Nucleic Acids Research, 2006, Vol. 00, No. 00 ***1–16***
doi:10.1093/nar/gkl888

The natural history of the WRKY–GCM1 zinc fingers and the relationship between transcription factors and transposons

M. Madan Babu[1,2,†,*], Lakshminarayan M. Iyer[1,†], S. Balaji[1] and L. Aravind[1,*]

[1]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA and [2]MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK

Received June 29, 2006; Revised October 4, 2006; Accepted October 9, 2006

ABSTRACT

WRKY and GCM1 are metal chelating DNA-binding domains (DBD) which share a four stranded fold. Using sensitive sequence searches, we show that this WRKY–GCM1 fold is also shared by the FLYWCH Zn-finger domain and the DBDs of two classes of Mutator-like element (MULE) transposases. We present evidence that they share a stabilizing core, which suggests a possible origin from a BED finger-like intermediate that was in turn ultimately derived from a C2H2 Zn-finger domain. Through a systematic study of the phyletic pattern, we show that this WRKY–GCM1 superfamily is a widespread eukaryote-specific group of transcription factors (TFs). We identified several new members across diverse eukaryotic lineages, including potential TFs in animals, fungi and *Entamoeba*. By integrating sequence, structure, gene expression and transcriptional network data, we present evidence that at least two major global regulators belonging to this superfamily in *Saccharomyces cerevisiae* (Rcs1p and Aft2p) have evolved from transposons, and attained the status of transcription regulatory hubs in recent course of ascomycete yeast evolution. In plants, we show that the lineage-specific expansion of WRKY–GCM1 domain proteins acquired functional diversity mainly through expression divergence rather than by protein sequence divergence. We also use the WRKY–GCM1 superfamily as an example to illustrate the importance of transposons in the emergence of new TFs in different lineages.

INTRODUCTION

Several studies have suggested that there are marked differences in the complement of DNA-binding domains DBDs in eukaryotic transcription factors (TFs) vis-à-vis their prokaryotic counterparts (1–5). In practically all prokaryotes studied to date, belonging to both the bacterial and archaeal super-kingdoms, the helix–turn–helix domain (HTH) is the most prevalent DBD of TFs. TFs with HTH domains constitute >90% of TFs found in any given prokaryotic genome and show a power-law scaling in their numerical distribution with respect to proteome size (3,6,7). In contrast, it has been noted that most eukaryotes show lineage-specific expansions of TFs with DBDs belonging to a wide variety of structural scaffolds. Although specific versions of the HTH domain, such as the homeodomain, are highly prevalent in crown group eukaryotes such as animals, fungi, slime moulds and plants, they are entirely absent or exceedingly rare in other eukaryotic lineages such as the diplomonads (*Giardia*), kinetoplastids, apicomplexans and ciliates (*Tetrahymena*) (3,8,9). Similarly, TFs with DBDs of the VP1 superfamily are currently only known from plants, whereas those of the POU-type of HTH domains are found only in animals (5,10). In addition to this lineage-specific diversity, TFs of earlier branching eukaryotic groups are poorly known due to lack of experimental studies on their transcription apparatus. These observations pose a general question regarding the origins of various eukaryotic TFs. Given their structural diversity, it is clear that their evolutionary history needs to be approached on a case-by-case basis. At the same time, it is worthwhile to investigate general trends in their evolutionary trajectories, which might throw light on the causes for the apparent diversity of DBDs recruited as TFs.

In structural terms, one prevalent structural category of DBDs, which appears to have been extensively used, primarily in eukaryotes, is the metal-chelating class. Examples include the classical C2H2 Zn-finger, versions of the treble clef fold, such as the GATA type Zn-finger and the nuclear hormone receptor Zn-finger, the double-sex domain, the fungal-type bi-nuclear (C6) Zn-finger and the plant-specific SBT domain (1,11–15). Some of these scaffolds, e.g. the C2H2 Zn-finger (found in low copy numbers in archaeal proteomes) (16) and the treble clef domain (found in the

*To whom correspondence should be addressed. Tel: +1 301 594 2445; Fax: +1 301 480 9241; Email: madanm@mrc-lmb.cam.ac.uk

†The authors wish it to be known that, in their opinion the first two authors should be regarded as joint First Authors

endonuclease VII/HNH fold DNAses), appear to be ancient and have representatives in prokaryotes (11,12,17). However, the above domains, as well as eukaryote-specific Zn-chelating DBDs underwent proliferation as TFs only much later in eukaryotic evolution (11,13). Another generic trend observed in eukaryotes is the relationship between DBDs of their TFs and those found in diverse transposases, integrases and other mobile selfish elements. For example, the AP2 DBD, which is highly prevalent in the TFs of plants, apicomplexans and diatom algae, is also found in the transposases of different elements and the integrase of lysogenic lambdoid phages (18). Similarly, other DBDs such as the BED finger (19), the THAP finger (20), the Paired domain (21) and the VP1 domain are also shared by several lineage-specific eukaryotic TFs and proteins of selfish elements such as transposases, integrases and restriction endonucleases (22).

The WRKY domain is a Zn-chelating DBD that is lineage-specifically expanded along with MADS, AP2, VP1 and Myb domains in plant TFs (1,15,23,24). It has also been detected in a few other eukaryotes such as *Dictyostelium* and *Giardia* (23–25), and recent structural studies have shown it to contain an unusual DBD that is believed to be distinct from most other well-characterized Zn-chelating domains (26). The only other Zn-chelating domain with a similar fold is the DBD of the Glial Cell Missing (GCM1) TFs of coelomate animals (26,27). Furthermore, anecdotal observations have pointed to possible relationship between the WRKY domain and a domain conserved in transposases with the MudR-type transposase domain. The unusual phyletic patterns, unique structure, availability of different high-throughput expression and ChIP-chip data (28–31) and possible links to transposases prompted us to systematically investigate the natural history of the WRKY TFs. We hoped that they might provide a general model for understanding evolution of lineage-specific DBDs and their expansions, as well as the rise of lineage-specific global regulatory hubs in transcriptional networks of eukaryotes. We also sought to better understand the more general connection between TFs and selfish elements by using the WRKY domain as a model.

We present below results of this study, which uncovered several novel points of interest regarding WRKY proteins. These include structural connections to other Zn-chelating domains, detection of novel versions of the WRKY domain and evidence for repeated rise of global regulatory hubs within this family in different organisms.

MATERIALS AND METHODS

The NR database of protein sequences (National Center for Biotechnology Information, NIH, Bethesda, MD) was searched using the BLASTP program (32). Iterative database searches were conducted using the PSI-BLAST program (33) with either a single sequence or an alignment used as the query, with the PSSM inclusion expectation value (*E*-value) threshold of 0.01 (unless specified otherwise); the searches were iterated until convergence. Hidden Markov models (HMMs) were built from alignments using the hmmbuild program and searches carried out using the hmmsearch program from the HMMER package (34). For all searches with compositionally biased proteins, the statistical correction for this bias was employed (35). Entropy analysis of proteins was carried out using the SEG program (36). Multiple sequence alignments were constructed using the T_Coffee (37) and MUSCLE (38) programs, followed by manual correction based on the PSI-BLAST results. Similarity-based clustering of proteins was carried out using the BLASTCLUST program (ftp://ftp.ncbi.nlm.nih.gov/blast/documents/blastclust.html). All large-scale sequence and structure analyses procedures were carried out using the TASS package (V. Anantharaman, S. Balaji, L. M. Iyer and L. Aravind, unpublished data), which operates similar to the SEALS package (39).

Protein secondary structure was predicted by using a multiple alignment to generate a HMM and PSSM, which were then used by the JPRED program to produce a final structural prediction with 72% or greater accuracy (40). Protein structure manipulations were performed using the Swiss-PDB viewer program (41) and the ribbon diagrams were constructed using the PYMOL program (http://www.pymol.org). Average structure for structures solved using NMR was determined using AVEPDB available from the Uppsala Software Factory (http://xray.bmc.uu.se/~gerard/manuals/avepdb_man.html). For structural searches of the PDB, the DALI and SSM programs were used (42–44). The studies on clustering-based DALI *Z*-scores have suggested that *Z*-scores > 10 are characteristic of obvious relationships, such as those between two closely related proteins of the same family. Between *Z*-scores 10 and 6, typically, the relationships correspond to more distant relationships that might be recovered through sequence profile analysis and searches using HMMs. *Z*-scores < 3 fall in the realm of remote structural relationships and require additional analysis, such as comparisons of topologies to make further inferences regarding these relationships.

Phylogenetic analysis was carried out using the neighbor-joining and minimum evolution (least squares) methods using the MEGA package (45). Gene expression data for the developmental stages and illumination conditions for *Arabidopsis thaliana* was obtained from the AtGenExpress expression atlas (28) (http://www.weigelworld.org/research/projects/resources/microarray/AtGenExpress/).

Expression levels for ~22 000 genes were available in triplicate from 79 samples covering many different developmental stages (embryogenesis to senescence) and from diverse organs (e.g. root, stem and leaves). Expression values were averaged to obtain the final estimate for a given gene. For this analysis, we only considered expression data available for the wild-type plant and excluded data available for the different mutants. Similarly, expression levels for the same set of genes were available as triplicates for eight different light conditions (e.g. continuous white light and continuous blue light) and two time points (45 min and 4 h) per condition. The final expression estimates were obtained by averaging these values for each time point and illumination condition. Expression profiles for the genes which were identified to contain the WRKY–GCM1 domain (from the three different families) were extracted from the above-mentioned datasets and were visualized using the program matrix2png (46). The genes in these matrices were ordered according to their sequence similarity between the WRKY–GCM1 domains for each of the families. The neighbor-joining tree was obtained using the MEGA package with distances calculated using the JTT matrix. All file manipulations and

data processing were carried out using custom written PERL scripts.

The transcriptional regulatory network for yeast was assembled from the results of genetic, biochemical and ChIP-chip experiments (29–31,47–50). This network consists of 4441 genes, which include 157 specific TFs, 4410 target genes and 12 873 regulatory interactions. Regulatory hubs were identified as proteins which regulate >150 target genes, i.e. the top 20% of the TFs with most number of regulated genes.

RESULTS AND DISCUSSION

Structural features and affinities

The recently published solution structure of the WRKY domain of the plant TF WRKY4 (PDB: 1WJ2) (26) revealed that it contains a four-stranded core with the characteristic WRKY signature mapping to the first strand. The domain is stabilized by a Zn atom chelated by two cysteines occurring, respectively, at the end of strand 2 and at the beginning of strand 3, and two conserved histidines occurring at the end of strand 4 (Figure 1a). The original comparative analysis of the WRKY structure revealed two other structures sharing the same core fold. The first of these, the GCM domain, shares an identical set of cysteine and histidine ligands. However, in the GCM domain, we noted that a copy of the evolutionarily mobile Zn-ribbon module is inserted between the two conserved N-terminal cysteines equivalent to those of the classical WRKY domain (Figure 1b). The second related structure is that of the *No apical meristem* (NAM) family of DNA-binding proteins which are exclusively found in plants (51). Although their DNA-binding modes are apparently similar, members of the NAM family lack a metal binding site and are entirely stabilized through hydrogen bonding. Given that the NAM DBD is exclusively found in plants, it is likely that they are a relatively recent offshoot of the classical metal binding WRKY domains in this lineage.

