

Genome-wide identification of C₂H₂ zinc-finger gene family in rice and their phylogeny and expression analysis

Pinky Agarwal · Rita Arora · Swatishmita Ray ·
Ashok K. Singh · Vijay P. Singh · Hiroshi Takatsuji ·
Sanjay Kapoor · Akhilesh K. Tyagi

Received: 1 February 2007 / Accepted: 27 May 2007 / Published online: 4 July 2007
© Springer Science+Business Media B.V. 2007

Abstract Transcription factors regulate gene expression in response to various external and internal cues by activating or suppressing downstream genes in a pathway. In this study, we provide a complete overview of the genes encoding C₂H₂ zinc-finger transcription factors in rice, describing the gene structure, gene expression, genome localization, and phylogenetic relationship of each member. The genome of *Oryza sativa* codes for 189 C₂H₂ zinc-finger transcription factors, which possess two main types of zinc-fingers (named C and Q). The Q-type zinc fingers contain a conserved motif, QALGGH, and are plant specific, whereas C type zinc fingers are found in other organisms as well. A genome-wide microarray based gene expression analysis involving 14 stages of vegetative and reproductive development along with 3 stress conditions has revealed that C₂H₂ gene family in *indica* rice could be involved during all the stages of reproductive development from panicle initiation till seed maturation. A total of 39 genes are up-regu-

lated more than 2-fold, in comparison to vegetative stages, during reproductive development of rice, out of which 18 are specific to panicle development and 12 genes are seed-specific. Twenty-six genes have been found to be up-regulated during three abiotic stresses and of these, 14 genes express specifically during the stress conditions analyzed while 12 are also up-regulated during reproductive development, suggesting that some components of the stress response pathways are also involved in reproduction.

Keywords ZOS · Zinc finger · Rice · Microarray · Panicle · Seed · Stress

Abbreviations

C ₂ H ₂	Cysteine2/Histidine2
EAR	ERF-associated amphiphillic repression
HMM	Hidden Markov Model
KOME	Knowledge-based Oryza Molecular biological Encyclopedia
MPSS	Massively parallel signature sequencing
NCBI	National Center for Biotechnology Information
NJ	Neighbour-joining
TIGR	The Institute for Genomic Research
ZF	Zinc-finger
ZFP	Zinc-finger protein
ZPT	C ₂ H ₂ zinc-finger protein TFIIIA type

Electronic Supplementary Material The online version of this article (doi:10.1007/s11103-007-9199-y) contains supplementary material, which is available to authorized users.

P. Agarwal · R. Arora · S. Ray · S. Kapoor ·
A. K. Tyagi (✉)
Interdisciplinary Centre for Plant Genomics and Department of
Plant Molecular Biology, University of Delhi South Campus,
Benito Juarez Road, New Delhi 110021, India
e-mail: akhilesht@genomeindia.org

A. K. Singh · V. P. Singh
Division of Genetics, Indian Agricultural Research Institute,
New Delhi 110012, India

H. Takatsuji
Plant Disease Resistance Research Unit, National Institute of
Agrobiological Sciences, 2-1-2 Kannondai, Tsukuba 305-8602
Ibaraki, Japan

Introduction

The Cys₂/His₂ zinc-finger proteins (ZFPs) constitute one of the largest transcription factor class regulatory protein families in eukaryotes (Englbrecht et al. 2004; Iuchi 2001; Takatsuji 1999). The first plant-specific ZFP, *EPF1* (later

renamed as *ZPT2-1*), was identified from *Petunia* that interacted with the promoter region of EPSP synthase gene (Takatsuji et al. 1992). Subsequently, by conventional approaches, approximately 40 C₂H₂ zinc-finger protein family members have been discovered in other plants and their contribution to important pathways during vegetative growth, reproductive development and stress responses elucidated (Davletova et al. 2005; Dinneny et al. 2006; Huang et al. 2005a; Huang et al. 2005b; Kapoor and Takatsuji 2006; Li et al. 2006; Takatsuji 1999). The petunia ZFPs were found to have a plant specific conserved domain, QALGGH, within the putative DNA-contacting surface in the zinc-finger (ZF) motif. The number of ZFs in these proteins varied from 1 to 4. The length of spacer between the ZFs in 2-fingered proteins was found to vary from 19 to 65 residues and in case of 3- and 4-fingered proteins was either short from 23 to 89 residues or long from 126 to 239 residues (Kubo et al. 1998). These ZFs were classified into A, B and M types (from hereon referred to as QA, QB and QM, respectively) based on the preferential occurrence of basic amino acid residues and conservation of residues in QALGGH domain (Kubo et al. 1998). Recently, using an *in silico* approach, 171 ZFP encoding genes were identified in Arabidopsis and their phylogenetic relationships described (Englbrecht et al. 2004). The classification of Arabidopsis ZFPs is based on tandem or dispersed nature of ZFs. The ZFPs containing tandem ZFs in one and more than one array are classified into sets A and B, respectively, while ZFPs with single or dispersed ZFs are grouped into set C. A subset of group C, C1 consists of 77 ZFPs that correspond to the petunia ZFPs with QALGGH domain.

The ZFPs not only interact with DNA or chromatin but their interactions with RNA and other proteins are also well documented in lower as well as higher eukaryotes (Gamsjaeger et al. 2007; Iuchi 2001; Yang et al. 2006). A number of plant ZFPs from petunia and Arabidopsis have also been functionally validated and are implicated in floral organogenesis, leaf initiation, lateral shoot initiation, gametogenesis and stress response (Davletova et al. 2005; Dinneny et al. 2006; Grigg et al. 2005; Kapoor et al. 2002; Kapoor and Takatsuji 2006; Li et al. 2006; Nakagawa et al. 2005). Besides, ZFPs are also involved in miRNA biogenesis (Yang et al. 2006) and regulation of flowering time (Kozaki et al. 2004). Some of these proteins function as repressors. The EAR (ERF-associated amphiphilic repression) motif, found in the C-terminal region of TFIIIA-type ZFPs from plants, has been shown to act as a repressor and has the conserved region L/F DLN L/F (X) P (Ohta et al. 2001). A hexapeptide motif 'DLELRL' in the C-terminal region of SUP also acts as a repression domain (Hiratsu et al. 2004). Presence of a similar conserved motif, DLNL, has been noted in C₂H₂ ZFPs in petunia

(Kobayashi et al. 1998; Kubo et al. 1998) and Arabidopsis (Meissner and Michael 1997).

An in-depth analysis of 189 C₂H₂ ZFP genes of rice is reported here. The members of this gene family are extremely diverse. They possess both tandem as well as dispersed ZFs. A comparison with Arabidopsis (Englbrecht et al. 2004) and yeast (Bohm et al. 1997) proteins shows that QALGGH motif containing Q-type ZFs are specific to plants, whereas, ZFPs with C-type ZFs group with yeast (Bohm et al. 1997). The zinc-finger proteins (TFIIIA class) have previously been named as ZPT (Kubo et al. 1998), hence we have named rice proteins as ZOS (ZPTs of *Oryza sativa*) and numbered the corresponding genes in the order of their physical presence on the chromosomes. By using microarray-based expression analysis of ZOS-encoding genes at 14 stages involving vegetative and reproductive development and 3 stress conditions in *indica* rice, we have identified genes that could specifically regulate panicle and seed development or be involved in stress responses. Since expression profiles have strong correlation with the gene function, this study would lay a foundation for functional validation exercises aimed at understanding the role of this class of transcription factors in plant growth and development.

Materials and methods

Data mining

A Hidden Markov Model (HMM) profile was formed from 37 most divergent C₂H₂ ZF domain sequences from Arabidopsis, yeast and humans using the HMMER software (<http://www.hmmer.wustl.edu/>). This profile was used to search TIGR Rice Pseudomolecules release 4 (The TIGR Rice Genome Annotation Database and Resource, Rockville, USA) and the KOME full length cDNA database (Kikuchi et al. 2003), followed by removal of same sequences from the two databases. All the protein sequences thus obtained were analyzed in PROSITE (Hulo et al. 2006) to validate the presence of C₂H₂ ZFs. All those sequences which were found to contain other domains or were altogether rejected by PROSITE were scanned in INTERPRO (Quevillon et al. 2005). Further, a name search for the same was performed in both the databases. The resulting new sequences were also analyzed in PROSITE and INTERPRO. A manual inspection of the rejected sequences was also performed. The fingers which conformed with the proposed pattern of TFIIIA type fingers ϕ -X-C-X_{2,4,5}-C-X₃- ϕ -X₅- ϕ -X₂-H-X_{3,4}-H where X is any amino acid and ϕ is a hydrophobic residue (Klug and Schwabe 1995) were included during manual inspection. All the confirmed C₂H₂ ZFPs were also examined in the

Rice Genome Browser (<http://www.tigr.org/tigr-scripts/osa1web/gbrowse/rice>).

Chromosomal localization and gene duplication

A total of 179 ZOS genes were localized on 12 chromosomes of rice and named according to their positions on pseudomolecules. The first one or two digits following the letters ZOS represent the number of the chromosome followed by a hyphen and the gene number. For example, the first gene on chromosome 1 is *ZOS1-01* and that on chromosome 12 is *ZOS12-01*. The genes were depicted on the chromosomes depending on the direction of transcription. The duplicated genes were elucidated from the segmental genome duplication of rice (http://www.tigr.org/tdb/e2k1/osa1/segmental_dup/100). The DAGchainer program (Haas et al. 2004) had been used to determine segmental duplications with parameters $V = 5$ $B = 5$ $E = 1e-10$ -filter seg and distance = 100 kb. Genes separated by 5 or less genes were considered to be tandem duplicates. The distance between these genes on the chromosomes was calculated and the percentage of sequence similarity between the proteins coded by these genes was determined by MegAlign software 4.03© (DNASTAR Inc.).

Phylogenetic analysis of ZFPs

Full-length ZFPs from rice, Arabidopsis and yeast were aligned using Clustal X version 1.83 (Thompson et al. 1994; Thompson et al. 1997) and two separate unrooted NJ plotted phylogenetic trees comparing rice ZOS proteins with Arabidopsis (Englbrecht et al. 2004) and yeast (Bohm et al. 1997) ZFPs, respectively, were constructed from these in TreeView 1.6.5. An NJ plot tree was also made exclusively with all the zinc-finger domains from the first cysteine to the last histidine.

Gene and protein structure and their organization

All the available details for these genes in the public databases of TIGR and KOME were obtained. For genes with a corresponding cDNA in the KOME database, the intron analysis was performed by aligning the cDNA sequence with the corresponding genomic/BAC sequence in SIM4 (Florea et al. 1998). Alternative splicing for these genes was determined by comparing the number of introns, their lengths and their positions on the cDNAs. Using a similar methodology, differences in the sequences between three different databases, viz. KOME, TIGR and RefSeq database of NCBI (Pruitt et al. 2005) were obtained. The number of C₂H₂ zinc-fingers in the proteins coded for by these genes was determined by scanning in PROSITE and INTERPRO and also by manual inspection. Once the

zinc-fingers were finalized, the type of C₂H₂ motif was determined by manual examination.

