Large-Scale, Lineage-Specific Expansion of a Bric-a-Brac/Tramtrack/Broad Complex Ubiquitin-Ligase Gene Family in Rice

Derek J. Gingerich, a Kousuke Hanada, b Shin-Han Shiu, b and Richard D. Vierstra a,1

a Department of Genetics, University of Wisconsin, Madison, Wisconsin 53706
b Department of Plant Biology, Michigan State University, East Lansing, Michigan 48824-1312

Selective ubiquitination of proteins is directed by diverse families of ubiquitin-protein ligases (or E3s) in plants. One important type uses Cullin-3 as a scaffold to assemble multisubunit E3 complexes containing one of a multitude of bric-a-brac/tramtrack/broad complex (BTB) proteins that function as substrate recognition factors. We previously described the 80-member BTB gene superfamily in Arabidopsis thaliana. Here, we describe the complete BTB superfamily in rice (Oryza sativa spp japonica cv Nipponbare) that contains 149 BTB domain–encoding genes and 43 putative pseudogenes. Amino acid sequence comparisons of the rice and Arabidopsis superfamilies revealed a near equal repertoire of putative substrate recognition module types. However, phylogenetic comparisons detected numerous gene duplication and/or loss events since the rice and Arabidopsis lineages split, suggesting possible functional specialization within individual BTB families. In particular, a major expansion and diversification of a subset of BTB proteins containing Meprin and TRAF homology (MATH) substrate recognition sites was evident in rice and other monocots that likely occurred following the monocot/dicot split. The MATH domain of a subset appears to have evolved significantly faster than those in a smaller core subset that predates flowering plants, suggesting that the substrate recognition module in many monocot MATH-BTB E3s are diversifying to ubiquitinate a set of substrates that are themselves rapidly changing. Intriguing possibilities include pathogen proteins attempting to avoid inactivation by the monocot host.

INTRODUCTION

Covalent attachment of ubiquitin (Ub) to specific proteins is an important mechanism for posttranslational control in both plants and animals. Its best-known function is to target specific proteins for breakdown (Smalle and Vierstra, 2004; Varshavsky, 2005). Here, numerous short-lived proteins within the cytoplasm, nucleus, or membranes that face these compartments become modified with polymeric chains of Ub.s, primarily linked internally through Lys-48. The resulting polyubiquitinated proteins are then recognized by the 26S proteasome, a 2-MD protease complex that degrades the target but releases the Ub moieties intact for reuse. Other functions of Ub attachment include roles in chromatin structure and transcriptional regulation, DNA repair, and endocytosis (Aguilar and Wendland, 2003; Pickart, 2004). Often these substrates are either monoubiquitinated or polyubiquitinated through Lys residues other than Lys-48. Via these proteolytic and nonproteolytic functions, Ub has profound effects on the physiology, development, and homeostasis of all eukaryotes. For plants in particular, Ub conjugation has been connected to the cell cycle, embryogenesis, most, if not all, hormone responses, photomorphogenesis, circadian rhythms, floral development, self incompatibility, environmental adaptation, disease resistance, and programmed cell death to name a few (Moon et al., 2004; Smalle and Vierstra, 2004).

The functions of Ub are primarily determined by a reaction cascade that attaches the first Ub and then assembles the polyUb chains. This ATP-dependent process is performed by the sequential action of three enzyme classes, Ub-activating enzymes (or E1s), Ub-conjugating enzymes (or E2s), and Ub-protein ligases (or E3s) (Pickart, 2004; Smalle and Vierstra, 2004; Varshavsky, 2005). E3s selectively bind the target and catalyze transfer of the Ub moiety from the E2 and as such determine both substrate specificity and how the Ub is linked (monoubiquitination versus polyubiquitination through specific Lys residues). Not surprisingly, genomic analyses revealed that large collections of E3s exist to handle the myriad of expected intracellular substrates. In Arabidopsis thaliana, mice, and Caenorhabditis elegans, for example, >1300, 610, and 590 E3s are predicted from scans of each proteome (Furukawa et al., 2003; Geyer et al., 2003; Semple, 2003; Moore and Boyd, 2004; Smalle and Vierstra, 2004; Willems et al., 2004).

The largest families of E3s are multisubunit complexes containing a core subcomplex comprised of a Cullin (CUL) that serves as the backbone and RBX1 (or ROC1/HRT1) that associates with the E2-Ub intermediate (Smalle and Vierstra, 2004; Varshavsky, 2005). One of a variable collection of substrate
substrate adaptor components of E3s in plants and to help explain this diversity is that each lineage mixed and compared it to our previous analysis of the BTB superfamily in Arabidopsis (Gingerich et al., 2005). These species diverged from a common ancestor ~150 million years ago (Wikstrom et al., 2001) and thus represented an opportunity to identify general sets of E3 adapters (and their targets) that have been preserved in angiosperms as well as those that emerged after the monocot/dicot split. Here, we describe a collection of 149 predicted BTB proteins in rice. Comparison of the rice and Arabidopsis superfamilies showed that the BTB domains in both species are linked to the same general sets of putative substrate binding motifs. However, substantial diversification within various BTB protein families is evident in addition to a large-scale expansion of a subset of MATH-BTB proteins that appears to be common among monocots. This expansion is combined with evidence of both rapid birth-and-death evolution and diversifying selection of a large collection of monocot-specific MATH-BTB proteins, suggesting that a subset of rice BTB E3s is rapidly changing to cope with targets that may also be rapidly changing.

RESULTS

Identification and Characterization of the Rice BTB Superfamily

Prior phylogenetic analysis of the CUL family in rice identified a set of 13 CUL-type proteins, with three (Os CUL3a-c) showing strong similarity to Arabidopsis CUL3a/b (Gingerich et al., 2005). In particular, the adapter interface predicted to be used by At CUL3a/b to bind BTB proteins is well conserved in Os CUL3a-c, strongly suggesting that Os CULa-c also help assemble BTB E3 complexes in rice. To identify the array of rice CUL3/BTB complexes that potentially exist, we defined the full complement of BTB proteins in the O. sativa ssp japonica cv Nipponbare sequence database. Iterative BLAST searches using 48 sequences encompassing the ~110-amino acid core BTB domain from yeast, plants, and animals (Gingerich et al., 2005; Stogios et al., 2005) recovered 192 open reading frames (ORFs) encoding one or more rice BTB domains. Subsequent analysis of this set categorized 43 loci as putative pseudogenes, based on the presence of one or more in-frame stop codons or frame shifts disrupting the coding region (see below). After removing these pseudogenes, a final set of 149 potentially functional BTB genes was predicted in rice. This collection is noticeably larger (86%) than the Arabidopsis BTB superfamily (Gingerich et al., 2005), indicating that either the rice superfamily had significantly expanded and/or the Arabidopsis counterpart had experienced significant gene loss since the split of monocots and dicots ~150 million years ago (Wikstrom et al., 2001).

Similar to previous descriptions of the Arabidopsis BTB protein superfamily (Dieterle et al., 2005; Gingerich et al., 2005), analysis of the rice BTB sequences both upstream and downstream of the BTB domain identified a collection of other protein–protein interaction motifs that likely represent substrate recognition sites, including ankyrin, armadillo, and TPR repeats, MATH, coiled-coil, and transcriptional adaptor zinc finger (TAZ) (Figure 1).
Similar to Arabidopsis, the rice superfamily also contains a large collection (25 members) with the plant-specific NPH3 domain, an ~250-residue motif first found in the blue light photoreceptor NPH1-interacting protein NPH3 (Motchoulski and Liscum, 1999; Sakai et al., 2000); two BTB proteins with pentapeptide repeats whose function(s) are unknown; and one with an F5/8-type C (discoidin) domain, which has been implicated in phospholipid interaction (Foster et al., 1990; Baumgartner et al., 1998). As with the Arabidopsis collection (Gingerich et al., 2005), SMART failed to detect previously described domains in the sequences flanking the BTB domain for a number of rice proteins. However, alignments of several subsets identified conserved regions that could represent new interaction motifs (e.g., B4, C3, E1, and H families; Gingerich et al., 2005; data not shown). Two of the predicted rice BTB proteins (Os11g02070 and Os12g02030) are significantly shorter (275 residues in length each) and appear to contain just the BTB domain (Figure 1). Such BTB-only proteins have been identified previously in Schizosaccharomyces pombe, C. elegans, and Arabidopsis (Geyer et al., 2003; Xu et al., 2003; Dieterle et al., 2005; Gingerich et al., 2005).