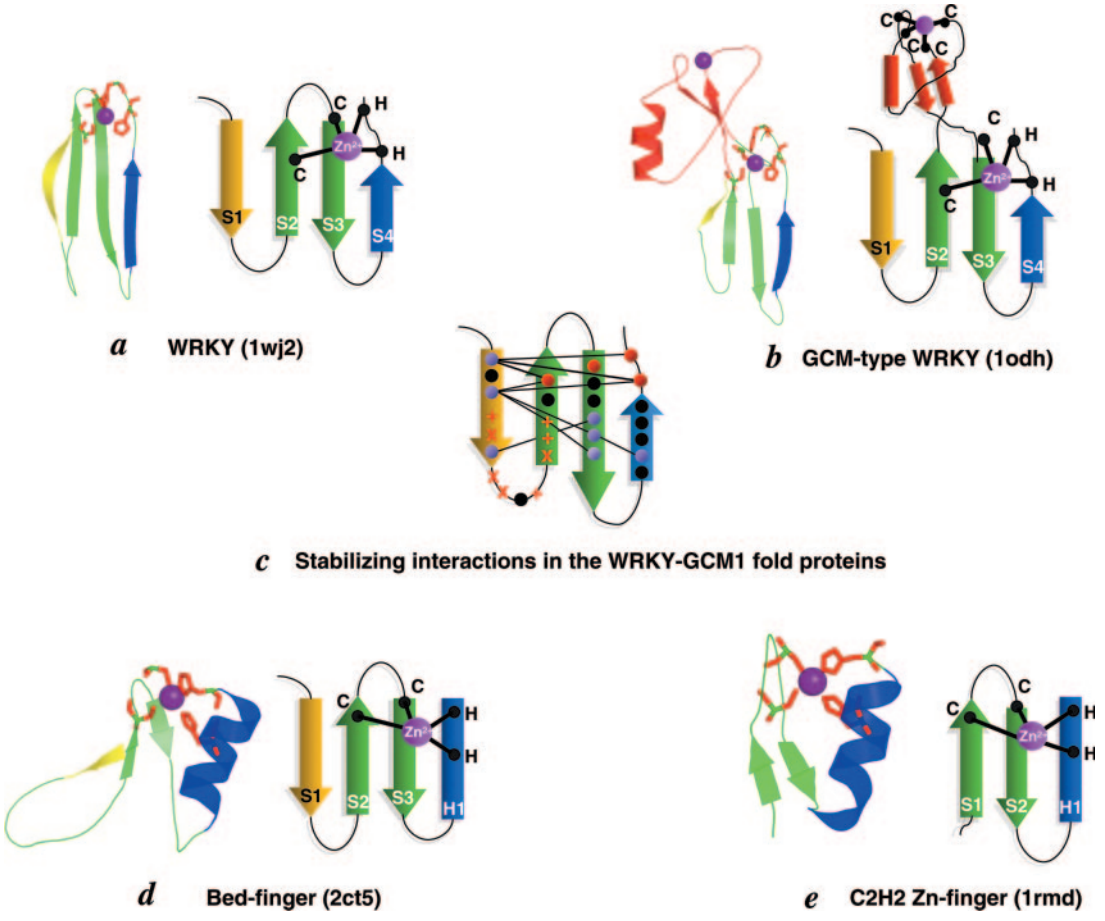
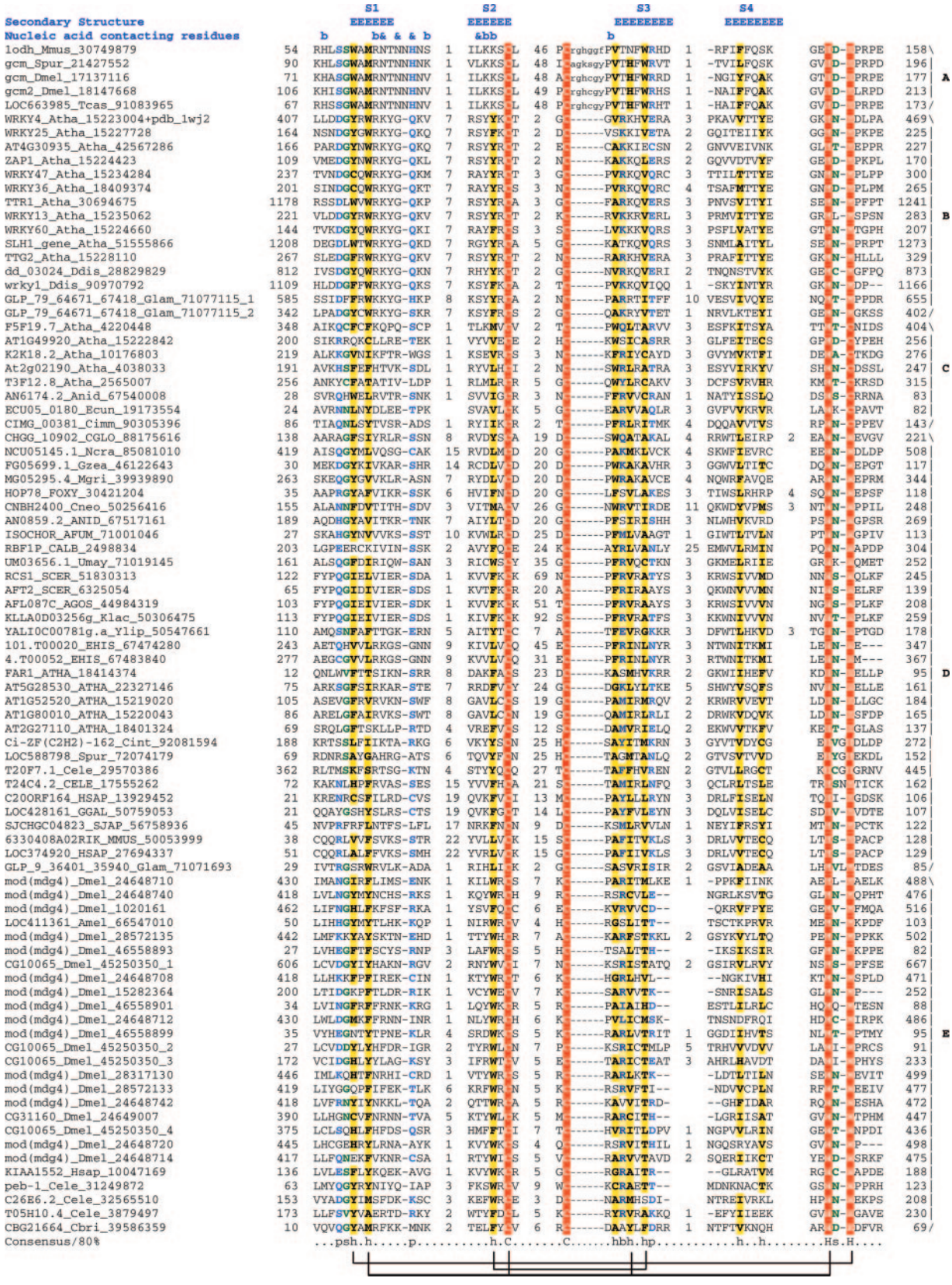


Figure 1. Topology diagram and cartoon representation of Zn-chelating DBDs. (**a**) WRKY domain from the plant TF WRKY4 (*Arabidopsis thaliana*, PDB:1wj2), which is primarily expressed in the leaf, root and seed. (**b**) The DBD from Glial Cell Missing 1 (*Mus musculus*, PDB:1odh) protein. The Zn-ribbon module inserted between the two conserved cysteines in the WRKY–GCM1 domain (shown in red) facilitates the binding of another Zn atom which is coordinated through the four conserved cysteines in this module. (**c**) The set of conserved intramolecular interactions which stabilize the fold in the classical WRKY proteins and the GCM1 protein. Lines represent interactions between the amino acids; metal chelating residues are shown in red; residue positions which participate in critical stabilizing interactions are shown in purple; + represent the position which contacts the backbone of the DNA; X represents the position which contacts the base. (**d**) The BED finger protein Zbed1x (*Homo sapiens*, PDB: 2CT5) and (**e**) the classical C2H2 Zn-finger domain from the RAG1 protein (PDB:1rmd). The first strand (and equivalent strands from the other structures) containing the WRKY motif in WRKY4 is shown in yellow. The two strands which house the conserved cysteine that participates in co-coordinating the Zn atom is shown in green. The last secondary structural element containing the pair of histidine residues is shown in blue.


S1
S2
S3
S4
Secondary Structure
Nucleic acid contacting residues
EEEEEE
EEEEEE
EEEEEEEE
EEEEEEEE
b
b& & & b
&bb
b
Consensus/80%
...psh.h.....p.....h.C.....C.....hbh.hp.....h..h.....Hs.H....

A number of critical stabilizing interactions could be identified using the structure of the classical WRKY domain (Figures 1c and 2). The conserved W of the WRKY motif is involved in a set of critical stabilizing interactions, which include (i) hydrophobic interactions with the first cysteine and histidine of the two metal chelating dyads and (ii) an interaction with the side chain of the residue in the fourth position, downstream of the second cysteine of the metal chelating dyad. The position occurring four residues upstream of the first metal chelating histidine participates in another stabilizing contact with the above-mentioned residue in the fourth position downstream of the second cysteine in the core. A further key stabilizing interaction is between an aromatic residue two positions upstream of the conserved W of the WRKY motif and the two histidines of the second metal-chelating dyad. This position and the first of the histidines typically interact via an aromatic stacking interaction. A fourth stabilizing interaction in the classical WRKY domain is mediated by a hydrophobic or an aromatic stacking interaction between the position two residues immediately upstream of the first cysteine and a well-conserved hydrophobic position in the middle of the third strand in the fold. One notable feature is that the majority of these interactions connect the metal-chelating residues with the stabilizing hydrophobic interaction network associated with the strands. Superposition of the structure of the classical WRKY domain with that of the GCM1 DBD shows that all the equivalent positions are present in the latter structure and participate in comparable potentially stabilizing interactions. This suggests that these positions and their interactions are a common conserved feature of the fold shared by these proteins. Accordingly, we hereinafter refer to the fold as the WRKY–GCM1 fold and the corresponding superfamily of protein domains the WRKY–GCM1 superfamily (Figures 1c and 2).

Additionally, in classical WRKY domains the conserved Y in the WRKY motif forms another stabilizing interaction on the opposite face to the above-described constellation of residues. Its principal interacting partner is a bulky position occurring three residues downstream of the second metal chelating cysteine with which it forms a hydrophobic or a PI–PI stacking interaction (Figures 1c and 2). This extensive stabilization of the WRKY domain through a hydrophobic core might have rendered the stabilizing metal superfluous, thus favoring its loss in certain versions, such as the derived NAM DBD.

To identify potential distant relationships, we scanned the Protein Data Bank (PDB) database for structures with similar arrangements of the key stabilizing positions of the fold. We observed that there are two domains with structural features related to the WRKY domain. These domains are the C2H2 Zn finger and the BED finger (Figure 1d and e), both of which are previously characterized Zn-chelating DBDs (19). The core conserved feature shared by these two domains and the WRKY–GCM1 domain is a hairpin of two strands which bear the N-terminal cysteine dyad in structurally congruent positions. The three domains also share the spatial similarity in the location of the two C-terminal metal liganding histidines. However, in the latter two structures, unlike the WRKY–GCM1 domain, they are derived from a helical segment. Furthermore, all the structures bear a key hydrophobic residue two positions upstream of their first cysteine which might participate in hydrophobic interactions stabilizing the core. The BED finger has certain additional shared features with the WRKY–GCM1 domain such as a conserved aromatic or hydrophobic position (19) equivalent to the W in latter domain. In the BED finger, this aromatic residue is in the context of a strand, which is positioned equivalent to the first strand of the WRKY–GCM1 domains. The key contacts of this aromatic residue in the BED finger also closely resemble those made by cognate positions in the WRKY domain. A systematic survey of a range of all currently characterized Zn-chelating modules with available experimentally determined structures (12) do not show the distinctive features shared by these three groups of Zn-chelating domains.

