Arabidopsis Transcription Factors: Genome-Wide Comparative Analysis Among Eukaryotes

J. L. Riechmann,* J. Heard, G. Martin, L. Reuber, C.-Z. Jiang, J. Keddie, L. Adam, O. Pineda, O. J. Ratcliffe, R. R. Samaha, R. Creelman, M. Pilgrim, P. Broun, J. Z. Zhang, D. Ghandehari, B. K. Sherman, G.-L. Yu

The completion of the *Arabidopsis thaliana* genome sequence allows a comparative analysis of transcriptional regulators across the three eukaryotic kingdoms. *Arabidopsis* dedicates over 5% of its genome to code for more than 1500 transcription factors, about 45% of which are from families specific to plants. *Arabidopsis* transcription factors that belong to families common to all eukaryotes do not share significant similarity with those of the other kingdoms beyond the conserved DNA binding domains, many of which have been arranged in combinations specific to each lineage. The genome-wide comparison reveals the evolutionary generation of diversity in the regulation of transcription.

Regulation of gene expression at the level of transcription influences or controls many of the biological processes in a cell or organism, such as progression through the cell cycle, metabolic and physiological balance, and responses to the environment. Development is based on the cellular capacity for differential gene expression and is often controlled by transcription factors acting as switches of regulatory cascades (*1*). In addition, alterations in the expression of genes coding for transcriptional regulators are emerging as a major source of the diversity and change that underlie evolution (*2*).

With the completion of the *Arabidopsis thaliana* genome sequence, the entire complement of genes coding for transcription factors from a plant can be identified and described. Together with the three other eukaryotic genomes that have already been sequenced, it also allows investigation of the similarities and differences in transcriptional regulators among the three eukaryotic kingdoms: plants, animals (*Caenorhabditis elegans* and *Drosophila melanogaster*) (*3, 4*), and fungi (*Saccharomyces cerevisiae*) (*5*). We present such a description and analysis here.

Gene Content and Organization

To characterize the entire complement of transcription factors encoded by the genomes of *Arabidopsis*, *Drosophila*, *C. elegans*, and *S. cerevisiae*, we used a comprehensive list of proteins, domains, and motifs to query the corresponding sequence databases. Transcription factors are usually defined as proteins that show sequence-specific DNA binding and are capable of activating and/or repressing transcription. Although most of the proteins and protein families that were considered in our study fit these criteria, we have also included some other types of transcriptional regulators. Most known transcription factors can be grouped into families according to their DNA binding domain (*6*). Protein domains that are sometimes present in transcription factors, but not necessarily associated with them, have not been included in this genome survey, for example, some zinc coordinating motifs that either are involved in protein-protein interactions or have not yet been functionally characterized.

We searched the *Drosophila*, *C. elegans*, and yeast encoded protein complements (proteomes) using BLAST and motif-finding programs (*7*). Because the complete predicted proteome of *Arabidopsis* was not available at the time of the analysis, we used the entire set of genomic sequences (*7*).

The *Arabidopsis* genome codes for at least 1533 transcriptional regulators, which account for ~5.9% of its estimated total number of genes (Table 1). We identified 635, 669, and 209 transcriptional regulators in the proteomes of *Drosophila*, *C. elegans*, and yeast, respectively (4.5, 3.5, and 3.5%). Thus, the *Arabidopsis* content of transcription factors is 1.3 times that of *Drosophila* and 1.7 times that of *C. elegans* and yeast. These results represent an underestimate of the total number of transcription factors in these organisms. Approximately 40 to 50% of the proteins encoded by each of those genomes cannot be assigned to functional categories on the basis of sequence similarity to proteins of known function (*3*, *8*–*11*). Some of those uncharacterized proteins are expected to be transcriptional regulators (*12, 13*). The large number and diversity of transcription factors in *Drosophila* were proposed to be related to its substantial regulatory complexity (*4*). Applying the same logic to *Arabidopsis* suggests that the regulation of transcription in plants is as complex as that in *Drosophila*. In contrast to *Drosophila* and *C. elegans*, for which a sizable (>25%) fraction of their known transcription factors have been characterized genetically (*14*), only ~5% of those from *Arabidopsis* have been defined by mutation analysis (*15*).

Arabidopsis contains many tandem gene duplications and large-scale duplications on different chromosomes, which might account for >60% of the genome (*9, 10, 16*). Whereas some of these duplications have been followed by rearrangements and divergent evolution, up to 40% of the *Arabidopsis* genes might comprise pairs of highly related sequences (*16*). In that respect, *Arabidopsis* is similar to the three other eukaryotic organisms. The *S. cerevisiae* genome is the result of a complete ancient genome duplication that was followed by extensive gene rearrangements and deletions (*17*). In yeast, ~30% of the genes form duplicate gene pairs. Similarly, duplicated genes account for ~48 and ~40% of the total gene content of *C. elegans* and *Drosophila*, respectively (*11*).

All of the *Arabidopsis* transcription factor gene families are scattered throughout the genome. On average, closely related genes account for ~45% of the total number in the major families (Table 2) (*18*). Gene duplications on different chromosomes are most common (~65%), but duplicated genes are also frequently found at large distances in the same chromosome (~22%) as well as organized in tandem repeats (~13%) (*19*). Clusters of three or more highly related genes are very rare (Table 2).

