Journal of Integrative Plant Biology 2008, **50** (8): 1031–1045

Phylogenetic Analysis of the Plant-specific *Zinc Finger-Homeobox* and *Mini Zinc Finger* Gene Families

Wei Hu†, Claude W. dePamphilis and Hong Ma*

(*Department of Biology and the Huck Institutes of the Life Sciences, Pennsylvania State University*, University Park, PA 16802, USA)

Abstract

Zinc finger-homeodomain proteins (ZHD) are present in many plants; however, the evolutionary history of the *ZHD* gene family remains largely unknown. We show here that *ZHD* genes are plant-specific, nearly all intronless, and related to *MINI ZINC FINGER* (*MIF*) genes that possess only the zinc finger. Phylogenetic analyses of *ZHD* genes from representative land plants suggest that non-seed plant *ZHD* genes occupy basal positions and angiosperm homologs form seven distinct clades. Several clades contain genes from two or more major angiosperm groups, including eudicots, monocots, magnoliids, and other basal angiosperms, indicating that several duplications occurred before the diversification of flowering plants. In addition, specific lineages have experienced more recent duplications. Unlike the *ZHD* genes, *MIF*s are found only from seed plants, possibly derived from *ZHD*s by loss of the homeodomain before the divergence of seed plants. Moreover, the *MIF* genes have also undergone relatively recent gene duplications. Finally, genome duplication might have contributed substantially to the expansion of family size in angiosperms and caused a high level of functional redundancy/overlap in these genes.

Key words: gene duplication; gene family evolution; *Mini Zinc Finger*; *Physcomitrella*; poplar; *Selaginella*; zinc finger-homeodomain proteins.

Hu W, dePamphilis CW, Ma H (2008). Phylogenetic analysis of the plant-specific *Zinc Finger-Homeobox* and *Mini Zinc Finger* gene families. *J. Integr. Plant Biol.* **50**(8), 1031–1045.

Available online at www.jipb.net

The homeodomain (HD) is a conserved DNA-binding domain that has about 60 amino acids (Burglin 1994). Homeodomain-containing proteins play crucial and diverse roles in both plant and animal development (Williams 1998; Ito et al. 2002; Akin and Nazarali 2005; Hunter and Rhodes 2005). The *Arabidopsis* proteome contains about 100 homeodomain proteins (The Arabidopsis Information Resource (TAIR), http://www.arabidopsis.org/). Most homeodomain proteins have additional domain(s) for protein-protein interaction or other functions (Burglin 1994). The HD-Zip family is a large class of plant-specific homeodomain proteins (Ariel et al. 2007). They have a leucine zipper domain at the C-terminal side of homeodomain for homo- and heterodimer formation. The TALE (three-amino acid extension loop) family includes two classes of homeodomain proteins, KNOX and BELL, with eight and 13 members in *Arabidopsis*, respectively. Members from these two classes can form heterodimer complexes via domains conserved within each class (Bellaoui et al. 2001; Muller et al. 2001; Smith and Hake 2003; Bhatt et al. 2004; Hake et al. 2004).

Zinc fingers are important motifs widely present in many regulatory proteins (Takatsuji 1999; Krishna et al. 2003). A typical zinc finger has two pairs of conserved cysteine and/or histidine residues coordinating a single zinc ion to stabilize the motif as a finger-shaped loop (Klug and Schwabe 1995). A protein can possess one or more zinc fingers and other domains. Zinc fingers are involved in DNA binding, protein-protein interaction, and more rarely in RNA-binding and protein folding (von Arnim and Deng 1993; Mackay and Crossley 1998; Takatsuji 1998). Zinc fingers can be classified into different types on the basis of the nature, number, and spacing pattern of zinc-binding residues. For example, C_2H_2, C_2C_2 and C_3H zinc fingers interact with one zinc ion, whereas the really interesting

Received 29 Sept. 2007 Accepted 21 Nov. 2007

†Present address: Section of Molecular and Cellular Biology, College of Biological Sciences, University of California, Davis, CA 95616, USA.

Supported by a National Science Foundation Plant Genome Grant for the Floral Genome Project (DBI-0115684) and by the Biology Department and the Huck Institutes of the Life Sciences, Pennsylvania State University. This study was conducted using material generated in part with support from the National Science Foundation (No. 0215923).

*Author for correspondence.

Tel: +1 814 863 6414;

Fax: +1 814 863 1357;

E-mail: <hxm16@psu.edu>.

doi: 10.1111/j.1744-7909.2008.00681.x

new gene (RING) finger, plant homeodomain (PHD) zinc finger and Lin-11/Isl-1/Mec-3 (LIM) domain coordinate two zinc ions (Halbach et al. 2000; Li et al. 2001; Kosarev et al. 2002; Englbrecht et al. 2004; Yanagisawa 2004).

A group of novel zinc finger-homeodomain (ZF-HD) proteins were first identified from the C_4 plant *Flaveria* by Windhovel et al. (2001) as potential regulators of the gene encoding C_4 phosphoenolpyruvate carboxylase (PEPCase). The ZF-HD proteins have an N-terminal conserved domain that contains several cysteine and histidine residues for potential zinc binding and a C-terminal domain that is distantly related to canonical homeodomains. Windhovel et al. (2001) showed that the putative ZF domain functions in homo- and heterodimer formation, which requires the conserved cysteines. The novel HD domain is able to bind DNA, particularly the regulatory regions of C_4 PEPCase genes (Windhovel et al. 2001). The ZF domain is not involved in DNA binding, but can enhance the protein-DNA interaction mediated by the HD domain (Windhovel et al. 2001).

Arabidopsis ZF-HD proteins also can bind to DNA sequences with a core consensus of ATTA, and form homo- and heterodimers (Tan and Irish 2006). More recently, an *Arabidopsis* ZF-HD protein (ZFHD1) was reported to specifically bind to the promoter of *EARLY RESPONSE TO DEHYDRATION STRESS 1* (*ERD1*) (Tran et al. 2007). The expression of *ZFHD1* is inducible by dehydration, salt stress and abscisic acid (ABA). In addition, ZFHD1 can interact with some NAM/ATAF1,2/CUC2 (NAC) proteins and the simultaneous overexpression of *ZFHD1* and *NAC* genes improved *Arabidopsis* tolerance to drought stress (Tran et al. 2007). Two soybean ZF-HD proteins were also shown to bind to the promoter of the gene encoding calmodulin isoform 4 (*GmCaM4*) and induce its expression upon pathogen stimulation (Park et al. 2007). These findings strongly support the notion that ZF-HD proteins are transcriptional regulators.

We previously identified three *Arabidopsis MINI ZINC FINGER* (*MIF*) genes and their homologs that encode proteins with high degrees of sequence similarity with the ZF domain of ZF-HD proteins but lack the HD domain (Hu and Ma 2006). Analyses of the pleiotropic and dramatic phenotypes conferred by the overexpression of *MIF1* in *Arabidopsis* suggest that *MIF1* may be involved in the regulation of plant development by multiple hormones (Hu and Ma 2006). It is possible that MIF1 interact with ZF-HD proteins via the ZF domain and, when overexpressed, interfere with the normal functions of ZF-HD proteins. The phenotypes of *35S::MIF1* transgenic plants could result from functional disruption of ZF-HD-containing protein complexes. If this is true, then other ZF-HD proteins might also play important roles in regulating plant development and physiology.

Zinc finger-homeodomain proteins form a monophyletic group distinct from other homeodomain proteins (Windhovel et al. 2001; Tan and Irish 2006). However, the origin and evolutionary history of the *ZF-HD* and *MIF* genes and the relationship between these two types of genes remain unclear. We have carried out extensive phylogenetic and sequence analyses of the *ZF-HD* and *MIF* genes. Available complete genomes of *Arabidopsis*, rice, poplar, the seedless vascular plant *Selaginella moellendorffii* and the nonvascular plant *Physcomitrella patens* make it possible to examine gene duplication and family size spanning a broad evolutionary scale in the plant kingdom. In addition, the Floral Genome Project (FGP; www.floralgenome.org) offers a great resource to identify ZF-HD homologs from basal angiosperms and a gymnosperm that occupy key positions in the plant phylogenetic tree (Albert et al. 2005). Gene expression pattern and its correlation with the gene phylogeny were also investigated in the model plant *Arabidopsis*, independent of the study by Tan and Irish (2006). In summary, we report that ZF-HD proteins are (land) plant-specific and encoded primarily by intronless genes. Furthermore, we present a comprehensive phylogenetic tree for this family and describe sequence features for non-angiosperm and seven classes of angiosperm ZF-HDs. Our analysis also suggests that the *MIF* gene family may have originated from a *ZF-HD* gene by loss of the homeodomain and then diversified in seed plants. Finally, a high level of functional redundancy among this family was implicated by evidence of gene duplication, expression pattern and normal development of null mutants.

Results

The zinc finger-homeobox gene family in *Arabidopsis*

We previously identified *Arabidopsis ZF-HD* genes as florally expressed cDNAs (Hu et al. 2003). To explore the evolutionary history of this family, we examined the *Arabidopsis* genome for *ZF-HD* genes. Following the *Arabidopsis* gene nomenclature guidelines and for systematically naming members in other plants, we propose to rename the *ZF-HD* genes as *ZHD* (zinc finger-homeodomain). The *Arabidopsis* genome contains 14 *AtZHD* genes and three *AtMIF* genes (Table 1). Phylogenetic analysis showed that *AtZHDs* form five well-supported clades (*AtZHD1*/*2*, *AtZHD3*/*4*, *AtZHD5-7*, *AtZHD8-12*, and *AtZHD13*/*14*) (Figure 1A). Three pairs of *AtZHD*s were further found to be derived from a proposed large-scale segmental or whole-genome duplication (Vision et al. 2000; Schoof et al. 2002); members of each of these pairs possess linker regions of similar lengths between the ZF and HD domains (Figure 1A,B).

