Expansion Mechanisms and Evolutionary History on Genes Encoding DNA Glycosylases and Their Involvement in Stress and Hormone Signaling

Shu-Ye Jiang and Srinivasan Ramachandran*

Genome Structural Biology Group, Temasek Life Science Laboratory, The National University of Singapore, Singapore

*Corresponding author: E-mail: sri@tll.org.sg.

Accepted: March 19, 2016

Abstract

DNA glycosylases catalyze the release of methylated bases. They play vital roles in the base excision repair pathway and might also function in DNA demethylation. At least three families of DNA glycosylases have been identified, which included 3′-methyladenine DNA glycosylase (MDG) I, MDG II, and HhH-GPD (Helix–hairpin–Helix and Glycine/Proline/aspartate (D)). However, little is known on their genome-wide identification, expansion, and evolutionary history as well as their expression profiling and biological functions. In this study, we have genome-widely identified and evolutionarily characterized these family members. Generally, a genome encodes only one *MDG II* gene in most of organisms. No *MDG I* or *MDG II* gene was detected in green algae. However, *HhH-GPD* genes were detectable in all available organisms. The ancestor species contain small size of *MDG I* and *HhH-GPD* families. These two families were mainly expanded through the whole-genome duplication and segmental duplication. They were evolutionarily conserved and were generally under purifying selection. However, we have detected recent positive selection among the *Oryza* genus, which might play roles in species divergence. Further investigation showed that expression divergence played important roles in gene survival after expansion. All of these family genes were expressed in most of developmental stages and tissues in rice plants. High ratios of family genes were downregulated by drought and fungus pathogen as well as abscisic acid (ABA) and jasmonic acid (JA) treatments, suggesting a negative regulation in response to drought stress and pathogen infection through ABA- and/or JA-dependent hormone signaling pathway.

Key words: DNA glycosylase, abiotic/biotic stresses, evolution, gene expansion, hormone signaling, transcriptome.

Introduction

Genomic DNA molecules continuously suffer damages due to their exposure to internal and external environment and man-made toxins, such as radiation, chemical mutagens, biotic and abiotic stresses. The effects of these damages on organisms were determined by the chemical nature of the lesion and reparability. Evidence from microorganisms and mammals suggested that some base modification, for example, 7-methylguanine (7-MeG) by methylating agents, might not be harmful as they did not inhibit or alter normal base pairing (Larson et al. 1985). Another class of damages is O^6-methylguanine, which is directly mutagenic and leads to mispairing with thymine (Loechler et al. 1984). The third class of lesions is 3-methyladenine (3-MeA), which acts as blocks to DNA replication and transcription (Larson et al. 1985) so is cytotoxic. 3′-Methyladenine DNA glycosylase (MDG) I specifically catalyzes the release of 3-methylated adenine and, to a lesser extent, guanosine bases from alkylated-DNA by hydrolysis of the deoxyribose *N*-glycosidic bond (Sakumi et al. 1986; Drohat et al. 2002). Thus, the glycosylase plays a vital role in the base excision repair (BER) (Wyatt et al. 1999).

In addition to MDG I, MDG II catalyzes the release of not only 3-MeA but also a variety of other methylated bases including 3-MeG, 7-MeG, O^2-MeT, and O^2-MeC (Sakumi and Sekiguchi 1990). Both *MDG I* and *II* have been cloned and functionally and structurally characterized from a variety of microbial and mammalian sources (Lee et al. 2009; Calvo et al. 2013; Ebrahimkhani et al. 2014; Admiraal and O'Brien 2015; Taylor and O'Brien 2015). However, in plants, limited data are available on their characterization and biological functions. Santerre and Britt (1994) first cloned an *Arabidopsis* gene encoding an MDG and it complemented the methyl methanesulfonate-sensitive phenotype of an *Escherichia coli* double mutant deficient in 3-MeA glycosylases. This protein belongs to MDG II. Expression analysis showed that the *Arabidopsis* methyladenine DNA glycosylase gene is preferentially expressed in rapidly dividing tissues such as meristematic tissue, the developing embryo and endosperm, and organ primordial (Shi et al. 1997).

Besides MDG I and II, other DNA glycosylases have also been reported (Wyatt et al. 1999). These proteins belong to the HhH-GPD superfamily as they contain Helix–hairpin–Helix and Glycine/Proline rich loop followed by a conserved aspartate (one letter code is D) (Nash et al. 1996). The superfamily is the largest and most functionally diverse group of DNA glycosylases. In microorganisms and mammals, HhH-GPDs catalyze the release of 3-MeA, 3-MeG, and 7-MeG (Wyatt et al. 1999). In plants, the *Arabidopsis* Repressor of Silencing 1 (Ponferrada-Marín et al. 2011) showed the similarity to the HhH-GPD proteins. A few other members were also reported such as genes encoding methyl-binding domain protein 4 (MBD4) or MBD4-like (Ramiro-Merina et al. 2013; Nota et al. 2015). Instead of 3-MeA, 3-MeG and 7-MeG in microorganisms and mammals, these DNA glycosylases excise 5-methylcytosine (Zhu 2009; Ponferrada-Marín et al. 2011; Ramiro-Merina et al. 2013). Thus, these enzymes also function in DNA demethylation through the BER pathway (Zhu 2009). DNA demethylation has been proved to be related to various abiotic stresses (Lukens and Zhan 2007). Recent studies showed that the glycosylases exhibited DNA-binding activities (Malhotra and Sowdhamini 2013).

Generally, only a few reports have been published on MDG I and II and little is known on their biological functions of these enzymes in plants. For the *HhH-GPD* superfamily, only several members were cloned and functionally characterized. However, little is known on the genome-wide identification and characterization of these genes encoding methyladenine DNA glycosylases and their expansion and evolution. Here, we have genome-widely identified these three gene families in 15 genomes. They exhibited different expansion and evolution histories. Both the whole-genome duplication and segmental duplication significantly contributed to the expansion of both *MDG I* and *HhH-GPD* families. In the rice genome, we have identified 20 genes encoding methylation-related DNA glycosylases including 6 *MDG I*, 1 *MDG II*, and 13 *HhH-GPD* genes. All of these genes were expressed in most of stages of rice development with differential expression abundance. Their expression was also regulated by drought and hormone treatments. In general, our data suggested that both *MDG I* and *HhH-GPD* family members might play a role in the drought stress and hormone signaling pathway in plants.

Materials and Methods

DNA/cDNA and Protein Databases for Genome-Wide Identification and Characterization

The all annotated rice gene and protein sequences were downloaded from the latest version (release 7) of the rice genome annotation database (Kawahara et al. 2013; http://rice.plantbiology.msu.edu/, last accessed March 31, 2016). For *Arabidopsis thaliana*, the latest version of the *Arabidopsis* genome annotation (TAIR10; http://www.arabidopsis.org, last accessed March 31, 2016) was used for retrieving all annotated gene and protein sequences (Lamesch et al. 2012). The gene and protein sequences from remaining 48 species were downloaded from the release v10.2 of Phytozome database (http://phytozome.jgi.doe.gov/, last accessed March 31, 2016).

Besides sequences from both indica and japonica rice databases, additional rice DNA/cDNA, and protein sequences from nine other rice species including *Oryza barthii*, *Oryza brachyantha*, *Oryza glaberrima*, *Oryza glumaepatula*, *Oryza longistaminata*, *Oryza meridionalis*, *Oryza nivara*, *Oryza punctate*, and *Oryza rufipogon* were downloaded from the Ensembl Plants database (http://plants.ensembl.org/index.html, last accessed March 31, 2016). The resequencing data of 1,402 rice accessions were obtained from the RiceVarMap database (http://ricevarmap.ncpgr.cn/, last accessed March 31, 2016).

Profile Hidden Markov Model Searches

Protein sequences of the MDG I, MDG II and HhH-GPD families contain a conserved domain structure with Pfam (http://pfam.xfam.org, last accessed March 31, 2016) ID PF03352, PF02245 and PF00730, respectively. The seed domain amino acid sequences were downloaded from the Pfam database (http://pfam.xfam.org/) and were used for building a hidden Markov model (HMM) profile with the HMMER 2.3.2 (http://hmmer.org/, last accessed March 31, 2016). We used the profile HMMs to scan the above mentioned 50 protein databases with *E*-value cut-off of 1.0. We then manually inspected the resulted sequences by domain detection to remove any artifacts. The obtained protein sequences were also used as queries for BLASTP searches with *E*-value less than 0.01 followed by domain verification to achieve more family members.

Protein Domain Alignment and Phylogenetic Analysis

As only one member was detected for the MDG II family in each species, no alignment was carried out for this family. For the MDG I and HhH-GPD families, domain amino acid sequences were achieved from 15 species, which included 6 dicot plants (*Arabidopsis thaliana*, *Brassica rapa*, *Malus domestica*, *Prunus persica*, *Populus trichocarpa*, and *Ricinus communis*), 6 monocot plants (*Brachypodium distachyon*, *Musa acuminata*, *Oryza sativa*, *Sorghum bicolor*, *Triticum aestivum*, *Zea mays*), 1 spikemoss (*Selaginella moellendorffii*), 1 moss (*Physcomitrella patens*), and 1 green alga (*Chlamydomonas reinhardtii*). The domain amino acid sequences were aligned using ClustalX 2.0 (http://www.clustal.org/; Thompson et al. 1997) and the alignment was manually edited with Jalview (version 2, Waterhouse et al. 2009). The aligned sequences were used for phylogenetic tree construction and analysis according to the previous description by Jiang and Ramachandran (2006).

Estimation of *Ka* (Nonsynonymous Substitutions per Site)/*Ks* (Synonymous Substitutions per Site) and Detection of Positive/Purifying Selection

To calculate the *Ka/Ks* ratios, domain or full-length protein sequences were aligned first using ClustalX 2.0 as mentioned above. The PAL2NAL program (Suyama et al. 2006) was used to convert a multiple sequence alignment of proteins and the corresponding cDNA sequences into a codon alignment. The aligned cDNA sequences were used to calculate the value of *Ka* and *Ks* as well as their ratios using the yn00 program of the PAML4b package (Yang and Nielsen 2000). The program "sitewise likelihood-ratio" (SLR; Massingham and Goldman 2005) was used to detect purifying/positively selected amino acid sites in a family using both phylogenetic trees and codon alignment.

Detection of Gene Expansion Mechanisms

To explore the mechanisms of *MDG I* and *HhH-GPD* family expansion, we investigated the contribution of the whole-genome duplication, tandem and segmental duplication, as well as mobile elements to the family expansion. The whole-genome duplication data were achieved from the plant genome duplication database (PGDD; http://chibba.agtec.uga.edu/duplication/ [last accessed March 31, 2016], Lee et al. 2013). Tandemly duplicated *MDG I/HhH-GPD* genes in 15 species were identified by three criteria: 1) Within ten genes apart, 2) belong to the same family, and 3) within 100 kb for *Arabidopsis*, moss and green algae or 350 kb for the remaining species. Segmentally duplicated chromosome blocks were identified using the flanking regions (50 kb upstream and downstream) of *MDG/HhH-GPD* genes according to the description by Kong et al. (2007). These genes that were located on segmentally duplicated chromosome blocks were regarded as segmentally duplicated genes. To determine the contribution of mobile elements to the expansion of the *MDG/HhH-GPD* family, the flanking genomic sequences of the 50-kb upstream and downstream of these genes were achieved from corresponding genomes. These sequences were used to identify major transposon family members according to the description (Jiang et al. 2009). We identified the following mobile elements including mutator-like transposable element (*MULE*), *hobo/Ac/ Tam3 (hAT)*, *CACTA*, retrotransposons and *Helitron* families as well as retrogenes.

Expression Databases Used in This Study

Several expression data sets were achieved for profiling transcriptome of *MDG I* and *HhH-GPD* genes in rice. The data set with GEO (Gene Expression Omnibus; Barrett et al. 2013) accession number GSE21396 (Sato et al. 2013) was used to evaluate the spatiotemporal gene expression of various tissues in the whole rice life cycle. The data set GSE6901 (Jain et al. 2007) was used to investigate the stress regulation under drought, high salinity and cold stresses. The third data set with GEO accession number GSE39429 (Sato et al. 2013) was employed to analyze the gene expression profile in response to various plant hormones. We also investigated the effects of fungus and bacterium pathogens on gene expression of *MDG I* and *HhH-GPD* families by using the data sets with accession numbers GSE62894 and GSE63047. The expression patterns in different tissue types in rice roots were carried out by using the data set GSE30136 (Takehisa et al. 2012). The data sets GSE12508 (Schreiber et al. 2009) and GSE29303 were used to analyze the expression divergence of duplicated genes in wheat and poplar, respectively. Expression divergence among expanded genes was determined according to their expression abundance among different tissues or under different abiotic/biotic stresses. Genes with at least two times difference in their processed signal value based on computing geometric mean between tissues/treatments were submitted for Student's *t*-test. These genes with a statistical difference at $P < 0.05$ were regarded as divergent genes in their expression. Similarly, the method was also applied to the identification of up- or downregulated genes under various abiotic and biotic stresses.

