Genome-Wide Phylogenetic Comparative Analysis of Plant Transcriptional Regulation: A Timeline of Loss, Gain, Expansion, and Correlation with Complexity

Daniel Lang[¶,1], Benjamin Weiche†[¶,2], Gerrit Timmerhaus‡[,1,2], Sandra Richardt§[,2], Diego M. Riaño-Pachón[3], Luiz G. G. Corrêa[‖,4], Ralf Reski[1,5], Bernd Mueller-Roeber[4,6], and Stefan A. Rensing*[,2,5]

[1]Plant Biotechnology, Faculty of Biology, University of Freiburg, Freiburg, Germany

[2]Faculty of Biology, University of Freiburg, Freiburg, Germany

[3]GabiPD team, Bioinformatics Group, Max Planck Institute of Molecular Plant Physiology, Potsdam-Golm, Germany

[4]Department of Molecular Biology, Institute of Biochemistry and Biology, GoFORSYS, University of Potsdam, Potsdam-Golm, Germany

[5]Freiburg Initiative for Systems Biology, Faculty of Biology, University of Freiburg, Freiburg, Germany

[6]Cooperative Research Group, Max Planck Institute of Molecular Plant Physiology, Potsdam-Golm, Germany

[†]Present address: Life & Medical Sciences Institute, Laboratory of Chemical Biology, Gerhard-Domagk-Straße 1, 53121 Bonn, Germany

[‡]Present address: Nofima Marin, Postboks 5010, 1432 Ås, Norway

[§]Present address: QIAGEN, Qiagen Strasse 1, 40724 Hilden, Germany

[‖]Present address: Fermentas, Opelstraße 9, 68789 St. Leon-Rot, Germany

[¶]These authors contributed equally to this work.

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

All sequence sources are listed in table 1.

Accepted: 4 June 2010

Downloaded from http://gbe.oxfordjournals.org/ at MPI Molec Plant Physiology on March 28, 2012

Abstract

Evolutionary retention of duplicated genes encoding transcription-associated proteins (TAPs, comprising transcription factors and other transcriptional regulators) has been hypothesized to be positively correlated with increasing morphological complexity and paleopolyploidizations, especially within the plant kingdom. Here, we present the most comprehensive set of classification rules for TAPs and its application for genome-wide analyses of plants and algae. Using a dated species tree and phylogenetic comparative (PC) analyses, we define the timeline of TAP loss, gain, and expansion among Viridiplantae and find that two major bursts of gain/expansion occurred, coinciding with the water-to-land transition and the radiation of flowering plants. For the first time, we provide PC proof for the long-standing hypothesis that TAPs are major driving forces behind the evolution of morphological complexity, the latter in Plantae being shaped significantly by polyploidization and subsequent biased paleolog retention. Principal component analysis incorporating the number of TAPs per genome provides an alternate and significant proxy for complexity, ideally suited for PC genomics. Our work lays the ground for further interrogation of the shaping of gene regulatory networks underlying the evolution of organism complexity.

Key words: transcription factor, evolution, Plantae, phylogenetic comparative methods, morphological complexity.

Introduction

The regulated expression of genes is essential for defining morphology, functional capacity, and developmental fate of both solitary living cells as well as cells inhabiting the social environment of a multicellular organism. In this regard, the regulation of transcription, that is, the synthesis of messenger RNA from a genomic DNA template, plays a crucial role. It contributes to the control of temporal and spatial RNA and protein levels in a cell and therefore has an essential function in all living organisms. Transcriptional regulation is primarily achieved by transcription-associated proteins (TAPs, comprising transcription factors [TFs] and

© The Author(s) 2010. 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/2.5), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

other transcriptional regulators [TRs]), which are especially attractive for the investigation of gene regulatory networks. The evolutionary development of organisms throughout all kingdoms of life seems to be tightly linked to the evolution and expansion of TAP gene families (Hsia and McGinnis 2003; Levine and Tjian 2003; Gutierrez et al. 2004; Carroll 2005). It has been proposed before that there is a direct correlation between the genomic fraction of TAPs and the morphological complexity of an organism (e.g., Levine and Tjian 2003) and that expansions of TAP families contribute to the evolution of morphological diversification (Lespinet et al. 2002; Richardt et al. 2007).

There are multiple lines of evidence suggesting that many, if not most, eukaryotic genomes underwent one to several large-scale duplication events in their evolutionary history (e.g., Paterson et al. 2006; Edger and Pires 2009; Van de Peer et al. 2009). Which genes are retained after such an event seems to be critical in terms of gene dosage balance, especially among members of signaling and regulatory networks. TAPs and DNA-binding TFs, in particular, belong to the functional classes of genes that have been found to be retained preferentially after duplication in most studies investigating large-scale and other "balanced duplications" events (Edger and Pires 2009). This finding has been discussed frequently as an indication for the importance of genome duplication events for the observable gain of morphological complexity in the animal and plant lineages (Freeling and Thomas 2006). Although the basic transcription machinery in different eukaryotes is essentially similar, the gene families regulating this machinery often show lineage-specific expansions (Lespinet et al. 2002) and are therefore ideally suited for analyzing taxonomic diversity (Coulson and Ouzounis 2003). In plants, due to the above-mentioned higher retention rate of TAPs after duplication events (Lespinet et al. 2002; Shiu et al. 2005), the contribution of TAPs to the total number of genes is more pronounced than in other eukaryotes (Riano-Pachon et al. 2008). Hence, especially in plants, the cross-species comparison of TAPs is expected to yield interesting insights into the evolution of regulatory networks.

Phylogenetic comparative methods (PCMs) allow the identification of evolutionary correlations across taxa (Quader et al. 2004). Statistical correlation of traits between different species without incorporation of their evolutionary relationships suffers from the problem of phylogenetic nonindependence as taxa may be similar simply due to shared ancestry. Thus, comparative data often violate the statistical assumption of independence. Given a species tree, PCMs can be used to correct comparative data for phylogenetic nonindependence (Felsenstein 1985; Pagel 1994; Garland and Ives 2000; Martins 2000). Although the hypothesis of a direct evolutionary correlation between TAPs and organismal complexity seems very intuitive and indirectly supported by findings from other fields of research, like, for example, the specific paleolog (i.e., paralogs retained after a paleoduplication event) retention pattern described earlier, it has not yet been put to test using comparative phylogenetic methods. Moreover, it is important for our understanding of TAP gene family evolution to elucidate whether all or only certain TAP families are correlated with complexity.

The availability of several annotated genomes covering most of the major clades along the red/green lineage (the Archaeplastida or Plantae; which have acquired their plastid by primary endosymbiosis; [Cavalier-Smith 1998; Adl et al. 2005]) now allows us to evaluate which traits have been important for the observed gain of morphological complexity. The findings reviewed above suggest that the TAP complement is an important trait to test in this context. Necessary prerequisites for these analyses are a detailed classification of the TAP complements of the genomes under investigation, a species phylogeny with branch lengths, divergence time estimates, and traits to describe organismal complexity.

TAPs can be divided into 1) TFs binding to *cis*-regulatory DNA elements in a sequence-specific manner, directly enhancing or repressing the transcription of their target genes; 2) TRs with indirect regulatory functions, assisting in the assembly of the RNA polymerase II complex (general TFs), functioning as scaffold proteins in enhancer/repressor complexes or controlling the chromatin structure by, for example, modifying histones or the DNA methylation state, respectively; and, finally, 3) putative TAPs (PTs) have so far not been functionally investigated but in silico prediction suggests a role in transcriptional regulation. For all groups, numerous members have been identified and described in different organisms, for example, there are extensive studies dealing with TAPs in the flowering plants *Arabidopsis thaliana* (Riechmann et al. 2000; Guo et al. 2005), *Oryza sativa* (Gao et al. 2006), *Populus trichocarpa* (Zhu et al. 2007), and *Nicotiana tabacum* (Rushton et al. 2008). Those studies revealed functional networks of transcriptional regulation in a particular system. However, for comparative evolutionary analyses as well as for gaining insight into the development of a wide variety of TAP families, a broader approach, covering a larger set of divergent species, is necessary. In order to identify TAPs in genomes, we classified proteins into families involved in transcriptional regulation. To this end, we applied an approach that exploits the domain structure of the protein for its characterization. Because protein domains fulfill a crucial role, mutations within domains are often deleterious. Therefore, such regions are strongly conserved, leading to highly similar sequences originating from a common evolutionary ancestor. Previous studies already applied a domain-based approach for defining protein families (Riano-Pachon et al. 2007; Guo et al. 2008). Domains relevant for TAP family classification were retrieved from those publications and used to establish rules defining domains mandatory or forbidden in proteins of a certain family. Further

Downloaded from http://gbe.oxfordjournals.org/ at MPI Molec Plant Physiology on March 28, 2012

relevant domains and rules were extracted from PlanTAPDB (Richardt et al. 2007) as well as from literature (Fernandez-Silva et al. 1997; Shuai et al. 2002; Hackbusch et al. 2005; Whitcomb et al. 2007; Yamada et al. 2008). By combining the existing rules and resolving potential conflicts between rules from different sources, we were able to enlarge the previously used sets of TAP domains and families. Here, we report the so far most extensive comparative phylogenetic analysis of TAP gene families in land plants and algae, revealing insights into the evolution of transcriptional regulation from unicellular to highly complex photosynthetic organisms.

Organismal or morphological complexity is often measured by the number of cell types or tissue types (Bell and Mooers 1997; Adami 2002; Hedges et al. 2004). However, the publication record of exact cell and tissue type estimates is scarce (Bell and Mooers 1997). Extended literature searches revealed no peer reviewed, experimentally determined absolute cell type or tissue type numbers for any of the sequenced green model species. Online resources like Plant Ontology (http://www.plantontology.org/) or BioNumbers (http://bionumbers.hms.harvard.edu) are beginning to conquer this dilemma but cannot yet guarantee completeness and accuracy. Thus, for more detailed taxon-rich analyses, an alternative proxy for organismal complexity is needed.

Using the phylogenetic framework of a 20 species phylogeny (based on a concatenated alignment of 14 nuclear-encoded markers) and subsequent molecular divergence time estimates, we here present the first comparative phylogenetic approach to better understand the evolutionary relationship and dependence of transcriptional regulation and morphological complexity in Viridiplantae. Employing a combination of principal component analysis (PCA) and PCMs, we derived a novel proxy for organismal complexity that allows us to assess more detailed evolutionary questions on broader taxonomic scale like, for example, which particular TAP families did expand in correlation with the general increase in morphological complexity.

