Institut für Biochemie und Biologie
Arbeitsgruppe Prof. Dr. Bernd Müller-Röber

Identification of transcription factor genes in plants

Dissertation
zur Erlangung des akademischen Grades
"doctor rerum naturalium"
(Dr. rer. nat.)
in der Wissenschaftdisziplin "Molekularbiologie"

eingereicht an der
Mathematisch-Naturwissenschaftlichen Falkultät
der Universität Potsdam

von
Diego Mauricio Riaño-Pachón
aus Bogotá, Kolumbien

Potsdam-Golm Summer 2008

Gutachter: Prof. Dr. Bernd Müller-Röber

Gutachter: Prof. Dr. Erich Grotewold

Gutachter: P.D. Dr. Stefan Rensing

Tag der mündlichen Prüfung: November 28^{th}, 2008

I love fools' experiments. I am always making them.
— Charles Darwin

Erklärung

I hereby declare that this Ph.D. thesis is the result of my own work carried out between the winter semester of 2005 and May 2008 in the group of Prof. Dr. Bernd Mueller-Roeber at the University of Potsdam in Golm, Germany. It has not been submitted for any degree or Ph.D. at any other university.

Potsdam, 14.08.2008

Diego Mauricio Riaño Pachón

Acknowledgements

I have managed to spend a bit more than four years working towards a PhD. In that time I have had the opportunity to interact with many people at the University of Potsdam and the Max Planck Institute of Molecular Plant Physiology. One of the rewards of finally finishing it is to take the opportunity to thank them.

First and foremost, to my thesis advisor Prof. Bernd Mueller-Roeber, for giving me the opportunity to develop my ideas in his group, for his continuous guidance, support and interest in my varied endeavours. To Judith Lucia Gomez Porras, for letting me know about the opportunities in Golm and offering her hospitality at my arrival and always ever since. To Ingo Dreyer, for offering his help and interest in my research, and all the small annoying 'favours' dealing with living in Deutschland. To Slobodan Ruzicic for fruitful discussions regarding the classification of transcription factors and the aesthetic appearance of our TF web sites. To Luiz Gustavo Guedes Correa for all the discussions we had about almost everything, his unconditional help and his friendship. To Marco Ende and Aixa Baumgärtel for their help regarding computer matters. To Babette Regierer and all MÜRÖS, for fruitful discussions and interesting joint projects.

The analysis that I present here would have been impossible without public access to data from different genome sequencing projects, and the effort of the several annotation communities. I am deeply grateful.

I want to acknowledge the funding agencies and projects through which I was funded over this years. The Center for Advanced Protein Technologies and the International PhD programme "Integrative Plant Science" at the University of Potsdam. The European Union (NICIP; EU CT-2002-00245) for supporting my participation in ISMB 2006 in Fortaleza, Brazil.

To the Latinamerican connection in Germany: to Flavia for being so loving; to Fernando Arana, María Inés, los Ticos: Rafa y Raúl. To CALEIDOSCOPIO LATINO 'en pleno', for great frienships and good moments.

To my parents Jorge and María, my brother David, my sister Adriana, my niece Daniela and my nephew Camilo, and to Catalina, for all their love and giving me the strength to achieve my goals.

Contents

List of Figures

List of Tables

List of Abbreviations

λ	Rate of gene gain and loss per million years
ω	Ratio of non-synonymous mutations to synonymous mutations
CRE	*cis*-regulatory element
DBD	DNA binding domain
MRCA	Most recent common ancestor
mya	Million years ago
qRT-PCR	Quantitative reverse transcription-polymerase chain reaction
TF	Transcription factor
TR	Transcription regulator
TSS	Transcription start site

Summary

In order to function properly, organisms have a complex control mechanism, in which a given gene is expressed at a particular time and place. One way to achieve this control is to regulate the initiation of transcription. This step requires the assembly of several components, i.e., a basal/general machinery common to all expressed genes, and a specific/regulatory machinery, which differs among genes and is the responsible for proper gene expression in response to environmental or developmental signals. This specific machinery is composed of transcription factors (TFs), which can be grouped into evolutionarily related gene families that possess characteristic protein domains.

In this work we have exploited the presence of protein domains to create rules that serve for the identification and classification of TFs. We have modelled such rules as a bipartite graph, where families and protein domains are represented as nodes. Connections between nodes represent that a protein domain should (required rule) or should not (forbidden rule) be present in a protein to be assigned into a TF family. Following this approach we have identified putative complete sets of TFs in plant species, whose genome is completely sequenced: *Cyanidioschyzon merolae* (red algae), *Chlamydomonas reinhardtii* (green alga), *Ostreococcus tauri* (green alga), *Physcomitrella patens* (moss), *Arabidopsis thaliana* (thale cress), *Populus trichocarpa* (black cottonwood) and *Oryza sativa* (rice). The identification of the complete sets of TFs in the above-mentioned species, as well as additional information and reference literature are available at `http://plntfdb.bio.uni-potsdam.de/`. The availability of such sets allowed us performing detailed evolutionary studies at different levels, from a single family to all TF families in different organisms in a comparative genomics context. Notably, we uncovered preferential expansions in different lineages, paving the way to discover the specific biological roles of these proteins under different conditions.

For the basic leucine zipper (bZIP) family of TFs we were able to infer that in the most recent common ancestor (MRCA) of all green plants there were at least four bZIP genes functionally involved in oxidative stress and unfolded protein responses that are

bZIP-mediated processes in all eukaryotes, but also in light-dependent regulations. The four founder genes amplified and diverged significantly, generating traits that benefited the colonization of new environments.

Currently, following the approach described above, up to 57 TF and 11 TR families can be identified, which are among the most numerous transcription regulatory families in plants. Three families of putative TFs predate the split between rhodophyta (red algae) and chlorophyta (green algae), i.e., G2-like, PLATZ, and RWPRK, and may have been of particular importance for the evolution of eukaryotic photosynthetic organisms. Nine additional families, i.e., ABI3/VP1, AP2-EREBP, ARR-B, C2C2-CO-like, C2C2-Dof, PBF-2-like/Whirly, Pseudo ARR-B, SBP, and WRKY, predate the split between green algae and streptophytes. The identification of putative complete sets of TFs has also allowed the delineation of lineage-specific regulatory families. The families SBP, bHLH, SNF2, MADS, WRKY, HMG, AP2-EREBP and FHA significantly differ in size between algae and land plants. The SBP family of TFs is significantly larger in *C. reinhardtii*, compared to land plants, and appears to have been lost in the prasinophyte *O. tauri*. The families bHLH, SNF2, MADS, WRKY, HMG, AP2-EREBP and FHA preferentially expanded with the colonisation of land, and might have played an important role in this great moment in evolution. Later, after the split of bryophytes and tracheophytes, the families MADS, AP2-EREBP, NAC, AUX/IAA, PHD and HRT have significantly larger numbers in the lineage leading to seed plants. We identified 23 families that are restricted to land plants and that might have played an important role in the colonization of this new habitat.

Based on the sets of TFs in different species we have started to develop high-throughput experimental platforms (in rice and *C. reinhardtii*) to monitor gene expression changes of TF genes under different genetic, developmental or environmental conditions. In this work we present the monitoring of *Arabidopsis thaliana* TFs during the onset of senescence, a process that leads to cell and tissue disintegration in order to redistribute nutrients (e.g. nitrogen) from leaves to reproductive organs. We show that the expression of 185 TF genes changes when leaves develop from half to fully expanded and finally enter partial senescence. 76% of these TFs are down-regulated during senescence, the remaining are up-regulated.

The identification of TFs in plants in a comparative genomics setup has proven fruitful for the understanding of evolutionary processes and contributes to the elucidation of complex developmental programs.

1

General introduction

1.1 Eukaryotic transcription

Transcription is the process in which the genetic information encoded by the DNA is transferred into RNA. This process is catalysed by an RNA polymerase (RNA pol), and controlled or assisted by a large number of other proteins, such as sequence-specific DNA-binding proteins and chromatin remodelling factors.

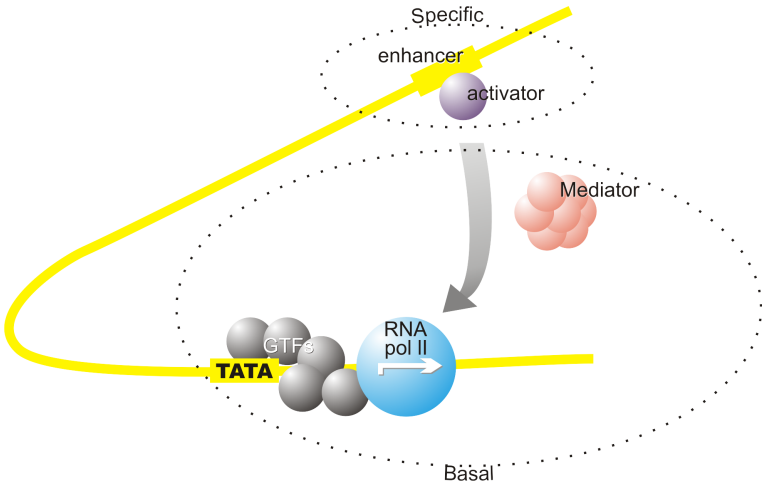


Figure 1.1: *Schematic representation of the eukaryotic transcriptional machinery for genes transcribed by RNA pol II (modified from* KORNBERG *2007).*

The transcriptional machinery can be divided in two main components, one general (or basal) and one specific (or regulatory) (Fig. 1.1). The basal apparatus is common to all genes that undergo transcription and is composed of the RNA polymerase (RNA pol) and general transcription factors (GTFs). Three types of RNA pol are present in all eukaryotes: **RNA pol I** transcribes most ribosomal RNAs, **RNA pol II** transcribes all protein coding genes, most of the small nuclear RNAs and micro RNAs, **RNA pol III** transcribes transfer RNAs, some ribosomal RNAs and small nuclear RNAs. In plants, an additional

RNA polymerase is found, **RNA pol IV**, which is required for the production of small interfering RNAs, that are involved in posttranscriptional gene silencing (ONODERA *et al.* 2005, ZHANG *et al.* 2007a). The following refers to RNA pol II alone.

The binding of the RNA pol II to the template DNA at the correct location is required for transcription initiation, however the RNA polymerase alone is not capable of recognising the DNA sequences around the transcription start site (TSS), GTFs, i.e., TFIIA, -B, -D, -E, -F and -H, accomplish this (ORPHANIDES *et al.* 1996). The Mediator protein, another important component of the transcription machinery, transduces regulatory information from distal promoter elements (e.g., enhancers) to the basal apparatus (KORNBERG 2007, LATCHMAN 2005, and references therein).

The specific apparatus consists mainly of transcription factors (*trans*-acting factors, TFs), proteins that regulate the initiation of transcription, and thus its rate, in a spatiotemporal manner (LATCHMAN 2005). TFs exert gene-specific and/or tissue-specific functions by binding to specific DNA sequences (*cis*-regulatory elements, CREs, e.g. enhancers, insulators) in the promoter of target genes, thereby enhancing or repressing their transcriptional rates. They can bind not only near or far away, but also up- or downstream, of the TSS of the gene they control. They are in charge of regulating transcriptional levels in response to different stimuli, through their interaction with the basal apparatus. In addition to TFs, other transcriptional regulators (TRs herein) are involved in transcriptional regulation, e.g., by controlling DNA packaging into chromatin.

1.2 Transcriptional regulation

The expression of a gene can be controlled at different stages: at the moment of transcription, after transcription when the mRNA is being processed, when the mature mRNA is exported from the nucleus to the cytosol, in the cytosol by means of small RNAs that might target the mRNA for degradation, at the moment of translation and post-translationally (for a detailed description see e.g., LATCHMAN 2005). Similarly, the transcription of a gene can be regulated at several distinct steps, e.g., chromatin remodelling in order to allow access to the promoter, RNA pol II recruiting to the gene promoter, transcription initiation, RNA pol II clearing of the promoter, elongation of the nascent RNA molecule and termination of transcription (ORPHANIDES and REINBERG 2002). The study of gene regulation has been focused predominantly on the initiation of transcription, however further steps in the process might be equally important, e.g., transcript elongation or promoter-proximal pausing (reviewed by CORE and LIS 2008, SIMS *et al.* 2004).

Recent genome-wide studies have challenged the widespread assumption that the pro-

moter of a gene is the immediate region upstream of the TSS. These studies have clearly shown that TFs can bind to CREs located downstream of the TSS, in introns or even exons (ENCODE PROJECT CONSORTIUM 2007, LEE *et al.* 2007, LI *et al.* 2008, ZHANG *et al.* 2007b). Furthermore, CREs can be located hundreds or even thousands of bases away (in either direction, up- or downstream) of the TSS. They can appear as single elements, or as modules, where TFs can bind cooperatively. In a similar way to TFs, CREs can be of two main types, basal (or general) and specific. Basal CREs need to be present in all genes that undergo transcription. Specific CREs are present only in the promoters of genes that should be transcribed in response to diverse stimuli. Therefore gene promoters with similar patterns of CREs will have identical or highly similar expression patterns, and will likely be regulated by common TFs.

The binding of a TF to a CRE can result in the reorganisation of histones in the neighbourhood, allowing the binding of further TFs which in turn modifies the transcriptional status. Bound TFs can interact with the basal transcriptional machinery directly or indirectly, e.g., through the Mediator protein complex. However they cannot promote transcription initiation by themselves. The bound TF can significantly increase the rate of transcription initiation. In that case the CRE is called an enhancer. If the bound TF inhibits or decreases the rate of transcription, the CRE is called a silencer. A third type of CRE, the insulator, blocks the effect of enhancers or silencers on neighboring genes when occupied by a TF, confining their effect to their intended targets (Fig. 1.2; reviewed by MASTON *et al.* 2006). Enhancers and silencers are found in plants whereas insulators appear to be absent (CHEN and ZHU 2004).

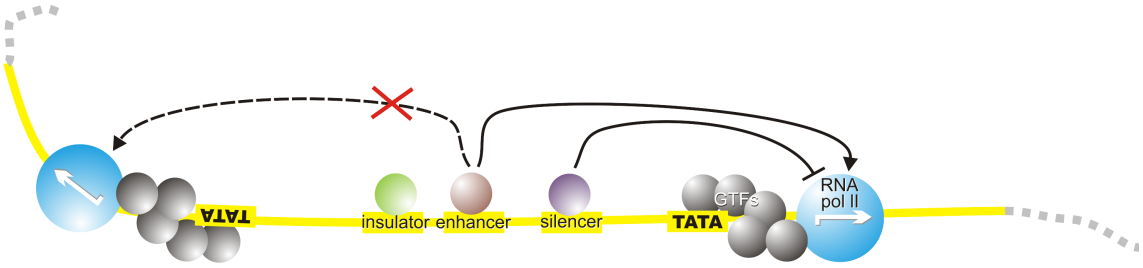


Figure 1.2: *Schematic representation of eukaryotic CREs: enhancers, silencers and insulators. Enhancers increase transcriptional rates (arrows), silencers inhibit or decrease transcriptional rates (flat arrow-heads). Insulators restrict the effect of either enhancers or silencers to their target genes.*

TFs are modular proteins; in order to interact with the DNA, they have a DNA-biding domain (DBD) that allows sequence-specific binding to CREs. An additional domain, trans-activation domain, is required for signal transduction to the basal apparatus. TFs can be grouped into classes responding to the different types of DBDs they have.

1.3 Transcription factor DNA-binding domains

DNA-binding domains (DBDs) have been classified according to their three-dimensional structural properties. Basic description of the domains can be found in LATCHMAN (2005). A more systematic and current classification of DNA-binding domains was carried out by STEGMAIER *et al.* (2004), in which DNA-binding domains were divided in superclasses and classes, families and subfamilies. According to this, five main structural superclasses can be distinguished (see Fig. 1.3 for a schematic representation and Table 1.1 for the classification of plant transcription factor families into DBD superclasses):

- Basic domain
- Helix-turn-Helix domain
- Zinc coordinating domain
- β-scaffold with minor groove contacts domain
- other domains

Basic domains are characterized by a region rich in basic amino acid residues in α-helix conformation that can interact directly with the DNA. DNA-binding specificity is determined by the sequence of the basic region. This domain is usually accompanied by an additional domain, e.g., leucine zipper, helix-loop-helix or helix-span-helix, that does not interact directly with DNA, but that is important for dimerisation and for the correct positioning of the DNA-binding regions of the dimer. Examples of this superclass are the bZIPs: 'human heterodimer c-Fos-c-Jun' (Fig. 1.3a), and Arabidopsis 'HY5', 'GBF4' and 'ABF1'; and the bHLHs: Yeast 'Pho4' (Fig. 1.3b) and Arabidopsis 'HFR1' and 'PIF3'.

The **helix-turn-helix** domain consists of two α-helical regions arranged at right angles to each other. It has been shown that one of the two helices lies partly within the major groove of DNA (recognition helix), where the sequence specific interaction takes place. The 'repressor protein of phage 434' (Fig. 1.3c) and yeast 'HSF' (Fig. 1.3d) represent this superclass.

In **zinc coordinating domains** the presence of zinc (Zn^{2+}) is required for sequence-specific DNA-binding. The zinc ion can be tetrahedrally liganded by either two cystein and two histidine residues (C2H2 zinc finger, not included in Stegmaier classification; STEGMAIER *et al.* 2004) or by multiple cysteine residues (C4 and C6 zinc fingers), allowing the formation of a structure called the zinc finger, which is responsible for sequence-specific DNA-binding. Examples of this superclass are the C2H2 zinc-fingers: mouse 'Zif268' (Fig. 1.3e) and 'GCM' (Fig. 1.3f), and the Arabidopsis WRKY TF 'ZAP1'.

β-scaffold domains with minor groove contacts is a very diverse superclass, without a structural characteristic shared by all members. Their overall mode of interaction

consists of inserting into the minor groove and causing a tight twist in the DNA. Human 'TBP' (Fig. 1.3g) and 'p53' (Fig. 1.3h) represent this superclass.

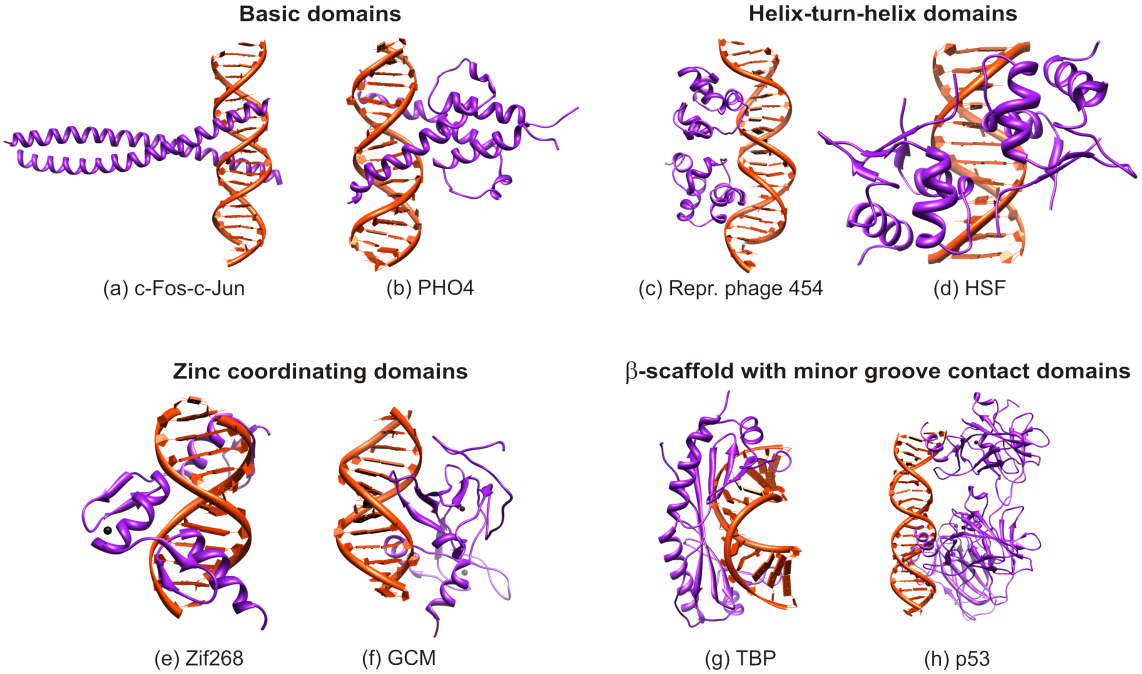


Figure 1.3: *Superclasses of DNA-binding domains. TFs are shown in purple, DNA in red and Zinc ions in black.* ***Basic domains****: (a) human c-Fos-c-Jun (PDB:1FOS) and (b) yeast PHO4 (PDB:1A0A).* ***Helix-turn-helix domains****: (c) the repressor protein of phage 434 (PDB:2OR1) and (d) yeast HSF (PDB:3HTS).* ***Zinc coordinating domains****: (e) mouse Zif268 (PDB:1ZAA) and (f) GCM (PDB:1ODH).* ***β-scaffold with minor groove contacts domains****: (g) human TBP (PDB:1TGH) and (h) p53 (PDB:1TSR).*

Domain superclass	TF families
Basic domain	BES1, bHLH, bZIP, EIL, GeBP, TCP
Helix-turn-helix domain	ARR-B, E2F-DP, FHA, G2-like, HB, HSF, MYB, MYB-related, RWP-RK, Sigma70-like, zf-HD
Zinc coordinating domain	Alfin-like, C2C2-CO-like, C2C2-Dof, C2C2-GATA, C2C2-YABBY, C2H2, C3H, CPP, GRF, HRT, LIM, PHD, PLATZ, SBP, SRS, TAZ, VOZ, WRKY, ZIM
β-scaffold with minor groove contacts domain	CCAAT, CSD, GRAS, HMG, MADS
Others	AP2-EREBP, ARF, ARID, BBR/BPC, CAMTA, DBP, DDT, Jumonji, LFY, NAC, NOZZLE, PBF-2-like, RB, S1Fa-like, Trihelix, TUB, ULT, ABI3VP1

Table 1.1: *Classification of plant transcription factor families into DBD superclasses according to their characteristic DBD.*

As described in the next section, the evolution of gene expression programs is important for generating the biodiversity in the biosphere. One crucial step towards understanding the evolution of these regulatory programs is the identification of their components,

i.e., TFs and CREs. The presence and type of a DBD can be used to identify TFs and further classify them into families, as described in Chapter 2.

1.4 Evolution of regulatory programs

Complex biological systems exhibit a large variety of lifestyles as they differ in their morphology, their behavior, and their physiology. Understanding the origins of such diversity is a quest that biologists have been after for centuries. After the decade of 1970s, the development of new technologies, such as DNA sequencing and gene expression profiling, allowed us to have a close look into the genome structure and function of a wide variety of organisms (e.g., *Methanococcus jannaschii*, *Saccharomyces cerevisiae*, *Homo sapiens*, *Arabidopsis thaliana*). It was expected that such approach would help explaining the patterns of diversity of the biological world. Soon it was realised that there is not a single mechanism to account for all the observed diversity, but instead an ensemble of molecular mechanisms that contribute to its generation. One example of such a mechanism is the evolutionary modification of gene expression programs. This concept was proposed more than 30 years ago by KING and WILSON (1975) who observed, when comparing protein sequences from chimpanzee and human, that mere sequence dissimilarity could not account for their observed differences in morphology and behavior. Several studies have provided support for this hypothesis, although with different points of view on which is the most important player (*cis*-variation vs. *trans*-variation; for reviews see CARROLL 2005, CHEN and RAJEWSKY 2007, HOEKSTRA and COYNE 2007, HSIA and MCGINNIS 2003, PRUD'HOMME *et al.* 2007, WRAY 2007, WRAY *et al.* 2003).

As mentioned before, the evolution of gene expression programs has two well known important players, TFs and short regulatory DNA sequences (i.e. CREs), to which TFs bind. CREs usually appear as modules in the promoters of genes. This *cis-trans* interaction allows fine tuning of gene expression due to the diversity of TFs and the myriad of potentially available *cis*-elements. Additionally, differential spatiotemporal control can also be achieved by TFs, in a way that TFs with similar DNA binding properties can control different biological processes (for a review see DE FOLTER and ANGENENT 2006). Beside these top players, microRNAs (miRNAs) recently received attention. miRNAs are small RNAs encoded by the genome that regulate gene expression programs post-transcriptionally (for reviews see CHEN and RAJEWSKY 2007, JONES-RHOADES *et al.* 2006). They have been just started to be catalogued (GRIFFITHS-JONES *et al.* 2008). Deciphering the relationships among these players in the control of developmental programs is one of the goals of functional genomics and of systems biology.

As mentioned in Section 1.2, TFs are modular at the sequence level, one module cor-

responds to the DBD, while another is a transactivation domain that mediates gene activation. Each module can evolve in a semi-independent manner. The concept of modularity is central in the evolution of regulatory programs. Another aspect of modularity can arise through gene duplication followed by changes in the coding sequence and/or the CREs, that can result in the origin of a new regulatory module. In Fig. 1.4, following gene duplication, one of the copies of the gene can accumulate mutations at a higher rate, which might eventually lead to the emergence of a new function, i.e., neofunctionalisation; to the split of the ancestral function among the duplicates, i.e., subfunctionalisation; or to the loss of one of the gene copies, i.e., pseudogenisation (MOORE and PURUGGANAN 2005). As a result, changes in regulatory factors, and consequently gene expression, would appear in different compartments, or tissues or at different times (for reviews see HOEKSTRA and COYNE 2007, PRUD'HOMME *et al.* 2007).

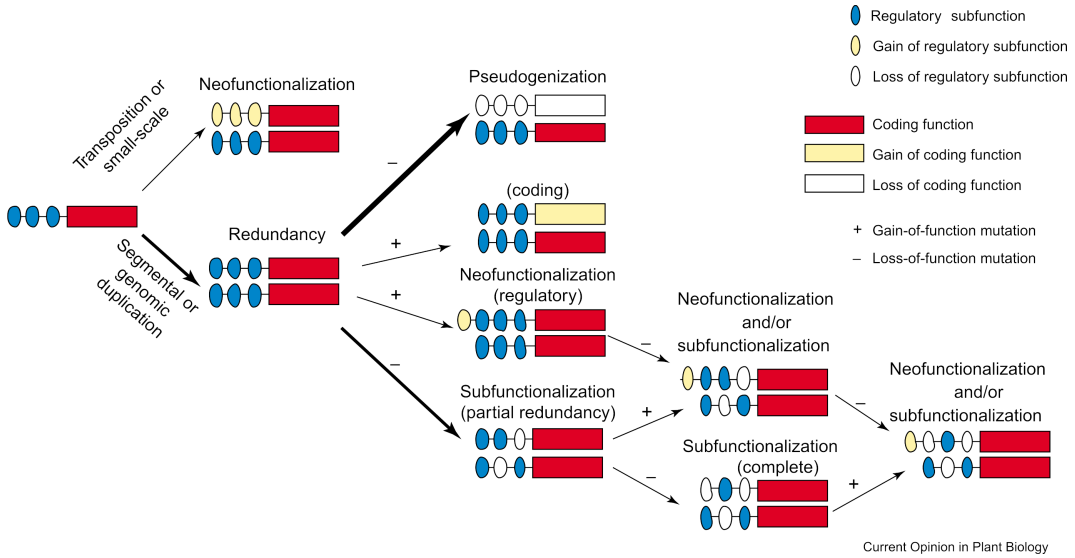


Figure 1.4: MOORE *and* PURUGGANAN *(2005) model for the evolutionary fate of duplicated genes. After a duplication event one of the gene copies can be lost by accumulating deleterious mutations, pseudogenisation; or, it can acquire a completely new function by accumulating neutral or useful mutations in, either or both, its promoter or in its protein coding region, neofunctionalisation; or the ancestral function can be split among the duplicates, subfunctionalisation.*

The evolution of regulatory programs has been widely documented (for reviews see HOEKSTRA and COYNE 2007, PURUGGANAN 2000, WRAY 2007). In flowering plants a clear example of morphological diversification due to evolutionary changes in regulatory genes is the evolution of floral development (reviewed by SOLTIS *et al.* 2007). As reviewed by BENLLOCH *et al.* (2007) the *LEAFY* (*LFY*) gene in Arabidopsis is responsible for conferring floral meristem identity, a role that is conserved in Angiosperms. The *lfy* mutant produces phenotypes where flowers are replaced by shoot-like structures. *LFY* is present in all land plants: as a single copy gene in Angiosperms, and with two copies in Bryophytes and Gymnosperms. It has been shown that the bryophyte orthologues of

LFY do not complement the Arabidopsis *lfy* mutant, while the gymnosperm orthologues complement it partially, and angiosperm homologues complement it fully. This example shows a correlation between phylogenetic relatedness and the potential for complementation, suggesting that the ancestral *LFY* gene had a different function and was recruited in flowering plants for the specification of floral meristem identity (BENLLOCH *et al.* 2007, and references therein). As seen in the previous example the identification of orthologous genes can provide insights into the ancestral functions played by those genes, and it is extremely useful to transfer knowledge about gene function between species, i.e., model plants to crop plants; however, if gene duplication precedes speciation, the function can be conserved by paralogues instead of orthologous genes (CAUSIER *et al.* 2005, VAN DE PEER 2006).

1.5 Overview of plant evolutionary relationships

One of the goals of this work is the identification of TFs in plants. I have restricted my analyses mainly to the monophyletic clade of green plants and one red alga, the genomes of which are completely sequenced and in an advanced or close-to-finish state of gene annotation. Basic information about the genomes and proteomes of the studied species can be found in Table 1.2.

Plants are essential organisms for sustaining most of life in the biosphere. Through the process of photosynthesis they get the energy required for growth directly from sunlight. By photosynthesis, which some bacteria are able to realise as well, plants convert water, CO_2 and light into organic compounds, i.e., chemical energy. This process, in eukaryotic organisms, takes place in the plastid.

The plastid is a subcellular organelle, product of an ancient endosymbiotic event (primary endosymbiosis), that might have occurred about 1.500 million years ago (mya) (YOON *et al.* 2004). It is hypothesised that an eukaryotic cell phagocyted and kept a cyanobacteria, a photosynthetically active bacteria. This event resulted in the lineage leading to the super group of Archaeplastida (*sensu* ADL *et al.* 2005). A second endosymbiotic event (secondary endosymbiosis), in which a red alga became the plastid of a non-photosynthetic protist, gave origin to the supergroup of Chromoalveolata (*sensu* ADL *et al.* 2005). A third, secondary endosymbiosis, gave rise to Rhizaria and Excavata probably in two independent events, in which a green alga turned into the plastid. Over time, the retained photosynthetic cell was reduced, becoming an organelle of the host cell. Most of the genetic machinery from the original photosynthetic cell has been transferred to the nucleus of the host cell (reviewed by REYES-PRIETO *et al.* 2007, see NOZAKI 2005 for an alternative hypothesis on plastid evolution).

Archaeplastida is a monophyletic group characterized by the presence of double membrane-bound plastids, that are free in the cytosol. It can be further divided into Glaucophyta, Rhodophyceae and Chloroplastida (ADL *et al.* 2005, RODRÍGUEZ-EZPELETA *et al.* 2005). The Glaucophyta is an early diverging small group of algae with a plastid resembling the engulfed cyanobacterium. They retained the peptidoglycan wall between their two membranes and an organelle-like body involved in CO_2 fixation, the carboxysome (BHATTACHARYA *et al.* 2004, RODRÍGUEZ-EZPELETA and PHILIPPE 2006). The red algae, Rhodophyceae, is a large group of algae characterized by the lack of flagella and the presence of phycobiliproteins within the plastid (COLE and SHEATH 1990). The Chloroplastida (green plants, *syn.* Viridiplantae *sensu* CAVALIER-SMITH 1981) consists of the Chlorophyta and the Streptophyta. Most of the green algae belong to the Chlorophyta, while Streptophyta consist of a diverse paraphyletic ensemble of freshwater algae and all land plants, the latter being the best known group of plants, including the mosses, the ferns, and the flowering plants, among others.

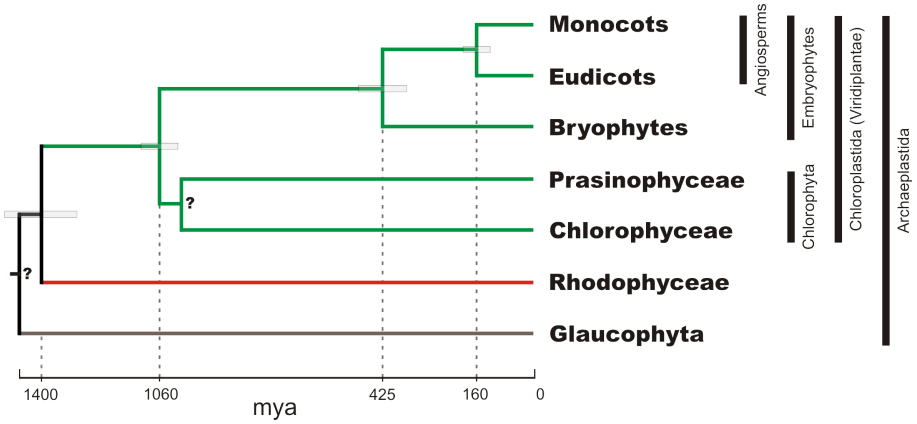


Figure 1.5: *Schematic representation of the evolutionary relationships among some of the groups of plants. Divergence times correspond to estimations and/or fossil records. The gray boxes at the nodes represent the range of possible divergence times according to literature.*

Figure 1.5 shows schematically the divergence times of the main lineages of plants. Viridiplantae and Rhodophyceae shared their most recent common ancestor (MRCA) between 1.600 and 1.474 mya (LEWIS and MCCOURT 2004, YOON *et al.* 2006, 2004, ZIMMER *et al.* 2007). The oldest known rhodophycean fossil dates from 1.200 mya (BUTTERFIELD 2000). This is therefore the youngest date for the divergence between this two groups. Viridiplantae might have split into Chlorophyta and Streptophyta around 1.111 to 1.010 mya (HECKMAN *et al.* 2001, SANDERSON *et al.* 2004, YOON *et al.* 2004). Soon after, Prasinophytes diverged from the main branch of Chlorophyta, while the streptophyte lineage split 360 to 490 mya into Tracheophyta and Bryophyta (KENRICK and CRANE 1997, NICKRENT *et al.* 2000, SANDERSON 2003, SHAW and RENZAGLIA 2004). Monocotyledoneous and dicotyledoneus plants, representatives of tracheophytes, shared their

MRCA between 200 and 120 mya (BELL *et al.* 2005, CHAW *et al.* 2004, SANDERSON and DOYLE 2001, YOON *et al.* 2004).

1.5.1 Species studied

Currently the genome sequences of several species of Archaeplastida are known and publicly available. In this thesis I intended to have a broad phylogenetic coverage. However, important groups as Monilophytes (ferns) and the Coniferophytes (e.g., pines) could not be included, since there is no annotated genome sequence available. The following species have been included: the red alga *Cyanidioschyzon merolae*, a member of the Rhodophyceae, is a small unicellular organism, found in sulfate-rich hot springs (MATSUZAKI *et al.* 2004). The remaining species are all members of the Viridiplantae. *Chlamydomonas reinhardtii* P. A. Dangeard and *Ostreococcus tauri* C. Courties & M. -J. Chrétiennot-Dinet are unicellular organisms as well, members of the Chlorophyta (green algae). *C. reinhardtii* is a member of the Chlorophyceae, soil-dwelling organism with two anterior flagella employed for motility and mating (MERCHANT *et al.* 2007). *O. tauri*, one of the smallest known free-living organisms ($\sim$1 μm in diameter), belongs to the Prasinophyceae, a group at the base of the green algal lineage and thought to be as the cell form most closely representing the first green algae, or "ancestral green flagellate" (AGF) (DERELLE *et al.* 2006, LEWIS and MCCOURT 2004). See MISUMI *et al.* (2008) for further details on this algal species.

Table 1.2: *Basic information about the species analysed in this work. G: Genome size (Mb), P_{TOTAL}: Total number of proteins encoded by the genome, C: Chromosome number.*

Species	G	P_{TOTAL}	C	Reference	Annotation
C. merolae	16.52	5014	20	MATSUZAKI *et al.* 2004 NOZAKI *et al.* 2007	Uni-Tokyo v07.2007[a]
O. tauri	12.56	7725	20	DERELLE *et al.* 2006	JGI v2.0[b]
C. reinhardtii	120	15143	17	MERCHANT *et al.* 2007	JGI v3.1[c]
P. patens	480	35938	27	RENSING *et al.* 2008	JGI v1.1[d]
A. thaliana	125	31921	5	AGI 2000 SWARBRECK *et al.* 2008	TAIR v7.0[e]
P. trichocarpa	485	45555	19	TUSKAN *et al.* 2006	JGI v1.1[f]
O. sativa	420	66710	12	GOFF *et al.* 2002 YUAN *et al.* 2005	TIGR v5.0[g]

[a] http://merolae.biol.s.u-tokyo.ac.jp/ [b] http://genome.jgi-psf.org/Ostta4/
[c] http://genome.jgi-psf.org/Chlre3/ [d] http://genome.jgi-psf.org/Phypa1_1/
[e] http://www.arabidopsis.org/ [f] http://genome.jgi-psf.org/Poptr1_1/
[g] http://www.tigr.org/tdb/e2k1/osa1/

Streptophytes are represented in this study by the bryophyte (moss) *Physcomitrella patens* ssp. *patens* (Hedw.) Bruch & Schimp. in B.S.G. , and the angiosperms *Ara-*

bidopsis thaliana (L.) Heynh. (thale cress), *Populus balsamifera* ssp. *trichocarpa* (Torr. & Gray ex Hook.) Brayshaw (synonym *Populus trichocarpa* Torr. & Gray ex Hook.) (black cottonwood) and *Oryza sativa* L. ssp. *japonica* (rice). *Arabidopsis* and *Populus* are eudicotyledons (eudicots), while *Oryza* is a monocotyledon (monocot).

1.6 Aims and structure of the thesis

The first step towards a systems-level understanding of the complex mechanisms that plants and other organisms employ to regulate their gene expression programs is to have a comprehensive list of parts, i.e., of the components of these programs. The first objective of this thesis is the identification and classification of one component of these regulatory programs, namely TFs; the questions that I wanted to tackle here were: Can the existing knowledge regarding the identification of TFs in *A. thaliana* (e.g., RIECHMANN *et al.* 2000) be applied to other plant species? Can we develop an automated or semi-automated pipeline for the identification and classification of TFs that has similar accuracy to current approaches in *A. thaliana*? The second objective is the evolutionary analysis of TFs families, which relies on the identification of complete lists of TFs in different species; the questions that I wanted to approach here were: What were the regulatory families present in the MRCA of green plants? Are there any lineage-specific family expansions? Can the evolution of individual TF families be correlated with great moments in green plant evolution? Finally, the third objective is to use the generated knowledge regarding the identification of TFs to approach the dynamics of the regulatory programs in which they play a role; the underlying question was: Can we uncover individual TF families playing preferential roles in some biological processes?

The results that I am presenting here are the fruits of collaborative work with several members of the group lead by Prof. Dr. Mueller-Roeber and are divided in the following way: Chapter 2, describes the strategy that, together with Dr. Ruzicic, P.D. Dr. Dreyer and Prof. Dr. Mueller-Roeber, we developed for the identification of TFs in plants (published in *BMC Bioinformatics*). We have identified the complement of TFs in the unicellular green alga *Chlamydomonas reinhardtii*, these data were included in the genome annotation of this organism, which was published in *Science* (MERCHANT *et al.* 2007); in Chapter 3 I present the analyses of the TFs present in this alga in a comparative genomics setup, result of a joint effort with fellow PhD students Luiz Gustavo Guedes Corrêa and Raúl Trejos-Espinosa, and Prof. Dr. Mueller-Roeber (published in *Genetics*). In Chapter 4, together with fellow PhD students Luiz Correa, Prof. Dr. Mueller-Roeber and our collaborator from the University of Campinas in Brazil Prof. Dr. Michel Vincentz, we have inferred the phylogenetic relationships among the bZIP TF family in the whole green

plant tree in a very detailed way (published in *PLoS ONE*). Chapter 5, presents an experimental approach lead by PhD student Salma Balazadeh and Prof. Dr. Mueller-Roeber, to analyse the role of TFs in plant senescence, where I have collaborated identifying gene expression clusters and evaluating the contribution of different TF families to different clusters (published in *Plant Biology*).

1.7 References

ADL, S. M., A. G. B. SIMPSON, M. A. FARMER, R. A. ANDERSEN, O. R. ANDERSON, *et al.*, 2005 The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. J Eukaryot Microbiol **52**: 399–451.

ARABIDOPSIS GENOME INITIATIVE, 2000 Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature **408**: 796–815.

BELL, C. D., D. E. SOLTIS, and P. S. SOLTIS, 2005 The age of the angiosperms: a molecular timescale without a clock. Evolution Int J Org Evolution **59**: 1245–1258.

BENLLOCH, R., A. BERBEL, A. SERRANO-MISLATA, and F. MADUEÑO, 2007 Floral initiation and inflorescence architecture: a comparative view. Ann Bot (Lond) **100**: 659–676.

BHATTACHARYA, D., H. S. YOON, and J. D. HACKETT, 2004 Photosynthetic eukaryotes unite: endosymbiosis connects the dots. Bioessays **26**: 50–60.

BUTTERFIELD, N. J., 2000 *Bangiomorpha pubescens* n. gen., n. sp.: implications for the evolution of sex, multicellularity, and the Mesoproterozoic/Neoproterozoic radiation of eukaryotes. Paleobiology **26**: 386–404.

CARROLL, S. B., 2005 Evolution at two levels: on genes and form. PLoS Biol **3**: e245.

CAUSIER, B., R. CASTILLO, J. ZHOU, R. INGRAM, Y. XUE, *et al.*, 2005 Evolution in action: following function in duplicated floral homeotic genes. Curr Biol **15**: 1508–1512.

CAVALIER-SMITH, T., 1981 Eukaryote kingdoms: seven or nine? Biosystems **14**: 461–481.

CHAW, S.-M., C.-C. CHANG, H.-L. CHEN, and W.-H. LI, 2004 Dating the monocot-dicot divergence and the origin of core eudicots using whole chloroplast genomes. J Mol Evol **58**: 424–441.

CHEN, K., and N. RAJEWSKY, 2007 The evolution of gene regulation by transcription factors and microRNAs. Nat Rev Genet **8**: 93–103.

CHEN, W. J., and T. ZHU, 2004 Networks of transcription factors with roles in environmental stress response. Trends Plant Sci **9**: 591–596.

COLE, K. M., and R. G. SHEATH, editors, 1990 *Biology of the Red Algae*. Cambridge University Press.

CORE, L. J., and J. T. LIS, 2008 Transcription regulation through promoter-proximal pausing of RNA polymerase II. Science **319**: 1791–1792.

DE FOLTER, S., and G. C. ANGENENT, 2006 *trans* meets *cis* in MADS science. Trends Plant Sci **11**: 224–231.

DERELLE, E., C. FERRAZ, S. ROMBAUTS, P. ROUZÉ, A. Z. WORDEN, *et al.*, 2006 Genome analysis of the smallest free-living eukaryote *Ostreococcus tauri* unveils many unique features. Proc Natl Acad Sci U S A **103**: 11647–11652.

ENCODE PROJECT CONSORTIUM, 2007 Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature **447**: 799–816.

GOFF, S. A., D. RICKE, T.-H. LAN, G. PRESTING, R. WANG, *et al.*, 2002 A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). Science **296**: 92–100.

GRIFFITHS-JONES, S., H. K. SAINI, S. VAN DONGEN, and A. J. ENRIGHT, 2008 miRBase: tools for microRNA genomics. Nucleic Acids Res **36**: D154–D158.

HECKMAN, D. S., D. M. GEISER, B. R. EIDELL, R. L. STAUFFER, N. L. KARDOS, *et al.*, 2001 Molecular evidence for the early colonization of land by fungi and plants. Science **293**: 1129–1133.

HOEKSTRA, H. E., and J. A. COYNE, 2007 The locus of evolution: evo devo and the genetics of adaptation. Evolution Int J Org Evolution **61**: 995–1016.

HSIA, C. C., and W. MCGINNIS, 2003 Evolution of transcription factor function. Curr Opin Genet Dev **13**: 199–206.

JONES-RHOADES, M. W., D. P. BARTEL, and B. BARTEL, 2006 MicroRNAS and their regulatory roles in plants. Annu Rev Plant Biol **57**: 19–53.

KENRICK, P., and P. R. CRANE, 1997 The origin and early evolution of plants on land. Nature **389**: 33–39.

KING, M. C., and A. C. WILSON, 1975 Evolution at two levels in humans and chimpanzees. Science **188**: 107–116.

KORNBERG, R. D., 2007 The molecular basis of eukaryotic transcription. Proc Natl Acad Sci U S A **104**: 12955–12961.

LATCHMAN, D. S., 2005 *Gene Regulation*. BIOS Advanced Text. Taylor & Francis Group, fifth edition.

LEE, J., K. HE, V. STOLC, H. LEE, P. FIGUEROA, *et al.*, 2007 Analysis of transcription factor HY5 genomic binding sites revealed its hierarchical role in light regulation of development. Plant Cell **19**: 731–749.

LEWIS, L. A., and R. M. MCCOURT, 2004 Green algae and the origin of land plants. Am J Bot **91**: 1535–1556.

LI, X., S. MACARTHUR, R. BOURGON, D. NIX, D. A. POLLARD, *et al.*, 2008 Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm. PLoS Biol **6**: e27.

MASTON, G. A., S. K. EVANS, and M. R. GREEN, 2006 Transcriptional regulatory elements in the human genome. Annu Rev Genomics Hum Genet **7**: 29–59.

MATSUZAKI, M., O. MISUMI, T. SHIN-I, S. MARUYAMA, M. TAKAHARA, *et al.*, 2004 Genome sequence of the ultrasmall unicellular red alga *Cyanidioschyzon merolae* 10D. Nature **428**: 653–657.

MERCHANT, S. S., S. E. PROCHNIK, O. VALLON, E. H. HARRIS, S. J. KARPOWICZ, *et al.*, 2007 The Chlamydomonas genome reveals the evolution of key animal and plant functions. Science **318**: 245–250.

MISUMI, O., Y. YOSHIDA, K. NISHIDA, T. FUJIWARA, T. SAKAJIRI, *et al.*, 2008 Genome analysis and its significance in four unicellular algae, *Cyanidioshyzon merolae*, *Ostreococcus tauri*, *Chlamydomonas reinhardtii*, and *Thalassiosira pseudonana*. J Plant Res **121**: 3–17.

MOORE, R. C., and M. D. PURUGGANAN, 2005 The evolutionary dynamics of plant duplicate genes. Curr Opin Plant Biol **8**: 122–128.

NICKRENT, D. L., C. L. PARKINSON, J. D. PALMER, and R. J. DUFF, 2000 Multigene phylogeny of land plants with special reference to bryophytes and the earliest land plants. Mol Biol Evol **17**: 1885–1895.

NOZAKI, H., 2005 A new scenario of plastid evolution: plastid primary endosymbiosis before the diver-

gence of the "Plantae," emended. J Plant Res **118**: 247–255.

NOZAKI, H., H. TAKANO, O. MISUMI, K. TERASAWA, M. MATSUZAKI, *et al.*, 2007 A 100%-complete sequence reveals unusually simple genomic features in the hot-spring red alga *Cyanidioschyzon merolae*. BMC Biol **5**: 28.

ONODERA, Y., J. R. HAAG, T. REAM, P. C. NUNES, O. PONTES, *et al.*, 2005 Plant nuclear RNA polymerase IV mediates siRNA and DNA methylation-dependent heterochromatin formation. Cell **120**: 613–622.

ORPHANIDES, G., T. LAGRANGE, and D. REINBERG, 1996 The general transcription factors of RNA polymerase II. Genes Dev **10**: 2657–2683.

ORPHANIDES, G., and D. REINBERG, 2002 A unified theory of gene expression. Cell **108**: 439–451.

PRUD'HOMME, B., N. GOMPEL, and S. B. CARROLL, 2007 Emerging principles of regulatory evolution. Proc Natl Acad Sci U S A **104 Suppl 1**: 8605–8612.

PURUGGANAN, M. D., 2000 The molecular population genetics of regulatory genes. Mol Ecol **9**: 1451–1461.

RENSING, S. A., D. LANG, A. D. ZIMMER, A. TERRY, A. SALAMOV, *et al.*, 2008 The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants. Science **319**: 64–69.

REYES-PRIETO, A., A. P. M. WEBER, and D. BHATTACHARYA, 2007 The origin and establishment of the plastid in algae and plants. Annu Rev Genet **41**: 147–168.

RIECHMANN, J. L., J. HEARD, G. MARTIN, L. REUBER, C. JIANG, *et al.*, 2000 Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. Science **290**: 2105–2110.

