Genome-Wide Analysis of PHOSPHOLIPID:
DIACYLGLYCEROL ACYLTRANSFERASE (PDAT)
Genes in Plants Reveals the Eudicot-Wide PDAT Gene Expansion and Altered Selective Pressures Acting on the Core Eudicot PDAT Paralogs

Xue Pan, Fred Y. Peng, and Randall J. Weselake*

Agricultural Lipid Biotechnology Program, Department of Agricultural, Food, and Nutritional Science, University of Alberta, Edmonton, Alberta, Canada T6G 2P5

PHOSPHOLIPID:DIACYLGLYCEROL ACYLTRANSFERASE (PDAT) is an enzyme that catalyzes the transfer of a fatty acyl moiety from the sn-2 position of a phospholipid to the sn-3 position of sn-1,2-diacylglycerol, thus forming triacylglycerol and a lysophospholipid. Although the importance of PDAT in triacylglycerol biosynthesis has been illustrated in some previous studies, the evolutionary relationship of plant PDATs has not been studied in detail. In this study, we investigated the evolutionary relationship of the PDAT gene family across the green plants using a comparative phylogenetic framework. We found that the PDAT candidate genes are present in all examined green plants, including algae, lowland plants (a moss and a lycophyte), monocots, and eudicots. Phylogenetic analysis revealed the evolutionary division of the PDAT gene family into seven major clades. The separation is supported by the conservation and variation in the gene structure, protein properties, motif patterns, and/or selection constraints. We further demonstrated that there is a eudicot-wide PDAT gene expansion, which appears to have been mainly caused by the eudicot-shared ancient gene duplication and subsequent species-specific segmental duplications. In addition, selection pressure analyses showed that different selection constraints have acted on three core eudicot clades, which might enable paleoduplicated PDAT paralogs to either become nonfunctionalized or develop divergent expression patterns during evolution. Overall, our study provides important insights into the evolution of the plant PDAT gene family and explores the evolutionary mechanism underlying the functional diversification among the core eudicot PDAT paralogs.

Plant oils and TAGs are important agricultural commodities. Triacylglycerols (TAGs), glycerol esters of fatty acids, are the major components of plant oils in nearly all commercially important crops. In the last decade, the TAG biosynthetic pathway has received substantial attention from the research community. In 1960, using chicken (Gallus gallus) liver as an enzyme source, Weiss et al. (1960) demonstrated the use of diacylglycerol (DAG) and fatty acyl-CoAs as substrates for TAG biosynthesis. The enzyme catalyzing this reaction is named ACYL-COA:DIACYLGLYCEROL ACYLTRANSFERASE (DGAT; EC 2.3.1.158). The elucidation of the DGAT reaction laid the groundwork for numerous studies on the role of DGAT in TAG biosynthesis. The molecular identity of DGAT, however, remained elusive until 1998, when Cases et al. (1998) isolated the first DGAT gene in mouse (Mus musculus) by homology searches of the EST databases using the coding sequences from ACYL-COA:DIACYLGLYCEROL ACYLTRANSFERASE (EC 2.3.1.26), an enzyme that catalyzes the synthesis of choleseryl esters. At least six types of DGAT genes (DGAT1, DGAT2, DGAT3, DEFECTIVE IN CUTICULAR RIDGES, DIACYLGLYCEROL ACETYLTRANSFERASE, and a bifunctional DGAT/wax synthase) have been identified in plants (Hobbs et al., 1999; Lardizabal et al., 2001; Saha et al., 2006; Li et al., 2008; Durrett et al., 2010; Rani et al., 2010). The biochemistry, molecular biology, evolutionary relationship, structure-function relationship, and technological applications of DGAT have been discussed in some recent publications (Cao, 2011; Turchetto-Zolet et al., 2011; Liu et al., 2012).

Recently, it has become obvious that TAG synthesis also can be catalyzed by PHOSPHOLIPID:DIACYLGLYCEROL ACYLTRANSFERASE (PDAT; EC 2.3.1.158). PDAT enzyme activity was first described in microsomal preparations from sunflower (Helianthus annuus), castor bean (Ricinus communis), and Crepis palaestina by...
Pan et al.

Dahlqvist et al. (2000) during their observation of the use of phospholipids as acyl donor and DAG as acceptor for TAG biosynthesis. They further found that PDAT activity is also present in yeast (Saccharomyces cerevisiae) and identified the first PDAT gene (YNR008w) from yeast as a homolog of human LECITHIN:CHO-LESTEROL ACYLTRANSFERASE (LCAT; EC 2.3.1.43). LCAT is a soluble acyltransferase that catalyzes cholesteryl ester synthesis in blood plasma. Knowledge of the yeast PDAT sequence led to the discovery of two PDAT orthologs in Arabidopsis (Arabidopsis thaliana), referred to as AthPDAT1 (At5g13640) and AthPDAT2 (At3g44830; Ståhl et al., 2004). Although the RNA interference-based approach provides evidence that AthPDAT1 and AthDGAT1 have an overlapping function for TAG biosynthesis in both seed and pollen (Zhang et al., 2009), overexpression or knockout of AthPDAT1 in Arabidopsis only led to significant changes in oil phenotype (oil content and fatty acid composition) in developing leaves (Fan et al., 2013) but not in seeds (Mhaske et al., 2005). The other ortholog, AthPDAT2, has no role in TAG biosynthesis, even though the gene encoding PDAT2 is highly expressed in seeds. In castor bean, three PDAT orthologs have been identified (Kim et al., 2011; van Erp et al., 2011). One particular PDAT, RcoPDAT1A, appears to be ricinoleic acid specific seed-specific overexpression of this PDAT in Arabidopsis resulted in an enhanced proportion of hydroxy fatty acids in the seed oil. Recently, we found that flax (Linum usitatissimum) contains six PDATs (Pan et al., 2013). Four out of the six PDATs (LusPDAT1/LusPDAT3 and LusPDAT2/LusPDAT4) have the unique ability to preferentially channel α-linolenic acid into TAG, whereas another two PDATs (LusPDAT3/LusPDAT6) do not show TAG-synthesizing ability. In addition to the PDATs from higher plants, a single PDAT with multiple catalytic functions has been characterized in the unicellular green alga Chlamydomonas reinhardtii (Yoon et al., 2012). It is worth noting that the PDAT-mediated TAG-forming mechanism also has been detected in the bacterium Streptomyces coelicolor (Arabolaza et al., 2008), but it has no counterpart in mammals.

These previous studies reveal that (1) PDAT can exist as multiple copies in plant genomes, (2) different PDAT gene paralogs can encode enzymes with different TAG-synthesizing ability, and (3) certain PDATs can have unique substrate selectivity. All these findings shed new light on TAG biosynthetic mechanisms in plants and highlight the need for a deeper understanding of the complexity of plant PDATs. In this study, we have sought to provide further insights into the present-day diversity and ortholog/paralog relationship of plant PDATs via a genome-wide comparative analysis.

RESULTS

Identification of the PDAT Gene Family in Plants

The growing number of fully sequenced plant genomes makes it possible to perform a comparative genomic analysis of the PDAT gene family across a wide range of plant species. To identify PDATs in different plant species, a genome-wide search was performed using both Arabidopsis AthPDAT1 and AthPDAT2 amino acid sequences as queries to BLAST against 40 genomes listed in the Phytozome database. Candidate PDAT genes were found in all examined plant genomes, including algae, lowland plants (a moss and a lycophyte), monocots, and eudicots. Multiple hits were identified in each land plant genome, with the exception of Brachypodium distachyon. Only one hit was identified in each of the algal genomes. In total, 139 sequences were identified, and sequence information is provided in Supplemental Table S1. Among the 139 sequences, six from five species (potato [Solanum tuberosum], Populus trichocarpa, Medicago truncatula, apple [Malus domestica], and Arabidopsis lyrata) encode fewer than 200 amino acid residues, which is most likely due to genome annotation errors (Supplemental Table S2). These short sequences were eliminated from further analysis. In addition, the predicted transcripts from the apple genome have multiple stop codons, and the predicted transcript from the Ostreococcus lucimarinus genome does not start with a start codon; thus, the sequences from these two species were excluded. In the end, a total of 128 sequences were included for the analysis. To verify the reliability of BLAST results, these 128 protein sequences were subjected to InterPro and Pfam analysis (Supplemental Table S3), and all of them were classified into the LCAT family (Pfam: 02450).

