Plant Physiology Preview. Published on September 5, 2008, as DOI:10.1104/pp.108.121921

Running head: F-box Genes in *Arabidopsis*, *Oryza*, *Populus*, *Carica* and *Vitis*

Corresponding author:

Dr. Gerald A. Tuskan
Environmental Sciences Division
Oak Ridge National Laboratory
Oak Ridge, TN 37831-6422, USA
Telephone: 865-576-8141
E-mail: tuskanga@ornl.gov

Research area: Genome Analysis

F-box Gene Family is Expanded in Herbaceous Annual Plants Relative to Woody Perennial Plants

Xiaohan Yang, Udaya C. Kalluri, Sara Jawdy, Lee E. Gunter, Tongming Yin, Timothy J. Tschaplinski, David J. Weston, Priya Ranjan and Gerald A. Tuskan

Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831

FOOTNOTES

Financial source: Funding for this research was provided by the U.S. Department of Energy, Office of Science, Biological and Environmental Research Carbon Sequestration Program. ORNL is managed by UT-Battelle, LLC for the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

Corresponding author: Gerald A. Tuskan; email: tuskanga@ornl.gov

ABSTRACT

F-box proteins are generally responsible for substrate recognition in the Skp1-Cullin-F-box complexes that are involved in protein degradation via the ubiquitin-26S proteosome pathway. In plants, F-box genes influence a variety of biological processes such as leaf senescence, branching, self-incompatibility and responses to biotic and abiotic stresses. The number of F-box genes in *Populus* (~320) is less than half that found in *Arabidopsis* (~660) or *Oryza* (~680), even though the total number of genes in *Populus* is equivalent to that in *Oryza* and 1.5 times that in *Arabidopsis*. We performed comparative genomics analysis between the woody perennial plant *Populus* and the herbaceous annual plants *Arabidopsis* and *Oryza* in order to explicate the functional implications of this large gene family. Our analyses reveal interspecific differences in genomic distribution, orthologous relationship, intron evolution, protein domain structure, and gene expression. The set of F-box genes shared by these species appear to be involved in core biological processes essential for plant growth and development; lineage-specific differences primarily occurred because of an expansion of the F-box genes via tandem duplications in *Arabidopsis* and *Oryza*. The number of F-box genes in the newly sequenced woody species *Vitis* (156) and *Carica* (139) is similar to that in *Populus*, supporting the hypothesis that F-box gene family is expanded in herbaceous annual plants relative to woody perennial plants. The present study provides insights into the relationship between the structure and composition of the F-box gene family in herbaceous and woody species and their associated developmental and physiological features.

INTRODUCTION

The ubiquitin-proteasome-dependent pathway is one of the most elaborate protein-degradation systems known. Ubiquitin and ubiquitin-like proteins are important in several cellular processes including targeted protein degradation. Ubiquitination of proteins is commonly carried out by the E3-ubiquitin protein ligase complex specified through an isopeptide linkage between target protein (E3-bound) and ubiquitin (E2-bound). E3 ligases occur in monomeric or multimeric complexes (Mazzucotelli et al., 2006). A well-characterized multi-subunit E3 ligase in plants is the Skp1-Cullin-F-box (SCF) protein complex (Kipreos and Pagano, 2000; Jin et al., 2005). The multiple steps required for protein ubiquitination, specificity, and de-ubiquitination are subject to control at many levels. SCF complexes are known to be regulated by action of the COP9 signalosome, RUB, CAND1, miRNA, and transcriptional/post-transcriptional modification of various component complexes (Chang and Schwechheimer, 2004; Jones-Rhoades et al., 2006).

The distinguishing 50-amino-acid F-box domain is a protein motif that functions as a site of protein-protein interaction (Kipreos and Pagano, 2000). F-box proteins are the substrate-recognition components of SCF ubiquitin-protein ligases. In plants, F-box proteins influence leaf senescence and branching (Woo et al., 2001; Stirnberg et al., 2007), flowering (Durfee et al., 2003; Imaizumi et al., 2005), circadian rhythms (Han et al., 2004; Kevei et al., 2006), self-incompatibility (Qiao et al., 2004; Sijacic et al., 2004; Wang et al., 2004; Takayama and Isogai, 2005), phytochrome signaling (Dieterle et al., 2001), and responses to plant growth regulators (ABA, auxin, ethylene and gibberellin) (Dill et al., 2004; Lai et al., 2004; Badescu and Napier, 2006; Binder et al., 2007) and abiotic (Calderon-Villalobos et al., 2007) and biotic (Kim and Delaney, 2002) factors.

Given the diverse set of developmental traits that F-box proteins are known to influence, it could be argued that long-lived woody plants would require a more abundant or elaborate system of protein degradation, when compared with short-lived herbaceous plants. That is, developmental changes in long-lived woody plants associated with juvenile vs. mature, vegetative vs. reproductive, and dormant vs. non-dormant states would lead to a more abundant set of F-box proteins. An alternate hypothesis may be that short-lived herbaceous plants would require a more strict, coordinated control of ontology in order to successfully complete development over a brief period of time. As such, short-lived plants would contain a more diverse set of gene regulation mechanisms including ubiquitin-proteasome-dependent protein degradation than would long-lived plants. In fact, the F-box gene number is twice as prevalent in the herbaceous annuals *Arabidopsis* and *Oryza* than it is in the perennial *Populus* (~620 and ~690 versus ~300, respectively), even though the number of genes in the *Populus* genome (45,555) is equivalent to that in the *Oryza* genome (42,653) and 1.5 times that in the *Arabidopsis* genome (27,000) (Haas et al., 2003; Tuskan et al., 2006; Ouyang et al., 2007). To illuminate the functional and comparative consequences of the aforementioned observation we compared F-box containing genes in *Arabidopsis*, *Populus* and *Oryza* by analysis of phylogenetic relationships, protein domains, gene expression patterns, gene duplication, and intron evolution.

RESULTS

Genome-wide identification of F-box genes

A HMMER search of a customized database containing the annotated proteins of *Arabidopsis* (TAIR release 7), *Oryza* (TIGR release 5) and *Populus* (JGI release 1.1) using the Pfam HMM profile built from 510 representative seed F-box proteins of diverse organisms including animals and plants identified 656 *Arabidopsis*, 678 *Oryza* and 320 *Populus* predicted proteins (Suppl. Table S1).

In *Populus*, F-box genes were found evenly distributed across all chromosomes in the genome, with the exception of chromosome XIX, on which the density of F-box genes is significantly lower in comparison with the other chromosomes (Table I). Of the 320 F-box genes in *Populus*, 74 (23% of the total) occur as tandem repeats with the largest array containing four genes. An additional 22% of the total number of F-box genes in *Populus* were found within segmental duplications that arose as a result of the salicoid whole-genome duplication event experienced by all members of the genus (Tuskan et al., 2006). Moreover, eight F-box genes that are part of two tandem arrays occurred as the result of at least one paralogous duplication.

The number of F-box genes occurring as tandem repeats in *Arabidopsis* and *Oryza*, 236 (36% of the total) and 291 (43%), respectively, is higher than that in *Populus,* whereas the number of F-box genes occurring as segmental duplicates in *Arabidopsis* and *Oryza*, 46 (7%) and 54 (8%), respectively, is substantially lower than that in *Populus* (Table II). Interestingly, there are two tandem repeats in *Arabidopsis* that occur as homologs in *Populus* and two additional tandem repeats that are homologous in all three species. Each of these arrays contains four genes in tandem order. This suggests that these genomic segments were present in the last shared common ancestor and that this gene family has experienced tandem expansions over the past 120 million years. Finally, in 9% and 18% of the duplications in *Arabidopsis* and *Oryza*, respectively, the F-box motifs were missing in one copy of the two duplicates (data not shown), implying that gene diversification and domain loss has occurred after gene duplication.

Phylogeny and orthologous clustering

To examine the relationship among the 1654 analyzed F-box proteins in *Arabidopsis*, *Oryza* and *Populus*, a gene-based phylogenetic tree was created using full-length protein sequences (Fig. 1). The F-box proteins were divided into 50 distinct phylogenetic groups (designated G01-G50) based on manual delineation of the phylogenetic tree.

To identify orthologous clades (*i.e.*, genes originating from a single ancestral gene in the last common ancestor of the compared genomes) among the F-box proteins in the three plant species a reconciled phylogenetic tree (Suppl. Fig. S1) was constructed by combining the gene tree (Fig. 1) and species tree (*i.e.*, (*Arabidopsis*, *Populus*), *Oryza*)). The F-box proteins were then divided into 7 clades: **AOP** (*Arabidopsis-Oryza-Populus*), **AO** (*Arabidopsis-Oryza*), **OP** (*Oryza-Populus*), **AP** (*Arabidopsis-Populus*), **A** (*Arabidopsis*-specific), **O** (*Oryza*-specific) and **P** (*Populus*-specific). The **AOP** clade contains genes having orthologs in *Arabidopsis*, *Oryza* and *Populus*; the **AP** clade contains genes having orthologs in *Arabidopsis* and *Populus*, etc. It is noteworthy that the number of genes in the **A** clade is equivalent to that in the **O** clade and about six times that in the **P** clade (Fig. 2A), suggesting lineage-specific F-box gene expansions in the annual herbaceous species.

The F-box genes in the **A** clade occurred more often than expected by chance alone in phylogenetic groups G02, G06, G22b and G49 ($p \leq 0.001$) (Table III; Fig. 1), indicating that these groups of genes may have experienced expansion in *Arabidopsis*. Examples of well-characterized genes of **A** clade include CEGENDUO and SON1 in the group G06 and FBX7 in the group G22b (Suppl. Table S2). F-box genes in the **P** clade occurred more often than expected by chance alone in the phylogenetic groups G02, G27, G35 and G39 ($p \leq 0.001$). We hypothesize that these groups of genes may be uniquely related to perennial or woody habit. The **AOP** clade is over-represented in the phylogenetic groups G09, G17, G23, G27, G41, G43, G44 and G48a ($p \leq 0.001$), indicating that these groups of genes, shared by the three plant species, may be involved in basic biological processes required for general plant growth and development. Some well-characterized genes of **AOP** clade associated with common plant growth and development include ARABIDILLO1 and ARABIDILLO2 in the group G12 and AtFBP7 in the group G13 (Suppl. Table S2).

Homologs in other herbaceous monocot, herbaceous eudicot and woody eudicot

To test validity of the hypotheses stated above, we investigated the homology of the F-box proteins in *Arabidopsis*, *Populus* and *Oryza* with genes in other plants by blast search against transcript assemblies of more than 193 plant species (Childs et al., 2007); Suppl. Table S3). Among all herbaceous monocot, herbaceous eudicot and woody eudicot EST datasets, the homologs of clade **A** or **AP** were significantly overrepresented in both sets of eudicots while underrepresented in herbaceous monocots; the homologs of clade **O** were overrepresented in herbaceous monocot dataset while underrepresented in all eudicots; the homologs of clade **P** were overrepresented in woody eudicots including *Vitis* and *Eucalyptus* while underrepresented in herbaceous monocots; and the homologs of clade **AO** were overrepresented in herbaceous eudicots while underrepresented in woody eudicots (Table IV). These data clearly supports the ortholog classification based on the phylogenetic tree and indicates that majority of the genes in the species-specific clade (*i.e.*, **A**, **O** or **P**) share genomic/genic features with other monocots vs. eudicots and/or herbaceous vs. woody species.

