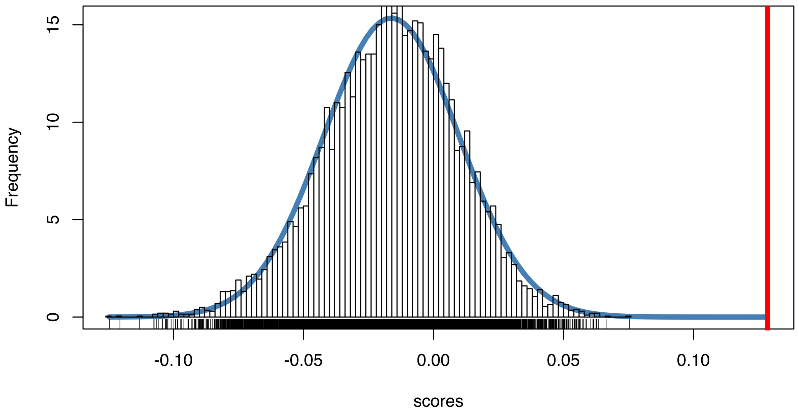
B

Stage	Z	Q	G	H	T	B
Q	0.458					
G	0.021	0.029				
H	0.008	0.007	0.356			
T	0.009	0.008	0.291	0.421		
B	0.006	0.007	0.313	0.440	0.463	
M	0.275	0.320	0.116	0.037	0.023	0.028

Supplementary figure S8: Expression patterns of essential genes during *A. thaliana* embryogenesis. **A**, Mean expression levels of essential genes (embryo defective genes = EDGs) throughout *A. thaliana* embryogenesis. A Kruskal-Wallis rank sum test was performed to test the statistical significance of different gene expression levels between developmental stages ($P < 0.005$). **B**, Results of Dunn‘s test of multiple comparison using Benjamini-Hochberg adjustment.



Supplementary figure S9: Transcriptome divergence index profiles across *D. rerio* ontogeny starting from unfertilized egg to adult stages based on the complete developmental data set of Domazet-Lošo and Tautz (2010). DS computations are based on *D. rerio* vs *A. mexicanus* (P-value red. hourgl. test = 6.49e-19). The blue shaded area marks the predicted phylotypic period. The grey lines represent the standard deviation estimated by permutation analysis.



Supplementary figure S10: Frequency distribution of 10,000 randomly permuted reductive hourglass scores D_{min} that has been used to compute the P-value returned by the reductive hourglass test for the TAI profile of *A.thaliana*. The corresponding frequency distribution was fitted by a gaussian distribution and the red line visualizes the reductive hourglass score of the observed TAI profile of *A.thaliana*.

10.7 Suppl. Material: Post-embryonic Hourglass Patterns Mark Ontogenetic Transitions in Plant Development

Supplementary Materials

Materials and Methods

Germination experiment

Plant material and growth conditions

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.

RNA extraction

Total RNA was extracted according to a modified hot borate method modified (Wan and Wilkins 1994), as described previously (Maia et al. 2011). RNA quality and concentration were assessed by agarose gel electrophoresis (0.1g mL^{-1}) and NanoDrop® measurements.

Microarray hybridization

The quality control, RNA labeling, hybridization and data extraction were performed at ServiceXS B.V. (Leiden, The Netherlands). Labelled ss-cDNA was synthesized using the Affymetrix NuGEN Ovation PicoSL WTA v2 kit and Biotin Module using 50 ng total RNA as template. The fragmented ss-cDNA was utilized for the hybridization on the Affymetrix ARAGene 1.1ST Array plate. The Affymetrix HWS Kit was used for the hybridization, washing and staining of the plate. Scanning of the Array Plates was performed using the Affymetrix GeneTitan scanner. All procedures were performed according to the instructions of the manufacturers (http://www.nugen.com and http://www.affymetrix.com). The resulting data were analysed using the R statistical programming environment and the Bioconductor packages (Gentleman et al. 2004). The data was normalized using the RMA algorithm (Irizarry et al. 2003) using the TAIRG v17 cdf file (http://brainarray.mbni.med.umich.edu). Expression data can be downloaded from the NCBI GEO database (accession number GSE65394; accessible here: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?token=onyxsyycjtaxxux&acc=GSE65394). Normalized expression data are included in Supplementary Dataset 1. See also Silva et al. (2016) for a detailed description of the microarray experiment.

Floral transition experiment

Plant material and growth conditions

To achieve synchronization of flowering times, we adapted a previously published cultivation regime (Schmid et al. 2003). 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 (8h 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 (16h light/8h 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

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.

RNA-Seq Analysis

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) (Kim et al. 2013). Uniquely mapped reads were counted using the featureCounts (v1.4.6) (Liao et al. 2014) with the annotation file from TAIR10. The normalized RPKM values were calculated by the function rpkm() from the Bioconductor package edgeR (Zhou et al. 2014) 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.

Flower development experiment

Plant material, growth conditions, generation of expression data and data analysis are described in detail elsewhere (Ryan et al. 2015). Expression data can be downloaded from the NCBI GEO database (accession number GSE64581). Normalized expression data are included in Supplementary Dataset 1.

Phylotranscriptomic analyses

Scripts for complete reproduction of all data presented in this manuscript including transcriptomic data (Supplementary Dataset 1), computation of TAI, relative expression patterns, and permutation testing of their statistical significance are available via the GitHub repository (https://github.com/HajkD/post-embryo). Detailed instructions for applications of the same analyses to any expression data set and species with sufficient genome information can be found in the R packages myTAI (https://github.com/HajkD/myTAI) and orthologr (https://github.com/HajkD/orthologr).

In brief, phylostratigraphy and TAI analyses have been performed based on and as described previously (Quint et al. 2012; Drost et al. 2015). The phylostratigraphic approach (Domazet-Lošo et al. 2007) uses the BLASTP algorithm (E-value < 1E-5) to detect the evolutionary most distant homolog within a phylogenetically categorized tree of life (Fig. S1). The evolutionary age of a gene is then assigned according to the phylogenetic category of the most distant homolog. This age assignment of each protein coding gene of *A. thaliana* is then stored in a phylostratigraphic map. The gene age distribution of *A. thaliana* genes ranges from PS1 to PS12 where PS1 represents the evolutionary most distant age category (cellular org.) and PS12 the evolutionary most recent age category (*A. thaliana* specific).

The phylostratigraphic map of *A. thaliana* is then matched with the developmental process of interest and TAI computations are performed according to the formula:

$$TAI_s = \frac{\sum_{i=1}^{n} ps_i e_{is}}{\sum_{i=1}^{n} e_{is}},$$

where ps_i denotes the PS of gene i, and n denotes the number of genes. A small value of ps_i represents an old PS, and a high value of ps_i a young PS. Together, a small value of TAI_s represents a high mean evolutionary age of the transcriptome at stage s, and a high value of TAI_s a low mean evolutionary age.

To quantify the statistical significance of phylotranscriptomic hourglass patterns, we applied the *flat line test* (Drost et al. 2015) (quantifying the significant deviation of the observed TAI pattern from a flat line) and the *reductive hourglass test* (Drost et al. 2015) (quantifying the significance of a TAI hourglass pattern). By definition, TAI profiles are computed with absolute expression levels (Domazet-Lošo and Tautz 2010; Quint et al. 2012). However, according to suggestions by Piasecka et al. (Piasecka et al. 2013), we also computed TAI profiles with log2 and squareroot-transformed expression values (Fig. S3).

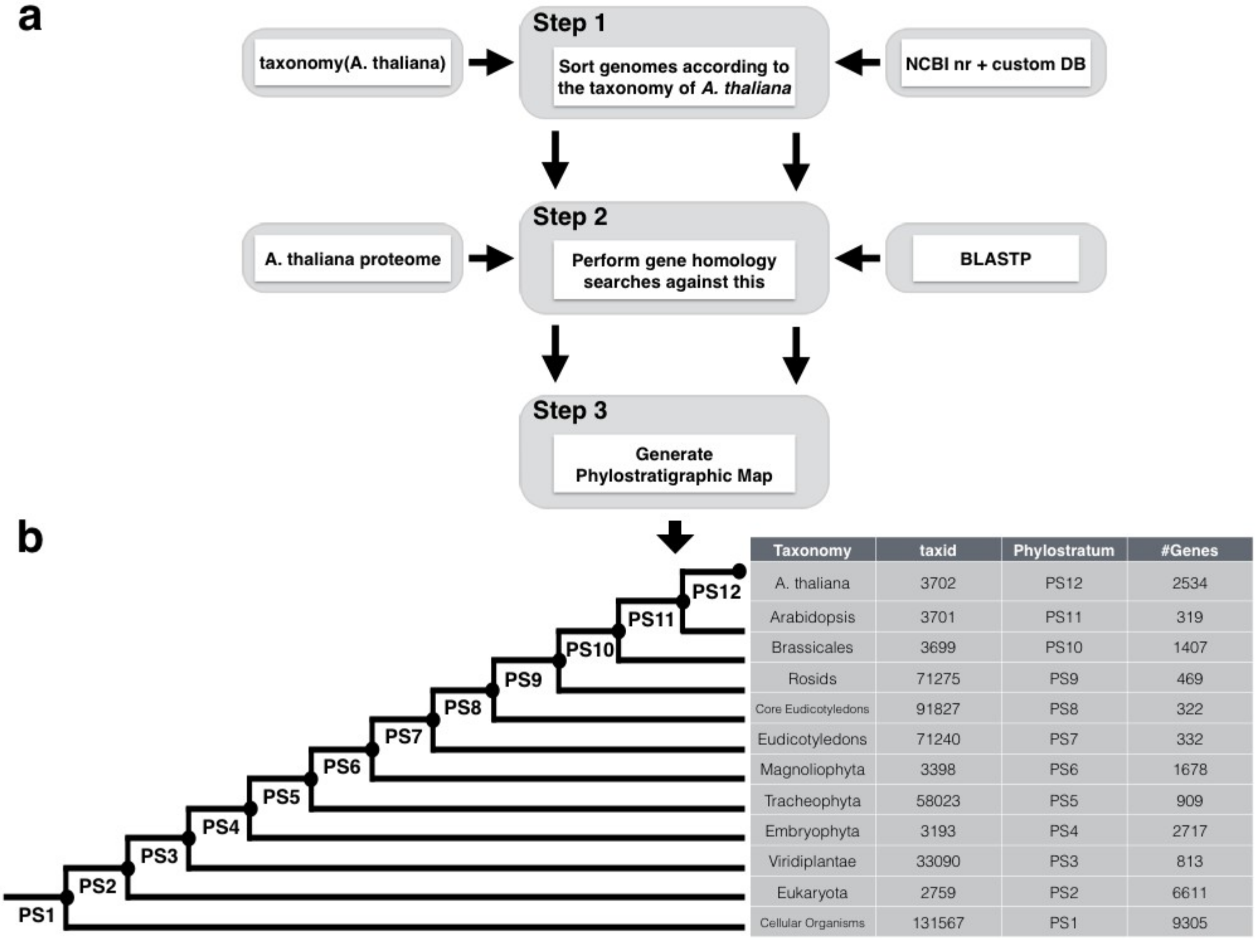
Computation of relative expression levels has been performed as described previously (Quint et al. 2012; Drost et al. 2015) and according to the formula:

$$RE_s = \frac{\bar{f} - f_{min}}{f_{max} - f_{min}}$$

where $\bar{f}$ denotes the mean expression level of phylostratum i and developmental stage s and f_{min}/f_{max} is the minimum/maximum mean expression level of phylostratum i over all developmental stages s.