RODRÍGUEZ-EZPELETA, N., H. BRINKMANN, S. C. BUREY, B. ROURE, G. BURGER, *et al.*, 2005 Monophyly of primary photosynthetic eukaryotes: green plants, red algae, and glaucophytes. Curr Biol **15**: 1325–1330.

RODRÍGUEZ-EZPELETA, N., and H. PHILIPPE, 2006 Plastid origin: replaying the tape. Curr Biol **16**: R53–R56.

SANDERSON, M. J., 2003 Molecular data from 27 proteins do not support a Precambrian origin of land plants. Am J Bot **90**: 954–956.

SANDERSON, M. J., and J. A. DOYLE, 2001 Sources of error and confidence intervals in estimating the age of angiosperms from rbcL and 18S rDNA data. Am J Bot **88**: 1499–1516.

SANDERSON, M. J., J. THORNE, N. WIKSTROM, and K. BREMER, 2004 Molecular Evidence of Plant Divergence Times. American Journal of Botany **91**: 1656–1665.

SHAW, J., and K. RENZAGLIA, 2004 Phylogeny and diversification of bryophytes. Am J Bot **91**: 1557–1581.

SIMS, R. J., R. BELOTSERKOVSKAYA, and D. REINBERG, 2004 Elongation by RNA polymerase II: the short and long of it. Genes Dev **18**: 2437–2468.

SOLTIS, D. E., H. MA, M. W. FROHLICH, P. S. SOLTIS, V. A. ALBERT, *et al.*, 2007 The floral genome: an evolutionary history of gene duplication and shifting patterns of gene expression. Trends Plant Sci **12**: 358–367.

STEGMAIER, P., A. E. KEL, and E. WINGENDER, 2004 Systematic DNA-binding domain classification of transcription factors. Genome Inform **15**: 276–286.

SWARBRECK, D., C. WILKS, P. LAMESCH, T. Z. BERARDINI, M. GARCIA-HERNANDEZ, *et al.*, 2008

The Arabidopsis Information Resource (TAIR): gene structure and function annotation. Nucleic Acids Res **36**: D1009–D1014.

TUSKAN, G. A., S. DIFAZIO, S. JANSSON, J. BOHLMANN, I. GRIGORIEV, *et al.*, 2006 The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science **313**: 1596–1604.

VAN DE PEER, Y., 2006 Evolutionary genetics: when duplicated genes don't stick to the rules. Heredity **96**: 204–205.

WRAY, G. A., 2007 The evolutionary significance of cis-regulatory mutations. Nat Rev Genet **8**: 206–216.

WRAY, G. A., M. W. HAHN, E. ABOUHEIF, J. P. BALHOFF, M. PIZER, *et al.*, 2003 The evolution of transcriptional regulation in eukaryotes. Mol Biol Evol **20**: 1377–1419.

YOON, H. S., J. D. HACKETT, and D. BHATTACHARYA, 2006 A genomic and phylogenetic perspective on endosymbiosis and algal origin. Journal of Applied Phycology **18**: 475–481.

YOON, H. S., J. D. HACKETT, C. CINIGLIA, G. PINTO, and D. BHATTACHARYA, 2004 A molecular timeline for the origin of photosynthetic eukaryotes. Mol Biol Evol **21**: 809–818.

YUAN, Q., S. OUYANG, A. WANG, W. ZHU, R. MAITI, *et al.*, 2005 The institute for genomic research Osa1 rice genome annotation database. Plant Physiol **138**: 18–26.

ZHANG, X., I. R. HENDERSON, C. LU, P. J. GREEN, and S. E. JACOBSEN, 2007a Role of RNA polymerase IV in plant small RNA metabolism. Proc Natl Acad Sci U S A **104**: 4536–4541.

ZHANG, Z. D., A. PACCANARO, Y. FU, S. WEISSMAN, Z. WENG, *et al.*, 2007b Statistical analysis of the genomic distribution and correlation of regulatory elements in the ENCODE regions. Genome Res **17**: 787–797.

ZIMMER, A., D. LANG, S. RICHARDT, W. FRANK, R. RESKI, *et al.*, 2007 Dating the early evolution of plants: detection and molecular clock analyses of orthologs. Mol Genet Genomics **278**: 393–402.

2

Identification and classification of transcription factors

PlnTFDB: an integrative plant transcription factor database

Diego Mauricio Riaño-Pachón[1,2], Slobodan Ruzicic[1,2], Ingo Dreyer[1,2] and Bernd Mueller-Roeber[1,2]

[1]Department of Molecular Biology, Institute for Biochemistry and Biology, University of Potsdam, Golm, Germany and [2]Cooperative Research Group, Max Planck Institute for Molecular Plant Physiology, Golm, Germany

Published in *BMC Bioinformatics* (2007) **8**:42. doi:10.1186/1471-2105-8-42

Highly accessed paper according to its age and number of views.

Author contributions

BMR, SR and ID participated in the design and coordination of the project. SR and DMRP participated in the definition of the rules for the classification of TFs, and in the design of the web interface. DMRP made all the computational analyses and implemented the web databases.

BMC Bioinformatics

BioMed Central

Database

Open Access

PlnTFDB: an integrative plant transcription factor database

Diego Mauricio Riaño-Pachón[1,2], Slobodan Ruzicic[1,2], Ingo Dreyer[1,2] and Bernd Mueller-Roeber*[1,2]

Address: [1]Department of Molecular Biology, Institute for Biochemistry and Biology, University of Potsdam, Karl-Liebknecht-Str. 25 Haus 20, D-14476, Golm, Germany and [2]Cooperative Research Group, Max Planck Institute for Molecular Plant Physiology, Am Muehlenberg 1, D-14476, Golm, Germany

Email: Diego Mauricio Riaño-Pachón - diriano@uni-potsdam.de; Slobodan Ruzicic - ruzicic@mpimp-golm.mpg.de; Ingo Dreyer - dreyer@uni-potsdam.de; Bernd Mueller-Roeber* - bmr@uni-potsdam.de

* Corresponding author

Published: 7 February 2007

BMC Bioinformatics 2007, **8**:42 doi:10.1186/1471-2105-8-42

Received: 22 December 2006

Accepted: 7 February 2007

This article is available from: http://www.biomedcentral.com/1471-2105/8/42

Abstract

Background: Transcription factors (TFs) are key regulatory proteins that enhance or repress the transcriptional rate of their target genes by binding to specific promoter regions (i.e. *cis*-acting elements) upon activation or de-activation of upstream signaling cascades. TFs thus constitute master control elements of dynamic transcriptional networks. TFs have fundamental roles in almost all biological processes (development, growth and response to environmental factors) and it is assumed that they play immensely important functions in the evolution of species. In plants, TFs have been employed to manipulate various types of metabolic, developmental and stress response pathways. Cross-species comparison and identification of regulatory modules and hence TFs is thought to become increasingly important for the rational design of new plant biomass. Up to now, however, no computational repository is available that provides access to the largely complete sets of transcription factors of sequenced plant genomes.

Description: PlnTFDB is an integrative plant transcription factor database that provides a web interface to access large (close to complete) sets of transcription factors of several plant species, currently encompassing *Arabidopsis thaliana* (thale cress), *Populus trichocarpa* (poplar), *Oryza sativa* (rice), *Chlamydomonas reinhardtii* and *Ostreococcus tauri*. It also provides an access point to its daughter databases of a species-centered representation of transcription factors (OstreoTFDB, ChlamyTFDB, ArabTFDB, PoplarTFDB and RiceTFDB). Information including protein sequences, coding regions, genomic sequences, expressed sequence tags (ESTs), domain architecture and scientific literature is provided for each family.

Conclusion: We have created lists of putatively complete sets of transcription factors and other transcriptional regulators for five plant genomes. They are publicly available through http://plntfdb.bio.uni-potsdam.de. Further data will be included in the future when the sequences of other plant genomes become available.

Background

Transcription factors (TFs) are proteins (*trans*-acting factors) that regulate gene expression levels by binding to specific DNA sequences (*cis*-acting elements) in the promoters of target genes, thereby enhancing or repressing their transcriptional rates. The identification and functional characterization of TFs is essential for the reconstruction of transcriptional regulatory networks, which govern major cellular pathways in the response to biotic (e.g. response against pathogens or symbiotic relationships) and abiotic (e.g. light, cold, salt content) stimuli, and intrinsic developmental processes (e.g. growth of organs). Two global types of TFs can be distinguished: basal or general, and regulatory or specific TFs. Basal TFs belong to the minimal set of proteins required for the initiation of transcription (e.g. TATA-box binding protein). Together with RNA polymerase they form the basal transcription apparatus, representing the core of each transcriptional process. In contrast, regulatory TFs bind proximal or distal (up or downstream) of the basal transcription apparatus and act either as constitutive or inducible factors. These proteins influence the initiation of transcription by contacting members of the basal apparatus. Regulatory TFs exert gene-specific and/or tissue-specific functions and influence the transcriptional levels of their target genes in response to different stimuli. In the following when using the term TF, we refer to regulatory TFs.

The large diversity of TFs and *cis*- acting elements they bind to are the source for an enormous combinatorial complexity which allows fine-tuning gene expression control, and gives rise to a huge spectrum of developmental and physiological phenotypes. Therefore, it is not surprising that the manipulation of the expression of TFs often results in drastic phenotypic changes in the organism. This makes them extremely interesting candidates for biotechnological approaches (e.g. [1]). It is widely acknowledged that the evolution of regulatory networks is an important actor in the development of evolutionary novelties, consequently in shaping biological diversity. A deep understanding of transcription factors and their regulatory networks would also improve our understanding of organism diversity [2,3].

The cataloguing of eukaryotic transcription factors started more than a decade ago and has e.g. resulted in the generation of TRANSFAC®, a database of *cis*-acting elements and *trans*-acting factors [4]. However, TRANSFAC® includes *A. thaliana* as the only plant species that is extensively represented. Other plant species are covered to a lesser extent (e. g. *Zea mays, Nicotiana tabacum, Lycopersicum esculentum*). Additionally, other TF databases focusing on single plant species are available (for *A. thaliana* [5-7], or *O. sativa* [8]). Kummerfeld and Teichmann [9], have created a server for the prediction of TFs in organisms with sequenced genomes. Up to date, however, none of the currently available databases provides a uniform platform to review plant TF families across several species, encompassing descriptions of each TF family and links to the appropriate literature, and cross-references between the databases by means of orthologous relationships.

Today, nuclear genome sequences are available for several hundreds of organisms, and the sequencing of many more is currently underway. This provides a huge opportunity for making comparisons along different evolutionary branches of the tree of life for various kinds of genes. In this study we have focused on plants and transcription factors. We have predicted the putatively complete sets of transcription factors in five plant species, i.e. the vascular plants *Arabidopsis thaliana* [10], *Populus trichocarpa* [11], *Oryza sativa* [12] and the algae *Chlamydomonas reinhardtii* [13] and *Ostreococcus tauri* [14], and made the data available through a uniform web resource. Currently, various other plant genomes are being sequenced, including genomes from crops and experimental model species (see [15]). Plant Transcription Factor Databases at Uni-Potsdam.de provides an easily usable platform for the incorporation of new TF sequences from these and additional plant species.

Construction and content

Source datasets

Sequence data for *A. thaliana* were downloaded from TAIR [16,17], annotation release version 6.0, for *P. trichocarpa* they were downloaded from JGI/DOE [18], annotation release version 1.1, for *O. sativa* from TIGR [19], annotation release version 4.0, for *C. reinhardtii* from JGI/DOE [13], annotation release version 3.1, and for *O. tauri* from the University of Ghent [20], annotation release version August 2006.

Identification and classification of transcription factors

Transcription factors can be identified and grouped into different families according to their domain architecture, mainly taking into account their DNA-binding domains, as described by Riechmann et al. [21] for *A. thaliana*. We have extended this approach by including new TF families and applied it in a systematic manner to other plant species.

Therefore, in a first step, we identified – using current literature – the list of all domains, which are known to occur in TFs and that are generally employed to classify proteins as transcriptional regulators. The list was established from available PFAM profile Hidden Markov Models (HMMs) (v20.0, [22]), additionally we generated new models for further TF families, as indicated below.

To group TF proteins into families, we identified – based on previously published data – those domains, or in some cases domain combinations, that were specific for each family ('Literature survey' in Fig. 1). Then, we established a set of rules for each TF family. The rules can be depicted as a bipartite graph with two types of nodes and two types of edges (Fig. 2).

One set of nodes (blue squares) represents protein families (i.e. transcription factors, solid color, or other transcriptional regulators, shaded) and the other set of nodes (yellow circles) represents protein domains. The edges indicate the connections between protein domains and families. A continuous edge represents a required relationship, i.e. the indicated domain must be present in a protein to be assigned to the respective TF family. A discontinuous edge represents a forbidden relationship, i.e. the definition of such a family excludes the presence of the given domain. Rules were implemented in a PERL script as "IF . . . THEN" statements ('Classifier' in Fig. 1).

The general pipeline we have developed for the identification and classification of TFs is shown in Fig. 1. Typically, the process starts with retrieving the complete set of predicted proteins for a given species, followed by a profile-HMM search with all available PFAM HMMs (v20.0, [22])

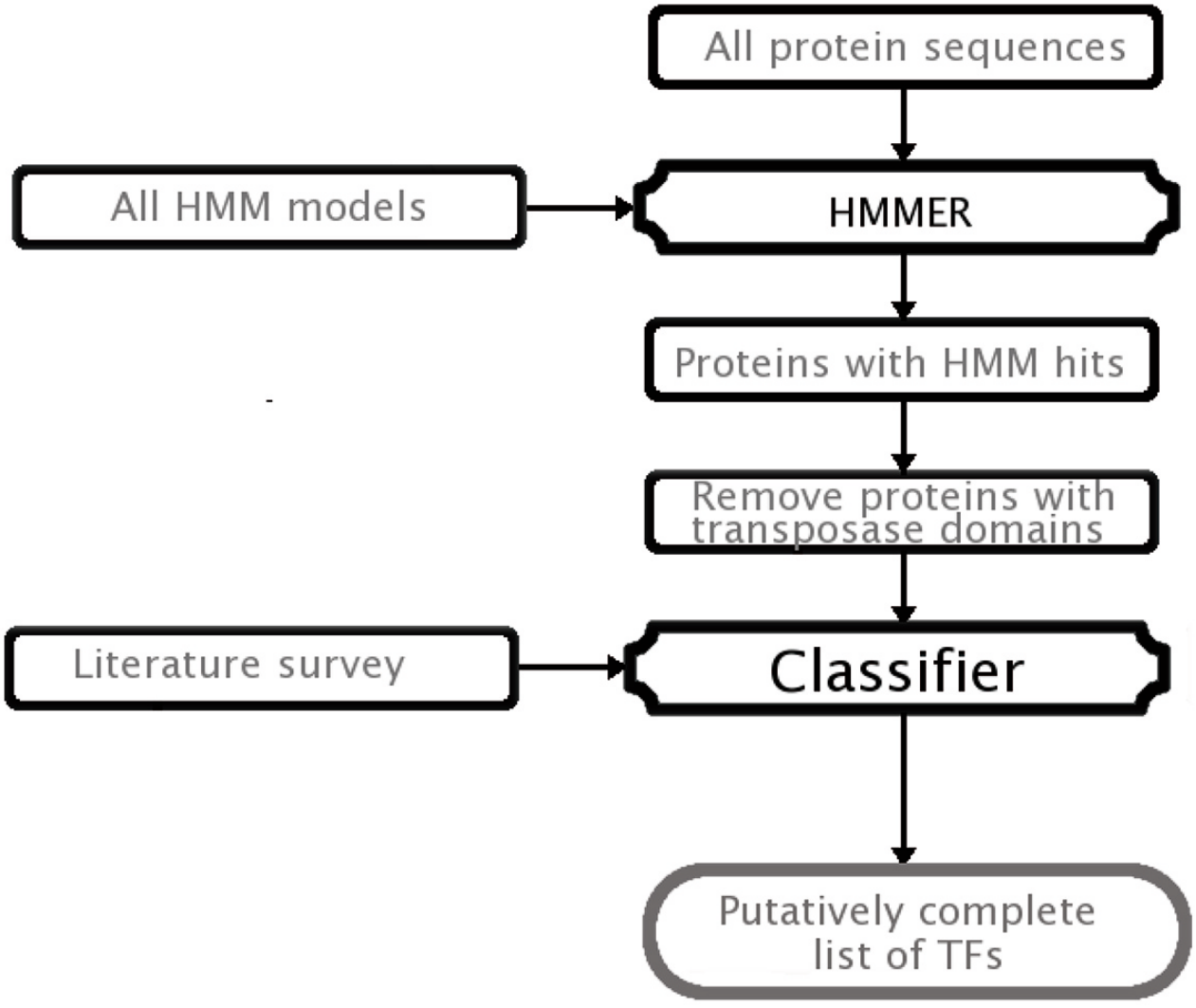


Figure 1
Pipeline for the identification and classification of TFs. The pipeline starts with the complete collection of predicted proteins for a given species. Then an HMM search is conducted over this collection keeping all significant hits and discarding all proteins containing a transposase-related domain. Finally the Classifier produces a list of putative TFs grouped into families.

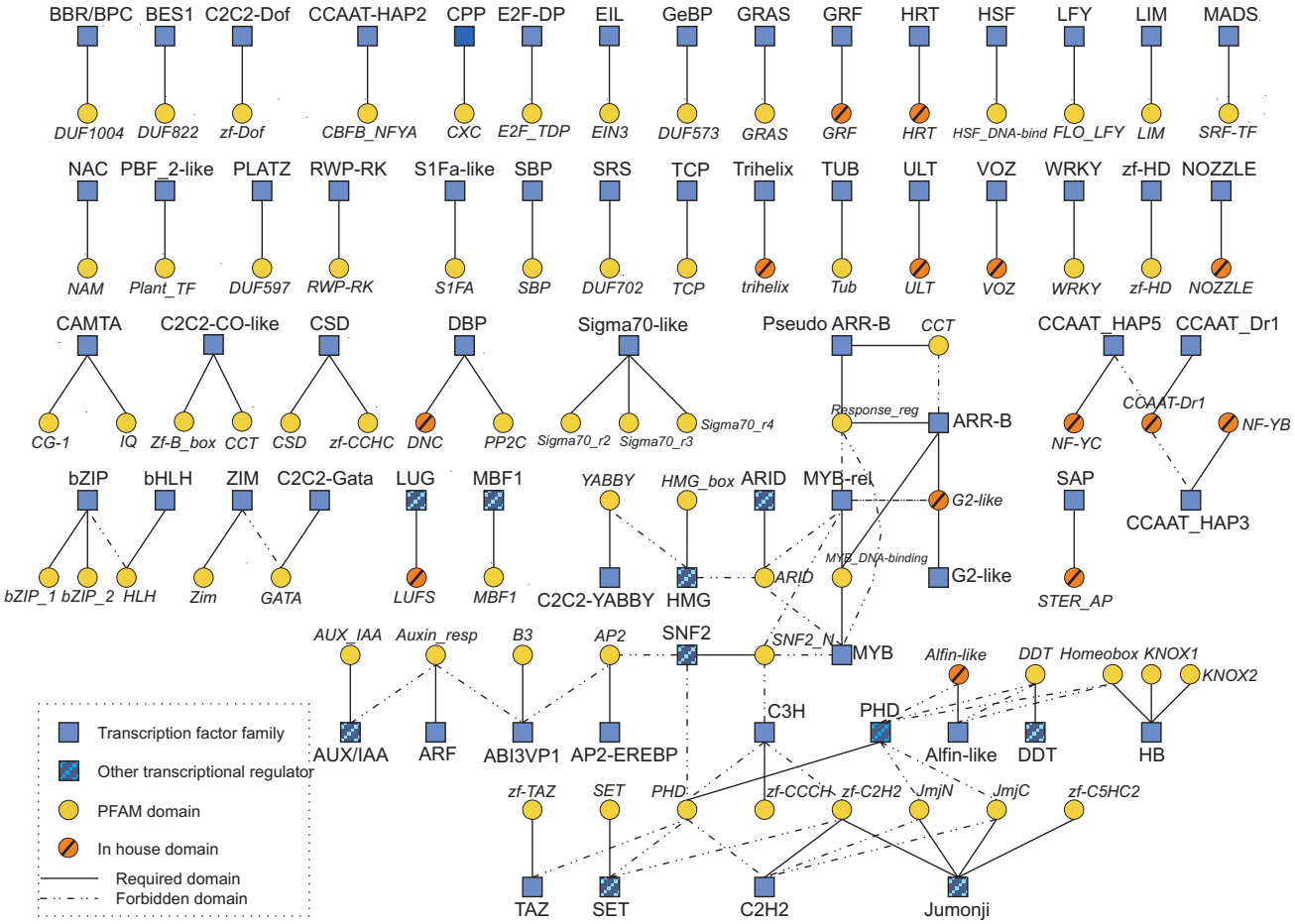


Figure 2

Rules for the classification of TF families. Rules for the classification of TFs and other transcriptional regulators depicted as a bipartite graph. Blue squares represent families, TFs are indicated in solid color, other transcription regulators are indicated by shaded squares. Yellow circles represent protein domains from the PFAM database, orange circles represent domains generated in-house. Continuous edges appear when a domain must be present in members of the family. Discontinuous edges indicate that the domain must not appear in members of the family. The profile-HMMs representing the domains Alfin-like and NOZZLE were created based on outputs derived from PSI-BLAST searches at the NCBI protein database; profile-HMMs for the domains CCAAT-Dr1, DNC, G2-like, GRF, HRT, LUFS, NF-YB, NF-YC, STER_AP, trihelix, ULT and VOZ were created from published multiple sequence alignments. All remaining domains were represented by profile-HMMs downloaded from the PFAM database. This figure is accessible via the Plant Transcription Factor Database http://plntfdb.bio.uni-potsdam.de/v1.0/rules.php, and links are provided to the respective TF families and domains.

and the models that we have generated for further TF families. The search is carried out using the software package HMMER (v2.3.2, [23]). All significant HMM hits are kept. For the PFAM models, only those hits with a bit-score larger than the gathering score reported for the HMM were considered significant. For our own HMMs, hits with an e-value smaller than 10^{-3} and a bit-score threshold that differed for each HMM were considered significant. From this set of significant HMM hits, we discarded all proteins that contained domains having DNA-related activity but not generally regarded as being parts of transcriptional regulators (such as e.g. transposase-related domains). Thereby, we eliminated potential false positives right at the beginning. Finally, we applied the PERL script implementing the set of established rules for the identification and classification of TFs on the remaining set of proteins ('Classifier' in Fig. 1). The script produces as output a list of proteins that belong to the different classes of transcriptional regulators and their classification into the identified families.

For 31 out of 68 families the presence of a single domain was sufficient to assign membership (two out of the 31 families belong to the category of other transcriptional

regulators). The remaining families were characterized by combinations of different domains. In this way we were able to classify transcription factors into 58 families plus 10 families for other types of transcriptional regulators, such as chromatin remodeling factors.

Table 1 summarizes the total number of TFs per species identified through the procedure outlined above. We detected 7597 different proteins classified as transcription factors or other transcriptional regulators in the five species analyzed. It is not surprising that the number of TFs generally increases with the number of genes in the genome (e.g. [24]). On average there are 4.2 ± 2.5 TFs per 100 genes. The INPARANOID software implements a variation of the best-reciprocal-BLAST-hits method to search for orthologs between pairs of species [25]. In finding functionally equivalent orthologous proteins INPARANOID has been shown to be the best ortholog identification method [26]. We used INPARANOID to detect orthologs between the analyzed species in a pairwise manner, starting from the complete sets of predicted proteins in each species. The predicted orthologous relationships were used to create cross-references between the species-centered databases.

New HMMs for TF families

For the families Alfin-like, CCAAT-Dr1, CCAAT-HAP3, CCAAT-HAP5, DBP, G2-like, GRF, HRT, LUG, NOZZLE, SAP, Trihelix, ULT and VOZ no appropriated models were found in the PFAM (v20.0) database. Consequently we created our own profile-HMMs based on either published multiple sequence alignments, or on alignments we created based on outputs of PSI-BLAST searches run against the NCBI protein database. The alignments used to build the HMMs are available through our web interfaces.

Database schemes

Data of the different TF families are stored in five MySQL relational databases, one for each species, and in a further, global database for PlantTFDB. To uniformly structure the databases two different schemes were implemented (Fig. 3). The first scheme (Fig. 3A) was applied for each of the five independent species-specific databases. The second scheme (Fig. 3B) was implemented for PlantTFDB, which was generated as an entry site to allow access to the species-specific databases.

The basic information in each species-specific database is structured in two sets of tables. One set (right side of the **TF** table) contains in several tables the information about the TF family: literature references, family description and domains relevant for their classification. The field relating the information in these tables is the **family_id**. The second set (left side of **TF** table) contains five tables with the information related to the TFs themselves: sequences, domains present, domain alignments, expressed sequence tags (ESTs), orthologs. The main field here is the **cds_id** that unequivocally identifies every TF. One additional table, the **TF** table relates the two sets of tables. This table has both keys, i.e., **cds_id** and **family_id**, and contains the information about the classification of the transcription factors into families. The PlantTFDB consists of a single table with the following fields: coding sequence identifier, locus identifier, transcription factor family, md5sum of the protein sequence, description of the protein sequence, species name and TF family. The field **md5sum_pep** contains the md5sum of the protein sequence, which is a sequence of 32 hexadecimal digits that identifies unequivocally each protein sequence in the database.

Web databases

A web resource with a uniform look-and-feel was developed in PHP (i) for each of the species studied, and (ii) for the PlantTFDB. We have taken care to follow W3 standards regarding HTML v4.01 and CSS v2.1 to assure browser interoperability as much as possible. Data can be downloaded from the databases as plain text files (Fig. 4).

The information provided in the species-specific web databases is linked through the gene identifiers or domain names to different external resources, when available and appropriate: TAIR [17], TIGR's rice genome annotation [19], JGI/DOE's poplar genome [18], and *C. reinhardtii* genome annotation [13], University of Ghent's *O. tauri* genome annotation [20], AthaMap [27], PlantGDB [28], Gramene [29], INPARANOID [30], SIMAP [31], and PFAM [22]. Additional external links to other databases and computational tools will continually be included.

Table 1: Number of TFs per species

Species	Total number of proteins	TFs	TF families	Percentage of TFs
Ostreococcus tauri	8236	174 (173)	33	2.1
Chlamydomonas reinhardtii	15256	229 (228)	38	1.5
Arabidopsis thaliana	30690	2304 (2147)	68	7.5
Populus trichocarpa	45555	2723 (2697)	67	6.0
Oryza sativa	62827	2516 (2352)	66	4.0

The number of TFs and other transcriptional regulators and the number of different families identified for each of the species studied. Numbers in parenthesis indicate unique protein sequences.

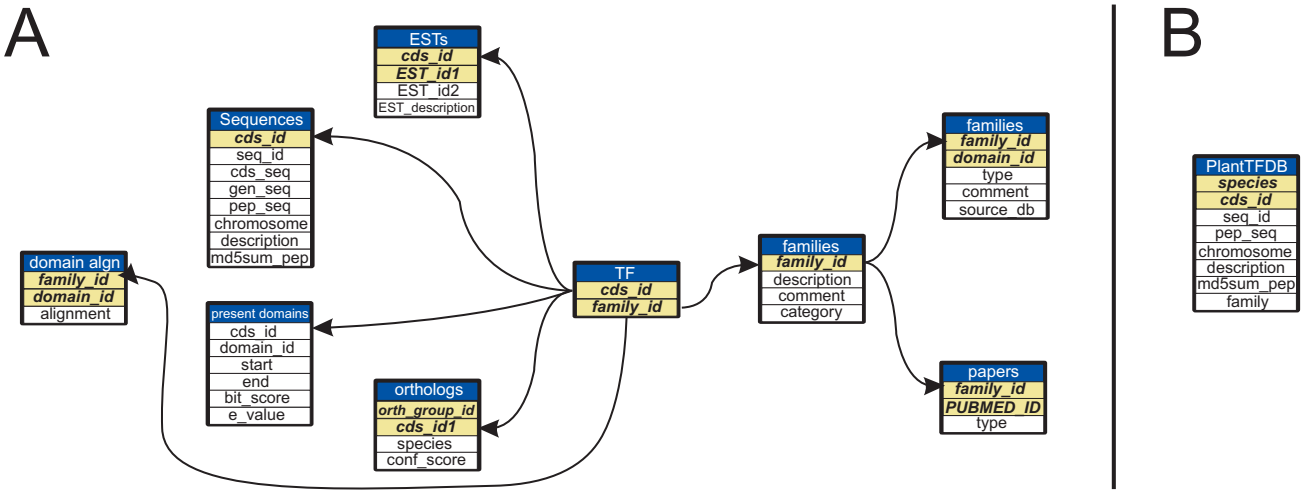


Figure 3
Database schemes. Panel A shows the scheme of the species-specific databases. Panel B shows the scheme followed by PlantTFDB. Nine tables structure the information stored in the species-centered databases. **A:** The tables **sequences, present domains, orthologs** and **ESTs** are connected to each other and to the table **TFs** by means of the **cds_id** field. The table **domain_algn** stores the alignments at the domain level for the members of a given family. All five tables contain information about the TFs. The tables **families, relevant domains** and **papers** are connected to each other and to the table **TFs** by means of the field **family_id**. They store the information concerning the TF families. **B**: A single table structures the information for Plant TFDB. Table names appear in blue background, and main keys in green background.

Quality control

To evaluate the confidence in our lists of putatively complete sets of transcription factors, we decided to compare our predictions to published data sets on detailed phylogenetic single-family analyses in *A. thaliana*. In this way the published analyses were taken as the *gold standard*. We measured the sensitivity and the positive predicive value (PPV) of our approach- in a similar fashion as done by Iida et al. [6] (The terminus 'specificity' used by Iida et al. [6] is in fact the PPV, see [32,33]).

The sensitivity is defined as:

$$Sensitivity = \frac{TP}{TP + FN},$$

where, *TP* is the number of true positives, i.e. the number of TFs listed in our database that are also found in the gold standard, and *TP* + *FN*, is the number of true positives plus the number of false negatives, i.e. *TP* + *FN* is equivalent to the total number of TFs in the gold standard.

The PPV is defined as:

$$PPV = \frac{TP}{TP + FP},$$

with the same notation as before, and *FP* being the number of false positives. Thus, *TP* + *FP* is equivalent to the total number of TFs listed in our database.

According to these definitions, the sensitivity gives an idea of the probability not to miss a true TF: a high sensitivity implies a low number of false negatives. The PPV, in contrast, gives an idea of the goodness of our method at only reporting true TFs: a high PPV implies a low number of false positives. The results of this evaluation are shown in Table 2. For 10 out of 12 tested TF families we obtained sensitivity and PPV values larger than 0.90 for both measurements (bold face in Table 2). Therefore the numbers of false negatives and false positives, respectively, are very low. Thus, the agreement with published results is still acceptable. For the remaining two families the agreement is still reasonable since both values are larger than 0.80, however at least one of them is smaller than 0.90.

The computational identification and classification of TFs is a very dynamic process that relies on the available computational models and tools, which in turn rely on the accumulated biological knowledge. This fact is reflected by the calculated Sensitivity and PPV values. As more experimental data become available over time, further improvements in HMMs are expected helping to mini-

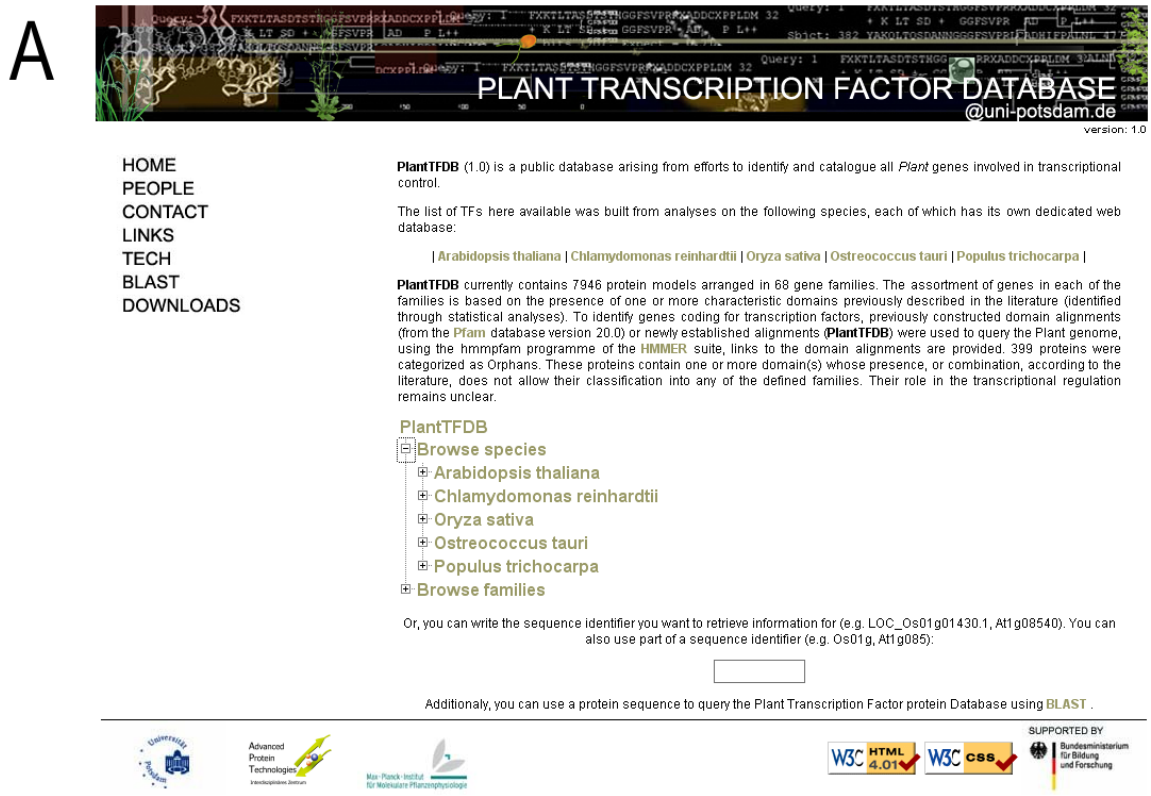


Figure 4
Web interface. Panel A shows the starting page for PlantTFDB. The tree menu in the center of the page allows browsing by species or by TF families. Panel B shows part of a typical page for a TF family; a short description and the domains that are important for the definition of the family are shown. Panel C shows part of the page for gene details, which is typical for each member of the DB. Alternative gene names are listed. Links to the genome databases and to the sister TFDBs where orthologs were found are provided.

Table 2: Quality control

Family	Reference	PPV	Sensitivity
AP2-EREBP	[39]	146/146 = 1.00	146/147 = 0.99
ARF	[40]	21/22 = 0.95	21/23 = 0.91
AUX/IAA	[40]	28/28 = 1.00	28/29 = 0.97
bHLH	[41]	122/132 = 0.92	122/154 = 0.80
bZIP	[42]	68/70 = 0.97	68/74 = 0.92
C2C2-Dof	[43]	35/36 = 0.97	35/36 = 0.97
C2C2-GATA	[44]	29/29 = 1.00	29/29 = 1.00
GRAS	[45]	32/33 = 0.97	32/33 = 0.97
MADS	[46]	99/104 = 0.95	99/108 = 0.92
MYB + MYB-related	[47]	184/209 = 0.88	184/198 = 0.93
NAC	[48]	100/101 = 0.99	100/100 = 1.00
WRKY	[49]	71/72 = 0.99	71/72 = 0.99

The Positive Predictive Value (PPV) and the Sensitivity were determined for arbitrarily selected *A. thaliana* TF families. For the PPV a deviation from 1.00 means the inclusion of false positives. For the Sensitivity deviations from 1.00 indicate exclusion of true members (false negatives). Families with both values larger than 0.90 appear in bold face.

mize further the existing gaps between the *gold standards* and the reported data in the database.

Utility and discussion

Users can start their data-mining either browsing by species, selecting one species and looking at all TF families found in that genome, or browsing by families, selecting one family and looking at the species where this TF family is present. In either case the number of proteins found is shown (see Fig. 4A). When a TF family of interest is located (e.g. Alfin-like family in rice), a click on the name of the family will lead the user to the appropriate species-centered database showing detailed information for that family (see Fig. 4B), where detailed information for each of the protein members can be accessed (e. g. LOC_Os01g66420.1; Fig. 4C). From there the user can navigate to any of the other species for which orthologs have been found. Alternatively, the user can use a preferred protein sequence to search the whole set of TFs in PlnTFDB@Uni-Potsdam, or the species-centered databases, using BLAST.

The availability of all members of a family in several species will facilitate the study of their biological functions, phylogenetic relationships, and the evolution of the DNA-binding domains. For example, Yang *et al*. [34] employed the sequences available in RiceTFDB, which is part of PlnTFDB@uni-potsdam.de, to perform an evolutionary study of DOF TFs from three different species, i.e. Arabidopsis, poplar and rice. Information extracted from our database is currently being used to establish an oligonucleotide-based microarray representing all predicted rice transcription factors (Christophe Perin, CIRAD, Montpellier, personal communication). In our own experiments we recently used the TF sequences listed in RiceTFDB to establish a large-scale quantitative real-time polymerase chain reaction (PCR) platform allowing us to test the expression of more than 2.500 rice TF genes in high throughput (manuscript in preparation). Using this platform we discovered rice TF genes responding to salt and/or drought stress, including, besides others, the genes LOC_Os04g45810 (HB TF), LOC_Os01g68370.3 (ABI3VP1 TF). Notably, the orthologous Arabidopsis genes, i.e. At2g46680.1 and At3g24650, respectively, are known to be affected by salt/drought stress [35,36].

Future plans and releases

The number of sequenced and annotated plant genomes is rapidly increasing. The computational pipeline described in this article will be applied to new plant genomes as soon as they become available and the new information will be added to future releases of PlnTFDB@uni-potsdam.de. Upcoming versions of the database will also include additional structural data about the domains employed for the identification and classification of TFs, and detailed information about the hierarchical family classification of DNA-binding domains [4,37,38].

We are currently extending the TF discovery pipeline towards large EST collections. The next release of PlnTFDB@uni-potsdam.de will include such information and will classify TFs from plant species whose genomes have not yet been sequenced but for which large EST collections are available.

Conclusion

We constructed PlnTFDB@uni-potsdam.de, the first database of its kind that provides a centralized putatively complete list of transcription factors and other transcriptional regulators from several plant species. Its daughter databases (OstreoTFDB, ChlamyTFDB, ArabTFB, PoplarTFDB,

and RiceTFDB) provide detailed information for individual members of each TF family, including orthologs present in the other species. The latest version of PlantTFDB (v1.0) contains 7597 different protein sequences, grouped into a total of 58 different TF families and 10 additional transcriptional regulator families. The web interface provides access from different starting points, from a gene ID, a protein sequence or a TF family.

Availability and requirements

All databases can be freely accessed through the WWW using any modern web browser.

PlnTFDB@uni-potsdam.dehttp://plntfdb.bio.uni-potsdam.de

RiceTFDB http://ricetfdb.bio.uni-potsdam.de

ArabTFDB http://arabtfdb.bio.uni-potsdam.de

PoplarTFDB http://poplartfdb.bio.uni-potsdam.de

OstreoTFDB http://ostreotfdb.bio.uni-potsdam.de

ChlamyTFDB http://chlamytfdb.bio.uni-potsdam.de

Abbreviations

BLAST, Basic Local Alignment Search Tool. bp, Base pair.

JGI/DOE, Joint Genome Institute/Department of Energy.

NCBI, National Center for Biotechnology Information.

TAIR, The Arabidopsis Information Resource.

TIGR, The Institute for Genomic Research.

Authors' contributions

BMR, SR and ID participated in the design and coordination of the project. SR and DMRP participated in the definition of the rules for the classification of TFs, and in the design of the web interface. DMRP made all the computational analyses and implemented the web databases. BMR supervised the group as a whole. All authors read and approved the final manuscript.

Acknowledgements

This work was financially supported by the Interdisciplinary Center 'Advanced Protein Technologies' of the University of Potsdam, coordinated by Dr. Babette Regierer, and the German Federal Ministry of Education and Research. The authors are grateful to Camila Caldana and Masood Soltaninajafabadi (Max-Planck Institute of Molecular Plant Physiology, Potsdam) for providing data about salt and drought stress regulated rice genes identified through quantitative RT-PCR, to Dr. Judith Lucia Gomez Porras and Luiz Gustavo Guedes Correa (University of Potsdam) for helpful comments on an outline version of this manuscript, to the student workers Cindy Ast and Zvonimir Marelja for their assistance during the set-up phase of this project, and to the anonymous reviewers for their valuable comments that helped to improve the article. Bernd Mueller-Roeber thanks the Fond der Chemischen Industrie for funding (No. 0164389).

References

1. Holmes-Davis R, Li G, Jamieson AC, Rebar EJ, Liu Q, Kong Y, Case CC, Gregory PD: **Gene regulation in planta by plant-derived engineered zinc finger protein transcription factors.** *Plant Mol Biol* 2005, **57(3):**411-423.
2. Tautz D: **Evolution of transcriptional regulation.** *Curr Opin Genet Dev* 2000, **10(5):**575-579.
3. Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, Rockman MV, Romano LA: **The evolution of transcriptional regulation in eukaryotes.** *Mol Biol Evol* 2003, **20(9):**1377-1419.
4. Matys V, Kel-Margoulis OV, Pricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, Voss N, Stegmaier P, Lewicki-Potapov B, Saxel H, Kel AE, Wingender E: **TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes.** *Nucleic Acids Res* 2006:D108-D110.
5. Guo A, He K, Liu D, Bai S, Gu X, Wei L, Luo J: **DATF: a database of Arabidopsis transcription factors.** *Bioinformatics* 2005, **21(10):**2568-2569.
6. Iida K, Seki M, Sakurai T, Satou M, Akiyama K, Toyoda T, Konagaya A, Shinozaki K: **RARTF: Database and Tools for Complete Sets of Arabidopsis Transcription Factors.** *DNA Res* 2005, **12(4):**247-256.
7. Davuluri RV, Sun H, Palaniswamy SK, Matthews N, Molina C, Kurtz M, Grotewold E: **AGRIS: Arabidopsis gene regulatory information server, an information resource of Arabidopsis cis-regulatory elements and transcription factors.** *BMC Bioinformatics* 2003, **4:**25.
8. Gao G, Zhong Y, Guo A, Zhu Q, Tang W, Zheng W, Gu X, Wei L, Luo J: **DRTF: a database of rice transcription factors.** *Bioinformatics* 2006, **22(10):**1286-1287.
9. Kummerfeld SK, Teichmann SA: **DBD: a transcription factor prediction database.** *Nucleic Acids Res* 2006:D74-D81.
10. Initiative AG: **Analysis of the genome sequence of the flowering plant Arabidopsis thaliana.** *Nature* 2000, **408(6814):**796-815.
11. Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, Schein J, Sterck L, Aerts A, Bhalerao RR, Bhalerao RP, Blaudez D, Boerjan W, Brun A, Brunner A, Busov V, Campbell M, Carlson J, Chalot M, Chapman J, Chen GL, Cooper D, Coutinho PM, Couturier J, Covert S, Cronk Q, Cunningham R, Davis J, Degroeve S, Déjardin A, Depamphilis C, Detter J, Dirks B, Dubchak I, Duplessis S, Ehlting J, Ellis B, Gendler K, Goodstein D, Gribskov M, Grimwood J, Groover A, Gunter L, Hamberger B, Heinze B, Helariutta Y, Henrissat B, Holligan D, Holt R, Huang W, Islam-Faridi N, Jones S, Jones-Rhoades M, Jorgensen R, Joshi C, Kangasjärvi J, Karlsson J, Kelleher C, Kirkpatrick R, Kirst M, Kohler A, Kalluri U, Larimer F, Leebens-Mack J, Leplé JC, Locascio P, Lou Y, Lucas S, Martin F, Montanini B, Napoli C, Nelson DR, Nelson C, Nieminen K, Nilsson O, Pereda V, Peter G, Philippe R, Pilate G, Poliakov A, Razumovskaya J, Richardson P, Rinaldi C, Ritland K, Rouzé P, Ryaboy D, Schmutz J, Schrader J, Segerman B, Shin H, Siddiqui A, Sterky F, Terry A, Tsai CJ, Uberbacher E, Unneberg P, Vahala J, Wall K, Wessler S, Yang G, Yin T, Douglas C, Marra M, Sandberg G, de Peer YV, Rokhsar D: **The genome of black cottonwood, Populus trichocarpa (Torr. & Gray).** *Science* 2006, **313(5793):**1596-1604.
12. Project IRGS: **The map-based sequence of the rice genome.** *Nature* 2005, **436(7052):**793-800.
13. **Chlamydomonas reinhardtii genome annotation – JGI/DOE** [http://genome.jgi-psf.org/Chlre3/Chlre3.home.html]
14. Derelle E, Ferraz C, Rombauts S, Rouzé P, Worden AZ, Robbens S, Partensky F, Degroeve S, Echeynié S, Cooke R, Saeys Y, Wuyts J, Jabbari K, Bowler C, Panaud O, Piegu B, Ball SG, Ral JP, Bouget FY, Piganeau G, Baets BD, Picard A, Delseny M, Demaille J, de Peer YV, Moreau H: **Genome analysis of the smallest free-living eukaryote Ostreococcus tauri unveils many unique features.** *Proc Natl Acad Sci USA* 2006, **103(31):**11647-11652.
15. **JGI – Sequencing Plans and Progress** [http://www.jgi.doe.gov/sequencing/seqplans.html]

16. Haas BJ, Wortman JR, Ronning CM, Hannick LI, Smith RK, Maiti R, Chan AP, Yu C, Farzad M, Wu D, White O, Town CD: **Complete reannotation of the Arabidopsis genome: methods, tools, protocols and the final release.** *BMC Biol* 2005, **3:**7.
17. Rhee SY, Beavis W, Berardini TZ, Chen G, Dixon D, Doyle A, Garcia-Hernandez M, Huala E, Lander G, Montoya M, Miller N, Mueller LA, di SM, Reiser L, Tacklind J, Weems DC, Wu Y, Xu I, Yoo D, Yoon J, Zhang P: **The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community.** *Nucleic Acids Res* 2003, **31:**224-228.
18. **Populus trichocarpa genome annotation – JGI/DOE** [http://genome.jgi-psf.org/Poptr1_1/Poptr1_1.home.html]
19. Yuan Q, Ouyang S, Wang A, Zhu W, Maiti R, Lin H, Hamilton J, Haas B, Sultana R, Cheung F, Wortman J, Buell CR: **The institute for genomic research Osal rice genome annotation database.** *Plant Physiol* 2005, **138:**18-26.
20. **Ostreococcus tauri genome annotation -Ghent University** [http://bioinformatics.psb.ugent.be/genomes/ostreococcus_tauri/]
21. Riechmann JL, Heard J, Martin G, Reuber L, Jiang C, Keddie J, Adam L, Pineda O, Ratcliffe OJ, Samaha RR, Creelman R, Pilgrim M, Broun P, Zhang JZ, Ghandehari D, Sherman BK, Yu G: **Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes.** *Science* 2000, **290(5499):**2105-2110.
22. Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, Eddy SR, Sonnhammer ELL, Bateman A: **Pfam: clans, web tools and services.** *Nucleic Acids Res* 2006:D247-D251.
23. **HMMER: profile HMMs for protein sequence analysis** [http://hmmer.janelia.org/]
24. van Nimwegen E: **Scaling laws in the functional content of genomes.** *Trends Genet* 2003, **19(9):**479-484.
25. Remm M, Storm C, Sonnhammer E: **Automatic clustering of orthologs and in-paralogs from pairwise species comparisons.** *J Mol Biol* 2001, **314(5):**1041-1052.
26. Hulsen T, Huynen MA, de Vlieg J, Groenen PMA: **Benchmarking ortholog identification methods using functional genomics data.** *Genome Biol* 2006, **7(4):**R31.
27. Steffens NO, Galuschka C, Schindler M, Bülow L, Hehl R: **AthaMap: an online resource for in silico transcription factor binding sites in the Arabidopsis thaliana genome.** *Nucleic Acids Res* 2004:D368-D372.
28. Dong Q, Lawrence CJ, Schlueter SD, Wilkerson MD, Kurtz S, Lushbough C, Brendel V: **Comparative plant genomics resources at PlantGDB.** *Plant Physiol* 2005, **139(2):**610-618.
29. Jaiswal P, Ni J, Yap I, Ware D, Spooner W, Youens-Clark K, Ren L, Liang C, Zhao W, Ratnapu K, Faga B, Canaran P, Fogleman M, Hebbard C, Avraham S, Schmidt S, Casstevens TM, Buckler ES, Stein L, McCouch S: **Gramene: a bird's eye view of cereal genomes.** *Nucleic Acids Res* 2006:D717-D723.
30. O'Brien KP, Remm M, Sonnhammer ELL: **Inparanoid: a comprehensive database of eukaryotic orthologs.** *Nucleic Acids Res* 2005:D476-D480.
31. Arnold R, Rattei T, Tischler P, Truong MD, Stümpflen V, Mewes W: **SIMAP-The similarity matrix of proteins.** *Bioinformatics* 2005, **21(Suppl 2):**ii42-ii46.
32. Altman DG, Bland JM: **Diagnostic tests 1: Sensitivity and specificity.** *BMJ* 1994, **308(6943):**1552.
33. Altman DG, Bland JM: **Diagnostic tests 2: Predictive values.** *BMJ* 1994, **309(6947):**102.
34. Yang X, Tuskan GA, Cheng MZM: **Divergence of the Dof gene families in poplar, Arabidopsis, and rice suggests multiple modes of gene evolution after duplication.** *Plant Physiol* 2006, **142(3):**820-830.
35. Söderman E, Mattsson J, Engström P: **The Arabidopsis homeobox gene ATHB-7 is induced by water deficit and by abscisic acid.** *Plant J* 1996, **10(2):**375-381.
36. Nakashima K, Fujita Y, Katsura K, Maruyama K, Narusaka Y, Seki M, Shinozaki K, Yamaguchi-Shinozaki K: **Transcriptional regulation of ABI3- and ABA-responsive genes including RD29B and RD29A in seeds, germinating embryos, and seedlings of Arabidopsis.** *Plant Mol Biol* 2006, **60:**51-68.
37. Stegmaier P, Kel AE, Wingender E: **Systematic DNA-binding domain classification of transcription factors.** *Genome Inform* 2004, **15(2):**276-286.
38. Qian Z, Cai YD, Li Y: **Automatic transcription factor classifier based on functional domain composition.** *Biochem Biophys Res Commun* 2006, **347:**141-144.
39. Feng JX, Liu D, Pan Y, Gong W, Ma LG, Luo JC, Deng XW, Zhu YX: **annotation update via cDNA sequence analysis and comprehensive profiling of developmental, hormonal or environmental responsiveness of the Arabidopsis AP2/EREBP transcription factor gene family.** *Plant Mol Biol* 2005, **59(6):**853-868.
40. Remington DL, Vision TJ, Guilfoyle TJ, Reed JW: **Contrasting modes of diversification in the Aux/IAA and ARF gene families.** *Plant Physiol* 2004, **135(3):**1738-1752.
41. Bailey PC, Martin C, Toledo-Ortiz G, Quail PH, Huq E, Heim MA, Jakoby M, Werber M, Weisshaar B: **Update on the basic helix-loop-helix transcription factor gene family in Arabidopsis thaliana.** *Plant Cell* 2003, **15(11):**2497-2502.
42. Jakoby M, Weisshaar B, Droege-Laser W, Vicente-Carbajosa J, Tiedemann J, Kroj T, Parcy F: **bZIP transcription factors in Arabidopsis.** *Trends in Plant Science* 2002, **7:**106-111.
43. Lijavetzky D, Carbonero P, Vicente-Carbajosa J: **Genome-wide comparative phylogenetic analysis of the rice and Arabidopsis Dof gene families.** *BMC Evol Biol* 2003, **3:**17.
44. Reyes JC, Muro-Pastor MI, Florencio FJ: **The GATA family of transcription factors in Arabidopsis and rice.** *Plant Physiol* 2004, **134(4):**1718-1732.
45. Bolle C: **The role of GRAS proteins in plant signal transduction and development.** *Planta* 2004, **218(5):**683-692.
46. Parenicová L, de Folter S, Kieffer M, Horner DS, Favalli C, Busscher J, Cook HE, Ingram RM, Kater MM, Davies B, Angenent GC, Colombo L: **Molecular and phylogenetic analyses of the complete MADS-box transcription factor family in Arabidopsis: new openings to the MADS world.** *Plant Cell* 2003, **15(7):**1538-1551.
47. Yanhui C, Xiaoyuan Y, Kun H, Meihua L, Jigang L, Zhaofeng G, Zhiqiang L, Yunfei Z, Xiaoxiao W, Xiaoming Q, Yunping S, Li Z, Xiaohui D, Jingchu L, Xing-Wang D, Zhangliang C, Hongya G, Li-Jia Q: **The MYB transcription factor superfamily of Arabidopsis: expression analysis and phylogenetic comparison with the rice MYB family.** *Plant Mol Biol* 2006, **60:**107-124.
48. Ooka H, Satoh K, Doi K, Nagata T, Otomo Y, Murakami K, Matsubara K, Osato N, Kawai J, Carninci P, Hayashizaki Y, Suzuki K, Kojima K, Takahara Y, Yamamoto K, Kikuchi S: **Comprehensive Analysis of NAG Family Genes in Oryza sativa and Arabidopsis thaliana.** *DNA Res* 2003, **10(6):**239-247.
49. Ulker B, Somssich IE: **WRKY transcription factors: from DNA binding towards biological function.** *Curr Opin Plant Biol* 2004, **7(5):**491-498.