Protein motif structure

InterProScan identified more than 90 types of protein motif structures in the 1654 studied F-box proteins (Suppl. Table S4). Thirty-five percent of the F-box proteins (579 out of 1654) contained only a single motif, *i.e.*, the F-box domain. Among the remaining 1075 F-box proteins, 793 proteins (~74%) contain one or more of the 10 most common protein motif structures (Fig. 3). Protein motif structure types 1, 5 and 6 containing F-box associated domains, leucine-rich repeat 2 domains and FBD domains, respectively, occurred more often than expected by chance alone in genes in the **A** clade ($p \leq 0.001$). Protein motif structure types 2 and 9 containing Kelch-related and leucine-rich repeat domains, respectively, occur more often than expected in genes in the **AOP** clade ($p \leq 0.001$), indicating that these motifs may be associated with the basic biological processes shared by all three species.

Intron-exon structure

To contrast gene structure among the examined species, we compared the intron composition of the F-box genes by dividing gene structures into 4 bins: intronless, 1 intron, 2 introns and 3 or more introns per gene. In general, the F-box genes in *Arabidopsis*, *Oryza* and *Populus* contain more intronless genes and fewer 3-or-more-intron genes than expected by chance alone when compared to all other genes in each examined genome ($p \leq 0.0001$). Moreover, F-box genes in the **A**, **P** and **AP** clades contain more intronless gene structure ($p \leq 0.001$), and the **AOP** clade contain more genes with 3-or-more introns ($p \leq 0.001$) than expected by chance alone when compared to all other F-box genes (Table V). Carmel *et al*. (2007) have suggested that the loss of introns is associated with recent evolutionary expansion in large gene families. Our data supports their conclusion and suggests that recent lineage-specific expansion of F-box gene family members has occurred among the three examined species.

Gene expression and predicted function

In *Arabidopsis*, *Oryza* and *Populus*, 333 (51% of the total), 414 (61%) and 141 (44%), respectively, of the predicted F-box genes have expression evidence (*i.e.*, ESTs and/or full-length cDNA data). Among the genes with expression evidence, the **A**, **O** and **P** clades are significantly under-represented and the **AOP** clade is over-represented when compared to all genes (Fig. 2B), demonstrating that genes common to all three species are more frequently represented in such databases and genes uniquely found in *Arabidopsis*, *Populus* or *Oryza* are less common in publicly-available gene expression databases. This observation could be due to sampling error within the tested libraries or differences in expression of recently evolved lineage-specific members of the F-box family, where lineage-specific genes may be infrequently expressed and as of yet uncataloged.

In addition to the F-box containing genes, there are several other genes associated with the SCF complex including CAND1, COP9, Cul1, E1, E2, RBX1, ROC1, RUB1/2 and SKP1/ASK1/ASK2 (Lechner et al., 2006). A Spearman's rank correlation indicated that the 320 *Populus* F-box genes and 146 SCF-associated genes are expressed in a coordinated manner across 9 different *Populus* tissue types (r=0.97, $p \leq 0.0001$) (Fig. 4). Similarly, expression patterns for F-box and SCF complex-related genes in *Arabidopsis* in both the developmental and environmental data sets are correlated (r=0.79, $p \leq 0.002$). These data indicate that there is transcriptional relationship between F-box genes and their associated protein complexes in

Arabidopsis and *Populus*. In addition to the F-box members of the SCF complex there are alternate members of the substrate-specific E3 ligase pathways include HECT, RING, and Ubox proteins. Both *Arabidopsis* and *Populus* have significantly more RING proteins than *Oryza*, and *Oryza* has more CUL3-BTB3 proteins than *Arabidopsis* and *Populus* (Table VI), suggesting that the large differences in numbers of F-box genes in *Arabidopsis* vs. *Populus* or *Oryza* vs. *Populus* are not being compensated for by alternate ubiquitination pathways.

A GO ontology was performed to further characterize the predicted functions of the F-box proteins. Essential biological processes, including signal transduction, flower development, regulation of circadian rhythm, lateral root formation and actin filament-based processes occurred significantly more frequently in the **AOP** clade, whereas the genes associated with responses to biotic stresses were significantly enriched in the **A** clade (Table VII), suggesting that 1) *Arabidopsis*, *Oryza* and *Populus* share some essential biological pathways mediated by F-box proteins and 2) the lineage-specific expansion of F-box genes in *Arabidopsis*.

Vitis and *Carica*

A phylogenetic analysis was performed on the F-box genes in *Populus*, *Vitis* and *Carica* (Suppl. Table S5; Suppl. Fig. S2). Based on the previously described HMMER search criteria we identified 156 and 139 F-box genes in the newly sequenced *Vitis* (Jaillon et al., 2007) and *Carica* genomes, respectively. The *Populus* genome has experienced a whole-genome duplication event (Tuskan et al., 2006) that is not shared by *Vitis* or *Carica* and thus the 320 F-box genes are in agreement with the detected F-box genes in *Vitis* and *Carica*. Interestingly, among the *Populus* F-box genes found in the **AOP** clade 54% had no homologs in the *Vitis* and *Carica* genomes (Fig. 5). In contrast, among the *Populus* F-box genes found uniquely in the **P** clade 75% had no homologs in the *Vitis* and *Carica* genomes. These data clearly show that even though *Populus* experienced a whole-genome duplication that was not shared by *Vitis* nor *Carica*, there are significantly fewer F-box genes in all woody perennials compared to *Arabidopsis* and *Oryza*. These results support our hypothesis that woody perennial plants have few F-box genes relative to herbaceous annuals.

DISCUSSION

F-box proteins represent a large gene family in most eukaryotic organisms and appear to be under-represented in *Populus*, *Vitis* and *Carica* relative to *Arabidopsis* and *Oryza*. The present study has explored the 1) extent of lineage/species specific differential distribution of F-box genes among various subgroups within this gene family and 2) functional implications of differential representation towards cellular and biological processes. For example, the genes involved in actin filament-based process were found uniquely within the **AOP** clade. Alternatively, the up-regulated expression of 38 **A** clade self-incompatibility genes, mainly in pollen, points towards lineage-specific expansion that has played an important role in flower development and successful reproduction in *Arabidopsis* (Suppl. Table S6). Self-incompatibility genes were not found in the dioecious *Populus* (Yin et al., 2008).

In addition to the role of the F-box proteins play in mediating innate signals for developmental transition, another aspect for protein turnover may be related to rapid response to external signals such as environmental cues and stressors. The presence of a much larger F-box gene family in plants (*i.e.*, *Arabidopsis*, *Oryza*, *Populus*, *Vitis* and *Carica*) when compared to less than 100 genes in animals (*i.e.*, human, mouse and *Drosophila*) suggests a predominant role for members of this gene family in management of response to environmental signals in immobile organisms.

Although *Populus* has half as many F-box genes as *Arabidopsis*, our results also confirm that certain F-box genes associated with developmental roles in organ boundary determination (*e.g.*, HAWAIIAN SKIRT), floral organ development (*e.g.*, UFO), and photoperiod and plant growth response signaling [*e.g.*, vernalization-response (FKF1), circadian rhythm signaling (ZTL), EMPFINDLICHER IM DUNKELROTEN LICHT 1 (EID1), ethylene perception (EBF1), gibberellin signaling (SLEEPY1) and auxin signaling (TIR1)] have expanded in *Populus* relative to *Oryza* and *Arabidopsis* (Suppl. Fig. S3).

Yet another distinctive feature of the F-box gene family is the relatively high proportion of intronless genes. Carmel *et al*. (2007) have inferred that high intron density was reached in the early evolutionary history of plants and the last common ancestor of multicellular life forms harbored approximately 3.4 introns per kb, a greater intron density than in most of the extant fungi and in some animals. A recent report also implies that rates of intron creation were higher during earlier periods of plant evolution (Roy and Penny, 2007). Our result support these hypotheses in that the **A**, **P** and **AP** clades are over-represented by intronless gene structure and the **AOP** clade is over-represented by genes with 3-or-more introns. From this perspective, we can conjecture that the F-box gene family members in *Arabidopsis*, *Oryza* and *Populus* have experienced expansion since they last shared a common ancestor.

The SCF ubiquitin-proteasome-dependent pathway is one of the most elaborate and common protein degradation systems. There are alternative pathways to ubiquitination in plants (Jin et al., 2005). The single subunit ubiquitination complex, HECT, is twice as common in *Populus*, which counter-argues the link between reduced F-box gene number and the extent of ubiquitination, although the HECT pathway is thought to be used less frequently in most organisms. Members of the RING and APC complexes are nearly twice as abundant in *Arabidopsis* and *Populus* as they are in *Oryza*, and members of the CUL3-BTB3 complex are nearly twice as abundant in *Oryza* as they are in *Arabidopsis* and *Populus*, suggesting that dicots and monocots utilize these pathways in a differential manner.

The relatively fewer F-box genes in *Populus* must be integral to biological processes in *Populus*. In *Arabidopsis*, functional redundancy among members of this large gene family appears to account for some of the expansion. *TIR1*, *AFB1*, *AFB2* and *AFB3* quadruple mutants are reported to be viable (Dharmasiri et al., 2005). The smaller number of F-box genes in *Populus* suggests that in *Populus* F-box proteins may have evolved to recognize multiple substrates, (*i.e.*, multifunctional/multi-affinity F-box proteins), contain fewer conserved pathways, and/or have alternate pathways that are functionally redundant. The recent expansion of the F-box gene family in *Arabidopsis* and *Oryza* compared to *Populus*, *Vitis* and *Carica* may reflect a comparatively reduced need for ubiquitination-mediated protein turnover in long-lived perennial plants. However, because it is difficult to compare developmental stages between perennial and annual plants, we cannot conclusively determine the extent of the proteome that is ubiquitinated in *Populus* or *Vitis* relative to *Arabidopsis* or *Oryza* for any given ontogeny. Future proteomics investigations may shed light on the extent and prevalence of the ubiquitination pathway in *Populus*, and in particular, whether the SCF pathway is employed to a lesser extent in long-lived plants.

