Evolutionary and Expression Signatures of Pseudogenes in Arabidopsis and Rice[1][C][W][OA]

Cheng Zou, Melissa D. Lehti-Shiu, Françoise Thibaud-Nissen, Tanmay Prakash, C. Robin Buell, and Shin-Han Shiu*

Department of Plant Biology (C.Z., M.D.L.-S., C.R.B., S.-H.S.) and Department of Statistics and Probability (C.Z.), Michigan State University, East Lansing, Michigan 48824; J. Craig Venter Institute, Rockville, Maryland 20850 (F.T.-N.); National Center for Biotechnological Information, National Institutes of Health, Bethesda, Maryland 20894 (F.T.-N.); and Novi High School, Novi, Michigan 48375 (T.P.)

Pseudogenes (Ψ) are nonfunctional genomic sequences resembling functional genes. Knowledge of Ψs can improve genome annotation and our understanding of genome evolution. However, there has been relatively little systemic study of Ψs in plants. In this study, we characterized the evolution and expression patterns of Ψs in Arabidopsis (*Arabidopsis thaliana*) and rice (*Oryza sativa*). In contrast to animal Ψs, many plant Ψs experienced much stronger purifying selection. In addition, plant Ψs experiencing stronger selective constraints tend to be derived from relatively ancient duplicates, suggesting that they were functional for a relatively long time but became Ψs recently. Interestingly, the regions 5' to the first stops in the Ψs have experienced stronger selective constraints compared with 3' regions, suggesting that the 5' regions were functional for a longer period of time after the premature stops appeared. We found that few Ψs have expression evidence, and their expression levels tend to be lower compared with annotated genes. Furthermore, Ψs with expressed sequence tags tend to be derived from relatively recent duplication events, indicating that Ψ expression may be due to insufficient time for complete degeneration of regulatory signals. Finally, larger protein domain families have significantly more Ψs in general. However, while families involved in environmental stress responses have a significant excess of Ψs, transcription factors and receptor-like kinases have lower than expected numbers of Ψs, consistent with their elevated retention rate in plant genomes. Our findings illustrate peculiar properties of plant Ψs, providing additional insight into the evolution of duplicate genes and benefiting future genome annotation.

Pseudogenes (Ψs) are defined as nonfunctional genomic sequences with significant sequence similarity to functional RNA or protein-coding genes (Li, 1983; Vanin, 1985; Balakirev and Ayala, 2003). The first described Ψ has similarity to the *Xenopus laevis* 5S ribosomal RNA gene but is truncated and not expressed (Jacq et al., 1977). Protein-coding sequences are defined as Ψs if degenerative features are present, such as premature stops, frameshift mutations, and truncations of the full-length gene. In this study, we focus on protein-coding Ψs derived from previously functional genes or duplication of existing Ψs (Li, 1983). Depending on the mechanism of the duplication event that created the Ψ copy, Ψs can be classified into two categories. Processed Ψs are derived from retro-transposition events where double-stranded cDNAs derived from reverse transcription events are integrated into the genome. Nonprocessed Ψs are derived from duplication of genomic DNA by whole-genome, tandem, and/or segmental duplication.

Ψs are by definition nonfunctional and therefore are expected to be evolving neutrally, consistent with the finding that approximately 95% of human Ψs are evolving neutrally (Torrents et al., 2003). Lack of function at the protein level, however, does not preclude the possibility that some Ψs may still function as RNA genes (Balakirev and Ayala, 2003; Zheng and Gerstein, 2007). It has been suggested that Ψ transcripts may act as intracellular inhibitors by hybridizing to sense RNA derived from their target genes (McCarrey and Riggs, 1986). The most prominent example is a nitric oxide synthase (NOS) Ψ that contains an approximately 140-bp region that forms a heteroduplex with the functional NOS transcript and suppresses the translation of NOS protein (Korneev et al., 1999). Another example of Ψ function at the RNA level is the *Makorin1-p1* Ψ, which potentially regulates the stability of its homologous coding gene transcript (Hirotsune et al., 2003) and experiences nonneutral evolution (Podlaha and Zhang, 2004); however, its function remains controversial (Gray et al., 2006). Although there are very few examples of Ψs that clearly function at the RNA level, recent large-

---

scale transcriptome sequencing projects and global expression studies using genome tiling arrays indicate that some Ψs may be expressed. For example, close to 10% of distinct transcripts in the FANTOM collection of mouse full-length cDNAs are likely derived from Ψs (Frith et al., 2006). Tiling array studies of the Arabidopsis (Arabidopsis thaliana) transcriptome suggest that approximately 20% of annotated Ψs may be expressed (Yamada et al., 2003). The functional relevance of Ψ expression remains to be explored. But even if Ψs are not functional at the protein or RNA level, they may still aid in the evolution of genes by serving as reservoirs for generating genetic diversity (Balakirev and Ayala, 2003).

Ψs are evolutionary relics of functional components in the genome and provide substantial information regarding the history of gene and genome evolution (Li, 1983; Balakirev and Ayala, 2003). For example, an understanding of the process of pseudogenization is important for estimating how frequently duplicate genes are retained in genomes (Sakai et al., 2007). In addition, studying Ψs and understanding their properties will aid genome annotation. Because Ψs are similar to functional genes, a large number of Ψs are misidentified as potentially functional genes during the genome annotation process (Arabidopsis Genome Initiative, 2000; Lander et al., 2001; Mounsey et al., 2002; Mouse Genome Sequencing Consortium, 2002). For example, in a detailed analysis of the Bric-a-Brac/Tramtrack/Broad domain family in rice (Oryza sativa), 43 out of 192 annotated genes were found to contain frameshifts and/or premature stops (Gingerich et al., 2007). Therefore, distinguishing Ψs from functional genes is important for primary genome annotation. For these reasons, Ψs have been extensively studied in various animal genomes and yeast (Torrents et al., 2003; Zhang et al., 2003, 2004; Lafontaine et al., 2004; Zheng et al., 2007).

In plants, despite a body of literature describing individual Ψs or Ψs in a limited number of gene families, there have been no systematic studies of Ψs. The only genome-wide analysis so far concerned the identification of hundreds of processed Ψs in the Arabidopsis genome (Benovoy and Drouin, 2006). It remains unclear if Ψ evolution in plants is similar to that in animals in terms of Ψ abundance, selection, expression, and patterns of preferential pseudogenization among gene families. In addition, although Ψs contain frameshifts and/or premature stops, they may still lead to the generation of truncated forms of proteins that remain functional (Zheng and Gerstein, 2007). Furthermore, some plant Ψs are likely expressed, but it is unclear if these expressed Ψs have properties distinct from nonexpressed Ψs. Finally, Ψs are remnants of duplicates that were not retained. Since plant gene family sizes vary widely and there is substantial bias in what kinds of duplicates were retained (Maere et al., 2005; Hanada et al., 2008), it is anticipated that the relative abundance of Ψs among plant gene families should complement studies on

gene gains. To address these questions, we identified and examined the properties of thousands of Ψs from two model plant species, Arabidopsis and rice, focusing on the strength of purifying selection on these Ψs, their expression, and their representation in various protein domain families.

RESULTS AND DISCUSSION

Abundant Ψs in the Genomes of Rice and Arabidopsis

The overall Ψ analysis pipeline and the number of sequences found during each step are shown in Figure 1. We identified over 21,000 and 123,000 intergenic sequence regions with significant similarity to known plant proteins from GenBank in Arabidopsis and rice, respectively (Fig. 1). These regions are referred to as "pseudoexons" (Zhang and Chasin, 2004). Note that it is possible that we missed some intergenic regions resembling protein sequences from nonplant species. After repeat masking, the pseudoexons in close proximity to each other and having the same plant protein matches were joined together to form contigs (see "Materials and Methods" and Supplemental Methods S1). These contigs are regarded as putative Ψs (set I). In Arabidopsis, a set of annotated genes has been designated as Ψs. These annotated Arabidopsis Ψs were combined with the intergenic Ψs identified in this study for all subsequent analyses. An independent analysis has been conducted recently to identify Ψs among annotated rice genes, but this data set is not included in our study (Thibaud-Nissen et al., 2009). The locations and the pseudocoding sequences of these Ψs are provided in Supplemental Tables S1 to S3.

A total of 28,330 set I Ψs were identified in rice, versus only 4,771 in Arabidopsis. Based on our earlier analysis of the Bric-a-Brac/Tramtrack/Broad ubiquitin ligase family (Gingerich et al., 2007), we found that a number of the annotated genes in rice may be Ψs. Therefore, the number of rice Ψs is likely even higher than what we have presented here. In most cases, a Ψ is derived from nonfunctionalization of one of a pair of duplicate genes while the other copy maintains its ancestral function (Little, 1982; Li, 1983). However, a substantial number of set I Ψs (Fig. 1) do not have significant within-species matches (paralogs), even though they were originally identified based on significant similarities to protein sequences from other plant species. We further filtered set I Ψs by requiring that these Ψs have one or more non-transposable-element paralogs with alignment lengths of 50 amino acids or greater, eliminating approximately 1,000 Arabidopsis and approximately 20,000 rice putative Ψ sequences. These remaining Ψs are referred to as set II Ψs (Fig. 1). Note that 10 times more set I rice Ψs were eliminated compared with Arabidopsis Ψs. Therefore, it appears that more rice genes either were fast evolving, precluding paralog identification, or were single-copy genes that became Ψs.
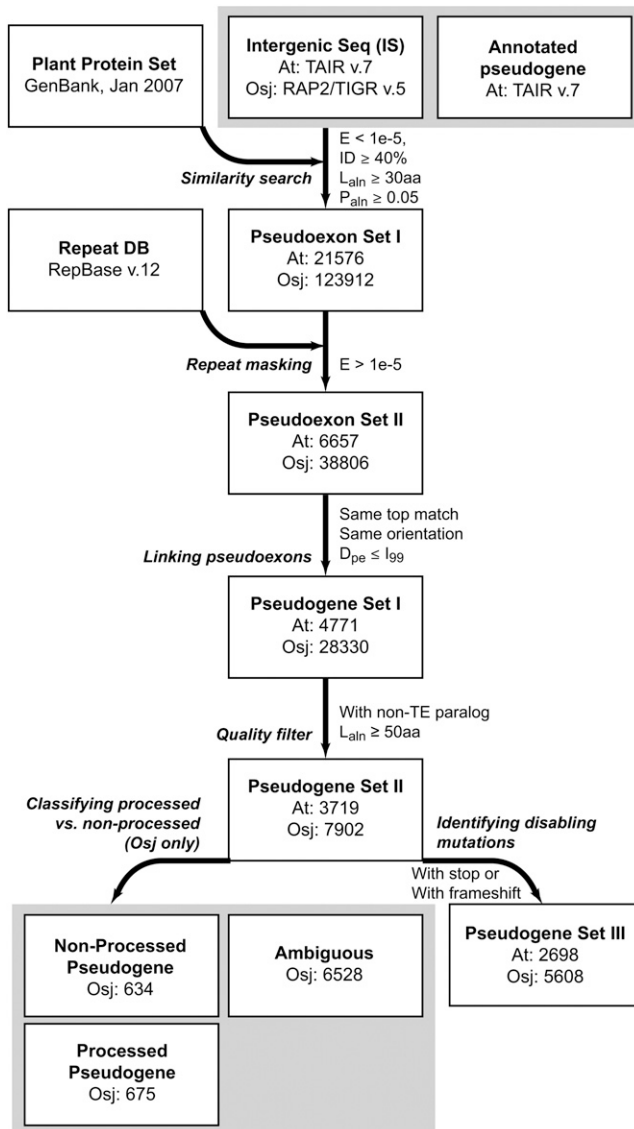
Figure 1. Ψ identification pipeline. The overall procedure for identifying Ψs from Arabidopsis (At) and rice (*O. sativa* subsp. *japonica*; Osj). D_{pe}, Distance between pseudoexons; E, BLAST Expect value; ID, identity; I_{99}, intron length at the 99th percentile; L_{aln}, alignment length; P_{aln}, proportion aligned; TE, transposable element.