Domain architecture comparison between the Arabidopsis and rice BTB superfamilies revealed two nearly identical clusters of four BTB genes on chromosomes 11 and 12. The pairs in these clusters (Os11g02070/Os12g02030 [C4 subfamily], Os11g02610/Os12g02530 and Os11g02620/Os12g02540 [F subfamily], and Os11g04600/Os12g04410 [E4 subfamily]) share 97 to 99% amino acid sequence identity and likely arose from a well-documented recent (4 to 14 million years ago) segmental duplication involving the ends of chromosomes 12 and 11 (Wu et al., 1998; Goff et al., 2002; Wang et al., 2005). Comparative Analysis of the Rice and Arabidopsis BTB Genes

To help identify common BTB proteins that were established during angiosperm evolution (or perhaps even earlier) and likely recognize substrates widely distributed in plants versus those that arose later and likely recognize targets specific to either rice or Arabidopsis, phylogenetic trees were generated with all 149 rice and 80 Arabidopsis members either separately or together (Figure 2; see Supplemental Figure 1 online). The trees were generated from BTB domain alignments (see Supplemental Figures 2 and 3 online) by the distance-based neighbor-joining (NJ) method (Saitou and Nei, 1987) and then color coded based on the other associated domains. Bootstrap values of most deep interior branches were low because of the large number of sequences and the small size of the BTB domain. However, more significant bootstrap values in the distal branches allowed us to group the rice and Arabidopsis BTB proteins into distinct families. The trees consistently clustered proteins with similar
BTB domains and target binding motifs, thus providing independent support for these family classifications (Figure 2; see Supplemental Figure 1 online). Similar trees were also generated by character-based maximum parsimony (MP) analysis (see Supplemental Figure 4 online). Although there was variation in the topology of the very deep interior branches and a few of the outermost branches, well-supported branches in the MP tree had identical compositions to those in the NJ tree. The lone exception was the A1 rice MATH-BTB subfamily (see below); ambiguity for the interior branches in the MP tree meant that members of this subfamily did not group into a single distinct clade.

For simplicity, the NJ trees were subdivided into alphabetical families (A to H) and subfamilies (e.g., A1 to A3) based on both tree topologies and predicted protein domain compositions (Figure 2; see Supplemental Figure 1 online). Even in cases where SMART failed to predict additional motifs (e.g., B4, C3, and E1 subfamilies and H family), alignments of the rice and Arabidopsis subfamilies revealed that their amino acid sequences are also conserved outside of the BTB domain. Apparent exceptions include the ankyrin-BTB proteins Os6g21330 and At2g04740 (D family), which did not group with the E4 subfamily, and the MATH-containing protein Os01g70670, which did not group with the A1 or A2 subfamilies. These outliers have substantially different domain structures, suggesting that they are not evolutionarily related to the other ankyrin-BTB or MATH-BTB members. We also detected a few BTB proteins within the A1, C1, and F subfamilies with dissimilar architectures compared with others in the cluster. Almost all appear to be shortened proteins when compared with other family members, with the BTB domain as the only recognizable motif (the exceptions being Os11g40670 and Os11g41260 [see above]). Although these loci may be pseudogenes, their putative coding regions harbor no obvious in-frame stops or frame shifts; consequently, we retained them as possibly functional. We also identified five A1 subfamily members in rice that contain motifs N-terminal to the...
BTB domain that were not identified as MATH domains by SMART/PFAM. While alignments suggest that these regions are related to MATH domains, they lack several conserved residues that are characteristic of canonical MATH domains and/or contain significant deletions within the domain. Consequently, we have designated these domains as “MATH related.”

Most rice and Arabidopsis BTB proteins with similar domain compositions clustered phylogenetically, suggesting that they were derived from the same ancestral genes. Many of these clades also contained similar numbers of Arabidopsis and rice sequences, indicating that major expansions/contractions have not occurred since the monocot/dicot split (see Supplemental Figure 1 online). For instance, all three armadillo repeat-BTB proteins (two Arabidopsis and one rice) clustered in the same B1 subfamily clade, all six TPR-BTB proteins (three Arabidopsis and three rice) clustered in the same B2 clade, and all three penta-peptide repeat-BTB proteins (one Arabidopsis and two rice) clustered in the same B3 clade. However, there were two notable exceptions (Figure 2). One was an expansion of the BTB-only type (C4 subgroup) in Arabidopsis, which contains eight members versus only two in rice. The second was the dramatic expansion and separation of the MATH-BTB family in rice. Whereas Arabidopsis encodes only six MATH-BTB proteins, rice encodes at least 69 MATH-BTB proteins, with five additional BTB proteins with MATH-related domains, and another 41 genomic loci predicted to be MATH-BTB pseudogenes (Figure 2; see below). In the rice/Arabidopsis combined tree, most of the rice MATH-BTB and MATH-related BTB proteins (70 members total) formed a large rice-specific group (A1) distinct from the remaining four from rice and all six MATH-BTB proteins from Arabidopsis that clustered together in a separate A2 subfamily (see Supplemental Figure 1 online). The presence of these two rice MATH-BTB clades implied that the MATH-BTB family can be divided into two groups with contrasting evolutionary histories: a small core set that is common to both rice and Arabidopsis and a larger expanded set present in rice.

To better understand how the BTB families evolved in the rice and Arabidopsis lineages, we inferred the number of BTB genes in the most recent common ancestor (MRCA) based on the rice/Arabidopsis NJ BTB domain tree (see Supplemental Figure 1 online). Because the relatively short sequence (>110 residues of the BTB domain) used for the alignments generated ambiguity for some subgroups, we further clarified the relationships within the groups by NJ analysis with the full-length proteins. (see Supplemental Figure 5 online). Well-supported clades (bootstrap value ≥65%) containing both Arabidopsis and rice sequences were first identified in the trees (see black dots in Supplemental Figures 1 and 5 online). Combination and reconciliation of the gene trees and species relationships allowed the identification of 41 putative orthologous groups where a single progenitor BTB gene in the MRCA appeared to generate both the rice and Arabidopsis descendants. At least one of these groups was present in most of the individual families and subfamilies, providing further support that most plant BTB gene subtypes appeared prior to the monocot/dicot split. Eleven rice or Arabidopsis–specific genes or groups of genes were also detected that form well-supported sister groups (bootstrap value ≥65%) with the orthologous groups defined above (see gray dots in Supplemental Figures 1 and 5 online). Each of these may reflect an additional ancestral gene in the MRCA where the descendents were lost in one of the two lineages. In addition, we identified three rice families/subfamilies (A1, A3, and G) and one Arabidopsis subfamily (E2) that appear to be species specific. These unique clades may represent BTB gene types lost in one of the two lineages and/or acquisition of completely new BTB types following speciation.

Collectively, under the assumption that at least one ancestor should be assigned to each BTB family or subfamily, a minimum of 56 BTB genes in the MRCA was estimated. When the number of BTB genes in the MRCA is compared with the number of functional BTB genes in Arabidopsis and rice, it appears that the Arabidopsis superfamily has increased slightly (42%), while the rice superfamily has almost tripled in size since the divergence of monocots and dicots. The greater expansion in the rice superfamily is almost completely the result of the dramatic rice-specific amplification of the MATH-BTB type (from three estimated in the MRCA to 74 MATH and MATH-related BTBs in rice). Assuming that this expansion occurred entirely following the split of monocots and dicots, the birth rate of MATH-BTB genes is at least 47 genes per 100 million years, which is ~50 times the estimated typical gene duplication rate of one gene per 100 million years in eukaryotic genomes (Lynch and Conery, 2000). This rice expansion would be even larger if the 43 BTB-related pseudogenes are included (see below).

