Transcription associated proteins in plant development and evolution

Dissertation

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

dem Fachbereich Biologie der Philipps Universität Marburg vorgelegt von

Per K. I. Wilhelmsson

aus Malmö, Schweden

Marburg an der Lahn

2019

Philipps Universität Marburg

Von der Philipps-Universität Marburg als Dissertation angenommen am

Ersgutachter: Prof. Dr. Stefan A. Rensing

Zweitgutachter: Prof. Dr. Uwe Maier

Drittgutachter: Prof. Dr. Anke Becker

Viertgutachter: Prof. Dr. Dominik Heider

Tag der Disputation am _____

"I do not fear computers.
I fear the lack of them."
Isaac Asimov

Abstract

Gene expression, the process in which DNA information is conveyed into a functional unit, is fundamental to cellular life. The extent to which gene expression can be regulated corresponds to a cells potential to modify its ability. One example is the progression through life stages of e.g. plants, going from seed to a tree, all achieved through different application of gene regulation upon the same identical DNA information. Transcriptional regulation is the process of regulating the initial step in gene expression, the transcribing of DNA into RNA. This is carried out by transcription associated proteins (TAPs).

The work in this thesis aims to increase our knowledge of TAP involvement in plant development and to shed new light on TAP evolution in plants.

By first providing an up-to-date method to screen for TAPs in plants (TAPscan), it was possible to screen a wide selection of plant genomes and transcriptomes. Using the data, ancestral states as well as gains, losses, expansion and contractions of TAPs, throughout the evolution of plants, could be calculated. The results suggest that many previously thought to be land plant specific TAPs actually predates the emergence of land plants.

By analyzing RNA-sequence (RNA-seq) libraries of the dimorphic seed producing plant *Aethionema arabicum* (provided through the SeedAdapt consortium) it was possible to investigate TAP influence on seed development. In addition, a study evaluating the usefulness of a *de novo* assembly compared to a reference genome when identifying differentially expressed genes was conducted. The RNA-seq analysis, with TAP annotations, showed a clear distinction between the two seed morphs. The dehiscent (short term) seed being geared towards faster maturation and the indehiscent (long term) seed being geared towards

dormancy was evident using both the *de novo* and reference approach.

Zusammenfassung

Die Genexpression ist ein fundamentaler Prozess der Zellbiologie. In diesem werden die in der DNA enthaltenen Informationen in funktionale Einheiten umgesetzt. Das Maß, in dem die Genexpression reguliert werden kann, korreliert dabei mit dem Potential einer Zelle verschiedenste Funktionen auszubilden. Als ein Beispiel dafür kann das Durchschreiten des pflanzlichen Lebenszyklus gesehen werden, wie es sich z.B. bei der Entwicklung eines Baumes aus einem Samen vollzieht. Dies wird durch unterschiedliche Anwendungen der Genregulation auf identische DNA Informationen möglich. Dabei bildet die Transkription, das Umschreiben der DNA in RNA, den initialen Schritt der Genexpression. Die Regulation der Transkription erfolgt durch transkriptions-assoziierte Proteine (*transcription associated proteins*, TAPs).

Die vorliegende Arbeit erweitert das Wissen über die Rolle von TAPs in der pflanzlichen Entwicklung und zeigt neue Aspekte ihrer Evolution in Pflanzen auf.

In einem ersten Schritt wurde eine Methode zur Detektion von TAPs in Pflanzen (TAPscan) etabliert. Unter Anwendung dieser Methode wurde eine große Anzahl pflanzlicher Genome und Transkriptome auf das Vorhandensein verschiedener TAP Gruppen untersucht. Auf Grundlage der erhobenen Daten ließen sich ursprüngliche Zustände sowie die Entstehung, Expansion und der vollständige oder teilweise Verlust verschiedener TAP Familien über die Evolution der Pflanzen hinweg nachvollziehen. Die Ergebnisse lassen darauf schließen, dass sich viele TAP Familien, deren Entstehung bisher mit dem Landgang der Pflanzen in Verbindung gebracht wurde, bereits vor den ersten Landpflanzen entwickelten.

Des Weiteren wurde der Einfluss von TAPs auf die Samenentwicklung untersucht. Diese Untersuchungen wurden anhand von *Aethionema arabicum* (zur Verfügung gestellt durch das SeedAdapt Konsortiums), einer Pflanze, die dimorphe Samen ausbildet, durchgeführt. Dazu wurden Daten aus RNA Sequenzierungen (RNA-seq) analysiert und TAPscan auf diese Daten angewendet. Im Zuge dessen konnte auch der Nutzen eines *de novo* Transkriptoms im Vergleich zu einem Referenzgenom bei der Identifizierung differenziell exprimierter Gene gezeigt werden. Die Analysen der Sequenzierungsdaten, auch unter Anwendung von TAPscan, konnten deutliche Unterschiede zwischen den beiden untersuchten Samenformen belegen. In den Analysen wurden, sowohl unter Anwendung des *de novo* Transkriptoms als auch des Referenzgenoms, die Anlagen der dehiszenten (Kurzzeit-) Samenform für schnelle Samenreife und die der indehiszenten (Langzeit-) Samenform für erhöhte Dormanz deutlich.

Contents

Abbreviations

ABA	Abscisic acid
bHLH	basic Helix-Loop-Helix
DEG	Differentially Expressed Gene
DNA	Deoxyribonucleic Acid
GA	Gibberellin
Gbp	Giga base pairs
GO	Gene Ontology
kbp	Kilo base pairs
M+	Mucilaginous (dehiscent seed)
NM	Non-Mucilaginous (indehiscent seed)
RNA	Ribonucleic Acid
TAP	Transcription Associated Protein
TF	Transcription Factor
TR	Transcriptional Regulator

Chapter 1

General introduction

1 General introduction

1.1 Bioinformatics and the age of sequencing

The use of sequencing technologies has throughout the last decades grown to become an essential part of biological research. In the early years of sequence analysis, in the wake of Watson and Crick 1953 [1] providing the three-dimensional structure of DNA, techniques were developed to determine nucleotide sequence. In 1961 Nirenberg et al. [2] demonstrated the codon nature of the DNA, showing that a triplet of DNA corresponds to an amino acid in the final protein. Holley et al. [3] developed a method to determine in which order the nucleotides were positioned and could in 1965 present the first nucleotide sequence, the alanine tRNA sequence from Saccharomyces cerevisiae [4]. In 1972 Min-Jou et al. [5] were able to determine the protein coding RNA sequence of the bacteriophage MS2 and thus opened the door to transcriptomics, being the study of a given samples RNA-transcripts (transcriptome), making it possible to investigate gene expression. In 1977 Sanger et al. [6] presented the 5,386 bp long DNA genome of the bacteriophage phi X, being the first sequenced DNA genome.

At this point, the sequencing process was tedious and resulted in very few sequences, which is reflected in there only being 606 sequences (680 kbp) available in the first public release of the GenBank sequence database of 1982 [7]. Through the following years much progress was made with regards to optimizing, automating and parallelizing sequencing driven by the goal of sequencing the human genome [8, 9]. In 2000 the first draft of the human genome, with its ~3 billion base pairs, was sequenced at a total cost of approximately 0.1$ per base pair [10]. In 2017 the price for sequencing a human genome was already below 1,000$, with sequencing instrument manufactures having the goal to go below the 100$ mark [11]. As for RNA-sequencing, sequencing the active protein coding parts of the genome (transcriptome), the possibility to generate large amount of sequences has been of great use for the field of

comparative genomics e.g. when studying gene expression between different traits, conditions, tissues etc.

With the reduce in effort and costs required to determine DNA sequences the total accumulated amount of publicly available nucleotide sequence doubles approximately every 18 month with the GenBank release of June 2018 containing more than 850 million sequences (3,200 Gbp) [7]. With this growing pile of data new possibilities and challenges have emerged. The need to handle the large supply of data, to analyze and make the data accessible, not only requires well established routines but also creates the opportunity to develop and improve analytic tools and pipelines. Thanks to the decreasing costs and increasing availability of sequenced DNA more large-scale projects can now be realized, such as the TAPscan online resource (https://plantcode.online.uni-marburg.de/tapscan/) (chapter 2) and the SeedAdapt collaborative research project (www.seedadapt.eu) (chapter 3).

1.2 Transcription associated proteins

Gene expression is the process in which the information encoded in DNA is converted into a functional unit, of either a protein or a functional RNA. This process starts with the transcription, where RNA-polymerase breaks open the DNA, runs through the stretch of the gene while creating a RNA copy. This copy can either act as a functional unit in itself or in cooperation with the ribosomes, where codon matching amino acids are linked together, result in a protein. The extent to which the process of gene expression can be regulated is connected to the potential complexity of the gene network organization and ultimately to the potential complexity of the organism.

Transcription associated proteins (TAPs) comprise proteins that are involved in regulation of transcription. This is done either through sequence specific binding to cis-regulatory elements by transcription factors (TFs), which can enhance or repress transcription, or through unspecific binding, protein-protein interactions or chromatin modification by transcription regulators (TRs).

Figure 1. Wild type *A. thaliana* (A) and blr (HD TF) mutant phenotype (B) [13].

TFs and TRs make it possible to, through transcription, regulate the expression of genes. One example are the well-studied homeodomain (HD) TFs that were first discovered in 1984 [12] and were shown to be involved in body plane/pattern formation in *Drosophila melanogaster*. These TFs were later also discovered to be conserved in all vertebrates as well as other animals, fungi and plants. HD involvement in body plane/pattern formation for plants has also been shown [13] (Fig. 1). Homeodomain TFs are one of the most abundant TF amongst metazoans (animals) with about 15-30% of all know TFs being homeodomain [14] making up for approximately 0.5 – 1.25% of all proteins in any given species [15].

It has been shown that TAPs played a key role in the acquisition of multicellularity and morphologic complexity amongst eukaryotes [14, 16, 17]. Metazoans and embryophytes (land plants) have the richest TF repertoire amongst eukaryotes and it is thought to be required to orchestrated the embryonic development [14, 18]. These complex repertoires were acquired in a stepwise manner with bursts of TF innovation in respective lineages unicellular ancestors followed by

further TF expansion at the origin of both metazoans and embryophytes [14, 16, 19, 20]. In these studies it is apparent that some TAP profiles are lineage/clade specific making them a very interesting group of proteins to study from an comparative phylogenomic perspective.

In many parts of the tree of life, there is an ongoing work filling the gaps of not yet sequenced species with some grand initiatives in both the animal (The Genome 10K Project) [21] and plant kingdom (10K Plants) [22]. For plants there has up until recent times been an understandable bias towards sequencing the more socioeconomically important flowering plants. This has left the field of early plant evolution struggling with less than the desirable amount of available sequenced species [23]. When land was conquered by plants, approximately 500 million years ago [24], the plant kingdom ventured into a completely new environment. The opportunity was exploited and the result is the approximately 500,000 species of embryophytes living today [25]. A complete picture of events that occurred during the terrestrialization cannot be drawn. Though, the most likely scenario is that a green algae, that either evolved in freshwater or adapted to freshwater from a marine environment, transitioned to land [26].

When looking into TAP evolution in Viridiplantae (green plants), comprising green algae and Embryophyta, it has until recently not been possible to get a clear view due to the missing data points. In 2000 the model organism *Arabidopsis thaliana* was sequenced [27] which opened the door to the sequencing of other flowering plants. In 2006 the first green algae was sequenced [28] which was followed by the green algae model organism *Chlamydomonas reinhardtii* in 2007 [29]. In 2008 the first non-vascular plant, *Physcomitrella patens*, was sequenced [30] which initiated the gap-closing between green algae and vascular plants. With genomes from these clades, vascular, non-vascular plants and green algae, Lang et al. [16] were able to detect an increasingly complex TAP repertoire being positively correlated with the morphologic

complexity (number of cell types). An evolutionary interesting group of algae, of which there were no large-scale sequence data of at the time, was the paraphyletic group of streptophyte algae. These are morphologically complex algae phylogenetically placed in between Chlorophyta and Embryophyta. It is thought that these closest living relatives of land plants harbors the key to understanding the colonization of land [31]. The first emerging streptophyte algae genomes, *Klebsormidium flaccidum* 2014 [32], directly raised the point that before thought to be land specific proteins (including many TAPs) has to be revised [32-34]. It is within the scope of this thesis to shed light and hopefully bring new knowledge to the current view of TAP evolution in Viridiplantae (chapter 2, paper 1).

1.3 Seed development

The aspiration to understand seeds have been a key to part in the prosperity of mankind. What initially was a struggle to understand and maintain a stable source of food, bringing the fruitful wild into your own backyard developed into breeding industry with the latest genetic engineering making it possible to find new ways to improve the yield.

Yield and quality was something the domesticators sought for when domesticating sorghum (Africa year ~6,000 years ago [35]), soybean (East Asia year 5-9,000 years ago [36]), sugar beet (Europe 18th century [37]), maize (North American year ~4,200 years ago [38]) and potato (South America year 5,4-4,200 years ago [39]). The underlying molecular mechanisms that these domesticated plants ended up to have altered were discovered far later, e.g. in maize 1939 [40]. Today, the importance of plants and seeds is reflected in the more than 1,000 existing seed banks around the world that aims to maintain a seed backup in case of crisis.

Plant abiotic stresses (including heat and drought) are major factors limiting the chances for plants to propagate and carry out offspring. To face this, plants have evolved mechanisms that, through sensing the

environment, adapt its seed formation and seed germination [41]. The units (diaspores) that gets dispersed by a plant, seed or tissue covered seed (fruit), are then provided with the mechanisms to find its most optimal window to germinate, based on its own sensing. Controlling the germination process makes it possible for seeds to avoid germination in shorter temporary favorable conditions in overall less favorable conditions to instead remain in the soil seed bank waiting for better conditions, thus picking the right moment to take this crucial step in the plant's life cycle.

For seed development as with many other well characterized biological systems in plants, *A. thaliana* is the model in which they have been studied the most. *A. thaliana* produces homomorphic diaspores that splits open (dehiscent) along a predetermined line to release its seeds. Once the seed is dispersed it is faced with either remaining dormant or breaking its dormancy mechanisms (germination block) to initiate germination. The plant hormone abscisic acid (ABA) has been shown to play a major part in both inducing the dormancy, through ABA synthesis in the embryo during its development, and prolonging the dormancy through ABA production by the seed itself during development [42]. Depending on the environmental cues the dormancy is broken and the seed moves towards germination. This starts with the seed increasing its water uptake (imbibition) and ends with the radicle part of the embryo bursting out of its protective coats (endosperm and testa). Gaining more insight into the molecular mechanisms of these fruit/seed traits are of great importance for both ecology and evolution research and for the seed industry and crop breeding as well.

There are multiple angiosperm families where the ability to produce heteromorphic diaspores (fruits and seeds) have evolved as adaptive traits [43]. With the different morphs comes different properties. Germination time, due to dissimilar dormancy mechanisms between the morphs, and differences in dispersal distance, due to differing

morphologies, works as a bet-hedging strategy enabling the progeny to escape both time and space. Thus, a plant with heteromorphic diaspores would be the optimal system with regards conducting a comparative analysis to study the differing traits since genetic differences due to studying separate individuals are erased.

Aethionema arabicum belongs to the genus Aethionema, the early diverging clade of the Brassicaceae and can be found throughout the middle east in arid and semiarid environments [43]. What makes *Ae. Arabicum* special is that it has dimorphic diaspores developing on the same plant [44]. One route resembles the default-pathway of *A. thaliana* (orange colored route Fig. 2).

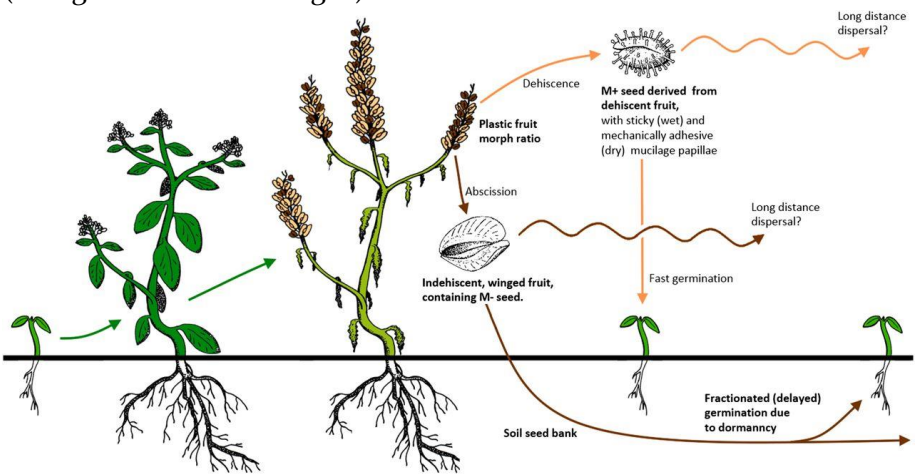


Figure 2. Visualization of the seed dispersal strategies employed by *Ae. Arabicum* [45].

Here the fruit splits open along its dehiscent zone and the seeds becomes mucilaginous upon imbibition with the radicles emerging first during germination. This represents a short-term dispersal strategy. In the second, novel, strategy (brown colored route Fig. 2) the fruit does not fully develop its dehiscent zone, does not become mucilaginous upon imbibition and once germination occurs the cotyledon emerges first. This approach represents the more long-term dispersal strategy. The fruit-morph ratio that the plant produces has been shown to be

connected to environmental factors with fruit morphs not being randomly distributed on the plant. Instead, side branches have been shown to produce more indehiscent fruits in comparison to the main branch [45].

Using two accessions of *Ae. arabicum*, one representing a cold/wet environment (Turkey), and one representing a warm/dry environment (Cyprus), the SeedAdapt consortium set out to investigate and gain insight into the regulatory mechanisms behind the fruit, seed, and seedling traits that evolved as adaptions to abiotic stresses. The whole consortium encompasses comparative analysis on epigenetic, hormonal and transcriptional level as well as studies involving abiotic stress physiology and biochemistry, identifying Quantitative Trait Loci (QTL) and seed bio-mechanics. Separate case studies were carried out to investigate fruit development and the effect of light on the inhibition of germination. In addition, a pilot case study was carried out to identify differentially expressed genes between dehiscent and indehiscent seeds (chapter 2, paper 1), all together resulting in the SeedAdapt consortium generating more than 300 RNA-seq libraries.

1.4 Question and objectives

Will increasing the species sample size change the current view on TAP evolution in Viridiplantae?

By using a wide selection of genomic and transcriptomic sequence data a comprehensive classification of TAPs will be carried out to investigate the TAP gains, losses, expansion and contractions throughout Viridiplantae. With the inclusion of streptophyte algae there is hope to gain further insight into the clouded parts of the phylogenetic tree of plants and shed new light on the evolution of TAPs in plants.

Are there different gene expression profiles between the dimorphic seeds of *Ae. arabicum* and to what extent do these coincide depending on using a genome or a *de novo* assembly approach?

The aim is to conduct a DEG analysis on the dimorphic seeds of *Ae. arabicum* and to investigate the reliability of using *de novo* assembled transcriptomes compared to having a reference genome. This will yield insight into the molecular differences between the two seed morphs as well as into the usability of a *de novo* transcriptome in comparison to a reference genome on the basis of DEG detection and functional annotation. The resulting DEG-pipeline will also be applied to additional projects within the SeedAdapt consortium.

1.5 Thesis structure

Chapter 1 ("General introduction") contains a general introduction to my research topics. This serves to introduce the reader to the bioinformatic realm of science as well as to my research topics.

In chapter 2 ("Comprehensive Genome-Wide Classification Reveals That Many Plant-Specific Transcription Factors Evolved in Streptophyte Algae"), a wide range of Viridiplantae were screened for 122 TAPs. The results were then used to elucidate the ancestral states, expansions/contractions and gains and losses of TAPs throughout Viridiplantae.

In chapter 3 ("Usability of reference-free transcriptome assemblies for detection of differential expression: a case study on *Ae. arabicum* dimorphic seeds"), mRNA libraries of dehiscent and indehiscent seeds were used to identify differentially expressed genes using both the available genome of *Ae. arabicum* as well as with a *de novo* assembled transcriptome of the RNA libraries. The annotated information, TAPs and Gene Ontology terms, gained using both approaches was compared and evaluated.

Chapter 4 ("TAPs and *Ae. arabicum*"), contains an overview of the TAP expression during the developmental stages of *Ae. arabicum* seeds using both published and unpublished data.

Chapter 5 ("Outlook") contains a reflection where my research objectives are treated separately as well as together as a whole. The strengths and weaknesses of the thesis work is highlighted and proposals for future actions are given.

Chapter 2

Paper I

2 Comprehensive Genome-Wide Classification Reveals That Many Plant-Specific Transcription Factors Evolved in Streptophyte Algae (Paper I)

2.1 Zusammenfassung

Transkriptions-assoziierte Proteine (*transcription associated proteins*, TAPs) nehmen sowohl direkten als auch indirekten Einfluss auf den fundamentalen zellulären Prozess der Transkription. Aufgrund dieser Eigenschaft sind sie von besonderer Bedeutung für die Entwicklung und Differenzierung eines Organismus. Daher stellen TAPs eine Schlüsselgruppe der Proteine dar, um ein besseres Verständnis dieser bedeutenden Prozesse zu entwickeln. Im Rahmen dieser Studie wurden durch die Erstellung einer aktualisierten TAP Klassifikation und die Betrachtung einer Vielzahl pflanzlicher Genome und Transkriptome spezies- und gruppenspezifische TAP Profile erstellt. Durch Vergleich der verschiedenen TAP Profile wurde deutlich, dass zahlreiche TAPs, deren Auftreten bisher mit dem Landgang der Pflanzen in Verbindung gebracht wurde, bereits vor diesem evolvierten. Diese Studie zeigt also eine primäre TAP Expansion im gemeinsamen Vorfahren der Streptophyta und nicht in dem der Landpflanzen. Alle erstellten TAP Profile sind über eine benutzerfreundliche Web-Oberfläche öffentlich zugänglich.

2.2 Summary

Transcription associated proteins (TAPs) are known for having both direct and indirect effect upon the fundamental cellular mechanism of transcription. This property has made them intrinsic for development and differentiation in organisms, and they are thus a key group of proteins to study to understand these processes. Using an up to date TAP classification scheme in combination with a wide selection of plant genomes and transcriptomes, each species, as well as clade, could be

assigned a TAP profile. By comparing TAP profiles it could be concluded that many TAPs, thought to have emerged with land plants or during land plant evolution, predated the terrestrialization. This study suggesting that the primary burst of TAP gains occurred in the common ancestor of Streptophyta and not within the common ancestor of land plants. All of the TAP profiles are publicly available through a user-friendly web interface.

2.3 Own contribution

The original pipeline developed by Lang et al. [16] was used as a foundation and further improved upon. The custom-made domain models were re-built by me using a phylogenetically guided species selection resulting in an increased scope of detection. In cooperation with Kristian K. Ullrich and Stefan A. Rensing, current literature was screened for novel TAP families which resulted in the addition of 12 new sub-family classifications. In cooperation with Cornelia Mühlich, a broad selection of Viridiplantae genomes were collected and then screened using the new updated pipeline. Joint with Stefan A. Rensing the results were analyzed to identify expansions, contractions, gains and losses of TAPs throughout the phylogenetic tree. I contributed to the writing of the manuscript as well as prepared most of its figures. Establishment of the web-interface was solely done by Cornelia Mühlich.

2.4 Paper

Following is the electronic publication.

Comprehensive Genome-Wide Classification Reveals That Many Plant-Specific Transcription Factors Evolved in Streptophyte Algae

Per K.I. Wilhelmsson[1], Cornelia Mühlich[1], Kristian K. Ullrich[1,3], and Stefan A. Rensing[1,2,*]

[1]Plant Cell Biology, Faculty of Biology, University of Marburg, Germany

[2]BIOSS Center for Biological Signaling Studies, University of Freiburg, Germany

[3]Present address: Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Biology, Ploen, Germany

*Corresponding author: E-mail: stefan.rensing@biologie.uni-marburg.de.

Accepted: December 4, 2017

Abstract

Plant genomes encode many lineage-specific, unique transcription factors. Expansion of such gene families has been previously found to coincide with the evolution of morphological complexity, although comparative analyses have been hampered by severe sampling bias. Here, we make use of the recently increased availability of plant genomes. We have updated and expanded previous rule sets for domain-based classification of transcription associated proteins (TAPs), comprising transcription factors and transcriptional regulators. The genome-wide annotation of these protein families has been analyzed and made available via the novel TAPscan web interface. We find that many TAP families previously thought to be specific for land plants actually evolved in streptophyte (charophyte) algae; 26 out of 36 TAP family gains are inferred to have occurred in the common ancestor of the Streptophyta (uniting the land plants—Embryophyta—with their closest algal relatives). In contrast, expansions of TAP families were found to occur throughout streptophyte evolution. 17 out of 76 expansion events were found to be common to all land plants and thus probably evolved concomitant with the water-to-land-transition.

Key words: Charophyta, Streptophyta, Embryophyta, evolution, transcription, land plant.

Introduction

Transcriptional regulation is carried out by transcription associated proteins (TAPs), comprising transcription factors (TFs, binding in sequence-specific manner to *cis*-regulatory elements to enhance or repress transcription), transcriptional regulators (TRs, acting as part of the transcription core complex, via unspecific binding, protein–protein interaction or chromatin modification) and putative TAPs (PTs), the role of which needs to be determined (Richardt et al. 2007).

The complexity of transcriptional regulation (as measured by the genomes' potential to encode TAPs, i.e., total number of TAP genes per genome) coincides with the morphological complexity (typically measured by number of cell types) of plants and animals (Levine and Tjian 2003; Lang et al. 2010; de Mendoza et al. 2013; Lang and Rensing 2015). Comparative studies in plants and animals have revealed gains, losses and expansions of key gene families, and demonstrated the unicellular ancestors of plants and animals had already gained much of the families known as important and typical for these lineages (Lang et al. 2010; de Mendoza et al. 2013; de Mendoza et al. 2015; Catarino et al. 2016). The recent initial analysis of data from streptophyte algae (sharing common ancestry with land plants) suggested that the origin of TAPs considered to be specific for land plants needs to be revised (Hori et al. 2014; Delaux et al. 2015; Wang et al. 2015), which we set out to do here by including more data of streptophyte algae and bryophytes than previously available.

Transcription associated proteins, and in particular TFs, are important signaling components and as such often key regulators of developmental progressions. They evolve via duplication, paralog retention and subsequent sub- and neofunctionalization (Rensing 2014), leading to a high abundance and combinatorial complexity of these proteins in the

© The Author(s) 2017. 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

most complex multicellular lineages (that perform embryogenesis)—namely plants and animals (de Mendoza et al. 2013; Lang and Rensing 2015; Rensing 2016).

Many plant TFs have initially been described as regulators of organ development or stress responses of flowering plants. However, by broadening the view to other plants it became clear that, for example LFY, initially described in *Arabidopsis thaliana* as determining the floral fate of meristems and regulating flower patterning, controls the first division of the zygote in the moss *Physcomitrella patens* (Tanahashi et al. 2005). Also, the flowering plant meristem controlling WOX genes have orthologs in moss that are involved in apical stem cell formation (Sakakibara et al. 2014). Such homeodomain (HD) TFs have deep eukaryotic roots and control important developmental progressions, for example, in embryogenesis, in plants, and animals (Hudry et al. 2014; Catarino et al. 2016). The KNOX and BELL subfamilies of plant HD proteins control mating types of green algae (Lee et al. 2008) and evolved into controlling cell fate determination of flowering plant stem cells (Hay and Tsiantis 2010). TF gene regulatory network kernels that were present in the earliest land plants are often modified and coopted during evolution (Pires et al. 2013), and plant TF paralogs are preferentially retained after whole genome duplication (WGD) events (De Bodt et al. 2005; Lang et al. 2010). TRs do not show the same tendency as TFs to expand with complexity, but they are important regulators nevertheless. For example, epigenetic control of important developmental steps like body plan control is maintained via components of the Polycomb Group (PcG) proteins throughout land plants (Mosquna et al. 2009; Okano et al. 2009; Bouyer et al. 2011).

Transcription associated proteins are thus key to understanding development and evolution of plant form and function. Access to reliable, up-to-date classification of TAPs is important, and enables comparative analyses informing our knowledge of plant transcriptional regulation. In a previous study (Lang et al. 2010) we combined rule sets of three studies (Riano-Pachon et al. 2007; Richardt et al. 2007; Guo et al. 2008) to generate the comprehensive TAPscan tool, encompassing sensitive domain-based classification rules for 111 TAP families. Similar approaches were undertaken by other studies, for example, PlnTFDB (Perez-Rodriguez et al. 2010), iTAK (Zheng et al. 2016), or PlantTFDB (Jin et al. 2016). We have now expanded our methodology by switching to HMMER v3, by updating the Hidden Markov Models (HMM) of many of the domains, and by including novel subfamily classification for several families. Moreover, we have included 92 more genomes than were available 7 years ago, dramatically improving taxon sampling. Here, we present an updated comprehensive analysis of TAP evolution of the green lineage as well as the TAPscan v2 web interface (http://plantco.de/tapscan/), including precomputed gene trees. This interface is a successor to PlnTFDB v3.0 (Perez-Rodriguez et al. 2010), encompasses the most comprehensive set of plant TAPs, and represents a novel tool for the plant community to access, screen and download genome-wide TAP annotations.

Materials and Methods

Data Set

In our previous analysis (Lang et al. 2010) no streptophyte algae, no gymnosperms and only a single bryophyte genome were covered. Here, we collected a set of 110 genomes and 13 transcriptomes with the purpose of covering as many major clades as possible within the Viridiplantae (green lineage, table 1 and supplementary table S5, Supplementary Material online), and to close the previous taxonomic holes.