Transcription Factors Across the Eukaryotic Kingdoms

Two features stand out when comparing the *Arabidopsis* complement of transcriptional regulators with that of the other organisms (Table 3). First, <22% of the *Arabidopsis* transcription factors are zinc-coordinating proteins [belonging to several different families that are thought to have evolved independently (*20*)]. In contrast, zinc-coordinating proteins constitute most of the transcription factors in the three other eukaryotes: ~51% in *Drosophila*, ~64% in *C. elegans*, and 56% in yeast. Second, in *Arabidopsis*, there is no single family of transcription factors that has been so disproportionately am-

Mendel Biotechnology, 21375 Cabot Boulevard, Hayward, CA 94545, USA.

*To whom correspondence should be addressed. E-mail: jriechmann@mendelbio.com

plified as the nuclear hormone receptors in *C. elegans* (~38% of its transcription factors), the C2H2 zinc finger proteins in *Drosophila* (~46%), or the C6 and C2H2 families in yeast (~25% each one). The three largest families of transcription factors in *Arabidopsis*, AP2/EREBP (APETALA2/ethylene responsive element binding protein), MYB-(R1)R2R3, and bHLH (basic helix-loop-helix), each represent only ~9% of the total, and there are several other families with comparable numbers of genes.

Each eukaryotic lineage has its own set of particular transcription factor families and genes [comparing such a small number of genomes represents a limitation for this type of analysis (*21*)] (Table 3). The lineage-specific families are of interest from an evolutionary point of view. According to molecular phylogenetic analyses, plants, animals, and fungi all diverged from a common ancestor during a short period of time, ~1.5 billion years ago (*15*). Thus, it would be expected that most of the transcription factor families would either be shared by the three lineages, if they were present in the common ancestor, or specific to each lineage, if they arose independently following divergence. This is indeed the case (Table 3). Members of lineage-specific families represent 45% of the *Arabidopsis* transcription factors, 47% in *C. elegans*, and 32% in yeast (but only 14% in *Drosophila*, because of its extensive use of the C2H2 zinc finger proteins). Families that are present in all four organisms account for most of the remaining transcription factors in each case.

There are, however, a few exceptions to this expected pattern: some genes and gene families are present in two of the three lineages. Transcription factors and transcription factor families that are present in *Drosophila*, *C. elegans*, and yeast (but are absent from *Arabidopsis*) include the SOX/TCF (SRY-related HMG box/T cell factor) group, the fork head–type/winged-helix proteins, and homologs of the human transcription factor RFX1 (Table 3). The SOX/TCF group, which includes developmental regulators like human SRY (sex-determining region Y) and TCF and the yeast hypoxic-gene regulator ROX1, forms part of the HMG-box (high-mobility group) superfamily of proteins (*22*). In contrast to other HMG-box proteins that act as architectural components of chromatin and have no sequence specificity on their own, the SOX/TCF factors show sequence-specific DNA binding and transactivation activities. There are 14 genes in the *Arabidopsis* genome encoding HMG box–containing proteins, but phylogenetic analyses indicate that none of these proteins belong to the SOX/TCF group (*15*).

In contrast to the examples described above, there does not appear to be any case of transcriptional regulators that are present in both yeast and *Arabidopsis* but absent from animals. This distribution of genes and gene families in the three eukaryotic lineages is in agreement with the notion that animals and fungi are more closely related to each other than to plants (*23*). There are at least three classes of transcription factors that are present in plants and animals but absent from yeast: TUBBY-like (TUB), CPP-like (cysteine-rich polycomb-like protein), and E2F/DP proteins (*13*, *24*, *25*) (Table 3). It remains to be determined whether these classes of genes were specifically lost from the *S. cerevisiae* genome or if they are really absent from the fungal lineage.

There are many transcription factor families that are found only in plants, some of which have been greatly amplified. These include the AP2/EREBP (*26*), NAC (*27*), and WRKY families (*28*); the trihelix DNA binding proteins (*29*); the auxin response factors (ARFs); the Aux/IAA proteins [which do not bind to DNA by themselves, but interact with the ARF proteins (*30*)]; and other smaller families (Table 3). Similarly, animals and yeast have many families of transcription factors that are not found in plants (Table 3).

A lingering question when considering protein families that appear to be exclusive to one lineage is whether their signature domains are true evolutionary innovations or whether their relationships with other proteins have been blurred because their amino acid sequences (but not their three-dimensional structures) have diverged substantially over time. Some of the plant-specific families of transcriptional regulators are characterized by domains that appear to be genuine novelties. For example, the AP2 domain exhibits a new mode of DNA recognition by a β-sheet structure (*31*). Other transcription factors classified as specific to plants, however, might be related to proteins found in other organisms. The plant-specific GRAS proteins might be distant relatives of the animal-spe-

Table 1. Content of transcriptional regulator genes in eukaryotic genomes. The number of genes in each of the eukaryotic genomes is given as an approximate number. This is because the number of genes predicted at the time that a genome is sequenced is always an estimate that is refined over time (*7*).

Organism	Total number of genes	Genes coding for transcriptional regulators	
		Total number	Percentage of total number of genes
A. thaliana	~26,000	1533	5.9
S. cerevisiae	~6,000	209	3.5
C. elegans	~19,000	669	3.5
D. melanogaster	~14,000	635	4.5

Table 2. Gene duplications in *Arabidopsis* transcription factor families. The major families of *Arabidopsis* transcription factors were analyzed for the presence of pairs or groups of highly related genes (*18*). The families analyzed together comprise over 1000 genes. Tandem duplications are arbitrarily defined as those that occur within a sequence distance of 50 kb. If two genes are duplicated in the same chromosome but reside >50 kb apart from each other, they are counted in the "Duplications in the same chromosome" column. (Zn) indicates a zinc coordinating DNA binding motif.