Predicted orthologous relationships between individual rice and Arabidopsis BTB proteins are detailed in Supplemental Table 1 online. Eighteen rice and Arabidopsis proteins have a one-to-one correspondence, where a single protein from each species shares a node on the tree. Included in this list are (1) Arabidopsis ARIA, which assists in abscisic acid responses (Kim et al., 2004) and groups with the rice protein Os05g33050; (2) NPH3 (RPT3), which helps mediate blue light photoreponses (Motchoulski and Liscum, 1999; Inada et al., 2004), and groups with rice CPT1 previously shown to mediate coleoptile and root phototropism (Haga et al., 2005); (3) ETO1, which targets the ethylene biosynthetic type-2 1-aminocyclopropane-1-carboxylate synthases for breakdown (Wang et al., 2004a), and groups with rice Os03g18360; and (4) AtB3, a calmodulin binding protein (Du and Poovaiah, 2004), which groups with Os01g66890. In some cases, additional BTB orthologs are evident in one of the two species that could reflect expansion. As examples, Arabidopsis BOP1 and BOP2, which act redundantly to regulate development of lateral organs (Ha et al., 2003, 2004; Norberg et al., 2005), are co-orthologs of a single rice BTB protein (Os01g72020), and NPR1, which regulates pathogen response gene expression (Dong, 2004), and NPR2, are co-orthologs of a single rice protein, OsNH1.

The Rice Genome Contains Numerous MATH-BTB Pseudogenes

Whereas our previous analysis of the Arabidopsis BTB superfamily failed to identify any obvious pseudogenes (Gingerich et al., 2005), 43 of the 192 rice BTB domain–containing loci (22%) contained frame shifts or in-frame premature stop codons characteristic of pseudogenes. When reconstructed, 36 of these pseudogenes were also predicted to encode all or part of a
MATH domain and clustered by NJ analysis with the A1 subclade containing the rice-specific expanded set of MATH-BTB proteins (see Supplemental Figure 6 online). Even though an additional four pseudogenes did not encode obvious MATH domains, they also fell within the A1 subgroup phylogenetically, suggesting that they descended from A1 progenitor(s) as well, and either the MATH sequence was eliminated or degenerated beyond recognition. These 40 pseudogenes were distributed throughout the A1 subfamily clade, suggesting a dynamic history of gene duplication and loss that is characteristic of active birth-and-death gene evolution (Nei and Rooney, 2005). We were unable to assemble full BTB domain–encoding sequences from the remaining three pseudogenes for phylogenetic analysis. However, alignments, BLAST best hits, and partial coding region reconstructions suggested that Os08g12970 is also part of the A1 subfamily, whereas Os05g12030 and Os12g31320 are part of the F (BTB-NPH3) subfamily. We also note that tBLASTn searches recovered eight locations in the rice genome where short ORFs encode ~20 to 50 amino acid fragments of the consensus BTB domain. These seemingly random remnants could not be even remotely assembled into functional ORFs and likely reflect more ancient pseudogenization events.

Analysis of the MATH-BTB Family in the Plant Kingdom

To further describe the evolutionary path of MATH-BTB proteins, we expanded our phylogenetic analysis to encompass homologs from an evolutionary diverse spectrum of land plant species, including other monocots (sorghum [Sorghum bicolor] and wheat [Triticum monococcum]) and dicots (Medicago truncatula and poplar [Populus trichocarpa]), a gymnosperm (pine [Pinus taeda]), a moss (Physcomitrella patens), and a bryophyte (Selaginella moellendorffii). These complete or near-complete MATH-BTB protein sequences were identified in both the genomic and EST databases, using the rice and Arabidopsis MATH-BTB protein sequences as queries. The number of identified MATH-BTB sequences for each species and their gene designations can be found in Figure 3 and Supplemental Table 2.

![Phylogenetic Trees of 135 MATH-BTB Proteins from Representative Land Plant Species.](image)

Alignments of the ~110-amino acid BTB domains or the ~110-amino acid MATH domains were used to generate NJ phylogenetic trees. The core group is highlighted by the gray ovals. Numbers at the internal branches leading to the core group indicate percentage of bootstrap support from 1000 replicates. Expanded views of the trees with the branches labeled with sequence identifiers are in Supplemental Figure 11 online as well as the color code for the species.

(A) BTB domain tree color-coded by the species.
(B) MATH domain tree color-coded by the species.
(C) BTB domain tree color-coded by the number of introns.
online. Whereas the MATH-BTB loci appeared to be intact in most of the species, 27 examples of apparent MATH-BTB pseudogenes were evident in sorghum, suggesting that this family has experienced similar evolutionary dynamics as their rice MATH-BTB A1 subfamily counterparts (data not shown). Similar to rice, we also identified two loci in sorghum that appear to encode degenerate MATH-related domains C-terminal to the BTB motif. Rice and sorghum genes containing these MATH-related sequences along with the possible BTB-MATH-BTB sequence Os11g41260, as well as one rice MATH-BTB sequence (Os08g31450), which has a slightly truncated BTB domain, were not included in further analysis (see below) because their inclusion produced large gaps and other aberrations in sequence alignments. All additional analysis focused on the remaining 135 canonical MATH-BTB sequences.

When the predicted canonical MATH-BTB proteins were analyzed phylogenetically by NJ analysis using either BTB or MATH domain alignments (see Supplemental Figure 7 online), we noticed a striking separation into two clades that strictly followed the core and expanded rice groups (Figure 3). All the Physcomitrella, Selaginella, Medicago, poplar, and pine MATH-BTBs and two of the sorghum MATH-BTBs clustered into a distinct clade along with all six from Arabidopsis and the core group of four from rice. The 64 canonical rice MATH-BTB sequences from the expanded A1 group were joined by 39 of the 41 MATH-BTB proteins from sorghum and the single wheat protein, demonstrating that the expanded group is likely to be monocot specific. This clustering was especially obvious in the MATH domain-based NJ tree and implied that the MATH domain in particular has had separate evolutionary histories between the core and monocot expanded groups (Figure 3B). This distinction was even more striking in sequence alignments of the MATH-BTB protein collection. Whereas the MATH domain alignment of the core group (which included sequences ranging from bryophytes to monocots) has no gaps and only 25 out of 113 positions with <80% sequence identity, the alignment of the expanded group (only monocot sequences) revealed much more limited conservation, with 149 out of 159 positions having <80% sequence identity, and a substantially greater number of sequences with insertions/deletions (Figure 7; see Supplemental Figure 8 online). By contrast, the BTB domain alignments of the core (46/127 positions with >80% identity) and expanded groups (38/167 positions with >80% identity) displayed more equivalent levels of diversity (see Supplemental Figure 8 online).

Upon further analysis of the genomic data, we found several other criteria that distinguished the core and expanded MATH-BTB groups. With respect to gene structure, most of the core MATH-BTB genes (where information is available) have four exons (Figure 3). The only exceptions are the two from Physcomitrella where only two exons are evident. Remarkably, the positions of the intron/exon junctions were absolutely conserved to the nucleotide in this highly diverse collection of plant species (Figure 4A; see Supplemental Figure 9 online). By contrast, the coding regions from a majority of the expanded rice MATH-BTB loci, along with the sorghum and the single wheat MATH-BTB loci, contained only one predicted exon uninterrupted by obvious introns (Figure 3), a configuration not detected in any of the nonmonocot species.

Differences between the core and expanded MATH-BTB groups were also evident based on chromosomal locations. Unlike most BTB genes in Arabidopsis and rice, which appear as singletons (65 of 80 in Arabidopsis and 63 of 72 non-A1 subfamily members in rice are at least 10 genes away from another BTB gene; Gingerich et al., 2005; data not shown), most of the expanded MATH-BTB genes and others in the A1 subfamily were concentrated in tandem duplication blocks in rice. These blocks include 54 of the 70 expanded MATH-BTB and MATH-related BTB genes that appear active and 32 of 36 MATH-BTB pseudogenes (Figure 5). As first noticed by Song et al. (2002), the largest blocks are the middle of chromosome 10. Here, 24 functional and 12 MATH-BTB pseudogenes (out of 69 total predicted ORFs) are clustered within a 325-kb region, and six functional and four pseudogenes (out of 27 ORFs) are clustered in a nearby 132-kb region (Figure 5). Smaller arrays of expanded MATH-BTB genes were found in chromosomes 8 and 11. Also in these clusters are three loci that encode BTB proteins that group phylogenetically with the rice A1 subfamily (Figure 2) but do not encode an associated MATH domain as well as Os11g41260, which may encode a BTB-MATH-BTB protein. Notably, these blocks include transposable element coding sequences that have been proposed to drive gene duplication events by inadvertently carrying copies of genes during transposition and/or by facilitating unequal crossovers (Hancock, 2005).