We previously reported expression patterns of *AtMIF* genes (Hu and Ma 2006). To obtain clues about the function of *ZHD* genes, reverse transcription-polymerase chain reaction (RT-PCR) experiments were carried out. Figure 1C shows that *AtZHD1*, *3-5*, *7* and *13* were relatively highly expressed in the inflorescence and/or open flower. In contrast, *AtZHD6*, *8-10*, and *14* were expressed similarly in the tissues tested. The *AtZHD12* mRNA was only detected weakly in the root after an increased number of PCR cycles (Figure 1C). In general, the

Table 1. *Zinc finger homeobox* (*ZHD*) and *MINI ZINC FINGER* (*MIF*) genes from five plant species with complete genome information

Annotated genome		Un-annotated genome	
Arabidopsis thaliana		*Populus trichocarpa* (poplar)[a]	
AtZHD1	At5g65410	PtZHD1	Chr II 7399622-7400410
AtZHD2	At4g24660	PtZHD2	Chr V 8372076-8372849
AtZHD3	At2g02540	PtZHD3	Scaffold 86 677268-678158
AtZHD4	At1g14440	PtZHD4	Scaffold 150 79633-80475
AtZHD5	At1g75240	PtZHD5	Chr IV 11561024-11561830
AtZHD6	At2g18350	PtZHD6	Chr II 2257696-2258577
AtZHD7	At3g50890	PtZHD7	Chr V 15831484-15832356
AtZHD8	At5g15210	PtZHD8	Chr VII 10862058-10863053
AtZHD9	At3g28920	PtZHD9	Scaffold 57 1470513-1471538
AtZHD10	At5g39760	PtZHD10	Scaffold 41 1309886-1310731
AtZHD11	At1g69600	PtZHD11	Chr XIX 9214242-9215039
AtZHD12	At5g60480[b]	PtZHD12	Chr VII 5356166-5357140
AtZHD13	At5g42780	PtZHD13	Chr X 15435018-15436001
AtZHD14	At1g14687	PtZHD14	Chr IV 12779967-12780971
AtMIF1	At1g74660	PtZHD15	Chr XVII 4563457-4564479
AtMIF2	At3g28917	PtZHD16	Scaffold 122 609134-609706
AtMIF3	At1g18835	PtZHD17	Chr XII 345336-345884
Oryza sativa (rice)		PtMIF1	Chr IV 12830248-12830660
OsZHD1	Os09g29130	PtMIF2	Chr XVII 4492679-4493217
OsZHD2	Os08g37400	*Physcomitrella patens* (moss)	
OsZHD3	Os12g10630	PpZHD1	DQ099428[d]
OsZHD4	Os11g13930[c]	PpZHD2	TI 713864897[e]
OsZHD5	Os01g44430[c]	PpZHD3	TI 857912325[e]
OsZHD6	Os05g50310	PpZHD4	TI 863018018[ef]
OsZHD7	Os02g47770[c]	PpZHD5	TI 815607583[e]
OsZHD8	Os04g35500	PpZHD6	TI 893568082[e]
OsZHD9	Os09g24820	PpZHD7	TI 1023221209[e]
OsZHD10	Os08g34010[c]	*Selaginella moellendorffii*	
OsZHD11	Os03g50920	SmZHD1	TI 725442290[e]
OsMIF1	Os11g03420	SmZHD2	TI 719942535[e]
OsMIF2	Os12g03110[c]	SmZHD3	TI 759773047[e]
OsMIF3	Os09g24810	SmZHD4	TI 1166989312[eg]
OsMIF4	Os08g33990[c]		

[a]Currently the poplar genome is not completely assembled. [b]AtZHD12 is proposed to be a pseudogene. [c]Annotation from the The Institute for Genomic Research (TIGR) database is incorrect. [d]Cloned in this study. [e] National Center for Biotechnology Information (NCBI) trace identifier (TI); only one representative TI is provided. This TI may contain some sequencing errors and may not cover the full coding region of the corresponding *ZHD*; TI sequences could be obtained from <http://www.ncbi.nlm.nih.gov/Traces/>. [f]The 3′ region of CDS is represented by only one TI that contains sequencing errors, PpZHD4 is thus not used for ZHD phylogenetic analysis, but it is very similar with PpZHD3. [g]SmZHD4 contains an intron, and thus is represented by an expressed sequence tag (EST) TI sequence.

most related genes exhibited similar expression patterns, such as *AtZHD3* and *AtZHD4*, and *AtZHD8*, *9*, and *10*. These RT-PCR results are in good agreement with results from microarray experiments using Affymetrix technology (Schmid et al. 2005; Zhang et al. 2005), and they are also largely consistent with the results reported by Tan and Irish (2006), although there are some differences.

As an example, we also carried out RNA *in situ* hybridization to determine the spatial expression pattern of *AtZHD5,* which was initially found to be preferentially expressed in flowers compared with leaves (Hu et al. 2003). Figure 2A shows that *AtZHD5* was expressed in the inflorescence meristem (except the most apical region), the stem cortex beneath the inflorescence meristem, the flower meristem, and flower buds. During early flower development, *AtZHD5* was expressed in all floral organs; at later stages, the *AtZHD5* expression was gradually restricted to the petal and gynoecium (Figure 2A,B). In addition, it was expressed in the axillary bud and the basal region of young leaves (Figure 2C).

Zinc finger-homeobox genes are land plant-specific

To identify *ZHD* homologs in other species, an extensive search of public sequence databases was carried out using BLAST. *ZHD* homologs were found in land plants, including angiosperms, gymnosperms, the seedless vascular plant *Selaginella* and the nonvascular plant (moss) *Physcomitrella*. However, BLAST search using different ZHD sequences as the queries did not result in any significant match (E-value < 0.1) from animals, fungi, yeast, green algae (*Chlamydomonas* and *Volvox*) or prokaryotes. These results strongly suggest that this family is land plant-specific, consistent with previous studies (Windhovel et al. 2001; Tan and Irish 2006). In addition, ZHD proteins were distantly related to the mammalian LIM-HD family of homeodomain proteins, which could only be identified by a position specific iterative-basic local alignment search tool (PSI-BLAST) with E values > 0.1. A comparison of genomic and cDNA sequences indicates that nearly all *ZHD* genes are intronless in the coding region. The only exception is the *Selaginella SmZHD4* gene, which has a small intron. The intronless feature greatly facilitated the identification and annotation of *ZHD* homologs from the recently sequenced genomes of *Populus*, *Physcomitrella* and *Selaginella* (Table 1) and from partial genomic sequences of other plants. We have also used sequence similarity to correct the likely mistakes in the annotation of six rice *ZHD* genes (Table 1).

The numbers of *ZHD* genes in *Arabidopsis* (120 Mb), *Populus* (550 Mb) and rice (430 Mb) are 14, 17 and 11, respectively (Table 1). These *ZHD* genes were named according to their phylogenetic placement (see below). We identified six *ZHD* genes from pine expressed sequence tag (EST) projects and one *Welwitschia* gene from the Floral Genome Project (Albert et al. 2005), but the total number in a gymnosperm

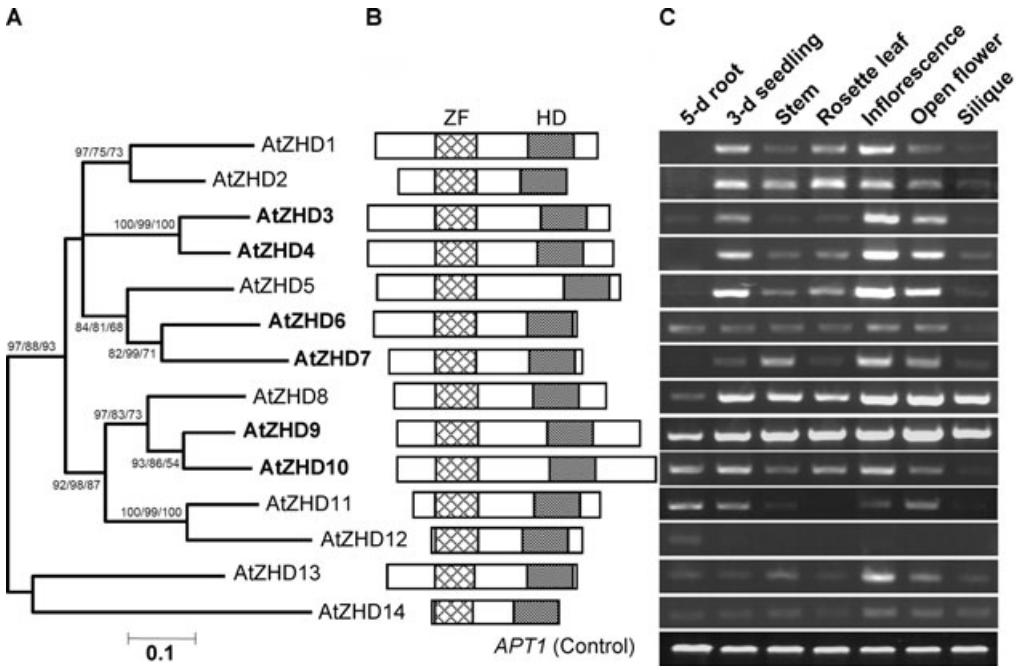

Figure 1. Phylogenetic relationship, primary protein structures and expression patterns of *AtZHD* genes.

(**A**) Neighbor-joining unrooted phylogenetic tree. Bootstrap values greater than 50% for branch support are shown in the order of neighbor-joining/maximum likelihood/maximum parsimony (NJ/ML/MP). Three pairs of genes derived from segmental genome duplication are in bold.

(**B**) Schematic representation of protein length and domain location. HD, homeodomain; ZF, zinc finger.

(**C**) Expression patterns examined by reverse transcription-polymerase chain reaction (RT-PCR). *APT1* was used as the internal control, all PCR reactions were performed for 28 cycles except for *AtZHD12*, which was 38 cycles.

species awaits a complete genome sequence. The genomes of *Selaginella* (estimated 106 Mb) and *Physcomitrella* (460 Mb) contain four and seven *ZHD* genes, respectively (Table 1).

The ZHD-type homeodomain consists of 64 amino acid residues (not counting extra residues that are present in class VI ZHDs) (Figure 3A). It is predicted to have the three helices typical for homeodomains, with a highly conserved third helix. Except for the poplar PtZHD4, all ZHD proteins possess an invariant tryptophan (W_{52}) residue in the third helix, a characteristic of all known homeodomains (Burglin 1994). In addition, all ZHDs are conserved at the 25th residue for W and at the 55th residue for asparagine (N). An alignment between the consensus of the ZHD homeodomain and that of typical homeodomains indicated limited similarity (Figure 3B), indicating that the ZHD homeodomain is atypical. In particular, a canonical phenylalanine (F) after W_{52} in the third helix is uniformly altered to a methionine (M) in the ZHD-type homeodomain, suggesting a possible change in DNA-binding specificity.

