Theor Appl Genet (2013) 126:1305–1319
DOI 10.1007/s00122-013-2054-4

ORIGINAL PAPER

Genome-wide identification and analysis of the B3 superfamily of transcription factors in Brassicaceae and major crop plants

Fred Y. Peng · Randall J. Weselake

Received: 28 June 2012 / Accepted: 9 January 2013 / Published online: 2 February 2013

Abstract The plant-specific B3 superfamily of transcription factors has diverse functions in plant growth and development. Using a genome-wide domain analysis, we identified 92, 187, 58, 90, 81, 55, and 77 B3 transcription factor genes in the sequenced genome of *Arabidopsis*, *Brassica rapa*, castor bean (*Ricinus communis*), cocoa (*Theobroma cacao*), soybean (*Glycine max*), maize (*Zea mays*), and rice (*Oryza sativa*), respectively. The B3 superfamily has substantially expanded during the evolution in eudicots particularly in Brassicaceae, as compared to monocots in the analysis. We observed domain duplication in some of these B3 proteins, forming more complex domain architectures than currently understood. We found that the length of B3 domains exhibits a large variation, which may affect their exact number of α-helices and β-sheets in the core structure of B3 domains, and possibly have functional implications. Analysis of the public microarray data indicated that most of the *B3* gene pairs encoding *Arabidopsis*-rice orthologs are preferentially expressed in different tissues, suggesting their different roles in these two species. Using ESTs in crops, we identified many *B3* genes preferentially expressed in reproductive tissues. In a sequence-based quantitative trait loci analysis in rice and maize, we have found many *B3* genes associated with traits such as grain yield, seed weight and number, and protein content. Our results provide a framework for future studies into the function of *B3* genes in different phases of plant development, especially the ones related to traits in major crops.

Introduction

The B3 domain was first identified in VIVIPAROUS1 (VP1) of maize (*Zea mays*; McCarty et al. 1991), and in its *Arabidopsis* ortholog ABSCISIC ACID INSENSITIVE 3 (ABI3; Giraudat et al. 1992). Three basic regions, named B1, B2 and B3, exist in VP1, among which B1 and B2 are exclusive to VP1 while B3 is also present in other transcription factors (Giraudat et al. 1992; Swaminathan et al. 2008; Romanel et al. 2009). The B3 domain has not been found in other kingdoms, thus a B3-containing transcription factor is considered plant specific. In addition to the B3 domain, which is involved in DNA binding, other domains can coexist in the multidomain B3 proteins and are thought to mediate protein–protein interaction and/or dimerization. These additional domains include APETALA2 (AP2), auxin response factor (ARF), auxin/indole-3-acetic acid (Aux/IAA), and zinc finger CW domain (zf-CW), which have different structures, functions, and evolutionary histories from the common B3 domain, a typical attribute of a superfamily. Based on the presence of these five domains in a protein, the B3 superfamily can be classified into five families: ABI3-VP1, ARF, high-level expression of sugar-inducible (HSI), related to ABI3-VP1 (RAV), and reproductive meristem (REM). In addition to B3 proteins, some of these domains (except B3) serve as the signature motif for other families, likely providing a mechanism to expand their regulatory roles through

Communicated by A. Schulman.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-013-2054-4) contains supplementary material, which is available to authorized users.

F. Y. Peng · R. J. Weselake (✉)
Agricultural Lipid Biotechnology Program,
Department of Agricultural, Food and Nutritional Science,
University of Alberta, Edmonton, AB T6G 2P5, Canada
e-mail: randall.weselake@ualberta.ca

interactions among different classes of transcription factors. For example, AP2 is the defining domain for the AP2/EREBP (ethylene response element binding protein) family (Riechmann and Meyerowitz 1998), whereas the Aux/IAA domain defines the Aux/IAA family (Riechmann et al. 2000), both of which are also specific to plants.

A number of B3-containing transcription factors have been shown to regulate a multitude of biological processes in plants, controlling or influencing both vegetative and reproductive development (Yamasaki et al. 2004, 2008; Swaminathan et al. 2008; Romanel et al. 2009; Agarwal et al. 2011). Three ABI3-VP1 family members, ABI3, FUS3 (FUSCA 3) and LEC2 (LEAFY COTYLEDON 2), for instance, are known to regulate seed development and storage reserve accumulation (Monke et al. 2004; Braybrook and Harada 2008; Weselake et al. 2009; Le et al. 2010). Their B3 domains bind to the Sph/RY motif (CATGCA) in the promoter region of genes under regulation (Suzuki et al. 1997; Reidt et al. 2000; Monke et al. 2004; Le et al. 2010). Our recent bioinformatic analyses indicated that the Sph/RY elements are overrepresented in the promoters of genes encoding oleosins and seed storage proteins in developing *Arabidopsis* seeds (Peng and Weselake 2011). Monke et al. (2012) identified a set of 98 putative target genes of ABI3, most of which require the presence of abscisic acid for activation and exhibit seed maturation-specific expression patterns. The ARF family is also well characterized in *Arabidopsis*, and their B3 domains can bind to auxin responsive elements (AuxREs; TGTCTC) in the upstream of auxin responsive genes (Ulmasov et al. 1999; Ellis et al. 2005). ARFs are involved in various auxin-mediated physiological processes, including apical dominance, tropic responses, lateral root formation, vascular differentiation, embryo patterning, and shoot elongation (Okushima et al. 2005a, 2005b; Guilfoyle and Hagen 2007). For example, *AT1G19850*, which encodes ARF5 (IAA24), was shown to regulate embryo axis formation and lateral root development (Smet 2010). A number of *ARF* genes in maize were recently predicted to be potential targets of small RNAs (Xing et al. 2011). By contrast, the functions of other *B3* genes are less well characterized, but some of them are also implicated in important biological processes. In the HSI family, for example, both HSI2 (VAL1; AT2G30470.1) and HSL1 (HSI2-like 1; AT4G32010.1) function as repressors in the LEAFY COTYLEDON 1 (LEC1)-B3 regulatory network in plant embryo development and regulate the transition from seed maturation to seedling growth (Tsukagoshi et al. 2005, 2007; Suzuki et al. 2007). Over-expression of an RAV family gene, *RAV1* (*AT1G13260*), resulted in lateral root retardation and rosette leaf development in *Arabidopsis*, and its under-expression caused an earlier flowering phenotype, suggesting that RAV1 is a negative component in the regulation of plant development (Hu et al. 2004). VERDANDI (AT5G18000.1), which is a REM family protein, has recently been demonstrated to be a direct target of an ovule identity complex that includes MADS-box proteins and affects embryo sac differentiation in *Arabidopsis* (Matias-Hernandez et al. 2010).

With the availability of a growing number of sequenced plant genomes, genome-wide analysis of transcriptional regulators has emerged as an active research area in comparative genomics. Such studies can shed light on the origin and evolution of transcription factor families among different species and help to unravel the evolutionary basis of regulatory diversification in transcription (Riechmann et al. 2000; Li et al. 2006; Romanel et al. 2009; Carretero-Paulet et al. 2010; Xing et al. 2011). Swaminathan et al. (2008) identified the B3-encoding genes in the genome of *Arabidopsis* and rice (*Oryza sativa*), and defined four B3 families: ARF, LAV (including ABI3-VP1 and HSI families), RAV and REM. Romanel et al. (2009) analyzed the *B3* genes in six species including two green algae *Chlamydomonas reinhardtii* and *Volvox carteri*, moss *Physcomitrella patens*, and three higher plant species *Arabidopsis*, poplar (*Populas trichocarpa*), and rice (*O. sativa*), to study the evolution of the B3 superfamily, with a focus on its REM family. Following *Arabidopsis* and rice, of which the genome was sequenced about a decade ago (*Arabidopsis* Genome Initiative 2000; Goff et al. 2002; Yu et al. 2002), advances in the whole-genome sequencing technology have made it possible to sequence larger, more complex genomes of major crops, such as soybean (*Glycine max*; Schnable et al. 2009) and maize (Schmutz et al. 2010). Genome-wide analyses of gene families including important crop species will enhance our understanding of physiological and agronomic diversity arising from plant evolution and crop domestication.

In this study, we identified a comprehensive list of *B3* genes using a genome-scale domain analysis in seven plant genomes including Brassicacea and major crop species: *Arabidopsis*, *Brassica rapa*, castor bean (*Ricinus communis*), cocoa (*Theobroma cacao*), soybean (*G. max*), maize, and rice. In these crops, the *B3* genes have not been studied in detail and, to our knowledge none of them have been functionally characterized. Our results indicate that the B3 superfamily has substantially amplified in the genome of eudicots, especially in the Brassicaceae lineage. In addition, our in silico expression analyses of the *B3* genes using public microarray and expressed sequence tag (EST) data uncovered *B3* genes that are preferentially expressed in reproductive organs (flower and seed) as compared with vegetative tissues (root, stem, and leaf). Of the 19 *Arabidopsis*-rice orthologous *B3* gene pairs existing in the two public Affymetrix microarray datasets, 16 of them were preferentially expressed in different tissues. Using the known quantitative trait loci (QTL) data in rice and maize, we found many *B3* genes in known QTL regions, providing a link of *B3* genes to important traits in crop breeding. Our results will be useful for genetic improvement of major crops.

