The 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*

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

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 proteasome 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 \textit{Populus} (\textit{Populus trichocarpa}; approximately 320) is less than half that found in \textit{Arabidopsis} (\textit{Arabidopsis thaliana}; approximately 660) or \textit{Oryza} (\textit{Oryza sativa}; approximately 680), even though the total number of genes in \textit{Populus} is equivalent to that in \textit{Oryza} and 1.5 times that in \textit{Arabidopsis}. We performed comparative genomics analysis between the woody perennial plant \textit{Populus} and the herbaceous annual plants \textit{Arabidopsis} and \textit{Oryza} in order to elucidate 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 \textit{Arabidopsis} and \textit{Oryza}. The number of F-box genes in the newly sequenced woody species \textit{Vitis} (\textit{Vitis vinifera}; 156) and \textit{Carica} (\textit{Carica papaya}; 139) is similar to that in \textit{Populus}, supporting the hypothesis that the F-box gene family is expanded in herbaceous annual plants relative to woody perennial plants. This 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.
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 (Arabidopsis thaliana) and Oryza (Oryza sativa) than it is in the perennial Populus (Populus trichocarpa; approximately 620 versus approximately 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 (The Arabidopsis Information Resource [TAIR] release 7), Oryza (The Institute for Genomic Research [TIGR] release 5), and Populus (U.S. Department of Energy Joint Genome Institute [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 (Supplemental 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 was 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 1,654 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 (Supplemental Fig. S1) was constructed by combining the gene tree (Fig. 1) and the species tree (i.e., [[Arabidopsis, Populus], Oryza]). The F-box proteins were then divided into seven 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, et cetera. 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 ≤ 0.001; Table III; Fig. 1), indicating that these groups of genes may have experienced expansion in Arabidopsis. Examples of well-characterized genes of the A clade include CEGENDUO and SON1 in group G06 and FBX7 in group G22b (Supplemental 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 ≤ 0.001). We hypothesize that these groups of genes may...
be uniquely related to perennial or woody habit. The AOP clade is overrepresented 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 the AOP clade associated with common plant growth and development include ARABIDILLO1 and ARABIDILLO2 in group G12 and AtFBP7 in group G13 (Supplemental Table S2).

**Homologs in Other Herbaceous Monocot, Herbaceous Eudicot, and Woody Eudicot Species**

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

**Protein Motif Structure**

InterProScan identified more than 90 types of protein motif structures in the 1,654 studied F-box proteins (Supplemental Table S4). Thirty-five percent of the F-box proteins (579 of 1,654) contained only a single motif (i.e. the F-box domain). Among the remaining 1,075 F-box proteins, 793 proteins (approximately 74%) contained 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, Leu-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 Leu-rich repeat domains, respectively, occurred 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 structures among the examined species, we compared the intron composition of the F-box genes by dividing gene structures into four bins: intronless, one intron, two introns, and three or more introns per gene. In general, the F-box genes in Arabidopsis, *Oryza*, and *Populus* contain more intronless genes and fewer three-or-more-intron genes than expected by chance alone when compared with 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 genes and fewer three-or-more-intron genes than expected by chance alone when compared with all other genes in each examined genome ($P \leq 0.0001$). However, the AOP clade contains more genes with three or more introns ($P \leq 0.001$), than expected by chance alone when compared with all other F-box genes (Table V). Carmel et al. (2007) suggested that the loss of introns is associated with recent evolutionary ex-
pansion in large gene families. Our data support their conclusion and suggest 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 underrepresented and the AOP clade is overrepresented when compared with 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 the 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 nine different Populus tissue types (r = 0.97, P ≤ 0.001; 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 ≤ 0.002). These data indicate that there is a 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 alternative

---

The text continues with detailed analysis and discussion of the results, including statistical methods and biological implications.
members of the substrate-specific E3 ligase pathways, include HECT, RING, and U-box 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 versus Populus or Oryza versus Populus are not being compensated for by alternative ubiquitination pathways.

A Gene Ontology (GO) analysis 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.

**DISCUSSION**

F-box proteins represent a large gene family in most eukaryotic organisms and appear to be underrepre-

---

**Table IV. Overrepresentation or underrepresentation in herbaceous monocot, herbaceous eudicot, or woody eudicot species of homologs obtained by a tBLASTn search of the plant transcript assemblies (Childs et al., 2007; Supplemental Table S3) using the F-box proteins of clades A, O, P, AO, AP, or OP, as compared with the AOP clade, with an e-value cutoff of 1E−30.**

$P$ value was calculated using the cumulative Poisson distribution.