After eliminating Ψs without paralog(s), the ratio of set II Ψs between rice and Arabidopsis is approximately 2:1, which correlates with the ratio of genome sizes between Arabidopsis (125 Mb; Arabidopsis Genome Initiative, 2000) and rice (372 Mb; International Rice Genome Sequencing Project, 2005) and the ratio of annotated protein-coding gene numbers (Arabidopsis, 27,029 versus rice, 41,030, after excluding annotated Ψs and repetitive elements). This ratio remains the same for Ψs with apparent "disabling mutations" (premature stops and/or frameshifts, referred to as set III). Given that, the larger set II and III Ψ numbers in rice may simply reflect the fact that a larger pool of

rice genes can become Ψs. In addition to the difference in Ψ number between Arabidopsis and rice, set II Ψs in Arabidopsis are much shorter relative to their paralogs ($P < 0.01$), while the length differences between rice set II Ψs and their paralogs are not significant ($P > 0.09$), potentially reflecting a stronger deletion pressure in Arabidopsis. Taken together, we have identified thousands of putative Ψs from Arabidopsis and rice. A subset of these putative Ψs contains disabling mutations that potentially disrupt protein function. It is not clear if putative Ψs without disabling mutations represent false-positive Ψs. To address this possibility further and to examine the selection pressure imposed on these putative Ψs, we evaluated the strength of purifying selection on Ψs in both plant species.

Strength of Purifying Selection on Plant Ψs

Ψs are expected to evolve neutrally. Therefore, after a sufficient amount of time, the signature of purifying selection at the amino acid level will ultimately be erased. We determined the strength of selection on the Arabidopsis and rice Ψs by estimating ω, the ratio of the nonsynonymous substitution rate (*Ka*) to the synonymous substitution rate (*Ks*) between each set II Ψ and its closest "functional" paralog. Here, a functional paralog (designated "*FP*") is defined as an annotated protein-coding gene that is not a repetitive element or a Ψ according to Arabidopsis and rice genome annotation. The ω value for each pair of *FP* reciprocal best non-Ψ matches (*FP-FP*) was generated as well to represent the selection strength on presumably functional protein-coding genes. In general, set II Ψ-*FP* pairs from both Arabidopsis (Fig. 2A) and rice (Fig. 2B) have significantly higher ω values compared with those of *FP* pairs (Wilcoxon rank sum test, $P < 2.2e-16$ for both; Fig. 2, A and B). Assuming that, immediately after gene duplication, one duplicate becomes nonfunctional and evolves neutrally (ω approximately 1) and the other duplicate is still subject to strong purifying selection with ω approximately 0.2, the ω value for a Ψ-*FP* pair is expected to be approximately 0.6. However, a substantial number of Ψ-*FP* pairs in both plants, particularly in Arabidopsis, have ω values that resemble those of nonneutrally evolving sequences ($P < 2.2e-16$, Wilcoxon rank sum test; Fig. 2, A and B).

One explanation for this difference between Arabidopsis and rice is that a substantial number of Arabidopsis Ψs may be derived from relatively more recent duplication events than rice Ψs and therefore had less time for neutral evolution. A more recent whole genome duplication (WGD) has occurred in the Arabidopsis lineage (20–40 million years ago [Blanc et al., 2003]) compared with the rice lineage (53–94 million years ago [Yu et al., 2005]). This is consistent with the presence of a clear peak at *Ks* approximately 0.8 among *FP-FP* pairs in Arabidopsis but not in rice (Fig. 2, C and D). Using *Ks* as a proxy of time, if the
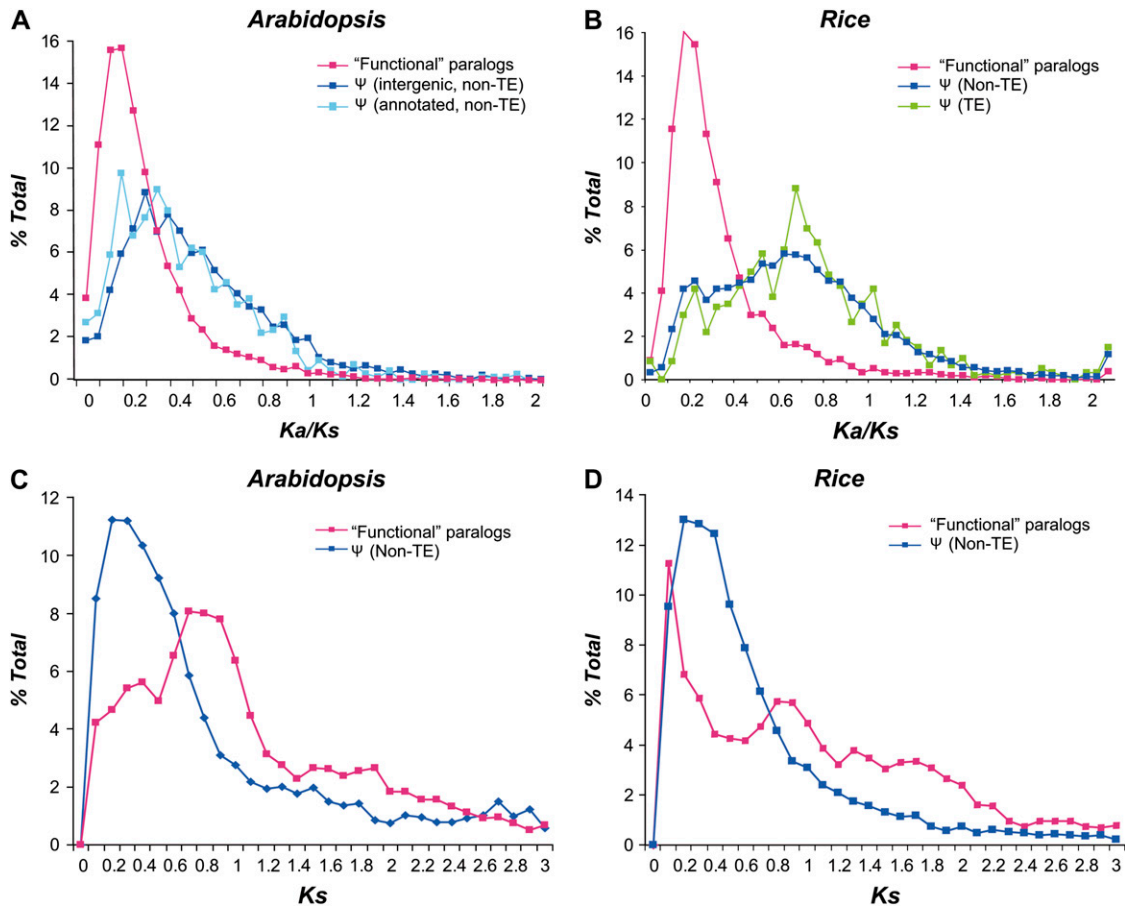
Figure 2. Strength of purifying selection on annotated genes and Ψs. A and B, Frequency distributions of ω, the ratio between the nonsynonymous substitution rate (Ka) and the synonymous substitution rate (Ks) of sequence pairs. A, Arabidopsis sequence pairs. Red symbols indicate annotated "functional" gene (FP) pairs that are not known transposable elements (TE) or Ψs. Blue symbols indicate Ψ-FP pairs; Ψs are those identified in this study (set II; Fig. 1) that do not resemble known TEs. Cyan symbols indicate non-TE annotated Ψ-FP pairs. B, Rice sequence pairs. Red symbols indicate FP-FP pairs that are not known TEs. Blue symbols indicate Ψ-FP pairs; Ψs are those identified in this study that do not resemble known TEs. Green symbols indicate TE Ψ-FP pairs. C and D, Distributions of Ks values of FP-FP pairs (red) and Ψ-FP pairs (blue) in Arabidopsis (C) and rice (D).

relatively more recent WGD in Arabidopsis contributed to the apparently lower ω among its Ψ-FP pairs, we would expect a Ks peak of Ψ-FP pairs following the Ks peak representing the WGD. However, there is no apparent Ψ-FP peak after WGD (Fig. 2, C and D). In addition, the Ks frequency distributions of Arabidopsis and rice Ψ-FP pairs are very similar, indicating that the lower ω values in Arabidopsis Ψs cannot be clearly attributed to its more recent WGD event. The absence of a Ks peak near the WGD also indicates that many Ψs derived from WGD duplicates likely were deleted or became too degenerated for detection within 20 to 40 million years.

Another explanation for the more relaxed selection of rice Ψs compared with those in Arabidopsis is that the larger retrogene pool in rice contributed to an overall higher ω value in rice. A much larger number of retroelements and retrogenes are present in rice compared with those in Arabidopsis (Zhang et al., 2005; Benovoy and Drouin, 2006; Wang et al., 2006). In

addition, given that retrogenes tend to be inserted in genomic regions with mostly irrelevant regulatory contexts, they are expected to be "dead on arrival" (Li et al., 1981; Brosius, 1991). We classified rice Ψs into retro and nonretro categories and determined the strength of purifying selection on these two sets. Although the classification rate is low (18%; Fig. 1), the stringent criteria we used (see "Materials and Methods") ensure low false identification rates of both retro and nonretro Ψs. We found that nonretro Ψs tend to have significantly lower ω values compared with retro Ψs ($P < 4.3e-06$, Wilcoxon rank sum test; Supplemental Fig. S1), indicating that the difference in strength of past selection between Arabidopsis and rice Ψs can be attributed, at least in part, to the differences in selection on duplicated and retro Ψs. In an earlier study of human Ψs, approximately 95% of the 19,724 Ψs were found to be neutrally evolving based on ω values (Torrents et al., 2003). The abundance of retrogenes in the human genome (Kazazian,

2004; Sakai et al., 2007) and the finding that 72% (7,819 of 10,834) human Ψs identified are derived from retrotransposition events (Zheng et al., 2007) likely significantly contributed to the apparently more relaxed selection on human Ψs compared with plant Ψs.