MATH-BTB Proteins in the Expanded Group Interact with CUL3

It has been previously shown that Arabidopsis members of the core MATH-BTB group interact with CUL3 proteins (Dieterle et al., 2005; Figueroa et al., 2005; Gingerich et al., 2005; Weber et al., 2005) and thus likely act as E3 Ub-ligase target adapters. To demonstrate that members of the rice expanded group likewise interact with CUL3s, we performed directed yeast two-hybrid (Y2H) analysis with four representatives. The four fell within different subclades and thus reflect a diversity of BTB sequences within the expanded MATH-BTB group (see Figure 2). We tested full-length rice MATH-BTB proteins against full-length Arabidopsis CUL3a, CUL3b, and CUL1 as well as a C-terminal truncated form (amino acids 1 to 364) of rice CUL3b that included helices II and V predicted to form the BTB binding interface (Geyer et al., 2003; Pintard et al., 2003; Xu et al., 2003; Gingerich et al., 2005; Figure 6B). As can be seen in Figure 6A, all four expanded group MATH-BTB proteins interacted with truncated Os CUL3b as well as full-length At CUL3a and At CUL3b. They did not interact with At CUL1, thus demonstrating their CUL isoform specificity. Full-length AtCUL3a interacted strongly with Os04g53410 and Os10g29110 but weakly with Os08g13070 and Os10g29310. Whether this distinction reflects a CUL3 isoform preference is not yet known.

Expression of the Rice MATH-BTB Family

To assess the transcriptional activity of the rice MATH-BTB genes, we examined three publicly available rice expression databases. First, we identified spp japonica full-length (FL) cDNAs or ESTs using the gene expression evidence search page at TIGR (http://
and searches of the EST database at the Gramene website (http://www.gramene.org/). FL-cDNAs and/or ESTs were identified for 26 of 64 loci in the expanded \textit{MATH-BTB} group loci as well as all four loci in the core \textit{MATH-BTB} group (Figure 4B). The four core genes were represented by between 44 and 59 ESTs/FL-cDNAs, while the 26 expanded genes were represented by between 1 and 80 ESTs/FL-cDNAs, with 15 represented by three or less, suggesting that the core \textit{MATH-BTB} genes are expressed at significantly higher levels than most of their expanded counterparts. A cursory search of the sorghum EST collection identified numerous cDNAs for the two core \textit{MATH-BTB} genes (12 for Sb MBTB5 and 14 for Sb MBTB45), whereas only four (Sb MBTB18, Sb MBTB20, Sb MBTB30, Sb MBTB31) of the 39

Figure 4. Gene Structure and Expression of the Core and Expanded \textit{MATH-BTB} Groups in Land Plants.


(B) Expression analysis for rice \textit{MATH-BTB} core and expanded groups and rice A1 subfamily pseudogene loci. The value in each box indicates the number of different tissues in which full-length cDNAs (FL-cDNAs) or ESTs were identified, the number of libraries in which significant expression was detected in the rice MPSS database, or the RICEATLAS whole-genome oligoarray expression data set. Shade of the boxes denotes level of expression (see Methods). A question mark indicates that significant expression was detected but that tag sequence (MPSS) or oligonucleotide (RICEATLAS) matched multiple locations in the genome. A slash indicates a locus was not represented in the data set.
expanded genes had ESTs, and then with only one or two representatives for each.

Next, we analyzed the rice massively parallel signature sequencing (MPSS) database (http://mpss.udel.edu/rice/). This database identified 17- or 20-bp sequence tags, each representing the 3' end of a single mRNA detected in transcript libraries isolated from a various tissue and treatments. Here, data were analyzed from the full set of 72 20-base signature libraries, representing 12 different tissue types and plants exposed to a variety of different biotic and abiotic stresses. Significant signatures, which matched only one gene, were found for 11 of the 64 expanded MATH-BTB genes and all four core MATH-BTB genes (Figure 4B). As with the EST analysis, expression of the core group loci was significantly higher than most of the expanded group loci, with 113, 177, 193, and 460 tags per million (tpm) identified in the highest-expressing tissue for individual members of the core group, and 7 to 195 tpm (and eight loci less than 32 tpm) identified in the highest-expressing tissue for each member of the expanded group. Moreover, expression of the four core group loci was identified in most libraries examined, while the expanded group loci were typically restricted to a much smaller subset of tissue/treatments.

Finally, we examined the RICEATLAS whole-genome oligoarray expression data set, which analyzed 43 libraries representing 25 different rice tissues and stages of development (http://plantgenomics.biology.yale.edu/riceatlas). A gene was assigned as expressed if signal intensity exceeded a cutoff value based on negative control oligonucleotides in at least three of four biological replicates for each cell type/stage. Based on this criterion, 38 of 64 expanded MATH-BTB loci were judged as expressed in at least one tissue. Two of the four core MATH-BTB loci were also judged as expressed (Os03g57854 and Os07g26160), but predicted cross-hybridization of the oligonucleotides precluded confirmation of the expression for the other two. Taken together, 48 of 64 (75%) of expanded MATH-BTB loci appear to be expressed. These transcriptionally active genes are widely distributed throughout the expanded MATH-BTB clusters assigned phylogenetically (Figure 2), indicating that expression is not restricted to only a subset of groups in the A1 subfamily.

We also performed similar analysis on the 41 BTB pseudogene loci predicted to be members of the rice A1 subfamily and were surprised to detect low but significant evidence of expression within this group (23 of 41 loci). This transcription was suggested primarily by the RICEATLAS data set with expression of the

---

**Figure 5.** Chromosomal Clustering of the Expanded MATH-BTB Genes and Related Pseudogenes.

The clusters include 54 of the 70 functional expanded MATH-BTB and MATH-related-BTB genes and 32 of 36 pseudogenes based on The Institute for Genomic Research (TIGR) rice pseudomolecules (Osa1, release 4). The positions of the segments in the rice chromosomes 8, 10, and 11 are indicated on the left. An expanded view of this figure with sequence identifiers for each locus is in Supplemental Figure 12 online.
That the divergence within the expanded group MATH domains was caused by reduced purifying selection or increased positive (or diversifying) selection and that this selection was significantly different from the selective pressures on the BTB domain, we evaluated the ratio of nonsynonymous distance (K_{\text{a}}) to synonymous distance (K_{\text{s}}) of the two domains in the core and expanded groups (Graur and Li, 2002). First, NJ trees were constructed for all 135 plant MATH-BTB proteins based on either the BTB or MATH domains alone to identify clades conserved in both trees. From this analysis, we removed 30 sequences from the expanded group and one from the core group where the phylogenetic relationships were different between the MATH and BTB domain analysis. K_{\text{a}}/K_{\text{s}} ratios for each branch of the MATH or BTB domain–based phylogenetic trees were then calculated for the remaining 30 core and 74 expanded MATH-BTB sequences.