Materials and methods

Identification and classification of *B3* genes in the seven genomes

The data source and characteristics of the proteome sequence used in this study are summarized in Table 1, for the seven species including *Arabidopsis* (Arabidopsis Genome Initiative 2000; Swarbreck et al. 2008), *B. rapa* (Cheng et al. 2011; The Brassica rapa Genome Sequencing Project Consortium et al. 2011), castor bean (*R. communis*; Chan et al. 2010), cocoa (*T. cacao*; Argout et al. 2011), soybean (*G. max*; Schmutz et al. 2010), maize (*Z. mays*; Duvick et al. 2008; Schnable et al. 2009), and rice (*O. sativa*). Two subspecies of rice (Japonica and Indica) have been independently sequenced (Goff et al. 2002; Yu et al. 2002), and we chose to use the genome sequence of the Japonica for this analysis, which has been more widely used as the reference genome for rice. The computational pipeline for genome-scale domain analysis to identify *B3* genes and classify them into the five families is shown in Fig. 1. Briefly, InterProScan (Mulder and Apweiler 2007), which combines multiple domain recognition methods, was installed locally to facilitate our genome-scale domain analysis. The InterPro database (Mulder et al. 2007; Hunter et al. 2009), which integrates multiple domain signature databases including the protein family database Pfam (Finn et al. 2008), was used for domain recognition by InterProScan with default parameters, including an *E* value threshold of 0.001. To enhance accuracy, we added a refinement step following computational analysis to eliminate putative *B3* genes without sufficient evidence for certain domains. For example, AT1G05930.1 contains a domain of unknown function (IPR005508; DUF313) and thus was not included in the final list of *B3* genes in *Arabidopsis*, even though it also contains a B3 domain (IPR003340). We also examined the annotation of candidate *B3* genes using BLAST (Altschul et al. 1997), and removed those that hit a pseudogene in *Arabidopsis* genome. For example, Bra027159, a putative *B3* gene in *B. rapa*, matches a pseudogene in *Arabidopsis* (AT5G24050) and was thus eliminated. For newly sequenced genomes, we also evaluated the sequence of each *B3* candidate gene. For example, we found two candidate *B3* genes in *B. rapa*, *Bra018098* and *Bra025779*, which contain large stretches of Ns in their DNA sequences (representing ambiguous base calls in sequence reads or genomic gaps yet to be sequenced), and were thus excluded from subsequent analyses. In addition, in the genome of *Arabidopsis* or rice, one gene can encode more than one protein isoform, and we retained only one of them if they belong to the same family. The final B3 transcription factors identified in each species were classified into each of the five families on the basis of the domain architecture described in Romanel et al. (2009).

Multiple sequence alignment of domain sequences and structural analyses of B3 domains

We extracted the sequences for B3, AP2, ARF, Aux/IAA, and zf-CW domains in the B3 proteins, and multiple sequence alignments were performed using ClustalX 2.0 (Larkin et al. 2007). Sequence conservation for each domain was analyzed using the plotcon tool in the EMBOSS suite (Rice et al. 2000), with a sliding window of four residues.

Expression analysis of *B3* genes in the seven species

To compare tissue-specific expression patterns of *B3* genes in *Arabidopsis* and rice, we used two public Affymetrix GeneChip datasets (Schmid et al. 2005; Jain et al. 2007; Jain and Khurana 2009; Sharma et al. 2009; Deveshwar et al. 2011), which include a large amount of microarray gene expression data in the five tissues of the two species: seed, flower, root, leaf, as well as stem (*Arabidopsis*) or SAM (shoot apical meristem; rice). The raw data (*.CEL* files) were downloaded from TAIR (Swarbreck et al. 2008; http://Arabidopsis.org/servlets/TairObject?type=hyb_descr_collection&id=1006710873) and the NCBI Gene Expression Omnibus (GEO; Barrett et al. 2011; http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE6893). To mitigate the potential impact of different analysis methods used in the two studies, we analyzed the raw data separately for each tissue of the two species with a consistent approach using Bioconductor packages in the R statistical computing environment (Gentleman et al. 2004; R Development Core Team 2010), which has been described in Peng and Weselake (2011). Only *B3* genes represented in the *Arabidopsis* or rice genome array and expressed in at least one of the five tissues (cut-off value of 5.0 in a $\log_2$ scale) were considered, and the highest expression value for each *B3* gene was chosen if multiple samples for each tissue were profiled. To visualize the expression pattern of *B3* genes in different tissues, their normalized expression data were standardized and analyzed in GenePattern (Reich et al. 2006).

For the remaining species, we used the expressed sequence tag (EST) data to identify *B3* genes preferentially expressed in reproductive and vegetative organs. All EST sequences for each species were retrieved from GenBank

Table 1 Summary of the proteome sequences used in the study

Species	Sequence file	Total # proteins	Online source
At	TAIR10_pep_20101214	35,386	ftp://ftp.arabidopsis.org/home/tair/Genes/
Br	Brapa_gene_v1.1.pep	41,174	http://brassicadb.org/brad/
Rc	TIGR_castorWGS_release_0.1.aa.fsa	31,221	http://castorbean.jcvi.org/downloads.php
Tc	cacao_v0.9_gene_peptides.fasta	34,996	http://www.cacaogenomedb.org/genome-sequence
Gm	Glyma1_highConfidence_longest.pep	46,430[a]	http://www.phytozome.net/soybean
Zm	Zea_mays.Protein.fasta	42,531	ftp://ftp.plantgdb.org/download/MaizeData/MaizeGDB1286198429/FASTA/
Os	Oryza_sativa.MSU6.55.pep.all.fa	68,682[b]	ftp://ftp.gramene.org/pub/gramene/CURRENT_RELEASE/data/fasta/oryza_sativa/pep/

[a] The soybean genome may contain >60,000 protein-coding genes (Schmutz et al. 2010), but we only used its 46,430 proteins with the highest level of predictive confidence

[b] Two subspecies (Indica and Japonica) of rice have been sequenced (Goff et al. 2002; Yu et al. 2002), and we used the Japonica genome sequence in this study

dbEST at NCBI on March 4 2011 (Benson et al. 2008). Because GenBank dbEST is an uncurated repository for the deposition of raw ESTs, some EST sequences could be contaminated, too short, or contain low-complexity reads (Lee and Shin 2009), therefore we cleaned all ESTs with SeqClean (http://compbio.dfci.harvard.edu/tgi/software/) and only retained high-quality ESTs for subsequent analysis. The final EST data were then clustered into vegetative and reproductive ESTs according to the tissue source in the EST library description. The vegetative tissues include root, stem, shoot, leaf, and seedling. The reproductive organs include anther, ovary, embryo, flower, fruit, gametophyte, grain, spore, stigma, pistil, pollen, carpel, seed, inflorescence, and strobilus. ESTs from apex, callus, meristem, cell culture (which could differentiate into different tissues), or ESTs with ambiguous tissue description were removed. *B3* genes with a singleton EST were excluded from this analysis, due to its minimal statistical value.

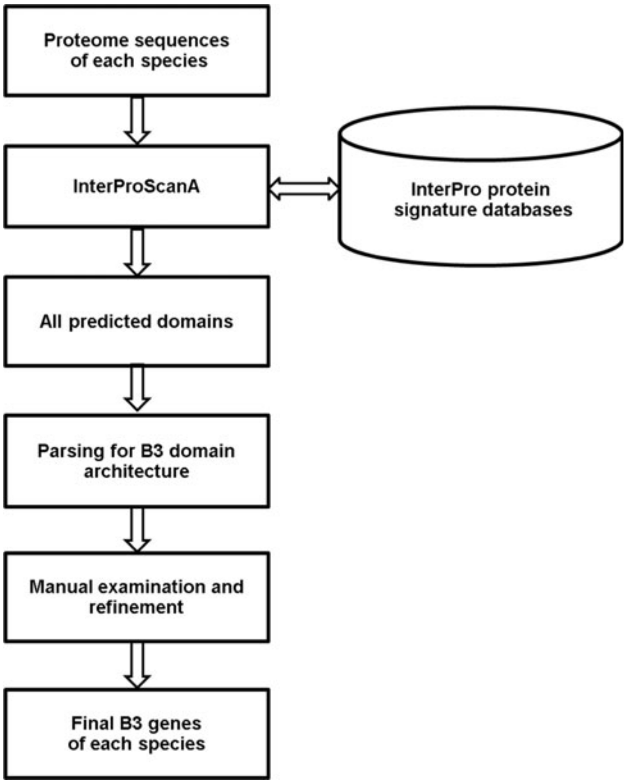

Proteome sequences of each species
InterProScanA
InterPro protein signature databases
All predicted domains
Parsing for B3 domain architecture
Manual examination and refinement
Final B3 genes of each species

Fig. 1 A schematic diagram for genome-wide identification of *B3* genes in Brassicacea and major crop species. To allow for genome-scale domain discovery, we locally installed InterProScan and its companion Interpro database (Mulder and Apweiler 2007; Mulder et al. 2007; Hunter et al. 2009). The input proteome data are summarized in Table 1. The domain identifiers used in this analysis were extracted from the Pfam database (Finn et al. 2008), as follows: B3, PF02362; AP2, PF00847; ARF, PF06507; Aux/IAA, PF02309, and zf-CW, PF07496. The candidate *B3* genes were classified into the five families according to the domain architecture in the B3 superfamily (Romanel et al. 2009)

Identification of *B3* genes overlapping with known QTLs in rice and maize

The QTL data for rice and maize, respectively, were retrieved from Gramene QTL database (Liang et al. 2008; Ni et al. 2009) and MaizeGDB (Cannon et al. 2011; Schaeffer et al. 2011). For the QTLs with known physical positions of their flanking markers, we extracted their corresponding genomic sequences. We then identified *B3* genes residing in these known QTL regions in the two crops using BLAST (Altschul et al. 1997), with an *E* value threshold of 10^{-15}. For each *B3* gene, up to three top hits were retained if their *E* value is below the cutoff.