CONCLUSIONS

The present study was undertaken to explore, through comparative bioinformatics, the qualitative and quantitative differences among the F-box genes present in three sequenced plant genomes. We further explored how the relative disparity of the F-box gene family in *Populus* may reflect the biology of this organism. Our results have shed light on several key differences in F-box gene family evolution between the three species, provided insights into the structure and composition of F-box gene family in relation to distinguishing developmental and physiological features, and demonstrate that though the overall family size is smaller in *Populus*, certain subgroups containing genes with known roles in light response and plant growth signaling have expanded in *Populus*, while those related to floral organ function have not. The modes of evolution of the gene families also varied among the examined species, where the F-box gene family appears to have predominantly expanded due to tandem duplication events in annual plants compared to the perennial *Populus*. Future studies employing proteomics and functional genomics approaches will be required to define the overall impact of gene family size, subgroup composition, and individual F-box genes on ubiquitination activity at the cellular-level and the associated plant processes at the whole-organism level.

MATERIAL AND METHODS

Gene identification and annotation

A HMM profile multiple sequence alignment of 510 protein sequences for the F-box domain (PF00646) was downloaded from Pfam. HMMER (Eddy, 1998) was used to search a customized database containing the genome annotations of *Arabidopsis* (TAIR release 7, http://www.arabidopsis.org/), *Oryza* (TIGR release 5, http://rice.plantbiology.msu.edu/), *Populus* (JGI release 1.1, http://genome.jgi-psf.org/Poptr1_1/Poptr1_1.home.html) *Vitis* (http://www.genoscope.cns.fr/externe/GenomeBrowser/Vitis/) (Jaillon et al., 2007), and *Carica* (ftp://asgpb.mhpcc.hawaii.edu/papaya/annotation.genbank_submission/) (Ming et al., 2008) for matches to the HMM profile with the threshold set at 1/100 of the Pfam GA gathering cutoff. The HMMER-selected proteins were then scanned for F-box domains using HMMPfam, HMMsmart, and ProfileScan implemented in InterPro (Mulder et al., 2007). The F-box-containing proteins identified by InterPro scan were used for a BLASTp query (with an e-value cutoff of $1x10^{-20}$) of the original protein database used for the HMMER search. Finally, the BLASTp hits were scanned for F-box domains using HMMPfam, HMMsmart, and ProfileScan implemented in InterPro (Mulder et al., 2007).

Our HMMER-BLASTp-InterProScan strategy initially identified 656, 699 and 336 F-box-containing genes in the genomes of *Arabidopsis*, *Oryza*, and *Populus*, respectively. Of the 699 *Oryza* F-box genes, 21 were transposable elements according to TIGR annotation (http://rice.plantbiology.msu.edu/), and they were excluded from the list of F-box proteins used for downstream analyses. Of the 336 *Populus* F-box genes, 17 genes were deleted because they appeared to represent gene duplicates found on small, unassembled scaffolds with no representation on the JGI *Populus* v1.1 VISTA browser (http://pipeline.lbl.gov/cgi-bin/gateway2?bg=ptr2filt&selector=vista) or because the gene model sequences were truncated by captured gaps. The 319 *Populus* F-box genes (represented by a Jamboree gene model in the JGI official release) were checked manually using the JGI *Populus* genome browser (http://genome.jgi-psf.org/Poptr1_1/Poptr1_1.home.html) to determine whether or not an alternative gene model better represented each gene. The final gene model was chosen based on the criteria of full-length (with start and stop codons), longer transcript/coding region, and, most importantly, higher homology with *Arabidopsis* proteins. As such, 120 *Populus* Jamboree-predicted gene models were replaced by 121 better alternative gene models (Note, the genomic region of a predicted gene model, fgenesh4_pm.C_LG_VI000041, overlapped two alternative F-box genes in tandem and was consequently replaced those models eugene3.00060123 and eugene3.00060124). Therefore, the final *Populus* F-box gene list contains 320 genes (Suppl. Table S1).

For other substrate-specific E3 ligase gene families such as HECT, RING, Ubox, CDC20, and BTB, the *Arabidopsis* genes documented by Mazzucotelli *et al*. **(2006)** were used as queries to search a customized database containing the genome annotations of *Arabidopsis* (TAIR release 7, http://www.arabidopsis.org/), *Oryza* (TIGR release 5, http://rice.plantbiology.msu.edu/) and *Populus* (JGI release 1.1, http://genome.jgi-psf.org/Poptr1_1/Poptr1_1.home.html) by BLASTp with an e-value cutoff of 1 × 10^{-20}. The protein sequences of the BLASTp hits were scanned by InterPro (Mulder et al., 2007) for the signature protein domains: IPR000569 (HECT) for the HECT family, IPR001841 (Zinc finger, RING-type) or IPR013083 (Zinc finger, RING/FYVE/PHD-type) for the RING family, IPR003613 (U-box) for the Ubox family, IPR000002 (Cdc20/Fizzy) for the CDC20 family, and IPR000210 (BTB/POZ-like), IPR013069 (BTB/POZ) or IPR011333 (BTB/POZ fold) for the BTB family.

Phylogenetic tree construction

Sequence alignments were performed with MAFFT (Katoh et al., 2005). The phylogenetic tree was constructed using the relaxed neighbor joining method (Evans et al., 2006). Bootstrap analysis was performed with SEQBOOT and CONSENSUS in the PHYLIP package (Felsenstein, 1989). The gene tree was reconciled with a species tree (*Oryza*, (*Arabidopsis*, *Populus*)) or (*Vitis*, (*Carica*, *Populus*)) using Notung (Chen et al., 2000) to estimate upper and lower bounds on the time of duplication. The tree was displayed using MEGA version 4.0 (Tamura et al., 2007). Orthologs, the genes originating from a single ancestral gene in the last common ancestor of the compared genomes (Koonin, 2005), were identified according to the reconciled phylogenetic trees.

Localization of F-box genes in the genome

F-box gene distribution among chromosomes was evaluated by the observed number of F-box genes compared with their expected number under a Poisson distribution. The expected gene number λ_i on chromosome *i* would be a sample from a Poisson distribution, $\lambda_i = mL_i/\sum L_i$, where, *m* is the total number of genes detected within the assembled sequences; and L_i is the length of chromosome *i*. The probabilities $p(m_i < \lambda_i)$ and $p(m_i > \lambda_i)$ were evaluated under the cumulative Poisson distribution at $\alpha \leq 0.05$ and $\alpha \leq 0.01$ significance level.

Identification of duplicated genes

Identification of homologous chromosome segments in *Populus* resulting from whole-genome duplication events was described in Tuskan *et al*. (2006). Blocks of the same color represent the homologous chromosome segments. The information for *Arabidopsis* gene duplication was obtained from ftp://ftp.tigr.org/pub/data/a_thaliana/ath1/DATA_RELEASE_SUPPLEMENT/. The information for *Oryza* segmental duplication was obtained from http://rice.plantbiology.msu.edu/segmental_dup/100kb/segdup_100kb.shtml. The tandemly duplicated genes in *Oryza* were identified and defined as an array of two or more genes with Smith-Waterman alignment E-value$\leq 1\text{x}10^{-25}$ that were enclosed within a 100-kb window. The analysis of *Populus* tandem gene duplication, obtained from Tuskan *et al.* (2006), used the same criteria as for *Oryza* with added inclusion of maximum 4FTV=1. Segmental duplications in *Populus* were identified by BLASTp as described for *Oryza*, but the expectation value was raised to $E=1e^{-25}$. Protein alignments <50 AA in length were excluded. Segmentally-duplicated pairs of *Populus* genes identified by BLASTp were verified as true paralogs using the VISTA browser (http://pipeline.lbl.gov/cgi-bin/gateway2?bg=ptr2filt&selector=vista) *Populus* duplicates track with default settings (min cons width = 100 bp; conservation identity = 70%) to confirm homology.

Homology search in other plant species

The 1654 F-box protein sequences identified in *Arabidopsis*, *Populus* and *Oryza* were used to query against transcript assemblies from more than 250 plant species (Childs *et al*. 2007) using tBLASTn with e-value cutoffs of $1e^{-10}$, $1e^{-20}$, $1e^{-30}$, $1e^{-40}$, $1e^{-50}$, $1e^{-60}$, $1e^{-70}$ and $1e^{-80}$ (Suppl. Table S7). Difference in distribution of F-box genes among the ortholog clades (*i.e.*, **AOP** (*Arabidopsis-Oryza-Populus*), **AO** (*Arabidopsis-Oryza*), **OP** (*Oryza-Populus*), **AP** (*Arabidopsis-Populus*), **A** (*Arabidopsis*-specific), **O** (*Oryza*-specific) and **P** (*Populus*-specific)) between the initial query F-box genes and the queries that have blast hits decreased with an increase of stringency in the e-value cutoff (from $1e^{-10}$ to $1e^{-80}$). This pattern in ortholog clade distribution was caused by a faster decrease in the percentage of the ortholog clades of F-box proteins in the species-specific clade (**A**, **O** or **P**) suggesting that the phylogenetic signal was decaying quicker in these clades (Suppl. Fig. S4). The ortholog clade distribution of the F-box proteins having blast hits became significantly ($p \leq 1e^{-7}$) different than the ortholog clade distribution of all the query F-box proteins at an e-value cutoff of $1e^{-30}$. This e-value cutoff was used to investigate the distribution of blast hits (the F-box gene homologs) among herbaceous monocot, herbaceous eudicot and woody eudicot which were derived from the transcript assemblies of more than 250 plant species (Childs et al., 2007).

Identification of protein motifs

Protein sequences were scanned for domains using BlastProDom, FPrintScan, HMMPIR, HMMPfam, HMMsmart, HMMTigr, ProfileScan, ScanRegExp and SuperFamily implemented in InterPro (Mulder et al., 2007).

Intron analysis

Intron information was obtained from the TAIR *Arabidopsis* annotation release 7 (http://www.arabidopsis.org/), TIGR *Oryza* annotation release 5 (http://rice.plantbiology.msu.edu/) and JGI *Populus* annotation release 1.1 (http://genome.jgi-psf.org/Poptr1_1/Poptr1_1.home.html).

Expression evidence of F-box genes

Expression evidence from ESTs or full-length cDNAs (FL-cDNA) for *Arabidopsis* genes was obtained from TAIR release 7 (http://www.arabidopsis.org/). Expression evidence from ESTs or FL-cDNAs for *Oryza* genes was obtained from TIGR release 5 (http://rice.plantbiology.msu.edu/). Expression evidence from ESTs or FL-cDNAs for *Populus* genes was determined by a minimum of 97% identity over an alignment of at least 100 bp and at least 80% length of the shorter sequences.

Analysis of gene expression

Two *Arabidopsis* microarray datasets were compiled from AtGenExpress (Schmid et al., 2005; Kilian et al., 2007). The developmental data set is represented by the following organs/tissues: cotyledons, hypocotyl, roots, shoot apex, rosette leaf, senescing leaves, 2nd internode, flowers, sepals, petals, stamens, carpels, siliques and seeds. The gene expression levels are expressed as $LOG_2(x/y)$, where x is the detection signal from the above tissue types and y is the detection signal from seedling. The environmental data set is represented by the following treatments: cold, salt, drought, oxidative, UV-B, heat, pathogen stresses and blue light, far-red light, red light and white light environments. Dark treatment was used as a control for the light experiments. See Kilian *et al*. (2007) and Schmid *et al*. (2005) for further details. K-means clustering of the *Arabidopsis* microarray data was performed using EPCLUST (http://ep.ebi.ac.uk/EP/EPCLUST/) with correlation distance (uncentered).