All TAI, relative expression level, and statistical test computations were performed using the R package *myTAI* (Drost 2016) and can be reproduced by following the instructions presented at https://github.com/HajkD/post-embryo.

Supplementary Figures


a
Step 1
taxonomy(A. thaliana)
Sort genomes according to the taxonomy of A. thaliana
NCBI nr + custom DB
Step 2
A. thaliana proteome
Perform gene homology searches against this
BLASTP
Step 3
Generate Phylostratigraphic Map
b
PS12
PS11
PS10
PS9
PS8
PS7
PS6
PS5
PS4
PS3
PS2
PS1

Taxonomy	taxid	Phylostratum	#Genes
A. thaliana	3702	PS12	2534
Arabidopsis	3701	PS11	319
Brassicales	3699	PS10	1407
Rosids	71275	PS9	469
Core Eudicotyledons	91827	PS8	322
Eudicotyledons	71240	PS7	332
Magnoliophyta	3398	PS6	1678
Tracheophyta	58023	PS5	909
Embryophyta	3193	PS4	2717
Viridiplantae	33090	PS3	813
Eukaryota	2759	PS2	6611
Cellular Organisms	131567	PS1	9305

Figure S1. Flow chart illustrating the phylostratigraphic procedure. **a,** First, the taxonomy of *A. thaliana* is retrieved from NCBI Taxonomy and genomes stored in the reference database are sorted into age categories according to this taxonomy. Next, each protein coding gene of *A. thaliana* is blasted against this categorized database using the *blastp* algorithm with an E-value cutoff < 1E-5. Each protein coding gene fulfilling the E-value criterium is then sorted into the age category (phylostratum) for which the most distant homolog could be detected. Genes fulfilling the blast criteria without having detectable homologs in other species are denoted as *A. thaliana* specific genes and were sorted in PS12. **b,** The table shown in this figure illustrates an overview of the taxonomy and the number of genes classified according to this taxonomy for *A. thaliana*. In detail, this table shows the taxonomic name, taxonomy id, phylostratum (PS) information, and number of genes corresponding to each PS for *A. thaliana*. The number of genes in each PS reflect all protein coding genes of *A. thaliana* fulfilling the phylostratigraphy criteria. This phylostratigraphic map is then matched with each expression dataset resulting in different numbers of genes that can be found on the microarray chip (in case of germination) or genes fulfilling the RPKM > 1 criterium (in case of floral transition).

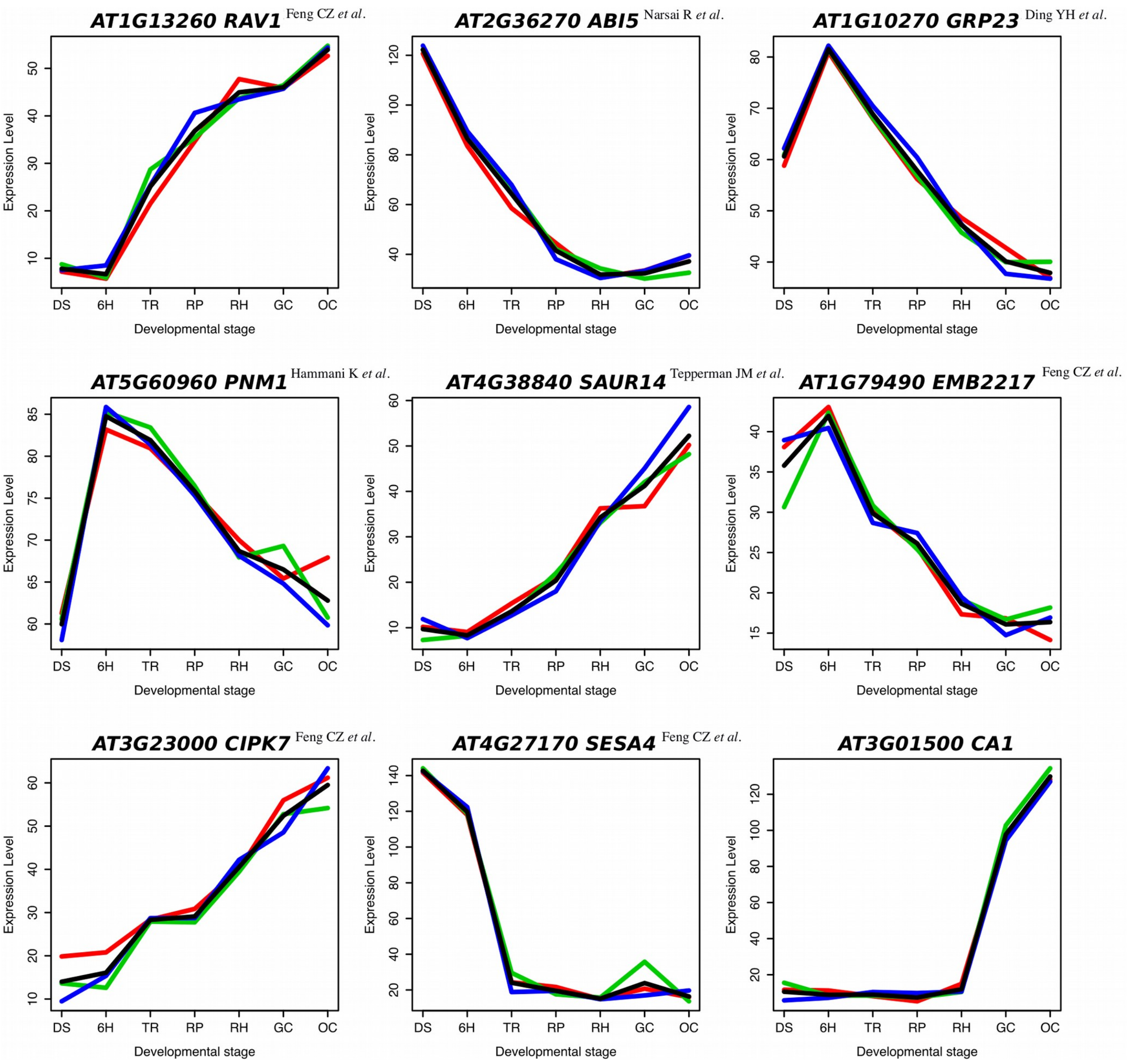


Figure S2. Expression profiles of germination (embryo-vegetative transition) reference genes. Absolute expression levels for selected seed germination genes are shown for all time points. Red, green, and blue lines represent data from three biological replicates, black lines the mean values of the replicate experiments. Expression profiles for each gene match their known expression profiles, validating the dataset. DS, dry seed; 6h, six-hours 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.