As can be seen in Figure 8A, the distributions of the K_{\text{a}}/K_{\text{s}} ratios for the MATH domains from the core and expanded groups were dramatically different, implying that the MATH domains in the two groups were under different selective pressures. Whereas the MATH domain K_{\text{a}}/K_{\text{s}} ratios for the core group were below 0.2 for 27 of 31 branches, the K_{\text{a}}/K_{\text{s}} ratios of the expanded group were above 0.2 for 57 of 63 branches and above 0.4 for 39 branches. In fact, the K_{\text{a}}/K_{\text{s}} ratios for several of the expanded MATH domain branches were significantly >1.0, suggesting strong positive selection of their MATH domains. By contrast, the distribution of K_{\text{a}}/K_{\text{s}} ratios for the BTB domains for the core and expanded groups were more similar (Figure 8A). We note that the distribution of the K_{\text{a}}/K_{\text{s}} ratios suggests the expanded group BTB domains experienced somewhat reduced purifying selection compared with the core group, though much less pronounced than the MATH domain. The K_{\text{a}}/K_{\text{s}} ratios suggest that the MATH domains in the core group were under significantly stronger purifying selection compared with the BTB domains, whereas in the expanded group, purifying selection in the MATH domains has been relaxed compared with the BTB domains. To confirm that this difference between the expanded and core groups in the relative evolutionary rates of the MATH domain compared with the BTB domain is significant, we subtracted the BTB K_{\text{a}}/K_{\text{s}} ratios from the MATH K_{\text{a}}/K_{\text{s}} ratios at each branch and plotted the distributions of the subtracted ratios for the core and expanded groups (Figure 8B). A statistically significant separation of the two groups was evident by nonparametric U-test (P = 0.00005502).

Such comparisons indicated that the MATH domains in the core group were under stronger purifying selection, while the MATH domains in the expanded group were under significantly reduced purifying selection. To investigate this evolutionary dichotomy further, synonymous and nonsynonymous substitution rates (K_{\text{a}}/K_{\text{s}}) at each codon were calculated to identify the selection regimes of individual residues. MATH and BTB domains from the core and monocot expanded groups were
aligned (Figure 7; see Supplemental Figure 8 online), and KA/KS ratios were calculated for each position. In the final analysis, only those positions with residues present in all sequences were used to avoid statistical bias or artifacts generated by alignment gaps.

Out of 46 testable positions in the expanded MATH domain, we detected seven positions (15.2%) with KA/KS ratios significantly above 1.0 (P < 0.1) (Figure 7; see Supplemental Figure 8 online). It is likely that more potentially positively selected sites would be uncovered if we also include regions of the MATH domain with insertions/deletions. Three of the five sequence conservation blocks had at least one position under apparent positive selection. By contrast, none of 113 positions in the core MATH domain was predicted to be under possible positive selection (KA/KS > 1.0, P < 0.1) (Figure 7; see Supplemental Figure 8 online). Given the role of the MATH domain in substrate recognition (Xu et al., 2003; La et al., 2004; Hernandez-Munoz et al., 2005; Kwon et al., 2006), the overrepresentation of positively selected sites in the expanded group suggests pressure for diversification, potentially to encourage recognition of fast evolving targets, while the core group recognizes targets that have had more stable evolutionary histories.

Like the core group MATH domains, no positions in the BTB domains of the core group (out of 105) were predicted to be under positive selection (KA/KS > 1.0, P < 0.1) (see Supplemental Figure 8 online), suggesting that both domains have evolved under the influence of purifying selection. Surprisingly, considering the generally strong sequence conservation in this domain, we did detect six positions with the BTB domain under possible positive selection in the expanded MATH-BTB group (see Supplemental Figure 8 online). While clearly not as obvious as the MATH domain, the BTB domain in the expanded group may also be under diversifying selection.

**DISCUSSION**

The BTB protein superfamily comprises a highly diverse collection of substrate recognition factors that when assembled with CUL3 and RBX1 help promote selective ubiquitination of various eukaryotic intracellular proteins. Prior descriptions of the Arabidopsis BTB proteins (Dieterle et al., 2005; Figueroa et al., 2005; Gingerich et al., 2005) and our analysis of rice homologs presented here show that their signature BTB domains are fused to a wide range of recognition modules presumably capable of identifying similarly varied targets in plants. The types of recognition motifs used are largely conserved between rice and Arabidopsis but are substantially different from those found in animal BTB proteins (Aravind and Koonin, 1999; Stogios et al., 2005). This diversity supports the view of the BTB domain as a self-contained...
members of the expanded MATH-BTB group in rice have been which represent eight of the 16 (Pintard et al., 2004), it should be noted that non-E3 functions for their prime function is to act as target adapters for CUL3 E3s (Dharmasiri et al., 2005a, 2005b; Kepinski and Leyser, 2005). Whether duplicated BTB genes retained identical functions, divided functions, or evolved new ones is not yet known. Sub- or neofunctionalization may be a common feature of plant E3 target recognition factors. For example, EBF1 and EBF2, a pair of LRR-F-box proteins that assemble related SCF E3 complexes in Arabidopsis, both regulate ethylene signaling by targeting the EIN3/EIL1 transcription factors in rice (Ha et al., 2003, 2004; Norberg et al., 2005). Consequently, our phylogenetic comparisons should provide an important template for predicting BTB protein functions and targets once the function of one ortholog is defined.

Our phylogenetic comparisons also revealed that many BTB gene families have undergone significant changes since the monocot/dicot split, with strong evidence for birth-and-death gene evolution (Nei and Rooney, 2005). With the exceptions of the Arabidopsis C4 (BTB-only) subfamily and the rice A1 (MATH-BTB) subfamily, limited rounds of duplication are inferred for each family or subfamily following the separation of the Arabidopsis and rice lineages (one to two rounds), which is consistent with proposed limited large-scale, possibly whole-genome duplications of each species during their evolution (Simillion et al., 2002; Bowers et al., 2003; Paterson et al., 2004; Wang et al., 2005). Whether duplicated BTB genes retained identical functions, divided functions, or evolved new ones is not yet known. Sub- or neofunctionalization may be a common feature of plant E3 target recognition factors. For example, EBF1 and EBF2, a pair of LRR-F-box proteins that assemble related SCF E3 complexes in Arabidopsis, both regulate ethylene signaling by targeting the EIN3/EIL1 transcription factors for breakdown (Guo and Ecker, 2003; Potuschak et al., 2003; Gagne et al., 2004). Consequently, our phylogenetic comparisons should provide an important template for predicting BTB protein functions and targets once the function of one ortholog is defined.

Our phylogenetic comparisons also revealed that many BTB gene families have undergone significant changes since the monocot/dicot split, with strong evidence for birth-and-death gene evolution (Nei and Rooney, 2005). With the exceptions of the Arabidopsis C4 (BTB-only) subfamily and the rice A1 (MATH-BTB) subfamily, limited rounds of duplication are inferred for each family or subfamily following the separation of the Arabidopsis and rice lineages (one to two rounds), which is consistent with proposed limited large-scale, possibly whole-genome duplications of each species during their evolution (Simillion et al., 2002; Bowers et al., 2003; Paterson et al., 2004; Wang et al., 2005). Whether duplicated BTB genes retained identical functions, divided functions, or evolved new ones is not yet known. Sub- or neofunctionalization may be a common feature of plant E3 target recognition factors. For example, EBF1 and EBF2, a pair of LRR-F-box proteins that assemble related SCF E3 complexes in Arabidopsis, both regulate ethylene signaling by targeting the EIN3/EIL1 transcription factors for breakdown (Guo and Ecker, 2003; Potuschak et al., 2003; Gagne et al., 2004). However, kinetic analyses reveal that the resulting SCFEBF1 and SCFEBF2 complexes work in temporally distinct ways to fine-tune EIN3/EIL1 turnover and, thus, ethylene perception (Binder et al., 2007). Likewise, the multiple SCF complexes assembled with the LRR-F-box proteins TIR1 and AFB1-5 play isoform-specific roles in auxin signaling possibly by directing the auxin-dependent ubiquitination of different members of the AUX/IAA family of repressors (Dharmasiri et al., 2005a, 2005b; Kepinski and Leyser, 2005; Walsh et al., 2006).

