EVIDENCE FOR A LARGE-SCALE, LINEAGE-SPECIFIC EXPANSION OF A BRIC-A-BRAC/TRAMTRACK/BROAD COMPLEX (BTB) UBIQUITIN-LIGASE GENE SUBFAMILY IN MONOCOTS.

Running Title: BTB Ubiquitin-Protein Ligases in Rice

By Derek J. Gingerich¹, Kou Hanada², Shin-Han Shiu², Richard D. Vierstra¹,³

¹Department of Genetics, 425 Henry Mall, University of Wisconsin-Madison
Madison, WI 53706, USA

²Department of Plant Biology, S-306 Plant Biology Building, Michigan State University
East Lansing, MI 48824-1312 USA

³ Corresponding author: Department of Genetics, 425 Henry Mall
University of Wisconsin-Madison, Madison, WI 53706, USA
(608)-262-8215, office; (608)-262-4743, fax; vierstra@wisc.edu, email

Keywords: ubiquitin, protein degradation, Arabidopsis, rice, MATH-BTB, phylogeny
ABSTRACT

Selective ubiquitination of proteins is directed by diverse families of ubiquitin-protein ligases (or E3s) in plants. An important E3 type uses one of several Cullins (CULs) as scaffolds to assemble multisubunit E3 complexes containing a RBX1 protein and one of a myriad of substrate recognition factors. We have previously shown that CUL3 in association with individuals within the 80-member Bric-a-Brac/Tramtrack/Broad Complex (BTB) protein superfamily form essential BTB-E3 complexes in Arabidopsis thaliana. Here, we describe the complete BTB superfamily in rice (Oryza sativa spp. japonica cv Nipponbare) that contains 149 BTB-domain containing 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 loss events after the rice and Arabidopsis BTB lineages split, suggesting specialization within individual BTB families. Most dramatically, a major expansion and diversification in the rice Meprin-and-TRAF-homology (MATH)-BTB family is evident that likely occurred following the appearance of monocots. A large set of monocot-specific MATH-BTB proteins in particular appears to be evolving faster within their substrate-recognition module than the smaller core set that may predate flowering plants. This rapid evolution raises the possibility that monocot MATH-BTB E3s are diversifying to ubiquitinate a set of targets that are themselves rapidly changing.

Gingerich et al., 2
INTRODUCTION

Covalent attachment of ubiquitin (Ub) to specific proteins is an important mechanism for post-translational control in both plants and animals. Its best-known function is to target specific proteins to (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 Ubs, primarily linked internally through Lys48. The resulting poly-ubiquitinated proteins are then recognized by the 26S proteasome, a 2-MDa 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. Often these substrates are either monoubiquitinated or polyubiquitinated through lysines other than Lys48. Via these proteolytic and non-proteolytic 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 response, photomorphogenesis, floral development, self incompatibility, disease resistance, environmental adaptation, circadian rhythms, and programmed cell death to name a few (Smalle and Vierstra, 2004).

The functions of Ub are primarily determined by the reaction cascades that attach the first Ub and then assemble the poly-Ub 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 ligase (or E3s) (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 (mono versus polyubiquitination through specific lysines). Not surprisingly, genomic analyses reveal that large collections of E3s exist to handle the myriad of expected intracellular substrates. In Arabidopsis thaliana, Caenorhabditis elegans, and mice for example, over 1,300, 500, and 400 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. One of a variable collection of substrate-binding adapter proteins delivers appropriate targets to this CUL/RBX subcomplex. Two large families of CUL-based E3s in plants are the SCF (Skp1, CDC53, and F-Box (or FBX)) and BTB (Bric-a-Brac/Tramtrack/Broad Complex) complexes. For SCF E3 complexes, the adaptor moiety includes of an F-Box protein that directly binds the target. It associates with one of a family of SKPs (or ASKs in Arabidopsis thaliana) through its signature F-Box motif; the SKP protein in turn links the heterodimer to the CUL1/RBX1 subcomplex. For BTB E3 complexes, the BTB protein is the substrate adaptor; it has both the substrate recognition site and a signature BTB domain that promotes interaction with the CUL3/RBX1 subcomplex. The F-Box and BTB proteins have an array of protein-protein interaction motifs, most of which are presumed to participate in target recognition, including kelch, leucine-rich repeat (LRR), armadillo, ankyrin, tryptophan-aspartic acid (WD)-40, tryptophan-tryptophan (WW), Tubby, lectin-binding, tetratricopeptide repeats (TPR), and Meprin-and-TRAF-homology (MATH) motifs (Aravind and Koonin, 1999; Gagne et al., 2002; Risseeuw et al., 2003; Gingerich et al., 2005; Stogios et al., 2005).

The F-Box and BTB proteins in yeast (Saccharomyces cerevisiae) are encoded by relatively small gene families (21 (Willems et al., 2004) and 3 members (Geyer et al., 2003), respectively), suggesting a limited repertoire of targets. However, in many multicellular eukaryotic lineages, these genes families have expanded greatly in size and complexity. For instance the BTB and FBX genes comprise superfamilies in the eudicot Arabidopsis (80 and ~700 members, respectively (Gagne et al., 2002; Gingerich et al., 2005), Caenorhabditis elegans (105 and 326 members, respectively (Furukawa et al., 2003; Geyer et al., 2003; Willems et al.,
Gingerich et al., 2004), and humans (208 and 109 members, respectively (Furukawa et al., 2003; Geyer et al., 2003; Willems et al., 2004)). Family-specific expansion is also evident in certain lineages. For example, the Drosophila melanogaster F-Box protein superfamily has remained comparatively small (30-31 members) while the BTB protein superfamily includes 141 members (Furukawa et al., 2003; Geyer et al., 2003; Willems et al., 2004). Comparisons among the superfamilies also showed that the plant and animal kingdoms underwent very different evolutionary paths (Winston et al., 1999; Andrade et al., 2001; Gagne et al., 2002). A large collection of BTB proteins in vertebrates, for example, include zinc finger and kelch motifs, combinations that are completely absent in higher plants like Arabidopsis. There is also evidence for large-scale expansions within specific BTB protein subtypes. C. elegans, for instance, has a much larger MATH-BTB family than is present in either insects or vertebrates (Stogios et al., 2005) and there have even been independent expansions of different subsets of the MATH-BTB protein family within individual Caenorhabditis species (Thomas, 2006). One hypothesis to explain this diversity is that each lineage mixed and matched various substrate-binding domains with different E3 complex-interacting motifs, and then expanded and diversified specific subgroups to handle their own particular sets of Ub targets as they evolved.

To begin to understand the complexity and evolution of the substrate adaptor components of E3s in plants and to help identify individual E3s that would direct general versus species-specific functions, we initiated a phylogenetic analysis of the BTB superfamily in monocot rice (Oryza sativa) and compared it to our previous analysis of the BTB superfamily in Arabidopsis (Gingerich et al., 2005). These species diverged from common ancestor approximately 150 mya (Wikstrom et al., 2001), and thus represented an opportunity to identify general sets of E3 adapters (and their targets) that were preserved in these two angiosperm species as well as those that emerged after the monocot/dicot split. Here, we describe the 149-member BTB protein superfamily 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 MATH-BTB family in monocots. This expansion is combined with evidence of both rapid birth-death evolution and diversifying selection of a large collection of MATH-BTB proteins, suggesting that a subset of rice of BTB E3s is rapidly changing to cope with targets that are also rapidly changing.