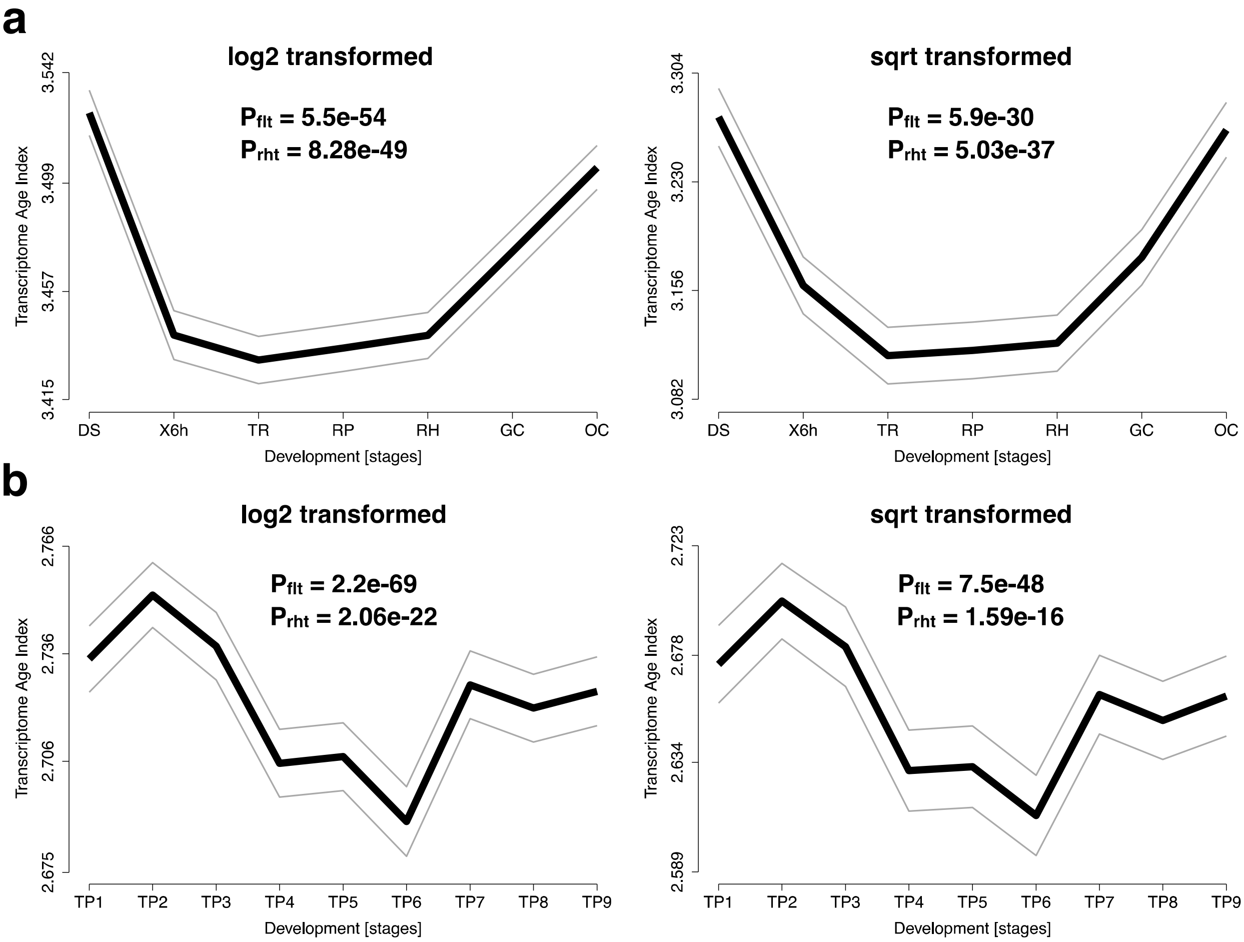


Figure S3. Transformed transcriptome age index analysis for germination (embryo-vegetative transition) and floral transition (vegetative-reproductive transition) in *A. thaliana*. Piasecka *et al.* (Piasecka et al. 2013) previously discussed the influence of gene expression transformation on the global TAI pattern. They found that different gene expression transformations can result in qualitatively different TAI patterns. To test whether or not the observed hourglass patterns in this study are influenced by such gene expression level transformations, we transformed expression levels using *log2* and *sqrt* as transformation functions. We find for both germination (**a**) and floral transition (**b**) that *sqrt* and *log* transformed expression levels result in qualitatively similar and statistically significant hourglass patterns, suggesting robust evolutionary signals in *A. thaliana* germination and floral transition. **a+b** The gray lines represent the standard deviation estimated by permutation analysis. P-values were derived by application of the flat line test (P_{flt}) and the reductive hourglass test (P_{rht}) (Drost et al. 2015).

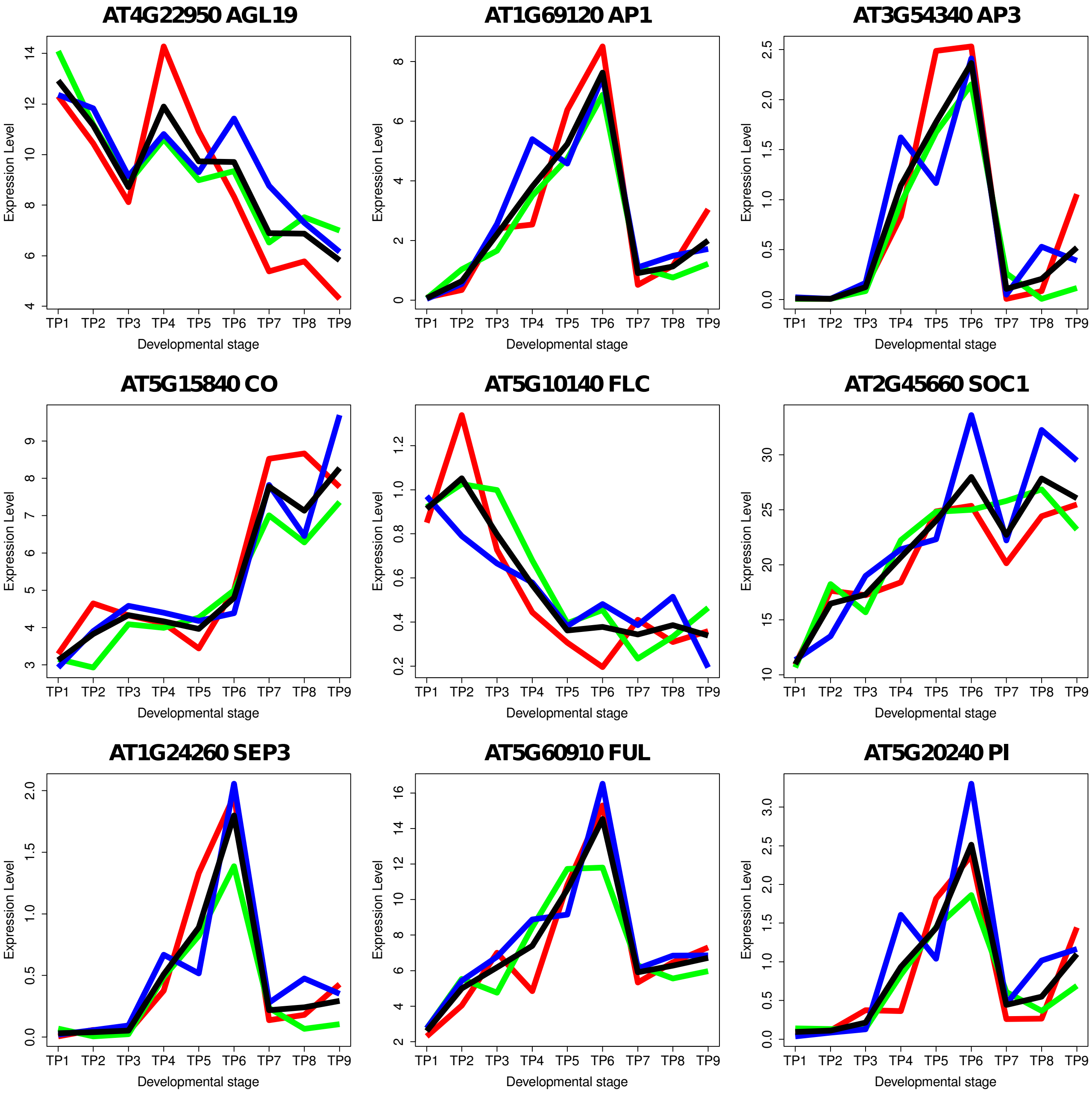


Figure S4. Expression profiles of floral transition (vegetative-reproductive transition) reference genes. Absolute expression levels for selected flowering genes are shown for all time points. Red, green, and blue lines represent data from three biological replicates, black lines the mean values of the replicate experiments. Expression profiles for each gene match their known expression profiles (Schmid et al. 2003), validating the dataset.

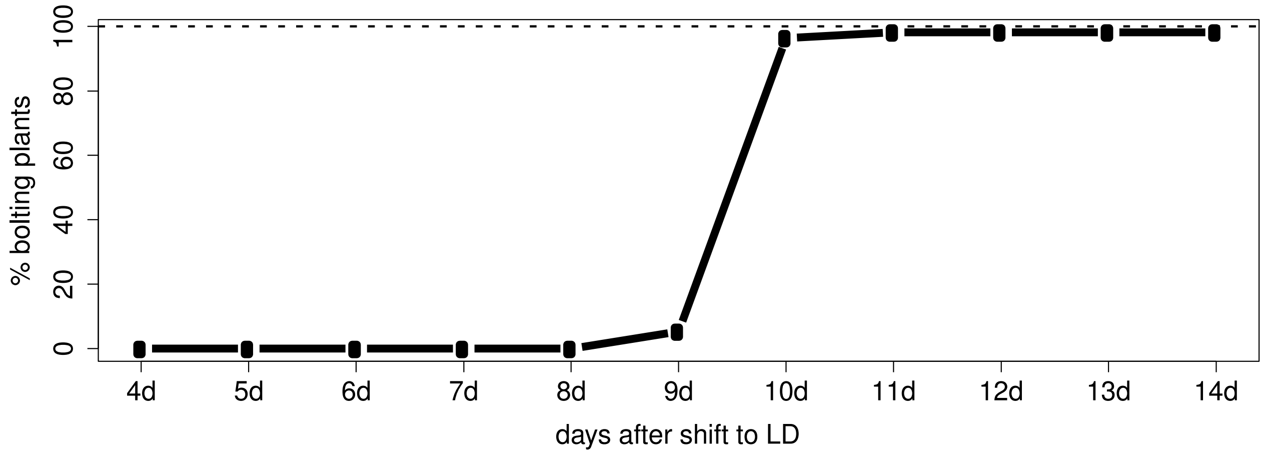


Figure S5. Synchronization of bolting for generating the floral transition (vegetative-reproductive transition) dataset. Almost 100% of the control plants started to bolt within the same day after the flowering time inducing shift from short to long days, demonstrating the high degree of developmental synchronization achieved by the cultivation regime described in the methods section.

Supplementary Dataset S1. Normalized gene expression data for *A. thaliana* flowering, germination, and flower development. The first column stores the phylostratum assignment of the corresponding gene, the second column the *gene id*, and all other columns store the expression levels of the corresponding developmental time points / stages. For germination (sheet 1), the absolute expression levels are shown for the developmental stages (TP); DS, mature dry seeds; 6h, six-hours 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. For flowering (sheet 2), the absolute expression levels are shown for the developmental time points (TP); 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. For flower development, (sheet 3) the absolute expression levels are shown for the developmental time points (TP); 0d, 1d, 1.5d, 2d, 2.5d, 3d, 3.5d, 4d, 4.5d, 5d, 7d, 9d, 11d, 13d after treatment with a solution containing 10 µM dexamethasone. Detailed scripts for reproducing the main figures with this data can be found at https://github.com/HajkD/post-embryo.

References Supplementary Material

Ding YH, Liu NY, Tang ZS, Liu J, Yang WC. 2006. *Arabidopsis GLUTAMINE-RICH PROTEIN23* Is Essential for Early Embryogenesis and Encodes a Novel Nuclear PPR Motif Protein That Interacts with RNA Polymerase II Subunit III. *Plant Cell* 18:815-830.

Domazet-Lošo T, Brajković J, Tautz D. 2007. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends Genet.* 23:533-9.

Domazet-Lošo T, Tautz D. 2010. A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature* 468:815-8.

Drost HG. 2016. Performing Evolutionary Transcriptomics with R. R package version 0.0.4. Available from: http://CRAN.R-project.org/package=myTAI.

Drost HG, Gabel A, Grosse I, Quint M. 2015. Evidence for active maintenance of phylotranscriptomic hourglass patterns in animal and plant embryogenesis. *Mol Biol Evol.* 32:1221-1231.