Additional Explanations for the Signature of Nonneutral Evolution among Plant Ψs

There are 685 and 926 Ψ-FP pairs with $\omega \leq 0.2$ in Arabidopsis and rice, respectively. These Ψs have apparently experienced selective constraints as strong as those experienced by most functional genes. To evaluate if these Ψs are false-positive predictions, we compared the distributions of ω values of Ψs with and without disabling mutations in Arabidopsis and rice to those of FP-FP pairs (Fig. 2). Although the median ω values for Ψs with disabling mutations are slightly higher than those without disabling mutations in both plants (Supplemental Fig. S2), Ψs without disabling mutations have much higher ω values compared with those of FP-FP pairs (Supplemental Fig. S2). Therefore, although we cannot rule out the possibility that a few of these truncated sequences may still be functional as protein-coding genes, the ω value distributions strongly suggest that most Ψs without disabling mutations are likely nonfunctional at the protein level and that the absence of disabling mutations may be explained by their recent pseudogenization.

An implicit assumption made when using Ψs with and without disabling mutations to assess the extent of Ψ false positives is that Ψs with disabling mutations are true Ψs, at least in the context of protein function. If a gene has acquired a premature stop or frameshift but remains functional at the protein level, we would expect the ω value of the segment 5' to the first stop to be significantly lower than the value of the 3' segment (after the first stop, in the original reading frame alignable to its paralog) of the same Ψ (set III; Fig. 1). To test this, we eliminated stops and frameshifted positions and determined ω values of 5' and 3' regions. Interestingly, the median ω value of the 5' half of set III Ψs (Arabidopsis, 0.40; rice, 0.48) is indeed significantly smaller than that of the 3' half (Arabidopsis, 0.46; rice, 0.58; Wilcoxon rank sum tests: Arabidopsis, $P < 9.2e-05$; rice, $P < 2.2e-16$; Fig. 3, A and B). Importantly, there is no significant difference between the ω value distributions of 5' regions and 3' regions of FP-FP pairs in both species when we assume that there is a premature stop in the same position as in the pseudogenes (Arabidopsis, $P > 0.12$; rice, $P > 0.82$; Supplemental Fig. S4). These findings indicate that the regions 3' to the disabling mutation may become nonfunctional but the 5' regions escaped nonsense-mediated decay (Chang et al., 2007) and continued to experience purifying selection for some time. It should be noted that 5' segments have almost three times as many sequences with ω values between 0 and 0.05 compared with 3' segments. Therefore, this pattern of differential selection on 5'

and 3' segments of Ψs is noticeable only for very young duplicates with small Ks values.

In addition to differential selection between the 5' and 3' parts of the Ψs, another possibility for strong selective constraints on Ψs is that they were derived from relatively ancient duplicates that became Ψs recently. The rationale behind this hypothesis is that, although some duplicate genes have persisted in the genome for tens to hundreds of millions of years, the Ks frequency distributions of duplicate genes indicate that most duplicates will eventually be lost (Fig. 2, C and D). Consistent with our expectation, Ψs derived from older duplication events (larger Ks) tend to have experienced stronger purifying selection (lower Ka/Ks) than Ψs derived from younger duplication events (Fig. 3, C and D). The selection on younger duplicates is much more relaxed, presumably due to the presence of two functionally identical sequences that are free to accumulate mutations (Hughes, 1994). Therefore, the relatively recent pseudogenization of Ψs derived from old duplicates and the insufficient time for accumulation of neutral mutations in these Ψs likely contributed to the signature of selection of plant Ψs.

Evidence of Ψ Expression

Although the Ψs we have identified may not be functional at the protein level, it remains an open question if they are still useful as RNA genes. To evaluate this possibility, we first determined if set II Ψs are transcribed using EST and massively parallel signature sequencing (MPSS) data sets. Among annotated protein genes, 73% and 49% have either EST and/or MPSS evidence in Arabidopsis and rice, respectively (Table I). In contrast, significantly fewer Ψs have evidence of expression (2%–5% and 2%–3% in Arabidopsis and rice, respectively; Fisher's exact test, $P < 2e-16$ in both cases; Table I). Our findings indicate that the majority of Ψs are no longer expressed at a sufficiently high level to be detected by EST sequencing and MPSS approaches. Interestingly, studies of mammalian Ψs have shown that 2% to 5% of Ψs are expressed based on similar expression tag analysis (Yano et al., 2004; Harrison et al., 2005; Zheng et al., 2005; Frith et al., 2006). The consistency in the proportion of Ψs with sequence tags between mammals and plants suggests that Ψs contribute similarly to the noncoding RNA gene repertoire in these divergent taxa.

In addition to the Ψ expression tags, we analyzed tiling array data from Arabidopsis and rice to compare the levels of Ψ expression with that of other sequence features (Fig. 4). Intron sequences in general have significantly lower hybridization intensities compared with those of exons and Ψs in both rice and Arabidopsis (Wilcoxon rank sum tests, all $P < 2e-16$), indicating that a large number of Ψs may be expressed but at a relatively low level. In addition, in Arabidopsis, exon sense expression is significantly higher than Ψ expression in either the sense or antisense direction
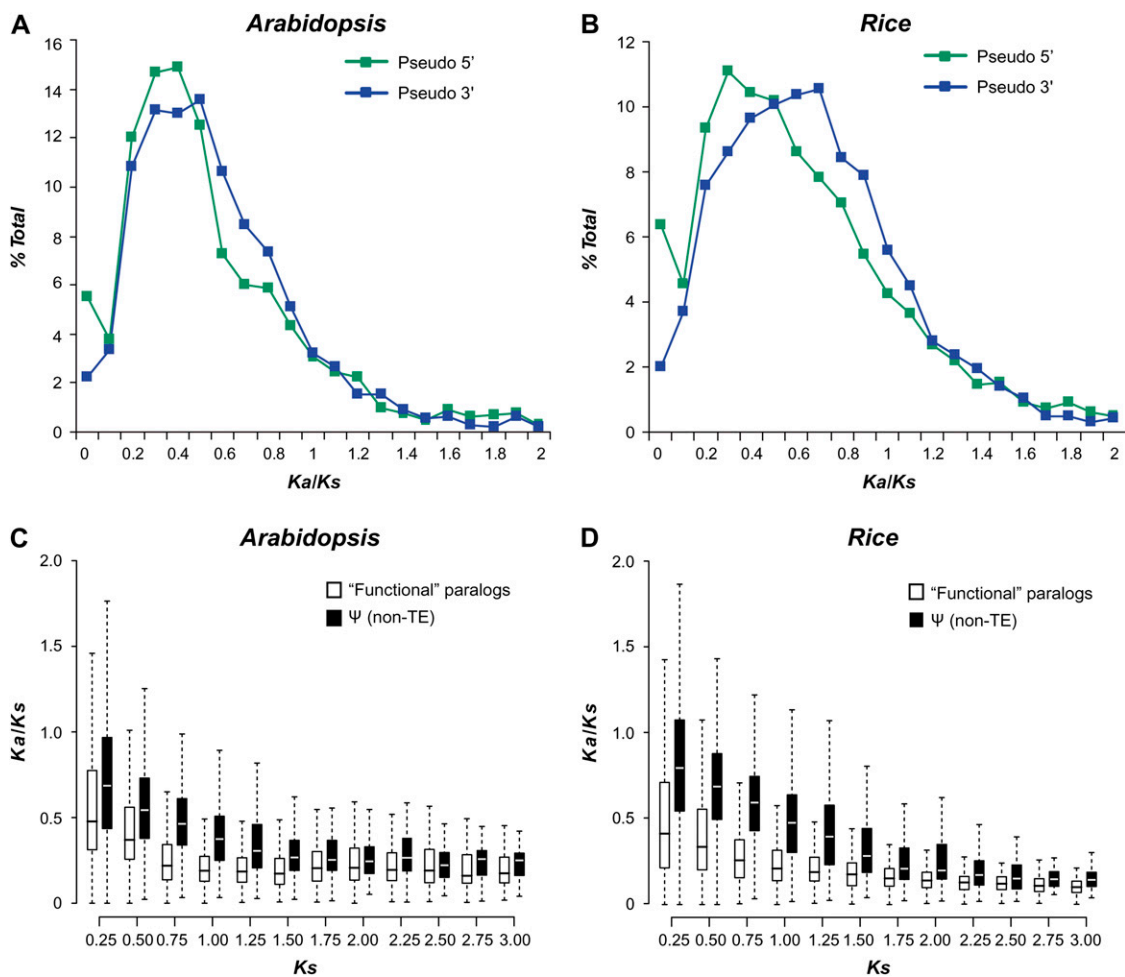
Figure 3. Differential selection on 5′ and 3′ regions of Ψs and time of duplication prior to pseudogenization. A and B, Frequency distributions of *Ka/Ks* on regions 5′ (green symbols) and 3′ (blue symbols) to the first stops in non-transposable-element Ψ-*FP* pairs in Arabidopsis (A) and rice (B). C and D, Relationship between *Ka/Ks* and *Ks* for *FP-FP* pairs (white boxes) and Ψ-*FP* pairs (black boxes) in Arabidopsis (C) and rice (D). Here, *FP*s and Ψs are as defined in Figure 2. *Ks* values were binned, and box plots are shown with the horizontal lines indicating median values, the lower and upper boundaries of the boxes indicating 25th and 75th percentiles, respectively, and the lower and upper boundaries of the dotted lines indicating 1st and 99th percentiles, respectively. non-TE, Non-transposable element. [See online article for color version of this figure.]

(Wilcoxon rank sum test, $P < 2e{-}16$; Fig. 4A). In rice, however, the intensity distributions of exons (sense direction) and Ψs are similar (Wilcoxon rank sum test, $P > 0.7$; Fig. 4B). This is inconsistent with the finding that there are more than 10 times more expression tags for annotated genes than for Ψs in rice. This does not appear to be due to cross-hybridization, since we have eliminated probes with potential to cross-hybridize (see "Materials and Methods"). We suspect this may be partly due to the overall lower hybridization intensity in the rice tiling array data sets (Li et al., 2006), but we do not have a definitive explanation.

Using the 95th percentile of the intron probe intensity level distribution as the threshold, we found that 610 Arabidopsis (16.79%) and 1,047 rice (22.91%) Ψs may be considered sense expressed (Table I), consistent with earlier studies (Yamada et al., 2003). We also

found that 523 Arabidopsis (14.42%) and 922 rice (20.17%) Ψs are likely expressed in the antisense direction (Table I). Another interesting finding is that the difference in the intensity distributions of Ψ probes between sense and antisense directions (Wilcoxon rank sum test, $P < 3e{-}4$) is not as significant as those for exons (Wilcoxon rank sum test, $P < 2.2e{-}16$; Fig. 4, A and B). Therefore, Ψ sense and antisense expression appear to be uncoupled, suggesting that Ψs may not be subjected to regulatory control at a level similar to functional genes.