A previous study (Yoon et al., 2012) showed that the LCAT-like family proteins from plants can be divided into four major groups, including PDAT, LCAT, PHOSPHOLIPID:STEROL ACYLTRANSFERASE (PSAT), and PHOSPHOLIPASE A (PLA) proteins. Therefore, some candidate sequences identified by BLAST may not encode PDAT. To clarify if the sequences obtained from BLAST are PDAT genes, phylogenetic analysis of the 128 full-length LCAT-like gene sequences was carried out. The maximum likelihood (ML) tree (Fig. 1; Supplemental Fig. S1) shows that all algal candidates are grouped together into a single clade (algal group), whereas the land plant sequences are partitioned into four major clades, designated as groups A, B, C, and D, with 91, three, four, and 26 identified sequences, respectively. The sequences with an expectation value (E value) < 1e-15 fall into group A, while the remaining ones (E value > 1e-15) branch into groups B, C, and D. Group B is more closely related to group A than either group C or group D is. Concerning the genes already characterized, group A contains all genes that were previously experimentally characterized as PDAT, including Arabidopsis, flax, and castor bean PDATs (Ståhl et al., 2004; Zhang et al., 2009; Kim et al., 2011; van Erp et al., 2011; Fan et al., 2013). The AthLCAT-like2 sequence (ATIG04010; Banas et al., 2005) from group B was previously identified and experimentally characterized as PSAT. Two MtrLCAT-like sequences (Medtr7g080450 and Medtr4g083980) from group C were predicted previously to be PLA, while the AthLCAT-like1 sequence (ATIG27480) from group D was classified as LCAT (Yoon et al., 2012).
Next, all protein sequences were classified using the InterPro and Protein Analysis through Evolutionary Relationships (PANTHER) classification systems. The results (Supplemental Table S3) show that all sequences are classified as LCAT-related proteins (PTHR:11440). PANTHER subfamily classification further reveals that the algal sequences and land plant sequences in group A belong to the PDAT subfamily (PTHR:11440:SF4), while the land plant sequences in groups B and C are classified into the PSAT subfamily (PTHR:11440:SF7) and the LCAT-like 4-related subfamily (PTHR:11440:SF3), respectively. The sequences in group D have no PANTHER subfamily classification.

The phylogeny results combined with the PANTHER classification suggest that the genes in groups B, C, and D very likely encode PSAT, PLA, and LCAT, respectively, rather than PDAT. Therefore, the sequences from group A are named PDAT, while the sequences from groups B, C, and D are named LCAt-like sequences. Only the sequences from group A (PDAT group) and the algal clade were included for further analyses. To avoid incomplete sampling of PDAT paralogs within species, the species with short candidate PDATs (potato, P. trichocarpa, M. truncatula, and A. lyrata) were eliminated. In the end, 86 full-length PDAT candidates from 34 species were selected for further analyses. The obvious annotation errors, including incorrect stop codon predictions and splicing errors, in six of the final selected sequences were manually corrected based on the EST database and the intron phases (Supplemental Table S2).

In summary, algal and lowland plant species possess a single copy of the PDAT gene. In monocots, one duplication event appears to have occurred in maize (Zea mays) and Panicum virgatum, resulting in a duplicated gene pair, while remaining monocots contain only one copy of PDAT. The PDAT copy number varied from two to six among eudicots, suggesting that multiple duplication events may have occurred in eudicots. It is worth noting that the number of PDAT paralogs within Arabidopsis, castor bean, flax, and C. reinhardtii identified in this study is consistent with previous studies (Ståhl et al., 2004; van Erp et al., 2011; Yoon et al., 2012; Pan et al., 2013).

To explore the evolutionary relationship of the plant PDAT gene family, we further constructed ML trees using complementary DNAs (cDNAs) of 86 full-length candidate PDATs. A gene encoding PDAT was first identified in S. cerevisiae; therefore, this sequence was included as an outgroup in phylogenetic analyses. The ML trees were built by using two phylogenetic programs, MrBayes and Randomized Accelerated Maximum Likelihood (RAxML). The trees based on both programs are topologically identical. Figure 2 shows the tree produced by MrBayes. Based on the topology and clade support values (85% or greater), the PDAT gene family can be classified into seven clades designated as clades I to VII (Fig. 2). The algal PDATs are phylogenetically divergent from the land plant PDATs and form a monophyletic group. Inside the land plants, PDATs from a moss, a lycophyte, and monocots diverged from each other and form three distinct clades, assigned as clades II to IV. The eudicots can be divided into the basal and the core eudicots. Within the core eudicots, PDATs are grouped into three clades: clades V, VI, and VII. Clades V and VI are more closely related to each other than they are to clade VII. For the two PDATs found in the basal eudicot A. coerulea, one (AcoPDAT1) is sister to the core eudicot clade VII, while the other (AcoPDAT2) forms a sister clade to the core eudicot clades V and VI. The number of PDAT paralogs in each species and their clade distributions are shown in Figure 3.

It is important to point out that taxa in the phylogenetic tree are very unevenly distributed among the clades, ranging from one to 28 sequences (Fig. 2), which may have a negative impact on phylogenetic accuracy (Heath et al., 2008). To examine the phylogenetic accuracy, we first compared the phylogenetic tree (Fig. 2) with the species tree shown in Phytozome (Fig. 3). As shown, the phylogenetic tree accords exactly with the evolutionary pathway from algae to basal eudicots. Algal PDATs are grouped at the base of the tree. The PDATs from Physcomitrella patens (moss) and Selaginella moellendorffii (lycophyte), two basal lineages of land plants, form monophyletic clades after the algal clade. Monocot PDATs form a monophyletic clade, with the more related species being closer on the phylogenetic tree. The basal eudicot (A. coerulea) is placed as sister to the core eudicots. When it comes to the core eudicots, the
Figure 2. Phylogenetic relationship of the PDAT gene family. The ML tree was generated using the MrBayes program. PDATs are grouped into seven distinct clades (I–VII). Numbers above branches represent the support values (Bayesian posterior probabilities). The tree is rooted using PDAT sequences from *S. cerevisiae* as the outgroup. The scale bar represents the number of nucleotide replacements per site.
existence of multiple PDAT copies makes the comparison between the species tree and the PDAT gene tree complicated. As shown in Figure 2, three core eudicot PDAT clades, each containing copies from a mixture of species, differ in topology from one another and from the species tree. To further test if the topological discordance between the core eudicot PDAT gene trees and the species tree is the result of the uneven taxonomic sampling, we pruned down the data set and reconstructed the phylogenetic tree with the sequences from clades V, VI, and VII, which have more balanced taxa sampling. The trees generated from the pruned and complete data sets are topologically identical (Fig. 2; Supplemental Fig. S2), suggesting that the phylogenetic separation of core eudicot PDATs was not affected by the very unevenly distributed taxa among the clades. The causes of discordance between the multicopy PDAT gene trees and the species tree remain unknown, but it is a well-known phenomenon that gene trees do not necessarily agree with the species tree, and this discordance can be the result of many evolutionary processes, such as gene duplication and loss, and incomplete lineage sorting (Maddison, 1997; Page and Charleston, 1997).