Feng CZ, Chen Y, Wang C, Kong YH, Wu WH, Chen YF. 2014. Arabidopsis RAV1 transcription factor, phosphorylated by SnRK2 kinases, regulates the expression of *ABI3*, *ABI4*, and *ABI5* during seed germination and early seedling development. *Plant J.* 80:654-668.

Gentleman R. 2004. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* 5:R80.

Hammani K, Gobert A, Hleibieh K, Choulier L, Small I, Giegé P. 2011. An *Arabidopsis* Dual-Localized Pentatricopeptide Repeat Protein Interacts with Nuclear Proteins Involved in Gene Expression Regulation. *Plant Cell* 23:730-740.

Irizarry RA, Bolstadt BM, Collin F, Cope LM, Hobbs B, Speed TP. 2003. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.* 31:e15.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* 14:R36.

Liao Y, Smith GK, Shi W. 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30:923-30.

Maia J, Dekkers BJW, Provart NJ, Ligterink W, Hilhorst HWM. 2011. *PLoS ONE* 6: e29123.

Narsai R, Law SR, Carrie C, Xu L, Whelan J. 2011. In-Depth Temporal Transcriptome Profiling Reveals a Crucial Developmental Switch with Roles for RNA Processing and Organelle Metabolism That Are Essential for Germination in Arabidopsis. *Plant Physiol.* 157:1342-1362.

Piasecka B, Lichocki P, Moretti S, Bergmann S, Robinson-Rechavi M. 2013. *PLoS Genet.* 9:e1003476.

Quint M, Drost HG, Gabel A, Ullrich KK, Boenn M, Grosse I. 2012. A transcriptomic hourglass in plant embryogenesis. *Nature* 490:98-101.

Ryan PT, Ó'Maoiléidigh DS, Drost HG, Kwasniewska K, Gabel A, Grosse I, Graciet E, Quint M, Wellmer, F. 2015. Patterns of gene expression during Arabidopsis flower development from the time of initiation to maturation. *BMC Genomics* 16:488.

Schmid M, Uhlenhaut NH, Godard F, Demar M, Bressan R, Weigel D, Lohmann JU. 2003. Dissection of floral induction pathways using global expression analysis. *Development* 130:6001-6012.

Silva AT, Ribone PA, Chan RL, Ligterink W, Hilhorst HWM. 2016. A predictive co-expression network identifies novel genes controlling the seed-to-seedling phase transition in *Arabidopsis thaliana*. *Plant Physiol.* (accepted)

Smyth DR, Bowman JL, Meyerowitz EM. 1990. Early flower development in Arabidopsis. *Plant Cell* 2:755-767.

Tepperman JM, Hwang YS, Quail PH. 2006. phyA dominates in transduction of red-light signals to rapidly responding genes at the initiation of Arabidopsis seedling de-etiolation. *Plant J*. 48:728-742.

Wan CY, Wilkins TA. 1994. A Modified Hot Borate Method Significantly Enhances the Yield of High-Quality RNA from Cotton (*Gossypium hirsutum* L.). *Anal Biochem.* 223:7-12.

Zhou X, Lindsay H, Robinson MD. 2014. Robustly detecting differential expression in RNA sequencing data using observation weights. *Nucleic Acids Res.* 42:e91.

10.8 Suppl. Material: Transcriptional Dynamics of Two Seed Compartments with Opposing Roles in Arabidopsis Seed Germination

Supplemental Data file accompanying the Dekkers et al manuscript entitled "Transcriptional Dynamics of Two Seed Compartments with Opposing Roles in Arabidopsis Seed Germination"

SUPPLEMENTAL DATA

Comparisons with other seed transcriptome datasets (with Supplemental Fig. S3).

1. Endosperm/Embryo Dataset. Taking the 18 .cel files published as part of the article (Penfield et al., 2006), [http://www.ncbi.nlm.nih.gov/projects/geo/query/acc.cgi?acc=GSE5751], we renormalize the chips using the RMA normalization procedure and the custom CDF as detailed in the material & methods, to ensure compatibility with our data. The resulting probe set distribution suggests that the noise region of the data is <5 ($\log_2$ RMA, Supplemental Fig. S3A). We calculate which probe sets are differentially expressed at a 5-fold level, thresholding the data at 4 and then perform a t-test to check each expressed probe set is significantly different at a p-value of 0.05. This resulted in 445 (434) genes which are 5 fold up-regulated in the post-radicle emergence Endosperm (Embryo), when using our methods of analysis (see Supplemental materials & methods). Compared to a list of genes which are 5 fold different between our MCE 38 HAS ER and RAD 38 HAS ER samples (the equivalent to their radicle protrusion at 24 HAS post-stratification), 277/445 of genes were in both Endosperm up lists, and 145/432 genes are in both Embryo up lists (Supplemental Fig. S3B). This represents a significant overlap certainly if one takes into account that the Penfield data compared embryo vs whole endosperm of stratified seeds while in our case non-stratified were used and compared and RAD vs halved endosperm, the MCE. This also provides an explanation why the overlap between EMB Up list is smaller, as this is likely due to the absence of COT specific genes in our RAD sample. There are only a few genes in the EMB Up/MCE 38 ER Up and the END Up / RAD 38 ER Up overlap (7 and 2 genes, respectively), mostly genes which are in the Cotyledons but not the Radicle (or vice versa). Compared to our compartment-specific lists, (which include the entire time course, not just the post-germination time point), we find there is less overlap with 91/452 and 76/432 genes that are 5-fold up-regulated in the Penfield dataset being tissue specific to our stringent definition (Supplemental Fig. S3B).

2. Microdissected Seed Development Dataset.

Taking the 87 .cel files published as part of the article on microdissected seeds (Le et al., 2010), [http://www.ncbi.nlm.nih.gov/projects/geo/query/acc.cgi?acc=GSE15165], we renormalize the chips using the RMA normalization procedure and the custom CDF as in the Supplemental materials and methods. The resulting probe set distribution suggests that the noise region of the data is <3 ($\log_2$ RMA, Supplemental Fig. S3C). To generate the tissue-specific lists we find genes which are 2-fold different between the two samples, having thresholded the data at 3, then perform a t-test on each gene found significant by this method (at a p-value of 0.05). As each tissue only has two replicates, a significant number of genes are called non-significant by this t-test and discarded. Comparing with the

Mature Green (MG) stage of the microdissected data from (Le et al., 2010), we find that some of the genes which are 2-fold higher expressed between our MCE or PE samples at 3, 16, 31 HAS are also 2-fold higher in the appropriate microdissected samples. Many of the genes which are specific in our time course are not expressed (no mean over 4) in the MG endosperm data (Le et al., 2010), and vice versa. Of the genes which are expressed in the seed development data, approximately 10% of the genes which are up-regulated in the time course are specific to the same part of the developing seed endosperm (Supplemental Fig. S3D). The data of (Le et al., 2010) separates the endosperm into Micropylar, Chalazal and Peripheral samples, and thus we need to make additional comparisons to compare with our data.

Similarly, when the END- and EMB- lists (without genes that are only expressed over 6 at 38 HAS) are investigated, we find 28 of the genes are also specific in the comparison between the microdissected Embryo and all three Endosperm samples (14 specific to the Endosperm and 14 specific to the Embryo, see Supplemental Fig. S3E), with only one gene (AT5G42200) which is higher in the Endosperm in the Mature Green sample but specific to the Embryo in our time course. These 28 genes are therefore specific to the Embryo/Endosperm in both our time course and the seed development data. If we compare our Endosperm/Embryo specific genelists with the comparisons between the Embryo and the three Endosperm samples, we find 51.4% of our Endosperm specific genes that are expressed in the developing seeds are higher in the Endosperm than Embryo in the developing seeds (Supplemental Fig. S3E). Lower overlap is seen for the Embryo specific list, but only a few genes are specific to the opposite tissue than in our data.

Confirmation of tissue specific gene expression by RT-qPCR (with Supplemental Fig. S4).

To confirm tissue specific expression found in the microarray data we performed gene expression analysis using RT-qPCR. Therefore seeds were isolated at the 31 HAS time point at which all seeds showed TR (Supplemental Fig. S4A) and dissected in the tissues described in Supplemental materials & methods (see also Fig. 1D) except that the MCE tissue was further dissected in the micropylar endosperm (ME) and the chalazal endosperm (CE). In total 20 genes were tested, the majority of which were either specific to the MCE or higher expressed in the MCE compared to the PE. Relative expression levels from microarray data and RT-qPCR data were compared (Supplemental Fig. S4C) and both analysis showed similar expression patterns and thereby confirmed the gene expression patterns found in the microarray dataset. Interestingly, the majority of the genes that are either specifically or highly expressed in the MCE tissue at 31 HAS are much more prominently expressed in the ME compared to the CE tissue (Supplemental Fig. S4C).

Correlation networks (with Supplemental Fig. S5 and S6).

To further investigate the topological features of these networks we have used TopoGSA (Glaab et al., 2010), available at http://www.topogsa.net/. We computed four topological features and their distributions: node degree (number of connections to other nodes), length of the shortest paths (how far from all other nodes in the network a given node is), local clustering coefficient (how interconnected a group of nodes is to each other) and node betweenness centrality (how many network shortest paths go through a given node) (Supplemental Fig. S5). Clusters of special interest are identified by comparing the average distribution of a given topological feature (e.g. node degree) to the distribution found for that feature in the entire network. While almost 90% of the clusters have higher mean node degree (Supplemental Fig. S5A) than their respective networks and while cluster 1 in both networks is the most connected (with mean node degree 4 times greater than the network average) the clusters node degree distributions are noticeable different for RadNet and EndoNet with the latter's average lower than the former. The mean length of shortest paths is higher for RadNet than for EndoNet (Supplemental Fig. S5B). Over 35% of the RadNet clusters have an average shortest path length greater than their network's average while this figure is only 20% for EndoNet's clusters. Taking together the average node degree (Supplemental Fig. S5A) in both networks (and their clusters) with the average clustering coefficients for clusters in RadNet and EndoNet (Supplemental Fig. S5C) suggests that, although clusters are well-defined, they are not internally dense (average clustering coefficient is never higher than 0.7) and cluster members have many connections outside the clusters they belong to. Approximately half of the clusters in both networks have higher betweenness than their networks averages (see Supplemental Fig. S5D). Notably, within EndoNet, clusters 7, 14, 15, 17 and 19 have the top five largest betweenness centrality score (thus are the most important hubs). For RadNet the clusters with highest betweenness centrality are 15, 21, 22, 25, 30. GO classes overrepresented in these important hubs are depicted in Supplemental Fig. S5E,F.