Gene family	Percentage of genes with close homolog	Tandem duplications (%)	Duplications in same chromosome (%)	Duplications in different chromosomes (%)	Number of gene clusters/ number of genes in cluster (chromosome)
MYB-(R1)R2R3	44	7	28	65	0
AP2/EREBP	45	11	39	50	1/3 (4)
bHLH	42	13	13	74	0
NAC	42	27	10	63	1/5 (1)
C2H2 (Zn)	40	9	23	68	1/3 (3)
HB	50	5	24	71	0
MADS	50	30	32	38	1/4 (5)
bZIP	53	13	22	65	1/3 (5)
WRKY (Zn)	33	12	17	71	1/3 (1)
GARP	48	0	8	92	0
Dof (Zn)	37	33	17	50	1/4 (4)
CO-like (Zn)	52	13	13	74	0
GATA (Zn)	50	0	0	100	0
Total	44	13	22	65	NA

cific STATS, based on a similar arrangement of related functional domains (*32*). The trihelix DNA-binding domain, present only in plants, might have evolved from the MYB domain, found in all eukaryotes (*29*).

The two transcription factor families that have been more substantially amplified in *Arabidopsis*, as compared to animals and yeast, are the MYB and the MADS families. The MYB motif consists of a helix-turn-helix structure with three regularly spaced Trp residues. In *Arabidopsis*, almost all of the MYB proteins belong to the MYB-R2R3 class (131 members): they contain two imperfect repeats of the MYB motif (*33*). MYB-R1R2R3 proteins, which are the norm in animals, are rare in *Arabidopsis* (five proteins). The plant-specific R2R3 organization is thought to have evolved from an R1R2R3-type ancestral gene from which the first repeat was lost (*34*). Because the plant MYB-R1R2R3 proteins are more closely related to the animal MYB proteins than to the plant proteins of the R2R3 type, it has been suggested that they might have functions related to those of the MYB proteins in animals, such as the control of cell proliferation (*34, 35*). Conversely, MYB-R2R3 proteins might have evolved to regulate processes specific to plants, including secondary metabolism, responses to plant hormones, and the identity of specific cell types.

In addition to the MYB-(R1)R2R3 proteins, *Arabidopsis* contains additional transcription factors characterized by a more divergent MYB domain, which is present either as a single copy or as a repeat. These proteins form a heterogeneous group and are often referred to as "MYB related." For the purpose of clarity, we have divided the *Arabidopsis* MYB-related proteins into several subclasses in Fig. 1 (*15*).

More distant but also related to the MYB superfamily is a previously unidentified group of proteins that we propose to name "GARP," after maize GOLDEN2, the ARR B-class proteins from *Arabidopsis*, and *Chlamydomonas* Psr1 (*36–39*) (Fig. 1). These proteins appear to be involved in plant-specific processes: GOLDEN2 controls the differentiation of a photosynthetic cell type of the maize leaf, whereas Psr1 is a regulator of phosphorus metabolism.

Arabidopsis also contains many more heat

Table 3. Eukaryotic transcriptional regulators. Number of transcriptional regulators in *Arabidopsis* (*A.t.*), *Drosophila* (*D.m.*), *C. elegans* (*C.e.*), and *S. cerevisiae* (*S.c.*), classified by families on the basis of sequence similarity. The table is nonredundant: proteins are counted only once, regardless of whether they have more than one signature motif. The way in which proteins combine different DNA binding motifs were organized into families is reflected in Fig. 1. Families that are specific to one lineage are indicated in color. Families are listed under "Transcription factors" or "Other transcriptional regulators," as described in the text. However, this distinction is not without problems (for example, the ARID and HMG-box families). Information about the signature motif(s) or sequences that define each family is provided as an InterPro (IPR) or GenBank accession number (*56*). (Zn) indicates a zinc coordinating DNA binding motif. In the bHLH class, only proteins with a discernible basic region were included. "Other" includes some single-copy genes and small families that are not individually mentioned in the text. The results of the database searches (P, motif searches; B, BLAST) and sequence comparisons were inspected by eye. The numbers reported here might therefore differ from other large-scale classifications that are performed automatically (*11*).