Overall, the high confidence of our phylogenetic separation of the PDAT gene family is achieved through the high bootstrap support obtained from multiple phylogenetic reconstruction methods, comparisons between the phylogenetic tree and the species tree, as well as consistent phylogenetic topologies inferred from the complete and pruned data sets.

Gene Structure Analysis Reveals Highly Conserved Exon/Intron Structure and Intron Phase Pattern throughout Land Plant PDATs

To further investigate the structural diversity of plant PDAT genes, we analyzed the exon/intron organization for each individual gene (representative PDATs are shown in Fig. 4, and details are shown in Supplemental Fig. S3). Diverse gene structure has been found in the algal PDATs (clade I): Micromonas pusilla CCMP1545 PDAT (MpuCMPDPAT) has no intron, while the rest of the PDAT genes have nine to 14 introns. By contrast, land plant PDATs (clades II–VII) are remarkably well conserved in terms of exon/intron structure. Approximately 92% (76 out of 82) of the land plant PDATs have six exons and five introns. Six exceptions to this exon/intron pattern are VviPDAT1, PnuPDAT1, GmaPDAT5, and GraPDAT1, with seven exons and six introns, and LusPDAT3 and

Figure 3. Number of PDAT paralogs in each species and their clade distributions. The species tree is based on information in Phytozome (http://www.phytozome.net). The star on the branch point within eudicot species indicates the divergence point between a basal eudicot (A. coerulea) and the core eudicots. The tree is not scaled.
LusPDAT6, with five exons and four introns. In addition, we also investigated intron phases across all PDATs. Intron phase can be classified into three categories (0, 1, and 2) depending on the position of the intron relative to the codon: phase 0 intron does not interrupt the reading frame and lies between two consecutive codons; phase 1 intron inserts and interrupts the reading frame between the first and second nucleotides; and phase 2 intron inserts and interrupts the reading frame between the second and third nucleotides. The analysis shows that the intron phase pattern (2, 0, 2, 0, 2) is strikingly conserved across 75 out of 82 land plant PDATs (Fig. 4; Supplemental Fig. S3).

Evaluation of PDAT Protein Properties Reveals That PDATs Belonging to the Core Eudicot Clade VI Had a Tendency to Maintain Acidic pI Values during Evolution

After the evaluation of gene structure, we continued our analysis with a focus on the protein properties of 86 PDATs, including protein length, molecular mass, and pI values. According to our analyses (Table I; Supplemental Table S4), the length and molecular mass of PDATs from clade I (algae clade) varied substantially. CrePDAT with 1,041 amino acid residues and 104.5 kD is the longest and largest PDAT, while CsuPDAT with 509 amino acid residues and 56.9 kD is the shortest and smallest PDAT of all 86 PDATs.
In contrast, the variation of protein length and molecular mass is small in land plant PDATs, ranging from 572 to 716 amino acid residues and 62.8 to 80.3 kD, with a mean of 671 amino acid residues and 74.7 kD. For the pI values, MpuCMPPDAT has the highest value of 9.53. Except for MpuCMPPDAT, PDATs within clades I, II, III, and IV have very close pI values, ranging from 5.96 to 6.5, with an average of 6.21. Interestingly, PDATs from clade VI (except GraPDAT5) have maintained acidic pI values with an average of 6.35, while more alkaline pI values (greater than 7) have been observed in 31 out of 42 PDATs belonging to clades V and VII.

**Table I. Summary of PDAT protein properties**

| Clade | Total No. of PDATs | Protein Length (Amino Acid Residues) | Molecular Mass (kD) | pI | No. of PDATs with pI > 7 |
|-------|--------------------|-------------------------------------|---------------------|----|------------------------|
| I     | 4                  | 757 ± 223                           | 79.6 ± 19.6         | 6.90 ± 1.76 | 0                      |
| II    | 1                  | 693                                 | 76.9                | 6.50 | 0                      |
| III   | 1                  | 647                                 | 71.8                | 6.21 | 0                      |
| IV    | 8                  | 682 ± 4.87                          | 75.4 ± 0.4          | 6.25 ± 0.13 | 0                      |
| V     | 17                 | 670 ± 21.96                         | 74.5 ± 2.33         | 7.65 ± 0.94 | 12                     |
| VI    | 28                 | 676 ± 21.66                         | 75.1 ± 2.52         | 6.34 ± 0.61 | 1                      |
| VII   | 25                 | 663 ± 40.7                          | 74.1 ± 4.76         | 7.94 ± 0.97 | 19                     |
| Minimum | 509            | 56.9                                | 5.27                |                |
| Maximum | 1,041           | 104.5                               | 9.53                |                |

Plant PDATs Contain the Conserved Amino Acids in LCAT

To gain more insights about the structure/function features of PDATs, multiple sequence alignment was further used to identify the conserved amino acid residues. The alignment shows that, besides the initial Met residue, 39 amino acid residues are completely conserved in 86 PDATs. Among the completely conserved amino acid residues, nine of them are located at the N-terminal portion and the rest are concentrated within the 320 amino acid residues following the TMD.

It is known that PDAT belongs to the LCAT-like family. The first PDAT gene was isolated based on its homology to human LCAT, which is a soluble protein with no TMD. It was reported previously that human LCAT contains several structurally conserved elements (Peelman et al., 1998, 1999), including a catalytic triad of Ser-181-His-377-Asp-345, a salt bridge between Asp-145 and Arg-147, and a so-called lid region. The Trp-61 within the lid region was proposed to play an important role in binding the cleaved fatty acid into the active site for an optimal acylation process. Aligning plant PDATs with human LCAT reveals (examples are shown in Fig. 6, and details are shown Supplemental Fig. S5) that the Trp-61, Asp-145, Arg-147, Ser-181, and Asp-345 of LCAT are completely conserved in 86 PDATs, while His-377 is conserved in 83 out of 86 PDATs, with the exception of GmaPDAT5, GmaPDAT6, and PvuPDAT1. Since the available EST sequences of GmaPDAT5,
and PvuPDAT1 were not long enough to cover the coding regions for His-377, we are not sure if these mismatches are the result of genome sequencing errors.

Conservation and Variation in the Motif Composition and Arrangement of PDATs Provides Further Support for the Grouping of Phylogenetic Clades

We further analyzed the motifs in PDATs. InterPro search identified two signature protein motifs in all PDATs, which are IPR003386 for the LCAT family and IPR029058 for the α/β-hydrolase fold family. InterPro, however, is limited to the known motifs present in PDATs.

In order to further identify the conservation and variation in the motif arrangements among PDATs, all PDATs were subjected to a Multiple Expectation Maximization for Motif Elicitation (MEME) analysis. A total of 51 distinct motifs were identified. The occurrences of the motifs in representative PDATs from seven major clades are shown in Figure 7. More detailed information is provided in Supplemental Figures S6 and S7. The analysis shows that the motif composition of PDATs in algae is very different from that in land plants, which corresponds to their divergent gene structure. Land plant PDATs were found to share many of the motifs.

Among 51 motifs, 11 motifs are classified into the LCAT family. All PDATs contain four LCAT-like motifs occurring after the TMD, with the first three arranged sequentially and the fourth one separated. Six major LCAT motif patterns (Fig. 7; Supplemental Fig. S7) have been identified based on the MEME combined block diagram. These patterns include the following: motifs 4, 50, 33, and 51 in clade I; motifs 1, 12, 10, and 5 in clade III; motifs 1, 15, 10, and 5 or motifs 1, 12, 10, and 5 in clade VI; motifs 1, 12, 7, and 5 or motifs 1, 12, 7, and 3 in clade VII; and motifs 1, 15, 10, and 5 shared among clades II, V, and VI.