The largest 30 clusters of the EndoNet network were investigated using overrepresentation analysis (ORA (Keller et al., 2008)), revealing cluster-specific overrepresentation of specific biological processes (Supplemental Fig. S6). For example, clusters 7 and 14 from EndoNet contain almost exclusively ribosome and translation related genes (Supplemental Fig. S6). Investigation of promoter elements in these clusters identified a strong enrichment of a telomere motif (TELO-box), a promoter element found in the Arabidopsis eEF1A (elongation factor) gene promoter and known to be present in numerous genes related to translation (Tremousaygue et al., 1999). Almost all connections in EndoNet clusters 7 and 14 (98% and 88% respectively) are also present in the RadNet (Fig. 3B), showing that genes related to the ribosome and translation are strongly co-expressed in both compartments. Despite this strong co-expression within both networks, the expression profile is different between the two compartments, being induced in both but subsequently repressed in the endosperm.

SUPPLEMENTAL MATERIAL & METHODS

Seed material. For this experiment the *Arabidopsis thaliana* accession Columbia-0 (Col-0, N60000) was used. Arabidopsis plants were grown on rockwool in a climate cell at 22°C and 70% humidity in a 16h light/8h dark cycle for seed production. Plants were watered with a Hyponex nutrient solution (1g/L, www.hyponex.co.jp). For germination and water content measurements, seeds were sown on 0.7% water agarose (Eurogentec) and incubated in a germination cabinet at 22°C with continuous lighting.

Water content measurements. To obtain the initial water content of the "dry" seeds, 5-7 mg of seeds were weighed on an AD-4 Autobalance (Perkin-Elmer). These seeds were dried in an oven at 104˚C for 17h (ISTA, 2009) and weighed again. To measure the water content of imbibed seeds a sample of weighed dry seeds were sown on 0.7% water agarose. After the indicated time points seeds were removed from the agarose plate and dried on filter paper to remove the access of water on the outside of the seeds. After that all seeds were collected and weighed on a balance and from this weight value (taking into account the initial dry weight) the water content was calculated on a dry weight basis.

Seed dissections and RNA isolation. After the indicated hours of imbibition seeds were harvested and dissected using forceps and a scalpel knife. To obtain the micropylar end of the endosperm the Arabidopsis seeds were dissected transversely (slightly out of the middle towards the micropylar end). This endosperm sample includes both the micropylar endosperm (endosperm layer over the radicle tip) as well as the chalazal endosperm (over the cotyledon tips). Therefore we call these samples the micropylar and chalazal endosperm (MCE). The remainder of the endosperm was sampled as peripheral endosperm (PE). Since the endosperm and seed coat are difficult to separate the endosperm was isolated including the seed coat tissue. Since the seed coat is a dead tissue we assumed that this does not interfere with endosperm transcriptome analysis. To obtain the embryo parts the seeds were carefully opened and the embryo was gently removed from the endosperm/seed coat tissue. To obtain the RAD sample the axis was cut just underneath the cotyledons meaning that the RAD sample includes the root tissue and the majority of the hypocotyl. Therefore this sample included the region that has been shown to elongate during germination (Sliwinska et al., 2009). The remainder of the embryo was collected as the COT sample which, next to the cotyledons, also consisted of remainder of the axis, i.e. the top of the hypocotyl and the shoot apical meristem. For the embryo parts approx. 100 seeds and for the endosperm sections 200 seeds were dissected per individual sample. Material was flash frozen in liquid nitrogen and ground in a dismembrator (Mikro-dismembrator U, B. Braun Biotech International) using stainless steel beads. For the isolation of RNA a commercial kit of Stratagene (Agilent Technologies, Absolutely RNA Nanoprep kit, 50 preps, cat# 400753) was used according to the manual. The only modification was the addition of polyvinylpolypyrrolidone (PVPP, 60mg/ml) to the extraction buffer for RNA extraction of endosperm samples, to inactivate phenolic compounds present in the seed coat.

RNA concentration of the samples was measured using a Nanodrop ND1000 spectrophotometer. To assess quality and integrity of the RNA the samples were analyzed using the Shimadzu MultiNA and Agilent Bioanalyzer. In total 100ng of RNA was used to synthesize Biotin-labelled cRNA (using the Affymetrix 3” IVT-Express Labelling Kit) and the concentration and size of the cRNA was assessed. Denaturized cRNA was hybridized on the Affymetrix GeneChips Arabidopsis ATH1 Genome Array.

Normalization of microarray data. The raw .cel files were background corrected and normalized using the Robust Microarray Averaging (RMA) procedure (Irizarry et al., 2003), with a custom chip definition file (.cdf) from the CustomCDF project (Ath1121501_At_TAIRG.cdf v14.0.0, released 22nd March 2011 (Dai et al., 2005)), using the Bioconductor ‘affy’ package in the programming language R. This CDF maps the individual probes on the Affymetrix chip, using recent sequencing information contained in The Arabidopsis Information Repository (TAIR), with their corresponding genes. This eliminates the many-many relationship which exists between the Affymetrix probe sets and gene targets as is traditionally used. In particular this bijective mapping ensures that gene AGI codes may be used as the primary identifier in the correlation networks with no question of how to deal with multiple probe sets, with sometimes markedly differing behaviours, corresponding to the same gene. The resulting probe sets have varying numbers of probes with a minimum of three, although the majority of probe sets have the eleven probes from an original Affymetrix probe set. After removing the control probes, 21313 genes remain.

Fold Changes. Throughout this paper, when the fold changes are calculated, the data means are first clipped at level of 4 ($\log_2$) in expression level – replacing anything less than four with four. This may slightly underestimate the number of differentially expressed genes (or their level of fold change), but helps prevent the noise region, between 2 and ~4.5-5 in this case, from heavily influencing the results of the fold changes.

Differential Gene Expression. A gene is considered differentially expressed between two conditions if the difference between the condition means is sufficiently large (with the clipping as detailed above), and the values are statistically significant at a p-value of 0.05.

Comparison with the seed development data set (Le et al., 2010). The 20 Affymetrix GeneChip microarrays from the dataset (GEO Accession GSE680) (Le et al., 2010) were normalized using RMA with the CustomCDF in the same way as detailed above. The histogram of the normalized data suggests a noise level of 5, and so the means were clipped at 4. The means of the two replicates of the WT Cotyledon Stage and the WT Post-Mature Green stage (PMG) samples were analyzed. Genes with a mean at least 5-fold higher in one condition were tested using a t-test for significance at a p-value of 0.05, with no False Discovery Rate applied. This process resulted in 907 genes which were higher expressed in the Cotyledon stage, and 602 which were higher in the later PMG stage. Without applying the t-test an additional 301 and 139 genes respectively would be considered expressed.

Comparison with the touch data set (Lee et al., 2005). In the original experiment by (Lee et al., 2005) for the response of leaves to touch or darkness, any genes which were not expressed in all conditions were ignored. We therefore obtained the original MAS normalized data (.cel files were not available) from the original authors, thresholded the data at 20 (not $\log_2$) and generated a list of genes which were both at least 2 fold differentially regulated and had a p-value of less than 0.05 in a t-test.

Correlation Networks. For both tissue types, endosperm (MCE and PE combined) and radicle (RAD), we filter the genes by keeping those probe sets which have at least one sample with mean expression (averaged over the four replicates) greater than or equal to 6. This means we only consider genes which have a significant amount of expression in at least one time point, both reducing the number of genes under consideration and removing those genes whose expression is noisy. This results in 11,525 and 11,645 expressed genes in the endosperm and RAD samples respectively. The two types of endosperm samples were combined to give more information into the Endosperm network, as they are very similar, whereas we decided not to combine the cotyledon samples with the radicle samples due to their significant expression as well as functional differences.

To identify interactions between the expressed genes the Pearson correlation coefficient between all pairs of genes is calculated. A cutoff, y, may then be applied to the resulting correlation matrix to produce a set of edges which are above this cutoff. In order to choose the value of this cutoff we calculate, for a range of y, the cumulative frequency of the edge degree of each node in the resulting graph. This may then be plotted, resulting in an approximately straight line in a log-log plot for many values of y (Supplemental Fig. S10). This suggests that the underlying network is obeying a power law distribution over several orders of magnitude, and over a wide range of number of edges.

The node degree distribution for a given cutoff y is approximately scale-free (Clauset et al., 2009), giving a straight line in a log-log plot of the node degree distribution. A simple log-log regression may be fitted to the log-transformed node degree distribution (excepting the degree 1 nodes), and the resulting adjusted r-squared value used as a measure of the linearity of the fit, for each value of y. For the MCE and PE samples (in the combined endosperm network, EndoNet) we choose a cutoff of y = 0.932, resulting in 577,846 edges. This choice balances the conflicting demands of the number of edges and the linearity of the power law fit, leading to an adjusted r-squared value of 0.986 (Supplemental Fig. S10). Using the RAD samples on their own to create a network (RadNet) results in higher correlations due to fewer chips being included, so we choose a cutoff such that approximately the same number of edges as in the MCE network are retained, that is $y = 0.946$ and 586,746 edges, in order to make the resulting networks somewhat comparable. The adjusted r-squared value is 0.996. We note that these choices of cutoff are essentially arbitrary, and very similar networks are generated by increasing or decreasing the cutoff, with more/fewer edges in the order of 20,000-30,000 for each 0.01 the cutoff is adjusted by.

Once the cutoff has been determined, each correlation above that number is considered as an edge and a table of edges between nodes is exported into Cytoscape v2.8.1 (Shannon et al., 2003; Smoot et al., 2011), along with the correlation value between each edge. The yGraph Organic or the Edge-Weighted Force Directed Biolayout methods for arranging the nodes were then used to display the resulting networks.

From these correlation networks, the Cytoscape plug-in ClusterMaker (Morris et al., 2011) is used to partition the overall network into distinct clusters. In particular, the Transitivity Clustering method is used (Wittkop et al., 2010) (with parameters Max Subcluster Size =400, Max Time = 10, using the correlation values as an edge weight) to generate small, well-connected clusters in the network. These resulting clusters contain almost all of the possible edges between the nodes involved, ensuring that all the genes considered correlate well with each other. For example, cluster 1 in EndoNet contains 18862 edges between 195 genes, 99.7% of the possible edges, and therefore these genes therefore have very similar expression profiles.