Upgrade to HMMER3 and New PFAM Profiles

The extensive update of HMMER from v2 to v3 included improvements in both sensitivity and run time. With this new version, HMMER abandoned its glocal (global/local) approach, the alignment of a complete model to a subsection of a protein, to exclusively use local alignments. This change made it possible to make use of how much of the respective HMM profile was matched per alignment. This information was implemented in our TAPscan pipeline as a dynamic coverage cutoff aimed to introduce a higher level of strictness to maintain sequence and functional conservation. For our custom-built profiles we set this cutoff to 75% based on manual inspection of the alignments (cf. Results). For the PFAM profiles we calculated the proportion of 100% conservation in each profiles' seed alignment and used this as minimum coverage cutoff (listed in supplementary table S3, Supplementary Material online). Out of the 124 HMM profiles published in 2010 (Lang et al. 2010), 108 had been obtained from the PFAM database (PFAM 23.0) and were again downloaded directly from the PFAM database (PFAM 29.0).

Updating the Custom-Built HMM Profiles

The 16 domains represented by custom-made profiles had to be updated separately. They were first checked against the PFAM database to see if any equivalent profiles could be found, which was true only for NAC/NAM (supplementary table S1, Supplementary Material online). To increase the sequence diversity underlying the HMM profiles we decided to not directly reuse the profile multiple sequence alignments published earlier (Lang et al. 2010), but instead to use the output of these profiles when run against a database of 46 genomes representing 12 diverse groups of organisms (supplementary table S2, Supplementary Material online; 2x animals, 1x bryophyte, 8x chlorophytes, 1× conifer, 9× dicots, 1× lycophyte, 6× fungi, 1× glaucophyte, 4× monocots, 1× charophyte, 7× protoctista [5× nongreen algae, 1× Mycetozoa, 1× Heterolobosea], and 5× rhodophytes).

Table 1
Included Species

Taxonomic Group	Lang et al. (2010) V1	V2 2017		Unpublished Genomes	Unpublished Transcriptomes
	Genomes	Genomes	Transcriptomes		
Angiosperm—Core Eudicots/Core Rosids	7	46	0	1 (*Salix purpurea*)	
Angiosperm—Core Eudicots/Asterids	0	11	0		
Angiosperm—Core Eudicots/Stem Rosids	0	4	0	1 (*Kalanchoe laxiflora*)	
Angiosperm—Stem Eudicots	0	2	0	1 (*Aquilegia coerulea*)	
Angiosperm—Monocots	3	22	0	3 (*Brachypodium stacei*, *Panicum virgatum*, *Setaria viridis*)	
Angiosperm—ANA grade (stem angiosperms)	0	1	0		
Sub total angiosperms	**10**	**86**	**0**		
Gymnosperm—Conifer	0	2	1		
Gymnosperm—Ginkgophyte	0	1	0		
Monilophytes—Leptosporangiate	0	2	2	2 (*Azolla filiculoides* and *Salvinia cucullata*)	1 (*Microlepia* cf. *marginata*)
Lycophytes	1	1	0		
Mosses	1	2	2	1 (*Sphagnum fallax*)	
Liverworts	0	1	1		
Sub total nonseed plants and gymnosperms	**2**	**9**	**6**		
Streptophytic Algae—Zygnematales	0	1	2		
Streptophytic Algae—Coleochaetales	0	0	3		
Streptophytic Algae—Charales	0	1	1	1 (*Chara braunii*)	
Streptophytic Algae—Klebsormidiales	0	1	0		
Streptophytic Algae—Chlorokybales	0	0	1		
Green Algae—Chlorophyta	7	13	0	1 (*Dunaliella salina*)	
Sub total algae	**7**	**16**	**7**		
Total	**19**	**111**	**13**		
		124			

NOTE.—Species are divided into angiosperms, nonseed plants and algae, with sub totals and totals in bold. The data used in TAPscan v1 (Lang et al. 2010) is compared with the present v2, divided into genomes and transcriptomes. Unpublished genomes and transcriptomes, which will be made available via the web interface upon publication, are listed.

To avoid sequences not encompassing the major part of the domain of interest, hit length and model usage had to be at least 75% of the model length, as mentioned above. For each of the 12 clades four sequences were sampled (if possible), before random sampling collected the remaining sequences to reach 50 sequences. If 50 sequences could not be sampled, due to too few hits in the 2010 (v1) output, all hits were used for building a new model. To measure the variability in the phylogenetically guided sampling approach it was repeated nine times. The detected domains from the chosen sampling run were then aligned using clustalw-2.1 (Larkin et al. 2007) and a new hmm3 model was built using hmmbuild. The new models were run against the same set of 46 genomes and the output scores were plotted (fig. 1*a*) and compared with the 2010 profile's findings (green in fig. 1*a*). To remain conservative, the sampling run that generated the profile that had the least amount of previously undetected sequences scoring higher than previously detected (diamond shaped in fig. 1*a*) was chosen for further processing. Defining the gathering cutoffs (ga_cut) of the profiles was done with the help of score-ordered multiple sequence alignments (fig. 1*b* and supplementary fig. S1, Supplementary Material online) visualized with Jalview v2.8.2 (Waterhouse et al. 2009). This made it possible to investigate each profiles' window of uncertainty with the aim to maintain physiochemical properties/conservation above the set ga_cut (cf. Results).

Updating Family Classification Rules

Using published detailed studies (see supplementary table S1, Supplementary Material online and Results for details) more subfamilies could be distinguished using both PFAM and novel custom profiles. By incorporating 9 new PFAM profiles and adding 5 new custom profiles, 11 additional TAP subfamilies could be added. This includes an expansion of the Homeodomain (HD) family from four to 12 subfamilies, an additional Jumonji subfamily, an additional Polycomb Group (PcG) subfamily, and being able to distinguish the MADS subclass MIKC. If no PFAM profile was available, custom profiles were made using existing multiple sequence alignments: BEL (Hamant and Pautot 2010; Sharma et al. 2014b), KNOX_C and PINTOX (Mukherjee et al. 2009) and WOX (van der Graaff et al.

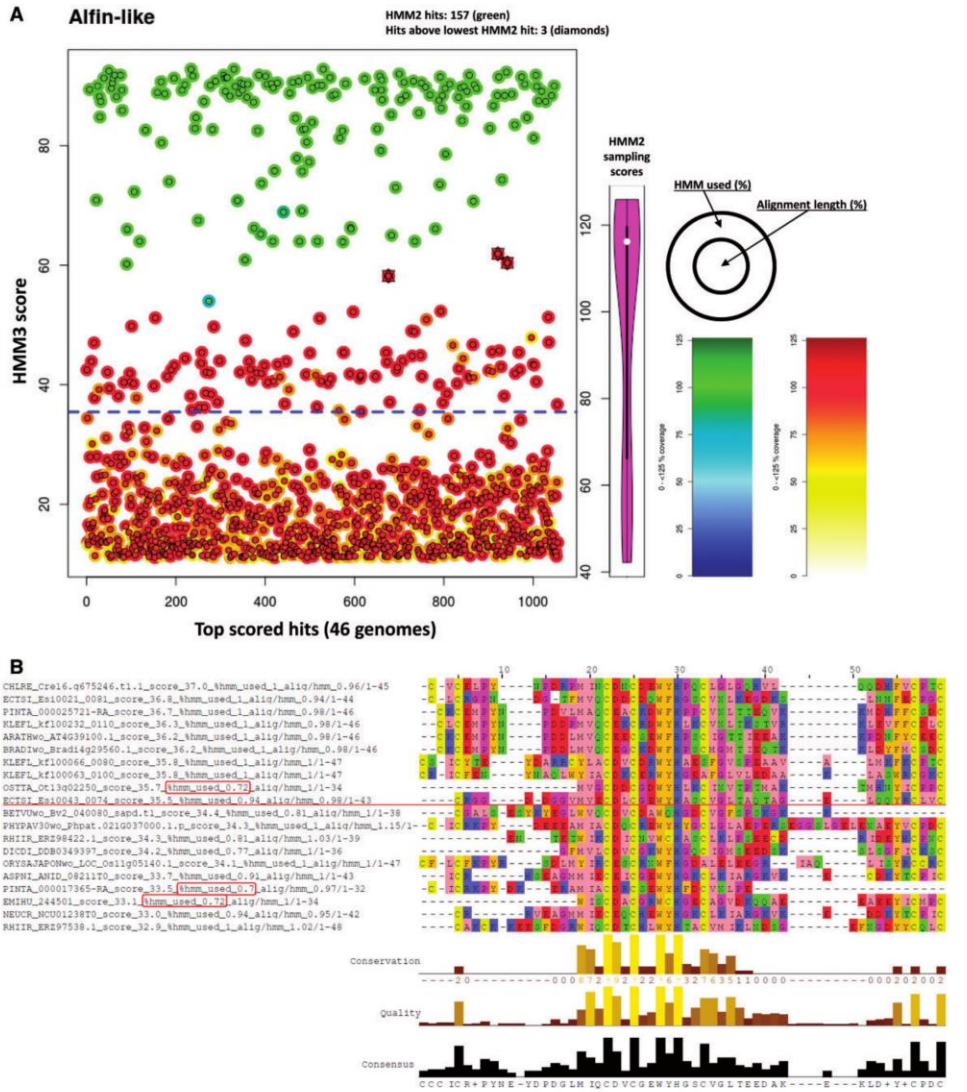


FIG. 1.—Determining gathering cutoffs for new custom profiles. (*A*) Plotted scores of the new profile (example: Alfin-like) run against 46 phylogenetically diverse genomes (supplementary table S2, Supplementary Material online). Sequences that were previously detected using the v1 profiles are colored in a green-blue gradient. New hits are colored in a red-yellow gradient. Each sequence's hit score is represented by an outer and inner area of the circle that represent the percentage hmm usage and alignment length, respectively. The dashed blue line represents the novel gathering cutoff, including sequences not previously captured (red circles above the line). The violin plot shows the old hmm2 score distribution of the sequences used to build the v1 model. If the new profile scored previously undetected sequences higher than previously detected sequences these are shown with diamond shapes. (*B*) A subsection of the sequence alignment of all hits (supplementary fig. S1, Supplementary Material online), highlighting where the gathering cutoff was set (red line, 34.5 in this example) based on manual inspection. The sequence names to the left of the alignment contain the five letter species code (supplementary table S2, Supplementary Material online) as well as the information of hmmsearch score and percentage of HMM used. Sequences later removed due to insufficient coverage (<75%) are marked with red boxes.

2009). When screening known PcG_EZ proteins (Pu and Sung 2015) the prosite CXC pattern (http://prosite.expasy.org/PS51633; last accessed December 8, 2017) was found and the underlying alignment used to build a custom model, replacing the SANTA domain (supplementary table S1, Supplementary Material online).

Inference of Ancestral States and Expansions/Contractions/Gains/Losses

We modified the ML phylogeny inferred by (Wickett et al. 2014) and placed our species into the clades included in their study. The tree was then pruned to only contain clades for which we had representative species (fig. 5). Our data included representatives of all major clades but hornworts, Magnoliids and Chloranthales, for which no appropriate data was available. This tree served as the basis for the inferences outlined below. Averages, fold changes between taxonomic groups and *q*-values (Mann–Whitney *U* test with Bonferroni correction for multiple testing) were calculated in Microsoft Excel (supplementary table S6, Supplementary Material online). Expansion/contractions and gains/losses were calculated with the count package (Csurös 2010). Their implementation of ancestral reconstruction by asymmetric Wagner parsimony was used to calculate expansions/contractions and their implementation of PGL (propensity for gene loss) was used to calculate gains/losses, both with default settings. All detected changes are shown in supplementary table S7, Supplementary Material online. The count predictions were entered into supplementary table S6, Supplementary Material online (tab Groups, column O-R) and manually reviewed; changes detected in (mainly) transcriptomic data/lineages with a low number of samples were disregarded, since they have a high chance of being due to incomplete data. Reviewed gains/losses/expansions/contractions were imposed onto the tree (fig. 5).

Phylogenetic Inference

The multiple sequence alignment of the DUF 632/PLZ family case study was calculated using muscle v3.8.31 (Edgar 2004) and visualized with Jalview v2.9.0b2. Sequences representing $<50\%$ of the alignment columns were removed and alignment columns with high entropy and low alignment quality as calculated by Jalview (Waterhouse et al. 2009) were manually clipped before Bayesian inference (BI) with MrBayes v3.2.5 x64 (Ronquist et al. 2012). The appropriate prior model was selected based on AIC/BIC using Prottest v3.4.2 (Darriba et al. 2011) and turned out to be JTT + G+F. BI was run with two hot and two cold chains until the standard deviation of split frequencies dropped < 0.01 at 756,600 generations, 200 trees were discarded as burn-in. The tree was visualized using FigTree v1.4.3pre (http://tree.bio.ed.ac.uk/software/figtree/; last accessed December 8, 2017).

For the gene trees shown in the TAPscan interface, we used several alignment tools as follows. Phylogenetic trees were generated for all TAPs appearing in more than one species of Archeaplastida. The protein sequences were downloaded using the TAPscan web interface and alignments were generated using MAFFT v7.310 (Katoh and Standley 2013). Alignments containing up to 500 input sequences were generated using MAFFT-linsi and MAFFT-fftnsi, whereas bigger alignments were generated only by MAFFT-fftnsi. The alignments were trimmed using two trimAl (Capella-Gutierrez et al. 2009) runs, one for trimming the alignments using the "-automated1" parameter and one for removing fragmentary sequences ("-resoverlap 0.75 -seqoverlap 50"). The trimmed mafft alignment was selected for inference if it was at least 100 columns long. If both linsi and fftnsi alignment were present and featured >100 columns, the longer one was selected.

If no suitable alignment could be generated, muscle v3.8.31 was run with two iterations and trimal applied. If that did not lead to a suitable trimmed alignment, ProbCons v1.12 (Do et al. 2005) was applied for alignments of up to 2, 100 input sequences. If that failed as well, muscle was applied with 16 iterations. In cases where trimAl produced empty/too short alignments, the automated trimming step was omitted. If all trimmed alignments were too short, the shortest untrimmed alignment was selected.

Alignments were formatted to Stockholm format using sreformat from the HMMer package. For neighbor-joining (NJ) tree inference, quicktree-SD (Frickenhaus and Beszteri 2008) was used applying using 100 bootstrap iterations. We used NJ inference due to the large to very large size of most of the gene families; in future trees generated with other methods of inference will be added. For visualization, the trees were formatted from Newick format to PhyloXML using the phyloxml (Han and Zmasek 2009) converter provided by the forester package (https://sites.google.com/site/cmzmasek/home/software/forester; last accessed December 8, 2017). The trees are presented on the TAPscan webpage using Archaeopteryx.js (https://sites.google.com/site/cmzmasek/home/software/archaeopteryx-js; last accessed December 8, 2017).

Visualization of Family Profiles and Column Charts

Using the R environment (R_Core_Team 2016) the family per species data (supplementary table S6, Supplementary Material online) was first log2 transformed and then hierarchically clustered on the *x* axis using complete linkage with euclidean distances to generate TAP clusters, and visualized as a heatmap using R gplots v3.0.1 (https://cran.r-project.org/web/packages/gplots/index.html; last accessed December 8, 2017). The *y* axis was ordered to follow our adaption of the (Wickett et al. 2014) phylogeny (fig. 5). The family per species data was also used to create stacked column charts

(supplementary figs. S3 and S4, Supplementary Material online). Each TAP value was log2 transformed and then grouped by either TAP-class (supplementary fig. S3, Supplementary Material online) or amount of multiple domain TAPs (supplementary fig. S4, Supplementary Material online), maintaining the species separation.

Implementation of the TAPscan Online Resource

The web page was setup using a LAMP architecture (Lawton 2005) implemented with Linux Debian 9.1, Apache 2.4.25 (Debian), MariaDB 10.1.26 and PHP 7.0.19-1. PHP additionally uses HTML5, CSS3, Javascript and jQuery v3.1.1 for dynamic web page creation. The data used for the web page is saved as 18 tables which are normalized to avoid redundancy of the data. For example, there are five tables storing taxonomy information for the species table and two tables storing the domain rules for the domain and TAP family table. Access to the database is provided using PHP which also generates the HTML code sent to the user. The databases' entity relationship model is visualized in supplementary figure S7, Supplementary Material online. The gene trees and the underlying alignments (see above) were made available on the TAP family view pages for viewing and download.

Results and Discussion

Availability of accurate and state-of-the-art genome-wide TAP annotation is considered to be of high value, in particular for the plant science community. TAPscan v2 presents a framework for comparative studies of TAP function and evolution. The availability of new software tools, protein domain circumscriptions, and plant genomes triggered the updating of our previous rule sets and resources, and allowed to draw novel important conclusions on plant TAP evolution.

TAPscan v2 Uses More and Better Profiles

TAPscan relies on HMMs to detect domains. We updated our approach from using HMMER2 to its accelerated successor HMMER3 (http://hmmer.org/; last accessed December 8, 2017), making use of the novel local alignment of HMMs to define better coverage cutoffs. Moreover, we updated all used PFAM (Finn et al. 2016) profiles from version 23.0 to 29.0 and included nine new PFAM profiles (supplementary table S1, Supplementary Material online, columns "Additional Profiles" in the rule change tabs). Eight of those were added due to our novel diversified classification rules, and one previous custom profile, NAC_plant, was replaced with the now available PFAM profile NAM (supplementary table S1, Supplementary Material online). Among the updated PFAM HMMs, seven were renamed and two merged into other existing domain models. Out of nine name changes that occurred due to the PFAM updates, five affected domains of (previously) unknown function (supplementary table S1, Supplementary Material online, tab "name change").

We also added/exchanged five new custom-built profiles (BEL, KNOXC, PINTOX, WOX_HD, and CXC; cf. Methods) due to our expanded classification rules (supplementary table S1, Supplementary Material online, rule change tabs HD and PcG). All custom HMMs were updated using a phylogenetic sampling approach. For that, previously used HMMs (Lang et al. 2010) were run against a database of 46 genomes with broad phylogenetic sampling (supplementary table S2, Supplementary Material online). Using the 2010 (v1) profiles, hit sequences were sampled from each of the 12 groups that the 46 genomes represent, and then used to rebuild each custom HMM. The resulting HMMs were run against the same set of 46 genomes, and the outputs were compared to determine how previously undetected sequences scored now (fig. 1a). By manual inspection of all aligned hit sequences we defined the individual score cutoffs to lie above sequences of uncertain functional conservation (fig. 1*b* and supplementary fig. S1, Supplementary Material online). In order to represent a functionally relevant hit, the major part of the HMM should be detected. Based on manual inspection of all custom profile alignments we decided to employ a global cutoff of 75% HMM used (fig. 1*b* and supplementary fig. S1, Supplementary Material online).

Improved Taxon Sampling, Subfamily Definition, and Specificity

In the past 7 years, a multitude of plant and algal genome sequences became available, allowing for a much better taxon sampling. There are now 82 more plant genomes included in TAPscan v2, and nine more algal genomes, bringing the total up to 110 (table 1). To improve taxonomic resolution we also included a selection of 13 transcriptomes, reaching a final set of 123 species. We have also included 11 genomes and 1 transcriptomes that are not yet published. Data for those will be quickly made available via the web interface as soon as they are publicly available. For example, PlantTFDB v4 (Jin et al. 2016) includes more angiosperm genomes, we took care to include as much as possible nonseed plants and streptophyte algae, to be able to take a close look at the early evolution of plant TAPs. In addition to the Viridiplantae that are the focus of this study, we have included Rhodophyta and the glaucophyte alga *Cyanophora paradoxa* as outgroup representatives within the Archaeplastida (supplementary table S5/S6, Supplementary Material online).

To update our classification rules (supplementary table S3, Supplementary Material online), we screened the literature for novel (sub) classifications of TAPs and checked them for applicability to our domain-based classification scheme. In total, 11 new subfamily classification rules were established, and some families renamed due to changes in domain or family names (fig. 2). In particular, we subdivided homeodomain

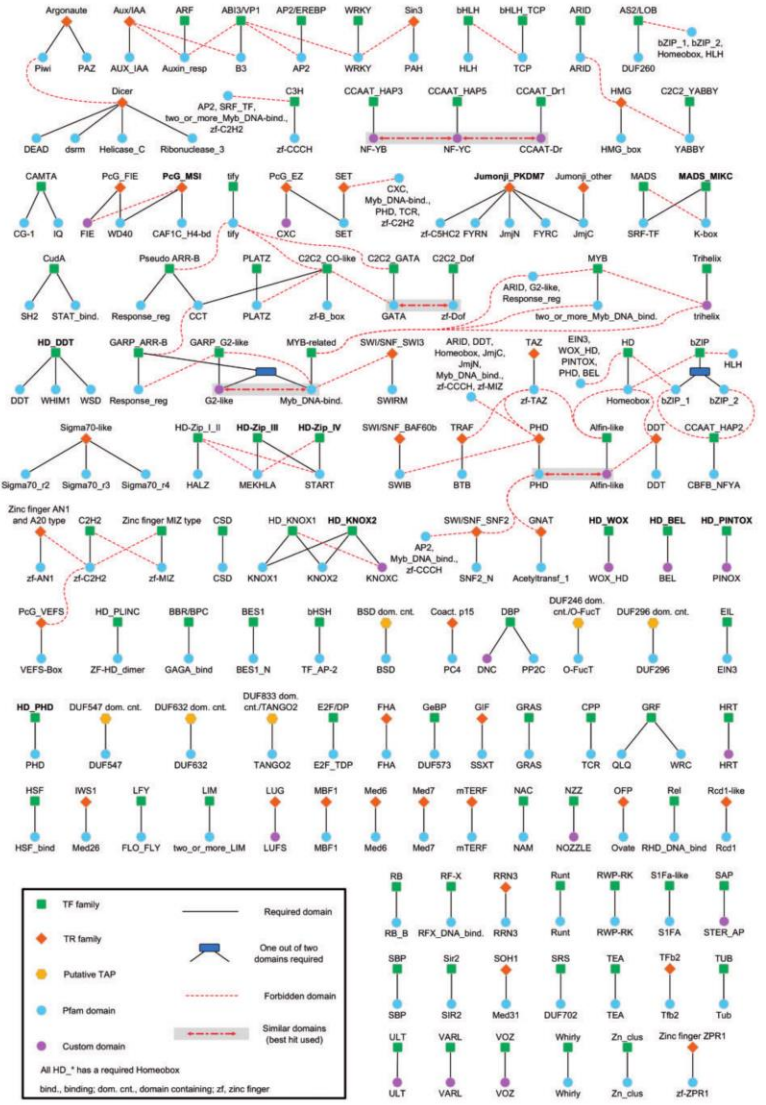


Fig. 2.—TAPscan classification rules. The name of each family or subfamily is shown on top of each classification rule set; novel (in v2) rule sets are shown in bold face. TF (green), TR (orange), and PT (yellow) are marked by different symbols and in the same color code that is used throughout the manuscript. Required ("should") domains (represented by corresponding HMMs) are connected to the family symbol by lines; forbidden ("should not") domains are connected via dotted red lines. Similar domains that are selected via the best hit are shown with red dotted double arrows on grey background, if one out of two domains are required this is denoted as a blue box with two required lines. Custom domains are depicted as purple circles, PFAM domains as blue circles. For brevity, the homeobox should rule for all HD_families was omitted. Compare supplementary table S3, Supplementary Material online for more detailed classification rules.

(HD) TFs into DDT, PHD, PINTOX, PLINC, WOX, HD-ZIP I/II, III, IV, and into the TALE class subfamilies BEL, KNOX 1, and 2 (Mukherjee et al. 2009; van der Graaff et al. 2009; Hamant and Pautot 2010; Sharma et al. 2014b) (supplementary table S1, Supplementary Material online, 1st sheet). Also, MADS-box TFs were divided into general and MIKC-type (Gramzow and Theissen 2010), Jumonji into PKDM7 and other (Qian et al. 2015), and the Polycomb Group (PcG) TR MSI was added (supplementary table S1, Supplementary Material online). Similar to (Zheng et al. 2016) we reclassified mTERF, Sigma70-like, FHA and TAZ as TR instead of TF; TAPs containing the DDT domain are subdivided into the TR DDT and the TF HD_DDT in TAPscan v2. With a total of 124 families and subfamilies (supplementary table S3, Supplementary Material online, fig. 2; 81 of them TFs) our rule set is the most comprehensive one for plant TAPs, since other approaches have significantly less resolution, for example, 58 in PlantTFDB 4.0 (Jin et al. 2016) and 72 in iTAK (Zheng et al. 2016).

We compared the TAPscan v1 and v2 annotations with a number of *A. thaliana* and *P. patens* phylogeny-based family classifications defined as gold standard (Mosquna et al. 2009; Mukherjee et al. 2009; Paponov et al. 2009; Martin-Trillo and Cubas 2010). We find that the average sensitivity of TAPscan v2 (87.76%) is only slightly lower than of v1 (89.31%), whereas the specificity of v2 (100.00%) is much higher than in the old version (92.31%; supplementary table S4, Supplementary Material online). The combined sensitivity and specificity of the new version is therefore 6.1% improved. It should be noted that the comparatively low sensitivity for some of the HD sub classes is balanced by the fact that all HD family members are detected as such, yet in cases where domain scores are below cutoffs are sometimes binned into HD_other. The weighted sensitivity, taking into account gene family sizes, is strongly improved to 87.03% as compared with 78.27% in (Lang et al. 2010).

The TAPscan Online Resource

In order to make the domain-based classification available to the scientific community in an easy to use way, we implemented a web-based resource that allows a user to browse the data either in a species-centric or a TAP family-centric view (http://plantco.de/tapscan/). The interface (fig. 3) includes taxonomic information as expandable trees and an intuitive click-system for selection of sequences of interest that can subsequently be downloaded in annotated FASTA format. TAPscan FASTA headers contain the species, TAP family information and domain positions. It is possible to either download all proteins of a custom set of species containing a specific TAP, or to download all proteins for a specific family and species. The latter makes it possible to download isoforms, if available.

The TAP overview pages show the domain rules a protein has to meet in order to be classified as belonging to that family. Domain names are linked to PFAM entries or custom domain alignments and HMM profiles. Locations of domains within the sequence are shown in sequence view. Precomputed phylogenies (gene trees) are available for viewing and download on the overview pages. These trees are intended as a first glimpse, allowing users to quickly access gene relationships without having to infer a tree on their own.

In the case of not yet published sequence data (table 1) a disclaimer is shown, mentioning that the data will be made available immediately upon publication. Such unpublished information is excluded from species or protein counts in the web interface. By including these data into the interface we are able to quickly release TAP annotation for these genomes as soon as the data become public.

Taxonomic Profiling of TAPs

Heatmap representation of the data shows that TAP family size generally increased during land plant evolution (fig. 4 and see supplementary fig. S2, Supplementary Material online for expanded version). Cluster 5 contains families (such as bZIP, bHLH, or MYB) that were already abundant in the algal relatives of land plants, whereas cluster 3 contains TAPs that expanded in land plants and again in seed plants, such as NAC or ABI3/VP1. The intervening cluster 4 contains families that show high abundance throughout, like HD or RWP-RK. The biggest cluster (1) contains families that show either only gradual expansion from algae (bottom of figure) to flowering plants (top of figure), or no expansion at all. Consequently, cluster 1 contains many TRs, which have previously shown not to be subject to as much expansion as TFs (Lang et al. 2010). The small cluster 2 next to cluster 1 harbors families of spurious presence, like those that evolved in vascular plants (Tracheophyta; like NZZ or ULT). In general, the heatmap visualizes a principal gain of (primarily) TF paralogs within existing families concomitant with the terrestrialization of plants (cf. supplementary fig. S3, Supplementary Material online). Interestingly, the propensity of TAPs to comprise of more than one functional domain increases in a very similar pattern (supplementary fig. S4, Supplementary Material online), akin to the domain combination tendency generally seen for plants (Kersting et al. 2012). Hence, the combinatorial potential of TFs clearly coincides with increasing morphological complexity (as measured by number of cell types), corroborating earlier results (Lang et al. 2010; Lang and Rensing 2015).

TAP Family Evolution

The taxonomic sampling of our data is visualized as a cartoon tree (fig. 5) derived from a recent phylogenomics study (Wickett et al. 2014). We plotted the gains, losses, expansions and contractions of TAP families onto this tree to enable a global view of plant TAP evolution (cf. supplementary table

Fig. 3.—TAPscan web interface main features. Upper left: Family-centric view—table of TAP families covered by TAPscan; the number of proteins per family is given in brackets. TAPs are colored according to their TAP class (TF, TR, and PT). Upper right: Species-centric view—part of the species tree; different levels can be expanded and collapsed. Numbers of published species per taxonomy level are given in brackets. Only species with published protein data can be accessed. Bottom left: Species view for TAP family bZIP in *Ceratodon purpureus*. The species' lineage, the bZIP domain rules, and the protein sequences are shown. One protein is marked for downloading. Bottom right: Species tree for the bZIP family with expanded SAR kingdom. Species belonging to Alveolata are marked for downloading; the resulting file will contain 54 proteins. TAP distribution is given in a table-like manner, with a dark green background: minimum, maximum, average, median, and standard deviation of proteins per species for the selected taxonomy level.