Results

B3 genes in the genome of *Arabidopsis*, *Brassica rapa*, castor bean, cocoa, soybean, maize, and rice

To identify candidate *B3* genes in the sequenced genome of *Arabidopsis*, *B. rapa*, castor bean, cocoa, soybean, maize, and rice, we performed a genome-wide domain analysis using each proteome as input (Table 1; Fig. 1). The numbers of *B3* genes in each of these seven species, and those in the five B3 families are shown in Table 2. The annotation of these *B3* genes identified in Brassicacea and the major crop species is given in Online Resource 1. We compare the number of *B3* genes identified in this study with previous surveys in *Arabidopsis* and rice, and found both agreement and discrepancy. We found the *Arabidopsis* genome encodes 92 B3 transcription factors, accounting for about 0.26 % of the total number of predicted proteins in this model species. This number of *B3* genes in *Arabidopsis* is between those in the two previous surveys: 118 in Swaminathan et al. (2008) and 87 in Romanel et al. (2009). Strikingly, we identified 187 *B3* genes in the *B. rapa* genome, and estimated that this represents nearly 0.46 % of its coding capacity. In other eudicots, we uncovered 58, 90, and 81 *B3* genes in castor bean, cocoa, and soybean, accounting for about 0.19, 0.26, and 0.17 % of the coding capacity, respectively. By contrast, only 77 (~0.11 %) and 55 (~0.13 %) *B3* genes were detected, respectively, in rice and maize, the two monocot genomes in the study. Our number of *B3* genes found in the rice genome is smaller than that in the two earlier studies: 91 in Swaminathan et al. (2008) and 86 in Romanel et al. (2009). Relative to the estimated gene content in each of these seven genomes, our results suggest that the B3 superfamily was substantially amplified during evolution in eudicot plants, particularly in *Arabidopsis* and *B. rapa*. In a *B. rapa* chromosome, a cluster of five *B3* genes belonging to the ABI3-VP1 family was found within a 40 kb genomic region (Online Resource 2). This expansion, however, is not uniform in all the five B3 families or across the seven plant genomes analyzed.

In most of these seven genomes, ABI3-VP1 is the largest family in the B3 superfamily, and this is particularly the case for cocoa, *Arabidopsis*, and *B. rapa*. The notable exception is soybean, in which 54 ARF members form the largest family, similar to the 55 soybean ARFs reported in Zhang et al. (2011). We identified 37 and 28 ABI3/VP1 family proteins in *Arabidopsis* and rice, respectively. By contrast, only six in *Arabidopsis* and seven in rice were identified in the ABI3-VP1 and HSI families combined in Romanel et al. (2009), which is much lower than that in earlier surveys for the ABI3-VP1 family alone. For example, the databases of transcription factors in *Arabidopsis* (DATF) and rice (DRTF) reported, respectively, 60 and 57 ABI3-VP1 members in *Arabidopsis* and Japonica rice (54 in Indica; Guo et al. 2005; Gao et al. 2006). In an early genome-wide transcription factor study, 14 ABI3-VP1 family members were identified in *Arabidopsis* (Riechmann et al. 2000). Interestingly, we found an example in *Arabidopsis* for one gene encoding two proteins that were classified into two different families in the B3 superfamily. *AT2G16210* can encode two protein forms: AT2G16210.1 and AT2G16210.2. AT2G16210.1 contains one extra exon at the 5′ end and its protein product was assigned to the REM family, whereas the protein encoded by AT2G16210.2 was classified into the ABI3-VP1 family. This indicates that the extra exon in AT2G16210.1 encodes an additional N-terminal B3 domain.

We identified 23 ARFs in *Arabidopsis*, in agreement with previous studies (Riechmann et al. 2000; Remington et al. 2004; Guo et al. 2005; Romanel et al. 2009). In rice, we identified 28 ARFs, the same as Romanel et al. (2009), but lower than the 41 ARFs identified in the Japonica rice genome in another previous analysis (Gao et al. 2006), although only 24 ARFs were found in Indica rice in the same study. In maize, we identified 24 ARFs, which is lower than the 31 ARFs reported recently (Xing et al. 2011). For the REM family in *Arabidopsis* and rice, we identified 24 and 19 members, respectively, compared to 45 and 39 REMs identified in Romanel et al. (2009). In castor bean and rice, ABI3-VP1 is roughly the same size as ARF, whereas in rice, the ABI3-VP1 family is slightly larger than ARF. Notably, we identified no REM family members in soybean, most likely because we used only the peptides predicted with highest confidence for this analysis (Table 1). For this reason, the number of soybean *B3* genes in Table 1 may represent an underestimation of the B3 superfamily in this species. For genes in other B3 families, we identified six RAV family members in *Arabidopsis*, in line with an early survey (Riechmann et al. 2000). These authors, however, did not define a B3 family or

Table 2 Number of *B3* genes and their classification in the B3 superfamily identified in the genome of *Arabidopsis*, *Brassica rapa*, castor bean, cocoa, soybean, maize, and rice

Species	ABI3-VP1	ARF	HSI	RAV	REM	Total number of *B3* genes	% in each genome[a]
At	37	23	2	6	24	92	0.26
Br	68	33	4	14	68	187	0.46
Rc	21	18	3	4	12	58	0.19
Tc	50	18	2	5	15	90	0.26
Gm	16	54	6	5	0	81	0.17
Zm	23	24	2	2	4	55	0.13
Os	28	26	2	4	17	77	0.11

ABI3-VP1 abscisic acid-insensitive 3-viviparous 1, *ARF* auxin response factor, *HSI* high-level expression of sugar-inducible gene, *RAV* related to ABI3-VP1, *REM* reproductive meristem

[a] The percentage of *B3* genes in each species was determined using the total number of *B3* genes we identified in the study divided by the total number of proteins listed in Table 1

superfamily, and instead classified those six proteins into a RAV-like subfamily in the AP2/EREBP family, presumably using the AP2 domain in these proteins.

Domain duplication and location of the five domains in B3 proteins

According to our analysis, the domain architecture in these B3 proteins is more complex than currently known. In the REM family proteins, some of them contain more than two (instead of two) B3 domains (Table 3). In *Arabidopsis* and *B. rapa*, we identified eight and 12 B3 proteins, each containing over two B3 domains, several of which contain more than five B3 domains. For example, seven predicted B3 domains were found in AT2G24650.1, whereas nine predicted B3 domains in Bra032075. Likewise, REM family members with more than two B3 domains were identified in castor bean (one), cocoa (two), and rice (two), but not in maize. In addition to the B3 domain, duplication events of the ARF domain were observed in two ARF family members, Bra002327 in *B. rapa* and LOC_Os07g08520 in rice. In contrast, we found no duplication events for AP2, Aux/IAA, or zf-CW domains in these B3 proteins.

To examine the locations of the five domains in the B3 proteins, we divided each B3 protein into three equal segments in a protein sequence: N-terminal, middle, and C-terminal, and determined where a domain (or the majority of its sequence) is located. We found that the five domains exhibit distinct distribution patterns in the B3 proteins (Table 4). The majority of the B3 domains in the seven species were found in the N-terminal regions, followed by C-terminal regions, and the least number of B3 domains exist in the middle regions. Conversely, the AP2 domains were located exclusively in the N-terminal regions, whereas the ARF domains exist predominantly in the middle regions. The Aux/IAA domains were found exclusively in the C-terminal region, which was also the case for the zf-CW domains, despite only a small number of zf-CW domains found in these B3 proteins and an equal number of zf-CW domains present in the middle region of soybean B3 proteins.