Gene family	Predicted # proteins *A.t.*	*D.m.*	*C.e.*	*S.c.*	InterPro or GenBank	Search
Transcription factors						
MYB superfamily						
MYB-(R1)R2R3	136	3	2	3	IPR001005	P, B
MYB-related	54	3	1	7	IPR000818	P, B
AP2/EREBP					IPR001471	B
AP2 subfamily	14	0	0	0		
EREBP subfamily	124	0	0	0		
RAV-like	6	0	0	0		
bHLH	139	46	25	8	IPR001092	B
NAC	109	0	0	0	BAB10725	B
C2H2 (Zn)	105	291	139	53	IPR000822	P, B
HB	89	103	84	9	IPR001356	B, P
MADS	82	2	2	4	IPR002100	B
bZIP	81	21	25	21	IPR001871	B
WRKY (Zn)	72	0	0	0	S72443	B
GARP						
G2-like	44	0	0	0	AAD55941	B
ARR-B class	12	0	0	0	BAA74528	B
C2C2 (Zn)						
Dof	37	0	0	0	CAA66600	B
CO-like	33	0	0	0	A56133	B
GATA	28	6	9	10	IPR000679	B, P
YABBY	6	0	0	0	AAD30526	B
CCAAT						
HAP2 type	10	1	2	1	A26771	B
HAP3 type	11	2	2	1	P13434	B
HAP4 type	0	0	0	1	S37936	B
HAP5 type	13	3	2	2	Q02516	B
Dr1	2	1	1	1	AAB51375	B
GRAS	32	0	0	0	AAB06318	B
Trihelix	28	0	0	0	S39484	B, P
HSF	26	1	1	5	IPR000232	B
TCP	25	0	0	0	AAC26786	B
ARF	23	0	0	0	AAC49751	B
C3H-type 1 (Zn)	17	3	15	3	IPR000571	P, B
C3H-type 2 (Zn)	16	0	0	0	CAA65242	B
SBP	16	0	0	0	CAB56581	B
Nin-like	15	0	0	0	CAB61243	B
ABI3/VP1	14	0	0	0	CAA48241	B
TUB	11	2	1	0	IPR000007	B
E2F/DP	8	3	4	0	O00716/Q64163	B
CPP (Zn)	8	1	1	0	CAA09028	B
Alfin-like	7	0	0	0	AAA20093	B
EIL	6	0	0	0	AAC49750	B
LFY	1	0	0	0	AAA32826	B
Other	20	0	0	0	_	B
NHR (C8) (Zn)	0	21	252	0	IPR001628	B
Adf-1	0	26	3	0	AAA28325	B
T-BOX	0	8	21	0	IPR001699	B
ETS	0	8	10	0	IPR000418	B
DM (Zn)	0	4	9	0	IPR001275	B, P
PAIRED (w/o HB)	0	5	7	0	IPR001523	B
Runt/CBFα	0	4	1	0	IPR001527	B
NF-kB/Rel/dorsal	0	3	0	0	IPR000451	P, B
Smad	0	3	3	0	BAA76956	B
NTF-1/grainyhead	0	2	1	0	CAA33692	B
STAT	0	1	1	0	IPR001217	B
AP-2	0	1	4	0	CAA36842	B
Olf-1/EBF	0	1	1	0	AAA41759	B
TSC-22/Dip/Bun	0	1	1	0	IPR000580	B
NF-I	0	1	1	0	CAA35853	B
p53	0	1	0	0	CAA42629	B
brinker	0	1	0	0	BAA76710	B
C6 (Zn)	0	0	0	52	IPR001138	B, P
Swi4/Swi6	0	0	0	5	CAA35949	B
Copper fist	0	0	0	3	IPR001083	B, P
SP23/MGA2	0	0	0	2	CAA81855	B
ABF1/AZF1	0	0	0	2	CAA81951	B
RAP1	0	0	0	1	IPR001357	B
Fork head	0	18	15	4	IPR001766	B
RFX	0	1	1	1	NP_002909	B
Other transcriptional regulators						
Aux/IAA	26	0	0	0	AAC39440	B
HMG-box	10	21	15	7	IPR000910	B
ARID	4	5	4	2	IPR001606	B
JUMONJI	9	2	1	1	T30254	B
PcG; E(z) class	3	1	1	0	_	B
PcG; Esc class	1	2	1	0	_	B
CBFβ	0	2	0	0	Q08024	B

shock transcription factors (HSFs) than does *Drosophila*, *C. elegans*, or yeast. Plant HSFs exhibit structural and functional characteristics specific to that lineage (*40*, *41*).

For those transcription factor families that are common to all eukaryotes, how similar are the *Arabidopsis* proteins to those from the other organisms? Each *Arabidopsis* transcription factor was compared to the proteomes of *Drosophila*, *C. elegans*, and yeast by using the BLASTX and BLASTP programs. The analysis revealed that *Arabidopsis* transcription factors do not share significant similarity with those from the other lineages, except in the conserved DNA binding domains that define the respective families. The only *Arabidopsis* proteins that showed similarity beyond the threshold of significance established in the comparison (*42*) were some homologs of the HAP3 subunit of the CCAAT-box binding factor and a MYB-related protein known to be homologous to the *S. cerevisiae* CEF1 and *S. pombe* Cdc5 proteins (*43*, *44*).

Domain Shuffling

The modular nature of transcription factors and the importance of domain shuffling in protein evolution are both well established. The characterization of the entire complement of *Arabidopsis* transcription factors allows consideration of the extent of domain accretion, shuffling, and divergence in these proteins and reveals the relationships among the different families at a genome-wide scale (Fig. 1).

Shuffling of some of the DNA binding domains that are present in all eukaryotes has generated novel transcription factors with plant-specific combinations of modules. This is well illustrated by the homeodomain proteins. In ~50% of the members of the *Arabidopsis* homeobox family, the homeodomain is followed by a leucine zipper (Fig. 1). This combination of motifs is not observed in the yeast or animal homeodomain proteins. The only *Arabidopsis* homeodomain proteins that have an additional motif also found in animal homeodomain proteins are those of the KNOX class, which contain a MEINOX domain (Fig. 1) (*45*). On the other hand, homeodomains in animals are associated with a large variety of motifs, such as the paired and POU-specific domains, the LIM motif, or C2H2 zinc fingers, in combinations that are not present in *Arabidopsis*. Some of these domains (paired and POU) are specific to animals.

Other examples of plant-specific arrangements of common domains include the MADS, YABBY, and ARID families. The ARID (for AT-rich interaction domain) motif is found in animals in a variety of developmental and cell-cycle regulators, like the *Drosophila* Dead ringer and Osa proteins (*46*). In animal ARID proteins, that domain is combined with other motifs, like PHD fingers or the jumonji domain (*47*). In the *Arabidopsis* ARID proteins, the ARID domain is associated with an HMG box, whereas PHD fingers and the jumonji domain form other combinations (Fig. 1). Some animal ARID proteins, like Bright, exhibit sequence-specific DNA binding, whereas others, like Osa, do not. Osa, however, modulates the activity of the SWI/SNF Brahma complex to promote the activation of specific target genes (*46*).