**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 (C. elegans, S. cerevisiae, D. melanogaster, S. pombe, Mus musculus, A. thaliana, and humans). These amino acid sequences were used as queries in BLASTP searches for possible homologs encoded by the O. sativa spp. japonica cv Nipponbare Build 3 genome available in the TIGR Rice Annotation database (http://rice.tigr.org/). These queries recovered 177 non-redundant sequences below an E-value cutoff of 9.4e-6. This cutoff value was sufficient to eliminate random sequences and was more stringent than similar domain-based searches in Arabidopsis and rice, including those for F-Box proteins (see (Gagne et al., 2002; Shiu and Bleecker, 2003; Shiu et al., 2004)). Gene/protein annotations were checked and refined by hand, and then rechecked and reconciled with the TIGR Build 4 rice genome. For Os11g41175 and Os10g29501, MATH-BTB coding regions were split off from existing annotations (Os11g41170 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 2 loci noted above and 2 more (Os03g13860 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 cut-off 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, 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 Suppl. Dataset 1 for the revised BTB protein sequences). The BTB domain itself 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, Os12g08720) were defined by alignments and hand analysis. Two separate BTB domains were recognized in three loci (Os06g21330, Os05g27880, 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 2 predicted BTB-NPH3 subfamily members) as
pseudogenes, based on the definition of a pseudogene as the clear presence of a coding sequence disrupted by frameshift(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 DOE 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 genome (v.1.0) (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). Sorghum bicolor sequences were identified from ESTs and genomic survey sequence (http://www.plantgdb.org/). Single MATH-BTB genes were identified in the bacterial artificial chromosome sequences from the wheat (Triticum monococcum) (5K14, accession #AF88415), and Medicago truncatula (mth2-10F21, accession #AC148965) (http://www.ncbi.nlm.nih.gov/blast/) genome databases.

Alignment and Phylogenetic Analysis

Full-length sequences, BTB domains, and MATH domains were aligned using CLUSTALW (Chenna et al., 2003) and hand-edited in Jalview (Clamp et al., 2004). Unrooted phylogenetic trees were generated in MEGA3.1 (Kumar, 2001 #176) by the NJ method, using the Poisson distance method and pairwise deletion, and a 1000x bootstrap replicate. The Arabidopsis-rice maximum parsimony tree was generated in MEGA3.1 with a 500x bootstrap replicate. Amino acid sequence alignments were calculated and displayed using MACBOXSHADE v2.15 (Institute of Animal Health, Pirbright, U.K.). 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).

**Evolutionary Association between MATH and BTB Domains**

Phylogenetic trees of the MATH and BTB domains were separately constructed for expanded MATH, core MATH, expanded BTB, and core BTB categories, using the NJ method and protein distances based on the Jones-Taylor-Thornton matrix (http://evolution.genetics.washington.edu/phylip.html). Genes phylogenetically clustered via the MATH and BTB domains were manually chosen by comparison of the separate MATH and BTB domain phylogenetic trees. We constructed new phylogenetic trees for each cluster and inferred ancestral sequences of all the nodes by the maximum likelihood method (http://abacus.gene.ucl.ac.uk/software/paml.html). Synonymous ($K_S$) and nonsynonymous ($K_A$) distances were calculated in all the branches by the modified Nei-Gojobori method (Zhang et al., 1998). The ratio of nonsynonymous distance to synonymous distance ($K_A/K_S$) for MATH and BTB domains was then estimated for each branch. First, $K_A/K_S$ ratios of the branches were compared between MATH and BTB domains. Second, to compare the different selective pressures of the two domains between the core and expanded categories, the deviation of MATH $K_A/K_S$ ratio from BTB $K_A/K_S$ ratio was estimated in all the branches. The difference of the deviations was statistically examined by the two-tailed T-test.

**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). The number of sequences used for each category were 102 (expanded MATH), 28 (core MATH), 108 (expanded BTB), and 29 (core BTB) (see Suppl. Fig. ??). The amino acid sites without any gaps were used in the following analyses. Positively selected amino acid sites were identified
by a modification of that described by Suzuki and Gojobori (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 (http://abacus.gene.ucl.ac.uk/software/paml.html). Then, the average number of synonymous (S_S) and nonsynonymous (S_N) sites and the total number of synonymous (C_S) and nonsynonymous (C_N) substitutions throughout the phylogenetic tree were estimated for each amino acid site by the modified Nei-Gojobori method (Zhang et al., 1998). The probability (P) of obtaining the observed or more biased number of synonymous and nonsynonymous substitutions was computed for each amino acid site assuming a binomial distribution. In the computation, S_S/(S_S + S_N) and S_N/(S_S + S_N) were used as the expected probabilities of synonymous and nonsynonymous substitutions, respectively. The significance level was set at 0.1. A significantly larger value of C_N over C_S was used to infer positive selection.

RESULTS

Identification and Characterization of the Rice BTB Superfamily

Prior phylogenetic analysis of the CUL family in rice (Oryza sativa) identified a set of thirteen CUL-type proteins with three (OsCUL3a-c) showing strong similarity to Arabidopsis CUL3a/b (Gingerich et al., 2005). In particular, the adaptor interface predicted to be used by AtCUL3a/b to bind BTB proteins was well conserved in OsCUL3a-c, strongly suggesting that OsCULa-c also help assemble BTB E3 complexes in rice. To identify the array of CUL3/BTB complexes that potentially exist in rice, we defined the full complement of BTB proteins in this species. Reiterative BLAST searches of Oryza sativa spp. japonica cv Nipponbare sequence database using 48 sequences encompassing the ~110-amino acid core BTB domain (Gingerich et al., 2005; Stogios et al., 2005) from yeast, plants, and animals recovered 192 open-reading frames (ORFs) encoding one or more BTB domains. Subsequent analysis of this set categorized 43 loci as putative pseudogenes, based on the presence of an in-frame stop codon or frameshift(s)
disrupting the coding region (see below). After removing these pseudogenes, a final family of 149 functional genes was predicted to encode BTB proteins. 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 evolutionary split of monocots and dicots approximately 150 Mya (Wikstrom et al., 2001).

Like previous descriptions of the *Arabidopsis* BTB protein superfamily (Dieterle et al., 2005; Gingerich et al., 2005), analysis of the rice BTB sequences both up and downstream of the BTB domains identified a collection of other protein-protein interaction motifs that likely represent substrate recognition motifs, including ankyrin repeats, armadillo repeats, MATH, coiled-coil, Transcriptional Adaptor Zinc Finger (TAZ), and tetratricopeptide (TPR) repeats (Fig. 1). Like *Arabidopsis*, the rice superfamily also contained a large collection (25 members) with the plant-specific NPH3 domain, a ~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 a 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 any previously described domains outside of the BTB domain in 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) and data not shown). Two of the predicted rice BTB proteins (Os11g02070, Os12g02030) are significantly smaller shorter (275 residues in length, each) and appear to contain just the BTB domain (Fig. 1). Such BTB-only proteins have been identified previously in *S. 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 strikingly similar sets, with only six rice BTB proteins (Os01g70670, Os09g16850, Os09g16870, Os10g28990, Os11g40670, and Os11g41260) and a single *Arabidopsis* BTB protein (At1g04390) predicted to contain architectures not present in the other species. When compared to other eukaryotes, the rice and *Arabidopsis* superfamilies are strikingly distinct. Whereas *Arabidopsis* and rice have BTB domains connected to Armadillo, TPR, NPH3, TAZ, and F 5/8 type C motifs, none of these combinations have been detected in animals or yeast. Conversely, both rice and *Arabidopsis* lack the BTB-zinc finger and BTB-BACK-kelch combinations that comprise a substantial percentage of the vertebrate BTB collections (Aravind and Koonin, 1999; Prag and Adams, 2003; Stogios et al., 2005). Rice appears more like *C. elegans* with respect to the MATH-BTB family (Huang et al., 2004; Stogios et al., 2005; Thomas, 2006), which is substantially larger relative to *Arabidopsis*, yeast and vertebrates (Fig. 1).