3

Transcription factors in *Chlamydomonas reinhardtii*

Green transcription factors: a Chlamydomonas overview

Diego Mauricio Riaño-Pachón[1,2,†], Luiz Gustavo Guedes Corrêa[1,3,†], Raúl Trejos-Espinosa[1,3], Bernd Mueller-Roeber[1,3]

[†] These authors contributed equally to this work and should thus both be considered as first authors.

[1]Department of Molecular Biology, University of Potsdam, Potsdam-Golm, Germany, [2]Bioinformatics Research Group, GabiPD Team, Max-Planck Institute of Molecular Plant Physiology, Potsdam–Golm, Germany,[3]Cooperative Research Group, Max-Planck Institute of Molecular Plant Physiology, Potsdam–Golm, Germany

Published in *Genetics* (2008) **179**(1):31-39. doi:10.1534/genetics.107.086090

Author contributions

BMR, LGGC and DMRP conceived and designed the study. BMR coordinated the project. LGGC and RTJ identified groups of orthologues by phylogenetic analyses using NJ. DMRP identified groups of orthologues by symmetrical BLAST hits, compared the results of phylogenetics and symmetric similarity matches, identified TFs and TRs in both photosynthetic and non-photosynthetic species. All authors discussed and analysed the data.

Note: The identification of transcription factors in *C. reinhardtii* presented here was part of the genome annotation project for this organism, which was published in MERCHANT *et al*. Science **318**: 245-250.

DOI: 10.1534/genetics.107.086090

Green Transcription Factors: A Chlamydomonas Overview

Diego Mauricio Riaño-Pachón,*,†,[1] Luiz Gustavo Guedes Corrêa,*,‡,[1] Raúl Trejos-Espinosa*,‡ and Bernd Mueller-Roeber*,‡,[2]

**Department of Molecular Biology, University of Potsdam, 14476 Potsdam-Golm, Germany, †Bioinformatics Research Group, Max-Planck Institute of Molecular Plant Physiology, GabiPD Team, 14476 Potsdam-Golm, Germany and ‡Cooperative Research Group, Max-Planck Institute of Molecular Plant Physiology, 14476 Potsdam-Golm, Germany*

Manuscript received December 15, 2007
Accepted for publication January 29, 2008

ABSTRACT

Transcription factors (TFs) control gene expression by interacting with *cis*-elements in target gene promoters. Transcription regulators (TRs) assist in controlling gene expression through interaction with TFs, chromatin remodeling, or other mechanisms. Both types of proteins thus constitute master controllers of dynamic transcriptional networks. To uncover such control elements in the photosynthetic green alga *Chlamydomonas reinhardtii*, we performed a comprehensive analysis of its genome sequence. In total, we identified 234 genes encoding 147 TFs and 87 TRs of ~40 families. The set of putative TFs and TRs, including their transcript and protein sequences, domain architectures, and supporting information about putative orthologs, is available at http://plntfdb.bio.uni-potsdam.de/v2.0/. Twelve of 34 plant-specific TF families were found in at least one algal species, indicating their early evolutionary origin. Twenty-two plant-specific TF families and one plant-specific TR family were not observed in algae, suggesting their specific association with developmental or physiological processes characteristic to multicellular plants. We also analyzed the occurrence of proteins that constitute the light-regulated transcriptional network in angiosperms and found putative algal orthologs for most of them. Our analysis provides a solid ground for future experimental studies aiming at deciphering the transcriptional regulatory networks in green algae.

THE regulation of growth and development and the coordination of these processes in response to hormonal or environmental stimuli, including adverse conditions, requires a dynamic control of the expression of hundreds to thousands of genes in each organism (Lemon and Tjian 2000; Chen *et al.* 2002; Li *et al.* 2007). Transcription factors (TFs) are master control proteins that regulate gene expression levels by binding to specific DNA sequences, so-called *cis*-acting elements, in the promoters of target genes, thereby enhancing or repressing their transcriptional rates. The genomewide identification of TF genes through computational methods, and genomewide comparative studies, are important tasks that not only provide an insight into existing TF families within individual species or organism lineages but also help to understand how evolution shaped developmental and physiological diversification. TFs, as well as other transcriptional regulators (TRs) that generally do not directly bind DNA but assist in gene expression regulation through interaction with *cis*-element-binding proteins, can be grouped into different protein families according to their primary and/or three-dimensional structure similarities in the DNA-binding and multimerization domains. TF genes represent a considerable fraction of the genomes of all eukaryotic organisms, including angiosperms (Riechmann *et al.* 2000; Goff *et al.* 2002). In *Oryza sativa* (rice), for example, ~2.6% of the identified genes encode TFs (Goff *et al.* 2002). Currently, the genome sequences of four angiosperms (*Arabidopsis thaliana*, *O. sativa*, *Populus trichocarpa*, and *Vitis vinifera*) are in the public domain (Arabidopsis Genome Initiative 2000; Goff *et al.* 2002; Yu *et al.* 2002; Tuskan *et al.* 2006; Jaillon *et al.* 2007). Additionally, the genomes of various algae, including the red alga *Cyanidioschyzon merolae* (Nozaki *et al.* 2007), the green algae *Ostreococcus tauri* (Derelle *et al.* 2006), *Chlamydomonas reinhardtii* (Merchant *et al.* 2007), and the moss *Physcomitrella patens* (Rensing *et al.* 2008) have become available.

To facilitate the analysis of plant TFs and TRs, we have recently established the Plant Transcription Factor Database (PlnTFDB) (Riano-Pachon *et al.* 2007) and updated it by including additional plant species (available at http://plntfdb.bio.uni-potsdam.de/v2.0). Here we report about the occurrence of putative transcriptional regulators in Chlamydomonas. We identified 147 putative TFs that belong to 29 different protein families and 87 putative TRs that are members of 10 families. Of 34 plant-specific families, 3 (G_2-like, PLATZ, RWP-RK) predate the split between green and red algae. Nine

[1]These authors contributed equally to this work.

[2]*Corresponding author:* Department of Molecular Biology, University of Potsdam, Karl-Liebknecht-Strasse 24-25, Haus 20, 14476 Potsdam-Golm, Germany. E-mail: bmr@uni-potsdam.de

additional families, *i.e.*, ABI3/VP1, AP2-EREBP, ARR-B, C2C2-CO-like, C2C2-Dof, PBF-2-like/Whirly, Pseudo ARR-B, SBP, and WRKY, predate the split between chlorophyta (green algae) and streptophyta (land plants and charophycean algae). In total, 12 families were identified from algal groups onward. Interestingly, 22 plant-specific TF families and one TR family are not present in algae, indicating their particular importance for plant multicellularity and tissue organization.

MATERIALS AND METHODS

Identification of transcription factors: Putative complete sets of transcription factors of the following species were retrieved from the Plant Transcription Factor Database v2.0, PlnTFDB (http://plntfdb.bio.uni-potsdam.de/v2.0; Riano-Pachon *et al.* 2007): the red alga *C. merolae* (Nozaki *et al.* 2007), the green algae *O. tauri* (Derelle *et al.* 2006) and *C. reinhardtii* (Merchant *et al.* 2007), the moss *P. patens* (Rensing *et al.* 2008), and the angiosperms *A. thaliana* (Arabidopsis Genome Initiative 2000), *P. trichocarpa* (black cottonwood) (Tuskan *et al.* 2006), and *O. sativa* (rice) (Goff *et al.* 2002). PlnTFDB has two divisions: one providing information about transcription factors, defined as proteins that directly bind to DNA and affect the level of transcription (called TFs here), and the other providing information about transcriptional regulators that, for example, exert regulatory control through interaction with TFs or through chromatin remodeling (called TRs). Additionally, we identified TF and TR families common to all eukaryotes using the following model organisms: the protozoan *Giardia lamblia* (Best *et al.* 2004), the yeast *Saccharomyces cerevisiae* (Goffeau *et al.* 1996), the nematode *Caenorhabditis elegans* (*C. elegans* Sequencing Consortium 1998), the insect *Drosophila melanogaster* (Adams *et al.* 2000), and *Homo sapiens* (International Human Genome Sequencing Consortium 2004). For the identification of nonplant TFs, we used the procedure described by Riano-Pachon *et al.* (2007), using PFAM (Finn *et al.* 2006) release 20.0 for domain identification. Sequences were downloaded from Integr8 (http://www.ebi.ac.uk/integr8/; Kersey *et al.* 2005), except for *G. lamblia* sequences, which were downloaded from GiardiaDB (http://www.giardiadb.org).

Phylogenetic analysis: Protein sequences corresponding to the defining conserved domain of each TF and TR family were extracted from whole-protein sequences of the photosynthetic eukaryotes using the domain coordinates identified by the PFAM search described above. Alignment of protein sequences was performed employing ClustalX (Thompson *et al.* 1997), using default parameters. Phylogenetic analyses based on amino acid sequences were conducted using MEGA v3.1 (Kumar *et al.* 2004). Unrooted phylogenetic tree topologies were reconstructed by neighbor-joining (NJ), the distances were obtained using p-distances (Nei and Kumar 2000), and the resampling of the original protein set was a 1000-bootstrap repetition. These NJ analyses provide an overview of the general patterns of TF and TR evolution. All sequences and alignments used in this study are available upon request.

Identification of orthologs among green plants: We identified orthologs through pairwise comparisons of protein sequences in whole-protein sets of the green plants Chlamydomonas, Ostreococcus, Physcomitrella, rice, Arabidopsis, and black cottonwood, using a variation of the best BLAST bidirectional hit approach implemented in the program InParanoid (Remm *et al.* 2001). Orthologs identified in this way are presented in PlnTFDB (http://plntfdb.bio.uni-potsdam.de/v2.0). Pairs of orthologs (and direct paralogs) allowed us to identify possible clusters of orthologs (and paralogs) comprising genes from more than two species. This was achieved using a graph-theoretic approach as follows: (i) only InParanoid clusters containing at least one protein annotated as a TF in PlnTFDB were kept; (ii) all proteins identified in this way were represented as nodes in a network of orthologous relationships; edges were drawn between nodes when the InParanoid confidence score for the orthologous relationship was ≥ 0.9; (iii) connected components were extracted from the network; by definition, a connected component is a subgraph in which every node can be reached from every other node. The connected components (subgraphs) represent putative clusters of orthologs. Network visualization and analysis were carried out using the software package Pajek (De Nooy *et al.* 2005). The identification of orthologs through BLAST searches can lead to false positives; consequently, we made use of a phylogenetic approach to largely compensate for this fact. In addition to that, and affecting both approaches for ortholog detection (phylogenetics and InParanoid), false negatives can arise due to incomplete genome sequence information (gaps in the sequence) or misannotated genes.

As mentioned above, in addition to the BLAST approach, we performed phylogenetic analyses of each family, which allowed the identification of possible groups of orthologs (PoGOs). A PoGO is defined by the following criteria: (i) members of a PoGO have a monophyletic origin, indicated by a bootstrap support of $>50\%$; (ii) a PoGO conserved in all green plants possesses at least one representative gene of each of the main lineages analyzed here, including algae, bryophytes, and angiosperms, assuming that the putative complete sets of TF genes of these organisms were identified and no selective gene loss had occurred; (iii) the inferred phylogeny is consistent with the known phylogeny of plant species (Vincentz *et al.* 2003).

We evaluated the overlap between the clusters of orthologs identified by InParanoid and by phylogenetic analysis using the Adjusted Rand Index (Rand 1971; Hubert and Arabie 1985), implemented in the statistical package R (R Development Core Team 2007).

RESULTS AND DISCUSSION

Transcription factors in eukaryotes: We identified the putatively complete nonredundant sets of TFs and TRs in the algae Chlamydomonas and Ostreococcus, the moss Physcomitrella, and the angiosperms Arabidopsis, black cottonwood, and rice (Table 1). The genes were grouped into 66 gene families according to their characteristic conserved domains, as described by Riano-Pachon *et al.* (2007). We identified the putatively complete sets of genes for the same families in *G. lamblia* (protozoa), *S. cerevisiae* (yeast), *C. elegans* (nematodes), *D. melanogaster* (fruit flies), and *H. sapiens* (humans). Twenty TF and 11 TR families were also present in nonphotosynthetic eukaryotes. In contrast to the previous report by Riechmann *et al.* (2000), we observed that the Trihelix family is not restricted to the plant kingdom (Table 1). G_2-like and WRKY TFs are generally regarded as plant specific; our analysis largely confirms this view. However, we also identified genes encoding putative members of these families in the nonplant species *G. lamblia*, a

TABLE 1

Transcription factors and regulators present in different eukaryotic species

	Photosynthetic species							Nonphotosynthetic species				
Family	CME	OTA	CRE	PPA	OSAJ	ATH	PTR	HSA	DME	CEL	SCE	GLA
ABI3VP1			*1*	*30*	*59*	*59*	*81*					
Alfin-like				*7*	*12*	*8*	*9*					
AP2-EREBP		*8*	*11*	*150*	*174*	*160*	*206*					
ARF				*13*	*42*	*32*	*36*					
ARR-B		*1*	*1*	*5*	*9*	*15*	*15*					
BBR/BPC					*3*	*10*	*15*					
BES1				*6*	*6*	*10*	*12*					
bHLH	1	1	4	100	175	160	159	154	65	50	7	
bZIP	3	7	7	37	113	93	84	59	32	39	12	1
C2C2-CO-like		*3*	*1*	*11*	*21*	*19*	*14*					
C2C2-Dof		*2*	*1*	*20*	*33*	*42*	*41*					
C2C2-GATA	*6*	*4*	*6*	*12*	*37*	*30*	*36*	*15*	*7*	*14*	*10*	
C2C2-YABBY					*13*	*7*	*13*					
C2H2	7	4	5	56	103	104	113	644	312	150	39	5
CAMTA				1	7	6	7	2	1	2		
CCAAT	6	8	8	27	58	53	59	25	11	12	10	3
CPP	2	2	1	6	16	9	12	4	2	2		
CSD		4	1	3	3	4	7	16	4	5		
DBP					*7*	*5*	*10*					
E2F-DP	5	3	6	10	12	11	10	18	3	7		1
EIL				*2*	*7*	*6*	*6*					
FHA	2	7	12	15	19	17	18	44	22	12	14	2
G_2-like[a]	*1*	*2*	*4*	*41*	*52*	*48*	*66*					*1*
GeBP					*6*	*20*	*7*					
GRAS				*39*	*56*	*35*	*97*					
GRF				*2*	*17*	*9*	*9*					
HB	5	6	1	42	124	97	129	299	114	107	7	
HRT				*7*	*1*	*2*	*1*					
HSF	3	1	2	8	36	23	31	6	1	1	5	
LFY				2	1	1	1					
LIM	1	1		3	7	6	13	ND	ND	ND	ND	ND
MADS	1	1	2	22	82	122	108	9	3	2	4	
MYB	11	10	11	61	129	161	210	19	5	7	3	2
MYB-related	21	17	14	44	99	90	100	36	16	12	12	2
NAC				*32*	*140*	*115*	*163*					
NOZZLE						*1*						
PBF-2 like/Whirly		*1*	*1*		*3*	*4*	*3*					
PLATZ	*1*	*1*	*3*	*13*	*18*	*13*	*20*					
Pseudo ARR-B		*1*	*2*	*2*	*7*	*5*	*7*					
RWP-RK[b]	*1*	*4*	*14*	*8*	*13*	*14*	*18*					
S1Fa-like				*1*	*2*	*3*	*2*					
SAP						*1*	*1*					
SBP			*21*	*13*	*21*	*17*	*29*					
Sigma70-like[c]	4	1	1	5	9	6	9					
SRS				*2*	*5*	*11*	*10*					
TAZ			2	5	9	9	7	4	1	6		
TCP				6	22	26	33					
Trihelix				25	24	27	43	8	1			
TUB		1	3	6	17	12	11	6	3	2		
ULT					*2*	*2*	*2*					
VOZ				*2*	*2*	*2*	*4*					
WRKY[a]		*2*	*1*	*37*	*114*	*84*	*101*					*1*
zf-HD				*7*	*15*	*17*	*22*					
ZIM				*12*	*22*	*22*	*16*					
ARID[d]	*4*	*1*	*2*	*7*	*6*	*10*	*13*	*21*	*6*	*5*	*2*	
AUX/IAA[d]				*2*	*43*	*34*	*32*					

(*continued*)

TABLE 1

(Continued)

Family	Photosynthetic species							Nonphotosynthetic species				
	CME	OTA	CRE	PPA	OSAJ	ATH	PTR	HSA	DME	CEL	SCE	GLA
C3H[d]	7	18	15	44	97	75	96	85	33	40	7	5
DDT[d]	1		1	2	7	5	5	5	3	2	2	
HMG[d]		5	7	8	17	19	12	90	26	22	7	3
Jumonji[d]	3	5	7	10	17	19	20	38	11	15	3	
LUG[d]				1	12	3	5	6	1	1	1	
MBF1[d]			1	3	4	3	3	1	1	1	1	
PHD[d]	7	11	12	50	55	53	70	118	44	23	14	2
RB[d]	1	1	1	2	4	1	1	4	2	1		
SET[d]	6	10	22	26	32	38	44	49	17	29	7	3
SNF2[d]	13	20	19	35	44	43	48	48	22	23	17	6

Plant-specific TF and TR families are in italics; all other families are in roman. We also highlight TF families in italics that, in addition to plants, have members in early branching eukaryotes. Numbers represent distinct protein sequences. CME, *C. merolae*; OTA, *O. tauri*; CRE, *C. reinhardtii*; PPA, *P. patens*; OSAJ, *O. sativa* ssp. *japonica*; ATH, *A. thaliana*; PTR, *P. trichocarpa*; HAS, *H. sapiens*; DME, *D. melanogaster*; CEL, *C. elegans*; SCE, *S. cerevisiae*; GLA, *G. lamblia*. ND, not determined.

[a] Present in *G. lamblia*.

[b] Present in *D. discoideum* and *E. histolytica* (according to PFAM website).

[c] Present in bacteria.

[d] Transcription regulators (TRs).

protozoan that arose early in eukaryote evolution. In general, the number of TFs and TRs increases with the number of genes in the genome, following a power law as observed before (Van Nimwegen 2003). TFs and TRs were found to be similarly abundant in algae and yeast; however, in these lineages they are considerably less frequent than in animals. Numbers of TFs and TRs in many cases were similar in mosses and animals, whereas gene numbers were often greater in angiosperms. In Chlamydomonas, we identified 147 putative TF and 87 putative TR coding sequences from 29 and 10 protein families, respectively, totaling 234 distinct proteins involved in the regulation of transcription (Table 1; protein sequences are available at http://plntfdb.bio.uni-potsdam.de/v2.0/index.php?sp_id=CRE). A schematic of the transcriptional regulatory proteins identified in Chlamydomonas, including their defining domains, is given in supplemental Figure 1. To date, however, the biological functions of only a small number of these proteins have been analyzed (supplemental Table 1).

Chlamydomonas transcription factors: In animals, TFs of the C2H2 and HB families play important roles in growth-related and development processes (Wu 2002) and body-plan formation (Deutsch and Mouchel-Vielh 2003). These two families are the largest in animals, with >100 members each in humans, Drosophila, and Caenorhabditis. In animals, HB TFs function as homeotic genes that control the formation and differentiation of different body parts (Garcia-Fernandez 2005; Negre and Ruiz 2007). In contrast, in plants homeotic functions are carried out by TFs of the MADS-box family (Irish 2003). Typically, angiosperms have ~80–120 MADS-box proteins, whereas such TFs are largely absent from animals (<10). Similarly, MADS-box TFs are present in only small numbers in Chlamydomonas (two genes) and in all other unicellular organisms. In contrast, in these organisms, members of the C2H2 family are slightly more abundant than members of the MADS-box family, with five, four, and seven genes, respectively, in the algae Chlamydomonas, Ostreococcus, and Cyanidioschyzon, and 39 members in Saccharomyces (Table 1). C2H2 TFs contain a zinc-finger domain. The recruitment of this domain for transcriptional regulation occurred in prokaryotes, and members of the Ros family may have been the origin of C2H2 in eukaryotes (Bouhouche *et al.* 2000). In general, TFs bearing a zinc-finger domain have significantly contributed to the evolution of eukaryotic organisms (Riechmann *et al.* 2000) either through gene duplication leading to an increased gene number or through the modulation of other domains present in these proteins, resulting in the formation of new families of TFs.

The acquisition of chloroplasts represents an important step in the evolutionary path that separated plants from animals and fungi. Evidently, new regulatory networks had to be established through evolution to achieve an optimal integration of photosynthetic functions with other cellular processes. TFs and TRs constitute important elements of such networks. Three families of putative TFs predate the split between rhodophyta (red algae) and chlorophyta, *i.e.*, G_2-like, PLATZ, and RWP-RK. These families appear to be of particular importance for the evolution of eukaryotic photosynthetic organisms, as they are the only plant-specific TFs (with perhaps the exception of G_2-like, which might also be present in Giardia; see above) that are present in both red and green algae. Both algal groups derived from the original

primary endosymbiotic event that led to the establishment of plastids (Reyes-Prieto *et al.* 2007). Nine additional families, *i.e.*, ABI3/VP1, AP2-EREBP, ARR-B, C2C2-CO-like, C2C2-Dof, PBF-2-like/Whirly, Pseudo ARR-B, SBP, and WRKY (Table 1), predate the split between green algae and streptophytes. Plant-specific TF families might have important roles in the control of light-dependent processes and related biochemical pathways such as those involved in sugar production or starch accumulation.

TFs of the G_2-like family (a distinct group within the GARP superfamily of TFs; Rossini *et al.* 2001) are present in all plants, including red and green algae, and in *G. lamblia*, suggesting a deep evolutionary origin, but they are not found in animals or fungi. G_2-like TFs regulate chloroplast development in diverse plant species (*e.g.*, Physcomitrella, Arabidopsis, and *Zea mays*) through a process that requires a close coordination between plastidial and nuclear genomes. More specifically, GOLDEN 2-like (GLK) TFs are required for correct stacking of thylakoids within chloroplasts, although it is not known in detail how they exert their function in this process. One possible model is that GLKs regulate the transcription of genes encoding thylakoid-stabilizing factor(s) (Yasumura *et al.* 2005). We did not detect the ortholog of *GLK* in the sequenced Chlamydomonas genome, which is consistent with the fact that chloroplast thylakoid stacking is less advanced in this alga as compared to bryophytes and angiosperms, as previously discussed (Yasumura *et al.* 2005). *PHOSPHORUS STARVATION RESPONSE1* (*PSR1*) from Chlamydomonas and its ortholog *PHOSPHATE STARVATION RESPONSE1* (*PHR1*) from Arabidopsis encode TFs that control cellular responses to phosphate deprivation (Wykoff *et al.* 1999; Rubio *et al.* 2001). Both proteins were originally thought to be members of the MYB TF family, but subsequently were placed within the GARP superfamily (G_2-like) (Fitter *et al.* 2002). PSR1 targets include genes encoding chloroplast-localized proteins involved in photosynthesis, regulation of gene expression, and other processes (Moseley *et al.* 2006). Recently, PSR1 has also been shown to control the accumulation of chloroplast RNA under phosphorous limitation through control of the expression of ribonuclease polynucleotide phosphorylase (Yehudai-Resheff *et al.* 2007). Whether the angiosperm ortholog exerts a similar function is currently unknown.

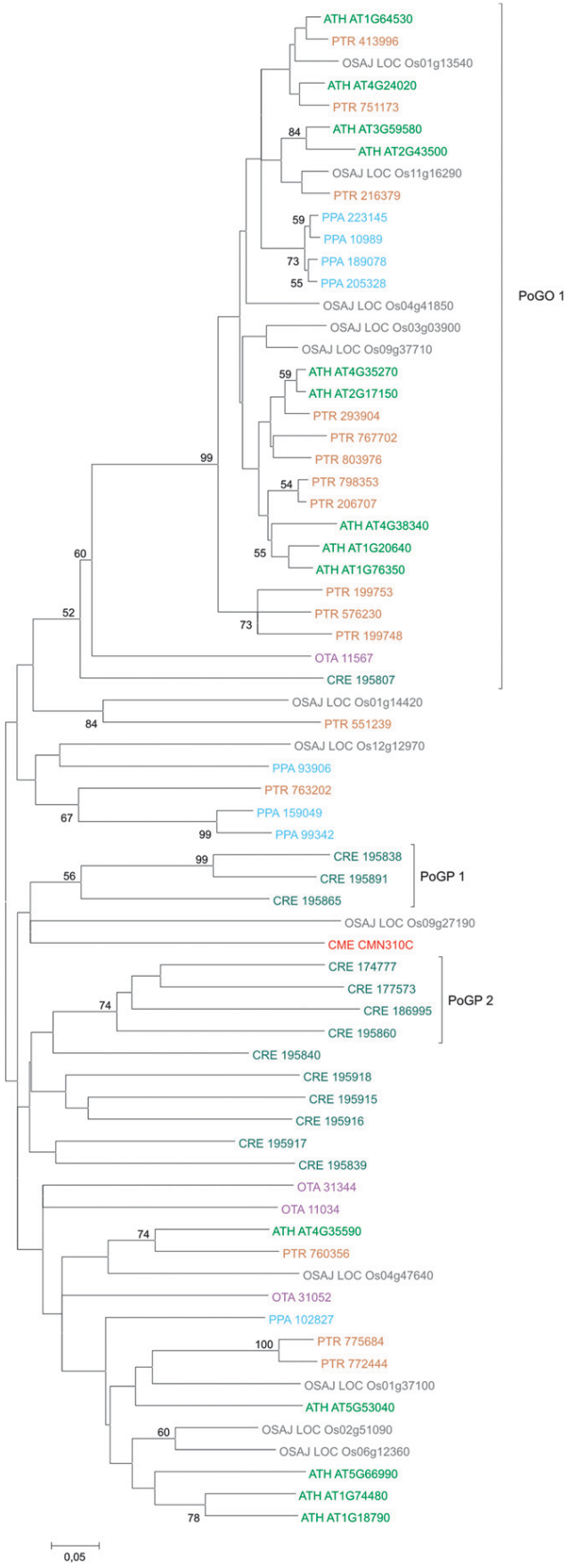
RWP-RK (Figure 1) is a TF family present in all green plants, as well as in red algae. It is also present in the early diverging amoebozoa *Dictyostelium discoideum* and *Entamoeba histolytica*, but not in animals or fungi, suggesting a deep evolutionary origin. In vascular plants, this family is involved in the regulation of genes in response to nitrogen status and nodule development in legumes (Schauser *et al.* 1995; Borisov *et al.* 2003). In Chlamydomonas, the gene *minus dominance* (*MID*; GenBank accession no. U92071; specific to mt^- strains and consequently not present in the sequenced strain, *i.e.*, CC-503 cw92 mt^+; Goodenough *et al.* 2007) is required for expression of *minus*-specific gamete-specific genes in response to nitrogen deprivation (Ferris and Goodenough 1997; Lin and Goodenough 2007). Another TF in this family, *NIT2*, is a positively acting regulatory gene of the nitrate assimilation pathway (Camargo *et al.* 2007) (GenBank accession no. DQ311647; this gene is mutated in the sequenced Chlamydomonas strain; Fernandez and Matagne 1984); the most similar entry in PlnTFDB is protein ID 195807).

Information regarding SBP TFs, Jumonji, and SET TRs, as well as microRNAs that often control TF genes in angiosperms but appear to be of minor importance in Chlamydomonas, is given in the supplemental text.

Transcription factors involved in hormone signaling: Phytohormones coordinate a vast spectrum of developmental and physiological processes in angiosperms. In contrast, knowledge about the occurrence of hormones in algae and their possible functions in cellular signaling is extremely limited. Some evidence indicates that auxins and cytokinins are present in algae (Tarakhovskaya *et al.* 2007), indicating their functional importance early in plant evolution. TF families known to participate in hormone signaling in angiosperms are also found in Chlamydomonas (Table 1). Recent work on ABSCISIC ACID INSENSITIVE 3 (ABI3) from Arabidopsis has indicated a possible role in cross talk of abscisic acid and auxin response pathways (Brady *et al.* 2003; Rock and Sun 2005). A similar observation was made for VP1 from maize, the ortholog of ABI3 (Suzuki *et al.* 2001). We observed a single *ABI3/VP1* gene in Chlamydomonas, whereas Physcomitrella has 30 *ABI3/VP1* genes, and angiosperms have ~60–80 (Table 1). To our knowledge the role of the *ABI3/VP1* gene in Chlamydomonas has not been characterized yet. TFs of the ARR-B and AP2-EREBP families are involved in cytokinin response pathways in angiosperms (Rashotte *et al.* 2006; Ishida *et al.* 2008). We detected one *ARR-B* gene and 11 *AP2-EREBP* genes in Chlamydomonas (Table 1). The role of these TFs has not been analyzed.

TF families absent from algae: Interestingly, 22 plant-specific TF families and 1 TR family are not present in algae (Table 1). These families may be related to the acquisition of multicellularity and tissue organization, invasion of the terrestrial environment, and long-distance trafficking. NAC TFs could be identified only from bryophytes onward. Functional studies have shown that several *NAC* genes play an important role in cell differentiation (Olsen *et al.* 2005). As we did not find any *NAC* gene in the *Volvox carteri* genome (not shown), we assume that TFs of this family were not important for establishing multicellularity in this organism.

Orthologs across green plants: The green plant lineage is a monophyletic group, its members having split from the red algal lineage ~1142 ± 167 million years ago (Zimmer *et al.* 2007). Tracing gene orthology relations across lineages provides a way to assess, to some



extent, the forces driving the functional diversification of multigene families. As reported previously, TF genes in plants have a higher retention rate after duplication than other genes (Seoighe and Gehring 2004; De Bodt *et al.* 2005). Additionally, genes functionally related to stress responses tend to undergo a more intense duplication process (Shiu *et al.* 2004). Therefore, TF gene families are well suited to trace back important events in evolution.

In our NJ analyses (for examples, see supplemental Figures 2–7), we have identified 120 clusters of orthologs with 1183 genes in total. Seventy-one of them are conserved in all green plants, and 26 are also common to red algae (see supplemental Table 2). Clusters to which functions could be assigned are involved mainly in light perception/response, control of plastidial gene expression, regulation of circadian rhythm, and the transition from the vegetative to the reproductive phase of growing plants (data not shown). Moreover, 38 of these clusters were found to have a one-to-one relationship (they do not possess any paralog inside the same group of orthologs). Such genes tend to exert key biological functions (Shiu *et al.* 2004). The greatest number of clusters, *i.e.*, 20, among all green plants is represented by the *SWI2/SNF2* gene family that encodes proteins involved in chromatin remodeling and thus the regulation of transcription, replication, and DNA recombination and repair. In plants, some Swi2/Snf2 proteins have been studied (Shaked *et al.* 2006), but a detailed functional analysis is missing for most of them. In the RWP-RK family, we found only one PoGO (Figure 1) in representatives from all green plants. The position of the *C. merolae* sequence is not evident from this analysis. In addition, groups of paralogs of Chlamydomonas are shown.

We also made a comparison between the clusters of orthologs obtained by phylogenetic analysis and by the InParanoid-Graph theoretic approach (see supplemental Table 3). In total, 446 genes from both classifications overlap, representing 99 of the 120 clusters obtained by the NJ analysis, and 98 of 168 clusters identified by the InParanoid approach (see supplemental Table 4). Thus, a large number of clusters was identified irrespective of the detection method used. We computed the Adjusted

Figure 1.—Phylogenetic tree of RWP-RK TFs in plants. We identified one PoGO (PoGO 1) conserved in all green plants, which includes the NIT2 TF (CRE 195807), a regulatory factor of genes involved in the nitrate assimilation pathway. Additionally, there are two possible groups of paralogs (PoGP 1 and PoGP 2) of Chlamydomonas genes. Red, *C. merolae* (CME); violet, *O. tauri* (OTA); light green, *C. reinhardtii* (CRE); light blue, *P. patens* (PPA); green, *A. thaliana* (ATH); brown, *P. trichocarpa* (PTR); gray, *O. sativa* ssp. *japonica* (OSAJ). The first three letters of the sequence name indicate the species (the first four letters in the case of OSAJ), and the remaining letters or numbers represent the accession code through which the respective sequence can be retrieved from the PlnTFDB (http://plntfdb.bio.uni-potsdam.de/v2.0).

Rand Index (*ar*) on the subset of common genes. The obtained value ($ar = 0.912$) indicates that the composition of the common clusters (gene membership) obtained by both methods is similar.

Evolution of photosynthetic networks: A recent review by Jiao *et al.* (2007) provides a good backbone for comparison of the light-regulated transcriptional networks of angiosperms and Chlamydomonas. The perception of light signals in dicots occurs through three cryptochromes and two phototropins, for which we found orthologs in Chlamydomonas (see supplemental Table 5). In contrast, phytochromes involved in the absorption of red and far-red light do not have homologs in green algae, consistent with previous findings (Mittag *et al.* 2005). One putative ortholog of the angiosperm bZIP protein COMMON PLANT REGULATORY FACTORS 1 (CPRF1), CMJ034C, was found in the red alga Cyanidioschyzon, although with a low InParanoid confidence score. The same Cyanidioschyzon protein is also orthologous to G-BOX BINDING FACTOR 1 (GBF1), suggesting subfunctionalization of the original multifunctional algal gene during angiosperm evolution; however, more detailed analyses are required to substantiate this hypothesis. GBF1 is phosphorylated by CASEIN KINASE II (CKII), which allows it to bind to target promoters containing the G-box, a well-defined light-response element. We found a putative CKII ortholog in Chlamydomonas, suggesting that light-dependent post-translational protein modification of the GBF1 ortholog was established early in plant evolution. Another important regulatory mechanism is the ubiquitin-mediated degradation of the bZIP TF ELONGATED HYPOCOTYL 5 (HY5) that is triggered by CONSTITUTIVE PHOTOMORPHOGENIC 1 (COP1) and its associated protein PHYTOCHROME A SUPPRESSOR 1 (SPA1). Both *HY5* and *SPA1* orthologs were found in green algae (see supplemental Table 5), whereas *COP1* has so far been found only in red algae. As the bZIP degradation mechanism triggered by COP1 is conserved throughout the plant kingdom (Yi and Deng 2005), one might speculate that COP1 is also present in Chlamydomonas. In summary, most of the components of the light-regulated transcriptional networks are shared between Chlamydomonas and seed plants, although phytochromes are missing in green algae. PHYTOCHROME INTERACTING FACTOR (PIF) TFs represent a subgroup of the bHLH family in angiosperms. No *PIF* ortholog could be found in Chlamydomonas. In addition, most other factors that are directly involved in phytochrome activity in land plants, such as FHY3, FAR1, HFR1, ATHB4, and PAR1, are absent from Chlamydomonas (see supplemental Table 5).

Conclusions: We have identified 147 putative TFs and 87 putative TRs in Chlamydomonas. Three TF families predate the rhodophyta–viridiplantae divide, while nine more of the TF families predate the chlorophyta–streptophyta divide and diversified further in bryophytes and angiosperms. However, we also observed that 22 plant-specific TF and 1 plant-specific TR family were not present in algae, highlighting their importance for the evolution of multicellular plants. Many of the elements of light-regulated transcriptional networks known from bryophytes and angiosperms are also present in Chlamydomonas, indicating an early evolutionary origin. Exceptions are elements of the phytochrome-mediated signaling pathways that are missing in algae. Our analysis provides a basis for further experimental studies on Chlamydomonas transcriptional regulators.

We thank the three anonymous reviewers for comments that helped to improve our manuscript. We are grateful to the Department of Energy Joint Genome Institute and the Chlamydomonas research community for sequencing and annotating the Chlamydomonas genome. L.G.G.C. and B.M.-R. thank the Interdisciplinary Research Centre, Advanced Protein Technologies, of the University of Potsdam and the International Ph.D. Programme, Integrative Plant Science [supported by the Deutscher Akademischer Austauschdienst and the Deutsche Forschungsgemeinschaft (DAAD), no. DAAD D/04/01336] for financial support. L.G.G.C. thanks the DAAD for providing a scholarship (no. A/04/34814). B.M.-R. thanks the Fonds der Chemischen Industrie for financial support (no. 0164389). D.M.R.-P. acknowledges financial support by the Bundesministerium fuer Bildung und Forschung (BMBF) (GABI-future grant 0315046). B.M.-R. and R.T.-E. thank the BMBF for funding of the systems biology research unit GoFORSYS–Potsdam-Golm BMBF Forschungseinrichtung zur Systembiologie [(Photosynthesis and Growth: A Systems Biology Based Approach (FKZ 0313924)].

LITERATURE CITED

Adams, M. D., S. E. Celniker, R. A. Holt, C. A. Evans, J. D. Gocayne *et al.*, 2000 The genome sequence of *Drosophila melanogaster*. Science **287:** 2185–2195.

Arabidopsis Genome Initiative, 2000 Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature **408:** 796–815.

Best, A. A., H. G. Morrison, A. G. McArthur, M. L. Sogin and G. J. Olsen, 2004 Evolution of eukaryotic transcription: insights from the genome of *Giardia lamblia*. Genome Res. **14:** 1537–1547.

Borisov, A. Y., L. H. Madsen, V. E. Tsyganov, Y. Umehara, V. A. Voroshilova *et al.*, 2003 The Sym35 gene required for root nodule development in pea is an ortholog of Nin from *Lotus japonicus*. Plant Physiol. **131:** 1009–1017.

Bouhouche, N., M. Syvanen and C. I. Kado, 2000 The origin of prokaryotic C2H2 zinc finger regulators. Trends Microbiol. **8:** 77–81.

Brady, S. M., S. F. Sarkar, D. Bonetta and P. McCourt, 2003 The ABSCISIC ACID INSENSITIVE 3 (ABI3) gene is modulated by farnesylation and is involved in auxin signaling and lateral root development in Arabidopsis. Plant J. **34:** 67–75.

Camargo, A., A. Llamas, R. A. Schnell, J. J. Higuera, D. Gonzalez-Ballester *et al.*, 2007 Nitrate signaling by the regulatory gene NIT2 in Chlamydomonas. Plant Cell **19:** 3491–3503.

C. elegans Sequencing Consortium, 1998 Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science **282:** 2012–2018.

Chen, W., N. J. Provart, J. Glazebrook, F. Katagiri, H. S. Chang *et al.*, 2002 Expression profile matrix of Arabidopsis transcription factor genes suggests their putative functions in response to environmental stresses. Plant Cell **14:** 559–574.

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

de Nooy, W., A. Mrvar and A. Batagelj, 2005 *Exploratory Social Network Analysis With Pajek (Structural Analysis in the Social Sciences)*. Cambridge University Press, New York.

Derelle, E., C. Ferraz, S. Rombauts, P. Rouze, A. Z. Worden *et al.*, 2006 Genome analysis of the smallest free-living eukaryote

Ostreococcus tauri unveils many unique features. Proc. Natl. Acad. Sci. USA **103:** 11647–11652.

Deutsch, J. S., and E. Mouchel-Vielh, 2003 Hox genes and the crustacean body plan. BioEssays **25:** 878–887.

Fernandez, E., and R. F. Matagne, 1984 Genetic analysis of nitrate reductase-deficient mutants in *Chlamydomonas reinhardtii*. Curr. Genet. **8:** 635–640.

Ferris, P. J., and U. W. Goodenough, 1997 Mating type in Chlamydomonas is specified by mid, the minus-dominance gene. Genetics **146:** 859–869.

Finn, R. D., J. Mistry, B. Schuster-Bockler, S. Griffiths-Jones, V. Hollich *et al.*, 2006 Pfam: clans, web tools and services. Nucleic Acids Res. **34:** D247–D251.

Fitter, D. W., D. J. Martin, M. J. Copley, R. W. Scotland and J. A. Langdale, 2002 GLK gene pairs regulate chloroplast development in diverse plant species. Plant J. **31:** 713–727.

Garcia-Fernandez, J., 2005 The genesis and evolution of homeobox gene clusters. Nat. Rev. Genet. **6:** 881–892.

Goff, S. A., D. Ricke, T. H. Lan, G. Presting, R. Wang *et al.*, 2002 A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). Science **296:** 92–100.

Goffeau, A., B. G. Barrell, H. Bussey, R. W. Davis, B. Dujon *et al.* 1996 Life with 6000 genes. Science **274:** 546, 563–567.

Goodenough, U., H. Lin and J. H. Lee, 2007 Sex determination in Chlamydomonas. Semin. Cell Dev. Biol. **18:** 350–361.

Hubert, L., and P. Arabie, 1985 Comparing partitions. J. Classification **2:** 193–218.

International Human Genome Sequencing Consortium, 2004 Finishing the euchromatic sequence of the human genome. Nature **431:** 931–945.

Irish, V. F., 2003 The evolution of floral homeotic gene function. BioEssays **25:** 637–646.

Ishida, K., T. Yamashino, A. Yokoyama and T. Mizuno, 2008 Three type-B response regulators, ARR1, ARR10 and ARR12, play essential but redundant roles in cytokinin signal transduction throughout the life cycle of *Arabidopsis thaliana*. Plant Cell Physiol. **49:** 47–57.

Jaillon, O., J. M. Aury, B. Noel, A. Policriti, C. Clepet *et al.*, 2007 The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature **449:** 463–467.

Jiao, Y., O. S. Lau and X. W. Deng, 2007 Light-regulated transcriptional networks in higher plants. Nat. Rev. Genet. **8:** 217–230.

Kersey, P., L. Bower, L. Morris, A. Horne, R. Petryszak *et al.*, 2005 Integr8 and Genome Reviews: integrated views of complete genomes and proteomes. Nucleic Acids Res. **33:** D297–D302.

Kumar, S., K. Tamura and M. Nei, 2004 MEGA3: Integrated software for molecular evolutionary genetics analysis and sequence alignment. Brief. Bioinform. **5:** 150–163.

Lemon, B., and R. Tjian, 2000 Orchestrated response: a symphony of transcription factors for gene control. Genes Dev. **14:** 2551–2569.

Li, L., X. Wang, R. Sasidharan, V. Stolc, W. Deng *et al.*, 2007 Global identification and characterization of transcriptionally active regions in the rice genome. PLoS ONE **2:** e294.

Lin, H., and U. W. Goodenough, 2007 Gametogenesis in the *Chlamydomonas reinhardtii* minus mating type is controlled by two genes, MID and MTD1. Genetics **176:** 913–925.

Merchant, S. S., S. E. Prochnik, O. Vallon, E. H. Harris, S. J. Karpowicz *et al.*, 2007 The Chlamydomonas genome reveals the evolution of key animal and plant functions. Science **318:** 245–250.

Mittag, M., S. Kiaulehn and C. H. Johnson, 2005 The circadian clock in *Chlamydomonas reinhardtii*. What is it for? What is it similar to? Plant Physiol. **137:** 399–409.

Moseley, J. L., C. W. Chang and A. R. Grossman, 2006 Genome-based approaches to understanding phosphorus deprivation responses and PSR1 control in *Chlamydomonas reinhardtii*. Eukaryot. Cell **5:** 26–44.

Negre, B., and A. Ruiz, 2007 HOM-C evolution in Drosophila: Is there a need for Hox gene clustering? Trends Genet. **23:** 55–59.

Nei, M., and S. Kumar, 2000 *Molecular Evolution and Phylogenetics*. Oxford University Press, New York.

Nozaki, H., H. Takano, O. Misumi, K. Terasawa, M. Matsuzaki *et al.*, 2007 A 100%-complete sequence reveals unusually simple genomic features in the hot-spring red alga *Cyanidioschyzon merolae*. BMC Biol. **5:** 28.

Olsen, A. N., H. A. Ernst, L. L. Leggio and K. Skriver, 2005 NAC transcription factors: structurally distinct, functionally diverse. Trends Plant Sci. **10:** 79–87.

Rand, M. W., 1971 Objective criteria for the evaluation of clustering methods. J. Am. Stat. Assoc. **66:** 846–850.

Rashotte, A. M., M. G. Mason, C. E. Hutchison, F. J. Ferreira, G. E. Schaller *et al.*, 2006 A subset of Arabidopsis AP2 transcription factors mediates cytokinin responses in concert with a two-component pathway. Proc. Natl. Acad. Sci. USA **103:** 11081–11085.

R Development Core Team, 2007 *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.

Remm, M., C. E. Storm and E. L. Sonnhammer, 2001 Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J. Mol. Biol. **314:** 1041–1052.