In terms of phyletic patterns, the WRKY–GCM1 and BED finger domains are more restricted than the classical C2H2 finger whose origin predates the common ancestor of the extant archaea and eukaryotes (19). These observations taken together suggest the possible derivation of the WRKY–GCM1 domain and BED finger domains from the C2H2 finger. An examination of the structures of the three domains suggests that the region N-terminal to the core C2H2 domain which is chiefly stabilized by the chelated metal, might have initially acquired an extended conformation with certain key hydrophobic/aromatic residues providing an additional stabilizing extension to the metal-chelating residues. This configuration would have resembled the state

Figure 2. Multiple sequence alignment of the WRKY domains. Proteins are denoted by their gene names, species abbreviations and GenBank identifier (gi) numbers. The secondary structure derived from the average solution structure of WRKY4 is shown above the alignment, where E represents a β-strand. Residues involved in contacting DNA in the structure of the WRKY domain in the GCM protein (PDB: 1odh) are shown below the alignment. Positions which contact the DNA are shown below the secondary structure profile. 'b' represents a position which contacts the DNA backbone and '&' mark positions which contact the base. Conserved interactions which are critical for stabilizing the fold are shown at the bottom of the alignment. The coloring reflects the conservation profile at 80% consensus. (A) GCM-type WRKY–GCM1 domains from mammals and *Drosophila*. Note the large insertion between strand 2 and strand 3, which normally contains a copy of the evolutionarily mobile Zn-ribbon module (see Figure 1). (B) Representative members of the classical WRKY family seen in the TFs of plants, *Dictyostelium* and *Giardia lamblia*. Members in this family do not show any major insertion between the conserved cysteines and typically contain a WRKY motif in the first strand. (C) The HxC family of WRKY–GCM1 domain family. (D) WRKY–GCM1 domain of the insert containing type. (E) FLYWCH-type WRKY domains seen primarily in animals. The coloring scheme and consensus abbreviations are as follows: h, hydrophobic (h: ACFILMVWY) and a, aromatic (a: FWY) residues shaded yellow; b, big (LIYERFQKMW) residues shaded gray; s, small (AGSVCDN) residues colored green; and p, polar (STEDKRNQHC) residues colored magenta. Species abbreviations are as follows: Afum: *Aspergillus fumigatus*; Agos: *Ashbya gossypii*; Amel: *Apis mellifera*; Anid: *Aspergillus nidulans*; Atha: *Arabidopsis thaliana*; Calb: *Candida albicans*; Cbri: *Caenorhabditis briggsae*; Cele: *Caenorhabditis elegans*; Cglo: *Chaetomium globosum*; Cimm: *Coccidioides immitis*; Cneo: *Cryptococcus neoformans*; Cint: *Ciona intestinalis*; Ddis: *Dictyostelium discoideum*; Dmel: *Drosophila melanogaster*; Ecun: *Encephalitozoon cuniculi*; Ehis: *Entamoeba histolytica*; Foxy: *Fusarium oxysporum*; Ggal: *Gallus gallus*; Glam: *Giardia lamblia*; Gzea: *Gibberella zeae*; Hsap: *Homo sapiens*; Klac: *Kluyveromyces lactis*; Mgri: *Magnaporthe grisea*; Mmus: *Mus musculus*; Ncra: *Neurospora crassa*; Scer: *Saccharomyces cerevisiae*; Sjap: *Schistosoma japonicum*; Spur: *Strongylocentrotus purpuratus*; Tcas: *Tribolium castaneum*; Umay: *Ustilago maydis*; Ylip: *Yarrowia lipolytica*.

seen in the extant BED fingers. Given the role of the N-terminal 'linker' region of C2H2 fingers in interacting with DNA (52), it is not surprising that the neomorphic N-terminal strand of the above-postulated evolutionary intermediate acquired a major DNA-binding role. This structural accretion might have re-organized the principal DNA–protein interface resulting in completely distinct DNA–protein interfaces between the classical C2H2 fingers and the WRKY fingers. Subsequently, the C-terminal helix might have developed an extended conformation resulting in the condition seen in the WRKY–GCM1 domains (Figure 1). Such an evolutionary scenario is reminiscent of the extensive structural modifications of the HTH domain to give rise to several distinctive domains with alternative DNA contacting elements such as the MetJ/Arc (ribbon–helix–helix) domain (3,5). The alternative scenario would imply *de novo* invention of the WRKY–GCM1 domains and convergent evolution of comparable structural features including the stabilizing network. Given that there are many alternative stabilization networks seen in the entire set of structurally distinct Zn-chelating domains of similar size as the above domains (12), such convergence is not a self-evident, likely outcome. This makes the alternative scenario less favored because it is a product of two low probability events (innovation and convergence).

A comparison of the available structures of two metal chelating domains with the WRKY–GCM1 fold, GCM1 and the plant TF, WRKY4, shows that the majority of DNA–protein contacts are made by the first two strands of the four stranded core. A series of large and polar residues including the R and the K of the classical WRKY domain project to the exterior and form both backbone and base contacts with the major groove of DNA. Though the positions of these DNA contacting residues are congruent and their orientations comparable between GCM1 and WRKY4, the actual residues themselves are poorly conserved. This suggests that the WRKY–GCM1 fold can possibly accommodate considerable diversity of binding sites mediated by residues unique to each family or subfamily (see below for further discussion). Additional subsidiary DNA contacts are made by the inserted Zn-ribbon, as well as C-terminal extensions in the case of the GCM1 protein.

Sequence analysis and identification of novel versions of the WRKY domain

In order to comprehensively identify all versions of the WRKY–GCM1 fold in the non-redundant (NR) Protein Data Bank (PDB), we initiated iterative sequence profile searches with different starting sequences using the PSI-BLAST program (33,35). In a parallel procedure, we also used alignments of known WRKY–GCM1 domains to search a database of all fully sequenced or nearly completed genomes using the hidden Markov models generated by the HMMER package (34). As a consequence of these searches, we identified several previously unknown versions of the WRKY–GCM1 fold from diverse eukaryotic organisms. For example, PSI-BLAST searches initiated with the C-terminal region of the *Entamoeba* protein 101.t00020 (gi: 67474280; region 220–347) retrieved the *Arabidopsis* FAR1 protein and its plant homologs ($E = 10^{-9}$, iteration 3), the WRKY family of TFs (iteration 5, $E = 10^{-3}$–10^{-9}), and the yeast TFs Rcs1p (iteration 3, $E = 10^{-3}$) and Aft2p (iteration 6, $E = 10^{-5}$). PSI-BLAST searches initiated with the N-terminus of the *Arabidopsis* FAR1 protein (gi: 5764395, region 1–210), retrieved the mod(Mdg4) protein from *Drosophila* (iteration 2, $E = 10^{-3}$). Similar transitive and iterative searches retrieved the Rbf1p from *Candida*, several uncharacterized proteins from various animals including vertebrates, *Entamoeba*, *Giardia* and the microsporidian *Encephalitozoon cuniculi* which have previously not been reported as containing WRKY domains. Amongst the significant hits in these searches were also regions of transposases of the Mutator-like element (MULE) from animals, plants and fungi with a pattern of conserved cysteines and histidines equivalent to that observed in the conventional WRKY domains. Although these transposases from different eukaryotic crown group lineages showed low sequence similarity to each other, they were unified by the presence of a common integrase domain similar to that of TnpA transposases of prokaryotes (see Supplementary Data). In plants, a subset of these transposases comprises the previously known MudR (Mutator) transposases. Fungal transposases which contain domains homologous to WRKY domains are the transposons of the Hop group from *Fusarium oxysporum* and the *Mutyl* group of *Yarrowia lipolytica*. Furthermore, these profile searches also identified Zn fingers of the mod(Mdg4) proteins, which are lineage-specifically expanded in insects (also called the FLYWCH Zn fingers) (53,54) and homologous Zn fingers in the *Caenorhabditis elegans* TF PEB-1 (55,56) and uncharacterized human proteins such as KIAA1552.

Having collected all the significant hits from these searches, we generated a multiple alignment of the WRKY–GCM1 superfamily using the T_coffee program (37) The alignment was further refined based on PSI-BLAST high scoring segment pairs and the structural alignment of the WRKY4 and GCM1 as a guide for the equivalent conserved positions. Most of the strong sequence conservation, other than the metal ligands, maps to the positions described above as mediating the key stabilizing interactions of the fold. Amongst the most striking of these are the two hydrophobic or aromatic positions associated with the core strands (Figure 2). There are a total of nine potential DNA contacting positions present in the core WRKY–GCM1 fold. Of these only two positions show consensus conservation of at least 80%, but even in these positions there is some variability in the actual residues present. However, the conservation of individual DNA contacting residues is considerably higher within the distinct families of the WRKY–GCM1 superfamily (Figure 2). Certain versions show characteristic inserts within the core fold. A prominent region displaying inserts of variable length is in the loop between the two conserved cysteines (Figure 2). Other than GCM1, which contains a Zn-ribbon in this region, a large sub-group of the WRKY–GCM1 domains including versions found in yeast TFs Rcs1p and Aft2p, the *Candida* TF Rbf1p and plant Far1-like TFs and related transposases show notable inserts of length 20–70 residues in this region. These residues are typically enriched in positively charged residues, which based on the precedence of GCM1, are suggested to make an additional DNA contact with the backbone phosphates and might affect the DNA-binding affinity of these domains. Another insert unique to the GCM1 family is seen

immediately after the second cysteine, and assumes the form of a loop projecting out of strand 3. This unusual outflow from the strand contributes to an interaction with the deoxyribose of the target DNA backbone, which is unique to this family of WRKY–GCM1 proteins.

Sequence and domain architectural diversity of the WRKY–GCM1 domain proteins

Examination of conserved sequence and structural motifs and clustering of sequences based on pair-wise BLAST bit scores helped us to identify several distinct families within the WRKY–GCM1 superfamily. These families show rather distinctive phyletic patterns and in certain cases domain architectures.

The classical WRKY family. This family is defined by the WRKY motif in strand-1 and a short spacing between the two conserved cysteines (Figure 2). It shows a lineage specific expansion (LSE) in plants with sporadic representatives in *Dictyostelium* and the early branching eukaryote *Giardia lamblia* (23–25). The majority of these proteins contain single or duplicate copies of the WRKY domain as the only detectable globular domains in the polypeptide (57,58). In another subset of classical plant WRKY domain proteins, we also identified a second previously uncharacterized Zn-chelating domain with a C[CH]CC set of metal ligands and a predicted C-terminal helix, which is highly enriched in positively charged residues (see Supplementary Data). This domain occurs immediately N-terminal to the WRKY domain and may constitute a second DBD of these proteins. A few unique lineage specific architectures appear to have emerged amongst these proteins which combine the WRKY domain with the plant disease resistant genes containing an AP-ATPase module. Additionally, as is typical of the plant disease resistant ATPases, these proteins may also contain the TIR and LRR modules associated with the AP-ATPase module (59,60) (Figure 3).