Genome-wide expression analysis of C₂H₂ ZFPs in indica rice

The RNA was isolated from panicles by TRIzol® reagent (Invitrogen Life Technologies, Carlsbad, CA) according to the manufacturer's protocol, whereas, for starchy rice seeds a modified protocol involving guanidine hydrochloride was followed (Singh et al. 2003). The RNA was further purified by RNeasy MinElute Cleanup Kit (Qiagen). The RNA samples with 260/280 ratios between 1.95 to 2.1 and even higher 260/230 ratios were used for hybridization with microarrays. The microarray analysis was performed using one-cycle target labeling and control reagents (Affymetrix, Santa Clara, USA) using five micrograms of total RNA as starting material for each sample as previously described (Jain et al. 2007). A total of 51 rice Affymetrix GeneChip® Rice Genome Array [Gene Expression Omnibus (GEO) platform accession number GPL2025] were hybridized to RNA isolated from 3 biological replicates each from three vegetative stages (mature leaf, root and seedling), 6 stages of panicle (P1–P6) and 5 stages of seed (S1–S5) development along with 3 abiotic stress conditions (cold, salt and desiccation) as described (Jain et al. 2007). The cDNA made from this RNA was in vitro transcribed into cRNA, which was fragmented and hybridized to microarrays. The washing and staining were performed according to the manufacturer's protocol using Fluidics Station 450. For data analysis, the raw data (*.cel) files were imported into Avadis prophetic (version 4.2) software (Strand Life Sciences, India). The normalization and probe summarization was performed by using GCRMA (GeneChip Robust Multi-array Analysis) algorithm (Wu et al. 2003). The three biological replicates of a sample with an overall correlation coefficient value of more than 0.95 were selected for final analysis. Any data set that did not meet this cut-off value was discarded. To identify differentially expressed genes, Student's t test was performed. The genes that are up- or down-regulated equal to or more than two-fold were considered to be differentially expressed significantly. The average of three biological replicates for each sample was used for analysis. We defined a gene as specifically expressed in a given organ only if the expression level of the gene in the organ was significantly higher (more than 2-fold) than the levels in all the other organs. The up- or down-regulated genes in any tissue were calculated from the average of log of normalized signal values. The expression of a particular gene was considered absent if the normalized signal value from corresponding probe set was < 7. The microarray data has been deposited at Gene Expression Omnibus (GEO) at

NCBI under the series accession numbers GSE6893 and GSE6901.

To validate the data obtained from microarray, real-time PCR was performed for biological duplicates. DNase (Qiagen) treated RNA was reverse-transcribed to form cDNA. SYBRGreen PCR Master Mix (Applied Biosystems, USA) was used to determine the expression levels for the genes in ABI Prism 7000 Sequence detection System (Applied Biosystems, USA).

Expression patterns of selective *ZOS* genes by MPSS

Massively parallel signature sequencing (MPSS) data (<http://www.mpss.udel.edu/rice/>) was consulted to determine the expression profiles of genes that either were not represented on the Affymetrix GeneChip® Rice or did not express in the developmental stages tissues used in the investigation. Data from the 17 bp signature and from the top strand was considered for analysis. A heat map was generated in Avadis prophetic 4.2 from the normalized value for the signatures in transcripts per million (TPM) for the selected genes.

Results

Identification of C₂H₂ zinc-finger protein genes in rice

The HMMER search resulted in a total of 395 sequences from TIGR pseudomolecule version 4, out of which 343 were unique. Also, 111 sequences were obtained from KOME by the same method. Of the KOME cDNA sequences, 10 did not find any significant match in the TIGR pseudomolecule database. They showed sequence similarity in the range of 54–61% with the available genomic DNA sequence. INTERPRO analysis of these sequences revealed the presence of ZF domain in all of them. After analyzing all the sequences from TIGR for their protein domains in PROSITE, it was concluded that 166 sequences had at least one ZF domain. Some of the other proteins selected by HMMER search were found to contain jmj, SET, hAT, transposase, GRAS, retrotransposon, helicase, Dof, LIM, BED, DnaJ, RING finger, RhoGAP, BED, C3HC, PHD zinc fingers, LIM, beige/BEACH, SNF2, F-box and B-box domains as determined by PROSITE. From the proteins designated with some other domain by PROSITE, only 1 was designated as a ZF on INTERPRO analysis. PROSITE did not show any domain for 59 proteins, out of which 3 were found to contain C₂H₂ ZFs after the INTERPRO analysis. This made a total of ZFPs to be 180 of which 170 had a corresponding gene model in TIGR database.

The name search for “C₂H₂ zinc-finger proteins” was also carried out since not all the known C₂H₂ zinc finger

motifs could be included during HMM profile generation and one may have missed out on some authentic candidate genes. This resulted in 92 and 102 gene models in TIGR and KOME databases, respectively. All but 9 gene models revealed by name search in KOME database could be accounted for in the list of 180 ZFP encoding genes mentioned above. PROSITE and INTERPRO scans of these 9 gene models confirmed the presence of ZF domains in 4 of them making the total of rice ZFPs to be 184. Additionally, 5 more ZOS encoding genes were identified for following reasons: (1) *ZOS1-09*'s 3 ZFs and (2) one ZF of *ZOS1-11* were apparent only after manual inspection (3) ORF2 of *ZOS2-05* had a ZF domain, (4) *ZOS6-10* had a ZF domain which was designated by INTERPRO as ‘family not named’ and (5) the homolog of *ZOS7-03* in the NCBI database showed a QALGGH containing ZF, though the cysteine residue at the 2nd position had been replaced by tyrosine (Y), both of which are polar amino acids. Earlier also an ambiguity over non-detection of certain yeast ZFPs led to the extension of PROSITE pattern (Bohm et al. 1997). Since some of the rice ZFs also could not be recognized by PROSITE, the PROSITE pattern may require further amendments to incorporate monocot ZFs as well. Collectively, this made the total of rice ZFPs to be 189 of which 179 had a corresponding locus ID in TIGR database (Supplementary Table 1). PROSITE scan of all the 189 proteins revealed the presence of additional domains in 25 proteins such as ankyrin-repeat region, DNAJ, EF hand, GRAS, tubulin, prokaryotic lipoprotein, post-set, integrase, RNaseH, basic leucine zipper, ferredoxin, aspartic acid and asparagine hydroxylation site, dihydrodipicolinate synthetase, ubiquitin associated domain, methyl-CpG binding domain and other types of zinc fingers viz., ring finger and BED-type finger (Supplementary Table 1).

It may further be noted that two public databases, viz. TIGR pseudomolecule version 4 and KOME full-length cDNA database, were searched for finding out the ZPT encoding rice genes and their counterparts were also searched for in the NCBI RefSeq database (Pruitt et al. 2005). Ten genes showed certain differences in annotation between any 2 databases essentially due to variation in predicting intron-exon boundaries (Supplementary Fig. 1).

Chromosomal localization and gene duplication

Out of 189 *ZOS* genes, 179 could be localized on the 12 chromosomes of rice by virtue of their presence on the TIGR pseudomolecules (Fig. 1). The rest are present as cDNAs in the KOME database only and are supposedly expressed from the physical gap regions of the rice genome sequence. Examination of all 179 genes in the Rice Genome Browser showed that six genes, viz. *ZOS1-10*, *1-12*, *1-22*, *6-07*, *7-09* and *12-02*, overlap with another cDNA

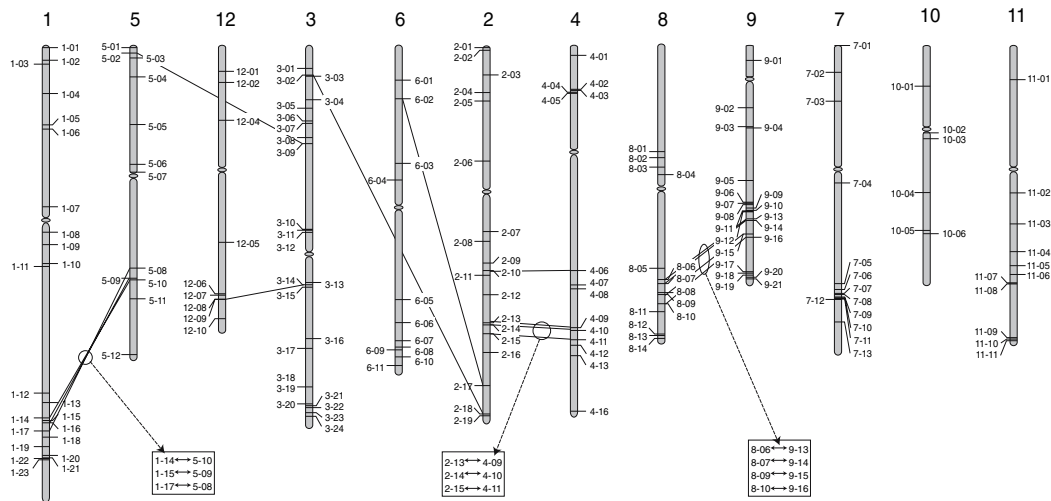


Fig. 1 Chromosomal localization of 179 C_2H_2 zinc-finger protein encoding genes (*ZOS*) on 12 chromosomes of rice. Only gene identity numbers are provided. The genes with open reading frames in opposite orientation have been marked on either side of the

chromosome. The chromosome order has been arranged to bring duplicated regions in the vicinity. The segmentally duplicated genes have been connected by line connectors. In case of duplicated genes present in clusters, the genes have been shown in inset

transcribed in antisense direction. cDNA is also available for all the *ZOS* genes, except for *ZOS1-10* (Supplementary Table 2). For this gene, the overlap with the antisense gene is in the 5' region. On the other hand, the structure of *ZOS6-07* is similar to murine *Fxyd3* and *Lgi4* genes (Runkel et al. 2003) with overlapping 3' regions. An ambiguity was also observed for *ZOS8-11* with a 1,944 bp long ORF. For this gene a full length cDNA (from KOMÉ) initiates at position 1,584 bp and extends 1,888 bp beyond the termination codon of the ORF (Fig. 2).

ZOS genes are present in all regions of a chromosome, i.e., at the telomeric ends, near centromere, scattered all over or in clusters (Fig. 1). Their distribution amongst the 12 rice chromosomes is highly variable. A maximum number of 24 genes are present on chromosome 3 closely followed by 23 on chromosome 1. On the other hand, only 7 genes are present on chromosome 10. Almost all the chromosomes have groups of *ZOS* genes in vicinity of each other. Interestingly, these clusters contain *ZOS* genes being transcribed in the reverse orientation as well. The lower region of the long arm of chromosome 9 has 2 clusters of *ZOS* encoding genes with the genes being transcribed in both the directions. *ZOS5-07* and *10-02* are localized very near the centromere. Recently, the rice centromeres have been found to contain actively transcribed genes (Bohm et al. 1997; Cooke 2004; Nagaki et al. 2004). Another gene located close to the centromere is *ZOS8-03*, which has RNase H and retrotransposon gag domains apart from a C_2H_2 domain. Additionally, *ZOS1-01*, *2-01*, *5-01* and *7-01* are located in the telomeric regions.

Fifteen pairs of *ZOS* genes could be assigned to TIGR segmental duplication blocks (Supplementary Table 3, Fig. 1). The overall sequence similarity of the cDNA sequences of these genes ranged from 27.9% to 70.9%. All these genes exhibited high sequence similarity in the ZF and the flanking regions as well. Between chromosomes 1 and 5, the whole duplicated cluster is reversed. In the duplicated cluster between chromosome 8 and 9, only *ZOS9-15* has reversed. Additionally, 16 groups of genes were found to be tandemly duplicated (Supplementary Table 4, Fig. 1). They were separated by a maximum of 5 intervening genes. Eleven of the gene pairs were placed juxtaposed with no intervening gene. The distance between these genes ranged from 2.4 kb in group IV to 36.2 kb in group XIII (Supplementary Table 4). There were 2 genes in tandem in most cases. However, groups VI, XII and XV had more than two genes.

Structural analysis of *ZOS* genes

The presence of an existing cDNA with a known corresponding position on the genome is required for the determination of the number and length of introns. This information was available for 102 genes (Supplementary Table 1). On analysis by SIM4, the number of introns varied from 0 to as many as 20. Thirty-two genes were intronless while 19 genes had a single intron. Out of 15 genes on chromosome 3, 12 did not contain any intron. The importance of intron in the 5' UTR in gene regulation has been well established (Jeong et al. 2006; Le Hir et al.