Rensing, S. A., D. Lang, A. D. Zimmer, A. Terry, A. Salamov *et al.*, 2008 The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants. Science **319:** 64–69.

Reyes-Prieto, A., A. P. Weber and D. Bhattacharya, 2007 The origin and establishment of the plastid in algae and plants. Annu. Rev. Genet. **41:** 147–168.

Riano-Pachon, D. M., S. Ruzicic, I. Dreyer and B. Mueller-Roeber, 2007 PlnTFDB: an integrative plant transcription factor database. BMC Bioinformatics **8:** 42.

Riechmann, J. L., J. Heard, G. Martin, L. Reuber, C. Jiang *et al.*, 2000 Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. Science **290:** 2105–2110.

Rock, C. D., and X. Sun, 2005 Crosstalk between ABA and auxin signaling pathways in roots of *Arabidopsis thaliana* (L.) Heynh. Planta **222:** 98–106.

Rossini, L., L. Cribb, D. J. Martin and J. A. Langdale, 2001 The maize golden2 gene defines a novel class of transcriptional regulators in plants. Plant Cell **13:** 1231–1244.

Rubio, V., F. Linhares, R. Solano, A. C. Martin, J. Iglesias *et al.*, 2001 A conserved MYB transcription factor involved in phosphate starvation signaling both in vascular plants and in unicellular algae. Genes Dev. **15:** 2122–2133.

Schauser, L., L. Christensen, S. Borg and C. Poulsen, 1995 PZF, a cDNA isolated from *Lotus japonicus* and soybean root nodule libraries, encodes a new plant member of the RING-finger family of zinc-binding proteins. Plant Physiol. **107:** 1457–1458.

Seoighe, C., and C. Gehring, 2004 Genome duplication led to highly selective expansion of the *Arabidopsis thaliana* proteome. Trends Genet. **20:** 461–464.

Shaked, H., N. Avivi-Ragolsky and A. A. Levy, 2006 Involvement of the Arabidopsis SWI2/SNF2 chromatin remodeling gene family in DNA damage response and recombination. Genetics **173:** 985–994.

Shiu, S. H., W. M. Karlowski, R. Pan, Y. H. Tzeng, K. F. Mayer *et al.*, 2004 Comparative analysis of the receptor-like kinase family in Arabidopsis and rice. Plant Cell **16:** 1220–1234.

Suzuki, M., C. Y. Kao, S. Cocciolone and D. R. McCarty, 2001 Maize VP1 complements Arabidopsis abi3 and confers a novel ABA/auxin interaction in roots. Plant J. **28:** 409–418.

Tarakhovskaya, E. R., Y. I. Maslov and M. F. Shishova, 2007 Phytohormones in algae. Russ. J. Plant Physiol. **54:** 163–170.

Thompson, J. D., T. J. Gibson, F. Plewniak, F. Jeanmougin and D. G. Higgins, 1997 The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. **25:** 4876–4882.

Tuskan, G. A., S. Difazio, S. Jansson, J. Bohlmann, I. Grigoriev *et al.*, 2006 The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science **313:** 1596–1604.

van Nimwegen, E., 2003 Scaling laws in the functional content of genomes. Trends Genet. **19:** 479–484.

Vincentz, M., C. Bandeira-Kobarg, L. Gauer, P. Schlogl and A. Leite, 2003 Evolutionary pattern of angiosperm bZIP factors homologous to the maize Opaque2 regulatory protein. J. Mol. Evol. **56:** 105–116.

Wu, L. C., 2002 ZAS: C2H2 zinc finger proteins involved in growth and development. Gene Expr. **10:** 137–152.

Wykoff, D. D., A. R. Grossman, D. P. Weeks, H. Usuda and K. Shimogawara, 1999 Psr1, a nuclear localized protein that regulates phosphorus metabolism in Chlamydomonas. Proc. Natl. Acad. Sci. USA **96:** 15336–15341.

Yasumura, Y., E. C. Moylan and J. A. Langdale, 2005 A conserved transcription factor mediates nuclear control of organelle biogenesis in anciently diverged land plants. Plant Cell **17:** 1894–1907.

Yehudai-Resheff, S., S. L. Zimmer, Y. Komine and D. B. Stern, 2007 Integration of chloroplast nucleic acid metabolism into the phosphate deprivation response in Chlamydomonas reinhardtii. Plant Cell **19:** 1023–1038.

Yi, C., and X. W. Deng, 2005 COP1: from plant photomorphogenesis to mammalian tumorigenesis. Trends Cell Biol. **15:** 618–625.

Yu, J., S. Hu, J. Wang, G. K. Wong, S. Li *et al.*, 2002 A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). Science **296:** 79–92.

Zimmer, A., D. Lang, S. Richardt, W. Frank, R. Reski *et al.*, 2007 Dating the early evolution of plants: detection and molecular clock analyses of orthologs. Mol. Genet. Genomics **278:** 393–402.

Communicating editor: S. Dutcher

4

bZIP transcription factors in plants

The Role of bZIP Transcription Factors in Green Plant Evolution: Adaptive Features Emerging From Four Founder Genes

Luiz Gustavo Guedes Corrêa[1,2,3,†], Diego Mauricio Riaño-Pachón[2,4,†], Carlos Guerra Schrago[5], Renato Vicentini dos Santos[1], Bernd Mueller-Roeber[2,3], Michel Vincentz[1]

† These authors contributed equally to this work and should thus both be considered as first authors.

[1] Centro de Biologia Molecular e Engenharia Genética, Departamento de Genética e Evolução, Instituto de Biologia, Universidade Estadual de Campinas, Campinas SP, Brazil, [2] Department of Molecular Biology, University of Potsdam, Potsdam–Golm, Germany, [3] Cooperative Research Group, Max Planck Institute of Molecular Plant Physiology, Potsdam–Golm, Germany, [4]Bioinformatics Research Group, GabiPD Team, Max-Planck Institute of Molecular Plant Physiology, Potsdam–Golm, Germany, [5] Laboratório de Biodiversidade Molecular, Departamento de Genética, Universidade Federal do Rio de Janeiro, Cidade Universitária, Rio de Janeiro RJ, Brazil

Published in *PLoS ONE* (2008) **3**(8):e2944. doi:10.1371/journal.pone.0002944

Author contributions

MV, LGGC and DMRP conceived and designed the study. BMR and MV coordinated the project. LGGC carried out phylogenetic analyses by NJ, made the comparison based on expression profiles, identified conserved motifs using MEME. DMRP carried out phylogenetic analyses by ML, made the comparison based on MPSS data, carried out bZIP searches on EST data collections, identified putative pseudogenes of bZIP TFs. DMRP and CGS perform the analysis about gene family expansions. All authors discussed and analysed the data.

OPEN ACCESS Freely available online

PLoS one

The Role of bZIP Transcription Factors in Green Plant Evolution: Adaptive Features Emerging from Four Founder Genes

Luiz Gustavo Guedes Corrêa[1,2,3☯], Diego Mauricio Riaño-Pachón[2,4☯], Carlos Guerra Schrago[5], Renato Vicentini dos Santos[1], Bernd Mueller-Roeber[2,3], Michel Vincentz[1]*

1 Centro de Biologia Molecular e Engenharia Genética, Departamento de Genética e Evolução, Instituto de Biologia, Universidade Estadual de Campinas, Campinas, Brazil, **2** Department of Molecular Biology, University of Potsdam, Potsdam-Golm, Germany, **3** Cooperative Research Group, Max-Planck Institute of Molecular Plant Physiology, Potsdam-Golm, Germany, **4** GabiPD Team, Bioinformatics Group, Max-Planck Institute of Molecular Plant Physiology, Potsdam-Golm, Germany, **5** Laboratório de Biodiversidade Molecular, Departamento de Genética, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil

Abstract

Background: Transcription factors of the basic leucine zipper (bZIP) family control important processes in all eukaryotes. In plants, bZIPs are regulators of many central developmental and physiological processes including photomorphogenesis, leaf and seed formation, energy homeostasis, and abiotic and biotic stress responses. Here we performed a comprehensive phylogenetic analysis of bZIP genes from algae, mosses, ferns, gymnosperms and angiosperms.

Methodology/Principal Findings: We identified 13 groups of bZIP homologues in angiosperms, three more than known before, that represent 34 Possible Groups of Orthologues (PoGOs). The 34 PoGOs may correspond to the complete set of ancestral angiosperm bZIP genes that participated in the diversification of flowering plants. Homologous genes dedicated to seed-related processes and ABA-mediated stress responses originated in the common ancestor of seed plants, and three groups of homologues emerged in the angiosperm lineage, of which one group plays a role in optimizing the use of energy.

Conclusions/Significance: Our data suggest that the ancestor of green plants possessed four bZIP genes functionally involved in oxidative stress and unfolded protein responses that are bZIP-mediated processes in all eukaryotes, but also in light-dependent regulations. The four founder genes amplified and diverged significantly, generating traits that benefited the colonization of new environments.

Citation: Guedes Corrêa LG, Riaño-Pachón DM, Guerra Schrago C, Vicentini dos Santos R, Mueller-Roeber B, et al. (2008) The Role of bZIP Transcription Factors in Green Plant Evolution: Adaptive Features Emerging from Four Founder Genes. PLoS ONE 3(8): e2944. doi:10.1371/journal.pone.0002944

Editor: Shin-Han Shiu, Michigan State University, United States of America

Received February 18, 2008; **Accepted** July 22, 2008; **Published** August 13, 2008

Funding: L.G.G.C. thanks the DAAD for providing a scholarship (A/04/34814). D.M.R.P. acknowledges financial support from the BMBF (FKZ 0315046). This work was supported in part by grants from the Fundação de Amparo a Ciência do Estado de São Paulo (FAPESP), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes) (to L.G.G.C and M.V.), the University of Potsdam Interdisciplinary Research Centre 'Advanced Protein Technologies' (to B.M.-R.), the DAAD/DFG International PhD Programme 'Integrative Plant Science' (DAAD D/04/01336; to B.M-R.), and the Fonds der Chemischen Industrie (N° 0164389; to B.M.-R.).

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: mgavince@unicamp.br

☯ These authors contributed equally to this work.

Introduction

Growth and development of all organisms depend on proper regulation of gene expression. The control of transcription initiation rates by transcription factors (TF) represents one of the most important means of modulating gene expression [1–4]. TFs can be grouped into different protein families according to their primary and/or three-dimensional structure similarities in the DNA-binding and multimerization domains [4–6]. The interplay between the amplification of the ancestral repertoire of TFs, the emergence of new TFs, the combination of protein domains and sequence divergence constitutes an important driving force towards the evolution of organismic complexity [7–10]. Understanding the detailed evolutionary history of these TFs and their corresponding functions is therefore crucial to reveal the changes and/or innovations in transcriptional regulatory circuits that underlie the biological diversity found among eukaryotes.

Large scale genomic comparisons revealed that angiosperm TF families undergo more intense gene expansion when compared to animals and fungi, possibly reflecting the ability of flowering plants to efficiently adapt to different and unstable environmental conditions. Moreover, gene expansion rates vary among plant TF families, indicating lineage-differential specializations [11,12]. For instance, MADS-box and homeodomain families, which exert similar functions in developmental control, expanded preferentially in the angiosperm and human lineages, respectively [13,14]. Contrariwise, the basic leucine zipper (bZIP) TF family apparently expanded to a similar extent in angiosperms and humans [15]. Currently we do not well understand why individual TF families underwent differential evolutionary expansions in the different

eukaryotic lineages. Therefore, a deep evolutionary analysis of TF families including the identification of the founding (ancestral) gene sets in combination with functional assignments will greatly assist in addressing this issue [16,17].

To our knowledge, however, only four families that are present in all green plants have until today been studied in a deep evolutionary scale, Dof [18], homeodomain [19], MADS-box [20,21] and WRKY [22]. As a matter of fact, groups of orthologues, for which functional equivalence is often assumed, are rarely identified in a systematic and direct manner, with the exception of the HD-Zip class III subfamily [23,24]. It is thus often difficult to infer ancestral functions at different time points of the evolutionary process. Here we performed a comprehensive analysis of the evolutionary relationships of TFs of the green plant bZIP family; homologous and orthologous relationships among bZIP TFs were established and ancestral functions were inferred.

The bZIP TFs are characterized by a 40- to 80-amino-acid-long conserved domain (bZIP domain) that is composed of two motifs: a basic region responsible for specific binding of the TF to its target DNA, and a leucine zipper required for TF dimerization [5,25]. Genetic, molecular and biochemical analyses indicate that bZIPs are regulators of important plant processes such as organ and tissue differentiation [26–30], cell elongation [31,32], nitrogen/carbon balance control [33,34], pathogen defence [35–40], energy metabolism [41], unfolded protein response [42,43], hormone and sugar signalling [44–47], light response [48–50], osmotic control [34,51], and seed storage protein gene regulation [52]. Initially, 50 plant bZIP proteins were classified into five families, taking into account similarities of their bZIP domain [53]. An original investigation of the complete *Arabidopsis thaliana* genome sequence indicated the presence of 81 putative *bZIP* genes [54,55]. However, further detailed studies revealed 75 to 77 bZIP proteins to be encoded by the Arabidopsis nuclear genome, representing members of ten groups of homologues [55,56].

The availability of the rice (*Oryza sativa*) [57,58], black cottonwood (*Populus trichocarpa*) [59] and Arabidopsis genomic sequences [54] provides an exciting opportunity for the large-scale investigation of the genetic bases that underlies the extensive physiological and morphological diversity amongst the two main angiosperm divisions: monocots and eudicots. A possible comparative approach involves the establishment of relationships between different genomes in a homologous gene system [60–62], in which each group of orthologues is derived from an ancestral gene that underwent numerous modifications throughout evolution, including duplication and subsequent functional diversification. Considering that all genes of a given group of orthologues have the same ancestral origin, the establishment of this classification should allow the transfer of biochemical, structural and functional information from one protein to another, inside the same group [63]. Moreover, the relationships within a group of orthologues constitute the basis for a better understanding of the evolution of ancestral functions (conservation versus neo- or sub-functionalization through duplication) [64–66].

In this study, we identified the possible non-redundant complete sets of bZIPs in rice, comprising 92 proteins, and in black cottonwood, comprising 89 proteins. These collections of bZIPs together with the 77 bZIPs from Arabidopsis [56] could be divided, based on bZIP domain and other conserved motifs similarities, into 13 groups of bZIP homologues in angiosperms, three more than previously reported [55]. The identified groups constituted a backbone for a more detailed analysis of each group, to which additional bZIP sequences reported from other plants, including those deduced from expressed sequence tags (ESTs), were added. In total, we defined 34 Possible Groups of Orthologues (PoGOs), which may represent 34 ancestral functions in angiosperms. Interestingly, one PoGO was found exclusively in monocots, whereas a Possible Group of Paralogues (PoGP) appears to be restricted to Arabidopsis.

To extend our bZIP analysis to all major lineages of green plants we additionally identified and incorporated bZIP sequences not only from two algal (*Chlamydomonas reinhardtii* [67] and *Ostreococcus tauri* [68]) and moss (*Physcomitrella patens* [69]) genomes, but also from ESTs of the ferns *Selaginella moellendorffii* and *Adiantum capillus-veneris* and the gymnosperms *Pinus taeda* and *Picea glauca*. Based on this investigation, a model for the evolution of *bZIP* genes in green plants, based on four founder genes representing an ancestral tool kit, was established. Its main points are discussed here. We also propose an updated classification of plant *bZIP* genes which should facilitate functional studies.

Results and Discussion

Groups of Homologues of Angiosperm *bZIP* Genes

The Arabidopsis genome encodes for a possible complete set of 77 unique bZIP proteins, representing an update of previous results [55,56,70]. *AtbZIP73* contains a premature stop codon and was thus not considered further in our analyses. As it appears to be a pseudogene it should be referred to as *ΨAtbZIP73*. Through iterated searches with tblastn and blastx algorithms, and PFAM bZIP Hidden Markov Models (HMM), we identified 92 *bZIP* genes in rice (Text S1a). Recently, Nijhawan *et al.* [71] reported the presence of 89 *bZIP* genes in rice and their phylogenetic relationship to the Arabidopsis *bZIPs*. Of the 89 bZIPs, 86 are also present in this study. Careful sequence analyses of both gene sets revealed complete sequence identity of the Os06g50480 and Os06g50830 TFs, and complete identity with TF Os06g50600 (OsbZIP14) along amino acids 1–143, indicating that these sequences were redundant in the Nijhawan *et al.* data set. *Os03g59460* has also been identified in our studies, however, the protein it encodes contains a proline residue at the beginning of its leucine zipper, precluding dimerization [25]; thus it may not function like other known bZIPs. Despite *OsbZIP24* and *OsbZIP75* being classified as retrotransposons in TIGR, we included them in our analysis as they possess a standard bZIP sequence in their open reading frame. Table S1 gives a summary of this information.

We identified 89 bZIP sequences in *P. trichocarpa*, some of which were incomplete. We therefore performed a more refined analysis of genomic data sets taking into account gene structures and conserved motifs. This allowed us to resolve the entire *bZIP* gene sequences in nine cases (Datasets S1 and S2).

Through Neighbour-Joining (NJ) analysis of the minimum bZIP domain (44 amino acids; Text S1a) of 257 unique bZIPs from Arabidopsis, rice and black cottonwood (bZARP data set) we identified seven clusters of proteins with bootstrap support greater than or equal to 50%, defining the groups of homologous genes B, D, F, G, H, J and K. The topology of the phylogenetic tree and a bootstrap support of 50% indicate that Groups D and F are sister groups that share a common ancestor (Figures 1A and S1). Although Group A has a weaker bootstrap support in NJ analyses (34% using PAM matrix data, and 58% using p-distance values), its members were kept together for two main reasons: (i) all its member genes share a common motif in accordance with previous results from Jakoby *et al.* [55]; (ii) all genes but *Gbf4* (*AtbZIP40*) and *AtbZIP13* from Arabidopsis share common intron positions, suggesting a single evolutionary origin (Text S1b, and Figure S2). In Group F a clear tendency for loss of introns was observed. None of the rice *bZIP* genes contains introns, nor do the black cottonwood genes *PtrbZIP39* and *PtrbZIP40*. Although *PtrbZIP38* and *PtrbZIP41* have introns, they lost it from the conserved basic

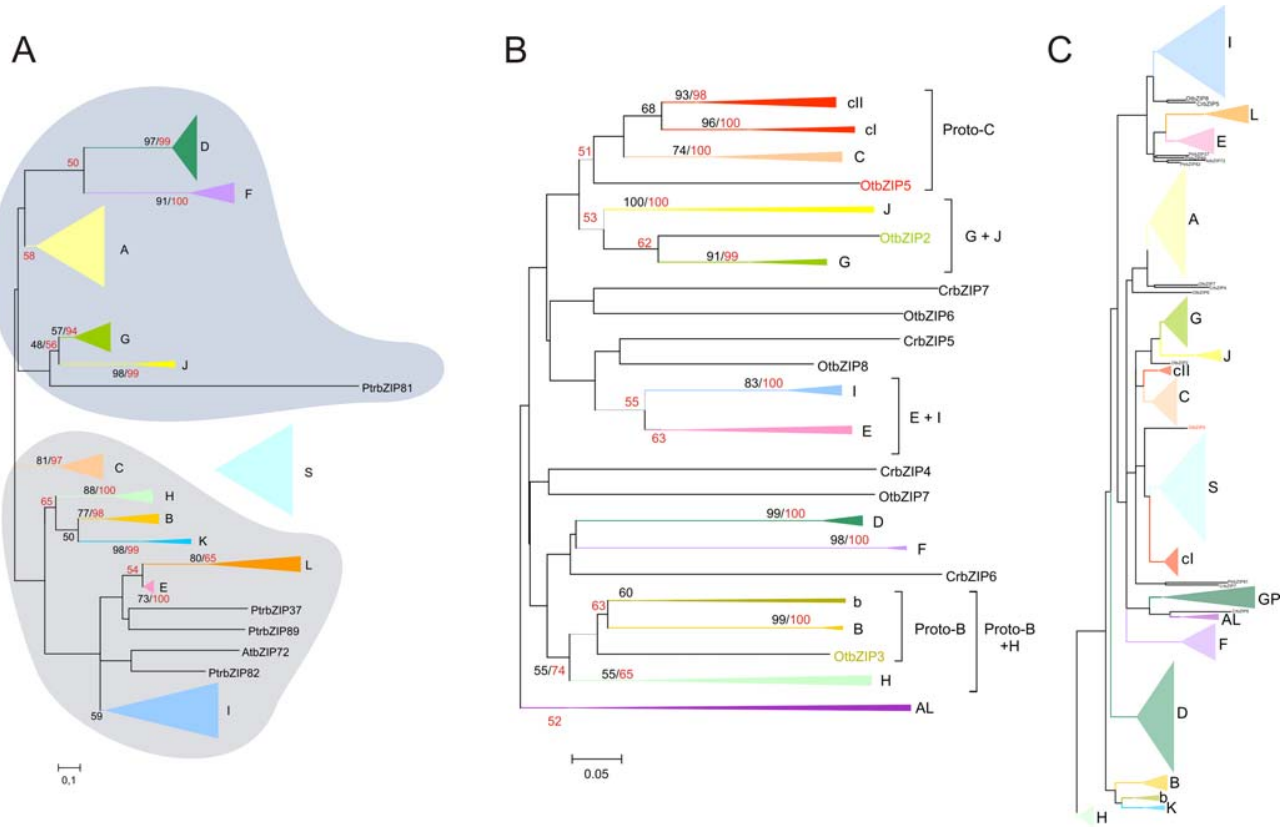


Figure 1. Phylogeny of bZIP transcription factors in green plants. (A) Model of angiosperm bZIP evolution with two large clades, one including groups A, D, F, G and J, and the other including groups B, C, E, H, I and L. Sister groups B and K, E and L, D and F, and G and J, respectively, were defined based on bootstrap support of >50%. The position of Group S could not be clearly defined. (B) Consensus tree inferred from NJ analyses of bryophyte and algal bZIP sequences. This tree reveals new evolutionary relationships among green plant bZIPs, which were not observed when the complete ViridiZIP set was analyzed. Group C appears to be related to two other groups (cI and cII) and members of these three groups are orthologues of *OtbZIP5*, constituting the Group Proto-C. Group b was identified as a sister group of Group B and genes of both groups are orthologous to the algal *OtbZIP3* gene, forming the Group Proto-B. Groups Proto-B and H have a common ancestral origin. Similarly, Groups G and J diverged from the same ancestor and are both orthologous to the algal gene *OtbZIP2*. Finally, Groups E and I show a sisterhood relation but no ancestral link to a bZIP from algae could be established. (C) Tree inferred from NJ analyses of the ViridiZIP data set (bZIPs from algae to angiosperms). This tree indicates that Group S probably originated from Proto-C, and Group K from Proto-B. Tree topology and functional data support these hypotheses. Bootstrap values were calculated from NJ analyses. Red, values obtained with p-distances and, black, with PAM matrix.
doi:10.1371/journal.pone.0002944.g001

motif. The only gene that possesses an intron in this motif is *AtbZIP24* from Arabidopsis.

Members of Groups A and D have a bZIP domain of only 44 amino acids. To refine our analysis we created a subset-of-bZARP (sbZARP) dataset that excluded groups A and D members but included all remaining 172 proteins with a bZIP domain of 60 amino acids (53, 60 and 59 bZIPs from Arabidopsis, rice and black cottonwood, respectively). NJ analyses revealed four new groups of homologues, Groups C, E, I and L, all supported by bootstrap values of >50% (Figure S3; note that Group L members harbor an atypical basic motif; see Figure S2, and Text S1c). The overall organization into twelve groups is further supported by the presence of at least one shared intron position among the members of each group, confirming a common ancestral origin of all its members (Figures 1A, 2 and S2). The twelve groups encompass 199 of the 257 bZIPs of the bZARP data set. Fifty-three of the remaining bZIPs (17, 17 and 19 from Arabidopsis, rice, and black cottonwood, respectively) tended to form a separate group, defined as Group S in agreement with previous data [55]. However, this group did not have significant bootstrap support. Members of Group S bZIPs share two characteristics: they harbor a long leucine zipper (eight to nine heptads) and are encoded by intron-less genes. Finally, *AtbZIP72* (Arabidopsis) and *PtrbZIP37*, *81*, *82* and *89* (black cottonwood) could not be classified into any of the above groups (Figure 1A).

In summary, our data suggest 13 groups of homologous angiosperm *bZIP* genes (A, B, C, D, E, F, G, H, I, J, K, L, and S), representing a unified classification of angiosperm bZIPs (Figure 3) [55,56,71]. This result is in agreement with previous analyses, but additionally revealed three new groups (J, K and L) (Figure S3). The name of each group of homologues follows the classification established by Jakoby *et al.* [55]. Similar conclusions were reached using Maximum Likelihood analyses.

Possible Groups of Orthologues (PoGOs) in Angiosperms

We next aimed at identifying Possible Groups of Orthologues (PoGOs) among the 13 groups of homologues. By definition, each PoGO represents a group of genes that diverged from an ancestral gene through speciation and duplication. Members of a given PoGO typically have closely related biological functions, and this allows making predictions for poorly characterized genes and rationalizes functional studies of the proteins they encode [72]. PoGOs also establish a basis for the definition of functional

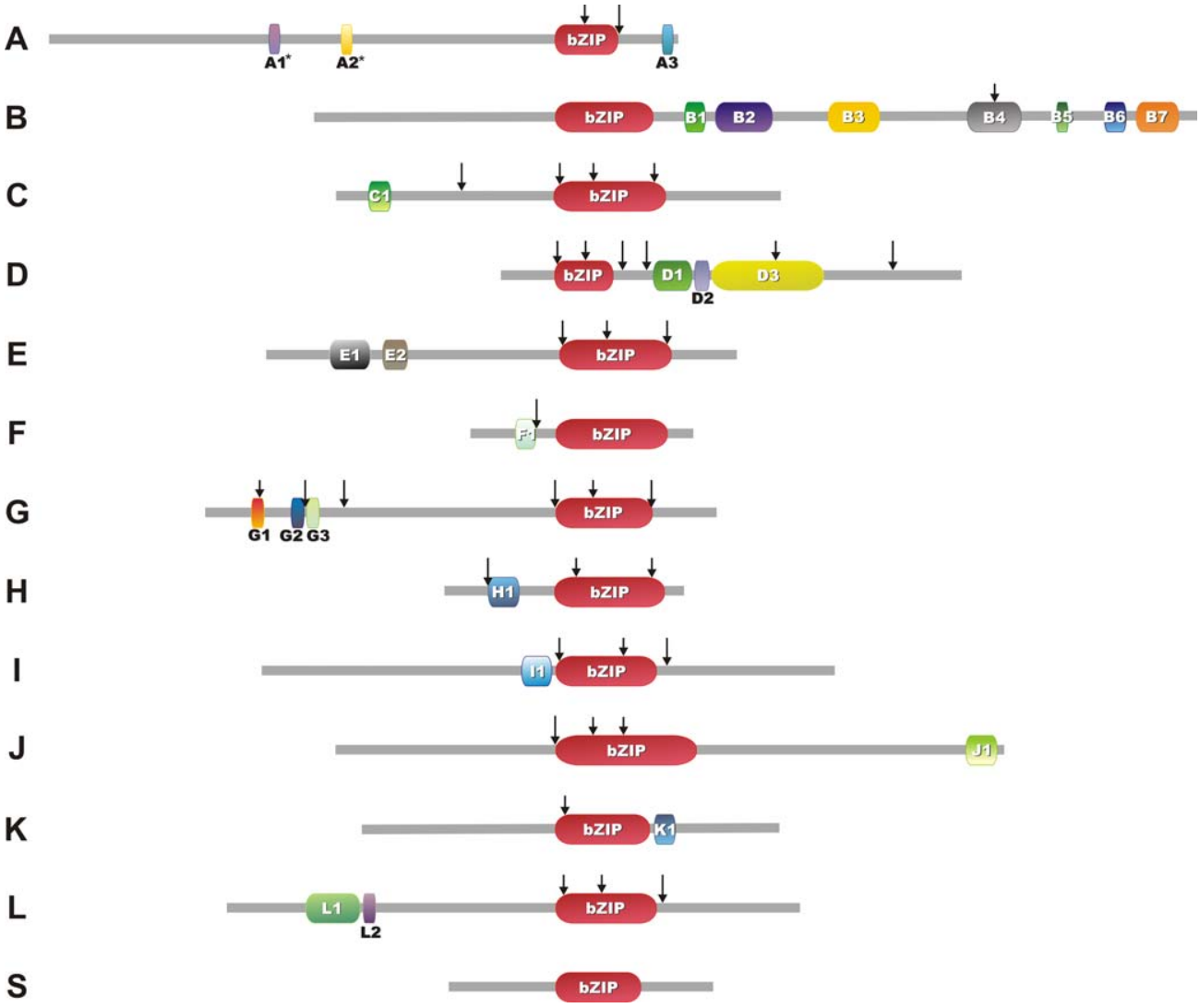


Figure 2. Motifs conserved in angiosperm bZIPs. A summary of the motif sequences is given in Table S2. Arrows indicate intron positions conserved among most members of each group. Representative bZIP sizes and positions of conserved motifs are shown. (*) Group A has two motifs (A1 and A2), that are important putative kinase phosphorylation sites involved in ABA responses. Both motifs appear to be conserved in most members of this group of homologues, except for OsbZIP8, 13, 14 and 15, and PtrbZIP5 and 10, which lack motif A1. The same sequences and also PtrbZIP9 lack motif A2. Due to the lack of complete sequences, no structures are shown for Groups AL, GP, b, cI and cII.
doi:10.1371/journal.pone.0002944.g002

diversification among genes. Here, we identified PoGOs by NJ analysis of each group of homologues separately, using the criteria defined in Material and Methods. To optimize the resolution of the evolutionary relationships, alignment lengths were extended by including conserved motifs specific to each group of homologues (Figure 2, and Table S2). Additionally, 636 further bZIP sequences, 260 from eudicots and 376 from monocots (Table S3), were extracted from EST databases. These new bZIPs were included in the respective groups of homologous genes according to their tblastn best matches against members of an upgraded Angiotot dataset that contained the rice and black cottonwood bZIPs.

Our analysis revealed 31 PoGOs distributed among Groups A to L (Figures 3 and S4, S5, S6, S7, S8, S9, S10, S11, S12, S13, S14 and S15). In all PoGOs except D2, at least one black cottonwood bZIP sequence could be included (Figure 3) further supporting the organization into PoGOs. The lack of a black cottonwood *bZIP* gene in PoGO D2 could be due to an absence of such a gene in its genome or to incomplete genome sequence availability. OsbZIP24, PtrbZIP86, 87 and 88 lack some of the motifs conserved in Group D members and were therefore assigned to the PoGO to which they showed the highest overall sequence similarity (as identified through blastp analysis).

We identified only one eudicot-monocot PoGO, S1, in Group S (Figure S16). The remaining sequences could be clustered into three PoGOs each restricted to either eudicots (SE1, SE2 and SE3) or monocots (SM1, SM2 and SM3) (Figure S16). Arabidopsis bZIP TFs of groups SE2 and SE3 are involved in energy metabolism and hypoosmolarity signaling (Table S4) further supporting the evolutionary relationship deduced from the phylogenetic analysis. Similarly, SM2 members play a role in cold signaling (Table S4), thus providing function-based support also for this group. Although further efforts to more precisely uncover the relationship between the three monocot (SM1, SM2 and SM3) and eudicot (SE1, SE2 and SE3) groups of orthologues

bZIP no.	Gene code	Synonym	GenBank	
OsbZIP1	Os03g20650			
OsbZIP2	Os07g48660		AK103188	
OsbZIP3	Os01g59760	DPBF3		
OsbZIP4	Os05g41070	AREB3os	AK063398	
AtbZIP12	At2g41070	DPBF4	AF334209	
AtbZIP66	At3g56850	AREB3	AB017162	
PtrbZIP8	754658			PoGO A1
PtrbZIP13	549022			
PtrbZIP14	558204			
PtrbZIP15	560286			
PtrbZIP16	808328			
PtrbZIP17	770717			
PtrbZIP18	803082			
PtrbZIP19	594286			
OsbZIP8	Os09g36910			
OsbZIP13	Os02g58670		AK061086	
OsbZIP14	Os06g50600		AK108991	
OsbZIP15	Os08g43600			
AtbZIP14	At4g35900	FD	BN000021	PoGO A2
AtbZIP15	At5g42910		AJ419599	
AtbZIP27	At2g17770	FDP	BN000022	
PtrbZIP5	550249			
PtrbZIP9	818828			
PtrbZIP10	642918			
OsbZIP5	Os01g64730	OSE2	AK067919	
OsbZIP6	Os05g36160	OSE2-like	AK120656	
AtbZIP13	At5g44080		BN000023	PoGO A3
AtbZIP40	At1g03970	GBF4	U01823	
PtrbZIP11	651568			
PtrbZIP12	754448			
OsbZIP7	Os01g64000	ABI5-2	AK070998	
AtbZIP39	At2g36270	ABI5	AF334206	
AtbZIP67	At3g44460	DPBF2	AJ419600	PoGO A4
PtrbZIP6	767006			
PtrbZIP7	801922			
OsbZIP9	Os09g28310	ABI5os	AK065873	
OsbZIP10	Os08g36790	TRAB1	NM001068553	
OsbZIP11	Os02g52780		AK072062	
OsbZIP12	Os06g10880		AK103188	
AtbZIP35	At1g49720	ABF1	AF093544	
AtbZIP36	At1g45249	ABF2	AF093545	PoGO A5
AtbZIP37	At4g34000	ABF3	AF093546	
AtbZIP38	At3g19290	ABF4	AF093547	
PtrbZIP1	551849			
PtrbZIP2	677861			
PtrbZIP3	267872			
PtrbZIP4	767577			
OsbZIP16	Os07g44950		AK121898	
OsbZIP17	Os05g34050		AK073142	
AtbZIP17	At2g40950		AV441374	PoGO B1
AtbZIP28	At3g10800		AJ419850	
AtbZIP49	At3g56660		AJ419851	
PtrbZIP20	255215			
OsbZIP22	Os03g58250	REB	AB021736	
OsbZIP23	Os07g08420	RISBZ1	AB053472	
AtbZIP63	At5g28770	BZO2H3		PoGO C1
PtrbZIP24	294737			
PtrbZIP25	729825			
OsbZIP18	Os12g40920	RBZO2H		
AtbZIP10	At4g02640	BZO2H1		
AtbZIP25	At3g54620	BZO2H4		PoGO C2
PtrbZIP22	551106			
PtrbZIP23	559630			
OsbZIP19	Os02g07840	RISBZ4	AB053473	
OsbZIP20	Os02g16680	RITA1	L34551	
OsbZIP21	Os06g45140	RISBZ5	AB053474	PoGO C3
AtbZIP9	At5g24800	BZO2H2	AF310223	
PtrbZIP21	271607			
OsbZIP24*	Os02g22280		AK103347	
OsbZIP25	Os09g10840			
OsbZIP26	Os09g31390		AK103174	
OsbZIP27	Os06g41100			PoGO D1
OsbZIP28	Os02g10140			
AtbZIP65	At5g06839		AJ314787	
PtrbZIP32	272608			
PtrbZIP86*	255651			
OsbZIP29	Os01g64020		AK101903	
OsbZIP30	Os05g37170		AK109520	
OsbZIP35	Os11g05480		AK102690	PoGO D2
OsbZIP36	Os12g05680	TGA-2.1	AK101620	
AtbZIP21	At1g08320		AJ314757	
OsbZIP32	Os08g07970	STGA	AK107028	
OsbZIP37	Os04g54474		AK066906	
AtbZIP22	At1g22070	TGA3	L10209	
AtbZIP47	At5g65210	TGA1	X68053	
AtbZIP50	At1g77920	TGA7	AJ315736	PoGO D3
AtbZIP57	At5g10030	OBF4	X69899	
PtrbZIP26	207609			
PtrbZIP27	217692			
PtrbZIP28	716556			
PtrbZIP29	830210			

bZIP no.	Gene code	Synonym	GenBank	
OsbZIP31	Os06g15480		AK109719	
AtbZIP46	At1g68640	PAN	AF111711	PoGO D4
PtrbZIP30	564507			
PtrbZIP88*	769678			
OsbZIP33	Os03g20310	NIF1	AB051294	
OsbZIP34	Os07g48820	NIF2	AB051295	
OsbZIP38	Os01g59350	NIF4	AB051297	
OsbZIP39	Os05g41280			
OsbZIP40	Os01g17260	NIF3	AB051296	
AtbZIP20	At5g06950	TGA2	D10042	PoGO D5
AtbZIP26	At5g06960	HBP-1b	X69900	
AtbZIP45	At3g12250	TGA6	AJ320540	
PtrbZIP31	712010			
PtrbZIP33	825048			
PtrbZIP87*	652586			
OsbZIP41	Os01g55150		AK108553	
OsbZIP42	Os01g11350	RF-2b like	AK100944	
OsbZIP43	Os02g14910			
AtbZIP34	At2g42380		AY074657	PoGO E1
AtbZIP61	At3g58120		AF401300	
PtrbZIP34	582775			
PtrbZIP35	836130			
PtrbZIP36	176347			
OsbZIP49	Os01g58760			
OsbZIP50	Os05g41540		AK104986	
OsbZIP51	OsIFCC032062		BX000502	
OsbZIP52	Os11g04390		AK103113	PoGO F1
AtbZIP24	At3g51960		AI994442	
PtrbZIP40	554977			
PtrbZIP41	812021			
OsbZIP48	Os06g50310		AK071639	
AtbZIP19	At4g35040		N65677	
AtbZIP23	At2g16770		AV544638	PoGO F2
PtrbZIP38	648793			
PtrbZIP39	649217			
OsbZIP53	Os01g46970	OSBZ8	U42208	
OsbZIP54	Os05g49420	Gbf	AK065440	
AtbZIP54	At4g01120	GBF2	AF053228	PoGO G1
AtbZIP55	At2g46270	GBF3	U51850	
PtrbZIP42	244814			
PtrbZIP43	411188			
OsbZIP56	Os03g13614	HBP-1a	AK066563	
AtbZIP41	At4g36730	GBF1	X63894	PoGO G2
PtrbZIP44	424322			
PtrbZIP45	719452			
OsbZIP57	Os02g03580		AK112009	PoGO G3
OsbZIP58	Os12g13170	osZIP-1a	U04295	
AtbZIP16	At2g35530		NM_179917	
AtbZIP68	At1g32150			
PtrbZIP46	757220			
PtrbZIP47	826637			
OsbZIP55	Os07g10890			PoGO G4
OsbZIP60	Os01g07880	THY5	BAB62558	
OsbZIP61	Os06g39960			PoGO H1
AtbZIP64	At3g17609	HY5-like	AF453477	
PtrbZIP50	657788			
OsbZIP59	Os02g10860			
AtbZIP56	At5g11260	HY5	AB005295	PoGO H2
PtrbZIP48	717128			
PtrbZIP49	809109			
OsbZIP67	Os11g06170		AY224425	
OsbZIP68	Os12g06520	RSG	AK065995	
AtbZIP51	At1g43700	VIP1	AF225983	PoGO I1
PtrbZIP53	204863			
PtrbZIP54	411874			
OsbZIP69	Os04g41820		AK064429	
OsbZIP70	Os09g34060	RF2a	AF005492	
AtbZIP59	At2g31370	PosF21	X61031	PoGO I2
AtbZIP69	At1g06070		AJ419854	
PtrbZIP55	718317			
PtrbZIP56	292756			
OsbZIP71	Os03g21800	RF2b	AY466471	
OsbZIP72	Os07g48180		AK102562	
AtbZIP18	At2g40620		AY0744269	
AtbZIP52	At1g06850		AAF63137	PoGO I3
PtrbZIP57	242954			
PtrbZIP58	739018			
PtrbZIP59	239991			
OsbZIP62	Os09g34880			
OsbZIP63	Os04g10260			
OsbZIP64	Os08g43090	vsf-1	AF467732	
OsbZIP65	Os03g03550			
OsbZIP66	Os10g38820		AK108607	PoGO I4
AtbZIP29	At4g38900		AF401297	
AtbZIP30	At2g21230		AF401298	
PtrbZIP51	556549			
PtrbZIP52	721835			
AtbZIP31	At2g13150		AF401301	
AtbZIP32	At2g12940	UNE4	AV566578	
AtbZIP33	At2g12900			PoGP I1
AtbZIP71	At2g24340			
AtbZIP74	At2g21235			

bZIP no.	Gene code	Synonym	GenBank	
OsbZIP73	Os11g05640	OsZIP-2a	U04296	
OsbZIP75	Os12g06010	OsZIP-2b	U04297	PoGO J1
AtbZIP62	At1g19490			
PtrbZIP60	826496			
OsbZIP74	Os06g41770		AK107021	
AtbZIP60	At1g42990		AY045964	PoGO K1
PtrbZIP61	818888			
OsbZIP44	OsIFCC014214			
OsbZIP45	Os02g33560		AK063644	
AtbZIP77	At1g35490		AY087319	PoGO L1
PtrbZIP83	820200			
PtrbZIP84	822688			
OsbZIP46	Os12g09270			
OsbZIP47	Os11g11100		AK072267	
AtbZIP76	At1g58110		BT015864	PoGO L2
AtbZIP78	at?		BT002467	
PtrbZIP85	584476			
OsbZIP88	Os08g38020		AK107150	
OsbZIP89	Os09g29820		AK108319	
OsbZIP90	Os02g49560			
OsbZIP91	Os06g42690			
OsbZIP92	Os02g09830			
AtbZIP3	At5g15830		AV549429	
AtbZIP8	At1g68880		AF400621	
AtbZIP42	AT3g30530		BAB01020	
AtbZIP43	At5g38800			PoGO S1
AtbZIP48	At2g04038		AC007178	
AtbZIP58	At1g13600		AF332430	
AtbZIP70	At5g60830			
AtbZIP75	At5g08141			
PtrbZIP72	251247			
PtrbZIP77	266015			
PtrbZIP78	590335			
PtrbZIP79	564461			
PtrbZIP80	566729			
AtbZIP4	At1g59530		AF400619	
AtbZIP5	At3g49760			
AtbZIP6	At2g22850			
AtbZIP7	At4g37730			SE1
PtrbZIP73	572012			
PtrbZIP74	754888			
PtrbZIP75	764916			
PtrbZIP76	774123			
AtbZIP2	At2g18160	GBF5	AF53939	
AtbZIP11	At4g34590	ATB2		
AtbZIP44	At1g75390		AV566155	
PtrbZIP62	710131			
PtrbZIP63	715285			SE2
PtrbZIP64	424048			
PtrbZIP65	719591			
PtrbZIP66	649375			
PtrbZIP67	818112			
AtbZIP1	At5g49450		AF400618	
AtbZIP53	At3g62420		AF400620	
PtrbZIP68	564400			SE3
PtrbZIP69	659068			
PtrbZIP70	245573			
PtrbZIP71	816720			
OsbZIP80	Os07g03220			
OsbZIP81	Os03g56010			SM1
OsbZIP82	Os12g43790			
OsbZIP83	Os03g47200			
OsbZIP84	Os01g36220		AK110526	
OsbZIP85	Os03g19370		AK109929	SM2
OsbZIP86	Os05g03860	LIP19	X57325	
OsbZIP87	Os12g37410	OBF1	AB185280	
OsbZIP76	Os08g26880		AK100580	
OsbZIP77	Os09g13570		AK064903	SM3
OsbZIP78	Os02g03960		AK070887	
OsbZIP79	OsIFCC038657			
AtbZIP72	At5g07160			
PtrbZIP37	767814			
PtrbZIP81	751080			
PtrbZIP82	767813			
PtrbZIP89	777882			

Figure 3. Classification of bZIPs from Arabidopsis, black cottonwood and rice. Thirteen groups of homologues (A to L, and S) were defined through NJ phylogenetic analyses with the bZARP set (Figures S1 and S3). The organization into Possible Groups of Orthologues (PoGOs) was done by more refined NJ phylogenetic analyses inside each group of homologues, including also sequences from other eudicots and monocots. The

alignment used for these analyses corresponds to a concatenated sequence of the group-specific conserved motifs identified employing MEME (http://meme.sdsc.edu/meme/website/intro.html; Figure 2). (*) Represents genes that lack group-wise conserved motifs, thus they were included inside a PoGO according to their best hit to another bZIP. Because the relation of AtbZIP72, PtrbZIP37, 81, 82 and 89 could not be clarified, they were not included in any of the groups of homologous or orthologous genes. One Possible Group of Paralogues (PoGP I1) was found in Arabidopsis. Column 'Gene code' provides the gene identifiers for Arabidopsis, black cottonwood and rice bZIP sequences taken from TAIR (http://www.arabidopsis.org/), JGI (http://www.jgi.doe.gov/) or TIGR (http://www.tigr.org/), respectively. 'Synonym' indicates published and often cites names of bZIP genes. The GenBank accession numbers of nucleotide sequences are given.
doi:10.1371/journal.pone.0002944.g003

proved unsuccessful, we propose that up to three additional eudicot-monocot PoGOs, besides S1, exist in Group S (as a minimal representation of the three possible monocot and eudicot PoGOs). The difficulty of organizing Group S bZIPs into PoGOs that comprise both eudicots and monocots sequences may reflect an increased evolutionary rate after their emergence. Rapid evolution can mainly be explained by relaxation of purifying selection or by positive selection. We used the Yang algorithm [73] to verify whether lineage-specific dN/dS ratios in Arabidopsis, black cottonwood and rice (the ω parameter, [74,75]) of Group S were different from that of all other groups. The ω value for Group S (0.12) was found to be significantly different from the average ω calculated for all other groups (0.03, likelihood ratio test $\chi^2_{df=1}$, $p<0.01$). Despite being under purifying selection ($\omega<1$), the value of ω for Group S is four times higher than the average. Thus it can be concluded that purifying selection is relaxed in this group, explaining the higher rate of sequence divergence among its members. Low selective constraint (i.e., low purifying selection) is a hallmark of more recently duplicated genes and can be correlated with functional diversification [76]. The extensive amplification of Group S members in angiosperms (see below) further supports the notion that functional diversification partly related to the control of energy metabolism is operating among Group S genes.

In Group G, we observed one PoGO that is restricted to monocots (PoGO G4; Figure S10). This may be explained by gene gain at an early phase of monocot radiation, or alternatively by gene loss in the ancestor of the eudicot lineage. Our analysis also revealed the existence of a Possible Group of Paralogues (PoGP) restricted to Group I in Arabidopsis (PoGP I1, Figure S12). This PoGP most probably reflects a recent duplication event followed by rapid divergence in the Arabidopsis lineage. As PoGO G4 and PoGP I1 are restricted to distinct evolutionary lineages, they probably do not play essential (common) roles in angiosperms as a whole. This conclusion is supported by the fact that EmBP from maize and wheat, both assigned to PoGO G4, control reserve protein (prolamin) production [77] which can be considered a monocot-specific function.

Gene duplication is an important means of evolutionary diversification. Therefore, PoGOs that preferentially expanded during angiosperm evolution are expected to include genes that were particularly important for establishing angiosperm-specific physiological or functional characteristics. Of the 13 groups of homologous genes, Groups A, D, E, I and S contain more genes per PoGO than the average (approximately six genes per PoGO, Figure S17), indicating their preferential contribution to the evolution of adaptive characteristics in angiosperms. Interestingly, Groups A, D and S include genes for responses and adaptation to environmental factors (abiotic and biotic stresses in Groups A/S and D, respectively; Table S4) and the control of energy use (Group S; Table S4). These observations raise the possibility that genes of these groups were particularly important for the colonization of new habitats and consequently for the radiation and expansion of angiosperms (Text S1d). Additionally, some PoGOs have a conserved one-to-one gene relationship, indicating that their genes may play a pivotal role during development (Text S1e)

In summary, we propose the existence of 31 monocot-eudicot PoGOs in Groups A to L, one monocot-specific PoGO (G5), one PoGP (I1) in Arabidopsis, and possibly three PoGOs in Group S. The 34 PoGOs are likely to be related to 34 possible ancestral functions of bZIPs in angiosperms (Figure 3, and Text S1d).

Tracing the Origin and Diversification of *bZIP* Genes in Green Plants

Based on the phylogenetic analyses and the *bZIP* gene structures from Arabidopsis, black cottonwood and rice, we propose a model for the evolution of angiosperm bZIPs (Figure 1A). This model proposes two large clades encompassing Groups A, D, F, G and J, and Groups B, C, E, H, I, K and L, respectively. Groups B, H and K, Groups E and L, and Groups D and F are sister groups, as evidenced by their bootstrap support. Furthermore, the conserved intron position in the bZIP domain shared by Groups A, D, G and J, as well as the one shared by Groups C, E, H, I, K and L (Figure S3) supports the hypothesis that these groups diverged from a common ancestor. We were not able to establish a clear relationship of Group S to any of the two larger groups. It may have an independent ancestral origin, constituting a third group, or may have evolved from one of the two large groups (Figure 1A).

