Comprehensive genome-wide classification reveals that many plant-specific transcription factors evolved in streptophyte algae

Per K.I. Wilhelmsson¹, Cornelia Mühlich¹, Kristian K. Ullrich¹,² and Stefan A. Rensing¹,³,*

¹ Plant Cell Biology, Faculty of Biology, University of Marburg, Karl-von-Frisch-Str. 8, 35043 Marburg, Germany
² Current: Max Planck Institute for Evolutionary Biology, August-Thienemann-Straße 2, 24306 Ploen, Germany
³ BIOSS Centre for Biological Signalling Studies, University of Freiburg, Schaezlestr. 18, 79104 Freiburg, Germany

* Author for correspondence: Stefan A. Rensing, Plant Cell Biology, Faculty of Biology, University of Marburg, Marburg, Germany. Phone: +4964212821940, Fax: +4964212822190, stefan.rensing@biologie.uni-marburg.de
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.

Keywords: 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 (de Mendoza, et al. 2013; Lang and Rensing 2015; Lang, et al. 2010; Levine and Tjian 2003). 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 (Catarino, et al. 2016; de Mendoza, et al. 2013; de Mendoza, et al. 2015; Lang, et al. 2010). 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 (Delaux, et al. 2015; Hori, et al. 2014; Wang, et al. 2015), which we set out to do here by including more data of streptophyte algae and bryophytes than previously available.

TAPs, 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 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 e.g. 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, e.g. in embryogenesis, in plants and animals (Catarino, et al. 2016; Hudry, et al. 2014). The KNOX and BELL sub-families 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 co-opted 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 (Bouyer, et al. 2011; Mosquina, et al. 2009; Okano, et al. 2009).

TAPs 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 (Guo, et al. 2008; Riano-Pachon, et al. 2007;
Richardt, et al. 2007) to generate the comprehensive TAPscan tool, encompassing sensitive
domain-based classification rules for 111 TAP families. Similar approaches were undertaken by
other studies, e.g. PInTFDB (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 sub-family classification for several families. Moreover, we have included 92 more
genomes than were available seven 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 pre-computed gene trees. This
interface is a successor to PInTFDB 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

Dataset

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, Table S5), 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 Table S3). 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 (Table S1). 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 (Table S2; 2x
animals, 1x bryophyte, 8x chlorophytes, 1x conifer, 9x dicots, 1x lycophyte, 6x fungi, 1x
glaucophyte, 4x monocots, 1x charophyte, 7x protoctista [5x non-green algae, 1x Mycetozoa, 1x
Heterolobosea] and 5x rhodophytes). 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 9 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 (Figure
1a) and compared to the 2010 profiles findings (green in Figure 1a). 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 Figure 1a) 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 (Figure 1b, Figure S1) 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 Table S1 and Results for details) more sub-families 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 sub-families could be added. This includes an expansion of the Homeodomain (HD) family from four to 12 sub-families, an additional Jumonji sub-family, an additional Polycomb group (PcG) sub-family, and being able to distinguish the MADS subclass MIKC. If no PFAM profile was available, custom profiles were made using existing MSAs: BEL (Hamant and Pautot 2010; Sharma, et al. 2014b), KNOX_C and PINTOX (Mukherjee, et al. 2009) and WOX (van der Graaff, et al. 2009). When screening known PcG_EZ proteins (Pu and Sung 2015) the prosite CXC pattern (http://prosite.expasy.org/PS51633) was found and the underlying alignment used to build a custom model, replacing the SANTA domain (Table S1).

Inference of ancestral states & 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 (Figure 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 (Table S6). Expansion/contractions and gains/losses were calculated with the count package (Csuros 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 Table S7. The count predictions were entered into Table S6 (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 (Figure 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 less than 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/).

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). The trees are presented on the TAPscan webpage using Archaeopteryx.js (https://sites.google.com/site/cmzmasek/home/software/archaeopteryx-js).

Visualization of family profiles and column charts

Using the R environment (R_Core_Team 2016) the family per species data (Table S6) 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
The y-axis was ordered to follow our adaption of the (Wickett, et al. 2014) phylogeny (Figure 5). The family per species data was also used to create stacked column charts (Figure S3 and Figure S4). Each TAP value was log2 transformed and then grouped by either TAP-class (Figure S3) or amount of multiple domain TAPs (Figure S4), maintaining the species separation.