The WRKY–GCM1 domain family with insert between conserved cysteines. In addition to the eponymous insert these proteins also typically show a conserved W in the fourth strand of the core fold (Figure 2). This family is widely distributed in plants, animals, fungi and *Entamoeba*. The most common of these versions occur in MULE transposases, which have proliferated in various genomes such as the fungi *Chaetomium globosum*, plants such as *Arabidopsis*

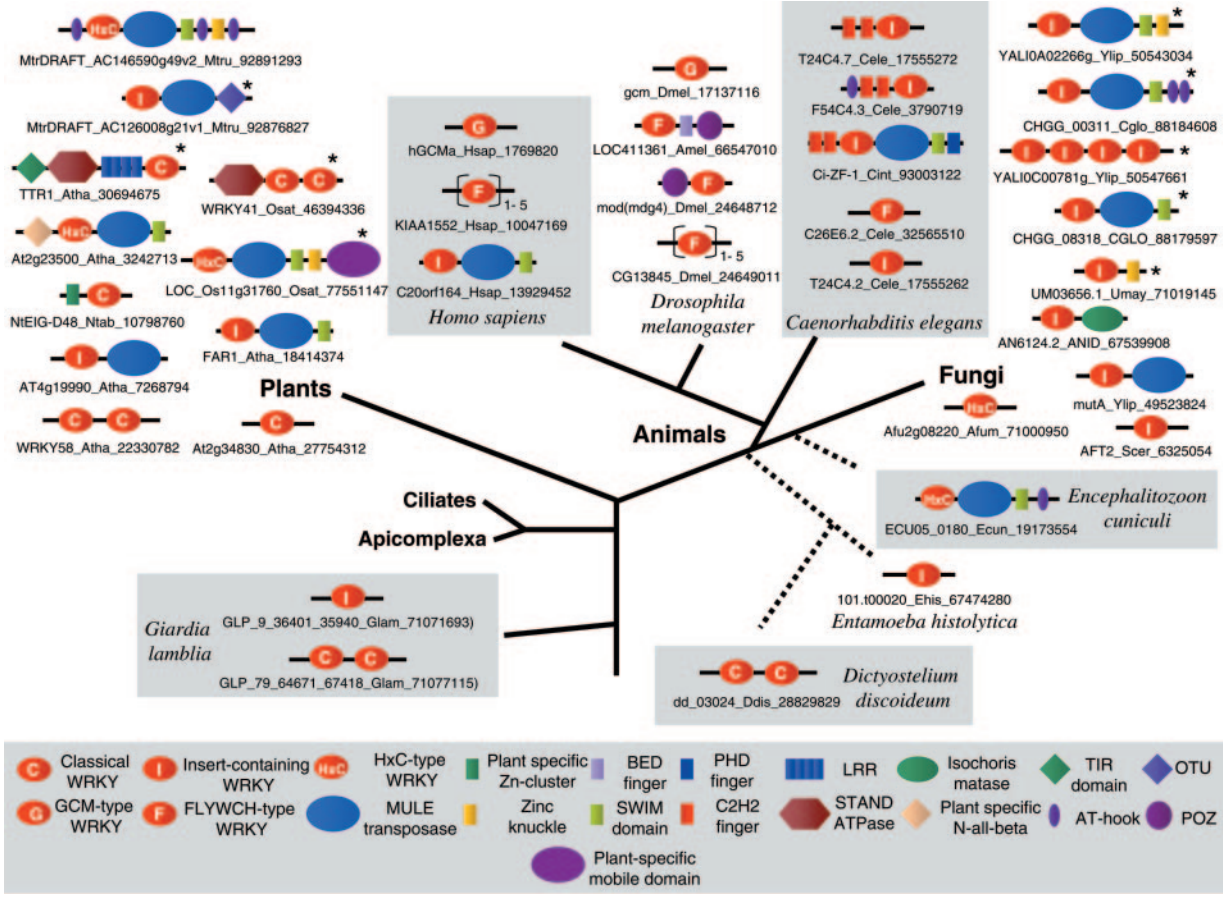


Figure 3. Domain architectures of proteins that contain the WRKY–GCM1 domain in the different lineages. A representative member for each distinct architectural class is denoted by its gene name, species abbreviation and GenBank identifier (gi) number. Species abbreviations are as follows. Afum: *Aspergillus fumigatus*; Anid: *Aspergillus nidulans*; Atha: *Arabidopsis thaliana*; Cele: *Caenorhabditis elegans*; Cglo: *Chaetomium globosum*; Cint: *Ciona intestinalis*; Ddis: *Dictyostelium discoideum*; Dmel: *Drosophila melanogaster*; Ecun: *Encephalitozoon cuniculi*; Ehis: *Entamoeba histolytica*; Glam: *Giardia lamblia*; Hsap: *Homo sapiens*; Mtru: *Medicago truncatula*; Ntab: *Nicotiana tabacum*; Osat: *Oryza sativa*; Scer: *Saccharomyces cerevisiae*; Spur: *Strongylocentrotus purpuratus*; Umay: *Ustilago maydis*; Ylip: *Yarrowia lipolytica*. Asterisk denotes distinct domain architectures seen within a single species in the lineage.

and *Medicago* and animals such as *Schistosoma japonica and C.elegans*. The catalytic domain of this class of transposases is structurally poorly characterized and is also improperly defined in publicly available domain databases such as CDD and Pfam. Hence we carried out systematic sequence searches to define the catalytic domain. Iterative sequence profile searches recovered the prokaryotic TnpA transposases with significant *E*-values (TnpA from *Escherichia coli*; $E = 10^{-5}$ in iteration 7) and this enabled us to precisely define the boundaries of the minimal catalytic domain of all these transposases. Secondary structure prediction showed that it contains a five stranded α+β core with the first three strands occurring in tandem succession and the remaining two in α/β elements. This unit is followed by a variable α-helical region and a conserved helix. The α+β unit has several conserved residues of which the most striking are two conserved acidic residues (typically aspartate) after strands 1 and 4. Additionally, the helix after the variable helical region contains a strictly conserved acidic residue (mostly glutamate) (Supplementary Data). This secondary structure progression and contexts of the three conserved acidic residues is very similar to a number of other transposases such as Activator/Hermes, Mariner, Transib, Mu, Tn5 and retroviral integrases that are known to adopt an RNAseH fold (17,61–65). Analysis of the predicted secondary structures of the TnpA family of transposases suggest that it is most similar to the Activator/Hermes-type of transposases as both these families share an α-helical insert between the core RNaseH fold and the C-terminal conserved helix that contributes the acidic residue (62). However, we observed that the TnpA family of transposases differ in having a strongly conserved acidic residue (typically glutamate) after strand 3 and an absolutely conserved histidine after strand 5 of the core RNaseH fold (Supplementary Data).

Majority of transposases with this version of the WRKY–GCM1 domain contain a core unit having an N-terminal WRKY–GCM1 domain fused with a C-terminal transposase catalytic domain. Typically, a metal-chelating SWIM domain is found fused with the C-terminus of the above core. Several transposases from the different eukaryotic crown group lineages may also contain further C-terminal extensions with a nucleic-acid binding Zn-knuckle (37) and/or the AT-hook motif (66) (Figure 3). A group of recently expanded transposases from the genome of *Medicago trunculata* show a C-terminal fusion to a papain-like protease domain of the OTU superfamily. This might suggest that the transposon encoded polyproteins are post-translationally processed by the protease domain. Another group of these transposases present in sea urchins, *Ciona* and expanded in *Caenorhabditis* contain two classic Zn fingers N-terminal to the WRKY–GCM1 domain. A few of these proteins from *Ciona* (gi: 93003122) and *C.elegans* (gi: 17544214) additionally contain a PHD finger at the C-terminus, suggesting that this form might be another widespread architecture. Members of this family without associated transposase domains are found in fungi, animals and *Entamoeba* and are closely related to transposase-containing variants. This raises the possibility that they could have emerged as fragments of transposases which have been probably recruited as TFs (See below for details).

The HxC family of WRKY–GCM1 domains. This family is typified by the presence of two distinctive features, namely the presence of a short spacer between the two N-terminal conserved cysteines and the presence of a H×C in place of the H×H motif of the C-terminal metal ligands. This family is found in plants, fungi and *Encephalitozoon cuniculi*, and similar to the previous family, they are typically found in transposases of MULE transposons. Versions from several transposases in plant genomes contain additional fusions beyond the core WRKY–GCM1 transposase catalytic domain unit (Figure 3). These include the DNA binding AT-hook motifs (66) on either side of the core domain and other C-terminal Zn-chelating modules such as the SWIM (67) and the Zn-knuckle (37). Moreover, in fungi like *Coccidiodes* and *Aspergillus*, the transposase domains might be truncated or entirely lost. These transitions again suggest a possible recruitment of erstwhile transposases as DBDs as in the case of the previous family. Some of these proteins from *Oryza sativa* contain an additional previously uncharacterized globular domain C-terminal to the transposase, SWIM and Zn-knuckle domain (e.g. LOC_Os11g31760). This domain is thus far found only in plants (partially overlapping with PFAM alignment DUF1723), is predicted to assume an α/β structure with 8 strands and at least 10 conserved helices and contains several conserved charged/polar positions. Stand-alone versions of this domain have proliferated to a greater or lesser degree in different plant genomes (e.g. ~22 in *Arabidopsis*, >150 in *Medicago* and >400 in *Oryza*) and are dispersed throughout the genome, often in identical or highly similar copies, including fragmentary versions. This suggests that it might define a novel mobile element in plants, which might have fused with the H×C family containing transposons.

In addition to the association of both the H×C family and the above-described family of the WRKY–GCM1 superfamily with MULE transposons, they are unified by the presence of a shared helix N-terminal to the first strand of the core fold. This shared helix might be involved in further interactions with DNA unique to these two families. These features suggest that the two families might form a higher-order group within the WRKY–GCM1 superfamily.

GCM1 family. This family is found in low copy number exclusively in the animal lineage and appears to have been recruited for a specific regulatory role in neural development early in animal evolution, with a secondary adaptation related to placental development in mammals (68,69). This phyletic pattern suggests that it might have been derived via the ingression of a Zn-ribbon into the pre-existing linker of a member of the insert-containing family.

The FLYWCH family. The FLYWCH fingers were defined originally based on the mod(Mdg4) proteins from *Drosophila*, which contain a characteristic W in the conserved hydrophobic position two residues upstream of the first cysteine. In *Drosophila*, the mod(Mdg4) proteins are produced by a distinctive locus containing an exon for an N-terminal POZ domain followed by several mutually exchangeable alternative C-terminal exons, each encoding a different Zn-finger domain. Similar loci are found in other insects such as beetles and hymenopterans, and vertebrates contain

homologous proteins with up to five tandem repeats of this Zn-finger domain (53,54). In *C.elegans*, members of this family similar to the TF Peb-1 contain a single copy of the ZnF (55,56). These observations suggest that tandem duplications of the domains of the FLYWCH family might have occurred before the divergence of the coelomates and might have been incorporated into a single polypeptide or utilized as alternative exons.

NAM family. The NAM family of DBDs is exclusively found in plants, where it may be lineage-specifically expanded to up to 100 or more representatives (70). As the basic fold of this domain is identical to the WRKY–GCM1 fold and it has a limited phyletic pattern, it is likely to have emerged early in the plant lineage followed by a massive LSE.

Emergence of transcription factors from MULE transposases

Presence of WRKY–GCM1 domains in transposases suggests a potential for their lateral mobility across different lineages of eukaryotes. The sporadic distribution of the classical WRKY family in phylogenetically distant eukaryotes is also suggestive of its possible spread through intra-eukaryotic lateral transfers. Different families of WRKY–GCM1 domains contain closely related versions that are either associated with transposases or occur as stand-alone forms which might be TFs. For example, two major yeast TFs Rcs1p and Aft2p, which form regulatory hubs of the yeast transcriptional network (71–75) and Rbf1p, which is critical for the yeast hyphal transition in *Candida* (76,77), belong to the insert containing family, which also includes several transposases. These observations raised the question regarding the evolutionary relationship between the TFs and transposases, as well as the rise of lineage-specific TF, to function as global regulatory hubs. On a related note, previous preliminary results from a comprehensive study on the expression patterns of plant genes suggested that there is potential tissue specific partitioning of different classical WRKY family proteins in *Arabidopsis* (28). As different families of WRKY–GCM1 superfamily have shown lineage-specific expansion to form multiple closely related groups, we were interested in investigating how they may have acquired a potential diversity of regulatory roles.