One feature of the rice BTB superfamily was 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 97-99% amino acids sequence identity and likely arose from a well-documented recent (4-14 mya) 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 recognize substrates widely distributed in plants (or even eukaryotes) versus those that recognize targets specific to rice or *Arabidopsis*, phylogenetic trees were generated with all 149 rice and 80 *Arabidopsis* members either separately or together (Fig. 2 and Suppl. Fig. 1). The trees were generated with the BTB domain alone by the distance-based neighbor-joining (NJ) method with MEGA3.1 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 outer branches allowed us to group the *Arabidopsis* and rice BTB proteins into distinct families. The trees consistently clustered proteins with similar BTB domains and target-binding motifs, thus providing independent support for the groups (Fig. 2, Suppl. Fig. 1). Similar trees were also generated by character-based maximum parsimony (MP) analysis (Suppl. Fig. 2). While 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; ambiguity in the interior branches did not cluster all members of this subfamily into a single distinct clade in the MP tree.

For simplicity, the NJ trees were further subdivided into alphabetical families and subfamilies based on both tree topologies and predicted protein domain structures (Fig. 2). Even in cases where SMART failed to predict additional motifs (*e.g.*, B4, C3, E1, and H subfamilies), alignments of the rice and *Arabidopsis* subfamilies confirmed that their members are also well-conserved outside of the BTB domain sequence. 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 had substantially different domain structures, suggesting 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 to others in the cluster. While it is possible that they are pseudogenes, the putative coding regions harbored no obvious in-frame stops or frameshifts and thus appear to be functional loci. Two genes in the rice A1 subfamily are unusual in possibly containing an additional domain linked to the MATH-BTB sequence. One was predicted to encode a BTB-MATH-BTB configuration (Os11g41260), while the other may encode a MATH-BTB sequence linked to a transposase family tnp2 motif (Os10g28990).
Most rice and *Arabidopsis* BTB subfamilies clustered phylogenetically and contained similar numbers of genes, suggesting that they were derived from the same ancestral genes and that major lineage-specific expansions have not occurred. For instance, all 3 Armadillo repeat-BTB proteins (2 *Arabidopsis*, 1 rice) clustered in subfamily B1, all 6 TPR-BTB proteins (3 *Arabidopsis*, 1 rice) clustered in subfamily B2, and all 3 pentapeptide repeat-BTB proteins (1 *Arabidopsis*, 2 rice) clustered in subfamily B3. However, two notable exceptions are obvious (Fig. 2). One is an expansion of the BTB-only type (C4 subgroup) in *Arabidopsis*, which contains 8 members versus only 2 in rice. The second is the dramatic expansion and separation of the MATH-BTB family in rice. Whereas *Arabidopsis* encodes only 6 MATH-BTB proteins, rice encodes at least 75 (Fig. 2) with another 41 loci predicted to be MATH-BTB pseudogenes (see below). In the rice/*Arabidopsis* combined tree, most of the rice MATH-BTB proteins (71 members) formed a large rice-specific group (A1) distinct from the remaining 4 rice members and all 6 *Arabidopsis* MATH-BTB proteins which clustered together in a separate A2 subfamily (Suppl. Fig. 1). This separation implied that the MATH-BTB family is divided to two separate evolutionary groups, one small set that is common to both rice and *Arabidopsis* (designated “core”), and a second larger set resulting from a rice-specific, large-scale expansion (designated “expanded”).

To better understand how the BTB families evolved independently in the rice and *Arabidopsis* lineages, we inferred the number of *BTB* genes in the most recent common ancestor (MRCA) based the rice/*Arabidopsis* NJ BTB domain tree (Suppl. Fig. 1). 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. (Suppl. Fig. 3). Well-supported clades (bootstrap value ≥65%) containing both *Arabidopsis* and rice sequences were first identified in the trees (black dots, Suppl. Figs 1, 2). Combination and reconciliation of the clades identified 41 groups where a single progenitor *BTB* gene in the MRCA appeared to generate both rice and *Arabidopsis*
descendants. At least one of these clades is 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. In addition, we identified 11 rice or Arabidopsis-specific genes or groups of genes with well-supported relationships (bootstrap value ≥65%) with the shared clades (grey dots, Suppl. Figs. 1 and 3). Each of these may reflect an additional ancestral gene in the MRCA where the descendents(s) were lost in one of the two lineages. In addition there are 3 rice-specific groups (A1, A3, and G) and one Arabidopsis specific-group (E2), which 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 subgroup, a minimum of 56 BTB genes in the MRCA was estimated. When compared to the actual 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 rice total is even larger, if the 43 BTB-related pseudogenes are included.

The phylogenetic relationships also suggested orthologous relationships between a number of rice and Arabidopsis proteins, which could in turn direct similar functions (Suppl. Table 1). For example, 18 rice and Arabidopsis BTB sequences have a one-to-one correspondence where a single BTB protein with the same overall architecture from each species clusters. Included in this list are (i) Arabidopsis ARIA, which assists in ABA responses (Kim et al., 2004) and groups with the rice protein Os05g33050; (ii) NPH3 (RPT3), which helps mediate blue-light photoresponses (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); (iii) ETO1 which targets the ethylene biosynthetic type-2 1-aminocyclopropane-1-carboxylate synthases for breakdown (Wang et al., 2004) and groups with rice Os03g18360; and (iv) AtBT3, 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 diversification and possible functional subspecialization in one lineage. As examples, *Arabidopsis* BOP1 and BOP2, which act redundantly to regulate development of lateral organs by repressing class 1 *knox* and *JAGGED* transcription factor gene expression (Ha et al., 2003; Ha et al., 2004; Norberg et al., 2005), are co-orthologs of a single rice BTB protein (Os01g72020), and At4g26120 and NPR1, which regulates pathogen-response gene expression (Dong, 2004), are co-orthologs of a single rice protein, OsNH1.