Co-expression of F-box related genes

In addition to the F-box genes there are also several other F-box related genes involved in the SCF complex including CAND1, COP9, Cul1, E1, E2, RBX1, ROC1, RUB1/2 and SKP1/ASK1/ASK2 (Lechner et al., 2006). In order to compare F-box related gene expression across 12 *Populus* tissues with the expression of the 320 F-box genes [Gene Expression Omnibus Database under the Accession Numbers: GSM146141-GSM146299; Series: GSE6422 and Platform: GPL2618], a Spearman's rank correlation was performed. F-box related genes in *Arabidopsis* were first identified by querying the gene names in the TAIR database (Rhee et al.,

2003) and subsequent sequence information was used to perform a BLASTp on the JGI *Populus* v1.1 browser (http://genome.jgi-psf.org/Poptr1_1/Poptr1_1.home.html). This resulted in a list of 146 *Populus* F-box related genes, the names of which were used to query NimbleGen whole-genome microarray data from 12 different *Populus* tissues. Those genes expressed significantly above background ($Q \leq 0.05$) were said to be expressed in a particular tissue. Microarray data from the 320 F-box genes and F-box related genes were then compared to see if the number of genes expressing in each tissue occurred in a similar pattern across the 12 tissues. A Spearman's rank correlation was calculated.

GO ontology analysis

Gene Ontology annotation of the F-box proteins was performed using BLAST2GO, with parameters optimized for the annotation of *Arabidopsis* sequences (NCBI non-redundant DB, 20 hits maximum and 33 AA minimum HSP-length, e-value-hit-filter of $1e^{-06}$, annotation cutoff value of 55 and GO weight of 5) (Conesa et al., 2005). GO enrichment analysis was performed using FatiGO+ (Al-Shahrour et al., 2007).

ACKNOWLEDGEMENT

We thank Mark Schuster for technical assistance. We thank Stan Wullschleger and Nancy Engle for comments on the manuscript. The authors wish to thank Matias Kirst, Chris Dervinis, Amy Brunner and Andrew Groover for allowing us to use their microarray data.

LITERATURE CITED

Al-Shahrour F, Minguez P, Tarraga J, Medina I, Alloza E, Montaner D, Dopazo J (2007) FatiGO +: a functional profiling tool for genomic data. Integration of functional annotation, regulatory motifs and interaction data with microarray experiments. Nucleic Acids Res **35:** W91-96

Badescu GO, Napier RM (2006) Receptors for auxin: will it all end in TIRs? Trends Plant Sci **11:** 217-223

Binder BM, Walker JM, Gagne JM, Emborg TJ, Hemmann G, Bleecker AB, Vierstra RD (2007) The *Arabidopsis* EIN3 binding F-Box proteins EBF1 and EBF2 have distinct but overlapping roles in ethylene signaling. Plant Cell **19:** 509-523

Calderon-Villalobos LI, Nill C, Marrocco K, Kretsch T, Schwechheimer C (2007) The evolutionarily conserved *Arabidopsis thaliana* F-box protein AtFBP7 is required for efficient translation during temperature stress. Gene **392:** 106-116

Carmel L, Wolf YI, Rogozin IB, Koonin EV (2007) Three distinct modes of intron dynamics in the evolution of eukaryotes. Genome Res **17:** 1034-1044

Chang EC, Schwechheimer C (2004) ZOMES III: the interface between signalling and proteolysis. Meeting on The COP9 Signalosome, Proteasome and eIF3. EMBO Rep **5:** 1041-1045

Chen K, Durand D, Farach-Colton M (2000) NOTUNG: a program for dating gene duplications and optimizing gene family trees. J Comput Biol **7:** 429-447

Childs KL, Hamilton JP, Zhu W, Ly E, Cheung F, Wu H, Rabinowicz PD, Town CD, Buell CR, Chan AP (2007) The TIGR Plant Transcript Assemblies database. Nucleic Acids Res **35:** D846-851

Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics **21:** 3674-3676

Dharmasiri N, Dharmasiri S, Estelle M (2005) The F-box protein TIR1 is an auxin receptor. Nature **435:** 441-445

Dieterle M, Zhou YC, Schafer E, Funk M, Kretsch T (2001) EID1, an F-box protein involved in phytochrome A-specific light signaling. Genes Dev **15:** 939-944

Dill A, Thomas SG, Hu J, Steber CM, Sun TP (2004) The *Arabidopsis* F-box protein SLEEPY1 targets gibberellin signaling repressors for gibberellin-induced degradation. Plant Cell **16:** 1392-1405

Durfee T, Roe JL, Sessions RA, Inouye C, Serikawa K, Feldmann KA, Weigel D, Zambryski PC (2003) The F-box-containing protein UFO and AGAMOUS participate in antagonistic pathways governing early petal development in *Arabidopsis*. Proc Natl Acad Sci U S A **100:** 8571-8576

Eddy SR (1998) Profile hidden Markov models. Bioinformatics **14:** 755-763

Evans J, Sheneman L, Foster J (2006) Relaxed neighbor joining: a fast distance-based phylogenetic tree construction method. J Mol Evol **62:** 785-792

Felsenstein J (1989) PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics **5:** 164-166

Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK, Jr., Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD, et al. (2003) Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res **31:** 5654-5666

Han L, Mason M, Risseeuw EP, Crosby WL, Somers DE (2004) Formation of an SCF(ZTL) complex is required for proper regulation of circadian timing. Plant J **40:** 291-301

Imaizumi T, Schultz TF, Harmon FG, Ho LA, Kay SA (2005) FKF1 F-box protein mediates cyclic degradation of a repressor of CONSTANS in *Arabidopsis*. Science **309:** 293-297

Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, et al. (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature **449:** 463-467

Jin J, Ang XL, Shirogane T, Wade Harper J (2005) Identification of substrates for F-box proteins. Methods Enzymol **399:** 287-309

Jones-Rhoades MW, Bartel DP, Bartel B (2006) MicroRNAS and their regulatory roles in plants. Annu Rev Plant Biol **57:** 19-53

Katoh K, Kuma K, Toh H, Miyata T (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res **33:** 511-518

Kevei E, Gyula P, Hall A, Kozma-Bognar L, Kim WY, Eriksson ME, Toth R, Hanano S, Feher B, Southern MM, et al. (2006) Forward genetic analysis of the circadian clock separates the multiple functions of ZEITLUPE. Plant Physiol **140:** 933-945

Kilian J, Whitehead D, Horak J, Wanke D, Weinl S, Batistic O, D'Angelo C, Bornberg-Bauer E, Kudla J, Harter K (2007) The AtGenExpress global stress expression data set: protocols, evaluation and model data analysis of UV-B light, drought and cold stress responses. Plant J **50:** 347-363

Kim HS, Delaney TP (2002) *Arabidopsis* SON1 is an F-box protein that regulates a novel induced defense response independent of both salicylic acid and systemic acquired resistance. Plant Cell **14:** 1469-1482

Kipreos ET, Pagano M (2000) The F-box protein family. Genome Biol **1:** REVIEWS3002

Koonin EV (2005) Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet **39:** 309-338

Lai CP, Lee CL, Chen PH, Wu SH, Yang CC, Shaw JF (2004) Molecular analyses of the Arabidopsis TUBBY-like protein gene family. Plant Physiol **134:** 1586-1597

Lechner E, Achard P, Vansiri A, Potuschak T, Genschik P (2006) F-box proteins everywhere. Curr Opin Plant Biol **9:** 631-638

Mazzucotelli E, Belloni S, Marone D, De Leonardis AM, Guerra D, Fonzo N, Cattivelli L, Mastrangelo AM (2006) The E3 ubiquitin ligase gene family in plants: Regulation by degradation. Current Genomics **7:** 509-522

Ming R, Hou S, Feng Y, Yu Q, Dionne-Laporte A, Saw JH, Senin P, Wang W, Ly BV, Lewis KL, et al. (2008) The draft genome of the transgenic tropical fruit tree papaya (Carica papaya Linnaeus). Nature **452:** 991-996

Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Buillard V, Cerutti L, Copley R, et al. (2007) New developments in the InterPro database. Nucleic Acids Res **35:** D224-228

Ouyang S, Zhu W, Hamilton J, Lin H, Campbell M, Childs K, Thibaud-Nissen F, Malek RL, Lee Y, Zheng L, et al. (2007) The TIGR Rice Genome Annotation Resource: improvements and new features. Nucleic Acids Res **35:** D883-887

Qiao H, Wang F, Zhao L, Zhou J, Lai Z, Zhang Y, Robbins TP, Xue Y (2004) The F-box protein AhSLF-S2 controls the pollen function of S-RNase-based self-incompatibility. Plant Cell **16:** 2307-2322

Rhee SY, Beavis W, Berardini TZ, Chen GH, Dixon D, Doyle A, Garcia-Hernandez M, Huala E, Lander G, Montoya M, et al. (2003) The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to *Arabidopsis* biology, research materials and community. Nucleic Acids Research **31:** 224-228

Roy SW, Penny D (2007) Patterns of intron loss and gain in plants: intron loss-dominated evolution and genome-wide comparison of *O. sativa* and *A. thaliana*. Mol Biol Evol **24:** 171-181

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

Sijacic P, Wang X, Skirpan AL, Wang Y, Dowd PE, McCubbin AG, Huang S, Kao TH (2004) Identification of the pollen determinant of S-RNase-mediated self-incompatibility. Nature **429:** 302-305

Stirnberg P, Furner IJ, Ottoline Leyser HM (2007) MAX2 participates in an SCF complex which acts locally at the node to suppress shoot branching. Plant J **50:** 80-94

Takayama S, Isogai A (2005) Self-incompatibility in plants. Annu Rev Plant Biol **56:** 467-489

Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) Software Version 4.0. Mol Biol Evol

Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, et al. (2006) The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). Science **313:** 1596-1604

Wang L, Dong L, Zhang Y, Zhang Y, Wu W, Deng X, Xue Y (2004) Genome-wide analysis of S-Locus F-box-like genes in *Arabidopsis thaliana*. Plant Mol Biol **56:** 929-945

Woo HR, Chung KM, Park JH, Oh SA, Ahn T, Hong SH, Jang SK, Nam HG (2001) ORE9, an F-box protein that regulates leaf senescence in *Arabidopsis*. Plant Cell **13:** 1779-1790

Yin T, DiFazio SP, Gunter LE, Zhang X, Sewell MM, Woolbright SA, Allan GJ, Kelleher CT, Douglas CJ, Wang M, et al. (2008) Genome structure and emerging evidence of an incipient sex chromosome in *Populus*. Genome Res **18:** 422-430

FIGURE LEGENDS

Figure 1. A phylogeny created with the full-length F-box protein sequences in *Arabidopsis*, *Oryza* and *Populus*. Ortholog clade: **A** (*Arabidopsis*-specific), **O** (*Oryza*-specific), **P** (*Populus*-specific), **AO** (*Arabidopsis*–Oryza), **OP** (*Oryza-Populus*) **AP** (*Arabidopsis–Populus*), and **AOP** (*Arabidopsis–Oryza–Populus*). The numbers in the Venn diagram represent the number of genes per ortholog clade. Domain structure type defined in Figure 3. The interior colored ring corresponds to the distribution of orthologous clades; the colors are as depicted in the Venn diagram. The outer colored ring corresponds to the domain structure distribution.