To understand better the evolutionary relationship between the transposases and TFs of the insert-containing family, we performed a systematic survey to identify all members of this family in plants, animals and fungi (Figure 4a). Subsequently, we used a multiple alignment including the N-terminal helical extension found in this family for a phylogenetic analysis. Although the domain is relatively short and rapidly diverging, certain clear-cut statistically well-supported (BP > 80%) groups emerged from the analysis. At the highest level, these included four monophyletic clades which composed of proteins from animals, plants, fungi and *Entamoeba*. We then systematically identified the domain architectures of all proteins in the tree and superimposed this information on it. The association with the transposase domain was the most prevalent configuration seen in the animal, plant and fungal clusters.

This observation, taken together with the higher order relationship between the insert containing family and the H×C containing family, which is also associated with the MULE transposons, suggests that the association with the WRKY–GCM1 domain with the transposase was the ancestral condition. Two distinct, well-supported clades of fungal TFs, respectively, typified by the Rbf1p of *Candida* and Rcs1p/Aft2p of *Saccharomyces*, were nested amongst transposases within the fungal cluster. This strongly supported the occurrence of at least two independent transitions in fungi from transposons to TFs.

We then sought to understand, in greater detail, the potential rise of these transposon-derived TFs to global regulatory roles or as developmental switches in particular fungal lineages. Using the extensive genomic data for ascomycete yeasts, we were able to determine that homologs of Rcs1p were only present in yeasts of the order Saccharomycetales (Figure 4b). Within this order of yeasts, a single ortholog was found in *Candida albicans*, *Debaryomyces hansenii*, *Klayveromyces lactis*, *Klayveromyces waltii*, *Saccharomyces kluyveri* and *Ashbya gossypii*. The two paralogous versions, Rcs1p and Aft2p, were only detected in the crown group species of the genus *Saccharomyces*, namely *S.cerevisiae*, *S.paradoxus*, *S.mikate*, *S.castlli* and *S.kudriavzevii*. Phylogenetic analysis within the ascomycete yeasts supported this duplication of Rcs1p and Aft2p within the *Saccharomyces* crown group. It also suggested an independent duplication in the yeast *Candida glabrata*. Examination of the transcriptional network (71,78) for these TFs showed that they regulated a total of 478 target genes of which they shared only 41 (Figure 4c). Though both these regulatory hubs only co-regulate 41 target genes, they are not 'autonomous hubs' in that they do not uniquely regulate their other target genes. Instead, they act as integrators of signals by combinatorially regulating different sub-sets of their target genes with >50 other TFs. This suggested that there was potentially a massive acquisition of new interactions following the duplication or specialization through rapid loss of most shared interactions present in their unduplicated ancestor. The independent duplication in *C.glabrata* and the comparable recent duplications seen in the TFs of the Rbf1p clade suggest that such newly recruited TFs may rapidly specialize to acquire functional diversity within relatively small phylogenetic distances.

To identify other potential derivations of TFs from MULE transposon, we examined the domain architectures within this family. We observed that there were several proteins containing stand-alone copies of the WRKY–GCM1 domain, both in yeasts filamentous ascomycetes and basidiomycetes such as *Ustilago maydis* and *Cryptococcus neoformans*. Some of these versions showed distinctive domain architectures, for example, fusion to a Zn-knuckle in UM03656.1 of *U.maydis* and fusion to an enzymatic isochorismatase domain in several filamentous fungi (Figure 3). In particular, the fusion to isochorismatase could potentially use the catalytic domain as a sensor domain to respond to particular small molecule effectors. Several of these stand-alone proteins were conserved across different fungal genera, for example, between *Magnaporthe grisea*, *Neurospora crassa* and *Giberella zea*, occur as single copy genes, and showed no other transposase genes in their vicinity. Similarly, in *C.elegans* and *C.briggsae*, there were solo versions of the insert-containing WRKY–GCM1 family which are closely related to those

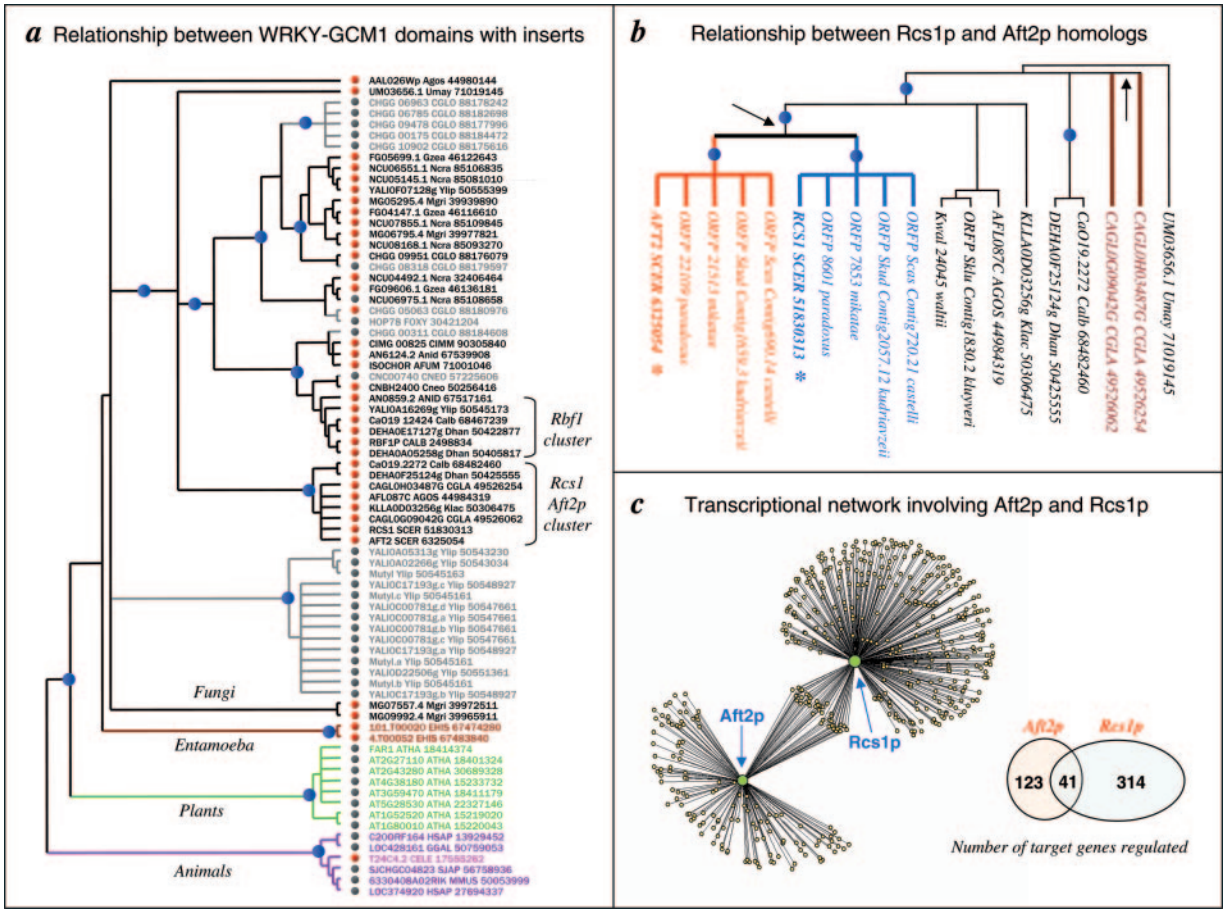


Figure 4. (**a**) Evolutionary relationship between the members of the insert containing WRKY–GCM1 domain family. Red circle represents known and potential TFs. Gray circle represents members which are transposases. A blue circle in the internal nodes of the tree indicates strong bootstrap support, (>80%). (**b**) Relationship between the Rcs1p and Aft2p homologs in the different fungal genomes. The tree has been rooted with the *Ustilago maydis* protein as the out-group. Arrowheads denote points where a potential gene duplication event occurred. A blue circle in the internal nodes of the tree indicates strong bootstrap support (>80%). (**c**) A section of the transcriptional regulatory network showing the target genes for the WRKY domain containing yeast TFs, Rcs1p (YGL071W) and Aft2p (YPL202C). The TFs are shown as green circles, and the target genes are shown as yellow circles. A line represents a direct transcriptional regulatory interaction between the TF and the target gene.

found in a specific sub-family of animal MULE transposases with two residues separating their conserved histidines (Figures 2 and 3 and see above). Taken together these observations suggest that the stand-alone copies may not be actively mobile or be dependent *in trans* on their transposase containing relatives. The conservation of such stand-alone versions across different species or related genera in both fungi and animals indicate that they might have possibly been fixed on account of their role as TFs.

The above observations hinted that there is a tendency for repeated transitions from transposases to TFs in this family. A possible clue for this is offered by the observation that transposases of different transposons are known to function as transcriptional regulators which regulate their own expression (79–81). In particular, there are examples in *Yarrowia* where the WRKY–GCM1 domain might occur as a stand-alone ORF in some versions of the Mutyl transposon. Other versions of the Mutyl transposon have no active transposase domain at all, but merely contain a single ORF with 3–4 repeats of the WRKY–GCM1 domain. Taken together, it appears to be likely that the original WRKY–GCM1 domain of the transposons possibly played a dual role in both recognizing the target sites as well as regulating the expression of the transposase itself. Subsequently, defective hyper-parasitic versions of the transposon containing only a single or multiple copies of the WRKY–GCM1 domain appear to have arisen, which propagated by using the active versions of the transposases *in trans* and regulated their own expression. This might have resulted in the observed proliferation of solo versions in various fungal genomes which provided the raw material for the evolution of new TFs.

Recruitment of lineage-specifically expanded WRKY–GCM1 domains in transcription factors that act as developmental switches and physiological regulators in plants

In plants, three major LSEs of WRKY–GCM1 domain proteins are seen. Two of these represent trivial cases of proliferation of two families of WRKY–GCM1 domains encoded by the MULE transposons. The third of these is the expansion of the classical WRKY domains. There are at least 60 classical WRKY members in the genome of

Arabidopsis, but they apparently show relatively low sequence diversity especially in terms of some of the key predicted DNA contacting residues. Previously, studies on gene expression of *Arabidopsis* had revealed that the classical WRKY family may show considerable potential for alternative regulation in different tissues (28). We sought to address, more precisely, the relationship between expression diversity and evolutionary diversification within the plant WRKY–GCM1 domain superfamily. For this purpose, we considered both the TFs of the classical WRKY family as well as the two transposon-associated families, with the hope of understanding their tissue preferences in expression and uncovering possible roles for the transposon-derived transcripts as regulators.

