Genome-wide characterisation of the HD-ZIP IV transcription factor family in maize: preferential expression in the epidermis

Marie Javelle\textsuperscript{1234}, Catherine Klein-Cosson\textsuperscript{123}, Vanessa Vernoud\textsuperscript{123}, Véronique Boltz\textsuperscript{123}, Chris Maher\textsuperscript{4}, Marja Timmermans\textsuperscript{4}, Nathalie Depège-Fargeix\textsuperscript{123} and Peter M. Rogowsky\textsuperscript{123§}

\textsuperscript{1} Université de Lyon, Ecole Normale Supérieure de Lyon, Université Lyon 1, Unité Reproduction et Développement des Plantes, F-69364 Lyon, France
\textsuperscript{2} INRA, UMR879 Reproduction et Développement des Plantes, F-69364 Lyon, France
\textsuperscript{3} CNRS, UMR5667 Reproduction et Développement des Plantes, F-69364 Lyon, France
\textsuperscript{4} Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA

Running title: HD-ZIP IV family in maize

\textsuperscript{§} corresponding author Peter.Rogowsky@ens-lyon.fr, Tel +33 4 72 72 86 07, FAX +33 4 72 72 86 00
ABSTRACT

Transcription factors of the plant-specific HD-ZIP IV family have been found from moss to higher plants and several family members have been associated with epidermis-related expression and/or function. In maize four of the five characterised HD-ZIP IV family members are expressed specifically in the epidermis, one contributes to trichome development and target genes of another one are involved in cuticle biosynthesis. Assessing the phylogeny, synteny, gene structure, expression and regulation of the entire family in maize, twelve novel ZmHDZIV (Homeo domain leucin zipper IV) genes were identified in the recently sequenced maize genome. Among the 17 genes, 8 form homeologous pairs duplicated after the split of maize and sorghum, whereas a fifth duplication is shared with sorghum. All 17 ZmHDZIV genes appear to be derived from a basic module containing 7 introns in the coding region. With one possible exception, all 17 ZmHDZIV genes are expressed and show preferential expression in immature reproductive organs. Fourteen of 15 ZmHDZIV genes with detectable expression in laser-dissected tissues exhibit a moderate to very strong expression preference for the epidermis suggesting that at least in maize the majority of HD-ZIP IV family members may have epidermis-related functions. 13 ZmHDZIV genes carry conserved motifs of 19 nt and 21 nt in their 3' UTR. The strong evolutionary conservation and the size of the conserved motifs in the 3' UTR suggests that the expression of HD-ZIP IV genes may be regulated by small RNAs.

KEYWORDS

development, epidermis, LCM, OCL genes, RNA secondary structure, RNA regulation, Zea mays

INTRODUCTION

Are transcription factor families expected to have conserved functions? At first sight the answer seems to be no because the classification of transcription factor families is based on DNA binding motifs (Pabo and Sauer, 1992), and there is no obvious reason to postulate a link between a particular type of DNA-protein interaction and a biological function. On the other hand, the enormous expansion of particular transcription factor families in certain eukaryotic taxa suggests that at least some families evolved from a common ancestor,
providing a basis for a possible functional conservation among extant family members. In the case of the homeo domain (HD), a 60 aa DNA binding domain found in all eukaryotic organisms, the different families containing this domain in plants regulate very diverse biological processes, but there seems to be at least some functional conservation within each of these families (Ariel et al., 2007).

The 14 distinct HD-containing families in plants are defined by the presence of additional functional protein domains (Mukherjee et al., 2009). Four of them are regrouped in the HD-ZIP super-family defined by the association of the HD with a leucine zipper (ZIP) domain, a combination found exclusively in the plant kingdom (Schena and Davis, 1992). The HD-ZIP families III and IV are defined by the presence of two additional domains, the steroidogenic acute regulatory protein-related lipid transfer (START) domain and the START adjacent domain (SAD). These two families can be distinguished by a fifth domain, the C-terminal MEKHLA motif, which is present in the HD-ZIP III and absent in the HD-ZIP IV family (Mukherjee and Burglin, 2006). The latter family might present an example for an evolutionary conserved transcription factor family with shared contributions to epidermal development (Javelle et al., 2011).

HD-ZIP IV transcription factors are likely present in all land plants. They have been described specifically in the dicots Arabidopsis (Nakamura et al., 2006), sunflower (Valle et al., 1997), cotton (Guan et al., 2008), tomato (Isaacson et al., 2009) and apple (Dong et al., 1999), in the monocots maize (Ingram et al., 2000), rice (Ito et al., 2003), sorghum (Swigonova et al., 2004) and Phalaenopsis (Nadeau et al., 1996), in the gymnosperm Norway spruce (Ingouff et al., 2003), the lycophyte Selaginella and the bryophyte Physcomitrella (Prigge and Clark, 2006). A comprehensive analysis of ten plant genomes established the number of genes belonging to the HD-ZIP IV family in the genomes of Arabidopsis, poplar, rice, maize, moss and spike moss, but failed to detect any HD-ZIP IV genes in the genomes of three green algae and one red alga (Mukherjee et al., 2009).

Interestingly, expression of nearly all HD-ZIP IV genes studied so far is associated with the epidermis in general or with specific epidermal structures. In Arabidopsis the founding member of the HD-ZIP IV family, GLABRA2 (GL2), is expressed specifically in leaf trichomes and in root atrichoblast cells (Rerie et al., 1994; Masucci et al., 1996), whereas Arabidopsis thaliana MERISTEM LAYER1 (AtML1) and PROTODERMAL FACTOR2
HD-ZIP IV family in maize

(PDF2) are both expressed throughout the epidermal layer of numerous organs (Lu et al., 1996; Abe et al., 2003). In contrast ANTHOCYANINLESS2 (ANL2) is mainly expressed in the subepidermal layer of the vegetative meristem, rosette leaves and seedling roots (Kubo et al., 2008). An exhaustive characterisation of the 12 remaining Arabidopsis HD-ZIP IV genes by promoter::GUS fusions revealed that the promoters of HOMEO DOMAIN GLABRA1 (HDG1), HDG2, HDG5, HDG7, HDG11 and HDG12 direct preferential expression in the epidermal layer of shoot organs (Nakamura et al., 2006). In rice, the five Rice outermost cell-specific (ROC) genes for which detailed expression data are available are all specifically expressed in the epidermis (Ito et al., 2002; Ito et al., 2003). Similarly, in maize, four of the five characterised Outer cell layer (OCL) genes (OCL1, OCL3, OCL4 and OCL5) are almost exclusively expressed in the epidermal layer, whereas OCL2 expression is restricted to the subepidermal layer of floral meristems (Ingram et al., 1999; Ingram et al., 2000). Expression data from other species is sparser, but Picea abies homeobox1 (PaHB1) in Norway spruce (Ingouff et al., 2001) provides yet another example of an HD-ZIP IV family member with epidermis-specific expression, whereas Ovule39 (O39) of Phalaenopsis, with broad expression in all ovule cell types, presents a counter-example (Nadeau et al., 1996).

Functional data based on the phenotypic characterisation of HD-ZIP IV mutants or transgenic knockouts have been obtained in Arabidopsis, maize and tomato and the great majority substantiate a link between HD-ZIP IV transcription factor activity and epidermal development and function. In Arabidopsis the gl2 mutant shows abnormal trichome expansion and ectopic root hair differentiation (Rerie et al., 1994; Di Cristina et al., 1996) and the hdg11 mutant is characterised by excessive trichome branching (Nakamura et al., 2006). Consistent with the sub-epidermal layer expression of ANL2, the anl2 mutant does not produce subepidermal anthocyanin pigments and has an aberrant cellular organisation of the primary root (Kubo et al., 1999). Single mutants in the remaining 13 HD-ZIP IV genes lack detectable phenotypes, but the atml1pdf2 double mutant has a very strong epidermal phenotype; it never forms an organized epidermal layer in the apical part of the pro-embryo and dies at the seedling stage under standard conditions (Abe et al., 2003). In maize, the ocl4 mutant is characterised by ectopic macrohairs (a class of maize trichomes) on the leaf margins and the formation of an additional, sub-epidermal cell layer in the anther wall (Vernoud et al., 2009), whereas a dominant negative form of OCL1 causes a strong delay in kernel development.
(Khaled et al., 2005). In tomato, a point mutation in Cutin deficient2 (Cd2) is likely responsible for a dramatically reduced cutin content in the tomato fruit (Isaacson et al., 2009). In addition, numerous direct or indirect targets of the maize HD-ZIP IV transcription factors are predicted to function in epidermal processes, which further supports the proposed link between this transcription factor family and epidermis-related functions (Abe et al., 2003; Javelle et al., 2010). In particular, OCL1 activates genes involved in lipid metabolism and cuticle biosynthesis in maize (Javelle et al., 2010).