Phylogenetic analyses and sequence motifs of *ZHD* genes

To investigate the evolutionary history of the *ZHD* gene family, 107 ZHD protein sequences (Table 1 and data available upon request) were used for phylogenetic analysis. Phylogeny based on the most conserved ZF and HD domains (120 residues) revealed several clades with low bootstrap support. To improve the resolution, two informative motifs, LALP and EDST (named for the four characteristic residues), which were found in most ZHDs in the linker region between the ZF and HD domains, were also included in our analysis. The tree generated using the neighbor-joining (NJ) method with an alignment of 143 residues suggested that angiosperm ZHDs formed several well-supported monophyletic groups (Figure 4). Several non-angiosperm ZHDs were clustered together, though without bootstrap support. Maximum likelihood analysis generated a tree with similar topology as the NJ tree. Maximum parsimony analysis also provided good bootstrap support for some clades shown in Figure 4. Additional phylogenetic analysis using DNA sequences did not improve the bootstrap support (data not shown). The phylogenetic tree indicates that the *ZHD* gene family likely underwent multiple duplication events before the divergence of major groups of angiosperms. The current diversity of *ZHD* genes represents the descendents of multiple distinct paralogs that likely were present in the common ancestor of most or all living angiosperms. For simplicity, we named the major clades of angiosperm genes as class I to VII. The

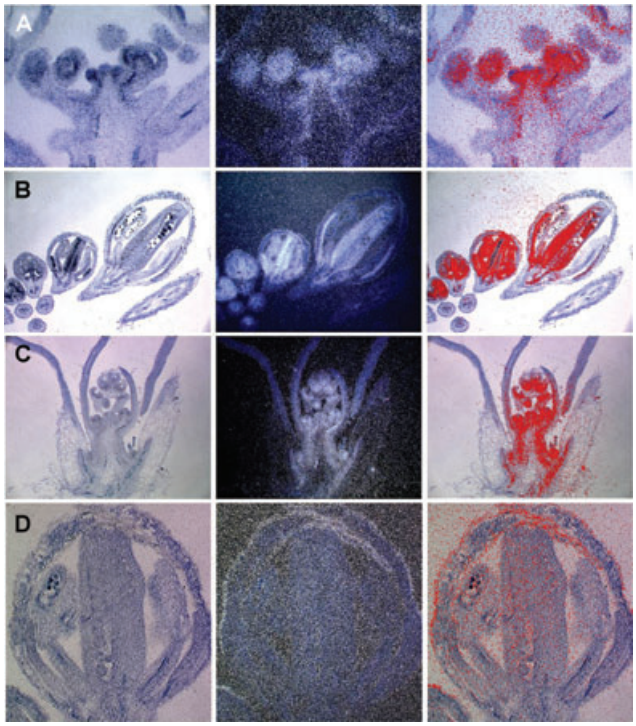

Figure 2. *AtZHD5* expression pattern examined by RNA *in situ* hybridization.

(**A**) An inflorescence apex and surrounding floral primordia and buds.

(**B**) Floral buds with developmental stages estimated from stage 5, 6, 8 to 10.

(**C**) A 2-week-old plant longitudinally sectioned.

(**D**) A floral bud at approximately stage 9 hybridized with sense probes showing no specific signal. Left panels are bright-field images, middle panels are dark-field images, and right panels are superimposed from left and middle panels using PHOTOSHOP.

(A–C) Hybridization with antisense probes.

presence or absence of additional conserved motifs in each class of ZHDs also supports these classes (Figure 5). The previously described five groups of *Arabidopsis* AtZHDs belong to classes I–III, V and VI, respectively (Figure 4).

Class I includes ZHDs from a magnoliid (SheZHD1), a monocot (OsZHD1 and OsZHD2), a basal eudicot (EcaZHD1), and a number of core eudicots. Class I ZHDs share high levels of similarity outside the ZF and HD domains, including the highly conserved LALP and EDST motifs (Figure 5). In addition, they possess a conserved N-terminal region that is rich in residues E and D, and have an invariant phenylalanine (F) rather than isoleucine (I) or leucine (L) commonly found in other ZHDs at the third position. Finally, most Class I ZHDs lack any sequences C-terminal of their HD domains, ending with the residues GKKP (Figure 5; and alignment is available upon request). Similar to Class I, Class II ZHDs cover a wide range of evolutionarily diverse angiosperms, including Magnoliids (SheZHD2), basal monocot (AamZHD1), other monocots (CloZHD1, YfiZHD1, AofZHD1, and three grass ZHDs), basal eudicot (EcaZHD2), and several core eudicots (Figure 4). Class II ZHDs also have a conserved EDST motif; however, the position corresponding to the Class I LALP motif is occupied by a related motif MIM(P/S) (Figure 5). In addition, a small N-terminal region is conserved among class II ZHDs. Class I and II occupy relative basal positions and include more homologs from basal angiosperms, basal monocots, and basal eudicots than other classes. In addition, they are weakly related to one or two gymnosperm ZHDs. Therefore, Class I and II ZHDs may represent two conserved lineages.

Class III and IV ZHDs consist of only eudicot and monocot genes, respectively (Figure 4). These two classes share a very similar EDST motif but have divergent LALP motifs. There are also other sequence differences between these two classes. Class III ZHDs are fairly conserved at the N-terminal region, whereas Class IV ZHDs are not (Figure 5). In addition, the central region of the ZF domain of Class IV ZHDs not only has a deletion of one to four residues, but also has different residue compositions compared with Class I to III ZHDs (Figure 5). Noticeably, Class IV ZHDs are rich in glutamine (Q) at the C termini. Although these two clades are sisters in the tree shown in Figure 4, this relationship does not have bootstrap support.

Class V includes the greatest number of detected *ZHD* genes and exhibits several sequence features different from Classes I–III. First, the central region of the Class V ZF domains is typically composed of 11 residues (S/T)P(S/T/A)$_3$P(S/T)DP(S/T)$_2$, rather than the nine residues (S/G)GEEG(S/T)(I/L/V)(E/D)A found in other ZHDs. Second, Class V ZHDs have a much longer C-terminal region that terminates with a conserved NGSS motif (Figure 5). Third, they have a highly conserved LALS motif in place of LALP, but lack an EDST motif. Finally, Class V ZHDs are highly conserved among themselves, with an FNGV motif in the N-terminal region and a poly HP motif (rich in histidine and proline) C-terminal of the ZF domain. Notably, Class V has many members from core eudicots (rosids and asterids) but only two members from monocot (Figure 4), suggesting that it has successfully expanded and possibly subfunctionalized in eudicots.

Class VI ZHDs form the best supported major clade in the phylogenetic tree and represent a lineage that is highly divergent from other classes (Figure 4). The protein length of this class is approximately two thirds that of other ZHDs, with very short N-terminal (except for AtZHD13) and linker regions (Figure 5). Furthermore, the central region of their ZF domain is distinct from other classes of ZHDs in terms of both length and residue composition. They also have one to six extra residues in the middle of the HD domain between the 31st and 32nd residues (Figure 3A). Finally, they do not have recognizable LALP and EDST motifs. However, a subset of VI ZHDs (excluding monocot ZHDs and *Arabidopsis* AtZHD13 and AtZHD14) share a high

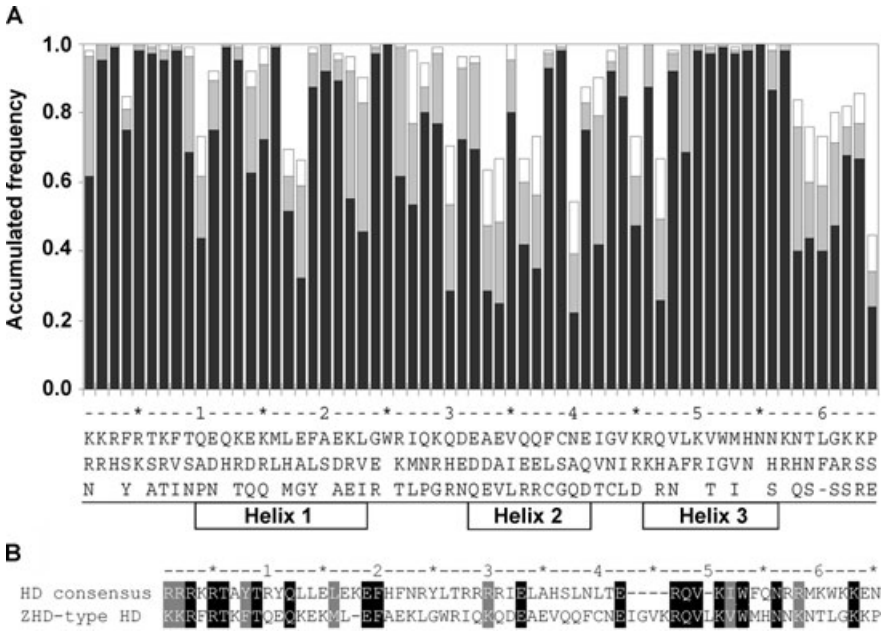

Figure 3. Analysis of zinc finger-homeodomain (ZHD) protein type homeodomain.

(**A**) Consensus of the homeodomain of ZHD proteins. The three most conserved residues of each position are listed along the x axis, with the most conserved one at the top; three helices of the homeodomain are indicated based on the solution structures of AtZHD1 and AtZHD2 (GenBank accession number for these two protein structures are 1WH5A and 1WH7A, respectively).

(**B**) Alignment between consensus sequences of classic homeodomains (Burglin 1994) and ZHD-type homeodomain. Gaps are introduced to maximize alignment; identical and similar residues are shaded.

level of similarity at the C-terminal region that is rich in serine and threonine. These distinct sequence features, coupled with the well-supported separation of this clade from all other clades in the phylogenetic tree, suggest that Class VI ZHDs may have lost the original ZHD function or obtained a new function (Figure 4). As this class includes ZHDs from *Liriodendron* (magnoliids), California poppy (basal eudicot), and ginger (mid-level monocot), this clade is likely the result of a duplication event before the separation of these major angiosperms groups.

In addition to Class I to VI, there is a small, poorly supported clade consisting of only four genes from magnoliids (SheZHD3), monocots and eudicots (Figure 4). For convenience, we call this clade Class VII. Except for a weak EDST motif, these ZHDs themselves do not share sequence similarity outside the ZF and HD domains, providing an explanation for its low support.

The non-angiosperm *ZHD* genes from *Physcomitrella*, *Selaginella* and gymnosperms do not form a well-supported group. These ZHDs possess LALP and EDST motifs similar to those in class I and II ZHDs, respectively (Figure 5). However, the EDST motif is not very well conserved among the non-angiosperm ZHDs. The *Physcomitrella* ZHDs apparently fall into two subgroups: PpZHD1-4, and PpZHD5-7 (PpZHD4 is very similar to PpZHD3, but was excluded from phylogenetic analysis because of its likely sequence inaccuracy). The pine PtaZHD5 and PtaZHD6 proteins are possibly related to Class I, and PtaZHD3 to Class III, although lacking statistical support (Figure 4). Although the evolutionary relationships between gymnosperm and angiosperm *ZHD* genes are not clear, it is possible that some of the gene duplication events leading to different ZHD classes occurred prior to the divergence of gymnosperms and angiosperms.