Next, we examined the non-LCAT motif composition in land plant PDATs. Based on the position of the TMD and LCAT-like motifs, we further divide PDATs
into four regions (Fig. 7): region 1 covers the segment before the TMD; region 2 spans the segment between the TMD and the first LCAT-like motif; region 3 defines the region between the third and fourth LCAT-like motifs; region 4 corresponds to the C-terminal segment. Among these four regions, regions 2 and 4, mainly made up of motifs 2 and 8, are highly conserved within the land plants. Region 3 is less conserved, mainly composed of motif 22 in clades II and III, motif 11 in clades IV, V, and VI, and motifs 14 and 22 in clade VII. Region 1 appears to be the most divergent region, in which the clade-specific motifs, including motif 38 in clade IV and motif 27 in clade VII, were found. It is worth mentioning that individual MEME motifs are gapless (no insertions or deletions), which means that motifs containing gaps still can be discovered, but they will be split into multiple ungapped motifs. It will be interesting to find out which motifs are really different and which motifs may be associated with specific functions.

Taken together, the identified LCAT and non-LCAT motif patterns match the clading pattern in the phylogenetic tree.

Eudicot-Wide PDAT Gene Expansion Arise Mainly from the Eudicot-Shared Ancient Gene Duplication followed by Species-Specific Segmental Duplications

The existence of multiple PDAT gene copies across eudicots suggests that the PDAT gene family expanded in eudicots. Gene copy number expansions can occur via three major evolutionary events: segmental duplication, tandem duplication, and transposition events (Kong et al., 2007). In this study, we focused on segmental and tandem gene duplications.

The phylogenetic analysis clearly divides the core eudicot PDATs into three distinct clades (Fig. 2). Each clade contains sequences from taxa across the core eudicots, including both the asterids and the rosids, indicating that PDAT paralogs among different clades were produced by a core eudicot-shared ancient gene duplication that predated the split of the two major clades of core eudicots. Based on our data, we cannot say with certainty whether this ancient gene duplication was shared with the basal eudicots. This will have to be confirmed in the future, when additional genome sequences of basal eudicots are available. The genes derived from the
ancient gene duplication are named paleoduplicated genes. Due to the fact that the ancient gene duplication was followed by species-specific gene duplication, gene loss, and chromosome rearrangements, only six species (grape [Vitis vinifera], Theobroma cacao, castor bean, Carica papaya, Eucalyptus grandis, and common bean [Phaseolus vulgaris]) have maintained the triplicated paleologous PDAT genes, each of which is present in one of the three core eudicot clades (clades V–VII; Figs. 2 and 3).

Besides paleoduplicated PDAT paralogs among different clades, we also found that some species contain duplicated gene pairs within the clades. These include three PDAT gene pairs in L. usitatissimum, soybean (Glycine max), and Gossypium raimondii, two in tomato (Solanum lycopersicum) and Manihot esculenta, and one in Mimulus guttatus and Brassica rapa (Fig. 2). To determine whether these within-clade gene pairs were derived from segmental duplication events, we analyzed 10 protein-coding genes from upstream and downstream of each PDAT gene pair. The results show that the genes flanking each PDAT gene pair are highly conserved in all species but G. raimondii, indicating that these PDAT gene pairs were formed via the segmental duplication event. The difference found in G. raimondii may be partly explained by the fact that the Gossypium spp. genome has a unique evolutionary history. The lineage-specific whole-genome multiplication event(s) occurred approximately 60 million years ago in Gossypium spp. genomes (Paterson et al., 2012).

The within-clade gene pairs tend to have higher sequence identity than the between-clade gene pairs (Tables II and III). Thus, we speculated that the within-clade gene pairs were derived from more recent duplication events. To confirm our hypothesis, we used the synonymous substitution rates ($K_s$) as a proxy for time to compare the date of gene duplications. Judging from the $K_s$ values, the within-clade gene pairs have a much lower $K_s$ than between-clade gene pairs, suggesting more recent duplications (Tables II and III). Because most of the $K_s$ values for between-clade gene pairs are saturated (greater than 2), such data can only provide a rough estimate.

Next, we investigated the role of tandem duplication in the evolution of the eudicot PDAT gene family. The previous literature indicated that a chromosome region consisting of two or more copies of a gene within 200 kb can be viewed as a gene cluster (Holub, 2001). Chromosome location analysis shows that the majority of the PDAT genes are located along scattered sites throughout the genome, and a single tandem duplication cluster
consisting of two genes has only been found in *Citrus clementina* (CclPDAT2 and CclPDAT3) and *Citrus sinensis* (CsiPDAT1 and CsiPDAT2). This suggests that tandem duplication does not play a dominant role in the expansion of the PDAT gene family in eudicots. Taken together, these analyses reveal that the eudicot-shared ancient gene duplication followed by species-specific segmental duplication primarily contributes to the expansion of the PDAT gene family in eudicots.

**Different Selection Pressures Have Acted on the Paleoduplicated PDAT Paralogos**

It has long been thought that gene duplication plays a crucial role in the evolution of gene diversity (Ohno, 1970; Hughes, 1994; Roth et al., 2007). To explore the evolutionary fate of paleoduplicated eudicot PDATs, we performed selection pressure analyses. The selective pressure acting on the core eudicot PDATs was estimated using the ratio (ω) of the nonsynonymous substitution rate versus the κ, as an indicator. Because two basal eudicot PDATs are sister to, rather than nested within, the core eudicot clades, they were not included in the analyses. Specifically, we extracted the core eudicot sequences and constructed a phylogenetic tree using *SmaPDAT1* as an outgroup (Supplemental Fig. S8).

To address the possibility of functional divergence among the core eudicot clades, we fit our data to the clade model C (CmC) implemented in PAML. In CmC, the entire target clade is set as the foreground partition, while the rest of the phylogeny comprises the background partition. In view of the above phylogenetic trees (Fig. 2; Supplemental Fig. S8), clade VII is more distant from clades V and VI. Therefore, we first applied CmC with the entire clade VII set as the foreground partition; clades V and VI along with the outgroup comprised the background partition. We call this analysis CmC VII. The CmC assumes that different selection pressures have acted on the foreground and background partitions, while the null model, M2a_rel, hypothesizes that there is no significant difference in selection pressures between the foreground and background partitions. The likelihood ratio test (LRT) comparing CmC versus the M2a_rel null model shows that CmC VII fits the data significantly better than the null model (P < 0.001; Table IV). Parameter estimates indicate that a larger set of sites (approximately 56%) evolved under stronger purifying selection (ω0 = 0.02764) and a smaller set of sites (approximately 43%) evolved under divergent selective pressures, with weaker purifying selection in clade VII (ω3 = 0.23272) and stronger purifying selection in the background (ω2 = 0.15745).