<table>
<thead>
<tr>
<th>Ortholog Clade of Query</th>
<th>Measure</th>
<th>Herbaceous Monocot</th>
<th>Herbaceous Eudicot</th>
<th>Woody Eudicot</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Observed</td>
<td>1,044</td>
<td>2,564</td>
<td>1,179</td>
</tr>
<tr>
<td></td>
<td>Expected</td>
<td>1,296</td>
<td>2,426</td>
<td>1,065</td>
</tr>
<tr>
<td></td>
<td>$P$</td>
<td>1.9E-13$^c$</td>
<td>2.6E-03$^b$</td>
<td>2.8E-04$^b$</td>
</tr>
<tr>
<td>O</td>
<td>Observed</td>
<td>4,669</td>
<td>1,288</td>
<td>642</td>
</tr>
<tr>
<td></td>
<td>Expected</td>
<td>1,841</td>
<td>3,445</td>
<td>1,513</td>
</tr>
<tr>
<td></td>
<td>$P$</td>
<td>0.0E+00$^b$</td>
<td>0.0E+00$^c$</td>
<td>0.0E+00$^c$</td>
</tr>
<tr>
<td>P</td>
<td>Observed</td>
<td>341</td>
<td>993</td>
<td>500</td>
</tr>
<tr>
<td></td>
<td>Expected</td>
<td>497</td>
<td>929</td>
<td>408</td>
</tr>
<tr>
<td></td>
<td>$P$</td>
<td>5.7E-14$^c$</td>
<td>1.8E-02</td>
<td>4.7E-06$^b$</td>
</tr>
<tr>
<td>AO</td>
<td>Observed</td>
<td>1,760</td>
<td>3,516</td>
<td>1,078</td>
</tr>
<tr>
<td></td>
<td>Expected</td>
<td>1,721</td>
<td>3,220</td>
<td>1,414</td>
</tr>
<tr>
<td></td>
<td>$P$</td>
<td>1.7E-01$^c$</td>
<td>0.0E+00$^b$</td>
<td>0.0E+00$^c$</td>
</tr>
<tr>
<td>AP</td>
<td>Observed</td>
<td>792</td>
<td>2,516</td>
<td>1,320</td>
</tr>
<tr>
<td></td>
<td>Expected</td>
<td>1,253</td>
<td>2,345</td>
<td>1,030</td>
</tr>
<tr>
<td></td>
<td>$P$</td>
<td>0.0E+00$^b$</td>
<td>2.1E-04$^b$</td>
<td>0.0E+00$^b$</td>
</tr>
<tr>
<td>OP</td>
<td>Observed</td>
<td>1,954</td>
<td>3,116</td>
<td>1,380</td>
</tr>
<tr>
<td></td>
<td>Expected</td>
<td>1,747</td>
<td>3,268</td>
<td>1,435</td>
</tr>
<tr>
<td></td>
<td>$P$</td>
<td>0.0E+00$^b$</td>
<td>3.7E-03$^c$</td>
<td>7.5E-02</td>
</tr>
<tr>
<td>AOP$^a$</td>
<td>Observed</td>
<td>8,449</td>
<td>15,810</td>
<td>6,941</td>
</tr>
</tbody>
</table>

$^a$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. $^b$Overrepresentation in herbaceous monocot, herbaceous eudicot, or woody eudicot species of homologs obtained by a tBLASTn search of the plant transcript assemblies. $^c$Underrepresentation in herbaceous monocot, herbaceous eudicot, or woody eudicot species of homologs obtained by a tBLASTn search of the plant transcript assemblies.

---

**Vitis and Carica**

A phylogenetic analysis was performed on the F-box genes in Populus, Vitis (Vitis vinifera), and Carica (Carica papaya; Supplemental Table S5; Supplemental 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; 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 or Carica, there are significantly fewer F-box genes in all woody perennials compared with Arabidopsis and Oryza. These results support our hypothesis that woody perennial plants have fewer F-box genes relative to herbaceous annuals.
sition of 38 A clade self-incompatibility genes, mainly in pollen, points toward lineage-specific expansion that has played an important role in flower development and successful reproduction in Arabidopsis (Supplementary Table S6). Self-incompatibility genes were not found in the dioecious Populus (Yin et al., 2008).