While most plant BTB genes have not expanded extensively in number, the monocot MATH-BTB genes appear to be notable exceptions. Our analysis of the relatives throughout the plant kingdom identified a small, ancient core group of plant MATH-BTBs. Their progenitors appear to predate bryophytes, indicating that this core type was established >400 million years ago,
during the early course of land plant evolution. MCRA analysis revealed that the last common ancestor of the monocots and dicots likely had three core MATH-BTB genes, a number similar to the family of four and six in the rice and Arabidopsis core groups, respectively (Figure 2). Despite the long evolutionary history of the plants examined here, the MATH domains of these core BTB proteins remained remarkably conserved (even relative to the BTB domain), suggesting that the counterpart sites recognized in their substrates have also been remarkably stable. The strong purifying selection acting on the core MATH-BTB sequences argues that their targets are similarly constrained and likely participate in basic plant cell processes with the MATH domain interface being important to their activities. Thus, it is tempting to speculate that the ubiquitination of these targets by this core MATH-BTB group is intrinsic to their proper function such that strong purifying selection of both binding interfaces was essential to maintain proper contact. Informative paradigms include the AUX/IAA proteins and the EIN3, ABI3, ABI5, and DELLA transcription factors whose functions are intimately intertwined with their turnover by the Ub/26S proteasome system in plants (McGinnis et al., 2003; Gagne et al., 2004; Zhang et al., 2005; Parry and Estelle, 2006; Stone et al., 2006). For AUX/IAA proteins in particular, the site recognized by the cognate F-box proteins (domain II) represents one of the most conserved regions in this family across a wide range of species (Ramos et al., 2001; Goldfarb et al., 2003).

At this time, the substrates of the core plant MATH-BTB proteins are unknown. The MATH-BTB domain configuration is present in animals (Aravind and Koonin, 1999; Geyer et al., 2003; Huang et al., 2004; Stogios et al., 2005), and identified substrates include the katanin AAA-type ATPase protein MEI-1 (Funukawa et al., 2003; Pintard et al., 2003; Xu et al., 2003), Ci/Gli2/Gli3 transcription factors (Zhang et al., 2006), the polycomb protein BMI1, and the MacroH2A histone (Hernandez-Munoz et al., 2005). Unfortunately, the animal MATH domain sequences have not been sufficiently diverged from those in plants to preclude target recognition and binding activity (Aravind and Koonin, 1999; Geyer et al., 2003; Stogios et al., 2005), and identified substrates include the katanin AAA-type ATPase protein MEI-1 (Funukawa et al., 2003; Pintard et al., 2003; Xu et al., 2003), Ci/Gli2/Gli3 transcription factors (Zhang et al., 2006), the polycomb protein BMI1, and the MacroH2A histone (Hernandez-Munoz et al., 2005). Nonetheless, the animal MATH domain sequences have sufficiently diverged from those in plants to preclude target predictions based solely on sequence similarity.

In contrast with the highly conserved core group, we also identified a large and potentially rapidly evolving family of expanded MATH-BTB genes that appear to be monocot specific. Both their presence in sorghum and possibly wheat in addition to rice argues that this group has ancient origins well before the appearance before the rice/sorghum split ~50 million years ago, and therefore considered to be a main avenue for the large-scale expansion of gene families (Zhang, 2003; Hancock, 2005). Higher K_{a}/K_{s} ratios is a general feature of duplicates in tandem arrays in the rice genome compared, for instance, with pairs generated by segmental duplication (Yu et al., 2005) and could be evidence that these tandem genes have been particularly susceptible to rapid evolution following duplication. Likewise, the concentration of related pseudogenes surrounding the apparently functional expanded MATH-BTB genes supports a rapid birth and death of these loci.

Both their lack of introns and extensive clustering within the rice genome suggest that progenitors of the expanded MATH-BTB group in monocots first appeared by retroposition and then expanded by tandem duplications. Given that loci created by this mechanism are often missing elements that direct expression, they are typically pseudogenes. However, at least with respect to the rice MATH-BTB family, it appears that most, if not all, are functional, for several reasons. First, despite their projected appearance before the rice/sorghum split ~50 million years ago, they appear to contain intact coding regions, which have not been disrupted by frame-shift or in-frame stop codons. In fact, we found that several expanded MATH-BTB proteins from different subclades have retained their ability to interact with CUL3s to assemble predicted E3 ligase complexes. Second, analysis of publicly available expression resources suggests as much as 75% of the loci are actively transcribed. Third, alignment and K_{a}/K_{s} analyses clearly demonstrate that the sequences of the expanded MATH-BTB proteins are not uniformly degenerating as would be expected of pseudogenes but are preferentially changing within the MATH domain to include both insertion/deletions and nonsynonymous amino acid substitutions. It is informative to note that some members of the SKP1 gene families in Arabidopsis and rice may have been created by similar events (i.e., retroposition followed by tandem duplication of the retrogenes) (Kong et al., 2007). Likewise, these duplicated SKP genes appear to be functional based on expression analysis and their ability to encode proteins that retain their interaction with CUL1 and F-box proteins to form SCF E3 complexes (Gagne et al., 2002; Risseeuw et al., 2003; Takahashi et al., 2004; Wang et al., 2004b). The results with SKP and expanded MATH-BTB proteins together with recent studies demonstrating that apparent retrogenes are actively transcribed (Boschan et al., 2002; Marques et al., 2005; Dae et al., 2006; Vinckenbosch et al., 2006; Zhao et al., 2007) suggests this type of gene expansion does not a priori synthesize nonfunctional loci, in contrast with previous assumptions (Graur and Li, 2000). For several of the obvious MATH-BTB pseudogenes in rice, several types of expression studies (ESTs/cDNAs, MPSS, and microarrays) indicate that they are still transcribed. Such transcription likely reflects the recent acquisition of mutations in the coding region prior to eventual changes that would impair expression.

The molecular evolution pattern (i.e., small, highly conserved stable sets of genes shared across lineages and lineage-specific groups that have undergone large-scale expansion and gene
loss and are often clustered in the genome) that we observe for the plant MATH-BTB gene family is very similar to those recently published for the MATH-BTB gene family in several Caenorhabditis species, in some FBX subfamilies in Caenorhabditis, Arabidopsis, and rice (Thomas, 2006; Jain et al., 2007), and in the aforementioned Arabidopsis and rice SKP1 gene families (Kong et al., 2007). Such similarities suggest that this type of evolution may be a common feature of E3 Ub-ligase subunit gene families, particularly those encoding the target recognition subunits of the complexes. Thomas (2006) proposed that these distinct family characteristics are driven by different evolutionary pressures exerted by two separate classes of substrates. In particular, the diversifying, expanded groups may recognize proteins themselves under strong diversifying selection pressures. With respect to the potentially diversifying expanded MATH-BTB group in monocots, we note that three of the MATH domain codons found under positive selection were the same ones identified as under positive selection in the Caenorhabditis collection. Based on a structural model of the TRAF6 MATH/RANK peptide ligand complex, Thomas (2006) suggested that these positively selected amino acids participate in ligand binding where diversification is exploited to expand the repertoire of partners. Consequently, it is possible that a similar dynamic may be occurring in this region for the expanded group MATH domains in monocots.

We also detected possible positive selection in the expanded group BTB domains. The significance of this is unclear, though the positively selected sites at positions 65 and 67 do flank amino acid positions important for CUL3–BTB interactions between C. elegans and S. pombe CUL3 and the BTB proteins MEL-26 and BTB3 (Geyer et al., 2003; Xu et al., 2003). However, we note that the four members of the expanded group that we show bind CUL3 contain a diverse set of residues at the six positions identified to be under positive selection, suggesting that these positions may not be important for assembly of the E3 ligase complex.

Diversification of a gene family can be driven by the need of an organism to adapt to a particular environment or to set up reproductive barriers. One classic example of diversifying selective pressure is the highly polymorphic S-locus in the Solanaceae (Clark and Kao, 1991), which encodes components involved in self-incompatibility. Interestingly, a gene within this locus encodes an F-box protein (self-incompatibility S-locus-encoded F-box protein, or SLF/SFB), which is thought to be the pollen self-incompatibility determinant (Qiao et al., 2004; Sijacic et al., 2004; Ushijima et al., 2004). Other recently published examples, which illustrate the diverse range of forces that can drive adaptive diversification in gene families, include the Dr family of adhesions in Escherichia coli, likely changed to provide functional variation in adhesive properties (Korotkova et al., 2007), the taste receptor family 1 (T1R) genes in Gasterosteus aculeatus, possibly altered to discriminate between different taste substances important for survival (Hashiguchi et al., 2007), and the lipase/feruloyl esterase A family in Euascomycetes, which appear to have acquired feruloyl esterase A activity to facilitate plant cell wall degradation in concurrence with the colonization of land by plants (Levasseur et al., 2006).