Properties of Expressed Ψs

The finding that a significant number of plant Ψs have evidence of expression raises the question if these Ψs are selected at the RNA level. Based on the findings

Table I. *Expression feature representation among annotated genes and Ψs*

Expression Tag Type	No. of Annotated Genes with Tags[a]	Percentage of Annotated Genes with Tags[b]	No. of Ψs with Tags[c]	Percentage of Ψs with Tags
PUT				
Arabidopsis	19,460	71.9	162	2.1
Rice	20,442	49.8	135	2.8
MPSS				
Arabidopsis	19,764	73.1	233	4.9
Rice	20,091	48.9	200	2.6

[a]Annotated genes are Arabidopsis or rice genes in TAIR version 7 and TIGR version 5 annotations. A PUT was matched to an annotated gene if the PUT-gene pair had 97% or greater identity over 50% or more of the shorter sequence with length of 300 bp or greater. Only MPSS tags with a 100% unique match to genes were considered. [b]The numbers of annotated protein genes in Arabidopsis (TAIR version 7) and rice (TIGR version 5) are approximately 41,030 and 27,029, respectively (after excluding annotated Ψs and repetitive elements). [c]The numbers of set II Ψs in Arabidopsis (TAIR version 7) and rice (TIGR version 5) are 3,697 and 7,762, respectively.

presented below, it seems that in most cases Ψ expression is not stable over a long evolutionary time and therefore may not be subjected to purifying selection. We found that expressed Ψs tend to have significantly lower Ks values, indicating that they were derived from relatively recent duplication events (Wilcoxon rank sum tests: $P < 1.0e{-}11$ for Arabidopsis [Fig. 4C] and $P < 0.05$ for rice [Fig. 4D]). In addition, expressed Ψs tend to be more "complete." That is, the alignment coverage of Ψs to their functional paralogs is significantly higher for Ψs with expression tags than for all Ψs (Wilcoxon rank sum tests: Arabidopsis, $P < 7.0e{-}10$; rice, $P < 9.0e{-}9$; Supplemental Fig. S3). A possible explanation for the tendency for younger Ψs to be expressed is that their cis-regulatory regions are not completely degenerated, assuming that the completeness of the Ψ coding region reflects the completeness of the associated promoter region. Another potential source of expressed Ψs is from those derived from ancient duplication events that have experienced strong purifying selection but have become pseudogenized relatively recently. Unfortunately, the number of expressed Ψs derived from ancient duplication is too small to assess this possibility statistically.

Note that the above explanations assume that Ψ expression evolved neutrally and that Ks can be used as a proxy for evolutionary time. These assumptions would not hold if in fact some expressed Ψs were functional as RNA genes and purifying selection acted on the majority of the nucleotide positions. In these cases, Ψ Ks would be an underestimate of the true neutral evolution rate and our earlier suggestion that expressed Ψs tend to be young would be invalid. It has been shown that the transcript from a NOS Ψ is a natural antisense of a paralogous NOS gene and is implicated in the transcriptional regulation of the latter (for review, see Zheng and Gerstein, 2007). Nevertheless, the heteroduplex between NOS Ψ and its functional paralog only involves an approximately 140-bp region over the greater than 2-kb NOS Ψ. This is similar to the findings of an analysis of trans-natural antisense transcripts in 10 eukaryotes, where the me-

dian sizes of regions of significant sequence similarity between the antisense transcripts and their target genes are 45 to 170 bp (Li et al., 2008). Therefore, if transcriptional regulatory roles of Ψs on their paralogs are mechanistically similar to NOS regulation, we do not expect the sequences outside of the relatively much smaller antisense regions to experience strong purifying selection. Taken together, our Ks estimate is likely not substantially influenced by purifying selection of the potential antisense regions important for RNA gene functions, and our suggestion that expressed Ψs tend to be derived from young duplicates is likely correct.

The Ks distribution of expressed Ψs in plants (Fig. 4, C and D) is similar to that of mouse Ψs, where there is a much higher frequency of Ψ-FP pairs with low Ks and a long tail of pairs with high Ks values (Frith et al., 2006). In the mouse study, it was suggested that the long tail indicated that some expressed Ψs may have persisted in the genome over a long evolutionary time scale. We have also found a similar pattern among plant Ψs. However, it is also possible that some of these expressed Ψs are simply derived from older duplicates that have become Ψs recently. Further studies examining the timing of pseudogenization would be necessary to assess these possibilities. Given the enrichment of expressed Ψs that are derived from younger duplicates, it seems that in most cases Ψ expression does not persist.

Correlation between the Number of Ψs and Family Size

Most of the Ψs we identified are remnants of duplicates that were not retained, and their presence can provide important clues to the past history of gene family evolution. Studies of duplicate genes in Arabidopsis have established that gene families vary greatly in size and that there is a substantial bias in the functions of the retained duplicates (Maere et al., 2005; Hanada et al., 2008). What is the relationship between the functional and Ψ members of a subfamily? Are the numbers of Ψs and functional members in a gene
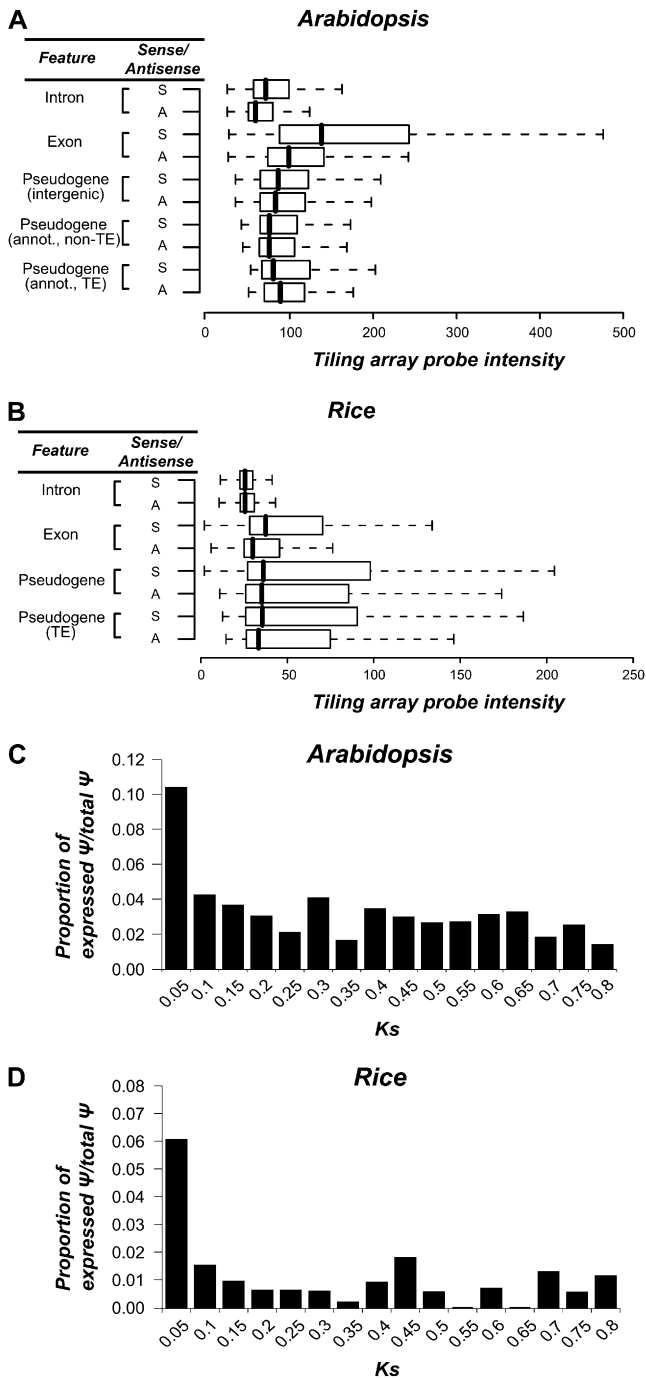
Figure 4. Expression of Ψs. A and B, Box plots of tiling array probe intensities of intron, exon, and Ψ features are shown for Arabidopsis (A) and rice (B). A, Probes complementary to the antisense strand; annot., annotated Ψs; S, probes complementary to the sense strand of features; TE, transposable elements. C and D, Proportion of the number of Ψs with expression tags (EST and/or MPSS) in different Ks bins in Arabidopsis (C) and rice (D).

family positively correlated, as would be expected assuming that there is no differential retention of duplicates?

To address this question, we first classified Arabidopsis and rice protein-coding genes into domain families based on Pfam domain designations. We then assigned Ψs according to the domain family designations of their top matching functional paralogs. Here, we assumed that the domain content of a Ψ and its functional paralog is identical. This is in general true, since few domain configurations have changed among functional duplicate pairs with $Ks \leq 2$ (Arabidopsis, 197 of 3,848 paralogous pairs; rice, 209 of 2,990). Based on these domain family assignments, we found that there are significant positive correlations between the numbers of Ψs and functional members in both Arabidopsis (Spearman rank, $\rho = 0.54$, $P < 2.2e$-16; Fig. 5A) and rice ($\rho = 0.56$, $P < 2.2e$-16; Fig. 5B). Our findings indicate that larger gene families tend to experience more gene loss than smaller gene families. Given that the most common fate of duplicate genes is pseudogenization (Lynch and Conery, 2000; Harrison et al., 2002), assuming that the gene loss rate is similar among families, larger gene families likely have proportionally higher numbers of duplicates at any given time. This may be the explanation for the observed correlation.

The same correlation has been observed in yeast (Lafontaine et al., 2004). However, in yeast the correlation is nearly perfect ($r^2 = 0.98$), which is in sharp contrast to our findings. Despite the highly significant correlation between numbers of Ψs and numbers of functional members in plant domain families, there is substantial variation that results in relatively moderate correlation coefficients (Fig. 5). Assuming a simple model of linear correlation between numbers of Ψs and family size, many gene families are either above the trend line, indicating excess loss, or below the trend line, indicating a higher than usual rate of retention. In addition to the correlation between numbers of Ψs and family size in each species, the numbers of Ψs per family in Arabidopsis and rice are also significantly correlated (Fig. 5C). This can be partly attributed to the fact that gene family sizes between species are correlated because of common ancestry and a similar degree of parallel retention (Hanada et al., 2008). This correlation also suggests that the proportion of family members that become pseudogenized after duplication is similar in rice and Arabidopsis.