Similar CmC analysis was further applied to clade V (referred to as CmC V) and clade VI (referred to as CmC VI). The LRT results show that CmC VI (P < 0.001) but not CmC V (P > 0.2) provides a significantly better fit than the null model (Table IV). However, including both clades VI and VII in the background partition in CmC V might be inappropriate, as the average of their ω ratio (clade VI ω1 = 0.14847 and clade VII ω3 = 0.23272) is close to the ω ratio for clade V (ω0 = 0.18669). To evaluate this possibility, we employed the extended clade model (Yoshida et al., 2011), which allows more than two partitions (foreground and background). We specified three partitions in our analysis: clades V, VI, and VII. We call this analysis Ex-CmC. The null hypothesis is that selection pressure is the same for clades V and VI. In null model testing, the phylogeny was divided into two partitions: clade VII and the combined clades V and VI. The null model for this test is named Ex-Null. The LRT result (Table IV) indicates that the null hypothesis is rejected, supporting that different selection pressures have acted on clades V and VI. To further confirm our result, we excluded clade VII from the analysis and only compared the functional divergence between clades V and VI. Using this data set, we found that setting either clade V or clade VI as the foreground partition yields a significant LRT result (P < 0.001; Table V), indicating that selective constraint indeed differs between clades V and VI.

| Table II. Sequence identity and κ, values of within-clade duplicated PDAT gene pairs |
|---------------------------------|----------------|----------------|----------------|----------------|
| Species                        | Gene Pair      | Clade | Sequence Identity | κ,  | Mean κ,  |
| Soybean                        | GmaPDAT3 versus GmaPDAT4 | V     | 96.1             | 0.08996 ± 0.01655 | 0.125 |
|                               | GmaPDAT1 versus GmaPDAT2 | VI    | 95.9             | 0.12586 ± 0.01946 | 0.125 |
|                               | GmaPDAT5 versus GmaPDAT6 | VII   | 92.0             | 0.15779 ± 0.02457 | 0.794 |
|                               | GraPDAT4 versus GraPDAT6 | V     | 79.1             | 0.70247 ± 0.06247 | 0.794 |
|                               | GraPDAT1 versus GraPDAT5 | VI    | 76.4             | 0.80740 ± 0.07259 | 0.794 |
|                               | GraPDAT2 versus GraPDAT3 | VII   | 76.9             | 0.87113 ± 0.07781 | 0.794 |
| Tomato                        | SlyPDAT3 versus SlyPDAT4 | VI    | 81.4             | 0.71665 ± 0.06707 | 0.755 |
|                               | SlyPDAT1 versus SlyPDAT5 | VII   | 79.7             | 0.79243 ± 0.06856 | 0.755 |
| M. esculenta                  | MesPDAT3 versus MesPDAT5 | VI    | 91.0             | 0.29142 ± 0.03067 | 0.316 |
|                               | MesPDAT2 versus MesPDAT4 | VII   | 87.6             | 0.34022 ± 0.03134 | 0.316 |
| Flax                          | LusPDAT1 versus LusPDAT5 | V     | 97.1             | 0.08181 ± 0.01528 | 0.100 |
|                               | LusPDAT2 versus LusPDAT4 | VI    | 95.8             | 0.12603 ± 0.01866 | 0.100 |
|                               | LusPDAT3 versus LusPDAT6 | VI    | 96.2             | 0.09326 ± 0.01564 | 0.100 |
| M. guttatus                   | MguPDAT2 versus MguPDAT3 | VI    | 76.7             | 1.35287 ± 0.14409 | 1.353 |
| B. rapa                       | BraPDAT1 versus BraPDAT2 | VI    | 91.5             | 0.3522           | 0.352 |
DISCUSSION

Despite the fact that many studies have revealed the crucial role of PDATs in TAG biosynthesis, our knowledge of PDATs is still very limited. To advance our understanding of the involvement of PDATs in TAG biosynthesis, it is essential to first understand their evolution and diversity. The goal of our study was to provide an overall picture of plant PDATs, including their gene family members, evolutionary history, present-day diversity, and structural similarities and differences.

Evolution, Conservation, and Variation of the PDAT Gene Family in Plants

Taking advantage of publicly available sequenced plant genomes, we surveyed 40 different plant species and identified 139 LCAT-like sequences. The results of phylogeny (Fig. 1) and PANTHER classification analyses (Supplemental Table S3), along with the previous findings, indicated that only sequences with E value < 1e-15 from land plants and sequences from algae belong to the PDAT family and, therefore, were included for the further analyses. PDAT candidates exist in all plants analyzed, including algae, lowland plants (a moss and a lycophyte), and highland plants (monocots and eudicots). The evolutionary analysis shows that the PDAT gene family can be clearly divided into seven major clades (Fig. 2).

Four algal PDATs form a separate well-supported clade (clade I) from land plant PDATs. This phylogenetic separation is supported by their different gene structure (Fig. 4), protein properties (Table I), and motif composition (Fig. 7). The observed differences between algal and land plant PDATs might be associated with different biological functions. Consistent with this hypothesis, the study of the microalga C. reinhardtii PDAT (Yoon et al., 2012) revealed that it has some unique features that have not been reported in land plant PDATs. For instance, the algal PDAT appears to be a chloroplast-localized protein with a higher preference for chloroplast membrane lipids (e.g. phosphatidylglycerol and phosphatidylinositol) over the other phospholipids (phosphatidylcholine and phosphatidylethanolamine). In addition, the algal PDAT is a multifunctional enzyme that has not only PDAT and DAG:DAG acyltransferase function but also galactolipid:DAG acyltransferase function.

Table IV. Parameter estimates, likelihood values, and LRT P values obtained from CmC and Ex-CmC analyses of the 70 core eudicot PDATs data set

For the clade model, \( \omega_0 \) is the estimated \( \omega \) value for site class 0, \( p_0 \) is the estimated proportion of sites in site class 0 (purifying selection), \( \omega_1 \) is the estimated \( \omega \) value for site class 1, \( p_1 \) is the estimated proportion of sites in site class 1 (neutral selection), \( \omega_2 \) is the estimated \( \omega \) value for divergent sites on the background partitions, \( \omega_3 \) is the estimated \( \omega \) value for divergent sites on the foreground partitions, and \( p_2 \) is the estimated proportion of sites in the evolutionary history of PDATs, present-day diversity, and structural similarities and differences.

Evolution, Conservation, and Variation of the PDAT Gene Family in Plants

Taking advantage of publicly available sequenced plant genomes, we surveyed 40 different plant species and identified 139 LCAT-like sequences. The results of phylogeny (Fig. 1) and PANTHER classification analyses (Supplemental Table S3), along with the previous findings, indicated that only sequences with E value < 1e-15 from land plants and sequences from algae belong to the PDAT family and, therefore, were included for the further analyses. PDAT candidates exist in all plants analyzed, including algae, lowland plants (a moss and a lycophyte), and highland plants (monocots and eudicots). The evolutionary analysis shows that the PDAT gene family can be clearly divided into seven major clades (Fig. 2).

Four algal PDATs form a separate well-supported clade (clade I) from land plant PDATs. This phylogenetic separation is supported by their different gene structure (Fig. 4), protein properties (Table I), and motif composition (Fig. 7). The observed differences between algal and land plant PDATs might be associated with different biological functions. Consistent with this hypothesis, the study of the microalga C. reinhardtii PDAT (Yoon et al., 2012) revealed that it has some unique features that have not been reported in land plant PDATs. For instance, the algal PDAT appears to be a chloroplast-localized protein with a higher preference for chloroplast membrane lipids (e.g. phosphatidylglycerol and phosphatidylinositol) over the other phospholipids (phosphatidylcholine and phosphatidylethanolamine). In addition, the algal PDAT is a multifunctional enzyme that has not only PDAT and DAG:DAG acyltransferase function but also galactolipid:DAG acyltransferase function.