Evolution of *MIF* genes

We recently reported the analysis of *MINI ZINC FINGER* (*MIF*) genes that has only the N-terminal zinc finger but lacks the C-terminal homeodomain (Hu and Ma 2006). In this study we identified one additional *MIF* gene each from rice, cotton, and *Welwitschia,* and five more from pine (data not shown), with a toal of 48 seed plant *MIF* genes. No MIF was found from seedless plants *Physcomitrella* and *Selaginella*. Similar to *ZHD* genes, *MIF* genes are intronless. We noted that six *MIF* genes have an in frame ATG codon upstream of the previously annotated initiation methione (Hu and Ma 2006); potentially extending the N-termini by seven to 60 residues (alignment is available upon request). Interestingly, when present, the extended sequences of pine and spruce MIFs are also conserved. Although the other pine and spruce *MIFs* appear to lack such

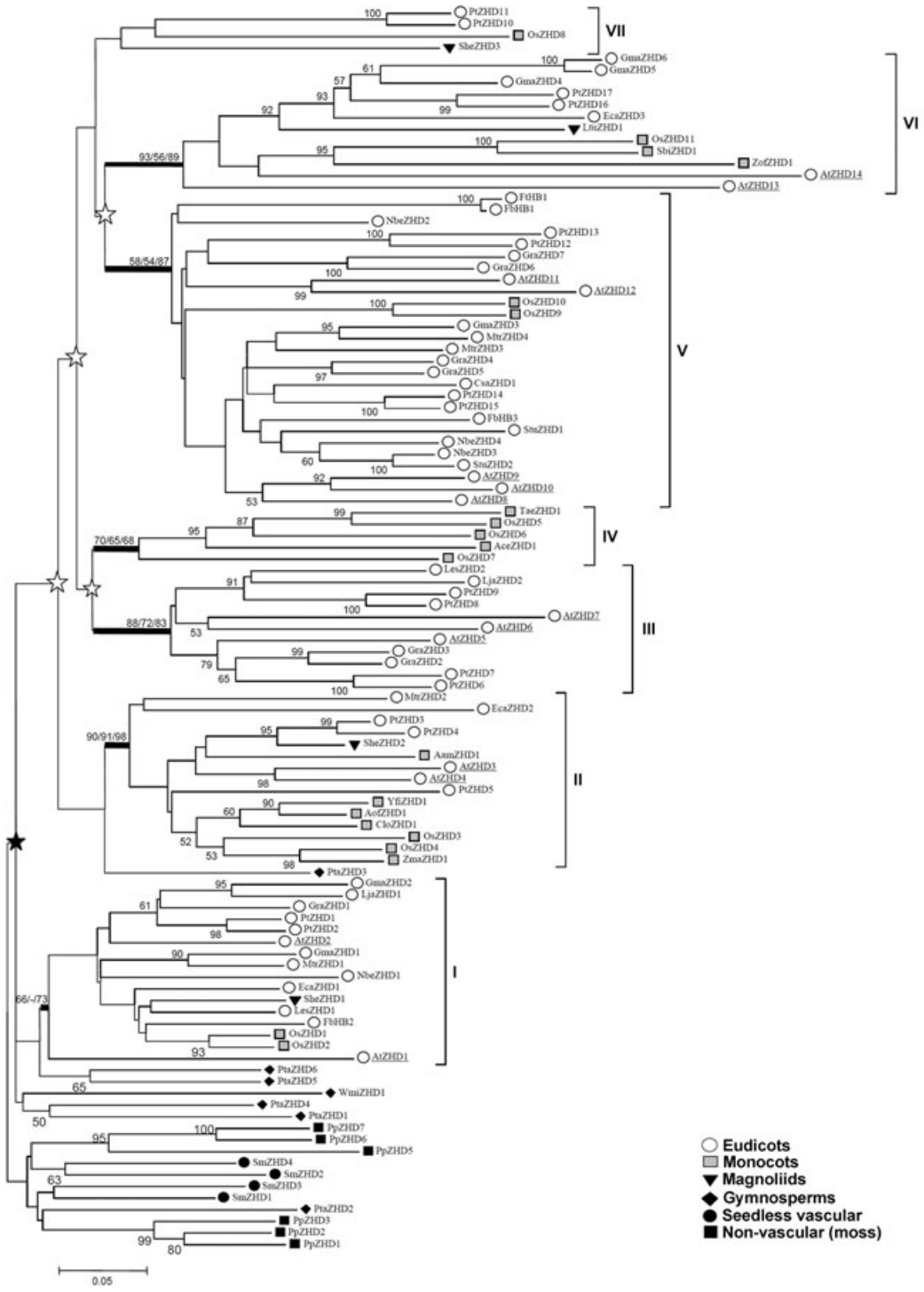

Figure 4. Neighbor-joining (NJ) phylogenetic tree of 107 zinc finger-homeodomain (ZHD) proteins constructed with the zinc finger (ZF) and homeodomain (HD) domains as well as the LALP and EDST motifs (143 AA).

Arabidopsis ZHDs are underlined. Bootstrap values greater than 50% for branch support are shown. Bootstrap values from maximum parsimony (MP) and maximum likelihood (ML) analysis are also provided for major clades in the order of NJ/MP/ML. The solid star indicates an early duplication event in the ancestor of seed plants, and open stars indicate likely multiple duplication events among early angiosperms.

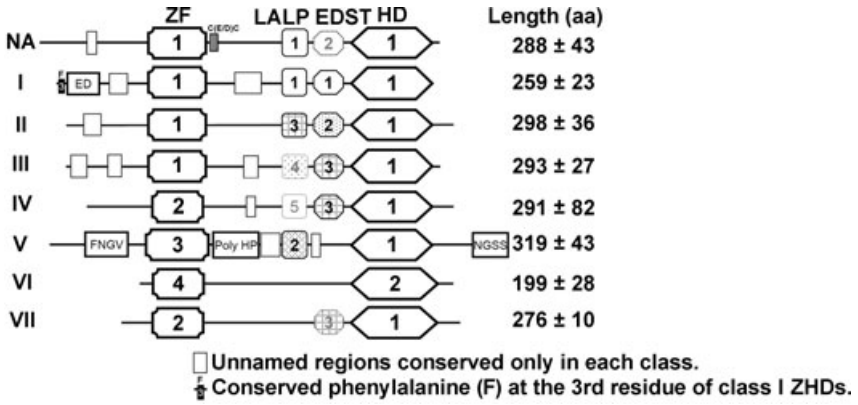

Figure 5. Domain and motif distribution in different zinc finger-homeodomain (ZHD) proteins.

Number and shading in the blocks indicate variant versions of the ZF and HD domains as well as the LALP and EDST motifs. Blocks drawn with dotted lines indicate low-level similarity of the motif within the corresponding class. Except for the linker region between the ZF and HD domains, lengths of the proteins, domains and motifs are approximately proportional. NA, non-angiosperm ZHDs.

an upstream ATG codon, their deduced sequences upstream of the initiation methionine still share high level similarity with the predicted N-terminal sequences of other MIFs (data not shown).

Phylogenetic analysis showed that gymnosperm MIFs form a monophyletic group distinct from angiosperm MIFs (Figure 6), suggesting a single common ancestor for all seed plant MIFs. Within the gymnosperm clade, PtaMIF7 and WmiMIF form a distinct branch; their most N-terminal regions remain most similar to those of non-angiosperm ZHDs (data not shown). The rice OsMIF3 and OsMIF4 contain an N-terminal region very similar to that characteristic of eudicot MIFs rather than that of other grass MIFs. It is thus not surprising that they are distantly related to the clade containing all other grass MIFs. In fact, the presence of OsMIF3 and OsMIF4 decreased the support for the eudicot clade from a bootstrap value of 82 to 61. Within the eudicot clade, AtMIF1 and AtMIF3 are immediate sisters; consistent with the finding that they are generated by a recent duplication (Hu and Ma 2006).

MIF genes might have originated from *ZHD* genes

The *MIF* and *ZHD* genes are related because they share a putative zinc finger that is distinct from other known zinc fingers. It is possible that MIFs were derived from ZHDs by losing the HD domain; alternatively, ZHDs might have originated from MIFs by gaining the HD domain. Furthermore, it was not clear whether there was a single/multiple origin(s) for *MIF* or *ZHD* genes if one is derived from the other. To address these questions, phylogenetic analysis was conducted using the ZF domain (56 residues) for all MIF and ZHD proteins. Figure 7A shows that all MIFs are clustered into one clade clearly separated from ZHDs though lacking statistical support due to a limited number of informative sites for the large sequence set. The single MIF clade suggests that *MIF* genes had a single origin, although another possibility cannot be completely ruled out. Nevertheless, it is unlikely that eudicot MIFs were directly derived from eudicot ZHDs or *vice versa*. Since the current dataset indicates that ZHDs originated before the split of land plants, much earlier than MIFs, it is favored that MIFs might be derived from ZHDs by loss of the HD domain but not *vice versa*. If we root the tree between *Physcomitrella* and *Selaginella* (and other) ZHDs, it seems that MIFs were derived from ZHDs after the emergence of the ZHD Classes I–III, but before or near the emergence of Classes IV–VI (Figure 7A).

Two additional lines of evidence further support the above speculation. First, no gene encoding the ZHD-type homeodomain and lacking the ZF domain has been identified, thus it is unlikely that MIFs evolved into ZHDs by gain of a pre-existing HD domain similar to the ones present in the ZHDs. Second, a C(E/D)C signature conserved immediately at the C-terminal side of the ZF domain of MIF proteins could be found in non-angiosperm ZHDs, but rarely in angiosperm ZHDs (Figures 5,7B). A four-residue spacer between the ZF domain and this signature is also conserved between MIFs and non-angiosperm ZHDs. Such a sequence feature suggests an evolutionary connection outside the ZF domain between MIFs and basal ZHDs.