Overrepresentation and Underrepresentation of Ψs among Domain Families

Although there is a significant correlation between numbers of Ψs and functional members in plant gene families, many gene families seem to have lost more or lost less than average (as indicated by deviation from the trend lines; Fig. 5). Which families have an overrepresented or underrepresented number of Ψs? In addition, do Ψs tend to be derived from genes with certain functions? To address these questions, we assigned Ψs to GeneOntology categories based on
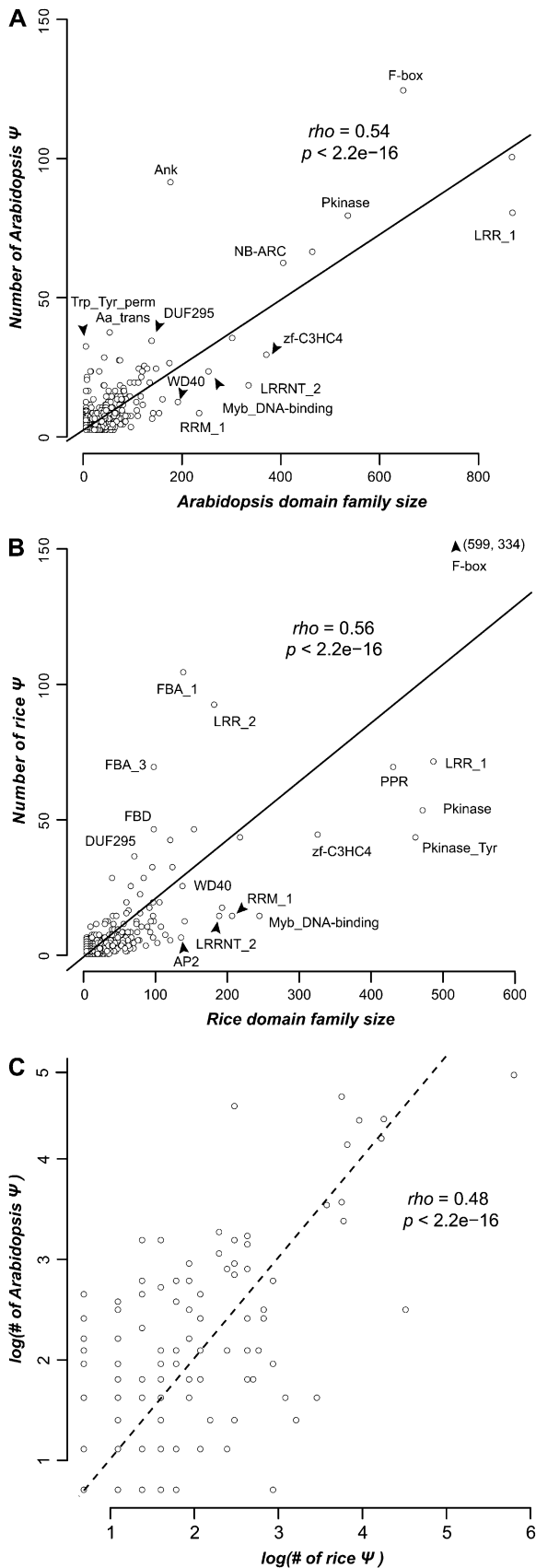
the annotations of each Ψ's best matching functional paralog to assess if a particular domain family has a significantly overrepresented or underrepresented number of Ψs (for original data and test statistics, see Supplemental Table S4). Among families with overrepresented numbers of Ψs, only 13 overlap between Arabidopsis and rice. In addition, only two families are underrepresented consistently between these two plants. Among Arabidopsis domain families with an overrepresented number of Ψs, four "functional classes" of domain families stand out (Fig. 6A): (1) defense genes: nucleotide-binding adaptor shared by APAF-1, certain R gene products, and CED-4 (NB-ARC), Toll/interleukin-1 receptor homology domain (TIR), leucine-rich repeat (LRR-3), S-domain in receptor-like kinases (PAN_2, S_locus_glycop, B_lectin); (2) cell wall-modifying enzymes: cellulose synthase, polygalacturonase (Glyco_hyrdo_28), pectin esterase; (3) enzymes involved in secondary metabolism: cytochrome P-450, terpene synthase; and (4) protein degradation machinery: F-box and various associated domains (FBD, FBA_1, FBA_3, and LRR_2), Seven in absentia (Sina), Meprin and TRAF-Homology domain (MATH), and U-box.

The excess of Ψs in these families is not simply due to the trivial explanation that larger families have more Ψs (this is accounted for in the statistical tests). Interestingly, genes harboring some of these domain families tend to be tandem duplicates. These domain families tend to experience repeated gene gain in one organismal lineage and undergo loss in another (single lineage expansion; Hanada et al., 2008). We found that there is a significantly higher ratio of Ψs that are in tandem compared with that of functional genes in both plants (Supplemental Table S5). In an earlier study, we also found that the genes derived from single-lineage expansion also tend to be responsive to environmental, particularly biotic, stresses (Hanada et al., 2008). We found that, as expected, genes in class 1 (as defined above; Fig. 6A) tend to be responsive to biotic stress (Fig. 6B; Supplemental Table S4). Our findings suggest that genes involved in defense responses, particularly those in tandem repeats, have high duplication rates but turn over rapidly, presumably due to selection imposed by rapid changes in biotic environments.

Figure 5. Relationship between the numbers of Ψs and annotated functional genes in Pfam domain families. A and B, The number of Ψs in each domain family is strongly and significantly correlated with the number of annotated genes, presumably non-Ψs, of the same family in Arabidopsis (A) and rice (B). The Spearman rank correlation coefficients (*rho*), their associated *P* values, and the linear model-fitted trend lines are shown. Selected domain families with significantly more or less Ψs than expected are labeled according to their Pfam identifiers. C, Relationship between domain family Ψ numbers in Arabidopsis and rice. The dotted line has a slope of 1 to illustrate the variation in Ψ numbers.

Figure 6. Ψ representation and stress responsiveness of domain families in Arabidopsis. A, Arabidopsis protein domain families with significantly overrepresented (shades of red) or underrepresented (shades of blue) numbers of Ψs. Ψ/NΨ tests determine if a domain family contains an underrepresented or overrepresented number of Ψs compared with non-Ψs (NΨ). The "Ψ/NΨ test" column indicates the significance of the test statistic (P value) in shades of blue and red (significant underrepresentation and overrepresentation, respectively). The "Funct. class" column indicates four groups of domain families with overrepresented numbers of Ψs: defense genes (class 1), cell wall-modifying enzymes (class 2), enzymes involved in secondary metabolism (class 3), and protein degradation machinery (class 4); and two groups with underrepresented numbers of Ψs: transcription factor-associated domains (class 5) and receptor-like kinases (class 6). B, Abiotic/biotic stress test. This test determines if the annotated members of a domain family tend to be responsive to a stress condition (see "Materials and Methods"). The color shading scheme is the same as that for A, with domain families containing genes that tend to be up-regulated under a particular stress condition compared to the genome average in shades of red and families with genes that tend not to be up-regulated in shades of blue.
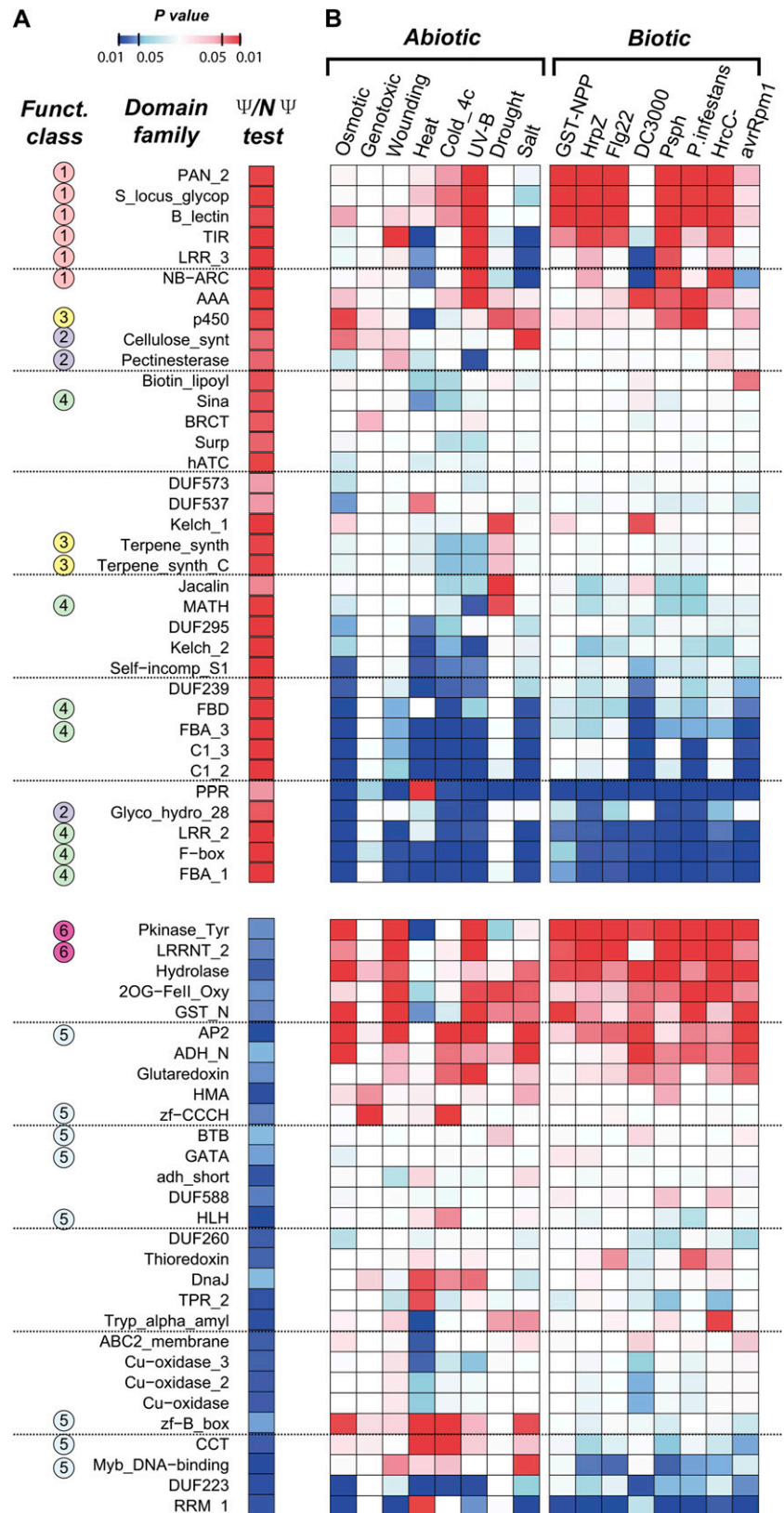
Aside from domain families with genes implicated in defense response, what factors explain elevated rates of pseudogenization? Most of the domain families in functional classes 2, 3, and 4 (Fig. 6A) do not have excess numbers of genes that are stress responsive (Fig. 6B). In class 2, the excess Ψs in the P450 and

terpene synthase families are likely indicative of rapid turnover due to selection for the production of a battery of secondary compounds serving as part of defense mechanisms and/or herbivore deterrents (Guengerich, 2004; Tholl, 2006). But it is puzzling why there are so many Ψs related to cell wall modification and protein degradation. The large family sizes among cell wall-modifying enzymes may be explained by selection for functional divergence for more flexible cell wall modification controls in myriad environments (Kim et al., 2006); however, it is still unclear why there is an overrepresentation of Ψs in these families. Similarly, it would appear that, given the complexity of cellular networks, a large complement of specificity determinants to accurately regulate protein degradation is essential. While this may explain why the F-box (Gagne et al., 2002), U-box (Mudgil et al., 2004), and MATH (Gingerich et al., 2007) domain families are much larger compared with most plant gene families, the selection pressures that contribute to the disproportionally larger numbers of Ψs in these families remain unclear.

There are two apparent functional classes among domain families with an underrepresented number of Ψs. We found a number of transcription factor-associated domains (class 5; Fig. 6A). We also found that there are significantly fewer transcription factor Ψs compared with all other Ψs in both Arabidopsis and rice (Supplemental Table S6). This is consistent with earlier studies showing that transcription factors tend to be retained after whole genome duplication in Arabidopsis (Blanc and Wolfe, 2004; Seoighe and Gehring, 2004) and that there is a higher rate of lineage-specific expansion of transcription factors in plants compared with other eukaryotes (Shiu et al., 2005). In addition to transcription factors, another domain class that behaves similarly in both Arabidopsis and rice is typically found in plant receptor-like kinases (RLKs; Shiu and Bleecker, 2001). The RLKs belong to the largest gene family in plants, with over 600 and 1,200 members in Arabidopsis and rice, respectively (Shiu et al., 2004). A number of RLKs with LRRs are important in controlling plant growth and development as well as in defense responses (Shiu and Bleecker, 2001; Morillo and Tax, 2006). The underrepresentation of RLK Ψs suggests that more RLK duplicates were retained than became Ψs. Aside from these examples, it is not entirely clear why some of the other domain families have underrepresented numbers of Ψs. It remains to be determined if this underrepresentation is due to a much reduced duplication rate or an elevated rate of duplicate retention.