**The Rice Genome Contains Numerous 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 also encoded a MATH domain and clustered phylogenetically by NJ analysis of the BTB domain with the A1 subclade containing the rice-specific expanded set of MATH-BTB proteins (Suppl. Fig. 4). Even though an additional four pseudogenes did not include obvious MATH domains, they also fell within the A1 subgroup phylogenetically, suggesting that they also descended from an A1 progenitor and either the MATH sequence was eliminated or degenerated beyond recognition. These 40 pseudogenes were distributed throughout the A1 subfamily tree, suggesting a dynamic history of gene duplication and loss that are characteristic of birth-death gene evolution (Nei and Rooney, 2005). We were unable to reconstruct the full BTB domains from the remaining three pseudogenes for phylogenetic analysis. However, alignments, BLAST best hits, and partial coding region reconstruction suggest that Os08g12970 is part of the A1 subfamily, and Os05g12030 and Os12g31320 are part of the F (BTB-NPH3) subfamily. We also note that tBLASTn searches recovered 8 locations in the rice genome where short ORFs encoded ~20-50 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 include homologs from other monocots (sorghum, *Sorghum bicolor* and wheat, *Triticum monococcum*) and dicots (*Medicago truncatula* and popular, *Populus trichocarpa*), a gymnosperm (pine, *Pinus taeda*), a moss (*Physcomitrella patens*) and a bryophyte (*Selaginella moellendorffii*). These sequences were identified in both the genomic and EST sequence databases, using the rice and *Arabidopsis* MATH-BTB proteins as queries. The number of MATH-BTB loci for each species and their gene designations can be found in Figure 3 and Supplemental Table 2. Whereas the MATH-BTB loci appeared to be intact for the other species, 27 examples of apparent MATH-BTB pseudogenes are evident in sorghum, suggesting this family has experienced similar evolutionary dynamics as their rice MATH-BTB A1 subfamily counterparts (data not shown).

When all 142 predicted MATH-BTB proteins were analyzed phylogenetically by NJ analysis using either the BTB and MATH domains alone, we notice a striking separation into two clades that followed the core and expanded rice groups (Fig. 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 4 from rice. The 71 rice MATH-BTB sequences in the expanded group was joined by 41 of the 43 MATH-BTBs from sorghum and the single wheat protein, demonstrating that the expanded group is likely to be monocot specific. This separation was particularly apparent in the MATH domain-based NJ tree and implied that the MATH domain specifically had distinct evolutionary histories in core and monocot expanded groups (Figs 3). In fact the distinction was even more obvious in sequence alignments (Suppl. Figs. 5-8). Whereas the MATH domain alignment of
the core group had no gaps and only 24 out of 113 positions with less than 80% sequence identity, the alignment of the expanded group revealed numerous insertions in several positions and identified 152 out of 162 positions with less than 80% sequence identity. In contrast, the BTB domain alignments of the core (41/131 positions with >80% identity) and expanded groups (39/170 positions with >80% identity) displayed more similar levels of diversity.

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 4 exons (Fig. 3). The only exceptions are the two from Physcomitrella where only two exons are evident. Remarkably, the positions of the intron/exon are junctions absolutely-conserved in this highly diverse collection of plant species (Fig. 4A and Suppl. Fig. 9). In contrast, the coding regions of a majority of the expanded rice MATH-BTB loci, along with 40 of the 41 sorghum and the single wheat MATH-BTB loci contained with a single predicted exon (Fig. 3), a configuration not detected in any of the non-monocot species. The estimated expression levels for the expanded and core genes also differed (Fig. 4B). Searches of the EST databases for rice and Arabidopsis identified numerous cDNAs for each member of the core set (10-39 sequences/gene) that were derived from a range of tissue types (Fig. 4B), implying that the core MATH-BTB genes are ubiquitously expressed at relatively high levels. In contrast, the rice expanded MATH-BTB group had much fewer ESTs (1-9 sequences per gene) with most having none (56 of 71 members), suggesting that they are not well expressed or have highly restricted expression patterns (Fig. 4B). A cursory analysis of the sorghum EST collection provided similar results; numerous ESTs were available for the 2 core MATH-BTB genes whereas only 4 of the 41 expanded genes had ESTs, and then with only 1 or 2 representatives each (data not shown).

Differences between the core and expanded MATH-BTB proteins were also evident based on chromosomal locations. Unlike a majority of the BTB genes in Arabidopsis and rice which appear as singletons ((Gingerich et al., 2005) and data not shown), most of the expanded
**BTB** genes (A1 subfamily) occur in tandem duplication blocks in rice. These blocks include 57 of 71 functional expanded **MATH-BTB** genes and 32 of 36 **MATH-BTB** pseudogenes (Fig. 5). As first noticed by Song *et al.* (Song *et al*., 2002), the largest blocks are the middle of chromosome 10. Here, 25 functional and 12 **MATH-BTB** pseudogenes (out of 69 total predicted open-reading frames) are clustered within a 325-kb region and 6 functional and 4 pseudogenes (out of 27 open-reading frames) are clustered in nearby 132-kb region (Fig. 5). Smaller arrays of expanded **MATH-BTB** genes were also detected in chromosomes 8 and 11. These blocks often contain a high concentration of transposable element (TE) coding sequences whose movements have been proposed to drive gene duplication events (Hancock, 2005).

**Monocot Expanded MATH-BTB Genes are Under Positive Selection**

Alignments of the MATH and BTB domains from the core and expanded groups clearly revealed a preferential divergence within a subfamily of MATH-BTB proteins in monocots, particularly in the target-recognition MATH domain (Fig. 6, Suppl. Figs. 5-8). To test if this divergence was a result of reduced level of purifying selection (*i.e.*, increased frequency of nonsynonymous codon changes versus synonymous codon changes (Hughes and Nei, 1988)), we first calculated the ratio of nonsynonymous distance (Kₐ) to synonymous distance (Kₛ) of each domain in each group. NJ trees were constructed for all 142 plant MATH-BTB proteins based on either the BTB or MATH domains to identify clades conserved in both trees. From this analysis, we removed 50 sequences from the expanded group and 8 from the core group where the phylogenetic relationships were ambiguous or different between the MATH and BTB domain analysis. Kₐ/Kₛ ratios for the full MATH and BTB domains at each node were calculated for the remaining 21 core and 63 expanded MATH-BTB sequences.

As can be seen in Figure 7A and B, the Kₐ/Kₛ ratios of the MATH domain were lower then that of the BTB domains at most nodes of the core group while the majority of nodes for the monocot expanded group had higher Kₐ/Kₛ ratios, suggesting that MATH domains are under
different selective pressures in the two groups. To further support this scenario, we subtracted the BTB $K_a/K_s$ ratios from the MATH $K_a/K_s$ ratios at each branch and graphed the distributions for the core and expanded groups (Fig. 7B). A statistically significant separation of the two groups was evident by two-tailed t-test ($P=0.000972$). Comparisons of absolute $K_a/K_s$ values suggested the MATH domains in the expanded group are under significantly reduced purifying selection (median $K_a/K_s = 0.4980$) as compared to the MATH domains of the core group (median $K_a/K_s = 0.0839$). In fact, the $K_a/K_s$ ratio for several of the expanded MATH domain nodes was greater than 1.0, indicating possible diversifying/positive selection of the MATH domain, while most of the nodes in the core group had extremely low $K_a/K_s$ ratios (19 of 22 nodes ≤ 0.2), suggesting strong purifying selection. These $K_a/K_s$ ratio differences between the core and expanded groups were largely eliminated when the BTB domain was used for the calculations. Here, the median $K_a/K_s$ values of the two were similar (0.2759 for the expanded and 0.2002 for the core) with the slight difference suggesting that the BTB domains of the expanded group may be under slightly reduced purifying selection.