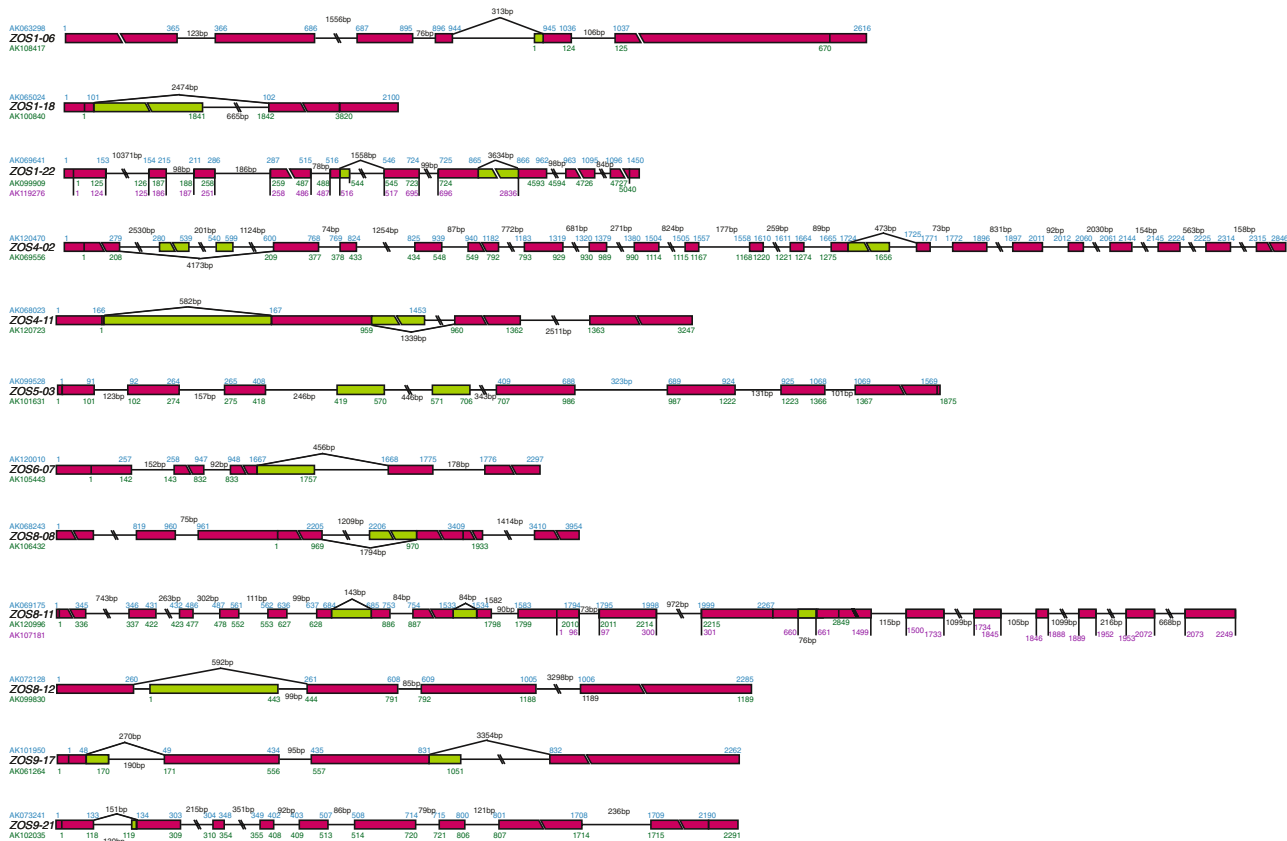


Fig. 2 Exon-intron structure of alternatively spliced ZOS proteins. The accession numbers of the genes and their names are mentioned on the left. The coloured bars represent the exons while the lines

represent the introns. The position of the exon and the length of the introns are mentioned. The green coloured bar represents the region showing alternative splicing

2003; Rose 2002). Twenty-one genes of the ZOS gene family contained at least 1 intron in the 5' UTR. Out of these, the intron was located just upstream of the start codon in ZOS4-03, 5-03, 8-01, 9-03, 12-06 and 12-07. Certain genes, viz. ZOS2-02, 4-02, 9-20 and 12-07, had more than 1 intron in the 5' UTR. Apart from this, 8 genes had 1 or 2 introns in the 3' UTR. Only 2 genes, viz. ZOS10-03 and 12-03, have an intron in both the UTRs. The intron length was also highly variable, ranging from 51 bp for ZOS10-01 to 10,371 bp for the 1st intron of ZOS1-22, which is in the 5' UTR. The intron was more than 1 kb in length for 35 genes and those long introns were the 1st intron for 21 genes. For 16 genes, the intron in the UTR region was more than 1 kb. The ORF length is also extremely variable. The exon length also ranges from a few bases (14 bp for ZOS3-04) to 2561 bases (Supplementary Table 1).

In TIGR database, 20 loci were interpreted to show more than one gene model attributed by alternative splicing out of 179 loci available for ZOS gene family. These are ZOS1-04, 1-16, 1-18, 1-22, 2-02, 2-15, 3-07, 3-17, 4-02, 4-11, 6-06, 6-07, 8-08, 8-09, 8-11, 8-12, 9-17,

9-19, 9-21 and 12-03. However, alternative splicing of only 10 of these loci was supported by cDNA evidence from KOME database (Fig. 2). ZOS1-06 and 5-03 had only single gene models in TIGR, although two alternatively spliced cDNAs existed in the KOME database. Overall, this led to a total of 22 alternatively spliced ZOS genes, with the maximum number of 6 transcripts being reported for ZOS4-11 by TIGR and with the actual evidence of a maximum of 3 transcripts through cDNA for ZOS1-06, AK108417 does not code for a ZF. In the case of ZOS1-22, out of 3 cDNAs, 2 (AK099909 and AK119276) do not code for a ZF due to alternative splicing. ZOS4-11 has 2 alternatively spliced cDNAs. AK068023 codes for C₂H₂ ZFP while AK120723 does not. In the case of ZOS5-03, AK099528 and AK101631 code for 3 fingers and 7 fingers, respectively. ZOS8-08 has 2 KOME clones showing alternative splicing. AK106432 with 2 exons does not have a ZF domain while AK068243 with 5 exons has this domain. Of the 3 cDNAs representing ZOS8-11, AK107181 also does not contain any zinc finger motif.

Types of zinc fingers in rice and their evolutionary relationship

A total of 307 individual ZFs were found in 189 ZOS proteins. Two extremely different kinds of motifs, bearing no sequence similarity whatsoever except for the presence

of two cysteine and two histidine residues were observed (Fig. 3). The first type contains a conserved sequence ‘QALGGH’ in the DNA-recognition motif and was designated as Q-type. This sequence has been known to be specific to plants (Takatsuji et al. 1994) and 99 such ZFs were found in rice. The other distinct kind of ZF was

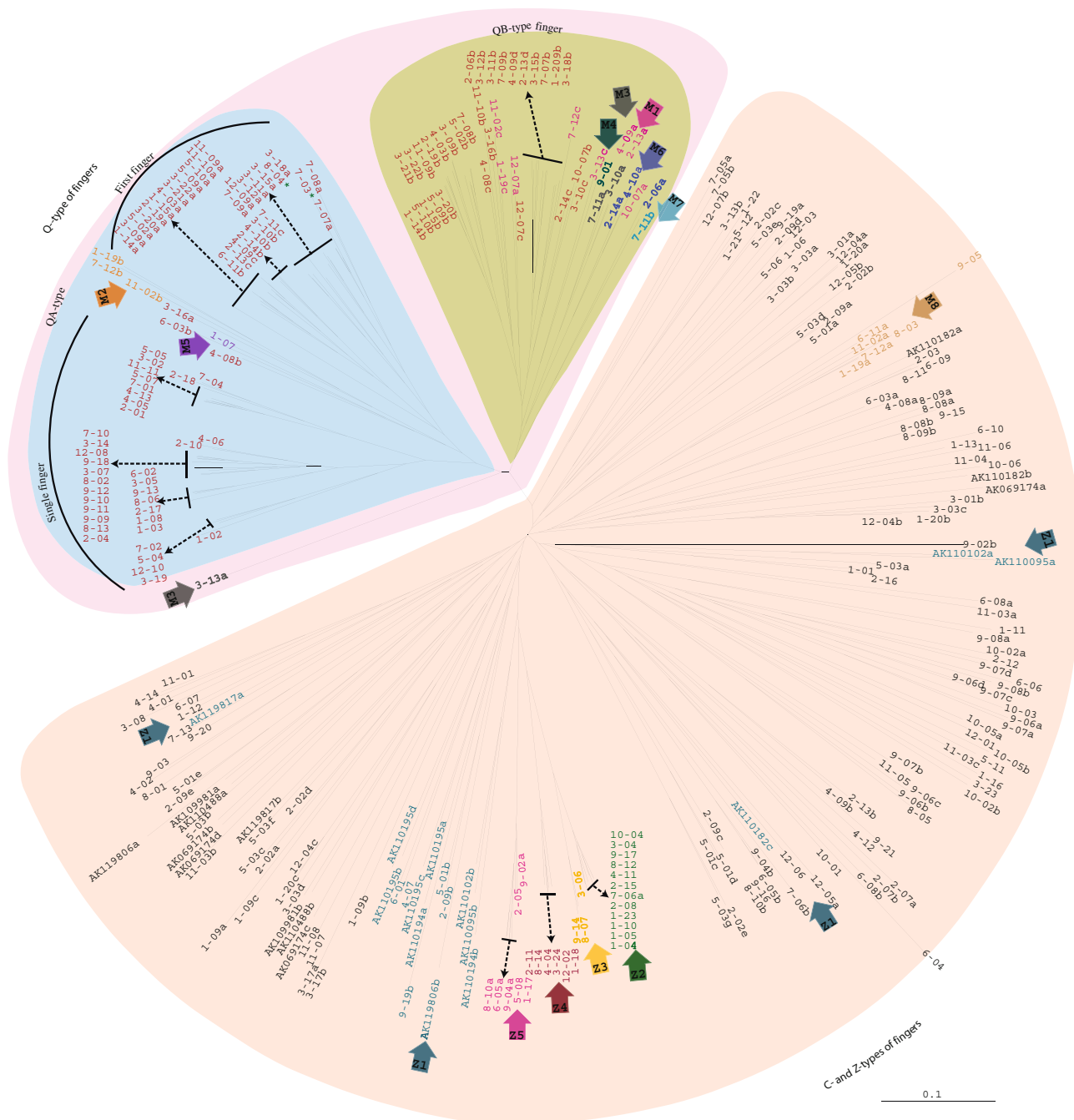


Fig. 3 Evolutionary relationship amongst different types of C₂H₂ ZF motifs of rice. An unrooted NJ tree is shown for 307 ZF motifs from the first cysteine to the second histidine. The number represents the ZOS protein from which the ZF sequence has been taken. The alphabets a, b, c etc. indicate the order of the ZF on the protein as first,

second, third and so on. The Q-type fingers have been coloured red. The ZFs not conforming to the general pattern, as indicated in the text, are marked with an asterisk. The M-type fingers are represented in various colours as indicated. The scale bar depicting 0.1 amino acid substitutions per site is shown

designated as C-type as these did not contain any conserved motif in the ZF region. Certain modifications observed in the Q-type ZFs were classified as M-type and were numbered from QM1-QM8, indicating the type of modification as shown in Supplementary Table 5. The M-type fingers have also been found in petunia (Kubo et al. 1998) and Arabidopsis (Englbrecht et al. 2004; Meissner and Michael 1997). Amongst C-type, some of the ZFs had highly conserved motifs in finger and the flanking regions and were designated as Z-type. Five variants of Z-type were named Z1 to Z5. In the NJ-plotted tree, QM1 to QM7 ZFs grouped with Q-type ZFs while QM8 clustered with the C-type ZFs. When an NJ plotted phylogenetic tree was constructed with full ZF sequences (data not shown), M8-type of ZFs that have AXXGH as the conserved motif, branched off as a sub-clade from the other Q-type ZFs. Most Z1-type ZFs formed a sub-clade amongst the C-type ZFs though some were scattered (Fig. 3). The members of this group showed the conservation of residues L, R and H, though the number of residues between these three was variable, representing an evolving clade. Each of the ZFs from Z2 to Z5 has a highly conserved motif extending through a large part of the ZF and with the exception of two proteins, all proteins which had Z2 to Z4 ZFs were single-fingered proteins.