Results

Genome-Wide Identification of Genes Encoding DNA Glycosylases in 15 Species

To genome-widely identify genes encoding DNA glycosylases, we first surveyed the conserved domains in representative protein sequences. We submitted all these protein sequences to the Pfam database (http://pfam.xfam.org/) for domain searches. We found that all available MDG I proteins contained a conserved domain structure with Pfam ID PF00352. Similarly, all the MDG II and HhH-GPD proteins have conserved domains with Pfam IDs PF02245 and PF00730, respectively. We then downloaded the representative domain sequences for building a profile HMM. Totally, we have built three HMM files based on three Pfam IDs. Subsequently, we executed the profile HMM searches against protein databases from 15 species. These species include 6 dicot, 6 monocot, 1 spikemosss, 1 moss, and 1 green alga.

By executing the profile HMM searches against the protein databases from 15 species, we have identified a total of 102 *MDG I*, 14 *MDG II*, and 173 *HhH-GPD* genes (supplementary tables S1–S3, Supplementary Material online). Neither *MDG I* nor *MDG II* gene was identified in the green alga genome. For the *MDG II* genes, only one member was encoded in each genome in the remaining 14 species and no duplication or expansion was found for the gene. For the *MDG I* genes, the 14 genomes encode varying numbers of members ranging from 2 to 16 genes. For the HhH-GPD family, the 15 genomes encode at least five members each and the wheat genome encodes the highest numbers (23) of *HhH-GPD* genes. In rice, we have identified 6 *MDG I* and 13 *HhH-GPD* genes. In general, during long evolution history, plant genomes have evolved into different sizes of DNA glycosylase families.

Both *MDG I* And *HhH-GPD* Families Exhibited Different Expansion and Evolutionary History

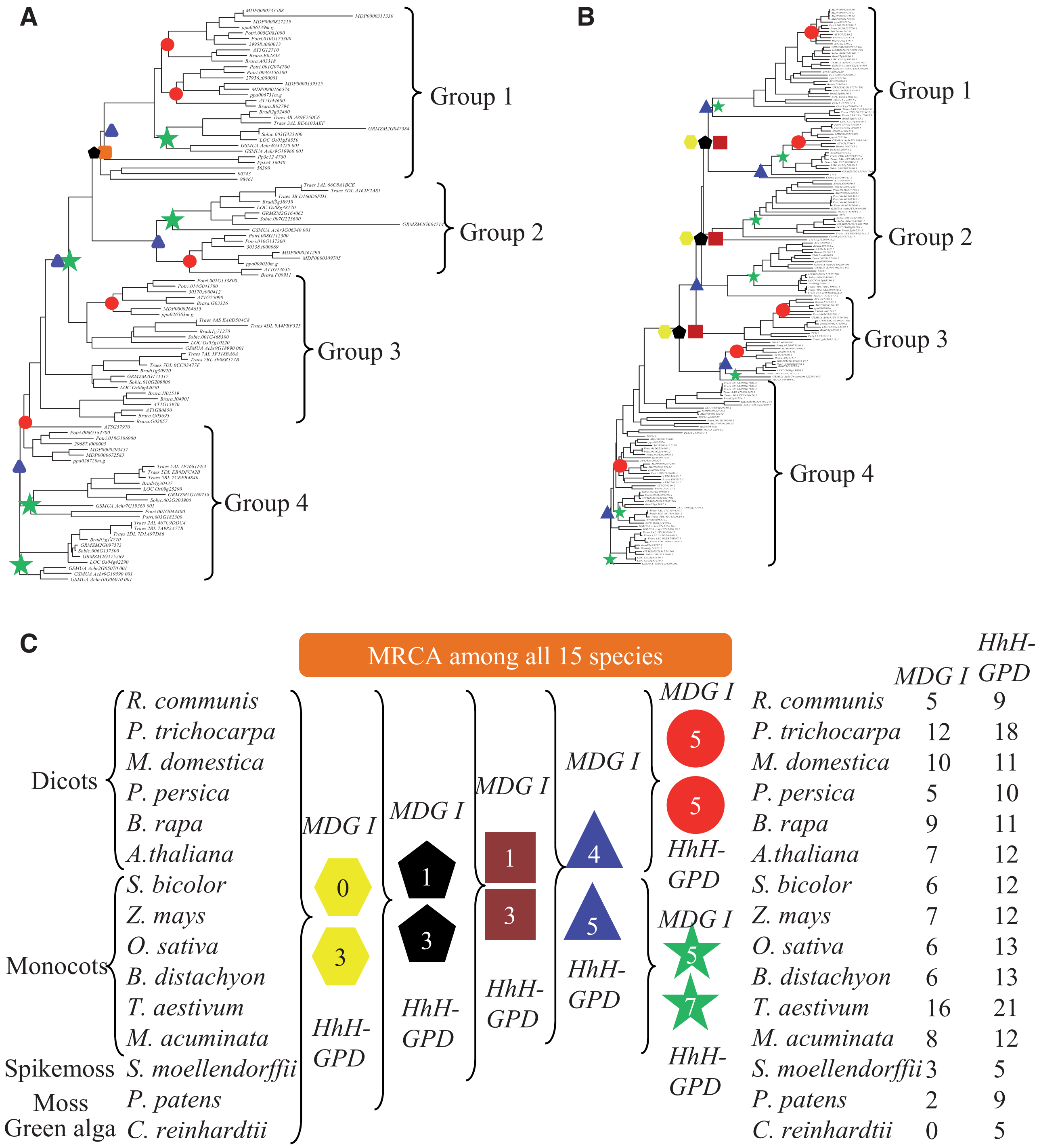
As only one member was identified in each genome for the *MDG II* genes, further investigation was focused on the remaining two gene families including *MDG I* and *HhH-GPD*. To classify the members of these two gene families and to facilitate their functional characterization, we achieved their corresponding protein domain sequences as described in the Materials and Methods and then reconstructed the phylogenetic trees for these two gene families (fig. 1*A* and *B*, supplementary fig. S1*A* and *B*, Supplementary Material online). Both *MDG I* and *HhH-GPD* families could be clustered into four groups. For the *MDG I* family, group 1 was the oldest one as it included all members from 14 species. The remaining three groups consisted of members from both dicot and monocot plants. In contrast, for the *HhH-GPD* family, groups 1, 2, and 3 contained all members from 15 species and group 4 consisted of members from both dicot and monocot plants.

As different species have evolved into different size of families, we further evaluated the patterns of expansion and evolutionary history of these two gene families. We broke down the phylogeny tree into ancestral units according to the method described by Shiu et al. (2004) and then estimated the most recent common ancestor (MRCA) among different species. As the lost genes and pseudogenes were not identified and were excluded for the phylogenetic tree construction, the MRCA members may be underestimated but the analysis could still be used to evaluate evolution histories. We first surveyed the *MDG I* family. As no member was detected in the green algae species *C. reinhardtii*, no MRCA exit among the 15 species as shown by the yellow hexagon (fig. 1*A* and *C*). We have detected only one MRCA among the remaining 14 species as shown by the black pentagon. No MRCA was expanded during the divergence of *Tracheophyta* species from moss (brown squares). Two more members were required during the divergence of dicot and monocot plants from *Lycopodiophyta* (blue triangles). During the divergence between dicot and monocot plants, one additional member was added in the MRCA of either dicot or monocot plants (red circles and green stars). After that, no expansion occurred for some species such as *R. communis* and *P. persica*, or one to three members were required during species divergence for other species such as *S. bicolor*, *A. thaliana*, *B. rapa*, and so on. For the remaining three species (*P. trichocarpa*, *M. domestica*, and *M. acuminate*), relatively higher expansion occurred during their species divergence and these species required double or more numbers of *MDG I* genes.

Different from the *MDG I* family, at least five *HhH-GPD* genes were detected in all 15 species (fig. 1*C*). We have identified three MRCA members among the 15 species. No additional member was required during the divergence of *Tracheophyta* from moss (brown squares) and two more members were added during the divergence of *Euphyllophyta* from *Lycopodiophyta* (blue triangle). During the divergence between dicot and monocot divergence, no other member was required for dicot plants (red circles); however, MRCA of monocots required two additional members (green stars). The large scale of expansion occurred during species divergence for both monocot and dicot plants. As result, 9–23 members of *HhH-GPD* genes have been evolved.

Contributions of Duplication and Transposition to Family Size

Both *MDG I* and *HhH*-GPD families exhibited different expansion histories. To explore the mechanisms of the family expansion, we further surveyed the contributions of both duplication and mobile elements to the family expansion. We have investigated the contribution of tandem, segmental and the whole-genome duplication, transposition, and retrotransposition to the family expansion. We first surveyed the contribution of tandem duplication to the gene expansion. We identified tandemly duplicated genes according to the description in the Materials and Methods. The survey showed that no tandem duplication was detected for the *MDG I* gene family. For the *HhH-GPD* family, only one pair of tandemly duplicated genes was detected in four species. They were *Bradi3g43692* and *Bradi3g43720* from *B. distachyon*, *LOC_Os05g37350* and *LOC_Os05g37410* from *O. sativa*, *Sobic.001G262700* and *Sobic.001G262900* from *S. bicolor*, *Traes_1BL_263DE6AA9* and *Traes_1BL_05EB7AD97* from *T. aestivum*. For the poplar species, the only tandem array was detected, which contained three genes including *Potri.014G187000*, *Potri.014G187300*, and *Potri.014G187500*. No tandemly duplicated genes were detected for the remaining ten species. We also surveyed the contribution of mobile elements to the family expansion. Similarly, for the 15 species, no gene was found to be expanded by mobile elements. Although one rice gene *LOC_Os12g10900* encodes the *HhH-GPD* domain, which might be expanded by a retrotransposon, the gene was annotated as a retrotransposon coding gene. As a result, the gene was excluded in this study. Thus, both tandem



	MDG I	HhH-GPD
R. communis	5	9
P. trichocarpa	12	18
M. domestica	10	11
P. persica	5	10
B. rapa	9	11
A.thaliana	7	12
S. bicolor	6	12
Z. mays	7	12
O. sativa	6	13
B. distachyon	6	13
T. aestivum	16	21
M. acuminata	8	12
S. moellendorffii	3	5
P. patens	2	9
C. reinhardtii	0	5

Fig. 1.—Phylogenetic and evolutionary analysis of the *MDG I* and *HhH-GPD* families. (*A*) and (*B*) Phylogenetic analyses and classification of the *MDG I* and *HhH-GPD* family members, respectively, from 15 species including six monocot (*A. thaliana*, *B. rapa*, *M. domestica*, *P. persica*, *P. trichocarpa*, *R. communis*), six dicot plants (*B. distachyon*, *M. acuminate*, *O. sativa*, *S. bicolor*, *T. aestivum*, *Z. mays*), one spikemoss (*S. moellendorffii*), one moss (*P. patens*), and one green algae (*C. reinhardtii*) species. Domain amino acid sequences from each family were aligned for phylogenetic tree construction using the bootstrap method with a heuristic search of the PAUP 4.0b8 program with 500 bootstrap tries. Ancestral units were defined according to the description from Shiu et al. (2004). Their enlarged phylogenetic trees and their analyses are shown in supplementary figure S1*A* and *B*, Supplementary Material online, respectively. No domain sequence was detected for the MDG I family in green algae. (*C*) Evolutionary history of the *MDG I* and *HhH-GPD* families in 15 organisms. Yellow hexagons represent the MRCA units among all 15 organisms;. black hexagons indicate the MRCA units among flowering plants, spikemoss, and moss; brown squares show the MRCA units among flowering plants and spikemoss. Blue triangles represent the MRCA units among flowering plants. Red circles and green stars show the MRCA units among dicots and monocots, respectively.

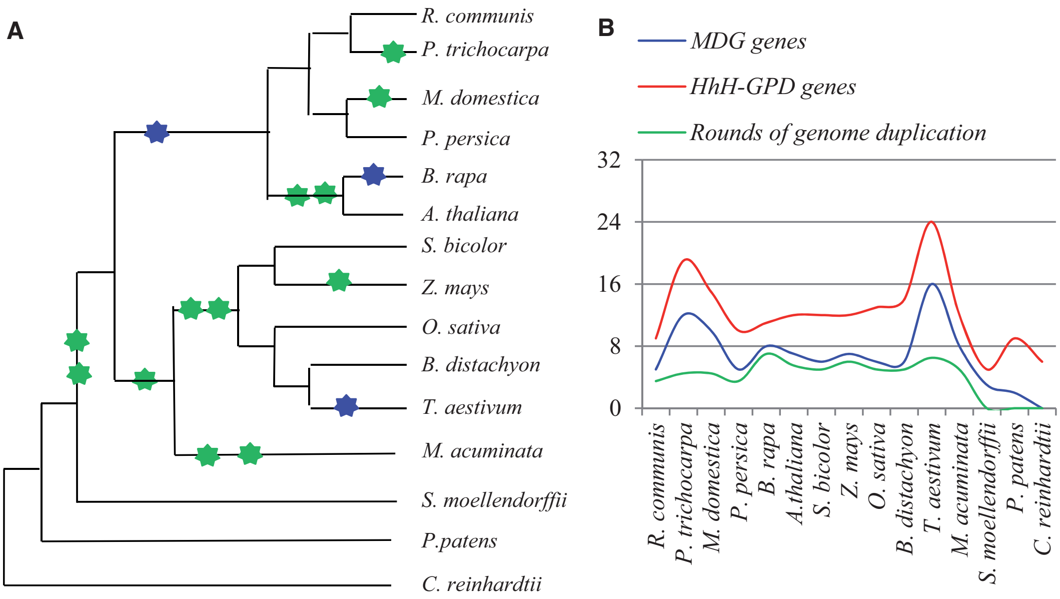


FIG. 2.—The whole-genome duplication history and its effect on gene expansion of the *MDG I/HhH-GPD* family. (*A*) Phylogenetic tree of 15 species and their whole-genome duplication history. This figure was constructed with the related data retrieved from the PGDD database. Green and blue stars indicate whole-genome duplication and triplication, respectively. (*B*) Correlation coefficient analysis of the *MDG I/HhH-GPD* gene family size and rounds of genome duplication in 15 species.

duplication and mobile elements might not be regarded as a major contributor for the family expansion.