Materials and Methods

Classification Rules

The rules for classifying the investigated proteins into TAP families define mandatory ("should") as well as forbidden ("should not") conserved protein domains in those families. The initial set of rules was adopted from three previous publications/databases, that is, PlantTFDB (Guo et al. 2008), PlnTFDB (Riano-Pachon et al. 2007), and PlanTAPDB (Richardt et al. 2007). In the latter case, domains present in more than 95% of the members of a given family were defined as a mandatory domain. Potential conflicts between the three sources were manually evaluated and subsequently solved, for example, via literature research, as in the case of PlnTFDB C2C2-GATA (tify; could) versus PlantTFDB (tify; should not). As the tify domain has been described to appear in GATA members (Reyes et al. 2004), the PlnTFDB rule was preferred. The sources for all rules are listed in supplementary table 1 (Supplementary Material online). The resulting combined set of preliminary classification rules was used in a test run with the *A. thaliana* TAIR 7 protein set. Comparison with the results of the above mentioned publications as well as reduction of the number of double-classified proteins by adding exclusion rules (see below) were steps to refine the rules and to create a unified classification set. This set was subsequently expanded based on literature reports of recently defined families or subfamilies. Eleven additional rules describing nine TAP families were derived from the literature (Andrianopoulos and Timberlake 1991; Burglin 1991; Kagoshima et al. 1993; Muller et al. 1995; Fernandez-Silva et al. 1997; Hackbusch et al. 2005; Da et al. 2006; Duncan et al. 2007; Whitcomb et al. 2007; Yamada et al. 2008). Furthermore, the examination of selected proteins from the Uniprot and PFAM databases led to the definition of five new families implemented via 12 rules. The sources of all rules are shown in detail in supplementary table 1 (Supplementary Material online). In cases where either 1 out of 2 domains was necessary and sufficient for assignment to the respective family (bZIP, HD-Zip, and GARP_ARR-B), an "OR" rule was applied (five rules in total). To render the rule set more robust, we furthermore implemented rules reducing the number of proteins classified into any two independent families. Based on the test run against *A. thaliana* and subsequent literature surveys, 30 rules were thus added. The rule set for each family consists of at least one entry defining a should rule, that is, a domain mandatory for that particular family. Additional entries may define further should or should not (forbidden) domains (fig. 2).

Hidden Markov Model Collection

All domains relevant for classifying the TAPs are represented by a "ls" ("glocal") hidden Markov model (HMM), that is, global with respect to the profile and local with respect to sequence (as compared with "fs" HMMs that are local with respect to both). Thus, the ls HMMs identify whole domains only. Because we used only fully sequenced genomes, identifying fragments of the relevant domains was neither necessary nor suitable for our approach. Accordingly, the use of ls HMMs reduced the number of false positive hits obtained during the search by limiting the identification of truncated domains (Perez-Rodriguez et al. 2010). If available, the HMMs were retrieved directly from the "PFAM_ls" database (Finn et al. 2008). For the remaining domains, HMMs were custom made using multiple sequence alignments (MSAs) to identify the conserved domains of interest.

Downloaded from http://gbe.oxfordjournals.org/ at MPI Molec Plant Physiology on March 28, 2012

The MSAs used for creating the custom-HMMs were downloaded from PlnTFDB (Riano-Pachon et al. 2007). For domains not represented in this database, MSAs were created as follows. Blast searches with a protein query containing the respective domain yielded homologous hits defined by having at least 30% sequence identity with the query over a minimum length of 80 amino acids (Rost 1999). Hits were aligned using MAFFT (Katoh et al. 2005) and manually curated using Jalview (Clamp et al. 2004). The conserved domain of interest was extracted and the HMM calculated with HMMER 2.3.2 (http://hmmer.janelia.org/) using "hmmbuild" with the default parameters to generate ls HMMs and subsequently "hmmcalibrate" with the option "–seed 0" for gaining reproducible results. Gathering (GA) cutoff values were defined for each custom-HMM. The GA was set as the lowest score of a domain-containing protein (true positive) after a "hmmpfam" search (using an E value cut off of 1×10^{-5}) against the full proteome sets of several different species and considering the alignments of all hits. In two cases (NOZZLE and NAC, substituting NAM), custom-HMMs were used despite their availability in the PFAM databases. Both custom-HMMs are able to detect more members of the corresponding families than the publicly available ones. All custom-HMMs are given in supplementary file 1 (Supplementary Material online).

Protein Sequences

In order to avoid sampling bias, only fully sequenced genomes were used in this study. For each organism, the complete set of proteins derived from conceptual translation of the nuclear gene models (using the filtered/selected model per locus) were combined with the proteins encoded by the respective mitochondrial and plastidal genomes, if available. All proteins can be unambiguously identified via their fasta id. We used a unique five-letter code for each organism (table 1) followed by "mt" (mitochondrial), "pt" (plastid), or "pl" (plasmid), if applicable, and the accession number of the gene model. In the case of splice variants (*A. thaliana*, *O. sativa*, *Medicago truncatula*, *Glycine max*, *Zea mays*, and *Carica papaya*), the model with the lowest index number per locus was chosen, as it usually represents the first model determined for that locus and therefore has the highest level of accuracy. The organisms, numbers of encoded proteins, genome versions/download times, institutions, and download links are mentioned in table 1.

Classification Procedure

Using all proteins of the investigated organisms as query, hmmpfam searches (from the HMMER v2.3.2 package, http://hmmer.janelia.org/) were performed against an HMM library containing all 124 domains necessary for the TAP classification (supplementary table 2, Supplementary Material online). The tool hmmpfam considers the HMM collection as the database to search against. Because we were focusing on TAPs, a restriction of the HMMs to the ones of interest reduced the number of observations to a reasonable size. Furthermore, the number of HMMs employed remained constant, in contrast to the number of proteins encoded by the genomes, therefore yielding comparable E values. In order to obtain a high level of specificity, hmmpfam searches were performed using GA (cutoff) values as the score cut off for domain hits. The PFAM GA is manually curated throughout the HMM building process. Besides the GA, two additional score cut offs are defined during the creation of a PFAM HMM. The noise cut-off (NC) represents a very relaxed criterion, whereas the trusted cut-off (TC) is very stringent. Those cut offs would lead to false positive assignments (NC) or the loss of identified true positive domains (TC), respectively. Therefore, TC and NC are regarded not suitable for genome-wide domain assignments (HMMer user guide). Because an arbitrary E value cut off would not yield trustworthy results as well, the GA is considered the best choice for our approach. GA values were either provided with the "PFAM" HMMs or defined as described above. The classification rules (fig. 2) were subsequently applied to all proteins for which at least one significant domain hit was found. In cases where the domain composition of a protein matched more than one classification rule, the should rule with the highest score determined the family into which the protein was categorized. Highly similar domains, which are often found in the same or overlapping regions of a protein, were treated in similar fashion, that is, the domain with the higher score was used for subsequent classification. This procedure was necessary in four cases, namely 1) Myb_DNA-binding and G2-like_Domain, 2) NF-YB, NF-YC, and CCAAT-Dr1_Domain, 3) PHD and Alfin-like, and 4) GATA and zf-Dof (fig. 2). In addition, a Boolean OR rule was applied to three families (bZIP, HD-Zip, and GARP_ARR-B) (fig. 2). In these cases, either 1 out of 2 domains was found to be necessary and sufficient for a protein to be classified into the corresponding family.

Statistical Testing

Significant expansion of individual families between different groups of organisms was analyzed using standard T-test with subsequent false discovery rate correction (Benjamini and Hochberg 1995). T-tests and PCA for figure 4 were performed using Expressionist Analyst v5.3.5 (GeneData).

Phylogenetic Methods

The predicted nuclear proteomes of the 20 Plantae species were clustered using BlastClust (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/) requiring 50% sequence identity and 70% coverage. Resulting clusters were filtered, selecting for clusters with only one gene in the genomes

Downloaded from http://gbe.oxfordjournals.org/ at MPI Molec Plant Physiology on March 28, 2012