S6/S7, Supplementary Material online). 32 losses were predicted that are scattered along the tree. The streptophyte alga *Klebsormidium nitens* apparently secondarily lost five TAP families, whereas the lycophyte *Selaginella moellendorffii* lost eight. Another eight families were lost during gymnosperm evolution, one of them (HD_Pintox) being absent from all studied gymnosperms, whereas two are lacking in conifers and five in *Ginkgo* (e.g., LFY—although a lacking gene model would be an alternative explanation). A total of 76 expansions were detected, of which the highest number (17.22%) are inferred to have occurred in the lineage that led to the last common ancestor of all land plants. All other expansions show a scattered distribution along the deep as well as distal nodes of the tree (fig. 5). The 13 inferred family contractions also display a patchy pattern. Strikingly, out of 36 TAP family gains 26 are predicted to have occurred in streptophyte algae (nodes 34-30). Another five are synapomorphic of land plants (Embryophyta), whereas only 2, 1, and 1 are evolutionary novelties of vascular plants, Euphyllophyta, and Eudicots, respectively.

Many TAP Families Were Gained in the Water

Previously, due to limited taxon sampling, many plant-specific TAPs were inferred to have been gained at the time of the water-to-land-transition of plant life (Lang et al. 2010). Streptophyte algae are sister to land plants and thus ideally suited, together with bryophyte sequences, to elucidate whether gains occurred prior or after terrestrialization. Although only two genomes of strepptophyte algae have yet been published (Hori et al. 2014; Delaux et al. 2015), there are transcriptome data available for seven species (Timme et al. 2012) that were included into TAPscan (table 1 and supplementary table S5/S6, Supplementary Material online). Similarly, although no other bryophyte genomes than *P. patens* are published yet, we included transcriptomes of the mosses *Ceratodon purpureus* (Szovenyi et al. 2015) and *Funaria hygrometrica* (Szovenyi et al. 2011), and of the liverwort *Marchantia polymoprha* (Sharma et al. 2014a). Out of 20 TAP families previously thought to have been gained with terrestrialization (Lang et al. 2010), only VOZ and bHLH_TCP

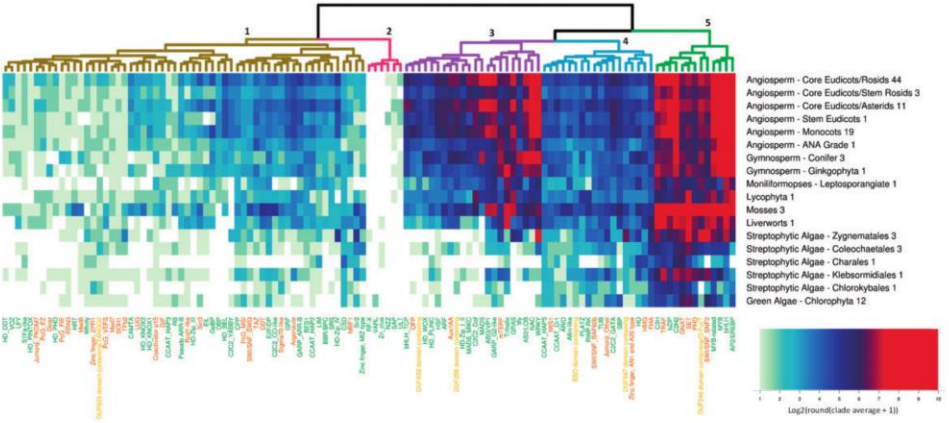


Fig. 4.—TAPfamily abundance heat map. Heatmap using log2 transformed average values of TAP abundance for each clade. The data was clustered on the *x* axis using complete linkage with euclidean distances. The *y* axis was kept to match the phylogeny as in Wickett et al. (2014), cf. figure 5. The logarithmic color scheme comprises white (absent) through blue to red (high abundance).

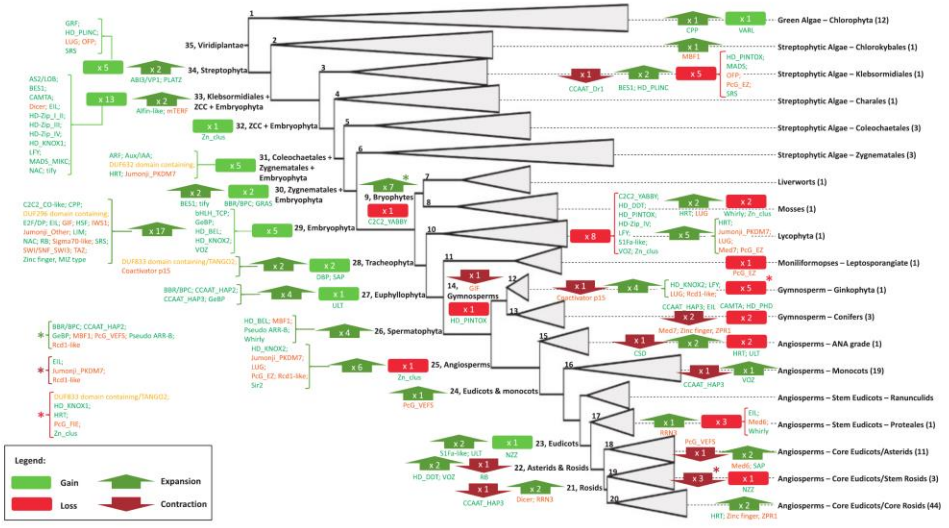


Fig. 5.—Cartoon tree illustrating the predicted ancestral states, expansion/contractions, and gains/losses of plant TAPs. The tree was modified from Wickett et al. (2014); number of data sets covered per clade is shown in brackets. Gains and losses were predicted using PGL, expansions and contractions using Wagner parsimony (cf. Methods and supplementary table S7, Supplementary Material online). These predictions were entered into supplementary table S6, Supplementary Material online (tab Groups, column O-R) and manually reviewed; changes detected in (mainly) transcriptomic data/lineages with a low number of samples were disregarded, since they have a high chance of being due to incomplete data. Reviewed gains/losses/expansions/contractions of TFs (green text), TRs (orange text), and PTs (yellow text) were imposed onto the tree: gains are shown as green boxes, losses as red boxes. Expansions are shown as green upward arrows, contractions as red downward arrows. Node numbers and names are as in supplementary table S7, Supplementary Material online; symbols are shown to the right of triangles if they concern a distal node, and to the left if they concern a deep node.

could be confirmed. Of the others, three (ARF, S1Fa-like, and O-FucT) are already present in Rhodophyta, Chlorophyta, or both. Strikingly, the vast majority of these 20 families (15) are present in Charophyta (comprising all lineages of streptophyte algae), but not Chlorophyta or Rhodophyta (supplementary table S6, Supplementary Material online). Hence, they were most probably gained during the evolution of the Streptophyta (uniting the Charophyta with the land plants). Out of these 15 families, 11 are already present in the KCM grade (encompassing Klebsormidiales, Chlorokybales, and Mesostigmatales and sister to the ZCC grade and land plants), whereas 4 (Aux/IAA, DUF632 domain containing, GRAS and HRT) are present only in the ZCC grade (encompassing Zygnematales, Coleochaetales, and Charales, together with the land plants comprising the Phragmoplastophyta). This finding is in line with the emerging evidence that in particular ZCC species share many unique features with land plants like polyplastidy (de Vries et al. 2016) or the phragmoplast (Pickett-Heaps et al. 1999; Buschmann and Zachgo 2016), and that Klebsormidium also possesses some "plant-like" features, like callose and the phenylpropanoid pathway (Herburger and Holzinger 2015; de Vries et al. 2017). Based on our findings, the last common ancestor of streptophytes had already evolved 11 TAP families previously thought to be land plant-specific, and the last common ancestor of Phragmoplastophyta (ZCC grade algae and land plants) another five families. Prominent examples of these families are the TF families LFY and NAC (present already in *K. nitens*), as well as GRAS and Aux/IAA (present in the ZCC grade). Most of what we know about function of these TF families stems from research in flowering plants, and many of them control development of organs unique to flowering plants. It will therefore be intriguing to determine the putative ancestral function of these genes in the last common ancestor of streptophytes. As an example, a recent study showed that a *P. patens* TCP TF is involved in suppressing branching of the moss sporophyte (which is determinate since it does not branch) (Ortiz-Ramirez et al. 2016).

Origin and Expansion Revisited

Several of the gains previously inferred to have occurred in vascular plants, angiosperms or eudicotelydons can now be dated back to the common ancestors with streptophyte algae, bryophytes, ferns or lycophytes (fig. 5 and supplementary table S6/S7, Supplementary Material online). Together with the families mentioned in the last paragraph, a total of 35 TAP families (most of them TFs) evolved at some point in the Archaeplastida, before the evolution of angiosperms, shifting the inferred gain dates back in time. Yet, out of 44 TAP families previously inferred to be expanded in land plants as compared with algae (Lang et al. 2010), 21 show a >2-fold increase in the data presented here, and all 44 significantly more members ($q < 0.05$, Mann–Whitney) in land plants than in algae (supplementary table S6, Supplementary Material online). These data suggest a primary burst of gain and expansion of TAPs concomitant with the origin of Streptophyta. The total numbers of TAPs, and in particular TFs, show a clear increase in the common ancestor of land plants, but also in some streptophyte algae (supplementary fig. S3, Supplementary Material online). We expect that with more genomes of streptophyte algae becoming available the gain and expansion of even more families will be inferred to have occurred at earlier time points.

Of 22 families previously inferred to have been expanded in angiosperms (Lang et al. 2010), the present data support 17 with a 2-fold change and 15 based on statistical testing (overlap 13; $q < 0.05$; supplementary table S6, Supplementary Material online). Six TAP families expanded at the basis of angiosperms (among them HD_KNOX2), and several families expanded subsequently (fig. 5). The subfunctionalization of such TAPs might be related to the more complex reproductive system of angiosperms. While most TF were already present in the earliest land plants, DBP and SAP appear first in vascular plants, ULT in the common ancestor of ferns and seed plants, and NZZ is unique to eudicots.

One of the major gaps in the previous sampling, besides the streptophyte algae, were gymnosperms. We have now included three conifers and *Ginkgo biloba*. If we consider the inferred expansions based on the tree (fig. 5), a total of 13 expansions occur between the land plant node (29) and the angiosperms (25). Four TF families (BBR/BPC, CCAAT_HAP2, CCAAT_HAP3, and GeBP) were apparently expanded in the Euphyllophyta (ferns and seed plants, node 27), another three (HD_BEL, Pseudo ARR-B, and Whirly) in the seed plants. All these TF families are thus presumably important for spermatophyte evolution and development.

In a recent study (Catarino et al. 2016), the authors had analyzed 48 plant TF families based on PlantTFDB classification rules (Jin et al. 2014) in 15 species. In general, their inference of TF family gain is consistent with our data: of 38 families that can be compared, 30 are placed at the same node. For the remaining eight, our study places six at earlier nodes of the tree, probably due to better taxon sampling. The study also did a subfamily analysis of HD TFs and concluded that almost all were already present in algae. In our study, we find that of 12 HD subfamilies all but two (HD_BEL and HD_KNOX2) are detected in algae. We also compared gain of 40 TF families from (Jin et al. 2016) with our data and can confirm their findings for 27 families. Out of the remaining 13, we detect 10 at earlier nodes in the tree, 4 of them in ZCC grade streptophyte algae instead of bryophytes, suggesting again that due to better sampling we infer family evolution more accurately. The *M. polymorpha* genome was published (Bowman et al. 2017) during the time this manuscript was under review. We have hence activated the previously computed data in the web interface and have added

corresponding columns to supplementary table S6, Supplementary Material online; the comparison of the transcriptomic and genomic data does not show any severe differences. The genome publication included an analysis of TFs that we compared with our data (supplementary table S6, Supplementary Material online). We detect 400 TFs, Bowman et al. 387 TFs; 33 out of 40 families are consistent; in the remaining seven cases the node of predicted origin varies due to different sampling.

Employing TAPscan Data

As an example on how the data presented with this study can be used, we selected the putative TAP family "DUF 632 domain containing." This domain of unknown function (http://pfam.xfam.org/family/PF04782; last accessed December 8, 2017) is described as representing a potential leucine zipper, which is why it was initially defined as a putative TAP, PT (Richardt et al. 2007). Our data show that this family first appears in the common ancestor of Coleochaetales, Zygnematales, and land plants (node 31) and is present throughout land plants (supplementary table S6, Supplementary Material online, and fig. 5). There are on average 19 family members in angiosperms, 7 in gymnosperms, 6 in bryophytes, and 3 in the streptophyte alga *Coleochaete orbicularis*. DUF 632 is part of cluster 3 (fig. 4) that shows expansion during land plant evolution. It is not detected to be expanded using Wagner parsimony (fig. 5), but shows significant size increase ($q < 0.05$, Mann–Whitney; supplementary table S6, Supplementary Material online) between nonseed plants and seed plants (fold change 2.95).

We selected protein sequences of this family using the TAPscan interface "family view" option, thus representing several angiosperm lineages as well as gymnosperms and nonseed plants. An alignment of the sequences (supplementary fig. S5, Supplementary Material online) shows several highly conserved blocks, all of which feature positively charged as well as regularly spaced Leucine residues, reinforcing the notion of a potential DNA-binding Leucine zipper. Given the proposed structure we suggest to call this family Plant Leucine Zipper (PLZ) TFs. Phylogenetic inference shows that all nonseed plant sequences are present in the same subclade (supplementary fig. S6, Supplementary Material online; the same can be derived from the tree automatically inferred and available via the TAPscan web interface), this subclade is sister to approximately half of the seed plant sequences. Based on the structure of the tree, duplication and paralog retention occurred several times during seed plant evolution. Most of the paralogs were already established in the lineage leading to the last common ancestor of seed plants, whereas some duplications occurred only in angiosperms.

In order to understand under which conditions members of this protein family are active, we conducted expression profiling using existing data for *P. patens* and *A. thaliana* (Hruz et al. 2008; Hiss et al. 2014, 2017, phytozome.org). Out of five *P. patens* genes detected by TAPscan, one appears to be a truncated pseudogene that was removed during alignment curation; another two genes are barely expressed. The remaining two genes (Pp3c16_15000V3.1 and Pp3c27_2840V3.1), however, show discrete expression profiles. The expression of both genes is higher under diurnal light and ammonia application. Pp3c16_15000V3.1 is more highly expressed upon heat stress, darkness and UV-B treatment, as well as in mature sporophytes and under biotic stimulus. Pp3c27_2840V3.1 is less expressed in gametophores (representing the late vegetative phase) as well as in mature sporophytes (i.e., adversely to the other gene). Similarly, ABA treatment leads to lower expression of Pp3c16_15000V3.1 and higher expression of Pp3c27_2840V3.1. The two *A. thaliana* genes most closely related to the nonseed plant clade, AT5G25590.1 and AT1G52320.2, show no particularly strong expression in any tissue or developmental stage, however, other members of the family show peaks in, for example, reproductive structures, xylem, or seed. AT1G52320.2 is induced, for example, under germination, drought and ABA, whereas AT5G25590.1 shows higher expression, for example, under UV-B, biotic stimulus, elevated CO_2 and drought. In summary, members of the streptophyte-specific PLZ family appear to be differentially regulated under a range of abiotic and biotic stimuli as well as in different development stages. Such an expression profile fits that of a TF family undergoing paralog retention followed by sub and neofunctionalization of expression domains (Birchler and Veitia 2010; Rensing 2014).

Outlook

Previous studies of land plant TAP evolution, like (Lang et al. 2010), suffered from severe sampling bias, leading to many gains and expansions being either associated with the water to land transition (because they were inferred to have occurred between green algae and the moss *P. patens*), or the angiosperm radiation (since they occurred between the lycophyte *S. moellendorffii* and angiosperms). Using better sampling, including streptophyte algae, more bryophytes, a fern and gymnosperms, we can now more accurately trace Viridiplantae TAP gains and expansions. Although we expect that we will have to again adjust our current understanding as more genomes become available, we can now say that much of what we considered to be specific for land plants or flowering plants already evolved in the water, in streptophyte algae, or in the course of preflowering land plant evolution.

The results of our improved genome-wide TAP annotation methodology, including annotated fasta files and gene trees, are now available online via an easy-to-use web interface. Species already sequenced but not yet published have already been included and will be made available immediately after publication. We trust that TAPscan v2 will be an important community resource for plant TAP analyses.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

We are grateful to Sven Gould, Günter Theißen, and two anonymous reviewers for providing helpful comments on the draft. P.W. was supported by the ERA-CAPS SeedAdapt consortium project (www.seedadapt.eu; last accessed December 8, 2017; Grant No. RE1697/8 to S.A.R.).

Author Contributions

S.A.R. conceived of the study, supervised it, wrote the paper and carried out evolutionary and phylogenetic analyses. K.K.U. was in charge of setting up the genomic data. P.K.I.W. adapted the TAPscan tool, carried out TAP classification and analyzed data. P.K.I.W. and K.K.U. implemented the phylogenetic sampling. C.M., K.K.U., and S.A.R. inferred gene trees. C.M. established the web interface. All authors contributed to writing the manuscript.

Literature Cited

Birchler JA, Veitia RA. 2010. The gene balance hypothesis: implications for gene regulation, quantitative traits and evolution. New Phytol. 186(1):54–62.

Bouyer D, et al. 2011. Polycomb repressive complex 2 controls the embryo-to-seedling phase transition. PLoS Genet. 7(3):e1002014.

Bowman JL, et al. 2017. Insights into land plant evolution garnered from the *Marchantia polymorpha* genome. Cell 171(2):287–304.

Buschmann H, Zachgo S. 2016. The evolution of cell division: from streptophyte algae to land plants. Trends Plant Sci. 21(10):872–883.

Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25(15):1972–1973.

Catarino B, Hetherington AJ, Emms DM, Kelly S, Dolan L. 2016. The stepwise increase in the number of transcription factor families in the precambrian predated the diversification of plants on land. Mol Biol Evol. 33(11):2815–2819.

Csurös M. 2010. Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. Bioinformatics 26(15):1910–1912.

Darriba D, Taboada GL, Doallo R, Posada D. 2011. ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics 27(8): 1164–1165.

De Bodt S, Maere S, Van de Peer Y. 2005. Genome duplication and the origin of angiosperms. Trends Ecol Evol. 20(11):591–597.

de Mendoza A, et al. 2013. Transcription factor evolution in eukaryotes and the assembly of the regulatory toolkit in multicellular lineages. Proc Natl Acad Sci U S A. 110(50):E4858–E4866.

de Mendoza A, Suga H, Permanyer J, Irimia M, Ruiz-Trillo I. 2015. Complex transcriptional regulation and independent evolution of fungal-like traits in a relative of animals. eLife 4. doi: 10.7554/eLife.08904

de Vries J, de Vries S, Slamovits CH, Rose LE, Archibald JM. 2017. How embryophytic is the biosynthesis of phenylpropanoids and their derivatives in streptophyte algae?. Plant Cell Physiol. 58(5):934–945.

de Vries J, Stanton A, Archibald JM, Gould SB. 2016. Streptophyte terrestrialization in light of plastid evolution. Trends Plant Sci. 21(6):467–476.

Delaux PM, et al. 2015. Algal ancestor of land plants was preadapted for symbiosis. Proc Natl Acad Sci U S A. 112(43):13390–13395.

Do CB, Mahabhashyam MS, Brudno M, Batzoglou S. 2005. ProbCons: probabilistic consistency-based multiple sequence alignment. Genome Res. 15(2):330–340.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32(5):1792–1797.

Finn RD, et al. 2016. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 44(D1):D279–D285.

Frickenhaus S, Beszteri B. 2008. Quicktree-SD, Software developed by AWI-Bioinformatics. Available from: http://hdl.handle.net/10013/epic.33164, last accessed December 8, 2017.

Gramzow L, Theissen G. 2010. A hitchhiker's guide to the MADS world of plants. Genome Biol. 11(6):214.

Guo AY, et al. 2008. PlantTFDB: a comprehensive plant transcription factor database. Nucleic Acids Res. 36(Database issue):D966–D969.

Hamant O, Pautot V. 2010. Plant development: a TALE story. C R Biol. 333(4):371–381.

Han MV, Zmasek CM. 2009. phyloXML: XML for evolutionary biology and comparative genomics. BMC Bioinformatics 10:356.

Hay A, Tsiantis M. 2010. KNOX genes: versatile regulators of plant development and diversity. Development 137(19):3153–3165.

Herburger K, Holzinger A. 2015. Localization and quantification of callose in the streptophyte green algae zygnema and klebsormidium: correlation with desiccation tolerance. Plant Cell Physiol. 56(11): 2259–2270.

Hiss M, et al. 2014. Large-scale gene expression profiling data for the model moss *Physcomitrella patens* aid understanding of developmental progression, culture and stress conditions. Plant J. 79(3):530–539.

Hiss M, et al. 2017. Sexual reproduction, sporophyte development and molecular variation in the model moss *Physcomitrella patens*: introducing the ecotype Reute. Plant J. doi: 10.1111/tpj.13501.

Hori K, et al. 2014. *Klebsormidium flaccidum* genome reveals primary factors for plant terrestrial adaptation. Nat Commun. 5:3978.

Hruz T, et al. 2008. Genevestigator v3: a reference expression database for the meta-analysis of transcriptomes. Adv Bioinformatics 2008: 420747.

Hudry B, et al. 2014. Molecular insights into the origin of the Hox-TALE patterning system. eLife 3:e01939.

Jin J, et al. 2016. PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. Nucleic Acids Res. doi: 10.1093/nar/gkw982.

Jin J, Zhang H, Kong L, Gao G, Luo J. 2014. PlantTFDB 3.0: a portal for the functional and evolutionary study of plant transcription factors. Nucleic Acids Res. 42(Database issue):D1182–D1187.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 30(4):772–780.

Kersting AR, Bornberg-Bauer E, Moore AD, Grath S. 2012. Dynamics and adaptive benefits of protein domain emergence and arrangements during plant genome evolution. Genome Biol Evol. 4(3):316–329.

Lang D, Rensing SA. 2015. The evolution of transcriptional regulation in the viridiplantae and its correlation with morphological complexity. In: Ruiz-Trillo I, Nedelcu AM, editors. Evolutionary transitions to multicellular life. Dordrecht: Springer Netherlands. pp. 301–333.

Lang D, et al. 2010. Genome-wide phylogenetic comparative analysis of plant transcriptional regulation: a timeline of loss, gain, expansion, and correlation with complexity. Genome Biol Evol. 2:488–503.

Larkin MA, et al. 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23(21):2947–2948.

Lawton G. 2005. LAMP lights enterprise development efforts. Computer 38(9):18–20.
Lee JH, Lin H, Joo S, Goodenough U. 2008. Early sexual origins of homeoprotein heterodimerization and evolution of the plant KNOX/BELL family. Cell 133(5):829–840.
Levine M, Tjian R. 2003. Transcription regulation and animal diversity. Nature 424(6945):147–151.
Martin-Trillo M, Cubas P. 2010. TCP genes: a family snapshot ten years later. Trends Plant Sci. 15(1):31–39.
Mosquna A, et al. 2009. Regulation of stem cell maintenance by the Polycomb protein FIE has been conserved during land plant evolution. Development 136(14):2433–2444.
Mukherjee K, Brocchieri L, Burglin TR. 2009. A comprehensive classification and evolutionary analysis of plant homeobox genes. Mol Biol Evol. 26(12):2775–2794.
Okano Y, et al. 2009. A polycomb repressive complex 2 gene regulates apogamy and gives evolutionary insights into early land plant evolution. Proc Natl Acad Sci U S A. 106(38):16321–16326.
Ortiz-Ramirez C, et al. 2016. A transcriptome atlas of *Physcomitrella patens* provides insights into the evolution and development of land plants. Mol Plant. 9(2):205–220.
Paponov IA, et al. 2009. The evolution of nuclear auxin signalling. BMC Evol Biol. 9:126.
Perez-Rodriguez P, et al. 2010. PlnTFDB: updated content and new features of the plant transcription factor database. Nucleic Acids Res. 38:D822–D827.
Pickett-Heaps JD, Gunning BE, Brown RC, Lemmon BE, Cleary AL. 1999. The cytoplast concept in dividing plant cells: cytoplasmic domains and the evolution of spatially organized cell. Am J Bot. 86(2):153–172.
Pires ND, et al. 2013. Recruitment and remodeling of an ancient gene regulatory network during land plant evolution. Proc Natl Acad Sci U S A. 110(23):9571–9576.
Pu L, Sung ZR. 2015. PcG and trxG in plants: friends or foes. Trends Genet. 31(5):252–262.
Qian S, Wang Y, Ma H, Zhang L. 2015. Expansion and functional divergence of jumonji C-containing histone demethylases: significance of duplications in ancestral angiosperms and vertebrates. Plant Physiol. 168(4):1321–1337.
R: A language and environment for statistical computing. [Internet]. R foundation for statistical computing, Vienna, Austria; 2016. Available from: https://www.r-project.org/, last accessed December 8, 2017.
Rensing SA. 2014. Gene duplication as a driver of plant morphogenetic evolution. Curr Opin Plant Biol. 17:43–48.
Rensing SA. 2016. (Why) does evolution favour embryogenesis?. Trends Plant Sci. 21(7):562–573.
Riano-Pachon DM, Ruzicic S, Dreyer I, Mueller-Roeber B. 2007. PlnTFDB: an integrative plant transcription factor database. BMC Bioinformatics 8:42.
Richardt S, Lang D, Frank W, Reski R, Rensing SA. 2007. PlanTAPDB: a phylogeny-based resource of plant transcription associated proteins. Plant Physiol. 143(4):1452–1466.
Ronquist F, et al. 2012. MrBayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space. Syst Biol. 61(3):539–542.
Sakakibara K, et al. 2014. WOX13-like genes are required for reprogramming of leaf and protoplast cells into stem cells in the moss *Physcomitrella patens*. Development 141(8):1660–1670.
Sharma N, Jung C-H, Bhalla PL, Singh MB, Sun M-x. 2014. RNA sequencing analysis of the gametophyte transcriptome from the liverwort, Marchantia polymorpha. PLoS One 9(5):e97497.
Sharma P, Lin T, Grandellis C, Yu M, Hannapel DJ. 2014. The BEL1-like family of transcription factors in potato. J Exp Bot. 65(2):709–723.
Szovenyi P, et al. 2015. De novo assembly and comparative analysis of the *Ceratodon purpureus* transcriptome. Mol Ecol Resour. doi: 10.1111/1755-0998.12284.
Szovenyi P, Rensing SA, Lang D, Wray GA, Shaw AJ. 2011. Generation-biased gene expression in a bryophyte model system. Mol Biol Evol. 28(1):803–812.
Tanahashi T, Sumikawa N, Kato M, Hasebe M. 2005. Diversification of gene function: homologs of the floral regulator FLO/LFY control the first zygotic cell division in the moss *Physcomitrella patens*. Development 132(7):1727–1736.
Timme RE, Bachvaroff TR, Delwiche CF. 2012. Broad phylogenomic sampling and the sister lineage of land plants. PLoS One 7(1):e29696.
van der Graaff E, Laux T, Rensing SA. 2009. The WUS homeobox-containing (WOX) protein family. Genome Biol. 10(12):248.
Wang C, Liu Y, Li SS, Han GZ. 2015. Insights into the origin and evolution of the plant hormone signaling machinery. Plant Physiol. 167(3):872–886.
Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. 2009. Jalview Version 2: a multiple sequence alignment editor and analysis workbench. Bioinformatics 25(9):1189–1191.
Wickett NJ, et al. 2014. Phylotranscriptomic analysis of the origin and early diversification of land plants. Proc Natl Acad Sci U S A. 111(45):E4859–E4868.
Zheng Y, et al. 2016. iTAK: a program for genome-wide prediction and classification of plant transcription factors, transcriptional regulators, and protein kinases. Mol Plant. 9(12):1667–1670.

Associate editor: John Archibald

2.5 Further applicability of this work

The results were published in a publicly available web interface maintained by AG Rensing, https://plantcode.online.uni-marburg.de/tapscan/, keeping it up to date by including the output of newly released genomes and making it possible for anyone to access and analyse the data.

2.6 TAPscan resource

With the ability to generate a TAP profile for any organism with available genome or proteome and to put it into a comparative perspective TAPscan is a great resource for newly sequenced genomes. This resulted in co-authorship of three impactful genome project collaborations, the Chara genome project [46], the Fern genomes project [47] and the Ulva genome project [48].

For each project, the respective novel genome(s) were screened for TAPs and the resulting annotation information was provided in itself as well as put in a comparative perspective to existing TAP knowledge.

In the Chara genome project the charophytic algae *Chara braunii* (Charophyceae) was sequenced (Fig. 3). It is the third sequenced streptophytic algae and considered the most morphologically complex one. As a Charophyceae it represents the earliest diverging Phragmoplastophyta, a monophyletic group uniting all plants sharing the same cytokinetic assembly

Figure 3. The Chara genome project making the Cell cover issue.

structure as well as other traits such as apical cell growth and branching. Due to the mosaic evolution of streptophytic algae, and though Charophyceae is not the sister group of land plants, Charophyceae is thought to be the group that can yield great insight into plant terrestrialization [49] making it a highly anticipated genome. In comparison to our genome-wide classification of 2017 (chapter 3), where Charophyceae was represented by a single transcriptome of *Nitella hyalina*, with the addition of the *C. braunii* genome we could get a much clearer picture with regards to the emergence of some TFs (ARF, HRT and TCP). ARF and HRT were previously placed to have emerged in the common ancestor of Coleochaetophyceae, Zygnematophyceae and land plants. With their presence in *C. braunii* these families should now be considered to have emerged in the common ancestor of all Phragmoplastophyta. Even bigger rearrangement concerns the TF TCP which was previously placed to have emerged in the common ancestor of all land plants. It should now be considered to have emerged in the common ancestor of all Phragmoplastophyta. TCP TFs are known to be involved in growth proliferation of organs and tissues in *A. thaliana* [50]. Them (so far) only being present in land plants and *C. braunii* speaks to the opinion that Charophyceae is the group that can yield the most insight into plant terrestrialization.