In addition to the role that the F-box proteins play in mediating innate signals for developmental transition, another aspect for protein turnover may be related to rapid responses 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 with less than 100 genes in animals (i.e. human, mouse, and Drosophila) suggests a predominant role for members of this gene family in the management of responses 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], phytochrome A-specific light signaling [EID1], ethylene perception [EBF1], GA signaling [SLEEPY1], and auxin signaling [TIR1]) have expanded in Populus relative to Oryza and Arabidopsis (Supplemental Fig. S3).

Yet another distinctive feature of the F-box gene family is the relatively high proportion of intronless genes. Carmel et al. (2007) suggested that high intron density was reached in the early evolutionary history of plants and that the last common ancestor of multicellular life forms harbored approximately 3.4 introns per kilobase, 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 results support these hypotheses in that the A, P, and AP clades are overrepresented by intronless gene structures and the AOP clade is over-represented by genes with three or more introns. From this perspective, we conjecture that the F-box gene

**Figure 3.** The 10 most common domain structures found in F-box proteins of Arabidopsis, Populus, and Oryza. The total number of genes associated with each domain structure is shown in parentheses.

**Table V.** Overrepresentation or underrepresentation of intron numbers per gene in each ortholog clade, as compared with all 1,654 F-box genes

<table>
<thead>
<tr>
<th>Ortholog Clade</th>
<th>Intronless</th>
<th>One Intron</th>
<th>Two Introns</th>
<th>Three or More Introns</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Genes</td>
<td>P</td>
<td>Genes</td>
<td>P</td>
</tr>
<tr>
<td>A</td>
<td>210 (145)</td>
<td>2.6E-07</td>
<td>76 (86)</td>
<td>0.1443</td>
</tr>
<tr>
<td>O</td>
<td>133 (137)</td>
<td>0.3878</td>
<td>87 (82)</td>
<td>0.2563</td>
</tr>
<tr>
<td>P</td>
<td>41 (21)</td>
<td>3.0E-05</td>
<td>10 (12)</td>
<td>0.2985</td>
</tr>
<tr>
<td>AO</td>
<td>15 (60)</td>
<td>6.6E-12</td>
<td>36 (36)</td>
<td>0.4260</td>
</tr>
<tr>
<td>AP</td>
<td>47 (33)</td>
<td>0.0084³</td>
<td>18 (20)</td>
<td>0.4089</td>
</tr>
<tr>
<td>OP</td>
<td>47 (68)</td>
<td>0.0038³</td>
<td>48 (41)</td>
<td>0.1172</td>
</tr>
<tr>
<td>AOP</td>
<td>94 (123)</td>
<td>0.0037³</td>
<td>75 (73)</td>
<td>0.3974</td>
</tr>
</tbody>
</table>

* Ortholog clades are depicted in Figure 1. ¹Overrepresentation of intron numbers per gene in each ortholog clade. ³Underrepresentation of intron numbers per gene in each ortholog clade.
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 argues against 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. The recent expansion of the F-box gene family in Arabidopsis and Oryza compared with 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 on whether the SCF pathway is employed to a lesser extent in long-lived plants.

CONCLUSION

This 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 members in relation to distinguishing developmental and physiological features, and demonstrated that although the overall family size is smaller in Populus, certain subgroups containing genes with known roles in light response and plant

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

<table>
<thead>
<tr>
<th>Complex</th>
<th>Gene Family</th>
<th>Domain for InterProScan</th>
<th>Arabidopsis</th>
<th>Oryza</th>
<th>Populus</th>
</tr>
</thead>
<tbody>
<tr>
<td>HECT</td>
<td>HECT</td>
<td>IPR0000569 (HECT)</td>
<td>7</td>
<td>8</td>
<td>16</td>
</tr>
<tr>
<td>RING</td>
<td>RING</td>
<td>IPR001841 (zinc finger, RING type) or IPR013083 (zinc finger, RING/FYVE/PHD type)</td>
<td>477</td>
<td>259</td>
<td>459</td>
</tr>
<tr>
<td>U-box</td>
<td>U-box</td>
<td>IPR003613 (U-box)</td>
<td>53</td>
<td>60</td>
<td>84</td>
</tr>
<tr>
<td>APC</td>
<td>CDC20</td>
<td>IPR000002 (Cdc20/Fizzy)</td>
<td>8</td>
<td>5</td>
<td>9</td>
</tr>
<tr>
<td>CUL3-BTB3</td>
<td>BTB</td>
<td>IPR00210 (BTB/POZ-like) or IPR013069 (BTB/POZ) or IPR011333 (BTB/POZ fold)</td>
<td>72</td>
<td>138</td>
<td>85</td>
</tr>
</tbody>
</table>
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 with 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.