All the solitary Q-type ZFs were highly similar to each other and shared a common origin with the exception of ZFs in *ZOS7-03* and *8-04* (marked with green asterisk, Fig. 3). Wherever two consecutive Q-type fingers were present, the QA-type and QB-type ZFs, as defined for petunia (Kubo et al. 1998), were apparent by the different clades formed from these fingers. The first finger clade shares a common ancestral node with the single Q-type fingers. Not more than two consecutive ZFs were found to be Q-type in rice. Only 5 three-fingered proteins were found which had both Q- and M-type of ZFs.

Numbers of ZFs per protein and the length of spacer region between them

ZOS proteins were found to possess 1 to 7 ZF domains per protein. A maximum of 113 proteins had a single ZF which included all four types of ZFs, i.e., Q-, M-, Z- and C-types. Of these, a maximum of 46 proteins had a C-type ZF, followed by 38 proteins with the Q type ZF. A significantly lesser number of 51 proteins had 2 ZFs out of which both the ZFs were either of Q-type or C-type or Z-type for 22 or 13 or 3 proteins, respectively. For the others, it was either Q and M ZF combination or C and Z combination. Fourteen proteins had 3 ZFs. The last 2 ZFs were Q-type for 3 proteins. All the ZFs were C-type for 4 proteins and M-type for 3 proteins. There were only seven 4-fingered proteins

out of which 2 proteins had the last 2 consecutive ZFs of Q-type with the 1st ZF being M1-type. Four proteins had all C-type ZFs and for 1 protein all ZFs were Z1-type. Only 3 proteins were found with 5 ZFs, all but one ZF in these were C-type. The single 7-fingered protein also had all C-type fingers.

The residues between the second histidine of the previous finger and the first cysteine of the next finger form the spacer or the linker between the two ZF domains (Takatsuji 1998). Proteins having more than two ZFs can have both cluster-type and separate-type arrangement of fingers (Takatsuji and Matsumoto 1996). The ZFPs of rice showed the existence of both tandemly placed or clustered and widely separated ZF domains. The spacer was as small as 0 amino acid (*ZOS1-09*) and reached a maximum of 325 residues (*ZOSa*). Though the spacer length was highly variable, the spacer length of 7 residues, which is similar to the H/C link in the Krüppel protein (Schuh et al. 1986) was observed for 25 spacers. Out of these, 11 spacers showed more than 50% sequence similarity to the H/C link of Krüppel, while the rest (i.e., 14) did not show any significant sequence similarity. The rest of the spacer lengths occurred in 1–3 proteins each. The spacer length between any two Q-type ZFs in rice, though arranged separately, was generally less than 100 residues, except for *ZOS2-13*. None of the Q-type ZFs was found to be clustered.

Evolutionary relatedness of rice, yeast and Arabidopsis ZFPs

An unrooted NJ plotted phylogenetic tree of all rice ZFPs along with 53 yeast ZFPs (Bohm et al. 1997) resulted in the formation of three distinct rice-specific clusters named group I, II and III, comprising 27, 35 and 37 proteins, respectively (Fig. 4). Notwithstanding few exceptions, group I contained single-fingered proteins, which had Z2- to Z5-type of ZFs; Group II proteins had two consecutive Q-type ZFs while the single Q-type fingered proteins clustered to form Group III. The other protein sequences, which mainly had C- and Z1 type of ZFs, were seemingly extremely diverse as they originated at or near the centre of the tree and formed Group IV. Groups I, II and III were specific to rice, while C- and Z1-type ZF-containing ZFPs grouped with the yeast ZFPs. To test if the ZOS proteins in groups I, II and III were plant-specific, another tree was constructed with ZOS proteins and 176 Arabidopsis ZFPs (Englbrecht et al. 2004); (Supplementary Fig. 2). Notably, most of the ZOS proteins specific to the three groups maintained their positions, only to be clustered with certain Arabidopsis proteins. However, there were few proteins that shuffled between groups I, II and III suggesting their common origin.

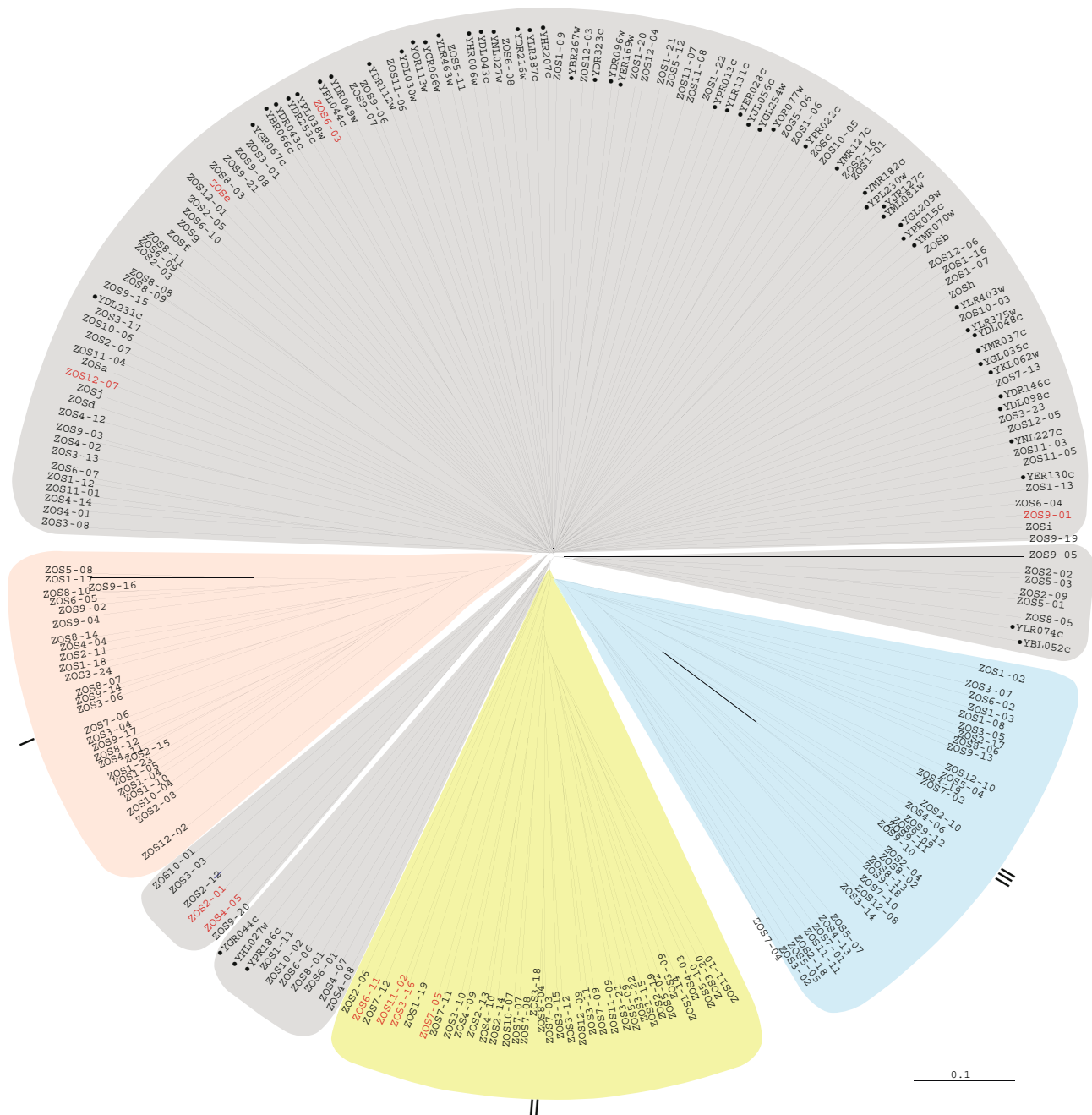


Fig. 4 Phylogenetic tree of rice and yeast C₂H₂ zinc finger proteins. An unrooted NJ tree made from 189 rice and 53 yeast ZF proteins is shown. The rice ZOS proteins can be distinguished from the yeast proteins which start with “Y” and have a dot as prefix. The three

clusters which are plant-specific are indicated by Roman numbers. The proteins which do not cluster in the same group on alignment with Arabidopsis proteins have been coloured in red

Expression of ZOS genes during panicle/seed development and under three abiotic stress conditions in indica rice

The clustering of expression profiles of 176 ZOS encoding genes were collated from a microarray-based transcriptome profiling of 14 stages of vegetative and reproductive development (Fig. 5) and three stress conditions (Fig. 6).

The genes exhibiting high and low expression were segregated based on MAS 5.0 software. Only 104 genes showed a “present” detection call at *P*-value of 0.05 in at least one of the stages analyzed, whereas 72 low expressing genes were either “absent” or “marginal” under these conditions. Two separate heat maps (Fig. 5A and B) were generated to emphasize the differential expression profiles of these two groups of genes.

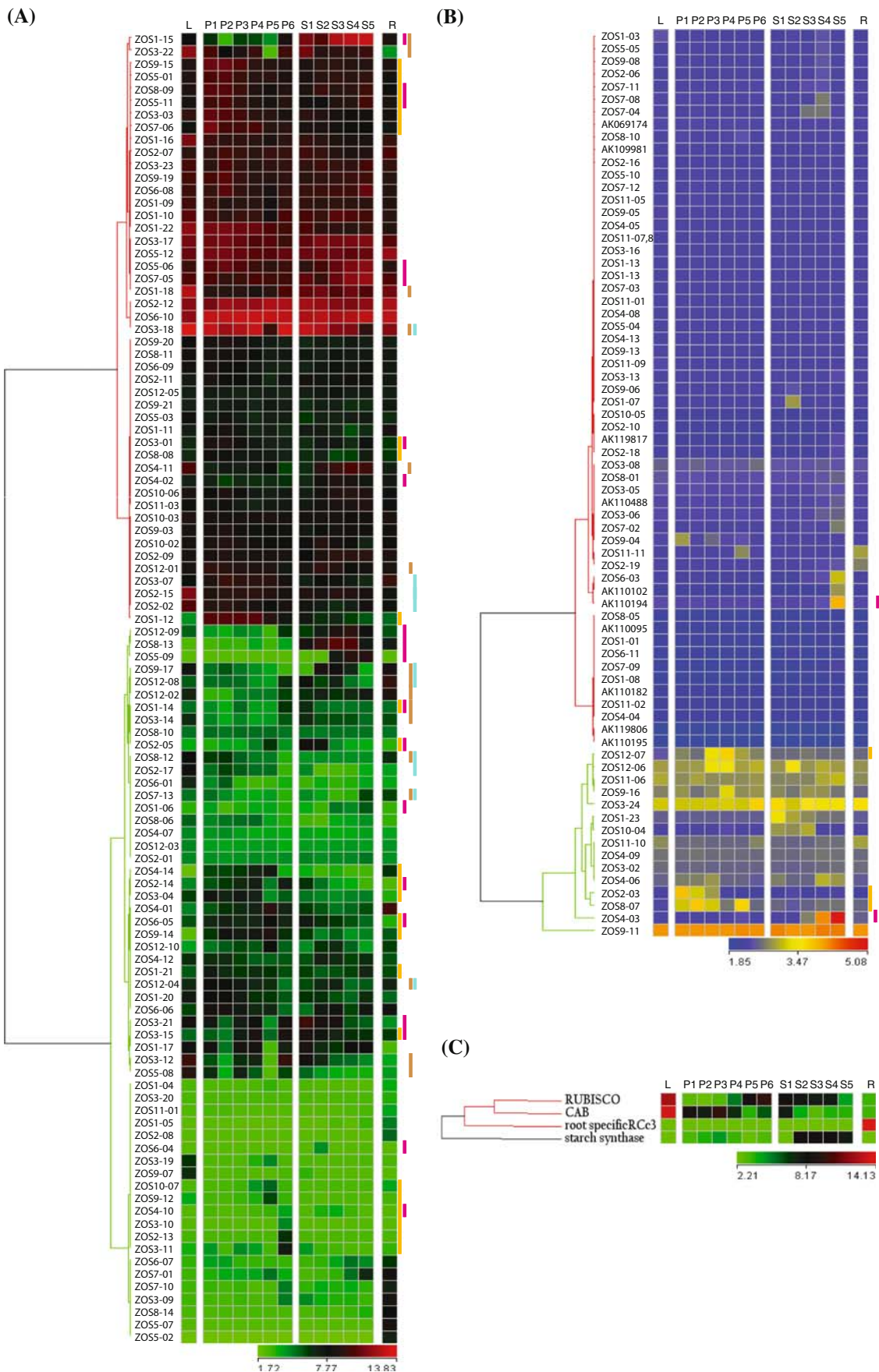


Fig. 5 Expression pattern of *ZOS* genes during panicle and seed development. (A) All the genes showing relatively high signal values (104 genes) in at least one of the samples were taken and hierarchical clustering was performed on the average of the log of the three replicates for each sample. Squared Euclidean distance metric and Ward's linkage rule were used for the analysis. The controls used were mature leaf (L) and root from a 7-day-old seedling (R). The stages of panicle development are P1 (0–3 cm), P2 (3–5 cm), P3 (5–10 cm), P4 (10–15 cm), P5 (15–22 cm) and P6 (22–30 cm). The stages of seed development are S1 (0–2 DAP), S2 (3–4 DAP), S3 (4–10 DAP), S4 (11–20 DAP) and S5 (21–29 DAP), (B) hierarchical clustering of the genes showing low levels of expression in all the tissues, (C) the expression pattern of some of the well known genes was analyzed as a validation exercise for microarray experiment. The color bars showing the range of \log_2 signal values are shown at the bottom of each heat map. The vertical lines on the right side of the map show the genes which are expressed more than 2-fold in any one stage of panicle development (yellow line) or seed development (pink line) w.r.t. leaf and root as controls. The genes which are more than 2-fold down-regulated in panicle (brown line) and seed (blue line) are also shown