Table 3 The REM family proteins containing more than two predicted B3 domains

Species	Gene ID	Protein length	Number of B3 domains
At	AT1G26680	920	6
	AT2G24690	748	4
	AT2G24700	555	4
	AT2G24650	1,045	7
	AT2G24680	851	5
	AT4G00260	528	4
	AT4G31650	493	4
	AT5G32460	530	4
Br	Bra032077	496	4
	Bra032075	1,157	9
	Bra024697	543	4
	Bra024696	530	4
	Bra000545	459	3
	Bra012445	691	5
	Bra011285	987	5
	Bra010225	807	5
	Bra007840	856	4
	Bra007836	1,063	8
	Bra007835	717	6
	Bra000564	1,015	5
Cr	29848.m004479	559	4
Ct	CGD0000402	939	4
	CGD0000405	771	4
Os	LOC_Os12g40070	661	4
	LOC_Os03g42370	1,029	5

Table 4 Domain distribution in the N-terminal, middle, or C-terminal region in the B3 superfamily of the seven species

Species	B3			AP2			ARF			Aux/IAA			zf-CW		
	N[a]	M[a]	C[a]	N	M	C	N	M	C	N	M	C	N	M	C
At	66	12	33	6	0	0	2	19	0	0	0	23	0	0	2
Br	127	30	77	14	0	0	5	20	1	0	1	25	0	0	4
Rc	28	7	21	4	0	0	2	11	0	0	0	15	0	0	3
Tc	59	5	26	5	0	0	2	10	1	0	0	15	0	0	2
Gm	47	5	4	5	0	0	7	27	1	0	0	46	0	3	3
Zm	21	3	20	2	0	0	1	10	4	0	0	10	0	0	2
Os	53	8	34	4	0	0	7	15	3	0	0	24	0	0	2

B3 basic domain 3, *AP2* APETALA2, *ARF* auxin response factor, *Aux/IAA* auxin: indole-3-acetic acid, *zf-CW* zinc finger domain with Cys (C) and Trp (W) residues

[a] The distribution of a domain was determined by dividing a protein sequence evenly into three segments and examining where the majority (or the entirety) of a domain sequence is located. 'N', 'M', and 'C' indicates N-terminal region, middle region and C-terminal region, respectively

The five domains exhibit different degree of sequence divergence in the B3 superfamily

We attempt to compare sequence divergence (or conversely conservation) between the B3 domains and other coexisting domains in the same B3 proteins. Our analyses showed that the B3 domains in the seven species exhibit considerable sequence divergence. In the RAV family, the B3 domains exhibited a higher degree of divergence than their coexisting AP2 domains (Fig. 2a, b), even though the sequence conservation was not uniform across different regions within the same domain. Many residues in the B3 domains did not align and large gaps existed, whereas most residues in the AP2 domains aligned well (Online Resource 3). Likewise, in the HSI family, the B3 domains also showed a higher degree of divergence than their zf-CW counterparts (Fig. 2c, d). Many residues in these B3 domains are not conserved, whereas in the zf-CW domains, most residues were conserved, especially the Cys and Trp residues as expected (Online Resource 3). Hence, the B3 domain might have been evolved independently from AP2 in the RAV family and zf-CW domain in the HSI family. Similar analyses in the ARF family, however, indicated that B3, ARF, and Aux/IAA domains display a comparable degree of divergence in the ARF family (Online Resource 3).

The length of the five domains varies in the B3 superfamily and the length of the B3 domain may affect its exact core structure

It has been shown that the B3 domain is relatively large, composed of 95–117 amino acid residues (Yamasaki et al. 2004; Waltner et al. 2005). According to our analysis, however, the length of the B3 can vary to a more substantial extent, ranging, respectively, from 62 to 118. Similar degree of length variation was observed for ARF and Aux/IAA domains in these seven species, ranging from 65 to 107, and 84 to 163 residues, respectively. In contrast, the length variation of the AP2 and zf-CW domains is small, ranging, respectively, from 46 to 49 and 43 to 45 residues. In total, we detected 1,054 domains in these B3 proteins, and the average domain length is 92, 48, 81, 117, and 44 for B3, AP2, ARF, Aux/IAA, and zf-CW, respectively. We found that the exact number of α helices and β sheets of the known core structure (two α helices and seven β sheets; Yamasaki et al. 2004; Waltner et al. 2005) of B3 domains can vary to a certain extent, largely depending on the domain length (data not shown). The large length variation observed in the B3, ARF, and Aux/IAA domains could have structural and/or functional implications as reported in Sandhya et al. (2009) showing 80 % of 'length-deviant' superfamilies possess distant internal structural repeats and nearly half of them acquired diverse biological function.

Most *B3* genes in *Arabidopsis* and rice were preferentially expressed in different tissues

The availability of two similar microarray data sets in *Arabidopsis* and rice allowed our comparison of tissue-specific expression patterns of *B3* genes in these two species. Both of these data sets were obtained using the Affymetrix GeneChip platform and included five tissues: seed, flower, root, leaf, and stem (*Arabidopsis*) or SAM (rice). Swaminathan et al. (2008) showed the expression patterns of most *B3* genes in *Arabidopsis* and rice, which did not include *ARF* genes. We included the ARF family genes for a similar analysis. Of the 92 *B3* genes identified in *Arabidopsis*, we found 56 *B3* genes (or ~60 %) represented in the ATH1 Genome Array and expressed in at least one of the five tissues. Of the 77 *B3* genes identified in

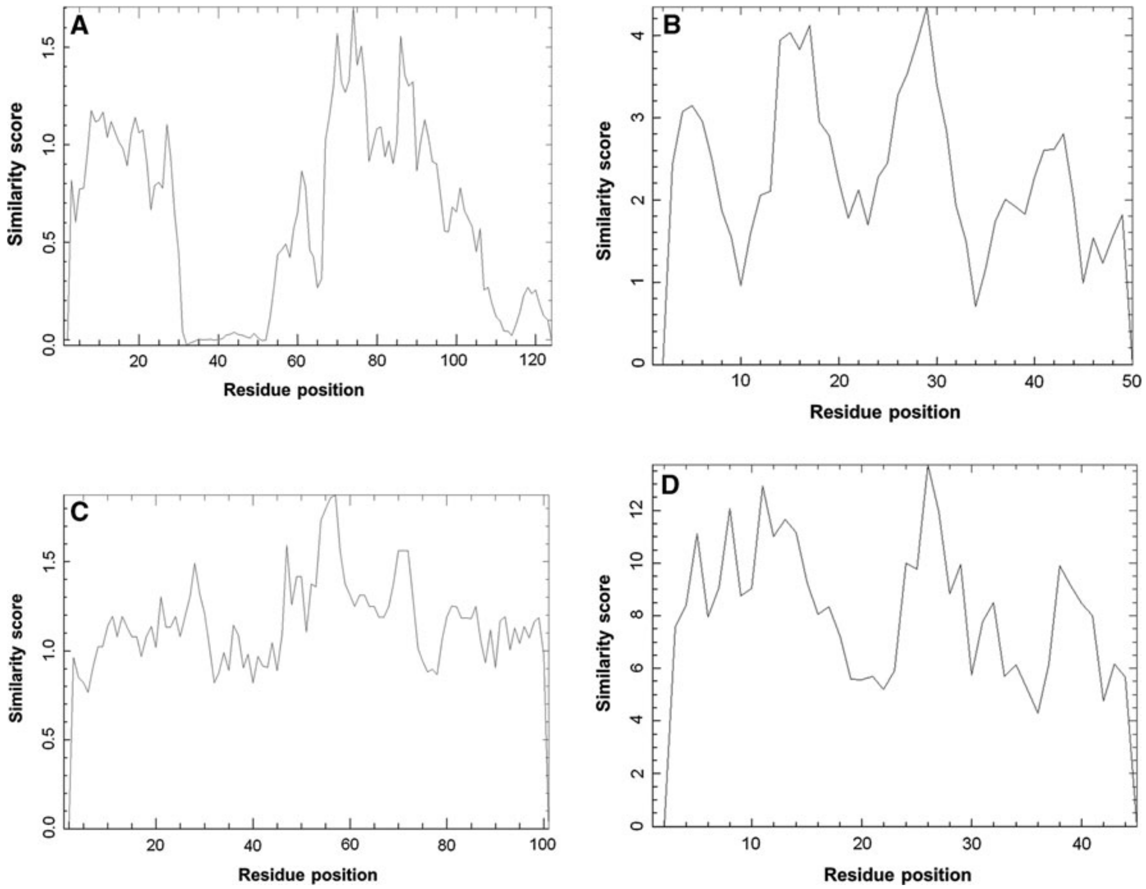


Fig. 2 Sequence conservation plots of the B3 and AP2, as well B3 and zf-CW domains in the B3 superfamily. Domain sequences were aligned using ClustalX 2.0 (Larkin et al. 2007), and the conservation plots were obtained using the plotcon tool in EMBOSS (Rice et al. 2000), using a sliding window size of four residues. In the sequence conservation plot for B3 domains (**a**) and AP2 domains (**b**), for the RAV family, the upper limit of the *y* axis is 1.7 in **a** and 4.2 in **b**. In the sequence conservation plot B3 domains (**c**) and zf-CW domains (**d**), for the HSI family, the upper limit of the *y* axis is 1.9 in **c** and 13.5 in **d**