MADS domain proteins in plants were first identified as regulators of floral organ identity and have since been found to control additional developmental processes, such as meristem identity, root development, fruit dehiscence, and flowering time (*48, 49*). A characteristic of the plant MADS domain proteins that sets them apart from their animal and fungal counterparts is a modular organization containing a distinct coiled-coil domain (K box). The *Arabidopsis* genome sequence, however, has revealed that there is an additional class of plant MADS domain proteins in which the K box is absent (*50*). Phylogenetic analyses indicate that a gene duplication event, ancestral to the divergence of plants and animals, generated two MADS-box gene lineages that are now present in all eukaryotes. In plants, one lineage resulted in MADS proteins with a K box, whereas the other resulted in proteins that lack it (*50*). This conclusion, which was based on sequence phylogeny, is also supported by the structure of the genes. K box–containing MADS-box genes have multiple exons, the

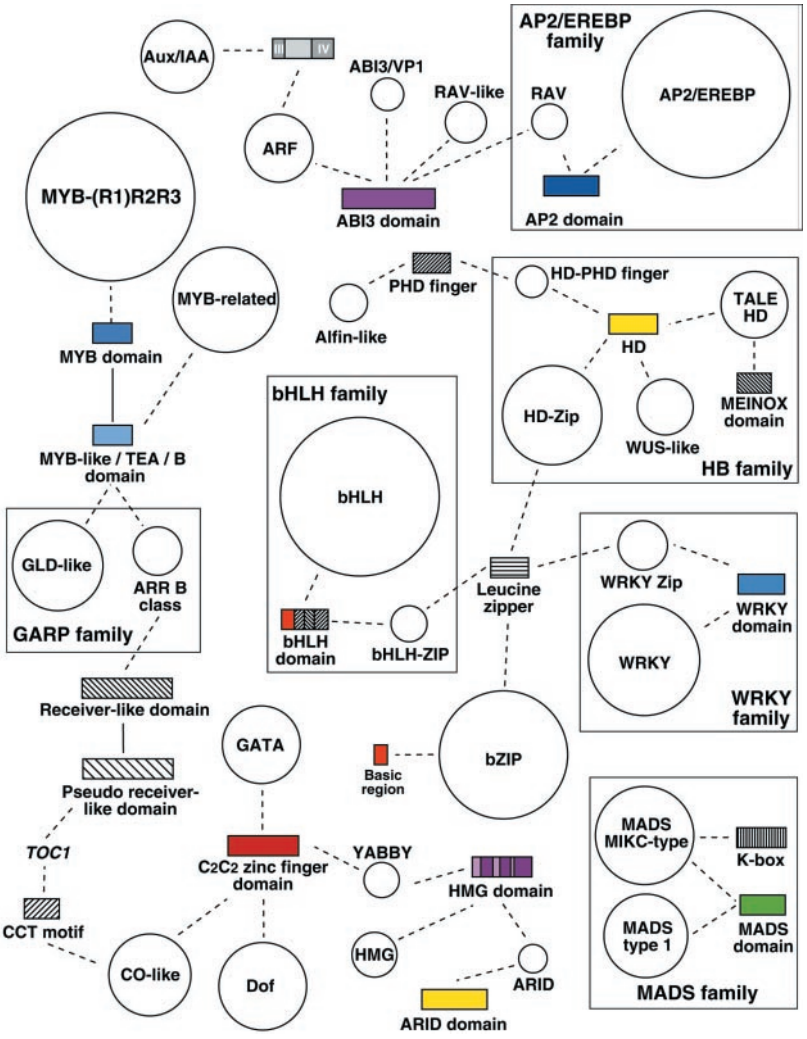

Fig. 1. Relationships and domain shuffling among the different *Arabidopsis* transcription factor families. Gene families are represented by circles, whose size is proportional to the number of members in the family. Domains that have been shuffled and that therefore "connect" different groups of transcription factors are indicated with rectangles, whose size is proportional to the length of the domain. DNA binding domains are colored; other domains (usually protein-protein interaction domains) are shown with hatched patterns. Dashed lines indicate that a given domain is a characteristic of the family or subfamily to which it is connected. Gene names are written in italics. Whereas many of the indicated domain-shuffling events are specific to plants, others likely predate the appearance of the three distinct eukaryotic lineages. For an expanded version of this figure and the information that was used to construct it, see supplemental material (*15*).

MADS box being completely encompassed in one of them. However, analysis of the *Arabidopsis* genomic sequence indicates that MADS-box genes lacking a K box have a simpler structure, with fewer or no introns. *Drosophila* and *C. elegans* each have two MADS-box genes, one per lineage. In *Arabidopsis*, in which at least 82 MADS-box genes can be identified, both classes have been substantially amplified (Fig. 1).

It has been proposed that the complexity in protein domain organization increases with the complexity of the organism (*11*). The above examples of domain shuffling and accretion suggest that, at least among transcription factors, plants are as complex as animals in this respect.