The differences in $K_a/K_s$ ratios between the MATH domains of the core and monocot expanded group could be driven simply by relaxation of purifying selection, or alternatively by positive selection (adaptive evolution) of the MATH domain from the expanded group. To investigate this further, synonymous and nonsynonymous substitution rates ($K_a/K_s$) at each codon were calculated to identify individual residues under positive selection. MATH and BTB domains from the core and monocot expanded groups were aligned (Suppl. Figs. 5-8) and $K_a/K_s$ ratios were calculated for each position. In the final analysis, only those residues present in all individuals were used to avoid statistical bias or artifacts generated by alignment gaps. Out of 40 testable positions in the expanded MATH domain, we detected 7 positions (17.5%) that appear to have been under positive selective pressure ($K_a/K_s >1.0$, $P>0.9$), with 2 more possible ($K_a/K_s >1.0$, $P>0.8$) (Fig. 7, Suppl. Fig. 7). More sites were likely had we included regions of the MATH domain that were excluded by the present of insertions/deletions in some

Gingerich et al., 20
representatives. Four of the five sequence conservation blocks had at least one position under apparent positive selection. In contrast, only 1 of 113 positions (0.09%) in the core MATH domain was predicted to be under possible positive selection (K_a/K_s > 1.0, P ≥ 0.8) (Fig. 6, Suppl. Fig. 5). Similarly low numbers of residues under positive selection were evident in the BTB domains of the core and expanded groups (Suppl. Fig 6,8). We detected only 1 out of 70 (1.4%) positions in the expanded group BTB domains under likely positive selection (Ka/Ks ratios > 1.0, P ≥ 0.9), though with six other positions possible (Ka/Ks ratios > 1.0, P ≥ 0.8). For the core group, only 1 of the 101 positions in the BTB domain may have been under positive selection (K_a/K_S > 1, P ≥ 0.8). Coupled with the high level of sequence conservation and overall low K_a/K_s ratios, we propose that purifying selection, and not adaptive evolution, has been the major force in the development of the core MATH-BTB group. This pressure is primarily seen in the MATH domain but some may also exist in the BTB domain. In contrast, the MATH domains in the expanded group have diversified under the apparent influence of positive selection. Given the role of the MATH domain in target recognition, one obvious possibility is that the MATH domains from the expanded group are diversifying to detect targets that are also under positive selection pressure.

DISCUSSION
The BTB protein superfamily comprises a highly diverse collection of target recognition factors that help promote selective ubiquitination of various eukaryotic intracellular proteins when assembled with CUL3-based E3 complexes. Prior descriptions of the Arabidopsis BTB proteins (Dieterle et al., 2005; Gingerich et al., 2005) and our analysis of rice homologs present here show that BTB proteins consist of a signature BTB domain fused to a wide range of different substrate recognition modules presumably capable of identifying similarly varied targets in plants. While the types of recognition motifs used are largely conserved between rice and Arabidopsis, they are substantially different from those found in animal BTB proteins. This
diversity supports the evolutionary view of the BTB domain as a self-contained CUL3-interaction module, which can be linked to a wide variety of other motifs to handle the sets of substrates in the different lineages (Aravind and Koonin, 1999; Stogios et al., 2005). While their prime functions are to act as target adapters for CUL3 E3 Ub-ligases (Pintard et al., 2004), it should be noted that non-E3 functions for several animal BTB proteins have been proposed via their dimerization or association with other non-CUL proteins (Stogios et al., 2005). To date, 16 different Arabidopsis BTB proteins have now been shown to interact with CUL3a/b, which represent 8 of the 16 Arabidopsis subfamilies (Wang et al., 2004; Dieterle et al., 2005; Figueroa et al., 2005; Gingerich et al., 2005; Weber et al., 2005)(D. J. Gingerich, unpublished data). Consequently, it is likely that most, if not all, plant BTB proteins act as CUL3-E3 target adapters for selective ubiquitination. Furthermore, based on BTB protein diversity in Arabidopsis and rice, it appears that plants have the capacity to ubiquitinate a number of distinct groups of proteins, some of which may be plant specific.

The overall repertoire of BTB protein types has been conserved between the rice and Arabidopsis superfamilies, suggesting the same general classes of proteins are targets. In fact, clear one-to-one orthologous relationships are evident for 18 loci, suggesting that the rice orthologs have the same targets as their Arabidopsis counterparts. Unfortunately, not enough BTB E3 targets are currently known in plants to confirm this premise. However, genetic analysis of Arabidopsis NPH3 and its rice ortholog CPT1 indicate that both participate in blue light perception, possibly via ubiquitination of the same target(s) (Motchoulski and Liscum, 1999; Inada et al., 2004; Haga et al., 2005). Likewise, we expect that rice Os05g33050 assists in ABA perception like Arabidopsis ARIA (Kim et al., 2004), that rice Os03g18360 controls the levels of type 2 ACC synthases like Arabidopsis ETO1 (Wang et al., 2004), and that rice Os01g72020 plays a role in controlling the expression of class I knox and JAG/JGL gene expression in rice like Arabidopsis BOP1 and 2 (Ha et al., 2003; Ha et al., 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/eudicot split, with strong evidence for birth-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 subfamily following the monocot/dicot split (1-2 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). Following duplication, unneeded genes were lost, while advantageous duplicates were retained. Whether this occurred because of a corresponding amplification of the targets and/or because the duplicates divided functions or evolved new ones is unknown. Such sub- and 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 (Gagne et al., 2004). However, kinetic analyses reveal that the resulting SCF_{EBF1} and SCF_{EBF2} complexes work in temporally distinct ways to fine tune EIN3/EIL1 turnover and thus ethylene perception. Likewise, the multiple SCF complexes assembled with the LRR-F-Box proteins TIR1 and AFB1-5 have isoform specific roles in auxin signaling by directing the auxin-dependent ubiquitination of members of the AUX/IAA family of auxin repressors (Dharmasiri et al., 2005a; Dharmasiri et al., 2005b; Kepinski and Leyser, 2005; Walsh et al., 2006).

While most BTB genes show limited evolutionary histories, the evolution of the MATH-BTB genes is particularly striking. Our extensive analysis of the relatives throughout the plant kingdom identified a small, ancient core group of plant MATH-BTBs. Their progenitor(s) appear to predate bryophytes, suggesting that this core type has been present at least since the origin
of land plants. MCRA analysis suggests that the last common ancestor of the monocots and eudicots likely had 3 core MATH-BTB genes, a number similar to the present family of 4 and 6 in the rice and Arabidopsis core groups, respectively (Fig. 2). Despite of 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 site(s) recognized in their binding partners have also been remarkably stable. Such strong purifying selection suggests that the target(s) participate in a basic plant cell process with the BTB-binding site being important to their activit(ies). Thus, it is tempting to speculate that the ubiquitination of these targets by this core BTB group is intrinsic to their proper function such that purifying selection of both binding interfaces is essential. Informative paradigms include the AUX/IAA proteins and EIN3, ABI3, and DELLA transcription factors whose functions are intimately intertwined with their turnover by the ubiquitin/26S proteasome system in plants (McGinnis et al., 2003; Zhang et al., 2005; Parry and Estelle, 2006) (Gagne et al., 2004). For AUX/IAA proteins in particular, the F-Box binding site (Domain II) represents one of the most conserved regions of these proteins across a wide range of species (Ramos et al., 2001; Goldfarb et al., 2003). At this time, the target(s) of the core plant MATH-BTB proteins is 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), however the core plant MATH domains are quite diverged from the animal MATH domains (data not shown). Therefore the known targets of animal MATH-BTBs, including the katanin AAA-type ATPase protein MEI-1 (Furukawa et al., 2003; Pintard et al., 2003; Xu et al., 2003), and Ci/Gli2/Gli3 transcription factors (Zhang et al., 2006) may not be substrates in plants.