Figure 2. Ortholog clades of F-box genes in *Arabidopsis*, *Oryza* and *Populus*. **A)** All F-box genes including expressed and unexpressed genes. **B)** F-box genes with expression evidence. "+", "++", and +++" indicates that genes with expression evidence are over-represented at $p \leq 0.05$, 0.01 and 0.001, respectively, while "-" and "---" indicate under-represented at $p \leq 0.05$ and 0.001, respectively, as compared to all F-box genes. Data label: **AOP**-83-13% represents 83 F-box genes (13% of the total F-box genes in *Arabidopsis*) found in the **AOP** clade.

Figure 3. The ten most common domain structures found in F-box proteins of *Arabidopsis*, *Populus* and *Oryza*. Note, the total number of genes associated with each domain structure are shown in parenthesis.

Figure 4. The number of F-box and SCF-complex-related genes expressed in **A)** *Populus* and **B)** *Arabidopsis* whole-genome microarray data. A Spearman's rank correlation with r=0.97 and r=0.79, respectively, indicates that F-box and SCF-complex-related genes appear to be expressed in a similar pattern across the tissue types included in the analysis ($p \leq 0.01$).

Figure 5. Comparison of F-box orthologs in *Vitis-Carica-Populus* within the *Arabidopsis-Oryza-Populus* context (Note: groups with 3 or more *Populus* genes were excluded). The values in parenthesis are the number of *Populus* F-box genes. Under the *Arabidopsis-Oryza-Populus* context: **AOP** represents *Populus* F-box genes having orthologs in *Arabidopsis* and *Oryza*; **AP** represents *Populus* genes having orthologs in *Arabidopsis* only; **OP** represents *Populus* genes having orthologs in *Oryza* only

and **P** represents *Populus* genes having no orthologs in either *Arabidopsis* or *Oryza*. “No ortholog” legend represents *Populus* F-box genes having no orthologs in *Vitis* and *Carica*; and the “Two orthologs” legend represents *Populus* genes having orthologs in *Vitis* and/or *Carica*.

Table I. Distribution of F-box genes in the *Populus* genome.

Chromosome	Physical length (Mb)	Observed number of F-box gene	Expected number of F-box gene	Distribution test [a] $P(m_{ij} < \lambda ij) < \alpha$ or $P(m_{ij} > \lambda ij) < \alpha$
LG I	32.16	36	30	0.137
LG II	23.44	25	22	0.235
LG III	17.45	13	17	0.235
LG IV	15.08	10	14	0.159
LG V	17.04	19	16	0.196
LG VI	17.68	19	17	0.242
LG VII	11.90	12	11	0.340
LG VIII	15.43	15	15	0.391
LG IX	12.41	12	12	0.395
LG X	19.21	15	18	0.273
LG XI	13.17	17	13	0.082
LG XII	13.03	13	12	0.353
LG XIII	11.50	13	11	0.207
LG XIV	13.68	13	13	0.421
LG XV	10.19	12	10	0.176
LG XVI	12.82	9	12	0.231
LG XVII	5.44	2	5	0.112
LG XVIII	12.44	12	12	0.398
LG XIX	10.23	2	10	0.004

[a] Note, distribution test $P(m_{ij} < \lambda_{ij}) \leq \alpha$ or $P(m_{ij} > \lambda_{ij}) < \alpha$, a Poisson distribution was used to determine the significance of the F-box gene distribution in the *Populus* genome.

Table II. Number of F-box genes resulting from segmental and tandem duplications in *Arabidopsis*, *Oryza*, and *Populus* genomes. Segmental duplications are sub-chromosomal DNA segments with high identity, microsynteny and shared total length.

Item	***Arabidopsis***	***Oryza***	***Populus***
Total number of F-box genes	656	678	320
Tandem duplicates			
Number of F-box genes	236	291	73
% of all F-box genes	36.0	42.9	22.8
Segmental duplicates			
Number of F-box genes	46	54	70
% of all F-box genes	7.0	8.0	21.9

Table III. Over- or under-representation of ortholog clades in each phylogenetic group, as compared to the average distribution across all the 1654 F-box genes. *P*-values were calculated using the cumulative Poisson distribution. The genes column represents the observed and expected (in parenthesis) number of genes.

Group [a]	A		O		P		AO		AP		OP		AOP	
	Genes	*P*-value	Genes	*P*-value	Genes	*P*-value	Genes	*P*-value	Genes	*P*-value	Genes	*P*-value	Genes	*P*-value
G02	79 (29)	0.E+00[b]	0 (28)	1.E-12	14 (4)	4.E-05[b]	0 (12)	6.E-06[c]	6 (7)	0.5054	14(14)	0.4051	5 (25)	2.E-06[c]
G03	19 (13)	0.0447	4 (12)	0.0059	3 (2)	0.1237	8 (5)	0.0960	6 (3)	0.0325	8 (6)	0.1724	5 (11)	0.0349
G05	0 (4)	0.0118	12 (4)	0.0004[b]	0 (1)	0.5262	0 (2)	0.1607	0 (1)	0.3635	2 (2)	0.6495	4 (4)	0.3276
G06	116(30)	0.E+00[b]	1 (28)	1.E-11	0 (4)	0.0129	0 (12)	4.E-06[c]	2 (7)	0.0329	0 (14)	7.E-07[c]	3 (26)	2.E-08[c]
G07	0 (6)	0.0016	5 (6)	0.4349	3 (1)	0.0149	5 (3)	0.0521	0 (1)	0.2318	0 (3)	0.0481	13 (5)	0.0016
G08	0 (8)	0.0004[c]	20 (7)	4.E-05[b]	0 (1)	0.3193	0 (3)	0.0388	0 (2)	0.1654	0 (4)	0.0239	12 (7)	0.0202
G09	0 (4)	0.0193	1 (4)	0.1131	4 (1)	0.9997	0 (2)	0.1969	0 (1)	0.4067	0 (2)	0.1546	11 (3)	0.0002[b]
G10	0 (2)	0.0849	6 (2)	0.0101	0 (0)	0.7000	0 (1)	0.3621	0 (1)	0.5699	0 (1)	0.3113	4 (2)	0.0619
G11	0 (14)	1.E-06[c]	2 (13)	0.0002[c]	1 (2)	0.4067	0 (6)	0.0034	3 (3)	0.6139	50 (7)	0.E+00[b]	0 (12)	8.E-06[c]
G12	1 (9)	0.0017	11 (8)	0.1244	0 (1)	0.2869	2 (4)	0.3108	4 (2)	0.0498	3 (4)	0.4172	14 (7)	0.0086
G13	1 (15)	5.E-06[c]	34(14)	2.E-06[b]	2 (2)	0.6292	20 (6)	2.E-06[b]	0 (3)	0.0324	0 (7)	0.0008[c]	4 (13)	0.0043
G14	2 (3)	0.4903	0 (3)	0.0768	0 (0)	0.6754	0 (1)	0.3272	2 (1)	0.0250	4 (1)	0.0101	3 (2)	0.2022
G15	7 (8)	0.4004	6 (8)	0.3214	4 (1)	0.0081	0 (3)	0.0316	4 (2)	0.0450	13 (4)	0.0001[b]	0 (7)	0.0008[c]
G16	0 (7)	0.0008[c]	25 (7)	1.E-08[b]	0 (1)	0.3554	0 (3)	0.0526	0 (2)	0.1958	0 (3)	0.0339	4 (6)	0.2740
G17	0 (14)	1.E-06[c]	26(13)	0.0004[b]	0 (2)	0.1406	1 (6)	0.0247	3 (3)	0.6265	0 (6)	0.0016	25(12)	0.0002[b]
G21	0 (5)	0.0072	18 (5)	5.E-07[b]	0 (1)	0.4900	2 (2)	0.6682	0 (1)	0.3248	0 (2)	0.0969	0 (4)	0.0151
G22a	0 (10)	0.0001[c]	39 (9)	3.E-14[b]	0 (1)	0.2488	0 (4)	0.0190	0 (2)	0.1116	0 (5)	0.0106	0 (8)	0.0003[c]
G22b	84(30)	0.E+00[b]	1 (29)	1.E-11	4 (4)	0.5536	0 (12)	4.E-06[c]	8 (7)	0.2600	3 (14)	0.0004[c]	23(26)	0.3346
G23	0 (9)	0.0002[c]	3 (8)	0.0378	0 (1)	0.2869	8 (4)	0.0108	0 (2)	0.1397	5 (4)	0.2281	19 (7)	0.0001[b]
G25	1 (4)	0.0642	14 (4)	3.E-05[b]	0 (1)	0.5262	0 (2)	0.1607	0 (1)	0.3635	3 (2)	0.1614	0 (4)	0.0229
G26	0 (7)	0.0008[c]	12 (7)	0.0214	0 (1)	0.3554	0 (3)	0.0526	0 (2)	0.1958	17 (3)	2.E-08[b]	0 (6)	0.0023
G27	1 (10)	0.0007[c]	2 (9)	0.0057	7 (1)	0.0001[b]	1 (4)	0.0945	5 (2)	0.0246	1 (5)	0.0586	22 (8)	2.E-05[b]
G28	0 (4)	0.0151	2 (4)	0.2429	0 (1)	0.5453	4 (2)	0.0313	2 (1)	0.0724	0 (2)	0.1376	9 (4)	0.0038
G29	1 (3)	0.1703	1 (3)	0.1941	1 (0)	0.0794	2 (1)	0.1476	2 (1)	0.0380	2 (2)	0.1954	4 (3)	0.1412
G30	3 (3)	0.6010	0 (3)	0.0481	0 (0)	0.6289	4 (1)	0.0113	6 (1)	1.E-05[b]	0 (2)	0.2194	0 (3)	0.0654
G32	0 (4)	0.0193	7 (4)	0.0368	0 (1)	0.5651	0 (2)	0.1969	0 (1)	0.4067	0 (2)	0.1546	9 (3)	0.0025
G33	0 (4)	0.0247	9 (4)	0.0033	0 (1)	0.5856	2 (2)	0.1971	0 (1)	0.4302	1 (2)	0.4778	3 (3)	0.6144
G34	0 (3)	0.0405	0 (3)	0.0481	1 (0)	0.0794	1 (1)	0.6196	3 (1)	0.0067	0 (2)	0.2194	8 (3)	0.0020
G35	3 (2)	0.2353	0 (2)	0.0969	3 (0)	0.0005[b]	2 (1)	0.0832	2 (1)	0.0196	0 (1)	0.3113	0 (2)	0.1227
G39	0 (4)	0.0151	2 (4)	0.2429	5 (1)	4.E-05[b]	3 (2)	0.0972	4 (1)	0.0030	3 (2)	0.1399	0 (4)	0.0283
G41	0 (8)	0.0005[c]	0 (7)	0.0007[c]	0 (1)	0.3309	2 (3)	0.3907	0 (2)	0.1750	3 (4)	0.5115	26 (7)	2.E-09[b]
G42	0 (10)	0.0001[c]	0 (9)	0.0001[c]	0 (1)	0.2488	0 (4)	0.0190	0 (2)	0.1116	23 (5)	1.E-10[b]	16 (8)	0.0046
G43	0 (7)	0.0008[c]	0 (7)	0.0012	0 (1)	0.3554	0 (3)	0.0526	0 (2)	0.1958	4 (3)	0.2528	25 (6)	2.E-09[b]
G44	0 (6)	0.0027	0 (6)	0.0037	0 (1)	0.4248	9 (2)	0.0002[b]	0 (1)	0.2594	0 (3)	0.0608	15 (5)	0.0001[b]
G46	0 (4)	0.0193	0 (4)	0.0239	0 (1)	0.5651	0 (2)	0.1969	14 (1)	6.E-14[b]	2 (2)	0.2874	0 (3)	0.0348
G48a	3 (10)	0.0136	0 (9)	0.0001[c]	1 (1)	0.5949	3 (4)	0.4411	5 (2)	0.0246	0 (5)	0.0106	27 (8)	5.E-08[b]
G48b	0 (13)	2.E-06[c]	53(12)	0.E+00[b]	0 (2)	2.E-01	0 (5)	0.0046	0 (3)	0.0508	0 (6)	0.0021	0 (11)	1.E-05[c]
G49	85 (29)	0.E+00[b]	7 (28)	3.E-06[c]	0 (4)	0.0149	6 (12)	0.0462	4 (7)	0.2090	8 (14)	0.0694	8 (25)	0.0001[c]
G50	0 (32)	2.E-14[c]	40(30)	0.0306	0 (5)	1.E-02	76(13)	0.E+00[b]	0 (7)	0.0007[c]	2 (15)	4.E-05[c]	10(27)	0.0002[c]