A total of 27 genes were found to be up-regulated by more than 2-fold during panicle development, in comparison to the vegetative stages. Out of these, 3 genes showed low level expression. Similarly, 21 genes were up-regulated during seed development and 9 out of these were up-regulated during both panicle and seed development (Fig. 5A). During panicle development, only 1 gene (*ZOS7-06*) was found to be up-regulated more than 2-fold in all the 6-stages, with the maximum expression at P1 and a subsequent decrease in the following stages. This gene was also up-regulated during dehydration stress. *ZOS1-12* expressed at high levels during the first 5 stages of panicle development. Both *ZOS1-21* and *9-14* expressed at moderate levels in consecutive expression windows, as *ZOS1-21* was up-regulated during P1 and *ZOS9-14* expressed from P2 to P5. A similar expression pattern was observed for *ZOS9-15* and *3-01* from P1 to P3, although the expression level of *ZOS9-15* was much higher. *ZOS5-01* and *8-08* also shared a similar expression pattern and were up-regulated in P1 and P2 with *ZOS5-01* having higher expression levels. *ZOS3-04* was up-regulated in P1 and P3 and also during dehydration and cold stress, which would imply the activation of similar genes in these cases. *ZOS8-07* also showed a biphasic expression pattern as its expression was up-regulated in P2 stage corresponding to 3–5 cm panicle and then in the P5 stage consisting of 15–20 cm panicles. Those genes, which were found to express at a particular development stage, are *ZOS2-03* and *3-03* at P1; *ZOS12-07* at P4; *ZOS4-14*, *9-12* and *10-07* at P5 and *ZOS2-13*, *3-10* and *3-11* at P6. Out of these, only *ZOS3-03* showed a high signal value, and rest had low or moderate expression. *ZOS3-11* was also expressed during dehydration and salt stresses.

During seed development, *ZOS1-15* and *12-09* transcripts accumulated at high levels in all 5 and 4 stages,

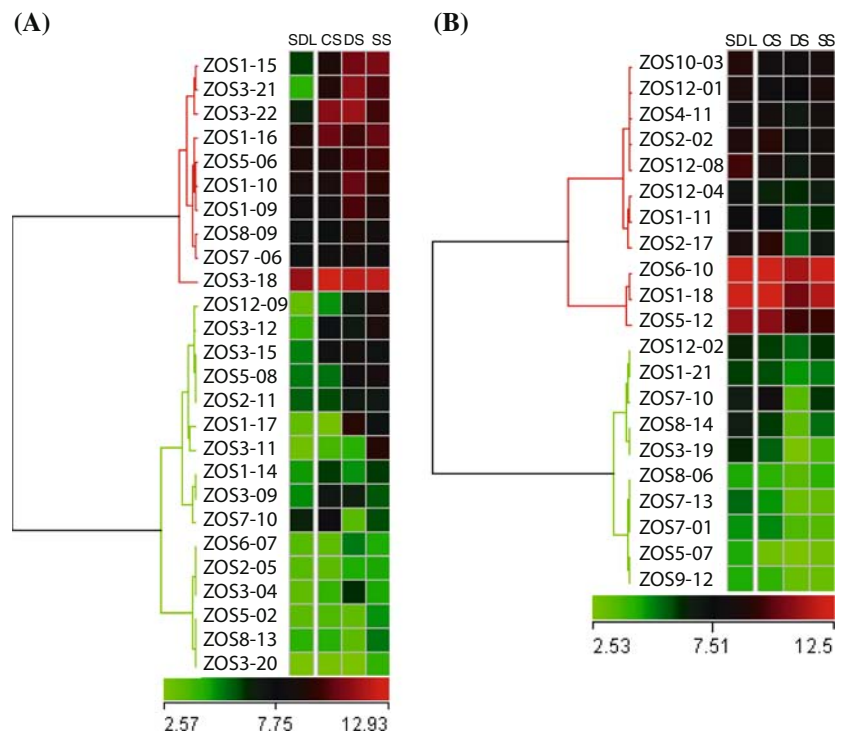
respectively, and also under all stress conditions. *ZOS8-13* was up-regulated in 4 of the 5 seed development stages. Starting from S1, its expression reached maximum level at S3, after which it declined. This gene was also expressed during salt stress. *ZOS1-06* and *5-09* were up-regulated from S3 to S5. *ZOS4-03* was up-regulated in S4 and S5 though at low levels. Surprisingly, the expression of *ZOS5-09* was over 100-fold in S4 and S5 as compared to all the vegetative stages. The expression of *ZOS5-06* increased in consecutive stages to be finally up-regulated in S4 and S5. This gene had a high signal value in all tissues tested and was also up-regulated during dehydration stress. *ZOS7-05* shared a similar expression level and pattern, though it was not up-regulated during any stress. Among the genes which were up-regulated at only one stage of seed development, viz., *ZOS3-21* at S1, *ZOS6-04* at S2, and *ZOS4-02* at S3 and *ZOS3-01* and *ZOSf* at S5, only *ZOS3-21* was up-regulated during all the 3 stresses. These genes also had low to moderate expression values just as for the genes up-regulated at one stage of panicle development.

Certain genes were up-regulated during panicle as well as seed development, probably because of the prevalence of similar processes in those stages. *ZOS8-09*, which had a moderate expression level, was up-regulated in P1 to P3 and also in S2 and S5. Apart from these, this gene was also up-regulated during dehydration stress. While *ZOS5-11* was up-regulated in P1, P2 and S5, *ZOS6-05* had a unique pattern and was up-regulated in P1, P2, P5, P6 and S1, i.e., during early and late stages of panicle development and early seed development. On the other hand was *ZOS3-15*, which was up-regulated in P3, P4, P6, S1 and S2. This gene was also up-regulated during all the 3 stresses. *ZOS2-14* was up-regulated from P1 to P4 and then again at P6 and S1. *ZOS1-14* was yet another gene which was up-regulated at P6 and S1. It was also up-regulated during cold and salt stress. *ZOS2-05* was up-regulated at P4, P5, S1 and S2 and during dehydration and salt stress. *ZOS4-10* was up-regulated at P6 and S3.

Sixteen and ten genes were found to be more than 2-fold down-regulated in at least 1 stage during panicle and seed development, respectively (Fig. 5). Out of these, *ZOS3-18*, *ZOS7-13*, *ZOS8-12*, *ZOS9-17*, *ZOS12-04* and *ZOS12-08* were found to be down-regulated during both panicle and seed development. Only *ZOS1-15* was up-regulated during seed development but down-regulated during panicle development.

Microarray analysis performed on 7-day-old seedlings subjected to three abiotic stresses, revealed 26 2-fold up-regulated genes in at least one of the stress conditions. Six genes were up-regulated under all the 3 stress conditions (Fig. 6A). The expression of *ZOS1-17*, *2-05*, *2-11*, *3-04*, *3-11*, *5-08* and *6-07* was up-regulated in response to dehydration as well as salt stress conditions, while

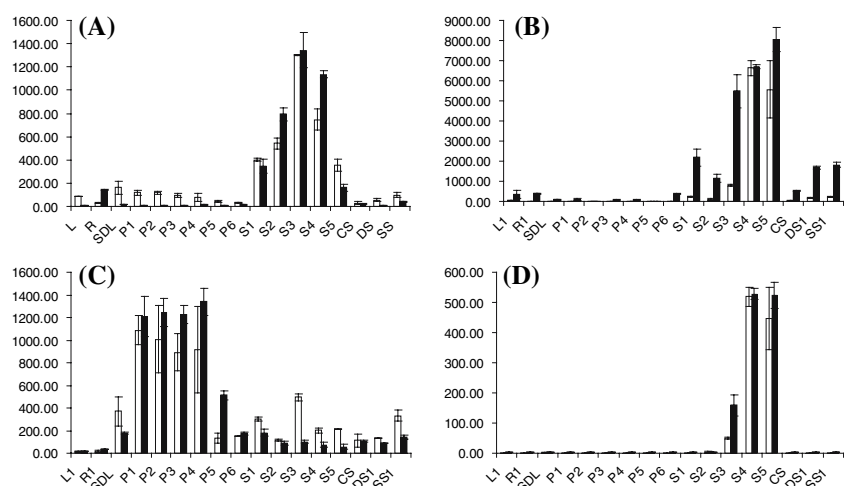
Fig. 6 The *ZOS* genes involved in stress responses. The genes up-regulated (**A**) and down-regulated (**B**) by more than 2-fold in any of the three abiotic stresses given to a 7-day-old seedling, viz. cold stress (CS), dehydration stress (DS) and salt stress (SS) as compared with an untreated seedling (SDL) have been shown by hierarchical clustering. The color bar representing \log_2 signal values is shown



ZOS1-14 and *1-16* were up-regulated under cold- and salt-stressed seedlings. Only *ZOS3-09* was up-regulated during both cold- and dehydration-stress conditions. *ZOS1-09*, *1-10*, *5-06*, *7-06* and *8-09* responded specifically to dehydration stress, similarly, *ZOS3-18* and *7-10* responded to cold stress while *ZOS 3-20*, *5-02* and *8-13* were distinctively up-regulated during salt stress. Apart from this, 21 genes were down-regulated in at least one of the abiotic stresses imparted to 7-day-old seedlings. Only 2 genes, viz. *ZOS12-08* and *ZOS5-07*, were down-regulated during all three kinds of stresses (Fig. 6B). A total of 8 genes, viz. *ZOS1-11*, *ZOS2-17*, *ZOS3-19*, *ZOS5-12*, *ZOS7-01*, *ZOS7-13*, *ZOS8-14*, and *ZOS9-12*, were simultaneously down-regulated during dehydration

and salt stress. Among the genes down-regulated in any one stress, a total of 11 genes were down-regulated during dehydration stress, viz. *ZOS1-18*, *ZOS1-21*, *ZOS2-02*, *ZOS4-11*, *ZOS6-10*, *ZOS7-10*, *ZOS8-06*, *ZOS10-03*, *ZOS12-01*, *ZOS12-02*, and *ZOS12-04*. *ZOS7-10* was up-regulated during cold stress but down-regulated during dehydration and salt stress. Also, *ZOS1-21* was up-regulated during P1 but down-regulated during dehydration stress. Not even a single gene which was up-regulated during seed development was down-regulated during any stress. Validation of microarray data was done by real-time PCR and the expression of three seed-specific and one panicle-specific genes was comparable by both the methods (Fig. 7).