In contrast to the highly conserved core group, we also identified a large and rapidly evolving family of expanded MATH-BTB proteins that appear to be monocot specific. This expanded group has all the hallmarks of a rapid birth-death evolution combined with diversifying/postive selection, suggesting strong pressure for amino acid sequence changes.
While most of the other rice BTB proteins are scattered throughout the genome, the 71 members of the expanded MATH-BTB are often clustered, with many present in tandem duplication blocks. Such blocks may reflect unequal crossing over, considered to be one cause of large-scale expansions of gene families (Zhang, 2003; Hancock, 2005). Given that the expanded MATH-BTB gene collection typically has a single-exon gene structure, it is possible that they first appeared as retrogens derived from a retrotransposition-mediated duplication of a core MATH-BTB gene parent. Since the expanded MATH-BTB genes from both sorghum (41 representatives) and wheat (1 representative) have the same single exon structures, the appearance of this retrogene would have occurred before the sorghum/rice split approximately 50 mya, (Kellogg, 2001)). A high proportion of the MATH-BTB group are pseudogenes (41 of 112 predicted expanded MATH-BTB genes) and likely represent unneeded or deleterious paralogs. As opposed to other BTB proteins, including those in the core MATH-BTB group, the expanded MATH-BTB group are represented at low levels in the EST databases, suggesting that they are either expressed in highly temporal or tissue-specific manners or some could be pseudogene precursors that are no longer expressed but have retained intact gene structures.

Coupled with a large increase in number, the expanded MATH-BTB gene family has undergone a high rate of sequence diversification, particularly within the MATH domain. This diversification may result from periods of reduced purifying selection or stronger diversifying selection, as evidenced by the higher fraction of non-synonymous (K_a) to synonymous (K_S) mutations detected for the monocot expanded group MATH domains, compared to the core MATH domains. Higher K_a/ K_S ratios is a general feature of duplicates in tandem arrays in the rice genome, as compared, for instance, to pairs generated by segmental duplication (Yu et al., 2005) and could be evidence that the expansion processes which result in these genomic arrangements create genes which are particularly susceptible to rapid evolution following duplication. In support, we detected 7 sites of likely positive selection in the expanded group MATH domain. Interestingly, two of these sites which are immediately downstream of a highly
conserved tryptophan correspond exactly to two sites identified as under positive selection in the MATH domains of a group of unstable nematode MATH-BTB proteins (Thomas, 2006). Based on a structural model of the TRAF6 MATH/RANK peptide ligand complex, Thomas predicted these positively-selected sites may be ligand-binding sites where diversification is used to recognize different partners (Thomas, 2006). Consequently, it is possible that a similar dynamic may be occurring in this region of the monocot expanded group MATH domains.

The molecular evolution pattern that we observe in the plant MATH-BTB family is very similar to recently published data on the MATH-BTB gene family and some F-box subfamilies in Caenorhabditis species (Thomas, 2006), with stable and unstable groups that correspond very closely in characteristics to our plant MATH-BTB core and expanded groups. Thomas suggested that these family characteristics (small groups of stable genes well-conserved across lineages and much larger, non-conserved groups resulting from rapid lineage-specific expansions) may result from the evolutionary pressures exerted by two distinct classes of target proteins: the stable proteins target endogenous targets involved in developmental control for degradation and the unstable proteins from the expanded families target exogenous proteins, possibly as part of a defense system (Thomas, 2006). Many of the characteristics of the expanded MATH-BTB group (large-scale expansion and gene loss, clustering in the genome, signatures of positive selection) are shared with other host defense gene families, including mammalian MHC genes (Hughes and Nei, 1988, 1989; Kelley et al., 2005), the defense-related RLK subfamilies (Shiu et al., 2004), the plant NBS-LRR family (Mondragon-Palomino et al., 2002; Meyers et al., 2003), and the Cladosporium fulvum resistance gene family (Parniske et al., 1997; Meyers et al., 1998). The characteristics of these families result from their rapid, dynamic evolution as they participate in an “arms race” with the pathogen proteins that they target. Interestingly, the one Arabidopsis F-box protein which has been implicated in defense responses (SON1) (Kim and Delaney, 2002) falls within a group that has experienced a large, lineage-specific expansion, relative to rice (Gingerich, unpublished data). That the expanded
monocot MATH-BTB proteins target foreign proteins for degradation is speculative, but it may not be unexpected considering increasing evidence ubiquitination has a key role in plant defense responses (Kawasaki et al., 2005; Gonzalez-Lamothe et al., 2006; Yang et al., 2006; Goritschnig et al., 2007) and that plant pathogens manipulate the host Ub system during infection (Abramovitch et al., 2006; Angot et al., 2006; Janjusevic et al., 2006; Navarro et al., 2006; Nomura et al., 2006).

FIGURE LEGENDS

Figure 1. Grouping of the BTB subfamilies based on the nature of additional motifs outside of the BTB domain, along with a protein diagram of representative members.

Figure 2. Phylogenetic trees of the complete BTB protein superfamilies in rice and Arabidopsis. Alignments of the ~110 amino-acid BTB domains were generated in CLUSTALW and refined by hand. The alignments were used to generate non-rooted phylogenetic trees in MEGA3.1, using the Poisson distance method and a bootstrap value of 1,000. The subfamilies identified from the phylogenetic analysis are marked on the bottom. Individual members of the tree are color-coded by the nature of the domains either N- or C-terminal to the BTB domain. Where possible, designations for proteins previously identified by other methods are marked on the bottom. Closed circles indicate rice-specific subfamilies. The closed diamond indicates an Arabidopsis-specific subfamily. Arrowheads indicate BTB proteins shown to interact with CUL3a and/or CUL3b. Expanded views of the trees with branches labeled with sequence identifiers and bootstrap values are in Supplemental Fig. 10.

Figure 3. Phylogenetic trees of 142 MATH-BTB protein sequences from representative land plant species. Alignments of the ~110 amino-acid BTB domains or the ~110 amino-acid MATH domain were generated in CLUSTALW and refined by hand. The alignments were used to
generate non-rooted phylogenetic trees in MEGA3.1, using the Poisson distance method and a bootstrap value of 1,000. The branches of the BTB and MATH domain trees are color-coded by the species. The BTB domain tree is also color coded to indicate the number of exons in the coding region of each gene. Expanded views of the trees with branches labeled with sequence identifiers are in Supplemental Figs. 11-13.