The sequencing of the maize genome (Schnable et al., 2009) offered us the unique opportunity to expand our initial work on 5 OCL genes to all members of the HD-ZIP IV family in maize. We present here a detailed analysis of the phylogeny, synteny, gene structure, expression and conserved regulatory motifs of the family. In particular, a comparison of the expression in the epidermal and sub-epidermal layer of two different organs provided further evidence for a specialisation of HD-ZIP IV transcription factors in epidermis-related functions.

RESULTS

Origin and structure of 17 ZmHDZIV genes in the maize genome

To detect all members of the HD-ZIP IV family present in the maize genome, tblastn searches of all maize BAC sequences of genotype B73 present in Genbank were performed separately with the HD-ZIP domain and the START domain sequences of all HD-ZIP IV family members in Arabidopsis (16) and rice (11), and of the published family members in maize (5). Based on an e-value < 10\(^{-5}\) and the absence of the MEKHLA domain characteristic of the HD-ZIP III family, 17 genes encoding HD-ZIP IV transcription factors were identified in release 4a.53 of the maize genome (Table 1), including the 5 published genes OCL1 to OCL5 (Ingram et al., 1999; Ingram et al., 2000). The additional genes were named ZmHDZIV6 (Zea mays homeo domain leucine zipper IV) to ZmHDZIV17 based on nomenclature recommendations for grass transcription factors (Gray et al., 2009). The gene model(s) predicted for these loci were manually verified and improved. The major modifications and the resulting amino acid sequences are presented in Supplementary Fig. S1. The annotated nucleotide sequence data are available in the Third Party Annotation (TPA)
section of the DDBJ/EMBL/GenBank databases under the accession numbers BK008026 to
BK008042.

In comparison to a previous analysis based on contigs of EST and GSS data (Mukherjee et al., 2009), we found only 17 rather than 21 genes in maize. This difference was likely due to duplicate entries for ZmHDZIV1_OCL1 (OCL1 and AZM459893) and ZmHDZIV2_OCL2 (OCL2 and AZM469787) and a triplicate entry for OCL9 (PUBHC67TD, AZM4101813, TF1) in the previous study. While the respective sequences were not identical, they clearly converged to a single gene model in the present B73 genome sequence and represented either variations between genotypes or alternatives in splicing. Similarly we excluded one rice HD-ZIP IV entry (P0409D09.32 = Os07g24350.1) from our study since the corresponding gene model contained only an HD domain but no START or SAD domain within 10 kb of downstream sequence, reducing the number of rice HD-ZIP IV genes to 11.

Analysis of the positions of the 17 ZmHDZIV genes in the maize genome did not reveal a strong clustering on particular chromosomes (Table 1). The only exception was ZmHDZIV11 and ZmHDZIV12, which were located within 153 kb of each other on chromosome 2. Since none of the neighbouring genes were duplicated, the two genes likely arose from a local rather than a whole genome duplication event.

One quarter of the genes in the maize genome is located in duplicated chromosomal segments, which are remnants of the allotetraploid origin of maize (Swigonova et al., 2004; Schnable et al., 2009). Since the whole genome duplication leading to tetraploidisation occurred after the split from sorghum and rice, these duplicate regions are defined by cosyneny of the respective homeologous maize genes to common reference genes in sorghum or rice. Using the CoGe website (http://synteny.cnr.berkeley.edu/CoGe/) we checked whether ZmHDZIV genes were located in duplicate regions of the maize genome and/or in syntenous regions with rice (Supplementary Fig. S2). We established that 13 of the 17 ZmHDZIV genes were present in duplicate regions and that ZmHDZIV5_OCL5/ZmHDZIV10, ZmHDZIV6/ZmHDZIV8, ZmHDZIV13/ZmHDZIV14 and ZmHDZIV15/ZmHDZIV17 likely formed homeologous gene pairs in four such duplicate chromosomal segments (Table 1). With the exception of ZmHDZIV1_OCL1, the same ZmHDZIV genes as well as ZmHDZIV4_OCL4 localised to chromosomal regions with syntenous regions in rice, which in all cases contained a rice HD-ZIP IV gene. The four duplicate gene pairs all showed
cosynteny with a single locus in rice, the gene pairs ZmHDZIV13/ZmHDZIV14 and ZmHDZIV15/ZmHDZIV17 sharing cosynteny with Roc8 (Table 1).

The phylogenetic relationship between all maize, sorghum, rice, Arabidopsis, Selaginalla and Physcomitrella HD-ZIP IV genes was established by the Bayesian method (Fig. 1). The corresponding tree divided the family into three clades (I to III), one of which contained 3 subclades of equal rank (Ia to Ic). As expected, each of the previously identified gene pairs ZmHDZIV5_OCL5/ZmHDZIV10, ZmHDZIV6/ZmHDZIV8, ZmHDZIV13/ZmHDZIV14 and ZmHDZIV15/ZmHDZIV17 clustered with a single sorghum gene. However, only in the last case the two maize genes were closer to each other than to the sorghum gene confirming that these gene pair ZmHDZIV15/ZmHDZIV17 arose during the whole genome duplication after the split of maize and sorghum. Each of the remaining maize genes had a sorghum ortholog, including the gene pair ZmHDZIV11/ZmHDZIV12, situating this local duplication event prior to the split of maize and sorghum.

Three branches of the tree had values below 75%. They all concerned the position of Arabidopsis sequences, which tended to cluster among each other, whereas sorghum and rice sequences were interspersed between maize sequences. And if the use of alternative settings or methods (maximum likelihood) yielded slightly different trees (data not shown), it did not change the clustering of monocot or dicot sequences. While the clustering can readily be explained with independent duplication events in dicots, it makes it very difficult to infer the function of particular cereal HD-ZIP IV genes from mutant data in the model species Arabidopsis. For example, the maize genes ZmHDZIV1_OCL1 and ZmHDZIV2_OCL2 are close neighbours of the functionally characterised Arabidopsis genes ANL2 (Kubo et al., 2008) and FWA (Kinoshita et al., 2004) in clade Ia, and yet it is impossible to determine from the phylogenetic tree, whether ANL2, FWA, HDG1 or HDG7 is the Arabidopsis ortholog of ZmHDZIV2_OCL2 or ZmHDZIV1_OCL1. More generally, the numerous duplications in cereals in general and maize in particular may hamper future functional analyses due to functional redundancy, as documented in Arabidopsis (Nakamura et al., 2006).

Analysis of the intron-exon structure of ZmHDZIV genes revealed substantial variation, with the intron number varying between 6 and 11. When focusing on the coding sequence a basic module of 8 exons and 7 introns emerged, from which all gene structures can be derived by the insertion of introns and/or the fusion of exons (Fig. 2). ZmHDZIV genes that
shared a particular novel intron tended to be clustered in the phylogenetic tree. For instance, the division of exon 3 into two exons 3a and 3b was found in all members of clade Ib but not in clades Ia, II or III; the additional intron in \textit{HDZIV9} was actually located at a different position in exon 3. Similarly, the division of exon 7 into exons 7a and 7b was not found in any member of clade Ia, but was shared by all members of clades Ib and II with the exception of \textit{ZmHDZIV7}, which might be explained by a secondary fusion event concerning not only exon 7 but also exons 5 and 6. Finally, the division of exon 8 and the presence of an intron in the 3'UTR were shared by all genes in clades II and III but were not found in any other clade (Fig. 2). It is further noteworthy that none of the new introns inserted within the highly conserved homeo and START domains.