In the Fern genomes project the two ferns *Azolla filiculoides* and *Salvinia cucullata* (Salviniales) were sequenced (Fig. 4). Ferns with their interesting position in the phylogenetic tree, being the sister group to all seed plants, while lacking any genome representation makes their genomes highly anticipated. *A. filiculoides*, with its unique symbiosis with N2-fixing cyanobacteria, has long been used as a "green manure" in rice fields [51]. It was also thought to be the main actor in the "Azolla event", ~50 million years ago, sequestering atmospheric CO_2 contributing to the transition from a greenhouse to modern icehouse earth [52], thus providing a socio-economic incentive to have this specific fern sequenced.

In our genome-wide classification of 2017 (chapter 3), where ferns were represented by the single transcriptome of *Pteridium aquilinum*, we detected a loss of the TR Polycomb group EZ (PcG_EZ). The Polycomb group has been shown to be involved in body plan control [53]. With these newly sequenced fern genomes the loss of PcG_EZ was further confirmed and the speculations for it to be a clade specific loss, for all ferns, is strengthened. In between the two newly sequenced ferns and the *P. aquilinum*, one interesting disparity was the discovery of the potential secondary loss within the Salviniales of the ULT TF, being involved in inflorescence and floral meristem regulation [54].

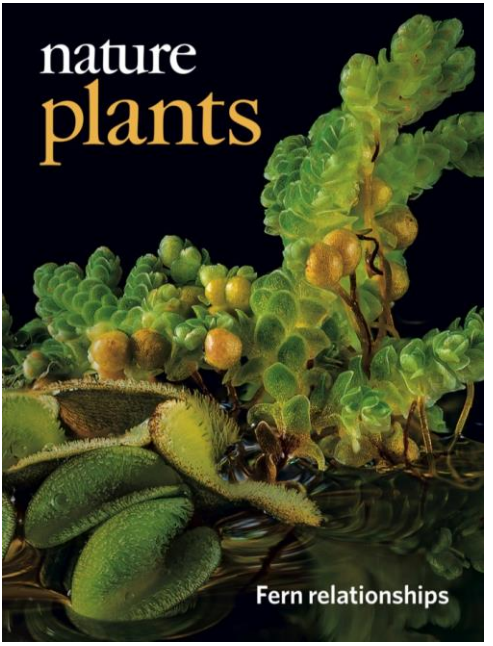
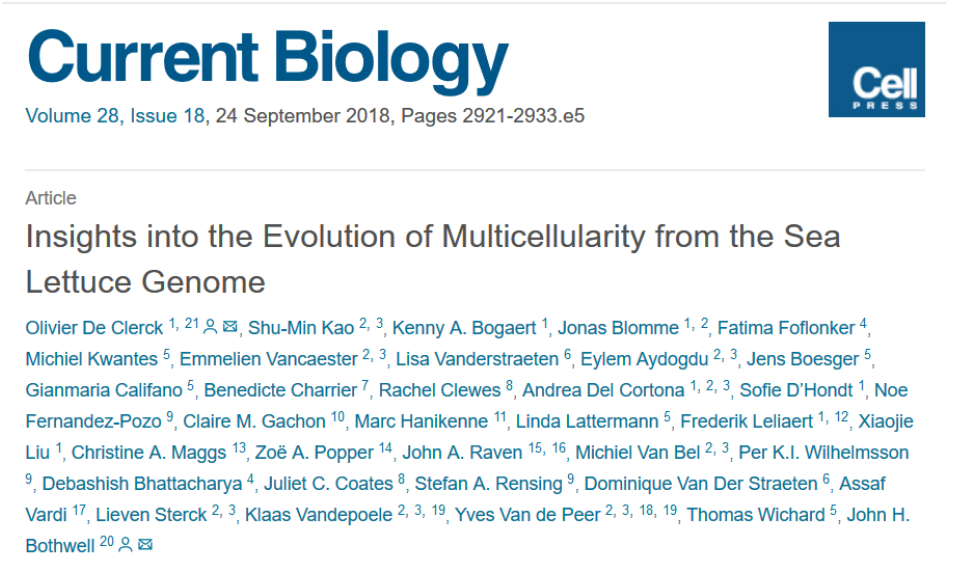


Figure 4. The Fern genome project making Nature Plants digital cover issue.

In the Ulva genome project the chlorophyte *Ulva mutabilis* was sequenced, being the first ever sequenced Ulvophyceae (Fig. 5). This green sea lettuce represents one of the multiple transitions from unicellularity to multicellularity that has occurred in Chlorophyta. *U. mutabilis* has a plant-like vegetative body, thallus, of sheet like structure. It relies on bacterial interactions to secure zoospore settlement [55] as well as for reaching a complete morphogenesis [56] making it a model organism for studying morphogenesis in sea lettuce [57]. In our genome-wide classification of 2017 (chapter 3), Chlorophyta was the most well represented clade outside of angiosperms. We detected, in total, less TAPs in *U. mutabilis* in comparison to other Chlorophytes, with only 1,94% of the genome encoding TAPs in comparison to the 2,66% average

of Chlorophyta. This was reflected in the absence of multiple TFs and TRs commonly present in other Chlorophytes. A notable exception was the high abundance of C2C2_CO-like TFs, known to be involved in the regulation of branching and shade avoidance [58] as well as flowering time [59]. We detected 5 C2C2_CO-like TFs in *U. mutabilis*, while other chlorophyte range between zero and two.



Current Biology

Volume 28, Issue 18, 24 September 2018, Pages 2921-2933.e5

Cell PRESS

Article

Insights into the Evolution of Multicellularity from the Sea Lettuce Genome

Olivier De Clerck [1, 21], Shu-Min Kao [2, 3], Kenny A. Bogaert [1], Jonas Blomme [1, 2], Fatima Foflonker [4], Michiel Kwantes [5], Emmelien Vancaester [2, 3], Lisa Vanderstraeten [6], Eylem Aydogdu [2, 3], Jens Boesger [5], Gianmaria Califano [5], Benedicte Charrier [7], Rachel Clewes [8], Andrea Del Cortona [1, 2, 3], Sofie D'Hondt [1], Noe Fernandez-Pozo [9], Claire M. Gachon [10], Marc Hanikenne [11], Linda Lattermann [5], Frederik Leliaert [1, 12], Xiaojie Liu [1], Christine A. Maggs [13], Zoë A. Popper [14], John A. Raven [15, 16], Michiel Van Bel [2, 3], Per K.I. Wilhelmsson [9], Debashish Bhattacharya [4], Juliet C. Coates [8], Stefan A. Rensing [9], Dominique Van Der Straeten [6], Assaf Vardi [17], Lieven Sterck [2, 3], Klaas Vandepoele [2, 3, 19], Yves Van de Peer [2, 3, 18, 19], Thomas Wichard [5], John H. Bothwell [20]

Figure 5. The Ulva genome project published in Current Biology.

Chapter 3

Paper II

3 Usability of reference-free transcriptome assemblies for detection of differential expression: a case study on *Aethionema arabicum* dimorphic seeds (Paper II)

3.1 Zusammenfassung

Ae. arabicum ist eine krautig wachsende, einjährige Pflanze, die in Teilen des östlichen Europas und des Mittleren Ostens heimisch ist. Sie ist eine der wenigen Pflanzen, die die Fähigkeit besitzen sowohl morphologisch als auch physiologisch voneinander verschiedene Früchte und Samen auf derselben Pflanze auszubilden (Diasporen Dimorphismus). Dies erlaubt der Pflanze eine sogenannte *bed-hedging* Strategie, eine Art Absicherungsstrategie, in der es den Pflanzensamen möglich ist in für das Wachstum günstigere Zeiten und/oder Habitate auszuweichen. Im Rahmen dieser Studie wurden für die beiden verschiedenen Samenformen RNA-Seq Daten erstellt und auf Grundlage des Referenzgenoms und eines selbst erstellten *de novo* Transkriptoms analysiert. Es zeigten sich deutliche Unterschiede im Expressionsmuster zwischen den beiden Samenformen. Die dehiszente (Kurzzeit-)Form ist eher auf schnellere Samenreifung ausgelegt, wohingegen die indehiszente (Langzeit-)Form eine stärkere Anlage zur Dormanz aufweist. Unter den differenziell exprimierten Genen (DEGs) konnte eine Vielzahl von Transkriptions-regulierender Proteine (TAPs) identifiziert werden, die in Samenreife und -dormanz involviert sind. Die vollständige funktionale Annotation (Gene Onthology) zeigte, trotz geringerer Überschneidungen in Bezug auf DEGs, eine große Übereinstimmung zwischen beiden Ansätzen. Dies zeigt auch den großen Stellenwert von *de novo* Transkriptomen für die Untersuchung von Arten ohne verfügbares Referenzgenom.

3.2 Summary

Ae. arabicum is an herbaceous annual native to parts of Eastern Europe

and the Middle East. It is one of the few plants that exhibits diaspore dimorphism, the ability to produce morphologically and physiologically distinct fruit and seed morphs on the same plant. This bet-hedging strategy makes it possible for the plant seeds to escape both space and time, to access more favorable growth conditions. RNA-seq libraries of the two seed morphs were sequenced and analyzed using both the available reference genome and a self-made *de novo* assembly. There are clear expressional differences between the two morphs with the dehiscent (short term) seed being geared towards faster maturation and the indehiscent (long term) seed being geared towards dormancy. Amongst the differentially expressed genes are a multitude of transcription associated proteins (TAPs) involved in regulating seed maturation and dormancy. Though the DEG overlap between the two approaches was low, the global functional annotations (Gene Ontology terms) overlap well, supporting the use of *de novo* assemblies when studying species with no available reference genome.

3.3 Own contribution

A robust DEG-calling pipeline, using the consensus of three separate DEG-calling packages, was developed in cooperation with Kristian K. Ullrich and Stefan A. Rensing. I then applied the pipeline to the RNA-sequence libraries of *Ae. arabicum* dry seeds. Both the available genome and a self-generated *de novo* assembly was annotated with Gene Ontology (GO) terms and screened for TAPs. Both assemblies were processed with the DEG-calling pipeline for the purpose of investigating the differences between the outcome of each. With the help of SeedAdapt collaborators Jake O. Chandler, Kai Graeber, Waheed Arshad, Safina Khan, Michael E. Schranz and Gerhard Leubner-Metzger, a biological interpretation was carried out. I contributed to the writing of the manuscript as well as prepared most of its figures.

3.4 Paper

Following is the electronic publication.

BMC Genomics

RESEARCH ARTICLE Open Access

Usability of reference-free transcriptome assemblies for detection of differential expression: a case study on *Aethionema arabicum* dimorphic seeds

Per K. I. Wilhelmsson[1], Jake O. Chandler[2], Noe Fernandez-Pozo[1], Kai Graeber[2], Kristian K. Ullrich[1,8], Waheed Arshad[2], Safina Khan[2], Johannes A. Hofberger[3], Karl Buchta[1], Patrick P. Edger[4], J. Chris Pires[5], M. Eric Schranz[3], Gerhard Leubner-Metzger[2,6*] and Stefan A. Rensing[1,7*]

Abstract

Background: RNA-sequencing analysis is increasingly utilized to study gene expression in non-model organisms without sequenced genomes. *Aethionema arabicum* (Brassicaceae) exhibits seed dimorphism as a bet-hedging strategy – producing both a less dormant mucilaginous (M^+) seed morph and a more dormant non-mucilaginous (NM) seed morph. Here, we compared de novo and reference-genome based transcriptome assemblies to investigate *Ae. arabicum* seed dimorphism and to evaluate the reference-free versus -dependent approach for identifying differentially expressed genes (DEGs).

Results: A de novo transcriptome assembly was generated using sequences from M^+ and NM *Ae. arabicum* dry seed morphs. The transcripts of the de novo assembly contained 63.1% complete Benchmarking Universal Single-Copy Orthologs (BUSCO) compared to 90.9% for the transcripts of the reference genome. DEG detection used the strict consensus of three methods (DESeq2, edgeR and NOISeq). Only 37% of 1533 differentially expressed de novo assembled transcripts paired with 1876 genome-derived DEGs. Gene Ontology (GO) terms distinguished the seed morphs: the terms translation and nucleosome assembly were overrepresented in DEGs higher in abundance in M^+ dry seeds, whereas terms related to mRNA processing and transcription were overrepresented in DEGs higher in abundance in NM dry seeds. DEGs amongst these GO terms included ribosomal proteins and histones (higher in M^+), RNA polymerase II subunits and related transcription and elongation factors (higher in NM). Expression of the inferred DEGs and other genes associated with seed maturation (e.g. those encoding late embryogenesis abundant proteins and transcription factors regulating seed development and maturation such as ABI3, FUS3, LEC1 and WRI1 homologs) were put in context with *Arabidopsis thaliana* seed maturation and indicated that M^+ seeds may desiccate and mature faster than NM. The 1901 transcriptomic DEG set GO-terms had almost 90% overlap with the 2191 genome-derived DEG GO-terms.

Conclusions: Whilst there was only modest overlap of DEGs identified in reference-free versus -dependent approaches, the resulting GO analysis was concordant in both approaches. The identified differences in dry seed transcriptomes suggest mechanisms underpinning previously identified contrasts between morphology and germination behaviour of M^+ and NM seeds.

Keywords: *Aethionema arabicum*, Dimorphic seeds, Reference and reference-free, RNA-seq, Transcriptome,

* Correspondence: gerhard.leubner@rhul.ac.uk; stefan.rensing@biologie.uni-marburg.de
[2]School of Biological Sciences, Royal Holloway University of London, Egham, Surrey TW20 0EX, UK
[1]Plant Cell Biology, Faculty of Biology, University of Marburg, 35043 Marburg, Germany
Full list of author information is available at the end of the article

© The Author(s). 2019 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. 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

RNA-sequencing (RNA-seq) technology is a valuable tool to investigate gene expression [1], especially in species where no reference genome is available. Without any prior molecular data about a particular species, de novo transcriptome assembly of RNA-seq data offers a unique opportunity to study gene expression on a transcriptome-wide scale of any trait of interest. Due to drops in library and sequencing costs, it is now widely utilized by many scientists to study traits of particular interest in a wide-range of species. However, there are limitations to using a de novo transcriptome assembly compared to a reference-genome guided approach. Since less sequence information is used in the creation of the transcripts in a de novo transcriptome, in comparison to a reference genome, low expressed genes are more difficult to detect. De novo assembled transcripts are also more likely to be fragmented.

Here, we apply a reference-free and a reference-dependent approach to compare the gene expression in the dry mature dimorphic seeds of *Aethionema arabicum*. This species represents the sister lineage to all other Brassicaceae, and is a herbaceous annual native to parts of Eastern Europe and the Middle East. It exhibits diaspore heteromorphism – i.e. the ability to produce multiple morphologically and physiologically distinct fruit or seed morphs on individual plants [2, 3]. *Ae. arabicum* produces two distinct fruits, a dehiscent (DEH) and an indehiscent (IND) fruit morph. The dehiscent fruit contains typically four seeds, shatters on maturity, and disperses mucilaginous seeds (M^+). Conversely, the indehiscent fruit contains a single non-mucilaginous seed (M^-) encased in a pericarp (fruit coat). Upon maturity, the entire IND fruit detaches, via abscission, from the parent plant leading to the fruit's dispersal [3, 4]. In addition to these morphological differences between the two morphs, the NM seeds appear to be more dormant compared to the M^+ seeds, with NM exhibiting much slower germination at 14°C [3]. The production of two contrasting seed/fruit morphs is proposed to constitute a bet-hedging strategy that increases long-term plant fitness in disturbed and unpredictable extreme environments. However, how this heteromorphism is reflected at the transcriptomic level is unknown. With its recently published genome sequence and its basal phylogenetic position within the Brassicaceae, *Ae. arabicum* has potential as a model species for diaspore heteromorphism [3, 5].

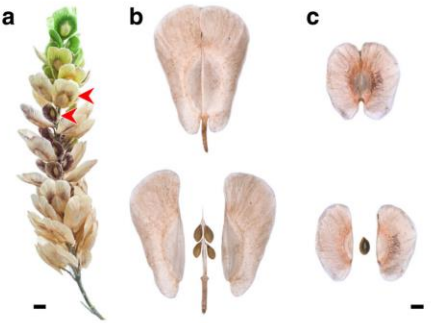
For many other non-model plant species, including other heteromorphic systems, a reference genome is not available. Thus, comparing the effectiveness of reference-free and reference-dependent transcriptome analyses is pertinent to future investigations into such non-model species. Comparison of the transcriptomes of the two *Ae. arabicum* seed morphs represents a realistic and interesting demonstration of both approaches. There are many genomes with accompanying large sets of microarray and qRT-PCR data, and it was early on concluded that de novo assembled transcriptome expression profiles positively correlate with corresponding microarrays and qRT-PCRs [6–8]. Due to the potential of RNA-seq, much work has been done on how to get the best results out of a de novo transcriptome assembly [9–13]. The Trinity suite [14] is one of the most cited de novo transcriptome assemblers exhibiting good performance metrics [13]. In order to generate a representative transcriptome, sequencing depth is important to be able to reconstruct as many genes as possible including those expressed at low levels. The ability to detect weakly expressed sequences can only be improved by increasing the sequencing depth. This highlights the diminishing investment returns (sequencing depth) in relation to yield (sequence resolution) for RNA-seq. Despite the known limiting factors of transcriptome assembly, the knowledge gained per investment makes reference-free gene expression profiling an obvious choice when working with non-model species.

To evaluate the knowledge that can be gained with reference-free gene expression profiling, a reference-dependent expression profiling was carried out using the existing genome assembly of *Ae. arabicum* [5]. To investigate the seed dimorphism of *Ae. arabicum*, we conducted a highly robust differentially expressed genes (DEGs) detection analysis and used it to compare DEGs derived from a transcriptome-based and a genome-based mapping approach. The aim of this study was to find DEGs between *Ae. arabicum* dimorphic seeds, and to compare the RNA-seq analysis performed using two different references, a de novo transcriptome assembly and the *Ae. arabicum* genome sequence V2.5.

Results and discussion

Overview of RNA-seq analysis of *Ae. arabicum* mature dimorphic seeds

The mature dimorphic seeds, M^+ from DEH fruits and NM from IND fruits (designated NM, for "non-mucilaginous", in our RNA-seq analysis), differed in size and mass but not in seed moisture content (Fig. 1). RNA was extracted from freshly harvested mature M^+ and NM seeds and the resultant RNA samples processed as described in the Methods section. As shown in Fig. 2, RNA-seq raw reads were processed and checked using FastQC



	M+ (derived from DEH fruit)	NM (derived from IND fruit)
Individual seed mass (μg)	378.0 ± 1.7	289.4 ± 2.2
Dry seed moisture content (% FW)	8.0 ± 0	9.0 ± 0

Fig. 1 Fruit and seed dimorphism in *Ae. arabicum*. Mature infructescence (**a**) of *Ae. arabicum*, showing distinct dehiscent (DEH) and indehiscent (IND) fruit morphs (marked by red arrows). Large DEH fruits (**b**) contain up to six mucilaginous (M^+) seeds, while small IND fruits (**c**) contain a single, non-mucilaginous (NM) seed. Both seed morphs differ in mean seed mass and moisture content. Values shown are means ± SEM for $n = 8$ each of 100 seeds (mass), and $n = 4$ each of 30 seeds (moisture content) replicate measurements. Scale bars = 4 mm (**a**), 1 mm (**b** and **c**). FW, fresh weight

(https://www.bioinformatics.babraham.ac.uk/projects/fastqc/), Trimmomatic version 0.32 [15] and PrinSeq [16]. Subsequently, cleaned reads were used for de novo transcriptome assembly for *Ae. arabicum* M^+ and NM seeds using Trinity [14]. The same set of cleaned reads was mapped to the gene models of the reference genome using GSNAP [17]. EdgeR, DESeq2 and NOISeq [18–20] were used to normalize read counts and to detect DEGs in a strict consensus approach, and Blast2GO [21] was used to assign Gene Ontology (GO) terms to the genes. Comparisons were performed between the transcriptome and the genome (Comparison 1, Fig. 2), the reads mapped to both the de novo transcriptome and reference-based genes (Comparison 2, Fig. 2), the DEGs found in both approaches (Comparison 3, Fig. 2), and between their GO terms (Comparison 4, Fig. 2).

Read filtering of RNA-seq raw data

To generate the raw reads, a total of four cDNA libraries were sequenced, with two biological replicates of *Ae. arabicum* dry mature dimorphic seeds, termed M^+1/M^+2 for the M^+ seeds and NM1/NM2 for the NM seeds. Raw reads were processed to remove adapters, organellar, ribosomal RNA (rRNA) and low-quality sequences (Fig. 3). Adapter sequences were removed and low-quality sequences were trimmed using Trimmomatic. Poly-A and poly-T tails were removed using PrinSeq. This process resulted in an average loss of 9.6% of all reads for the four libraries. To reduce the complexity of the assembly/mapping, and to check for correct poly-T selection, all data were filtered to remove reads with plastid, mitochondrial and ribosomal RNA origin resulting in an average loss of 12% of the reads for the four libraries. Visualization of these quality control steps provides a good measure of library quality making possible to see if there are any higher than average read losses in the individual steps. After passing all the filters, the sets of cleaned sequences contained between 20 and 30 million reads (Fig. 3), which is in the range of read numbers commonly used for RNA-seq analysis for DEG detection [22].

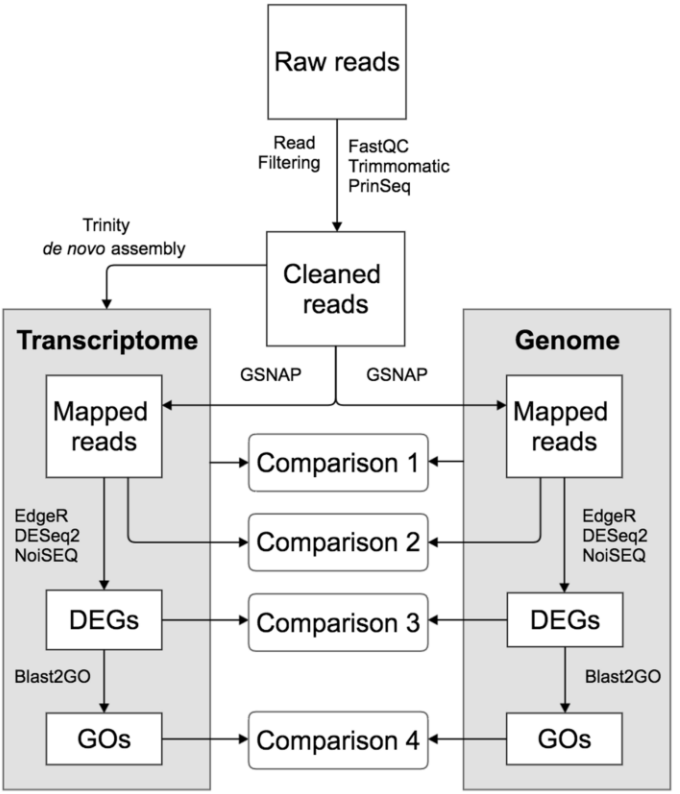


Fig. 2 RNA-Seq analysis pipeline. Raw RNA-seq reads were checked for quality control (FastQC) and processed to remove adapters and low-quality bases (Trimmomatic, PrinSeq). Cleaned reads were either: mapped to the genome (GSNAP); or were used for de novo *transcriptome* assembly (Trinity) and mapped to the resulting transcriptome (GSNAP). Transcriptome-mapped and genome-mapped reads were compared at each stage of analysis: After mapping; after differentially expressed gene (DEG) identification (EdgeR, DESeq2, NoiSEQ), and after gene-ontology (GO) analysis (Blast2GO)

De novo transcriptome assembly

Processed reads from all four samples combined were assembled de novo using Trinity to reconstruct the *Ae. arabicum* dry seed transcriptome. From a total of 30,742,186 reads, 27,407,363 reads (89.15%) could be assembled. This resulted in a total of 62,182 transcripts including potential splice variants or fragmentary sequences. The longest gene sequences from each Trinity gene cluster were selected to reduce redundancy, resulting in 34,784 transcripts (Additional file 1). To assess the quality and completeness of the *Ae. arabicum* dry seed de novo transcriptome, and to compare it to the gene models from the genome (Comparison 1, Fig. 2), it was analyzed using the Benchmarking Universal Single-Copy Orthologs (BUSCO) tool [23] (embryophyta *odb9*) which checks for the presence of Embryophyta "near-universal single-copy orthologs". For the de novo assembled transcriptome, 908 transcripts out of 1440 of the BUSCO genes were complete (63.1%). Of those, 885 were single copy and 23 duplicated. One hundred sixty-eight transcripts were fragmented and 364 missing (Fig. 4). The corresponding number of BUSCO completeness in the 23,594 gene models of the genome was 1309 (90.9%). Of those, 1274 were single copy and 35

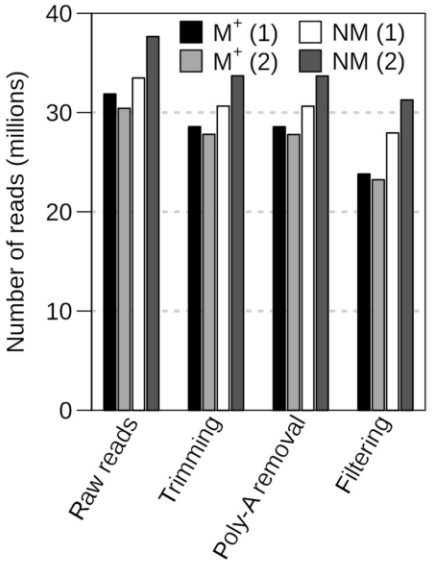


Fig. 3 From raw to filtered reads. Trimming of raw reads with Trimmomatic removed adapters and low-quality reads. Trimmed reads were further processed with poly-A / poly-T removal with PrinSeq. The resulting reads were then filtered to remove chloroplastic, mitochondrial and ribosomal RNA reads. The total number of reads left after each step is indicated for samples M^+ (1), M^+ (2), NM (1) and NM (2)

duplicated. Forty-one gene models were fragmented and 90 missing (Fig. 4). To compare these results with a well-annotated model species, *Arabidopsis thaliana* (TAIR10, [24]) was included in the BUSCO analysis. For *A. thaliana*, 1431 complete genes were found (99.3%), 1413 were single copy and 18 duplicated; five genes were fragmented and four missing. The relatively low number of complete genes in *Ae. arabicum* transcriptome is to be expected, since dry seeds represent an atypical tissue that lacks much of the transcription going on in photosynthetically/developmentally active tissue. Also, it is common that some genes are fragmented in de novo assemblies, as shown in Fig. 5a which indicates the length distribution of de novo assembled transcripts is skewed towards shorter lengths compared to the *Ae. arabicum* mRNAs predicted from the genome.

Mapping reads to the transcriptome and the genome

To determine read counts for subsequent DEG analysis, cleaned reads were mapped to the transcriptome and the genome using GSNAP [17] and counted using HTSeq-count [25] with the respective general feature format (GFF) file. Counted reads for the four samples are shown in Fig. 5b. This analysis showed that on average 84% of reads were mapped to the transcriptome and 94% to the genome. The drop from 89.15% of the reads being used for assembling to 84% mapping is to a large extent explained by the removal of redundancy keeping only the longest isoform of each transcript. On average, the cleaned reads had a read length of 83 bp. Mapping the reads to the 23,594 genomic gene models, 7814 models had a coverage lower than 1 (where 1 corresponds to an average 1-fold coverage of the gene length; see Methods for details) and 11,189 gene models had a coverage lower than 5 (Additional file 2: Table S1). This highlights the challenges to assemble full-length transcripts. Using reciprocal BLASTN with a coverage cut off of 50% for both transcriptomic (virtual transcripts) and genomic coding sequences (CDS), 6745 transcript-gene pairs could be identified (Additional file 2: Table S2). To compare the expression levels between the transcriptome- and the genome-based approach (Comparison 2, Fig. 2), the 6745 gene-transcript pairs were considered. Principal Component Analysis (PCA) using the Reads Per Kilobase per Million mapped reads (RPKM) of the 6745 genes (Additional file 3: Figure S1) showed, as expected [9], that replicates from the same seed morph clustered together and samples from different seed morphs are more distant. This is apparent in both the de novo and reference-genome approach. To assess gene family completeness, the predicted proteins of the reference genome and the de novo transcriptome were screened for Transcription Associated Proteins (TAPs, comprising transcription factors, TF, and transcriptional regulators, TR) using the TAPscan pipeline [26]. 1860 (113 unique families) and 1009 (105 unique families) TAPs were detected in the genome and transcriptome, respectively (Additional file 2: Table S3 and S4). Finding fewer TAPs in the transcriptome is to be expected due to the atypical tissue of the transcriptome in comparison to the whole genome. Genome-wide, 7.6% were multi domain TAPs (defined by more than one domain), while only 4.2% TAPs were multi domain in the transcriptome, due to the fragmented nature of the transcriptome.