Table IV. Parameter estimates, likelihood values, and LRT P values obtained from CmC and Ex-CmC analyses of the 70 core eudicot PDATs data set

For the clade model, \( \omega_0 \) is the estimated \( \omega \) value for site class 0, \( p_0 \) is the estimated proportion of sites in site class 0 (purifying selection), \( \omega_1 \) is the estimated \( \omega \) value for site class 1, \( p_1 \) is the estimated proportion of sites in site class 1 (neutral selection), \( \omega_2 \) is the estimated \( \omega \) value for divergent sites on the background partitions, \( \omega_3 \) is the estimated \( \omega \) value for divergent sites on the foreground partitions, and \( p_2 \) is the estimated proportion of sites in site class 2 (divergent selection). \( \text{ln} \) is the log likelihood value, \( \Delta \text{ln} \) is the LRT statistic for comparing the CmC and M2a_ref (null) models, and \( P \) is the value of the LRT. For the extended clade model, the first two site classes (class 0 and class 1) are the same as in the CmC. The final class (class 2) models divergent selection among three partitions, each with a separately estimated \( \omega \) ratio (\( \omega_0 \) for clade V, \( \omega_1 \) for clade VI, and \( \omega_2 \) for clade VII). The null model has only two partitions and two estimated \( \omega \) ratios (\( \omega_0 \) for clade VII and \( \omega_0 \) for both clades V and VI). \( \text{ln} \) is the log likelihood value, \( \Delta \text{ln} \) is the LRT statistic for comparing the Ex-CmC and Ex-Null models, and \( P \) is the value of the LRT.
and lipase activities. It is important to note that the sequences within the algal clade (clade I) have very long branches, suggesting that the algal PDATs are quite different from each other. Due to the limited number of algal PDATs, we did not further divide algal PDATs and considered them as a single clade in this study. It will be interesting to include more algal sequences to evaluate the extent of divergence among the algal PDATs.

Most land plant PDATs share four major structural features at both the gene and protein levels. First, the exon/intron structures (six introns/seven exons) and intron phase patterns (2, 0, 2, 0, 2) are remarkably conserved in most land plant PDAT genes (Fig. 4; Supplemental Fig. S3), suggesting that the PDAT gene structure in land plants has been established and retained after the divergence of land plants from algae. Second, a single TMD in the N terminus has been preserved in most land plant PDATs (Fig. 5; Supplemental Fig. S4). Third, all LCAT-like motifs (Fig. 7; Supplemental Fig. S7) and LCAT-conserved amino acid residues (Fig. 6; Supplemental Fig. S5) are located at the C-terminal end of the TMD, suggesting that the active and/or binding sites of land plant PDATs possibly face the luminal side of the ER. Fourth, the C-terminal portion and the region between the TMD and the first LCAT-like motif are highly conserved (Fig. 7; Supplemental Fig. S7). It has been reported that two Arabidopsis PDATs contain ER retrieval signals at their C termini (McCartney et al., 2004); therefore, it is possible that the C terminus is involved in assisting the association of PDAT with the ER.

Besides similarities, there are variations among land plant PDATs. The alignment of PDAT polypeptides shows that the hydrophilic N terminus preceding the TMD is the most divergent region (Fig. 5B). The motif occurrences in the N terminus are quite unique for PDATs from each clade (Fig. 7; Supplemental Fig. S7). The clade IV-specific motif 38 and clade VII-specific motif 27 were found within this region. Therefore, this region could serve as a candidate target to study the functional and structural divergence among land plant PDATs from different clades. Although deletion of the TMD along with the N terminus of yeast PDAT does not affect its catalytic activity and substrate selectivity (Ghosal et al., 2007), it is still possible that the N terminus of land plant PDATs is associated with specific functions, such as sorting PDATs to the ER (Pelham, 2000) and forming a multimeric complex, as demonstrated in DGAT1 from plants and animals (Pelham, 2000; Cheng et al., 2001; Weselake et al., 2006; McFie et al., 2010).

Our analysis also reveals a eudicot-wide PDAT gene expansion. Combined with evidence from the phylogenetic (Fig. 2) and Kc analyses (Tables II and III), the eudicot-shared ancient gene duplication followed by species-specific segmental duplications appears to be mainly responsible for the expansion of PDAT genes in eudicots. The duplicated core eudicot PDATs are grouped into three clades (clades V–VII). The MEME combined block reveals that the motif compositions of PDATs in clade VII are quite different from those in clades V and VI (Fig. 7; Supplemental Fig. S7). A tendency for the conservation of acidic pI values in clade VI (Table I; Supplemental Table S4) adds another distinct characteristic for the separation among the core eudicot clades. In addition, Kim et al. (2011) observed different subcellular localizations for the proteins encoded by paleoduplicated PDAT paralogs in castor bean. RcoPDAT1A and RcoPDAT1B were found to be ER localized, whereas RcoPDAT2 was proposed to be localized in the plasma membrane. It will be interesting to study the localization of PDAT protein paralogs in different species to determine whether the plasma membrane-localized RcoPDAT2 is an exclusive case.

| Model | lnL   | 0: Purifying | 1: Neutral | 2: Divergent | 2Δ   | P    |
|-------|-------|-------------|------------|-------------|------|------|
|       |       | ω0, p0      | ω1, p1     | ω2, ω3, p2  |      |      |
| M2a_ref | −33,307.980006 | 0.02467 | 0.6711 | 1.0000 | 0.02888 | 0.20335 | 0.30002 | 19.778246 | <0.001 |
| CmC V  | −33,298.090883 | 0.02432 | 0.66597 | 1.0000 | 0.02792 | 0.18046 | 0.30252 | 19.778246 | <0.001 |
| CmC VI | −33,301.896393 | 0.02506 | 0.67361 | 1.0000 | 0.02779 | 0.23597 | 0.29860 | 12.167226 | <0.001 |

Has Ancient Gene Duplication Led to Functional and Expression Divergence among PDAT Paralogs?

This study was motivated by the finding that some plant genomes contain multiple PDAT paralogs that show evidence of diverging TAG-synthesizing function. Our study showed that there is a eudicot-wide PDAT gene expansion, but questions remain. Why do eudicots contain multiple copies of PDATs in their genome? Is gaining functional divergence among PDAT paralogs a general trend in the evolution of the eudicot PDATs? And how may this happen? Now, we may be able to answer these questions from an evolutionary perspective.

Gene duplication is believed to be one of the major driving forces for evolutionary novelties, including neofunctionalization (Ohno, 1970; Force et al., 1999;
Roth et al., 2007) and subfunctionalization (Li et al., 2005; Wang et al., 2012), at the level of expression or coding sequence. A central theory of molecular evolution states that most genes evolved primarily under strong purifying constraints for functional conservation, and gene duplication allows a gene to be free from this selection pressure and eventually accumulate mutations that can lead to new function or complete loss of function (Ohno, 1970; Lynch and Conery, 2000). Our selection pressure analyses show that (1) strong purifying selection is a primary evolutionary mode for the core eudicot PDATs and (2) after ancient gene duplication, paleoduplicated PDAT genes have been subjected to different selective constraints (Tables IV and V).