Fig. 7 Validation of microarray data by Q-PCR. Biological duplicates were taken for Q-PCR for four C_2H_2 ZFP genes, (**A**) *ZOS8-13*, (**B**) *ZOS1-15*, (**C**) *ZOS6-04* and (**D**) *ZOS5-09* and the data were compared with the normalized signal values from microarray. The white bar represents the expression from microarray data while the black bar represents the signal from Q-PCR. The standard error bar is shown



The expression patterns for C₂H₂ genes present in segmentally duplicated regions and for tandemly duplicated genes were also examined. Though all the members of the segmentally duplicated pairs of genes coded for C₂H₂ zinc finger proteins, varying expression patterns were observed. Out of 15 pairs of genes located in the segmentally duplicated regions, probe sets were available for 14 pairs on Affymetrix GeneChip®. In most of the cases, one of the genes of the pair did not express at significant levels (*ZOS1-14*, *ZOS5-10*; *ZOS2-13*, *ZOS4-09*; *ZOS2-14*, *ZOS4-10*; *ZOS8-06*, *ZOS9-13*; *ZOS8-07*, *ZOS9-14* and *ZOS8-10*, *ZOS9-16*). This is an indication of the high level of pseudofunctionalization in this gene family. The expression pattern was found to be highly similar for 4 pairs of genes (Fig. 8), even though the expression of one of the duplicated genes was extremely low. This may be due to the fact that the gene with low level of expression will, in due course of evolution, slowly lose its function as is the case with other duplicated pairs of genes. Two gene pairs viz., *ZOS2-18*, *ZOS3-02* and *ZOS2-10*, *ZOS4-06* did not show any significant expression levels in the tissues tested. For 2 pairs of genes, *ZOS4-11*, *ZOS2-15* and *ZOS8-09*, *ZOS9-15*; the expression pattern was different for most of the tissues tested indicating neofunctionalization. The result for tandem duplication was found to be highly similar to that for segmentally duplicated genes. Here also, out of a total of 14 groups for which the probe sets were present, 1 of the genes for 6 pairs showed pseudofunctionalization (Groups I, II, IV, VII, VIII, IX and XII) The expression pattern of the tandem genes was found to be same only for 4 groups (Fig. 9). The expression was extremely low for group XV and neofunctionalization was apparent for groups III and XI. These results suggest that either the upstream regulatory regions of duplicated genes are considerably different or there is a general regulatory mechanism that silences one of the duplicates.

Thirteen genes did not have corresponding probe sets on the Rice GeneChip used in the present study. Therefore, the MPSS data was also analyzed for the expression patterns of all the genes and the expression was found to be generally similar in comparable stages. An investigation to the expression of 13 genes not represented on the Rice GeneChip in MPSS database revealed that three such genes showed the presence of 17 bp signatures, with *ZOS1-02* showing high expression in a 35-day callus tissue, while *ZOS8-04* expressed in roots and 14-day-old leaves, and *ZOS11-04* expressed in 60-day mature leaves (Fig. 10). Seven other genes that did not show any expression in microarray-based analysis were found to have low to moderate expression levels when analyzed in MPSS database (Fig. 10).

Discussion

The C₂H₂ ZFPs are known to play important regulatory roles in various organisms (Dinneny et al. 2006; Gamsjaeger et al. 2007; Takatsuji 1999). Rice codes for 189 ZFPs, which can be classified in various types and are distributed on all chromosomes.

The phylogenetic analyses reveals that Q-type (clades II and III) and Z-type (clade I) ZF-containing ZFPs may have evolved in a plant-specific manner as none of the yeast ZFPs group with these clades (Fig. 4; Supplementary Fig. 2). Our analysis is also in favour of the hypothesis that plant ZFPs containing at least two ZFs may have evolved from C-type ZFs by conserving the QALGGH sequence in at least two type of ancestral fingers which may have given rise to present day QA- and QB- types of 2-fingered Q-type ZFPs (Fig. 3). These ZF domains in 2-fingered Q-type ZFPs either duplicated to give rise to 4-fingered ZFPs or lost one domain (probably QB) to give rise to single fingered ZFPs (Kubo et al. 1998). The QA type domain in single fingered ZFPs further diversified to form a separate clade. The M1, M4, M6 and M7 further diversified from the QB ancestor while M2, M3 and M5 evolved from QA lineage giving rise to multi-fingered ZFPs (Fig. 3). A similar conclusion was recently drawn from the comprehensive analysis of Arabidopsis ZFPs (Englbrecht et al. 2004). The data also suggest that the Z class of ZFs, which otherwise group with C-type ZFs, have also evolved relatively recently in a plant-specific manner.

Out of 189 *ZOS* genes in rice, the expression profiles of 179 genes could be studied in vegetative, reproductive and stressed tissues of *indica* rice. Since each reproductive stage analyzed in this study represented a complex set of tissues and cell types, the magnitude of the change in expression values of individual genes in a particular cell type may not be evident completely. Therefore, even a 2-fold estimated increase in the expression value could have high significance, as it would actually magnify several folds if only a particular cell-type or tissue were considered (Dinkins et al. 2002; Luo et al. 1999). By analyzing the entire spectrum of the reproductive development, we have been able to identify the genes relevant to the panicle and seed development. These data show that the genes up-regulated in narrow windows of reproductive development do not have very high expression signals, implying that their expression could be limited to specific cell types (Kapoor et al. 2002; Kapoor and Takatsuji 2006).

To date, many identified ZFPs have been implicated in abiotic-stress tolerance in plants (Takatsuji 1999). A ZFP from soybean, SCOF-1, is shown to be induced in response to cold and ABA (Kim et al. 2001a; Kim et al. 2001b). The STZ/ZAT10 from Arabidopsis can complement yeast

Fig. 8 Expression patterns of ZOS genes found in segmentally duplicated regions of the rice genome in the controls, leaf (L), root (R) and 7-day-old seedling (SDL), and in the stages of panicle development (P1-P6), seed development (S1-S5) and cold stress (CS), dehydration stress (DS) and salt stress (SS) were compared from microarray data. Duplicated genes are as indicated in Supplementary Table 3. The duplicate gene pairs marked with * show similar expression patterns

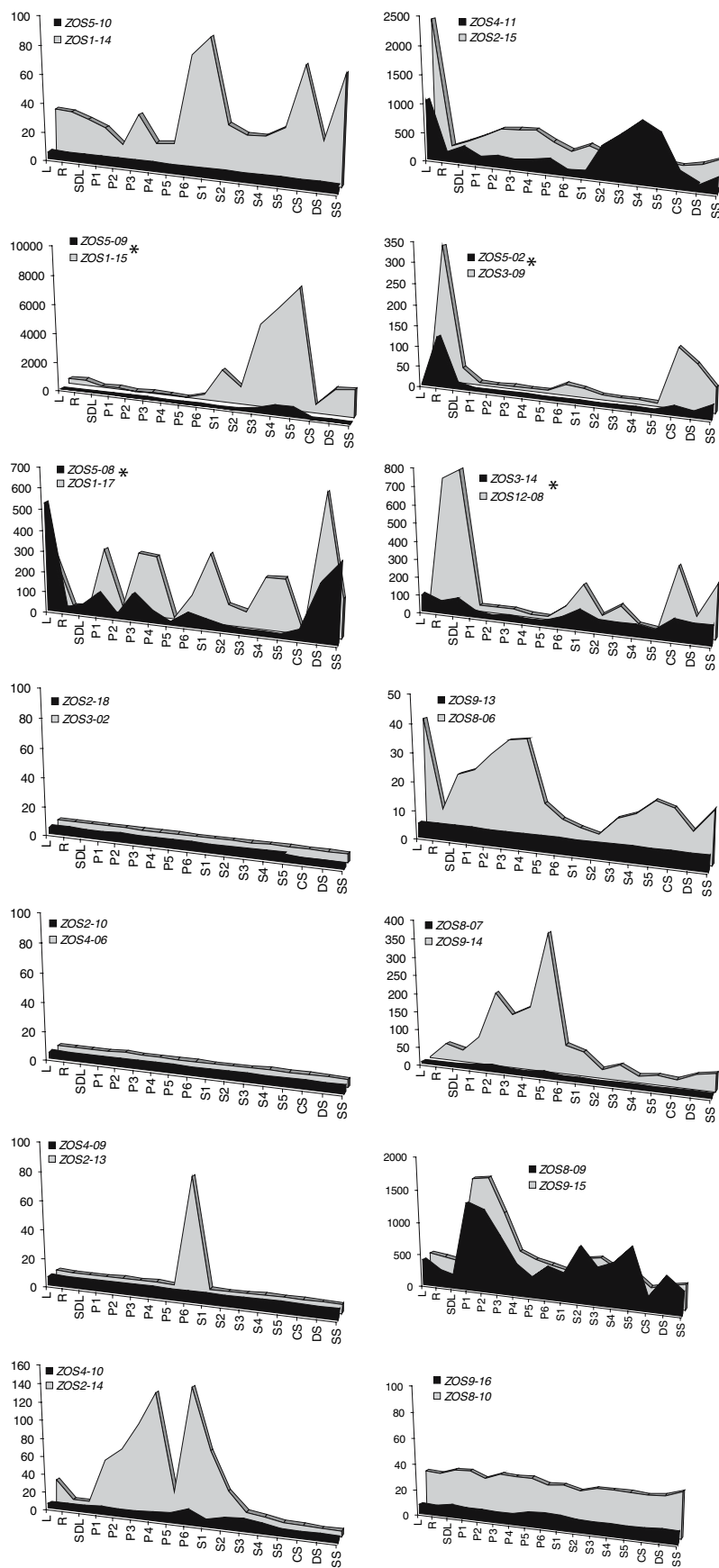


Fig. 9 Expression of tandemly duplicated *ZOS* genes in rice. The expression pattern of the 14 groups of tandemly duplicated genes (Supplementary Table 4) was examined in the controls, panicle and seed development tissues and in the stress samples. The pairs marked with * showed similar expression patterns