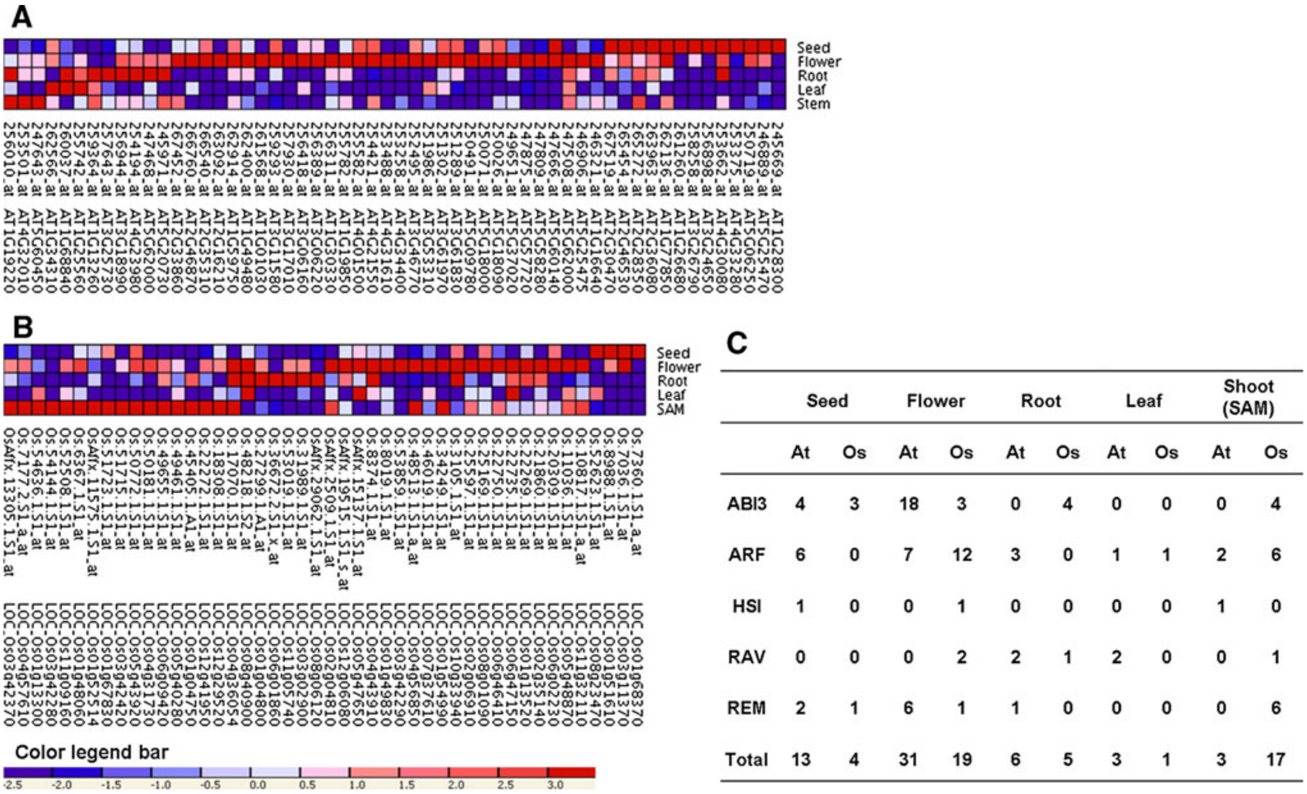
rice, we found 46 *B3* genes (or ~60 %) in the Affymetrix Rice Genome Array and expressed in at least one of the five tissues. The tissue-specific expression pattern for these *B3* genes is shown in Fig. 3, and their normalized expression values are provided in Online Resource 4. In *Arabidopsis*, most *B3* genes are preferentially expressed in flower and seed (Fig. 3a). Moreover, all *B3* genes highly expressed in *Arabidopsis* flowers and seeds belong to the ABI3-VP1 family (Fig. 3c). A total of 24 *ABI3-VP1* genes were preferentially expressed in flowers (18) and seeds (6), accounting for nearly 38 % of this transcription factor class in *Arabidopsis*. These include *ABI3-VP1* genes, *ABI3* (*AT3G24650*), *FUS3* (*FUSCA 3*; *AT3G26790*), and *LEC2* (*LEAFY COTYLEDON 2*; *AT1G28300*), which are well-characterized regulators of seed development and storage compound accumulation (Weselake et al. 2009; Le et al. 2010).

In rice, most *B3* genes were preferentially expressed in flower and SAM (Fig. 3b), totalling 19 and 17, respectively. Fourteen genes in the ABI3-VP1 family were found in the rice Affymetrix data, but only six of them were preferentially in the seed and flower (Fig. 3c). In comparison, 19 *ARF* genes were found, with 12 and 6 preferentially expressed in the flower and SAM, respectively.

Interestingly, we identified 19 *B3* gene pairs in these two microarray datasets that encode putative orthologs in *Arabidopsis* and rice. Of them, 16 gene pairs were preferentially expressed in different tissues (Table 5). For example, *AT3G24650* (*ABI3*) was preferentially expressed in *Arabidopsis* seeds, but its rice orthologous gene *LOC_Os08g01090* was preferentially expressed in flower.

B3 genes preferentially expressed in reproductive tissues of the crop species

For the *B3* genes identified in the remaining five crop species, no extensive microarray data for gene expression estimates can be found. Consequently, we used the expressed sequence tag (EST) data to identify *B3* genes expressed preferentially in the reproductive flowers and seeds. In all of the available ESTs for each of the six



	Seed		Flower		Root		Leaf		Shoot (SAM)	
	At	Os	At	Os	At	Os	At	Os	At	Os
ABI3	4	3	18	3	0	4	0	0	0	4
ARF	6	0	7	12	3	0	1	1	2	6
HSI	1	0	0	1	0	0	0	0	1	0
RAV	0	0	0	2	2	1	2	0	0	1
REM	2	1	6	1	1	0	0	0	0	6
Total	13	4	31	19	6	5	3	1	3	17

Fig. 3 The expression patterns of *B3* genes in the five tissue types of *Arabidopsis* and rice. **a** The *B3* genes represented in the ATH1 Arabidopsis Genome Array and expressed in at least one of these tissues. The AGI code and probe set identifier for each gene are also shown. **b** The *B3* genes represented in the Affymetrix Rice Genome Array and expressed in at least one of the five tissues. The rice gene and their probe set identifiers are also shown. **c** Number of *B3* genes in the five families expressed in different tissues of *Arabidopsis* and rice. The abbreviations for the five family names are: ABI3-VP1, abscisic acid insensitive 3-viviparous1; ARF, auxin response factor; HSI, high-level expression of sugar-inducible gene; RAV, related to ABI3-VP1; REM, reproductive meristem

species retrieved from GenBank dbEST (Benson et al. 2008), we found that certain ESTs would not be ideal for this type of analysis. Notably, 59,901 soybean ESTs and 263,993 maize ESTs were either too short or contain low-complexity nucleotides, accounting for ~7.5 and ~13 % of their total ESTs, respectively. Furthermore, some cDNA libraries were normalized (or subtracted), and others were constructed using pooled samples including multiple tissues and organs. Therefore, the number of ESTs with an unambiguous tissue designation was small for each of the five organs in most of the six species, did not permit us to carry out a digital gene expression analysis using EST tag counts. As such, we combined ESTs into vegetative and reproductive tissues for each species (Online Resource 5), and used a qualitative approach to identify *B3* genes with ESTs only in reproductive tissues but none in vegetative tissues, in each species, which represent *B3* genes preferentially expressed in reproductive tissues of these crop species (Table 6). As expected, several genes orthologous to known *B3* genes in *Arabidopsis* were found preferentially expressed in reproductive tissues in most of these species, including *LEC2*, *FUS3*, and *ABI3*. In addition, several *REM* genes were preferentially expressed in reproductive tissues in these crops. This includes the VDD-encoding gene in castor bean and cocoa (Table 6); its *Arabidopsis* ortholog (AT5G18000) has recently been shown to be a direct target of the MADS domain ovule identity complex, whose mutation affected embryo sac differentiation (Matias-Hernandez et al. 2010). Similarly, *B3* genes preferentially expressed in vegetative tissues of the six species can be identified (data not shown). This tissue-specific expression data of *B3* genes may be useful in designing knockout experiments in crops. Due to the constraint of EST data, genes listed in Table 6 likely represent an underestimation of the *B3* genes preferentially expressed in the reproductive tissues of the crop species.

B3 genes overlapping with known QTLs in rice and maize

We identified 8,216 and 3,564 QTLs in the rice and maize, respectively. All the rice QTLs have known physical locations of their flanking markers. In contrast, only 600 maize QTLs have inferred physical coordinates. The marker

Table 5 Tissues of preferential expression of the orthologous *B3* gene pairs in *Arabidopsis* and rice

B3 family name	AtGeneID	Tissue of preferential expression	OsGeneID	Tissue of preferential expression
ABI3-VP1	*AT3G24650*	Seed	*LOC_Os08g01090*	Flower
	AT5G06250	Seed	*LOC_Os12g06080*	Flower
	AT3G26790[a]	Seed	*LOC_Os01g51610*	Seed
	AT2G36080	Seed	*LOC_Os11g05740*	Root
	AT5G58280	Flower	*LOC_Os05g40280*	SAM
	AT3G61970	Flower	*LOC_Os06g01860*	Root
ARF	*AT1G19850*[a]	Flower	*LOC_Os04g56850*	Flower
	AT1G30330	Flower	*LOC_Os12g41950*	SAM
	AT1G59750	Flower	*LOC_Os04g36054*	SAM
	AT5G62000	Flower	*LOC_Os12g29520*	SAM
	AT2G33860	Flower	*LOC_Os05g43920*	SAM
	AT4G30080	Seed	*LOC_Os10g33940*	Flower
	AT2G28350	Seed	*LOC_Os04g43910*	Flower
HSI	*AT4G32010*	Stem	*LOC_Os07g37610*	Flower
RAV	*AT1G25560*	Leaf	*LOC_Os05g47650*	Flower
	AT1G68840	Leaf	*LOC_Os01g04750*	SAM
REM	*AT4G34400*	Flower	*LOC_Os01g67830*	SAM
	AT3G18990	Root	*LOC_Os03g42290*	Flower
	AT4G33280[a]	Seed	*LOC_Os08g23470*	Seed

[a] The B3 orthologous genes were preferentially expressed in the same tissue in both *Arabidopsis* and rice

positions of each QTL allowed us to extract its corresponding genomic sequence. The average length of genomic sequence spanning known QTLs in rice is approximately 2.4 Mb long (maximum nearly 39 Mb), whereas the average length covering each maize QTL is about 36 Mb (maximum QTL length nearly 173 Mb). We identified many *B3* genes in the known QTLs in rice and maize. In rice, we identified more than 400 QTLs associated with *B3* genes. Among these QTL links of *B3* genes in rice, one gene can be involved in multiple traits and one trait can be associated with multiple *B3* genes, forming a complex relationship. For example, LOC_Os01g04750.1, a putative RAV family protein (RAV2; Online Resource 1), is related to both root number (CQAI24-RTNB) and leaf senescence (CQN28-LFSNS) in rice. On the other hand, spikelet number (SPKNB) is associated with many rice *B3* genes (Online Resource 6). In maize, only 19 QTLs related to *B3* genes were found, because in a vast majority of maize QTLs, the physical location of their flanking markers has not been inferred. In both rice and maize, *B3* genes are associated with such important traits as kernel weight, kernel length, ear diameter, protein content, and dry matter (Table 7).