Genome Biology and Evolution

SMBE

Table 1
Data Sources

Organism	Five-Letter Code	Number of Proteins[a]	mt/pt/pl	Version/ Download Time	Institution	Download
Arabidopsis thaliana	*ARATH*	27,235	1/1/0	TAIR 8	TAIR	ftp://ftp.arabidopsis.org/home/tair/Sequences/blast_datasets/TAIR8_blastsets/
Carica papaya	*CARPA*	27,544	0/1/0	Date: 18/02/09	University of Hawaii	ftp://asgpb.mhpcc.hawaii.edu/papaya/annotation/
Glycine max	*GLYMA*	66,293	0/1/0	v 1.0	JGI	ftp://ftp.jgi-psf.org/pub/JGI_data/Glycine_max/Glyma1/annotation/
Medicago truncatula	*MEDTR*	44,337	0/1/0	v 2.0	Medicago.org	http://medicago.org/genome/downloads/Mt2/
Populus trichocarpa	*POPTR*	45,654	0/1/0	v 1.1	JGI	http://genome.jgi-psf.org/Poptr1_1/Poptr1_1.download.ftp.html
Ricinus communis	*RICCO*	31,221	0/0/0	Date: 26/02/09	J. Craig Venter Institute	http://castorbean.jcvi.org/downloads.php
Vitis vinifera	*VITVI*	30,434	0/1/0	v 1	Genoscope	http://www.genoscope.cns.fr/externe/GenomeBrowser/Vitis/
Oryza sativa	*ORYSA*	56,441	1/1/1	v 5.0	TIGR	ftp://ftp.tigr.org/pub/data/Eukaryotic_Projects/o_sativa/annotation_dbs/pseudomolecules/version_5.0/all.chrs/
Sorghum bicolor	*SORBI*	36,015	1/1/0	Date: 12/01/09	JGI	http://genome.jgi-psf.org/Sorbi1/Sorbi1.download.ftp.html
Zea mays	*ZEAMA*	45,271	1/1/1	v 3b.50	Maizesequence.org	http://ftp.maizesequence.org/current/
Selaginella moellendorfii	*SELMO*	22,285	0/0/0	v 1.0 FM3	JGI	http://genome.jgi-psf.org/Selmo1/Selmo1.download.ftp.html
Physcomitrella patens	*PHYPA*	28,093	1/1/0	v 1.2	JGI	http://www.cosmoss.org
Volvox carteri	*VOLCA*	15,544	0/0/0	Date: 12/01/09	JGI	http://genome.jgi-psf.org/Volca1/Volca1.download.ftp.html
Chlamydomonas reinhardtii	*CHLRE*	14,675	1/1/0	v 3.1	JGI	http://genome.jgi-psf.org/Chlre3/Chlre3.download.ftp.html
Chlorella sp.	*CHLSP*	9,965	0/1/0	Date: 12/01/09	JGI	http://genome.jgi-psf.org/ChlNC64A_1/ChlNC64A_1.download.ftp.html
Micromonas pusilla	*MICP1*	10,475	0/0/0	v 2.0	JGI	http://genome.jgi-psf.org/MicpuC2/MicpuC2.download.ftp.html
Micromonas pusilla NOUM 17	*MICP2*	9,815	0/0/0	v 2.0	JGI	http://genome.jgi-psf.org/MicpuN3/MicpuN3.download.ftp.html
Ostreococcus lucimarinus	*OSTLU*	7,651	0/0/0	v 2.0	JGI	http://genome.jgi-psf.org/Ost9901_3/Ost9901_3.download.ftp.html
Ostreococcus tauri	*OSTTA*	7,829	1/1/0	v 2.0	JGI	http://genome.jgi-psf.org/Ostta4/Ostta4.download.ftp.html
Cyanidioschyzon merolae	*CYAME*	5,255	1/1/0	v 1.0	*C. merolae* consortium	http://merolae.biol.s.u-tokyo.ac.jp/download/
Guillardia theta (endosymbiont)	*GUITH*	632[b]	0/1/0	v 1.0	*G. theta* consortium	http://gib.genes.nig.ac.jp/single/index.php?spid=Gthe_NUCLEOMORPH
Aureococcus anophagefferens	*AURAN*	11,501	0/0/0	Date: 05/02/09	JGI	http://genome.jgi-psf.org/Auran1/Auran1.download.ftp.html
Phaeodactylum tricornutum	*PHATR*	10,157	0/1/0	v 2.0	JGI	http://genome.jgi-psf.org/Phatr2/Phatr2.download.ftp.html
Thalassiosira pseudonana	*THAPS*	11,566	1/1/0	v 3.0	JGI	http://genome.jgi-psf.org/Thaps3/Thaps3.download.ftp.html
Ectocarpus siliculosus	*ECTSI*	16,377	0/0/0	v 2.0	Genoscope	Cock et al. (2010)
Emiliania huxleyi	*EMIHU*	39,265	1/1/0	Date: 12/01/09	JGI	http://genome.jgi-psf.org/Emihu1/Emihu1.download.ftp.html

[a] Number of proteins denotes the sum of nuclear-encoded proteins plus those encoded by mt and pt, where applicable.
[b] Reduced nucleomorph genome; counted as belonging to the Rhodophyta (Van de Peer et al. 1996).

without evidence of recent polyploidization events (Arath, Carpa, Chlre, Volca, Ostta, Ostlu, Micp1, Micp2, Cyame, Poptr, Ricco, Vitvi, Phypa, Selmo, Sorbi, Orysa; see table 1 for full species names and supplementary table 5 (Supplementary Material online) for a detailed taxonomic profile of all 14 gene families). The protein sequences resulting in 13 clusters (supplementary table 5, Supplementary Material online) were aligned using M.A.F.F.T. L-INSI (Katoh

Downloaded from http://gbe.oxfordjournals.org/ at MPI Molec Plant Physiology on March 28, 2012

et al. 2009) and manually curated using Jalview (Waterhouse et al. 2009). Based on individual genes, trees were inferred by Neighbor-Joining as implemented in quicktree-SD (Howe et al. 2002; Frickenhaus and Beszteri 2008) using the ScoreDist distance matrix (Sonnhammer and Hollich 2005) with 1,000 bootstrap replicates and rooted at the longest internal branch. Possible in-paralogs within the polyploid species were reduced to one representative sequence based on evolutionary distance. For the small subunit (SSU) alignment, available RNA alignments were downloaded from the SILVA rRNA database (Pruesse et al. 2007). The alignments (available from TreeBase at http://purl.org/phylo/treebase/phylows/study/TB2:S10409) were combined into a partitioned data set which was used to infer the final species topology and branch lengths using MrBayes (Ronquist and Huelsenbeck 2003) with a mixed model (all: ratepr = variable; SSU:GTR rates = invgamma ngamma = 8 statefreqpr = dirichlet(1,1,1,1); proteins: rates = invgamma ngamma = 8; aamodelpr = mixed). The topology was constrained to reflect relationships based on current taxonomic literature constraining Brassicales (Arath, Carpa), Malpighiales (Poptr, Ricco), Fabidae (Poptr, Ricco, Medtr, Glyma), and tracheophytes (Selmo, Orysa, Zeama, Sorbi, Vitvi, Arath, Carpa, Poptr, Ricco, Medtr, Glyma).

The resulting species tree (available from TreeBase at http://purl.org/phylo/treebase/phylows/study/TB2:S10409) was used to estimate divergence times with the r8s software (Sanderson 2003), employing the procedures and recommendations as previously described ([Sanderson et al. 2004; Bell and Donoghue 2005; Hug and Roger 2007; Magallon and Castillo 2009; Wang et al. 2009], r8s manual). Age constraints were derived from previous analyses or reviews of fossil records (Bowe et al. 2000; Zimmer et al. 2007; Lang et al. 2008; Wang et al. 2009): red/green (1,000–1,400 Ma), chlorophytes/streptophytes (500–1,200 Ma), bryophytes/tracheophytes (400–700 Ma), lycophytes/spermatophytes (423–475 Ma), rosids (minimum age 89.5 Ma), and Liliopsida/eudicotyledons (125–300 Ma). Divergence times and 95% confidence intervals (CIs) as presented in figure 5 and supplementary figure S1 (Supplementary Material online) are based on the application of several methods implemented in r8s (LF, PL Powell, PL NPRS) including variants with fixed root ages (1,200–1,500 Ma) and fossil-based cross-validation for model selection using penalized likelihood. The output of all methods was combined to calculate mean divergence times and 95% CIs, which are shown in supplementary figure S1 (Supplementary Material online). The individual divergence time estimates of all applied strategies and summary statistics including CIs are listed in supplementary table 6 (Supplementary Material online).

The character matrix used for the PCM analysis is provided in supplementary table 7 (Supplementary Material online) and has been submitted to BioNumbers (bion 105322). miRNA (family) annotations are based on miRBase release 13.0 (http://www.mirbase.org). Drawing of phylo- and chronograms, ancestral state reconstruction, PCM, and PCA were carried out in R (http://www.r-project.org) using the R packages APE (Paradis et al. 2004) and GEIGER (Harmon et al. 2008). Pairwise PCM comparisons of single traits were carried out by Pearson correlation of phylogenetically independent contrasts (PICs) (Felsenstein 1985) and the application of the Brownian, Martens, Grafen, and PIC generalized least square (GLS) models as implemented by APE and GEIGER using linear, linear/log, and log/log transformed data. The best model was selected for each comparison using the Akaike Information Criterion as implemented in APE. In cases of missing data, the species tree was truncated using the drop.tip function of APE. Scaled PICs were used to derive principal components in the search for a proxy of organismal complexity. Ancestral states were reconstructed for all traits of the character matrix (supplementary table 7, Supplementary Material online) comparing the GLS, PIC, and maximum likelihood methods implemented in APE. The reconstructed ancestral states of all methods were plotted together with the extant states at the nodes of the species tree (e.g., supplementary files 2 and 3, Supplementary Material online) and analyzed manually to derive the final set of gains and losses presented in figure 5.

Results

Generation of a Comprehensive Set of Classification Rules

For classifying proteins into TAP families, we applied a set of rules defining whether a certain domain is mandatory (should) or forbidden (should not) in a given family. The majority of these rules were extracted from three publications/databases dealing with TAPs in plants (Riano-Pachon et al. 2007; Richardt et al. 2007; Guo et al. 2008). Rules describing optional domains were excluded for brevity. In total, 83 rules describing 63 TAP families could be extracted from PlantTFDB (Guo et al. 2008), whereas 103 rules defining 68 families were taken from PlnTFDB (Riano-Pachon et al. 2007). PlantTAPDB (Richardt et al. 2007) was employed as the third major source of rules. TAP family members from this database were examined for the occurrence of domains present in more than 95% of the respective family members. Using this approach, 51 rules representing 48 families were obtained. The contribution of each source to the complete rule set (fig. 1) emphasizes the presence of common as well as unique rules derived from the different sources.

As a general meta-rule, we assigned higher importance to sequence-specific DNA-binding domains present in TFs than to domains classifying TRs. Thus, whenever a combination of domains leads to multiple possible family classifications, the TF family is favored over TR and PT, based on the

Downloaded from http://gbe.oxfordjournals.org/ at MPI Molec Plant Physiology on March 28, 2012

Genome Biology and Evolution

SMBE

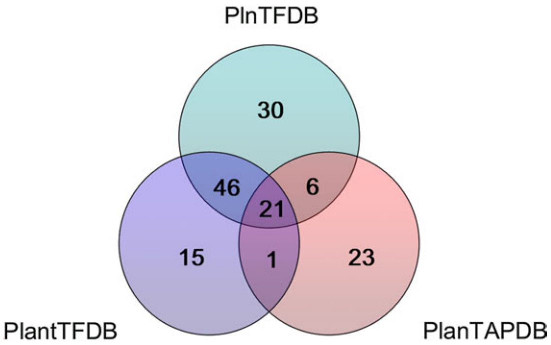

Fig. 1.—The sources of the TAP classification rules (supplementary table 1, Supplementary Material online) are depicted in a Venn diagram.

evolutionary perspective that the majority of *cis*-specific TF domains have been acquired as differentiations to introduce DNA specificity to more generally acting TRs with protein-interaction domains. This was encountered in 14 cases, resulting in 14 rules. For the homeobox TFs, several subfamilies (HB, HB-KNOX, HD-Zip) were defined by another nine rules. Taken together, the complete rule set (fig. 2) defines 111 TAP families by using 223 rules comprising 134 "mandatory" and 89 "forbidden" rules. In total, 124 domain HMMs are being used, 16 of which are custom made (supplementary file 1, Supplementary Material online), and 108 obtained from the PFAM database (supplementary table 2, Supplementary Material online).