Plotting the genes present in each cluster allows us to see that this method produces very similar looking clusters as expected, ensuring the genes involved do have very similar behaviour, avoiding the problem of an averaging clustering process like k-means. To determine how similar the correlations are between the two tissues, the edges in an EndoNet cluster are then investigated for correlation in the RAD samples (at varying levels of c), and *vice versa*. This gives a percentage expressing the similarity between these genes in the opposing tissue. In the 111 EndoNet clusters which contain at least 10 genes, 70 (63%) of them contain at least 70% of the corresponding edges at c=0.75 in the RAD samples, showing that many of the genes have correlate in the opposing tissue, although at a lower threshold. The clusters in which only a minority of edges exist in the opposing tissue are interesting, as they may either be tissue specific processes or contain genes which are only present in one tissue. Conversely, of the 95 RadNet clusters containing at least 10 genes, only 47 (49%) have at least 70% corresponding edges at c=0.75 in the endosperm samples.

Overrepresentation analysis. The gene lists of the 30 largest clusters of the EndoNet network (Supplemental Fig. S6) and the compartment specific gene sets (Supplemental Fig. S2) were analyzed using Genetrial (Keller et al., 2008) (http://genetrail.bioinf.uni-sb.de) for GO categories that are overrepresented. For ORA of the gene classes overrepresented in time and tissue were analyzed using Pageman (Usadel et al., 2006; Sreenivasulu et al., 2008) (http://mapman.mpimp-golm.mpg.de/pageman/). A mapping file described by (Joosen et al., 2011) was used. The only modification was addition of a bin containing genes related to aging which was obtained from TAIR (www.arabidopsis.org). The results of Pageman analysis were summarized and redrawn in Fig. 5, Fig. 7 and Supplemental Fig. S7.

Phylotranscriptomic analysis. The determination of the evolutionary age of the *A. thaliana* protein coding genes was performed as described in (Quint et al., 2012). The resulting phylostratigraphic map is identical to (Quint et al., 2012) with one exception: as phylostratum (PS) 10 contained only 18 genes, PS 9 and 10 were fused. Hence, instead of 13 PS, the resulting phylostratigraphic map contains only 12 PS. Relative expression levels were computed as described previously (Domazet-Loso and Tautz, 2010). In brief, the mean expression level e_{js} of phylostratum j and developmental stage s was computed for each j and s as the arithmetic mean of expression levels e_{is} of all genes i belonging to phylostratum j. The mean expression levels e_{js} were linearly transformed to the interval [0,1] according to

$$f_{js} = \frac{e_{js} - e_{jmin}}{e_{jmax} - e_{jmin}}$$

where $e_{j\min}/e_{j\max}$ is the minimum/maximum mean expression level of phylostratum j over the seven developmental stages s. This linear transformation corresponds to a shift by $e_{j\min}$ and a subsequent shrinkage by $e_{j\max} - e_{j\min}$. As a result, the relative expression level f_{js} of developmental stage s with minimum e_{js} is 0, the relative expression level f_{js} of the developmental stage s with maximum e_{js} is 1, and the relative expression levels f_{js} of all other stages s range between 0 and 1, accordingly (Quint et al., 2012). Mean relative expression levels of genes in PS1-PS2, PS3-PS5 and PS6-PS12 were computed in each sampled developmental stage. Error bars represent the standard error of the relative expression levels in PS1-PS2, PS3-PS5 and PS6-PS12 in each developmental stage. Statistical significance of the differences between mean relative expression levels of different phylostrata classes was tested by one-way ANOVA.

RT-qPCR. For RT-qPCR, RNA was isolated from radicle, cotyledon and endosperm tissue used as indicated above. Seeds were dissected using forceps and a scalpel knife. For the radicle and cotyledon samples, material of approximately 300 seeds and for the endosperm samples material of approximately 1300 seeds were used. Genomic DNA was removed using a DNase treatment (RNase-free DNase set, Qiagen). Absence of DNA was checked by comparing cDNA samples with RNA samples which were not reverse transcribed (minus RT control) and the difference was at least 5 Cq values as suggested (Nolan et al., 2006). RNA integrity of all samples was assessed by analysis on a 1% agarose gel. For all Arabidopsis samples clear ribosomal rRNA bands were visible and the OD 260/280 ratios (measured using a Nanodrop ND-1000, Nanodrop Technologies Inc.) were close to 2.0 for all samples used in this experiment.

cDNA synthesis, RT-qPCR conditions and primer design. RNA was reverse transcribed using the iScript^{TM} cDNA synthesis kit (Bio-Rad), with 500ng of total RNA being reverse transcribed according to the kit protocol. cDNA samples were diluted in a total volume of 360µl using sterile milliQ water. Per qPCR reaction 5µl sample, 12.5µl iQ SYBR Green Supermix (Bio-Rad), 0.5µl of primer (from a 10µM work solution) was added and supplemented with water to a final volume 25µl. The RT-qPCR reactions were run on a MyiQ (Bio-Rad). The qPCR program run consisted of a first step at 95°C for 3 min. and afterwards 40 cycles alternating between 15 sec. at 95°C and 1 min. at 60°C.

Primers for the target genes were designed preferably in the 3' part of the transcript. When possible the primer or primer pair was designed in such a way that it spanned an intron/exon border. The T_m of the primers was between 59 and 62°C. The primer sequences are described in Supplemental Table S2. Routinely a melting curve analysis was performed after the qPCR run (between 55°C and 95°C with 0.5°C increments for 10 sec. each) and for all primers a single peak was observed.

RT-qPCR data analysis. For analyzing our RT-qPCR data we used qbasePLUS (Hellemans et al., 2007) which is commercially available software (Biogazelle, Ghent, Belgium, www.biogazelle.com). For normalizing the data we mined our microarray data for stably expressed genes using a set that was recently tested for stable expression in seeds (Graeber et al., 2011; Dekkers et al., 2012). Six genes (AT1G13320, AT1G17210, AT2G28390, AT3G18780, AT4G34270 and AT5G25760) appeared to be stably expressed and their expression in our samples was confirmed using the geNORM program (Vandesompele et al., 2002), which is integrated into the qbasePLUS software. In the calculation we corrected for primer efficiency which was calculated from the amplification curve using LinReg PCR (Ramakers et al., 2003; Ruijter et al., 2009).

SUPPLEMENTAL FIGURES

Fig. S1. ATH1 Genechip quality assessment and reproducibility. A,B, All 116 ATH1 Genechip arrays showed similar patterns of raw probe intensity. Slide images were manually inspected, with no noticeable spatial artefacts. C, RNA degradation plot shows comparable slopes for all arrays. D, After RMA normalization (Irizarry et al., 2003) the data distributions become comparable, although lower median values are found for the dry and shortly imbibed seeds, which are samples that were isolated

from metabolically less active material. E, The histogram of the normalized data shows separated peaks for noise and signal, and the plot indicates a value of five (on a $\log_2$ scale) as being potentially expressed. F, The correlation between individual replicates are all above 0.980, with the majority (143 out of 174 comparisons) being over 0.990 (Supplemental Table S1). Six individual samples needed to be re-done, with RNA being isolated, labelled and hybridized at a later time. The correlation for these samples was slightly lower, but still above 0.980.

Fig. S2. General expression numbers and the identification and analysis of endosperm and embryo specific gene sets. A, Number of gene expressed (i.e. over 5 on $\log_2$ scale) in different tissues over the whole germination time course. The majority of the genes are shared by all seed compartments. B, Number of genes expressed increased during germination in all compartments. C, Small compartment specific gene sets were identified for the endosperm, embryo, MCE, PE, RAD and COT. D, Simplified reproduction of the ORA of the endosperm specific gene set. E, Simplified reproduction of the ORA of the embryo specific gene set. F, The endosperm and embryo specific gene sets are overrepresented for TFs. The table shows the TF classes and indicates the numbers of each family present on the chip, the number of expressed in the germination time course, and the number of TF genes expressed specifically in endosperm and embryo. p-value is calculated Chi-square test using a Yates correction. TF = transcription factor.

Fig. S3. Comparisons with two other seed microarray datasets. A, Histogram of the probe set values of the Penfield dataset (Penfield et al., 2006). B, Venn diagram showing the overlap between endosperm specific genes in our set at germination (MCE>RAD, 38 HAS ER) or over the whole time course compared the Penfield set (ENDO>EMB, using a 5 fold cutoff). C, Histogram of the probe set values of the Le dataset (Le et al., 2010). D, Table indicates overlap of expression in the endosperm between microdissected data at the post mature green stage and our set at 3, 16 and 31 HAS. E, Overlap of the endosperm and embryo specific sets from the germination time course compared to the microdissected seed development set (embryo and all three endosperm samples).

Fig. S4. RT-qPCR confirms tissue specific expression found in the microarray dataset. A, Indicates the different time points and stages that were sampled along the germination time course. B, The expression pattern of five example genes is depicted on pictograms that represent all 29 samples. Red indicates that the gene is expressed. C, The relative expression level in the different tissues at 31 HAS was calculated based on the microarray data. The seed compartment with the highest expression was set to 1 and indicated by the green colour and low expression was indicated by an orange to red colouring. Similarly the relative expression levels of the qPCR were depicted, with the micropylar (ME) and chalazal endosperm (CE) collected as separate samples. * = genes are part of the MCE specific gene list. Genes indicated in bold are also shown in B. HAS = hours after sowing

Fig. S5. Topological features of the EndoNet and RadNet. Four topological features were computed for both networks A, node degree; B, mean length of shortest paths; C, mean average clustering coefficients; and D, mean betweenness centrality score. The red line in both plots marks the mean value of the feature for entire network. We identified overrepresented GO classes for the five clusters with the highest mean betweenness centrality score (the most important hubs) in E, EndoNet and F, RadNet.

Fig. S6. Overrepresentation analysis of the 30 largest clusters from the EndoNet co-expression network. Clusters were grouped based on their expression pattern ('DOWN', 'UP and DOWN' or 'UP'). The graphs are divided in two parts and show the expression pattern of all genes in the cluster in both the endosperm (left side of each graph) as well as the expression pattern of the same set of genes in the RAD (right), see the schematic graph left of the legend for details. Clusters were analyzed by ORA using Genetrail. In total 25 out of 30 clusters showed overrepresented gene categories which are summarized underneath the graphs.