The ZF domain sequences of MIFs and ZHDs exhibit considerable differences, supporting their separate evolutionary histories after the origin of *MIF* genes (Figure 7C). Ten out of the 54 consensus residues in the ZF domain (not considering those of Class V and VI ZHDs) are different. In particular, the L7Q, H17Y, G23R and P27A differences between ZHDs and MIFs involve residues of distinct chemical properties and might contribute to functional divergence of these two groups of proteins. At other positions that MIF and ZHD proteins share

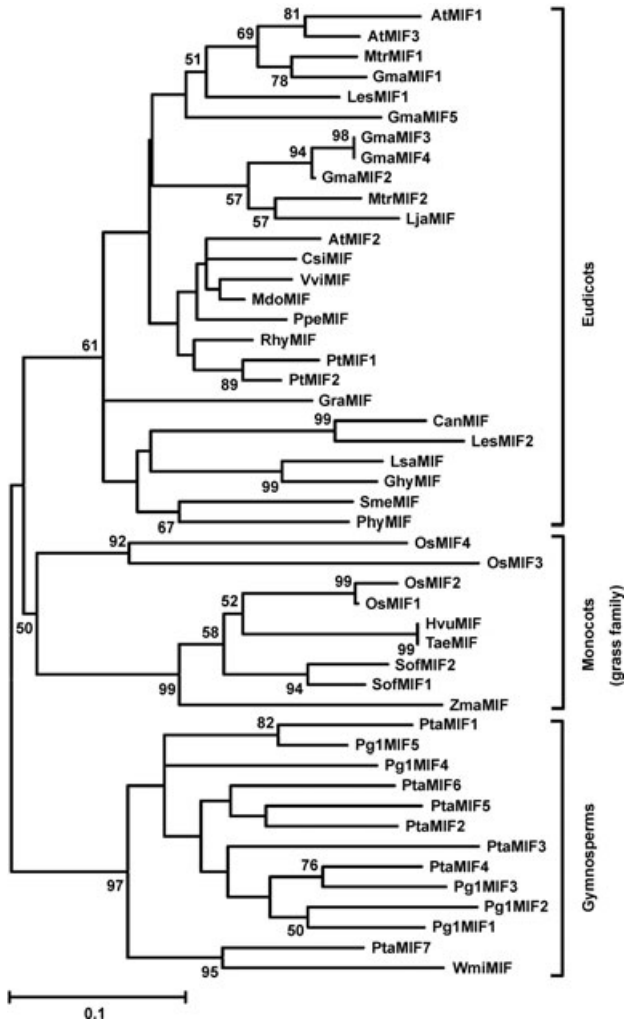

Figure 6. Neighbor-joining (NJ) phylogenetic tree of MINI ZINC FINGER (MIF) proteins.

Bootstrap values greater than 50% for branch support are shown.

the same consensus residue, MIFs frequently display a higher level of conservation (Figure 7C), consistent with a more recent history of *MIF* genes.

Arabidopsis T-DNA insertion lines

As an attempt to dissect the function of *ZHD* and *MIF* genes, we obtained available Salk T-DNA insertion lines (http://signal.salk.edu/cgi-bin/tdnaexpress). Homozygous mutants with the T-DNA insertion annotated in the coding region or the 5′ UTR close to the start codon were sequenced to determine the exact insertion position (Table 2). Putative null mutants were thereby identified for *AtZHD5*, *9*, *10* and *13* and *AtMIF3*. Mutations believed to cause downregulation of the gene expression or affect the activity of encoded protein were also identified for some *AtZHD* genes, such as *AtZHD1* and *2*. Under our normal growth conditions, these insertion lines appeared to develop normally. Double homozygous mutants that were generated from available single mutants did not display abnormal morphology either (data not shown). We have also generated overexpression and RNA interference *AtZHD5* transgenic plants, which also seemed normal (data not shown). Given that most *AtZHD* genes have a closely related homolog that shares a similar expression pattern, these results suggest a high level of functional redundancy in this family.

Discussion

Evolution of the *ZHD* gene family

Our extensive searches indicate that *ZHD* genes are plant-specific, having identified *ZHD* genes from major groups of land plants, including seed plants, *Selaginella* and *Physcomitrella*, but not from the single-cell green alga *Chlamydomonas*. Therefore, *ZHD* genes likely have originated prior to the divergence of all land plants, but possibly after the split of the land plant lineage from the algal groups. It is believed that the green alga Charales (stoneworts) is the sister of all land plants (Bhattacharya and Medlin 1998; Karol et al. 2001). Further investigation of Charales and other algae is needed to more precisely determine the origin of *ZHD* genes.

Our phylogenetic analysis further indicated that ZHDs have expanded considerably during angiosperm evolution (Figure 4). The angiosperm *ZHD* genes form six relatively well-supported clades (Classes I–VI) and one poorly supported clade (Class VII). Classes I, II, VI and VII each contain homologs from magnoliids, monocots and/or eudicots (Figure 4); these three major groups represent over 95% of angiosperm species. In addition, we identified a partial EST sequence of *ZHD* from *Amborella*, the basalmost angiosperm, which likely belongs to Class III according to sequence analysis (data not shown). Therefore, our results suggest that at least five, perhaps seven ZHD groups were present before the divergence of angiosperms. Since all *ZHD* genes of flowering plants likely originated from a single copy in the seed plant ancestor, the multiple groups present before the divergence of most of the angiosperms were due to duplication before or near the divergence of extent angiosperms. Because the placement of gymnosperm ZHDs is not certain, it is possible that ZHDs duplicated once or more before the divergence of angiosperms and gymnosperms.

The branch lengths in the phylogenetic tree and sequence similarity suggest that different classes of *ZHD* genes have evolved at different rates. Class I contains genes that are quite similar throughout the entire sequence and seems to be more basal than other angiosperm clades. In addition, the Class I members AtZHD1 and AtZHD2 show the strongest affinity for dimerization among the 14 *Arabidopsis* ZHDs (Tan and Irish 2006). These results support the idea that Class I

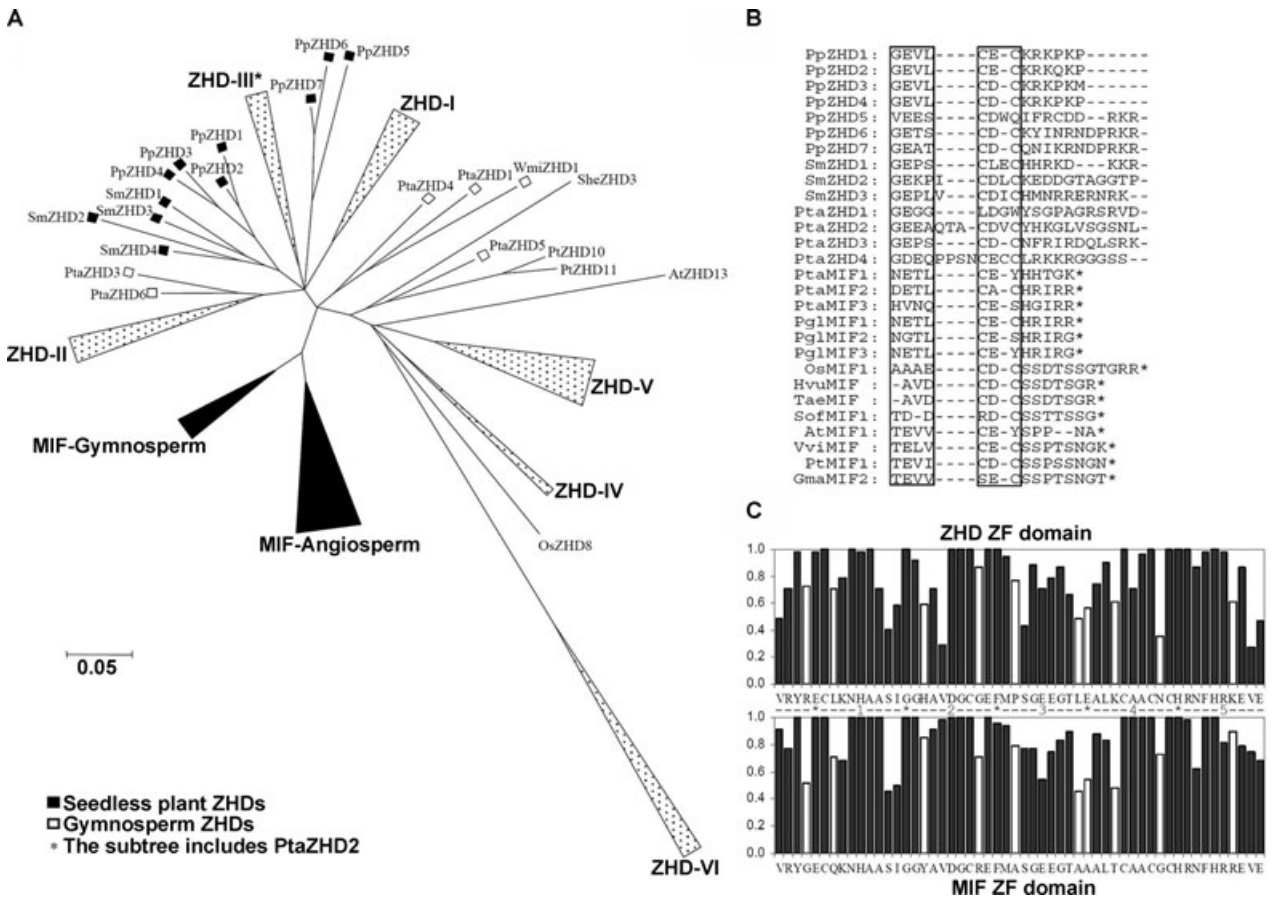

Figure 7. Relationship between zinc finger-homeodomain (ZHD) and MINI ZINC FINGER (MIF) proteins.

(**A**) Phylogenetic tree of ZHDs and MIFs constructed using the ZF domain (56 AA). Except for seedless plant and gymnosperm ZHDs, subtrees are compressed when they are consistent with previous analysis in Figures 4 and 6.

(**B**) Alignment of the region immediately at the ZF C-terminal side of basal ZHDs and representative MIFs. *C-terminal ends of MIFs.

(**C**) Comparison of the ZF-domain consensus sequences of ZHDs (68 sequences; excluding class V and VI) and MIFs (48 sequences). Insertional residues in some MIFs are excluded. Bars represent the frequency of the consensus residues. Bars in white highlight the residues different between ZHDs and MIFs.

ZHDs have conserved functions (under constraints) and have been subject to strong purifying selection.