The observed heterogeneity in selection pressure among the core eudicot clades might enable the changes in gene function and/or the development of expression-level divergence among duplicated genes. Consistent with this hypothesis, previous studies (Ståhl et al., 2004; Zhang et al., 2009; Kim et al., 2011; van Erp et al., 2011; Pan et al., 2013) showed that PDATs, including Arabidopsis, flax, and castor bean PDATs, from clade VII (AthPDAT2, LusPDAT3, LusPDAT6, and RcoPDAT2) do not have an apparent function in TAG biosynthesis. Clade VII under the weakest selection constraint (Table IV) seems to have evolved in a manner very different from clades V and VI and may have eventually lost the TAG-synthesizing function. In addition to the possible nonfunctionalization of clade VII, previous studies (Ståhl et al., 2004; Zhang et al., 2009; Kim et al., 2011; van Erp et al., 2011; Pan et al., 2013) also provide some lines of evidence suggesting that expression divergence may have occurred between the PDAT paralogs from another two core eudicot clades (clades V and VI). More interestingly, many PDATs in clade VI appear to have been subfunctionalized at the expression level into the nonseed tissues. Studies on flax (Pan et al., 2013) and castor bean (Kim et al., 2011) PDATs revealed that the PDAT paralogs sitting on clade VI (LusPDAT2, LusPDAT4, and RcoPDAT1B) are expressed at high levels in nonseed tissues, and very low expression levels are detected in seeds. In this study, we also used RNA-sequencing (RNA-Seq) data to examine the expression profiles of PDATs in soybean and common bean. The results show that soybean and common bean PDAT paralogs in clade VI (GmaPDAT1, GmaPDAT2, and PruPDAT3) have significantly higher expression in leaves, flowers, and roots than in developing seeds (Supplemental Figs. S9 and S10). Although similar expression levels of Arabidopsis PDAT1 (from clade VI) in leaves, roots, flowers, and developing seeds were reported, a semi-quantitative reverse transcription-PCR approach with seeds at a single developmental stage (mid stage) was used in that study (Ståhl et al., 2004). To obtain more detailed expression data, we extracted microarray expression data for AthPDAT1 from the AtGenExpress database (http://jsp.weigelworld.org/express/express.jsp). The result (Supplemental Fig. S11) shows that AthPDAT1 does appear to have followed the trend of tissue subfunctionalization, and the expression level is generally higher in other tissues than in seeds across the different developmental stages. In addition, previous studies revealed that the TAG-synthesizing function of AthPDAT1 has been detected only in rapidly developing leaves rather than in seeds under both overexpression and RNA interference approaches (Mhaske et al., 2005; Fan et al., 2013). A fairly recent study (Fan et al., 2014) indicated that AthPDAT1-mediated TAG synthesis is involved in the process of diverting fatty acids from membrane lipids toward peroxisomal β-oxidation, thereby maintaining membrane lipid homeostasis in Arabidopsis leaves. PDATs from clade VI, which are closely related to AthPDAT1, may have a similar protective role in maintaining membrane integrity in leaf tissues. Furthermore, studies of mammalian proteins suggest that the shifts in pI values may be due to the functional divergence of proteins (Khalidi and Shields, 2011) or an adaptation to the changed subcellular localization or tissue compartmentalization (Alendé et al., 2011). Therefore, a tendency for the conservation of acidic pI values in clade VI may be related to tissue sub-functionalization. At the same time, it must be noted that three PDATs from the third core eudicot clade (clade V) characterized to date, LusPDAT1, LusPDAT5, and RcoPDAT1A, have seed-specific expression patterns, and the encoded enzymes have unique substrate selectivity properties (Kim et al., 2011; Pan et al., 2013). The identified α-linolenic acid-prefering flax PDATs (Pan et al., 2013) and hydroxy fatty acid-selective castor bean PDAT (van Erp et al., 2011) support the speculations that (1) the substrate selectivity of PDAT has coevolved with the species’ fatty acid composition (Yoon et al., 2012) and (2) the contribution of PDAT from clade V to seed oil synthesis can be significant in some oilseeds that are high in polyunsaturated fatty acids or unusual fatty acids. It will be very exciting to learn whether other PDATs from clade V have been specialized in seeds and developed unique substrate selectivity. In contrast, the PDATs found in algae, moss, lycophytes, and monocots form monophyletic clades. Two PDAT paralogs from maize have similar expression patterns (Supplemental Fig. S12). The PDAT expression profile in rice (Oryza sativa) did not show any significantly differential expression among tissues (Supplemental Fig. S12). The functional and expression divergence of PDATs appears to be core eudicot specific.

Together, we proposed answers to the above questions. Functional and expression divergence of PDAT paralogs appears to be a general trend in the evolution of the core eudicot PDATs. Ancient gene duplication may enable one of the paleoduplicated PDAT paralogs to become nonfunctionalized and another two paralogs to develop divergent expression patterns. Since it has been well recognized that gain of functional diversification and expression-level divergence is a key process in promoting the retention of duplicated genes in the genome (Lynch and Force, 2000; Torgerson and Singh, 2004; Li et al., 2005; Roth et al., 2007), nonfunctionalization and expression divergence among paralogs may account for the retention of multiple copies of PDATs in the core eudicots.
CONCLUSION

In conclusion, our study provides a comprehensive genomic analysis of the PDAT gene family in plants, covering phylogeny, gene structure, protein properties, topology, critical amino acid identification, functional motifs, and selection pressure analyses. Phylogenetic analysis indicates that plant PDATs can be clustered into seven distinct clades, which is further supported by conservation and variation in gene structure, protein properties, motif occurrences, and/or functional divergence among clades. In addition, selection pressure analyses demonstrate that paleoduplicated core eudicot PDATs have evolved under different selection constraints. Combined with the insights of previous studies, the observed variation in selection constraints might have led to the nonfunctionalization and expression divergence of paleoduplicated PDAT paralogs. Our current knowledge regarding the functions of plant PDATs is limited to only four PDATs: one in the unicellular green alga C. reinhardtii and three in core eudicots (Ståhl et al., 2004; van Erp et al., 2011; Yoon et al., 2012; Pan et al., 2013). To obtain a more thorough understanding of the evolution of the plant PDAT family, further sampling, expression profile analyses, and functional characterization of PDATs in more species will be necessary.

MATERIALS AND METHODS

Identification of PDAT Genes and Their Homologs in Plants

To identify PDAT genes and their homologs, we performed a TBLASTN search using the Arabidopsis (Arabidopsis thaliana) AthPDAT1 and AthPDAT2 protein sequences as queries against the Phytozone databases (http://www.phytozone.net/). All 39 listed plant species, including algae (Chlamydomonas reinhardtii, Volvox carteri, Cocomappa subdiversipila C-169, Micrasterias pusilla CCMP1545, and Ostrococcus lucimarinus), the lycophyte Selaginella moellendorffii, the moss Physcomitrella patens, monocots (Sorghum bicolor, maize [Zea mays], Selaginella sertularioides [S. selaginoides], Lycopodium clavatum), dicots (Populus trichocarpa, Medicago truncatula, common bean [Phaseolus vulgaris], soybean [Glycine max], Cassumum sativum, Puerariae sinesis, apple [Malus domestica], Fragaria vesca, Arabidopsis lyrata, Capsella rubella, Brassica rapa, Thellungiella halophila, Carica papaya, Gossypium raimondii, and Zea mays), gymnosperms (Pinus sylvestris, pine [Pinus sylvestris], Populus trichocarpa, black pine [Pinus nigra var. sinica], the conifer Taxodium distichum, the conifer Sequoia sempervirens, the conifer Araucaria cunninghamii, the conifer Pinus nigra var. sinica), and the caryophyllaceae (Malus domestica, Fragaria vesca, Arabidopsis lyrata, Capsella rubella, Brassica rapa, Thellungiella halophila, Carica papaya, Gossypium raimondii, and Zea mays) were included in the analysis. Because the predicted trans-cDNA for the M. truncatula genome are not available in the Phytozone database, the LegumIP database was used instead for this species (http://plantgrn.noble.org/LegumIP/blast.do). The cDNA, genomic DNA, and amino acid sequences corresponding to each PDAT or putative PDAT were downloaded from the Phytozone database. The theoretical molecular mass and pI values were calculated using the Compute pI/Mw tool provided in ExPASy (http://web.expasy.org/compute_pi/). For the InterPro domain analysis, all candidate sequences (without ending asterisk symbols) were scanned with InterProScan version 5 (Jones et al., 2014), which was installed locally in a 32-bit Red Hat Linux environment. The default parameters were used, and its InterPro lookup option (iprlookup) was turned on to generate InterPro annotation. For protein classification, all sequences were subjected to the Pfam (Punta et al., 2012; http://pfam.sanger.ac.uk/search) and PANTHER (Mi et al., 2013; http://www.pantherrdb.org/) classification systems. All taxa were indicated by three-letter acronyms in which the first letter is the first letter of the genus and the next two letters are the first two letters of the species name (e.g. Ath corresponds to Arabidopsis thaliana). Extra numbers were added after taxon names to indicate individual gene copies. To avoid confusion, the names for the previously reported Arabidopsis, castor bean, flax, and C. reinhardtii PDATs followed the published names.