We then analyzed the contribution of the whole-genome duplication to the expansion of two gene families. As the studies on the whole-genome duplication events for many species were previously carried out, we collected these data and reconstructed the phylogenetic tree with these 15 species and their paleopolyploidy histories (fig. 2*A*). We then compared the duplication events (green stars for genome doubling or blue stars for tripling in fig. 2*A*) with the size of *MDG I* and *HhH-GPD* families. The comparison implied that the whole-genome duplication events might have contributed to the expansion of these two gene families for some species. For example, *P. trichocarpa* (poplar) underwent one more genome doubling event when compared with *R. communis* (castor) and as a result, the poplar genome encoded more than two times numbers of family members (12 *MDG I* and 18 *HhH-GPD* compared with 5 *MDG I* and 9 *HhH-GPD*, respectively, in the caster). In order to further confirm the contribution of genome duplication to the family expansion, we carried out the co-relationship analysis between rounds of genome duplication and encoded *MDG I/HhH-GPD* genes (fig. 2*B*). The correlation coefficient was calculated as 0.738 ($P < 0.01$) for the *MDG I* family and 0.662 ($P < 0.01$) for the *HhH-GPD* family. The data suggested that the whole-genome duplication significantly contributed to the gene expansion for both *MDG I* and *HhH-GPD* families.

Subsequently, we analyzed the contribution of segmental duplication to the family expansion. For the *MDG I* family, only ten species were selected for such analyses as the remaining five species showed no further expansion during species divergence from MRCA of monocots or dicots. Our data showed that segmental duplication significantly contributed to the expansion of *MDG I* genes in at least seven species including *A. thaliana*, *B. distachyon*, *B. rapa*, *M. acuminate*, *M. domestica*, *P. trichocarpa*, and *Z. mays* (fig. 3*A*). In these seven species, 28.6–100% of *MDG I* genes were located on segmental duplication blocks. In contrast, for the species *O. sativa*, *S. bicolor* and *T. aestivum*, no segmentally duplicated *MDG I* genes were detected. For the *HhH-GPD* gene family, 13 species were selected for segmental duplication analysis as only five *HhH-GPD* genes were identified in the remaining two species including *S. moellendorffii* and *C. reinhardtii*. Among the selected 13 species, segmental duplication significantly contributed to the family expansion in ten species and their contribution rates ranged from 15.4% to 50% (fig. 3*B*). Similar to the *MDG I* family, for three species *O. sativa*, *S. bicolor* and *T. aestivum*, no segmentally duplicated *HhH-GPD* genes were detected. For both *MDG I* and *HhH-GPD* families, up to 100% and 50% of *MDG I* and *HhH-GPD* genes have been involved in segmental duplication, respectively, in the species *P. trichocarpa*. For example, all the 12 *MDG I* genes in the species were located on segmental duplication region (fig. 3*C*). For most of these genes, they segmentally duplicated once. However, some of these genes segmentally duplicated two or three times. For example, the gene *Potri.001G044400* was segmentally related to *Potri.003G182300*, *Potri.018G106900*, and *Potri.006G184700* (fig. 3*C*).

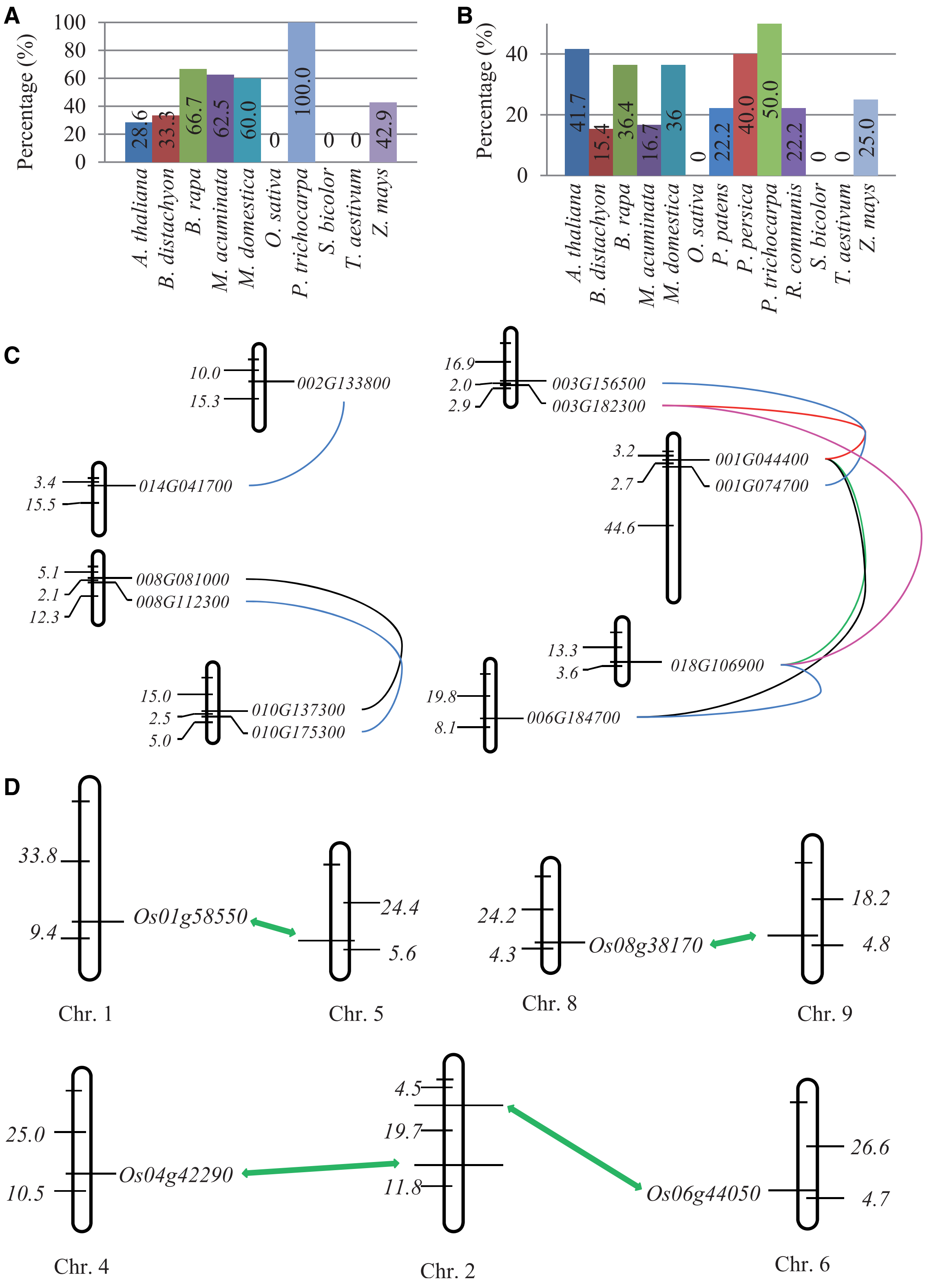


Fig. 3.—The effects of segmental duplication on gene expansion in the *MDG I* and *HhH-GPD* families. (*A*) and (*B*), The contribution of segmental duplication to gene expansion in the *MDG I* and *HhH-GPD* families, respectively. Detection of segmental duplication was carried out only in these species, where gene expansion was observed in either *MDG I* or *HhH-GPD* families. These species were listed in (*A*) for the *MDG I* and (*B*) for *HhH-GPD* family. (*C*) and (*D*), *MDG I* gene expansion by segmental duplication in *P. trichocarpa* and *O. sativa*, respectively. The prefix "*Potri.*" in the locus name in (*C*) was omitted for convenience.

On the other hand, as just mentioned, for some species, no segmentally duplicated gene was detected. However, further analysis showed that this might be due to the gene loss after segmental duplication. For example, in rice, no segmentally duplicated gene was identified but we did detect segmentally duplicated fragments (fig. 3*D*). The gene *LOC_Os01g58550* was detected to be segmentally duplicated and the duplicated fragment was integrated on Chromosome 5. However, no *MDG I* gene was encoded in the duplicated fragment, which might be due to gene loss. Similar situations were also observed for other three *MDG I* genes including *LOC_Os08g38170*, *LOC_Os04g42290*, and *LOC_Os06g44050* (fig. 3*D*).

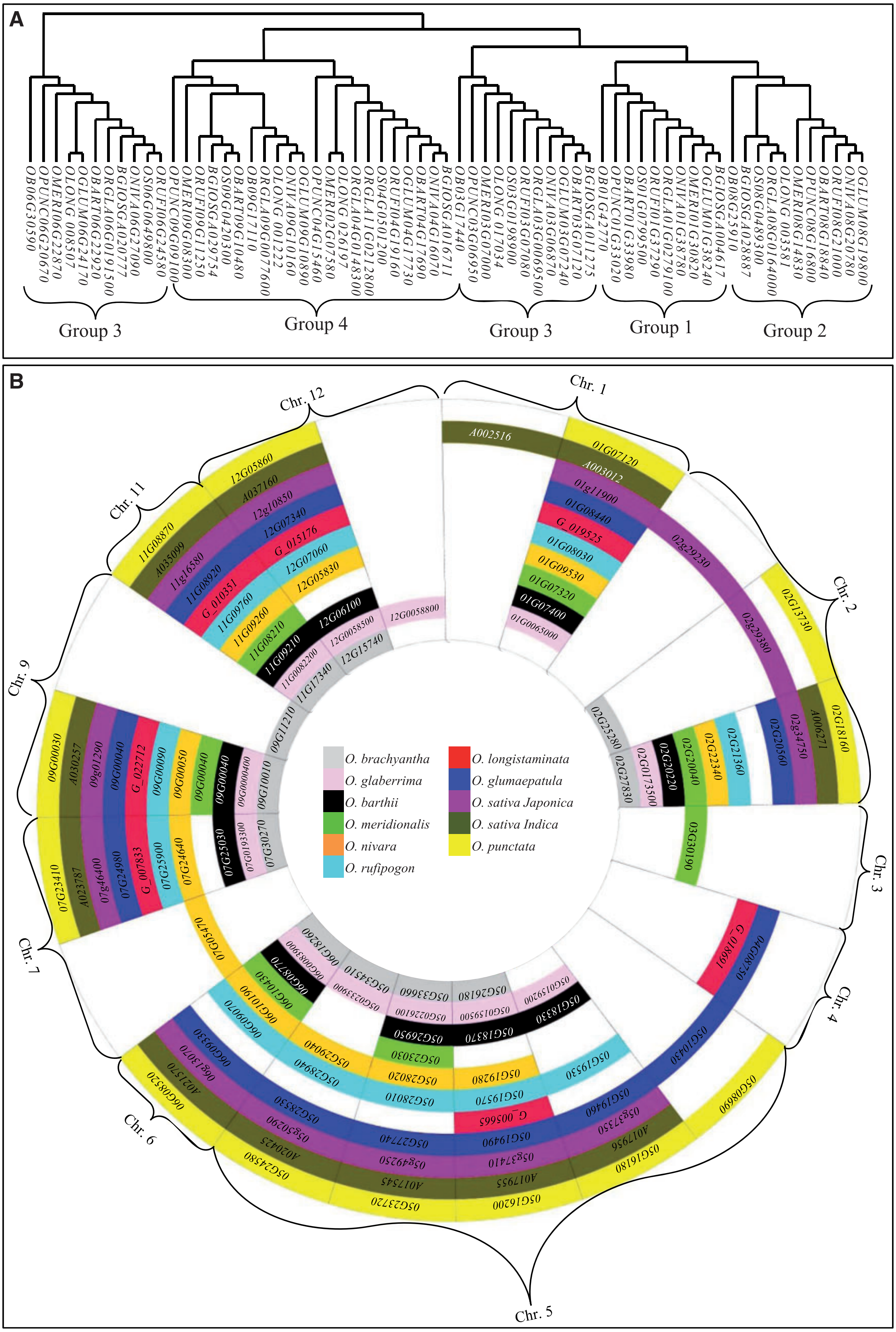
Evolutionary History of *MDG I* and *HhH-GPD* Gene Families during the Divergence from the Rice Genus *Oryza*

We have surveyed the expansion patterns and evolutionary history of these two gene families by analyzing all members from 15 species, which were from different genus. To better understand their evolutionary history, we further identified all family members from 11 rice species/subspecies, which were from the same genus *Oryza*, where all of their genomes have been sequenced. The genome-wide searches showed that these 11 species/subspecies encoded 5–7 members of *MDG I* genes (supplementary table S4, Supplementary Material online). We then constructed the phylogenetic tree using domain region of these members (fig. 4*A*). Similarly, a total of four groups were clustered. In group 1, no member was detected for the species *O. longistaminata*, implying the gene loss and only one member was identified for the remaining ten species. Similarly, group 2 also contained only one member in each species but no gene loss was observed for all species. For both groups 3 and 4, each species usually encodes two members. However, in the species *O. brachyantha*, only one member was clustered into the group 4, suggesting a gene loss event. In contrast, one additional member was required (gene gain) for the species *O. glaberrima*, as a result, three genes were detected in this group. Generally, for the *MDG I* gene family, 11 rice species/subspecies showed similar expansion and evolution history. However, these species exhibited the different patterns of gene gain and loss.