Together with the lineage-specific generation of novel classes of transcription factors or the specific amplification and divergence in one lineage of a common type of regulator, development of novel functions might also result from the organization of transcription factors in novel networks of protein-protein interactions, perhaps as a consequence of domain-shuffling events. For example, the animal-specific Smad proteins depend on interactions with other transcription factors to compensate for their relatively low DNA binding sequence specificity (*51*). These factors include the vertebrate winged-helix protein Fast-1 (winged-helix proteins are found in animals and in fungi) and the *Xenopus* homeodomain proteins Mixer and Milk. The Smad–Mixer/Milk interaction has been proposed to mediate mesoendodermal induction (*52*). All of these Smad-interacting proteins of different classes (Fast1, and Mixer and Milk) share a short Smad-interaction motif (*52*) that appears to be specific to vertebrates: it is not found in *Drosophila*, *C. elegans*, *Arabidopsis*, or yeast proteins. More examples of this kind will be uncovered as the networks of protein-protein interactions among transcription factors are deciphered.

Functional Diversity

The differences in transcription factor content, sequence, and structure among the three eukaryotic lineages are also accompanied by functional diversity. Equivalent or similar biological functions can be controlled by different families of transcription factors in each lineage. Conversely, DNA binding domains that are found in all three eukaryotic kingdoms often control different functions in each one. Developmental regulators illustrate this point. There are also cases, however, in which the involvement of a gene or family in a particular biological function has been maintained across the three lineages (for example, the HSF family).

Pattern formation is an obligate requirement in the development of complex multicellular organisms. In animals, determination of regional identity and specification of the body plan are achieved through the localized activities of homeodomain proteins. Similar functions in plants, meristem patterning and floral organ identity determination, rely on the domain-specific expression of a subset of MADS-box genes (*48*, *49*). Therefore, two different transcription factor families have been used for similar developmental functions in the two lineages.

Patterning depends on a system of axes. The dorsoventral polarity of *Drosophila* has been likened to the dorsoventral asymmetry of zygomorphic flowers and could also be conceptualized as being similar to the adaxial-abaxial polarity of the plant lateral organs. In all of these cases, polarity is established through the regionally localized expression or accumulation of transcription factors, but those belong to different classes. Floral asymmetry in *Antirrhinum* is dependent on the activities of CYC and DICH, two members of the plant-specific family of transcription factors TCP (*53, 54*). Transcription factors of another plant-specific family, YABBY, are involved in establishing the adaxial-abaxial polarity of the plant lateral organs, together with other genes like PHAN, a MYB-related protein (*55*). In *Drosophila*, embryonic dorsoventral polarity is established through a gradient of Dorsal, a transcription factor of the NF-κB/Rel/Dorsal group (NF-κB, nuclear factor κB). NF-κB/Rel/Dorsal proteins are found in *Drosophila* and mammals but not in *C. elegans*, yeast, or plants.

Conclusion

Each eukaryotic lineage has invented a sizable fraction of its own transcriptional regulators. DNA binding domains that are conserved in sequence and structure have been rearranged in different ways to create novel proteins. The degree of domain shuffling among transcription factors is large. In many instances, families that are common to the three kingdoms have been used for different or novel processes in each of the lineages. The picture that emerges from the comparison of the entire complement of transcription factors of *Arabidopsis*, *Drosophila*, *C. elegans*, and *S. cerevisiae* is one of diversity.