Figure 4. Analysis of the core and expanded MATH-BTB groups in land plants

(A) Gene structure diagrams for representative MATH-BTB genes. Black or grey boxes denote exons, white boxes UTRs, and solid lines indicate introns. Dashed lines indicate sequence encoding corresponding regions of the MATH-BTB proteins, based on alignments. At, Arabidopsis thaliana; Mt, Medicago truncatula; Os, Oryza sativa; Pp, Physcomitrella patens; Pt, Pinus taeda; Sb, Sorghum bicolor; Sm, Selaginella moellendorffii.

(B) Numbers of ESTs in the public databases for members of the Arabidopsis and rice core and expanded groups. Counts for each gene were taken from the TAIR Associated Transcript and TIGR Rice Transcript Assembly v.1 databases.

Figure 5. Chromosomal locations of 57 of 71 functional expanded MATH-BTB genes and 32 of 36 MATH-BTB pseudogenes, based on TIGR rice pseudomolecules (Osa1, release 4).

Figure 6. Inference of positively selected sites in core and expanded MATH domains.

Alignments of 28 core and 102 expanded ~110 amino-acid MATH domains were generated in CLUSTALW and displayed with MacBoxshade using a threshold of 55% (see Suppl. Figs. 5 and 7). Representative sequences from the alignments are shown here. Conserved and similar amino acids are shown in black and grey boxes, respectively. Dots denote gaps. The histograms show maximum-likelihood $K_a/K_s$ ratios that were calculated for each gap-free position in the alignments. The dotted lines indicate a $K_a/K_s$ ratio of 1.0. Sites under likely
positive selection (P ≥ 0.9) are marked with black asterisks. Sites under possible positive selection (P ≥ 0.8) are marked with grey asterisks. At, Arabidopsis thaliana; Mt, Medicago truncatula; Os, Oryza sativa; Pp, Physcomitrella patens; Pit, Pinus taeda; Pot, Populus trichocarpa; Sb, Sorghum bicolor; Sm, Selaginella moellendorfii.

**Figure 7.** KA/KS ratio analysis of plant core and monocot expanded group MATH and BTB domains.

(A) Phylogenetic trees of phylogenetically clustered MATH-BTB sequences were constructed and the ratio of nonsynonymous (KA) to synonymous (KS) distance was calculated in each branch. Each closed circle (expanded group) or square (core group) represents the estimated MATH and BTB domain KA/KS ratios for one branch in the trees.

(B) Distribution of MATH domain KA/KS ratios minus BTB domain KA/KS ratios for each branch for the core (black boxes) and expanded (grey boxes) groups. Two-tailed T-test (P=0.000972) indicates the relationship of selective pressures between the MATH and BTB domains is significantly different between core and expanded groups.

**Supplemental Figure 1.** Phylogenetic tree of the complete BTB protein superfamilies in rice and Arabidopsis. An alignment of all 229 ~110 amino-acid BTB domains was generated in CLUSTALW and refined by hand. The alignment was used to generate a non-rooted phylogenetic tree in MEGA3.1, using the Poisson distance method and a bootstrap value of 1,000. The subfamilies identified from the phylogenetic analysis are marked on the right. Individual members of the tree are color-coded by the nature of the domains either N- or C-terminal to the BTB domain. Where possible, designations for proteins previously identified by other methods are marked on the tree. Black asterisks indicate rice-specific subfamilies. The grey asterick indicates an Arabidopsis-specific subfamily. Black triangles indicate rice sequences, grey triangles indicate Arabidopsis sequences. Black dots indicate an inferred gene.
in the most recent common ancestor (MRCA) which gave rise to at least one descendant currently present in each species. Grey dots indicate an inferred gene in the MRCA giving rise to at least one descendant in one of the two species.

**Supplemental Figure 2.** Maximum-parsimony (MP) consensus tree of the complete BTB protein superfamilies in rice and *Arabidopsis*. An alignment of all 229 ~110 amino-acid BTB domains was generated in CLUSTALW and refined by hand. The alignment was used to generate non-rooted phylogenetic trees in MEGA3.1 by MP analysis with a bootstrap value of 500. The dataset was tested using the Close-neighbor-interchange (CNI) method with a search level of 1. Initial trees for CNI searches were built by random additional with 10 replicates. The tree is a consensus tree generated from 14 equally parsimonious trees. All sites in the alignments were used. The subfamilies identified from the phylogenetic analysis are marked on the right. Individual members of the tree are color-coded by the nature of the domains either N- or C-terminal to the BTB domain.

**Supplemental Figure 3.** Phylogenetic trees of BTB protein subfamilies in rice and *Arabidopsis* generated with full-length protein sequences. Subfamily alignments of predicted full-length protein sequences were generated in CLUSTALW. The alignments were used to generate non-rooted phylogenetic trees in MEGA3.1, using the Poisson distance method and a bootstrap value of 1,000. The subfamily designations are marked to the left of each tree. Black dots indicate an inferred gene in the most recent common ancestor (MRCA) which gave rise to at least one descendant currently present in each species. Grey dots indicate an inferred gene in the MRCA giving rise to at least one descendant in one of the two species.

**Supplemental Figure 4.** Phylogenetic tree of the complete BTB protein superfamilies in rice and *Arabidopsis*, including predicted pseudogenes. The BTB domains of 40 of 43 predicted rice
pseudogenes were reconstructed from genomic sequence. An alignment of all 269 (229 functional + 40 pseudogene) ~110 amino-acid BTB domains was generated in CLUSTALW and refined by hand. The alignment was used to generate a non-rooted phylogenetic tree in MEGA3.1, using the Poisson distance method and a bootstrap value of 1,000. The subfamilies identified from the phylogenetic analysis are marked on the right. Individual members of the tree are color-coded by the nature of the domains either N- or C-terminal to the BTB domain. Black triangles indicate rice sequences, grey triangles indicate *Arabidopsis* sequences. Black dots indicate pseudogenes.

**Supplemental Figure 5.** Inference of positively selected sites in core group MATH domains.

An alignment of 28 core ~110 amino-acid MATH domains was generated in CLUSTALW and displayed with MacBoxshade using a threshold of 55%. Conserved and similar amino acids are shown in black and grey boxes, respectively. The histograms show maximum-likelihood $K_a/K_s$ ratios that were calculated for each gap-free position in the alignments. The dotted line indicates a $K_a/K_s$ ratio of 1.0. One site under possible positive selection ($P \geq 0.8$) is marked with a black asterisk. *At*, *Arabidopsis thaliana*; *Mt*, *Medicago truncatula*; *Os*, *Oryza sativa*; *Pp*, *Physcomitrella patens*; *Pit*, *Pinus taeda*; *Pot*, *Populus trichocarpa*; *Sb*, *Sorghum bicolor*; *Sm*, *Selaginella moellendorffii*.

**Supplemental Figure 6.** Inference of positively selected sites in core group BTB domains.

An alignment of 29 core ~110 amino-acid BTB domains was generated in CLUSTALW and displayed with MacBoxshade using a threshold of 55%. Conserved and similar amino acids are shown in black and grey boxes, respectively. Dots denote gaps. The histograms show maximum-likelihood $K_a/K_s$ ratios that were calculated for each gap-free position in the alignments. The dotted line indicates a $K_a/K_s$ ratio of 1.0. One site under possible positive selection ($P \geq 0.8$) is marked with a black asterisk. *At*, *Arabidopsis thaliana*; *Mt*, *Medicago*
truncatula; Os, Oryza sativa; Pp, Physcomitrella patens; Pit, Pinus taeda; Pot, Populus trichocarpa; Sb, Sorghum bicolor; Sm, Selaginella moellendorffi.