Differential gene expression analysis

To learn more about the differences between the mature dimorphic seeds, gene expression was analyzed using both references: the de novo transcriptome assembly and the genome annotation. Since the combination of several methods minimizes false positives [27], DEGs were detected in a robust way using the strict consensus (overlap) of three different DEG analysis programs: edgeR, DESeq2 and NOISeq. This approach combines

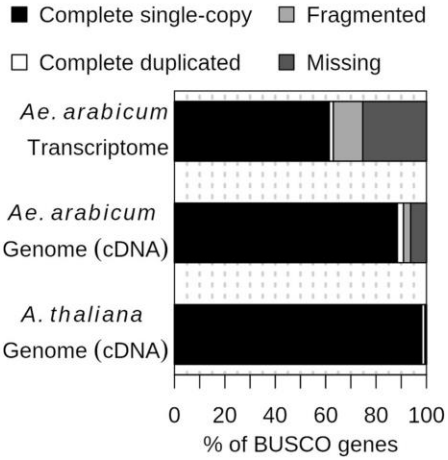


Fig. 4 BUSCO completeness analysis. cDNA from the *Ae. arabicum* de novo assembly, *Ae. arabicum* genome v2.5 and *A. thaliana* TAIR10 were compared to 1440 Embryophyta reference orthologs for completeness assessment

two parametric methods to detect DEGs (edgeR and DESeq2), and a non-parametric method (NOISeq). The intersection of the DEGs obtained by the three methods was considered the resulting DEGs (Fig. 6a, b). In all comparisons edgeR called the most DEGs while NOISeq was the most restrictive (Fig. 6a, b), thus the NOISeq set was representing the consensus DEG set best. This approach resulted in the exclusion of low expressed DEGs (Additional file 3: Figure S2) below RPKM 2, representing genes of low abundance that typically cannot be shown as expressed in a quantitative PCR approach [28].

One thousand five hundred thirty-three and one thousand eight hundred seventy-six DEGs were obtained, respectively, using the de novo transcriptome (Fig. 6a, Additional file 2: Table S4) and the reference genome (Fig. 6b; Additional file 2: Table S3). When comparing common DEGs detected in both approaches (Comparison 3, Fig. 2), 561 gene-transcript pairs were found to be differentially expressed in both. Thus, 561/1533 (37%) of the de novo transcriptome consensus DEGs were also well represented by transcripts identified as DEGs by the genome approach, all of them showing the same direction of expression (Additional file 2: Table S2). PCA for the 561 DEGs identified by both approaches showed that the biological differences between the dimorphic seeds are much greater than the differences deriving from the references used (Fig. 6c). All samples from the same seed morphs clearly clustered together, independently of the sequence reference (transcriptome or genome). The remaining 972 transcripts (63%) of the 1533 transcriptome DEGs did either not pass the 50% coverage cut-off (405/1533), only had a hit in one direction of the reciprocal BLAST (122/1533), their reciprocal hit was not a DEG in the genome (197/1533) or they did not produce any significant alignment at all (248/1533). Hence, approximately 40% of the DEGs from the de novo transcriptome assembly are equivalent to the DEGs found when a genome reference is available, and 60% of the DEGs were either fragmented or could not be clearly paired up with a gene model. This indicates that data for individual genes might not always be available when working with de novo transcriptome differential expression analysis. In cases like this, it might be important to perform other analyses that study the changes of global functions occurring in the samples, such as Gene Ontology bias. To verify the robustness of the expression pattern between the dimorphic seeds, we performed qRT-PCR on a selection of DEGs with varying levels of RPKM values in an independent biological experiment (Additional file 3: Figure S3). Despite the fact that the qRT-PCR results are derived from a completely independent experiment with different RNA samples, the expression patterns were confirmed for eight of the ten selected DEGs.

Gene ontology analysis

The number of GO terms associated with the genome and the de novo transcriptome, for all transcripts, for the DEGs and for the overlap between both approaches is summarized in Table 1 (and in more detail in Additional file 2: Table S5–S6) and is referred to as a GO-presence list. When comparing

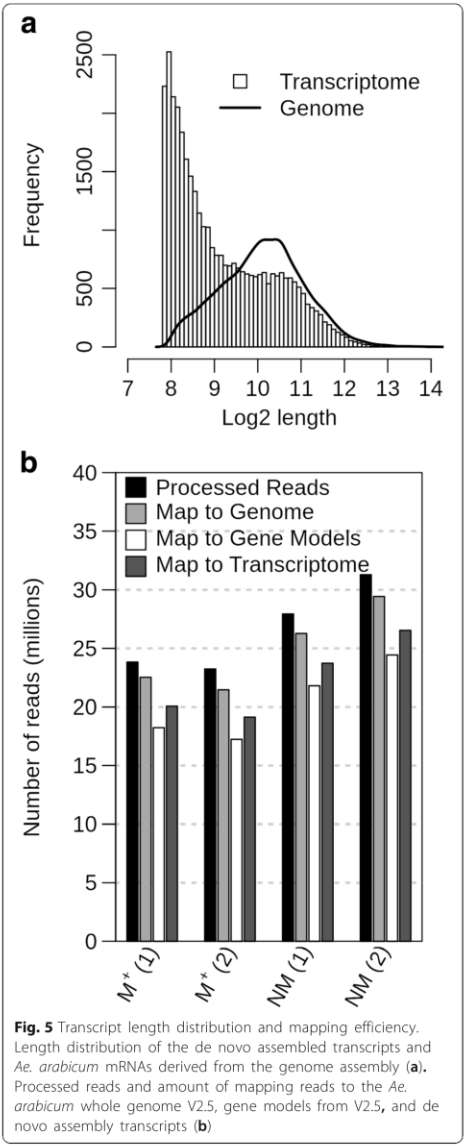


Fig. 5 Transcript length distribution and mapping efficiency. Length distribution of the de novo assembled transcripts and *Ae. arabicum* mRNAs derived from the genome assembly (**a**). Processed reads and amount of mapping reads to the *Ae. arabicum* whole genome V2.5, gene models from V2.5, and de novo assembly transcripts (**b**)

(Comparison 4, Fig. 2) what is shared between the GO-presence list of the reference genome and the de novo transcriptome (All Transcripts Overlapping GO terms from Table 1; using Fisher's exact test with an fdr corrected p value of 0.05), only 12 out of 5584 GO terms were shown to have significant differences in the number of transcripts associated to them (Additional file 2: Table S5). The GO-presence list of the DEGs (All DEGs Overlapping GO terms from Table 1) showed no significant differences at all between the genome and the transcriptome (Additional file 2: Table S6). Furthermore, having 1663 common GO terms present in the GO-presence lists of both DEG sets (Fig. 7) is a significant over-representation compared to the null hypothesis of selecting 1901 and 2191 GO terms randomly (Chi squared test, p = 2.2e-16). This suggests a biological signal, supporting that functional analysis of GO terms by transcriptome de novo assembly resembles the data derived by genomic analysis.

For both the 1256 overlapping GO-terms of the DEGs GO-presence lists with higher abundance in NM (NM) seeds ("NM-high") and 880 overlapping GO-terms of the DEGs GO-presence lists of with lower abundance in NM seeds ("NM-low"), none had significantly different quantities of underlying transcripts. The numbers and overlap of significantly over- and under-represented GO-terms of each class (Biological Process (BP), Molecular Function (MF) and Cellular Component (CC)) for all, NM-high and NM-low DEGs derived from the two approaches are summarized in Additional file 2: Table S7 and in more detail in (Additional file 4: Table S8) and are referred to as GO-bias lists. Overall, the NM-high and NM-low BP GO-bias lists are quite different. In the reference genome approach, NM-high has 340 unique BP terms, NM-low has 137 unique BP terms in the respective GO-bias list, with only 58 BP terms overlapping between both sets. Some of the most significant overlapping BP terms belong to high-level categories, such as 'protein metabolic process' and 'gene expression' (comprehensive lists of GO terms associated with the DEG sets are provided in Fig. 7). In agreement with this, ribonucleoprotein complex is the most significantly over-represented CC term in the genome approach, and structural constituent of ribosome is the most significantly over-represented MF term (Additional file 4: Table S8).

Many of the GO-terms found to be significantly over-represented and under-represented using the transcriptome approach were also found with the genomic approach: Out of the 321 BP terms found to be significantly over (255) and under (66) represented in the transcriptome-derived DEG set (GO-bias lists) (Fig. 7b and Additional file 4: Table S8), 258 (80%) were also found to be the same in the genome-derived DEG set (GO-bias lists) (Fig. 7a and Additional file 4: Table S8. On average, approximately 80% of the significantly over- and under-represented GO terms of the transcriptomic DEG sets (GO-bias lists) were also

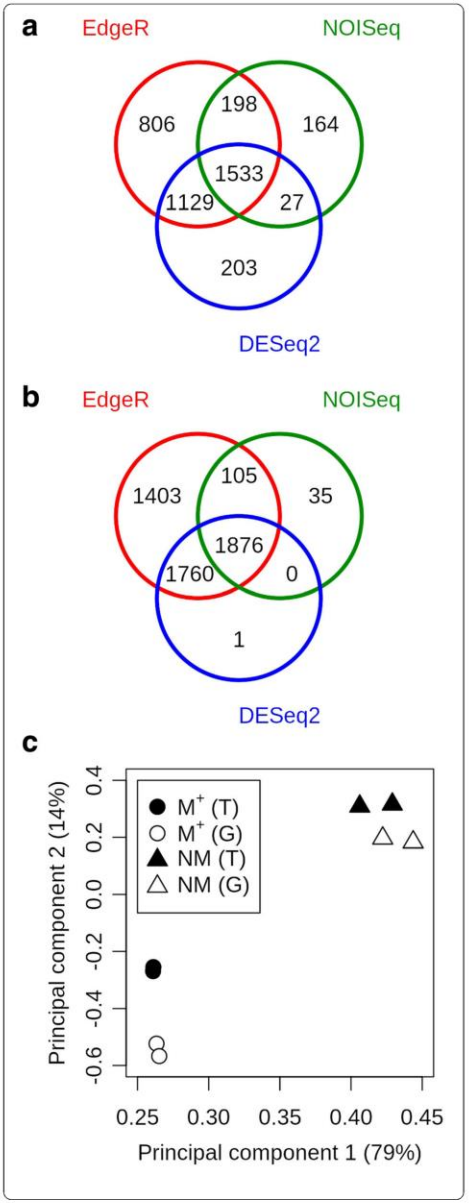


Fig. 6 Consensus of DEG calling and PCA of overlap of common DEGs. Venn diagram of the DEGs called between NM and M^+ seeds by the three DEG detection programs (edgeR, NOIseq and DESeq2) using the transcriptome (**a**) and genome (**b**) approach. Principal Component Analysis of RPKM (Reads Per Kilobase per Million reads) of the 561 DEGs common to the transcriptome, 'T' and genome, 'G' (**c**). Samples M^+ (circle) and NM (triangles), in black, show the results for the dehiscent and indehiscent seeds in the transcriptome approach. Samples M^+ (circle) and NM (triangles), in white, show the corresponding results in the genome approach. The percentage variance explained by each principal component is indicated on the axes

reported using the genomic approach. So, in comparison to the 40% overlap of DEGs on a gene-transcript pair level, we found a much higher overlap of differentially expressed functions between the *Ae. arabicum* M^+ and NM dimorphic seeds using GO term bias analysis, even though some of the genes involved in these functions are missing in the transcriptome DEG dataset. The genomic approach reports on average 37% more GO-terms to be significantly over- or under-represented, which can be explained by the 22% more DEGs and 10% more GO-terms per gene. Though a transcriptome de novo assembly approach gives less information, the information that is given overlaps very well with a genome-based approach. Taken together, this finding supports the view that analysis of GO terms by transcriptome de novo assembly is a useful tool when no

Table 1 Summary of GO terms associated with both the genome- and transcriptome-derived transcripts and respective DEG sets

		Transcriptome	Genome
All Transcripts	Total number	34,784	23,594
	Number with GO terms	18,845 (54%)	18,320 (78%)
	GO terms per transcript[a]	7.1	7.9
	Amount of GO terms	6091	6080
	Overlapping GO terms	5584	
All DEGs (M^+ + NM)	DEGs	1533	1876
	Amount of GO terms	1901	2191
	Overlapping GO terms	1663	
NM-high[b]	DEGs	745	998
	Amount of GO terms	1427	1673
	Overlapping GO terms	1256	
NM-low[c]	DEGs	788	878
	Amount of GO terms	1085	1185
	Overlapping GO terms	880	

[a]Average including only transcripts with at least 1 GO term
[b]DEGs where transcript is more abundant in NM dry seed than M^+ seed
[c]DEGs where transcript is less abundant in NM dry seed than M^+ seed

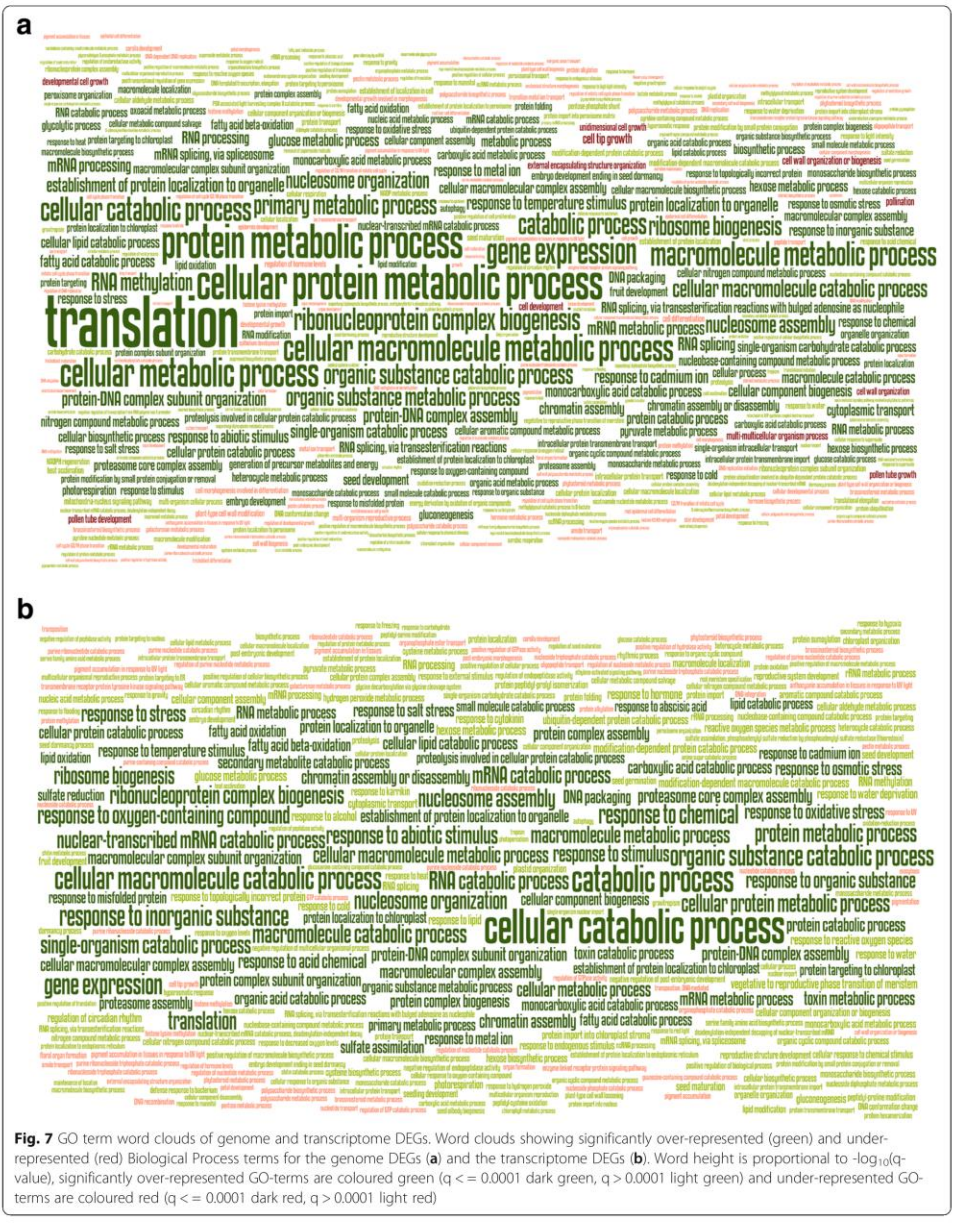


Fig. 7 GO term word clouds of genome and transcriptome DEGs. Word clouds showing significantly over-represented (green) and under-represented (red) Biological Process terms for the genome DEGs (**a**) and the transcriptome DEGs (**b**). Word height is proportional to $-\log_{10}$(q-value), significantly over-represented GO-terms are coloured green ($q <= 0.0001$ dark green, $q > 0.0001$ light green) and under-represented GO-terms are coloured red ($q <= 0.0001$ dark red, $q > 0.0001$ light red)

genome is available, and resembles the data derived by genomic analysis.

DEG analysis of mature dimorphic *Ae. arabicum* seeds

The most significantly over-represented BP terms unique to the NM-high DEGs GO-bias list (transcripts with a higher abundance in NM seed compared to M^+ seed) include mRNA metabolic process, mRNA processing and response to stimulus. On the other hand, the most significantly over-represented BP terms unique to the NM-low DEGs GO-bias lists (transcripts with a lower abundance in NM seed compared to M^+ seed) are translation, ribosome biogenesis and nucleosome assembly (Additional file 4: Table S8). This is also reflected in the CC and MF terms, with the nucleus CC term and RNA binding MF term being among the most significantly over-represented terms in the NM-high DEGs (GO-bias list) and the structural constituent of ribosome MF term and ribosome CC term being among the most significantly over-represented terms in the NM-low DEGs (GO-bias lists; Additional file 4: Table S8). Thus, it is generally indicative that the transcriptome of the M^+ "dry" mature seed morph transcriptome may be relatively more oriented towards translation of RNA and chromatin assembly, whereas the NM "dry" mature seed morph transcriptome may be more oriented to post-transcriptional processing of RNA. It is possible that these differences may reflect the stage which was sampled – the dry seed. Thus, transcriptomic differences may be due to differences in the stage of seed development or maturation the seed morphs have reached before desiccation. For this reason, we put the transcriptomic differences between *Ae. arabicum* NM and M^+ seed in context of the well-studied seed development and maturation of *A. thaliana*.

The *Ae. arabicum* M^+ seed morph as well as *A. thaliana* seeds are both dispersed from dehiscent fruits and seem to resemble each other in terms of morphology and physiology [3]. In Fig. 8, we compare the expression of selected *Ae. arabicum* key DEGs (which differ between the dimorphic M^+ and NM seeds, selected based on the prominent GO terms and genes with importance to seed development and maturation) with the expression of their putative orthologs derived from published transcriptomes of developing and mature *A. thaliana* seeds [29–31]. During the *A. thaliana* seed maturation and late maturation phases desiccation tolerance and dormancy are established in parallel with drying resulting in the low-hydrated dispersed seed state (Fig. 8a) [32, 33].

For the dry mature *Ae. arabicum* dimorphic seeds, we found that the abundance of at least 119 (reference approach) and 113 (de novo approach) ribosomal protein transcripts were 1.5- to 3-fold higher in M^+ seeds as compared to NM seeds (Fig. 8d, Additional file 3: Figure S4a). This seems to be a general pattern as there were no ribosomal protein genes with higher transcript abundances in NM seeds. The abundance of the putatively orthologous transcripts of these DEGs decreased during *A. thaliana* seed maturation (Fig. 8b). A genome-wide analysis of ribosomal protein gene expression during *A. thaliana* and *Brassica napus* seed maturation revealed the same temporal pattern [30, 34]. During maturation, ribosomal activity is required for processes such as seed storage compound accumulation which deceases upon late maturation drying. In dry seeds, ribosomes are mainly present in the monosome form [35]. Ribosomal profiles change with polysomes being formed during seed germination and subsequent seedling growth. Interestingly, during these processes, differential expression of ribosomal protein genes occurs and may affect ribosome composition and thereby the selection of translated mRNAs [31, 35–37]. 35–40% (reference approach) and ~ 30% (de novo approach) of the ribosomal protein genes in M^+ seeds show approximately 2-fold higher transcript abundances, which suggests that they dry out earlier during late maturation as compared to NM seeds. Considering their overall decrease over time during seed maturation (Fig. 8b), this would explain the higher abundance of transcripts for ribosomal protein genes in dry M^+ seeds. Alternatively, M^+ seeds could have a higher translational activity with a higher ribosome per seed content. In the latter case, we would also expect elevated rRNA biosynthesis in the larger M^+ seeds as compared to the smaller NM seeds. This is however not the case, as evident from the rRNA amounts estimated by filtering during the RNA-seq workflow (Figs. 2 and 3). We therefore conclude that the higher transcript abundance of a large number of ribosomal protein genes in M^+ seeds seems to be due to faster drying of M^+ seeds during late maturation. This conclusion is also consistent with the DEG patterns for histones and other genes as discussed later. We propose that the earlier drying out may preserve the mature M^+ seeds in a state with higher ribosome content and translational activity compared to the mature NM seeds. The distinct states are consistent with the distinct germination and dormancy behavior of the dimorphic *Ae. arabicum* seeds [3].

The NM-low DEGs of the reference approach related to nucleosome assembly include 21 *Ae. arabicum* histone genes, including seven H4, five H3, four H2B, five H2A, but no H1 homolog of *A. thaliana* histone variants. For the dry mature *Ae. arabicum* dimorphic seeds, we found that the abundance of these histone transcripts were 1.5- to 4-fold higher in M^+ seeds as compared to NM seeds (Fig. 8d, Additional file 3: Figure S4b). The NM-low DEGs of the de novo approach related to nucleosome assembly include nine histone genes, including

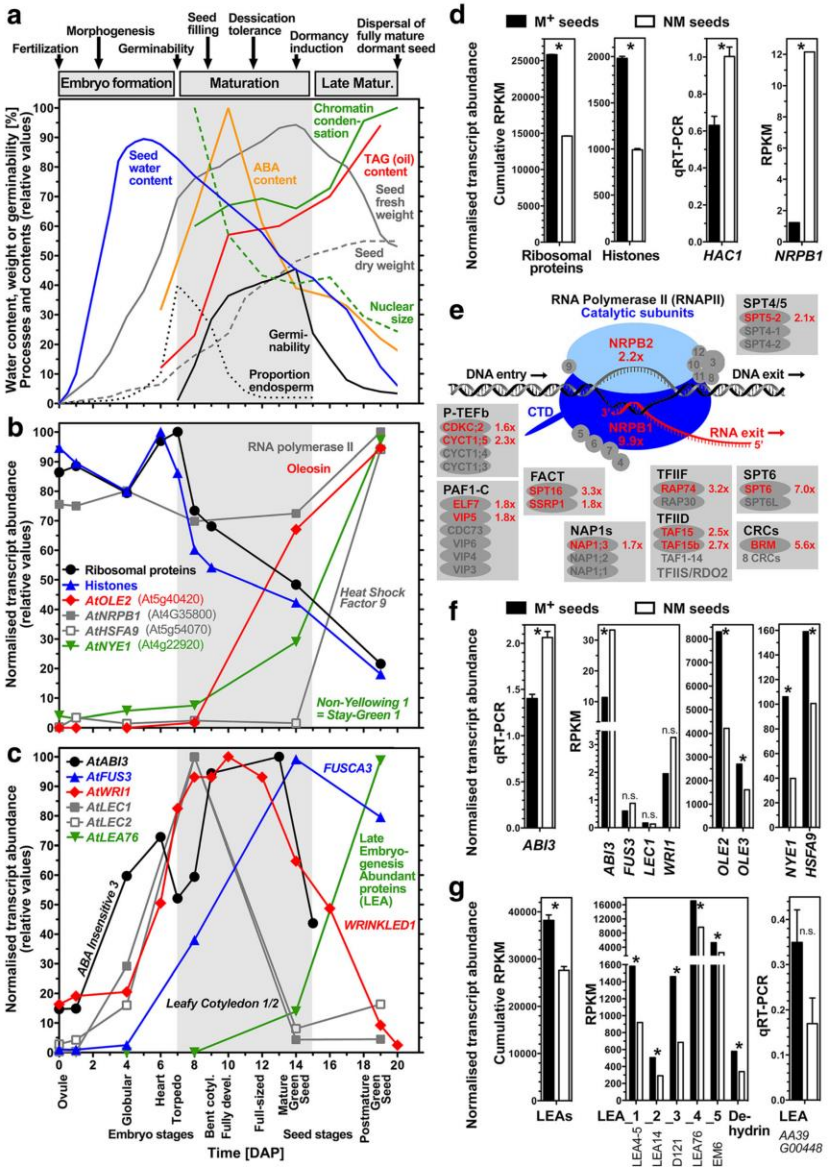


Fig. 8 (See legend on next page.)

(See figure on previous page.)
Fig. 8 Key processes and differentially expressed genes (DEGs) differ between *Ae. arabicum* M^+ and NM seeds. **a** Timing of key processes during development and maturation of *A. thaliana* seeds. Dormancy and desiccation tolerance coincides with changes in water, abscisic acid (ABA) and triacylglycerol (TAG) contents, seed weight, nuclear size and chromatin condensation, endosperm proportion and germinability; Data from [32, 41, 55]. **b** Selected *Ae. arabicum* DEG putative ortholog expression during *A. thaliana* seed development and maturation. Cumulative transcript abundances for *A. thaliana* putative orthologs of *Ae. arabicum* 21 histone and 119 ribosomal protein genes (Additional file 3: Figure S4); individual abundances for RNA polymerase II large subunit (*AtNRPB1*), oleosin *AtOLE2* (seed storage), heat shock factor *AtHSFA9* (longevity), and *AtNYE1* (chlorophyll degradation); data from Arabidopsis eFP browser [74] and [29–31]. **c** Expression of late embryogenesis abundant (*LEA*) proteins, seed maturation master regulators (*AtLEC1*, *AtLEC2*, *AtABI3*, *AtFUS3*) and *WRINKLED1* (*AtWRI1*), a transcription factor associated with enhanced fatty acid and TAG biosynthesis during *A. thaliana* seed maturation; data from Arabidopsis eFP browser and [29–31, 58]. **d** Expression of selected *Ae. arabicum* DEGs for ribosomal proteins, histones, *NRPB1* (RNAseq) and histone acetyltransferase *HAC1* (qRT-PCR) in M^+ and NM seeds. Cumulative RPKM values presented for 21 histone and 119 ribosomal protein genes of *Ae. arabicum* (Additional file 3: Figure S4). A * indicates a significant difference between M^+ and NM seeds based on using a t-test ($p < 0.05$); n.s. means 'not significant'. **e** Expression of RNA polymerase II complex and associated factors [50, 51] that mediate transcription including initiation, elongation and processing of transcripts in *Ae. arabicum* dry seed morphs. Red text indicates factor identified as NM-high DEG with expression ratio (NM / M^+) indicated. Note core NRPB1/2 transcript abundance and most factors are several-fold higher in NM seeds. **f** Seed maturation master regulators expression (RNAseq, *ABI3* also by qRT-PCR), oleosins, *NYE1* and *HSFA9* in dry M^+ and NM *Ae. arabicum* seeds. **g** Selected *Ae. arabicum LEA* expression in dry M^+ and NM seeds (RNAseq and qRT-PCR). The presented dehydrin is the putative ortholog of At4G39130. Error bars indicate mean ± SEM for qRT-PCR experiments. For the plotted RPKM values of single genes from the RNAseq data we used the result of the DEG detection pipeline (edgeR + NOISeq + DESeq2) as the indicator of significance

four H3, two H2B, three H2A, with transcript abundance of 1.5- to 4-fold higher in M^+ seeds as compared to NM seeds. Like the ribosomal protein DEGs, the transcript abundance of the *A. thaliana* histone homologs decreased during seed maturation (Fig. 8b). As with the ribosomal protein DEGs, the approximately 2-fold higher histone transcript abundance in M^+ seed could be due to faster drying of M^+ seeds during late maturation. However, as these DEGs represent only ca. 20% of the histones they may serve specific roles which define distinct processes in the dimorphic *Ae. arabicum* seeds. Differential expression of histone variants is linked to DNA replication and transcriptional regulation in response to developmental or environmental cues [38–40]. Histones are major components of chromatin, the protein-DNA complex involved in DNA packaging, chromatin remodeling and heterochromatin formation. *A. thaliana* seed maturation is characterized by nuclear size reduction and increased chromatin condensation (Fig. 8a) [41]. Chromatin condensation and heterochromatin formation involves the expression of specific histone H2B, H2A, and H3 variants [42–44], some of which we found to be *Ae. arabicum* DEGs with higher transcript abundance in M^+ compared to NM seeds (Fig. 8d). In contrast to those histone transcripts which are NM-low DEGs, genes which modify histones and facilitate transcription and RNA processing were found among the NM-high DEGs. Several genes encoding histone acyetyltransferases, deacetylases, and methyltransferases are among the NM-high DEGs, including for example putative orthologs of *A. thaliana HAC1* (At1g79000), *HAC12* (At1g16710), *HDA19* (At4g38130), *EFS* (At1g77300) and a SET7/9 family protein (At4g17080) (Fig. 8d, Additional file 3: Figure S4b), with HAC1, HAC12 and EFS putative orthologs being classified as transcriptional regulators by TAPscan (Additional file 2: Table S3). The NM-high DEGs of the de novo approach included *HAC1* (At1g79000), *HAC12* (At1g16710) and *EFS* (At1g77300), with HAC12 and EFS putative orthologs being classified as transcriptional regulators by TAPscan (Additional file 2: Table S4). These histone modifications are involved in regulating seed maturation and dormancy in response to environmental cues [43]. EFS for example is known to inhibit seed germination [45], HDA19 to repress seed maturation genes [46], and HAC1 to affect seed production and germination [47].