[a] Phylogenetic groups are depicted in Figure 1; **AOP** (*Arabidopsis*-*Oryza*-*Populus*), **AO** (*Arabidopsis*-*Oryza*), **OP** (*Oryza*-*Populus*), **AP** (*Arabidopsis*-*Populus*), **A** (*Arabidopsis*-specific), **O** (*Oryza*-specific) and **P** (*Populus*-specific). Groups with less than 10 genes are not shown.

[b] Over-representation of ortholog clades in each phylogenetic group.

[c] Under-representation of ortholog clades in each phylogenetic group.

Table IV. Over- or under-representation in herbaceous monocot, herbaceous eudicot, or woody eudicot, of homolog obtained by a tBLASTn search of the plant transcript assemblies (Childs et al. 2007, Suppl. Table S3) using the F-box proteins of clade **A**, **O**, **P**, **AO**, **AP** or **OP**, as compared to the **AOP** clade, with an e-value cutoff of $1E^{-30}$. *P*-value was calculated using the cumulative Poisson distribution.

Ortholog clade of query		Number of observed and expected F-box genes by category		
		Herbaceous monocot	Herbaceous eudicot	Woody eudicot
A	Observed	1044	2564	1179
	Expected	1296	2426	1065
	P-value	1.9E-13[c]	2.6E-03[b]	2.8E-04[b]
O	Observed	4869	1288	642
	Expected	1841	3445	1513
	P-value	0.0E+00[b]	0.0E+00[c]	0.0E+00[c]
P	Observed	341	993	500
	Expected	497	929	408
	P-value	5.7E-14[c]	1.8E-02	4.7E-06[b]
AO	Observed	1760	3516	1078
	Expected	1721	3220	1414
	P-value	1.7E-01	0.0E+00[b]	0.0E+00[c]
AP	Observed	792	2516	1320
	Expected	1253	2345	1030
	P-value	0.0E+00[c]	2.1E-04[b]	0.0E+00[b]
OP	Observed	1954	3116	1380
	Expected	1747	3268	1435
	P-value	0.0E+00[b]	3.7E-03[c]	7.5E-02
AOP [a]	Observed	8449	15810	6941

[a] The **AOP** clade was used as a reference for comparison and contains F-box genes that are homologous by clade that were initially identified in *Arabidopsis*, *Oryza* and *Populus*.

[b] Over-representation in herbaceous monocot, herbaceous eudicot, or woody eudicot, of homolog obtained by a tBLASTn search of the plant transcript assemblies.

[c] Under-representation in herbaceous monocot, herbaceous eudicot, or woody eudicot, of homolog obtained by a tBLASTn search of the plant transcript assemblies.

Table V. Over- or under-representation of intron numbers per gene in each ortholog clade, as compared to all the 1654 F-box genes. *P*-values were calculated using the cumulative Poisson distribution. The genes column represent the observed and expected (in parenthesis) number of genes.

Ortholog Clade [a]	Intronless		1 intron		2 introns		3 or more introns	
	Genes	*P*-value	Genes	*P*-value	Genes	*P*-value	Genes	*P*-value
A	210 (145)	2.E-07[b]	76 (86)	0.1443	54 (77)	0.0031[c]	68 (99)	0.0005[c]
O	133 (137)	0.3878	87 (82)	0.2563	76 (73)	0.3472	90 (94)	0.3628
P	41 (21)	3.E-05[b]	10 (12)	0.2985	4 (11)	0.0132	4 (14)	0.0014[c]
AO	15 (60)	6.E-12[c]	36 (36)	0.4260	68 (32)	9.E-09[b]	49 (41)	0.0932
AP	47 (33)	0.0084[b]	18 (20)	0.4089	17 (18)	0.5012	11 (23)	0.0053[c]
OP	47 (68)	0.0038[c]	48 (41)	0.1172	29 (37)	0.1165	69 (47)	0.0010[b]
AOP	94 (123)	0.0037[c]	75 (73)	0.3974	66 (66)	0.4612	112 (85)	0.0018[b]

[a] Ortholog clades are depicted in Fig. 1, such that, **AOP** (*Arabidopsis-Oryza-Populus*), **AO** (*Arabidopsis-Oryza*), **OP** (*Oryza-Populus*), **AP** (*Arabidopsis-Populus*), **A** (*Arabidopsis*-specific), **O** (*Oryza*-specific) and **P** (*Populus*-specific).

[b] Over-representation of intron numbers per gene in each ortholog clade.

[c] Under-representation of intron numbers per gene in each ortholog clade.

Table VI. Number of substrate-specific E3 ligase genes in *Arabidopsis*, *Oryza* and *Populus*.

Complex	Gene Family	Domain for InterProScan	*Arabidopsis*	*Oryza*	*Populus*
HECT	HECT	IPR000569 (HECT)	7	8	16
RING	RING	IPR001841 (Zinc finger, RING-type) or IPR013083 (Zinc finger, RING/FYVE/PHD-type)	477	259	459
U-Box	Ubox	IPR003613 (U-box)	53	60	84
APC	CDC20	IPR000002 (Cdc20/Fizzy)	8	5	9
CUL3-BTB3	BTB	IPR000210 (BTB/POZ-like) or IPR013069 (BTB/POZ) or IPR011333 (BTB/POZ fold)	72	138	85

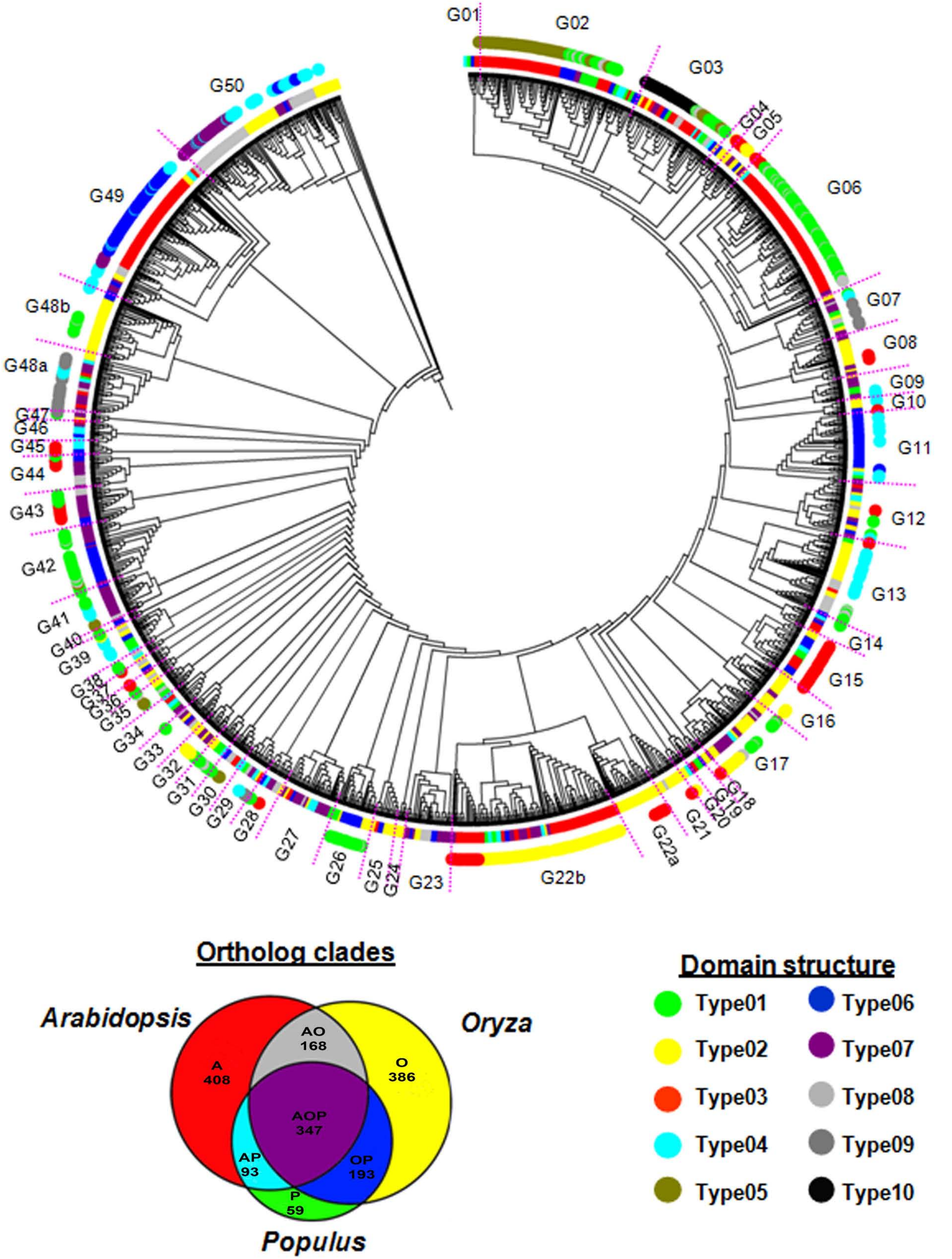
Table VII. GO term enrichment in F-box proteins.