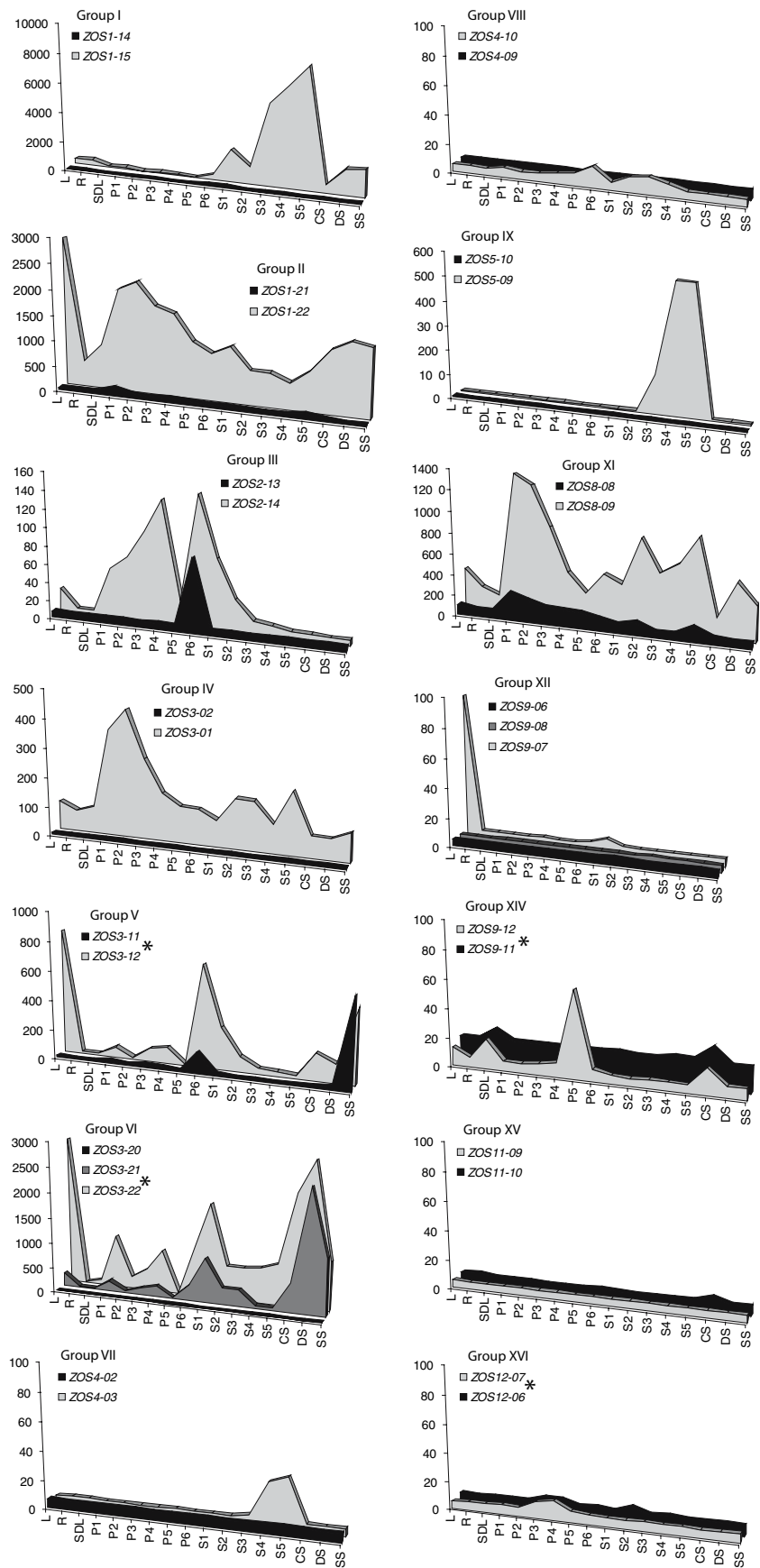
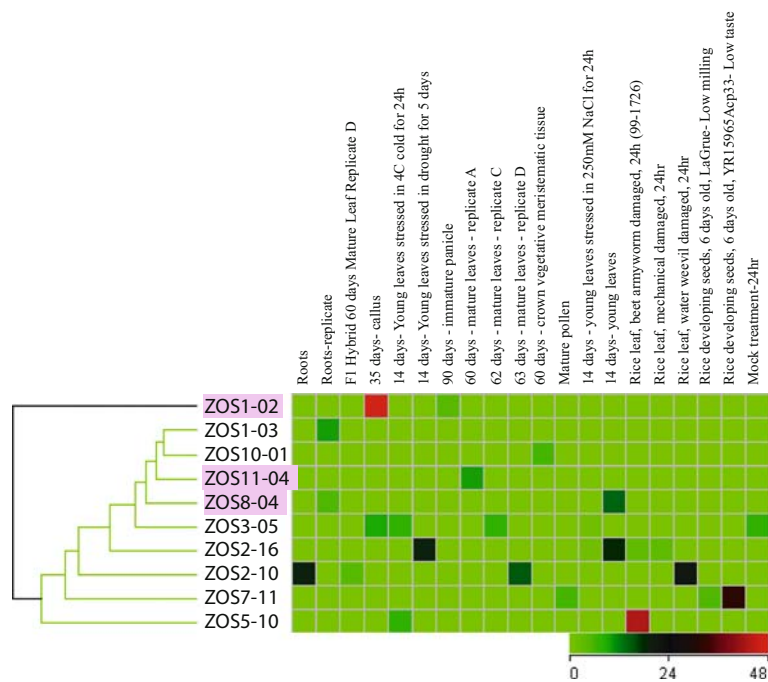


Fig. 10 MPSS expression of C_2H_2 ZFP genes. The expression pattern of 3 new genes (highlighted) and 8 genes showing negligible expression by microarray data is shown. The colour bar representing normalized transcripts per million (tpm) is shown



calcineurin mutants and increase salt tolerance of transgenic yeast (Lippuner et al. 1996; Sakamoto et al. 2000; Sakamoto et al. 2004). In petunia, *ZPT2-3* is induced in response to wounding, cold, drought and heavy metals and it improves drought tolerance of transgenic plants when expressed constitutively (Sugano et al. 2003). A *Medicago truncatula* ZFP Mt-ZIP1 was implicated in cytokinin, ABA and methyl jasmonate mediated stress response (Xu and Ma 2004). The microarray-based expression analysis presented here has revealed up-regulation of 26 rice ZFP genes in response to either drought and/or salt and/or cold stress. Recently, ZFP245 (named here as ZOS7-08) of rice has been shown to play a regulatory role in drought and cold stress via an ABA independent pathway (Huang et al. 2005b). In the present study, *ZOS7-08* expressed at very low levels in most of the stages analyzed with marginal increase in transcript abundance during S4 stage of seed development. It did not show any kind of stress inducibility under our experimental conditions. Since we exposed rice seedlings to stress conditions for only 3 hours, the time period may not have been sufficient for the induction of gene expression. Huang et al. (2005a, b) have shown the induction of gene expression after 6 h of stress. It may, therefore, suggest that most of the genes identified in our experimental conditions might be the early stress responsive genes. And, many of the genes that show very low expression values might in fact express in stages of development and under stress conditions other than those included here. This is clearly exemplified by the expression patterns of *ZOS3-05* and *ZOS7-11* which did not show any expression during microarray analysis but MPSS data

revealed that these genes expressed specifically in callus and mature pollen, respectively (Fig. 10). The down-regulation of 21 genes during abiotic stress clearly implies that they could be repressors of pathways important to stress. The down-regulation of these genes may also be a mechanism of adaptation to stress and hence a part of suppression of some other pathways (Blodner et al. 2007; Bona et al. 2007). Additionally, on the basis of domain analysis and MPSS small RNA database, it seems that *ZOS8-03* is really expressed (data not shown).

Our analysis did not reveal any correlation between the percentage homology and the expression profiles of the genes in the segmentally duplicated regions of rice. However, a pair of genes was found to have similar expression patterns even though they showed ~40% homology in their cDNA sequences (Supplementary Table 3; Fig. 8).

Apart from regulation at the level of transcription, genes with intron-exon gene structure can possibly also be regulated at the level of splicing. Genes with multiple introns can be alternatively spliced, leading in many cases to functional protein isoforms, thus increasing the coding potential of the genome (Smith and Valcarcel 2000). There are a number of well-known examples in *Drosophila*, *Caenorhabditis elegans* and vertebrates resulting in 2 to as many as 576 possible alternatively spliced forms (Black 1998; Blencowe 2006; Cahill 2004; Shao et al. 2006; Sharp 1994; Worch et al. 2006). Unlike the many examples in animal systems, there are very few examples of gene regulation involving alternative splicing in plants (Lorkovic et al. 2000; Lucyshyn et al. 2007; Reddy 2001). Ribulose-1,5-bisphosphate carboxylase/ oxygenase (Rubisco)

activase represents one of the best-known examples of alternative splicing in higher plants. Alternative splicing of Rubisco activase, in most plant species, yields two polypeptides of different sizes that are identical except for an additional number of amino acids at the C terminus of the larger isoform (Werneke et al. 1989). A list of 29 plant pre-mRNAs that undergo alternative splicing, many of which are known to encode proteins with different functions and/or different cellular or subcellular localization patterns, was presented by Reddy (2001). In rice, 21.2% of the expressed genes undergo alternative splicing (Wang and Brendel 2006). Amongst the genes involved in reproductive development, *FCA* is alternatively spliced to form four different transcripts in a spatially and temporally controlled manner to limit the amount of functional FCA protein (Macknight et al. 2002). Of the 189 *ZOS* genes described here 20 loci can possibly be alternatively spliced. cDNA evidence exists for 12 *ZOS* loci resulting in 22 genes which exhibit single or multiple exon exclusions and retention of partial or complete intron sequences. Six of the loci also produce gene models with missing ZF coding sequences and are likely to generate functionally distinct proteins. Studies aimed at understanding differential expression of these alternatively spliced mRNA in a development and temporal context could unravel some of the mysteries concerning role of post-transcriptional gene regulation in plants.

Additionally, six genes were found which had a KOME cDNA transcribed from the opposite direction (Supplementary Table 2). Though ‘nested genes’ or overlapping genes reported from organisms as diverse as bacteria, viruses, *Drosophila* and humans (Gibson et al. 2005), lie in opposite orientation in the intron of a larger gene, in rice, the exons of antisense genes overlap with the exons of the C_2H_2 genes (*ZOS1-12*, *ZOS1-22* and *ZOS12-02*). In plants, genes with overlapping ends have been reported from maize chloroplast (Schwarz et al. 1981). However, the analysis of our microarray data did not suggest any direct influence of the expression of one gene on the other (data not shown).

In conclusion, a comprehensive account of C_2H_2 ZF protein encoding gene family has been presented along with the expression profiles of 179 genes during stages of vegetative and reproductive development and three abiotic stress conditions. These data would be useful in selecting candidate genes for functional validation in relation to stress and various aspects of reproductive development in rice and other crops.

Acknowledgements The research work was supported by the Department of Biotechnology, Government of India. UGC Research Fellowship to P.A. and CSIR Research Fellowship to R.A. and S.R. are acknowledged.

References

- Black DL (1998) Splicing in the inner ear: a familiar tune, but what are the instruments? *Neuron* 20:165–168
- Blencowe BJ (2006) Alternative splicing: new insights from global analyses. *Cell* 126:37–47
- Blodner C, Goebel C, Feussner I, Gatz C, Polle A (2007) Warm and cold parental reproductive environments affect seed properties, fitness, and cold responsiveness in *Arabidopsis thaliana* progenies. *Plant Cell Environ* 30:165–175
- Bohm S, Frishman D, Mewes HW (1997) Variations of the C2H2 zinc finger motif in the yeast genome and classification of yeast zinc finger proteins. *Nucleic Acids Res* 25:2464–2469
- Bona E, Marsano F, Cavaletto M, Berta G (2007) Proteomic characterization of copper stress response in *Cannabis sativa* roots. *Proteomics* 7:1121–1130
- Cahill K (2004) Alternative splicing and genomic stability. *Phys Biol* 1:C1–C4
- Cooke HJ (2004) Silence of the centromeres—not. *Trends Biotechnol* 22:319–321
- Davletova S, Schlauch K, Coutu J, Mittler R (2005) The zinc-finger protein Zat12 plays a central role in reactive oxygen and abiotic stress signaling in *Arabidopsis*. *Plant Physiol* 139:847–856
- Dinkins R, Pflipsen C, Thompson A, Collins GB (2002) Ectopic expression of an *Arabidopsis* single zinc finger gene in tobacco results in dwarf plants. *Plant Cell Physiol* 43:743–750
- Dinneny JR, Weigel D, Yanofsky MF (2006) NUBBIN and JAGGED define stamen and carpel shape in *Arabidopsis*. *Development* 133:1645–1655
- 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
- Florea L, Hartzell G, Zhang Z, Rubin GM, Miller W (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res* 8:967–974
- Gamsjaeger R, Liew CK, Loughlin FE, Crossley M, Mackay JP (2007) Sticky fingers: zinc-fingers as protein-recognition motifs. *Trends Biochem Sci* doi:10.1016/j.tibs.2006.1012.1007
- Gibson CW, Thomson NH, Abrams WR, Kirkham J (2005) Nested genes: biological implications and use of AFM for analysis. *Gene* 350:15–23
- Grigg SP, Canales C, Hay A, Tsiantis M (2005) SERRATE coordinates shoot meristem function and leaf axial patterning in *Arabidopsis*. *Nature* 437:1022–1026
- Haas BJ, Delcher AL, Wortman JR, Salzberg SL (2004) DAGchainer: a tool for mining segmental genome duplications and synteny. *Bioinformatics* 20:3643–3646
- Hiratsu K, Mitsuda N, Matsui K, Ohme-Takagi M (2004) Identification of the minimal repression domain of SUPERMAN shows that the DLELRL hexapeptide is both necessary and sufficient for repression of transcription in *Arabidopsis*. *Biochem Biophys Res Commun* 321:172–178
- Huang J, Wang J, Zhang H (2005) Rice *ZFP15* gene encoding for a novel C2H2-type zinc finger protein lacking DLN box, is regulated by spike development but not by abiotic stresses. *Mol Biol Rep* 32:177–183
- Huang J, Wang JF, Wang QH, Zhang HS (2005) Identification of a rice zinc finger protein whose expression is transiently induced by drought, cold but not by salinity and abscisic acid. *DNA Seq* 16:130–136
- Hulo N, Bairoch A, Bulliard V, Cerutti L, De Castro E, Langendijk-Genevaux PS, Pagni M, Sigrist CJ (2006) The PROSITE database. *Nucleic Acids Res* 34:D227–230
- Iuchi S (2001) Three classes of C_2H_2 zinc finger proteins. *Cell Mol Life Sci* 58:625–635