Near the other end of the spectrum of divergence, phylogenetic and sequence analyses both suggest that class VI ZHDs have evolved more rapidly than others in angiosperms. Also, yeast two hybrid analysis of all possible pairs of the 14 *Arabidopsis* AtZHDs showed that the class VI ZHDs, AtZHD13 and 14, are unable to form homodimers or heterodimers with other AtZHD proteins (Tan and Irish 2006). Furthermore, the soybean GmaZHD5 and 6 (also named as GmZF-HD1 and 2 by Park et al. (2007)) are also class VI ZHDs. Truncated GmaZHD5 and 6 recombinant proteins lacking the entire ZF domain can still bind to ATTA repeats in the *GmCaM4* promoter, suggesting that dimerization may not be required for protein–DNA interaction (Park et al. 2007). Thus, class VI ZHDs could function as monomers, or interact with other non-ZHD proteins. Notably, AtZHD12 in the conserved class V is divergent in sequence, lacking the entire N-terminal region, and the NGSS and Poly HP motifs found in most class V ZHDs (Figure 5). AtZHD12 also could not dimerize (Tan and Irish 2006). Furthermore, the highly conserved methionine (M) at the 26th position of the ZF domain (Figure 7C) is replaced by other amino acids in all class VI ZHDs and AtZHD12. This methionine may be required for protein-protein interaction, since AtZHD12, 13 and 14 are unable to form protein dimers (Tan and Irish 2006).

AtZHD expression patterns and implications

Our expression analysis indicates that most *AtZHD* genes are widely expressed. There are two general patterns: *AtZHD1-5,7*

Table 2. Summary of screened T-DNA lines for *AtZHD* and *AtMIF* genes

Gene	T-DNA line	Insertion position and comments
AtZHD1	Salk_133857	5′UTR, −31 bp, sequenced[a]; expression greatly downregulated
	Salk_014031	Indicated in the middle of exon, but failed in genotyping
	Salk_014023	Indicated in the middle of exon, but failed in genotyping
AtZHD2	Salk_017963	Exon, 25 bp upstream of the stop codon, sequenced
AtZHD5[b]	Salk_097388[c]	Exon, +237 bp, sequenced
AtZHD7	Salk_096579	5′UTR, indicated at −175 bp
AtZHD9	Salk_085482[c]	Exon, +381 bp, sequenced
	Salk_123593	Indicated at −125 bp, but failed in genotyping
AtZHD10	Salk_059288[c]	Exon, +698 bp, sequenced
	Salk_006394	5′UTR, ∼ −100 bp
AtZHD13	Salk_051655[c]	Exon, +58 bp, sequenced
	Salk_092897	Exon, 26 bp upstream of the stop codon, sequenced
AtZHD14	Salk_068489	Promoter, indicated at −219 bp, but failed in genotyping
AtMIF1	Salk_038432	Promoter, −246 bp, sequenced, reported (Hu and Ma 2006)
AtMIF3	Salk_009428[c]	Exon, +65 bp

[a]The position of nucleotide A in the start codon is numbered as +1 bp. [b]Other three lines indicated to be inserted in the exon of *AtZHD5* failed in genotyping, and no morphological phenotype was observed from those plants. [c]The mutation is believed to disrupt the gene function.

(Classes I–III) are expressed in various organs of the shoot, with strong expression in the inflorescence, but not detectable in the root, whereas *AtZHD6, 8-14* (classes III, V and VI) are all expressed in the root, as well as organs of the shoot (except *AtZHD12,* which was only detected in the root). This is consistent with the previous observation that the Class I *ZHD* gene *Flaveria FbHB2* is also expressed in the shoot but undetectable in the root (Windhovel et al. 2001).

Tan and Irish (2006) reported that 13 *AtZHD* genes were expressed predominantly or exclusively in floral tissues. We found that only eight *AtZHD* genes had strong expression in the inflorescence and/or flower. In addition, some of these genes also showed strong expression in 3-d-old seedlings. It is likely that these *AtZHD*s are expressed preferentially in vegetative and reproductive shoot apical meristems. RNA *in situ* hybridization of *AtZHD5* indicated that it was indeed preferentially expressed in shoot apical meristems (Figure 2C). Affymetrix microarray data also indicate that *AtZHD1*, *3*, *4*, *5* and *7* are expressed predominantly in the shoot apex (Schmid et al. 2005) and (http://www.cbs.umn.edu/arabidopsis/). Therefore, some *AtZHD*s are active in and may be involved in regulating some aspects of the shoot apical meristem.

Origin and evolution of *MIF* genes

Our phylogenetic analysis of *MIF* genes suggests that the angiosperm and gymnosperm MIFs form separate clades. In addition, the phylogeny with both *ZHD* and *MIF* genes suggests that the *MIF* genes are phylogenetically separate from the *ZHD* genes. The simplest hypothesis is that the angiosperm and gymnosperm lineages each had a single ancestral *MIF* gene, which evolved from a *MIF* gene in the ancestor of seed plants. In addition, ZHDs were found in other major groups of land plants, apparently much earlier than the origin of MIFs. This suggests that the original *MIF* gene might be derived from a *ZHD* gene, perhaps by a pre-mature termination.

Regardless of the exact origin of *MIF* genes, our results indicate that there have been duplication events in both angiosperms and gymnosperms. As supported by the highly similar *AtMIF1* and *AtMIF3* genes, some of the duplication events were probably rather recent. The emergence and subsequent diversification of MIFs suggest that these proteins that presumably lack DNA-binding activity due to the absence of the HD domain offer an evolutionary advantage. Because the MIF proteins are quite similar to the ZHD zinc fingers, which can mediate dimerization of ZHD proteins, it is possible that the MIF proteins might interact with some ZHD proteins, thereby modify the activities of the latter.

The relationship between ZHDs and MIFs might be analogous to that between LIM-HD and LMO (LIM-only) proteins. Interestingly, ZHD proteins are most similar with, though still very distant from, mammalian LIM-HD proteins (Windhovel et al. 2001). The LIM zinc finger is a protein-protein interaction motif and it is present most often as two tandem repeats in LIM-HD and LMO proteins (Bach 2000; Retaux and Bachy 2002). The LIM domain interacts with co-factors to form regulatory complexes (Retaux and Bachy 2002). Nuclear LMO proteins, on the other hand, act as competitors to prevent the formation of functional complexes involving LIM-HD proteins (Milan et al. 1998; Shoresh et al. 1998; Zeng et al. 1998; Retaux and Bachy 2002). However, it should be noted that there is no direct LIM-to-LIM interaction, whereas the ZHD proteins directly form dimers (Windhovel et al. 2001; Retaux and Bachy 2002; Tan and Irish 2006). The MIF and ZHD proteins potentially provide an intriguing system to study co-evolution of potentially interacting proteins in future investigations.

Genome duplication and gene functional redundancy

It was proposed that the *Arabidopsis* genome had undergone three whole-genome duplication events (Maere et al. 2005). In particular, the date for the most recent genome duplication event was estimated to be 24–40 Mya, before the divergence of *Arabidopsis* and *Brassica* (Blanc et al. 2003). We found that

the duplication that produced *AtZHD3* and *AtZHD4,* as well as *AtZHD9* and *AtZHD10,* coincided with the most recent genome duplication event. In addition, *AtZHD1* and *AtZHD2* might also be the result of the same genome duplication, because *AtZHD2* is located in a recently duplicated segment. However, *AtZHD1* is not found in the corresponding duplicated segment, possibly due to a transposition to the present locus. *AtZHD6* and *AtZHD7* as well as *AtZHD11* and *AtZHD9* were likely derived from an earlier event of genome duplication (Blanc et al. 2003). Furthermore, an older genome duplication event proposed by Blanc et al. (2003) might have generated the *AtZHD5*/*7* and *AtZHD8*/*10* pairs. Therefore, more than half of *Arabidopsis ZHD* genes were probably generated by genome duplication.

AtZHD genes likely resulting from genome duplication were paired in the phylogenetic tree (Figure 4). Other angiosperm species also frequently have two *ZHD* genes paired as the terminal branches of the tree, indicating that they are from relatively recent duplication events. It was reported that genome duplication and polyploidy are widespread in angiosperms (Blanc and Wolfe 2004; Cui et al. 2006); therefore, genome duplication might also be a major force for the expansion of *ZHD* genes in these species. Functional redundancies between pairs of recently duplicated *Arabidopsis* genes have been reported a number of times (Liljegren et al. 2000; Pelaz et al. 2000; Albrecht et al. 2005; Hord et al. 2006). Functional redundancy is therefore reasonably expected for at least some *ZHD* genes. A similar situation could also be true for the *MIF* family. The observation that no single T-DNA mutants exhibited discernable morphological and/or developmental defects underpins the hypothesis of gene functional redundancy among this family (Tan and Irish 2006; and our results). Additional genetic studies, including the use of double or even higher-order mutants, will be necessary to uncover the *in vivo* functions of *ZHD* and *MIF* genes.

Materials and Methods

Sequence retrieval and designation

Arabidopsis sequences were obtained from TAIR (www.arabidopsis.org). Rice sequences were obtained from the The Institute for Genomic Research (TIGR) Rice Genome Annotation Project (http://www.tigr.org/tdb/e2k1/osa1). Several annotation errors of rice genes were recognized by sequence comparisons and thereby corrected. Putative poplar *PtZHD* and *PtMIF* sequences were first obtained from the Floral Genome Project (FGP; www.floralgenome.org), and each was used as the queries to perform MegaBLAST against *Populus trichocarpa* trace archives from the National Center for Biotechnology Information (NCBI) database (http://www.ncbi.nlm.nih.gov/BLAST/). More than five different trace files were collected for every putative gene, and these were built into one contig using CAP3 (Huang and Madan 1999) (http://pbil.univ-lyon1.fr/cap3.php), upon which the correct coding sequence (CDS) was obtained. Assembled *PtZHD* and *PtMIF* genes were further confirmed by BLAST search against preliminarily assembled *Populus* genomic sequences in JGI (http://genome.jgi-psf.org/Poptr1/Poptr1.home.html). *Physcomitrella patens PpZHD1* was cloned from genomic DNA based on EST information from the PHYSCObase (http://moss.nibb.ac.jp/) (see below). The other six *PpZHD* sequences and all *Selaginella moellendorffii ZHD* (*SmZHD*) sequences were obtained from released trace archives using similar strategies as for poplar. Homologs from the gymnosperm *Welwitschia mirabilis*, the Magnoliids *Saruma henryi* and *Liriodendron tulipifera*, the basal monocot *Acorus americanus* (sweet flag), the basal eudicot *Eschscholzia californica* (California poppy), the mid-level monocot *Yucca filamentosa* and *Asparagus officinalis*, and the eudicot *Cucumis sativus* (cucumber) were obtained from the FGP cDNA libraries. Internal primers were designed and used to complete the sequencing if needed. The 13 *ZHD* and *MIF* sequences of FGP species were deposited in the GenBank with accession numbers EU200152–EU200164. *ZHD* and *MIF* genes from other plant species were mainly obtained from the TIGR unigene indices (http://tigrblast.tigr.org/tgi) and NCBI EST and/or nr databases (http://www.ncbi.nlm.nih.gov/BLAST/). Some genes of *Solanaceae* species were obtained from the SOL Genomics Network (http://sgn.cornell.edu). Sequences from the C_4 plant *Flaveria* were obtained from Windhovel et al. (2001). Extensive BLAST search was also carried out against other public EST and/or genomic databases to find more potential ZHD homologs.