The absence of histone H2B mono-ubiquitination in the *A. thaliana hub1* and *hub2* mutants leads to altered chromatin remodeling and reduced seed dormancy [43, 44, 48], but the *HUB1/2* putative orthologs were not among the *Ae. arabicum* NM-high and NM-low DEGs. *HUB1/2* interacts with the FAcilitates Chromatin Transcription (FACT) complex, consisting of the SSRP1 and SPT16 proteins, for which mutants exhibit reduced seed production [49, 50]. The FACT complex is a histone chaperone that assists the progression of transcribing RNA polymerase II (RNAPII) on chromatin templates by destabilizing nucleosomes. The transcript abundance of the RNAPII catalytic subunit *NRPB1* increases during the late seed maturation of *A. thaliana* (Fig. 8b). Interestingly, putative *Ae. arabicum* putative orthologs of both RNAPII catalytic subunits were among the NM-high DEGs of the reference approach, with *NRPB1* approximately 10-fold and *NRPB2* 2-fold higher in NM seeds (Fig. 8d, e). *NRPB1* and *NRPB2* were also present with similar expression values in the NM-high DEGs of the de novo approach. Further to this, several key components of the RNAPII elongation complex [50–52] were also among the NM-high DEGs of both

approaches, including transcripts of subunits of almost all known factors known to be involved in regulating RNAPII-mediated transcription initiation, elongation and processing (Fig. 8e, Additional file 3: Figure S4b). In contrast to this, there were no such factors among the NM-low DEGs. Mutants for several of these key components are known for their developmental phenotypes including seed germination and dormancy traits [43, 48, 49, 52, 53]. Moreover, several other transcripts in downstream RNA processing were also among the NM-high DEGs of both approaches. Examples for this include factors with RNA binding, splicing and helicase activity (Additional file 3: Figure S4b). Among them is SMG7 (detected in both approaches) which is involved in nonsense-mediated mRNA decay (NMD) and regulates seed number in *B. napus* [54]. Taken together, these findings support the view that the transcriptome of NM seeds seems to be geared towards transcription which is important for dormancy and persistence. In contrast to this, seed maturation of M^+ seeds lead to a dry seed transcriptome in which translation is most dominant and is also most important during germination.

Dimorphic *Ae. arabicum* seeds differ in their maturation programmes

Seed-related processes were also amongst the BP terms significantly over-represented in the DEGs (GO-bias list), with the terms embryo development, fruit development, seed development and seed dormancy common to both the NM-high and NM-low DEG list (GO-presence list) (GO terms for each list can be found in Additional file 4: Table S8). However, the BP terms seed maturation, seed germination and seedling development were specific to the NM-high DEG GO-presence list. Additionally, the more specific BP terms positive regulation of seed maturation and negative regulation of seed germination were also identified in the NM-high DEG list. On the other hand, the term seed oil body biogenesis was only identified in the NM-down DEG GO-presence list. Thus, it appears that the M^+ and NM seed morphs differ in their expression of genes which determine seed traits during maturation. Seed maturation is associated with abscisic acid (ABA) regulated storage reserve accumulation such as oil (triacylglycerol, TAG) which requires gene expression [33, 55–58]. To achieve this fatty acid and TAG biosynthesis genes encoding proteins such as long chain acyl-CoA synthetase (LACS) and acyl-CoA:diacylglycerol acyltransferase (DGAT) are upregulated during *A. thaliana* seed maturation [59]. The TAGs are then transferred and accumulated into oil bodies which are covered on their surface with oleosins. Oleosins are the most abundant proteins found in the seed proteomes of oilseeds [57, 58]. Oleosin gene expression is also upregulated during *A. thaliana* seed maturation (Fig. 8b), but transcript abundances subsequently decline at the end of late maturation [57]. Their roles include to control oil body dynamics, size, and total oil accumulation during seed maturation. Interestingly, while putative orthologs of *A. thaliana LACS7*, *DGAT1*, a fatty acid alcohol dehydrogenase and a lipid transporter are among the NM-high DEGs of the reference approach (Additional file 3: Figure S4b), two oleosin homologs, *OLE2* and *OLE3*, are among the NM-low DEGs (Fig. 8f). In the de novo approach, putative orthologs of *LACS7* and *OLE2* are present among the NM-high and NM-low DEGs respectively, while the *DGAT1* putative ortholog was not detected as DEG and no *OLE3* homolog could be identified. That oleosin and TAG biosynthesis genes are in distinct DEG groups may either be due to distinct regulation during late seed maturation with TAG biosynthesis still up while oleosin expression is declining, or due to more profound differences between the dimorphic seeds in their maturation processes.

Four master regulators of seed maturation have been identified in *A. thaliana*: *ABSCISIC ACID INSENSITIVE3* (*ABI3*, At3g24650), *FUSCA3* (*FUS3*, At3g26790), *LEAFY COTYLEDON2* (*LEC2*, At1g28300), and *LEAFY COTYLEDON1* (*LEC1*, At1g21970) [33, 59, 60]. Whilst *LEC1* encodes the HAB3 subunit of a CCVAAT-box binding TF, ABI3, FUS3, and LEC2 are TFs with a B3 DNA binding domain. Corresponding TF classification was detected in the *Ae. arabicum* putative orthologs using TAPscan (Additional file 2: Table S3). In the de novo approach, orthologs of the ABI3/VP1 TFs *ABI3* and *FUS3* could be identified, with only *FUS3* being identified by TAPscan, probably because of the shorter length of the transcriptome based protein (577aa) vs. the reference based one (701aa) (Additional file 2: Table S4). These four master regulators control seed maturation including fatty acid and TAG biosynthesis, as well as oleosin expression and oil body formation. Enhancement of fatty acid and TAG biosynthesis by these master regulators is achieved, at least in part, by interaction of the *WRINKLED1* (*WRI1*, At3g54320) TF of the AP2/EREBP family [56, 58–61]. The temporal transcript patterns of these genes during *A. thaliana* seed maturation is depicted in Fig. 8c. Consistent with the *Ae. arabicum* fatty acid and TAG biosynthesis genes being among the NM-high DEGs, the putative *Ae. arabicum ABI3* ortholog is among the NM-high DEGs in the reference approach, with a putative *WRI1* ortholog also tending towards higher expression in NM seed (Fig. 8f). It should be noted that the *WRI1* transcript (TR24803|c0_g1_i1) is not represented by a gene model in the current genome version, demonstrating that occasionally the de novo transcriptome approach might outcompete the genomic approach. However, *FUS3* and

LEC1 are expressed roughly equal in dry M^+ and NM seeds (Fig. 8f). Also, if earlier drying of M^+ seeds is the only difference compared to NM seeds, *WRI1* and *ABI3* should be among the NM-low DEGs because their transcript abundances decline in *A. thaliana* during late maturation (Fig. 8c). It therefore seems that M^+ seeds not only dry out earlier, but also mature faster as compared to NM seeds. That M^+ seed maturation is faster is further supported by the finding that the *Ae. arabicum* NM-low DEG list of the reference approach contains the putative orthologs of *NON-YELLOWING1/STAY-GREEN1* (*NYE1/SGR1*, At4g22920), *HEAT SHOCK TRANSCRIPTION FACTOR9* (*HSFA9*, At5g54070) and of several Late Embryogenesis Abundant (LEA) protein genes which are upregulated during *A. thaliana* seed maturation (Fig. 8b, c) and are among the NM-low DEGs (Fig. 8f, g). The same findings were made using the de novo approach except that the *HSFA9* was not in the NM-low DEG list, but only trended towards lower expression in NM seeds. Efficient chlorophyll degradation during late seed maturation, in part mediated by the NYE1 protein, is critical for seed quality, longevity (storability), dormancy and germination properties [62]. During seed maturation, ABI3, through HSFA9, induces the accumulation of a subset of heat shock proteins (HSP) that contribute to seed longevity by protecting protein molecules and structures in the dry state [33, 63]. Among the *Ae. arabicum* DEGs, there are indeed *HSF9* and two other HSFs and several HSPs, but different HSPs are expressed in either a NM-low or a NM-high specific manner (Fig. 8f, Additional file 3: Figure S4b). A more distinctive pattern was obtained for the LEA proteins which were primarily found among the *Ae. arabicum* NM-low DEGs (Fig. 8g), supporting the view that M^+ seeds may mature faster and that M^+ and NM seeds may differ in their longevity.

Accumulation of LEA proteins is a landmark of seed maturation and several accumulate only during late maturation drying [33]. The 51 LEA protein encoding genes identified in *A. thaliana* cluster into 9 groups including LEA_1 to LEA_5, Seed Maturation Proteins (SMP) and dehydrins [64]. In the reference approach we found 13 putative LEA orthologs from all these groups in the *Ae. arabicum* NM-low and only two in the NM-high DEGs list (Fig. 8f, Additional file 3: Figure S4b). In the de novo approach, six LEA homologs were amongst the NM-low and only one in the NM-high DEGs list. The cumulative LEA transcript abundances were higher in M^+ compared to NM seeds, and the known most abundant LEA genes followed this pattern (Fig. 8f). Among them are the putative orthologs of *A. thaliana* LEA_1 *LEA76* (At5g06760), LEA_4 (At3g15670), LEA_5 *EM6* (At2g40170), the SMP *RAB28*, and dehydrins which are also most abundant in mature *A. thaliana* seeds [65]. The *A. thaliana* mutant *em6–1* is altered in seed hydration and desiccation tolerance during seed maturation [66]. LEA proteins are highly hydrophilic and intrinsically unstructured, and act by protecting proteins and enzyme activities in the desiccated state which, together with HSPs, may lead to maintaining seed longevity during dry storage [33, 63, 64]. In addition to their higher LEA transcript abundance (Fig. 8g), in both approaches, M^+ seeds also have higher transcript abundances of enzymes involved in detoxifying Reactive Oxygen Species (ROS) such as superoxide dismutase (SOD) and glutathione-*S*-transferase (GST) (Additional file 3: Figure S4b). ROS are produced during a number of seed related processes: with potentially deleterious effects during seed maturation, desiccation, ageing and germination; but also acting by controlling dormancy and germination [63, 67, 68]. Thus, the two seed morphs may differ in mechanisms by which seed longevity and dormancy are established and regulated. Whilst the GO term 'hormone metabolic process' was amongst 137 BP GO terms significantly under-represented in the reference approach DEGs (GO-bias list), the putative orthologs of genes involved in ABA and gibberellin signaling (*XERICO*), ethylene biosynthesis (*S-adenosylmethionine synthetase*, *SAMS3*) and signaling (*EIN3-binding F-box protein*, *EBF1*), and auxin and brassinosteroid signaling (*Auxin Response Factor 2*, *ARF2*) are amongst the DEGs (Additional file 3: Figure S4b), with all but *XERICO* also being among the de novo approach DEGs. The presence of these genes is consistent with previously observed differences in seed development and dormancy (described further in Additional file 3: Figure S5).

Conclusions

RNA-seq analysis of *Ae. arabicum* M^+ and NM dry seed transcriptomes using either a de novo assembled transcriptome approach or reference genome guided approach showed only a modest overlap in the DEGs identified, but much greater consistency in the GO terms identified. Thus, using global functional annotations such as GO terms, the de novo assembled transcriptome approach would result in similar conclusions being drawn from the data compared to the reference genome approach. Studying seeds, which are a well characterized biological system, allowed us to identify many well studied genes and put them into context using both a de novo assembled transcriptome approach and a reference genome guided approach. This highlights the potential usefulness of de novo transcriptome assembly in the study of species that do not have a reference genome. With the decreasing costs of RNA-seq one should aim for using at least three replicates, potentially bridging the gap between a de novo assembly and reference genome guided approach even further. However, our

results also highlight the limitations of de novo transcriptome analysis. Namely, if the goal is to pinpoint the DEGs underlying a trait, then reference based assemblies perform better.

Major differences in the seed morph transcriptomes were highlighted by GO analysis. In particular, genes associated with translation and histone assembly were more abundant in the less dormant M^+ dry seed, whereas genes associated with transcription and mRNA processing were more abundant in the more dormant NM dry seed. By putting the M^+ and NM dry seed transcriptomes in the context of transcriptomes from developing and maturing *A. thaliana* seeds, it was indicated that M^+ seeds may both desiccate earlier (M^+ has higher histone and ribosomal protein expression) and mature faster than NM seeds (compared to NM, M^+ seed have higher expression of genes that increase with maturation, such as homologs of LEAs, *NYE1* and *HSFA9*, and lower expression of genes that decrease during maturation such as *ABI3* and *WRI1*). The differences identified align with the known development and germination behaviour of the two seed morphs, but hint at other differences such as in longevity mechanisms (LEAs, ROS detoxification). However, the difference in longevity of M^+ and NM seed are so far unknown. It would also be valuable to study how the differences in dry seed lead to differences in transcription and germination physiology in the imbibed dimorphic seeds.

Methods

Plant material and RNA extraction

Aethionema arabicum (L.) A.DC. accession 0000309 (collected from Turkey and obtained from Kew's Millennium Seed Bank, UK) and ES1020 (collected from Turkey and obtained from Eric Schranz, Wageningen) [3] plants were grown on soil under long-day conditions (16 h light/20°C and 8 h dark/18°C). Freshly matured seeds from dehiscent (harboring M^+ seeds) and indehiscent (harboring NM seeds) fruits derived from several plants were harvested. Two replicates of 20 mg fresh dry M^+ and NM seeds, resulting in four samples in total, were pulverized in liquid N_2 using a mortar and pestle. RNA extraction was performed according to [69]. RNA integrity was checked by gel electrophoresis (Additional file 3: Figure S6) followed by quantity and purity determination with a Nanodrop spectrophotometer ND-1000 (Peqlab) showing sufficiently low levels of degradation for RNAseq and OD ratios of at least 2 (260/280 nm) and 1.8 (260/230 nm).

RNA-seq library preparation and sequencing

RNA libraries were prepared following instructions of the TruSeq™ RNA library prep kit (Illumina) using oligo-dT-based mRNA selection. Libraries were sequenced using a HiSeq-2000 sequencer (Illumina) generating 100 bp single-end reads.

RNA-seq data trimming and filtering

The raw RNA sequences were processed with trimmomatic [15] (ILLUMINACLIP:adaptors:2:20:8, SLIDINGWINDOW:4:15, TRAILING:15, HEADCROP:12, MINLENGTH:20) to remove poor quality stretches and adaptors. Poly-A and Poly-T tails were removed using PrinSeq [16]. To reduce the complexity of the dataset prior to mapping our reads to the genome/transcriptome rRNA, mitochondrial and chloroplast sequences were filtered. Since *Ae. arabicum* sequences for rRNA, mitochondria and chloroplast were not available in public repositories, sequences from closely related and well annotated *A. thaliana* were used. GSNAP version 2016–11-07 [17] with default settings was used to map the reads against the chloroplast (GenBank: AP000423.1), mitochondria (GenBank: Y08501.2) and rRNA (GenBank:X52320.1) sequences from *A. thaliana*.

De novo transcriptome assembly

Prior to the de novo transcriptome assembly, redundant duplicate reads, i.e. reads with the exact same length and sequence, were removed since they might constitute PCR artefacts. The trimmed, filtered and de-duplicated reads were assembled into a transcriptome using Trinity [14] with default settings. For each isoform group, the longest transcript was chosen as representative and its longest open reading frame was translated into protein using a custom python script.

Evaluation of assembly and comparison to genome

Genome scaffolds and accompanying GFF file of *Ae. arabicum* genome version 2.5 [5] was obtained from CoGe (genome id23428, https://genomevolution.org/coge/OrganismView.pl?gid=23428). The CDS of each gene was translated into proteins using the R package biostrings version 2.32.0. The completeness of the assembled transcriptome and the available genome of *Ae. arabicum* was evaluated using the Benchmarking Universal Single-Copy Orthologs tool BUSCO v3.0.1 [23] and their accompanying dataset of 1440 plant orthologs (embryophyta *odb9*). To investigate how well the assembled transcripts represented and paired up with the existing gene models from *Ae. arabicum* genome version 2.5, reciprocal BLAST (version 2.2.29+, [70]) searches were carried out. Reciprocal best hits (RBH) with a minimum query and subject coverage of 50% each were considered as a match and selected for comparison.

Read mapping and feature counting

Processed reads were mapped against the assembled transcriptome and the *Ae. arabicum* genome version 2.5 using GSNAP with default settings. Reads that mapped to multiple positions in the genome were discarded and only uniquely mapped reads were kept. Mapped reads per feature were counted using HTSeq-count (version 0.6.1 [25]) with the options "–s no –t gene –m union". For the transcriptome a custom GFF was generated with one feature for each transcript, while for the gene models the GFF mentioned above was used. The average coverage was calculated using the genome reference. The total amount of mapped reads (all libraries) for each gene was multiplied by the read length (83) and divided by gene length (Additional file 2: Table S1).

Differential gene expression analysis pipeline

Differentially expressed genes were identified using R [71] and the Bioconductor packages DESeq2 version 1.14.1 [19], edgeR version 3.16.5 [18] and NOISeq version 3.16.5 [20]. It is recommended to discard features with low counts for edgeR DEG analysis, so only genes with at least 10 read counts when summing up all the sample counts were selected for edgeR. Default parameters were used for DESeq2, edgeR (classic approach, "exactTest") and NOISeq with normalization method relative log expression for DESeq2, trimmed mean of M values for edgeR and RPKM for NOISeq. DESeq2 and edgeR make use of Benjamini-Hochberg [72] adjusted *p*-value (q-value) cut offs which were set to 0.001. For NOISeq, which uses probabilities of differential expression, a cutoff value of >0.9 was used. This is higher than the recommended 0.8 but has been shown to overlap well with experimental array data, representing a conservative (specific) selection of DEGs [28]. The overlap (strict consensus) of the three packages' outputs was used for further analysis.

Principal component analysis of expression values

To compare the feature counts of the two approaches (de novo transcriptome and reference genome), PCAs were carried out using the built in R package prcomp. RPKM normalized expression values of the 6745 paired de novo transcripts and reference genes were calculated and used as input, as well as the 561 DEGs identified by both approaches.

Annotation and GO-bias

The transcripts of the genome and assembled transcriptome were blasted against the nr database of NCBI (nucleotide release 13-05-2015), UniProtKB/Swiss-Prot (protein release 10–2015) and TAIR 10 (proteins release 20,110,103). GO-terms were retrieved using BLAST2GO version 2.5 [21] in combination with the NCBI nr blast results. GO-bias, i.e. over/under-representation of GO-terms in defined sets of genes as compared to all genes, was calculated as in [73] using Fisher's exact test with FDR correction [72]. Wordle (www.wordle.net) was used to build word clouds, with word height proportional to –log10(q-value), significantly over-represented GO-terms colored green ($q <= 0.0001$ dark green, $q > 0.0001$ light green) and under-represented GO-terms colored red ($q <= 0.0001$ dark red, $q > 0.0001$ light red). Transcripts of the genome and assembled transcriptome were screened for TAPs using the TAPscan pipeline [26].

qRT-PCR analysis

For technical as well as biological validation of RNA-seq derived gene expression data, RNA was extracted from separate batches of dry fresh mature M^+ and NM seeds (five biological replicates each) as described above, and quantitative RT-PCR analysis of selected candidate genes was performed as previously described [69]. As normalization factor the geometric mean of three reference genes, *Ae. arabicum* putative orthologs of *ACTIN2* (*ACT2*, AA26G00546), *POLYUBIQUITIN10* (*UBQ10*, AA6G00219) and *ANAPHASE-PROMOTING COMPLEX2* (*APC2*, AA61G00327) was used, which was found to show comparable stable expression in M^+ and NM seeds (Additional file 3: Figure S7). Primers for qRT-PCR are listed in (Additional file 2: Table S9).

Additional files

Additional file 1: de novo transcriptome assembly. The 34,784 longest gene sequences from each Trinity gene cluster. (FA 28331 kb)

Additional file 2: Gene coverage calculation (**Table S1**), reciprocal best BLAST paring (**Table S2**), full annotation and RPKM tables for the genome method (**Table S3**) and transcriptome method (**Table S4**). Comparison of abundance of transcripts (genome method vs. transcriptome method) belonging to: the 5584 GO terms shared between both methods (**Table S5**); or the 1663 overlapping GO terms of the DEG sets (**Table S6**). **Table S7** shows a summary of significantly under- and over-represented GO terms associated with DEG lists. **Table S9** contains a list of primers used for qRT-PCR. (XLSX 10033 kb)

Additional file 3: **Figure S1.** PCA of RPKM values for 6745 paired transcripts (identified in both genome and transcriptome methods) by method and morphotype. **Figure S2.** RPKM levels (reference genome approach) of the overlapping DEGs as well as of the non-overlapping DEGs called by NOISeq, edgeR and DESeq2. **Figure S3.** Expression of selected DEGs measured by qRT-PCR. **Figure S4.** showing abundances of *Ae. arabicum* ribonucleoprotein transcripts (a) and transcripts from selected gene categories (b) and **Figure S5.** showing the pattern of expression of select hormonal signaling related genes during *A. thaliana* seed maturation. Assessment of RNA integrity and purity (**Figure S6.**) and validation of reference genes used for qRT-PCR normalization (**Figure S7.**) (DOCX 2738 kb)

Additional file 4: **Table S8.** Excel document containing GO term analysis output for BP, CC and MF classes and all DEG lists. (XLSX 347 kb)

Abbreviations
ABA: Abscisic acid; ABI3: ABSCISIC ACID INSENSITIVE3; BP: Biological Process; BUSCO: Benchmarking Universal Single-Copy Orthologs; CC: Cellular Component; CDS: Coding sequence; DEGs: Differentially expressed genes; DEH: Dehiscent; DGAT: Acyl-CoA:diacylglycerol acyltransferase; FACT: FAcilitates Chromatin Transcription; FDR: False Discovery Rate; GFF: General feature format; GO: Gene ontology; HSFA9: HEAT SHOCK TRANSCRIPTION FACTOR9; HSP: Heat shock proteins; IND: Indehiscent; LACS: Long chain acyl-CoA synthetase; LEA: Late Embryogenesis Abundant; LEC1: LEAFY COTYLEDON1; LEC2: LEAFY COTYLEDON2; M^+: Mucilaginous; MF: Molecular function; NM: Non-mucilaginous; NYE1/SGR1: NON-YELLOWING1/STAY-GREEN1; PCA: Principal component analysis; RNAPII: RNA Polymerase II; RNA-seq: RNA-sequencing; ROS: Reactive Oxygen Species; RPKM: Reads Per Kilobase per Million mapped reads; rRNA: ribosomal RNA; SMP: Seed Maturation Proteins; TAG: Triacylglycerol; TAPs: Transcription Associated Proteins; WRI1: WRINKLED1

Acknowledgements
We thank the members of the SeedAdapt consortium for useful discussions on the biology of *Ae. arabicum*.

Funding
This work is part of the ERA-CAPS SeedAdapt consortium project (www.seedadapt.eu) and was supported by the Deutsche Forschungsgemeinschaft (grant no. RE 1697/8–1 to S.A.R.); by the Netherlands Organization for Scientific Research (grant no. 849.13.004 to M.E.S.); by the Biotechnology and Biological Sciences Research Council (grant nos. BB/M00192X/1 and BB/M000583/1 to G.L.-M.); and by a Natural Environment Research Council (NERC) Doctoral Training Partnership studentship to W.A. (grant no. NE/L002485/1).

Availability of data and materials
Single-ended Illumina raw reads from this study were uploaded to the NCBI Sequence Read Archive (SRA) and can be found under BioProject PRJNA413671 (https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA413671). The following accession numbers correspond to each one of the samples: SRR6157646 (Indehiscent: NM seed rep 1, NM1), SRR6157647 (Indehiscent: NM seed rep 2, NM2), SRR6157648 (dehiscent: M^+ seed rep 1, M^+ 1), SRR6157649 (dehiscent: M^+ seed rep 2, M^+ 2).

Authors' contributions
KG, MES, JAH and SK prepared *Ae. arabicum* material and RNA. KG and SK performed qRT-PCR. JCP and PPE synthesized sequencing libraries. JOC, PKIW, NF-P, GL-M, SAR, WA and KG prepared figures and wrote the manuscript. KG and WA provided plant and seed images, mass and moisture content data. PKIW, NF-P, KKU, KB, KG and SAR assembled RNA-seq data and analyzed data. GL-M and JOC provided biological interpretation of RNA-seq analysis. All authors read and approved the manuscript.

Ethics approval and consent to participate
The source of the *Ae. arabicum* seeds were accessions 0000309 (obtained from Kew's Millennium Seed Bank) and ES1020 (obtained from Eric Schranz, Wageningen) [3]. This study complies with institutional, national, and international guidelines.

Consent for publication
Not applicable

Competing interests
The authors declare that they have no competing interests.

Publisher's Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details
[1]Plant Cell Biology, Faculty of Biology, University of Marburg, 35043 Marburg, Germany. [2]School of Biological Sciences, Royal Holloway University of London, Egham, Surrey TW20 0EX, UK. [3]Biosystematics Group, Wageningen University, Wageningen 6708 PB, The Netherlands. [4]Department of Horticulture, Michigan State University, East Lansing, MI 48864, USA. [5]Division of Biological Sciences, University of Missouri, Columbia, MO 65211, USA. [6]Laboratory of Growth Regulators, Centre of the Region Haná for Biotechnological and Agricultural Research, Palacký University and Institute of Experimental Botany, Academy of Sciences of the Czech Republic, 78371 Olomouc, Czech Republic. [7]BIOSS Centre for Biological Signalling Studies, University of Freiburg, Freiburg, Germany. [8]Present Address: Max Planck Institute for Evolutionary Biology, August-Thienemann-Straße 2, 24306 Ploen, Germany.

Received: 16 May 2018 Accepted: 14 January 2019
Published online: 30 January 2019

References
1. Brautigam A, Gowik U. What can next generation sequencing do for you? Next generation sequencing as a valuable tool in plant research. Plant Biol (Stuttg). 2010;12(6):831–41.
2. Mohammadin S, Peterse K, van de Kerke SJ, Chatrou LW, Donmez AA, Mummenhoff K, Pires JC, Edger PP, Al-Shehbaz IA, Schranz ME. Anatolian origins and diversification of Aethionema, the sister lineage of the core Brassicaceae. Am J Bot. 2017;104(7):1042–54.
3. Lenser T, Graeber K, Cevik OS, Adiguzel N, Donmez AA, Grosche C, Kettermann M, Mayland-Quellhorst S, Merai Z, Mohammadin S, et al. Developmental control and plasticity of fruit and seed dimorphism in Aethionema arabicum. Plant Physiol. 2016;172(3):1691–707.
4. Arshad W, Sperber K, Steinbrecher T, Nichols B, Jansen VAA, Leubner-Metzger G, Mummenhoff K. Dispersal biophysics and adaptive significance of dimorphic diaspores in the annual Aethionema arabicum (Brassicaceae). New Phytol. 2019;221(3):1434–46. https://doi.org/10.1111/nph.15490. Epub 2018 Oct 25.
5. Haudry A, Platts AE, Vello E, Hoen DR, Leclercq M, Williamson RJ, Forczek E, Joly-Lopez Z, Steffen JG, Hazzouri KM, et al. An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. Nat Genet. 2013;45(8):891–U228.
6. t Hoen PA, Ariyurek Y, Thygesen HH, Vreugdenhil E, Vossen RH, de Menezes RX, Boer JM, van Ommen GJ, den Dunnen JT. Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. Nucleic Acids Res. 2008;36(21):e141.
7. Wilhelm BT, Marguerat S, Watt S, Schubert F, Wood V, Goodhead I, Penkett CJ, Rogers J, Bahler J. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. Nature. 2008;453(7199):1239–43.
8. Gibbons JG, Janson EM, Hittinger CT, Johnston M, Abbot P, Rokas A. Benchmarking next-generation transcriptome sequencing for functional and evolutionary genomics. Mol Biol Evol. 2009;26(12):2731–44.
9. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szczesniak MW, Gaffney DJ, Elo LL, Zhang X, et al. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016;17(1):13.
10. Gongora-Castillo E, Buell CR. Bioinformatics challenges in de novo transcriptome assembly using short read sequences in the absence of a reference genome sequence. Nat Prod Rep. 2013;30(4):490–500.
11. Li B, Fillmore N, Bai Y, Collins M, Thomson JA, Stewart R, Dewey CN. Evaluation of de novo transcriptome assemblies from RNA-Seq data. Genome Biol. 2014;15(12):553.
12. Surget-Groba Y, Montoya-Burgos JI. Optimization of de novo transcriptome assembly from next-generation sequencing data. Genome Res. 2010;20(10):1432–40.
13. Zhao QY, Wang Y, Kong YM, Luo D, Li X, Hao P. Optimizing de novo transcriptome assembly from short-read RNA-Seq data: a comparative study. BMC Bioinformatics. 2011;12(Suppl 14):S2.
14. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29(7):644–52.
15. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–20.
16. Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. Bioinformatics. 2011;27(6):863–4.
17. Wu TD, Nacu S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics. 2010;26(7):873–81.
18. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139–40.
19. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.
20. Tarazona S, Garcia-Alcalde F, Dopazo J, Ferrer A, Conesa A. Differential expression in RNA-seq: a matter of depth. Genome Res. 2011;21(12):2213–23.

21. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics. 2005;21(18):3674–6.
22. Liu Y, Zhou J, White KP. RNA-seq differential expression studies: more sequence or more replication? Bioinformatics. 2014;30(3):301–4.
23. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31(19):3210–2.
24. Berardini TZ, Reiser L, Li D, Mezheritsky Y, Muller R, Strait E, Huala E. The Arabidopsis information resource: making and mining the "gold standard" annotated reference plant genome. Genesis. 2015;53(8):474–85.
25. Anders S, Pyl PT, Huber W. HTSeq–a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015;31(2):166–9.
26. Wilhelmsson PKI, Muhlich C, Ullrich KK, Rensing SA. Comprehensive genome-wide classification reveals that many plant-specific transcription factors evolved in streptophyte algae. Genome Biol Evol. 2017;9(12):3384–97.
27. Zhang ZH, Jhaveri DJ, Marshall VM, Bauer DC, Edson J, Narayanan RK, Robinson GJ, Lundberg AE, Bartlett PF, Wray NR, et al. A comparative study of techniques for differential expression analysis on RNA-Seq data. PLoS One. 2014;9(8):e103207.
28. Perroud PF, Haas FB, Hiss M, Ullrich KK, Alboresi A, Amirebrahimi M, Barry K, Bassi R, Bonhomme S, Chen H, et al. The Physcomitrella patens gene atlas project: large scale RNA-seq based expression data. Plant J. 2018;95:168.
29. Le BH, Cheng C, Bui AQ, Wagmaister JA, Henry KF, Pelletier J, Kwong L, Belmonte M, Kirkbride R, Horvath S, et al. Global analysis of gene activity during Arabidopsis seed development and identification of seed-specific transcription factors. Proc Natl Acad Sci U S A. 2010;107(18):8063–70.
30. Xiang D, Venglat P, Tibiche C, Yang H, Risseeuw E, Cao Y, Babic V, Cloutier M, Keller W, Wang E, et al. Genome-wide analysis reveals gene expression and metabolic network dynamics during embryo development in Arabidopsis. Plant Physiol. 2011;156(1):346–56.
31. Nakabayashi K, Okamoto M, Koshiba T, Kamiya Y, Nambara E. Genome-wide profiling of stored mRNA in Arabidopsis thaliana seed germination: epigenetic and genetic regulation of transcription in seed. Plant J. 2005; 41(5):697–709.
32. Graeber K, Nakabayashi K, Leubner-Metzger G. Seed development and germination. In: Thomas B, Murray BG, Murphy DJ, editors. Encylopedia of applied plant sciences, vol. 1. Waltham: Academic Press; 2017. p. 483–9.
33. Leprince O, Pellizzaro A, Berriri S, Buitink J. Late seed maturation: drying without dying. J Exp Bot. 2017;68(4):827–41.
34. Fei H, Tsang E, Cutler AJ. Gene expression during seed maturation in Brassica napus in relation to the induction of secondary dormancy. Genomics. 2007;89(3):419–28.
35. Bai B, Peviani A, van der Horst S, Gamm M, Snel B, Bentsink L, Hanson J. Extensive translational regulation during seed germination revealed by polysomal profiling. New Phytol. 2017;214(1):233–44.
36. Galland M, Rajjou L. Regulation of mRNA translation controls seed germination and is critical for seedling vigor. Front Plant Sci. 2015;6:284.
37. Tatematsu K, Kamiya Y, Nambara E. Co-regulation of ribosomal protein genes as an indicator of growth status: comparative transcriptome analysis on axillary shoots and seeds in Arabidopsis. Plant Signal Behav. 2008;3(7):450–2.
38. Xiao J, Jin R, Wagner D. Developmental transitions: integrating environmental cues with hormonal signaling in the chromatin landscape in plants. Genome Biol. 2017;18:88.
39. Bonisch C, Hake SB. Histone H2A variants in nucleosomes and chromatin: more or less stable? Nucleic Acids Res. 2012;40(21):10719–41.
40. Boisnard-Lorig C, Colon-Carmona A, Bauch M, Hodge S, Doerner P, Bancharel E, Dumas C, Haseloff J, Berger F. Dynamic analyses of the expression of the HISTONE::YFP fusion protein in arabidopsis show that syncytial endosperm is divided in mitotic domains. Plant Cell. 2001;13(3):495–509.
41. van Zanten M, Koini MA, Geyer R, Liu Y, Brambilla V, Bartels D, Koornneef M, Fransz P, Soppe WJ. Seed maturation in Arabidopsis thaliana is characterized by nuclear size reduction and increased chromatin condensation. Proc Natl Acad Sci U S A. 2011;108(50):20219–24.
42. Yelagandula R, Stroud H, Holec S, Zhou K, Feng S, Zhong X, Muthurajan UM, Nie X, Kawashima T, Groth M, et al. The histone variant H2A.W defines heterochromatin and promotes chromatin condensation in Arabidopsis. Cell. 2014;158(1):98–109.
43. Footitt S, Muller K, Kermode AR, Finch-Savage WE. Seed dormancy cycling in Arabidopsis: chromatin remodelling and regulation of DOG1 in response to seasonal environmental signals. Plant J. 2015;81(3):413–25.
44. Liu Y, Koornneef M, Soppe WJJ. The absence of histone H2B monoubiquitination in the *Arabidopsis hub1 (rod4)* mutant reveals a role for chromatin remodeling in seed dormancy. Plant Cell. 2007;19:433–44.
45. Lee N, Kang H, Lee D, Choi G. A histone methyltransferase inhibits seed germination by increasing PIF1 mRNA expression in imbibed seeds. Plant J. 2014;78(2):282–93.
46. Zhou Y, Tan B, Luo M, Li Y, Liu C, Chen C, Yu CW, Yang S, Dong S, Ruan J, et al. HISTONE DEACETYLASE19 interacts with HSL1 and participates in the repression of seed maturation genes in Arabidopsis seedlings. Plant Cell. 2013;25(1):134–48.
47. Heisel TJ, Li CY, Grey KM, Gibson SI. Mutations in HISTONE ACETYLTRANSFERASE1 affect sugar response and gene expression in Arabidopsis. Front Plant Sci. 2013;4:245.
48. Graeber K, Nakabayashi K, Miatton E, Leubner-Metzger G, Soppe WJ. Molecular mechanisms of seed dormancy. Plant Cell Environ. 2012;35:1769–86.
49. Lolas IB, Himanen K, Gronlund JT, Lynggaard C, Houben A, Melzer M, Van Lijsebettens M, Grasser KD. The transcript elongation factor FACT affects Arabidopsis vegetative and reproductive development and genetically interacts with HUB1/2. Plant J. 2010;61(4):686–97.
50. Antosz W, Pfab A, Ehrnsberger HF, Holzinger P, Kollen K, Mortensen SA, Bruckmann A, Schubert T, Langst G, Griesenbeck J, et al. The composition of the Arabidopsis RNA polymerase II transcript elongation complex reveals the interplay between elongation and mRNA processing factors. Plant Cell. 2017;29(4):854–70.
51. Wang Y, Ma H. Step-wise and lineage-specific diversification of plant RNA polymerase genes and origin of the largest plant-specific subunits. New Phytol. 2015;207(4):1198–212.
52. Eom H, Park SJ, Kim MK, Kim H, Kang H, Lee I. TAF15b, involved in the autonomous pathway for flowering, represses transcription of FLOWERING LOCUS C. Plant J. 2018;93(1):79–91.
53. Liu Y, Geyer R, van Zanten M, Carles A, Li Y, Horold A, van Nocker S, Soppe WJ. Identification of the Arabidopsis REDUCED DORMANCY 2 gene uncovers a role for the polymerase associated factor 1 complex in seed dormancy. PLoS One. 2011;6(7):e22241.
54. Li S, Chen L, Zhang L, Li X, Liu Y, Wu Z, Dong F, Wan L, Liu K, Hong D, et al. BnaC9.SMG7b functions as a positive regulator of the number of seeds per silique in Brassica napus by regulating the formation of functional female gametophytes. Plant Physiol. 2015;169(4):2744–60.
55. Baud S, Boutin J-P, Miquel M, Lepiniec L, Rochat C. An integrated overview of seed development in Arabidopsis thaliana ecotype WS. Plant Physiol Biochem. 2002;40:151–60.
56. Baud S, Wuilleme S, To A, Rochat C, Lepiniec L. Role of WRINKLED1 in the transcriptional regulation of glycolytic and fatty acid biosynthetic genes in Arabidopsis. Plant J. 2009;60(6):933–47.
57. Miquel M, Trigui G, d'Andrea S, Kelemen Z, Baud S, Berger A, Deruyffelaere C, Trubuil A, Lepiniec L, Dubreucq B. Specialization of oleosins in oil body dynamics during seed development in Arabidopsis seeds. Plant Physiol. 2014;164(4):1866–78.
58. Ruuska SA, Girke T, Benning C, Ohlrogge JB. Contrapuntal networks of gene expression during Arabidopsis seed filling. Plant Cell. 2002;14(6):1191–206.
59. Baud S, Lepiniec L. Physiological and developmental regulation of seed oil production. Prog Lipid Res. 2010;49(3):235–49.
60. Devic M, Roscoe T. Seed maturation: simplification of control networks in plants. Plant Sci. 2016;252:335–46.
61. Cernac A, Andre C, Hoffmann-Benning S, Benning C. WRI1 is required for seed germination and seedling establishment. Plant Physiol. 2006;141(2):745–57.
62. Li Z, Wu S, Chen J, Wang X, Gao J, Ren G, Kuai B. NYEs/SGRs-mediated chlorophyll degradation is critical for detoxification during seed maturation in Arabidopsis. Plant J. 2017;92(4):650–61.
63. Sano N, Rajjou L, North HM, Debeaujon I, Marion-Poll A, Seo M. Staying alive: molecular aspects of seed longevity. Plant Cell Physiol. 2016;57(4):660–74.
64. Hundertmark M, Hincha DK. LEA (late embryogenesis abundant) proteins and their encoding genes in Arabidopsis thaliana. BMC Genomics. 2008;9:118.
65. Kimura M, Nambara E. Stored and neosynthesized mRNA in Arabidopsis seeds: effects of cycloheximide and controlled deterioration treatment on the resumption of transcription during imbibition. Plant Mol Biol. 2010;73(1–2):119–29.
66. Manfre AJ, LaHatte GA, Climer CR, Marcotte WR Jr. Seed dehydration and the establishment of desiccation tolerance during seed maturation is altered in the Arabidopsis thaliana mutant atem6-1. Plant Cell Physiol. 2009;50(2):243–53.
67. Bailly C. Active oxygen species and antioxidants in seed biology. Seed Sci Res. 2004;14:93–107.

68. Linkies A, Leubner-Metzger G. Beyond gibberellins and abscisic acid: how ethylene and jasmonates control seed germination. Plant Cell Rep. 2012; 31(2):253–70.
69. Graeber K, Linkies A, Wood AT, Leubner-Metzger G. 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. 2011;23(6):2045–63.
70. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25(17):3389–402.
71. R: A language and environment for statistical computing. https://www.r-project.org/.
72. Benjamini Y, Hochberg Y. Controlling the false discovery rate - a practical and powerful approach to multiple testing. J R Stat Soc Ser B Methodol. 1995;57(1):289–300.
73. Widiez T, Symeonidi A, Luo C, Lam E, Lawton M, Rensing SA. The chromatin landscape of the moss *Physcomitrella patens* and its dynamics during development and drought stress. Plant J. 2014;79(1):67–81.
74. Winter D, Vinegar B, Nahal H, Ammar R, Wilson GV, Provart NJ. An "electronic fluorescent pictograph" browser for exploring and analyzing large-scale biological data sets. PLoS One. 2007;2(8):e718.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

BMC

3.5 Further applicability of this work

Amongst the inferred DEGs TAPs of interest were chosen and further investigated in the context of the moss *P. patens* (BSc-thesis of group member Marlies Peter). The DEG pipeline developed was applied and used with all further SeedAdapt RNA-seq data including more than 300 RNA-seq libraries.

3.6 SeedAdapt experiments and expression atlas

Within the framework of the SeedAdapt consortium (www.seedadapt.eu) an extensive experimental design was set up to investigate the adaptive plasticity of the dispersal unit (seed diaspore syndrome) of *Ae. arabicum*. This included the generation of multiple RNA-sequence libraries of the two seed morphs, as well as the whole indehiscent fruit, sampled from different accessions in different conditions. Seeds and fruits were taken from maternal plants of Turkish and Cyprus accession grown in temperatures of 20 and 25°C. These were then let to germinate at temperatures 9, 14, 20 and 24°C and sampled for sequencing after different time points ranging from 0 to 125h. The sample comparisons proposed by the SeedAdapt consortium resulted in 193 pairwise comparisons. These comparisons were carried out using the methods developed in the paper (chapter 3.2), resulting in DEG-lists as well as GO-term bias analysis for each of the 193 comparisons. For visualization and analytical simplification, the Expression Atlas framework developed by Fernandez-Pozo et al. [60] was adopted to the SeedAdapt data (Fig. 6). The cube designed

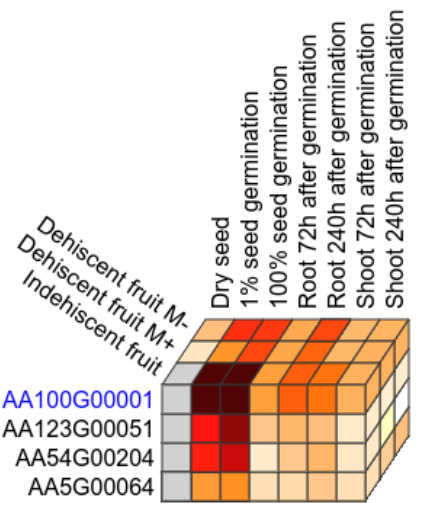


Figure 6. Example of the Expression Atlas framework employing SeedAdapt data.

interface enables easy look up of specific genes of interest and presents the expression values of the gene in the context of the whole experimental set up as well as showing other genes with correlating expression profiles. This resource is key in the ongoing interpretation and analysis carried out by the SeedAdapt consortium and will be made publicly available upon publication of the final consortium report.

Additional stand-alone experimental set ups were designed within the SeedAdapt framework to investigate the underlying molecular mechanisms of the plants fruit morph decision [61] as well as germination controlling factors [62]. The RNA-seq generated for these experiments were handled using the methods developed in the paper (3.2) and presented in the expression atlas application. These data, leading to one published and one submitted paper, contributed to significant findings with regards to their respective aims.

E. g. in Lenser et al. 2018 (Fig. 7), it could be demonstrated that it is a last-minute decision, occurring in the flowers of *Ae. arabicum*, that determines which seed morph is to develop [61]. It was also shown that

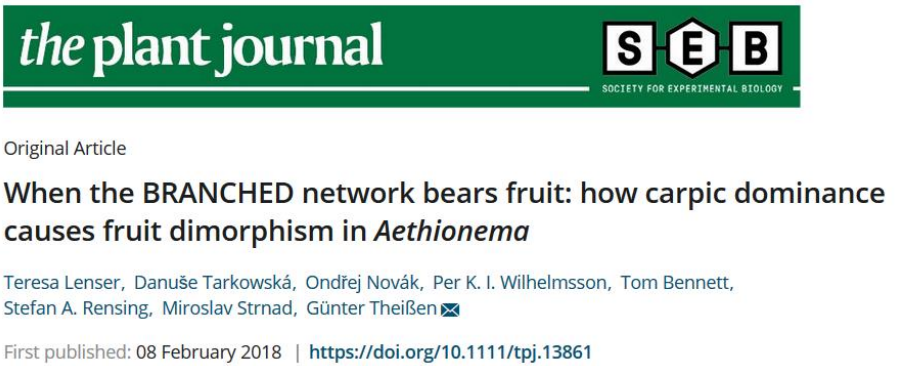


Figure 7. Lenser et al. 2018 published in the plant journal.

the decision making is made possible through the recruitment of the pre-existing shoot branching network involving the TAP (bHLH_TCP)

coding gene *BRC1* with *BRC1* expression level, in turn, being dependent on the ratios of auxin and cytokinin.

In Merai et al. (under revision) it was discovered that *Ae. arabicum* seeds of the Turkish accession germinated well under white light and dark conditions while the Cyprus accession germinated well in darkness but were strongly inhibited under white light conditions [62]. This photoblastic difference fits well with effects the different environments the plants inhabit have, with Cyprus accession colonizing more arid locations with higher light intensity which might be unfavorable for young seedlings. The expression profiles did not pinpoint any unique molecular switch mechanism responsible for the difference, though there were significant differences in respective accessions GA:ABA (gibberellin and abscisic acid) ratio. This works findings strongly suggests that *Ae. arabicum* could serve as a model plant for studying light-controlled germination in plants.

Chapter 4

TAPs and *Ae. arabicum*

4 TAPs and *Ae. arabicum*

The importance of transcription associated proteins, as orchestrators and key players in regulator networks, have been shown for many traits in many organisms. The SeedAdapt data resource, RNA-seq samples as well as the DEG comparisons, in combination with TAPscan makes it possible to investigate *Ae. arabicum* seed development in in the light of TAPs.

4.1 TAPs in *Ae. arabicum* seed development (bud, flower, fruit and seed)

In the first publication of the SeedAdapt consortium (Lenser at al 2016) [45] orthologs of eight *A. thaliana* fruit regulatory genes were investigated using quantitative reverse transcription-PCR (qRT-PCR) on indehiscent and dehiscent fruits. One gene that stood out from the rest was the ortholog of the INDEHISCENT (*IND*) gene (AT4G00120), *AearIND* AA32G00014. It was shown to have a 7-fold expression difference, being higher expressed in dehiscent fruits, suggesting it to be one of the key molecular mechanisms for establishing the dimorphism during *Ae. arabicum* fruit development. The *AearIND* protein model contains the helix-loop-helix domain and thus gets classified as a bHLH TF by TAPscan. In further studies, led by Lenser et al. [61] the plants decision making to produce the different seed morphs was investigated. No morph-specific difference could be detected when comparing fruit buds of to be dehiscent and dehiscent seeds. Differential expression of *AearIND* was only detected in late flower stage, suggesting that the decision has to be taken in the early flower stage. Gene expression analysis suggested another bHLH containing gene, *BRC1*, a central integrator in branching control [61]. This bHLH_TCP TAP showed a strong expression peak in early flowers of the indehiscent morph in comparison to the dehiscent morph and is suggested to act as a binary switch resulting in the two fruit and seed morphs.

Once the seed morphs are established, the most interesting TAP discrepancy between the morphs is found amongst histone modifying SET and TAZ TRs and the TF HSF (Chapter 3). SET and TAZ are upregulated in the indehiscent morph and are known for regulating seed maturation and dormancy as well as inhibiting germination. HSF is upregulated in dehiscent seeds and is thought to accelerate maturation.

4.2 TAPs in *Ae. arabicum* seed germination - Employing SeedAdapt expression atlas

Looking into the expression levels (RPKM) of each specific TAP for each sequenced SeedAdapt RNA-seq sample, we can measure their expression levels throughout the developmental stages of dry seed (6 samples), imbibing seed (53 samples) to germinated seed (3 samples) (Fig. 8). We see that the dry seed samples cluster together. Out of the three samples representing germinated seeds, two cluster together (68h and 72h) with the 200h sample falling outside.

Cumulatively adding the expression of the genes encoding each TAP we find 38 TAPs that increase and 25 that decrease in expression once the dry seed is let to imbibe. This suggests that imbibing seeds are more transcriptionally active tissues, compared to dry seeds. 9 TAPs (ABI3/VP1, Aux/IAA, bHLH, DUF296, HD_KNOX1, HD-Zip_I_II, OFP, Rcd1-like and ULT) show a significant continuous increase going from dry seed through to germination. Amongst these we find known players in the seed development regulatory network, such as ABI3/VP1 [63] and Aux/IAA [64] (unpublished Fig. 9 a/b).

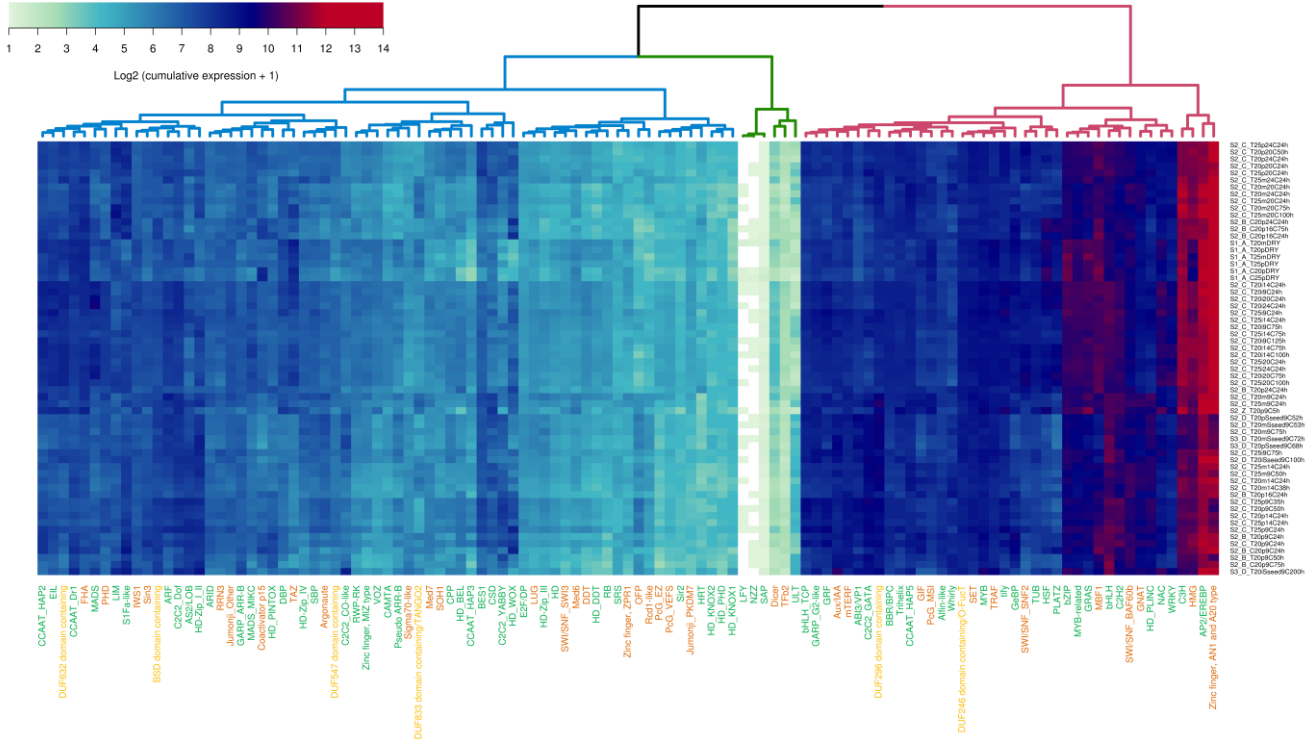


Figure 8 Heatmap using log2 transformed cumulative expression values of each TAP (x-axis) throughout the different developmental stages (y-axis), going from dry seed (S1), imbibing seed (S2) to germinated seed (S3). The data was clustered on the x and y axis using complete linkage with euclidean distances

HD_KNOX TFs, involved in body plan formation and cell fate determination of flowering plant stem cells [65], has an expression profile accompanied by OFP, a known interacting regulator [66].

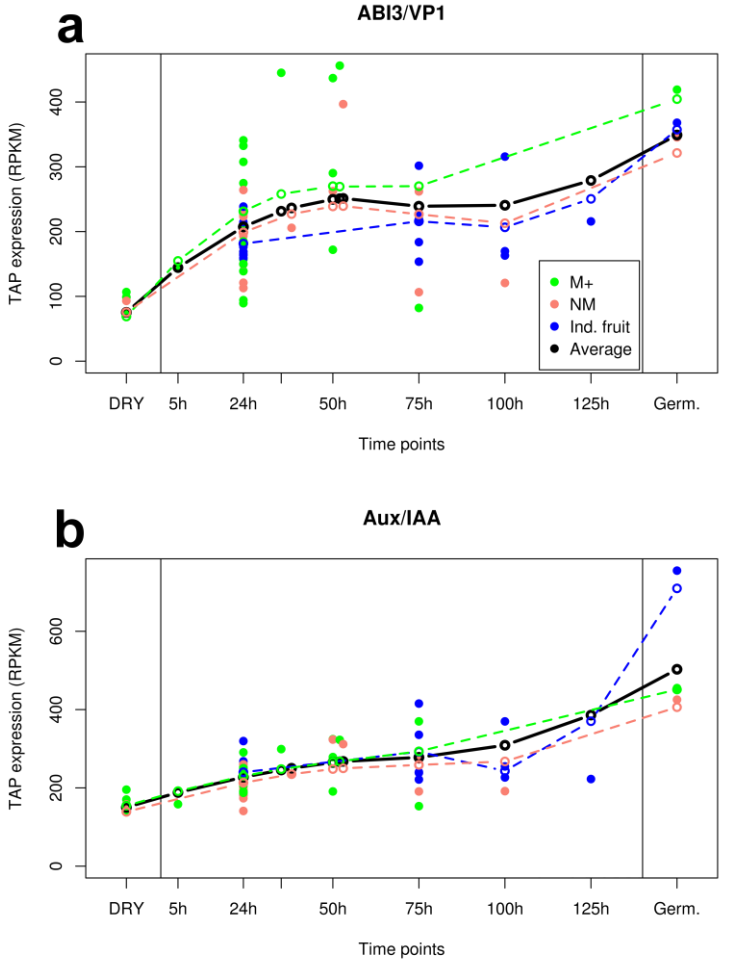


Figure 9. Cumulative expression of TAPs ABI3/VP1 (a) and Aux/IAA (b) throughout germination, going from dry seed to germinated seed.

Another TAP interacting with HD_KNOX is HD_BEL. HD_BEL also shows an increase in expression in the early imbibing seeds that levels out and is maintained through to germination. ULT TF is well studied

for its impact on shoot and flower development, and has been shown to regulate other TAPs such as MADS and HD_KNOX1 [67].

A few of the initially increasing TAPs seem to have their peak in the start of imbibition, such as HD_WOX (Fig. 10), to later drop off and return to

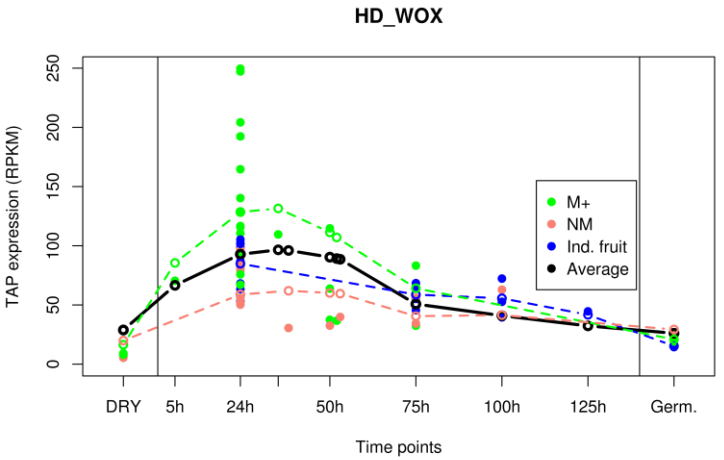


Figure 10. Cumulative expression of TAP HD_WOX throughout germination, going from dry seed to germinated seed.

the initial dry seed level once germination is reached. HD_WOX is known for its broad involvement in plant development, from early embryo patterning [68] to shoot and floral stem cell maintenance [69]. This initial early spike in expression could indicate a role in the early transition from dry seed to imbibing seed.

Though, as mentioned, the dry seed is a less transcriptionally active tissue, compared to imbibing, 25 TAPs still show a decrease in expression in the transitioning from dry seed to imbibing seed. 9 TAPs (DBP, EIL, FHA, GeBP, GRAS, Jumonji_PKDM7, MBF1, PcG_MSI and TUB) show a continuous decrease in expression going from imbibing seed to germinated seed. Amongst these we find known hormone response TFs, EIL and GeBP, each promoting the response of ethylene and cytokinin which are known for their involvement in germination, seedling

development and growth [70, 71].

A few TAPs have their decrease mainly going from dry seed to early stage imbibing. Amongst them we find HSF, TAZ and VOZ. HSF and TAZ, as previously mentioned, are involved in regulating seed maturation and dormancy by repressing germination and would thus expectedly decrease at the onset of germination. Not much is known about VOZ, with regards to seed development, though studies have shown it to be involved in flowering timing in *A. thaliana* [72]. Presented in that study, interestingly, was that VOZ double mutants showed an increase in seed abortants.

Though this is just a broad overview, scratching on the surface of the SeedAdapt data resource, we still identify varying levels of TAP expression throughout seed germination in *Ae. arabicum*. These changing expression levels overlap well with known TAP functions. The potential to delve further, taking the experimental design factors into account such as seed types, maternal growth temperatures as well as germination temperatures, will yield a clear picture of TAP influence during seed germination in *Ae. Arabicum*, all made possible through the SeedAdapt consortium.

4.4 Method

The cumulative expression of the genes encoding each TAP was calculated for all samples sequenced and analyzed through the SeedAdapt consortium. The samples were then divided into groups depending on respective seed stage (dry seed, imbibing seed and germinated seed). The data was log2 transformed and then hierarchically clustered on the x and y axis using complete linkage with euclidean distances, and visualized as a heatmap using R gplots v3.0.1 (https://cran.r-project.org/web/packages/gplots/index.html; last accessed December 8, 2017). The samples were further divided depending on the time point the sample was collected (dry, imbibing 5h,

24h, 50h, 75h, 100h, 125h and germinated) as well as morph (M+ seed, NM seed and indehiscent fruit). Each TAPs cumulative expression value was plotted and local estimated scatterplot smoothing (LOESS) was applied for each morph as well as for the total. Wilcoxon two-sampled tests were carried out to distinguish differences in TAP expression between the different stages.