On the other hand, the apparent intron deletions were specific to individual \textit{ZmHDZIV} genes. Whenever an intron at the same position was lost in more than one \textit{ZmHDZIV} gene, the resulting intron-exon structures were clearly distinct. For example, the two events involving intron 2 (\textit{ZmHDZIV4\_OCL4, ZmHDZIV15}) did not resemble each other. In the case of \textit{ZmHDZIV7} and \textit{ZmHDZIV16} two introns appeared to have been lost, leading to the most compact genes with only 6 introns.

**Preferential expression of \textit{ZmHDZIV} genes in immature reproductive organs**

To complement the existing expression data of \textit{ZmHDZIV1\_OCL1} to \textit{ZmHDZIV5\_OCL5} (Ingram et al., 2000) and to establish the expression patterns of the remaining \textit{ZmHDZIV} genes, qRT-PCR experiments were carried out on RNA extracted from major organs of the maize plant. Both vegetative organs such as young seedlings, juvenile and adult leaves, stems or roots, and reproductive organs comprising male (tassel) and female (ear) inflorescences, silks (maize styles) and kernels at 12 or 35 days after pollination (DAP) were examined.

Most \textit{ZmHDZIV} genes exhibited a rather broad expression profile. With the exception of roots and stems, in which the transcript levels of most \textit{ZmHDZIV} genes were low or undetectable, expression of several to many \textit{ZmHDZIV} genes was observed in the other organs tested (Fig. 3 and Supplementary Table S2). All the \textit{ZmHDZIV} genes, except \textit{ZmHDZIV12}, had a relatively high level of expression in reproductive compared to vegetative organs. Indeed, most showed a preferential expression either in immature tassels (\textit{ZmHDZIV4\_OCL4, ZmHDZIV5\_OCL5, ZmHDZIV6, ZmHDZIV8, ZmHDZIV10, ZmHDZIV16}).
ZmHDZIV13, ZmHDZIV14, ZmHDZIV15 and ZmHDZIV17) or immature ears (ZmHDZIV1_OCL1, ZmHDZIV2_OCL2 and ZmHDZIV7). Only the transcript levels for ZmHDZIV16 and ZmHDZIV9 were most abundant in young developing kernels and for ZmHDZIV3_OCL3 in silks (Fig. 3). Interestingly, ZmHDZIV expression levels were consistently higher in immature, developing organs compared to mature, differentiated ones, both in vegetative and reproductive structures. Consequently the relative richness of reproductive organs in young dividing tissues may have contributed to their generally high overall expression levels. One notable exception was ZmHDZIV11, which was almost exclusively expressed in mature tassels that included pollen. Expression of its paralog ZmHDZIV12 was not detected in any of the 13 organs tested, despite the use of several distinct primer pairs that all amplified genomic DNA. This result is substantiated by the absence of EST for this gene in Genbank. In conclusion, 16 of the 17 ZmHDZIV genes were expressed and the majority of them showed preferential expression in immature aerial organs and more precisely young male and female inflorescences.

**Preferential expression of ZmHDZIV genes in the epidermal/L1 cell layer**

Previous *in situ* hybridisation experiments had shown that ZmHDZIV1_OCL1, ZmHDZIV3_OCL3, ZmHDZIV4_OCL4 and ZmHDZIV5_OCL5 exhibited epidermis-specific expression notably in embryo, meristems and young organ primordia (Ingram et al., 1999; Ingram et al., 2000), whereas ZmHDZIV2_OCL2 expression was restricted to the sub-epidermal layer (L2) in floral meristems. To evaluate whether this cell layer specificity was shared by the remaining ZmHDZIV genes, we performed laser microdissection (LMD) followed by RT-PCR experiments on RNA extracted from epidermal (E) versus mesophyll cells (M) of the central part of fully expanded juvenile leaves, and on L1 versus L2 cells of the SAM (shoot apical meristem) (Fig. 4A and 4C). Of the 17 ZmHDZIV genes tested on leaf epidermal and mesophyll cells, amplification products were detected for 9 of them (Fig. 4B). As observed previously (Javelle et al., 2010), ZmHDZIV1_OCL1 transcripts accumulated preferentially but not exclusively in epidermal cells of juvenile leaves. Similar results were obtained for ZmHDZIV3_OCL3 and ZmHDZIV10. On the other hand, highly enriched epidermal expression was detected for ZmHDZIV5_OCL5 and epidermis-specific expression for ZmHDZIV6, ZmHDZIV7, ZmHDZIV8, ZmHDZIV14 and ZmHDZIV16. Experiments on RNA extracted from SAM L1 and L2 cells confirmed these results and gave us information
for five additional genes. \textit{ZmHDZIV4\_OCL4}, \textit{ZmHDZIV13}, \textit{ZmHDZIV15} and \textit{ZmHDZIV17} were also extremely enriched in the L1 layer of the meristem while \textit{ZmHDZIV9} was very weakly expressed in L2 cells but not detected in the L1. As in juvenile leaves, no expression of \textit{ZmHDZIV11} or \textit{ZmHDZIV12} was detected in the SAM samples.

The previous \textit{in situ} hybridisation data in embryos, shoot apices, inflorescences and flowers (Ingram et al., 1999; Ingram et al., 2000), the leaf LMD data and the SAM LMD data were generally in good agreement. However, some quantitative differences were observed between leaf and SAM, such as the increased outer cell layer specificity of \textit{ZmHDZIV3\_OCL3} or the decreased specificity of \textit{ZmHDZIV16}. In addition, notable qualitative differences were observed for \textit{ZmHDZIV2\_OCL2}. While \textit{in situ} hybridisation had shown that \textit{ZmHDZIV2\_OCL2} mRNA was excluded from the L1 in floral meristems (Ingram et al., 2000), the gene was preferentially expressed in this cell layer in the SAM. This exception demonstrated that the preferential or near-specific expression in one of the cell layers could not be automatically extrapolated from the leaf or SAM data to the entire plant, although the conservation of layer preference throughout the plant remained the most likely scenario.

Conservation of two short sequence motifs in the 3’ UTR of \textit{HD-ZIP IV} genes

Previously, the 3’ UTR of \textit{HD-ZIP IV} genes had attracted attention due to the conservation of a 17 nt motif between the two \textit{Picea} genes \textit{PaHB1} and \textit{PaHB2} and several \textit{HD-ZIP IV} genes of \textit{Arabidopsis}, sunflower and cotton (Ingouff et al., 2003). Analysis of the \textit{ZmHDZIV} 3’ UTR sequences detected this motif in 13 of the 17 \textit{ZmHDZIV} genes and allowed us to extend the consensus sequence to 21 nt, 5’ GGTGGTTCGGGTATTGACTTC 3’ (Supplementary Table S3). Further analysis of \textit{HD-ZIP IV} genes from moss, lycopods, gymnosperm, monocots and eudicots demonstrated that 9/13 sorghum, 8/11 rice, 10/16 \textit{Arabidopsis}, 4/4 \textit{Selaginella} and 4/4 \textit{Physcomitrella HD-ZIP-IV} genes contained this conserved motif in their 3’ UTR. Two maize genes, \textit{ZmHDZIV4\_OCL4} and \textit{ZmHDZIV7} carried a mismatch in the strongly conserved core region; in the case of \textit{ZmHDZIV4\_OCL4} this mismatch was conserved in the orthologs in sorghum and rice, hinting at functional implications of the primary sequence of the motif.

Moreover, detailed analysis of the 3’ UTR sequences revealed a second conserved motif of 19 nt (consensus sequence 5’ GGAGTCAAGAACGCACCTC 3’) which was located upstream the 21 nt-motif, in all the species under investigation (Fig. 5, Supplementary Table
Interestingly, the two conserved sequences were either both present or both absent in a given 3' UTR suggesting a functional link between the two motifs. The presence of the conserved motifs in all four HD-ZIP IV genes of Physcomitrella patens, which holds a basal position in the phylogeny of land plants, suggested that the ancestral HD-ZIP IV gene(s) possessed the two motifs.