To identify groups of homologues among the major eukaryotic lineages, i.e. animals, fungi, and plants, we performed a large-scale phylogenetic analysis using the conserved bZIP region of all bZIPs from *Homo sapiens* [78], *Caenorhabditis elegans* (http://www.wormbase.org/), *Drosophila melanogaster* [79], *Saccharomyces cerevisiae* (http://mips.gsf.de/genre/proj/yeast/), *A. thaliana* and *O. sativa*. This analysis revealed that *bZIPs* of each of these lineages share only one common ancestor (data not shown) which is in accordance with the fact that only a single *bZIP* sequence is present in the primitive eukaryote *Giardia lamblia* [80,81], perhaps representing the *bZIP* gene content prior to the plant/animal/fungal separation [80]. The function of this unique ancestral gene may be related to unfolded protein (UPR) and oxidative stress responses (see below). Deep evolutionary analyses have also been performed for the homeodomain and MADS-box families and it appears that their member TFs derived from at least two genes present in the last common ancestor of the three eukaryotic kingdoms [19,82]. It has been proposed that one of the ancestral functions of the MIKC^c class of MADS-box genes is an involvement in reproductive organ development [83,84]. Although this function appears to be conserved, it is still not clear whether it has a monophyletic origin.

We identified 7, 8, and 40 *bZIP* genes, respectively, in the genomes of the algae *Chlamydomonas reinhardtii* and *Ostreococcus tauri* and the moss *Physcomitrella patens* (however, a complete bZIP domain is missing in three of the moss proteins). Additionally, we identified bZIP sequences from assembled ESTs of species representing the most relevant divisions of the green plants from which sequences are available: four *bZIP* genes in the bryophyte *Marchantia polymorpha*, one each in the ferns *Selaginella moellendorffii* and *Adiantum capillus-veneris*, and 40 and nine, respectively, in the gymnosperms *Pinus taeda* and *Picea glauca* (Table S5). Although no complete genomic sequences were available for ferns or gymno-

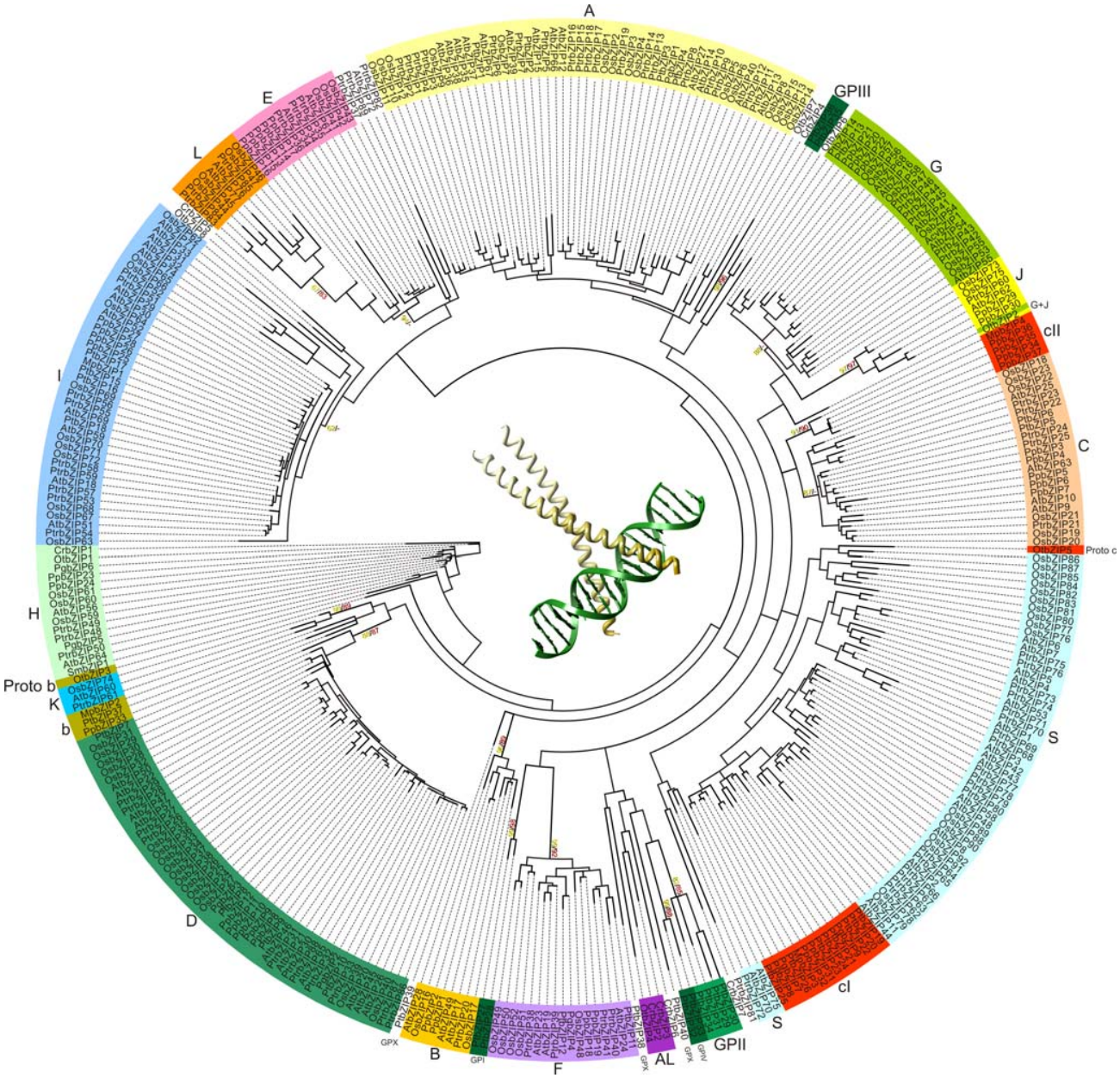


Figure 4. Global Phylogeny of bZIPs in green plants. This tree is a consensus of NJ analyses with p-distance performed with the ViridiZIP set. Bootstrap values in yellow were calculated from NJ analysis (PAM matrices, and with 44 and 60 amino acid alignments; only the highest bootstrap values are shown). Bootstrap values in red were calculated from ML analyses using the JTT+Γ evolutionary model (either with 44 or 60 amino acid alignments; only the highest bootstrap values are shown). GPX, GPI, GPII, GPIII, and GPIV indicate putative gymnosperm specific groups. Each group of homologues is colored following the same colour scheme used in Tables I and SV. The center of the tree depicts a typical bZIP dimer bound to DNA, representing the conserved bZIP domain (GCN4 from *Saccharomyces cerevisiae*; Protein Data Bank entry 2DGC).
doi:10.1371/journal.pone.0002944.g004

sperms, a considerable number of ESTs is available for the latter. We assembled a set of 345 bZIPs from algae to angiosperms (ViridiZIP set) for phylogenetic analyses (Figures 1B, 1C and 4).

Our study revealed that Group H is the most conserved group of *bZIP* homologues; members of this group are present in all green plant lineages. This observation is particularly interesting because Group H includes *HY5* and *HYH* that are important regulators of light responses and anthocyanin biosynthesis (Table S4). We therefore propose that Hy5-like bZIPs control light-dependent processes in all green plants. Similar to bZIPs in Group H, DOF transcription factors involved in light responses (subfamily A) also appear to be well conserved, suggesting that genes involved in light-related functions are under strong selective constraints [85]. In Arabidopsis Hy5-mediated photomorphogenesis is negatively regulated by the E3 ubiquitin ligase Cop1, which ubiquitylates Hy5 protein leading to its degradation [86]. We detected Cop1-related proteins in Physcomitrella, in agreement with previous results, as well as the Cop1-interaction motif in Physcomitrella Hy5-like bZIPs, suggesting that the genetic toolkit for photomorphogenesis described in angiosperms is also present

in mosses [87]. We also detected a single gene similar to *COP1* in Ostreococcus (ID 30007), but while in higher plants Cop1 protein contains a RING domain at the N-terminus, followed by multiple WD40 repetitions [88], this order is reversed in the Ostreococcus protein. Moreover, a Cop1 interaction site (Table S2) was not detected in the algal *HY5*-orthologues OtbZIP1 or CrbZIP1, or in any other green algae bZIP. Nevertheless, we found one Cop1-related protein in the red alga *Cyanidioschyzon merolae* (ID CMK039C; http://merolae.biol.s.u-tokyo.ac.jp/). Cop1-like proteins are also known in animals where they promote the degradation of the bZIP transcription factor c-Jun [88], suggesting Cop1-dependent protein degradation to be a regulatory scheme conserved in most eukaryotes.

Groups B, C, D, E, F, G, I and J were present in the most recent common ancestor (MRCA) of bryophytes and tracheophytes, indicating a functional connection to the colonization of the terrestrial environment (Figure 5). Some of these genes play a role in light responses (Group G), nitrogen/carbon balance control (Groups C and G), and ion responses (Group D), which are some of the important features that developed further in embryophytes (Table S4). Moreover, it appears that during the evolution from early land plants to angiosperms, Group D and I genes amplified

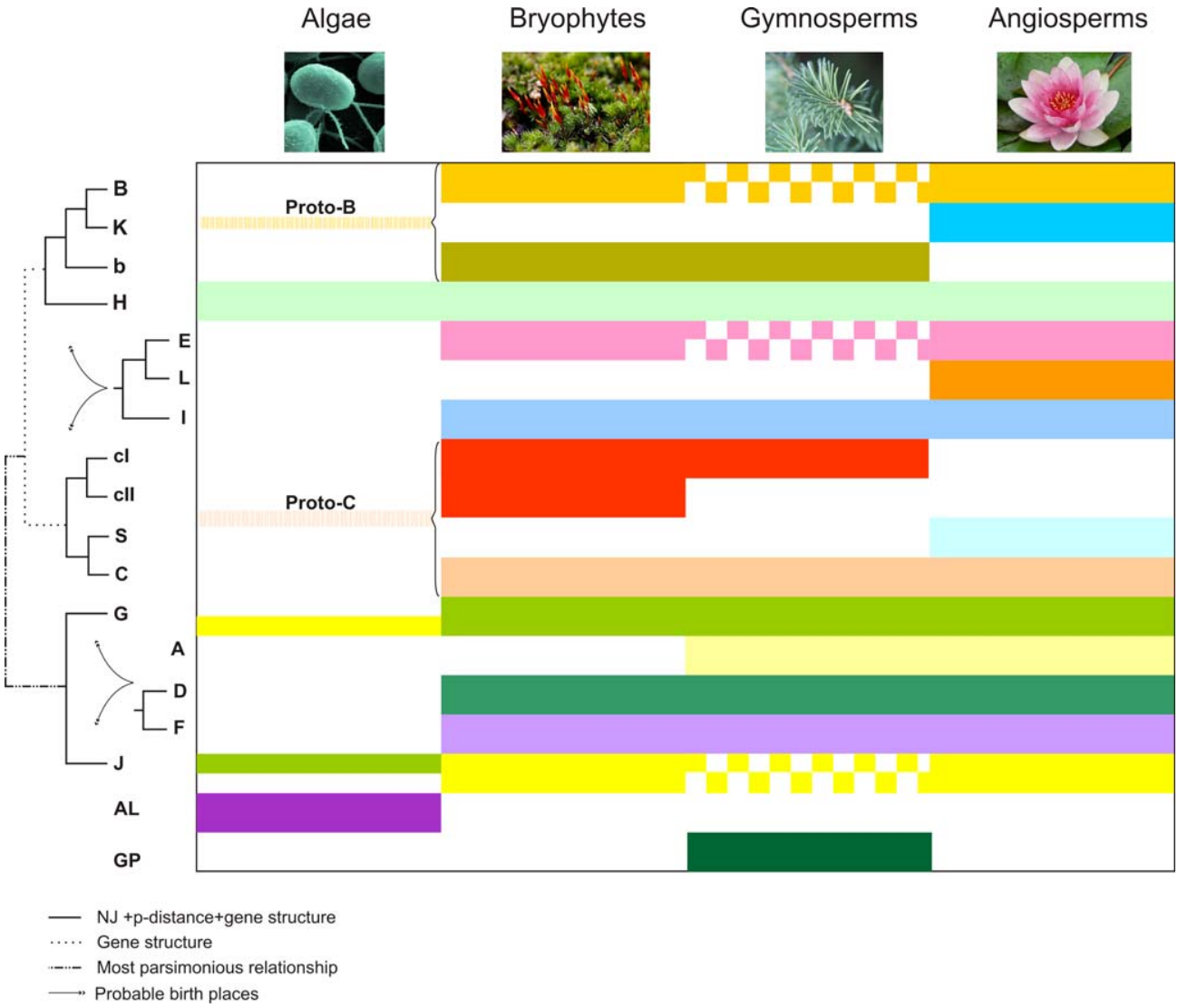


Figure 5. Phylogenetic profile and structure of bZIPs in green plants. Groups E, L and I belong to the same branch as Groups Proto-B, Proto-C and H but their exact position is not clear (Figure 1A). Similarly, Groups A, D and F do not have a clear position, though they belong to the same branch as Groups G and J (Figure 1A). The relation of Groups AL and GP to the other groups could not be established. bZIPs of the species studied here were grouped at the level of higher taxa, i.e., algae (represented by *C. reinhardtii* and *O. tauri*); bryophytes (*P. patens*); gymnosperms (*P. glauca* and *P. taeda*), and angiosperms (*O. sativa*, *A. thaliana* and *P. trichocarpa*). Solid boxes indicate that at least one bZIP was found for a given group of homologues in the respective taxon. Squared boxes indicate that homologous bZIP sequences were not yet observed in gymnosperms, possibly due to sampling limitations. Notably, however, sequences of the respective groups are conserved in bryophytes and angiosperms. Dashed lines with brackets shown in Groups Proto-B and Proto-C indicate that there is an orthologous bZIP in at least one of the algal species, although it does not strictly belong to any of the homologous groups. The half lines present in G and J indicate the presence of common orthologues in algae. Groups AL, GP, K, L and S appear to be lineage specific.
doi:10.1371/journal.pone.0002944.g005

more than genes of the other groups of homologues (5 to 10, and 4 to 11 genes in groups D and I, respectively), strongly suggesting that both groups were particularly important for this transition. Several Group D genes are involved in biotic stress responses (Table S4) indicating that improved pathogen defense was important for land plant evolution. Some *bZIP* genes of Group I control the expression of vascular genes (Table S4), which are central to vascular tissue development in tracheophytes.

Group A probably first appeared in the MRCA of spermatophytes and may thus be related to seed formation (Figure 5). As a matter of fact, Group A bZIPs often have functions in seed development, ABA responsiveness and fruit maturation (Table S4). Moreover, they are elements of ABA-dependent signaling pathways that coordinate responses to desiccation/dehydration and salt stress. ABA-mediated signaling is known in Physcomitrella [89,90], however, Group A bZIPs are not present in this organism (Figure 5), indicating a less developed ABA regulatory network (Text S1f).

According to our data Groups K, L and S are angiosperm-specific (Figure 5). However, due to sampling limitations we can not formally exclude the possibility that these groups are also present in gymnosperms. Additionally, this analysis eliminates the hypothesis that Group S has an independent ancestral origin (Figures 1A and 1C).

We also detected Group NA, a possible group of homologues exclusively present in non-angiosperm plants (Figure S18, and Text S1g). This finding is intriguing as genes conserved in mosses and gymnosperms are expected to represent general plant functions. Group NA bZIPs may thus have lineage-specific roles unimportant for angiosperms; the reduction of a dominant gametophyte during angiosperm evolution combined with a concomitant gene loss is an example for this. Alternatively, gene loss could have played a key role in the acquisition of important features in angiosperms, as seen for *KNOX* genes [91]; or, the roles played by bZIPs of Group NA could have been taken over by non-related but functionally analogous genes (non-orthologous gene displacement).

Ancestral Relationships in Groups B and C

The above analysis in combination with detailed sequential NJ analyses restricted to algal, moss and/or Arabidopsis sequences revealed two new groups, i.e. Groups Proto-B and Proto-C (Figure 1B). Group Proto-C encompasses Group C (Figure 1A) and two new Groups, cI and cII that correspond to the sequences previously identified in Group NA (Figure S18). While cI appears to be restricted to bryophytes, cII is found up to gymnosperms, and C is present up to angiosperms (Figures 1C and 5). Notably, in all phylogenetic analyses Group S appeared to be more attracted by Groups C, cI and cII (Figures 1C, 4 and 5), suggesting it originated from Group Proto-C, probably by gene duplication followed by rapid evolution. This finding is supported by the observation that bZIPs tend to dimerize with more similar partners, e.g. AtbZIP10 (Group C) with AtbZIP53 (Group S) [34,92]. Additionally, members of Group C (*AtbZIP63*) and S (*ATB2*, *GBF5*, *AtbZIP1* and *AtbZIP53*) participate in the control of energy metabolism and thus share similar functions (Table S4). Moreover, Group Proto-C possesses one *bZIP* gene, *OtbZIP5* from Ostreococcus, supporting the model that the biological functions played by bZIPs of Group C/S, such as oxidative stress responses associated with *AtbZIP10* [40] and energy metabolism control mediated for example by *GBF5* [41], are at least partially present in all green plants. Importantly, oxidative stress signaling involving bZIPs has been reported in yeast and men and thus appears to be conserved in all eukaryotes [93–97].

Group Proto-B consists of Group B, which includes members from bryophytes and angiosperms, a new group of homologues (Group b) that is apparently restricted to bryophytes and gymnosperms, and the Ostreococcus gene *OtbZIP3* (Figures 1B, 4 and 5). Based on our initial phylogenetic analysis of angiosperm sequences (Figure 1A) and tree topology (Figures 1C and 4) we concluded that angiosperm-specific Group K is not only a sister group of B, but very likely also emerged from Proto-B. Members of Group K are likely to have a role in the unfolded protein response (UPR), a cellular process involving the endoplasmic reticulum (ER) that counteracts cellular stress when incorrectly folded proteins accumulate [43]. bZIPs involved in this response are known in mammals and yeast and thus appear to be conserved in many lineages [98,99]. Recently, Liu et al. [42] demonstrated a role of Arabidopsis AtbZIP17 (Group B) in the UPR pathway, supporting the hypothesis that Group K emerged from Group B, and that *OtbZIP3* plays a similar role. Members of Groups B and K (like animal bZIP proteins involved in UPR) posses a trans-membrane domain for ER attachment (Table S2), but members of Group K lack the cleavage site recognized by the so-called site-1 protease (S1P). Most likely, the two groups function in different branches of the UPR pathway. Additionally, we looked for the presence of both trans-membrane and S1P interaction domains in other plant proteins. The trans-membrane domain is present in all Group B and K bZIPs from green plant lineages, whereas the S1P interaction domain was not found in some of them, perhaps due to missing sequence data.

Another important result of our analysis is that Ostreococcus sequences could be included, with significant bootstrap support, into Groups Proto-C (*OtbZIP5*) and Proto-B (*OtbZIP3*; Figure 1B). Moreover, Ostreococcus *OtbZIP2* was found to significantly cluster with Groups G and J, forming a new group named G+J (Figure 1B).

In conclusion, our results indicate that four Ostreococcus *bZIP* genes can be assigned to Groups Proto-C (*OtbZIP5*), Proto-B (*OtbZIP3*), G+J (*OtbZIP2*), and H (*OtbZIP1*), defining four orthologous relationships between algal and five groups of homologues from terrestrial plants (Figure 6). This data suggests the presence of at least four founder genes in the MRCA of green plants. Our analysis also indicates that Groups H (including *OtbZIP1* and *CrbZIP1*) and Proto-B (including *OtbZIP3*) originated from a common ancestral gene (Figure 1B). However, their relationship with Proto-C (*OtbZIP5*) and G+J (*OtbZIP2*), and the relationship of the four founder genes to the possible monophyletic origin of bZIPs in green plants could not be determined. The most parsimonious model that can explain the origin of the four ancestral bZIPs is shown in Figure 6. The assumption that Group Proto-C and Groups H/Proto-B share a common ancestral gene was inferred from the observation that angiosperm Groups C, B and H also cluster together (Figure 1A). Similarly, all DOF TFs appear to have originated from a single founder gene from subfamily A, which was present in the MRCA of green plants and might have played a role in light-regulated mechanisms [18]. In addition, MADS-box TYPE II (MIKC^{C}) and HD-Zip class III TF families each emerged from a single founder gene present in the MRCA of streptophytes that was possibly involved in haploid reproductive cell differentiation [84] or control of apical growth [23,24], respectively.

bZIP Evolution in Plants

Our data show that Group C and B members are elements of the oxidative stress signaling and UPR pathways, respectively, which appear to be crucial in all eukaryotes. This observation and the likely monophyletic origin of bZIPs of the main eukaryotic lineages (plants, animals, and fungi) suggest that the common bZIP ancestor was a multifunctional regulatory factor. An important

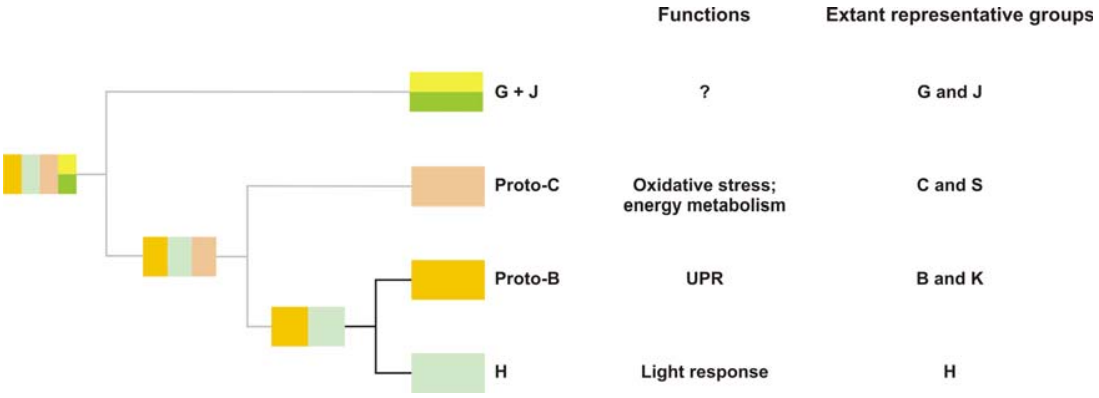


Figure 6. Most parsimonious model explaining the emergence of the four green plant founder *bZIP* genes. The four founder genes (in Groups G+J, Proto-C, Proto-B and H) are derived from a unique ancestral gene common to all eukaryotes. Groups Proto-B and Proto-C most likely derived from a multifunctional UPR/oxidative stress gene. Groups Proto-B and H are sister groups and their relationship to Group Proto-C was found by analyzing angiosperm bZIPs (Figure 1A). Group G+J is the ancestral group of a large set of *bZIP* genes included in Groups A, D and F, but the ancestral function played by this group is still largely unknown.
doi:10.1371/journal.pone.0002944.g006

consequence of this model is that Group H, which has a central role in light-mediated control, emerged from bZIPs of the oxidative stress and UPR regulatory modules. The integration of the branch leading to Group G+J, however, remains unclear which is partially due to the fact that functional information is limited and restricted to Group G that plays a role in light and ABA signaling.

From the extant algal sequences that do not cluster into any of the homologous groups of streptophytes, only a single group of homologues restricted to algae could be detected (Group AL; Figures 1C and 5). In most cases bZIP sequences from Chlamydomonas and Ostreococcus do not cluster together at all. This observation indicates that bZIPs evolved differently in the algal lineages, probably reflecting adaptations to different ecological niches; Chlamydomonas lives in fresh water, while Ostreococcus lives in sea water.

We estimated the number of bZIPs in the MRCA of all land plants (embryophytes), using the method of Hahn *et al.* [100]; the MRCA most likely had 64 bZIPs that expanded to 83 in the branch leading to seed plants. The rate of gene gain-loss, λ, in the seed plant lineage was found to be 2.01×10^{-3} per million years, which is similar to estimates for yeast (0.002) [100] and mammals (0.0016) [101]. We calculated expansions and contractions of the bZIP phylogenetic branches in the land plant lineage, using the estimated value for λ; this revealed a significant expansion ($p<0.05$) of the branch leading to the seed plant lineage. Finally, the evolution of the *bZIP* gene family is well explained by the random birth-and-death model in seed plants, i.e., no significant expansions/contractions occurred preferentially in any specific PoGO or group of homologues (Figure S19, and Text S1h).

Conclusions

In our analysis presented here we systematically classified bZIP TFs into PoGOs and considered existing knowledge about their biological functions to establish a robust methodology to reveal evolutionary relationships of this group of regulatory proteins. The moss Physcomitrella possesses almost five times more *bZIP* genes (37 genes, Table S5) than the alga Ostreococcus (8 genes), and half the number found in angiosperms (around 80 genes). Group A genes first appeared in the MRCA of spermatophytes and were recruited for seed development or germination but also to fine tune the responses to desiccation/dehydration and salt stress. Groups K, L and S are seemingly exclusive to angiosperms. Unexpectedly, Groups K and S control processes conserved in all eukaryotes, i.e. UPR and energy homeostasis. This apparent paradox can be explained by the fact that both, Groups K and S derived from the functionally related Groups Proto-B and Proto-C, respectively, that emerged early on during green plant evolution. Group S amplification likely contributed to refining the regulatory circuit controlling the organism's energy status. The most strongly conserved group of homologues in algae and angiosperms is Group H which includes light control factors *HY5* and *HYH*. Group H is representative of one of the four green plant founder *bZIP* genes. Our data thus establish the hypothesis that bZIP-controlled light responses of Group H emerged (through neofunctionalization) from a multifunctional ancestral gene of the UPR and oxidative stress response pathways (UPR/oxidative stress). The UPR/oxidative stress gene is also the ancestor of two other of the four founder genes, i.e. Groups Proto-B (UPR) and Proto-C (oxidative stress), which most likely diverged through subfunctionalization processes. The fourth founder gene, represented by Groups G and J, is the sister gene of the multifunctional UPR/oxidative stress gene. More functional data for Group G- and J-related bZIPs are required to further elaborate the model of green plant bZIP evolution.

Materials and Methods

Datasets of *bZIP* Genes

We generated a bZIP dataset (Angiotot) representing an updated version of the ABZ data set [56]. Plant bZIP sequences were identified as described by Riaño-Pachón et al. [102]. The whole proteomes deduced from the completely sequenced genomes of the algae *Ostreococcus tauri* [68] and *Chlamydomonas reinhardtii* [67], the bryophyte *Physcomitrella patens* [69], and the angiosperm *Populus trichocarpa* [59] were downloaded from the Joint Genome Institute/Department of Energy (JGI/DOE; http://www.jgi.doe.gov/). Protein sequences for the angiosperm *Arabidopsis thaliana* [54] were downloaded from The Arabidopsis Information Resource (TAIR; http://www.arabidopsis.org/), and from The J. Craig Venter Institute (http://www.tigr.org/) for the monocot *Oryza sativa* ssp. *japonica* [58].

Assembled ESTs from *Marchantia polymorpha*, *Physcomitrella patens*, *Adiantum capillus-veneris*, *Selaginella moellendorffii*, *Picea glauca*, *Pinus*

taeda, *Brassica napus*, *Glycine max*, *Heliathus annus*, *Medicago truncatula*, *Solanum lycopersicum*, *Solanum tuberosum*, *Hordeum vulgare*, *Saccharum ssp.*, *Sorghum bicolor*, *Triticum aestivum* and *Zea mays* were downloaded from the TIGR Plant Transcript Assemblies Database [103]. ESTs from *Oryza sativa* ssp. *indica* were downloaded from the Beijing Genomics Institute website (07.11.2006), and assembled into clusters using TGICL [104]. Additional rice bZIP sequences were obtained from the Full Length Rice cDNA Consortium [105]. Some sequences from completely sequenced genomes were re-annotated (Datasets S1 and S2), based on conserved protein motifs and gene structures of each family. The list of abbreviations of the organisms used is given in Table S6.

The tblastn program [106] was used to search for bZIP sequences in rice nucleotide databases (*Oryza sativa* ssp. *indica* [57]; Beijing Genomics Institute, http://btn.genomics.org.cn/rice, and *Oryza sativa* ssp. *japonica*; Syngenta, http://www.syngenta.com/; IRGSP, http://www.gramene.org/) using Angiotot as query. Sequences with an e-value $<10^{-4}$ were selected to form a subset (SeqZIP), from which false positive hits, corresponding mainly to low complexity regions, and hits that we initially identified using the above procedure were excluded. To identify the open reading frame and gene structure of each SeqZIP sequence, pairwise blastx analyses against their respective Angiotot best hits were performed. Gene structures were defined based on the alignments obtained, the conserved positions of introns in homologous bZIP genes, and the presence of canonical splicing sites (GT-AG). The protocol used for bZIP identification is described in Figure S20.

The procedure used to identify bZIPs in EST datasets was identical to that used for genomic sequences, except that the estwisedb program of the Wise2 package [107] was included to identify the most likely reading frames and its bZIP domains in a given cluster.

Phylogenetic Analyses

Alignment of bZIP protein sequences was performed by ClustalX [108], using default parameters, and subsequently adjusted manually. The alignments used for the analyses within each group of homologues represent a concatenated sequence of the different conserved motifs found within each group (Figure 2). The phylogenetic analyses based on amino acid sequences were conducted using MEGA v3.1 [109] and PHYLIP v3.6 [110]. Unrooted phylogenetic tree topologies were reconstructed by Neighbor-Joining (NJ), the distances were obtained using a PAM-like distance matrix [111], or alternatively, using p-distances [112], and the re-sampling of the original bZIP set was a 1,000 bootstrap repetition. Maximum Likelihood (ML) analyses of the bZIP domain (44 and 60 amino acids) were carried out using RAxML [113] with the distances computed using the JTT+Γ evolutionary model [114], and a re-sampling of the original bZIP set of 500 bootstrap repetitions. Bayesian approaches were not employed as they often lead to very liberal estimates of branch confidence that can result in wrong topologies [115]. Additionally, phylogenetic trees for nucleotide sequences, corresponding to the conserved motifs used for proteins, were inferred by means of the maximum likelihood method available in PAUP 4b10 [116]. The TrN+Γ [117] model of sequence evolution was used. Model choice was performed in MODELTEST 3.6 [118] by the likelihood ratio test with significance level set at 1%. ML trees are available upon request. Branch lengths of the tree comprising all species analyzed were estimated by Maximum Likelihood in TREE-PUZZLE v5.2 [119], using the consensus topology inferred by NJ analysis with PAM-like distances. All sequences and alignments used in this study are available upon request.

Identification of Conserved Motifs

The putative complete sets of unique bZIPs from Chlamydomonas, Ostreococcus, Physcomitrella, black cottonwood, Arabidopsis and rice served as input for a conserved motif analysis performed with MEME (http://meme.sdsc.edu/meme/meme.html) [120]. Whole protein sequences were employed for this search. A given motif was allowed to appear at any number of repetitions, the maximum width of a motif was set to 80, and the maximum number of motifs was set to 20. The other parameters were used as default. In a complementary approach, each group of homologues was analyzed individually with the parameters described above.

Phylogenetic Analyses and Identification of Possible Groups of Orthologoues (PoGOs)

The detailed evolutionary analysis of angiosperm bZIP sequence relationships within each group allowed the identification of PoGOs. A PoGO is defined by the following criteria: (i) members of a PoGO have a monophyletic origin, indicated by a bootstrap support greater than 50%; (ii) a PoGO possesses at least one representative gene each from *A. thaliana* and *O. sativa*, assuming that the putative complete sets of *bZIP* genes of these organisms were identified and no selective gene loss had occurred. In case a PoGO is found to be restricted to either monocots or eudicots, the presence of sequences from at least one other species of the same lineage in this PoGO is required; and (iii) the inferred phylogeny should be consistent with the known phylogeny of plant species [56].

Identification of Pseudogenes and Genomic Duplications

Search for pseudogenes in Chlamydomonas, Ostreococcus, black cottonwood, Arabidopsis and rice was performed by masking the genomic region for each identified bZIP. Blastx searches were performed against the masked sequences using the Angiotot bZIP database as query. A hit was considered as a pseudogene only if it possessed all or part of the bZIP domain; therefore all hits were compared against bZIP PFAM models [121] and manually cured, eliminating false positives. Genomic duplications in Arabidopsis were identified via "Paralogons in Arabidopsis thaliana" (http://wolfe.gen.tcd.ie/athal/dup) and 'MATDB: Segmental Duplications' from MIPS (Munich Information Center for Protein Sequences; http://www.mips.gsf.de/projects/plants) (Table S7).

Analysis of Gene Family Expansion and Contraction

The evolution of rates of *bZIP* gene gain and loss along the history of green plants was analyzed by the method of Hahn *et al.* [100], implemented in CAFÉ [122]. The method models gene family evolution as a stochastic birth-and-death process implemented as a probabilistic graphical model that allows for the inference of the most likely family sizes in the common ancestors of every branching point. In this way one can test the null hypothesis of random change in the family size. To avoid incomplete sampling, only plants with fully sequenced genomes were analyzed. The algorithm developed by Hahn *et al.* uses a birth-and-death parameter, λ, which was also estimated within CAFE. In addition to the parameter λ, CAFE needs divergence times to be entered along with the phylogeny of the organisms used. Since the inference of the size of gene families at deep evolutionary times is not reliable with any of the current methods available (Hahn, personal communication; [100]), we focused on land plants only. Tree topology and divergence times are shown in Figure S19. Significance of the contractions and expansions along branches was accessed by means of the three methodologies available in

CAFE: branch cutting, likelihood ratio test, and Viterbi assignments [122].

Gene Expression Analysis

Absolute signal intensity values from Arabidopsis ATH1_22K array (Affymetrix) was obtained through Meta-Analyzer from GENEVESTIGATOR (http://www.genevestigator.ethz.ch/) [123]. The developmental stages were as described by Boyes *et al.* [124]. Massively Parallel Signature Sequencing, MPSS, [125] was also verified for Arabidopsis and rice genes (Datasets S3 and S4).

Supporting Information

Figure S1 Definition of homologous gene groups A, D and F. This figure is a partial representation of the tree inferred from NJ analysis from the 258 non-redundant set of bZIPs from Arabidopsis, rice and black cottonwood using *p*-distance and 1000 bootstrap repetitions (indicated as percentages at the branch points). The alignment used corresponds to the minimum bZIP domain of 44 amino acids. Groups D and F are sister groups supported by a 50% bootstrap. Rice, black cottonwood and Arabidopsis sequences are represented in orange, dark blue and light blue, respectively.
Found at: doi:10.1371/journal.pone.0002944.s001 (1.01 MB TIF)

Figure S2 Conserved intron position in the basic motif region of angiosperm bZIP transcription factors. The first leucine of the leucine zipper is highlighted in green, and the conserved asparagine of the basic motif is shown in red. According to the position of the introns, indicated by arrows, four different groups can be observed (1 to 4). bZIPs from Group L have a basic motif five amino acids shorter than that of the other bZIPs, and the conserved asparagine, shown in red, is substituted either by lysine (K) or arginine (R). In bold, the first amino acid after the intron. The *bZIP* genes used in this figure are: *AtbZIP24* (Group F), *AtbZIP45* (Group D), *AtbZIP39* (Group A), *AtbZIP54* (Group G), *AtbZIP62* (Group J), *AtbZIP63* (Group C), *AtbZIP56* (Group H), *AtbZIP61* (Group E), *AtbZIP31* (Group I), *AtbZIP60* (Group K), *AtbZIP76* (Group L), *AtbZIP70* (Group S), and *AtbZIP49* (Group B).
Found at: doi:10.1371/journal.pone.0002944.s002 (1.85 MB TIF)

Figure S3 Unrooted phylogenetic tree inferred from a NJ analysis from a subset of 173 bZIPs of Arabidopsis, rice and black cottonwood using *p*-distance and 1000 bootstrap repetitions (indicated as percentages at the branches). The alignment used corresponds to the minimal bZIP domain extended by two leucine repetitions, totaling 60 amino acids. Groups B, K and H, as well as Groups E and L are sister groups supported by bootstrap analysis. Rice, black cottonwood and Arabidopsis sequences are represented in orange, dark blue and light blue, respectively.
Found at: doi:10.1371/journal.pone.0002944.s003 (1.11 MB TIF)

Figure S4 Phylogenetic tree of monocot and eudicot bZIPs of Group A. The unrooted tree was inferred by a NJ analysis from distances calculated with the PAM distance matrix. The bootstrap values correspond to 1000 repetitions and are indicated as percentage in every branch. The amino acid alignment used to generate this tree corresponds to the bZIP domain plus the conserved motif A1 (Figure 2 and Table S2). Rice, black cottonwood and Arabidopsis sequences are represented in orange, dark blue and light blue, respectively. Other eudicot sequences are shown in green. The organism from which the remaining monocot and eudicot bZIPs originated is indicated by the last two letters in each sequence. Abbreviations are explained in Table S6.
Found at: doi:10.1371/journal.pone.0002944.s004 (1.28 MB TIF)

Figure S5 Phylogenetic tree of Group B bZIPs from monocots and eudicots. An unrooted tree was inferred by a NJ analysis from distances obtained from the PAM distance matrix. The bootstrap values correspond to 1000 repetitions and are indicated as percentage in every branch. The amino acid alignment used to generate this tree corresponds to the bZIP domain plus the conserved motifs within this group (Figure 2 and Table S2). Rice, black cottonwood and Arabidopsis sequences are represented in orange, dark blue and light blue, respectively. Other monocot sequences are shown in red. The organism from which the remaining monocot and eudicot bZIPs originated is indicated by the last two letters in each sequence. Abbreviations are explained in Table S6.
Found at: doi:10.1371/journal.pone.0002944.s005 (0.31 MB TIF)

Figure S6 Phylogenetic tree of Group C bZIPs from monocots and eudicots. An unrooted tree was inferred by a NJ analysis from distances calculated with the PAM distance matrix. The bootstrap values correspond to 1000 repetitions and are indicated as percentage in every branch. The amino acid alignment used to generate this tree corresponds to the bZIP domain plus the conserved motif within this group (Figure 2 and Table S2). Rice, black cottonwood and Arabidopsis sequences are represented in orange, dark blue and light blue, respectively. Other eudicot and monocot sequences are show in green and red, respectively. The organism from which the remaining monocot and eudicot bZIPs originated is indicated by the last two letters in each sequence. Abbreviations are explained in Table S6.
Found at: doi:10.1371/journal.pone.0002944.s006 (2.03 MB TIF)

Figure S7 Phylogenetic tree of Group D bZIPs from monocots and eudicots. An unrooted tree was inferred by a NJ analysis from distances calculated with the PAM distance matrix. The bootstrap values correspond to 1000 repetitions and are indicated as percentage in every branch. The amino acid alignment used to generate this tree corresponds to the bZIP domain plus the conserved motifs within this group (Figure 2 and Table S2). Rice, black cottonwood and Arabidopsis sequences are represented in orange, dark blue and light blue, respectively. Other eudicot and monocot sequences are show in green and red, respectively. The organism from which the remaining monocot and eudicot bZIPs originated is indicated by the last two letters in each sequence. Abbreviations are explained in Table S6.
Found at: doi:10.1371/journal.pone.0002944.s007 (1.31 MB TIF)

Figure S8 Phylogenetic tree of Group E bZIPs from monocots and eudicots. An unrooted tree was inferred by a NJ analysis from distances calculated with the PAM distance matrix. The bootstrap values correspond to 1000 repetitions and are indicated as percentage in every branch. The amino acid alignment used to generate this tree corresponds to the bZIP domain plus the conserved motifs within this group (Figure 2 and Table S2). Rice, black cottonwood and Arabidopsis sequences are represented in orange, dark blue and light blue, respectively. The organism from which the remaining monocot and eudicot bZIPs originated is indicated by the last two letters in each sequence. Abbreviations are explained in Table S6.
Found at: doi:10.1371/journal.pone.0002944.s008 (0.31 MB TIF)

Figure S9 Phylogenetic tree of Group F bZIPs from monocots and eudicots. An unrooted tree was inferred by a NJ analysis from distances calculated with the PAM distance matrix. The bootstrap values correspond to 1000 repetitions and are indicated as percentage in every branch. The amino acid alignment used to generate this tree corresponds to the bZIP domain plus the conserved motif within this group (Figure 2 and Table S2). Rice,

black cottonwood and Arabidopsis sequences are represented in orange, dark blue and light blue, respectively. Other eudicot and monocot sequences are show in green and red, respectively. The organism from which the remaining monocot and eudicot bZIPs originated is indicated by the last two letters in each sequence. Abbreviations are explained in Table S6.
Found at: doi:10.1371/journal.pone.0002944.s009 (0.83 MB TIF)

Figure S10 Phylogenetic tree of Group G bZIPs from monocots and eudicots. An unrooted tree was inferred by a NJ analysis from distances calculated with the PAM distance matrix. The bootstrap values correspond to 1000 repetitions and are indicated as percentage in every branch. The amino acid alignment used to generate this tree corresponds to the bZIP domain plus the conserved motifs within this group (Figure 2 and Table S2). Rice, black cottonwood and Arabidopsis sequences are represented in orange, dark blue and light blue, respectively. Other eudicot and monocot sequences are show in green and red, respectively. The organism from which the remaining monocot and eudicot bZIPs originated is indicated by the last two letters in each sequence. Abbreviations are explained in Table S6.
Found at: doi:10.1371/journal.pone.0002944.s010 (1.03 MB TIF)

Figure S11 Phylogenetic tree of Group H bZIPs from monocots and eudicots. An unrooted tree was inferred by a NJ analysis from distances obtained from a PAM distance matrix. The bootstrap values correspond to 1000 repetitions and are indicated as percentage in every branch. The amino acid alignment used to generate this tree corresponds to the bZIP domain plus the conserved motif within this group (Figure 2 and Table S2). Rice, black cottonwood and Arabidopsis sequences are represented in orange, dark blue and light blue, respectively. Other eudicot and monocot sequences are show in green and red, respectively. The organism from which the remaining monocot and eudicot bZIPs originated is indicated by the last two letters in each sequence. Abbreviations are explained in Table S6.
Found at: doi:10.1371/journal.pone.0002944.s011 (0.85 MB TIF)

Figure S12 Phylogenetic tree of Group I bZIPs from monocots and eudicots. An unrooted tree was inferred by a NJ analysis from distances obtained from a PAM distance matrix. The bootstrap values correspond to 1000 repetitions and are indicated as percentage in every branch. The amino acid alignment used to generate this tree corresponds to the bZIP domain plus the conserved motifs within this group (Figure 2 and Table S2). Rice, black cottonwood and Arabidopsis sequences are represented in orange, dark blue and light blue, respectively. Other eudicot sequences are show in green. The organism from which the remaining monocot and eudicot bZIPs originated is indicated by the last two letters in each sequence. Abbreviations are explained in Table S6.
Found at: doi:10.1371/journal.pone.0002944.s012 (1.12 MB TIF)

Figure S13 Phylogenetic tree of Group J bZIPs from monocots and eudicots. An unrooted tree was inferred by a NJ analysis from distances obtained from a PAM distance matrix. The bootstrap values correspond to 1000 repetitions and are indicated as percentage in every branch. The amino acid alignment used to generate this tree corresponds to the bZIP domain plus the conserved motifs within this group (Figure 2 and Table S2). Rice, black cottonwood and Arabidopsis sequences are represented in orange, dark blue and light blue, respectively. The organism from which the remaining monocot and eudicot bZIPs originated is indicated by the last two letters in each sequence. Abbreviations are explained in Table S6.
Found at: doi:10.1371/journal.pone.0002944.s013 (0.14 MB TIF)

Figure S14 Phylogenetic tree of Group K bZIPs from monocots and eudicots. An unrooted tree was inferred by a NJ analysis from distances obtained from a PAM distance matrix. The bootstrap values correspond to 1000 repetitions and are indicated as percentage in every branch. The amino acid alignment used to generate this tree corresponds to the bZIP domain plus the conserved motif within this group (Figure 2 and Table S2). Rice, black cottonwood and Arabidopsis sequences are represented in orange, dark blue and light blue, respectively. Other eudicots and monocot sequences are show in green and red, respectively. The organism from which the remaining monocot and eudicot bZIPs originated is indicated by the last two letters in each sequence. Abbreviations are explained in Table S6.
Found at: doi:10.1371/journal.pone.0002944.s014 (0.82 MB TIF)

Figure S15 Phylogenetic tree of Group L bZIPs from monocots and eudicots. An unrooted tree was inferred by a NJ analysis from distances obtained from a PAM distance matrix. The bootstrap values correspond to 1000 repetitions and are indicated as percentage in every branch. The amino acid alignment used to generate this tree corresponds to the bZIP domain plus the conserved motifs within this group (Figure 2 and Table S2). Rice, black cottonwood and Arabidopsis sequences are represented in orange, dark blue and light blue, respectively. The organism from which the remaining monocot and eudicot bZIPs originated is indicated by the last two letters in each sequence. Abbreviations are explained in Table S6.
Found at: doi:10.1371/journal.pone.0002944.s015 (0.47 MB TIF)

Figure S16 Phylogenetic tree of Group S bZIPs from monocots and eudicots. An unrooted tree was inferred by a NJ analysis from distances obtained from a PAM distance matrix. The bootstrap values correspond to 1000 repetitions and are indicated as percentage in every branch. The amino acid alignment used to generate this tree corresponds to the bZIP domain. Rice, black cottonwood and Arabidopsis sequences are represented in orange, dark blue and light blue, respectively. Other eudicot and monocot sequences are show in green and red, respectively. The organism from which the remaining monocot and eudicot bZIPs originated is indicated by the last two letters in each sequence. Abbreviations are explained in Table S6.
Found at: doi:10.1371/journal.pone.0002944.s016 (2.04 MB TIF)

Figure S17 Gene amplification pattern in each angiosperm group of bZIP homologues.
Found at: doi:10.1371/journal.pone.0002944.s017 (0.77 MB TIF)

Figure S18 Identification of Groups cI and cII. Both trees are a partial representation of the whole tree obtained by NJ analyses. (A) In the initial phylogenetic analysis with the complete ViridiZIP set, we were able to identify two clusters of genes that did not posses any member from angiosperms; therefore, we called them NA (non-angiosperm). (B) Restricted analyses including bZIPs from algae and mosses uncovered the relationship of Groups NA and C; both groups share the same homologue in Ostreococcus (*OtbZIP5*), indicating it to be a common ancestor. Group NA was re-classified into Groups cI and cII. Their relation to members of Group NA shown in (A) is indicated by stars (* for Group cII, or ** for Group cI). Groups cI, cII, C and OtbZIP5 form the Group Proto-C. The bootstrap support of each group is shown in the figure.
Found at: doi:10.1371/journal.pone.0002944.s018 (2.44 MB TIF)

Figure S19 Evolution of the bZIP family of transcription factors in land plants. We estimated the birth-and-death parameter (λ) using CAFE, as described in Materials and Methods. (A) The examined values of λ ranged from 1.0×10^{-4} to 6.8×10^{-3}. The log probabilities obtained for each assayed value are shown. The

shadowed region is displayed at a higher scale in the inset, where a peak at $\lambda = 0.002011$ is observed. (B) Evolutionary relationships of land plants with divergence time points (Arabidopsis - black cottonwood, 100–120 million years ago (mya) (47); monocot - eudicot, 140–150 mya (57); Physcomitrella - angiosperms, 450 mya (58)). Numbers at the branch end points indicate the numbers of bZIPs observed in the extant species. Numbers at the nodes represent the expected number of bZIPs in the ancestral species. Using the three methods available in CAFE, i.e., Viterbi assignments, branch cutting and likelihood ratio test, we identified branches deviating from the background model. According to all three methods, the branch leading to angiosperms significantly deviates from the null model ($p<0.05$), which implies that there was a significant increase in the number of bZIPs in the lineage leading to that group. Similarly, the Viterbi and branch cutting methods identify the branch leading to bryophytes (Physcomitrella) exhibiting a significant reduction in the number of bZIPs ($p<0.05$). Finally, we did not observe any significant deviation of the model for the extant group of angiosperms which can be interpreted as an even diffusion of the number of bZIPs in each branch. However, one cannot exclude the effect of natural selection in accounting for the differences that are nevertheless occurring. The increased number of bZIPs in the branch leading to angiosperms might be, at least partly, related to the several genome-wide duplication events that took place in the history of that lineage.
Found at: doi:10.1371/journal.pone.0002944.s019 (1.62 MB TIF)

Figure S20 Scheme of the pipeline for bZIP identification in genomic sequences and ESTs. (I) Input genomic and EST sequences are compared by tblastn with the Angiotot protein dataset, generating a group of sequences that putatively code for bZIPs (SeqZIP). (II) Manual curation allowed subtracting sequences already present in Angiotot (redundancies) and false positives, which mainly correspond to low-complexity sequences. (III) The remaining sequences (true positives) are compared by tblastx against the best hit from Angiotot obtained in step I, allowing to identify the most probable ORF, and in the case of genomic sequences, to identify their gene structure, taking into account conserved intron positions and the presence of canonic splicing sites (GT-AG).
Found at: doi:10.1371/journal.pone.0002944.s020 (0.75 MB TIF)

Table S1 Comparison between bZIPs reported in this manuscript and in Nijhawan et al. (2008)
Found at: doi:10.1371/journal.pone.0002944.s021 (0.04 MB XLS)

Table S2 Conserved motifs in bZIP PoGOs.
Found at: doi:10.1371/journal.pone.0002944.s022 (0.01 MB PDF)

Table S3 Accession numbers and classification into groups of homologues of non-sequenced angiosperms.
Found at: doi:10.1371/journal.pone.0002944.s023 (0.03 MB PDF)

Table S4 Biological functions of genes in PoGOs.
Found at: doi:10.1371/journal.pone.0002944.s024 (0.02 MB PDF)

Table S5 Classification of non-angiosperm bZIPs.
Found at: doi:10.1371/journal.pone.0002944.s025 (0.02 MB XLS)

Table S6 Organism abbreviations.
Found at: doi:10.1371/journal.pone.0002944.s026 (0.03 MB XLS)

Table S7 Gene pairs resulting from segmental duplications of the Arabidopsis genome.
Found at: doi:10.1371/journal.pone.0002944.s027 (0.03 MB DOC)

Dataset S1 Re-annotated nucleotide sequences from rice and black cottonwood.
Found at: doi:10.1371/journal.pone.0002944.s028 (0.02 MB TXT)

Dataset S2 Re-annotated amino acid sequences from rice and black cottonwood.
Found at: doi:10.1371/journal.pone.0002944.s029 (0.01 MB TXT)

Dataset S3 MPSS Expression data for bZIP genes from rice.
Found at: doi:10.1371/journal.pone.0002944.s030 (0.02 MB PDF)

Dataset S4 MPSS Expression data for bZIP genes from Arabidopsis.
Found at: doi:10.1371/journal.pone.0002944.s031 (0.01 MB PDF)

Text S1 Supporting texts including further results and discussion.
Found at: doi:10.1371/journal.pone.0002944.s032 (0.06 MB DOC)

Acknowledgments

We thank Amanda Bortolini Silveira (Universidade Estadual de Campinas, Brazil) for nuclear localisation experiments on Group L bZIPs, and Stefanie Hartmann (University of Potsdam) for critical comments on our manuscript, Liam Childs (MPI of Molecular Plant Physiology, Potsdam) for improving our English and the two reviewers for their helpful comments on the manuscript.