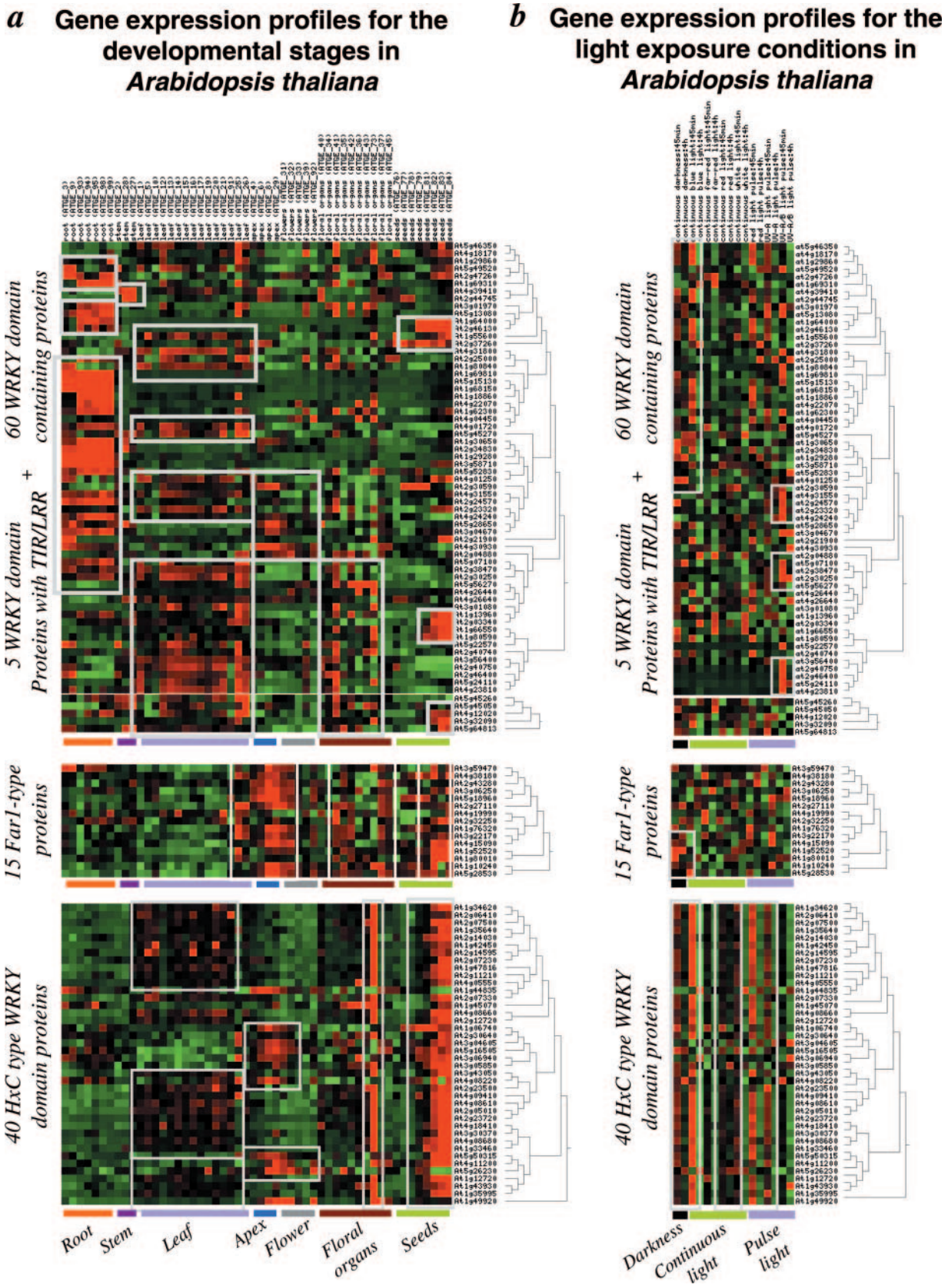
To investigate the gene expression patterns, we mined the recently deduced gene expression map of *Arabidopsis thaliana* development to obtain expression estimates from the different tissues and developmental stages (28). Additionally, we also used the expression dataset where plants were exposed to different light conditions (28), to understand the physiological effects under different illumination conditions on the gene expression patterns of this superfamily.

In both experiments, the authors had used the ATH1 Affymetrix array containing 22 746 probe sets which corresponds to >80% of the known genes in *Arabidopsis*. Expression estimates were available in triplicates from 79 different samples covering many stages (embryogenesis to senescence) and from diverse organs such as root, leaf, apex, tissue, etc. For the expression dataset pertaining to the different light conditions, gene expression estimates were available for eight different light conditions, with data for two different time points (45 min and 4 h) after exposure to a particular light condition (28).

A reasonably plain pattern which emerged within each of the three families of the WRKY–GCM1 domain proteins was a correlation between phylogenetic relationship and expression similarity (Figure 5). This was evident both from the clustering of similar expression states in an expression matrix ordered as per the neighbor-joining phylogenetic tree (Figure 5) as well as a general positive correlation observed within most of the tissue types between maximum-likelihood phylogenetic distance and expression similarity. Effectively, this implies that as the LSE of WRKY–GCM1 domains diversified in sequence, they potentially acquired new functions in different spatial and temporal segments of the physiological and developmental program. For instance, WRKY proteins that are primarily expressed in the root tend to be largely expressed in continuous darkness and in blue light (e.g. At2g34830, At1g30650; Figure 5) whereas the WRKY proteins which are largely expressed in the leaves and apex show expression in different light conditions such as UV-A/B light pulse and expressed at relatively lower levels in darkness (e.g. At4g23810, At4g30930; Figure 5). In part, the co-expression of phylogenetically close genes within the same tissue, similar developmental stage or physiological condition is suggestive of a strong tendency for direct backup for key regulatory functions. It might also additionally reflect some level of local functional specialization undergone by the family. Alternatively, it could simply be a result of neutral drift where recently duplicated proteins tend to retain similar expression patterns due of the lower number of mutations that could have accumulated in their regulatory regions. On the other hand, the partitioning of some of the phylogenetically closely related members in very different tissue types might also be seen. This suggests that occasionally there are major shifts in the expression pattern of genes with respect to their closest relatives. Such shifts might provide new regulatory switches in different tissue or developmental context suggesting a role in the evolutionary diversification of morphology or tissue physiology (Figure 5a).

We then used the comprehensive alignment of the classical WRKY proteins to assess the relationship between divergence in expression patterns and the residues predicted to be critical for DNA binding. This comparison revealed that some of the clusters with very distinct expression patterns (e.g. At5g13080, primarily expressed in the root and At1g64000, principally expressed in seed) share identical or very similar set of DNA contacting residues (Figures 2 and 5). This observation implies that a significant proportion of the lineage-specific expansion of the classical WRKY domains in *Arabidopsis* has primarily diversified in terms of their expression patterns rather than in their target DNA-binding sites. We also found that there were at least some members from each tissue-specific expression cluster that showed complementary expression pattern upon exposure to different light conditions, suggesting that regulatory elements upstream of duplicated genes, which are expressed in the same tissue have been fine-tuned such that a majority of the genes are expressed primarily in one of the two conditions (light or dark). Thus, the principal tinkering after the expansion does not appear to have happened at the level of DNA–protein interactions, but more likely the upstream regulatory elements of the classical WRKY genes.

Expression patterns of the two families of WRKY–GCM1 domains associated with *Arabidopsis* transposons show an interesting expression signal that is pretty uniform across the phylogenetic diversity of these transposases. Both show a strong tendency to avoid expression in the roots and stem. Additionally, the Far1-type of transposases belonging to the insert-containing family of WRKY–GCM1 domains, are also largely excluded from the leaves in course of their development. A subset of the H×C family is also similarly strongly excluded from the leaves. But, another subset shows a low level of expression. Far1-type WRKY–GCM1 genes tend to show a strong expression in developing apical tissues, whereas majority of the H×C family tend to show a strong expression in the pollen. The few H×C members which are not expressed in the pollen are instead expressed in apical tissues, such as the former family. Practically all Far1-type transposons show some expression in the ovules and female organs. Both families show a strong expression in later stages of germinating seeds (i.e. late torpedo to early walking-stick embryos, walking-stick to early curled cotyledons embryos, curled cotyledons to early green cotyledons embryos and green cotyledons embryos). One interpretation of this striking expression pattern of the selfish elements is that they are likely to maximize their propagation to the subsequent generation potentially by transposing to new sites in cells that are likely to spawn reproductive organs or gametes. A by-product of this phenomenon could be the utilization of the transposase DBDs as transcriptional regulators of plant genes in the conditions in which they are


a
Gene expression profiles for the developmental stages in Arabidopsis thaliana
60 WRKY domain containing proteins + 5 WRKY domain Proteins with TIR/LRR
15 Far1-type proteins
40 HxC type WRKY domain proteins
Root
Stem
Leaf
Apex
Flower
Floral organs
Seeds
b
Gene expression profiles for the light exposure conditions in Arabidopsis thaliana
60 WRKY domain containing proteins + 5 WRKY domain Proteins with TIR/LRR
15 Far1-type proteins
40 HxC type WRKY domain proteins
Darkness
Continuous light
Pulse light

strongly expressed. For example, transposons of the H×C family show reasonably strong expression in particular illuminating conditions, such as continuous darkness, continuous blue light, UV-light pulse and red-light pulse and could act as potential transcriptional regulators in these conditions (Figure 5b).

General implications of the relationship between transcription factors and transposases

The discovery of the evolutionary relationship between the classical WRKY TFs, GCM1 and Rcs1p/Aft2p with transposons adds to the growing set of connections found between transposons and TFs. Such a link appears to be an ancient one. In bacteria, the majority of transposases and resolvases utilize the different types of HTH DBDs. Consistent with this several bacterial TFs with HTH domain, including sigma factors such as Sigma-54, appear to have ultimately emerged from such DBDs of selfish elements (3,82). Similarly, a number of HTH domains of transposase origin have been utilized as TFs in eukaryotes, such as the Paired domain, which is a key developmental regulator in animals, Tigger DBD (seen in CENP-B) and the Pipsqueak (PSQ) domain (3,83–86). More recently, studies by others and us have suggested that the AP2-integrase DBD superfamily associated with different transposases and integrases of bacterial and eukaryotic selfish elements have spawned the principal TFs in both plants and apicomplexans (18,87,88). Similarly, the VP1 family of TFs which are massively expanded in plants appear to have acquired their distinctive DBD from prokaryotic selfish elements which include the mobile operons encoding restriction modification systems (22).

Eukaryotic selfish elements are distinctive in possessing multiple Zn-chelating DBDs, namely the BED, WRKY–GCM1 and THAP families of DBDs. This tendency to use Zn-chelating DBDs is consistent with the general overrepresentation of such metal-supported nucleic acid binding domains in eukaryotes. There is evidence, in each of the above superfamilies, for the derivation of cellular regulatory proteins from transposon-derived Zn-binding domains (19,20). Often, after acquisition of such a domain from a transposon source, there might be a LSE of the DBD in a particular eukaryotic lineage (19,20). The classical WRKY family of plants appears to represent one such example of a LSE following the emergence of an ancestral TF of this family from a transposon. This is also mirrored in the case of the THAP superfamily in which we observe two independent LSEs even with the closely related nematode species *C.elegans* and *Caenorhabditis briggsae*, respectively. The *C.elegans* specific expansion of THAP domains has spawned the developmental regulator lin-15A. It is also likely that in the WRKY–GCM1 superfamily, the FLYWCH and GCM versions emerged from a single seeding of the animal lineage by a transposon-derived domain. It is of interest to note that in insects, another lineage specifically expanded family of proteins which combine the POZ domain with a C2H2 Zn finger domain has on multiple occasions combined with DBDs of transposase origin such as the WRKY–GCM1 [mod(mdg4), PSQ type HTH (Pipsqueak) and BED finger (fruitless)]. This observation suggests that the recombination between members of LSEs and potential transposon-derived domains could be a means of generating new TFs.

The WRKY–GCM1 superfamily offers evidence for the phenomenon of multiple derivations of TFs from transposons within relatively short phylogenetic ranges. This phenomenon, best illustrated by the different fungal TFs of the WRKY–GCM1 domain superfamily, underscores the importance of transposons as a potential source for novel TFs in cellular genomes. In particular, the fact that transposons might regulate their own expression or transposition via the binding of specific internal sequences by their DBDs also hints at the possibility that they might provide the raw material for regulatory elements (89,90).