Since the two motifs not only co-evolved but also were partially complementary, we investigated the possibility that the two sites formed a specific secondary RNA structure within the 3' UTR using the secondary structure prediction software RNAfold (Gruber et al., 2008). Of the 13 ZmHDZIV genes containing the two motifs, ZmHDZIV1_OCL1, ZmHDZIV2_OCL2, ZmHDZIV3_OCL3, ZmHDZIV6, ZmHDZIV7, ZmHDZIV8, ZmHDZIV10 and ZmHDZIV17 were predicted to form a stem-loop structure with considerable base-pairing probability between the two motifs (Supplementary Fig. S3). The predicted secondary structures were also well conserved in Arabidopsis and rice despite the fact that the two motifs were not always located at the same distance from the stop codon or separated by the same number of nucleotides (data not shown). In conclusion, the presence of evolutionary conserved 19 and 21 nt motifs suggested that a post-transcriptional regulatory mechanism controlled ZmHDZIV gene expression, possibly via base pairing of the two motifs.

**DISCUSSION**

**Duplications, losses and transpositions of cereal HD-ZIP IV genes**

Cereal genome evolution has been marked by two whole genome duplications. The first concerns all cereals and took place ~90 Mya in the common ancestor of rice, sorghum and maize, while the second occurred ~12 Mya specifically in the lineage leading to maize (Salse et al., 2009). As a consequence of the second duplication event, approximately 25% of the genes in the maize genome possess closely related paralogs (Schnable et al., 2009). Thus it is not surprising that the maize genome with 17 members contains more HD-ZIP IV genes than the sorghum and rice genomes with 13 and 11 members, respectively. Unexpectedly in three cases one of the maize genes was slightly more similar to its sorghum ortholog than to its maize homeolog (Fig. 1) raising the possibility of a more ancient duplication in the common ancestor of maize and sorghum after its split from rice and subsequent gene loss in sorghum. While this possibility cannot be excluded, several arguments favour the hypothesis
that the homeologous pairs ZmHDZIV5/OCL5/ZmHDZIV10, ZmHDZIV6/ZmHDZIV8 and ZmHDZIV13/ZmHDZIV14 arose during the second whole genome duplication event ~12 Mya, just like the pair ZmHDZIV15/ZmHDZIV17 for which the two maize genes are closer to each other than to their sorghum ortholog. Firstly, all 6 maize genes concerned are located in the 25% of the genome duplicated during the maize-specific whole genome duplication. Secondly, the 6 maize regions show synteny only to the three sorghum regions carrying HD-ZIP IV genes and not to any other part of the genome. Thirdly, there is no documentation of a major duplication event concerning the common ancestor of maize and sorghum but not rice. Finally, a strong bias in gene loss and expression suggests that maize homeologs evolve at different rates, providing a possible explanation for the stronger conservation of one of them (Schnable et al., 2011).

With regard to rice the gene pairs ZmHDZIV5/OCL5/ZmHDZIV10 and ZmHDZIV6/ZmHDZIV8 cluster as expected with single gene from rice. However, the phylogenetic clade encompassing the two maize gene pairs ZmHDZIV13/ZmHDZIV14 and ZmHDZIV15/ZmHDZIV17 contains two sorghum genes but only a single rice gene, Roc8, suggesting either gene loss in rice or a specific duplication in the branch leading to sorghum and maize. We favour the first hypothesis, and more precisely a gene loss on rice chromosome 2, since there is not only strong synteny between rice chromosome 2 (position 32 600 000 containing no Roc) and rice chromosome 6 (region around Roc8) but also between this region of rice chromosome 2 and maize chromosome 4 (region around ZmHDZIV13), maize chromosome 5 (region around ZmHDZIV14), maize chromosome 6 (region around ZmHDZIV15) and maize chromosome 9 (region around ZmHDZIV17). In fact the data converge to a model in which a single ancestor gene was duplicated during the first whole genome duplication ~90 Mya in the ancestor of rice, sorghum and maize. During the second whole genome duplication ~12 Mya both copies were duplicated in maize, whereas the copy on chromosome 2 was lost in rice.

Duplication of the maize ZmHDZIV11 and ZmHDZIV12 genes likely involved a different evolutionary path. ZmHDZIV11 and ZmHDZIV12 are located near each other on maize chromosome 2, share less extensive sequence similarity than the paralogs described previously, and each gene has a sorghum counterpart. While they are phylogenetically related to ZmHDZIV16, neither the pair ZmHDZIV11/ZmHDZIV12 nor ZmHDZIV16 is located in
duplicate regions of the maize genome. These data lead to a hypothesis in which a copy of ZmHDZIV16 was inserted elsewhere in the genome of the common ancestor to maize and sorghum, associated with a local duplication either during this insertion event or afterwards. One might speculate that this insertion was possibly mediated by transposable elements, helitrons being prime candidates (Yang and Bennetzen, 2009).

The remaining 6 ZmHDZIV genes all possess a single sorghum and rice ortholog. While the high sequence similarity between these maize and rice genes is generally reflected by synteny between their chromosomal locations, there is one exception. ZmHDZIV1_OCL1 and Roc5 share strong sequence similarity and are closest neighbours in the phylogenetic tree, and yet ZmHDZIV1_OCL1 is located in a region of maize chromosome 3 that has no synteny to the rice genome; in fact this region corresponds to a gap of 2.92 Mbp in the otherwise strong synteny between maize chromosome 3 and rice chromosome 1. On the other hand the region of rice chromosome 2 containing Roc5 shows extensive synteny to regions of maize chromosome 4 and maize chromosome 5, which do not contain ZmHDZIV genes. Consequently one of the two genes must have been displaced in the genome, ZmHDZIV1_OCL1 being the better candidate due to the well documented plasticity of the maize genome (Wang and Dooner, 2006), which seems to be higher than that of rice (Sweredoski et al., 2008).

As reported previously, whole genome duplications affecting HD-ZIP IV genes also occurred independently in the eudicot branch leading to Arabidopsis where 14 of the 16 family members have been defined as paralogous gene pairs by Nakamura et al (2006): HDG1/ANL2, HDG2/HDG3, HDG4/HDG5, FWA/HDG7, HDG8/HDG9, HDG11/HDG12 and ATML1/PDF2. Consequently it is not surprising to find more Arabidopsis than rice, maize or sorghum genes in some clades. For example, the ZmHDZIV4_OCLA clade includes the Arabidopsis HDG4/HDG5 pair and only one gene from rice (Roc3), maize (ZmHDZIV4_OCLA) and sorghum (Sb01g028160). More unexpectedly, even the Arabidopsis gene pairs tend to cluster together in the phylogenetic tree rather than being interspersed among the cereal genes. This suggests the presence of a rather limited number of HD-ZIP IV genes in the common ancestor to monocots and dicots and numerous duplication events after the separation of the two lineages. Taking into account the clustering of the eight extant HD-ZIP IV genes from Physcomitrella and Selaginella in a single clade, one may want to
postulate gene loss in these two species and the presence of three HD-ZIP IV genes corresponding to the clades I, II and III (Fig. 1) in the ancestor of land plants. These findings raise the question, why the majority of duplicated genes have been retained in the respective genomes. Previous findings in eudicots suggest that this is a general feature of transcription factor genes, which show significantly higher retention than genes fulfilling basic cellular functions (Tang et al., 2008). In addition one may speculate that the specialisation of epidermal functions in land plants necessitated additional control mechanisms and it may not be a coincidence that all the HD-ZIP IV genes involved in trichome development, which are GL2, HDG11 and HDG12 from Arabidopsis (Nadeau et al., 1996), ZmHDZIV4_OCL4 from maize (Vernoud et al., 2009) and GaHOX1 from cotton (Guan et al., 2008), are members of clade Ib (data not shown for GaHOX1).