Author Contributions

Conceived and designed the experiments: LGGC CGS RVRVdS MV. Performed the experiments: LGGC DMRP RVRVdS. Analyzed the data: LGGC DMRP CGS MV. Contributed reagents/materials/analysis tools: BMR. Wrote the paper: LGGC DMRP BMR MV.

References

1. Meshi T, Iwabuchi M (1995) Plant transcription factors. Plant Cell Physiol 36: 1405–1420.
2. Beckett D (2001) Regulated assembly of transcription factors and control of transcription initiation. J Mol Biol 314: 335–352.
3. Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, et al. (2003) The evolution of transcriptional regulation in eukaryotes. Mol Biol Evol 20: 1377–1419.
4. Warren AJ (2002) Eukaryotic transcription factors. Curr Opin Struct Biol 12: 107–114.
5. Wingender E, Chen X, Fricke E, Geffers R, Hehl R, et al. (2001) The TRANSFAC system on gene expression regulation. Nucleic Acids Res 29: 281–283.
6. Riechmann JL, Ratcliffe OJ (2000) A genomic perspective on plant transcription factors. Curr Opin Plant Biol 3: 423–434.
7. Hsia CC, McGinnis W (2003) Evolution of transcription factor function. Curr Opin Genet Dev 13: 199–206.
8. Chen K, Rajewsky N (2007) The evolution of gene regulation by transcription factors and microRNAs. Nat Rev Genet 8: 93–103.
9. Lawton-Rauh A (2003) Evolutionary dynamics of duplicated genes in plants. Mol Phylogenet Evol 29: 396–409.
10. Levine M, Tjian R (2003) Transcription regulation and animal diversity. Nature 424: 147–151.
11. 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.
12. Riaño-Pachón DM, Corrêa LGG, Trejos-Espinosa R, Mueller-Roeber B (2008) Green transcription factors: a chlamydomonas overview. Genetics 179: 31–39.
13. Irish VF (2003) The evolution of floral homeotic gene function. Bioessays 25: 637–646.
14. Garcia-Fernandez J (2005) The genesis and evolution of homeobox gene clusters. Nat Rev Genet 6: 881–892.
15. Deppmann CD, Acharya A, Rishi V, Wobbes B, Smeekens S, et al. (2004) Dimerization specificity of all 67 B-ZIP motifs in *Arabidopsis thaliana*: a comparison to Homo sapiens B-ZIP motifs. Nucleic Acids Res 32: 3435–3445.
16. Floyd SK, Bowman JL (2007) The ancestral developmental tool kit of land plants. Int J Plant Sci 168: 1–35.

17. Bowman JL, Floyd SK, Sakakibara K (2007) Green genes-comparative genomics of the green branch of life. Cell 129: 229–234.
18. Moreno-Risueno MA, Martinez M, Vicente-Carbajosa J, Carbonero P (2007) The family of DOF transcription factors: from green unicellular algae to vascular plants. Mol Genet Genomics 277: 379–390.
19. Derelle R, Lopez P, Le GH, Manuel M (2007) Homeodomain proteins belong to the ancestral molecular toolkit of eukaryotes. Evol Dev 9: 212–219.
20. Martinez-Castilla LP, Alvarez-Buylla ER (2003) Adaptive evolution in the Arabidopsis MADS-box gene family inferred from its complete resolved phylogeny. Proc Natl Acad Sci U S A 100: 13407–13412.
21. Parenicova L, de Folter S, Kieffer M, Horner DS, Favalli C, et al. (2003) Molecular and phylogenetic analyses of the complete MADS-box transcription factor family in Arabidopsis: new openings to the MADS world. Plant Cell 15: 1538–1551.
22. Zhang Y, Wang L (2005) The WRKY transcription factor superfamily: its origin in eukaryotes and expansion in plants. BMC Evol Biol 5: 1.
23. Prigge MJ, Clark SE (2006) Evolution of the class III HD-Zip gene family in land plants. Evol Dev 8: 350–361.
24. Floyd SK, Zalewski CS, Bowman JL (2006) Evolution of class III homeodomain-leucine zipper genes in streptophytes. Genetics 173: 373–388.
25. Hurst HC (1995) Transcription factors 1: bZIP proteins. Protein Profile 2: 101–168.
26. Walsh J, Waters CA, Freeling M (1998) The maize gene liguleless2 encodes a basic leucine zipper protein involved in the establishment of the leaf blade-sheath boundary. Genes Dev 12: 208–218.
27. Chuang CF, Running MP, Williams RW, Meyerowitz EM (1999) The *PERIANTHIA* gene encodes a bZIP protein involved in the determination of floral organ number in *Arabidopsis thaliana*. Genes Dev 13: 334–344.
28. Abe M, Kobayashi Y, Yamamoto S, Daimon Y, Yamaguchi A, et al. (2005) FD, a bZIP protein mediating signals from the floral pathway integrator FT at the shoot apex. Science 309: 1052–1056.
29. Silveira AB, Gauer L, Tomaz JP, Cardoso PR, Carmello-Guerreiro S, et al. (2007) The Arabidopsis AtbZIP9 protein fused to the VP16 transcriptional activation domain alters leaf and vascular development. Plant Sci 172: 1148–1156.
30. Shen H, Cao K, Wang X (2007) A conserved proline residue in the leucine zipper region of AtbZIP34 and AtbZIP61 in *Arabidopsis thaliana* interferes with the formation of homodimer. Biochem Biophys Res Commun 362: 425–430.
31. Yin Y, Zhu Q, Dai S, Lamb C, Beachy RN (1997) RF2a, a bZIP transcriptional activator of the phloem-specific rice tungro bacilliform virus promoter, functions in vascular development. EMBO J 16: 5247–5259.
32. Fukazawa J, Sakai T, Ishida S, Yamaguchi I, Kamiya Y, et al. (2000) Repression of shoot growth, a bZIP transcriptional activator, regulates cell elongation by controlling the level of gibberellins. Plant Cell 12: 901–915.
33. Ciceri P, Locatelli F, Genga A, Viotti A, Schmidt RJ (1999) The activity of the maize Opaque2 transcriptional activator is regulated diurnally. Plant Physiol 121: 1321–1328.
34. Weltmeier F, Ehlert A, Mayer CS, Dietrich K, Wang X, et al. (2006) Combinatorial control of Arabidopsis proline dehydrogenase transcription by specific heterodimerisation of bZIP transcription factors. EMBO J 25: 3133–3143.
35. Zhang B, Foley RC, Singh KB (1993) Isolation and characterization of two related Arabidopsis ocs-element bZIP binding proteins. Plant J 4: 711–716.
36. Despres C, DeLong C, Glaze S, Liu E, Fobert PR (2000) The Arabidopsis NPR1/NIM1 protein enhances the DNA binding activity of a subgroup of the TGA family of bZIP transcription factors. Plant Cell 12: 279–290.
37. Pontier D, Miao ZH, Lam E (2001) Trans-dominant suppression of plant TGA factors reveals their negative and positive roles in plant defense responses. Plant J 27: 529–538.
38. Niggeweg R, Thurow C, Kegler C, Gatz C (2000) Tobacco transcription factor TGA2.2 is the main component of as-1-binding factor ASF-1 and is involved in salicylic acid- and auxin-inducible expression of as-1-containing target promoters. J Biol Chem 275: 19897–19905.
39. Thurow C, Schiermeyer A, Krawczyk S, Butterbrodt T, Nickolov K, et al. (2005) Tobacco bZIP transcription factor TGA2.2 and related factor TGA2.1 have distinct roles in plant defense responses and plant development. Plant J 44: 100–113.
40. Kaminaka H, Nake C, Epple P, Dittgen J, Schutze K, et al. (2006) bZIP10-LSD1 antagonism modulates basal defense and cell death in Arabidopsis following infection. EMBO J 25: 4400–4411.
41. Baena-Gonzalez E, Rolland F, Thevelein JM, Sheen J (2007) A central integrator of transcription networks in plant stress and energy signalling. Nature 448: 938–943.
42. Liu JX, Srivastava R, Che P, Howell SH (2007) Salt stress responses in Arabidopsis utilize a signal transduction pathway related to endoplasmic reticulum stress signaling. Plant J 51: 897–909.
43. Iwata Y, Koizumi N (2005) An Arabidopsis transcription factor, AtbZIP60, regulates the endoplasmic reticulum stress response in a manner unique to plants. Proc Natl Acad Sci U S A 102: 5280–5285.
44. Finkelstein RR, Lynch TJ (2000) Abscisic acid inhibition of radicle emergence but not seedling growth is suppressed by sugars. Plant Physiol 122: 1179–1186.
45. Uno Y, Furihata T, Abe H, Yoshida R, Shinozaki K, et al. (2000) Arabidopsis basic leucine zipper transcription factors involved in an abscisic acid-dependent signal transduction pathway under drought and high-salinity conditions. Proc Natl Acad Sci U S A 97: 11632–11637.
46. Niggeweg R, Thurow C, Weigel R, Pfitzner U, Gatz C (2000) Tobacco TGA factors differ with respect to interaction with NPR1, activation potential and DNA-binding properties. Plant Mol Biol 42: 775–788.
47. Nieva C, Busk PK, Dominguez-Puigjaner E, Lumbreras V, Testillano PS, et al. (2005) Isolation and functional characterisation of two new bZIP maize regulators of the ABA responsive gene rab28. Plant Mol Biol 58: 899–914.
48. Wellmer F, Kircher S, Rugner A, Frohnmeyer H, Schafer E, et al. (1999) Phosphorylation of the parsley bZIP transcription factor CPRF2 is regulated by light. J Biol Chem 274: 29476–29482.
49. Osterlund MT, Hardtke CS, Wei N, Deng XW (2000) Targeted destabilization of HY5 during light-regulated development of Arabidopsis. Nature 405: 462–466.
50. Ulm R, Baumann A, Oravecz A, Mate Z, Adam E, et al. (2004) Genome-wide analysis of gene expression reveals function of the bZIP transcription factor HY5 in the UV-B response of Arabidopsis. Proc Natl Acad Sci U S A 101: 1397–1402.
51. Satoh R, Fujita Y, Nakashima K, Shinozaki K, Yamaguchi-Shinozaki K (2004) A novel subgroup of bZIP proteins functions as transcriptional activators in hypoosmolarity-responsive expression of the *ProDH* gene in Arabidopsis. Plant Cell Physiol 45: 309–317.
52. Lara P, Onate-Sanchez L, Abraham Z, Ferrandiz C, Diaz I, et al. (2003) Synergistic activation of seed storage protein gene expression in Arabidopsis by ABI3 and two bZIPs related to OPAQUE2. J Biol Chem 278: 21003–21011.
53. Vettore AL, Yunes JA, Cord NG, da Silva MJ, Arruda P, et al. (1998) The molecular and functional characterization of an Opaque2 homologue gene from Coix and a new classification of plant bZIP proteins. Plant Mol Biol 36: 249–263.
54. Arabidopsis Genome Iniciative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408: 796–815.
55. Jakoby M, Weisshaar B, Droge-Laser W, Vicente-Carbajosa J, Tiedemann J, et al. (2002) bZIP transcription factors in Arabidopsis. Trends Plant Sci 7: 106–111.
56. Vincentz M, Bandeira-Kobarg C, Gauer L, Schlogl P, Leite A (2003) Evolutionary pattern of angiosperm bZIP factors homologous to the maize Opaque2 regulatory protein. J Mol Evol 56: 105–116.
57. Yu J, Hu S, Wang J, Wong GK, Li S, et al. (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). Science 296: 79–92.
58. Goff SA, Ricke D, Lan TH, Presting G, Wang R, et al. (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). Science 296: 92–100.
59. Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, et al. (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science 313: 1596–1604.
60. Bennetzen J (2002) The rice genome. Opening the door to comparative plant biology. Science 296: 60–63.
61. Pennacchio LA (2003) Insights from human/mouse genome comparisons. Mamm Genome 14: 429–436.
62. Vincentz M, Cara FA, Okura VK, da Silva FR, Pedrosa GL, et al. (2004) Evaluation of monocot and eudicot divergence using the sugarcane transcriptome. Plant Physiol 134: 951–959.
63. Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. Science 278: 631–637.
64. Adams KL (2007) Evolution of duplicate gene expression in polyploid and hybrid plants. J Hered 98: 136–141.
65. Rijpkema AS, Gerats T, Vandenbussche M (2007) Evolutionary complexity of MADS complexes. Curr Opin Plant Biol 10: 32–38.
66. Woolfe A, Elgar G (2007) Comparative genomics using Fugu reveals insights into regulatory subfunctionalization. Genome Biol 8: R53.
67. Merchant SS, Prochnik SE, Vallon O, Harris EH, Karpowicz SJ, et al. (2007) The Chlamydomonas genome reveals the evolution of key animal and plant functions. Science 318: 245–250.
68. Derelle E, Ferraz C, Rombauts S, Rouze P, Worden AZ, et al. (2006) Genome analysis of the smallest free-living eukaryote *Ostreococcus tauri* unveils many unique features. Proc Natl Acad Sci U S A 103: 11647–11652.
69. Rensing SA, Lang D, Zimmer AD, Terry A, Salamov A, et al. (2008) The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants. Science 319: 64–69.
70. Corrêa LGG (2004) Análise Filogenética de Fatores de Transcrição bZIP em Angiospermas. (Phylogenetic analyses of bZIP transcription factors in angiosperms) [dissertation]. Universidade Estadual de Campinas, Campinas, Brazil.
71. Nijhawan A, Jain M, Tyagi AK, Khurana JP (2008) Genomic survey and gene expression analysis of the basic leucine zipper transcription factor family in rice. Plant Physiol 146: 333–350.
72. Thornton JW, DeSalle R (2000) Gene family evolution and homology: genomics meets phylogenetics. Annu Rev Genomics Hum Genet 1: 41–73.
73. Yang Z (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol 15: 568–573.
74. Kellogg EA (2004) Evolution of developmental traits. Curr Opin Plant Biol 7: 92–98.
75. Nam J, Kim J, Lee S, An G, Ma H, et al. (2004) Type I MADS-box genes have experienced faster birth-and-death evolution than type II MADS-box genes in angiosperms. Proc Natl Acad Sci U S A 101: 1910–1915.

76. Lynch M, Katju V (2004) The altered evolutionary trajectories of gene duplicates. Trends Genet 20: 544–549.
77. Carlini LE, Ketudat M, Parsons RL, Prabhakar S, Schmidt RJ, et al. (1999) The maize EmBP-1 orthologue differentially regulates opaque2-dependent gene expression in yeast and cultured maize endosperm cells. Plant Mol Biol 41: 339–349.
78. Vinson C, Myakishev M, Acharya A, Mir AA, Moll JR, et al. (2002) Classification of human B-ZIP proteins based on dimerization properties. Mol Cell Biol 22: 6321–6335.
79. Fassler J, Landsman D, Acharya A, Moll JR, Bonovich M, et al. (2002) B-ZIP proteins encoded by the Drosophila genome: evaluation of potential dimerization partners. Genome Res 12: 1190–1200.
80. Deppmann CD, Alvania RS, Taparowsky EJ (2006) Cross-species annotation of basic leucine zipper factor interactions: Insight into the evolution of closed interaction networks. Mol Biol Evol 23: 1480–1492.
81. Best AA, Morrison HG, McArthur AG, Sogin ML, Olsen GJ (2004) Evolution of eukaryotic transcription: insights from the genome of *Giardia lamblia*. Genome Res 14: 1537–1547.
82. Alvarez-Buylla ER, Pelaz S, Liljegren SJ, Gold SE, Burgeff C, et al. (2000) An ancestral MADS-box gene duplication occurred before the divergence of plants and animals. Proc Natl Acad Sci U S A 97: 5328–5333.
83. Singer SD, Krogan NT, Ashton NW (2007) Clues about the ancestral roles of plant MADS-box genes from a functional analysis of moss homologues. Plant Cell Rep 26: 1155–1169.
84. Tanabe Y, Hasebe M, Sekimoto H, Nishiyama T, Kitani M, et al. (2005) Characterization of MADS-box genes in charophycean green algae and its implication for the evolution of MADS-box genes. Proc Natl Acad Sci U S A 102: 2436–2441.
85. Shigyo M, Tabei N, Yoneyama T, Yanagisawa S (2007) Evolutionary processes during the formation of the plant-specific Dof transcription factor family. Plant Cell Physiol 48: 179–185.
86. Holm M, Ma LG, Qu LJ, Deng XW (2002) Two interacting bZIP proteins are direct targets of COP1-mediated control of light-dependent gene expression in Arabidopsis. Genes Dev 16: 1247–1259.
87. Richardt S, Lang D, Reski R, Frank W, Rensing SA (2007) PlanTAPDB, a phylogeny-based resource of plant transcription-associated proteins. Plant Physiol 143: 1452–1466.
88. Yi C, Deng XW (2005) COP1 - from plant photomorphogenesis to mammalian tumorigenesis. Trends Cell Biol 15: 618–625.
89. Kamisugi Y, Cuming AC (2005) The evolution of the abscisic acid-response in land plants: comparative analysis of group 1 *LEA* gene expression in moss and cereals. Plant Mol Biol 59: 723–737.
90. Marella HH, Sakata Y, Quatrano RS (2006) Characterization and functional analysis of ABSCISIC ACID INSENSITIVE3-like genes from *Physcomitrella patens*. Plant J 46: 1032–1044.
91. Singer SD, Ashton NW (2007) Revelation of ancestral roles of KNOX genes by a functional analysis of Physcomitrella homologues. Plant Cell Rep 26: 2039–2054.
92. Vinson C, Acharya A, Taparowsky EJ (2006) Deciphering B-ZIP transcription factor interactions *in vitro* and *in vivo*. Biochim Biophys Acta 1759: 4–12.
93. Lawrence CL, Maekawa H, Worthington JL, Reiter W, Wilkinson CR, et al. (2007) Regulation of *Schizosaccharomyces pombe* Atf1 protein levels by Sty1-mediated phosphorylation and heterodimerization with Pcr1. J Biol Chem 282: 5160–5170.
94. Rodrigues-Pousada CA, Nevitt T, Menezes R, Azevedo D, Pereira J, et al. (2004) Yeast activator proteins and stress response: an overview. FEBS Lett 567: 80–85.
95. Jaiswal AK (2004) Nrf2 signaling in coordinated activation of antioxidant gene expression. Free Radic Biol Med 36: 1199–1207.
96. Warabi E, Takabe W, Minami T, Inoue K, Itoh K, et al. (2007) Shear stress stabilizes NF-E2-related factor 2 and induces antioxidant genes in endothelial cells: role of reactive oxygen/nitrogen species. Free Radic Biol Med 42: 260–269.
97. Makino C, Sano Y, Shinagawa T, Millar JB, Ishii S (2006) Sin1 binds to both ATF-2 and p38 and enhances ATF-2-dependent transcription in an SAPK signaling pathway. Genes Cells 11: 1239–1251.
98. Yoshida H, Haze K, Yanagi H, Yura T, Mori K (1998) Identification of the *cis*-acting endoplasmic reticulum stress response element responsible for transcriptional induction of mammalian glucose-regulated proteins. Involvement of basic leucine zipper transcription factors. J Biol Chem 273: 33741–33749.
99. Cox JS, Walter P (1996) A novel mechanism for regulating activity of a transcription factor that controls the unfolded protein response. Cell 87: 391–404.
100. Hahn MW, De Bie T, Stajich JE, Nguyen C, Cristianini N (2005) Estimating the tempo and mode of gene family evolution from comparative genomic data. Genome Res 15: 1153–1160.
101. Demuth JP, Wade MJ (2007) Maternal expression increases the rate of bicoid evolution by relaxing selective constraint. Genetica 129: 37–43.
102. Riaño-Pachón DM, Ruzicic S, Dreyer I, Mueller-Roeber B (2007) PlnTFDB: an integrative plant transcription factor database. BMC Bioinformatics 8: 42.
103. Childs KL, Hamilton JP, Zhu W, Ly E, Cheung F, et al. (2007) The TIGR Plant Transcript Assemblies database. Nucleic Acids Res 35: D846–D851.
104. Pertea G, Huang X, Liang F, Antonescu V, Sultana R, et al. (2003) TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. Bioinformatics 19: 651–652.
105. Kikuchi S, Satoh K, Nagata T, Kawagashira N, Doi K, et al. (2003) Collection, mapping, and annotation of over 28,000 cDNA clones from japonica rice. Science 301: 376–379.
106. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403–410.
107. Birney E, Clamp M, Durbin R (2004) GeneWise and Genomewise. Genome Res 14: 988–995.
108. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucl Acids Res 25: 4876–4882.
109. Kumar S, Tamura K, Nei M (2004) MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. Brief Bioinform 5: 150–163.
110. Felsenstein J (1989) PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics 5: 164–166.
111. Dayhoff MO, Schwartz RC, Orcutt BC (1978) A model of evolutionary change in proteins. In: Dayhoff MO, ed. Atlas of protein sequence and structure. Silver Spring, MD: National Biomedical Research Fundation Silver. pp 301–310.
112. Nei M, Kumar S (2000) Molecular Evolution and Phylogenetics. New York: Oxford University Press.
113. Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22: 2688–2690.
114. Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 8: 275–282.
115. Suzuki Y, Glazko GV, Nei M (2002) Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. Proc Natl Acad Sci U S A 99: 16138–16143.
116. Swofford DL (2003) PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Sinauer Associates).
117. Tamura K, Nei M (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 10: 512–526.
118. Posada D, Crandall KA (1998) MODELTEST: testing the model of DNA substitution. Bioinformatics 14: 817–818.
119. Schmidt HA, Strimmer K, Vingron M, von Haeseler A (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18: 502–504.
120. Bailey TL, Elkan C (1995) The value of prior knowledge in discovering motifs with MEME. Proc Int Conf Intell Syst Mol Biol 3: 21–29.
121. Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, et al. (2006) Pfam: clans, web tools and services. Nucleic Acids Res 34: D247–D251.
122. De Bie T, Cristianini N, Demuth JP, Hahn MW (2006) CAFE: a computational tool for the study of gene family evolution. Bioinformatics 22: 1269–1271.
123. Zimmermann P, Hirsch-Hoffmann M, Hennig L, Gruissem W (2004) GENEVESTIGATOR. Arabidopsis microarray database and analysis toolbox. Plant Physiol 136: 2621–2632.
124. Boyes DC, Zayed AM, Ascenzi R, McCaskill AJ, Hoffman NE, et al. (2001) Growth stage-based phenotypic analysis of Arabidopsis: a model for high throughput functional genomics in plants. Plant Cell 13: 1499–1510.
125. Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, et al. (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat Biotechnol 18: 630–634.

5

Transcription factors in plant senescence

Transcription factors regulating leaf senescence in *Arabidopsis thaliana*

Salma Balazadeh[1,2], Diego Mauricio Riaño-Pachón[1,2], and Bernd Mueller-Roeber[1,2]
[1] Max-Planck Institute of Molecular Plant Physiology, Potsdam-Golm, Germany, [2] University of Potsdam, Institute of Biochemistry and Biology, Potsdam-Golm, Germany

Published in *Plant Biology* (2008) **10**(s1):63-75. doi:10.1111/j.1438-8677.2008.00088.x

Author contributions

BMR conceived and designed and coordinated the study. SB performed the expression profiling experiments using the qRT-PCR resource at the MPIMP, and identified differentially expressed TFs. DMRP clustered expression profile patterns and computed TF family over-representation in the expression profile clusters. All authors discussed and analysed the data.

Plant Biology ISSN 1435-8603

REVIEW ARTICLE

Transcription factors regulating leaf senescence in *Arabidopsis thaliana*

S. Balazadeh[1,2], D. M. Riaño-Pachón[1,2] & B. Mueller-Roeber[1,2]

1 Max-Planck Institute of Molecular Plant Physiology, Potsdam-Golm, Germany
2 University of Potsdam, Institute of Biochemistry and Biology, Potsdam-Golm, Germany

Keywords
Abiotic stress; *Arabidopsis;* expression profiling; leaf senescence; transcription regulators.

Correspondence
B. Mueller-Roeber, Max-Planck Institute of Molecular Plant Physiology, Am Mühlenberg 1, Potsdam-Golm 14476, Germany.
E-mail: bmr@uni-potsdam.de

Guest editor
K. Krupinska

Received: 7 January 2008; Accepted: 12 February 2008

doi:10.1111/j.1438-8677.2008.00088.x

ABSTRACT

Senescence is a highly regulated process, eventually leading to cell and tissue disintegration: a physiological process associated with nutrient (*e.g.* nitrogen) redistribution from leaves to reproductive organs. Senescence is not observed in young leaves, indicating that repressors efficiently act to suppress cell degradation during early leaf development and/or that senescence activators are switched on when a leaf ages. Thus, massive regulatory network re-wiring likely constitutes an important component of the pre-senescence process. Transcription factors (TFs) have been shown to be central elements of such regulatory networks. Here, we used quantitative real-time polymerase chain reaction (qRT-PCR) analysis to study the expression of 1880 TF genes during pre-senescence and early-senescence stages of leaf development, using *Arabidopsis thaliana* as a model. We show that the expression of 185 TF genes changes when leaves develop from half to fully expanded leaves and finally enter partial senescence. Our analysis identified 41 TF genes that were gradually up-regulated as leaves progressed through these developmental stages. We also identified 144 TF genes that were down-regulated during senescence. A considerable number of the senescence-regulated TF genes were found to respond to abiotic stress, and salt stress appeared to be the major factor controlling their expression. Our data indicate a peculiar fine-tuning of developmental shifts during late-leaf development that is controlled by TFs.

INTRODUCTION

Senescence is an important phase of leaf development. It supports the redistribution of micro- and macro-nutrients, including nitrogen, sulphur, phosphorus and potassium, to growing and reproductive organs (young leaves, developing seeds, fruits) (Buchanan-Wollaston 1997; Quirino *et al.* 2000; Hörtensteiner & Feller 2002). Biochemically, senescence is characterised by the degradation of chlorophyll, proteins, lipids and RNA. The progression through later stages of the senescence process is visible as leaf yellowing resulting from chlorophyll loss and chloroplast disassembly (Woolhouse 1984; Thomson & Platt-Aloia 1987). Various factors participate in triggering and modulating the senescence process, including nutrient availability (Crafts-Brandner *et al.* 1998; Diaz *et al.* 2006), hormones (van der Graaff *et al.* 2006), sugars (Pourtau *et al.* 2006; Wingler *et al.* 2006) and extended darkness (in individual leaves; Weaver & Amasino 2001). Also, abiotic and biotic stresses (drought, salt stress, high temperature, pathogen attack and others) can trigger and affect senescence to various extents (*e.g.* Buchanan-Wollaston *et al.* 2003). Transcriptional control mechanisms leading to differential gene expression are believed to play important roles in coordinating the senescence process. In senescing leaves, many of the genes expressed in green leaves, *e.g.* those encoding photosynthetic proteins, are down-regulated (senescence down-regulated genes, SDGs), while other genes are up-regulated (generally referred to as senescence-associated genes, SAGs). Recently, different experimental approaches, including microarray-based expression profiling and suppression subtractive hybridisation revealed hundreds of genes changing their expression during developmentally-regulated leaf senescence in

Arabidopsis or when senescence was artificially induced through prolonged dark incubation or leaf detachment (*e.g.* Buchanan-Wollaston *et al.* 2003; Gepstein *et al.* 2003; Guo *et al.* 2004; Lin & Wu 2004; Buchanan-Wollaston *et al.* 2005; van der Graaff *et al.* 2006). Reprogramming of transcriptomes during senescence has also been studied in other plant species, such as free-growing aspen (*Populus tremula*; Andersson *et al.* 2004) and wheat (Gregersen & Holm 2007). Genes encoding transcription factors (TFs) often represent a sizable fraction of the senescence-associated expression clusters, supporting the notion that this group of regulatory proteins is particularly important in coordinating the progression towards and through this final stage of leaf development.

TRANSCRIPTION FACTORS CONTROLLING LEAF SENESCENCE IN *ARABIDOPSIS THALIANA*

Transcription factors (TFs) are master-control proteins in all living cells. They regulate gene expression by binding to distinct *cis*-elements generally located in the 5′ upstream regulatory regions of target genes, resulting in their activation and/or suppression. Of the more than 25,000 genes that have been annotated in the *Arabidopsis* nuclear genome (http://www.arabidopsis.org) approximately 5–6% code for TFs (Riechmann *et al.* 2000; Davuluri *et al.* 2003). Although much has been learned about transcriptional control in plants in recent years, the biological roles of many TFs remain enigmatic. Of the large number of TFs encoded by the *Arabidopsis* genome, surprisingly few have been functionally related to senescence, although many are known to be induced, and some repressed, in senescing tissues. Among the largest groups of senescence-regulated TFs are the NAC, WRKY, MYB, C2H2 zinc-finger, bZIP and AP2/EREBP families (*e.g.* Chen *et al.* 2002; Guo *et al.* 2004; Lin & Wu 2004; Buchanan-Wollaston *et al.* 2005).

Although approximately 20 *NAC* genes in *Arabidopsis* exhibit elevated expression in senescing leaves (Guo *et al.* 2004; Lin & Wu 2004; and data extractable via the eFP browser website at http://www.bar.utoronto.ca/efp/cgi-bin/efpWeb.cgi), only one of them, *AtNAP* (At1g69490; also called NAC2 or ANAC029) has been shown to control leaf senescence (Guo & Gan 2006). NAC TFs have been found to be encoded by the genome of vascular plants but not of unicellular green algae, such as *Chlamydomonas reinhardtii* and *Ostreococcus tauri* (http://plntfdb.bio.uni-potsdam.de; Riaño-Pachón *et al.* 2007). Their N-terminal region encompasses the highly conserved NAC domain that was originally identified in the proteins NAM from petunia and ATAF1, ATAF2 and CUC2 from *Arabidopsis* (Aida *et al.* 1997). The highly divergent C-terminal parts of NAC TFs are putative transcriptional activation domains. *AtNAP* was previously found in microarray-based transcriptome studies to be strongly expressed in senescent, but less so in non-senescent, *Arabidopsis* leaves (Guo *et al.* 2004). To prove that *AtNAP* controls leaf senescence, Guo and colleagues took advantage of *atnap* null (T-DNA insertion) mutants. They observed that leaf senescence was strongly delayed in the mutants and that expression of the highly senescence-specific marker gene *SAG12* was reduced. The mutant phenotype could be complemented by a homologue from *Phaseolus vulgaris* (kidney bean). This was also possible with an *AtNAP* homologue from a monocot plant (rice; Guo & Gan 2006), indicating it faithfully retained its *cis*-element recognition specificity in a dicot plant (*Arabidopsis*). This observation argues for a significant degree of evolutionary conservation of the function of *AtNAP* homologues in the regulatory pathway controlling senescence, and underscores its importance in this physiological process. Further evidence for a role in senescence regulation was provided by transgenic plants expressing *AtNAP* under the control of a chemically inducible promoter. After application of the chemical (dexamethasone, a synthetic glucocorticoid) precocious leaf senescence and a significant reduction of the F_v/F_m ratio, reflecting a lowered photochemical quantum efficiency of photosystem II, was observed (Guo & Gan 2006). None of the other *Arabidopsis NAC* genes has been shown to regulate the onset or progression of senescence. However, a functional role for NAC TFs in relation to senescence has recently been demonstrated in wheat. Positional cloning of a quantitative trait locus (QTL) that is associated with increased grain protein, zinc and iron content, *Gpc-B1*, identified NAM-B1 as a TF that accelerates leaf senescence when present in a functional form. The ancestral (wild) wheat allele encodes such a functional NAC TF, which is, however, missing in modern wheat cultivars. Inhibition of NAM homologues through RNA interference in transgenic wheat resulted in delayed leaf senescence and reduced grain protein, zinc and iron content (Uauy *et al.* 2006). Regulation of senescence is, however, not the only function of NAC TFs. They have previously been shown to be involved in a number of other crucial developmental and physiological processes, including, for example, abscisic acid inducible gene expression (Fujita *et al.* 2004), lateral root development (He *et al.* 2005), secondary wall synthesis (Zhong *et al.* 2006), regulation of cell division (Kim *et al.* 2006a), responses to pathogen attack (Collinge & Boller 2001) and regulation of salt tolerance (Nakashima *et al.* 2007).

Besides NAC TFs, several members of the WRKY family of transcription factors have been shown to exert a prominent role in regulating *Arabidopsis* senescence, besides being central in disease-resistance pathways (Eulgem & Somssich 2007). The WRKY family comprises zinc finger-type transcription factors. *WRKY53* (At4g23810) is a senescence-induced transcription factor gene (Hinderhofer & Zentgraf 2001). Inhibiting *WRKY53* function through T-DNA insertion or RNA interference retards leaf senescence in low-light conditions in long-day culture (Miao *et al.*, 2004). Importantly, more than 60 putative target genes of WRKY53 have been identified, including at least six other members of the *WRKY* gene family, suggesting that it acts as an upstream control element in a

transcription factor signalling cascade leading to leaf senescence (Miao *et al.* 2004). The senescence marker gene *SAG12*, encoding a cysteine protease, was among the targets of WRKY53. *SAG12* expression only occurs during the senescence of older leaves and is not generally regarded as a marker for early senescence stages (Noh & Amasino 1999). Therefore, although WRKY53 appears to adopt a function at the beginning of the leaf senescence cascade, it apparently also regulates the expression of genes playing a role at a later stage of senescence. Recently, a jasmonic acid (JA)-inducible protein called EPITHIOSPECIFYING SENESCENCE REGULATOR (ESR/ESP) was discovered to interact with WRKY53. Expression of the *ESR/ESP* and *WRKY53* genes is antagonistically regulated by salicylic acid (SA) and JA. Leaf senescence is accelerated in *ESR/ESP* mutants, indicating that the physical interaction with WRKY53 protein is indeed of functional relevance in this process (Miao & Zentgraf 2007). Another intriguing observation was recently made by the same group: searching for proteins that are upstream of *WRKY53*, they discovered a mitogen-activated protein kinase kinase kinase (MEKK1) binding to its promoter (Miao *et al.* 2007). MEKK1 also interacted with WRKY53 protein *in vivo* and phosphorylated it *in vitro*, enhancing its DNA-binding activity towards the *WRKY53* promoter and its transcriptional activation.

WRKY4, 6, 7 and *11* are other members of the family that were shown to be strongly up-regulated during leaf senescence. Expression of *WRKY6* (At1g62300) was analysed in more detail and found to be induced by wounding and treatment with SA, JA or ethylene (Robatzek & Somssich 2001). Strong over-expression of *WRKY6* under control of the cauliflower mosaic virus 35S promoter induced a pleiotropic plant phenotype (dwarfing, necrotic leaves, reduction of apical dominance, early flowering) (Robatzek & Somssich 2002). Evidence was obtained indicating that WRKY6 negatively regulates its own promoter function, pointing to it having repressor activity. However, it exerts positive regulatory activity on other genes, such as the senescence- and pathogen defence-associated *PR1* gene, although this activation might be indirect through the involvement of NPR1, a key regulator of SAR-dependent signalling (Robatzek & Somssich 2002). Target genes of WRKY6 have been identified, and *SIRK*, encoding a receptor-like protein kinase, is one of these. Recently, the function of another *WRKY* gene, *WRKY70* (At3g56400), was analysed. Microarray expression data, as well as promoter-β-glucuronidase (GUS) fusions, showed it to be expressed throughout leaf development, with enhanced expression in senescing leaves. Loss of *WRKY70* function in two independent T-DNA insertion lines promoted both developmentally and dark-induced leaf senescence, indicating that it constitutes a negative regulator of senescence (Ülker *et al.* 2007).

Another interesting observation was recently made by Ellis *et al.* (2005) and Okushima *et al.* (2005). Both groups found that ARF2 (At5g62000), a member of the AUXIN RESPONSE FACTOR family of TFs that mediate responses to the plant hormone auxin, functions as a repressor of age-dependent and dark-induced rosette leaf senescence and several other age-related processes in *Arabidopsis*, including floral organ abscission. Overall, ARF2 appears to be a pleiotropic developmental regulator that also affects leaf size, flower morphology and hypocotyl length. Mutations in several other *ARF* genes, *i.e. ARF1*, *NPH4/ARF7* and *ARF19*, typically enhanced *arf2* mutant phenotypes. Mutations in these genes alone, however, did not affect senescence (Ellis *et al.* 2005). As expression of all genes overlaps in some tissues (*e.g.* expression of *ARF2*, *NPH4/ARF7* and *ARF19* increases in response to senescence; all three genes including *ARF1* are expressed at the flower base, including the abscission zone), their protein products might interact to exert age-dependent functions.

Cytokinins are plant hormones that have profound effects on many developmental and physiological processes, including the regulation of leaf longevity. Recently, it was demonstrated in *Arabidopsis* that ARR2, a B-type response regulator of the cytokinin receptor AHK3, controls leaf longevity (Kim *et al.* 2006b). Overexpression of wild-type ARR2 TF delays dark-induced and age-dependent senescence, whereas overexpression of a mutant version of ARR2 that is not phosphorylated through the AHK3-dependent signalling pathway does not affect leaf longevity (Hwang & Sheen 2001; Kim *et al.* 2006b). These observations suggest that cytokinin-induced phosphorylation of ARR2 has a positive role in cytokinin-mediated control of leaf longevity. However, an early senescence phenotype was not observed in *arr2* knockout plants, suggesting that other ARR TFs or other senescence control systems compensate for the loss of ARR2 TF activity (Kim *et al.* 2006b). Other recently discovered TFs that appear to play a role in cytokinin-mediated processes are the GeBP/GPL proteins. *GeBP* (*GLABROUS1 enhancer Binding Protein*) and *GPL* (*GeBP-like*) genes encode a newly-defined class of TFs containing a non-canonical leucine zipper motif. A triple loss-of-function mutant of the three closely-related genes *GeBP*, *GPL1* and *GPL2*, exhibited lowered sensitivity to exogenously applied cytokinins. Typically, in detached leaves, chlorophyll loss occurs during dark-induced senescence, a response that is normally inhibited by cytokinins such as 6-benzyl-adenine (BA). Chlorophyll loss was found to be more severe in the mutant than in the wild type, indicating that part of the senescence-delaying property of cytokinins is mediated through a signalling pathways involving GeBP/GPL TFs (Chevalier *et al.* 2007). Inhibition of chlorophyll loss by cytokinins during dark-induced senescence was also impaired in an *arr1 arr12* double mutant (*ARR1* and *ARR12* encode B-type ARRs) (Chevalier *et al.* 2007).

Besides TFs, the chromatin architecture-controlling AT hook protein ORE7/ESC has also been shown to control leaf senescence in *Arabidopsis*. Overexpression of the *ORE7/ESC* gene extends leaf longevity, alters chromatin structure and globally triggers changes in the

transcriptome that are consistent with a more juvenile status of the mutant leaves (Lim *et al.* 2007).

TRANSCRIPTION FACTOR EXPRESSION PROFILING

TF genes are often expressed in a cell- or tissue-specific manner, or at low levels. Due to technical limitations, accurate TF expression profiling with microarrays is difficult (*e.g.* Czechowski *et al.* 2004). Most importantly, the down-regulation (repression) of already weakly expressed genes can not be reliably studied using current macro- or microarray-based technologies. In contrast, quantitative reverse transcription–polymerase chain reaction (qRT-PCR or real-time RT-PCR) allows even weakly expressed genes to be accurately quantified (Pfaffl *et al.* 2002). Thus, whilst array-based hybridisation typically allows the detection of one transcript per cell (Holland 2002; Horak & Snyder 2002), qRT-PCR can detect one transcript per 1000 cells (Czechowski *et al.* 2004).

We employed qRT-PCR to identify TF genes changing their expression level during *Arabidopsis thaliana* leaf growth and the beginning of senescence. We used an advanced version of an expression profiling platform that was originally described by Czechowski *et al.* (2004). The current TF profiling platform covers 1880 TF genes. For our study, we chose leaves of three developmental stages: 50% expanded (L50: ~1.5 cm leaf length); 100% expanded (L100: ~3 cm leaf length, no visible senescence); and fully expanded with approximately 20% of the leaf blade showing senescence, starting at the tip region, where leaves turned yellow due to chlorophyll loss (S20). Leaves were harvested from approximately ten individual plants for each stage. We chose leaf number 11 of the *Arabidopsis* Colombia-0 ecotype for all experiments, as it is one of the first leaves of the rosette growing to full size (leaves produced earlier generally remained smaller, even at full development). Focusing on a distinct leaf also helped us to precisely define the developmental stage of the tissue analysed and to reduce the risk of potential confounding effects that might otherwise occur when whole plants are sampled in senescence studies.

To follow the progression of senescence in the leaf samples, chlorophyll content was monitored. A steady decline in chlorophyll levels (normalised to leaf fresh weight) was observed (Fig. 1A). Notably, although fully expanded leaves did not visibly appear senescent, their chlorophyll content was significantly lower than that of 50% expanded (L50) leaves. We also analysed the photosynthetic efficiency of leaves from the different developmental stages. Photosynthetic efficiency was slightly lower in L100 than in L50 leaves, and further declined in S20 leaves and upon further progression of senescence (in S50 leaves) (Fig. 1B). Thus, photosynthetic efficiency followed the chlorophyll concentration. Based on these results, we considered fully expanded (L100) leaves as representing a physiological stage of the start of senescence. Analysing this leaf stage will likely help to discover important information on the molecular and biochemical processes that prepare leaf physiology for the processes leading to (visible) senescence.

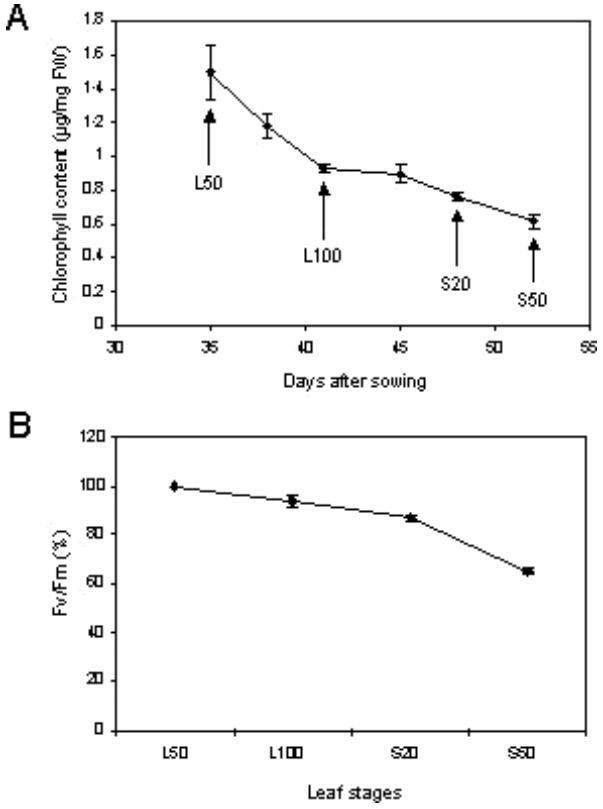


Fig. 1. A: Chlorophyll concentration, and B: F_v/F_m ratio reflecting photochemical quantum efficiency of photosystem II of leaf number 11 of *Arabidopsis thaliana,* accession Columbia-0. Plants were grown in soil (Einheitserde GS90; Gebrüder Patzer, Sinntal-Jossa, Germany) in a growth chamber with a 16-h daylength provided by fluorescent light at 120 $\mu mol \cdot m^{-2} \cdot s^{-1}$ and a day/night temperature of 20/16 °C and relative humidity of 60/75%. The following developmental stages of leaf number 11 (*i.e.* the 11th leaf that emerged after the cotyledons) were used: L50 (50% fully expanded leaf, 15 ± 3 mm long; harvested from 35-day-old plants); L100 (fully expanded leaf; 30 ± 3 mm long; plants were 41 days old); S20 (fully expanded leaves with ~20% leaf yellowing; plants were ~50 days old); and S50 (leaves with ~50% leaf yellowing; plants were ~53 days old). For chlorophyll measurements leaves were ground in liquid nitrogen, re-suspended in 1 ml of 96% (v/v) ethanol, and homogenised. After centrifugation, chlorophyll (chl) was determined photometrically at 650 nm. Chlorophyll content is given as µg chl per 1 mg leaf fresh weight. The photochemical efficiency of photosystem II (PSII) was deduced from the characteristics of chlorophyll fluorescence using a pulse-amplitude modulated portable fluorometer PAM-2000 (Heinz Walz, Effeltrich, Germany) with the leaf clip holder 2030-B following the manufacturer's instructions. The leaf was held by the leaf clip holder without dark adaptation and then a brief and strong light pulse at a frequency of 600 or 20000 Hz was applied for 3 µs to induce fluorescence excitation.

The available TF qRT-PCR platform allowed us to score the expression level of 1880 TF genes with high

confidence. Transcriptional changes were calculated based on the comparative C_T method. Briefly, the C_T value of each TF was normalised to the C_T value of the reference gene *UBQ10 (*At4g05320), revealing ΔC_T. To calculate fold changes of TF expression levels, the ΔC_T of each two stages were subtracted from each other, resulting in $\Delta\Delta C_T$. Genes were considered differentially expressed when the change was more than fivefold ($\log_2 > 2.3$) between any of the three leaf stages analysed. The expression level of these genes was further investigated in two additional independent biological replicates. A total of 185 TF genes displayed differential expression among the three developmental leaf stages, representing ~10% of all TF genes tested here. Analysis of transcript profiles revealed that the expression of 144 TF genes declined when leaves expanded from L50 to L100 stages and became visibly senescent at S20 stage, or had lower expression in L100 leaves in comparison to L50 and S20 leaves, potentially reflecting functions during early, but not late, senescence. The expression of 30 TF genes increased throughout phases L50 to L100 and S20, and the expression of 11 TF genes peaked in L100 leaves,while being lower at L50 and S20 stages. The list of TF genes identified by our screen is given in Table 1 (genes sorted according to family membership; the complete list of all expression data is available from the authors upon request).

In order to uncover groups of genes with similar expression patterns, we performed cluster analysis of the senescence-related 185 TF genes. We performed *K*-means clustering on correlation values (Pearson); six clusters were determined to be the optimal number of groups for the data. The expression profiles of genes in each cluster are shown in Fig. 2. Cluster B includes 73 genes whose expression decreased steadily throughout leaf development, from stage L50 to stages L100 and S20. Similarly, cluster C includes 36 genes whose expression decreases towards the S20 stage (genes of clusters B and C are collectively called senescence down-regulated, SDGs, here). Cluster F includes 30 TF genes whose transcript abundance generally increased in later stages of leaf development (SAGs); the cluster also includes genes exhibiting a more prominent change in expression after full leaf size has been reached in the L100 stage. Genes of the remaining clusters, *i.e.* clusters A (11 genes, early SAGs, ESAGs), D (18 genes) and E (17 genes; genes of clusters D and E are collectively called early SDGs, ESDGs here), show additional patterns which may be important for fine-tuning gene expression during senescence.