CONCLUSION

In this study, we identified Ψs that likely were remnants of protein-coding genes in the genomes of Arabidopsis and rice. We found that a large number of plant Ψs seem to be subjected to a substantial period of purifying selection before pseudogenization. A number of the plant Ψs are likely expressed but at a lower level compared with functional genes. Our findings also indicate that Ψ expression is found mostly among young duplicates, suggesting that many of the expressed Ψs may not be under selection as noncoding RNA. In addition, we found that the number of Ψs in a domain family is significantly correlated with the number of presumably functional members in the family. This correlation illustrates that pseudogenization at the gene family level is to some extent a neutral process. However, the correlation is far from perfect, because many families either have significantly more or significantly less than the expected numbers of Ψs, consistent with the substantial functional bias of retained duplicates in mammals (Shiu et al., 2006) and plants (Blanc and Wolfe, 2004; Seoighe and Gehring, 2004). Genes in families with underrepresented numbers of Ψs likely have a higher than usual retention rate of their family members (Shiu et al., 2004, 2005). On the other hand, genes in families with overrepresented numbers of Ψs may have experienced rapid birth-and-death evolution (Nei and Rooney, 2005). In several cases, for example the F-box family, rapid turnover is observed but the selection pressure driving the rapid birth and death is not clear. Our study represents, to our knowledge, the first comprehensive study of plant Ψ evolution and expression. The findings here highlight multiple properties of plant Ψs that are important for our understanding of genic evolution in plants and for guiding future annotation efforts of plant genomes.

MATERIALS AND METHODS

Identification of Ψs in the Rice and Arabidopsis Genomes

The intergenic sequences used for identifying putative Ψs were defined according to The Institute of Genome Research (TAIR) version 7 annotations for Arabidopsis (*Arabidopsis thaliana*) and Rice Annotation Project 2 (RAP2) and The Institute of Genome Research (TIGR; currently the J. Craig Venter Institute) version 5 annotation for rice (*Oryza sativa*). The overall pipeline for Ψ identification is outlined in Figure 1 and is generally based on the pseudopipe workflow (Zhang et al., 2006) with modifications. The pipeline consists of six major steps: (1) identifying intergenic regions with sequence similarity to known proteins; (2) repeat masking; (3) linking Ψ fragments (pseudoexons) into contigs (set I Ψ); (4) quality filtering (set II Ψ); (5) identifying features that disrupt contiguous protein sequences (set III Ψ); and (6) distinguishing retro and nonretro Ψ (for details, see Supplemental Methods S1). The intergenic regions that qualify for the first three steps are referred to as set I Ψ. Those passing through the first four steps are referred to as set II Ψ. Ψs with disabling mutations identified during step 5 are referred to as set III Ψ. The coordinates and the pseudocoding sequences of these Ψs are included in Supplemental Tables S1 to S3.

Analysis of Expression

For each plant species, we determined the number of annotated genes that are presumably functional and the number of Ψs that have evidence of expression based on (1) putative unique transcript (PUT), (2) MPSS tags, and (3) tiling array data. Rice and Arabidopsis PUTs were downloaded from PlantGDB (version 163a; Duvick et al., 2008) and were used to search against annotated genes and set II Ψs. PUTs are regarded as cognate transcripts for

Zou et al.

annotated genes and Ψs if (1) these PUTs do not have a better match to other genes, (2) their identities are 97% or greater, (3) the aligned regions are 300 nucleotides or greater, and (4) the matched region is greater than 50% of the shorter sequence length. The MPSS tags for rice and Arabidopsis were downloaded (http://mpss.udel.edu/; Nobuta et al., 2007) and mapped to the rice and Arabidopsis pseudomolecules (100% identity and 100% coverage). MPSS tags that mapped uniquely to functional genes or Ψs were regarded as expression tags for the respective genic sequences.

The third type of expression data set we examined was tiling array data for Arabidopsis (GEO: GSE601;Yamada et al., 2003) and for rice (GEO: GSE6996; Li et al., 2007). In both studies, the genomes were covered with the use of multiple arrays. A between-array normalization procedure was applied to each data set using the affyPLM package of Bioconductor (Gentleman et al., 2004). Among probe sequences with perfect matches to the genome assemblies, probes with one or more match with 85% or greater identity were discarded due to their potential contribution to cross-hybridization. Intensities for probes enclosed within the regions defined by the coordinates of exons and introns in the annotations as well as Ψs identified in this study were averaged to serve as summary statistics of expression level.

Gene expression data under eight abiotic and eight biotic stress conditions were obtained from AtGenExpress (http://www.uni-tuebingen.de/plantphys/AFGN/atgenex.htm) and processed as described previously (Hanada et al., 2008). Significantly up- and down-regulated genes under each condition were identified using LIMMA (Wettenhall and Smyth, 2004). Overrepresentation and underrepresentation of each domain family (D) member under each condition (C) were determined by setting up a two-by-two contingency table comparing numbers of D and non-D members that are up- or down-regulated in C and numbers of D and non-D members without significant expression change under C using Fisher's exact test.

Evolutionary Rate Calculation

To evaluate the level of selective constraint on Ψs, we calculated the synonymous and nonsynonymous substitution rates (Ks and Ka, respectively) between each Ψ and its closest paralog in Arabidopsis and rice. The rates were determined using the yn00 program in the PAML program package (Yang, 1997) based on pairwise alignments of the sequence pairs. Very few pairs had run errors (e.g. NAN in PAML output) and were excluded. Sequence pairs that were too similar ($Ks \leq 0.005$) or too divergent ($Ks > 3$) were excluded as well. Note that these filtered Ψs were used for comparing the strength of selection between 5′ segments before the first disabling mutation and 3′ segments after the last disabling mutation. However, we did not apply the Ks cutoffs on the segments. The rates for within-species reciprocal best match gene pairs (FP pairs) were determined as mentioned above.

Classification of Annotated Genes and Ψs into Domain Families and Tandem/Nontandem Classes

Annotated genes in Arabidopsis and rice were first classified into families based on their Pfam domain composition. First, the protein sequences from these two plants were compared against hidden Markov models using the HMMER program package (http://hmmer.wustl.edu/) with the trusted cutoff threshold scores defined by Pfam (build May 2007; Bateman et al., 2004). For genes with multiple annotations due to alternatively spliced forms, the longest annotated protein sequences were used for domain identification. Because in most cases Ψs are truncated and/or degenerate, they were assigned to domain families based on the assumption that their domain compositions were the same as their closest paralogs identified in step 4 of the Ψ identification pipeline. Tandemly duplicated genes were defined as genes in any gene pair, T1 and T2, that (1) belong to the same gene family, (2) are located within an average distance of 10 genes, and (3) are separated by 10 or fewer nonhomologous spacer genes.

Supplemental Data

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

Supplemental Figure S1. Frequency distributions of ω values (Ka/Ks) of rice annotated genes, retro-Ψs, and nonretro-Ψs.

Supplemental Figure S2. Frequency distributions of ω values (Ka/Ks) of rice annotated genes, all Ψs, and Ψs with premature stops and/or frameshifts.

Supplemental Figure S3. Proportion of Ψ sequences covered by their closest functional paralogs in Arabidopsis and rice.

Supplemental Figure S4. Strength of purifying selection on regions 5′ and 3′ to the first stops in non-transposable-element Ψ-FP pairs and FP-FP pairs.

Supplemental Table S1. Ψ annotation data.

Supplemental Table S2. Arabidopsis Ψ coding sequences.

Supplemental Table S3. Rice Ψ coding sequences.

Supplemental Table S4. GeneOntology categories with consistently overrepresented or underrepresented numbers of Ψs in Arabidopsis and rice.

Supplemental Table S5. Overrepresentation of tandem Ψs.

Supplemental Table S6. Underrepresentation of transcription factor Ψs.

Supplemental Methods S1.

LITERATURE CITED

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

Balakirev ES, Ayala FJ (2003) Pseudogenes: are they "junk" or functional DNA? Annu Rev Genet **37:** 123–151

Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, et al (2004) The Pfam protein families database. Nucleic Acids Res **32:** D138–D141

Benovoy D, Drouin G (2006) Processed pseudogenes, processed genes, and spontaneous mutations in the Arabidopsis genome. J Mol Evol **62:** 511–522

Blanc G, Hokamp K, Wolfe KH (2003) A recent polyploidy superimposed on older large-scale duplications in the Arabidopsis genome. Genome Res **13:** 137–144

Blanc G, Wolfe KH (2004) Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. Plant Cell **16:** 1679–1691

Brosius J (1991) Retroposons: seeds of evolution. Science **251:** 753

Chang YF, Imam JS, Wilkinson MF (2007) The nonsense-mediated decay RNA surveillance pathway. Annu Rev Biochem **76:** 51–74

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

Frith MC, Wilming LG, Forrest A, Kawaji H, Tan SL, Wahlestedt C, Bajic VB, Kai C, Kawai J, Carninci P, et al (2006) Pseudo-messenger RNA: phantoms of the transcriptome. PLoS Genet **2:** e23

Gagne JM, Downes BP, Shiu SH, Durski AM, Vierstra RD (2002) The F-box subunit of the SCF E3 complex is encoded by a diverse superfamily of genes in Arabidopsis. Proc Natl Acad Sci USA **99:** 11519–11524

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, et al (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol **5:** R80

Gingerich DJ, Hanada K, Shiu SH, Vierstra RD (2007) Large-scale, lineage-specific expansion of a bric-a-brac/tramtrack/broad complex ubiquitin-ligase gene family in rice. Plant Cell **19:** 2329–2348

Gray TA, Wilson A, Fortin PJ, Nicholls RD (2006) The putatively functional Mkrn1-p1 pseudogene is neither expressed nor imprinted, nor does it regulate its source gene in trans. Proc Natl Acad Sci USA **103:** 12039–12044

Guengerich FP (2004) Cytochrome P450: what have we learned and what are the future issues? Drug Metab Rev **36:** 159–197

Hanada K, Zou C, Lehti-Shiu MD, Shinozaki K, Shiu SH (2008) Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli. Plant Physiol 148: 993–1003

Harrison PM, Hegyi H, Balasubramanian S, Luscombe NM, Bertone P, Echols N, Johnson T, Gerstein M (2002) Molecular fossils in the human genome: identification and analysis of the pseudogenes in chromosomes 21 and 22. Genome Res 12: 272–280

Harrison PM, Zheng D, Zhang Z, Carriero N, Gerstein M (2005) Transcribed processed pseudogenes in the human genome: an intermediate form of expressed retrosequence lacking protein-coding ability. Nucleic Acids Res 33: 2374–2383