Implementation of the TAPscan online resource

The web page was setup using a LAMP architecture (Lawton 2005) implemented with Linux Ubuntu 14.04.4 LTS, Apache 2.4.7, MySQL 14.14 Distrib 5.5.49 and PHP 5.5.9-1ubuntu4.17. 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. E.g. there are 5 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 Figure S7. 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/), 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 (Table S1, 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 (Table S1). 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 (Table S1, 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 (Table S1, 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 (Table S2). Using the 2010 (v1) profiles, hit sequences were sampled from each of the 12 groups that the 46 genomes represent, and then used to re-build 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 (Figure 1a). By manual inspection of all aligned hit sequences we defined the individual score cutoffs to lie above sequences of uncertain functional conservation (Figure 1b, Figure S1). 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 (Figure 1b, Figure S1).

**Improved taxon sampling, sub-family definition and specificity**

In the past seven 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 13 genomes and 5 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. While e.g. PlantTFDB v4 (Jin, et al. 2016) includes more angiosperm genomes, we took care to include as much as possible non-seed 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 (Table S5/S6).

To update our classification rules (Table S3), 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 sub-family classification rules were established, and some families renamed due to changes in domain or family names (Figure 2). In particular, we sub-divided
homeodomain (HD) TFs into DDT, PHD, PINTOX, PLINC, WOX, HD-ZIP I/II, III, IV, and into the TALE class sub-families BEL, KNOX 1 and 2 (Hamant and Pautot 2010; Mukherjee, et al. 2009; Sharma, et al. 2014b; van der Graaff, et al. 2009) (Table S1, 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 (Table S1). 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 sub-divided into the TR DDT and the TF HD_DDT in TAPscan v2. With a total of 124 families and sub-families (Table S3, Figure 2; 81 of them TFs) our rule set is the most comprehensive one for plant TAPs, since other approaches have significantly less resolution, e.g. 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 (Martin-Trillo and Cubas 2010; Mosquina, et al. 2009; Mukherjee, et al. 2009; Paponov, et al. 2009). We find that the average sensitivity of TAPscan v2 (87.76 %) is only slightly lower than of v1 (89.31 %), while the specificity of v2 (100.00 %) is much higher than in the old version (92.31 %; Table S4). 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 to 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 (Figure 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 (Figure 4, see Figure S2 for expanded version). Cluster 5 contains families (such as bZIP, bHLH or MYB) that were already abundant in the algal relatives of land plants, while 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. Figure S3). Interestingly, the propensity of TAPs to comprise of more than one functional domain increases in a very similar pattern (Figure S4), 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 and Rensing 2015; Lang, et al. 2010).

**TAP family evolution**

The taxonomic sampling of our data is visualized as a cartoon tree (Figure 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. Table S6/S7). 32 losses were predicted that are scattered along the tree. The streptophyte alga
Klebsormidium nitens apparently secondarily lost five TAP families, while the lycophyte Selaginella moellendorffii lost 8. Another 8 families were lost during gymnosperm evolution, one of them (HD_Pintox) being absent from all studied gymnosperms, whereas 2 are lacking in conifers and 5 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 (Figure 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), while 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. While only two genomes of streptophyte algae have yet been published (Delaux, et al. 2015; Hori, et al. 2014), there are transcriptome data available for seven species (Timme, et al. 2012) that were included into TAPscan (Table 1, Table S5/S6). Similarly, although no other bryophyte genomes than P. patens are published yet, we included transcriptomes of the mosses Ceratodon purpureus (Szovenyi, et al. 2014) and Funaria hygrometrica (Szovenyi, et al. 2010), and of the liverwort Marchantia polymorpha (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 could be confirmed.
Of the others, three (ARF, S1Fa-like, 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 (Table S6). 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), while 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 (Buschmann and Zachgo 2016; Pickett-Heaps, et al. 1999), and that Klebsormidium also possesses some “plant-like” features, like callose and the phenylpropanoid pathway (de Vries, et al. 2017; Herburger and Holzinger 2015). 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 5 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. 2015).
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 (Figure 5, Table S6/S7). 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 to algae (Lang, et al. 2010), 21 show a more than 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 (Table S6). 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 (Figure S3). 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; Table S6). Six TAP families expanded at the basis of angiosperms (among them HD_KNOX2), and several families expanded subsequently (Figure 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 (Figure 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, GeBP) were apparently expanded in the Euphyllophyta (ferns and seed plants, node 27), another three (HD_BEL, Pseudo ARR-B, 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 analysed 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 8, our study places 6 at earlier nodes of the tree, probably due to better taxon sampling. The study also did a sub-family analysis of HD TFs and concluded that almost all were already present in algae. In our study, we find that of 12 HD sub-families 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 Table S6; 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 (Table S6). We detect 400 TFs, Bowman et al. 387; 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](http://pfam.xfam.org/family/PF04782)) 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 (Table S6, Figure 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 (Figure 4) that shows expansion during land plant evolution. It is not detected to be expanded using Wagner parsimony (Figure 5), but shows significant size increase (q < 0.05, Mann Whitney; Table S6) between non-seed 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 non-seed plants. An alignment of the sequences (Figure S5) 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 non-seed plant sequences are present in the same subclade (Figure S6; 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, while 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* [(Hiss, et al. 2014; Hiss, et al. 2017; Hruz, et al. 2008), phytozome.org]. Out of 5 *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 non-seed 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 e.g. reproductive structures, xylem or seed. AT1G52320.2 is induced e.g. under germination, drought and ABA, while AT5G25590.1 shows higher expression e.g. under UV-B, biotic stimulus, elevated CO₂ 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 lycoplyte *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 pre-flowering 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.
Acknowledgements