Fig. S7. ORA using Pageman of genes that are either higher expressed in the MCE or the RAD. Pageman analysis was comparing both tissues at each time point along the time course. Selected classes of the Pageman output were redrawn showing the most obvious differences between both tissues. Red colour indicates gene classes that are overrepresented while the blue colour indicates the underrepresented ones.

Fig. S8. Seed tissues differentiate during germination. A, The number of endosperm and embryo specific genes expressed increase along the germination time course. B, Graphs show the expression along the germination time course of exemplar genes related to stomatal development (Bergmann and Sack, 2007; Liu et al., 2010) and root development (Blilou et al., 2005; Overvoorde et al., 2010; Petricka et al., 2012) including examples of the core auxin biosynthetic pathway (Mashiguchi et al., 2011), auxin transport (Blakeslee et al., 2005) and auxin perception (Mockaitis and Estelle, 2008). The genes related to stomatal development were detected in the COT at 3, 16 and 31 HAS. The other genes were detected in the RAD throughout the whole time course (from 1 to 38 HAS).

Fig. S9. Expression of evolutionary old and young genes during Arabidopsis seed germination. A, The genes encoded on Arabidopsis genome are subdivided in 12 evolutionary age classes (phylostrata) depicted in a phylostratigrapic map. B,C,D, Mean relative expression in the MCE of PS1 and 2, PS3-5 and PS6-12 respectively. E,F,G, Mean relative expression in the RAD of PS1 and 2, PS3-5 and PS6-12 respectively.

Fig. S10. The node degree distribution for the correlation networks, showing power-law behaviour. A, A log-log cumulative frequency plot of the node degree distribution for the combined endosperm network, EndoNet and B, the radicle network, RadNet.

LITERATURE CITED

Bergmann DC, Sack FD (2007) Stomatal development. Annu Rev Plant Biol **58:** 163-181

Blakeslee JJ, Peer WA, Murphy AS (2005) Auxin transport. Curr Opin Plant Biol **8:** 494-500

Blilou I, Xu J, Wildwater M, Willemsen V, Paponov I, Friml J, Heidstra R, Aida M, Palme K, Scheres B (2005) The PIN auxin efflux facilitator network controls growth and patterning in Arabidopsis roots. Nature **433:** 39-44

Clauset A, Shalizi CR, Newman MEJ (2009) Power-Law Distributions in Empirical Data. Siam Review **51:** 661-703

Dai M, Wang P, Boyd AD, Kostov G, Athey B, Jones EG, Bunney WE, Myers RM, Speed TP, Akil H, Watson SJ, Meng F (2005) Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. Nucleic Acids Res **33:** e175

Dekkers BJ, Willems L, Bassel GW, van Bolderen-Veldkamp RP, Ligterink W, Hilhorst HW, Bentsink L (2012) Identification of reference genes for RT-qPCR expression analysis in Arabidopsis and tomato seeds. Plant Cell Physiol **53:** 28-37

Domazet-Loso T, Tautz D (2010) A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. Nature **468:** 815-818

Glaab E, Baudot A, Krasnogor N, Valencia A (2010) TopoGSA: network topological gene set analysis. Bioinformatics **26:** 1271-1272

Graeber K, Linkies A, Wood AT, Leubner-Metzger G (2011) A guideline to family-wide comparative state-of-the-art quantitative RT-PCR analysis exemplified with a Brassicaceae cross-species seed germination case study. Plant Cell **23:** 2045-2063

Hellemans J, Mortier G, De Paepe A, Speleman F, Vandesompele J (2007) qBase relative quantification framework and software for management and automated analysis of real-time quantitative PCR data. Genome Biol **8:** R19

Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics **4:** 249-264

ISTA (2009) International rules for seed testing. International Seed Testing Association, Basserdorf

Joosen RVL, Ligterink W, Dekkers BJW, Hilhorst HWM (2011) Visualization of molecular processes associated with seed dormancy and germination using MapMan. Seed Science Research **21:** 143-152

Keller A, Backes C, Al-Awadhi M, Gerasch A, Kuntzer J, Kohlbacher O, Kaufmann M, Lenhof HP (2008) GeneTrailExpress: a web-based pipeline for the statistical evaluation of microarray experiments. BMC Bioinformatics **9:** 552

Le BH, Cheng C, Bui AQ, Wagmaister JA, Henry KF, Pelletier J, Kwong L, Belmonte M, Kirkbride R, Horvath S, Drews GN, Fischer RL, Okamuro JK, Harada JJ, Goldberg RB (2010) Global analysis of gene activity during Arabidopsis seed development and identification of seed-specific transcription factors. Proc Natl Acad Sci U S A **107:** 8063-8070

Lee D, Polisensky DH, Braam J (2005) Genome-wide identification of touch- and darkness-regulated Arabidopsis genes: a focus on calmodulin-like and XTH genes. New Phytol **165:** 429-444

Liu YK, Liu YB, Zhang MY, Li DQ (2010) Stomatal development and movement: the roles of MAPK signaling. Plant Signal Behav **5:** 1176-1180

Mashiguchi K, Tanaka K, Sakai T, Sugawara S, Kawaide H, Natsume M, Hanada A, Yaeno T, Shirasu K, Yao H, McSteen P, Zhao Y, Hayashi K, Kamiya Y, Kasahara H (2011) The main auxin biosynthesis pathway in Arabidopsis. Proc Natl Acad Sci U S A **108:** 18512-18517

Mockaitis K, Estelle M (2008) Auxin receptors and plant development: a new signaling paradigm. Annu Rev Cell Dev Biol **24:** 55-80

Morris JH, Apeltsin L, Newman AM, Baumbach J, Wittkop T, Su G, Bader GD, Ferrin TE (2011) clusterMaker: a multi-algorithm clustering plugin for Cytoscape. BMC Bioinformatics **12:** 436

Nolan T, Hands RE, Bustin SA (2006) Quantification of mRNA using real-time RT-PCR. Nat Protoc **1:** 1559-1582

Overvoorde P, Fukaki H, Beeckman T (2010) Auxin control of root development. Cold Spring Harb Perspect Biol **2:** a001537

Penfield S, Li Y, Gilday AD, Graham S, Graham IA (2006) Arabidopsis ABA INSENSITIVE4 regulates lipid mobilization in the embryo and reveals repression of seed germination by the endosperm. Plant Cell **18:** 1887-1899

Petricka JJ, Winter CM, Benfey PN (2012) Control of Arabidopsis root development. Annu Rev Plant Biol **63:** 563-590

Quint M, Drost HG, Gabel A, Ullrich KK, Bonn M, Grosse I (2012) A transcriptomic hourglass in plant embryogenesis. Nature **490:** 98-101

Ramakers C, Ruijter JM, Deprez RH, Moorman AF (2003) Assumption-free analysis of quantitative real-time polymerase chain reaction (PCR) data. Neurosci Lett **339:** 62-66

Ruijter JM, Ramakers C, Hoogaars WM, Karlen Y, Bakker O, van den Hoff MJ, Moorman AF (2009) Amplification efficiency: linking baseline and bias in the analysis of quantitative PCR data. Nucleic Acids Res **37:** e45

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res **13:** 2498-2504

Sliwinska E, Bassel GW, Bewley JD (2009) Germination of Arabidopsis thaliana seeds is not completed as a result of elongation of the radicle but of the adjacent transition zone and lower hypocotyl. J Exp Bot **60:** 3587-3594

Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T (2011) Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics **27:** 431-432

Sreenivasulu N, Usadel B, Winter A, Radchuk V, Scholz U, Stein N, Weschke W, Strickert M, Close TJ, Stitt M, Graner A, Wobus U (2008) Barley grain maturation and germination: metabolic pathway and regulatory network commonalities and differences highlighted by new MapMan/PageMan profiling tools. Plant Physiol **146:** 1738-1758

Tremousaygue D, Manevski A, Bardet C, Lescure N, Lescure B (1999) Plant interstitial telomere motifs participate in the control of gene expression in root meristems. Plant J **20:** 553-561

Usadel B, Nagel A, Steinhauser D, Gibon Y, Blasing OE, Redestig H, Sreenivasulu N, Krall L, Hannah MA, Poree F, Fernie AR, Stitt M (2006) PageMan: an interactive ontology tool to generate, display, and annotate overview graphs for profiling experiments. BMC Bioinformatics **7:** 535

Vandesompele J, De Preter K, Pattyn F, Poppe B, Van Roy N, De Paepe A, Speleman F (2002) Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. Genome Biol **3:** RESEARCH0034

Wittkop T, Emig D, Lange S, Rahmann S, Albrecht M, Morris JH, Bocker S, Stoye J, Baumbach J (2010) Partitioning biological data with transitivity clustering. Nat Methods **7:** 419-420