Hirotsune S, Yoshida N, Chen A, Garrett L, Sugiyama F, Takahashi S, Yagami K, Wynshaw-Boris A, Yoshiki A (2003) An expressed pseudogene regulates the messenger-RNA stability of its homologous coding gene. Nature 423: 91–96

Hughes AL (1994) The evolution of functionally novel proteins after gene duplication. Proc R Soc Lond B Biol Sci 256: 119–124

International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. Nature 436: 793–800

Jacq C, Miller JR, Brownlee GG (1977) A pseudogene structure in 5S DNA of Xenopus laevis. Cell 12: 109–120

Kazazian HH Jr (2004) Mobile elements: drivers of genome evolution. Science 303: 1626–1632

Kim J, Shiu SH, Thoma S, Li WH, Patterson SE (2006) Patterns of expansion and expression divergence in the plant polygalacturonase gene family. Genome Biol 7: R87

Korneev SA, Park JH, O'Shea M (1999) Neuronal expression of neural nitric oxide synthase (nNOS) protein is suppressed by an antisense RNA transcribed from an NOS pseudogene. J Neurosci 19: 7711–7720

Lafontaine I, Fischer G, Talla E, Dujon B (2004) Gene relics in the genome of the yeast Saccharomyces cerevisiae. Gene 335: 1–17

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al for the International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. Nature 409: 860–921

Li JT, Zhang Y, Kong L, Liu QR, Wei L (2008) Trans-natural antisense transcripts including noncoding RNAs in 10 species: implications for expression regulation. Nucleic Acids Res 36: 4833–4844

Li L, Wang X, Sasidharan R, Stolc V, Deng W, He H, Korbel J, Chen X, Tongprasit W, Ronald P, et al (2007) Global identification and characterization of transcriptionally active regions in the rice genome. PLoS One 2: e294

Li L, Wang X, Stolc V, Li X, Zhang D, Su N, Tongprasit W, Li S, Cheng Z, Wang J, et al (2006) Genome-wide transcription analyses in rice using tiling microarrays. Nat Genet 38: 124–129

Li WH (1983) Evolution of duplicate genes and pseudogenes. In M Nei, RK Koehn, eds, Evolution of Genes and Proteins. Sinauer Associates, Sunderland, MA, pp 14–37

Li WH, Gojobori T, Nei M (1981) Pseudogenes as a paradigm of neutral evolution. Nature 292: 237–239

Little PF (1982) Globin pseudogenes. Cell 28: 683–684

Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science 290: 1151–1155

Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, Kuiper M, Van de Peer Y (2005) Modeling gene and genome duplications in eukaryotes. Proc Natl Acad Sci USA 102: 5454–5459

McCarrey JR, Riggs AD (1986) Determinator-inhibitor pairs as a mechanism for threshold setting in development: a possible function for pseudogenes. Proc Natl Acad Sci USA 83: 679–683

Morillo SA, Tax FE (2006) Functional analysis of receptor-like kinases in monocots and dicots. Curr Opin Plant Biol 9: 460–469

Mounsey A, Bauer P, Hope IA (2002) Evidence suggesting that a fifth of annotated Caenorhabditis elegans genes may be pseudogenes. Genome Res 12: 770–775

Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420: 520–562

Mudgil Y, Shiu SH, Stone SL, Salt JN, Goring DR (2004) A large complement of the predicted Arabidopsis ARM repeat proteins are members of the U-box E3 ubiquitin ligase family. Plant Physiol 134: 59–66

Nei M, Rooney AP (2005) Concerted and birth-and-death evolution of multigene families. Annu Rev Genet 39: 121–152

Nobuta K, Venu RC, Lu C, Beló A, Vemaraju K, Kulkarni K, Wang W, Pillay M, Green PJ, Wang GL, et al (2007) An expression atlas of rice mRNAs and small RNAs. Nat Biotechnol 25: 473–477

Podlaha O, Zhang J (2004) Nonneutral evolution of the transcribed pseudogene Makorin1-p1 in mice. Mol Biol Evol 21: 2202–2209

Sakai H, Koyanagi KO, Imanishi T, Itoh T, Gojobori T (2007) Frequent emergence and functional resurrection of processed pseudogenes in the human and mouse genomes. Gene 389: 196–203

Seoighe C, Gehring C (2004) Genome duplication led to highly selective expansion of the Arabidopsis thaliana proteome. Trends Genet 20: 461–464

Shiu SH, Bleecker AB (2001) Receptor-like kinases from Arabidopsis form a monophyletic gene family related to animal receptor kinases. Proc Natl Acad Sci USA 98: 10763–10768

Shiu SH, Byrnes JK, Pan R, Zhang P, Li WH (2006) Role of positive selection in the retention of duplicate genes in mammalian genomes. Proc Natl Acad Sci USA 103: 2232–2236

Shiu SH, Karlowski WM, Pan R, Tzeng YH, Mayer KF, Li WH (2004) Comparative analysis of the receptor-like kinase family in Arabidopsis and rice. Plant Cell 16: 1220–1234

Shiu SH, Shih MC, Li WH (2005) Transcription factor families have much higher expansion rates in plants than in animals. Plant Physiol 139: 18–26

Thibaud-Nissen F, Ouyang S, Buell CR (2009) Identification and characterization of pseudogenes in the rice gene complement. BMC Genomics 10: 317

Tholl D (2006) Terpene synthases and the regulation, diversity and biological roles of terpene metabolism. Curr Opin Plant Biol 9: 297–304

Torrents D, Suyama M, Zdobnov E, Bork P (2003) A genome-wide survey of human pseudogenes. Genome Res 13: 2559–2567

Vanin EF (1985) Processed pseudogenes: characteristics and evolution. Annu Rev Genet 19: 253–272

Wang W, Zheng H, Fan C, Li J, Shi J, Cai Z, Zhang G, Liu D, Zhang J, Vang S, et al (2006) High rate of chimeric gene origination by retroposition in plant genomes. Plant Cell 18: 1791–1802

Wettenhall JM, Smyth GK (2004) limmaGUI: a graphical user interface for linear modeling of microarray data. Bioinformatics 20: 3705–3706

Yamada K, Lim J, Dale JM, Chen H, Shinn P, Palm CJ, Southwick AM, Wu HC, Kim C, Nguyen M, et al (2003) Empirical analysis of transcriptional activity in the Arabidopsis genome. Science 302: 842–846

Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13: 555–556

Yano Y, Saito R, Yoshida N, Yoshiki A, Wynshaw-Boris A, Tomita M, Hirotsune S (2004) A new role for expressed pseudogenes as ncRNA: regulation of mRNA stability of its homologous coding gene. J Mol Med 82: 414–422

Yu J, Wang J, Lin W, Li S, Li H, Zhou J, Ni P, Dong W, Hu S, Zeng C, et al (2005) The genomes of Oryza sativa: a history of duplications. PLoS Biol 3: e38

Zhang XH, Chasin LA (2004) Computational definition of sequence motifs governing constitutive exon splicing. Genes Dev 18: 1241–1250

Zhang Y, Wu Y, Liu Y, Han B (2005) Computational identification of 69 retroposons in Arabidopsis. Plant Physiol 138: 935–948

Zhang Z, Carriero N, Gerstein M (2004) Comparative analysis of processed pseudogenes in the mouse and human genomes. Trends Genet 20: 62–67

Zhang Z, Carriero N, Zheng D, Karro J, Harrison PM, Gerstein M (2006) PseudoPipe: an automated pseudogene identification pipeline. Bioinformatics 22: 1437–1439

Zhang Z, Harrison PM, Liu Y, Gerstein M (2003) Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. Genome Res 13: 2541–2558

Zheng D, Frankish A, Baertsch R, Kapranov P, Reymond A, Choo SW, Lu Y, Denoeud F, Antonarakis SE, Snyder M, et al (2007) Pseudogenes in the ENCODE regions: consensus annotation, analysis of transcription, and evolution. Genome Res 17: 839–851

Zheng D, Gerstein MB (2007) The ambiguous boundary between genes and pseudogenes: the dead rise up, or do they? Trends Genet 23: 219–224

Zheng D, Zhang Z, Harrison PM, Karro J, Carriero N, Gerstein M (2005) Integrated pseudogene annotation for human chromosome 22: evidence for transcription. J Mol Biol 349: 27–45
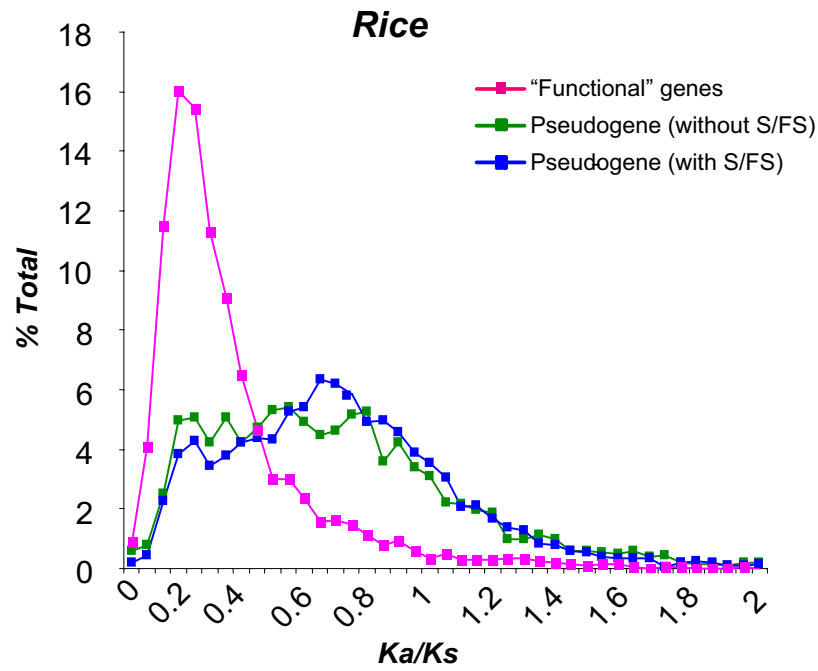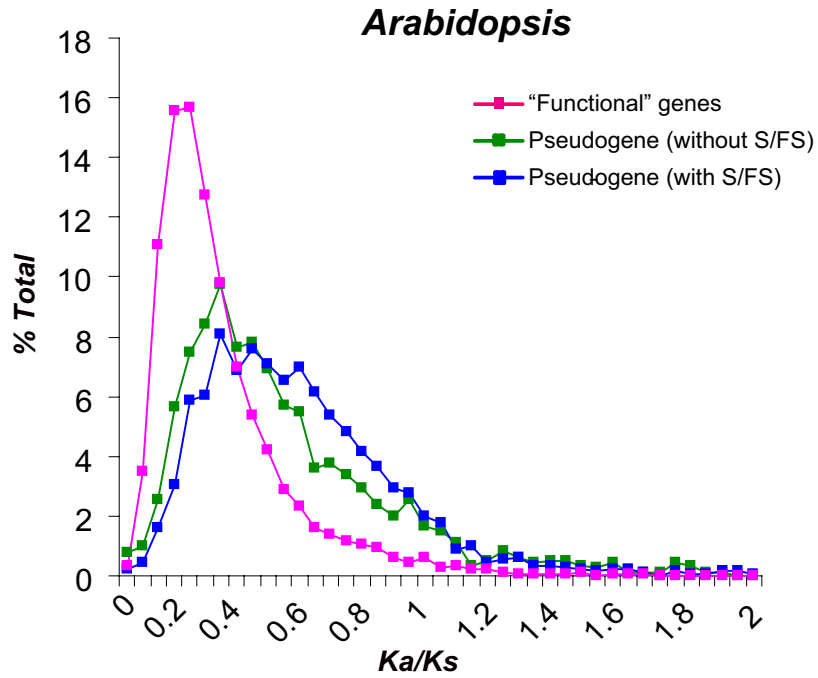
Supplemental Figure S1