Chapter 5

Outlook

5 Outlook

5.1 TAP evolution in Viridiplantae and the TAPscan resource

With a higher resolution picture of TAP evolution in Viridiplantae we can provide evidence that a majority of TAP gains occurred in between streptophytic algae and embryophytes, some of which formerly thought to be land plant specific gains. This confirms the expected scenario that with the emergence of new sequence data from species situated between land plants and Chlorophyta the view on some land plant specific proteins has to be revised [32-34]. Though our data points toward a stepwise acquisition of TAP families, going from the early branching streptophytic algae until the embryophytes, these results might still be due to the few and of poor quality data sets representing these clades. More streptophytic algae sequence data is required to more precisely assert the points where the major plant TAP family gains occurred, such as the Chara genome project [46]. Within the group of streptophytic algae we also find the emergence of multiple plant characteristics, such as the plant cell wall cellulose synthase complex being present in the KCM clade (Klebsormidiophyceae, Chlorokybophyceae and Mesostigmatophyceae) [32], polyplastidy in ZCC clade (Zygnematophyceae, Coleochaetophyceae and Charophyceae) [73] as well as the phragmoplast and the preprophase band in Zygnematales [74]. Independently if the common ancestor of all streptophytic algae coincides with the point of the major TAP family gains within Viridiplantae, venturing into the less explored giant group of green algae (Chlorophyta), such as the Ulva genome project [48], would be required to confirm it and to further trace the evolution of TAPs. To further deepen our understanding of TAP evolution in land plants specifically, of which there were recently distinct clades that completely lacked representation (Ferns), the low genome representation in some fundamentally important clades needs to be solved. The relatively plentifully sequenced group of flowering plants makes up for the largest amount of land plant species known to-date, ~330,000 out of ~380,000

[22], and is the one clade group that have evolved all the typical land plant characteristics (flowers, seeds, leaves, vascularity, embryos). Conducting comparative studies with the aim to elucidate the emergence of these significant traits, it is important to look outside of flowering plants. As for which genomes that will be sequenced in the near future, the 10KP initiative declares that ~2,500 genomes, from 357 different families, are going to be non-seed plants, said to result in covering every genus of the Viridiplantae [75]. Once these data are published and available, it will make for a great opportunity to be analyzed with TAPscan giving us an even more detailed picture of TAP evolution in Viridiplantae.

5.2 Dimorphic Seeds of *Ae. arabicum*

The differences in the expression profiles of the two morphological seed types of *Ae. arabicum* coincide well with the expected seed characteristics when compared in the context of *A. thaliana* seeds. The less dormant mucilaginous morph (M+) showed higher expression of genes that increase during maturation in comparison to the more dormant non-mucilaginous morph (NM). The expression profiles propose that the M+ seed are genetically wired to dry out and mature faster in comparison to the NM seeds whose expression profile is geared towards transcription, important for dormancy. Though this study only included a small set of samples from a specific seed stage, being the dry seed, the detected expression differences fits well into the proposed bet hedging strategy employed by *Ae. arabicum*. The plants decision of development of the distinct seed types was shown to be connected to the expression levels of bHLH_TCP coding *BRC1* [61]. The development of the dehiscent seed is default up until the early flower stage where, if unfavorable conditions occur, the seed ratio switches towards the "low risk" indehiscent seed.

Having developed the analytic pipeline in chapter 2 and then applied it to all of the data generated by SeedAdapt (chapter 4) a tremendous

resource has been created. In the short run, this will constitute the fundament enabling a deeper understanding of the diaspore syndrome, and will, also in the long run, act as a resource making it possible to investigate any homologous gene of interest and compare its expression profile throughout the different provided conditions.

5.3 Seed development TAPs in an evolutionary perspective

Looking at the accumulated expression of TAPs during seed development up until germination (chap 4) we see extensive changes in expression occurring. Out of all the highlighted TAPs it is only ULT that is unique to seed plants (spermatophyte). ULT is involved in shoot development [54], which is expected to be initiated at the start of germination. Moving further back to the common ancestor of all land plants we find the emergence of plant embryogenesis, giving rise to the embryo. At that point we can pinpoint the further gain of the two TAPs GeBP and VOZ. The majority of TAP gains through the evolution of Viridiplantae, occured at the common ancestor of Embryophytes, the ZCC-clade and Klebsormidiales. At this point the majority of the TAPs highlighted in chapter 4.2 were present. Within the seed we find the plant embryo, an ancestral trait to all land plants. The embryo is formed during embryogenesis from the zygote. That the large portion of the highlighted TAPs, with changing expression during seed development, are present outside of Spermatophyta emphasizes their fundamental involvement in the establishment, development and propagation of the zygote and embryo. Homologs of genes involved in the delay of seed germination have been shown, through knock-out studies (BSc-thesis of group member Marlies Peter), to also affect the germination time of spores in the moss *P. patens*. Though seed and spores are not homologous this suggests a conservation of germination regulation with regards to respective dispersal unit (seeds and spores).

To single out the TAP signal that could play a unique role in the development of seeds, the vast TAP expression signal that is devoted to

general embryo and/or zygote development has to be removed. This could be done by looking into the TAP expression during development of the dispersal units of non-seed embryophytes, such as the ferns, lycophytes and bryophytes. Assuming identification of the analogous stage, suggestively the sporophyte embryo, it would then provide a general TAP expression profile that could be compared to seeds. Additionally, generating a TAP expression profile of zygote development in non-embryogenic plants, such as the streptophytic algae, would further elucidate the role of the TAP repertoire in the development of dispersal units.

5.4 Sequence analysis and annotation

Though sequence data becomes cheaper and cheaper there is still more effort and sequence data required to assemble a genome in comparison to a transcriptome. Following the development of sequencing techniques and computational tools, assembling a genome will at some point become as easy and cheap as generating a transcriptome is today. While scientific research is moving more into the unexplored branches of life, where little to no prior work has been done, continuous work on a transcriptomic level, proven to conform well with when using a genome approach, will provide the possibility to conduct comparative studies in non-model and novel species. This is exemplified in the 1KP [76] as well as in the Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP) [77]. Here, transcriptomic data was generated for 1,000 and 650 species respectively and has since their release in 2014 been cited more than 200 and 340 times respectively.

Annotation tools, such as TAPscan, depend on implementing the latest scientific discoveries to provide the most up to date results. This not only includes the need to be up to date with the latest literature to implement newly discovered families and subfamilies, but also to make sure the most appropriate protein screening tools are being used. Further, the idea of a comprehensive eukaryote-wide TAPscan classification tool is

appealing. Unifying the Archaeplastida with the Excavates, Opisthokonts, Amoebozoans, SARs and every unclassified species in between these clades, under the same extensive, all inclusive, TAP classification scheme, would yield the most comprehensive eukaryote TAP resource. With it, the influence of gene expression regulation, in the form of TAPs, on cellular evolution could be further understood.

References

1. WATSON JD, CRICK FH: Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* 1953, 171(4356):737-738.
2. NIRENBERG MW, MATTHAEI JH: The dependence of cell-free protein synthesis in E. coli upon naturally occurring or synthetic polyribonucleotides. *Proc Natl Acad Sci U S A* 1961, 47:1588-1602.
3. HOLLEY RW, EVERETT GA, MADISON JT, ZAMIR A: NUCLEOTIDE SEQUENCES IN THE YEAST ALANINE TRANSFER RIBONUCLEIC ACID. *J Biol Chem* 1965, 240:2122-2128.
4. Holley RW: Structure of an alanine transfer ribonucleic acid. *JAMA* 1965, 194(8):868-871.
5. Min Jou W, Haegeman G, Ysebaert M, Fiers W: Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein. *Nature* 1972, 237(5350):82-88.
6. Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, Hutchison CA, Slocombe PM, Smith M: Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* 1977, 265(5596):687-695.
7. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Ostell J, Pruitt KD, Sayers EW: GenBank. *Nucleic Acids Res* 2018, 46(D1):D41-D47.
8. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W *et al*: Initial sequencing and analysis of the human genome. *Nature* 2001, 409(6822):860-921.
9. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA *et al*: The sequence of the human genome. *Science* 2001, 291(5507):1304-1351.
10. The Cost of Sequencing a Human Genome. In. https://www.genome.gov/27565109/the-cost-of-sequencing-a-human-genome/: National Institutes of Health (NIH); July 6, 2016.

11. Press Release: Illumina Introduces the NovaSeq Series—a New Architecture Designed to Usher in the $100 Genome. In. https://www.illumina.com/company/news-center/press-releases/press-release-details.html?newsid=2236383: Illumina Inc.; January 9, 2017.
12. McGinnis W, Levine MS, Hafen E, Kuroiwa A, Gehring WJ: A conserved DNA sequence in homoeotic genes of the Drosophila Antennapedia and bithorax complexes. *Nature* 1984, 308(5958):428-433.
13. Byrne ME, Groover AT, Fontana JR, Martienssen RA: Phyllotactic pattern and stem cell fate are determined by the Arabidopsis homeobox gene BELLRINGER. *Development* 2003, 130(17):3941-3950.
14. de Mendoza A, Sebé-Pedrós A, Šestak MS, Matejcic M, Torruella G, Domazet-Loso T, Ruiz-Trillo I: Transcription factor evolution in eukaryotes and the assembly of the regulatory toolkit in multicellular lineages. *Proc Natl Acad Sci U S A* 2013, 110(50):E4858-4866.
15. Bürglin TR, Affolter M: Homeodomain proteins: an update. *Chromosoma* 2016, 125(3):497-521.
16. Lang D, Weiche B, Timmerhaus G, Richardt S, Riaño-Pachón DM, Corrêa LG, Reski R, Mueller-Roeber B, Rensing SA: Genome-wide phylogenetic comparative analysis of plant transcriptional regulation: a timeline of loss, gain, expansion, and correlation with complexity. *Genome Biol Evol* 2010, 2:488-503.
17. Lang D, Rensing SA: The Evolution of Transcriptional Regulation in the Viridiplantae and its Correlation with Morphological Complexity. In: *Evolutionary Transitions to Multicellular Life: Principles and mechanisms.* Edited by Ruiz-Trillo I, Nedelcu AM. Dordrecht: Springer Netherlands; 2015: 301-333.
18. Rensing SA: (Why) Does Evolution Favour Embryogenesis? *Trends Plant Sci* 2016, 21(7):562-573.
19. de Mendoza A, Suga H, Permanyer J, Irimia M, Ruiz-Trillo I: Complex transcriptional regulation and independent evolution of fungal-like traits in a relative of animals. *Elife* 2015, 4:e08904.

20. Catarino B, Hetherington AJ, Emms DM, Kelly S, Dolan L: The Stepwise Increase in the Number of Transcription Factor Families in the Precambrian Predated the Diversification of Plants On Land. *Mol Biol Evol* 2016, 33(11):2815-2819.
21. Koepfli KP, Paten B, O'Brien SJ, Scientists GKCo: The Genome 10K Project: a way forward. *Annu Rev Anim Biosci* 2015, 3:57-111.
22. Cheng S, Melkonian M, Smith SA, Brockington S, Archibald JM, Delaux PM, Li FW, Melkonian B, Mavrodiev EV, Sun W *et al*: 10KP: A phylodiverse genome sequencing plan. *Gigascience* 2018, 7(3):1-9.
23. Rensing SA: Why we need more non-seed plant models. *New Phytol* 2017, 216(2):355-360.
24. Morris JL, Puttick MN, Clark JW, Edwards D, Kenrick P, Pressel S, Wellman CH, Yang Z, Schneider H, Donoghue PCJ: The timescale of early land plant evolution. *Proc Natl Acad Sci U S A* 2018, 115(10):E2274-E2283.
25. Corlett RT: Plant diversity in a changing world: Status, trends, and conservation needs. *Plant Divers* 2016, 38(1):10-16.
26. Rensing SA: Plant Evolution: Phylogenetic Relationships between the Earliest Land Plants. *Curr Biol* 2018, 28(5):R210-R213.
27. Initiative AG: Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* 2000, 408(6814):796-815.
28. Derelle E, Ferraz C, Rombauts S, Rouzé P, Worden AZ, Robbens S, Partensky F, Degroeve S, Echeynié S, Cooke R *et al*: Genome analysis of the smallest free-living eukaryote Ostreococcus tauri unveils many unique features. *Proc Natl Acad Sci U S A* 2006, 103(31):11647-11652.
29. Merchant SS, Prochnik SE, Vallon O, Harris EH, Karpowicz SJ, Witman GB, Terry A, Salamov A, Fritz-Laylin LK, Maréchal-Drouard L *et al*: The Chlamydomonas genome reveals the evolution of key animal and plant functions. *Science* 2007, 318(5848):245-250.
30. Rensing SA, Lang D, Zimmer AD, Terry A, Salamov A, Shapiro H, Nishiyama T, Perroud PF, Lindquist EA, Kamisugi Y *et al*: The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants. *Science* 2008, 319(5859):64-69.
31. Becker B, Marin B: Streptophyte algae and the origin of embryophytes. *Ann Bot* 2009, 103(7):999-1004.

32. Hori K, Maruyama F, Fujisawa T, Togashi T, Yamamoto N, Seo M, Sato S, Yamada T, Mori H, Tajima N *et al*: Klebsormidium flaccidum genome reveals primary factors for plant terrestrial adaptation. *Nat Commun* 2014, 5:3978.
33. Delaux PM, Radhakrishnan GV, Jayaraman D, Cheema J, Malbreil M, Volkening JD, Sekimoto H, Nishiyama T, Melkonian M, Pokorny L *et al*: Algal ancestor of land plants was preadapted for symbiosis. *Proc Natl Acad Sci U S A* 2015, 112(43):13390-13395.
34. Wang C, Liu Y, Li SS, Han GZ: Insights into the origin and evolution of the plant hormone signaling machinery. *Plant Physiol* 2015, 167(3):872-886.
35. Winchell F, Stevens CJ, Murphy C, Champion L, Fuller D: Evidence for Sorghum Domestication in Fourth Millennium BC Eastern Sudan: Spikelet Morphology from Ceramic Impressions of the Butana Group. *Current Anthropology* 2017, 58(5):673-683.
36. Sedivy EJ, Wu F, Hanzawa Y: Soybean domestication: the origin, genetic architecture and molecular bases. *New Phytol* 2017, 214(2):539-553.
37. H F: Origin of the 'Weisse Schlesische Rübe' (white Silesian beet) and resynthesis of sugar beet. In., vol. 41. *Euphytica*; 1989: 75–80.
38. Benz BF: Archaeological evidence of teosinte domestication from Guilá Naquitz, Oaxaca. *Proc Natl Acad Sci U S A* 2001, 98(4):2104-2106.
39. Rumold CU, Aldenderfer MS: Late Archaic-Early Formative period microbotanical evidence for potato at Jiskairumoko in the Titicaca Basin of southern Peru. *Proc Natl Acad Sci U S A* 2016, 113(48):13672-13677.
40. Beadle GW: Teosinte and the origin of maize. *Journal of Heredity* 1939, 30(6):245-247.
41. Alonso-Blanco C, Aarts MG, Bentsink L, Keurentjes JJ, Reymond M, Vreugdenhil D, Koornneef M: What has natural variation taught us about plant development, physiology, and adaptation? *Plant Cell* 2009, 21(7):1877-1896.
42. Finch-Savage WE, Leubner-Metzger G: Seed dormancy and the control of germination. *New Phytol* 2006, 171(3):501-523.
43. Imbert E: Ecological consequences and ontogeny of seed heteromorphism. *Perspectives in Plant Ecology, Evolution and Systematics* 2002, 5(1):13 - 36.

44. Solms-Laubach HG: Über die Arten des Genus Ae*thionema,* die Schließfrüchte hervorbringen. In. Botanische Zeitung: Verlag von Arthur Felix; 1901: 61–78.
45. Lenser T, Graeber K, Cevik Ö, Adigüzel N, Dönmez AA, Grosche C, Kettermann M, Mayland-Quellhorst S, Mérai Z, Mohammadin S *et al*: Developmental Control and Plasticity of Fruit and Seed Dimorphism in Aethionema arabicum. *Plant Physiol* 2016, 172(3):1691-1707.
46. Nishiyama T, Sakayama H, de Vries J, Buschmann H, Saint-Marcoux D, Ullrich KK, Haas FB, Vanderstraeten L, Becker D, Lang D *et al*: The Chara Genome: Secondary Complexity and Implications for Plant Terrestrialization. *Cell* 2018, 174(2):448-464.e424.
47. Li FW, Brouwer P, Carretero-Paulet L, Cheng S, de Vries J, Delaux PM, Eily A, Koppers N, Kuo LY, Li Z *et al*: Fern genomes elucidate land plant evolution and cyanobacterial symbioses. *Nat Plants* 2018, 4(7):460-472.
48. De Clerck O, Kao SM, Bogaert KA, Blomme J, Foflonker F, Kwantes M, Vancaester E, Vanderstraeten L, Aydogdu E, Boesger J *et al*: Insights into the Evolution of Multicellularity from the Sea Lettuce Genome. *Curr Biol* 2018, 28(18):2921-2933.e2925.
49. Delwiche CF: The Genomes of Charophyte Green Algae. In., vol. 78. Advances in Botanical Research; 2016: 255-270.
50. Nicolas M, Cubas P: TCP factors: new kids on the signaling block. *Curr Opin Plant Biol* 2016, 33:33-41.
51. Lumpkin TA, Plucknett DL: Azolla: botany, physiology, and use as a green manure. In., vol. 34. Economic Botany; 1980: 111–153.
52. Speelman EN, Van Kempen MM, Barke J, Brinkhuis H, Reichart GJ, Smolders AJ, Roelofs JG, Sangiorgi F, de Leeuw JW, Lotter AF *et al*: The Eocene Arctic Azolla bloom: environmental conditions, productivity and carbon drawdown. *Geobiology* 2009, 7(2):155-170.
53. Bouyer D, Roudier F, Heese M, Andersen ED, Gey D, Nowack MK, Goodrich J, Renou JP, Grini PE, Colot V *et al*: Polycomb repressive complex 2 controls the embryo-to-seedling phase transition. *PLoS Genet* 2011, 7(3):e1002014.
54. Fletcher JC: The ULTRAPETALA gene controls shoot and floral meristem size in Arabidopsis. *Development* 2001, 128(8):1323-1333.

55. Tait K, Joint I, Daykin M, Milton DL, Williams P, Cámara M: Disruption of quorum sensing in seawater abolishes attraction of zoospores of the green alga Ulva to bacterial biofilms. *Environ Microbiol* 2005, 7(2):229-240.
56. Spoerner M, Wichard T, Bachhuber T, Stratmann J, Oertel W: Growth and Thallus Morphogenesis of Ulva mutabilis (Chlorophyta) Depends on A Combination of Two Bacterial Species Excreting Regulatory Factors. *J Phycol* 2012, 48(6):1433-1447.
57. Wichard T, Charrier B, Mineur F, Bothwell JH, Clerck OD, Coates JC: The green seaweed Ulva: a model system to study morphogenesis. *Front Plant Sci* 2015, 6:72.
58. Wang H, Zhang Z, Li H, Zhao X, Liu X, Ortiz M, Lin C, Liu B: CONSTANS-LIKE 7 regulates branching and shade avoidance response in Arabidopsis. *J Exp Bot* 2013, 64(4):1017-1024.
59. Xiao G, Li B, Chen H, Chen W, Wang Z, Mao B, Gui R, Guo X: Overexpression of PvCO1, a bamboo CONSTANS-LIKE gene, delays flowering by reducing expression of the FT gene in transgenic Arabidopsis. *BMC Plant Biol* 2018, 18(1):232.
60. Fernandez-Pozo N, Zheng Y, Snyder SI, Nicolas P, Shinozaki Y, Fei Z, Catala C, Giovannoni JJ, Rose JKC, Mueller LA: The Tomato Expression Atlas. *Bioinformatics* 2017, 33(15):2397-2398.
61. Lenser T, Tarkowská D, Novák O, Wilhelmsson PKI, Bennett T, Rensing SA, Strnad M, Theißen G: When the BRANCHED network bears fruit: how carpic dominance causes fruit dimorphism in Aethionema. *Plant J* 2018, 94(2):352-371.
62. Merai Z, Graeber K, Wilhelmsson PKI, Ullrich KK, Arshad W, Grosche C, Tarkowska D, Tureckova V, Strnad M, Rensing SA *et al*: A novel model plant to study the light control of seed germination. *BioRxiv 470401 [Preprint]* 2018.
63. Santos-Mendoza M, Dubreucq B, Baud S, Parcy F, Caboche M, Lepiniec L: Deciphering gene regulatory networks that control seed development and maturation in Arabidopsis. *Plant J* 2008, 54(4):608-620.
64. Locascio A, Roig-Villanova I, Bernardi J, Varotto S: Current perspectives on the hormonal control of seed development in Arabidopsis and maize: a focus on auxin. *Front Plant Sci* 2014, 5:412.

65. Hay A, Tsiantis M: KNOX genes: versatile regulators of plant development and diversity. *Development* 2010, 137(19):3153-3165.
66. Wang S, Chang Y, Ellis B: Overview of OVATE FAMILY PROTEINS, A Novel Class of Plant-Specific Growth Regulators. *Front Plant Sci* 2016, 7:417.
67. Monfared MM, Carles CC, Rossignol P, Pires HR, Fletcher JC: The ULT1 and ULT2 trxG genes play overlapping roles in Arabidopsis development and gene regulation. *Mol Plant* 2013, 6(5):1564-1579.
68. Haecker A, Gross-Hardt R, Geiges B, Sarkar A, Breuninger H, Herrmann M, Laux T: Expression dynamics of WOX genes mark cell fate decisions during early embryonic patterning in Arabidopsis thaliana. *Development* 2004, 131(3):657-668.
69. Dolzblasz A, Nardmann J, Clerici E, Causier B, van der Graaff E, Chen J, Davies B, Werr W, Laux T: Stem Cell Regulation by Arabidopsis WOX Genes. *Mol Plant* 2016, 9(7):1028-1039.
70. Corbineau F, Xia Q, Bailly C, El-Maarouf-Bouteau H: Ethylene, a key factor in the regulation of seed dormancy. *Front Plant Sci* 2014, 5:539.
71. Chevalier F, Perazza D, Laporte F, Le Hénanff G, Hornitschek P, Bonneville JM, Herzog M, Vachon G: GeBP and GeBP-like proteins are noncanonical leucine-zipper transcription factors that regulate cytokinin response in Arabidopsis. *Plant Physiol* 2008, 146(3):1142-1154.
72. Celesnik H, Ali GS, Robison FM, Reddy AS: Arabidopsis thaliana VOZ (Vascular plant One-Zinc finger) transcription factors are required for proper regulation of flowering time. *Biol Open* 2013, 2(4):424-431.
73. de Vries J, Stanton A, Archibald JM, Gould SB: Streptophyte Terrestrialization in Light of Plastid Evolution. *Trends Plant Sci* 2016, 21(6):467-476.
74. Buschmann H, Zachgo S: The Evolution of Cell Division: From Streptophyte Algae to Land Plants. *Trends Plant Sci* 2016, 21(10):872-883.
75. Species distribution and Phylogenetic Diversity. In. https://db.cngb.org/10kp/: 10KP Homepage; 2018.
76. Matasci N, Hung LH, Yan Z, Carpenter EJ, Wickett NJ, Mirarab S, Nguyen N, Warnow T, Ayyampalayam S, Barker M *et al*: Data access for the 1,000 Plants (1KP) project. *Gigascience* 2014, 3:17.

77. Keeling PJ, Burki F, Wilcox HM, Allam B, Allen EE, Amaral-Zettler LA, Armbrust EV, Archibald JM, Bharti AK, Bell CJ *et al*: The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. *PLoS Biol* 2014, 12(6):e1001889.

Acknowledgements

Stefan A. Rensing, I am forever grateful that you believed in me and took me in as a PhD student. During the time in your group, my personal development as a scientist excelled far beyond anything I could imagine. Without your guidance, support and inspiration my achievements would not have been possible, the very least this thesis.

I would like to thank Kristian K. Ullrich for helping me through my first years in Marburg, contributing a lot to the steep learning curve I went through. Noe Fernandez-Pozo, thanks for helping me close out my final project and getting my mind set on finalizing my thesis. Christopher Grosche, thanks for your help and consultation throughout my studies.

Thanks to all group members for making no day in the lab go by without laughter.

I would like to thank the members of the SeedAdapt consortium for the many joy and fruitful conferences and collaborations.

Thanks to all my friends in Marburg, making me quickly refer to Marburg as home.

Eva, Jan and Erik, your encouragement gave me strength to accomplish this.

List of publications

Wilhelmsson, P. K. I., C. Muhlich, K. K. Ullrich & S. A. Rensing (2017) Comprehensive Genome-Wide Classification Reveals That Many Plant-Specific Transcription Factors Evolved in Streptophyte Algae. Genome Biol Evol, 9, 3384-3397.

Wilhelmsson, P. K. I., J. O. Chandler, N. Fernandez-Pozo, K. Graeber, K. K. Ullrich, W. Arshad, S. Khan, J. A. Hofberger, K. Buchta, P. P. Edger, J. C. Pires, M. E. Schranz, G. Leubner-Metzger & S. A. Rensing (2019) Usability of reference-free transcriptome assemblies for detection of differential expression: a case study on Aethionema arabicum dimorphic seeds. BMC Genomics, 20, 95.

Lenser, T., D. Tarkowská, O. Novák, **P. K. I. Wilhelmsson**, T. Bennett, S. A. Rensing, M. Strnad & G. Theißen (2018) When the BRANCHED network bears fruit: how carpic dominance causes fruit dimorphism in Aethionema. Plant J, 94, 352-371.

Nishiyama, T., H. Sakayama, J. de Vries, H. Buschmann, D. Saint-Marcoux, K. K. Ullrich, F. B. Haas, L. Vanderstraeten, D. Becker, D. Lang, S. Vosolsobě, S. Rombauts, **P. K. I. Wilhelmsson**, P. Janitza, R. Kern, A. Heyl, F. Rümpler, L. I. A. C. Villalobos, J. M. Clay, R. Skokan, A. Toyoda, Y. Suzuki, H. Kagoshima, E. Schijlen, N. Tajeshwar, B. Catarino, A. J. Hetherington, A. Saltykova, C. Bonnot, H. Breuninger, A. Symeonidi, G. V. Radhakrishnan, F. Van Nieuwerburgh, D. Deforce, C. Chang, K. G. Karol, R. Hedrich, P. Ulvskov, G. Glöckner, C. F. Delwiche, J. Petrášek, Y. Van de Peer, J. Friml, M. Beilby, L. Dolan, Y. Kohara, S. Sugano, A. Fujiyama, P. M. Delaux, M. Quint, G. Theißen, M. Hagemann, J. Harholt, C. Dunand, S. Zachgo, J. Langdale, F. Maumus, D. Van Der Straeten, S. B. Gould & S. A. Rensing (2018) The Chara Genome: Secondary Complexity and Implications for Plant Terrestrialization. Cell, 174, 448-464.e24.

De Clerck, O., S. M. Kao, K. A. Bogaert, J. Blomme, F. Foflonker, M. Kwantes, E. Vancaester, L. Vanderstraeten, E. Aydogdu, J. Boesger, G. Califano, B. Charrier, R. Clewes, A. Del Cortona, S. D'Hondt, N. Fernandez-Pozo, C. M. Gachon, M. Hanikenne, L. Lattermann, F. Leliaert, X. Liu, C. A. Maggs, Z. A. Popper, J. A. Raven, M. Van Bel, **P. K. I. Wilhelmsson,** D. Bhattacharya, J. C. Coates, S. A. Rensing, D. Van Der Straeten, A. Vardi, L. Sterck, K. Vandepoele, Y. Van de Peer, T. Wichard & J. H. Bothwell (2018) Insights into the Evolution of Multicellularity from the Sea Lettuce Genome. Curr Biol, 28, 2921-2933.e5.

Li, F. W., P. Brouwer, L. Carretero-Paulet, S. Cheng, J. de Vries, P. M. Delaux, A. Eily, N. Koppers, L. Y. Kuo, Z. Li, M. Simenc, I. Small, E. Wafula, S. Angarita, M. S. Barker, A. Bräutigam, C. dePamphilis, S. Gould, P. S. Hosmani, Y. M. Huang, B. Huettel, Y. Kato, X. Liu, S. Maere, R. McDowell, L. A. Mueller, K. G. J. Nierop, S. A. Rensing, T. Robison, C. J. Rothfels, E. M. Sigel, Y. Song, P. R. Timilsena, Y. Van de Peer, H. Wang, **P. K. I. Wilhelmsson**, P. G. Wolf, X. Xu, J. P. Der, H. Schluepmann, G. K. Wong & K. M. Pryer (2018) Fern genomes elucidate land plant evolution and cyanobacterial symbioses. Nat Plants, 4, 460-472.

Curriculum Vitae

Personal information

Name	Per Karl Ivar Wilhelmsson
Born	6th Nov. 1986
in	Malmö, Sweden

Higher education

2007 - 2011	Bachelor of Science Lunds Universitet, Sweden
2011 - 2013	Master in Bioinformatics Lunds Universitet, Sweden
2014 – pres.	Doctoral studies Philipps Universität, Marburg

Declarations

I hereby assure that I have written this dissertation

Transcription associated proteins in plant development and evolution

I wrote it myself without any external assistance aids and used no sources or aids other than those indicated. This dissertation has not been submitted to any other domestic or foreign university in connection with a doctoral application or for other examination purposes.

_____

Per K. I. Wilhelmsson