**MATERIALS 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 thaliana*; TAIR release 7; http://www.arabidopsis.org/), *Oryza sativa*; TIGR release 5; http://rice.plantbiology.msu.edu/), *Populus trichocarpa*; JGI release 1.1; http://genome.jgi-psf.org/Poptr1_1/Poptr1_1.home.html), *Vitis vinifera*; http://www.genoscope.cns.fr/externe/GenomeBrowser/Vitis/; Jaillon et al., 2007), and *Carica papaya*; http://asgpb.mhcc.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-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 lobbyist.org/cgi-bin/gateway2?bg=ptr2filt&selector=vista) or because the gene
model sequences were truncated by captured gaps. The 319 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 that the genomic region of a predicted gene model, fgenesh4.pm.C_LG.VIII00041, overlapped two alternative F-box genes in *Populus* and was consequently replaced by those models, eugene3.0006122 and eugene3.0006124.) Therefore, the final *Populus* F-box gene list contains 320 genes (Supplemental Table S1).

For other substrate-specific E3 ligase gene families, such as HECT, RING, U-box, CDC20, and BTB, the *Arabidopsis* genes documented by Mazuccotelli 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 $1 \times 10^{-5}$. 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 IPR013803 (zinc finger, RING/FYVE/PH domain type) for the RING family, IPR003613 (U-box) for the U-box family, IPR000002 (Cdc20/Fizzy) for the CDC20 family, and IPR002120 (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 tree was reconciled with a species tree ([*Arabidopsis*, *Populus*, *Oryza*] or [*[Carica*, *Populus*, *Vitis*]) using Notung (Chen et al., 2000) to estimate upper and lower bounds of 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 $\lambda_i$ on chromosome $i$ would be a sample from a Poisson distribution, $\lambda_i = mL_i/L$, where, $m$ is the total number of genes detected within the assembled sequences and $L_i$ is the length of chromosome $i$. The probabilities $p(m < \lambda_i)$ and $p(m > \lambda_i)$ were evaluated under the cumulative Poisson distribution at $\alpha = 0.05$ and $\alpha = 0.01$ significance levels.

Identification of Duplicated Genes

The identification of homologous chromosome segments in *Populus* resulting from whole-genome duplication events was described by 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-values $\leq 1 \times 10^{-7}$ 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 $e = 10^{-5}$ to be at least 10-fold higher than the random expectation. Protein alignments of fewer than 50 amino acids 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?g=p-trz2&f=de&l=sector=vista) *Populus* duplicate track with default settings (minimum conserved region width = 100 bp; conservation identity $\geq 70\%$) to confirm homology.

Homology Search in Other Plant Species

The 1,654 F-box protein sequences identified in *Arabidopsis*, *Populus*, and *Oryza* were used to query against transcript assemblies from 193 plant species (Childs et al., 2007) using tblASTh with e-value cutoffs of $1 \times 10^{-10}$, $1 \times 10^{-20}$, $1 \times 10^{-30}$, $1 \times 10^{-40}$, $1 \times 10^{-50}$, $1 \times 10^{-60}$, and $1 \times 10^{-70}$ (Supplemental Table S7). Differences in the distribution of F-box genes among the ortholog clades (i.e. AOP, AO, AP, A, O, and P) 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 $1 \times 10^{-10}$ to $1 \times 10^{-70}$). 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 clades (A, O, and P), suggesting that the phylogenetic signal was decaying more quickly in these clades (Supplemental Fig. S4). The ortholog clade distribution of the F-box proteins having BLAST hits became significantly ($P \geq 10^{-1}$) different from the ortholog clade distribution of all the query F-box proteins at an e-value cutoff of $1 \times 10^{-40}$. 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 samples 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, HMMfam, HMMmsmart, 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 for *Arabidopsis* genes was obtained from TAIR release 7 (http://www.arabidopsis.org/). Expression evidence from ESTs or full-length cDNAs for *Oryza* genes was obtained from TIGR release 5 (http://rice.plantbiology.msu.edu/). Expression evidence from ESTs or full-length cDNAs for *Populus* genes was determined by a minimum of 92% identity over an alignment of at least 100 bp and at least 80% length of the shorter sequences.