Discussion

The plant-specific B3 superfamily of transcription factors is defined by the presence of one or more B3 domains, or a combination of the B3 domain and one or more additional domains including AP2 (APETALA2), ARF, AUX/IAA, and zf-CW. This superfamily includes five families, among which ABI3-VP1 and ARF families are well studied in *Arabidopsis* and have diverse functions in plant growth and development (Yamasaki et al. 2004, 2008; Swaminathan et al. 2008; Agarwal et al. 2011). In contrast, few *B3* genes have been identified and characterized in major crops, and new insight into this superfamily could be gained from genome analysis including crop species. Using a genome-scale domain analysis approach, we identified a comprehensive list of B3-encoding genes in the seven plant genomes including both model species and economically important crops. The *B3* genes have been analyzed in *Arabidopsis* and rice (Swaminathan et al. 2008; Romanel et al. 2009), and our numbers of *B3* genes in these two genomes show agreement and discrepancy with previous studies. The discrepancy in the number of *B3* genes identified in different studies may be attributed to different database sources, approaches, and parameters being used. For example, previous studies assigned B3 proteins without typical AP2 domains to the RAV family, and some REMs lacked additional B3 domains (Magnani et al. 2004; Kim et al. 2006; Swaminathan et al. 2008; Romanel et al. 2009); in such cases our domain analysis classified them into the ABI3-VP1 family. This is also true for other multidomain B3 families; if no typical domain (other than B3) was found in a protein, we assigned it to the ABI3-VP1 family.

Table 6 *B3* genes differentially expressed in the flower and seed in *Brassica rapa*, castor bean, cocoa, soybean, and maize

Species	Gene ID	AGI	*Arabidopsis* ortholog name/description
Br	Bra032890	AT1G28300	*LEC2* (*LEAFY COTYLEDON 2*)
	Bra030087		
	Bra025229	AT3G26790	*FUS3*
	Bra020417	AT5G57720	AP2/B3-like protein
	Bra018530	AT4G34400	AP2/B3-like protein
	Bra017692	AT3G53310	AP2/B3-like protein
	Bra017651	AT4G34400	AP2/B3-like protein
	Bra017650	AT4G34400	AP2/B3-like protein
	Bra017649	AT4G34400	AP2/B3-like protein
	Bra017648	AT3G06160	AP2/B3-like protein
	Bra011110	AT4G34400	AP2/B3-like protein
	Bra011086	AT3G06160	AP2/B3-like protein
	Bra006989	AT3G53310	AP2/B3-like protein
	Bra003130	AT3G53310	AP2/B3-like protein
	Bra002509	AT5G60142	AP2/B3-like protein
	Bra037132	AT5G66980	AP2/B3-like protein
	Bra012121	AT5G66980	AP2/B3-like protein
Cr	29887.m000243	AT1G49480	*RTV1* (*RELATED TO VRN1*)
	29887.m000238	AT3G18990	*VRN1* (*REM39*)
	29887.m000236		
	29676.m001675		
	29585.m000597	AT5G18000	*VDD* (*VERDANDI*)
	29801.m003200	AT3G19184	AP2/B3-like protein
	29801.m003201	AT3G19184	AP2/B3-like protein
	29945.m000088	AT5G58280	AP2/B3-like protein
	29851.m002498	AT3G18990	*VRN1* (*REM39*)
Ct	CGD0000997	AT3G24650	*ABI3* (*ABSCISIC ACID INSENSITIVE 3*)
	CGD0008599	AT5G18000	*VDD* (*VERDANDI*)
	CGD0008603	AT1G49480	*RTV1* (*RELATED TO VRN1*)
Gm	Glyma08g47240.1	AT3G24650	*ABI3* (*ABSCISIC ACID INSENSITIVE 3*)
	Glyma18g38490.1		
	Glyma16g05480.1	AT3G26790	*FUS3* (*FUSCA 3*)
	Glyma16g05480.1		
	Glyma20g04730.1	AT1G28300	*LEC2* (*LEAFY COTYLEDON 2*)
Zm	B4FSX9_MAIZE	AT5G58280	AP2/B3-like protein
	C0PGW2_MAIZE	AT5G58280	AP2/B3-like protein
	C0PJB5_MAIZE	AT3G19184	AP2/B3-like protein
	B4FB68_MAIZE	AT5G66980	AP2/B3-like protein
	B4FGA0_MAIZE	AT4G33280	AP2/B3-like protein
	B6T3X9_MAIZE	AT4G33280	AP2/B3-like protein
	B6TXS3_MAIZE	AT5G66980	AP2/B3-like protein

The ortholog AGI ID and name/description were omitted if a gene has the exactly same *Arabidopsis* ortholog as the previous row

This may explain why we identified a large number of members in the ABI3-VP1 family in *Arabidopsis*.

Our analysis suggested a substantial expansion of the B3 superfamily in the dicot genomes, particularly in the genome of Brassicaceae (*Arabidopsis* and *B. rapa*). In the *B. rapa* gneome, we observed many tandem arrayed *B3* genes, suggesting tandem duplication as a primary mechanism for the expansion of the B3 superfamily in this species. An example is shown in Online Resource 2, and other duplicated *B3* genes were also observed in the *B.*

Table 7 A selected set of *B3* genes associated with quality and yield traits in rice and maize

B3 gene	QTL accession	QTL name	*E* value
LOC_Os11g05740	AQE046	100-seed weight	0
LOC_Os01g48060	CQAS10	1,000-seed weight	0
LOC_Os01g48060	CQAS12	Seed number	0
LOC_Os08g06120	AQEO017	Seed width	1e−135
LOC_Os12g06080	AQFA015	Grain length/width ratio	0
LOC_Os02g41800	AQHE088	Total biomass yield	0
LOC_Os01g67830	AQFF020	Harvest index	0
LOC_Os02g45850	CQJ3	Panicle number	0
LOC_Os02g04810	CQAS26	Yield	0
B4FPB4_MAIZE	q300k3	300-kernel weight	0
B4FXE4_MAIZE	qproc3	Protein content	1e−134
B4FSX9_MAIZE	qgyld8	Grain yield	1e−111
B4G1A0_MAIZE	qgrdm1	%grain dry matter	2e−61
B4FEK2_MAIZE	qproc3	Protein content	8e−32

rapa genome, with some consisting of two copies, whereas others including more than two copies. Further observation indicated that most of the duplicated *B3* genes belong to ABI3-VP1 or REM families, which helps explain the large size of both ABI3-VP1 and REM families in *B. rapa* (Table 2). Duplicated *B3* genes (Romanel et al. 2009) and other transcription factors (Riechmann et al. 2000) have also been found in the *Arabidopsis* genome.

We observed B3 and ARF domain duplications in some B3 proteins, forming more complex domain organizations than currently known. The functional implication of these duplicated domains remains elusive, although domain duplication tends to introduce functional diversification among related proteins in a gene family (Sandhya et al. 2009; Carretero-Paulet et al. 2010). In a gene family, new members could arise via domain duplication or loss. For example, in the MYB family, R2R3-MYB proteins may originate from MYB3R proteins through the loss of R1, or MYB3R proteins emerged through the gain of R1 in an ancient R2R3 predecessor, but R2R3 has been proposed to be a precursor of MYB3R (Braun and Grotewold 1999; Riechmann et al. 2000; Dias et al. 2003; Jiang et al. 2004; Feller et al. 2011). In the B3 superfamily, these duplicated domains we observed led us to hypothesize that domain duplication is the dominant event, suggesting ABI3-VP1 was the founding family in this superfamily. The other four families were then formed via duplication of the B3 domain and/or emergence of other domains. This hypothesis is supported by a higher sequence similarity between *B3* genes in algae and ABI3-VP1 genes (and HSIs) in the land plants (Romanel et al. 2009). In addition, the B3 domain duplication may have contributed to the relative large number of REM proteins (containing two or more B3 domains) identified in *B. rapa* and *Arabidopsis* (Table 2). Domain duplication has been reported previously in the RAV family and other transcription factor families. For example, a RAV protein in the liverwort *Marchantia polymorpha* contains two B3 domains (Swaminathan et al. 2008). And among the 167 basic helix-loop-helix (bHLH) proteins identified in the rice genome, one (OC_Os01g09930) was predicted to contain two duplicated bHLH domains (Li et al. 2006).