TF families preferentially contributing to the senescence transcriptome

Leaf senescence is a higher plant-specific developmental process and thus it appears possible that some of the plant TF gene families selectively expanded throughout evolution to accommodate the specific functions needed for fine-tuning this process. Thus, we were interested to know whether any of the TF gene families that we analysed by qRT-PCR-based expression profiling preferentially contributes to the senescence transcriptome (Table 2). We found that members of the *NAC* TF family are significantly over-represented ($P_c << 0.05$) in the SAG group of TFs. Twelve out of 66 *NAC* TF genes that were expressed in leaves (in the three developmental stages tested) belong to this group, showing at least fivefold up-regulation in L100 and S20 stages of leaf development, compared to stage L50. Although *NAC* TFs were not over-represented in the ESAG group, and *WRKY* TFs were not statistically over-represented in SAG or ESAG groups (referenced to TF family sizes), approximately half, *i.e.* 22 out of the 41 SAG and ESAG TFs, are members of the *NAC* and *WRKY* TF families, indicating their important role in leaf senescence.

Among the group of early senescence down-regulated genes (ESDGs), TFs of the *AP2-EREBP* and *bHLH* families were significantly over-represented, ($P_c \leq 0.05$) (Table 2). Members of the *bHLH* and *GATA* families were moderately over-represented ($P_c < 0.1$) in the SDG group. Collectively, 38% (17 out of 45) of all leaf-expressed *bHLH* TFs belong to the SDG/ESDG groups, whereas only two are in the SAG/ESAG groups. Similarly, 27 out of 117 *AP2-EREBP* TFs are in the SDG/ESDG groups, but only one is in the SAG/ESAG groups. Eight out of 26 TFs (31%) of the *GATA* family belong to the SDG group, whereas no GATA TF was found in the SAG/ESAG groups. One of the *GATA* genes, called *GNC* (for GATA, nitrate-inducible, carbon metabolism-involved; At3g50870), has been shown to have a role in the regulation of carbon/nitrogen metabolism. Mutants deficient in this gene have lower chlorophyll levels and are hypersensitive to exogenous glucose (Bi *et al.* 2005). Notably, expression profiling identified only two TF genes (NAC At1g52890 and WRKY At3g01970) that were significantly repressed in the *gnc* mutant compared to the wild type. Both genes were found here to be up-regulated during leaf senescence.

TF genes down-regulated during natural leaf senescence

In our study, focusing on leaf 11 of the *Arabidopsis* rosette, we detected more TF genes being down-regulated (clusters B and C) than up-regulated (cluster F) during senescence. TFs induced during senescence are generally assumed to actively participate in regulating the senescence process, whereas down-regulated TF genes might reflect a more general reduction of the leaf maintenance machinery rather than being an active part of the senescence regulation network itself. Using multi-parallel qRT-PCR we faithfully detected expression in leaves of 1430 of the 1880 TF genes covered by the whole platform. This indicates that, not unexpectedly, a certain fraction of these TFs is not expressed to any detectable level in leaves, at least not in the developmental stages we tested under our experimental conditions. From all TFs found to be expressed in leaves, ~13% (185 TFs) exhibited a

Table 1. Transcription factor genes exhibiting differential expression in leaf stages L50, L100 and S20.

		ΔC_T L50		ΔC_T L100		ΔC_T S20				ΔC_T L50		ΔC_T L100		ΔC_T S20	
locus	family	mean	SD	mean	SD	mean	SD	locus	family	mean	SD	mean	SD	mean	SD
At1g01030	ABI3VP1	10.44	0.71	14.75	1.02	13.97	1.09	At3g58120	bZIP	5.90	1.8	7.37	1.07	16.76	1.29
At4g01500	ABI3VP1	13.50	0.68	16.61	1.30	16.00	1.57	At1g26610	C2H2	13.79	3.40	15.39	1.48	17.85	1.63
At5g06250	ABI3VP1	13.35	1.10	15.23	0.27	18.02	0.88	At1g75710	C2H2	8.20	0.63	10.39	1.32	13.30	0.94
At1g03800	AP2 ER	13.71	1.72	17.81	0.58	17.03	1.27	At2g28710	C2H2	11.24	0.77	7.38	0.86	7.41	0.22
At1g12610	AP2 ER	12.51	2.25	16.46	2.62	11.76	1.40	At3g49930	C2H2	7.41	0.34	8.35	0.17	12.18	0.85
At1g21910	AP2 ER	6.86	1.14	8.08	1.14	12.21	1.33	At3g58070	C2H2	9.66	0.36	11.25	0.56	12.91	0.81
At1g43160	AP2 ER	7.90	2.04	14.12	0.69	9.11	1.03	At4g02670	C2H2	10.55	0.39	12.76	1.82	13.85	0.84
At1g63040	AP2 ER	12.21	1.49	15.15	2.04	14.44	1.53	At4g16610	C2H2	10.07	1.76	13.01	0.85	14.53	2.31
At1g77200	AP2 ER	11.05	0.42	14.37	0.84	13.95	0.75	At5g03510	C2H2	9.96	0.51	12.23	0.99	13.90	1.05
At2g35700	AP2 ER	10.46	0.34	13.15	1.61	13.27	1.06	At5g04340	C2H2	6.57	1.26	7.33	0.64	4.45	0.60
At2g44840	AP2 ER	8.17	0.97	11.95	1.86	10.54	1.89	At5g16540	C2H2	7.07	0.37	8.51	1.20	10.42	1.10
At2g44940	AP2 ER	6.91	0.74	7.71	1.19	9.36	0.90	At5g54630	C2H2	8.25	0.99	10.20	1.32	12.04	1.00
At4g11140	AP2 ER	10.55	0.60	13.15	0.84	15.04	0.79	At5g60470	C2H2	12.99	0.48	14.79	0.76	10.98	0.76
At4g17490	AP2 ER	8.91	0.96	11.16	1.03	12.79	1.05	At1g72830	CCHAP2	11.98	0.43	10.63	0.75	8.63	0.45
At4g23750	AP2 ER	6.76	1.39	9.60	1.86	12.36	0.63	At2g13570	CCHAP3	11.95	0.59	14.54	0.79	18.54	1.42
At4g32800	AP2 ER	5.23	0.51	6.11	0.54	9.87	0.48	At4g14540	CCHAP3	5.01	1.91	5.92	1.93	8.45	1.68
At4g34410	AP2 ER	12.11	2.26	14.21	0.99	10.71	1.13	At5g27910	CCHAP5	13.86	1.21	16.57	1.45	18.11	2.29
At4g37750	AP2 ER	9.12	1.30	12.49	2.42	14.90	1.21	At5g43250	CCHAP5	8.44	0.58	10.68	0.71	13.46	1.54
At5g07580	AP2 ER	3.99	0.55	4.61	0.79	6.25	0.90	At5g63470	CCHAP5	4.37	0.29	4.81	0.66	7.16	0.51
At5g10510	AP2 ER	13.02	0.93	17.36	2.58	15.38	2.54	At3g22760	CPP (Zn)	10.43	0.71	14.19	1.22	15.14	0.33
At5g11190	AP2 ER	14.13	1.03	16.61	0.89	19.86	1.02	At1g69570	DOF	10.22	0.84	7.26	0.18	7.01	0.51
At5g11590	AP2 ER	6.58	0.75	8.09	1.53	9.88	0.74	At2g37590	DOF	12.33	1.45	16.51	1.70	17.32	1.22
At5g13330	AP2 ER	8.43	1.34	9.88	0.59	5.57	0.96	At3g45610	DOF	8.39	0.46	11.42	0.97	14.81	2.69
At5g25190	AP2 ER	5.41	0.21	6.37	1.26	10.31	1.37	At5g60200	DOF	7.61	0.57	10.44	1.37	11.59	1.41
At5g25390	AP2 ER	9.90	0.41	12.53	1.10	14.21	1.17	At5g62940	DOF	9.91	0.63	13.17	1.96	15.01	3.22
At5g25810	AP2 ER	8.95	0.52	9.55	1.68	12.13	2.07	At5g65590	DOF	8.35	0.46	9.58	0.99	11.81	0.77
At5g51990	AP2 ER	13.39	2.17	16.40	0.31	12.91	2.40	At3g01330	E2F-DP	10.97	1.47	13.35	0.56	15.21	0.44
At5g57390	AP2 ER	12.09	1.18	15.12	1.35	16.92	1.16	At3g48160	E2F-DP	8.33	0.44	11.10	0.77	13.09	0.89
At5g61890	AP2 ER	12.69	0.61	12.73	1.49	9.72	0.77	At2g18380	GATA	7.76	0.40	10.14	1.49	11.64	0.87
At5g64750	AP2 ER	9.72	1.11	12.48	0.83	9.26	0.50	At2g45050	GATA	10.88	0.89	13.05	1.78	16.34	0.30
At5g67180	AP2 ER	11.91	1.21	15.29	0.78	13.45	1.15	At3g60530	GATA	5.11	0.32	6.65	0.49	8.66	0.74
At1g04250	ARP	5.38	0.32	6.95	0.97	9.35	0.69	At4g32890	GATA	6.82	0.61	8.15	0.72	11.12	0.70
At1g15580	ARP	6.74	1.32	8.94	1.37	13.62	3.39	At4g36240	GATA	5.20	0.42	7.64	1.33	8.64	1.44
At1g19220	ARP	12.68	0.30	15.35	0.59	13.50	1.28	At5g25830	GATA	8.83	0.37	10.15	1.40	12.65	0.42
At1g52830	ARP	11.12	1.35	12.92	1.48	17.94	1.45	At5g26930	GATA	11.54	0.67	14.07	0.87	15.53	0.44
At2g22670	ARP	2.87	0.40	5.47	0.76	6.81	0.58	At5g56860	GATA	3.36	0.26	4.97	0.80	6.29	0.61
At3g15540	APR	5.19	1.01	5.61	0.37	9.17	1.04	At4g26170	general	15.25	1.15	18.19	1.48	19.74	0.38
At3g17600	ARP	13.30	1.60	16.49	0.30	18.09	0.65	At2g02540	HB	11.31	3.23	13.47	3.16	17.34	3.77
At3g23050	ARP	3.31	0.29	4.48	0.55	8.36	0.80	At2g44910	HB	16.47	0.54	13.06	0.97	14.71	0.35
At3g62100	ARP	13.58	0.57	17.79	0.15	18.29	0.48	At2g46680	HB	10.12	0.37	9.06	1.06	5.46	0.99
At4g14550	ARP	9.72	0.62	10.84	0.86	14.02	0.71	At3g03260	HB	15.62	0.63	14.40	1.51	17.15	1.00
At4g29080	ARP	7.48	0.34	8.79	0.65	10.97	0.66	At3g11260	HB	11.64	0.79	11.33	0.35	17.97	1.88
At5g43700	ARP	3.88	0.61	4.98	1.54	7.69	0.59	At3g18010	HB	13.52	0.46	16.69	2.72	17.90	1.11
At2g01760	ARR-B	7.54	0.73	9.33	1.72	11.06	0.58	At3g50890	HB	8.79	1.47	10.80	2.39	13.29	0.98
At1g02340	bHLH	12.04	1.25	7.00	0.30	6.86	0.57	At3g61890	HB	8.61	0.33	9.10	0.62	5.58	1.50
At1g12860	bHLH	5.12	1.07	7.66	1.97	9.14	0.90	At4g03250	HB	9.45	1.23	9.85	0.98	11.09	1.46
At1g63650	bHLH	13.87	0.91	17.40	2.40	18.97	1.68	At5g46880	HB	10.80	1.45	13.83	1.55	14.45	1.45
At1g68810	bHLH	10.80	0.32	14.19	0.90	16.12	1.70	At5g65310	HB	3.90	0.24	5.39	1.47	7.14	0.52
At1g72210	bHLH	11.27	4.69	15.53	2.91	15.74	1.80	At4g00480	HLH	12.97	1.91	15.61	1.83	18.52	0.82
At1g73830	bHLH	11.52	1.77	13.75	1.41	18.26	1.83	At4g18870	HSF	15.34	1.13	11.97	0.76	12.64	0.26
At2g22770	bHLH	9.93	0.91	14.66	1.64	11.57	0.49	At5g03720	HSF	10.07	0.57	11.12	1.11	13.52	0.60
At2g41130	bHLH	10.21	0.08	11.61	1.09	14.81	0.99	At5g43840	HSF	14.78	1.22	13.17	0.12	10.48	2.89
At3g56970	bHLH	11.22	3.79	13.05	0.40	18.42	1.27	At5g45710	HSF	7.22	1.90	8.22	1.47	11.11	1.40
At3g61950	bHLH	12.10	0.96	16.75	1.36	19.36	1.09	At1g47760	MADS	8.91	1.23	12.00	1.58	14.79	0.14
At4g01460	bHLH	5.96	0.70	9.04	2.87	14.18	1.57	At5g26870	MADS	4.92	1.21	4.66	1.43	6.97	1.17

Table 1. Continued.

		ΔC_T L50		ΔC_T L100		ΔC_T S20				ΔC_T L50		ΔC_T L100		ΔC_T S20	
locus	family	mean	SD	mean	SD	mean	SD	locus	family	mean	SD	mean	SD	mean	SD
At4g30980	bHLH	12.86	0.51	14.91	1.24	15.97	1.63	At5g27070	MADS	15.31	2.47	15.17	1.09	18.75	1.35
At5g41315	bHLH	11.83	1.26	14.57	1.24	11.44	1.35	At5g27580	MADS	16.51	2.55	17.86	2.47	19.74	2.42
At5g43650	bHLH	11.21	0.77	12.09	1.37	8.08	0.63	At1g01380	MYB	11.56	0.89	13.15	1.18	15.50	1.01
At5g46690	bHLH	6.79	0.46	11.05	0.77	13.81	1.18	At1g06180	MYB	11.00	1.45	12.27	2.29	8.90	1.20
At5g46830	bHLH	11.72	0.18	16.56	1.07	16.65	0.30	At1g08810	MYB	8.61	0.53	10.00	0.21	12.66	1.05
At5g65320	bHLH	12.33	0.64	17.28	1.29	15.40	2.02	At1g48000	MYB	14.13	1.76	12.04	0.67	7.99	1.07
At5g65640	bHLH	6.59	0.62	10.06	2.75	8.35	1.17	At1g56650	MYB	5.87	0.62	10.79	1.95	3.89	0.37
At5g67110	bHLH	10.27	0.86	11.36	1.06	13.24	0.31	At1g63380	MYB	13.10	1.27	14.04	0.22	9.92	0.49
At1g08320	bZIP	17.67	0.49	16.53	0.47	13.13	1.22	At1g66230	MYB	6.77	0.59	8.21	1.65	10.43	1.86
At1g22070	bZIP	9.56	0.33	7.03	1.08	8.26	0.38	At1g66390	MYB	11.90	0.68	13.80	0.44	4.86	0.83
At1g75250	MYB	8.79	0.54	6.60	0.81	12.00	1.38	At1g71930	NAC	13.34	0.78	15.84	1.16	18.06	1.22
At2g31180	MYB	14.96	2.79	17.09	1.19	12.46	2.23	At2g18060	NAC	13.49	1.61	16.58	2.02	18.31	3.14
At2g37630	MYB	5.91	1.12	6.47	0.59	8.68	0.95	At2g43000	NAC	16.75	2.28	12.32	2.52	14.29	2.32
At2g39880	MYB	9.27	0.55	12.21	1.17	14.40	0.73	At3g04070	NAC	10.98	0.39	8.98	0.29	5.12	0.39
At2g46830	MYB	13.66	1.24	14.78	0.98	9.80	0.63	At3g04420	NAC	13.98	0.45	10.65	0.49	11.43	0.29
At2g47190	MYB	12.54	2.34	10.05	2.90	6.51	0.79	At3g15500	NAC	11.26	1.30	10.87	0.36	5.04	0.85
At2g47460	MYB	14.51	0.73	19.06	0.32	21.06	0.81	At3g15510	NAC	10.96	1.40	8.89	0.58	6.89	0.19
At3g01140	MYB	9.23	0.55	11.39	2.04	13.23	0.85	At3g17730	NAC	13.12	0.54	16.49	0.90	16.55	0.50
At3g06490	MYB	14.01	3.12	15.21	2.62	12.44	3.69	At3g29035	NAC	13.21	0.30	8.93	0.08	7.03	0.63
At3g16350	MYB	7.47	0.14	8.91	1.19	10.17	0.76	At4g27410	NAC	5.16	1.18	5.58	0.72	2.60	1.09
At3g27810	MYB	14.44	1.35	17.91	1.31	10.88	0.24	At4g28530	NAC	15.87	1.13	12.45	0.70	10.39	0.83
At3g50060	MYB	11.18	0.85	13.67	0.59	15.86	0.50	At5g07680	NAC	14.66	1.85	13.92	1.50	11.07	2.56
At4g01680	MYB	9.42	0.53	11.06	1.47	12.86	0.70	At5g39610	NAC	10.65	0.71	8.07	0.64	4.90	1.02
At4g05100	MYB	8.29	0.95	9.72	0.23	5.44	1.07	At5g61430	NAC	15.80	1.64	16.28	1.84	12.98	2.55
At4g21440	MYB	12.82	2.30	13.85	1.32	10.30	1.72	At5g64060	NAC	11.31	4.18	15.18	5.62	14.80	3.98
At5g11510	MYB	11.36	0.50	14.95	1.18	17.65	1.69	At3g57920	SBP	15.29	0.44	18.33	1.28	21.26	1.33
At5g40330	MYB	9.93	0.65	11.35	0.60	13.39	0.90	At1g18860	WRKY	17.30	0.69	14.09	1.16	9.52	1.24
At5g54230	MYB	17.73	1.63	15.66	1.56	12.66	0.37	At1g29860	WRKY	16.14	2.87	13.92	3.52	12.21	1.70
At1g13300	MYB-L	18.05	0.76	16.28	1.09	12.06	0.40	At2g37260	WRKY	15.27	0.51	18.85	0.24	13.51	0.95
At2g30420	MYB-L	5.90	0.69	8.83	0.78	13.34	2.23	At2g45190	YABBY	8.04	0.82	10.86	1.64	13.12	1.65
At5g18240	MYB-L	9.94	1.37	11.40	1.08	13.44	0.78	At3g01970	WRKY	11.09	0.52	8.11	0.97	4.70	0.49
At1g02220	NAC	14.85	0.65	14.27	1.74	11.87	0.92	At4g23810	WRKY	3.91	0.30	1.46	0.52	4.28	0.63
At1g12260	NAC	10.44	0.95	12.95	1.53	16.11	1.20	At5g07100	WRKY	13.25	0.40	9.78	0.84	7.91	1.34
At1g52890	NAC	8.99	4.21	9.61	3.43	5.29	3.66	At5g13080	WRKY	16.67	4.62	14.69	4.48	9.91	1.33
At1g54330	NAC	15.88	0.35	18.54	0.54	20.58	0.29	At2g46790		13.92	0.78	13.89	0.90	9.81	0.68
At1g56010	NAC	9.54	0.18	7.25	0.33	4.82	0.65	At3g11110		9.29	0.91	9.92	0.10	15.18	0.80
At1g62700	NAC	14.19	0.69	17.18	0.70	20.64	0.93	At5g63780		4.40	0.328	5.19	1.19	9.90	0.77
At1g69490	NAC	8.88	0.22	6.21	0.41	2.87	0.21								

Arabidopsis thaliana (L.) Heynh., accession Col-0, was used for expression profiling. Plants were grown as indicated in the legend to Fig 1. Leaves were harvested at around 9 a.m. (*i.e.* 3 h into the light period). Expression profiling was essentially done as described in Caldana *et al.* (2007) using an extended version of the *Arabidopsis* TF qRT-PCR platform originally described by Czechowski *et al.* (2004). Absence of genomic DNA was verified by PCR using primers targeting an intron of the control gene At5g65080 (forward 5′-TTTTTTGCCCCCTTCGAATC; reverse 5′-ATCTTCCGCCACCACATTGTAC). Efficiency of cDNA synthesis was controlled by qRT-PCR checking transcripts of three housekeeping genes (At2g28390: forward 5′-AACTCTATGCAGCATTTGATCCACT; reverse 5′-TGATTGCATATCTTTATCGCCATC; At4g26410: forward 5′-GAGCTGAAGTGGCTTCCATGAC; reverse 5′-GGTCCGACATACCCATGATCC; At4g05320: forward 5′-CACACTCCACTTGGTCTTGCGT; reverse 5′-TGGTCTTTCCGGTGAGAGTCTTCA). Triplicate measurements were carried out to determine mRNA abundance of each gene in each leaf sample. Mean and SD (standard deviation) are given. Gene annotations are according to Czechowski *et al.* (2004). Regular updates will be provided through the Plant Transcription Factor Database at http://plntfdb.bio.uni-potsdam.de/v2./.

senescence-dependent shift in expression level. Sixteen per cent of these were found to be up-regulated and 59% to be down-regulated during the transition from L50 to L100 leaves, and when senescence became visible (S20 stage). The remaining TFs exhibited transient increases (6%) or decreases (19%), respectively, upon the transition from L50 to S20 leaves (Table 1 and Fig. 2). Although we cannot exclude at the present stage that transcript abundance of the group of down-regulated TFs diminishes simply because of a general breakdown of macromole-

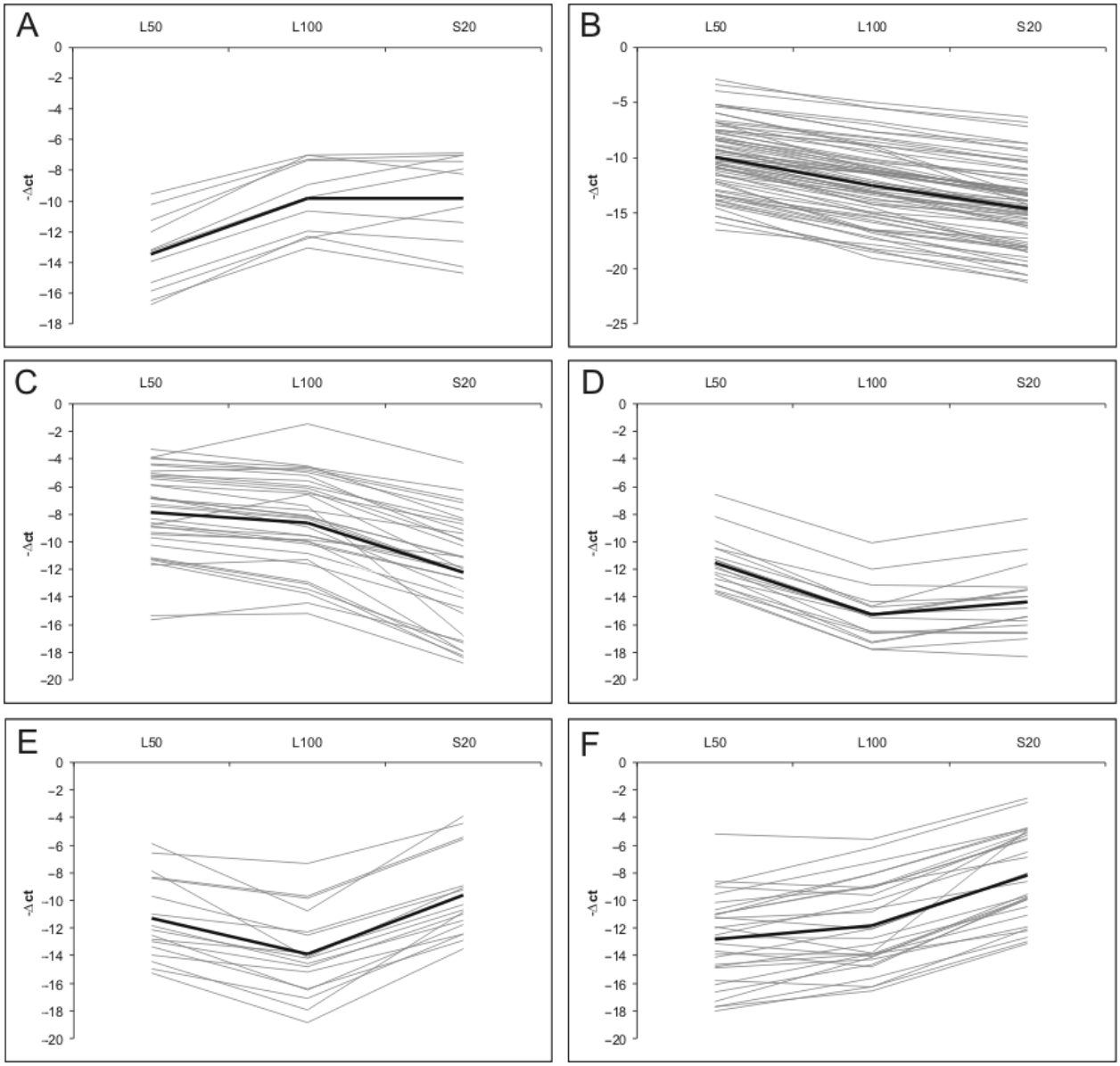


Fig. 2. Cluster analysis of expression data of senescence-related TF genes. Clustering was performed using the *K*-means algorithm on Pearson correlations between genes. The best number of clusters was determined by the Figure of Merit (FOM). The average pattern for each cluster is shown in bold. All analyses were run in the MultiExperiment Viewer (MeV) part of the TM4 software from TIGR (Saeed *et al.* 2003).

cules at senescence, we do not expect this to be the case for all TFs of this group, because, in total, only ~8% (109 TFs) of all leaf-expressed TFs (1430 genes), or 10% if also genes from cluster C are included, followed this expression trend.

It is well recognised that senescence does not occur in young leaves but requires aging before it can start. Although it is not known how the plant manages to exclude senescence from young leaves, it is not astonishing that evolution has established tight control over this process; part of this control might rely on repressor functions, perhaps exerted by some of the TFs that exhibit high expression in L50 leaves and lowered expression at L100 and S20 stages. It remains to be tested whether any of the senescence down-regulated TFs indeed functions as a repressor of leaf senescence. In fact, at least one TF with senescence repressor functions has already been identified (*WRKY70*; see above), although this gene – in contrast to the senescence down-regulated genes discussed here – is expressed throughout leaf development and exhibits even higher expression in senescent leaves (Ülker *et al.* 2007).

In general, senescence down-regulated TF genes have not been intensively studied thus far. van der Graaff *et al.* (2006) observed a relatively large number of both senes-

Table 2. Statistical analysis of over-representation of TF families contributing to the senescence transcriptome.

		SAG				ESAG				SDG				ESDG				
family	M	n	OR	P	P_c	n	OR	P	P_c	n	OR	P	P_c	n	OR	P	P_c	others
ABI3VP1	17	0	0.00	1.00	1.00	0	0.00	1.00	1.00	1	0.50	0.87	1.00	2	3.76	0.12	1.00	14
AP2 ER	117	1	0.25	0.98	1.00	0	0.00	1.00	1.00	14	1.10	0.42	1.00	13	4.78	7.6×10^{-5}	1.9×10^{-3}*	89
ARP	50	0	0.00	1.00	1.00	0	0.00	1.00	1.00	10	2.10	0.04	0.34	2	1.14	0.54	1.00	38
ARR-B	16	0	0.00	1.00	1.00	0	0.00	1.00	1.00	1	0.53	0.85	1.00	0	0.00	1.00	1.00	15
bHLH	45	1	0.71	0.76	1.00	1	2.11	0.40	1.00	11	2.77	0.01	0.09	6	4.81	3.9×10^{-3}	0.05*	26
bZIP	57	1	0.55	0.84	1.00	1	1.63	0.48	1.00	1	0.14	1.00	1.00	0	0.00	1.00	1.00	54
C2H2	92	0	0.00	1.00	1.00	1	0.97	0.66	1.00	9	0.86	0.72	1.00	2	0.58	0.86	1.00	80
CCHAP2	10	1	3.61	0.27	1.00	0	0.00	1.00	1.00	0	0.00	1.00	1.00	0	0.00	1.00	1.00	9
CCHAP3	8	0	0.00	1.00	1.00	0	0.00	1.00	1.00	2	2.70	0.22	0.72	0	0.00	1.00	1.00	6
CCHAP5	9	0	0.00	1.00	1.00	0	0.00	1.00	1.00	3	4.09	0.07	0.42	0	0.00	1.00	1.00	6
CPP	8	0	0.00	1.00	1.00	0	0.00	1.00	1.00	1	1.15	0.61	1.00	0	0.00	1.00	1.00	7
DOF	31	0	0.00	1.00	1.00	1	3.13	0.30	1.00	5	1.57	0.25	0.72	0	0.00	1.00	1.00	25
E2F-DP	7	0	0.00	1.00	1.00	0	0.00	1.00	1.00	2	3.24	0.18	0.72	0	0.00	1.00	1.00	5
GATA	26	0	0.00	1.00	1.00	0	0.00	1.00	1.00	8	3.76	0.01	0.09	0	0.00	1.00	1.00	18
General	16	0	0.00	1.00	1.00	0	0.00	1.00	1.00	3	1.87	0.26	0.72	0	0.00	1.00	1.00	13
HB	72	2	0.90	0.66	1.00	1	1.27	0.57	1.00	8	1.00	0.56	1.00	0	0.00	1.00	1.00	61
HLH	8	0	0.00	1.00	1.00	0	0.00	1.00	1.00	1	1.15	0.61	1.00	0	0.00	1.00	1.00	7
HSF	17	1	2.02	0.41	1.00	0	0.00	1.00	1.00	2	1.07	0.58	1.00	0	0.00	1.00	1.00	14
MADS	70	0	0.00	1.00	1.00	0	0.00	1.00	1.00	4	0.47	0.96	1.00	0	0.00	1.00	1.00	66
MYB	156	6	1.34	0.34	1.00	0	0.00	1.00	1.00	13	0.69	0.91	1.00	7	1.34	0.31	1.00	130
MYB-L	24	1	1.39	0.53	1.00	0	0.00	1.00	1.00	2	0.72	0.77	1.00	0	0.00	1.00	1.00	21
NAC	66	12	11.03	1.0×10^{-7}	2.6×10^{-6}*	4	8.34	0.00	0.11	5	0.64	0.88	1.00	2	0.84	0.70	1.00	43
SBP	10	0	0.00	1.00	1.00	0	0.00	1.00	1.00	1	0.89	0.69	1.00	0	0.00	1.00	1.00	9
WRKY	47	4	3.25	0.05	0.63	2	4.56	0.09	1.00	1	0.17	1.00	1.00	1	0.58	0.83	1.00	39
YABBY	4	0	0.00	1.00	1.00	0	0.00	1.00	1.00	1	2.68	0.17	0.72	0	0.00	1.00	1.00	3

M = total number of genes in each family, irrespective of their behaviour in the senescence transcriptome; n = number of genes in each family in the respective expression group. OR: conditional maximum likelihood estimate of the odds ratio; P = P-value from Fisher Exact Test; P_c: P-value after correcting for multiple testing using the Benjamini–Hochberg approach for the control of the False Discovery Rate (FDR; Benjamini & Hochberg 1995). Evaluation of the association of TF families with the expression groups SAG, ESAG, SDG and ESDG. Association was evaluated by means of the Fisher Exact Test on 2 × 2 contingency tables. The total number of genes used was 983. All statistics were computed in the statistical package R (R Development Core Team, 2007). *, highlights P_c-values with FDR ≤ 0.05.

cence-induced and -repressed plant-specific TF genes (which have no relatives in other organisms). However, their data are not directly comparable with our dataset, as largely different experimental setups were used in the two studies. Whereas we devoted our analysis to a rather narrow window of developmental stages of naturally regulated senescence (L50, L100 and S20 leaves), van der Graaff and colleagues chose to investigate a much broader spectrum of stages, including leaves with a much more progressed senescence phenotype (*i.e.* 50% and 75% yellow leaf surface), with the oldest plants being in the silique ripening phase. Nevertheless, even with these divergent experimental conditions, we found 32 TF genes to be commonly up-regulated in both studies. In contrast, only 15 genes were found commonly down-regulated in the two analyses. Notably, although six different leaf stages were analysed by van der Graaff *et al.* (2006), only 79 senescence down-regulated TF genes were discovered in total, whereas we observed 144 by comparing three leaf stages (see above). The lower number of TFs discovered in the former study probably reflects the lower sensitivity of microarray-based expression platforms in comparison to qRT-PCR (Czechowski *et al.* 2004).

SENESCENCE AND ABIOTIC STRESS

Developmentally-regulated senescence is assumed to play an important role for nutrient recycling, supporting the formation of reproductive organs (seeds). Therefore, to maximise seed production, and hence reproductive fitness, disintegration of leaf tissue for the supply of nutrients has to be balanced against the already existing leaf biomass. Under optimal growth conditions, in the absence of longer-lasting external stress, initiation of leaf senescence is dependent on age and developmental stage, and under stable environmental conditions is relatively constant and predictable (Hensel *et al.* 1993; Nooden & Penny 2001). However, it is well known that environmental stresses can induce precocious senescence, including energy deprivation, darkness, excess light, drought, salinity, nutrient limitation and wounding (*e.g.* Whitehead *et al.* 1984; Becker & Apel 1993; Lutts *et al.* 1996; Bucha-

nan-Wollaston *et al.* 2005; Munns 2005). In rice leaves, for example, it has been proposed that many salt stress-triggered processes, such as a decline in photosynthetic activity or an increase in membrane damage, reflect a hastening of the naturally occurring senescence process (Dwidedi *et al.* 1979; Dhindsa *et al.* 1981). Many of the genes changing their expression during leaf senescence are also known to be affected by environmental stresses, both abiotic and biotic in nature, indicating at least a partial disconnection from the age-dependent senescence pathway. However, from an evolutionary perspective it would appear disadvantageous if intermittent or short-term stresses induce leaf senescence. This might be particularly harmful in the case of TFs as they regulate a whole suite of downstream target genes that, once affected, might be difficult to reset through cellular mechanisms to the original ′stress-free′ status.

Here, we examined the effect of abiotic stresses, in particular drought, salt stress and wounding, on the expression of senescence-related TFs detected through qRT-PCR analysis. Table 3 provides a list of TFs responding to at least one of the three stresses. The data shown were extracted from microarray studies using the Response Viewer tool of the GENEVESTIGATOR database (Zimmermann *et al.* 2004). We observed that approximately 30% of the senescence-related TF genes identified by our study (*i.e.* 52 out of 185 TF genes) also responded to at least one type of abiotic stress. Of the stresses analysed, salt stress appeared to have the most prominent effect on most of the TF genes (Table 3). For example, almost all of the 18 abiotic stress-responsive SAGs respond more strongly to salt stress than to drought or wounding. In various cases (*e.g.* HSF At5g43840, MYBs At2g47190 and At1g66390, NACs At1g52890, At3g15500 and At4g27410) induction by salt stress was severe (18- to 34-fold), whereas drought or wounding affected expression of SAGs generally not much more than twofold. Similarly, many of the senescence down-regulated TFs responded more strongly to salt stress than to drought and wounding. We conclude that TFs play a prominent role in salt stress-induced plant senescence. Commonalities in the molecular expression signatures were also observed for genes induced by salt stress and during senescence in rice (Chao *et al.* 2005), supporting the conclusion of at least partly shared response pathways.

Table 3. Effect of abiotic stresses on the expression levels of senescence-regulated TF genes.

locus	family	drought	salt	wounding
SAG				
At5g61890	AP2ER	1.27	7.08	1.38
At2g46680	HB	1.12	7.3	0.98
At3g61890	HB	1.43	9.21	1.29
At5g43840	HSF	1.11	33.44	1.22
At1g48000	MYB	2	5.08	1.4
At1g66390	MYB	1.55	26.36	2.45
At2g47190	MYB	1.59	18.3	1.74
At5g54230	MYB	0.75	5.21	1.11
At1g13300	MYB-L	0.8	0.47	1.09
At1g52890	NAC	3.2	34.12	2.72
At1g69490	NAC	1.79	5.06	1.72
At3g04070	NAC	1.22	2.74	1.54
At3g15500	NAC	1.68	22.68	2.39
At4g27410	NAC	1.28	18.31	1.64
At5g39610	NAC	1.87	5.85	2.02
At1g29860	WRKY	0.92	4.28	0.83
At3g01970	WRKY	2.41	2.86	2.22
At5g13080	WRKY	1.55	6.7	2.07
ESAG				
At2g44910	HB	0.73	2.43	0.91
At2g43000	NAC	1.47	2.86	1.26
At4g23810	WRKY	1.1	12.52	1.59
At5g07100	WRKY	1.14	1.56	2.08
SDG				
At1g21910	AP2ER	0.84	6.36	1.41
At4g17490	AP2ER	1.25	24.39	1.73
At4g32800	AP2ER	1.6	3.55	1.1
At4g37750	AP2ER	0.75	0.45	0.81
At5g25390	AP2ER	0.97	5.77	0.68
At5g25810	AP2ER	0.67	0.32	1
At3g15540	ARP	0.86	3.02	0.77
At4g29080	ARP	1.13	0.37	1.21
At2g13570	CCHAP3	0.48	0.55	0.91
At4g14540	CCHAP3	0.9	0.45	0.83
At3g01330	E2F-DP	0.82	0.47	0.81
At5g26870	MADS	1.5	6.5	1.5
At1g75250	MYB	0.8	0.08	0.54
At3g50060	MYB	0.89	12.16	1.1
At2g45190	YABBY	0.95	0.49	0.79
At5g63780		0.81	0.28	0.78
ESDG				
At1g12610	AP2ER	0.88	28.24	3.24
At1g43160	AP2ER	2.72	28.16	3.01
At2g44840	AP2ER	1.12	140.4	2.32
At4g34410	AP2ER	1.32	95.17	3.81
At5g13330	AP2ER	1.55	2.99	1.54
At5g51990	AP2ER	0.86	5.62	1.87
At3g62100	ARP	1.07	9.06	1.03
At5g46830	bHLH	0.95	2.23	0.42
At5g04340	C2H2	1.71	4.45	1.28
At1g56650	MYB	1.71	7.1	1.27
At2g31180	MYB	1.39	2.1	1.31
At3g06490	MYB	1.3	16.55	1.48

SUMMARY AND OUTLOOK

A large number of senescence-activated and senescence down-regulated TF genes have been discovered using various technologies in recent years, including microarray-based expression profiling and suppression subtractive hybridisation in previous studies, and qRT-PCR in this report. Thus, it appears that TFs play a prominent role in controlling leaf senescence. However, until today, functional studies on senescence-regulated TF genes are very limited, and knowledge about their integration into

Table 3. Continued.

locus	family	drought	salt	wounding
At4g05100	MYB	0.99	5.76	1.14
At4g21440	MYB	1.13	6.76	1.09

TF genes undergoing expression changes during developmentally-regulated senescence (see Table 1) were analysed for abiotic stress-dependent expression changes using public microarray data. Data were retrieved through the Response Viewer tool of the GENEVESTIGATOR micro-array database (Zimmermann *et al.* 2004). Ratios are calculated as treatment *versus* control, averaging expression data of 6, 12 and 24 h of shoots and roots. Only genes for which a significant abiotic stress effect could be detected (at least twofold expression change in at least one of the stresses analysed) are listed.

molecular networks is vague at best. It appears that transcriptional control occurring at different phases of leaf development is an enormously complex but fascinating phenomenon that will reveal its secrets only with continued research. Importantly, work on senescence control can also be expected to benefit plant cultivation in an agricultural setting, as recently suggested by work on wheat senescence (Uauy *et al.* 2006). As more and more plant genome sequences become available, and genomic technologies with ever increasing throughput and sensitivity contribute to data collection and analysis, we can expect that research addressing the molecular wirings of senescence control circuits will see major leaps forward in the near future.

ACKNOWLEDGEMENTS

We also thank Aleksandra Skirycz, Camila Caldana, Matthew Hannah and Armin Schlereth from the MPI of Molecular Plant Physiology for their support while running experiments and performing data analyses.

CONFLICTS OF INTEREST

This work was supported by a research grant from the BMBF (German Federal Ministry of Education and Research; FKZ 0312854). Salma Balazadeh is member of the International PhD Programme 'Integrative Plant Science' (IPP-IPS) funded by the DAAD (Deutscher Akademischer Austauschdienst) and the DFG (Deutsche Forschungsgemeinschaft) under DAAD No. D/04/01336. Bernd Mueller-Roeber thanks the Fonds der Chemischen Industrie for funding (No. 0164389). Funding through the BMBF for financial support of Diego Mauricio Riaño-Pachón (GABI-Future grant 0315046) is greatly acknowledged.

REFERENCES

Aida M., Ishida T., Fukaki H., Fujisawa H., Tasaka M. (1997) Genes involved in organ separation in *Arabidopsis*: an analysis of the cup-shaped cotyledon mutant. *Plant Cell*, **9**, 841–857.

Andersson A., Keskitalo J., Sjödin A., Bhalerao R., Sterky F., Wissel K., Tandre K., Aspeborg H., Moyle R., Ohmiya Y., Bhalerao R., Brunner A., Gustafsson P., Karlsson J., Lundeberg J., Nilsson O., Sandberg G., Strauss S., Sundberg B., Uhlen M., Jansson S., Nilsson P. (2004) A transcriptional timetable of autumn senescence. *Genome Biology*, **5**, R24.

Becker W., Apel K. (1993) Differences in gene expression between natural and artificially induced leaf senescence. *Planta*, **189**, 74–79.

Benjamini Y., Hochberg Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, **57**, 289–300.

Bi Y.M., Zhang Y., Signorelli T., Zhao R., Zhu T., Rothstein S. (2005) Genetic analysis of *Arabidopsis* GATA transcription factor gene family reveals a nitrate-inducible member important for chlorophyll synthesis and glucose sensitivity. *The Plant Journal*, **44**, 680–692.

Buchanan-Wollaston V. (1997) The molecular biology of leaf senescence. *Journal of Experimental Botany*, **48**, 181–199.

Buchanan-Wollaston V., Page T., Harrison E., Breeze E., Lim P.O., Nam H.G., Lin J.F., Wu S.H., Swidzinski J., Ishizaki K., Leaver C.J. (2005) Comparative transcriptome analysis reveals significant differences in gene expression and signalling pathways between developmental and dark/starvation-induced senescence in *Arabidopsis*. *The Plant Journal*, **42**, 567–585.

Buchanan-Wollaston V., Earl S., Harrison E., Mathas E., Navabpour S., Page T., Pink D. (2003) The molecular analysis of leaf senescence – a genomic approach. *Plant Biotechnology Journal*, **1**, 3–22.

Caldana C., Scheible W.-R., Mueller-Roeber B., Ruzicic S. (2007) A quantitative RT-PCR platform for high-throughput expression profiling of 2500 rice transcription factors. *Plant Methods*, **3**, 7.

Chao D.Y., Luo Y.H., Shi M., Luo D., Lin H.X. (2005) Salt-responsive genes in rice revealed by cDNA mircoarray analysis. *Cell Research*, **15**, 796–810.

Chen W., Provart N.J., Glazebrook J., Katagiri F., Chang H.S., Eulgem T., Mauch F., Luan S., Zou G., Whitham S.A., Budworth P.R., Tao Y., Xie Z., Chen X., Lam S., Kreps J.A., Harper J.F., Si-Ammour A., Mauch-Mani B., Heinlein M., Kobayashi K., Hohn T., Dangl J.L., Wang X., Zhu T. (2002) Expression profile matrix of *Arabidopsis* transcription factor genes suggests their putative functions in response to environmental stresses. *Plant Cell*, **14**, 559–574.

Chevalier F., Perazza D., Laporte F., Le Henanff G., Hornitschek P., Bonneville J.M., Herzog M., Vachon G. (2008) GeBP and GeBP-like proteins are non-canonical leucine-zipper transcription factors that regulate cytokinin response in *Arabidopsis thaliana*. *Plant Physiology*, **146**, 1142–1154.

Collinge M., Boller T. (2001) Differential induction of two potato genes, *Stprx2* and *StNAC*, in response to infection by *Phytophthora infestans* and to wounding. *Plant Molecular Biology*, **46**, 521–529.

Crafts-Brandner S.J., Holzer R., Feller U. (1998) Influence of nitrogen deficiency on senescence and the amounts of RNA and proteins in wheat leaves. *Physiologia Plantarum*, **102**, 192–200.

Czechowski T., Bari R.P., Stitt M., Scheible W.R., Udvardi M.K. (2004) Real-time RT-PCR profiling of over 1400 *Arabidopsis* transcription factors: unprecedented sensitivity reveals novel root- and shoot-specific genes. *The Plant Journal*, **38**, 366–379.

Davuluri R.V., Sun H., Palaniswamy S.K., Matthews N., Molina C., Kurtz M., Grotewold E. (2003) AGRIS: *Arabidopsis* Gene Regulatory Information Server, an information resource of *Arabidopsis cis*-regulatory elements and transcription factors. *BMC Bioinformatics*, **4**, 25.

Dhindsa R.S., Plumb-Dhindsa P., Thorpe T.A. (1981) Leaf senescence: correlated with increased level of membrane permeability and lipid peroxidation, and decreased levels of superoxide dismutase and catalase. *Journal of Experimental Botany*, **32**, 93–101.

Diaz C., Saliba-Colombani V., Loudet O., Belluomo P., Moreau L., Daniel-Vedele F., Morot-Gaudry J.F., Masclaux-Daubresse C. (2006) Leaf yellowing and anthocyanin accumulation are two genetically independent strategies in response to nitrogen limitation in *Arabidopsis thaliana*. *Plant Cell Physiology*, **47**, 74–83.

Dwidedi S., Kar M., Mishra D. (1979) Biochemical changes in excised leaves of *Oryza sativa* subjected to water stress. *Physiologia Plantarum*, **45**, 35–40.

Ellis C.M., Nagpal P., Young J.C., Hagen G., Guilfoyle T.J., Reed J.W. (2005) *AUXIN RESPONSE FACTOR1* and *AUXIN RESPONSE FACTOR2* regulate senescence and floral organ abscission in *Arabidopsis thaliana*. *Development*, **132**, 4563–4574.

Eulgem T., Somssich I.E. (2007) Networks of WRKY transcription factors in defense signaling. *Current Opinion in Plant Biology*, **10**, 366–371.

Fujita M., Fujita Y., Maruyama K., Seki M., Hiratsu K., Ohme-Takagi M., Tran L.S., Yamaguchi-Shinozaki K., Shinozaki K. (2004) A dehydration-induced NAC protein, RD26, is involved in a novel ABA-dependent stress-signaling pathway. *The Plant Journal*, **39**, 863–876.

Gepstein S., Sabehi G., Carp M.J., Hajouj T., Nesher M.F., Yariv I., Dor C., Bassani M. (2003) Large-scale identification of leaf senescence-associated genes. *The Plant Journal*, **36**, 629–642.

van der Graaff E., Schwacke R., Schneider A., Desimone M., Flügge U.-I., Kunze R. (2006) Transcription analysis of *Arabidopsis* membrane transporters and hormone pathways during developmental and induced leaf senescence. *Plant Physiology*, **141**, 776–792.

Gregersen P.L., Holm P.B. (2007) Transcriptome analysis of senescence in the flag leaf of wheat (*Triticum aestivum* L.). *Plant Biotechnology Journal*, **5**, 192–206.

Guo Y., Gan S. (2006) AtNAP, a NAC family transcription factor, has an important role in leaf senescence. *The Plant Journal*, **46**, 601–612.

Guo Y., Cai Z., Gan S. (2004) Transcriptome of *Arabidopsis* leaf senescence. *Plant Cell and Environment*, **27**, 521–549.

He X.J., Mu R.L., Cao W.H., Zhang Z.G., Zhang J.S., Chen S.Y. (2005) AtNAC2, a transcription factor downstream of ethylene and auxin signaling pathways, is involved in salt stress response and lateral root development. *The Plant Journal*, **44**, 903–916.

Hensel L.L., Gribic V., Baumgarten D.A., Bleecker A.B. (1993) Developmental and age-related processes that influence the longevity and senescence of photosynthetic tissues in *Arabidopsis*. *Plant Cell*, **5**, 553–564.

Hinderhofer K., Zentgraf U. (2001) Identification of a transcription factor specifically expressed at the onset of leaf senescence. *Planta*, **213**, 469–473.

Holland M.J. (2002) Transcript abundance in yeast varies over six orders of magnitude. *Journal of Biological Chemistry*, **277**, 14363–14366.

Horak C.E., Snyder M. (2002) Global analysis of gene expression in yeast. *Functional & Integrative Genomics*, **2**, 171–180.

Hörtensteiner S., Feller U. (2002) Nitrogen metabolism and remobilisation during senescence. *Journal of Experimental Botany*, **53**, 927–937.

Hwang I., Sheen J. (2001) Two-component circuitry in *Arabidopsis* cytokinin signal transduction. *Nature*, **413**, 383–389.