For five plants with complete genome information, they are represented by the first letter of the genus name in capital and the first letter of the species name. Other plants are represented by the first letter of the genus name in capital and the first two letters of the species name. The following lists the plant species in full name and abbreviation name as well as common name if applicable:

Acorus americanus (Aam; sweet flag), *Allium cepa* (Ace; onion), *Asparagus officinalis* (Aof), *Arabidopsis thaliana* (At), *Capsicum annuum* (Can; pepper), *Curcuma longa* (Clo), *Cucumis sativus* (Csa; cucumber), *Citrus sinensis* (Csi; orange), *Eschscholzia californica* (Eca), *Gerbera hybrida* (Ghy), *Glycine max* (Gma; soybean), *Gossypium raimondii* (Gra; cotton), *Hordeum vulgare* (Hvu; barley), *Lycopersicon esculentum* (Les; tomato), *Lotus japonicus* (Lja), *Lactuca sativa* (Lsa; lettuce), *Liriodendron tulipifera* (Ltu; tuliptree), *Malus domestica* (Mdo; apple), *Medicago truncatula* (Mtr), *Nicotiana benthamiana* (Nbe; wild tobacco), *Oryza sativa* (Os; rice), *Picea glauca* (Pgl; spruce), *Petunia hybrida* (Phy; petunia), *Physcomitrella patens* (Pp; moss), *Prunus persica* (Ppe; peach), *Populus trichocarpa* (Pt; poplar), *Pinus taeda* (Pta; pine), *Rosa hybrida* (Rhy; rose), *Sorghum bicolor* (Sbi; sorghum), *Saruma henryi* (She), *Selaginella moellendorffii* (Sm), *Solanum melongena* (Sme; eggplant), *Saccharum officinarum* (Sof, sugarcane), *Solanum*

tuberosum (Stu; potato), *Triticum aestivum* (Tae; wheat), *Vitis vinifera* (Vvi; grape), *Welwitschia mirabilis* (Wmi), *Yucca filamentosa* (Yfi), *Zea mays* (Zma; maize), *Zingiber officinale* (Zof).

Sequence alignment, motif identification, and phylogenetic analyses

Deduced amino acid sequences containing at least the putative zinc finger domain (ZF) and the homeodomain (HD) were first aligned using CLUSTALX version 1.81 (Thompson et al. 1997) with default parameters. A preliminary NJ phylogenetic tree was produced using the most conserved ZF and HD domains. The order of the sequences was rearranged based on their phylogenetic placement. The whole set of sequences in every single phylogenetic clade or class were then re-aligned using MUSCLE (Edgar 2004). Alignments generated by this approach displayed much better sequence similarity outside the ZF and HD domains. Manual adjustment further improved the alignment and helped identification of class-specific conserved motifs. These refined alignments from individual classes were then aligned together using the profile-alignment option in MUSCLE (Edgar 2004). Assembled alignments were further manually refined. The entire ZF and HD domains plus the LALP and EDST motifs of the ZHD proteins ($56 + 64 + 13 + 10 = 143$ AA) were used for phylogenetic analyses of ZHDs. A similar strategy was applied to align the MIF proteins. The ZF domain, 14 conserved residues at the N-terminal region and 12 residues conserved at the C-terminal region of MIF proteins were used for phylogenetic analysis of MIFs. Only the ZF domain was used when analyzing the phylogeny of both ZHDs and MIFs together.

Three approaches of phylogenetic analyses, neighbor-joining, maximum parsimony and approximate maximum likelihood, were carried out using MEGA version 3.1 (Kumar et al. 1994), PAUP (Swofford 1998) and PHYML (Guindon and Gascuel 2003; Guindon et al. 2005), respectively. Neighbor-joining method was carried out with 1 000 bootstrap replicates using the *p*-distances for distance measures, and the pairwise deletion option for gaps. Maximum parsimony was carried out with 500 bootstrap replicates using the branch-and-bound or full heuristic search, stepwise addition, and tree-bisection-reconnection (TBR) algorithm. Maximum likelihood was carried out using JTT as the amino acid substitution model, BIONJ as the starting tree, with optimized parameters (gamma = 1.00, proportion of invariant = 0.089) and 1 000 bootstrap replicates. Phylogenetic trees were edited using TreeExplorer that is integrated in MEGA3.1.

RT-PCR examination of the expression pattern of *AtZHD* genes

Table 3 lists RT-PCR primers for the 14 *AtZHD* genes. RNA extraction, RT-PCR, and the internal control gene *APT1* were as described previously (Hu et al. 2003). PCR reactions were repeated two or three times.

RNA *in situ* hybridization of *AtZHD5*

Sections of wild-type inflorescence and 2-week-old plants were prepared and hybridized with radioactively-labeled probes as described previously (Flanagan and Ma 1994). Primers oMC712 and oMC713 (Table 3) were used to clone a fragment of the *AtZHD5* coding region into the pGEM-T vector (Promega, Madison, WI, USA), yielding the plasmid pMC2995. pMC2995 was linearized with *Nco*I and used for *in vitro* transcription with the SP6 polymerase to synthesize the antisense probe. For the control sense probe, pMC2995 was linearized with *Not*I and transcribed with the T7 polymerase.

Cloning of the *Physcomitrella patens PpZHD1* gene

PpZHD1 was first identified from partially sequenced full-length clones at PHYSCObase (http://moss.nibb.ac.jp/). Primers

Table 3. Primers for reverse transcription-polymerase chain reaction (RT-PCR) examination of the expression of *AtZHD* genes

Gene[a]	Primer	Sequence ($5' \rightarrow 3'$)
At1g14440 (*AtZHD4*)	oMC866	CCACCTCCAATGCCGTTACATG
	oMC867	CGGTGGTTACGCCATCTTCCTC
At1g14687 (*AtZHD14*)	oMC868	GGCTGCCGTGAATACTCTCAAC
	oMC869	ATCCTTCAACGTCCATCCCAAC
At1g69600 (*AtZHD11*)	oMC870	TTTCCAACCAGCTTTCTCTGCG
	oMC871	CCCATCGCCGGAGAACTAACTG
At1g75240 (*AtZHD5*)	oMC712	TCCATCTCCGCCGCAGCTAAAC
	oMC713	TCCTTCTCCGCTTGATTGTCCG
At2g02540 (*AtZHD3*)	oMC874	GGTGGGCATGGGAACATGAACC
	oMC875	CAAGCTTCCTCCGCCTTCTTCC
At2g18350 (*AtZHD6*)	oMC876	CCGCCAACAAAGAGAAACCCAC
	oMC877	AGATCCTCCGTCGATGACTCCG
At3g28920 (*AtZHD9*)	oMC878	AACCCGAATCCGAAACTCCGAC
	oMC879	TCTCGCCGCCGATACCGTTATC
At3g50890 (*AtZHD7*)	oMC880	CCCGAATCAGATCCATCCATG
	oMC881	TCTTTCTTCTGCATCCTCCACC
At4g24660 (*AtZHD2*)	oMC882	TAAGCGGTGAGGGAGCCACATC
	oMC883	ACCGCCACGTCATCATGCTTC
At5g15210 (*AtZHD8*)	oMC884	GCCCGCGAAGCCTATTTCTTTC
	oMC885	TTCGGCATCCTCCACCCAACTC
At5g39760 (*AtZHD10*)	oMC886	CCACCGCCGTGATCCAGATAAC
	oMC887	GTGGTTTCCTCCGCCGTTATCG
At5g42780 (*AtZHD13*)	oMC888	CCCATTTGCCGTGAAACTGG
	oMC889	CAGGTCTCACCTTCCACCGAAG
At5g60480 (*AtZHD12*)	oMC890	ACACCAAAGTCAACCACCATCC
	oMC891	TCTGCTCCTCCCGTTGTTAAAG
At5g65410 (*AtZHD1*)	oMC892	ACGACGACGCCGTTTACGACTC
	oMC893	TTGTGACGGTGGTGGTCGGTAG

[a]Genes are listed in the order of chromosomal locus.

oMC1629 (5′-GGGACGAGTTGCAGAGTGACCTG-AG-3′) and oMC1630 (5′-AAGACTTGTCAACCGCATGGGAAG-3′) were used to clone *PpZHD1* directly from *Physcomitrella patens* genomic DNA into the pGEM-T vector (Promega), yielding the plasmid pMC2994. The internal primer oMC1672 (5′-TGAGATGCCAGGTGCTGCGAAG-3′) was used to complete the sequencing. *PpZHD1* sequence was deposited in the GenBank with an accession number DQ099428.

Screen of T-DNA insertion lines

T-DNA insertion lines for *AtZHD* genes were obtained from the *Arabidopsis* Biological Research Center (ABRC). Plants were grown in a greenhouse or growth chamber under long-day conditions. Gene-specific primers were designed based on the indicated T-DNA insertion position, and used in combination with the T-DNA left border primer LBb1 (5′-GCGTGGACCGCTTGCTGCAACT-3′) for genotyping to identify homozygous mutants. The precise T-DNA insertion position of some lines was determined by sequencing the PCR products amplified with LBb1 and one gene-specific primer. Homozygous mutants were further grown and compared with the wild-type plants under normal growth conditions to look for possible growth defects or morphological phenotypes.

Acknowledgements

We thank Kerr Wall for providing an initial genomic sequence dataset of *PtZHD* genes, Lena L. Landherr for sequencing most clones from the FGP species, Yi Hu for helping perform *AtZHD5* RNA *in situ* hybridization experiment, and Robert Carey for providing *Physcomitrella* genomic DNA. We are grateful to Zhenguo Lin, Hongzhi Kong, Laura Zahn and James H. Leebens-Mack for help and discussion on phylogenetic analyses.

References

Akin ZN, Nazarali AJ (2005). Hox genes and their candidate downstream targets in the developing central nervous system. *Cell Mol. Neurobiol.* **25**, 697–741.

Albert VA, Soltis DE, Carlson JE, Farmerie WG, Wall PK, Ilut DC et al. (2005). Floral gene resources from basal angiosperms for comparative genomics research. *BMC Plant Biol.* **5**, 5.