The expression level of a gene in a tissue or at a particular developmental stage is a crucial indication for its potential function. Many *B3* genes exhibit tissue-specific expression patterns in the seven plant species, which is especially evident in *Arabidopsis* and rice, two species with large amounts of publicly available microarray data. Some well-studied *Arabidopsis B3* genes, such as *ABI3*, *FUS3*, and *LEC2* were found preferentially expressed in seed and several *REM* genes in reproductive flowers and seeds, as expected. The function of many other *ABI3-VP1* genes preferentially expressed in reproductive tissues remains to be elucidated, and may also play important roles in reproductive development. Seventeen *ARF* genes were found to be expressed in the five tissues of *Arabidopsis*, including seven preferentially expressed in flowers and seeds. It is interesting to note that *ARF2* (*AT5G62000*) has two sets of probes in the Affymetrix ATH1 Genome Array, 247468_at and 247508_at, with the former detecting higher expression in roots and the latter detecting higher expression in flower (Fig. 3a). Nevertheless, this gene was expressed in a relatively constitutive manner across these five tissue types. ARF2 has been shown to promote transitions between multiple stages of *Arabidopsis* development, and regulates leaf senescence and floral organ abscission (Ellis et al. 2005). One HSI family gene, *HSI2* (*AT2G30470*), was also preferentially expressed in *Arabidopsis* seeds, supporting the experimental study of HSI2 (Tsukagoshi et al. 2007), which showed HSI2 and HSL1 repress the sugar-inducible expression of the seed maturation program in seedlings and play an essential role in regulating the transition from seed maturation to seedling growth. Our analysis also indicated that the HSL1-encoding gene, *AT4G32010*, was preferentially expressed in stem (shoot), even though it was also highly expressed in root, flower, and seed, and modestly expressed in leaf (Online Resource 4), indicating its relatively stable expression in these tissues. In additiona, six and two *Arabidopsis B3* genes encoding REM proteins were found preferentially expressed in reproductive organ flowers and seeds, respectively. *AtREM1* (*AT4G31610*), for example, is preferentially expressed in flower (Fig. 3a), consistent with a previous study showing its preferential expression in reproductive meristems (Franco-Zorrilla et al. 2002). One

REM gene, *REM39 (REDUCED VERNALIZATION RESPONSE 1*; *AT3G18990*), however, was preferentially expressed in root, possibly related to the response of root meristems to cold treatment during vernalization. Nonetheless, *REM39* was also highly expressed in flower, seed, and stem, and moderately expressed in leaf (Online Resource 4), and therefore *REM39* is another constitutively expressed gene in these organs. Furthermore, we found that most of gene pairs encoding putative orthologs in *Arabidopsis* and rice were preferentially expressed in different tissues (Table 5). The different expression patterns of *B3* genes among the five major tissue types suggest that many of them may have evolved different functions in the growth and development of *Arabidopsis* (eudicot) and rice (monocot).

With the QTL data in rice (Liang et al. 2008; Ni et al. 2009) and maize (Cannon et al. 2011; Schaeffer et al. 2011), we performed a sequence-based analysis and found many *B3* genes within these known QTL regions. QTL data are often placed on genetic maps, and lack of physical positions of their flanking markers for many QTLs hampered identification of candidate genes in QTL regions. For example, in the maize QTLs we downloaded from maizeGDB (Cannon et al. 2011; Schaeffer et al. 2011), over 80 % have no physical coordinates and were not suitable for such an analysis. The lack of physical positions for the soybean QTLs in SoyBase (Grant et al. 2010), prevented us from doing a similar analysis in a dicot species. Notwithstanding the constraint, some of the associations between *B3* genes and quality traits we presented here in rice and maize (Table 7; Online Resource 6), could be exploited in breeding for traits involving *B3* genes in major crops.

Acknowledgments We are grateful for the research support provided by Alberta Enterprise and Advanced Education, Alberta Innovates Bio Solutions, Genome Alberta, Genome Canada, the Natural Sciences and Engineering Research Council of Canada (Discovery Grant), and the Canada Research Chairs Program.

References

Agarwal P, Kapoor S, Tyagi AK (2011) Transcription factors regulating the progression of monocot and dicot seed development. BioEssays 33:189–202

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796–815

Argout X, Salse J, Aury JM et al (2011) The genome of *Theobroma cacao*. Nat Genet 43:101–108

Barrett T, Troup DB, Wilhite SE et al (2011) NCBI GEO: archive for functional genomics data sets—10 years on. Nucleic Acids Res 39:D1005–D1010

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL (2008) GenBank. Nucleic Acids Res 36:D25–D30

Braun EL, Grotewold E (1999) Newly discovered plant c-myb-like genes rewrite the evolution of the plant myb gene family. Plant Physiol 121:21–24

Braybrook SA, Harada JJ (2008) LECs go crazy in embryo development. Trends Plant Sci 13:624–630

Cannon EK, Birkett SM, Braun BL, Kodavali S, Jennewein DM, Yilmaz A, Antonescu V, Antonescu C, Harper LC, Gardiner JM, Schaeffer ML, Campbell DA, Andorf CM, Andorf D, Lisch D, Koch KE, McCarty DR, Quackenbush J, Grotewold E, Lushbough CM, Sen TZ, Lawrence CJ (2011) POPcorn: an online resource providing access to distributed and diverse maize project data. Int J Plant Genomics 2011:923035

Carretero-Paulet L, Galstyan A, Roig-Villanova I, Martinez-Garcia JF, Bilbao-Castro JR, Robertson DL (2010) Genome-wide classification and evolutionary analysis of the bHLH family of transcription factors in *Arabidopsis*, poplar, rice, moss, and algae. Plant Physiol 153:1398–1412

Chan AP, Crabtree J, Zhao Q et al (2010) Draft genome sequence of the oilseed species *Ricinus communis*. Nat Biotechnol 28: 951–956

Cheng F, Liu S, Wu J, Fang L, Sun S, Liu B, Li P, Hua W, Wang X (2011) BRAD, the genetics and genomics database for Brassica plants. BMC Plant Biol 11:136

Deveshwar P, Bovill WD, Sharma R, Able JA, Kapoor S (2011) Analysis of anther transcriptomes to identify genes contributing to meiosis and male gametophyte development in rice. BMC Plant Biol 11:78

Dias AP, Braun EL, McMullen MD, Grotewold E (2003) Recently duplicated maize R2R3 Myb genes provide evidence for distinct mechanisms of evolutionary divergence after duplication. Plant Physiol 131:610–620

Duvick J, Fu A, Muppirala U, Sabharwal M, Wilkerson MD, Lawrence CJ, Lushbough C, Brendel V (2008) PlantGDB: a resource for comparative plant genomics. Nucleic Acids Res 36:D959–D965

Ellis CM, Nagpal P, Young JC, Hagen G, Guilfoyle TJ, Reed JW (2005) AUXIN RESPONSE FACTOR1 and AUXIN RESPONSE FACTOR2 regulate senescence and floral organ abscission in *Arabidopsis thaliana*. Development 132:4563–4574

Feller A, Machemer K, Braun EL, Grotewold E (2011) Evolutionary and comparative analysis of MYB and bHLH plant transcription factors. Plant J 66:94–116

Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, Hotz HR, Ceric G, Forslund K, Eddy SR, Sonnhammer EL, Bateman A (2008) The Pfam protein families database. Nucleic Acids Res 36:D281–D288

Franco-Zorrilla JM, Cubas P, Jarillo JA, Fernandez-Calvin B, Salinas J, Martinez-Zapater JM (2002) AtREM1, a member of a new family of B3 domain-containing genes, is preferentially expressed in reproductive meristems. Plant Physiol 128:418–427

Gao G, Zhong Y, Guo A, Zhu Q, Tang W, Zheng W, Gu X, Wei L, Luo J (2006) DRTF: a database of rice transcription factors. Bioinformatics 22:1286–1287

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 5:R80

Giraudat J, Hauge BM, Valon C, Smalle J, Parcy F, Goodman HM (1992) Isolation of the *Arabidopsis* ABI3 gene by positional cloning. Plant Cell 4:1251–1261

Goff SA, Ricke D, Lan TH et al (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). Science 296:92–100

Grant D, Nelson RT, Cannon SB, Shoemaker RC (2010) SoyBase, the USDA-ARS soybean genetics and genomics database. Nucleic Acids Res 38:D843–D846

Guilfoyle TJ, Hagen G (2007) Auxin response factors. Curr Opin Plant Biol 10:453–460

Guo A, He K, Liu D, Bai S, Gu X, Wei L, Luo J (2005) DATF: a database of *Arabidopsis* transcription factors. Bioinformatics 21:2568–2569

Hu YX, Wang YX, Liu XF, Li JY (2004) *Arabidopsis* RAV1 is down-regulated by brassinosteroid and may act as a negative regulator during plant development. Cell Res 14:8–15

Hunter S, Apweiler R, Attwood TK et al (2009) InterPro: the integrative protein signature database. Nucleic Acids Res 37:D211–D215

Jain M, Khurana JP (2009) Transcript profiling reveals diverse roles of auxin-responsive genes during reproductive development and abiotic stress in rice. FEBS J 276:3148–3162

Jain M, Nijhawan A, Arora R, Agarwal P, Ray S, Sharma P, Kapoor S, Tyagi AK, Khurana JP (2007) F-box proteins in rice. Genome-wide analysis, classification, temporal and spatial gene expression during panicle and seed development, and regulation by light and abiotic stress. Plant Physiol 143:1467–1483

Jiang C, Gu J, Chopra S, Gu X, Peterson T (2004) Ordered origin of the typical two- and three-repeat Myb genes. Gene 326:13–22