Kim Y.S., Kim S.G., Park J.E., Park H.Y., Lim M.H., Chua N.H., Park C.M. (2006a) A membrane-bound NAC transcription factor regulates cell division in *Arabidopsis*. *Plant Cell*, **18**, 3132–3144.

Kim H.J., Ryu H., Hong S.H., Woo H.R., Lim P.O., Lee I.C., Sheen J., Nam H.G., Hwang I. (2006b) Cytokinin-mediated control of leaf longevity by AHK3 through phosphorylation of ARR2 in *Arabidopsis*. *Proceedings of the National Academy of Sciences USA*, **103**, 814–819.

Lim P.O., Kim Y., Breeze E., Koo J.C., Woo H.R., Ryu J.S., Park D.H., Beynon J., Tabrett A., Buchanan-Wollaston V., Nam H.G. (2007) Overexpression of a chromatin architecture-controlling AT-hook protein extends leaf longevity and increases the post-harvest storage life of plants. *The Plant Journal*, **52**, 1140–1153.

Lin J.F., Wu S.H. (2004) Molecular events in senescing *Arabidopsis* leaves. *The Plant Journal*, **39**, 612–628.

Lutts S., Kinet J.M., Bouharmont J. (1996) NaCl-induced senescence in leaves of rice (*Oryza sativa* L.) cultivars differing in salinity resistance. *Annals of Botany*, **78**, 389–398.

Miao Y., Zentgraf U. (2007) The antagonist function of *Arabidopsis* WRKY53 and ESR/ESP in leaf senescence is modulated by the jasmonic and salicylic acid equilibrium. *Plant Cell*, **19**, 819–830.

Miao Y., Laun T.M., Zimmerman P., Zentgraf U. (2004) Targets of WRKY53 transcription factor and its role during leaf senescence in *Arabidopsis*. *Plant Molecular Biology*, **55**, 853–867.

Miao Y., Laun T.M., Smykowski A., Zentgraf U. (2007) *Arabidopsis* MEKK1 can take a short cut: it can directly interact with senescence-related WRKY53 transcription factor on the

protein level and can bind to its promoter. *Plant Molecular Biology*, **65**, 63–76.

Munns R. (2005) Genes and salt tolerance: bringing them together. *New Phytologist*, **167**, 645–663.

Nakashima K., Tran L.S., van Nguyen D., Fujita M., Maruyama K., Todaka D., Ito Y., Hayashi N., Shinozaki K., Yamaguchi-Shinozaki K. (2007) Functional analysis of a NAC-type transcription factor OsNAC6 involved in abiotic and biotic stress-responsive gene expression in rice. *The Plant Journal*, **51**, 617–630.

Noh Y.-S., Amasino R.M. (1999) Identification of a promoter region responsible for the senescence-specific expression of *SAG12*. *Plant Molecular Biology*, **41**, 181–194.

Nooden L.D., Penny J.P. (2001) Correlative controls of senescence and plant death in *Arabidopsis thaliana* (Brassicaceae). *Journal of Experimental Botany*, **52**, 2151–2159.

Okushima Y., Mitina I., Quach H.L., Theologis A. (2005) AUXIN RESPONSE FACTOR 2 (ARF2): a pleiotropic developmental regulator. *The Plant Journal*, **43**, 29–46.

Pfaffl M.W., Daxenberger A., Hageleit M., Meyer H.H.D. (2002) Effects of synthetic progestagens on the mRNA expression of androgen receptor, progesterone receptor, oestrogen receptor alpha and beta, insulin-like growth factor-1 (IGF-1) and IGF-1 receptor in heifer tissues. *Journal of Veterinary Medicine Series A – Physiology, Pathology, Clinical Medicine*, **49**, 57–64.

Pourtau N., Jennings R., Pelzer E., Pallas J., Wingler A. (2006) Effect of sugar-induced senescence on gene expression and implications for the regulation of senescence in *Arabidopsis*. *Planta*, **224**, 556–568.

Quirino B.F., Noh Y.-S., Himelblau E., Amasino R.M. (2000) Molecular aspects of leaf senescence. *Trends in Plant Sciences*, **5**, 278–282.

R Development Core Team (2007) R: A language and environment for statistical computing. *R Foundation for Statistical Computing*. Vienna, Austria. http://www.R-project.org.

Riaño-Pachón D.M., Ruzicic S., Dreyer I., Mueller-Roeber B. (2007) PlnTFDB: an integrative plant transcription factor database. *BMC Bioinformatics*, **8**, 42.

Riechmann J.L., Heard J., Martin G., Reuber L., Jiang C., Keddie J., Adam L., Pineda O., Ratcliffe O.J., Samaha R.R., Creelman R., Pilgrim M., Broun P., Zhang J.Z., Ghandehari D., Sherman B.K., Yu G. (2000) *Arabidopsis* transcription factors: genome-wide comparative analysis among eukaryotes. *Science*, **290**, 2105–2110.

Robatzek S., Somssich I.E. (2001) A new member of the *Arabidopsis* WRKY transcription factor family, AtWRKY6, is associated with both senescence- and defence-related processes. *The Plant Journal*, **28**, 123–133.

Robatzek S., Somssich I.E. (2002) Targets of AtWRKY6 regulation during plant senescence and pathogen defense. *Genes and Development*, **16**, 1139–1149.

Saeed A.I., Sharov V., White J., Li J., Liang W., Bhagabati N., Braisted J., Klapa M., Currier T., Thiagarajan M., Sturn A., Snuffin M., Rezantsev A., Popov D., Ryltsov A., Kostukovich E., Borisovsky I., Liu Z., Vinsavich A., Trush V., Quackenbush J. (2003) TM4: a free, open-source system for microarray data management and analysis. *Biotechniques*, **34**, 374–378.

Thomson W.W., Platt-Aloia K.A. (1987) Ultrastructure and senescence in plants. In: Thomson W.W., Nothnagel E.A., Huftaker R.C. (Eds), *Plant Science: Biochemistry and Physiology*. The American Society of Plant Physiologists, Rockville, MD: p. 20–30.

Uauy C., Distelfeld A., Fahima T., Blechl A., Dubcovsky J. (2006) An *NAC* gene regulating senescence improves grain protein, zinc and iron content in wheat. *Science*, **314**, 1298–1301.

Ülker B., Shahid Mukhtar M., Somssich I.E. (2007) The WRKY70 transcription factor of *Arabidopsis* influences both the plant senescence and defense signalling pathways. *Planta*, **226**, 125–137.

Weaver L.M., Amasino R.M. (2001) Senescence is induced in individually darkened *Arabidopsis* leaves, but inhibited in whole darkened plants. *Plant Physiology*, **127**, 876–886.

Whitehead Ch.S., Halevy A.H, Reid M.S. (1984) Roles of ethylene and 1-aminocyclopropane-1-carboxylic acid in pollination and wound-induced senescence of *Petunia hybrida* flowers. *Physiologia Plantarum*, **61**, 643–648.

Wingler A., Purdy S., MacLean J.A., Pourtau N. (2006) The role of sugars in integrating environmental signals during the regulation of leaf senescence. *Journal of Experimental Botany*, **57**, 391–399.

Woolhouse H.W. (1984) Senescence in plants cells. In: Davies I., Sigee D.C. (Eds), *Cell aging and cell death*. Cambridge University Press, Cambridge: pp. 123–153.

Zhong R., Demura T., Ye Z.H. (2006) SND1, a NAC domain transcription factor, is a key regulator of secondary wall synthesis in fibers of *Arabidopsis*. *Plant Cell*, **18**, 3158–3170.

Zimmermann P., Hirsch-Hoffmann M., Hennig L., Gruissem W. (2004) GENEVESTIGATOR. Arabidopsis microarray database and analysis toolbox. *Plant Physiology*, **136**, 2621–2632.

6

General discussion and outlook

6.1 Genome annotation

The focus of this thesis was the identification of putative complete sets of TFs and TRs encoded by the genome of fully sequenced and annotated plant species. The first goal, identification and classification of TFs, is an effort of gene annotation in completed genomes. By establishing PlnTFDB, we made available putative complete sets of TF and TR families of several plant species, i.e., *C. merolae*, *O. tauri*, *C. reinhardtii*, *P. patens*, *A. thaliana*, *P. trichocarpa* and *O. sativa* (Fig. 6.1). Additionally, at the time of writing we are annotating the TF and TR genes in the spikemoss *Selaginella moellendorffii*. These data together with those of other species, e.g., *Vitis vinifera*, are to be included into PlnTFDB in the near future.

The accurate identification of complete TF and TR families presented in this thesis is subject to three factors. The quality of the genome sequence, the quality of gene annotation, and the ability of current profile-HMMs to identify remote homologues in a broad phylogenetic range of plant species.

The plant genomes that served as the basis of this study have been sequenced at different times (see Fig. 1.2). The quality of earlier sequenced genomes is usually higher, as most of the gaps in the genome sequence had been filled in the meantime. Newer genome sequences, i.e., that from *Chlamydomonas reinhardtii*, currently have variable lengths of sequence gaps that further rounds of genome sequencing promise to fill in. The existence of gaps might increase the rate of false negatives, i.e., a TF might be encoded by the genome but may not be identified if it is located in a region that has not been sequenced so far.

The quality of the genome annotation, after the quality of the genome sequence itself, is a crucial point in the identification of complete gene families. Different computational

approaches have been developed for the prediction of protein coding genes in eukaryotic organisms, exploiting sequence biases characteristic of protein coding regions and the presence of splice-site signals, among others (for reviews see DAVULURI and ZHANG 2003, DO and CHOI 2006, ZHANG 2002). Additional computational approaches are available for the prediction of non-coding RNAs (e.g., tRNA; LOWE and EDDY 1997). Moreover experimental evidence, e.g., ESTs, and manual curation of the predicted gene models are very important resources in any genome project, as they can uncover or provide support for genes that might have been missed by automated approaches (e.g., RIAÑO PACHÓN *et al.* 2005). With this work we contributed actively to the genome annotation of the green algae *C. reinhardtii*; and we are currently participating in the genome annotation of the spikemoss *Selaginella moellendorffii*. Furthermore, the rules for the identification of TFs and TRs, and their classification into families, as well as the existing classifications available in PlnTFDB, have been used to elucidate the TF and TR complements in different species, e.g., one strain of the grapevine *Vitis vinifera* (VELASCO *et al.* 2007).

The last crucial factor for the identification of complete sets of TFs and TRs is the ability of the current domain models (profile-HMMs; from the PFAM database) to detect all remote homologues in plants. On one hand, for some TF families, their characteristic DBD might not be represented in the collection of profile-HMMs; this was actually the case for the families: CCAAT-HAP3, CCAAT-HAP5, CCAAT-DR1, DBP, LUG, G2-like, GRF, HRT, NOZZLE, Trihelix, ULT, VOZ and Whirly. For them we have created new profile-HMM models. For two of these families (Whirly and GRF) new models have been recently included in the PFAM collection (Whirly: PF08536; GRF is characterized by two domains, WRC: PF08879 and QLQ: PF08880). On the other hand, as most of the current profile-HMMs have been trained with non-plant sequences, or only including *A. thaliana* or other angiosperms species, atypical family members in other groups of plants as algae and mosses could be missed. New plant-specific models can be trained with the member sequences that have been recovered so far encompassing a broad phylogenetic range and increasing their likelihood to detect all family members. The NOZZLE TF family, that appears to be derived from MADS-box TFs (WILSON and YANG 2004), is an example of this approach. Currently, the only identified member of this family is found in Arabidopsis; it was detected by a model derived from Arabidopsis sequences. After the availability of the poplar genome a putative orthologue has been found in this genome (PEP ID: 568986; 37% sequence identity). It also appears to be present in tobacco[1], which will make this an eudicot-specific family, among the angiosperm clade. Another example

[1] `http://compsysbio.achs.virginia.edu/tobfac/browse_family.pl?family=NZZ`

is found in the bZIP TF family, where some members of this family (e.g. AT4G35900) are not detected by the current PFAM model. In order to carry out the study presented in Chapter 4 we performed iterative tblastn and blastx searches, in addition to searches with the current PFAM bZIP HMMs, and manual curation to identify the complete sets of bZIPs in green plants.

With the information currently available, new profile-HMMs can be devised encompassing broad phylogenetic ranges in the plant kingdom, improving their sensitivity on the identification of green TFs and TRs.

6.2 Comparative genomic analyses of TF families in plants

In the current version of PlnTFDB[2], up to 57 TF and 11 TR families can be identified, which are among the most numerous transcription regulatory families in plants (summarised in Table 6.1). Further families will be added in the near future, e.g., mTERF, a mitochondrial transcription termination factor that appears to be present in all eukaryotes (FERNANDEZ-SILVA *et al.* 1997); and the VARL family where the *regA* gene cluster is involved in the control of cell differentiation in green algae (DUNCAN *et al.* 2007).

Table 6.1: *Updated numbers of TFs and TF families in plant species. P_{TOTAL}: Total number of proteins encoded by the genome, TFs: number of transcription factors and other transcriptional regulators, in parenthesis the number of distinct proteins, TF_{FAM}: Number of TF and other transcriptional regulator families identified, %TF: Number of TFs per 100 non-TF genes.*

Species	P_{TOTAL}	*TFs*	TF_{FAM}	*%TF*
C. merolae	5014 (5002)	130 (130)	27	2.7
O. tauri	7725 (7715)	183 (182)	36	2.4
C. reinhardtii	15143 (14920)	248 (246)	40	1.7
P. patens	35938 (35597)	1274 (1264)	59	3.7
A. thaliana	31921 (29988)	2437 (2250)	68	8.3
P. trichocarpa	45555 (44922)	2758 (2732)	66	6.4
O. sativa	66710 (62742)	2798 (2527)	65	4.4

The broad phylogenetic coverage in PlnTFDB, has allowed the delineation of lineage-specific regulatory families (see Chapter 3, and Figs. 6.1 and 6.2). Some of which are restricted to plants and were present in the MRCA of Archaeplastida, e.g., PLATZ and RWP-RK. Furthermore, families with a narrower phylogenetic distribution could be identified, and are thought to be involved in processes restricted to these clades (see

[2] `http://plntfdb.bio.uni-potsdam.de/v2.0/`

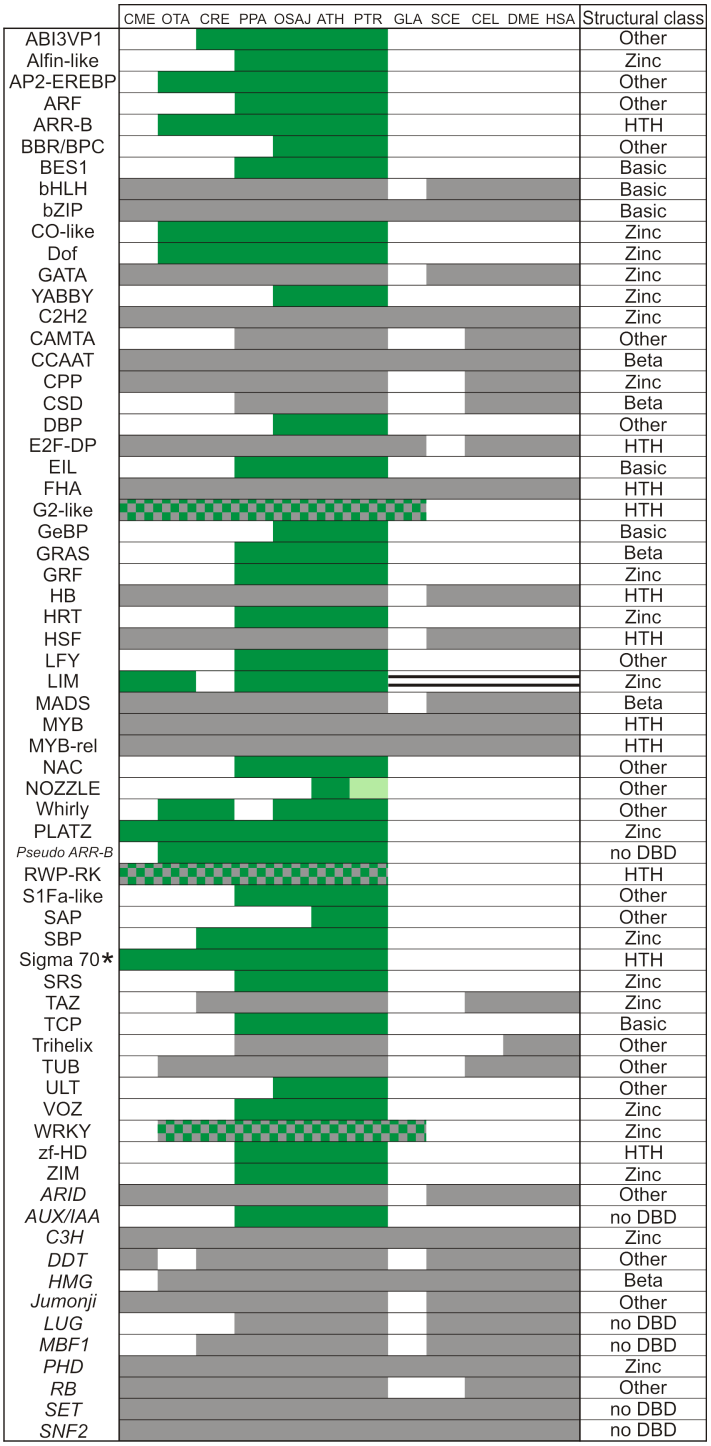


Figure 6.1: *Phylogenetic profile of TFs and TRs in photosynthetic and nonphotosynthetic eukaryotes. Dashed cells indicate that the identification of the family was not attempted. Families in light green are thought to be present in the species, see text. *The Sigma70-like TF family originated in bacteria, and appears in photosynthetic eukaryotes by means of the ancient primary endosymbiosis. Double-coloured cells indicate that the family is largely restricted to plants, i.e., absent in animals and fungi, but present in an early diverging eukaryotic clade, e.g, G. lamblia. The last column indicates the structural class of the characteristic DNA binding domain of each family (see Section 1.3). Basic: Basic domain, HTH: Helix turn helix, Zinc: Zinc-coordinating domain, Beta: β-scaffold domains with minor groove contacts, no DBD: Does not have a DNA-binding domain.*

Figs. 6.1 and 6.2), e.g., the family SAP, involved in maintaining floral meristem identity and megasporogenesis. SAP achieves its regulatory roles genetically interacting with members of the MADS and AP2-EREBP TF families, AG and AP2 respectively (BYZOVA *et al.* 1999). With a single member in the angiosperms *A. thaliana* and *P. trichocarpa*, and in the fern ally *S. moellendorffii*, but absent in the monocot *O. sativa*, as well as in algae and mosses, the SAP family must have been present in the MRCA of tracheophytes, likely as a single-copy gene involved in megasporogenesis, and lost at some time in the lineage leading to rice. The loss in the rice lineage is surprising as single-gene families with one-to-one orthology relationships and involved in macromolecular complexes tend to be well conserved in order to keep strict stoichiometry (reviewed by KOONIN 2005), nevertheless it is also possible that the SAP member in rice is located in a genomic region that has not been sequenced so far. Another family with restricted phylogenetic range is NOZZLE, that appears in the lineage leading to eudicots (see above), is required for the initiation of sporogenesis (SCHIEFTHALER *et al.* 1999, WILSON and YANG 2004), i.e., early anther cell division and differentiation (reviewed by MA 2005), and carpel development (reviewed by DINNENY and YANOFSKY 2004).

As shown in Table 3.1, most TF and TR families differ in the number of members that can be identified in the different species; as described in Section 1.4, these differences can arise through the processes of gene duplication, gene loss and horizontal gene transfer, and they are the prime source for evolutionary change (MOORE and PURUGGANAN 2005). A clear example of gene loss, actually family loss, might be represented by the Whirly family in the moss (see Fig. 6.1). This family is present in the whole green lineage with the exception of *P. patens*. Whirly is a small TF family with a single member in both green algae and up to three members in angiosperms. Two alternative explanations can account for the lack of Whirly TFs in the moss. First, it can actually be present in the genome, but in a yet to be sequenced region. Second, the gene present in the MRCA of land plants was lost in the lineage leading to moss. The biological role of the Whirly TF in unicellular algae is unknown so far. In angiosperms Whirly is involved in pathogen response (reviewed by DESVEAUX *et al.* 2005), like members of the bZIP, AP2-EREBP, MYB and WRKY families that in contrast to Whirly are all present in the moss (reviewed by EULGEM 2005).

The families SBP, bHLH, SNF2, MADS, WRKY, HMG, AP2-EREBP and FHA significantly differ in size between algae and land plants. The SBP family of TFs is significantly larger in *C. reinhardtii*, compared to land plants, and appears to have been lost in the prasinophyte *O. tauri*. So far, only a single SBP from *C. reinhardtii* has been characterized, the *COPPER RESPONSE REGULATOR1* (*CRR1*) required for activating and repressing target genes of a copper- and hypoxia-sensing pathway (KROPAT *et al.* 2005).

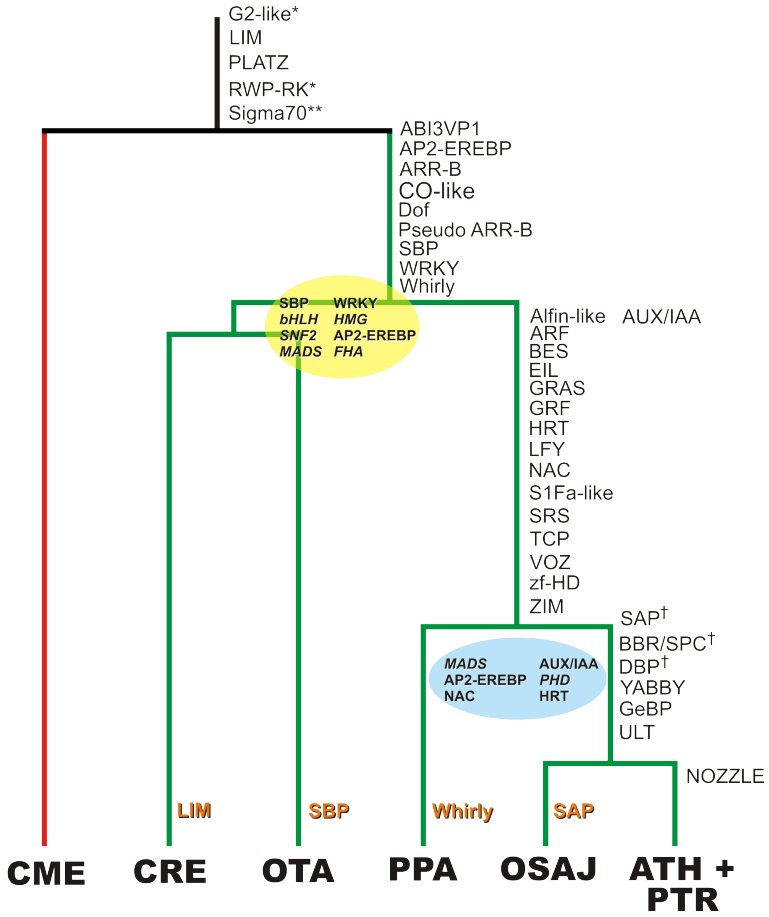


Figure 6.2: *Emergence of plant-specific TF families, and family bias among groups. Family names at the right side of the branches and in black denote the emergence of the family. Families that have been lost appear in orange. The yellow cloud indicates the families that significantly differ in size between algae and land plants. Families that significantly differ in size between seed plants and bryophytes appear in the blue cloud. Significant size differences were identified by a Fisher's exact test with FDR (q-value)* ≤ 0.01 *(*STOREY *and* TIBSHIRANI *2003). * Families that appear in early diverging non-photosynthetic eukaryotes, but that were lost in animals and fungi. ** Sigma70-like originates in bacteria; in eukaryotes it is restricted to photosynthetic clades.* † *Families that appear in the MRCA of tracheophytes, i.e., they are present in the spikemoss S. moellendorffii; the SAP family seems to have been lost in grasses. Families that are not plant-specific appear in italics. Species name abbreviation as in Table 3.1*

In land plants this family plays diverse roles e.g., leaf development, pathogen response and floral transition (reviewed by RIESE *et al.* 2007). The families bHLH, SNF2, MADS, WRKY, HMG, AP2-EREBP and FHA preferentially expanded with the colonisation of land, and might have played an important role in this great moment in evolution. They play a plethora of biological roles, e.g., regulation of the production of anthocyanin pigments, chromatin remodelling, regulation of development, responses to abiotic and biotic stresses and regulation of disease resistance pathways (EISEN *et al.* 1995, GUTTERSON and REUBER 2004, HEIM *et al.* 2003, NAM *et al.* 2003, SHIGYO *et al.* 2006, WU *et al.* 2005). Later, after the split of bryophytes and tracheophytes, the families MADS, AP2-

EREBP, NAC, AUX/IAA and PHD have significantly larger numbers in the lineage leading to seed plants, while HRT is significantly larger in the moss. MADS, AP2-EREBP, NAC and AUX/IAA are involved in the regulation of developmental programs (NAM *et al.* 2003, OLSEN *et al.* 2005, REED 2001, SHIGYO *et al.* 2006). HRT is involved in the response to the phytohormone gibberelin, and it is involved in development (RAVENTÓS *et al.* 1998). PHD is involved in chromatin remodelling and in response to cell stress (BIENZ 2006, SOLIMAN and RIABOWOL 2007).

Detailed phylogenetic analysis of TF families, i.e., phylogenetics and conserved protein motifs and intron positions, as the shown here for the bZIP TFs (see Chapter 4), lead to the identification of different clades or gene lineages inside the family. Similar to the families themselves, these clades arise at different stages of plant evolution (see Fig. 4.5), and are the result of sub- and neofunctionalisation. Their phylogenetic profile can be linked to great moments in plant evolution, i.e., the emergence of evolutionary novelties, e.g., land colonisation and seed formation.

6.3 Expression profiling of TF and TR families

The collection of TFs and TRs in PlnTFDB have been used to carry out genome-wide expression profiling experiments in order to assess the role of this genes in different processes.

The set of transcriptional regulators in rice described in PlnTFDB was used to develop a quantitative reverse transcription-polymerase chain reaction resource (qRT-PCR) (CALDANA *et al.* 2007), in order to track the response of TF genes under abiotic stresses, i.e., salt and drought (CALDANA *et al.* 2006, RUZICIC *et al.* 2005).

A similar qRT-PCR resource for Arabidopsis was also established before the release of PlnTFDB by a group from the Max Planck Institute of Molecular Plant Physiology (CZECHOWSKI *et al.* 2004). We have employed this resource to uncover TFs playing important roles in leaf development, i.e., the transition that undergoes a leaf from a net carbon importer (sink) to a net carbon exporter (source) known as the sink-to-source transition (CORREA *et al.* 2006) and the transition that a leaf undergoes at the end of its lifetime, where nutrients are redistributed to other organs known as leaf senescence (see Chapter 5; BALAZADEH *et al.* 2008).

In the genome-wide analysis of TFs involved in leaf senescence we have shown that the NAC family plays a preferential role in this developmental process. And members of AP2-EREBP and bHLH TF families are preferentially important in the early stages of senescence. Most of the TFs differentially expressed were down-regulated, which was explained as a, "... general reduction of the leaf maintenance machinery rather than being

an active part of the senescence regulation network itself". However, senescence-specific pathways can be 'turned on' by the down-regulation of key regulators. A clear example, from sea urchin embryo development, is the use of double-negative logic gates, as the one describe by OLIVERI *et al.* (2008), where the inactivation of a repressor leads to the activation of the pathway.

The use of such resources, as well as microarray platforms, will allow to uncover the topological features of gene regulatory networks, that in turn will help in the development of new hypothesis about the underlying regulatory logic.

6.4 Further resources for transcription factors

In addition to PlnTFDB, some additional resources for plant transcription factors in different organisms are available. Here is a list of the currently available databases providing information about TFs and TRs from plant species.

PlnTFDB – The Plant Transcription Factor Database (RIAÑO PACHÓN *et al.* 2007)
`http://plntfdb.bio.uni-potsdam.de/`

AGRIS: The Arabidopsis gene regulatory information server (DAVULURI *et al.* 2003)
`http://arabidopsis.med.ohio-state.edu/`

PlanTAPDB, a phylogeny-based resource of plant transcription-associated proteins (RICHARDT *et al.* 2007)
`http://www.cosmoss.org/bm/plantapdb/`

TOFBAC – The Database of Tobacco Transcription Factors (RUSHTON *et al.* 2008)
`http://compsysbio.achs.virginia.edu/tobfac/`

PlantTFDB – Plant Transcription Factor Databases (GUO *et al.* 2008)
`http://planttfdb.cbi.pku.edu.cn/`

DBD: Transcription factor prediction database (WILSON *et al.* 2008)
`http://www.transcriptionfactor.org/`

AGRIS (DAVULURI *et al.* 2003) was the first computational resource listing complete sets of TF and TR genes in *A. thaliana*, proteins were grouped into families according to their conserved DNA-binding domains. Additionally, AGRIS list putative *cis*-regulatory elements and links the TF information with putative target genes into gene regulatory networks. AGRIS was one of the motivations to develop a resource that encompassed a broad phylogenetic range of plant species with sequenced genomes, i.e., PlnTFDB.

PlanTAPDB (RICHARDT *et al.* 2007) maintained at the University of Freiburg, has a special focus on the automated phylogenetic inference of transcription associated proteins, i.e., TFs and TRs. This resource is family centered, even phylogeny centered, however identification of clusters of orthologues or orthologues pairs is difficult, in contrast

PlnTFDB is species centered, facilitating the retrieval and analysis of TFs and TRs from single species, and allowing easier cross-species comparison mediated by the identification of putative pairs of orthologues.

PlantTFDB (GUO *et al.* 2008) maintained at the Peking University share most functionalities with PlnTFDB. In its current version it includes the same sequenced species, but additionally has a larger list of species for which large EST collections are available, constituting a very useful resource for non-sequenced plant species.

DBD (WILSON *et al.* 2008) created at the Medical Research Council in the UK, covers a broader phylogenetic range, including animals and fungi, besides plants, but also bacteria and archaea. However no phylogenetic information is provided for members of the identified TF families.

In summary, several resources with different focus, functionalities and look and feel are freely available to the scientific community interested in the regulation of transcription.

6.5 Outlook

PlnTFDB will be updated regularly, including more sequenced species and increasing the number of identified families.

We have identified that some of the current protein domain models are not able to detect some family members that can be nevertheless detected via manual curation. This has prompted us to develop plant-specific domain models with improved sensitivity. A natural step will be to extend this approach for all TF and TR families.

The availability of complete sets of TF and TRs has facilitated the inference of phylogenetic relationships among this type of proteins along the green tree of life. We have started to carry out detailed phylogenetic analyses for individual families, e.g., bZIPs (see Chaper 4), further families are being analysed. Future releases of PlnTFDB will incorporate the main findings derived from such studies.

6.6 References

BALAZADEH, S., D. M. RIAÑO PACHÓN, and B. MUELLER-ROEBER, 2008 Transcription factors regulating leaf senescence in *Arabidopsis thaliana*. Plant Biol **10**: 63–75.

BIENZ, M., 2006 The PHD finger, a nuclear protein-interaction domain. Trends Biochem Sci **31**: 35–40.

BYZOVA, M. V., J. FRANKEN, M. G. AARTS, J. DE ALMEIDA-ENGLER, G. ENGLER, *et al.*, 1999 Arabidopsis STERILE APETALA, a multifunctional gene regulating inflorescence, flower, and ovule development. Genes Dev **13**: 1002–1014.

CALDANA, C., B. MUELLER-ROEBER, D. M. RIAÑO PACHÓN, and S. RUZICIC, 2006 Transcription factor networks in the initial phase of salt stress in rice. In *8th International Congress of Plant Molecular Biology, Adelaide, South Australia, August 20-25*.

CALDANA, C., W.-R. SCHEIBLE, B. MUELLER-ROEBER, and S. RUZICIC, 2007 A quantitative RT-PCR platform for high-throughput expression profiling of 2500 rice transcription factors. Plant Methods **3**: 7.

CORREA, L. G. G., S. BALAZADEH, D. M. RIAÑO PACHÓN, and B. MUELLER-ROEBER, 2006 Functional analysis of transcription factors that play important roles in leaf development and/or physiology at sink-to-source transition and the onset of senescence. POS-TUE-159. In *8th International Congress of Plant Molecular Biology, Adelaide, South Australia, August 20-25*.

CZECHOWSKI, T., R. P. BARI, M. STITT, W.-R. SCHEIBLE, and M. K. UDVARDI, 2004 Real-time RT-PCR profiling of over 1400 Arabidopsis transcription factors: unprecedented sensitivity reveals novel root- and shoot-specific genes. Plant J **38**: 366–379.

DAVULURI, R. V., H. SUN, S. K. PALANISWAMY, N. MATTHEWS, C. MOLINA, *et al.*, 2003 AGRIS: Arabidopsis gene regulatory information server, an information resource of Arabidopsis *cis*-regulatory elements and transcription factors. BMC Bioinformatics **4**: 25.

DAVULURI, R. V., and M. Q. ZHANG, 2003 Computer software to find genes in plant genomic DNA. Methods Mol Biol **236**: 87–108.

DESVEAUX, D., A. MARÉCHAL, and N. BRISSON, 2005 Whirly transcription factors: defense gene regulation and beyond. Trends Plant Sci **10**: 95–102.

DINNENY, J. R., and M. F. YANOFSKY, 2004 Floral development: an ABC gene chips in downstream. Curr Biol **14**: R840–R841.

DO, J. H., and D.-K. CHOI, 2006 Computational approaches to gene prediction. J Microbiol **44**: 137–144.

DUNCAN, L., I. NISHII, A. HARRYMAN, S. BUCKLEY, A. HOWARD, *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.

EISEN, J. A., K. S. SWEDER, and P. C. HANAWALT, 1995 Evolution of the SNF2 family of proteins: subfamilies with distinct sequences and functions. Nucleic Acids Res **23**: 2715–2723.

EULGEM, T., 2005 Regulation of the Arabidopsis defense transcriptome. Trends Plant Sci **10**: 71–78.

FERNANDEZ-SILVA, P., F. MARTINEZ-AZORIN, V. MICOL, and G. ATTARDI, 1997 The human mitochondrial transcription termination factor (mTERF) is a multizipper protein but binds to DNA as a monomer, with evidence pointing to intramolecular leucine zipper interactions. EMBO J **16**: 1066–1079.

GUO, A.-Y., X. CHEN, G. GAO, H. ZHANG, Q.-H. ZHU, *et al.*, 2008 PlantTFDB: a comprehensive plant transcription factor database. Nucleic Acids Res **36**: D966–D969.

GUTTERSON, N., and T. L. REUBER, 2004 Regulation of disease resistance pathways by AP2/ERF transcription factors. Curr Opin Plant Biol **7**: 465–471.

HEIM, M. A., M. JAKOBY, M. WERBER, C. MARTIN, B. WEISSHAAR, *et al.*, 2003 The basic helix-loop-helix transcription factor family in plants: a genome-wide study of protein structure and functional diversity. Mol Biol Evol **20**: 735–747.

KOONIN, E. V., 2005 Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet **39**: 309–338.

KROPAT, J., S. TOTTEY, R. P. BIRKENBIHL, N. DEPÈGE, P. HUIJSER, *et al.*, 2005 A regulator of nutritional copper signaling in Chlamydomonas is an SBP domain protein that recognizes the GTAC core of

copper response element. Proc Natl Acad Sci U S A **102**: 18730–18735.

LOWE, T., and S. EDDY, 1997 tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res **25**: 955–64.

MA, H., 2005 Molecular genetic analyses of microsporogenesis and microgametogenesis in flowering plants. Annu Rev Plant Biol **56**: 393–434.

MOORE, R. C., and M. D. PURUGGANAN, 2005 The evolutionary dynamics of plant duplicate genes. Curr Opin Plant Biol **8**: 122–128.

NAM, J., C. W. DEPAMPHILIS, H. MA, and M. NEI, 2003 Antiquity and evolution of the MADS-box gene family controlling flower development in plants. Mol Biol Evol **20**: 1435–1447.

OLIVERI, P., Q. TU, and E. H. DAVIDSON, 2008 Global regulatory logic for specification of an embryonic cell lineage. Proc Natl Acad Sci U S A **105**: 5955–5962.

OLSEN, A. N., H. A. ERNST, L. L. LEGGIO, and K. SKRIVER, 2005 NAC transcription factors: structurally distinct, functionally diverse. Trends Plant Sci **10**: 79–87.

RAVENTÓS, D., K. SKRIVER, M. SCHLEIN, K. KARNAHL, S. W. ROGERS, *et al.*, 1998 HRT, a novel zinc finger, transcriptional repressor from barley. J Biol Chem **273**: 23313–23320.

REED, J. W., 2001 Roles and activities of Aux/IAA proteins in Arabidopsis. Trends Plant Sci **6**: 420–425.

RIAÑO PACHÓN, D. M., I. DREYER, and B. MUELLER-ROEBER, 2005 Orphan transcripts in *Arabidopsis thaliana*: identification of several hundred previously unrecognized genes. Plant J **43**: 205–212.

RIAÑO PACHÓN, D. M., S. RUZICIC, I. DREYER, and B. MUELLER-ROEBER, 2007 PlnTFDB: An integrative plant transcription factor database. BMC Bioinformatics **8**: 42.

RICHARDT, S., D. LANG, R. RESKI, W. FRANK, and S. A. RENSING, 2007 PlanTAPDB, a phylogeny-based resource of plant transcription-associated proteins. Plant Physiol **143**: 1452–1466.

RIESE, M., S. HÖHMANN, H. SAEDLER, T. MÜNSTER, and P. HUIJSER, 2007 Comparative analysis of the SBP-box gene families in *P. patens* and seed plants. Gene **401**: 28–37.

RUSHTON, P. J., M. T. BOKOWIEC, T. W. LAUDEMAN, J. F. BRANNOCK, X. CHEN, *et al.*, 2008 TOBFAC: the database of tobacco transcription factors. BMC Bioinformatics **9**: 53.

RUZICIC, S., C. CALDANA, M. SOLTANINAJAFABADI, D. M. RIAÑO PACHÓN, and B. MUELLER-ROEBER, 2005 Comparative expression profiling of different rice varieties during initial phase of abiotic stress. Poster 329. In *5th International Rice Genetics Symposium and 3rd International Rice Functional Genomics Symposium. Manila, Philippines. November 19-23*. 220.

SCHIEFTHALER, U., S. BALASUBRAMANIAN, P. SIEBER, D. CHEVALIER, E. WISMAN, *et al.*, 1999 Molecular analysis of NOZZLE, a gene involved in pattern formation and early sporogenesis during sex organ development in *Arabidopsis thaliana*. Proc Natl Acad Sci U S A **96**: 11664–11669.

SHIGYO, M., M. HASEBE, and M. ITO, 2006 Molecular evolution of the AP2 subfamily. Gene **366**: 256–265.

SOLIMAN, M. A., and K. RIABOWOL, 2007 After a decade of study-ING, a PHD for a versatile family of proteins. Trends Biochem Sci **32**: 509–519.

STOREY, J. D., and R. TIBSHIRANI, 2003 Statistical significance for genomewide studies. Proc Natl Acad Sci U S A **100**: 9440–9445.

VELASCO, R., A. ZHARKIKH, M. TROGGIO, D. A. CARTWRIGHT, A. CESTARO, *et al.*, 2007 A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. PLoS ONE **2**:

e1326.

WILSON, D., V. CHAROENSAWAN, S. K. KUMMERFELD, and S. A. TEICHMANN, 2008 DBD–taxonomically broad transcription factor predictions: new content and functionality. Nucleic Acids Res **36**: D88–D92.

WILSON, Z. A., and C. YANG, 2004 Plant gametogenesis: conservation and contrasts in development. Reproduction **128**: 483–492.

WU, K.-L., Z.-J. GUO, H.-H. WANG, and J. LI, 2005 The WRKY family of transcription factors in rice and Arabidopsis and their origins. DNA Res **12**: 9–26.

ZHANG, M. Q., 2002 Computational prediction of eukaryotic protein-coding genes. Nat Rev Genet **3**: 698–709.

Allgemeinverständliche Zusammenfassung

Organismen weisen einen komplexen Steuerungsmechanismus auf, bei dem die Aktivität eines Gens räumlich und zeitlich reguliert wird. Ein wichtiger Schritt in diesem Mechanismus ist die sogenannte RNA Transkription. Hierbei wird die genetischen Information von der DNA (dem Molekül, das die Information speichert) in RNA (dem Molekül, das die Information weiter tragen kann) umgeschrieben. Zur Einleitung der RNA Transkription bedarf es mehrerer verscheidener Komponenten. Unter anderem werden Proteine benötigt, die die Aktivität der Gene in Abhängigkeit verschiedener Stimuli regulieren. Proteine mit solch einer Funktion werden spezifische/regulatorische Transkriptionsfaktoren (TFs) genannt. TFs können in evolutionär verwandte Genfamilien gruppiert werden, welche in ihren Proteinsequenzen charakteristische konservierte Regionen und Domänen aufweisen.

In dieser Arbeit habe ich unter Verwendung der Proteindomänen, die jede TF-Familie in den verschiedenen Pflanzenspezies von den einzelligen Rot- und Grünalgen zu den mehrzelligen blühenden Pflanzen kennzeichnen, komplette Sätze an TFs identifizieren können. Diese kompletten TF-Sätze (die Bandbreite reicht von 150 bis 2500 TFs pro Spezies), sowie weitergehende Informationen und Literaturhinweise wurden unter der Internetadresse `http://plntfdb.bio.uni-potsdam.de/` öffentlich zugänglich gemacht. Die Datensätze erlaubten es mir, detailliertere evolutionäre Studien mit unterschiedlichen Schwerpunkten durchzuführen. Diese reichten von der Analyse einzelner Familien bis hin zum genomweiten Vergleich aller TF-Familien in verschiedenen Organismen. Als Resultat besonders erwähnenswert ist, dass bevorzugt einige bestimmte TF-Familien in verschiedenen Spezies eine hervorgehobene Rolle spielen.

Eine wichtige TF-Familie in blühenden Pflanzen ist die bZIP Familie. Für diese konnte gezeigt werden, dass der letzte gemeinsame Vorfahr (LGV) aller Grünpflanzen mindestens vier bZIP Gene hatte. Darüber hinaus konnte gezeigt werden, dass der LGV aller

Grünpflanzen mit neun TF-Familien ausgestattet war und der LGV aller Grünpflanzen und Rotalgen drei zusätzliche TF-Familien aufwies. 23 TF-Familien wurden identifiziert, die es nur in Landpflanzen gibt. Sie könnten eine besondere Rolle bei der Besiedelung des neuen Lebensraum gespielt haben.

Aufbauend auf die Transkriptionsfaktordatensätze, die in dieser Arbeit erstellt wurden, wurde mittlerweile damit begonnen, experimentelle Plattformen zu entwickeln (für Reis und für *C. reinhardtii*), um Änderungen in der Genaktivität der TF-Gene unter verschiedenen genetischen oder Umweltbedingungen zu untersuchen.

Publications

Peer reviewed papers

- **Riaño-Pachón D.M.**, Nagel A., Neigenfind J., Wagner R., Basekow R., Weber E., Mueller-Roeber B., Diehl S., Kersten B. (2009) GabiPD: The GABI primary database - a plant integrative 'omics' database. *Nucleic Acids Res*, database issue.

- Arvidsson S., Kwasniewski M., **Riaño-Pachón D.M.**, Mueller-Roeber B. (2008) Quantprime - a flexible tool for reliable high-throughput primer design for quantitative PCR. *BMC Bioinformatics*, 9:465.

- Corrêa L.G.G., **Riaño-Pachón D.M.**, Schrago C.G., dos Santos R.V., Mueller-Roeber B., Vincentz M.. (2008) The role of bZIP transcription factors in green plant evolution: adaptive features emerging from four founder genes. *PLoS ONE*, 3:e2944.

- Balazadeh S., **Riaño-Pachón D.M.**, Mueller-Roeber B. (2008) Transcription factors regulating leaf senescence in *Arabidopsis thaliana*. *Plant Biol*, 10:63–75.

- **Riaño-Pachón D.M.**, Corrêa L.G.G., Trejos-Espinosa R., Mueller-Roeber B. (2008) Green transcription factors: a Chlamydomonas overview. *Genetics*, 179:31–39.

- Montoya-Solano J.D., Suarez-Moreno Z.R., **Riaño-Pachón, D.M.**, Montoya-Castaño D., Aristizabal-Gutierrez F.A. (2007) Diseño de oligonucleótidos para el estudio de genes celulolíticos y solventogénicos en cepas colombianas de *Clostridium* sp. (Clostridiaceae) (Oligonucleotide Probe Design for the Study of Cellulolytic and Solventogenic Genes in Colombian *Clostridium* sp. Strains (Clostridiaceae)). *Acta Biol Col*, 12S1: 55–74.

- Gómez-Porras J.L., **Riaño-Pachón D.M.**, Dreyer I., Mayer J.E., Mueller-Roeber B. (2007) Genome-wide analysis of ABA-responsive elements ABRE and CE3 reveals divergent patterns in Arabidopsis and rice. *BMC Genomics*, 8:260

- Merchant S.S., Prochnik S.E., Vallon O. *et al.* (2007) The Chlamydomonas genome reveals the evolution of key animal and plant functions. *Science*, 318: 245–251.

- **Riaño-Pachón D.M.**, Ruzicic S., Dreyer I., Mueller-Roeber B. (2007) An integrative plant transcription factor database. *BMC Bioinformatics*, 8: 42.

- **Riaño-Pachón D.M.**, Dreyer I., Mueller-Roeber B. (2005) Orphan transcripts in *Arabidopsis thaliana*: identification of several hundred previously unrecognized genes. *Plant J*, 43:205–212

- **Riaño-Pachón D.M.**, Valenzuela E.M., Mantilla J.R., Agudelo C. (2003) Diversidad genética y estructura de la población de *Vibrio cholerae* en Colombia. (Genetic diversity and population structure of *Vibrio cholerae* in Colombia) *Rev Col Biotec*, 5:36–44

Invited papers

- Kersten B.,, Nagel A., **Riaño-Pachón D.M.**, Neigenfind J., Weber E., Wagner R, Diehl S. (2008) Die GABI-Primärdatenbank GabiPD Komplexe Integration von GABI-Daten aus Modell- und Nutzpflanzen. *GenomXPres*, 1.08:17–19

- Mueller-Roeber B., **Riaño-Pachón D.M.**, Ruzicic S., Caldana C., Witt I., Zanor M.I. (2005) Pflanzliche Regulatorproteine. *BIO forum* 6:32–34

In preparation

- Caldana C., Ruzicic S., **Riaño-Pachón D.M.**, Mueller-Roeber B. Matrix of rice transcription factor genes during initial phase of salt stress.

- **Riaño-Pachón D.M.**, Dreyer I., Mueller-Roeber B. Evolution of protein domain co-occurrence networks in plants.

- Köhler B., Müller K., Schulz K., Kretschmer N., **Riaño-Pachón D.M.**, Stuckas H., Mueller-Roeber B. Recombination and regulation at the splicing level leading to functional diversification of AtCNGC12, AtCNGC11, and AtCNGC3.

Diego Mauricio Riaño Pachón

Contact Information

E-mail: diriano@uni-potsdam.de — diriano@gmail.com
WWW: `http://www.geocities.com/dmrp.geo/`

Personal Information

Nationality	Colombian.
Place of birth	Bogotá D. C.
Date of bith	August 16th, 1975.

Research Interests

Computational Biology, Statistical methods for classification, sequence analysis, machine learning, non-protein-coding RNAs, transcriptional regulation, phylogenetic analysis, complex network analysis, graph theory.

Education

University of Potsdam, Potsdam, Germany
Department of Molecular Biology

PhD Student (2005–2008).

- Dissertation Topic: "Identification of transcription factor genes in plants".
- Supervisor: Prof. Dr. Bernd Mueller-Roeber.

Universidad Nacional de Colombia, Bogotá D.C., Colombia
Department of Biology

Biologist, May 2001 (Equivalent to the German "Diplom Biologe")

- Dissertation Topic: "Molecular characterization of *Vibrio cholerae* isolates obtained in Colombia between 1991 and 1996".
- Supervisor: Asoc. Prof. Emilia Maria Valenzuela de Silva.
- Grade with Honors: Thesis Meritorious Mention.

Honors and Awards

Third Place in the **Contest Best Graduation Papers - Universidad Nacional de Colombia-XI version**. Category Health Sciences. 2001.

Diploma thesis Meritorious mention. **Universidad Nacional de Colombia**.

Computational Skills

- Statistical Packages: R, SPAD-N.
- Programming languages: Perl, BioPerl, PHP, SQL, HTML, CSS.
- Applications: Common Linux and Microsoft Windows software, LaTeX.
- Database design: Microsoft Access, MySQL, Oracle.
- Operating Systems: Unix/Linux (Server and workstation), MacOSX, MS Windows.