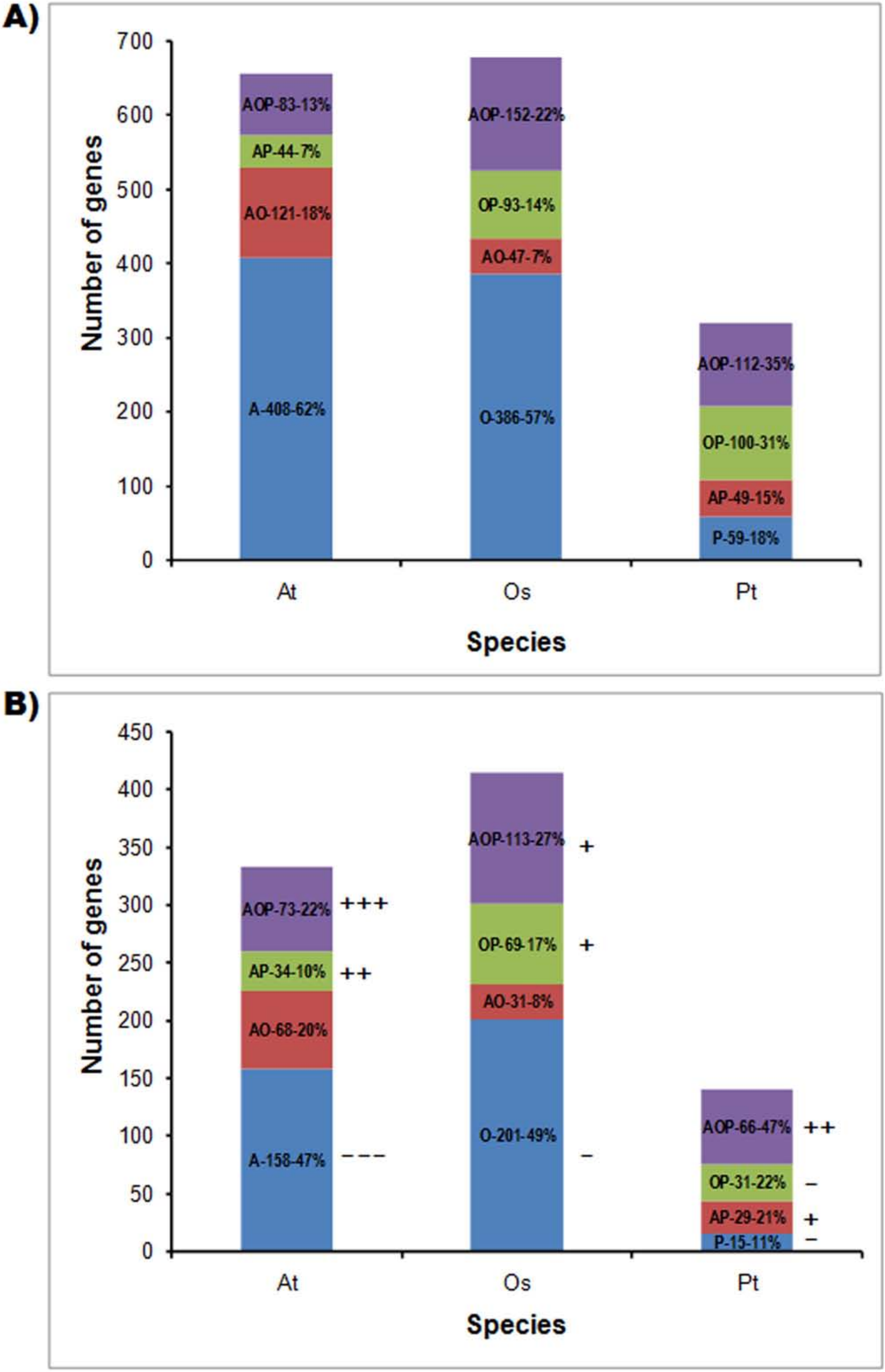
GO ID	Biological process	% total query genes [b]	% total reference genes [b]	Unadj. *P*-value [c]	Adj. *P*-value [c]
AOP vs non-AOP [a]					
GO:0007165	signal transduction	3.98	0	$7E^{-06}$	0.0003
GO:0006355	regulation of transcription	3.98	0	$7E^{-06}$	0.0003
GO:0006511	ubiquitin-dependent protein catabolic process	21.24	9.24	$1E^{-05}$	0.0003
GO:0009908	flower development	3.98	0.17	$5E^{-05}$	0.0009
GO:0043153	entrainment of circadian clock by photoperiod	3.98	0.17	$5E^{-05}$	0.0009
GO:0042752	regulation of circadian rhythm	3.98	0.17	$5E^{-05}$	0.0009
GO:0010114	response to red light	3.98	0.17	$5E^{-05}$	0.0009
GO:0048589	developmental growth	3.98	0.5	$7E^{-04}$	0.0050
GO:0045014	negative regulation of transcription by glucose	3.98	0.5	$7E^{-04}$	0.0050
GO:0009733	response to auxin stimulus	3.98	0.5	$7E^{-04}$	0.0050
GO:0002237	response to molecule of bacterial origin	3.98	0.5	$7E^{-04}$	0.0050
GO:0010311	lateral root formation	3.98	0.5	$7E^{-04}$	0.0050
GO:0031146	SCF-dependent proteasomal ubiquitin-dependent protein catabolic process	2.21	0	$1E^{-03}$	0.0097
GO:0030029	actin filament-based process	1.77	0	$5E^{-03}$	0.0330
A vs non-A [a]					
GO:0009620	response to fungus	5.76	0	$7E^{-08}$	$2.E^{-06}$
GO:0009617	response to bacterium	5.76	0	$7E^{-08}$	$2.E^{-06}$
GO:0044267	cellular protein metabolic process	4.71	0	$2E^{-06}$	$3.E^{-05}$
GO:0051707	response to other organism	3.66	0	$3E^{-05}$	0.0004
GO:0043283	biopolymer metabolic process	4.71	0.31	$5E^{-05}$	0.0006
P vs non-P [a]					
GO:0048589	developmental growth (The increase in size or mass of an entire organism)	28.57	1.21	$4E^{-03}$	$8.E^{-02}$
GO:0045014	negative regulation of transcription by glucose	28.57	1.21	$4E^{-03}$	$8.E^{-02}$
GO:0009733	response to auxin stimulus	28.57	1.21	$4E^{-03}$	$8.E^{-02}$
GO:0002237	response to molecule of bacterial origin	28.57	1.21	$4E^{-03}$	$8.E^{-02}$
GO:0010311	lateral root formation	28.57	1.21	$4E^{-03}$	$8.E^{-02}$

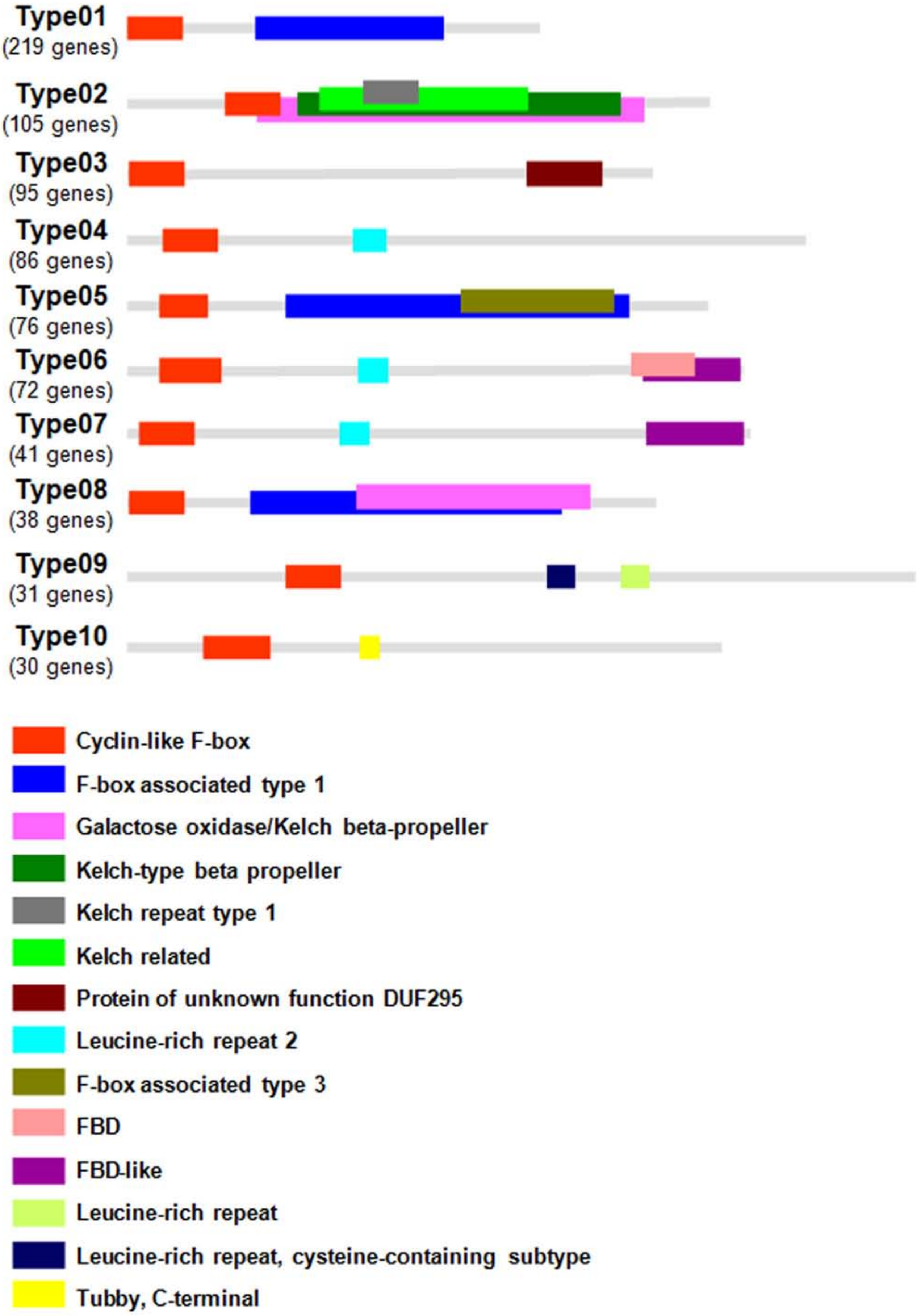
[a] Ortholog clades: **AOP** (*Arabidopsis-Oryza-Populus*), **A** (*Arabidopsis*-specific), **P** (*Populus*-specific), non-**AOP** (all F-box genes in *Arabidopsis*, *Oryza* and *Populus* excluding clade **AOP**), non-**A** (all F-box genes in *Arabidopsis*, *Oryza* and *Populus* excluding clade **A**) and non-**P** (all F-box genes in *Arabidopsis*, *Oryza* and *Populus* excluding clade **P**).