Kim S, Soltis PS, Wall K, Soltis DE (2006) Phylogeny and domain evolution in the APETALA2-like gene family. Mol Biol Evol 23:107–120

Larkin MA, Blackshields G, Brown NP et al (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23:2947–2948

Le BH, Cheng C, Bui AQ et al (2010) Global analysis of gene activity during *Arabidopsis* seed development and identification of seed-specific transcription factors. Proc Natl Acad Sci USA 107: 8063–8070

Lee B, Shin G (2009) CleanEST: a database of cleansed EST libraries. Nucleic Acids Res 37:D686–D689

Li X, Duan X, Jiang H, Sun Y, Tang Y, Yuan Z, Guo J, Liang W, Chen L, Yin J, Ma H, Wang J, Zhang D (2006) Genome-wide analysis of basic/helix-loop-helix transcription factor family in rice and *Arabidopsis*. Plant Physiol 141:1167–1184

Liang C, Jaiswal P, Hebbard C et al (2008) Gramene: a growing plant comparative genomics resource. Nucleic Acids Res 36:D947–D953

Magnani E, Sjolander K, Hake S (2004) From endonucleases to transcription factors: evolution of the AP2 DNA binding domain in plants. Plant Cell 16:2265–2277

Matias-Hernandez L, Battaglia R, Galbiati F, Rubes M, Eichenberger C, Grossniklaus U, Kater MM, Colombo L (2010) VERDANDI is a direct target of the MADS domain ovule identity complex and affects embryo sac differentiation in *Arabidopsis*. Plant Cell 22:1702–1715

McCarty DR, Hattori T, Carson CB, Vasil V, Lazar M, Vasil IK (1991) The Viviparous-1 developmental gene of maize encodes a novel transcriptional activator. Cell 66:895–905

Monke G, Altschmied L, Tewes A, Reidt W, Mock HP, Baumlein H, Conrad U (2004) Seed-specific transcription factors ABI3 and FUS3: molecular interaction with DNA. Planta 219:158–166

Monke G, Seifert M, Keilwagen J, Mohr M, Grosse I, Hahnel U, Junker A, Weisshaar B, Conrad U, Baumlein H, Altschmied L (2012) Toward the identification and regulation of the *Arabidopsis thaliana* ABI3 regulon. Nucleic Acids Res 40:8240–8254

Mulder N, Apweiler R (2007) InterPro and InterProScan: tools for protein sequence classification and comparison. Methods Mol Biol 396:59–70

Mulder NJ, Apweiler R, Attwood TK et al (2007) New developments in the InterPro database. Nucleic Acids Res 35:D224–D228

Ni J, Pujar A, Youens-Clark K, Yap I, Jaiswal P, Tecle I, Tung CW, Ren L, Spooner W, Wei X, Avraham S, Ware D, Stein L, McCouch S (2009) Gramene QTL database: development, content and applications. Database 2009:bap005

Okushima Y, Mitina I, Quach HL, Theologis A (2005a) AUXIN RESPONSE FACTOR 2 (ARF2): a pleiotropic developmental regulator. Plant J 43:29–46

Okushima Y, Overvoorde PJ, Arima K, Alonso JM, Chan A, Chang C, Ecker JR, Hughes B, Lui A, Nguyen D, Onodera C, Quach H, Smith A, Yu G, Theologis A (2005b) Functional genomic analysis of the AUXIN RESPONSE FACTOR gene family members in *Arabidopsis thaliana*: unique and overlapping functions of ARF7 and ARF19. Plant Cell 17:444–463

Peng FY, Weselake RJ (2011) Gene coexpression clusters and putative regulatory elements underlying seed storage reserve accumulation in *Arabidopsis*. BMC Genomics 12:286

R Development Core Team (2010) R: a language and environment for statistical computing. The R Foundation for Statistical Computing, Vienna, Austria

Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, Mesirov JP (2006) GenePattern 2.0. Nat Genet 38:500–501

Reidt W, Wohlfarth T, Ellerstrom M, Czihal A, Tewes A, Ezcurra I, Rask L, Baumlein H (2000) Gene regulation during late embryogenesis: the RY motif of maturation-specific gene promoters is a direct target of the FUS3 gene product. Plant J 21:401–408

Remington DL, Vision TJ, Guilfoyle TJ, Reed JW (2004) Contrasting modes of diversification in the Aux/IAA and ARF gene families. Plant Physiol 135:1738–1752

Rice P, Longden I, Bleasby A (2000) EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 16:276–277

Riechmann JL, Meyerowitz EM (1998) The AP2/EREBP family of plant transcription factors. Biol Chem 379:633–646

Riechmann JL, Heard J, Martin G et al (2000) *Arabidopsis* transcription factors: genome-wide comparative analysis among eukaryotes. Science 290:2105–2110

Romanel EA, Schrago CG, Counago RM, Russo CA, Alves-Ferreira M (2009) Evolution of the B3 DNA binding superfamily: new insights into REM family gene diversification. PLoS One 4:e5791

Sandhya S, Rani SS, Pankaj B, Govind MK, Offmann B, Srinivasan N, Sowdhamini R (2009) Length variations amongst protein domain superfamilies and consequences on structure and function. PLoS One 4:e4981

Schaeffer ML, Harper LC, Gardiner JM, Andorf CM, Campbell DA, Cannon EK, Sen TZ, Lawrence CJ (2011) MaizeGDB: curation and outreach go hand-in-hand. Database (Oxford) 2011:bar022

Schmid M, Davison TS, Henz SR, Pape UJ, Demar M, Vingron M, Scholkopf B, Weigel D, Lohmann JU (2005) A gene expression map of *Arabidopsis thaliana* development. Nat Genet 37:501–506

Schmutz J, Cannon SB, Schlueter J et al (2010) Genome sequence of the palaeopolyploid soybean. Nature 463:178–183

Schnable PS, Ware D, Fulton RS et al (2009) The B73 maize genome: complexity, diversity, and dynamics. Science 326:1112–1115

Sharma R, Mohan Singh RK, Malik G, Deveshwar P, Tyagi AK, Kapoor S, Kapoor M (2009) Rice cytosine DNA methyltransferases—gene expression profiling during reproductive development and abiotic stress. FEBS J 276:6301–6311

Smet ID (2010) Multimodular auxin response controls lateral root development in *Arabidopsis*. Plant Signal Behav 5:580–582

Suzuki M, Kao CY, McCarty DR (1997) The conserved B3 domain of VIVIPAROUS1 has a cooperative DNA binding activity. Plant Cell 9:799–807

Suzuki M, Wang HH, McCarty DR (2007) Repression of the LEAFY COTYLEDON 1/B3 regulatory network in plant embryo development by VP1/ABSCISIC ACID INSENSITIVE 3-LIKE B3 genes. Plant Physiol 143:902–911

Swaminathan K, Peterson K, Jack T (2008) The plant B3 superfamily. Trends Plant Sci 13:647–655

Swarbreck D, Wilks C, Lamesch P et al (2008) The Arabidopsis Information Resource (TAIR): gene structure and function annotation. Nucleic Acids Res 36:D1009–D1014

The Brassica rapa Genome Sequencing Project Consortium, Wang X, Wang H, Wang J et al (2011) The genome of the mesopolyploid crop species *Brassica rapa*. Nat Genet 43:1035–1039

Tsukagoshi H, Saijo T, Shibata D, Morikami A, Nakamura K (2005) Analysis of a sugar response mutant of *Arabidopsis* identified a novel B3 domain protein that functions as an active transcriptional repressor. Plant Physiol 138:675–685

Tsukagoshi H, Morikami A, Nakamura K (2007) Two B3 domain transcriptional repressors prevent sugar-inducible expression of seed maturation genes in *Arabidopsis* seedlings. Proc Natl Acad Sci USA 104:2543–2547

Ulmasov T, Hagen G, Guilfoyle TJ (1999) Activation and repression of transcription by auxin-response factors. Proc Natl Acad Sci USA 96:5844–5849

Waltner JK, Peterson FC, Lytle BL, Volkman BF (2005) Structure of the B3 domain from *Arabidopsis thaliana* protein At1g16640. Protein Sci 14:2478–2483

Weselake RJ, Taylor DC, Rahman MH, Shah S, Laroche A, McVetty PB, Harwood JL (2009) Increasing the flow of carbon into seed oil. Biotechnol Adv 27:866–878

Xing H, Pudake RN, Guo G, Xing G, Hu Z, Zhang Y, Sun Q, Ni Z (2011) Genome-wide identification and expression profiling of auxin response factor (ARF) gene family in maize. BMC Genomics 12:178

Yamasaki K, Kigawa T, Inoue M et al (2004) Solution structure of the B3 DNA binding domain of the *Arabidopsis* cold-responsive transcription factor RAV1. Plant Cell 16:3448–3459

Yamasaki K, Kigawa T, Inoue M, Watanabe S, Tateno M, Seki M, Shinozaki K, Yokoyama S (2008) Structures and evolutionary origins of plant-specific transcription factor DNA-binding domains. Plant Physiol Biochem 46:394–401

Yu J, Hu S, Wang J et al (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). Science 296:79–92

Zhang H, Jin J, Tang L, Zhao Y, Gu X, Gao G, Luo J (2011) PlantTFDB 2.0: update and improvement of the comprehensive plant transcription factor database. Nucleic Acids Res 39:D1114–D1117