We are grateful to Sven Gould, Günter Theißen and two anonymous reviewers for providing helpful comments on the draft. PW was supported by the ERA-CAPS SeedAdapt consortium project (www.seedadapt.eu, grant no. RE1697/8 to SAR).

Authors’ contributions

SAR conceived of the study, supervised it, wrote the paper and carried out evolutionary and phylogenetic analyses. KKU was in charge of setting up the genomic data. PKIW adapted the TAPscan tool, carried out TAP classification and analysed data. PKIW and KKU implemented the phylogenetic sampling. CM, KKU and SAR inferred gene trees. CM established the web interface. All authors contributed to writing the manuscript.
References


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 (Table S2). 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 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 (Figure S1), highlighting were 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 (Table S2) 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.
Fig. 2. TAPscan classification rules.

The name of each family of sub-family 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. Cf. Table S3 for more detailed classification rules.
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.
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 datasets covered per clade are shown in brackets. Gains and losses were predicted using PGL, expansions and contractions using Wagner parsimony (cf. Methods and Table S7). These predictions were entered into Table S6 (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 Table S7; symbols are shown to the right of triangles if they concern a distal node, and to the left if they concern a deep node.
### Table 1. Included species.

Species are divided into angiosperms, non-seed plants and algae. The data used in TAPscan v1 (Lang, et al. 2010) is compared to 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.

<table>
<thead>
<tr>
<th>Taxonomic group</th>
<th>Lang et al. (2010) V1</th>
<th>V2 2017</th>
<th>Unpublished genomes</th>
<th>Unpublished transcriptomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angiosperm - Core Eudicots/Core Rosids</td>
<td>Genomes: 7, Genomes: 46, Transcriptomes: 0</td>
<td>1 (Salix purpurea)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Angiosperm - Core Eudicots/Asterids</td>
<td>Genomes: 0, Genomes: 11, Transcriptomes: 0</td>
<td>1 (Kalanchoe laxiflora)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Angiosperm - Core Eudicots/Stem Rosids</td>
<td>Genomes: 0, Genomes: 4, Transcriptomes: 0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Angiosperm - Stem Eudicots</td>
<td>Genomes: 0, Genomes: 2, Transcriptomes: 0</td>
<td>1 (Aquilegia coerules)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Angiosperm - Monocots</td>
<td>Genomes: 3, Genomes: 22, Transcriptomes: 0</td>
<td>3 (Brachypodium stacei, Panicum virgatum, Setaria viridis)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Angiosperm – ANA grade (stem angiosperms)</td>
<td>Genomes: 0, Genomes: 1, Transcriptomes: 0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sub total angiosperms</td>
<td>Genomes: 10, Genomes: 86, Transcriptomes: 0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gymnosperm – Conifer</td>
<td>Genomes: 0, Genomes: 2, Transcriptomes: 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gymnosperm – Ginkgophyte</td>
<td>Genomes: 0, Genomes: 1, Transcriptomes: 0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Monilophytes – Leptosporangiote</td>
<td>Genomes: 0, Genomes: 2, Transcriptomes: 2</td>
<td>2 (Azolla filiculoides and Salvinia cucullata)</td>
<td>1 (Microlepia cf. marginata)</td>
<td></td>
</tr>
<tr>
<td>Lycophytes</td>
<td>Genomes: 1, Genomes: 1, Transcriptomes: 0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mosses</td>
<td>Genomes: 1, Genomes: 2, Transcriptomes: 2</td>
<td>1 (Sphagnum fallax)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Liverworts</td>
<td>Genomes: 0, Genomes: 1, Transcriptomes: 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sub total non-seed plants and gymnosperms</td>
<td>Genomes: 2, Genomes: 9, Transcriptomes: 6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Streptophytic Algae – Zygnematales</td>
<td>Genomes: 0, Genomes: 1, Transcriptomes: 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Streptophytic Algae – Coleochaetales</td>
<td>Genomes: 0, Genomes: 0, Transcriptomes: 3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Streptophytic Algae – Charales</td>
<td>Genomes: 0, Genomes: 1, Transcriptomes: 1</td>
<td>1 (Chara braunii)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Streptophytic Algae – Klebsormidiales</td>
<td>Genomes: 0, Genomes: 1, Transcriptomes: 0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Streptophytic Algae – Chlorokybales</td>
<td>Genomes: 0, Genomes: 0, Transcriptomes: 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Green Algae – Chlorophyta</td>
<td>Genomes: 7, Genomes: 13, Transcriptomes: 0</td>
<td>1 (Dunaliella salina)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sub total algae</td>
<td>Genomes: 7, Genomes: 16, Transcriptomes: 7</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| total                                  | Genomes: 19, Genomes: 111, Transcriptomes: 13 | | | 124
Supplementary Material