Intron insertions and losses in cereal HD-ZIP IV genes

In parallel to gene number, the structure of HD-ZIP IV genes underwent evolutionary changes. Applying the rules of parsimony and taking into account both the conservation of certain modifications of the intron-exon structure and the phylogenetic position of the corresponding genes, an ancestral HD-ZIP IV gene with 7 introns in the coding region can be deduced. This basic unit was modified by the insertion of introns into exons 3, 7 and/or 8, which in each case are shared between a group of closely related family members. Based on the published intron-exon structure in Arabidopsis (Nakamura et al., 2006) the complete absence of additional introns in exons 3, 7 and 8 of all genes belonging to clade Ia, the presence of an additional intron in both exon 3 and exon 7 of all members of clade Ib (with the exception of AtGL2 lacking the insertion in exon 3) and the restriction of an additional intron in exon 8 to clades II and III were evolutionary conserved between maize and Arabidopsis. In the case of exon 8, the new intron is positioned just downstream of the coding region in the 3' UTR. The splicing of this intron may play a role in transcript stability as suggested by the precedent of the Arabidopsis THIAMINE C SYNTHASE A gene (Bocobza et al., 2007).

On the other hand, the predicted intron losses seem to have occurred more recently, and as independent events, since these occurred in non-conserved positions and in distantly related ZmHDZIV genes. Beyond the coding sequence, introns of varying number and size are also found in the 5' UTR possibly provoking an intron-mediated enhancement phenomenon,
which enhances the level of translation. For example, in the AtEFIα-A3 gene, the presence of a long intron in the 5’ UTR is sufficient to enhance gene expression in plants in a size dependent manner (Chung et al., 2006). Neither the insertions nor the losses of introns can readily be linked to modifications in gene function, since even in the model species Arabidopsis only very few of the 16 hdg single or double mutants have been attributed detectable developmental phenotypes despite a rather exhaustive study (Nakamura et al., 2006).

Wide-spread epidermal expression among ZmHDZIV genes

The expression profiles of the 17 ZmHDZIV genes in 15 organs of the maize plant represent the first comprehensive analysis of the HD-ZIP IV family in monocots and complement an earlier study in the dicot model species Arabidopsis (Nakamura et al., 2006). No expression was detected for ZmHDZIV12 in any of the organs tested, suggesting that the gene is either not expressed or exhibits a highly restricted spatiotemporal expression pattern. All of the remaining 16 ZmHDZIV genes show strong expression in reproductive organs (ear, tassel) and half show moderate expression in leaves. This roughly reflects the situation in Arabidopsis, where only two of the 16 Arabidopsis genes, HDG3 and HDG8, are not expressed in flowers. Contrary to Arabidopsis, where approximately half of the 16 ZmHDZIV genes have substantial expression in root and stem, only ZmHDZIV1_OCL1, ZmHDZIV2_OCL2 and ZmHDZIV3_OCL3 are expressed above basal levels in these organs. Another difference with Arabidopsis resides in the fact that the four paralogous pairs in maize have very similar expression patterns, whereas only two of the 6 paralogous pairs in Arabidopsis behave in a similar fashion (Nakamura et al., 2006). This may indicate that the duplication events in Arabidopsis are generally more ancient than the ones in maize. All these observations lead to the conclusion that individual genes have lost part of the ancestral expression pattern and/or gained new expression territories.

The use of two different developmental stages for 5 of the maize organs allowed a comparison of ZmHDZIV expression between young, developing and old, differentiated organs. With the notable exception of ZmHDZIV11, which is expressed in mature but not in immature tassels, expression of most ZmHDZIV genes is more abundant in immature compared to mature organs. This observation is consistent with the functional analysis of ZmHDZIV4_OCL4, which demonstrated that morphological defects in ocl4 mutant leaves and
anthers become microscopically detectable at stages when ZmHDZIV4_OCL4 is no longer expressed, but these do occur precisely in those tissues that express ZmHDZIV4_OCL4 earlier in development (Vernoud et al., 2009). Extrapolating these observations, ZmHDZIV genes may set the stage in immature organs for developmental events manifesting themselves in mature organs.

Within a given organ, HD-ZIP IV genes often show predominant expression in a single tissue layer, typically the epidermis or, occasionally, the sub-epidermal cell layer. In Arabidopsis, in situ hybridisation and/or gene-specific promoter fusions to the GUS reporter gene, revealed that nine of the 16 HD-ZIP IV genes show strongly enriched expression in the epidermis, whereas expression of one family member is limited to the sub-epidermal layer (Rerie et al., 1994; Lu et al., 1996; Masucci et al., 1996; Nakamura et al., 2006; Kubo et al., 2008). Similar results were obtained for all those genes characterised by in situ hybridisation in rice (5 genes) and maize (5 genes) (Ingram et al., 1999; Ingram et al., 2000; Ito et al., 2002; Ito et al., 2003). Using laser micro-dissected outer and inner cell layers of juvenile leaves and of the SAM, we obtained layer-specific data for all ZmHDZIV genes but ZmHDZIV11 and ZmHDZIV12. The expression of all the genes was moderately to very strongly enriched in the outer cell layer of both structures, lending further evidence to the hypothesis that most members of the HD-ZIP IV family have epidermis-related functions. As in Arabidopsis, there was one exception, ZmHDZIV9 being weakly but specifically expressed in the L2 of the vegetative meristem.

Expression of almost all family members in the SAM and of only a limited number in juvenile or adult leaves further substantiates the previously mentioned preference for ZmHDZIV genes to be expressed in immature organs and suggests a role of the family very early in the development of the maize leaf in general and its epidermis in particular.

Regulation of HD-ZIP IV genes via a small RNA?

Earlier observations had identified a highly conserved 17 nt sequence motif in the 3’ UTR of several HD-ZIP IV genes from Picea, Arabidopsis, sunflower and cotton (Ingouff et al., 2003). Through an exhaustive analysis of 3’ UTR sequences from all maize, rice, sorghum, Arabidopsis, Selaginella and Physcomitrella HD-ZIP IV genes we established a 21 nt consensus sequence for this motif. In addition, we identified a second conserved motif of 19 nt located between the stop codon and the 21 nt motif. The two motifs are frequently
predicted to base pair in the stem of hairpins in the minimum free energy structure of the 3’ UTR.

Given the size and the high evolutionary conservation of the motifs, their primary sequence and/or the formation of a stem-loop structure through their base pairing may have regulatory implications. Considering the lengths of the motifs, it is interesting to postulate that one or both may be binding sites for a small RNA and consequently play a role in the regulation of *HD-ZIP IV* genes. Indeed, the majority of plant miRNAs are known to regulate gene expression through transcript cleavage and/or translation repression by near-perfect base pairing with their mRNA targets at conserved sites in the coding sequence or 3’ UTR (Voinnet, 2009). One of many examples, genes of the HD-ZIP III family are regulated post-transcriptionally by *miR165/166* binding to a conserved site in the coding sequence corresponding to the beginning of the START domain (Bowman, 2004). It seems unlikely that both motifs are miRNA target sites. Although some plant genes are known to be targets for multiple distinct small RNA, such examples are not conserved throughout evolution (Axtell et al., 2007; Howell et al., 2007). However, it remains possible that one of the conserved motifs corresponds to a small RNA target site, with the second motif regulating its accessibility. Such a situation exists in the *p27* 3’ UTR in *C. elegans* where the binding of Pumilio protein to the first element induces a local change in RNA structure that facilitates the recognition of the second element by *miR221/miR222* and leads to efficient down regulation of *p27* expression (Kedde et al., 2010). More generally, tight secondary structures surrounding a miRNA binding site, such as the paired stem structures predicted to form in the 3’ UTR of *ZmHDZIV* genes, may influence the accessibility of a miRNA to its target, and modulation of these secondary structures by binding of cofactors may control miRNA binding site accessibility (Long et al., 2007). Since the base-pairing between plant miRNAs and their targets is generally very strict, the sequence of a hypothetical miRNA binding to the 19 nt or the 21 nt motif can be deduced. However, no corresponding miRNA has been found in the present releases of miRNA databases (data not shown).

Alternatively one may assume that the secondary structure on its own plays a role without the involvement of a miRNA or other small RNA and that the role of the two motifs is limited to the formation of this hairpin. Indeed, such hairpins serve as binding sites for a variety of proteins. For instance, TAR RNA binding protein is involved in regulation of
mRNAs via binding to a hairpin (Svoboda and Di Cara, 2006). Independently of the underlying mechanism the biological function of the hypothetical regulation of HD-ZIP IV genes by the conserved motifs in their 3' UTR remains open. A link with the preferential expression of HD-ZIP IV genes in the outer cell layer seems to be compromised by the example of HDZIV2_OCL2 which is not L1-preferential but contains the conserved 19 and 21 bp elements in its 3’ UTR. Other potential roles include the fine tuning of HD-ZIP IV mRNA or HD-ZIP IV protein stability which may also contribute to the maintenance of layer-specific cell fates by HD-ZIP IV transcription factors.