Figure 5. Comparison of F-box orthologs in *Vitis-Carica-Populus* within the Arabidopsis-*Oryza-Populus* context (note that groups with three or more *Populus* genes were excluded). The values in parentheses are numbers 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" represents *Populus* F-box genes having no orthologs in *Vitis* and *Carica*; "Two orthologs" represents two *Populus* genes co-orthologous to one or two genes in *Vitis* and/or *Carica*. The expected gene number $\lambda_i$ on chromosome $i$ would be a sample from a Poisson distribution, $\lambda_i = mL_i/L$, where, $m$ is the total number of genes detected within the assembled sequences and $L_i$ is the length of chromosome $i$. The probabilities $p(m < \lambda_i)$ and $p(m > \lambda_i)$ were evaluated under the cumulative Poisson distribution at $\alpha = 0.05$ and $\alpha = 0.01$ significance levels.
Analysis of Gene Expression

Two Arabidopsis microarray data sets were compiled from AtGenExpress (Schmid et al., 2005; Kilian et al., 2007). The developmental data set is represented by the following organs/tissues: cotyledons, hypocotyls, roots, shoot apices, rosette leaves, senescing leaves, second internodes, 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 seedlings. The environmental data set is represented by the following treatments: cold, salt, drought, oxidative, UV-B light, 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://epibi.uci.edu/EP/EPCLUST/) with correlation distance (uncentered).

Coexpression 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 accession nos GSM146141 to GSM146299; series GSE6422; 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 analysis on the JGI Populus v1.1 browser (http://genome.jgi-psd.org/Ppopr1_1/Ppopr1_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 expressed in each tissue occurred in a similar pattern across the 12 tissues. A Spearman’s rank correlation was calculated.

GO Analysis

GO annotation of the F-box proteins was performed using Blast2GO, with parameters optimized for the annotation of Arabidopsis sequences (National Center for Biotechnology Information nonredundant database; 20 hits maximum and 33 amino acid minimum high scoring pair length; e-value hit filter of \( 10^{-6} \); annotation cutoff value of 55; GO weight of 5; Conesa et al., 2005). GO enrichment analysis was performed using Fatigo+ (Al-Shahrour et al., 2007).

Supplemental Data

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

Supplemental Figure S1. Phylogenetic tree of the F-box genes in Arabidopsis, Oryza, and Populus.

Supplemental Figure S2. Phylogenetic tree of the F-box genes in Carica, Populus, and Vitis.

Supplemental Figure S3. Ortholog clades of well-characterized Arabidopsis genes.

Supplemental Figure S4. The percentage of the ortholog clades of F-box proteins in Arabidopsis, Oryza, and Populus showing homology to other plant species.

Supplemental Table S1. F-box genes in Arabidopsis, Oryza, and Populus.

Supplemental Table S2. Well-characterized Arabidopsis F-box genes.

Supplemental Table S3. The herbaceous monocot, herbaceous eudicot, and woody eudicot species.

Supplemental Table S4. Domain structure of the F-box proteins in Arabidopsis, Populus, and Oryza.

Supplemental Table S5. F-box genes in Carica and Vitis.

F-Box Genes in Arabidopsis, Oryza, Populus, Carica, and Vitis

Supplemental Table S6. Arabidopsis S-locus F-box genes by phylogenetic group and orthologous clade.

Supplemental Table S7. Number of F-box proteins in Arabidopsis, Oryza, and Populus showing homology to other plant species.

ACKNOWLEDGMENTS

We thank Mark Schuster for technical assistance. We thank Stan Wullischger and Nancy Engle for comments on the manuscript. We also thank Mathias Kirst, Chris Dervinis, Amy Brunner, and Andrew Groover for allowing us to use their microarray data.

Received April 28, 2008; accepted August 24, 2008; published September 5, 2008.

LITERATURE CITED