References and Notes

1. M. P. Scott, *Cell* **100**, 27 (2000).
2. S. B. Carroll, *Cell* **101**, 577 (2000).
3. The *C. elegans* Sequencing Consortium, *Science* **282**, 2012 (1998).
4. M. D. Adams *et al.*, *Science* **287**, 2185 (2000).
5. A. Goffeau *et al.*, *Nature* **387** (suppl.), 5 (1997).
6. N. M. Luscombe, S. E. Austin, H. M. Berman, J. M. Thornton, review available at http://genomebiology.com/2000/1/1/reviews/001/.
7. The following sequence sets were used: *Drosophila*, 14,080 predicted protein sequences (file aa_gadfly.dros.Z, available at www.fruitfly.org/sequence/download.html); *C. elegans*, 19,101 predicted protein sequences (file WormPep 20, available at www.sanger.ac.uk/Projects/C_elegans/wormpep/); and *S. cerevisiae*, 6308 predicted protein sequences (file orf_trans.fasta.Z, available at ftp://genome-ftp.stanford.edu/pub/yeast/yeast_ORFs/). The complete set of *Arabidopsis* genomic sequences was retrieved from GenBank and analyzed at Mendel Biotechnology. Version 2.0a19MP-WashU of BLAST was used, including the following settings: with BLOSUM62 scoring matrix, with gapping on, without filter, and with other parameters set to default values. Additional information is available as supplemental material (*15*).
8. S. A. Chervitz *et al.*, *Science* **282**, 2022 (1998).
9. X. Lin *et al.*, *Nature* **402**, 761 (1999).
10. K. Mayer *et al.*, *Nature* **402**, 769 (1999).
11. G. M. Rubin *et al.*, *Science* **287**, 2204 (2000).
12. L. Schauser, A. Roussis, J. Stiller, J. Stougaard, *Nature* **402**, 191 (1999).
13. T. J. Boggon, W.-S. Shan, S. Santagata, S. C. Myers, L. Shapiro, *Science* **286**, 2119 (1999).
14. G. Ruvkun, O. Hobert, *Science* **282**, 2033 (1998).
15. Supplemental material is available at www.sciencemag.org/cgi/content/full/290/5499/2105/DC1.
16. G. Blanc, A. Barakat, R. Guyot, R. Cooke, M. Delseny, *Plant Cell* **12**, 1093 (2000).
17. K. H. Wolfe, D. C. Shields, *Nature* **387**, 708 (1997).
18. The complete set of *Arabidopsis* transcription factors was compared to itself (all against all) with the TBLASTX and BLASTP programs. The BLASTP comparison was used to generate the data summarized in Table 2. A pair of proteins was considered highly similar if they showed >60% amino acid sequence identity along at least two-thirds of the length of one of them.
19. The ordered list of *Arabidopsis* clones that have been used to sequence the genome was obtained from The Arabidopsis Information Resource (TAIR) (www.arabidopsis.org). Those genes that formed related pairs or groups were mapped to the clones, and if they were in the same chromosome, the distance between them was calculated.
20. J. M. Berg, Y. Shi, *Science* **271**, 1081 (1996).
21. E. M. Meyerowitz, *Trends Genet.* **15**, M65 (1999).
22. S. Soullier *et al.*, *J. Mol. Evol.* **48**, 517 (1999).
23. S. L. Baldauf, J. D. Palmer, *Proc. Natl. Acad. Sci. U.S.A.* **90**, 11558 (1993).
24. S. Cvitanich *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 8163 (2000).
25. N. Dyson, *Genes Dev.* **12**, 2245 (1998).
26. J. L. Riechmann, E. M. Meyerowitz, *Biol. Chem.* **379**, 633 (1998).
27. M. Aida, T. Ishida, H. Fukaki, H. Fujisawa, M. Tasaka, *Plant Cell* **9**, 841 (1997).
28. T. Eulgem, P. J. Rushton, S. Robatzek, I. E. Somssich, *Trends Plant Sci.* **5**, 199 (2000).
29. Y. Nagano, *Plant Physiol.* **124**, 491 (2000).
30. T. Guilfoyle, G. Hagen, T. Ulmasov, J. Murfett, *Plant Physiol.* **118**, 341 (1998).
31. M. D. Allen, K. Yamasaki, M. Ohme-Takagi, M. Tateno, M. Suzuki, *EMBO J.* **17**, 5484 (1998).
32. D. E. Richards, J. Peng, N. P. Harberd, *Bioessays* **22**, 573 (2000).
33. H. Jin, C. Martin, *Plant Mol. Biol.* **41**, 577 (1999).
34. E. L. Braun, E. Grotewold, *Plant Physiol.* **121**, 21 (1999).
35. H. Kranz, K. Scholz, B. Weisshaar, *Plant J.* **21**, 231 (2000).
36. L. N. Hall, L. Rossini, L. Cribb, J. A. Langdale, *Plant Cell* **10**, 925 (1998).
37. S. Makino *et al.*, *Plant Cell Physiol.* **41**, 791 (2000).
38. D. D. Wykoff, A. R. Grossman, D. P. Weeks, H. Usuda, K. Shimogawara, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 15336 (1999).
39. The similarity among these proteins was probably not realized before because the sequence of the published maize Golden2 is not available from GenBank.
40. E. Czarnecka-Verner, C.-X. Yuan, K.-D. Scharf, G. Englich, W. B. Gurley, *Plant Mol. Biol.* **43**, 459 (2000).
41. F. Schöffl, R. Prändl, A. Reindl, *Plant Physiol.* **117**, 1135 (1998).
42. Each *Arabidopsis* transcription factor was compared by BLASTX and/or BLASTP to a pooled data set that combined the proteomes of *Drosophila*, *C. elegans*, and yeast. A default threshold of $P < 10^{-15}$ was established for the comparison. HSPs with a *P* value

below that threshold were inspected by eye. To be considered significantly similar, the two proteins had to show >50% identity over a region of at least 75% of the length of one of them.
43. T. Hirayama, K. Shinozaki, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 13371 (1996).
44. C. G. Burns, R. Ohi, A. R. Krainer, K. L. Gould, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 13789 (1999).
45. T. R. Bürglin, *Dev. Genes Evol.* **208**, 113 (1998).
46. R. D. Kortschak, P. W. Tucker, R. Saint, *Trends Biochem. Sci.* **25**, 294 (2000).
47. D. Balciunas, H. Ronne, *Trends Biochem. Sci.* **25**, 274 (2000).
48. J. L. Riechmann, E. M. Meyerowitz, *Biol. Chem.* **378**, 1079 (1997).
49. G. Theissen *et al.*, *Plant Mol. Biol.* **42**, 115 (2000).
50. E. R. Alvarez-Buylla *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 5328 (2000).
51. Y. Shi *et al.*, *Cell* **94**, 585 (1998).
52. S. Germain, M. Howell, G. M. Esslemont, C. S. Hill, *Genes Dev.* **14**, 435 (2000).
53. D. Luo *et al.*, *Cell* **99**, 367 (1999).
54. P. Cubas, N. Lauter, J. Doebley, E. Coen, *Plant J.* **18**, 215 (1999).
55. J. Bowman, *Curr. Opin. Plant Biol.* **3**, 17 (2000).
56. InterPro (www.ebi.ac.uk/interpro/) is a database that integrates protein domain and motif sequence patterns from other databases, like PROSITE, Pfam, and PRINTS.
57. We acknowledge the work of all those who have participated in the Arabidopsis Genome Initiative (AGI), as well as the AGI policy of immediate release of sequence data, which made possible the analysis presented here. We thank all of our colleagues at Mendel Biotechnology for their input and work in our functional genomics research program and E. Meyerowitz and F. Ausubel for discussions and comments on the manuscript.