Albrecht C, Russinova E, Hecht V, Baaijens E, de Vries S (2005). The *Arabidopsis thaliana* somatic embryogenesis receptor-like kinases1 and 2 control male sporogenesis. *Plant Cell* **17**, 3337–3349.

Ariel FD, Manavella PA, Dezar CA, Chan RL (2007). The true story of the HD-ZIP family. *Trends Plant Sci.* **12**, 419–426.

Bach I (2000). The LIM domain: regulation by association. *Mech. Dev.* **91**, 5–17.

Bellaoui M, Pidkowich MS, Samach A, Kushalappa K, Kohalmi SE, Modrusan Z et al. (2001). The *Arabidopsis* BELL1 and KNOX TALE homeodomain proteins interact through a domain conserved between plants and animals. *Plant Cell* **13**, 2455–2470.

Bhatt AM, Etchells JP, Canales C, Lagodienko A, Dickinson H (2004). VAAMANA—a BEL1-like homeodomain protein, interacts with KNOX proteins BP and STM and regulates inflorescence stem growth in *Arabidopsis*. *Gene* **328**, 103–111.

Bhattacharya D, Medlin L (1998). Algal phylogeny and the origin of land plants. *Plant Physiol.* **116**, 9–15.

Blanc G, Hokamp K, Wolfe KH (2003). A recent polyploidy superimposed on older large-scale duplications in the *Arabidopsis* genome. *Genome Res.* **13**, 137–144.

Blanc G, Wolfe KH (2004). Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell* **16**, 1667–1678.

Burglin TR (1994). A comprehensive classification of homeobox genes. In: Duboule D, ed. *Guidebook to the Homeobox Genes*. Oxford University Press, New York. pp. 27–71.

Cui L, Wall PK, Leebens-Mack JH, Lindsay BG, Soltis DE, Doyle JJ et al. (2006). Widespread genome duplications throughout the history of flowering plants. *Genome Res.* **16**, 738–749.

Edgar RC (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797.

Englbrecht CC, Schoof H, Bohm S (2004). Conservation, diversification and expansion of C2H2 zinc finger proteins in the *Arabidopsis thaliana* genome. *BMC Genomics* **5**, 39.

Flanagan CA, Ma H (1994). Spatially and temporally regulated expression of the mads-box gene *AGL2* in wild-type and mutant *Arabidopsis* flowers. *Plant Mol. Biol.* **26**, 581–595.

Guindon S, Gascuel O (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* **52**, 696–704.

Guindon S, Lethiec F, Duroux P, Gascuel O (2005). PHYML online–a web server for fast maximum likelihood-based phylogenetic inference. *Nucleic Acids Res.* **33**, 557–559.

Hake S, Smith HM, Holtan H, Magnani E, Mele G, Ramirez J (2004). The role of KNOX genes in plant development. *Annu. Rev. Cell Dev. Biol.* **20**, 125–151.

Halbach T, Scheer N, Werr W (2000). Transcriptional activation by the PHD finger is inhibited through an adjacent leucine zipper that binds 14-3-3 proteins. *Nucleic Acids Res.* **28**, 3542–3550.

Hord CL, Chen C, Deyoung BJ, Clark SE, Ma H (2006). The BAM1/BAM2 receptor-like kinases are important regulators of *Arabidopsis* early anther development. *Plant Cell* **18**, 1667–1680.

Hu W, Ma H (2006). Characterization of a novel putative zinc finger gene *MIF1*: involvement in multiple hormonal regulation of *Arabidopsis* development. *Plant J.* **45**, 399–422.

Hu W, Wang Y, Bowers C, Ma H (2003). Isolation, sequence analysis, and expression studies of florally expressed cDNAs in *Arabidopsis*. *Plant Mol. Biol.* **53**, 545–563.

Huang X, Madan A (1999). CAP3: a DNA sequence assembly program. *Genome Res.* **9**, 868–877.

Hunter CS, Rhodes SJ (2005). LIM-homeodomain genes in mammalian development and human disease. *Mol. Biol. Rep.* **32**, 67–77.

Ito M, Sato Y, Matsuoka M (2002). Involvement of homeobox genes in early body plan of monocot. *Int. Rev. Cytol.* **218**, 1–35.

Karol KG, McCourt RM, Cimino MT, Delwiche CF (2001). The closest living relatives of land plants. *Science* **294**, 2351–2353.

Klug A, Schwabe JW (1995). Protein motifs 5. Zinc fingers. *FASEB J.* **9**, 597–604.

Kosarev P, Mayer KF, Hardtke CS (2002). Evaluation and classification of ring-finger domains encoded by the *Arabidopsis* genome. *Genome Biol.* **3**, research0016.1–research0016.12.

Krishna SS, Majumdar I, Grishin NV (2003). Structural classification of zinc fingers: survey and summary. *Nucleic Acids Res.* **31**, 532–550.

Kumar S, Tamura K, Nei M (1994). MEGA: molecular evolutionary genetics analysis software for microcomputers. *Comput. Appl. Biosci.* **10**, 189–191.

Li J, Jia D, Chen X (2001). *HUA1*, a regulator of stamen and carpel identities in *Arabidopsis*, codes for a nuclear RNA binding protein. *Plant Cell* **13**, 2269–2281.

Liljegren SJ, Ditta GS, Eshed Y, Savidge B, Bowman JL, Yanofsky MF (2000). *SHATTERPROOF* MADS-box genes control seed dispersal in *Arabidopsis*. *Nature* **404**, 766–770.

Mackay JP, Crossley M (1998). Zinc fingers are sticking together. *Trends Biochem. Sci.* **23**, 1–4.

Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, Kuiper M et al. (2005). Modeling gene and genome duplications in eukaryotes. *Proc. Natl. Acad. Sci. USA* **102**, 5454–5459.

Milan M, Diaz-Benjumea FJ, Cohen SM (1998). *Beadex* encodes an LMO protein that regulates Apterous LIM-homeodomain activity in *Drosophila* wing development: a model for LMO oncogene function. *Genes Dev.* **12**, 2912–2920.

Muller J, Wang Y, Franzen R, Santi L, Salamini F, Rohde W (2001). *In vitro* interactions between barley TALE homeodomain proteins suggest a role for protein-protein associations in the regulation of KNOX gene function. *Plant J.* **27**, 13–23.

Park HC, Kim ML, Lee SM, Bahk JD, Yun DJ, Lim CO et al. (2007). Pathogen-induced binding of the soybean zinc finger homeodomain proteins GmZF-HD1 and GmZF-HD2 to two repeats of ATTA homeodomain binding site in the calmodulin isoform 4 (GmCaM4) promoter. *Nucleic Acids Res.* **35**, 3612–3623.

Pelaz S, Ditta GS, Baumann E, Wisman E, Yanofsky MF (2000). B and C floral organ identity functions require *SEPALLATA* MADS-box genes. *Nature* **405**, 200–203.

Retaux S, Bachy I (2002). A short history of LIM domains. (1993–2002): from protein interaction to degradation. *Mol. Neurobiol.* **26**, 269–281.

Schmid M, Davison TS, Henz SR, Pape UJ, Demar M, Vingron M et al. (2005). A gene expression map of *Arabidopsis thaliana* development. *Nat. Genet.* **37**, 501–506.

Schoof H, Zaccaria P, Gundlach H, Lemcke K, Rudd S, Kolesov G et al. (2002). MIPS *Arabidopsis thaliana* database (MAtDB): an integrated biological knowledge resource based on the first complete plant genome. *Nucleic Acids Res.* **30**, 91–93.

Shoresh M, Orgad S, Shmueli O, Werczberger R, Gelbaum D, Abiri S et al. (1998). Overexpression *beadex* mutations and loss-of-function *heldup-a* mutations in *Drosophila* affect the 3′ regulatory and coding components, respectively, of the *Dlmo* gene. *Genetics* **150**, 283–299.

Smith HM, Hake S (2003). The interaction of two homeobox genes, *brevipedicellus* and *pennywise*, regulates internode patterning in the *Arabidopsis* inflorescence. *Plant Cell* **15**, 1717–1727.

Swofford DL (1998). *PAUP** : *Phylogenetic Analysis Using Parsimony* (*And Other Methods*). Sinauer Associates, Sunderland.

Takatsuji H (1998). Zinc-finger transcription factors in plants. *Cell Mol. Life Sci.* **54**, 582–596.

Takatsuji H (1999). Zinc-finger proteins: the classical zinc finger emerges in contemporary plant science. *Plant Mol. Biol.* **39**, 1073–1078.

Tan QK, Irish VF (2006). The *Arabidopsis* zinc finger-homeodomain genes encode proteins with unique biochemical properties that are coordinately expressed during floral development. *Plant Physiol.* **140**, 1095–1108.

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997). The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **25**, 4876–4882.

Tran LS, Nakashima K, Sakuma Y, Osakabe Y, Qin F, Simpson SD et al. (2007). Co-expression of the stress-inducible zinc finger homeodomain ZFHD1 and NAC transcription factors enhances expression of the *ERD1* gene in *Arabidopsis*. *Plant J.* **49**, 46–63.

Vision TJ, Brown DG, Tanksley SD (2000). The origins of genomic duplications in *Arabidopsis*. *Science* **290**, 2114–2117.

von Arnim AG, Deng XW (1993). Ring finger motif of *Arabidopsis thaliana* COP1 defines a new class of zinc-binding domain. *J. Biol. Chem.* **268**, 19626–19631.

Williams RW (1998). Plant homeobox genes: many functions stem from a common motif. *Bioessays* **20**, 280–282.

Windhovel A, Hein I, Dabrowa R, Stockhaus J (2001). Characterization of a novel class of plant homeodomain proteins that bind to the c4 phosphoenolpyruvate carboxylase gene of *Flaveria trinervia*. *Plant Mol. Biol.* **45**, 201–214.

Yanagisawa S (2004). Dof domain proteins: plant-specific transcription factors associated with diverse phenomena unique to plants. *Plant Cell Physiol.* **45**, 386–391.

Zeng C, Justice NJ, Abdelilah S, Chan YM, Jan LY, Jan YN (1998). The *Drosophila* LIM-only gene, *Dlmo*, is mutated in *beadex* alleles and might represent an evolutionarily conserved function in appendage development. *Proc. Natl. Acad. Sci. USA* **95**, 10637–10642.

Zhang X, Feng B, Zhang Q, Zhang D, Altman N, Ma H (2005). Genome-wide expression profiling and identification of gene activities during early flower development in *Arabidopsis*. *Plant Mol. Biol.* **58**, 401–419.

(Handling editor: Jianru Zuo)