Gene families involved in host defense and innate immunity are also particularly well represented among those shown to have diversified under positive selection and include the mammalian major histocompatibility complex genes (Hughes and Nei, 1988, 1989; Kelley et al., 2005), the nucleotide binding site-LRR genes (Mondragon-Palomino et al., 2002; Meyers et al., 2003), the Cladosporium fulvum resistance (Parniske et al., 1997; Meyers et al., 1998), and the defense-related receptor-like kinase (Shiu et al., 2004) gene families in plants. The characteristics of these host families (rapid expansion, sequence diversification, and evidence of positive selection), which are shared by the monocot expanded MATH-BTB group, result from participation in an arms race and/or trench warfare coevolution with the pathogen proteins that they recognize (Dawkins and Krebs, 1979; Stahl et al., 1999). Therefore, one possible group of substrates for the expanded group MATH-BTB proteins, as also suggested by Thomas (2006) for the MATH-BTB family in Caenorhabditis, could be pathogen proteins that are themselves under strong positive selection to avoid detection and inactivation by the host.

Consistent with a role of MATH-BTB proteins in host defense, an increasing number of reports have appeared in the last few years supporting a central role for host-directed ubiquitination in plant defense responses (Kawasaki et al., 2005; Gonzalez-Lamothe et al., 2006; Yang et al., 2006; Goritschnig et al., 2007). Alternatively, it has become clear that pathogens have also developed sophisticated methods to exploit or interfere with the host Ub system during infection (Munro et al., 2006; Angot et al., 2007). These methods even include introducing their own E3s, likely to target intracellular host proteins involved in pathogen surveillance for ubiquitination and subsequent turnover (Schrammeyer et al., 2001; Tzira et al., 2004; Abramovitch et al., 2006; Angot et al., 2006; Janjusevic et al., 2006; Nomura et al., 2006). A role of various E3 types in innate immunity may in turn help explain the preponderance of E3 genes in plants relative to other eukaryotes (Vierstra, 2003; Smalle and Vierstra, 2004). While we cannot rule out a role for these expanded MATH-BTB proteins in other adaptive functions, their rapid evolution in monocots as part of an innate immunity system may reflect another intriguing mechanism of defense in the continuing battle between pathogen and host.

METHODS

Identification of Rice Genes Encoding BTB Domain Proteins

The SMART database (http://smart.embl-heidelberg.de) was used to locate the core BTB/POZ domain in 48 BTB proteins from a variety of organisms (Caenorhabditis elegans, Saccharomyces cerevisiae, Drosophila melanogaster, Schizosaccharomyces pombe, Mus musculus, Arabidopsis italiana, and humans). These amino acid sequences were used as queries in BLASTP searches for possible homologs encoded by the Oryza sativa spp japonica cv Nipponbare Build 3 genome available in the TIGR Rice Annotation database (http://rice.tigr.org/). These queries recovered 177 nonredundant sequences below an E-value cutoff of 9.4e-5. This cutoff value was sufficient to eliminate random sequences and was more stringent than similar domain-based searches in Arabidopsis and rice (see Gagne et al., 2002; Shiu and Bliek, 2003; Shiu et al., 2004). Rice gene/protein annotations were checked and refined manually and then rechecked and reconciled with the TIGR Build 4 rice genome. For Os11g01175 and Os10g29501, MATH-BTB coding regions were split off from existing annotations (Os11g01170 and Os10g29510).
and predicted to be separate genes. TIGR reannotation after the release of Build 4 subsequently eliminated the predicted MATH-BTB coding regions from Os10g29510 (renumbered as Os10g29502) and Os11g41170. In three cases, our hand analysis split single loci in the database into two separate BTB-encoding genes (Os08g40495 split from Os08g40490, Os08g42135 from Os08g42130, and Os11g41315 from Os11g41310), resulting in a final collection of 180 predicted BTB protein sequences.

BLASTP searches of the rice Build 4 genome were repeated with SMART and PFAM (http://www.sanger.ac.uk/Software/Pfam/) predicted BTB domains from each of 180 predicted BTB loci. These searches recovered all previously identified sequences (with the exception of the two loci noted above and two more [Os03g13880 and Os11g40480] where reannotation following the release of Build 4 had incorrectly eliminated the predicted BTB domain), and an additional 14 loci with scores beneath the 9.4e^-6 cutoff. This process was repeated a third time with all 194 sequences; no additional sequences were recovered beneath the cutoff value. Finally, 11 representative rice BTB domains were used as tBLASTn queries against the six-frame translated rice (O. sativa) genome.

In addition to a number of random partial BTB encoding sequences, these searches recovered one additional, previously unannotated locus that was subsequently predicted to encode a BTB protein by hand analysis (numbered Os02g52313). Final hand analysis removed three loci (Os08g13130, Os07g44570, and Os04g53820) from the family because they were predicted to encode only part of the degenerate BTB domains, resulting in a final collection of 192 rice genes. Additional BLAST searches against the protein database with all 192 predicted BTB domains recovered no additional sequences beneath the cutoff score.

Hand analysis of the rice BTB family identified numerous adjustments and refinements of TIGR-predicted annotations (see Supplemental Data Set 1 online for the revised BTB protein sequences). The predicted BTB domain itself of each individual protein was refined by sequence alignments and hand analysis. Of the 192 BTB domain-containing rice sequences, SMART and PFAM predicted BTB coding domains in all but six loci. Three of these are missing part of the BTB-encoding region and have been categorized as pseudogenes. BTB domains in the remaining three (H-subfamily members Os04g20920, Os07g15600, and Os12g08720) were defined by alignments and hand analysis. Two separate BTB domains were recognized in three loci (Os06g21330, Os05g27880, and Os11g41260); the BTB domain with the lowest SMART or PFAM E-value was used in this study. We also classified 43 loci (41 predicted MATH-BTB and two predicted BTB-NPH3 subfamily members) as pseudogenes based on the definition of a pseudogene as the clear presence of a coding sequence disrupted by frame shift(s) or an in-frame stop codon(s).

Identification of Plant Genes Encoding MATH-BTB Proteins

Full-length MATH-BTB protein sequences from rice (this study) and Arabidopsis (Gingerich et al., 2005) were used as queries in BLASTP and tBLASTn searches to locate possible homologs in other plant species. Physcomitrella patens and Selaginella moellendorffii sequences were identified from the Department of Energy Joint Genome Institute raw whole genome shotgun sequences at databases http://moss.nibb.ac.jp/ and http://selaginella.genomics.purdue.edu/cgi-bin/blast_tmpl_s.cgi. Populus trichocarpa sequences were found in the assembled v.1.0 genome (www.jgi.doe.gov/poplar; Tuskan et al., 2006). Pinus taeda sequences were identified from EST databases (http://www.plantgdb.org/ and http://www.tigr.org/tigr-scripts/tgi/T_index.cgi?species=pinus). Medicago truncatula sequences were found in the assembled genome v.1.0 (http://www.medicago.org/genome/). Sorghum bicolor sequences were identified from ESTs and genomic survey sequences (http://www.plantgdb.org/). A single MATH-BTB gene was identified in a BAC sequence from wheat (Triticum monococcum) (SK14, accession number AF88415).

Alignment and Phylogenetic Analysis

Full-length sequences, BTB domains, and MATH domains were aligned using ClustalW (Chenna et al., 2003) and manually edited in Jalview (Clamp et al., 2004). Midpoint rooted phylogenetic trees were generated in MEGA3.1 (Kumar et al., 2004) by the NJ method, using the Poisson distance method, pairwise deletion of gaps, and the default assumptions that the substitution patterns among lineages and substitution rates among sites were homogeneous. The reliability of the trees was assessed by 1000 bootstrap replicates. The Arabidopsis-rice MP tree was generated in MEGA3.1 with a 1000x bootstrap replicate. The data set was tested using the close-neighbor-interchange method with a search level of 1. Initial trees for close-neighbor-interchange searches were built by random additional with 10 replicates. All sites in the alignments were used. The tree represents a consensus generated from 29 equally parsimonious trees. Amino acid sequence alignments were calculated and displayed using MACBOXSHADE v2.15 (Institute of Animal Health, Pirbright, UK). Additional domains were predicted by SMART and PFAM, BLAST searches, sequence alignments, and the COILS algorithm (http://www.ch.embnet.org/software/COILS_form.html).