Conclusions

By means of systematic sequence and structure comparisons, we unify three major families of DBDs (WRKY, GCM1 and FLYWCH) with DBDs of two distinct families of the widely distributed MULE transposons of eukaryotes. We also show that several key fungal global or developmental regulatory proteins (Rcs1p, Rbf1p) belong to the WRKY–GCM1 superfamily. Through structural comparisons, we identified a conserved core of interacting residues which are likely to be the principal stabilizing elements of this fold. We also suggest that a probable evolutionary pathway for emergence of the WRKY–GCM1 superfamily was from classical C2H2 fingers (Znf) via an intermediate that was structurally close to the BED Zn-finger. Having identified the principal DNA-binding positions in the superfamily, we present evidence for diversification of target sites between different families although they are predicted to maintain an overall similarity in their mode of DNA contact.

By integrating the sequence and structure analysis with high-throughput ChIP-chip data and gene expression data, we show that two major fungal global regulators are likely to have risen to that status relatively recently during the evolution of saccharomycetale yeasts. This provides evidence for the rapid acquisition of regulatory interactions in eukaryotic transcriptional regulatory networks. In the case of the plant WRKY family, we demonstrate that their LSE has acquired functional diversity mainly through expression divergence rather than acquisition of a wide array of DNA-binding specificities. Finally, we also use the WRKY–GCM1 superfamily as an example to illustrate the significant role of transposons in the emergence of new

Figure 5. Gene expression profile for the WRKY domain containing proteins from *Arabidopsis thaliana* across (**a**) different developmental stages and organs. The samples are ordered according to the organs (root, leaf, apex, flowers, floral organs and seeds; asterisk denotes pollen within the floral organs category), and progressively from embryogenesis to senescence. The WRKY domain containing genes shown on the right as rows have been ordered according to their evolutionary relationship as obtained from using the similarity between their sequences. The neighbor-joining tree was obtained using the distances calculated according to the JTT distance matrix in the MEGA package. Boxes denote similarly expressed tissue-specific clusters of organs and genes. (**b**) Different light exposures with two time points for each condition. The samples are ordered according to the dark/light source (continuous darkness, continuous blue light, continuous far-red light, continuous red light, continuous white light, pulse of red light, pulse of UV-A light and pulse of UV-A/B light), one after 45 min and another after 4 h of exposure. The WRKY domain containing genes are ordered in the same way. Boxes denote clusters of genes which show high expression for a give light condition or in darkness. The gene expression data were obtained from Schmid *et al.* (29), and the expression matrix was generated using matrix2png.

TFs. We hope that the findings presented here would serve as a platform for future investigations into the evolutionary process of TF innovation in eukaryotes.

NOTE ADDED IN PROOF

While this paper was being prepared for publication the genome of the chlorophyte alga, Ostreococcus tauri, an early branching representative of the viridiplantae lineage, became available. Analysis of this genome showed that it contained 3 proteins, each with a single classical WRKY domain. This suggests that the classical WRKY domain had emerged prior to the diversification of the viridiplantae lineage, but underwent its lineage specific expansion only in the line leading to the embryophytes or classical plants.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online and at http://www.mrc-lmb.cam.ac.uk/genomes/madanm/wrky-gcm1/.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the Intramural research program of National Institutes of Health, USA for funding their research. Funding to pay the Open Access publication charges for this article was provided by the National Institutes of Health, USA.

Conflict of interest statement. None declared.

REFERENCES

1. Lespinet,O., Wolf,Y.I., Koonin,E.V. and Aravind,L. (2002) The role of lineage-specific gene family expansion in the evolution of eukaryotes. *Genome Res.*, **12**, 1048–1059.
2. Lindahl,L. and Hinnebusch,A. (1992) Diversity of mechanisms in the regulation of translation in prokaryotes and lower eukaryotes. *Curr. Opin. Genet. Dev.*, **2**, 720–726.
3. Aravind,L., Anantharaman,V., Balaji,S., Babu,M.M. and Iyer,L.M. (2005) The many faces of the helix–turn–helix domain: transcription regulation and beyond. *FEMS Microbiol Rev.*, **29**, 231–262.
4. Babu,M.M., Luscombe,N.M., Aravind,L., Gerstein,M. and Teichmann,S.A. (2004) Structure and evolution of transcriptional regulatory networks. *Curr. Opin. Struct. Biol.*, **14**, 283–291.
5. Aravind,L. and Koonin,E.V. (1999) DNA-binding proteins and evolution of transcription regulation in the archaea. *Nucleic Acids Res.*, **27**, 4658–4670.
6. Ranea,J.A., Buchan,D.W., Thornton,J.M. and Orengo,C.A. (2004) Evolution of protein superfamilies and bacterial genome size. *J. Mol. Biol.*, **336**, 871–887.
7. Madan Babu,M., Teichmann,S.A. and Aravind,L. (2006) Evolutionary dynamics of prokaryotic transcriptional regulatory networks. *J. Mol. Biol.*, **358**, 614–633.
8. Ivens,A.C., Peacock,C.S., Worthey,E.A., Murphy,L., Aggarwal,G., Berriman,M., Sisk,E., Rajandream,M.A., Adlem,E., Aert,R. *et al.* (2005) The genome of the kinetoplastid parasite, *Leishmania major*. *Science*, **309**, 436–442.
9. Templeton,T.J., Iyer,L.M., Anantharaman,V., Enomoto,S., Abrahante,J.E., Subramanian,G.M., Hoffman,S.L., Abrahamsen,M.S. and Aravind,L. (2004) Comparative analysis of apicomplexa and genomic diversity in eukaryotes. *Genome Res.*, **14**, 1686–1695.
10. Phillips,K. and Luisi,B. (2000) The virtuoso of versatility: POU proteins that flex to fit. *J. Mol. Biol.*, **302**, 1023–1039.
11. Aravind,L., Iyer,L.M. and Koonin,E.V. (2006) Comparative genomics and structural biology of the molecular innovations of eukaryotes. *Curr. Opin. Struct. Biol.*, **16**, 409–419.
12. Krishna,S.S., Majumdar,I. and Grishin,N.V. (2003) Structural classification of zinc fingers: survey and summary. *Nucleic Acids Res.*, **31**, 532–550.
13. Chervitz,S.A., Aravind,L., Sherlock,G., Ball,C.A., Koonin,E.V., Dwight,S.S., Harris,M.A., Dolinski,K., Mohr,S., Smith,T. *et al.* (1998) Comparison of the complete protein sets of worm and yeast: orthology and divergence. *Science*, **282**, 2022–2028.
14. Sluder,A.E., Mathews,S.W., Hough,D., Yin,V.P. and Maina,C.V. (1999) The nuclear receptor superfamily has undergone extensive proliferation and diversification in nematodes. *Genome Res.*, **9**, 103–120.
15. Riechmann,J.L., Heard,J., Martin,G., Reuber,L., Jiang,C., Keddie,J., Adam,L., Pineda,O., Ratcliffe,O.J., Samaha,R.R. *et al.* (2000) Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. *Science*, **290**, 2105–2110.
16. Beja,O., Koonin,E.V., Aravind,L., Taylor,L.T., Seitz,H., Stein,J.L., Bensen,D.C., Feldman,R.A., Swanson,R.V. and DeLong,E.F. (2002) Comparative genomic analysis of archaeal genotypic variants in a single population and in two different oceanic provinces. *Appl. Environ. Microbiol.*, **68**, 335–345.
17. Aravind,L., Makarova,K.S. and Koonin,E.V. (2000) Holliday junction resolvases and related nucleases: identification of new families, phyletic distribution and evolutionary trajectories. *Nucleic Acids Res.*, **28**, 3417–3432.
18. Balaji,S., Babu,M.M., Iyer,L.M. and Aravind,L. (2005) Discovery of the principal specific transcription factors of Apicomplexa and their implication for the evolution of the AP2-integrase DNA binding domains. *Nucleic Acids Res.*, **33**, 3994–4006.
19. Aravind,L. (2000) The BED finger, a novel DNA-binding domain in chromatin-boundary-element-binding proteins and transposases. *Trends Biochem. Sci.*, **25**, 421–423.
20. Roussigne,M., Kossida,S., Lavigne,A.C., Clouaire,T., Ecochard,V., Glories,A., Amalric,F. and Girard,J.P. (2003) The THAP domain: a novel protein motif with similarity to the DNA-binding domain of P element transposase. *Trends Biochem. Sci.*, **28**, 66–69.
21. Breitling,R. and Gerber,J.K. (2000) Origin of the paired domain. *Dev. Genes Evol.*, **210**, 644–650.
22. Yamasaki,K., Kigawa,T., Inoue,M., Tateno,M., Yamasaki,T., Yabuki,T., Aoki,M., Seki,E., Matsuda,T., Tomo,Y. *et al.* (2004) Solution structure of the B3 DNA binding domain of the *Arabidopsis* cold-responsive transcription factor RAV1. *Plant Cell*, **16**, 3448–3459.
23. Wu,K.L., Guo,Z.J., Wang,H.H. and Li,J. (2005) The WRKY family of transcription factors in rice and *Arabidopsis* and their origins. *DNA Res.*, **12**, 9–26.
24. Zhang,Y. and Wang,L. (2005) The WRKY transcription factor superfamily: its origin in eukaryotes and expansion in plants. *BMC Evol. Biol.*, **5**, 1.
25. Eichinger,L., Pachebat,J.A., Glockner,G., Rajandream,M.A., Sucgang,R., Berriman,M., Song,J., Olsen,R., Szafranski,K., Xu,Q. *et al.* (2005) The genome of the social amoeba *Dictyostelium discoideum*. *Nature*, **435**, 43–57.
26. Yamasaki,K., Kigawa,T., Inoue,M., Tateno,M., Yamasaki,T., Yabuki,T., Aoki,M., Seki,E., Matsuda,T., Tomo,Y. *et al.* (2005) Solution structure of an Arabidopsis WRKY DNA binding domain. *Plant Cell*, **17**, 944–956.
27. Cohen,S.X., Moulin,M., Hashemolhosseini,S., Kilian,K., Wegner,M. and Muller,C.W. (2003) Structure of the GCM domain-DNA complex: a DNA-binding domain with a novel fold and mode of target site recognition. *EMBO J.*, **22**, 1835–1845.
28. Schmid,M., Davison,T.S., Henz,S.R., Pape,U.J., Demar,M., Vingron,M., Scholkopf,B., Weigel,D. and Lohmann,J.U. (2005) A gene expression map of *Arabidopsis thaliana* development. *Nature Genet.*, **37**, 501–506.
29. Lee,T.I., Rinaldi,N.J., Robert,F., Odom,D.T., Bar-Joseph,Z., Gerber,G.K., Hannett,N.M., Harbison,C.T., Thompson,C.M., Simon,I. *et al.* (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, **298**, 799–804.
30. Harbison,C.T., Gordon,D.B., Lee,T.I., Rinaldi,N.J., Macisaac,K.D., Danford,T.W., Hannett,N.M., Tagne,J.B., Reynolds,D.B., Yoo,J. *et al.*