On the contrary, obvious difference in gene expansion was observed in the *HhH-GPD* family among 11 rice species/subspecies. Both genomes *O. longistaminata* and *O. meridionalis* encoded only seven members of the family and the remaining nine genomes encoded 10–13 *HhH-GPD* genes (supplementary table S5, Supplementary Material online). We carried out genome-wide identification of orthologous genes among the 11 rice species/subspecies and presented in the figure 4*B*. A total of 20 orthologous loci have been detected to encode 120 *HhH-GPD* genes in the 11 species/subspecies. Among them, six loci encoded only one gene each without any orthologous gene in other species. They were *BGIOSGA002516* from *O. sativa* indica, *LOC_Os02g29230* from *O. sativa* japonica, *OMERI03G30190* from *O. meridionalis*, *ONIVA07G05470* from *O. nivara*, *OB09G11210* from *O. brachyantha*, and *ORGLA12G0058800* from *O. glaberrima*. Other three orthologous loci encoded 2–3 genes each. For example, three genes *OPUNC02G13730*, *LOC_Os02g29380* and *OB02G25280* were orthologous genes from *O. punctate*, *O. sativa* japonica and *O. brachyantha*, respectively. The remaining 11 orthologous loci encoded at least seven genes each and only two loci contain all 11 orthologous genes from 11 species. The data suggested the significantly differential gene expansion patterns among these 11 species.

Evolution Forces for Both the *MDG I* and *HhH-GPD* Families

As mentioned above, both the *MDG I* and *HhH-GPD* families showed difference in their expansion histories especially within the same *Oryza* genus during the evolution into different species from the same genus. To test whether the divergence was due to reduced purifying selection or increased positive (or diversifying) selection, we evaluated the ratio of nonsynonymous distance (*Ka*) to synonymous distance (*Ks*) of these two families among 15 different species or 11 species/subspecies from the same *Oryza* genus. As we surveyed the *Ka*/*Ks* ratios among different species from single-cell green alga to multiple-cell higher plants, only conserved domain regions were achieved for sequence alignment followed by *Ka*/*Ks* estimation through the SLR program (Massingham and Goldman 2005, Materials and Methods). For the *MDG I* family, the *Ka*/*Ks* ratios among 15 species ranged from 0 to 0.53 with the average ratio at 0.13 (fig. 5*A*). Similarly, the *Ka*/*Ks* ratios among 11 species/subspecies from the same *Oryza* genus ranged from 0 to 0.69 with the average ratio at 0.15 (fig. 5*B*). Thus, no significant difference was observed for the *MDG I* family between these two sets of data analysis. This result suggested the relatively consistent selection force under purifying selection during the long evolutionary history. We then analyzed the *Ka*/*Ks* ratios for the *HhH-GPD* family among the 15 species. The ratios ranged from 0 to 0.68 with the average ratio at 0.13 (fig. 5*C*). The *Ka*/*Ks* distribution was similar to those from the *MDG I* family. Their divergence was subjected to purifying selection. Similar results were observed for the *Ka*/*Ks* analysis among 11 rice species/subspecies for this gene family (fig. 5*D*). We further extended our analysis to the nondomain regions for the two families. For the *MDG I* family, many gaps were found during alignments among 15 species or 11 rice species/subspecies and the alignments were not suitable for *Ka*/*Ks* analyses. For the *HhH-GPD* family, many gaps were also found in the alignment among 15 species. However, the alignment among 11 rice species/subspecies was suitable for *Ka*/*Ks* analysis. Interestingly, positively selected sites were detected among 11 species/subspecies from the


A
OB06G30590
OPUNC06G20670
OMERI06G22870
OLONG 008507
OGLUM06G24170
OBART06G22920
ORGLA06G01915 00
BGIOSGA020777
ONIVA06G27090
OS06G0649800
ORUFI06G24580
OPUNC09G09100
OMERI09G08300
ORUFI09G11250
BGIOSGA029754
OS09G0420300
OBART09G10480
OB09G17110
ORGLA09G0077600
OLONG 001222
ONIVA09G10160
OGLUM09G10890
OPUNC04G15460
OMERI02G07580
OLONG 026197
ORGLA04G0148300
ORGLA11G0212800
OS04G0501200
ORUFI04G19160
OGLUM04G17730
OBART04G17690
ONIVA04G16070
BGIOSGA016711
OB03G17440
OPUNC03G06950
OMERI03G07000
OLONG 017034
OS03G0198900
ORUFI03G07080
ORGLA03G0069500
ONIVA03G06870
OGLUM03G07240
OBART03G07120
BGIOSGA011275
OB01G42770
OPUNC01G33020
OBART01G33980
OS01G0799500
ORUFI01G37290
ORGLA01G0279100
ONIVA01G38780
OMERI01G30820
OGLUM01G38240
BGIOSGA004617
OB08G25910
BGIOSGA028887
OS08G0489300
ORGLA08G0164000
OLONG 003581
OMERI08G14830
OPUNC08G16800
OBART08G18840
ORUFI08G21000
ONIVA08G20780
OGLUM08G19800
Group 3
Group 4
Group 3
Group 1
Group 2
B
Chr. 1
Chr. 2
Chr. 3
Chr. 4
Chr. 5
Chr. 6
Chr. 7
Chr. 9
Chr. 11
Chr. 12
A002516
01G07120
A003012
01g11900
01G08440
G_019525
01G08030
01G09530
01G07320
01G07400
01G0065000
02g29230
02g29380
02G13730
02G18160
A006271
02g34750
02G20560
02G21360
02G22340
02G20040
02G20220
02G0173500
02G27830
02G25280
03G30190
04G08750
G_018691
05G08690
05G16180
A017956
05g37350
05G19460
05G10430
05G19530
05G18330
05G0159200
05G16200
A017955
05g37410
05G19490
G_005665
05G19570
05G19280
05G18370
05G0159500
05G26180
05G23720
A017545
05g49250
05G27740
05G28010
05G28020
05G23030
05G26950
05G0226100
05G33660
05G0233900
05G34510
05G24580
A020425
05g50290
05G28830
05G28940
05G29040
06G08520
A021570
06g13070
06G09330
06G09070
06G10190
06G10430
06G08770
06G0083900
06G18260
07G05470
07G23410
A023787
07g46400
07G24980
G_007833
07G25900
07G24640
07G25030
07G0193300
07G30270
09G00030
A030257
09g01290
09G00040
G_022712
09G00090
09G00050
09G00040
09G00040
09G0000400
09G10010
09G11210
11G08870
A035099
11g16580
11G08920
G_010351
11G09760
11G09260
11G08210
11G09210
11G0082200
11G17340
12G05860
A037160
12g10850
12G07340
G_015176
12G07060
12G05830
12G06100
12G0058500
12G15740
12G0058800
O. brachyantha
O. glaberrima
O. barthii
O. meridionalis
O. nivara
O. rufipogon
O. longistaminata
O. glumaepatula
O. sativa Japonica
O. sativa Indica
O. punctata

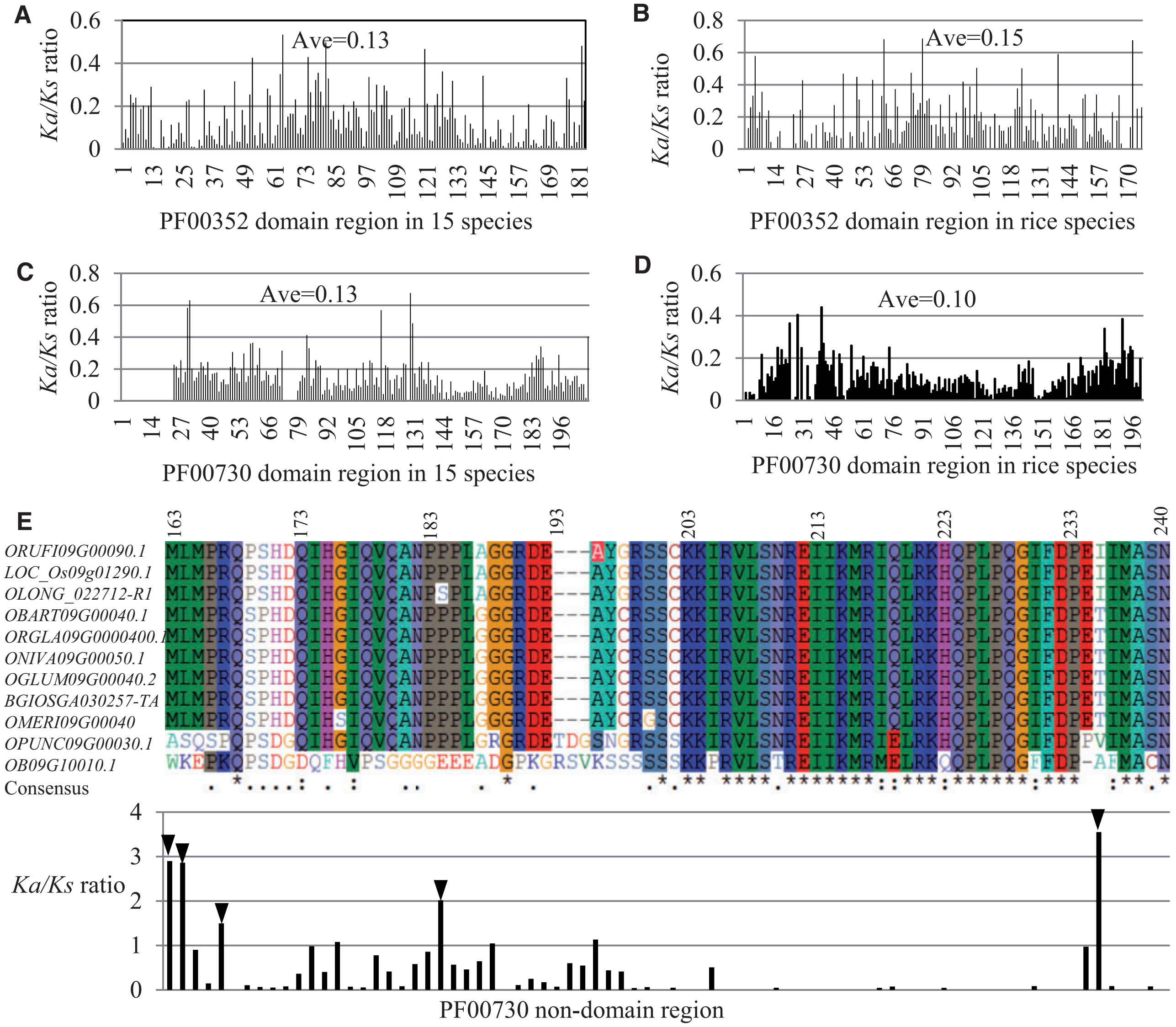


Fig. 5.—Frequency distributions of *Ka*/*Ks* ratios in domain regions of *MDG I* and *HhH-GPD* members and tests of sites with purifying/positive selection in the *HhH-GPD* family. (*A*) and (*B*) Frequency distributions of *Ka*/*Ks* ratios were calculated with *MDG I* domain regions from 15 species and 10 rice species, respectively. (*C*) and (*D*) *HhH-GPD* domain regions from 15 species and 10 rice species, respectively, were used for *Ka*/*Ks* calculation. The average (ave) *Ka*/*Ks* ratios were also calculated in (*A*)–(*D*). (*E*) Screening for amino acid sites with purifying/positive selection in nondomain regions of the *HhH-GPD* family by the SLR program as described in the Materials and Methods section. Sites under likely positive selection with *Ka*/*Ks* >1.0 by statistical analysis were marked with inverted triangles.