Comparison with Other Data (Quality Check)

In order to assess the quality of our identification and classification approach (detailed in Materials and Methods), we compared our results with selected publications in which detailed phylogenetic analyses had been carried out for TAP families of the plant species *A. thaliana* and *O. sativa* and of the green alga *Chlamydomonas reinhardtii.* We refer to the reported TAP classification as the "gold standard" in the following. In addition, we included results from a recent report describing TAPs in the two *Micromonas* strains so far fully sequenced (Worden et al. 2009). In cases where the genome annotation version was the same as the one employed in the current study, we compared the coincidences of protein IDs between the respective data sets (ours and the gold standard). In the case of deviating versions, we compared the actual protein sequences. Following these guidelines, we were able to compute the sensitivity and the positive predictive value (PPV) of our approach, as previously described (Iida et al. 2005; Riano-Pachon et al. 2007). The comparison of the classification results presented in this study with those chosen as gold standard is shown in supplementary table 3 (Supplementary Material online). For the two seed plants, *A. thaliana* and *O. sativa*, the average sensitivity is 0.94 and 0.86, respectively. For the green alga *C. reinhardtii*, it is 0.93, whereas for the two *Micromonas* strains, it is only 0.65 and 0.56, respectively. This might be due to differences in the methodology employed and the relatively small average size of the *Micromonas* TAP families, resulting in a comparatively large amplitude of sensitivity and PPV even if only a few classifications differ.

Genome-Wide TAP Annotation

The summary of TAP classification results for (formerly) plastid-bearing organisms (supplementary table 4, Supplementary Material online) enables us to identify trends during the evolution of photosynthetic eukaryotes. The absolute amount of TAPs per genome shows an extensive expansion of TAPs between algae and land plants and between the nonseed and seed plants (fig. 3*A*), which is congruent with previous results based on fewer organisms (Richardt et al. 2007). The haptophyte *Emiliania huxleyi* appears to be an exception from this trend. However, when displaying the data relative to the coding potential of the genome (fig. 3*B*), thus smoothing effects that might be due to large-scale gene duplication events, the *E. huxleyi* TAP complement seems to follow the mentioned trend as well. Furthermore, it can be deduced that TFs in particular were subject to expansion during plant evolution, which has been suggested earlier (Richardt et al. 2007). PCA of the phylogenetically uncorrected TAP family sizes results in separation according to taxonomic groups, as shown, for example, for seed and nonseed plants, or algae derived from primary and secondary endosymbiosis, respectively (fig. 4). Based on a smaller data set, a trend correlating multicellularity and global TAP amount had been suggested before (Richardt et al. 2007). Using the uncorrected, phylogenetically dependent comparative data, there is no clear-cut trend of this kind in any of the present visualizations.

A Timeline for the Evolution of Plant Transcriptional Regulation

To ensure a reliable framework for phylogenetic comparative (PC) analyses, we derived a novel marker set of 13 nuclear single-copy, protein-coding orthologs, which was employed together with the small ribosomal subunit DNA (SSU; supplementary table 5, Supplementary Material online) to infer a phylogenetic tree of the Plantae (red/green lineage; [Cavalier-Smith 1998; Adl et al. 2005]) comprising the 20 sequenced species analyzed in this study. Subsequently, the nodes on the tree were dated by relaxed molecular clock approaches (Sanderson 2003) using fossil constraints and protocols from the literature estimating CIs by averaging over time points obtained from multiple methods (supplementary table 6, Supplementary Material online; [Sanderson et al. 2004; Bell and Donoghue 2005; Hug and Roger 2007; Magallon and Castillo 2009; Wang et al. 2009]). The divergence

Downloaded from http://gbe.oxfordjournals.org/ at MPI Molec Plant Physiology on March 28, 2012

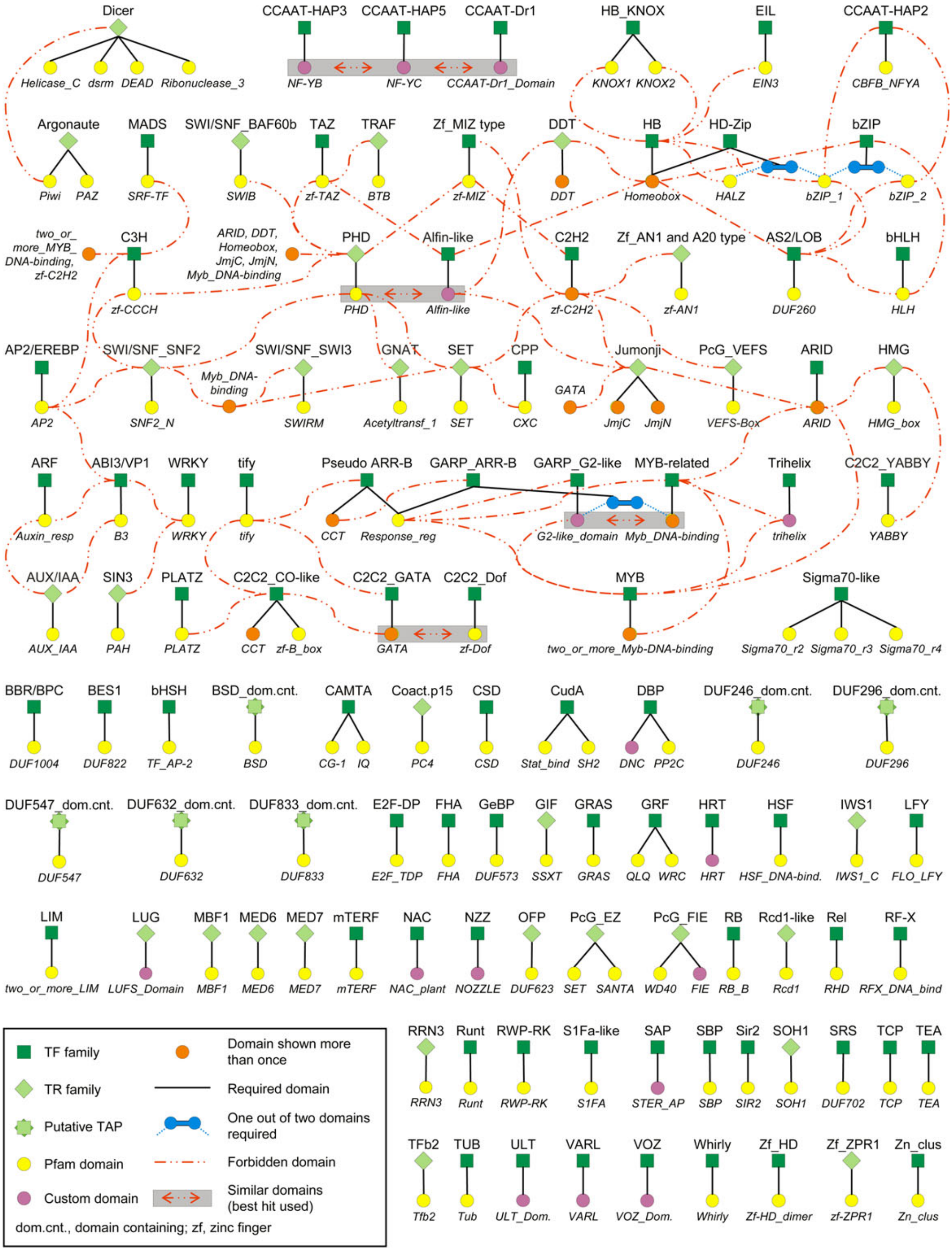

Fig. 2.—Visualization of the TAP classification rules (supplementary table 1, Supplementary Material online). Family names may contain blanks ( ), slashes (/), or dashes (-) as separation marks. Subfamily names are separated from the main family name by an underscore (_). See inset box for explanation of symbols.

time estimates of the resulting chronogram (supplementary fig. 1, Supplementary Material online) are in good agreement with previous reports (Kenrick and Crane 1997; Bowe et al. 2000; Hedges et al. 2004; Yoon et al. 2004; Zimmer et al. 2007; Lang et al. 2008; Wang et al. 2009). The tree was used to trace and visualize

Downloaded from http://gbe.oxfordjournals.org/ at MPI Molec Plant Physiology on March 28, 2012

Downloaded from http://gbe.oxfordjournals.org/ at MPI Molec Plant Physiology on March 28, 2012

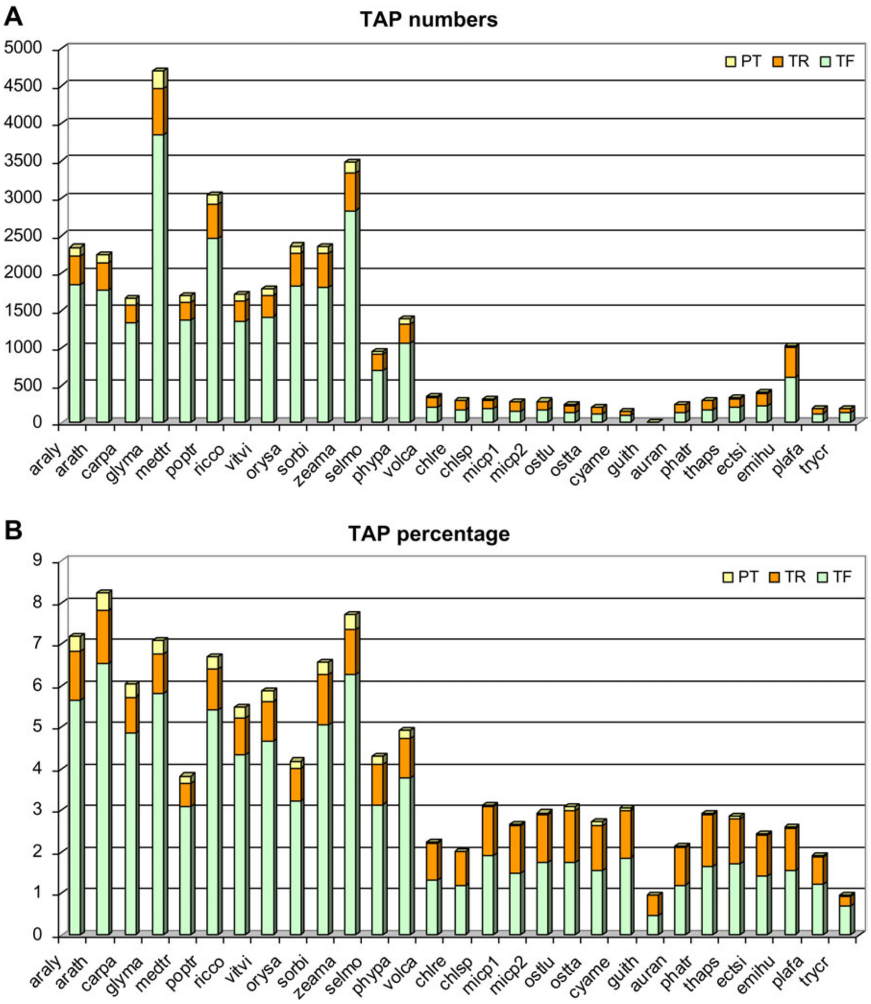

Fig. 3.—(*A*) Absolute numbers of TAPs, subdivided as stacked bars depicting TFs (green), TRs (orange), and PTs (yellow), are shown per genome. Species abbreviations: see table 1. (*B*) Relative (percentage of total number of encoded proteins) amounts of TAPs, subdivided as stacked bars depicting TFs (green), TRs (orange), and PTs (yellow), are shown per genome. Species abbreviations: see table 1.