(2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99–104.
31. Borneman,A.R., Leigh-Bell,J.A., Yu,H., Bertone,P., Gerstein,M. and Snyder,M. (2006) Target hub proteins serve as master regulators of development in yeast. *Genes Dev.*, **20**, 435–448.
32. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
33. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
34. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
35. Schaffer,A.A., Aravind,L., Madden,T.L., Shavirin,S., Spouge,J.L., Wolf,Y.I., Koonin,E.V. and Altschul,S.F. (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.*, **29**, 2994–3005.
36. Wootton,J.C. (1994) Non-globular domains in protein sequences: automated segmentation using complexity measures. *Comput. Chem.*, **18**, 269–285.
37. Notredame,C., Higgins,D.G. and Heringa,J. (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, **302**, 205–217.
38. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
39. Walker,D.R. and Koonin,E.V. (1997) SEALS: a system for easy analysis of lots of sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **5**, 333–339.
40. Cuff,J.A., Clamp,M.E., Siddiqui,A.S., Finlay,M. and Barton,G.J. (1998) JPred: a consensus secondary structure prediction server. *Bioinformatics*, **14**, 892–893.
41. Guex,N. and Peitsch,M.C. (1997) SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis*, **18**, 2714–2723.
42. Holm,L. and Sander,C. (1995) Dali: a network tool for protein structure comparison. *Trends Biochem. Sci.*, **20**, 478–480.
43. Krissinel,E. and Henrick,K. (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta. Crystallogr. D Biol. Crystallogr.*, **60**, 2256–2268.
44. Holm,L. and Sander,C. (1996) The FSSP database: fold classification based on structure–structure alignment of proteins. *Nucleic Acids Res.*, **24**, 206–209.
45. Kumar,S., Tamura,K. and Nei,M. (2004) MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Brief Bioinformatics*, **5**, 150–163.
46. Pavlidis,P. and Noble,W.S. (2003) Matrix2png: a utility for visualizing matrix data. *Bioinformatics*, **19**, 295–296.
47. Horak,C.E., Luscombe,N.M., Qian,J., Bertone,P., Piccirrillo,S., Gerstein,M. and Snyder,M. (2002) Complex transcriptional circuitry at the G_1/S transition in *Saccharomyces cerevisiae*. *Genes Dev.*, **16**, 3017–3033.
48. Svetlov,V.V. and Cooper,T.G. (1995) Review: compilation and characteristics of dedicated transcription factors in *Saccharomyces cerevisiae*. *Yeast*, **11**, 1439–1484.
49. Teichmann,S.A. and Babu,M.M. (2004) Gene regulatory network growth by duplication. *Nature Genet.*, **36**, 492–496.
50. Luscombe,N.M., Babu,M.M., Yu,H., Snyder,M., Teichmann,S.A. and Gerstein,M. (2004) Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, **431**, 308–312.
51. Ernst,H.A., Olsen,A.N., Larsen,S. and Lo Leggio,L. (2004) Structure of the conserved domain of ANAC, a member of the NAC family of transcription factors. *EMBO Rep.*, **5**, 297–303.
52. Ryan,R.F. and Darby,M.K. (1998) The role of zinc finger linkers in p43 and TFIIIA binding to 5S rRNA and DNA. *Nucleic Acids Res.*, **26**, 703–709.
53. Dorn,R. and Krauss,V. (2003) The modifier of mdg4 locus in *Drosophila*: functional complexity is resolved by trans splicing. *Genetica*, **117**, 165–177.
54. Krauss,V. and Dorn,R. (2004) Evolution of the *trans*-splicing *Drosophila* locus mod(mdg4) in several species of Diptera and Lepidoptera. *Gene*, **331**, 165–176.
55. Thatcher,J.D., Fernandez,A.P., Beaster-Jones,L., Haun,C. and Okkema,P.G. (2001) The *Caenorhabditis elegans peb-1* gene encodes a novel DNA-binding protein involved in morphogenesis of the pharynx, vulva, and hindgut. *Dev. Biol.*, **229**, 480–493.
56. Beaster-Jones,L. and Okkema,P.G. (2004) DNA binding and *in vivo* function of *C.elegans* PEB-1 require a conserved FLYWCH motif. *J. Mol. Biol.*, **339**, 695–706.
57. Ulker,B. and Somssich,I.E. (2004) WRKY transcription factors: from DNA binding towards biological function. *Curr. Opin. Plant. Biol.*, **7**, 491–498.
58. Eulgem,T., Rushton,P.J., Robatzek,S. and Somssich,I.E. (2000) The WRKY superfamily of plant transcription factors. *Trends Plant. Sci.*, **5**, 199–206.
59. Kalde,M., Barth,M., Somssich,I.E. and Lippok,B. (2003) Members of the Arabidopsis WRKY group III transcription factors are part of different plant defense signaling pathways. *Mol. Plant Microbe Interact.*, **16**, 295–305.
60. Dong,J., Chen,C. and Chen,Z. (2003) Expression profiles of the *Arabidopsis* WRKY gene superfamily during plant defense response. *Plant Mol. Biol.*, **51**, 21–37.
61. Liu,D., Wang,Y.S. and Wyss,D.F. (2003) Solution structure of the hypothetical protein YqgF from *Escherichia coli* reveals an RNAse H fold. *J. Biomol. NMR*, **27**, 389–392.
62. Hickman,A.B., Perez,Z.N., Zhou,L., Musingarimi,P., Ghirlando,R., Hinshaw,J.E., Craig,N.L. and Dyda,F. (2005) Molecular architecture of a eukaryotic DNA transposase. *Nature Struct. Mol. Biol.*, **12**, 715–721.
63. Rice,P. and Mizuuchi,K. (1995) Structure of the bacteriophage Mu transposase core: a common structural motif for DNA transposition and retroviral integration. *Cell*, **82**, 209–220.
64. Richardson,J.M., Dawson,A., O'Hagan,N., Taylor,P., Finnegan,D.J. and Walkinshaw,M.D. (2006) Mechanism of Mos1 transposition: insights from structural analysis. *EMBO J.*, **25**, 1324–1334.
65. Rice,P.A. and Baker,T.A. (2001) Comparative architecture of transposase and integrase complexes. *Nature Struct Biol.*, **8**, 302–307.
66. Aravind,L. and Landsman,D. (1998) AT-hook motifs identified in a wide variety of DNA-binding proteins. *Nucleic Acids Res.*, **26**, 4413–4421.
67. Makarova,K.S., Aravind,L. and Koonin,E.V. (2002) SWIM, a novel Zn-chelating domain present in bacteria, archaea and eukaryotes. *Trends Biochem. Sci.*, **27**, 384–386.
68. Jones,B.W., Fetter,R.D., Tear,G. and Goodman,C.S. (1995) Glial cells missing: a genetic switch that controls glial versus neuronal fate. *Cell*, **82**, 1013–1023.
69. Anson-Cartwright,L., Dawson,K., Holmyard,D., Fisher,S.J., Lazzarini,R.A. and Cross,J.C. (2000) The glial cells missing-1 protein is essential for branching morphogenesis in the chorioallantoic placenta. *Nature Genet.*, **25**, 311–314.
70. Olsen,A.N., Ernst,H.A., Leggio,L.L. and Skriver,K. (2005) NAC transcription factors: structurally distinct, functionally diverse. *Trends Plant Sci.*, **10**, 79–87.
71. Balaji,S., Babu,M.M., Iyer,L.M., Luscombe,N.M. and Aravind,L. (2006) Comprehensive analysis of combinatorial regulation using the transcriptional regulatory network of yeast. *J. Mol. Biol*, **360**, 213–227.
72. Blaiseau,P.L., Lesuisse,E. and Camadro,J.M. (2001) Aft2p, a novel iron-regulated transcription activator that modulates, with Aft1p, intracellular iron use and resistance to oxidative stress in yeast. *J. Biol. Chem.*, **276**, 34221–34226.
73. Rutherford,J.C., Jaron,S., Ray,E., Brown,P.O. and Winge,D.R. (2001) A second iron-regulatory system in yeast independent of Aft1p. *Proc. Natl Acad. Sci. USA*, **98**, 14322–14327.
74. Gil,R., Zueco,J., Sentandreu,R. and Herrero,E. (1991) RCS1, a gene involved in controlling cell size in *Saccharomyces cerevisiae*. *Yeast*, **7**, 1–14.
75. Yamaguchi-Iwai,Y., Dancis,A. and Klausner,R.D. (1995) AFT1: a mediator of iron regulated transcriptional control in *Saccharomyces cerevisiae*. *EMBO J.*, **14**, 1231–1239.
76. Ishii,N., Yamamoto,M., Yoshihara,F., Arisawa,M. and Aoki,Y. (1997) Biochemical and genetic characterization of Rbf1p, a putative transcription factor of *Candida albicans*. *Microbiology*, **143**, 429–435.
77. Ishii,N., Yamamoto,M., Lahm,H.W., Iizumi,S., Yoshihara,F., Nakayama,H., Arisawa,M. and Aoki,Y. (1997) A DNA-binding protein from *Candida albicans* that binds to the RPG box of *Saccharomyces*

cerevisiae and the telomeric repeat sequence of *C. albicans*. *Microbiology*, **143**, 417–427.
78. Balaji,S., Iyer,L.M., Aravind,L. and Babu,M.M. (2006) Uncovering a hidden distributed architecture behind scale-free transcriptional regulatory networks. *J. Mol. Biol*, **320**, 204–212.
79. Raizada,M.N., Brewer,K.V. and Walbot,V. (2001) A maize MuDR transposon promoter shows limited autoregulation. *Mol. Genet. Genomics*, **265**, 82–94.
80. Rudenko,G.N. and Walbot,V. (2001) Expression and post-transcriptional regulation of maize transposable element MuDR and its derivatives. *Plant Cell*, **13**, 553–570.
81. Benito,M.I. and Walbot,V. (1997) Characterization of the maize Mutator transposable element MURA transposase as a DNA-binding protein. *Mol. Cell. Biol*., **17**, 5165–5175.
82. Pietrokovski,S. and Henikoff,S. (1997) A helix–turn–helix DNA-binding motif predicted for transposases of DNA transposons. *Mol. Gen. Genet*., **254**, 689–695.
83. Izsvak,Z., Khare,D., Behlke,J., Heinemann,U., Plasterk,R.H. and Ivics,Z. (2002) Involvement of a bifunctional, paired-like DNA-binding domain and a transpositional enhancer in Sleeping Beauty transposition. *J. Biol. Chem*., **277**, 34581–34588.
84. Smit,A.F. and Riggs,A.D. (1996) Tiggers and DNA transposon fossils in the human genome. *Proc. Natl Acad. Sci. USA*, **93**, 1443–1448.
85. Czerny,T., Schaffner,G. and Busslinger,M. (1993) DNA sequence recognition by Pax proteins: bipartite structure of the paired domain and its binding site. *Genes Dev*., **7**, 2048–2061.
86. Tanaka,Y., Nureki,O., Kurumizaka,H., Fukai,S., Kawaguchi,S., Ikuta,M., Iwahara,J., Okazaki,T. and Yokoyama,S. (2001) Crystal structure of the CENP-B protein–DNA complex: the DNA-binding domains of CENP-B induce kinks in the CENP-B box DNA. *EMBO J*., **20**, 6612–6618.
87. Fujimoto,S.Y., Ohta,M., Usui,A., Shinshi,H. and Ohme-Takagi,M. (2000) *Arabidopsis* ethylene-responsive element binding factors act as transcriptional activators or repressors of GCC box-mediated gene expression. *Plant Cell*, **12**, 393–404.
88. Magnani,E., Sjolander,K. and Hake,S. (2004) From endonucleases to transcription factors: evolution of the AP2 DNA binding domain in plants. *Plant Cell*, **16**, 2265–2277.
89. Marino-Ramirez,L., Lewis,K.C., Landsman,D. and Jordan,I.K. (2005) Transposable elements donate lineage-specific regulatory sequences to host genomes. *Cytogenet. Genome Res*., **110**, 333–341.
90. Jordan,I.K., Rogozin,I.B., Glazko,G.V. and Koonin,E.V. (2003) Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends Genet*., **19**, 68–72.