**Supplemental Figure 7.** Inference of positively selected sites in expanded group MATH domains. An alignment 102 expanded ~110 amino-acid MATH domains was generated in CLUSTALW and displayed with MacBoxshade using a threshold of 55%. Conserved and similar amino acids are shown in black and grey boxes, respectively. Dots denote gaps. The histograms show maximum-likelihood $K_a/K_s$ ratios that were calculated for each gap-free position in the alignments. The dotted line indicates a $K_a/K_s$ ratio of 1.0. Sites under likely positive selection ($P \geq 0.9$) are marked with red asterisks. Sites under possible positive selection ($P \geq 0.8$) are marked with black asterisks. Os, Oryza sativa; Sb, Sorghum bicolor; Tm, Triticum monococcum.

**Supplemental Figure 8.** Inference of positively selected sites in expanded group BTB domains. An alignment 108 expanded ~110 amino-acid BTB domains was generated in CLUSTALW and displayed with MacBoxshade using a threshold of 55%. Conserved and similar amino acids are shown in black and grey boxes, respectively. Dots denote gaps. The histograms show maximum-likelihood $K_a/K_s$ ratios that were calculated for each gap-free position in the alignments. The dotted line indicates a $K_a/K_s$ ratio of 1.0. A site under likely positive selection ($P \geq 0.9$) is marked with a red asterick. Sites under possible positive selection ($P \geq 0.8$) are marked with black astericks. Os, Oryza sativa; Sb, Sorghum bicolor; Tm, Triticum monococcum.

**Supplemental Figure 9.** Conservation of intron/exon junctions of representative core group MATH-BTB genes. Exon nucleotide sequences immediately flanking the introns are shown, along with the encoded amino acid residues, below each. At, Arabidopsis thaliana; Mt,
Medicago truncatula; Os, Oryza sativa; Pp, Physcomitrella patens; Pot, Populus trichocarpa; Sb, Sorghum bicolor; Sm, Selaginella moellendorffii.

**Supplemental Figure 10.** Phylogenetic trees of the complete BTB protein superfamilies in rice and Arabidopsis. Alignments of the ~110 amino-acid BTB domains were generated in CLUSTALW and refined by hand. The alignments were used to generate non-rooted phylogenetic trees in MEGA3.1, using the Poisson distance method and a bootstrap value of 1,000. The subfamilies identified from the phylogenetic analysis are marked on the bottom. Individual members of the tree are color-coded by the nature of the domains either N- or C-terminal to the BTB domain. Where possible, designations for proteins previously identified by other methods are marked on the bottom. Closed circles indicate rice-specific subfamilies. The closed diamond indicates an Arabidopsis-specific subfamily. Arrowheads indicate BTB proteins shown to interact with CUL3a and/or CUL3b.

**Supplemental Figure 11.** Phylogenetic tree of 142 MATH-BTB protein sequences from representative land plant species. An alignment of the ~110 amino-acid BTB domain was generated in CLUSTALW and refined by hand. The alignment was used to generate a non-rooted phylogenetic tree in MEGA3.1, using the Poisson distance method and a bootstrap value of 1,000. The branches of the tree are color-coded by species.

**Supplemental Figure 12.** Phylogenetic tree of 142 MATH-BTB protein sequences from representative land plant species. An alignment of the ~110 amino-acid MATH domain was generated in CLUSTALW and refined by hand. The alignment was used to generate a non-rooted phylogenetic tree in MEGA3.1, using the Poisson distance method and a bootstrap value of 1,000. The branches of the tree are color-coded by species.
**Supplemental Figure 13.** Phylogenetic tree of 142 MATH-BTB protein sequences from representative land plant species. An alignment of the ~110 amino-acid BTB domain was generated in CLUSTALW and refined by hand. The alignment was used to generate a non-rooted phylogenetic tree in MEGA3.1, using the Poisson distance method and a bootstrap value of 1,000. The branches of the tree are color-coded by the number of exons in the coding regions of each gene.

**Supplemental Table 1.** Orthology relationships of *Arabidopsis* and rice MATH-BTB sequences. Predictions of orthologous relationships were based on phylogenetic analysis of the *Arabidopsis* and rice BTB superfamilies using the BTB domain and full-length predicted protein sequences (Suppl. Figs. 1 and 3). Unresolved relationships result from ambiguity or conflicting results between the individual analyses.

**Supplemental Table 2.** Sequence designations and predicted MATH-BTB protein sequences from *Medicago truncatula* (Mt), *Physcomitrella patens* (Pp), *Populus trichocarpa* (Pot), *Pinus taeda* (Pit), *Sorghum bicolor* (Sb), *Selaginella moellendorffii* (Sm), and *Triticum monococcum* (Tm).

**Supplemental Dataset 1.** Revised predicted protein sequences in the rice BTB superfamily, based on gene reannotations.

**BIBLIOGRAPHY**


Walsh, T.A., Neal, R., Merlo, A.O., Honma, M., Hicks, G.R., Wolff, K., Matsumura, W., and Davies, J.P. (2006). Mutations in an auxin receptor homolog AFB5 and in SGT1b confer resistance to synthetic picolinate auxins and not to 2,4-dichlorophenoxyacetic acid or indole-3-acetic acid in Arabidopsis. Plant Physiol 142, 542-552.


### Arabidopsis

<table>
<thead>
<tr>
<th># of Genes</th>
<th>Gene #</th>
<th>BTB</th>
<th>NPH3</th>
<th>CC</th>
<th>MATH</th>
<th>Ank</th>
<th>TAZ/CaM BD</th>
<th>TPR</th>
<th>NPH3/ coiled-coil</th>
<th>MATH</th>
<th>Ankyrin</th>
<th>TAZ Zn Finger</th>
<th>F5/8 type C</th>
<th>Ankyrin</th>
<th>BTB-only</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>23</td>
<td>At5g64330</td>
<td></td>
<td></td>
<td></td>
<td>MATH</td>
<td>Ank</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>At1g67900</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>At3g43700</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>NPR1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>At3g48360</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>ETO1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>At5g19330</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>At5g55000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>At2g30600</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>At2g04740</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>At2g40450</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Rice

<table>
<thead>
<tr>
<th># of Genes</th>
<th>Gene #</th>
<th>BTB</th>
<th>NPH3</th>
<th>CC</th>
<th>MATH</th>
<th>Ank</th>
<th>TAZ/CaM BD</th>
<th>TPR</th>
<th>NPH3/ coiled-coil</th>
<th>MATH</th>
<th>Ankyrin</th>
<th>TAZ Zn Finger</th>
<th>F5/8 type C</th>
<th>Ankyrin</th>
<th>BTB-only</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>Os03g55830</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Os03g10800</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>73</td>
<td>Os07g01140</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Os03g46440</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Os01g68020</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Os03g18360</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Os05g33050</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Os03g06160</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Os05g27880</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Os06g21330</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Os11g02070</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 1. Gingerich et al.
Figure 2. Gingerich et al.
Figure 3. Gingerich et al.
Figure 4. Gingerich et al.
Figure 5. Gingerich et al.
Figure 6. Gingerich et al.
Figure 7. Gingerich et al.