19 October 2000; accepted 14 November 2000

Orchestrated Transcription of Key Pathways in *Arabidopsis* by the Circadian Clock

Stacey L. Harmer,[1] John B. Hogenesch,[2] Marty Straume,[3] Hur-Song Chang,[4] Bin Han,[4] Tong Zhu,[4] Xun Wang,[4] Joel A. Kreps,[4] Steve A. Kay[1,2]*

Like most organisms, plants have endogenous biological clocks that coordinate internal events with the external environment. We used high-density oligonucleotide microarrays to examine gene expression in *Arabidopsis* and found that 6% of the more than 8000 genes on the array exhibited circadian changes in steady-state messenger RNA levels. Clusters of circadian-regulated genes were found in pathways involved in plant responses to light and other key metabolic pathways. Computational analysis of cycling genes allowed the identification of a highly conserved promoter motif that we found to be required for circadian control of gene expression. Our study presents a comprehensive view of the temporal compartmentalization of physiological pathways by the circadian clock in a eukaryote.

Circadian rhythms control processes ranging from human sleep-wake cycles to cyanobacterial cell division. This is made possible by the circadian clock, an internal biochemical oscillator. The circadian clock allows organisms to anticipate daily changes in the environment such as the onset of dawn and dusk, providing them with an adaptive advantage (*1*). Physiological processes regulated by the clock in higher plants include photoperiodic induction of flowering (*2*) and rhythmic hypocotyl elongation, cotyledon movement, and stomatal opening (*3*). A small number of genes regulated by the clock have been found in an essentially serendipitous fashion (*4, 5*). However, a global examination of genes controlled by the clock in plants, or in any eukaryote, has been lacking.

[1]Department of Cell Biology, Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, CA 92037, USA. [2]Genomics Institute of the Novartis Research Foundation, 3115 Merryfield Row, La Jolla, CA 92121, USA. [3]Center for Biomathematical Technology, NSF Center for Biological Timing, Division of Endocrinology and Metabolism, Department of Internal Medicine, University of Virginia, Charlottesville, VA 22904, USA. [4]Novartis Agricultural Discovery Institute, 3115 Merryfield Row, San Diego, CA 92121, USA.

*To whom correspondence should be addressed. E-mail: stevek@scripps.edu

The circadian clock regulates hundreds of genes. We have used highly reproducible oligonucleotide-based arrays (*6*) to determine steady-state mRNA levels in *Arabidopsis* at 4-hour intervals during the subjective day and night. We examined temporal patterns of gene expression in *Arabidopsis* plants under constant light conditions using GeneChip arrays representing about 8200 different genes. We hybridized duplicate microarrays with biotin-labeled probes derived from plant tissues harvested every 4 hours over 2 days (*7*). Reproducibility between arrays was excellent (Web fig. 1) (*8*). The mean hybridization signal strength and the standard error of the mean for each probe set at each time point were calculated from the duplicate hybridizations.

To objectively determine which genes exhibited a circadian pattern of expression, we empirically tested for statistically significant cross-correlation between the temporal expression profiles of each probe set and cosine waves of defined period and phase. Genes with a greater than 95% probable correlation with a cosine test wave with a period between 20 and 28 hours were scored as circadian-regulated (*9*). This analysis is independent of signal strength and imposes no minimal change in amplitude. According to this criterion, 494 probe sets, representing 453 genes or 6% of the genes on the chip, were classified as cycling (Web table 1) (*8*); 28% of these genes have not been characterized, and no conclusions can be drawn about their function. More than 20 of the known genes we found to be clock-regulated have been previously reported to be under circadian control (*3, 10*), validating our experimental methods.

We placed the cycling genes into phase clusters of peak expression time. All six possible phases (given our 4-hour time resolution) were well represented, although there were fewer genes peaking at CT16 (*11*) than in other phases [Web table 1 and Web fig. 2 (*8*)]. This is in contrast to cyanobacteria, in which 80% of circadian-regulated genes peak near subjective dusk (*12*). Many of the genes we found to cycle can be clustered into functional groups on the basis of their known and predicted physiological roles.

Clock-controlled anticipation of dawn and dusk. A large cluster of genes implicated in the light-harvesting reactions of photosynthesis were found to be under clock control. mRNAs encoding four LHCA and seven LHCB proteins, chlorophyll binding proteins that funnel light energy to the reaction centers of photosystems I and II, were cycling (Fig. 1A). Also, mRNA encoding an enzyme (protoporphyrin IX magnesium chelatase) involved in the synthesis of their ligand, chlorophyll, was cycling (Web table 1) (*8*). Seven photosystem I reaction center genes and three photosystem II reaction center genes were likewise cycling (Fig. 1B). These 22 photosynthesis genes exhibit striking coregulation, with most peaking around midday at CT4 (*9*). Two *LHC* genes, the reaction center gene *PSAD1*, and the magnesium chelatase gene have been previously reported to cycle (*10, 13*).

Light also regulates growth and development and resets the circadian clock. Genes encoding phytochrome B (*PHYB*), cryptochrome 1 (*CRY1*), cryptochrome 2 (*CRY2*), and phototropin (*NPH1*) (Web fig. 3A) (*8*) were clock-regulated. Homologs of the blue light photoreceptor genes *CRY1* and *CRY2* are also clock-controlled in animals (*14*). Downstream mediators of phototransduction pathways, *SPA1* and *RPT2*, were also clock-

Arabidopsis Transcription Factors: Genome-Wide Comparative Analysis Among Eukaryotes

J. L. Riechmann, J. Heard, G. Martin, L. Reuber, C.-Z. Jiang , J. Keddie, L. Adam, O. Pineda, O. J. Ratcliffe, R. R. Samaha, R. Creelman, M. Pilgrim, P. Broun, J. Z. Zhang, D. Ghandehari, B. K. Sherman and G. -L. Yu

Science **290** (5499), 2105-2110.
DOI: 10.1126/science.290.5499.2105