extant and ancestral character states of TAP gene family evolution in Plantae. The total number of TFs, TRs, and PTs encoded by the respective extant and ancestral genomes are visualized in figure 5. Manual inspection of the individual ancestral state reconstructions for all individual TAP families resulted in gain, loss, and expansion estimates for the individual nodes, which were integrated into a global view providing a detailed timeline of TAP gene family evolution in the green lineage (fig. 5). The data from the sole red alga *Cyanidioschyzon merolae* was used as an outgroup to elucidate TAP evolution in the green lineage (Viridiplantae). A total of 21 TAP families (of which 16 are TF) arose within the earliest land plants (500 Ma, megaannum) or in their aquatic ancestor. Three further TAP families arose in the last common ancestor (LCA) of vascular plants 470 Ma (TFs: BBR/BPC and DBP; PT: DUF246), and three more TFs (C2C2_YABBY, GeBP, and ULT) in the LCA of extant angiosperms (or seed plants) 210 Ma. Two TF families specifically arose within eudicotyledons (NZZ and SAP; 143 Ma), and one is an invention of the lineage leading to the extant Volvocales (VARL; 47 Ma). As has been noted before (Riano-Pachon et al. 2008), the TF families, CAMTA and Trihelix, were secondarily lost from the genomes of all algae around 600 Ma. Red algae and prasinophytes share the loss of Alfin-like, Argonaute, and MBF1 genes. Green algae and prasinophytes have apparently lost the TR families DDT, SWI/SNF_SWI3, and the green algae have lost the TF family LIM.

In terms of expansions, a total of 44 TAP families are larger in size in land plants than in algae (fig. 5). The size of three TF families (EIL, GRF, and SRS) increased with the onset of vascularity; 23 TAP families (of which 18 are TF) are expanded in angiosperms (or possibly seed plants). Within the green algae, the TF families, RWP-RK and SBP, were expanded and the TR family TRAF in the Volvocales. Among the angiosperms, the TR family Sin3 was found

Genome Biology and Evolution
SMBE

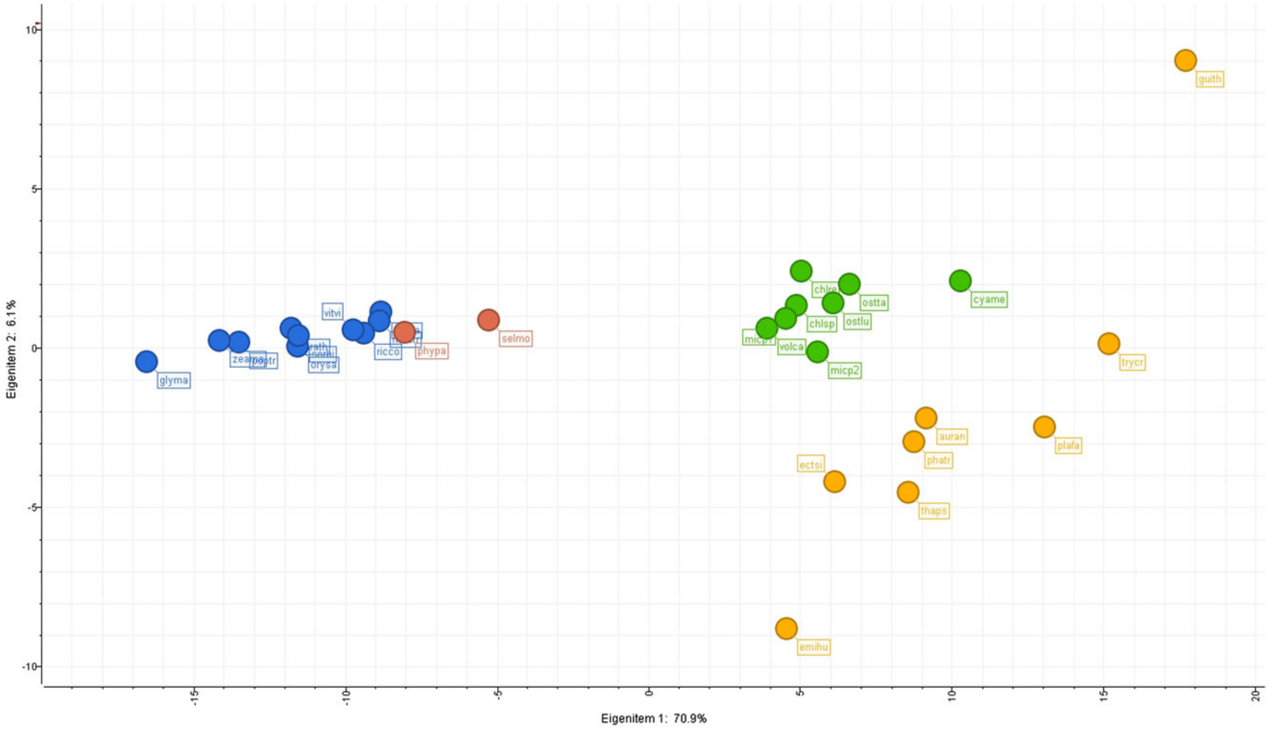

Fig. 4.—The PCA was performed on the absolute numbers of TAPs per family (uncorrected, phylogenetically dependent). A 2D plot of eigenitem 1 versus 2 is shown, with coloring according to taxonomic properties (blue: seed plants; red: nonseed plants; green: plastid derived from primary endosymbiosis; yellow: plastid derived from secondary endosymbiosis).

to be expanded among the rosids, the TR families TRAF and OFP in Poales, the TF family Alfin-like in the Panicoideae, and the MADS TF family in the Brassicales and Fabales.

Correlated Evolution of Transcriptional Regulation and Organismal Complexity

Both, the emergence and expansion of TAP families during land plant evolution, suggest a clear trend of increasing transcriptional complexity along with morphological complexity. An obvious and commonly used proxy for organismal complexity is the number of cell types (Bell and Mooers 1997; Carroll 2001; Hedges et al. 2004; McCarthy and Enquist 2005; Vogel and Chothia 2006; Xia et al. 2008). We were able to gather cell type count estimates for 12 of the organisms under study from literature, public databases, and through personal communications (supplementary table 7, Supplementary Material online). The application of this reduced taxon set to answer the question of evolutionary correlation between TAPs and morphological complexity (i.e., number of cell types), using both PC correlation analysis with PICs and regression analysis with the best of several phylogenetic GLSs models, confirms the initial hypothesis. The evolutionary pattern of the overall number of TAPs as well as the numbers of TFs and PTs show significant positive correlation with the number of cell types (TAPs: $R = 0.95$, P value best GLS model and correlation $<<0.01$; TF: $R = 0.96$, both P values $<< 0.01$; PT: $R = 0.94$, $P << 0.01$). TRs on the other hand show only weak correlation with the number of cell types ($R = 0.68$, $P < 0.05$).

Like multicellularity, increases in mean and maximal organismal size have occurred multiple times in the evolution of uni- and multicellular life forms (Carroll 2001), which has often been discussed as a general evolutionary trend toward an increase of body size. Although this trend could only be proven within some lineages (Carroll 2001), body size has been shown to be positively correlated with the number of cell types in metazoans (McCarthy and Enquist 2005) and therefore might provide an indirect proxy for complexity. Initial correlation analysis confirms this trend for the reduced Plantae set with available cell type estimates ($R = 0.84$, $P < 0.001$). If the reported maximum size is used as a proxy for complexity it mirrors the above demonstrated correlation of TAPs and number of cell types ($R = 0.83$, $P < 0.05$). However, if the analysis is extended to the full set of 20 Plantae taxa, this relationship is not significant anymore (e.g., TAPs $R = 0.31$, $P = 0.2$).