Frequency distributions of ω values (Ka/Ks) of annotated genes (purple), retropseudogenes (green), and non-retropseudogenes (blue) in rice.
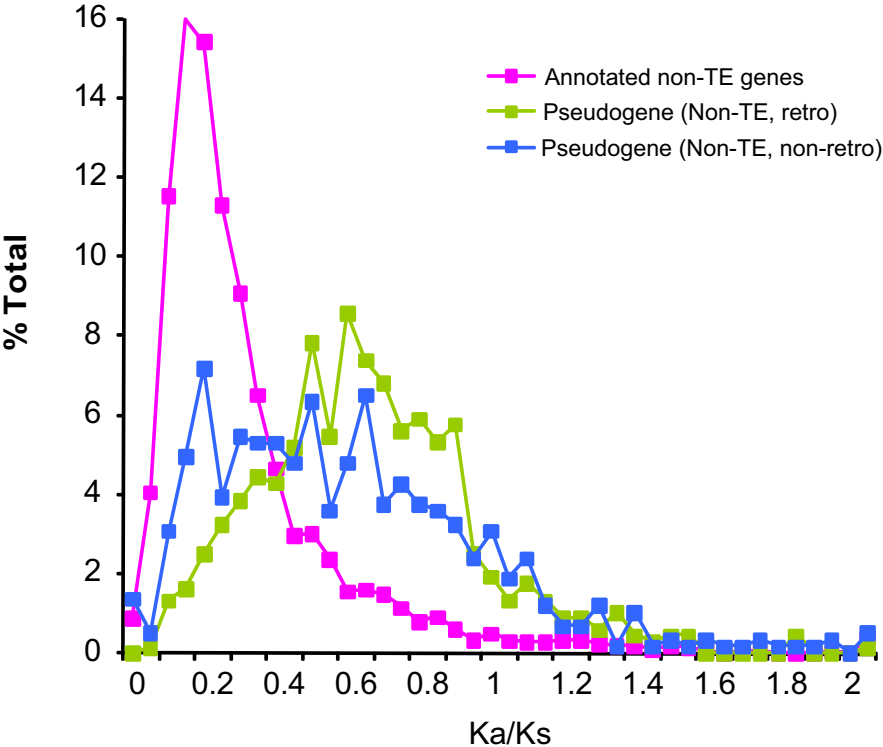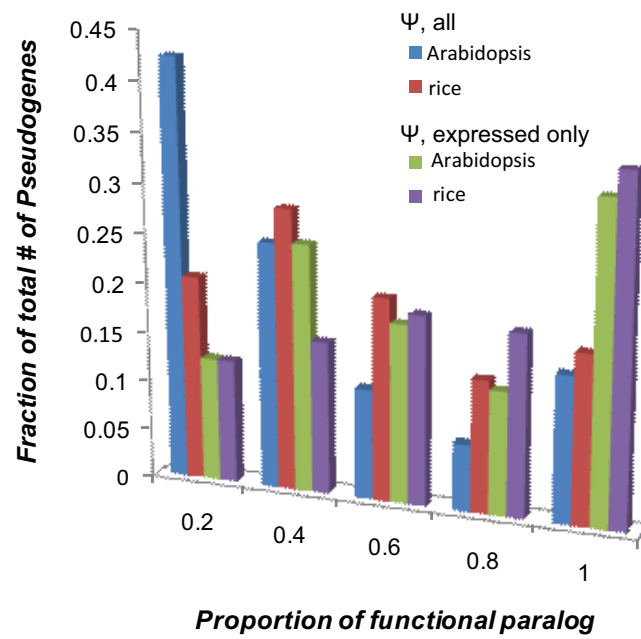
Supplemental Figure S2

Frequency distributions of ω values (Ka/Ks) of rice annotated genes (purple), all pseudogenes (green), and pseudogenes with premature stops (S) and/or frameshifts (FS) (blue).

Supplemental Figure S3

Proportion of pseudogene sequences covered by their closest functional paralogs in *Arabidopsis* and rice. All: all pseudogenes in each species in *A.thaliana* (blue) and rice (red). Expressed only: only pseudogenes with expression tags in Arabidopsis (green) and rice (purple).

Supplemental Figure S4

Strengths of purifying selection (Ka/Ks) on regions 5' and 3' to the first stops in non-TE Ψ-FP pairs and FP-FP pairs. Frequency distributions of Ka/Ks values of **(A)** A. thaliana Ψ-FP pairs, **(C)** A. thaliana FP-FP pairs, **(B)** Rice Ψ-FP pairs, and **(D)** rice FP-FP pairs.
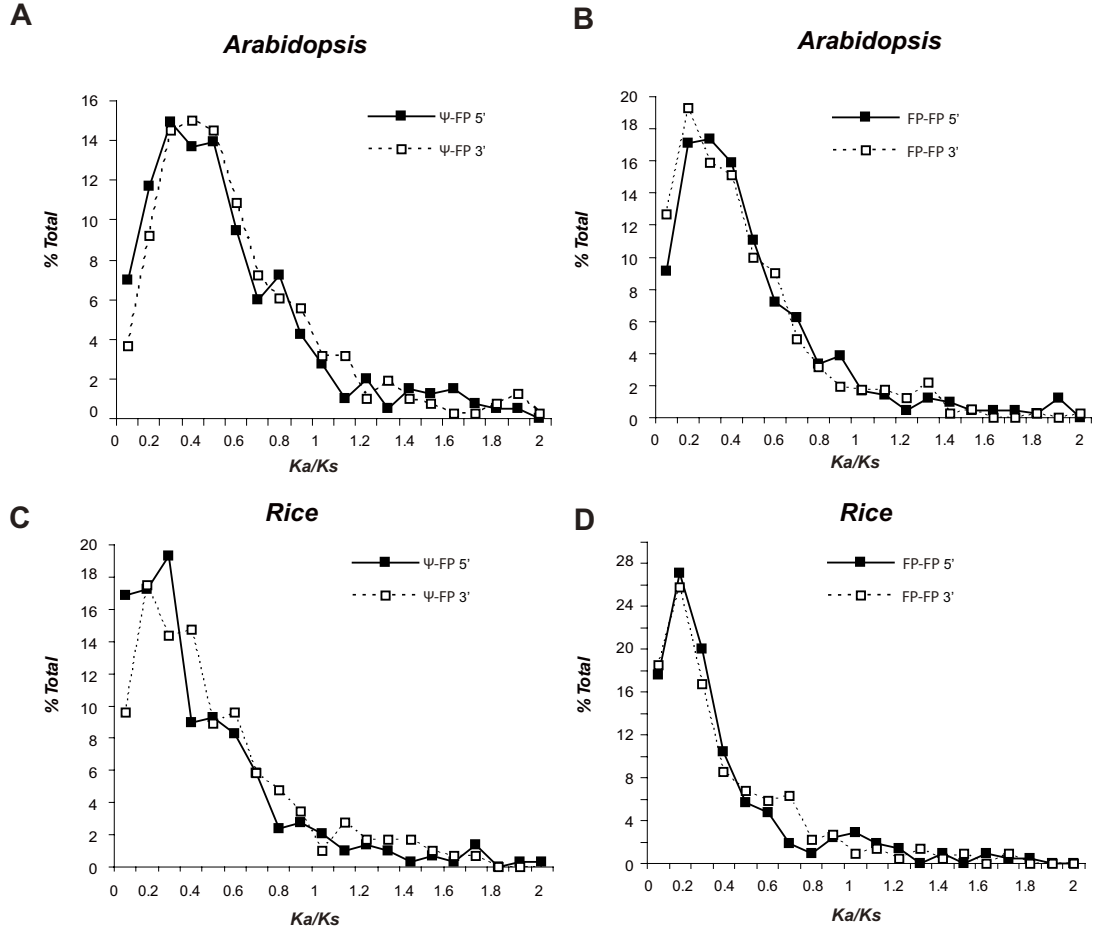
Supplemental Table S4. GeneOntology categories with consistently over- and under-represented numbers of Ψs in A. thaliana and rice

GeneOntology	Description	*A. thaliana*[1]	Rice[1]
Biological process			
GO:0016106	sesquiterpenoid biosynthetic process	+	+
GO:0006952	defense response	+	+
GO:0007165	signal transduction	-	-
GO:0006869	lipid transport	-	-
GO:0006886	intracellular protein transport	-	-
GO:0006950	response to stress	-	-
Cellular component			
GO:0005834	heterotrimeric G-protein complex	-	-
GO:0031225	anchored to membrane	-	-
GO:0016021	integral to membrane	-	-
GO:0009570	chloroplast stroma	-	-
GO:0009543	chloroplast thylakoid lumen	-	-
Molecular function			
GO:0003994	aconitate hydratase activity	+	+
GO:0005529	sugar binding	+	+
GO:0046933	hydrogen ion transporting ATP synthase activity, rotational mechanism	+	+
GO:0008289	lipid binding	-	-
GO:0046872	metal ion binding	-	-
GO:0004871	signal transducer activity	-	-
GO:0016301	kinase activity	-	-
GO:0003755	peptidyl-prolyl cis-trans isomerase activity	-	-
GO:0005199	structural constituent of cell wall	-	-
GO:0005198	structural molecule activity	-	-
GO:0004364	glutathione transferase activity	-	-
GO:0030508	thiol-disulfide exchange intermediate activity	-	-

[1]A "+" indicates that a GeneOntology category has an over-represented number of pseudogenes compared to GeneOntology categories in general. A "-" indicates that a GeneOntology category has an under-represented number of pseudogenes compared to GeneOntology categories in general. Significant over- and under-representation was determined by Fisher Exact test.

Supplemental Table S5. Over-representation of tandem Ψs

	Annotated, functional genes[1]		Ψs[1]		Significance[2]
	Tandem	Non-tandem	Tandem	Non-tandem	
A. thaliana	4428	11301	1023	953	*p* < 1e-92
Rice	6321	13458	844	1243	*p* < 1e-14

1: Genes with ≥ 1Pfam domain, excluding those containing transposable element-associated domains. Mitochondrial and chloroplast genes were also excluded.

2: Significance was evaluated by Fisher exact test.

Supplemental Table S6. Under-representation of transcription factor Ψs

	Annotated, functional genes[1]		Ψs[1]		Significance[2]
	Transcription factors	**Other genes**	**Transcription factors**	**Other genes**	
A. thaliana	1824	14075	152	1654	$p < 1e\text{-}4$
Rice	1924	17915	163	1864	$p < 0.02$

1: Genes with ≥ 1Pfam domain, excluding those containing transposable element-associated domains.

2: Significance was evaluated by Fisher exact test.