Oryza genus. We have detected a total of seven sites with positive selection. Figure 5*E* showed the five sites of them with *Ka*/*Ks* ratios ranging from 1.50 to 3.55. We also detected several positive positions with gapped residues during alignment, which were presented in 10 out of 11 sequences and with *Ka*/*Ks* ratio larger than 1. However, these positions were not regarded as positively selected positions as the alignment gaps might result in statistical bias or artifacts. Thus, we

Fig. 4.—Continued

Fig. 4.—The *MDG I*/*HhH-GPD* families in the *Oryza* genus and their phylogenetic/collinear analysis. (*A*) Phylogenetic tree of the *MDG I* family members from 11 rice species/subspecies and their classification. (*B*) Syntenic analysis of orthologous genes of the *HhH-GPD* family among ten species of the *Oryza* genus. The coordinate mapping was carried out using the GenomeRing program (Herbig et al. 2012). The locus prefixes in each species were omitted. These prefixes include "*OB*" for *O. barthii*, "*OBART*" for *O. brachyantha*, "*ORGLA*" for *O. glaberrima*, "*OGLUM*" for *O. glumaepatula*, "*OLON*" for *O. longistaminata*, "*OMERI*" for *O. meridionalis*, "*ONIVA*" for O. *nivara*, "*OPUNC*" for *O. punctata*, "*ORUFI*" for *O. rufipogon*, "*BGIOSG*" for *O. sativa* indica, "*LOC_Os*" for *O. sativa* japonica.

have detected a total of seven sites in the nondomain region under positive selection. The data suggested the different selection forces between domain and nondomain regions and also suggested the role of positive selection in the species divergence of the *HhH-GPD* family for the *Oryza* genus.

Expression Profiling of Both *MDG I* and *HhH-GPD* Families in the Whole Life Stages of Rice Development

We surveyed the expression patterns of 6 rice *MDG I* and 13 *HhH-GPD* genes among 48 rice samples from 12 different tissues including leaf blade, lead sheath, root, stem, inflorescence, anther, pistil, lemma, palea, ovary, embryo, and endosperm (fig. 6). We first examined the difference in transcript abundance among 48 different samples for each gene. The expression level in the sample LeafBlade_27DAT_12:00 was set as 0 (log2 value) for all genes and the relative mRNA amount in the remaining samples was calculated by comparing with the standard. Such analyses showed that no gene was evenly expressed among tested tissues and no tissue-specific gene was detected (fig. 6*A*). Even in the same tissue, differential expression was observed among different developmental stages. For example, the gene *LOC_Os01g58550* showed the higher expression in 27-day-old leaf sheath when compared with that in 76-day-old leaf sheath. The data suggested that both families should play roles in multiple tissues and developmental stages. On the other hand, we observed that some of genes showed significantly higher expression abundance in nonleaf tissues, for example, *LOC_Os01g58550* and *LOC_Os06g13070*. Others showed higher expression in leaf tissues such as *LOC_Os08g38170* and *LOC_Os11g16580*. In general, both families exhibited diverse expression patterns among multiple tissues.

We then compared the expression level among different genes in each sample. By comparing the average expression level among a total of analyzed 19 genes, we selected *LOC_Os09g01290* as a control gene to measure the relative expression abundance for the remaining genes. Our analyses showed that either *MDG I* or *HhH-GPD* genes distinguished themselves from other genes in their expression level in one or more tissues (fig. 6*B*). All genes exhibited no similar expression abundance each other (fig. 6*B*). Generally, most of *MDG I* genes exhibited higher expression level than those in *HhH-GPD* genes. Some of *HhH-GPD* genes, for example, both *LOC_Os02g29230* and *LOC_Os05g37410*, exhibited very low expression level in multiple tissues.

Expression Regulation of Both *MDG I* and *HhH-GPD* Genes under Abiotic and Biotic Stresses

To explore whether these 6 rice *MDG I* and 13 *HhH-GPD* genes were regulated by various abiotic and biotic stresses, we analyzed their expression patterns under three different abiotic stresses including drought, high salinity and cold stresses as well as two different pathogens (fig. 7). We first investigated the expression profiles under various abiotic stresses. Our data showed that two out of six *MDG I* genes, *LOC_Os01g58550* and *LOC_Os03g10220*, were significantly upregulated by drought stress (fig. 7*A*). However, the gene *LOC_Os08g38170* was downregulated by drought stress (fig. 7*A*). The remaining three *MDG I* genes were not regulated in their expression by any of three tested abiotic stresses (fig. 7*A*). Interestingly, no *MDG* gene was regulated by both high salinity and cold stresses. The data might suggest that some of *MDG* genes might play a specific role in response to drought stress. On the contrary, no *HhH-GPD* gene was upregulated by any of three abiotic stresses (fig. 7*A*). Interestingly, we detected a total of three *HhH-GPD* genes with downregulation under drought stress. These genes included *LOC_Os02g29230*, *LOC_Os05g49250*, and *LOC_Os12g10850*. Among them, the gene *LOC_Os12g10850* was also downregulated by high salinity stress (fig. 7*A*).

We then surveyed the expression profile after the inoculation of the blast fungus *Magnaporthe oryzae*. Three Nipponbare (NB) lines carrying the blast resistance genes *Pia* and *Pish* were designated as NB (*Pia*/*Pish*), NB (*Pish*), and NB (*ΔPish*). They were inoculated with two strains P91-15B (harboring *AVR-Pia*) and Kyu77-07A (harboring *AVR-Pish*). Among six *MDG I* genes, the gene *LOC_Os03g10220* was not regulated by both pathogens and the remaining five genes (80%) were downregulated by compatible or incompatible pathogens (fig. 7*B*). The gene *LOC_Os01g58550* was downregulated by two pathogen strains in all three rice lines. Three genes *LOC_Os04g42290*, *LOC_Os06g44050*, and *LOC_Os08g38170* were not in response to the pathogen P91-15B in the line NB (*Pia*/*Pish*) but were downregulated by either P91-15B or Kyu77-07A in the remaining two lines NB (*Pish*) and NB (*ΔPish*). The remaining one *MDG I* gene *LOC_Os09g25290* was downregulated only 5 days postinoculation of the pathogen Kyu77-07A in the line NB (*Pish*). Among 19 *HhH-GPD* genes, eight of them (42%) showed the response to pathogens (fig. 7*B*). We found one gene *LOC_Os12g10850* was upregulated by both pathogens in all three inoculated lines. On the contrary, the gene *LOC_Os11g16580* was downregulated by both pathogens in all three inoculated lines. The remaining six genes were all downregulated by the pathogen Kyu77-07A in NB (*Pish*)/NB (Δ*Pish*) or both lines.

To investigate the expression profile of both *MDG I* and *HhH-GPD* genes in response to the bacterium pathogen *Xanthomonas oryzae* pv. *oryzae* (*Xoo*), the wild-type (WT) strain T-7114R or mutated strain *ΔhrcV* in type III secretion (T3S) system was used for inoculation. Among the six *MDG I* genes, three of them were regulated only by the WT pathogen. The gene *LOC_Os01g58550* was upregulated only after 4 or 6 days of inoculation; the remaining two genes were downregulated after the same stages of inoculation (fig. 7*C*). For the *HhH-GPD* gene family, only two genes were

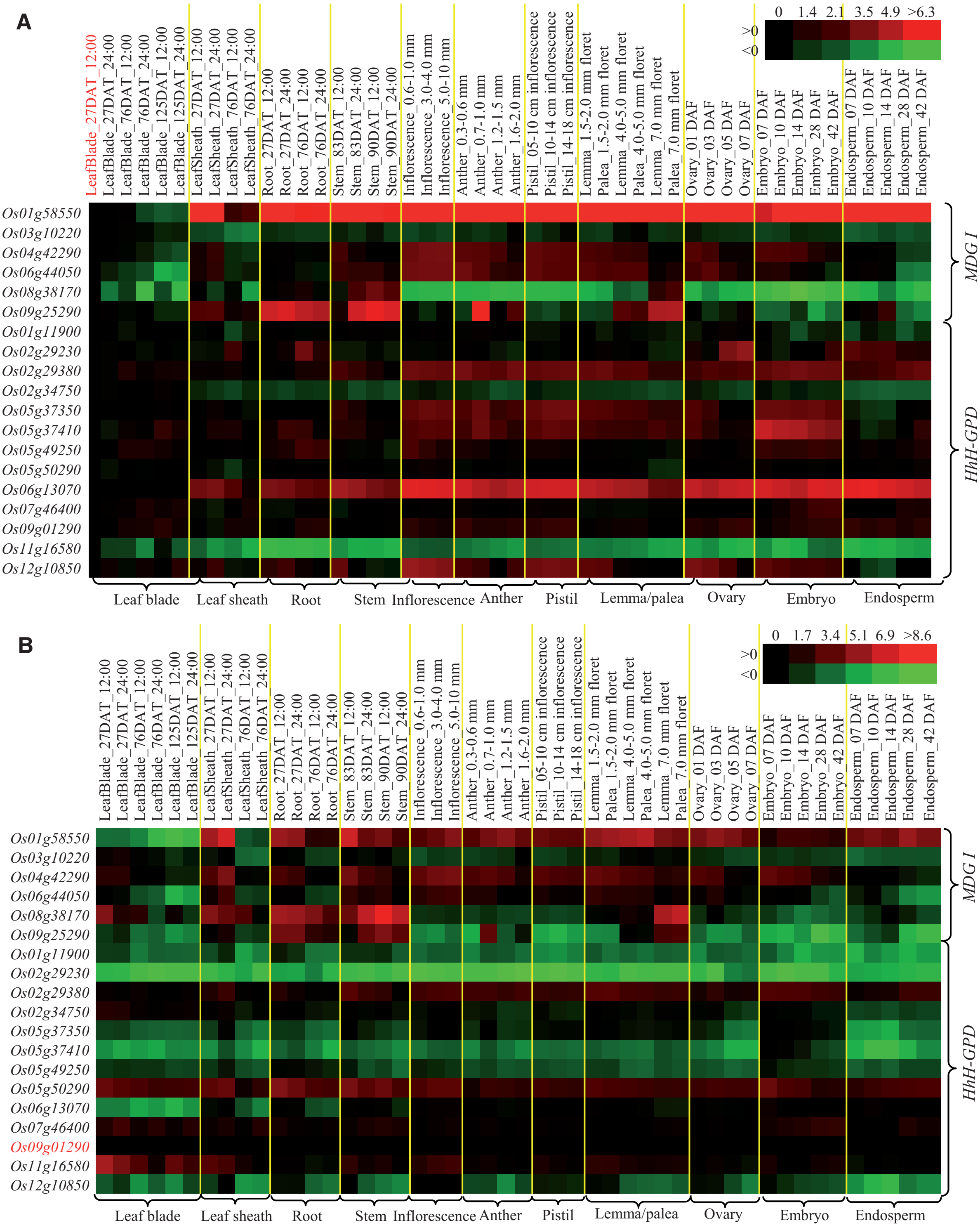


FIG. 6.—Spatiotemporal expression profile of 6 rice *MDG I* and 13 *HhH-GPD* genes as shown by heat map. (*A*) Relative expression level among 48 different developmental stages of samples from 11 tissues in each gene. The processed microarray expression value of the sample LeafBlade_27DAT_12:00 (labeled as red fonts) in each gene was set as control and the expression level in the remaining samples was calculated by comparing with the control. (*B*) Comparison of expression abundance among different genes in each sample. The gene *LOC_Os09g01290* with moderate transcript abundance (highlighted in red fonts) was set as control and the expression level in the remaining genes was calculated by comparing with the control. The values in (*A*) and (*B*) were then converted into $\log_2$ scale for heat mapping using the TreeView program (Eisen et al. 1998).

observed in response to these two pathogens (fig. 7*C*). The gene *LOC_Os02g29230* was downregulated by either T-7114R or ΔhrcV after 12 h inoculated. Another gene *LOC_Os11g16580* was downregulated by T-7114R after 6 days of inoculation.

Both Rice *MDG I* and *HhH-GPD* Genes Were Down- or upregulated by Some Hormone Treatments

As some of these *MDG I* and *HhH-GPD* genes showed abiotic/biotic stress-regulated expression profile, we are interested in their responses to various plant hormones. A total of six hormones were investigated and they were abscisic acid (ABA), brassinosteroid, gibberellin, auxin, jasmonic acid (JA), and cytokinin. We first focused on the *MDG I* gene family. Our data showed that six *MDG I* genes showed the difference between roots and shoots in response to plant hormones (fig. 8). In shoots, only one gene *LOC_Os01g58550* showed downregulation under the hormone auxin after 1-h treatment (fig. 8*A*). No other genes showed regulated expression under the remaining five phytohormones. However, in roots, four out of six genes were regulated by ABA, auxin or JA. The gene *LOC_Os01g58550* was downregulated by ABA after 3–6 h of treatments and it was also downregulated by JA after 30 min to 6 h of treatments. For the gene *LOC_Os06g44050*, it was only upregulated by auxin during 1- and 3-h treatments and no significant difference in its expression abundance was observed under other hormone treatments. The gene *LOC_Os08g38170* was downregulated by two hormones including ABA and JA. The gene *LOC_Os09g25290* was also downregulated by two hormones, which were ABA and auxin. Thus, a total of three genes were downregulated by ABA and two of them were also downregulated by JA.