It would certainly be premature to falsify the long-standing hypothesis based on this data only. Instead, we needed an alternative proxy that allowed us to cover the entire taxonomic range of Plantae. To achieve this, we combined PC analysis with PCA, which allows combining several,

Downloaded from http://gbe.oxfordjournals.org/ at MPI Molec Plant Physiology on March 28, 2012

Genome Biology and Evolution
SMBE

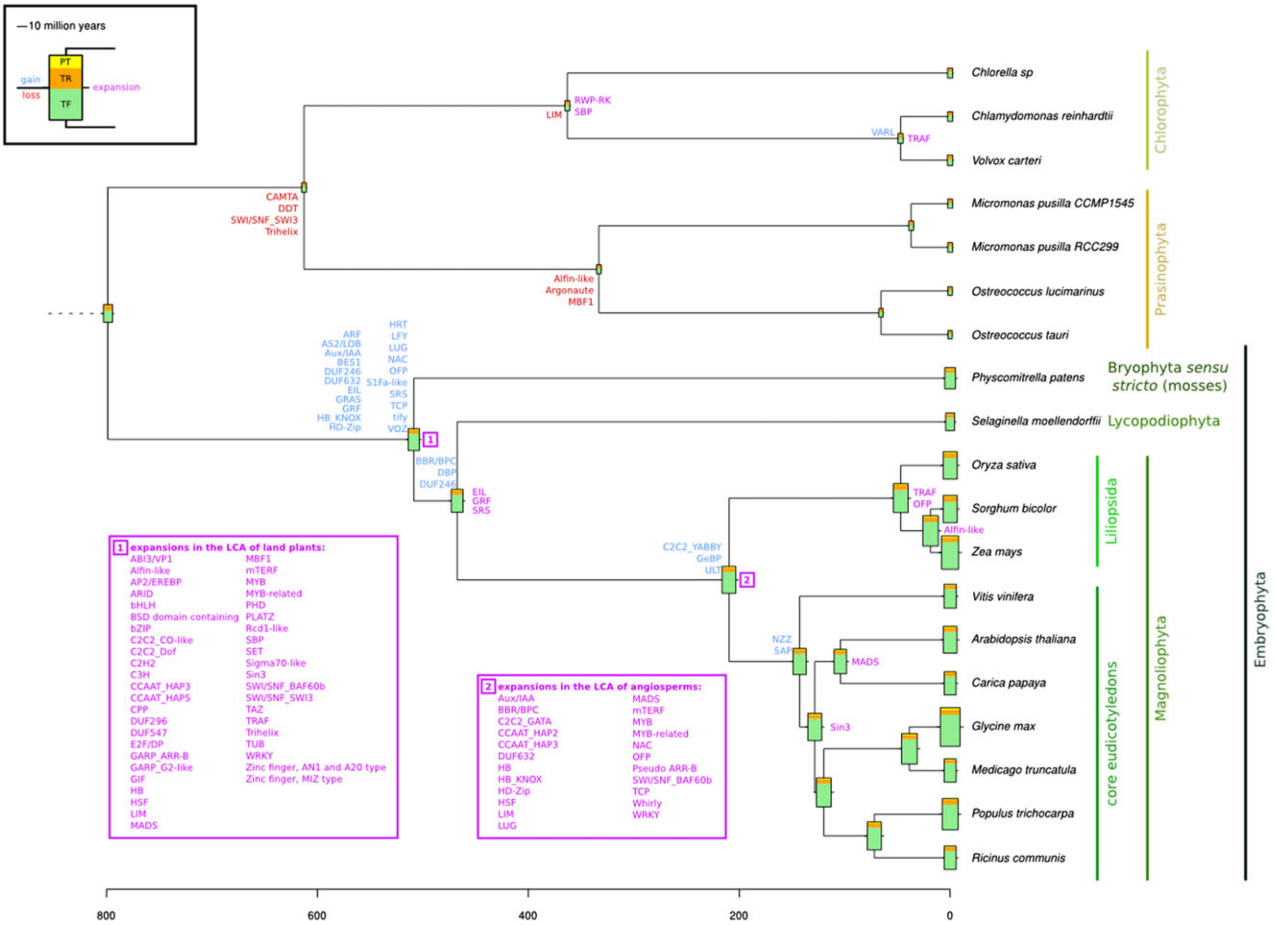

Fig. 5.—Extant and reconstructed ancestral numbers of TFs (green), TRs (orange), and PTs (yellow) per genome. Size of the boxes corresponds to the number of TAPs. The root used for the character tracing, *C. merolae*, was removed for brevity. The scale bar is in Ma.

possibly correlated traits into a smaller number of uncorrelated variables. To define traits suitable for establishing an improved proxy for organismal complexity, we first carried out PCA on the reduced data set using PICs of the following traits: cell types, numbers of TAPs, TFs, TRs, or PTs, genome size, number of genome duplications, and reported maximum body size (and length of the sporophyte). The number of cell types as well as the numbers of TFs, PTs, and TAPs contributed most to the variance represented by the first component (>95%). In the next step, the (difficult to determine) cell type trait was excluded from the PCA, and the first component was employed to test for correlation with the number of cell types. In fact the number of cell types shows strong correlation with the PCA proxy ($R = 0.95$, $P << 0.01$).

For the general applicability of this proxy, a reduction of incorporated variables would be desirable. Therefore, we repeated the PCA only with the three traits that contributed most to the first principal component (TAPs, TFs, and PTs). In this case, the positive correlation with the number of cell types is significant as well ($R = 0.95$, $P << 0.01$). Thus, the first component (combining the PICs of the total number of TAPs, TFs, and PTs) provides an excellent proxy for organismal complexity, ideally suited for PC genomics approaches on a larger taxonomic scale. If we apply this proxy to look for patterns of correlated evolution of TAP gene families and complexity on the complete Plantae data set, we find that all but 10 TF and 8 TR families show significantly correlated evolution with complexity ($q \leq 0.05$; supplementary table 8, Supplementary Material online). Among the TF families that seem to deviate from the general trend are the ABI3/VP3, CCAAT Dr1, and VARL families. Examples for TRs without obvious correlation are Dicer, Tfb2, and Sin3.

Importance of Genome Duplication Events

By employing the first principal component as a proxy for complexity, we find a significant correlation ($R = 0.78$, $P = 0.0046$) for the relationship between whole-genome duplication (WGD) events and organismal complexity among Plantae. Along the individual plant lineages, the expansion patterns of 27 TAP families (22 TFs, 3 TRs, and 2 PTs; supplementary table 9, Supplementary Material online) display significant correlation with the number of paleoploidy

Downloaded from http://gbe.oxfordjournals.org/ at MPI Molec Plant Physiology on March 28, 2012

Genome Biology and Evolution SMBE

events ($q < 0.01$). These finding clearly backs previous predictions about the importance of WGDs for organismal complexity and diversity (Crow and Wagner 2006; Freeling and Thomas 2006; Freeling 2008; Edger and Pires 2009; Soltis PS and Soltis DE 2009; Van de Peer et al. 2009). However, the number of WGDs alone was below the inclusion cut off for the PCA to derive a complexity proxy, and the relationship in the correlation analysis was found to be relatively weak (as compared with TAPs).

Discussion

Genome-Wide TAP Classification and Comparison

Previous comparative studies revealed important information about TAP evolution. They were, however, facing different problems. Due to the lack of sufficient numbers of fully sequenced genomes, such analyses could only be performed with a small number of organisms (Riano-Pachon et al. 2007; Richardt et al. 2007). Additionally, in some cases, it was necessary to retrieve sequences from expressed sequence tag databases (Richardt et al. 2007; Guo et al. 2008), which may lead to biased TAP classifications because of lacking sequence data for such organisms. Recently, however, many genomes of plants and algae became available providing a solid basis for the study of gene family evolution. The combined and updated TAP classification rule set presented here (fig. 2) is the most comprehensive described for plants so far. Our rigorous classification procedure, employing a set of PFAM domain-specific models and manually curated GA cut offs performs with high specificity and sensitivity, as shown by comparison with data reported for well-annotated Plantae genomes. The rule set presented here is expected to yield accurate results with regard to species of the red/green lineage. As there are no well-analyzed genomes of nongreen algae with regard to their TAP complement, no detailed comparison is possible yet. The uncorrected, phylogenetically dependent comparative studies show that there is an extensive expansion of TAPs, mainly TFs, during evolution from algae to land plants and from nonseed to seed plants (fig. 3*A*). PCA of the data is able to correctly separate taxonomic groups (fig. 4). However, these analyses fail to clearly correlate multicellularity with TAP complexity. The same conclusion was reached in a more detailed comparison of the TAPs encoded by the genome of the multicellular brown alga *Ectocarpus siliculosus* with those of unicellular heterokonts (Cock et al. 2010) and of the multicellular green alga *Volvox carteri* as compared with its unicellular sister taxa (Prochnik et al. 2010).

Phylogenetically Independent Tracing of Green Lineage TAP Evolution

The set of nuclear orthologs developed and employed here allows the robust dating of divergence times among Plantae. This data set can be expanded as further genomes become available and thus represents a valuable basis for acquiring divergence time estimates. The timeline of TAP evolution in the Viridiplantae (gain, loss, and expansion analyses, fig. 5) demonstrates that the major bursts of TF expansion occurred in the LCA of angiosperms 210 Ma. Although the lack of data for, for example, ferns and gymnosperms might convolute the picture, the interlinked increase of TF and flower complexity might well help to explain Darwin's "abominable mystery" (Busch and Zachgo 2009).

Correlation of TAPs with Morphological Complexity

Unfortunately, the exact number of cell types is not yet available for most of the organisms under study here. For most of the vascular plants (except *O. sativa*, *Z. mays*, and *A. thaliana*), we failed to get any estimates at all. Therefore, the resulting data set might provide a biased picture of complexity. Yet, by applying PC genomics, we can demonstrate for the first time that the total complement of TAPs is positively correlated with morphological complexity as measured by the number of cell types. This is in contrast to the analysis of the phylogenetically dependent, uncorrected data mentioned above (fig. 4) and demonstrates that genomic-scale PCM are necessary in order to detect otherwise convoluted evolutionary signals.

While TFs mirror this correlation, TRs do not. Therefore, paleolog retention of TFs (that bind in sequence-specific fashion to *cis*-regulatory elements) occurs more often than within TRs (that interact with DNA and proteins, including TFs, in order to regulate transcription). The fact that the size of the PT families described here is also positively correlated with the number of cell types suggests that they actually include TFs. As mentioned above, the number of cell types as a proxy for organism complexity is difficult to track down, and the maximum and average body size apparently represent a less than optimal proxy. Here, we can show that a principal component comprising the total number of TAPs, TFs, and PTs (as defined in this study) can be used as a proxy for organismal (morphological) complexity, as they are significantly positively correlated with the number of cell types in those organisms where data are available. Therefore, genome-wide determination of the TAP complement can serve as an indicator of morphological complexity, allowing to trace this trait in case it is hidden or, for example, in metagenomic studies where the identity of the contributing species is not always known.

Although there is an observable gain of morphological complexity in the evolution of Plantae and Metazoa, there have been discussions whether this actually represents a global, unidirectional, upward trend. Our ancestral state reconstructions indicate cases of both, increases and decreases. For example, the ancestral states of the maximal reported organism size and genome size (supplementary

Downloaded from http://gbe.oxfordjournals.org/ at MPI Molec Plant Physiology on March 28, 2012

files 2 and 3, Supplementary Material online) suggest secondary reductions along the lineages leading to the different unicellular algae under study here.

Correlation of TAP Expansion with Multicellularity?