Expression Analysis

Expression of the rice MATH-BTB loci was evaluated using the rice EST collection, the rice MPSS database (http://mpss.udel.edu/rice/), or the RICEATLAS whole-genome oligoarray expression data set (http://bioinformatics.med.yale.edu/rc/overview.jsp). ESTs were identified in the TIGR Rice Transcript Assembly v.1 databases and from BLAST searches of the Gramene (www.gramene.org) EST database. MPSS tags were identified from the 20-bp libraries and were only used if they uniquely identified one gene. Signal intensities were used from the RICEATLAS data set only if they exceeded a cutoff value based on negative control spots in at least three of four biological replicates for each cell type/stage and only if the sequence uniquely matched one gene. Levels of expression were categorized based on the number of ESTs/FLcDNAs [low [one to three], moderate [four to eight], and high [>8]], the number of transcripts per million (tpm) in the MPSS collection [low [1 to 25 tpm], moderate [26 to 150 tpm], and high [>151 tpm]], or normalized signal intensity (RICEATLAS) [low [0 to 500], moderate [501 to 2000], and high [>2000]] in the library with the highest level expression for that locus.

Evolutionary Association between MATH and BTB Domains

Initial phylogenetic trees of the MATH and BTB domains were separately constructed for the expanded MATH, core MATH, expanded BTB, and core BTB categories in MEGA3.1 (Kumar et al., 2004) by the NJ method, using the Poisson distance method, complete deletion of gaps, and the default assumptions that the substitution patterns among lineages and substitution rates among sites were homogeneous. The reliability of the trees was assessed by 1000 bootstrap replicates. Phylogenies generated with the MATH and BTB domains were compared, and branches with consistent topologies were manually chosen for further analysis. NJ phylogenetic trees of each cluster were constructed with NEIGHBOR in PHYLIP3.66 (distributed by J. Felsenstein, University of Washington, Seattle) from evolutionary distances calculated in PROTDIST using the Jones-Taylor-Thornton distance matrix (Jones et al., 1992). Ancestral sequences were inferred for all the nodes by the maximum likelihood method (Yang, 1997). Synonymous (Ks) and nonsynonymous (Ka) distances were estimated in all the branches by the modified Nei-Gojobori method (Zhang et al., 1998). The ratio of nonsynonymous distance to synonymous distance (Ka/Ks) for MATH and BTB domains was then calculated for each branch. Ka/Ks ratios of the branches were compared between MATH and BTB domains. A nonparametric U-test was used to determine the significance of the different Ka/Ks ratios between the core and expanded groups.
Inference of Positively Selected Amino Acid Sites

Multiple alignments were independently constructed for the expanded MATH, core MATH, expanded BTB, and core BTB domains by ClustalW (Chenna et al., 2003) and manually edited in Jalview (Clamp et al., 2004). The number of sequences used for each category were 104 (expanded MATH), 31 (core MATH), 104 (expanded BTB), and 31 (core BTB) (see Supplemental Figure 8 online). Positively selected amino acid sites were identified by a modification of that described by Suzuki and Gojobori (1999). A phylogenetic tree was first reconstructed by the NJ method, and the ancestral sequence was inferred at each node using the maximum likelihood method (Yang, 1997). Then, the average number of synonymous (SS) and nonsynonymous (SN) sites and the total number of substitutions, respectively. A significantly larger value of CN over CS (P < 0.1) was used to infer positive selection.

Y2H Analyses

The full coding regions for At CUL1, At CUL3a, and At CUL3b were obtained by RT-PCR from Arabidopsis ecotype Col-0 or PCR amplified from full-length ESTs provided by the ABRC. Truncated Os CUL3b was obtained by RT-PCR of RNA isolated from rice (O. sativa spp japonica cv Nipponbare) plants. The MATH-BTB genes were amplified from rice genomic DNA. Coding regions were reamplified with primers designed to add 33 to 36 additional nucleotides complementary to the pGADT7-rec and pGBK7 vectors to facilitate in-frame insertion of the products by yeast homologous recombination. The inserts and the pGBK7-rec (linearized with EcoRI and NotI) and pGADT7-rec (linearized with SmaI) vectors were transformed into the appropriate yeast strains at a 3:1 ratio using the Frozen-EZ Yeast Transformation II kit (Zymo Research).

The CUL and MATH-BTB coding regions in the respective pGBK7 and pGADT7-rec vectors were transformed into the haploid yeast strains YP82a and LB414 (Gray et al., 1999). All plasmids were verified as correct by DNA sequence analyses. Yeast were mated on YPAD for 24 h at 30°C and then grown on complete supplemental medium (Q-BIOgene) minus Leu and Trp. Growth was tested on complete supplemental medium containing 7 mM 3-amino-1,2,3-triazole but without Leu, Trp, and His. For each mating, 5 μL of cells with an OD600 of 10^{-2} were spotted and grown for 4 d at 23°C.

Supplemental Data

The following materials are available in the online version of this article.

**Supplemental Figure 1.** Phylogenetic Tree of the Complete BTB Protein Superfamilies in Rice and Arabidopsis.

**Supplemental Figure 2.** Sequence Alignment of all 229 BTB Domains from the Arabidopsis and Rice BTB Proteins.

**Supplemental Figure 3.** Sequence Alignments of all 149 and 80 BTB Domains from the Rice and Arabidopsis BTB Proteins, Respectively.

**Supplemental Figure 4.** Maximum Parsimony Consensus Tree of the Complete BTB Protein Superfamilies in Rice and Arabidopsis.

**Supplemental Figure 5.** Phylogenetic Trees of BTB Protein Families and Subfamilies in Rice and Arabidopsis Generated with Full-Length Protein Sequences.

**Supplemental Figure 6.** Phylogenetic Tree of the Complete BTB Protein Superfamilies in Rice and Arabidopsis, Including Predicted Pseudogenes.

**Supplemental Figure 7.** Sequence Alignments of All 135 MATH and BTB Domains from the Land Plant MATH-BTB Proteins.

**Supplemental Figure 8.** Inference of Positively Selected Sites in Core and Expanded Group MATH and BTB Domains.

**Supplemental Figure 9.** Conservation of Intron/Exon Junctions of Representative Core Group MATH-BTB Genes.

**Supplemental Figure 10.** Phylogenetic Trees of the Complete BTB Protein Superfamilies in Rice and Arabidopsis.

**Supplemental Figure 11.** Phylogenetic Trees of 135 MATH-BTB Protein Sequences from Representative Land Plant Species.

**Supplemental Figure 12.** Chromosomal Clustering of the Expanded MATH-BTB Genes and Related Pseudogenes.

**Supplemental Table 1.** Orthology Relationships of Arabidopsis and Rice MATH-BTB Sequences.

**Supplemental Table 2.** Sequence Designations and Predicted MATH-BTB Protein Sequences from Medicago truncatula, Physcomitrella patens, Populus trichocarpa, Pinus taeda, Sorghum bicolor, Selaginella moellendorffii, and Triticum monococcum.

**Supplemental Data Set 1.** Revised Predicted Protein Sequences in the Rice BTB Superfamily Based on Gene Reannotations.

**Supplemental Data Set 2.** Expression Analysis for Rice MATH-BTB Core and Expanded Groups and Rice A1 Subfamily Pseudogene Loci.

ACKNOWLEDGMENTS

Rice MATH-BTB expression data were obtained from the RICEATLAS Project website at http://plantgenomics.biology.yale.edu/riceatlas (under the support of the National Science Foundation Plant Genome Program Award DBI-0235821) with acknowledgments to Tim Nelson, Xing-Wang Deng, and Hongyu Zhao for providing access. Our work was supported by grants from the National Science Foundation Arabidopsis 2010 Program (MCB-01155870) to R.D.V. and the Plant Genome Comparative Sequencing Program (DBI-0638591) to S.H.S. and a National Institutes of Health postdoctoral fellowship to D.J.G. (F32-GM68361).

Received February 22, 2007; revised July 25, 2007; accepted August 3, 2007; published August 24, 2007.

REFERENCES