Supplementary Figures 1-7 provided as a file containing the legends and Fig. S2-S7, and Fig. S1 as an individual file.

Supplementary Tables: legends provided below, tables S1-S7 as individual files.

Supplemental Table 1. New and modified domains.
In the first four tabs, all new PFAM and custom domains used to enable the sub-family rules for HD, Jumonji, MADS and PcG are shown (column J in each tab). Tab five details the switch from a custom to a PFAM domain, and tab 6 lists all domain name changes.

Supplemental Table 2. Phylogenetically diverse genome set for custom domains.
The 46 genomes listed were used for the phylogenetic sampling runs described in Methods and Results.

Supplemental Table 3. TAPscan v2 rules.
Column A displays the name of the TAP family, B its type. C and D show the underlying rules, comments are added to column E. Altered and new rules are highlighted in dark green, name changes in orange. Sheet two lists the percentage HMM use cutoffs required for the sequence to be used (cf. Methods).
Supplemental Table 4. Gold standard comparison.

Several publications with phylogenetic definition of families were used as gold standard to determine sensitivity and specificity of TAPscan. Publications are listed in column C and in Results. The first tab contains data for TAPscan v2 (this study), the second tab the data for v1.

Supplemental Table 5. Datasets used.

All incorporated species are listed with type of resource (genome/transcriptome) and source.

Supplemental Table 6. TAPscan prediction of family members per species.

Sheet one lists all TAP families in column A, their type (TF/TR/PT) in column B, and their number of domains in column C. Number of family members per species is shown in the species column to the right. Sheet two uses the same columns A/B/C, followed by annotation of the corresponding family by several previous publications in columns D-AA, compared to the present study. In particular, column O-R contain the manually reviewed gain/loss/expansion/contraction data of the present study. Columns AC-BF contain groupwise comparisons of TAP data. Columns labeled in row five (like eudicots, column AC) contain average values calculated from the corresponding organisms from sheet one. Columns labeled in row 14 (like eudicots vs. monocots, column AD) contain q-values (if < 0.05) of Bonferroni-corrected one-side Mann Whitney U tests. Columns labeled in row 4-9 (like Amborella trichopoda, column AF) contain TAP families of individual species, if they are the only representative of that lineage. Fold change columns (blue header in row 18, like “fold angio- vs. before” in column AG) contain fold change calculations of the first mentioned group (here: angiosperms divided by the second group (here: all lineages that split off before). Fold changes >=2 (potential expansion in first group) and <= 0.5 (potential contraction on first group) are highlighted. Columns labeled “diff.”, like AI,
contain the difference of the fold change column to the left minus the fold change column to the right. Column AW (Charophyta / ZCC) and AZ (Charophyta / KCM) contains yellow highlighted cells for families that are present in the respective grade, but not in unicellular Archaeplastida that branch off before. Column AW contains red highlighted cells for families that are present in ZCC but not KCM algae; column AZ contains red highlighted cells for families that are present in KCM but not ZCC algae.

**Supplemental Table 7. Count output.**

Sheet one lists the node (leave) numbers and their names *(cf. Figure 5)*. Sheet two lists the count prediction of gains/losses/expansions/contractions per family.