In terms of TAP gain and expansion patterns, no clear marker for multicellularity emerges from our analyses. However, the pattern of initial expansion concomitant with the development of multicellularity might be obscured by subsequent expansions within the multicellular lineages as they developed more tissues and cell types. Yet, some families, upon close scrutiny, might offer hints for the development of land plant multicellularity, such as TAP families that are not encoded by the genomes of the green algae and prasinophytes analyzed in this study (fig. 5). The TR family DDT is not well characterized, it contains the DDT (DNA-binding homeobox and different TFs) domain that has been proposed to bind to DNA (Doerks et al. 2001). DDT is encoded in single copy by the genomes of some unicellular organisms (e.g., *C. merolae* and several heterokonts). The trihelix TFs appear to be involved in a plethora of specialized functions in seed plants, for example, abiotic stress tolerance (Xie et al. 2009), ploidy-dependent cell growth (Breuer et al. 2009), repression of seed maturation (Gao et al. 2009), and perianth architecture (Brewer et al. 2004). Next to the lack of these gene families in the green algal genomes studied here, the only genomes that encode more than one DDT and trihelix gene, respectively, are those of land plants and multicellular animals, suggesting a possible involvement in transcriptional regulation of cell-to-cell interactions within these groups of multicellular organisms.

The Importance of Whole-Genome Duplications

The comparatively weak correlation of WGD with organism complexity might be due to the current genome sampling bias that excludes major lineages like ferns, gymnosperms, and charophytes. However, it might also suggest that not only large-scale events but also small-scale or balanced segmental duplications are apparently important driving forces in the evolution of transcriptional regulation and complexity in Plantae. This ambivalence has been reported earlier (Crow and Wagner 2006) and might be a reflection of what is observed in animal evolution, where the inclusion of fossil taxa does not provide support for hypotheses linking genome duplications to the evolution of complexity in vertebrates (Donoghue and Purnell 2005). In Plantae, WGDs have been implicated before to be positively correlated with the rise of morphological complexity and the adaptive radiation of angiosperms (reviewed, e.g., by Soltis PS and Soltis DE [2009]; Van de Peer et al. [2009]). Moreover, they have been suggested to be correlated with geological upheaval periods such at the Cretaceous–Tertiary boundary (Fawcett et al. 2009), although this hypothesis might be disputable (Soltis and Burleigh 2009). Both might be realized through the potential for subfunctionalization and neofunctionalization that a WGD event allows for. Here, we show a significant positive correlation between the size of the TAP complement and the number of paleopolyploidizations that supports this hypothesis. The application of the TAP PCA proxy also reveals a weak correlation between genome size and complexity ($R = 0.69$, $P < 0.1$), which might be related to the trend observable for WGDs. The mitochondrial ($R = -0.33$, $P = 0.29$) and plastid ($R = 0.08$, $P = 0.77$) genome sizes do not follow this trend, neither do the reported maximum body sizes ($R = 0.28$, $P = 0.24$) nor the more detailed maximum sizes of the sporophyte ($R = 0.28$, $P = 0.24$), respectively, gametophyte ($R = -0.08$, $P = 0.83$).

Evolutionary Importance of miRNAs

The phylogenetic framework described here opens the door to address additional important evolutionary questions. For example, there is a growing body of evidence that miRNAs (which often target TAPs) are also a viable causal factor for the increase in morphological complexity (Li and Mao 2007; Lee et al. 2007; Heimberg et al. 2008). The evolutionary pattern of miRNA families was shown to coincide with the advent of morphological complexity in vertebrates (Heimberg et al. 2008). We find initial phylogenetic evidence for this pattern to be true for the evolution of Plantae as well. The number of miRNAs (supplementary table 6, Supplementary Material online) correlates with organismal complexity ($R = 0.93$, $P < 0.05$) and with the complement of TAPs ($R = 0.93$, $P < 0.05$). The data provide initial evidence only because miRNA annotations for the genomes under investigation vary in their completeness and do not provide the same level of coverage as we now have for the TAP gene families.

Correlated Evolution within Gene Families

Pairwise comparisons of the evolutionary pattern of individual miRNA and TAP families and other traits like, for example, sporophyte size can provide interesting hypotheses for further experiments. Examples for this are the miRNA family MIR390 and the TF families ABI3/VP3 and BBR/BPC, which show significant correlated evolution ($R > 0.6$, $P < 0.01$) with the size of the sporophytes. Furthermore, patterns of correlated evolution between gene families can be indicative of functional relationships (e.g., members of a protein complex) or regulatory roles (e.g., repressor or activator functions). As an example for such a correlation, we find a significant correlation between expansion patterns of the gene families coding for Aux/IAA TRs and ARF TFs ($R = 0.97$, best model and correlation $P << 0.01$), which are known to dimerize to regulate the transcription of auxin-responsive genes in plants (Paponov et al. 2009).

Downloaded from http://gbe.oxfordjournals.org/ at MPI Molec Plant Physiology on March 28, 2012

Conclusions and Outlook

Using a comprehensive set of classification rules and genomes, we show for the first time that the observable increase in morphological complexity in Plantae is positively correlated with the expansion of their TAP complement and especially TFs. Large-scale or WGD events are confirmed as major driving forces behind transcriptional and morphological complexity. The evolutionary pattern of miRNAs, which also act as important TRs and often regulate TFs, reveals correlated evolution with TAPs and morphological complexity. It will be exciting to test whether this pattern also holds true for the evolution of *cis*-regulatory elements and other proteins involved in signalling cascades. Together with the wealth of available and upcoming plant genome sequences, the TAP classification scheme and the phylogenetic framework developed in this study provide a powerful resource to address a plethora of evolutionary questions on a genome-wide scale. Yet, the currently available taxon sampling in terms of completely sequenced genomes is biased toward angiosperms, green algae, prasinophytes, and some groups within the heterokonts. In order to unravel how the lineage leading to extant land plants managed (in terms of transcriptional regulation) to become multicellular, we are in need of sequences from other (multicellular) algal genomes that are more closely related to extant land plants, that is, of Charophyta. In addition, other huge gaps have to be closed, namely within the red algae, liverworts, hornworts, ferns, and gymnosperms.

Supplementary Material

Supplementary files 1–3, figure S1, and tables 1–9 are available at *Genome Biology and Evolution* online (http://www.oxfordjournals.org/our_journals/gbe/).

Acknowledgments

We are indebted to the consortia that have made the current wealth of genome drafts available. Funding by German Research Foundation (RE 837/10-2 to R.R. and S.A.R.) and German Federal Ministry of Education and Research (Freiburg Initiative for Systems Biology, FKZ 0313921 to R.R. and S.A.R.; GoFORSYS, FKZ 0313924 to B.M.R.) is gratefully acknowledged. D.M.R.P. acknowledges financial support by the German Federal Ministry of Education and Research (GABI-FUTURE, FKZ 0315046). We are indebted to Jo Ann Banks, Virginia Walbot, and Paul Rushton for their help on the cell type number estimates.

Literature Cited

Adami C. 2002. What is complexity? Bioessays. 24:1085–1094.

Adl SM, et al. 2005. The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. J Eukaryot Microbiol. 52:399–451.

Andrianopoulos A, Timberlake WE. 1991. ATTS, a new and conserved DNA binding domain. Plant Cell. 3:747–748.

Bell CD, Donoghue MJ. 2005. Dating the dipsacales: comparing models, genes, and evolutionary implications. Am J Bot. 92:284–296.

Bell G, Mooers AO. 1997. Size and complexity among multicellular organisms. Biol J Linn Soc. 60:345–363.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate—a practical and powerful approach to multiple testing. J Roy Statist Soc Ser B-Methodological. 57:289–300.

Bowe LM, Coat G, dePamphilis CW. 2000. Phylogeny of seed plants based on all three genomic compartments: extant gymnosperms are monophyletic and Gnetales' closest relatives are conifers. Proc Natl Acad Sci U S A. 97:4092–4097.

Breuer C, et al. 2009. The trihelix transcription factor GTL1 regulates ploidy-dependent cell growth in the Arabidopsis trichome. Plant Cell. 21:2307–2322.

Brewer PB, et al. 2004. PETAL LOSS, a trihelix transcription factor gene, regulates perianth architecture in the Arabidopsis flower. Development. 131:4035–4045.

Burglin TR. 1991. The TEA domain: a novel, highly conserved DNA-binding motif. Cell. 66:11–12.

Busch A, Zachgo S. 2009. Flower symmetry evolution: towards understanding the abominable mystery of angiosperm radiation. Bioessays. 31:1181–1190.

Carroll SB. 2001. Chance and necessity: the evolution of morphological complexity and diversity. Nature. 409:1102–1109.

Carroll SB. 2005. Evolution at two levels: on genes and form. PLoS Biol. 3:e245.

Cavalier-Smith T. 1998. A revised six-kingdom system of life. Biol Rev Camb Philos Soc. 73:203–266.

Clamp M, Cuff J, Searle SM, Barton GJ. 2004. The Jalview Java alignment editor. Bioinformatics. 20:426–427.

Cock JM, et al. 2010. The Ectocarpus genome and the independent evolution of multicellularity in the brown algae. Nature. 465:617–621.

Coulson RM, Ouzounis CA. 2003. The phylogenetic diversity of eukaryotic transcription. Nucleic Acids Res. 31:653–660.

Crow KD, Wagner GP. 2006. Proceedings of the SMBE Tri-National Young Investigators' Workshop 2005. What is the role of genome duplication in the evolution of complexity and diversity? Mol Biol Evol. 23:887–892.

Da G, et al. 2006. Structure and function of the SWIRM domain, a conserved protein module found in chromatin regulatory complexes. Proc Natl Acad Sci U S A. 103:2057–2062.

Doerks T, Copley R, Bork P. 2001. DDT—a novel domain in different transcription and chromosome remodeling factors. Trends Biochem Sci. 26:145–146.

Donoghue PC, Purnell MA. 2005. Genome duplication, extinction and vertebrate evolution. Trends Ecol Evol. 20:312–319.

Duncan L, et al. 2007. The VARL gene family and the evolutionary origins of the master cell-type regulatory gene, regA, in Volvox carteri. J Mol Evol. 65:1–11.

Edger PP, Pires JC. 2009. Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes. Chromosome Res. 17:699–717.

Fawcett JA, Maere S, Van de Peer Y. 2009. Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event. Proc Natl Acad Sci U S A. 106:5737–5742.