CONCLUSIONS

With 17 family members, the HD-ZIP IV family in maize is larger than in sorghum rice or Arabidopsis. ZmHDZIV genes are expressed in most organs and show highest mRNA levels in immature reproductive organs. Highly preferential expression in the epidermal layer is widespread possibly suggesting a specialisation of this transcription factor family in epidermis-related functions. Finally, the regulation of numerous family members may be linked to the presence of two conserved motifs in their 3' UTR.

MATERIALS AND METHODS

Sequence analysis

Gene models were downloaded from release 4a.53 of the B73 maize genome assembly (http://www.maizesequence.org) and manually annotated by comparisons to maize EST data (http://www.maizesequence.org/blast) and alignments with Vector NTI software (Invitrogen). The resulting nucleotide sequence data are available in the Third Party Annotation (TPA) section of the DDBJ/EMBL/GenBank databases under the accession numbers BK008026 to BK008042. Partial cDNA sequences for ZmHDZIV6 to ZmHDZIV17 (excluding ZmHDZIV12) have been assigned the accession numbers JN003608 to JN003618. Duplications in the maize genome and synteny between maize and rice were established with the SynMap tool and confirmed with the GEvo tool at the CoGe website (http://synteny.cnr.berkeley.edu/CoGe/). Secondary structures of 3' UTR sequences were predicted using RNAfold (http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi).
Phylogenetic analysis

The cDNA coding sequences were aligned with MUSCLE v3.5 (Edgar, 2004) and phylogenetic trees established with the MrBayes v3.1 software for the Bayesian inference of phylogeny (Ronquist and Huelsenbeck, 2003) using standard settings (nst=6, rates=invgamma). After an initial analysis with 100,000 generations, additional generations were added until the standard deviation of split frequencies fell below 0.01. The trees were formatted with TreeDyn (Chevenet et al., 2006).

Quantitative reverse transcription polymerase chain reaction

Total RNA from different maize organs (described in Vernoud et al., 2009) was extracted using TRIzol reagent according to the manufacturer’s instructions (Invitrogen). Genomic DNA contaminants were removed from RNA samples using the TURBO DNase from Ambion. DNase treatment and inactivation were performed according to the manufacturer’s protocol (Ambion). Approximately 5 µg of total RNA were reverse transcribed using random hexamers (Amersham Biosciences) and a reverse transcriptase without RNaseH activity (Fermentas) in a final volume of 20 µl.

Quantitative PCR was conducted on an ABI StepOne Plus cycler (Applied Biosystems) using the FastStart Universal SYBRGreen Master mix (Roche). Reactions were prepared in a total volume of 20 µl with 5 µl of 50 times diluted cDNA, 0.6 µl of each primer to a final concentration of 300 nM and 10 µl of 2 x SYBRGreen mix. Amplification was achieved by 10 min initial enzyme activation at 95°C followed by 40 cycles composed of 30 s denaturation at 95°C and 1 min primer annealing/extension at 60°C. To verify the specificity of the amplification of each primer pair, a melting curve analysis was performed ranging from 60°C to 95°C with temperature increasing steps of 0.1 °C/s. Data were analysed with version 2.1 of the StepOne software (Applied Biosystems) and gene expression levels relative to the 18S rRNA reference gene were calculated by the ΔΔC_T method (Schmittgen and Livak, 2008).

Gene specific primers (Supplementary Table S1) were designed at the end of the coding sequence and/or in the 3’ UTR. In case of homeologous gene pairs, PCR product identity was verified by cloning and sequencing. Whenever possible, intron spanning primers
were chosen to ascertain the absence of genomic DNA contamination. For primers not fulfilling this criterion controls without reverse transcriptase were performed.

**Laser capture micro-dissection and RT-PCR**

From the region of maximum width of fully expanded leaf #4 1 cm² sections were fixed in acetone and paraffin-embedded as described (Ohtsu et al., 2007). Epidermal and mesophyll sub-epidermal cells were micro-dissected from 10 µm sections using the Arcturus XT infrared laser capture micro-dissection system with the following settings for epidermal/mesophyll cells, respectively: laser spot size 10/20 µm, laser pulse duration 20/30 ms and laser power 50/70 mW. About 5000 epidermal cells (predominantly adaxial) and 2500 mesophyll cells were collected. For SAM micro-dissection, maize shoot apices were dissected from 2-weeks-old plants in a block of 3 x 2 mm and fixed using the same procedures. Sections of 8 µm were spread on PEN 1.0 membrane slides (Zeiss). About 250000 µm² of L1 and 500000 µm² of L2 cells were captured into an AdhesiveCap 500 tube (Zeiss) using the PALM micro-Beam system with the following parameters: CUT at 36-50 UV-Energy/65 UV-Focus and LPC at 61-75 UV-Energy/63 UV-Focus. In both cases, RNA was extracted with the PicoPure RNA isolation kit (Arcturus). RNA samples were treated with DNase I (Qiagen) and amplified (two rounds) with the TargetAmpTM 2-Round aRNA Amplification kit 2.0 (Epicentre Biotechnologies). aRNA were reverse transcribed and amplified by PCR as described above.

**ACKNOWLEDGMENTS**

We thank Christine Arnould and the members of the Microscopy Centre at INRA Dijon for their essential help in the micro-dissection experiments and Doreen Ware for her precious advice on the bioinformatic analysis of putative small RNA precursors. Isabelle Desbouchages, Alexis Lacroix and Priscilla Angelot are acknowledged for maize culture, Hervé Leyral and Claudia Bardoux for the preparation of buffers and media. We also acknowledge the contribution of the AG (Bariza Blanquier) platform of the SFR Biosciences Gerland-Lyon Sud (UMS344/US8). *ZmHDZIV* expression analysis on laser microdissected SAM L1 and L2 layers was performed as part of NSF award 0820610 to MT. MJ and CKC were supported by PhD fellowships of the French Ministry of Higher Education.
REFERENCES


Ingram GC, Magnard JL, Vergne P, Dumas C, Rogowsky PM (1999) ZmOCL1, an HDGL2 family homeobox gene, is expressed in the outer cell layer throughout maize development. Plant Mol Biol 40: 343-354


FIGURE LEGENDS

Fig. 1: Phylogenetic tree of HD-ZIP IV genes from maize, soghum, rice, Arabidopsis, Physcomitrella and Selaginella

A Bayesian phylogenetic tree was generated using all Arabidopsis (At), maize (Zm, red), rice (Os), sorghum (Sb) Physcomitrella (Pp) and Selaginella (Sm) HD-ZIP IV coding sequences with MrBayes software (2,000,000 generations, average standard deviation of split frequencies = 0.0028). Coding sequences of HD-ZIP III genes from Arabidopsis behaved as an outgroup (HDZIII). Different clades (roman numbers) were annotated manually.

Fig. 2: Genomic structure of maize ZmHDZIV genes

Exons are indicated by thick boxes, introns by thin boxes. The regions coding for the HD and the START domain are coloured in blue and green, respectively. Vertical bars indicate the start and the stop codon, respectively. Conserved exons are given the same number in all genes, the numbering starting arbitrarily with the exon containing the beginning of the highly conserved homeo domain. The accession numbers of the sequences used are listed in Table 1.

Fig. 3: qRT-PCR analysis of ZmHDZIV transcript levels in different maize organs

Total RNA was extracted from the aerial part of seedlings (Sa), juvenile leaf blade (Lbj), juvenile leaf sheath (Lsj), adult leaf blade (Lba), adult leaf sheath (Lsa), stem (St),
seedling root (Rs), mature root (Rm), immature ear (Ei), mature ear (Em), immature tassel (Ti), mature tassel (Tm), silk (Si), 12 DAP kernel (K12) and 35 DAP kernel (K35). Organs are colour coded: vegetative in green, root in grey, reproductive in red and kernel in blue. For each gene, the expression levels obtained by normalization with the 18S RNA gene are presented on relative scales that differ from gene to gene. Data are average values ± SD from two experiments, each carried out in duplicate. For each sample, organs from two to six independent plants were pooled. No data are presented for ZmHDIV12 since the gene was not expressed in any of the organ tested.