For the *HhH-GPD* family members, they also exhibited obviously different expression patterns between shoots and roots under various hormone treatments. In hormone-treated shoots, three genes were downregulated by hormone treatments. One of them is *LOC_Os02g29230*, whose expression was downregulated by JA with the highest expression level after 12-h treatment (fig. 8*A*). *LOC_Os05g37410* was downregulated by three hormones including ABA, gibberellins, and auxin (fig. 8*A*). The remaining one is *LOC_Os12g10850*, which was downregulated by both ABA and JA (fig. 8*A*). In hormone-treated roots, we have detected four *HhH-GPD* genes and all of them were downregulated by hormones. The gene *LOC_Os02g34750* was not regulated by any hormone treatment in shoots but was downregulated by JA in roots (fig. 8*A* and *B*). *LOC_Os05g37410* was downregulated by ABA and JA in shoots but was downregulated by brassinosteroid and JA in roots (fig. 8*A* and *B*). For the gene *LOC_Os12g10850*, in both shoots and roots, it was downregulated by both ABA and JA and shoots responded more rapidly than roots (fig. 8*A* and *B*).

Root Expression Profiles at Different Tissue Types

To further evaluate the expression profiles of these two gene families, we examined the root expression specificity at cellular level (fig. 9). The 10-day-old crown roots were separated into eight different sections as indicated in figure 9*A*. Total RNA samples from these eight sections were submitted for expression analyses. Among six *MDG I* genes, two of them (*LOC_Os03g10220* and *LOC_Os06g44050*) showed very low expression level in these eight different samples and were omitted for further analysis. The remaining four genes exhibited obvious expression diversity (fig. 9*A*). The gene *LOC_Os01g58550* was mainly expressed in root cap, division, and elongation zones; *LOC_Os04g42290* was mainly in elongation zone and maturation zone I; *LOC_Os08g38170* was mainly in maturation zone I; and the gene *LOC_Os09g25290* was mainly expressed in both elongation zone and maturation zone I. Low expression level was observed for all tested four *MDG I* genes. For the *HhH-GPD* family, a total of 7 out of 13 genes showed very low expression level in the eight different root sections and were omitted for further analysis. The remaining six genes also exhibited diverse expression patterns (fig. 9*A*). All these genes showed the difference in their expression profiles either in expression abundance or in root cell types. For example, both *LOC_Os02g29380* and *LOC_Os11g16580* showed similar expression patterns but exhibited different abundance in maturation zones IV and V.

We further analyzed their expression specificity among three different cell types including epidermis/exodermis/sclerenchyma, cortex, and endodermis/pericycle/stele (fig. 9*B*). These RNA samples were isolated from either maturation zone V for both epidermis/exodermis/sclerenchyma and endodermis/pericycle/stele or between elongation zone and maturation zone I for all three cell types. Generally, for most of two family genes, they were mainly expressed in endodermis/pericycle/stele with higher expression level at elongation zone and maturation zone I when compared with the maturation zone V. Although these genes showed similar expression patterns at the root zones, they exhibited the difference in their expression abundance. Interestingly, the gene *LOC_Os08g38170* exhibited distinct difference from the remaining genes, where it was mainly expressed at epidermis/exodermis/sclerenchyma.

Discussion

Evolutionary Origins of Genes Encoding DNA Glycosylases

In this study, we have genome-widely identified three gene families in 15 sequenced genomes. We have also identified these families in 11 rice species/subspecies belonging to the same *Oryza* genus. The investigation showed that these gene families varied in family size and did not ubiquitously exist in all analyzed organisms. Here we further surveyed their distribution in additional 35 species including 2 moss, 6 algae and 27

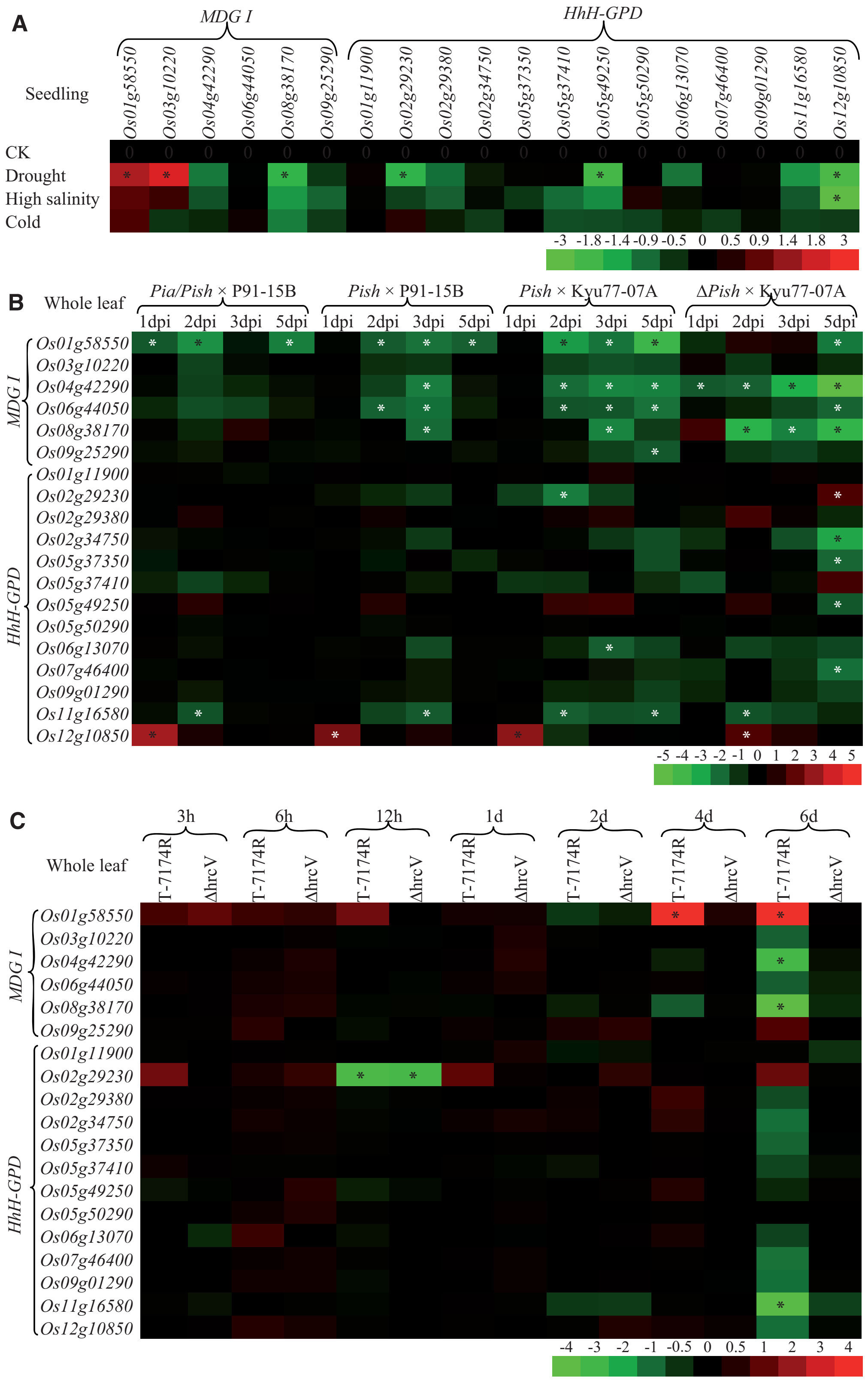


FIG. 7.—Expression profiling of rice *MDG I* and *HhH-GPD* genes under various abiotic and biotic stresses. (*A*) Heat map showing expression patterns of 6 *MDG I* and 13 *HhH-GPD* genes under drought, high salinity and cold stresses. The processed expression values were calculated from three biological repeats and were then converted into $\log_2$ scale with water stress as control. (*B*) Heat map showing expression regulation of 6 *MDG I* and 13 *HhH-GPD* genes by

monocot/dicot plant species, whose whole-genome sequences are available (supplementary tables S6 and S7, Supplementary Material online, for the *MDG I* and *HhH-GPD* families, respectively). Such a survey showed that the *HhH-GPD* family existed in all tested genomes and no *MDG I* or *II* gene was detected in all six green algae genomes including *C. reinhardtii*, *Coccomyxa subellipsoidea* C-169, *Micromonas pusilla* CCMP1545, *Micromonas* sp. RCC299, *Ostreococcus lucimarinus*, and *Volvox carteri*. Furthermore, we detected the distribution of these three gene families by BLAST (Basic Local Alignment Search Tool) searches against all available protein sequences deposited in both Pfam and Interpro databases. Based on the searches, the *HhH-GPD* family generally presents in all types of organisms including higher plant, moss, algae, animal, fungi, bacteria, archaea, and virus (supplementary table S8, Supplementary Material online). The *MDG II* genes were detected all organisms except for green algae and generally only one *MDG II* gene was encoded in each genome. Interestingly, green algae genomes also do not encode any *MDG I* gene and this family gene does also not present in animals and virus. No evidence showed why neither *MDG I* or *MDG II* was required for green algae. However, evidence showed that the green alga *Chlamydomonas* has the most unusual pattern of methylation (Feng et al. 2010). Thus, our data might provide some implication underlying the unusual methylation in the green algae.

Although the *MDG I* gene family was detected in higher plants, mosses, red algae, fungi, bacteria and archaea, it presents in only one or a few species of mosses, red algae, fungi and archaea. For example, only one sequence from either fungi or red algae was detected to encode this family protein. Thus, the gene family only ubiquitously exists in both higher plants and bacteria. Similar situation was also observed in the *MDG II* gene family, which is ubiquitous in animals, plants, and bacteria. Different from both *MDG I* and *II* families, the *HhH-GPD* family ubiquitously presents in most of organisms. On the other hand, no expansion was observed for the *MDG II* gene for all tested genomes. For the remaining two gene families *MDG I* and *HhH-GPD*, a genome from archaea, bacterium or fungus usually encodes only one member in each family and higher plant genomes encode various sizes of family members. At the early stage of evolutionary history, very low expansion occurred and a genome generally encodes one or a few members of these two gene families. This situation continuously existed until the divergence between monocots and dicots (fig. 1). A large scale of expansion occurred only for some species during or after the divergence from one species to another in monocot and dicot plants. As a result, different species encode different sizes of family members ranging from 5 to 16 for *MDG I* and from 9 to 21 for *HhH-GPD* families. However, in algae or moss, less expansion occurred during long evolutionary history. Thus, our data showed that these three gene families have gone through different origin and evolutionary histories.

Gene Expansion and Inside Mechanisms in the *MDG I* and *HhH-GPD* Families

In this study, we investigated a total of three families on a genome-wide level. Among them, the *MDG II* family contains only one member in all tested genomes. No expansion occurred during a long evolutionary history. For the remaining two families *MDG I* and *HhH-GPD*, they exhibited both similarity and difference in their family expansion. Before the divergence between monocot and dicot plants, two gene families experienced very low expansion and their MRCA genome encoded only four (for *MDG I*) and five (for *HhH-GPD*) genes. After the divergence between dicots and monocots, only one (for *MDG I*) or two (for *HhH-GPD*) more members were required for their ancestor species in this evolutionary stage. A relatively large scale of expansion of the *MDG I* family occurred during species divergence for some species of dicot and monocot plants (fig. 1). For some species, for example, *R. communis* and *P. persica*, no expansion was observed. However, for the *HhH-GPD* family, all genomes of monocot and dicot plants experienced a large scale of expansion during species divergence. These data suggested the recently expansion events for these two gene families. In addition, our data showed that the gene expansion and loss in the *HhH-GPD* family were also detected during the divergence of species from the same genus. We have investigated the *HhH-GPD* family in 11 different species/subspecies from the *Oryza* genus and the data showed that these species exhibited the difference in gene expansion and loss (fig. 4*B*). Furthermore, differentiated gene expansion and loss events were observed between indica and japonica rice genomes for the *HhH-GPD* family (fig. 4*B*). For example no japonica ortholog was detected for the indica member *BGIOSGA002516* (fig. 4*B*). Similarly, no indica orthologs were found for the japonica members *LOC_Os02g29230* and *LOC_Os02g29380* (fig. 4*B*).

To explore the mechanisms of gene expansion in these two families, we have investigated the contributions of multiple

Fig. 7.—Continued
fungus pathogen strains P91-15B (*AVR-Pia*) and Kyu77-07A (*AVR-Pish*) in the whole rice leaf at 4-leaf stage. (*C*) Heat map showing expression regulation of 6 *MDG I* and 13 *HhH-GPD* genes by bacterium pathogen strains T7114R, a WT strain, and or Δ*?rcV*, a mutant deficient in T3S system in the whole rice leaf at 42-day stage. In (*B*) and (*C*), signal intensity data were based on 75 percentile normalization and $\log_2$ transformation with the average relative value to control treatment (pathogen/H_2O). The star "*" in (*A*) and (*B*) indicated the genes with at least two times difference in their processed signal value based on computing geometric mean after treatment and showing significant difference by Student *t*-test at $P < 0.05$.