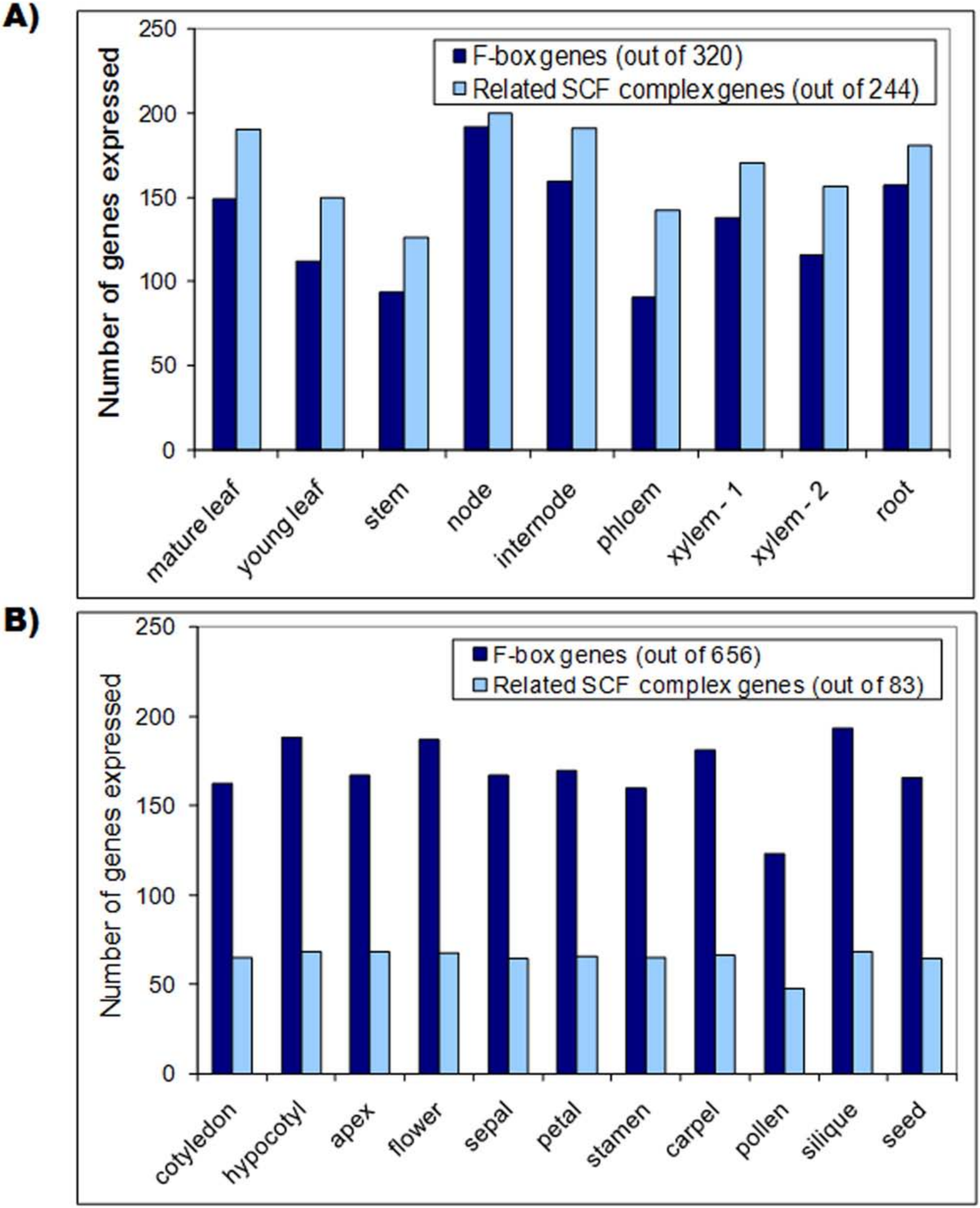
[b] %total query genes corresponding to a GO term is the proportion (%) of GO (biological process)-annotated genes having that specific GO term in the clade **AOP**, **A** or **P**. %total reference genes corresponding to a GO term is the proportion (%) of GO (biological process)-annotated genes having that specific GO term in the **non-AOP**, **non-A** or **non-P** F-box genes.

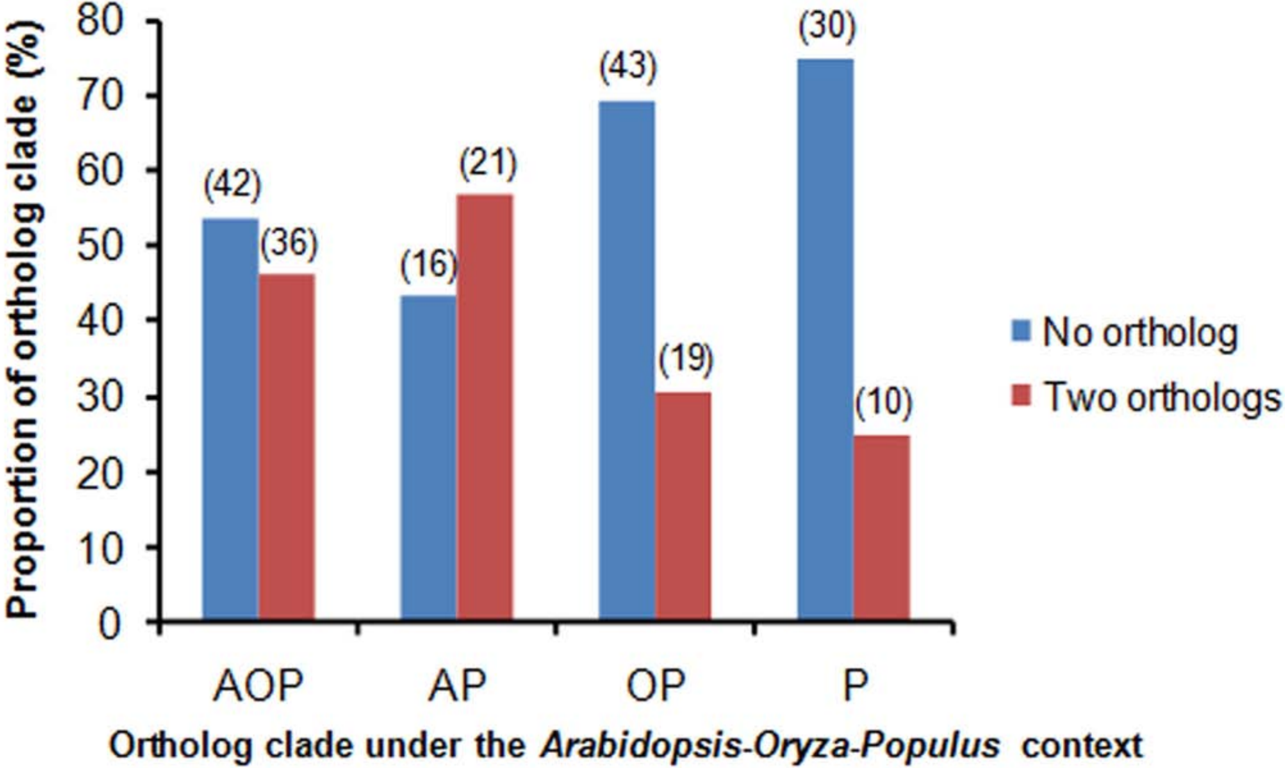
[c] Unadj. *P*-values were obtained by means of a Fisher exact test without adjusting for multiple comparisons to detect over-represented GO terms in the clade **AOP**, **A** and **P**, with **non-AOP**, **non-A** and **non-P**, respectively, as a reference. Adj. *P*-values were the FDR (False Discovery Rate corrections) adjusted *P*-values obtained by Fisher exact test.


G01
G02
G03
G04
G05
G06
G07
G08
G09
G10
G11
G12
G13
G14
G15
G16
G17
G18
G19
G20
G21
G22a
G22b
G23
G24
G25
G26
G27
G28
G29
G30
G31
G32
G33
G34
G35
G36
G37
G38
G39
G40
G41
G42
G43
G44
G45
G46
G47
G48a
G48b
G49
G50
Ortholog clades
Arabidopsis
Oryza
Populus
A 408
AO 168
O 386
AOP 347
AP 93
OP 193
P 59
Domain structure
Type01
Type02
Type03
Type04
Type05
Type06
Type07
Type08
Type09
Type10


A)
700
600
500
400
300
200
100
0
Number of genes
AOP-83-13%
AP-44-7%
AO-121-18%
A-408-62%
AOP-152-22%
OP-93-14%
AO-47-7%
O-386-57%
AOP-112-35%
OP-100-31%
AP-49-15%
P-59-18%
At
Os
Pt
Species
B)
450
400
350
300
250
200
150
100
50
0
Number of genes
AOP-73-22%
+++
AP-34-10%
++
AO-68-20%
A-158-47%
−−−
AOP-113-27%
+
OP-69-17%
+
AO-31-8%
O-201-49%
−
AOP-66-47%
++
OP-31-22%
−
AP-29-21%
+
P-15-11%
−
At
Os
Pt
Species


Type01
(219 genes)
Type02
(105 genes)
Type03
(95 genes)
Type04
(86 genes)
Type05
(76 genes)
Type06
(72 genes)
Type07
(41 genes)
Type08
(38 genes)
Type09
(31 genes)
Type10
(30 genes)
Cyclin-like F-box
F-box associated type 1
Galactose oxidase/Kelch beta-propeller
Kelch-type beta propeller
Kelch repeat type 1
Kelch related
Protein of unknown function DUF295
Leucine-rich repeat 2
F-box associated type 3
FBD
FBD-like
Leucine-rich repeat
Leucine-rich repeat, cysteine-containing subtype
Tubby, C-terminal


A)
Number of genes expressed
250
200
150
100
50
0
■ F-box genes (out of 320)
■ Related SCF complex genes (out of 244)
mature leaf
young leaf
stem
node
internode
phloem
xylem - 1
xylem - 2
root
B)
Number of genes expressed
250
200
150
100
50
0
■ F-box genes (out of 656)
■ Related SCF complex genes (out of 83)
cotyledon
hypocotyl
apex
flower
sepal
petal
stamen
carpel
pollen
silique
seed


Proportion of ortholog clade (%)
0
10
20
30
40
50
60
70
80
(42)
(36)
(16)
(21)
(43)
(19)
(30)
(10)
AOP
AP
OP
P
Ortholog clade under the *Arabidopsis-Oryza-Populus* context
No ortholog
Two orthologs