- Jain M, Nijhawan A, Arora R, Agarwal P, Ray S, Sharma P, Kapoor S, Tyagi AK, Khurana JP (2007) F-box proteins in rice: genome-wide analysis, classification and spatial and temporal gene expression during panicle and seed development and regulation by light and abiotic stress. *Plant Physiol* 143:1467–1483
- Jeong YM, Mun JH, Lee I, Woo JC, Hong CB, Kim SG (2006) Distinct roles of the first introns on the expression of Arabidopsis profilin gene family members. *Plant Physiol* 140:196–209
- Kapoor S, Kobayashi A, Takatsuji H (2002) Silencing of the tapetum-specific zinc finger gene *TAZI* causes premature degeneration of tapetum and pollen abortion in petunia. *Plant Cell* 14:2353–2367
- Kapoor S, Takatsuji H (2006) Silencing of an anther-specific zinc-finger gene, *MEZI*, causes aberrant meiosis and pollen abortion in petunia. *Plant Mol Biol* 61:415–430
- Kikuchi S, Satoh K, Nagata T, Kawagashira N, Doi K, Kishimoto N, Yazaki J, Ishikawa M, Yamada H, Ooka H, Hotta I, Kojima K, Namiki T, Ohneda E, Yahagi W, Suzuki K, Li CJ, Ohtsuki K, Shishiki T, Otomo Y, Murakami K, Iida Y, Sugano S, Fujimura T, Suzuki Y, Tsunoda Y, Kurosaki T, Kodama T, Masuda H, Kobayashi M, Xie Q, Lu M, Narikawa R, Sugiyama A, Mizuno K, Yokomizo S, Niikura J, Ikeda R, Ishibiki J, Kawamata M, Yoshimura A, Miura J, Kusumegi T, Oka M, Ryu R, Ueda M, Matsubara K, Kawai J, Carninci P, Adachi J, Aizawa K, Arakawa T, Fukuda S, Hara A, Hashizume W, Hayatsu N, Imotani K, Ishii Y, Itoh M, Kagawa I, Kondo S, Konno H, Miyazaki A, Osato N, Ota Y, Saito R, Sasaki D, Sato K, Shibata K, Shinagawa A, Shiraki T, Yoshino M, Hayashizaki Y, Yasunishi A (2003) Collection, mapping, and annotation of over 28,000 cDNA clones from japonica rice. *Science* 301:376–379
- Kim JC, Jeong JC, Park HC, Yoo JH, Koo YD, Yoon HW, Koo SC, Lee SH, Bahk JD, Cho MJ (2001) Cold accumulation of SCOF-1 transcripts is associated with transcriptional activation and mRNA stability. *Mol Cells* 12:204–208
- Kim JC, Lee SH, Cheong YH, Yoo CM, Lee SI, Chun HJ, Yun DJ, Hong JC, Lee SY, Lim CO, Cho MJ (2001) A novel cold-inducible zinc finger protein from soybean, SCOF-1, enhances cold tolerance in transgenic plants. *Plant J* 25:247–259
- Klug A, Schwabe JW (1995) Protein motifs 5. Zinc fingers. *FASEB J* 9:597–604
- Kobayashi A, Sakamoto A, Kubo K, Rybka Z, Kanno Y, Takatsuji H (1998) Seven zinc-finger transcription factors are expressed sequentially during the development of anthers in petunia. *Plant J* 13:571–576
- Kozaki A, Hake S, Colasanti J (2004) The maize ID1 flowering time regulator is a zinc finger protein with novel DNA binding properties. *Nucleic Acids Res* 32:1710–1720
- Kubo K, Sakamoto A, Kobayashi A, Rybka Z, Kanno Y, Nakagawa H, Takatsuji H (1998) Cys₂/His₂ zinc-finger protein family of petunia: evolution and general mechanism of target-sequence recognition. *Nucleic Acids Res* 26:608–615
- Le Hir H, Nott A, Moore MJ (2003) How introns influence and enhance eukaryotic gene expression. *Trends Biochem Sci* 28:215–220
- Li K, Yang J, Liu J, Du X, Wei C, Su W, He G, Zhang Q, Hong F, Qian X (2006) Cloning, characterization and tissue-specific expression of a cDNA encoding a novel *EMBRYONIC FLOWER 2* gene (*OsEMF2*) in *Oryza sativa*. *DNA Seq* 17:74–78
- Lippuner V, Cyert MS, Gasser CS (1996) Two classes of plant cDNA clones differentially complement yeast calcineurin mutants and increase salt tolerance of wild-type yeast. *J Biol Chem* 271:12859–12866
- Lorkovic ZJ, Wiczeorek Kirk DA, Lambermon MH, Filipowicz W (2000) Pre-mRNA splicing in higher plants. *Trends Plant Sci* 5:160–167
- Lucyshyn D, Busch BL, Abolmaali S, Steiner B, Chandler E, Sanjarian F, Mousavi A, Nicholson P, Buerstmayer H, Adam G (2007) Cloning and characterization of the ribosomal protein L3 (*RPL3*) gene family from *Triticum aestivum*. *Mol Genet Genomics* doi:10.1007/s00438-00006-00201-00431
- Luo M, Bilodeau P, Koltunow A, Dennis ES, Peacock WJ, Chaudhury AM (1999) Genes controlling fertilization-independent seed development in *Arabidopsis thaliana*. *Proc Natl Acad Sci USA* 96:296–301
- Macknight R, Duroux M, Laurie R, Dijkwel P, Simpson G, Dean C (2002) Functional significance of the alternative transcript processing of the Arabidopsis floral promoter *FCA*. *Plant Cell* 14:877–888
- Meissner R, Michael AJ (1997) Isolation and characterisation of a diverse family of *Arabidopsis* two and three-fingered C2H2 zinc finger protein genes and cDNAs. *Plant Mol Biol* 33:615–624
- Nagaki K, Cheng Z, Ouyang S, Talbert PB, Kim M, Jones KM, Henikoff S, Buell CR, Jiang J (2004) Sequencing of a rice centromere uncovers active genes. *Nat Genet* 36:138–145
- Nakagawa H, Jiang CJ, Sakakibara H, Kojima M, Honda I, Ajisaka H, Nishijima T, Koshioka M, Homma T, Mander LN, Takatsuji H (2005) Overexpression of a petunia zinc-finger gene alters cytokinin metabolism and plant forms. *Plant J* 41:512–523
- Ohta M, Matsui K, Hiratsu K, Shinshi H, Ohme-Takagi M (2001) Repression domains of class II ERF transcriptional repressors share an essential motif for active repression. *Plant Cell* 13:1959–1968
- Pruitt KD, Tatusova T, Maglott DR (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* 33:D501–504
- Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R (2005) InterProScan: protein domains identifier. *Nucleic Acids Res* 33:W116–120
- Reddy ASN (2001) Nuclear pre-mRNA splicing in plants. *Critical Reviews in Plant Sciences* 20:523–571
- Rose AB (2002) Requirements for intron-mediated enhancement of gene expression in Arabidopsis. *RNA* 8:1444–1453
- Runkel F, Michels M, Franz T (2003) Fxyd3 and Lgi4 expression in the adult mouse: a case of endogenous antisense expression. *Mamm Genome* 14:665–672
- Sakamoto H, Araki T, Meshi T, Iwabuchi M (2000) Expression of a subset of the *Arabidopsis* Cys₂/His₂-type zinc-finger protein gene family under water stress. *Gene* 248:23–32
- Sakamoto H, Maruyama K, Sakuma Y, Meshi T, Iwabuchi M, Shinozaki K, Yamaguchi-Shinozaki K (2004) Arabidopsis Cys₂/His₂-type zinc-finger proteins function as transcription repressors under drought, cold, and high-salinity stress conditions. *Plant Physiol* 136:2734–2746
- Schuh R, Aicher W, Gaul U, Cote S, Preiss A, Maier D, Seifert E, Nauber U, Schroder C, Kemler R, et al. (1986) A conserved family of nuclear proteins containing structural elements of the finger protein encoded by *Krippel*, a *Drosophila* segmentation gene. *Cell* 47:1025–1032
- Schwarz Z, Jolly SO, Steinmetz AA, Bogorad L (1981) Overlapping divergent genes in the maize chloroplast chromosome and in vitro transcription of the gene for tRNA His. *Proc Natl Acad Sci USA* 78:3423–3427
- Shao H, Zhu C, Zhao Z, Guo M, Qiu H, Liu H, Wang D, Xue L, Gao L, Sun C, Li W (2006) KRAB-containing zinc finger gene *ZNF268* encodes multiple alternatively spliced isoforms that contain transcription regulatory domains. *Int J Mol Med* 18:457–463
- Sharp PA (1994) Split genes and RNA splicing. *Cell* 77:805–815
- Singh G, Kumar S, Singh P (2003) A quick method to isolate RNA from wheat and other carbohydrate-rich seeds. *Plant Mol Biol Rep* 21:93a-f

- Smith CW, Valcarcel J (2000) Alternative pre-mRNA splicing: the logic of combinatorial control. *Trends Biochem Sci* 25:381–388
- Sugano S, Kaminaka H, Rybka Z, Catala R, Salinas J, Matsui K, Ohme-Takagi M, Takatsuji H (2003) Stress-responsive zinc finger gene *ZPT2-3* plays a role in drought tolerance in petunia. *Plant J* 36:830–841
- Takatsuji H, Mori M, Benfey PN, Ren L, Chua NH (1992) Characterization of a zinc finger DNA-binding protein expressed specifically in petunia petals and seedlings. *EMBO J* 11:241–249
- Takatsuji H, Nakamura N, Katsumoto Y (1994) A new family of zinc finger proteins in petunia: structure, DNA sequence recognition, and floral organ-specific expression. *Plant Cell* 6:947–958
- Takatsuji H, Matsumoto T (1996) Target-sequence recognition by separate-type Cys₂/His₂ zinc finger proteins in plants. *J Biol Chem* 271:23368–23373
- 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
- Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22:4673–4680
- 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
- Wang BB, Brendel V (2006) Genomewide comparative analysis of alternative splicing in plants. *Proc Natl Acad Sci USA* 103:7175–7180
- Werneke JM, Chatfield JM, Ogren WL (1989) Alternative mRNA splicing generates the two ribulosebisphosphate carboxylase/oxygenase activase polypeptides in spinach and *Arabidopsis*. *Plant Cell* 1:815–825
- Worch S, Fiedler E, Hansmann I, Schlote D (2006) Genomic organization and expression pattern of scapinin (PHACTR3) in mouse and human. *Cytogenet Genome Res* 115:23–29
- Wu Z, Irizarry RA, Gentleman R, Murillo FM, Spencer F (2003) A Model Based Background Adjustment for Oligonucleotide Expression Arrays. Technical Report, Department of Biostatistics Working Papers, Baltimore, MD
- Xu Y, Ma QH (2004) *Medicago truncatula Mt-ZFP1* encoding a root enhanced zinc finger protein is regulated by cytokinin, abscisic acid and jasmonate, but not cold. *DNA Seq* 15:104–109
- Yang L, Liu Z, Lu F, Dong A, Huang H (2006) SERRATE is a novel nuclear regulator in primary microRNA processing in *Arabidopsis*. *Plant J* 47:841–850