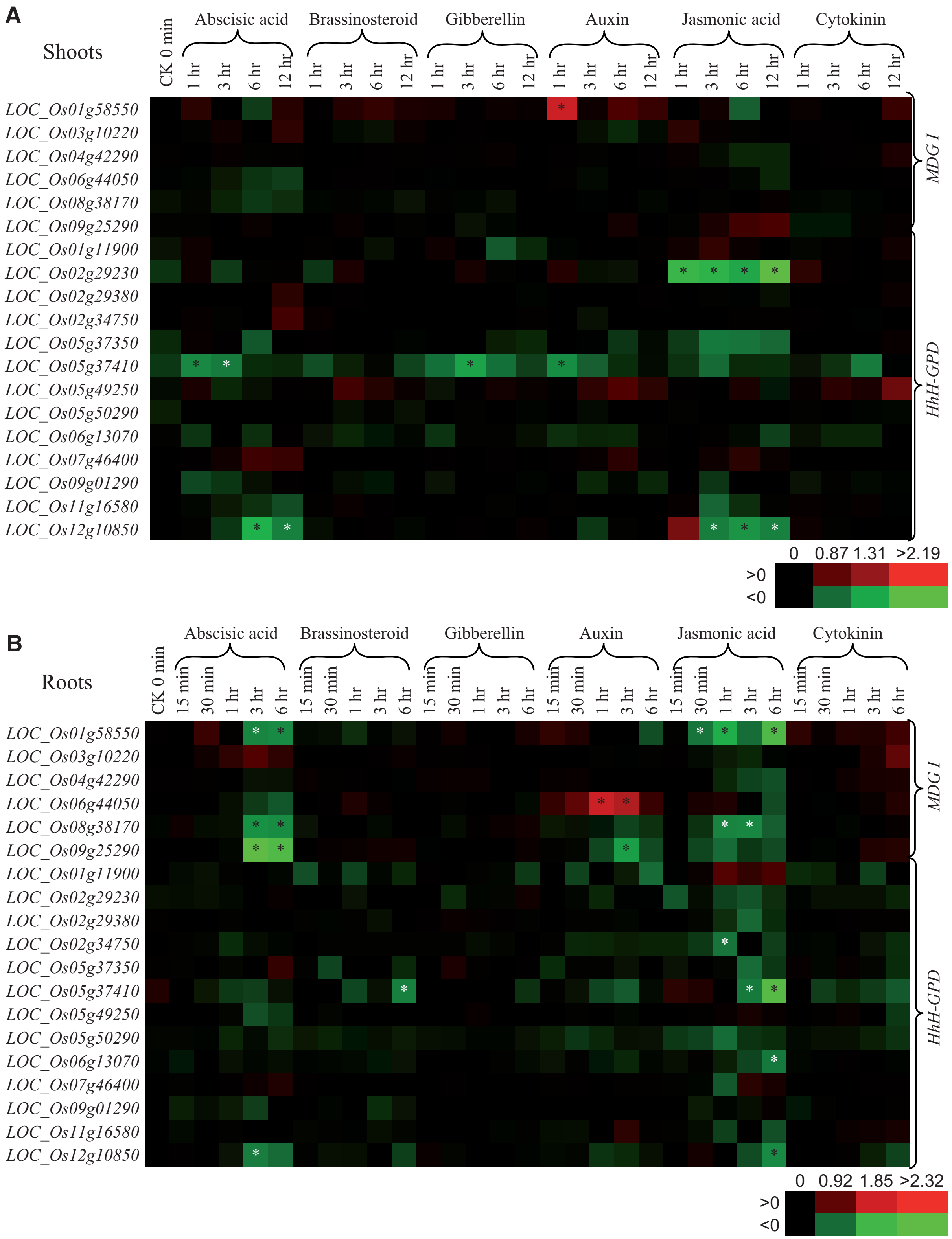


Fig. 8.—Expression regulation of rice *MDG I* and *HhH-GPD* genes by various hormone treatment. (*A*) and (*B*) showed the expression data in shoots and roots, respectively. The processed expression values were calculated from two or three biological repeats and were then converted into $\log_2$ scale with water stress as control. The star "*" in (*A*) and (*B*) indicated the genes with significant difference between control and hormone treatment with the standard as shown in figure 7.

DNA/RNA duplication events to the family expansion. Previous studies showed that genome-wide duplication significantly contributed to gene expansion (Meyer and Van de Peer 2003). Our data showed that rounds of genome duplication were co-related to either *MDG I* or *HhH-GPD* family size in some species, suggesting the contribution of the whole-genome duplication to the family expansion. In addition to the whole-genome duplication, we have also surveyed the contribution of both tandem and segmental duplications to these two family expansions as previous data showed the contribution of tandem and segmental duplication to gene family expansion in some species (Flagel and Wendel 2009; Freeling 2009). Our survey showed that segmental duplication significantly contributed to the expansion of both *MDG I* and *HhH-GPD* families in most of species (fig. 3*A* and *B*). However, tandemly duplicated genes were observed only in a few species, suggesting a limited contribution to the gene expansion. Additionally, we have also examined the contribution of DNA mobile elements to the expansion of *MDG I* and *HhH-GPD* genes. We have examined the presence of various mobile elements in the flanking genomic sequences of the 50 kb upstream and downstream of *MDG I/HhH-GPD* genes in different genomes. We have identified *LTR* (long terminal repeat)-retrotransposons, *MULE*, *hAT*, *CACTA*, *Helitron*, and retrogene in these flanking genomic sequences. However, our detailed analysis showed that no gene was expanded by any mobile element in most of species, suggesting the limited contribution of transposons/retrotransposons to the family expansion. Generally, we have investigated multiple molecular mechanisms for the family expansion and our data showed that both the whole-genome duplication and segmental duplication significantly contributed to the expansion of these two gene families.

Evidence showed that gene duplications occurred frequently; however, gene loss was also frequently observed during long evolutionary history due to redundant functions (Lynch and Conery 2000; Flagel and Wendel 2009; Freeling 2009). In this study, we have detected gene duplication by the-whole genome duplication and segmental duplication (figs. 2 and 3). We have also detected gene loss (fig. 3*D*) in the rice *MDG I* family. The fact might explain why the size of the *MDG I* family was smaller than that of the *HhH-GPD* family in the same genome. After gene expansion, duplicated genes might evolve into new genes with subfunctions/novel functions and they might also become pseudogenes due to redundant functions. The *Ka/Ks* analysis among different species showed that the ratios were low and these two families were under purifying selection. The newly born genes might be retained with the similar or subfunctions and they might also be survived by expression divergence. We have analyzed the expression patterns of these two rice gene families among different tissues and developmental stages (figs. 6 and 9). We have also investigated the expression regulation under various abiotic/biotic stresses or hormone treatments (figs. 7 and 8). Such analyses showed that no gene within the same gene family exhibited the same expression abundance or patterns. One gene differentiates from others by either expression abundance or patterns. Besides rice genes, we have also surveyed expression divergence of *MDG I* genes from both poplar and wheat, where higher rates of gene expansion were detected. We analyzed the expression divergence of 16 wheat and 12 poplar *MDG I* genes among different tissues or treatments. We first constructed phylogenetic trees and figured out closely related genes, which were used for investigating expression divergence (supplementary fig. S2*A* and *B*, Supplementary Material online). Such an investigation showed that no similar expression patterns were observed among closely related *MDG I* genes in both wheat and poplar. In fact, these genes differentiated each other and no gene showed the same expression pattern to any other genes. These data suggested that expression divergence significantly contributed to the gene survival after duplication.

Positive Selection Occurred Only during Intragenus Divergence in the *HhH-GPD* Family

We have analyzed the *Ka/Ks* ratios among 15 distantly related species using their domain regions for both *MDG I* and *HhH-GPD* gene families. Such analyses showed that the divergence of these genes among the 15 species was under purifying selection for both the *MDG I* and *HhH-GPD* gene families (fig. 5). We have also investigated the *Ka/Ks* ratios within the rice *Oryza* genus and similar results were obtained for these two gene families. However, positive selection was detected among 11 rice species/subspecies from the *Oryza* genus for the nondomain region of the *HhH-GPD* family. Further analysis showed that positively selected sites were within the rice gene *LOC_Os09g01290* and its orthologs (fig. 5*E*). As the positive selection was observed only in these species belonging to the *Oryza* genus, we further examined whether it occurred within subspecies such as indica and japonica rice accessions. We first examined the *Ka/Ks* ratio between the gene *LOC_Os09g01290* (from Nipponbare) and its indica ortholog *BGIOSGA030257* (from the rice variety 93-11). The ratio was 0.72 and no positive selection was observed between indica and japonica species. We then analyzed a total of 1,402 rice accessions, whose genomes were genome-widely resequenced. Based on the single nucleotide polymorphisms and Indels (1–10 bp insertions and deletions) identified from the resequencing data, we calculated their *Ka/Ks* ratios by comparing with either indica genome 93-11 (first sequenced indica variety) or japonica genome Nipponbare (first sequenced japonica variety). However, our data showed that no positive selection was detected either within or among indica and japonica lines. Thus, positive selection was only detected among species within the *Oryza* genus. The data suggested that the positive selection might play a role in the species divergence within the genus. Previous data

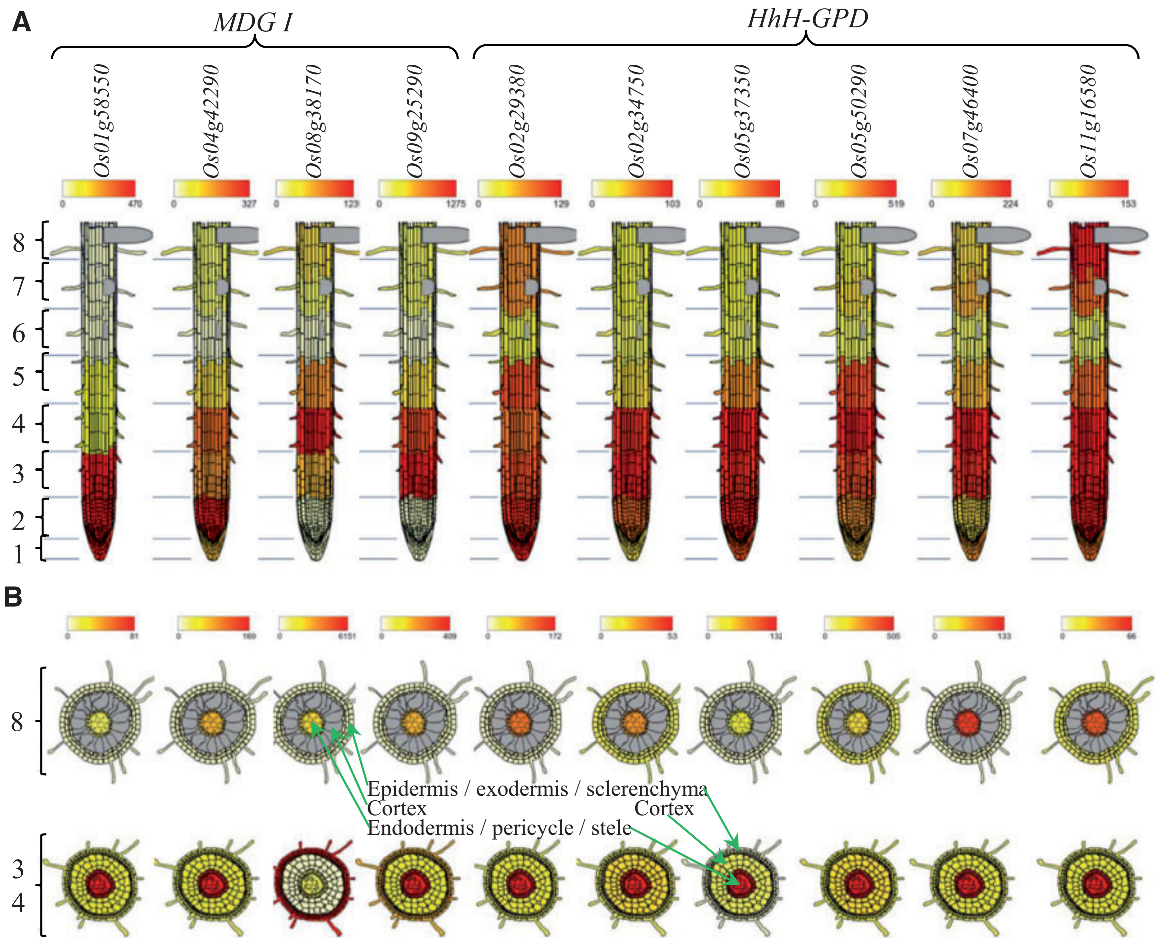


Fig. 9.—Expression profiling of rice *MDG I* and *HhH-GPD* genes in various developmental stages and tissue-types of roots. (*A*) showing the expression patterns of four *MDG I* and six *HhH-GPD* genes in eight developmental stages of roots. 1, Root cap; 2, division zone; 3, elongation zone; 4, maturation zone I; 5, maturation zone II; 6, maturation zone III; 7, maturation zone IV; 8, maturation zone V. (*B*) showing the expression patterns in different tissue types of roots. Top panel showed the transversal sections of roots at the maturation zone V region. Bottom panel showed the transversal sections of roots between elongation zone and maturation zone I region. In these root sections, samples for RNA extraction were taken from three parts including epidermis/exodermis/sclerenchyma, cortex and endodermis/pericycle/stele as indicated in the figure. However, the samples from cortex at maturation zone_V were excluded for microarray analysis as they mainly consisted of aerenchyma (nonliving cells) at this stage. More detailed description about collecting these samples in (*A*) and (*B*) is available on the link http://ricexpro.dna.affrc.go.jp/RXP_4001/index.php (last accessed March 31, 2016) (Takehisa et al. 2012). The numbers 3, 4 and 8 in (*B*) indicate the tissue types elongation zone, maturation zone I and maturation zone V, respectively.

showed that some adaptive phenotypes were due to the gene variants referring to positive selection (Koester et al. 2013). Studies also showed that positively selected genes might play roles in multiple biological processes, such as signal transduction, sexual reproduction, transporters, and so on (Castillo-Davis et al. 2004; Bustamante et al. 2005; Nielsen et al. 2005; Namroud et al. 2008; Li et al. 2009; Voolstra et al. 2011). We have further carried out the *Ka*/*Ks* analysis among rice lines from indica, japonica, and their intermediates. However, no positive selection was detected. Thus, positive selection in this gene might occur only during adaptive divergence within the *Oryza* genus.