**Fig. 4: Expression of ZmHDZIV genes in outer and inner cell layers of juvenile leaf and SAM**

(A) Leaf section showing epidermal cells (blue) and mesophyll cells (green) isolated using the infrared laser capture microdissection system. (B) RT-PCR experiments assessing expression of ZmHDZIV genes in micro-dissected epidermal (E) and mesophyll (M) cells using gene specific primers. (C) SAM section showing L1 cells (blue) and L2 cells (green) isolated using the UV laser capture microdissection system before (left) and after (right) capture of the L2 cells. (D) RT-PCR experiments assessing expression of ZmHDZIV genes in micro-dissected L1 and L2 cells using gene specific primers. In (B) and (D) the concentration of the cDNA templates was normalised according to the abundance of the Actin RT-PCR product; the number of PCR cycles is indicated in brackets.

**Fig. 5: Position of conserved 21 nt and 19 nt motifs in the 3' UTR of maize, sorghum, rice and Arabidopsis HD-ZIP IV genes**

The 3' UTR of maize (Zm), sorghum (Sb), rice (Os) and Arabidopsis (At) HD-ZIP IV genes are drawn to scale starting with the first nucleotide after the stop codon. 3' UTR with corresponding EST data are represented in blue and 3' UTR without EST data in white. Thin green boxes depict introns. Yellow and red boxes indicate the position of the 19 nt and 21 nt conserved motifs, respectively. Motifs with non-conserved nucleotides in their core region are depicted with the same filling. Sequences are grouped according to phylogenetic clades of Fig. 1.
### Table 1: *ZmHDZIV* genes in the maize genome

<table>
<thead>
<tr>
<th>Name¹</th>
<th>Genbank accession</th>
<th>ID in maize genome (release 4a.53)</th>
<th>Chromosome</th>
<th>Position (start in release 4a.53)</th>
<th>Position (end in release 4a.53)</th>
<th>Orientation</th>
<th>Protein size [aa]</th>
<th>Maize chromosome with syntenous region</th>
<th>ZmHDZIV gene in syntenous region²</th>
<th>Presence in sorghum and rice³</th>
<th>Rice chromosome with syntenous region</th>
<th>Roc gene in syntenous region²</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZmHDZIV1_OCL1</td>
<td>BK008026</td>
<td>GRMZM2G026643_T02</td>
<td>Zm3</td>
<td>27 197 008</td>
<td>27 204 687</td>
<td>forward</td>
<td>803</td>
<td>Zm8</td>
<td>NO</td>
<td>Sb, (Os)</td>
<td>NO</td>
<td>na</td>
</tr>
<tr>
<td>ZmHDZIV2_OCL2</td>
<td>BK008027</td>
<td>AC235534.1_FGT007_mod</td>
<td>Zm10</td>
<td>136 522 908</td>
<td>136 529 482</td>
<td>forward</td>
<td>834</td>
<td>Zm2</td>
<td>NO</td>
<td>Sb, Os</td>
<td>Os4</td>
<td>Roc4</td>
</tr>
<tr>
<td>ZmHDZIV3_OCL3</td>
<td>BK008028</td>
<td>GRMZM2G116658_T01</td>
<td>Zm7</td>
<td>136 100 609</td>
<td>136 106 796</td>
<td>forward</td>
<td>863</td>
<td>Zm2</td>
<td>NO</td>
<td>Sb, Os</td>
<td>Os9</td>
<td>Roc6</td>
</tr>
<tr>
<td>ZmHDZIV4_OCL4</td>
<td>BK008029</td>
<td>GRMZM2G123140_T03</td>
<td>Zm1</td>
<td>99 763 128</td>
<td>99 769 093</td>
<td>reverse</td>
<td>884</td>
<td>NO</td>
<td>na</td>
<td>Sb, Os</td>
<td>Os10</td>
<td>Roc3</td>
</tr>
<tr>
<td>ZmHDZIV5_OCL5</td>
<td>BK008030</td>
<td>GRMZM2G130442_T01</td>
<td>Zm4</td>
<td>28 833 054</td>
<td>28 839 237</td>
<td>reverse</td>
<td>796</td>
<td>Zm10</td>
<td>ZmHDZIV10</td>
<td>Sb, Os</td>
<td>Os8</td>
<td>Roc1</td>
</tr>
<tr>
<td>ZmHDZIV6</td>
<td>BK008031</td>
<td>GRMZM2G001289_T02</td>
<td>Zm2</td>
<td>8 237 590</td>
<td>8 243 890</td>
<td>reverse</td>
<td>795</td>
<td>Zm10</td>
<td>ZmHDZIV8</td>
<td>Sb, Os</td>
<td>Os4</td>
<td>Roc2</td>
</tr>
<tr>
<td>ZmHDZIV7</td>
<td>BK008032</td>
<td>GRMZM2G004957_T01</td>
<td>Zm10</td>
<td>84 135 045</td>
<td>84 138 429</td>
<td>forward</td>
<td>769</td>
<td>NO</td>
<td>NO</td>
<td>Sb, Os</td>
<td>Os8</td>
<td>Roc7</td>
</tr>
<tr>
<td>ZmHDZIV8</td>
<td>BK008033</td>
<td>GRMZM2G122897_T01</td>
<td>Zm10</td>
<td>142 981 287</td>
<td>142 986 049</td>
<td>reverse</td>
<td>742</td>
<td>Zm2</td>
<td>ZmHDZIV6</td>
<td>Sb, Os</td>
<td>Os4</td>
<td>Roc2</td>
</tr>
<tr>
<td>ZmHDZIV9</td>
<td>BK008034</td>
<td>GRMZM2G386276_T01</td>
<td>Zm3</td>
<td>187 939 546</td>
<td>187 944 608</td>
<td>reverse</td>
<td>701</td>
<td>Zm6</td>
<td>NO</td>
<td>Sb, Os</td>
<td>Os1</td>
<td>OsTF1</td>
</tr>
<tr>
<td>ZmHDZIV10</td>
<td>BK008035</td>
<td>GRMZM2G118063_T01</td>
<td>Zm10</td>
<td>42 115 584</td>
<td>42 124 124</td>
<td>forward</td>
<td>802</td>
<td>Zm4</td>
<td>ZmHDZIV5_OCL5</td>
<td>Sb, Os</td>
<td>Os8</td>
<td>Roc1</td>
</tr>
<tr>
<td>ZmHDZIV11</td>
<td>BK008036</td>
<td>GRMZM2G438260_T01</td>
<td>Zm2</td>
<td>87 878 884</td>
<td>87 882 865</td>
<td>forward</td>
<td>750</td>
<td>NO</td>
<td>na</td>
<td>(Sb)</td>
<td>NO</td>
<td>na</td>
</tr>
<tr>
<td>ZmHDZIV12</td>
<td>BK008037</td>
<td>GRMZM2G060444_T01_mod</td>
<td>Zm2</td>
<td>88 035 597</td>
<td>88 038 630</td>
<td>reverse</td>
<td>747</td>
<td>NO</td>
<td>na</td>
<td>(Sb)</td>
<td>NO</td>
<td>na</td>
</tr>
<tr>
<td>ZmHDZIV13</td>
<td>BK008038</td>
<td>GRMZM2G126646_T01</td>
<td>Zm4</td>
<td>176 668 879</td>
<td>176 673 174</td>
<td>reverse</td>
<td>698</td>
<td>Zm5</td>
<td>ZmHDZIV14</td>
<td>Sb</td>
<td>Os6</td>
<td>Roc8</td>
</tr>
<tr>
<td>ZmHDZIV14</td>
<td>BK008039</td>
<td>GRMZM2G145690_T01</td>
<td>Zm5</td>
<td>210 323 721</td>
<td>210 327 636</td>
<td>reverse</td>
<td>692</td>
<td>Zm4</td>
<td>ZmHDZIV13</td>
<td>Sb</td>
<td>Os6</td>
<td>Roc8</td>
</tr>
<tr>
<td>ZmHDZIV15</td>
<td>BK008040</td>
<td>GRMZM2G004334_T01</td>
<td>Zm6</td>
<td>111 001 775</td>
<td>111 004 906</td>
<td>reverse</td>
<td>732</td>
<td>Zm9</td>
<td>ZmHDZIV17</td>
<td>Sb, Os</td>
<td>Os6</td>
<td>Roc8</td>
</tr>
<tr>
<td>ZmHDZIV16</td>
<td>BK008041</td>
<td>GRMZM2G109252_T01</td>
<td>Zm7</td>
<td>36 835 558</td>
<td>36 838 693</td>
<td>forward</td>
<td>611</td>
<td>NO</td>
<td>na</td>
<td>Sb</td>
<td>NO</td>
<td>na</td>
</tr>
<tr>
<td>ZmHDZIV17</td>
<td>BK008042</td>
<td>AC201766.3_FGT002_mod</td>
<td>Zm9</td>
<td>8 971 612</td>
<td>8 975 297</td>
<td>forward</td>
<td>699</td>
<td>Zm6</td>
<td>ZmHDZIV15</td>
<td>Sb, Os</td>
<td>Os6</td>
<td>Roc8</td>
</tr>
</tbody>
</table>