Felsenstein J. 1985. Phylogenies and the comparative method. Am Nat. 125:1–15.

Fernandez-Silva P, Martinez-Azorin F, Micol V, Attardi G. 1997. The human mitochondrial transcription termination factor (mTERF) is a multizipper protein but binds to DNA as a monomer, with

Downloaded from http://gbe.oxfordjournals.org/ at MPI Molec Plant Physiology on March 28, 2012

Genome Biology and Evolution
SMBE

evidence pointing to intramolecular leucine zipper interactions. Embo J. 16:1066–1079.

Finn RD, et al. 2008. The Pfam protein families database. Nucleic Acids Res. 36:D281–D288.

Freeling M. 2008. The evolutionary position of subfunctionalization, downgraded. Genome Dyn. 4:25–40.

Freeling M, Thomas BC. 2006. Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. Genome Res. 16:805–814.

Frickenhaus S, Beszteri B. 2008. Quicktree-SD, Software developed by AWI-Bioinformatics. Available from: http://epic.awi.de/Publications/Fri2008j.pdf.

Gao G, et al. 2006. DRTF: a database of rice transcription factors. Bioinformatics. 22:1286–1287.

Gao MJ, et al. 2009. Repression of seed maturation genes by a trihelix transcriptional repressor in Arabidopsis seedlings. Plant Cell. 21:54–71.

Garland T, Ives AR. 2000. Using the past to predict the present: confidence intervals for regression equations in phylogenetic comparative methods. Am Nat. 155:346–364.

Guo A, et al. 2005. DATF: a database of Arabidopsis transcription factors. Bioinformatics. 21:2568–2569.

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

Gutierrez RA, Green PJ, Keegstra K, Ohlrogge JB. 2004. Phylogenetic profiling of the *Arabidopsis thaliana* proteome: what proteins distinguish plants from other organisms? Genome Biol. 5:15.

Hackbusch J, Richter K, Muller J, Salamini F, Uhrig JF. 2005. A central role of Arabidopsis thaliana ovate family proteins in networking and subcellular localization of 3-aa loop extension homeodomain proteins. Proc Natl Acad Sci U S A. 102:4908–4912.

Harmon LJ, Weir JT, Brock CD, Glor RE, Challenger W. 2008. GEIGER: investigating evolutionary radiations. Bioinformatics. 24:129–131.

Hedges SB, Blair JE, Venturi ML, Shoe JL. 2004. A molecular timescale of eukaryote evolution and the rise of complex multicellular life. BMC Evol Biol. 4:2.

Heimberg AM, Sempere LF, Moy VN, Donoghue PC, Peterson KJ. 2008. MicroRNAs and the advent of vertebrate morphological complexity. Proc Natl Acad Sci U S A. 105:2946–2950.

Howe K, Bateman A, Durbin R. 2002. QuickTree: building huge Neighbour-Joining trees of protein sequences. Bioinformatics. 18:1546–1547.

Hsia CC, McGinnis W. 2003. Evolution of transcription factor function. Curr Opin Genet Dev. 13:199–206.

Hug LA, Roger AJ. 2007. The impact of fossils and taxon sampling on ancient molecular dating analyses. Mol Biol Evol. 24:1889–1897.

Iida K, et al. 2005. RARTF: database and tools for complete sets of Arabidopsis transcription factors. DNA Res. 12:247–256.

Kagoshima H, et al. 1993. The Runt domain identifies a new family of heteromeric transcriptional regulators. Trends Genet. 9: 338–341.

Katoh K, Asimenos G, Toh H. 2009. Multiple alignment of DNA sequences with MAFFT. Methods Mol Biol. 537:39–64.

Katoh K, Kuma K, Toh H, Miyata T. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. 33:511–518.

Kenrick P, Crane PR. 1997. The origin and early evolution of plants on land. Nature. 389:33–39.

Lang D, Zimmer AD, Rensing SA, Reski R. 2008. Exploring plant biodiversity: the *Physcomitrella* genome and beyond. Trends Plant Sci. 13:542–549.

Lee CT, Risom T, Strauss WM. 2007. Evolutionary conservation of microRNA regulatory circuits: an examination of microRNA gene complexity and conserved microRNA-target interactions through metazoan phylogeny. DNA Cell Biol. 26:209–218.

Lespinet O, Wolf YI, Koonin EV, Aravind L. 2002. The role of lineage-specific gene family expansion in the evolution of eukaryotes. Genome Res. 12:1048–1059.

Levine M, Tjian R. 2003. Transcription regulation and animal diversity. Nature. 424:147–151.

Li A, Mao L. 2007. Evolution of plant microRNA gene families. Cell Res. 17:212–218.

Magallon S, Castillo A. 2009. Angiosperm diversification through time. Am J Bot. 96:349–365.

Martins EP. 2000. Adaptation and the comparative method. Trends Ecol Evol. 15:296–299.

McCarthy MC, Enquist BJ. 2005. Organismal size, metabolism and the evolution of complexity in metazoans. Evol Ecol Res. 7:681–696.

Muller CW, Rey FA, Sodeoka M, Verdine GL, Harrison SC. 1995. Structure of the NF-kappa B p50 homodimer bound to DNA. Nature. 373:311–317.

Pagel M. 1994. Detecting correlated evolution on phylogenies—a general-method for the comparative analysis of discrete characters. Proc R Soc Lond Ser B Biol Sci. 255:37–45.

Paponov IA, et al. 2009. The evolution of nuclear auxin signalling. BMC Evol Biol. 9:126.

Paradis E, Claude J, Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R language. Bioinformatics. 20:289–290.

Paterson AH, et al. 2006. Many gene and domain families have convergent fates following independent whole-genome duplication events in Arabidopsis, Oryza, Saccharomyces and Tetraodon. Trends Genet. 22:597–602.

Perez-Rodriguez P, et al. 2010. PlnTFDB: updated content and new features of the plant transcription factor database. Nucleic Acids Res. 38:D822–D827.

Prochnik SE, et al. 2010. Genomic analysis of organismal complexity in the multicellular green alga Volvox carteri. Science 329:223–226.

Pruesse E, et al. 2007. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res. 35:7188–7196.

Quader S, Isvaran K, Hale RE, Miner BG, Seavy NE. 2004. Nonlinear relationships and phylogenetically independent contrasts. J Evol Biol. 17:709–715.

Reyes JC, Muro-Pastor MI, Florencio FJ. 2004. The GATA family of transcription factors in Arabidopsis and rice. Plant Physiol. 134:1718–1732.

Riano-Pachon DM, Correa LG, Trejos-Espinosa R, Mueller-Roeber B. 2008. Green transcription factors: a chlamydomonas overview. Genetics. 179:31–39.

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. 143:1452–1466.

Riechmann JL, et al. 2000. Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. Science. 290: 2105–2110.

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics. 19:1572–1574.

Rost B. 1999. Twilight zone of protein sequence alignments. Protein Eng. 12:85–94.

Downloaded from http://gbe.oxfordjournals.org/ at MPI Molec Plant Physiology on March 28, 2012

Genome Biology and Evolution SMBE

Rushton PJ, et al. 2008. TOBFAC: the database of tobacco transcription factors. BMC Bioinformatics. 9:53.

Sanderson MJ. 2003. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. Bioinformatics. 19:301–302.

Sanderson MJ, Thorne JL, Wikstrom N, Bremer K. 2004. Molecular evidence on plant divergence times. Am J Bot. 91:1656–1665.

Shiu SH, Shih MC, Li WH. 2005. Transcription factor families have much higher expansion rates in plants than in animals. Plant Physiol. 139:18–26.

Shuai B, Reynaga-Pena CG, Springer PS. 2002. The lateral organ boundaries gene defines a novel, plant-specific gene family. Plant Physiol. 129:747–761.

Soltis DE, Burleigh JG. 2009. Surviving the K-T mass extinction: new perspectives of polyploidization in angiosperms. Proc Natl Acad Sci U S A. 106:5455–5456.

Soltis PS, Soltis DE. 2009. The role of hybridization in plant speciation. Annu Rev Plant Biol. 60:561–588.

Sonnhammer EL, Hollich V. 2005. Scoredist: a simple and robust protein sequence distance estimator. BMC Bioinformatics. 6:108.

Van de Peer Y, Maere S, Meyer A. 2009. The evolutionary significance of ancient genome duplications. Nat Rev Genet. 10:725–732.

Van de Peer Y, Rensing SA, Maier UG, De Wachter R. 1996. Substitution rate calibration of small subunit ribosomal RNA identifies chlorarachniophyte endosymbionts as remnants of green algae. Proc Natl Acad Sci U S A. 93:7732–7736.

Vogel C, Chothia C. 2006. Protein family expansions and biological complexity. PLoS Comput Biol. 2:e48.

Wang H, et al. 2009. Rosid radiation and the rapid rise of angiosperm-dominated forests. Proc Natl Acad Sci U S A. 106:3853–3858.

Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ. 2009. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. Bioinformatics. 25:1189–1191.

Whitcomb SJ, Basu A, Allis CD, Bernstein E. 2007. Polycomb group proteins: an evolutionary perspective. Trends Genet. 23:494–502.

Worden AZ, et al. 2009. Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes Micromonas. Science. 324:268–272.

Xia K, Fu Z, Hou L, Han JD. 2008. Impacts of protein-protein interaction domains on organism and network complexity. Genome Res. 18:1500–1508.

Xie ZM, et al. 2009. Soybean Trihelix transcription factors GmGT-2A and GmGT-2B improve plant tolerance to abiotic stresses in transgenic Arabidopsis. PLoS One. 4:e6898.

Yamada Y, Wang HY, Fukuzawa M, Barton GJ, Williams JG. 2008. A new family of transcription factors. Development. 135:3093–3101.

Yoon HS, Hackett JD, Ciniglia C, Pinto G, Bhattacharya D. 2004. A molecular timeline for the origin of photosynthetic eukaryotes. Mol Biol Evol. 21:809–818.

Zhu QH, et al. 2007. DPTF: a database of poplar transcription factors. Bioinformatics. 23:1307–1308.

Zimmer A, et al. 2007. Dating the early evolution of plants: detection and molecular clock analyses of orthologs. Mol Genet Genomics. 278:393–402.

Associate editor: Bill Martin

Downloaded from http://gbe.oxfordjournals.org/ at MPI Molec Plant Physiology on March 28, 2012