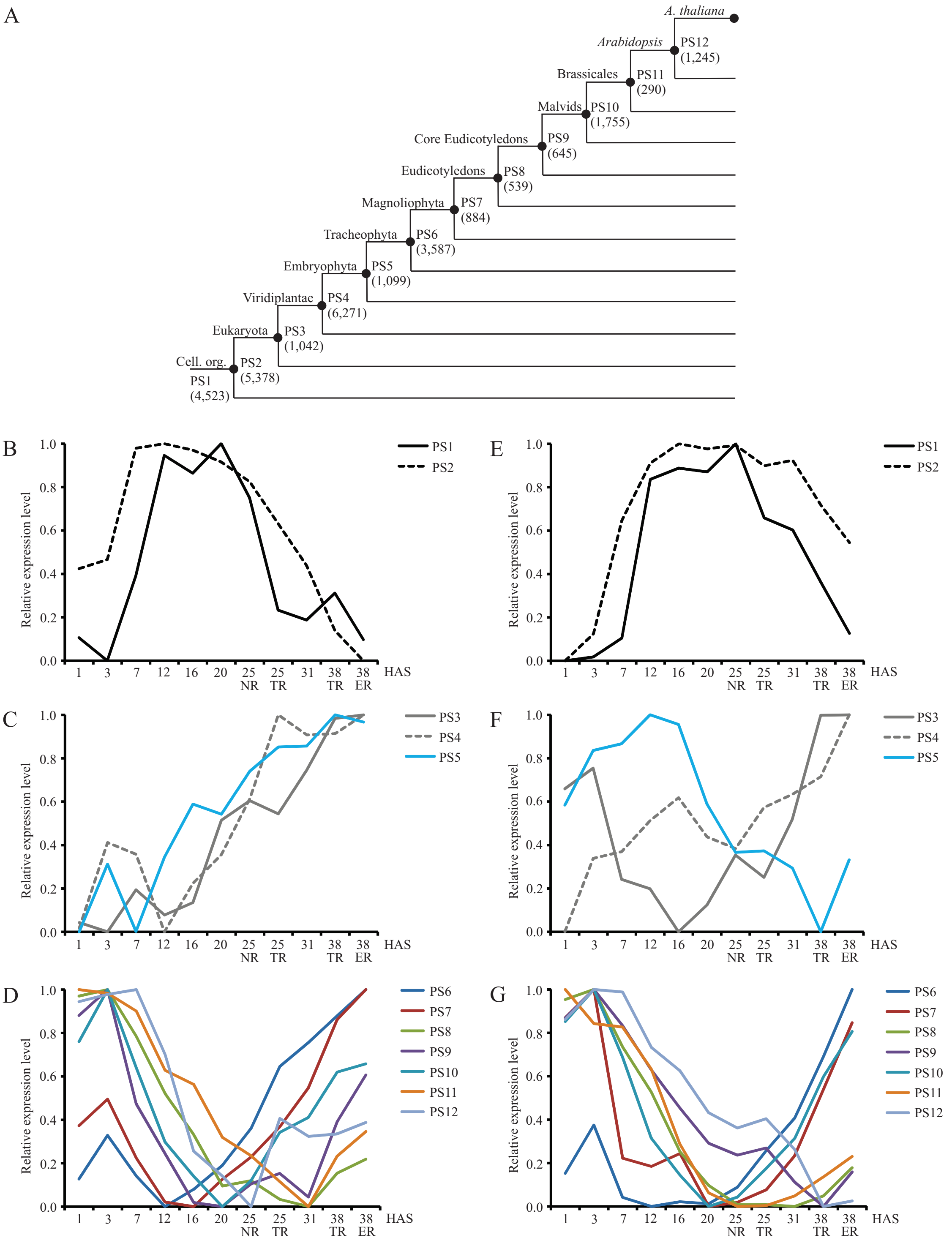


Fig. S9. Expression of evolutionary old and young genes during Arabidopsis seed germination. A, The genes encoded on Arabidopsis genome are subdivided in 12 evolutionary age classes (phylostrata) depicted in a phylostratigrapic map. B,C,D, Mean relative expression in the MCE of PS1 and 2, PS3-5 and PS6-12 respectively. E,F,G, Mean relative expression in the RAD of PS1 and 2, PS3-5 and PS6-12 respectively.

Hajk-Georg **Drost**

Curriculum Vitae

Personal Data

Birthday	04 December 1986 in Halle (Saale), Germany
Address	80A Norwich Street, CB2 1NE Cambridge, UK
Nationality	German
Mobile	+49 176 82053728
Email	hgd23@cam.ac.uk

Education

2013-2016	**Dr. rer. nat.**, *Martin Luther University*, Halle (Saale), Germany, Institute of Computer Science. *Thesis: A bioinformatic study on transcriptome conservation patterns in animal and plant development* *Advisors* Professor Ivo Grosse & Professor Marcel Quint
2011 - 2013	**Master of Science in Bioinformatics**, *Martin Luther University*, Halle (Saale), Germany, Institute of Computer Science. *Thesis: A bioinformatics approach to study the origin of embryogenesis in plants and animals* *Advisors* Professor Ivo Grosse & Professor Marcel Quint
2008 - 2011	**Bachelor of Science in Bioinformatics**, *Martin Luther University*, Halle (Saale), Germany, Institute of Computer Science. *Thesis: Development of a phylogenetic transcriptome atlas of Arabidopsis thaliana* *Advisors* Professor Ivo Grosse & Professor Marcel Quint
2007	**Abitur**, *Elisabeth Gymnasium*, Halle (Saale), Germany.
2003/04	**Foreign Exchange Student**, *Niagara Wheatfield Senior Highschool*, Sanborn, NY, USA.

Experience

2015–present **Research Associate**, *University of Cambridge*, Cambridge, UK, Sainsbury Laboratory.

Research Topics: Evolutionary Bioinformatics and Epigenetics

Lab - Jerzy Paszkowski

2015 **Research Collaboration**, *Sainsbury Laboratory Cambridge*, Cambridge, UK.

Detection of evolutionary signals in developmental transcriptomes of *Arabidopsis thaliana* organs with a special focus on splice variants and lncRNAs using Next-Generation Sequencing technologies and bioinformatics.

Collaborators Dr Christoph Schuster & Prof Elliot Meyerowitz

2013–2015 **Research Assistent**, *Martin Luther University*, Halle (Saale), Germany, Institute of Computer Science, Bioinformatics.

Research Topics: Performing comparative meta-genomics and evolutionary transcriptomics to decipher the developmental hourglass phenomenon in animals and plants.

Lab - Ivo Grosse

2013 **Research Collaboration**, *Wageningen Seed Laboratory, Wageningen University and Research Centre and Leibniz Institute of Plant Biochemistry*, Wageningen, NL and Halle (Saale), Germany.

Evolutionary transcriptomics of *Arabidopsis* seed germination.

Collaborators Professor Marcel Quint, Dr Bas Dekkers & Prof Leonie Bentsink

2012 **Research Intern**, *Medicinal Chemistry*, Martin Luther University, Halle (Saale), Germany, Institute of Pharmacy.

Benchmarking Molecular Docking algorithms using the PDBbind database and implementing an automated pipeline for quantifying the goodness of docking.

Advisor Professor Wolfgang Sippl

2011 **Research Intern**, *Helmholtz Centre for Environmental Research*, Halle (Saale), Germany.

Implementation of a software framework for computational population genetic analyses.

Advisor Dr Walter Durka

Publications

2016 **HG Drost**, A Gabel, T Domazet-Lošo, M Quint, I Grosse. *Capturing Evolutionary Signatures in Transcriptomes with myTAI* **bioRxiv** doi:http://dx.doi.org/10.1101/051565 (2016).

2016 **HG Drost**, J Bellstaedt, DS O'Maoileidigh, AT Silva, A Gabel, C Weinholdt, PT Ryan, BJ Dekkers, L Bentsink, HW Hilhorst, W Ligterink, F Wellmer, I Grosse, M Quint. *Post-embryonic hourglass patterns mark ontogenetic transitions in plant development.* **Mol. Biol. Evol.** 33 (5): 1158-1163 (2016). (co-corresponding author; journal cover story)

2015 **HG Drost**, A Gabel, I Grosse, M Quint. *Evidence for Active Maintenance of Phylotranscriptomic Hourglass Patterns in Animal and Plant Embryogenesis.* **Mol. Biol. Evol.** 32 (5): 1221-1231 (2015).

2015 PT Ryan, DS O'Maoileidigh, **HG Drost**, *et al. Patterns of gene expression during Arabidopsis flower development from the time of initiation to maturation*. **BMC Genomics** 16:488 (2015).

2013 BJW Dekkers, S Pearce, RP van Bolderen-Veldkamp, A Marshall, P Widera, J Gilbert, **HG Drost** *et al. Transcriptional dynamics of two seed compartments with opposing roles in Arabidopsis seed germination*. **Plant Physiology** 163 (1), 205-215 (2013).

2012 M Quint, **HG Drost**, A Gabel, KK Ullrich, M Boenn, I Grosse. *A transcriptomic hourglass in plant embryogenesis*. **Nature** 490 (7418), 89-101. (journal cover story)

2012 **HG Drost**. *Development of a phylogenetic transcriptome atlas of Arabidopsis thaliana embryo development*. **Lecture Notes in Informatics (LNI)** - Seminars, Vol. S-11, 175-178 (2012).

Teaching Experience

2016 **Supervision: Research Project**, *University of Cambridge*, Cambridge, UK.
Jonathan Williams, Project Title: *Towards a Predictive Model of DNA Cytosine Methylation in Arabidopsis thaliana*

2015 **Supervision: Master's Thesis**, *Martin Luther University*, Halle (Saale), Germany.
B.Sc. Norman Urban, Thesis Title: *Orthology Inference Methods and their Influence on Divergence Stratigraphy*

2015 **Supervision: Student Project** , *Martin Luther University*, Halle (Saale), Germany.
B.Sc. Anne Hoffmann (Master Student), Project Title: *Detection of evolutionary signals in biological cycles.*

2015 **Supervision: Student Project**, *Martin Luther University*, Halle (Saale), Germany.
B.Sc. Sebastian Wussow (Master Student), Project Title: *Phylotranscriptomics of the circadian cycle in animals and plants.*

2015 **Bachelor Course**, *Martin Luther University*, Halle (Saale), Germany.
Seminar: Selected Problems in Bioinformatics

2014 **PhD Course**, *Martin Luther University*, Halle (Saale), Germany.
Seminar and Tutorial: Expression Data Analysis using R

2014 **Master Course**, *Martin Luther University*, Halle (Saale), Germany.
Seminar: Analysis of Biological Networks

2013 **Master Course**, *Martin Luther University*, Halle (Saale), Germany.
Journal Club on Evolutionary Transcriptomics and Next-Generation Sequencing

2013 **Bachelor Course**, *Martin Luther University*, Halle (Saale), Germany.
Seminar: Selected Problems in Bioinformatics

Presentations

2014 **Conference Talk**, *Martin Luther University* , Halle (Saale), Germany.
Symposium on Novel Applications of Deep Sequencing in Medicine, Genomics, and Biodiversity Research. Talk: *Phylotranscriptomic Hourglasses in Animals and Plants.*

Open Source Software

2016 **biomartr:** Genomic Data Retrieval with R.
`https://github.com/HajkD/biomartr.`

2016 **philentropy:** Information Theory and Distance Quantification with R.
`https://github.com/HajkD/philentropy.`

2015 **orthologr:** Comparative Genomics with R.
`https://github.com/HajkD/orthologr`

2015 **myTAI:** Evolutionary Transcriptomics with R.
`https://github.com/HajkD/myTAI.`

2015 **seqreadr:** Read Genomic File Formats with R.
`https://github.com/HajkD/seqreadr`

Awards

2014 FERCHAU Graduation Award 2014

2013 SKW Piesteritz Research Award 2013

2012 Georg-Cantor Research Award 2012

Interests

History of Science

Collecting original scientific literature (books) published between 1700 - 2000

Football, Volleyball, and Running