¹ Paralogs are coded with the same colour
² na, not applicable
³ Os, *Oryza sativa*; Sb, *Sorghum bicolor*; ( ), not in syntenous position
ADDITIONAL FILES

Supplementary Fig. S1: Deduced amino acid sequences of HD-ZIP IV proteins in maize, Arabidopsis, rice, sorghum, Selaginella and Physcomitrella

Sequences of the HD-ZIP IV proteins from Zea mays (Zm), Arabidopsis thaliana (At), Oryza sativa (Os), Sorghum bicolor (Sb), Selaginella moellendorfii (Sm) and Physcomitrella patens (Pp) used for the phylogenetic analysis were retrieved from http://www.maizesequence.org, http://www.ncbi.nlm.nih.gov, http://rice.plantbiology.msu.edu and http://www.phytozome.net.

Maize gene models were chosen and modified as follows. In the case that multiple gene models were predicted for a single locus, preference was given to the model best supported by EST coverage and conservation with other HD-ZIP IV genes. For four loci, none of the predicted gene models was satisfactory and new models were established. In gene model AC235534.1_FGT007 (ZmHDZIV2_OCL2) the 5' and 3' boundaries of exon 3 were modified based on comparisons between the published genomic sequence (accession number AJ250984) and two EST sequences (accession numbers DR789442 and EE033045). In gene model GRMZ2G060444_T01 (ZmHDZIV12) the putative start codon was moved upstream (538 bp) as in gene model AC214481.2_FG013 present in a previous version of the maize genome (release 3b.50). In GRMZM2G004334_P01 (ZmHDZIV15) an intron was introduced in exon 1 based on the splicing of the closely related ZmHDZIV genes ZmHDZIV14 and ZmHDZIV15. Finally, in AC201766.3_FGT002 (ZmHDZIV17) a 3' UTR (538 bp) was added and exon 1 was expanded (45 bp) based on comparison a full length cDNA sequence (accession number BT066813).

Supplementary Fig. S2: Synteny between maize, rice and sorghum HD-ZIP IV genes

Coding sequences of the 10 maize (Zm) chromosomes showing synteny with the 12 rice (Os) chromosomes (A), the 10 maize (Zm) chromosomes (B) or the 10 sorghum (Sb) chromosomes (C) are indicated by green dots that are grouped into syntenous regions by black squares (http://synteny.cnr.berkeley.edu/CoGe/). The position of ZmHDZIV genes are indicated by red horizontal lines, positions of rice, maize or sorghum HD-ZIP IV genes by vertical lines (blue for rice or sorghum). Green arrows indicate synteny of maize/rice, maize/maize or
maize/sorghum gene pairs that have been validated by high resolution analysis (http://synteny.cnr.berkeley.edu/CoGe/GEvo.pl).

Supplementary Fig. S3: Prediction of base pairing between the conserved 19 nt and 21 nt motifs in the 3' UTR of ZmHDZIV genes

Schematic representation of the most stable RNAfold prediction of base pairing in the 3' UTR of ZmHDZIV genes. Only the part of the 3' UTR containing the 19 nt (blue line) and the 21 nt motif (red line) is presented. Numbers indicate to the position of the corresponding nucleotide in the primary sequence of the 3' UTR. Base-pair probability is expressed by a colour scale.

Supplementary Table S1: Primers used in this study

Supplementary Table S2: Expression of 17 ZmHDZIV genes in major organs of the maize plant as determined by qRT-PCR

Supplementary Table S3: Conservation of a 21 nt and a 19 nt sequence in the 3' UTR of HD-ZIP IV genes from different species
A Bayesian phylogenetic tree was generated using all Arabidopsis (At), maize (Zm, red), rice (Os), sorghum (Sb) Physcomitrella (Pp) and Selaginella (Sm) HD-ZIP IV coding sequences with MrBayes software (2,000,000 generations, average standard deviation of split frequencies = 0.0028). Coding sequences of HD-ZIP III genes from Arabidopsis behaved as an outgroup (HDZIII). Different clades (roman numbers) were annotated manually.
**Fig. 2.** Genomic structure of maize *ZmHDZIV* genes.

Exons are indicated by wide boxes, introns by narrow boxes. The regions coding for the HD and the START domain are coloured in blue and green, respectively. Vertical bars indicate the start and the stop codon, respectively. Conserved exons are given the same number in all genes, the numbering starting arbitrarily with the exon containing the beginning of the very conserved homeo domain. The accession numbers of the sequences used are listed in Table 1.
Fig. 3: qRT-PCR analysis of ZmHDIV transcript levels in different maize organs. Total RNA was extracted from aerial part of seedling (Sa), juvenile leaf blade (Lbj), juvenile leaf sheath (Lsj), adult leaf blade (Lba), adult leaf sheath (Lsa), stem (St), seedling root (Rs), mature root (Rm), immature ear (Ei), mature ear (Em), immature tassel (Ti), mature tassel (Tm), silk (Si), 12 DAP kernel (K12) and 35 DAP kernel (K35). Organs are colour coded: vegetative in green, root in grey, reproductive in red and kernel in blue. For each gene, the expression levels obtained by normalization with the 18S RNA gene are presented on relative scales that differ from gene to gene. Data are average values ± SD from two experiments, each carried out in duplicate. For each sample, organs from two to six independent plants were pooled and data are presented for ZmHDIV12 since the gene was not expressed in any of the organ tested.
Fig. 4: Expression of ZmHDZIV genes in outer and inner cell layers of juvenile leaf and SAM

(A) Leaf section showing epidermal cells (blue) and mesophyll cells (green) isolated using the infrared laser capture microdissection system.

(B) RT-PCR experiments assessing the expression of ZmHDZIV genes in micro-dissected epidermal (E) and mesophyll (M) cells using gene specific primers.

(C) SAM section showing L1 cells (blue) and L2 cells (green) isolated using the UV laser capture microdissection system before (left) and after (right) capture of the L2 cells.

(D) RT-PCR experiments assessing the expression of ZmHDZIV genes in micro-dissected L1 and L2 cells using gene specific primers.

In (B) and (D) the concentration of the cDNA templates was normalised according to the abundance of the Actin RT-PCR product; the number of PCR cycles is indicated in brackets.
Fig. 5: Position of conserved 21 nt and 19 nt motifs in the 3' UTR of maize, sorghum, rice and Arabidopsis HD-ZIP IV genes

The 3' UTR of maize (Zm), sorghum (Sb), rice (Os) and Arabidopsis (At) HD-ZIP IV genes are drawn to scale starting with the first nucleotide after the stop codon. 3' UTR with corresponding EST data are represented in blue and 3' UTR without EST data in white. Thin green boxes depict introns. Yellow and red boxes indicate the position of the 19 nt and 21 nt conserved motifs, respectively. Motifs with non conserved nucleotides in their core region are depicted with the same filling. Sequences are grouped according to phylogenetic clades of Fig. 1.