Rice *MDG I* and *HhH-GPD* Genes Might Play Roles in Abiotic/Biotic and Hormone Signaling Pathways

In this study, we have genome-widely identified a total of 6 *MDG* and 13 *HhH-GPD* genes in the rice genome. Full-length cDNA sequences have been detected in most of these 19 genes and all of them were expressed in multiple tissues or developmental stages, suggesting their roles in multiple tissues or stages (fig. 6). Interestingly, the expression of three of six *MDG I* genes was down- or upregulated by only drought but not high salinity or cold stress. On the other hand, 4 out of 13 *HhH-GPD* genes were downregulated by drought stress. These data suggested that both *MDG I* and *HhH-GPD* genes should play roles in the drought stress response. Drought response is a complex mechanism, which involves in both ABA-dependent and ABA-independent pathway (Nakashima et al. 2014). In our study, two *MDG I* genes (*LOC_Os01g58550* and *LOC_Os08g38170*) and one *HhH-GPD* gene *LOC_Os12g10850* were coregulated by both drought stress and ABA treatment. Thus, they might play a role in an ABA-dependent pathway. Studies showed that besides ABA, both JA and auxin might also regulate drought responses (Divi et al. 2010; Peleg and Blumwald 2011). Thus, our data imply the roles of some of *MDG I* genes in the drought stress through ABA/JA-dependent signaling pathway. To our surprise, up to 80% of *MDG I* genes were downregulated and 42% *of HhH-GPD* genes were down- or upregulated by rice blast fungus pathogens (fig. 7*B*). Most of them were also downregulated by ABA or JAs. More attention should be paid to the gene *LOC_Os01g58550*, which was highly expressed in multiple tissues (fig. 6*A*), upregulated by drought stress, downregulated by both pathogens *M. grisea* and Xoo as well as by ABA and JA (figs. 7 and 8). ABA is a negative regulator of disease resistance (Mauch-Mani and Mauch 2005) and plays a key role in modulating diverse plant–pathogen interactions (Fan et al. 2009). JA plays a central node in plant defense signaling network (Robert-Seilaniantz et al. 2011; Campos et al. 2014). Thus, the gene *LOC_Os01g58550* might play a key role in drought and *MG*/*XOO*-related stress regulation through ABA/JA signaling pathways.

Supplementary Material

Supplementary figures S1 and S2 and tables S1–S8 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgment

This work was supported by the Singapore National Research Foundation under the Competitive Research Programme Funding Scheme (CRP award no. NRF-CRP7-2010-02).

Literature Cited

Admiraal SJ, O'Brien PJ. 2015. Base excision repair enzymes protect abasic sites in duplex DNA from interstrand cross-links. Biochemistry 54:1849–1857.

Barrett T, et al. 2013. NCBI GEO: archive for functional genomics data sets–update. Nucleic Acids Res. 41:D991–D995.

Bustamante CD, et al. 2005. Natural selection on protein-coding genes in the human genome. Nature 437:1153–1157.

Calvo JA, et al. 2013. Aag DNA glycosylase promotes alkylation-induced tissue damage mediated by Parp1. PLoS Genet. 9:e1003413.

Campos ML, Kang JH, Howe GA. 2014. Jasmonate-triggered plant immunity. J Chem Ecol. 40:657–675.

Castillo-Davis CI, Kondrashov FA, Hartl DL, Kulathinal RJ. 2004. The functional genomic distribution of protein divergence in two animal phyla: coevolution, genomic conflict, and constraint. Genome Res. 14:802–811.

Divi U, Rahman T, Krishna P. 2010. Brassinosteroid-mediated stress tolerance in *Arabidopsis* shows interactions with abscisic acid, ethylene and salicylic acid pathways. BMC Plant Biol. 10:151.

Drohat AC, Kwon K, Krosky DJ, Stivers JT. 2002. 3-Methyladenine DNA glycosylase I is an unexpected helix-hairpin-helix superfamily member. Nat Struct Biol. 9:659–664.

Ebrahimkhani MR, et al. 2014. Aag-initiated base excision repair promotes ischemia reperfusion injury in liver, brain, and kidney. Proc Natl Acad Sci USA. 111:E4878–E4886.

Eisen MB, Spellman PT, Brown PO, Botstein D. 1998. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA. 95:14863–14868.

Fan J, Hill L, Crooks C, Doerner P, Lamb C. 2009. Abscisic acid has a key role in modulating diverse plant-pathogen interactions. Plant Physiol. 150:1750–1761.

Feng S, et al. 2010. Conservation and divergence of methylation patterning in plants and animals. Proc Natl Acad Sci USA. 107:8689–8694.

Flagel LE, Wendel JF. 2009. Gene duplication and evolutionary novelty in plants. New Phytol. 183:557–564.

Freeling M. 2009. Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. Annu Rev Plant Biol. 60:433–453.

Herbig A, Jäger G, Battke F, Nieselt K. 2012. GenomeRing: alignment visualization based on SuperGenome coordinates. Bioinformatics 28:i7–i15.

Jain M, et al. 2007. F-box proteins in rice. Genome-wide analysis, classification, temporal and spatial gene expression during panicle and seed development, and regulation by light and abiotic stress. Plant Physiol. 143:1467–1483.

Jiang SY, Christoffels A, Ramamoorthy R, Ramachandran S. 2009. Expansion mechanisms and functional annotations of hypothetical genes in the rice genome. Plant Physiol. 150:1997–2008.

Jiang SY, Ramachandran S. 2006. Comparative and evolutionary analysis of genes encoding small GTPases and their activating proteins in eukaryotic genomes. Physiol Genomics 24:235–251.

Kawahara Y, et al. 2013. Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. Rice 6:4.

Koester JA, Swanson WJ, Armbrust EV. 2013. Positive selection within a diatom species acts on putative protein interactions and transcriptional regulation. Mol Biol Evol. 30:422–434.

Kong H, et al. 2007. Patterns of gene duplication in the plant SKP1 gene family in angiosperms: evidence for multiple mechanisms of rapid gene birth. Plant J. 50:873–885.

Lamesch P, et al. 2012. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res. 40:D1202–D1210.

Larson K, Sham J, Shenkar R, Strauss B. 1985. Methylation-induced blocks to *in vitro* DNA replication. Mutat Res. 150:77–84.

Lee CY, et al. 2009. Recognition and processing of a new repertoire of DNA substrates by human 3-methyladenine DNA glycosylase (AAG). Biochemistry 48:1850–1861.

Lee TH, Tang H, Wang X, Paterson AH. 2013. PGDD: a database of gene and genome duplication in plants. Nucleic Acids Res. 41(Database issue):D1152–D1158.

Li YD, et al. 2009. Detecting positive selection in the budding yeast genome. J Evol Biol. 22:2430–2437.

Loechler EL, Green CL, Essigman JM. 1984. *In vivo* mutagenesis by O6-methylguanine built into a unique site in a viral genome. Proc Natl Acad Sci USA. 81:6271–6275.

Lukens LN, Zhan S. 2007. The plant genome's methylation status and response to stress: implications for plant improvement. Curr Opin Plant Biol. 10:317–322.

Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. Science 290:1151–1155.

Malhotra S, Sowdhamini R. 2013. Genome-wide survey of DNA-binding proteins in *Arabidopsis thaliana*: analysis of distribution and functions. Nucleic Acids Res. 41:7212–7219.

Massingham T, Goldman N. 2005. Detecting amino acid sites under positive selection and purifying selection. Genetics 169:1753–1762.

Mauch-Mani B, Mauch F. 2005. The role of abscisic acid in plant-pathogen interactions. Curr Opin Plant Biol. 8:409–414.

Meyer A, Van de Peer Y. 2003. "Natural selection merely modified while redundancy created"–Susumu Ohno's idea of the evolutionary importance of gene and genome duplications. J Struct Funct Genomics 3:7–9.

Nakashima K, Yamaguchi-Shinozaki K, Shinozaki K. 2014. The transcriptional regulatory network in the drought response and its crosstalk in abiotic stress responses including drought, cold, and heat. Front Plant Sci. 5:170.

Namroud MC, Beaulieu J, Juge N, Laroche J, Bousquet J. 2008. Scanning the genome for gene single nucleotide polymorphisms involved in adaptive population differentiation in white spruce. Mol Ecol. 17:3599–3613.

Nash HM, et al. 1996. Cloning of a yeast 8-oxoguanine DNA glycosylase reveals the existence of a base-excision DNA-repair protein superfamily. Curr Biol. 6:968–980.

Nielsen R, et al. 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol. 3:e170.

Nota F, Cambiagno DA, Ribone P, Alvarez ME. 2015. Expression and function of *AtMBD4L*, the single gene encoding the nuclear DNA glycosylase MBD4L in *Arabidopsis*. Plant Sci. 235:122–129.

Peleg Z, Blumwald E. 2011. Hormone balance and abiotic stress tolerance in crop plants. Curr Opin Plant Biol. 14:290–295.

Ponferrada-Marín MI, Parrilla-Doblas JT, Roldán-Arjona T, Ariza RR. 2011. A discontinuous DNA glycosylase domain in a family of enzymes that excise 5-methylcytosine. Nucleic Acids Res. 39:1473–1484.

Ramiro-Merina Á, Ariza RR, Roldán-Arjona T. 2013. Molecular characterization of a putative plant homolog of MBD4 DNA glycosylase. DNA Repair (Amst) 12:890–898.

Robert-Seilaniantz A, Grant M, Jones JD. 2011. Hormone crosstalk in plant disease and defense: more than just jasmonate-salicylate antagonism. Annu Rev Phytopathol 49:317–343.

Sakumi K, et al. 1986. Purification and structure of 3-methyladenine-DNA glycosylase I of *Escherichia coli*. J Biol Chem. 261:15761–15766.

Sakumi K, Sekiguchi M. 1990. Structures and functions of DNA glycosylases. Mutat Res. 236:161–172.

Santerre A, Britt AB. 1994. Cloning of a 3-methyladenine-DNA glycosylase from *Arabidopsis thaliana.* Proc Natl Acad Sci USA. 91:2240–2244.

Sato Y, et al. 2013. RiceFREND: a platform for retrieving coexpressed gene networks in rice. Nucleic Acids Res. 41:D1214–D1221.

Schreiber AW, et al. 2009. Comparative transcriptomics in the *Triticeae*. BMC Genomics 10:285.

Shi L, Kent R, Bence N, Britt AB. 1997. Developmental expression of a DNA repair gene in *Arabidopsis*. Mutat Res. 384:145–156.

Shiu SH, et al. 2004. Comparative analysis of the receptor-like kinase family in *Arabidopsis* and rice. Plant Cell 16:1220–1234.

Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 34:W609–W612.

Takehisa H, et al. 2012. Genome-wide transcriptome dissection of the rice root system: implications for developmental and physiological functions. Plant J. 69:126–140.

Taylor EL, O'Brien PJ. 2015. Kinetic mechanism for the flipping and excision of 1,N(6)-ethenoadenine by AlkA. Biochemistry 54:898–908.

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

Voolstra CR, et al. 2011. Rapid evolution of coral proteins responsible for interaction with the environment. PLOS One 6:e20392.

Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. 2009. Jalview version 2—a multiple a multiple sequence alignment editor and analysis workbench. Bioinformatics 25:1189–1191.

Wyatt MD, Allan JM, Lau AY, Ellenberger TE, Samson LD. 1999. 3-methyladenine DNA glycosylases: structure, function, and biological importance. Bioessays 21:668–676.

Yang Z, Nielsen R. 2000. Estimating synonymous and nonsynonymous substitution rates under evolutionary models. Mol Biol Evol. 17:32–43.

Zhu JK. 2009. Active DNA demethylation mediated by DNA glycosylases. Annu Rev Genet. 43:143–166.

Associate editor: Marta Barluenga