Construction of Phylogenetic Trees

The full-length nucleotide sequences were aligned based on their corresponding amino acid translations using the TranslatorX server (Abascal et al., 2010; http://translatorx.co.uk/). Then, ModelTest 0.1.1 analysis (Posada, 2008) was carried out to select the best-fit model under the Akaike Information Criterion framework (Akaike, 1974). The result of ModelTest indicates that the best-fit substitution model to determine the evolution for all data sets is the General Time Reversible model with the shape of the gamma distribution plus the proportion of invariable sites. According to the best-fit model, ML phylogenetic analysis was constructed via the CIPRES Web Portal http://www.phylo.org using MrBayes 3.2.2 (Huelsenbeck and Ronquist, 2001) with 1,000,000 generations, four Markov chains, and two runs. The first 25% of the tree from all runs was discarded as burn-in. To verify the reliability of phylogenetic analysis, an ML tree was also performed using the online program RAxML (Stamatakis, 2006; http://www.tres.uqam.ca/index.php?action=raml&project=trex) under the best-fit model with 100 bootstrap samples. The phylogenetic tree was visualized using FigTree (http://tree.bio.ed.ac.uk/software/figtree/). The same methods were used to carry out all phylogenetic analyses included in this study. Sequence alignments for all analyses used in the phylogenetic construction are provided as Supplemental Data S1 to S3.

Gene Structure Analysis

Intron/exon distribution and intron phase patterns were analyzed using the online Gene Structure Display Server (Guo et al., 2007; http://gsds.cbi.pku.edu.cn).

Detection of Transmembrane Domains and Conserved Motifs

The potential transmembrane domains in PDATs were predicted using the TMHMM (Krogh et al., 2001) program provided by the CBS Prediction Servers (http://www.cbs.dtu.dk/services/TMHMM-2.0/). Functional motifs of PDAT proteins were identified using the MEME program (Bailey and Elkan, 1994; http://meme.nbcr.net/meme/cgi-bin/meme.cgi) with the following parameters: distribution of motifs = any number of repetitions, maximum number of motifs = 100, and optimum motif width = three to 300 residues. The identified motifs were further subjected to Pfam analysis for protein classification (Punta et al., 2012; http://pfam.sanger.ac.uk/search).

Gene Expansion Pattern and Selective Pressure Analysis

Tandem duplication was identified as multiple gene family members clustering within a 200-kb region of a chromosome (Holub, 2001). The chromosomeal locations of PDAT genes were determined using Phytozome's GBrowse genome browser.

For calculating Ks amino acid sequences representing the duplicated PDATs were aligned using ClustalW (Thompson et al., 1994) implemented in Geneious Pro 5.3.6 (Drummond et al., 2013), and the obtained protein alignments were used to guide the conversion of the corresponding cdna sequences into the cdnoal alignments via PAL2NAL (Suyama et al., 2006; http://www.bork.embl.de/pal2nal/). The resulting cdnoal alignments were imported into the cdnoal substitution model (CodeML) implemented in the PAML program 4.4c software package (Yang, 1997) for Ks calculation. The Goldman and Yang ML method and the F3x4 model were used in the analyses.

The selective pressure operating on the core eudicot PDATs was estimated using the ratio (ω) of the nonsynonymous substitution rate versus the Ks, as an indicator (Yang and Bielawski, 2000; Anisimova and Kosiol, 2009): 0 < ω < 1 corresponds to purifying selection, ω = 1 indicates neutral selection, and ω > 1 suggests positive selection. The estimation of ω ratio was performed using the CodeML program within the PAML package. Nucleotide alignments were generated using TranslatorX. Phylogenetic analyses of the core eudicot PDATs were performed using RAxML, and the resulting trees without branch lengths were used as input trees for the simple one-ratio model (model = 0 and nonsynonymous site = 0) analyses. The trees with branch lengths generated by one-ratio model analyses were further used to investigate functional divergence. To test divergent selective pressures among the core eudicot clades, we used the CmC (model = 3 and nonsynonymous site = 2) of Bielawski and Yang
(2004) as modified by Yang et al. (2005). The CmC assumes that the phylogeny can be divided into foreground and background partitions. For each analysis, the clade of interest (all branches within the clade) was selected as the foreground partition and the remaining phylogeny was set as the background partition. The CmC contains three site classes across the entire phylogeny: Site class 0 is under purifying selection (0 < ω0 < 1), site class 1 is under neutral selection (ω1 = 1), and site class 2 is the divergent site class where independent ω is estimated to the background (ω0 > 0) and foreground (ω0 > 0) partitions. The null model M2a_rel hypothesizing the same ω across the foreground and background partitions also has three site classes. The first two site classes are the same as the ones in CmC, while the third site class is represented by a single ω ratio for all branches across the phylogeny (ω0 > 0). LRTs were used to compare the fit of the CmC against the null model M2a_rel (Weadick and Chung, 2012). LRTs were performed by comparing twice the difference in ln-likelihood scores of CmC and M2a_rel against a χ2 distribution with the degree of freedom equal to the difference in the number of parameters between the two models. The data set was run multiple times with different initial ω values to avoid local optima. Like CmC, the extended clade model (Yoshida et al., 2011) also assumes three site classes. The null model M2a_rel hypothesizing the same ω across the foreground and background partitions also has three site classes. The first two site classes are the same as the ones in CmC, while the third site class is represented by a single ω ratio for all branches across the phylogeny (ω0 > 0). LRTs were used to compare the fit of the CmC against the null model M2a_rel.

Accession numbers and gene identifiers for sequences used in this study are provided in Supplemental Table S1.

**Supplemental Data**

The following supplemental materials are available.

**Supplemental Figure S1.** Phylogenetic relationship of 128 LCAT-like sequences from 38 plant species.

**Supplemental Figure S2.** Phylogenetic relationship of core-eudicot PDATs.

**Supplemental Figure S3.** Schematic diagram of gene structures of 86 plant PDATs.

**Supplemental Figure S4.** Alignment of TMDs in 86 plant PDATs.

**Supplemental Figure S5.** Alignment of 86 plant PDAT polypeptides with the human LCAT.

**Supplemental Figure S6.** Sequence logos derived from the MEME analysis.

**Supplemental Figure S7.** Combined motif diagram of 86 PDAT polypeptides.

**Supplemental Figure S8.** Phylogenetic tree used for selection pressure analyses.

**Supplemental Figure S9.** RNA-Seq digital gene expression analysis of PDAT paralogs in soybean.

**Supplemental Figure S10.** RNA-Seq digital gene expression analysis of PDAT paralogs in common bean.

**Supplemental Figure S11.** Gene expression profile of PDAT paralogs in Arabidopsis.

**Supplemental Figure S12.** Gene expression profile of PDAT(s) in maize and rice.

**Supplemental Table S1.** Gene identifiers of protein sequences used in this study.

**Supplemental Table S2.** Gene identifiers from Phyre2 database for genes with annotation errors.

**Supplemental Table S3.** InterPro domain and PANTHER classification analysis of 128 LCAT-like protein sequences.

**Supplemental Table S4.** PDAT protein properties and their clade distributions.

**Supplemental Table S5.** Sequence identity and synonymous substitution rates of between-clade duplicated PDAT gene pairs.

**Supplemental Data S1.** Alignment of 128 full-length LCAT-like nucleotide sequences.

**Supplemental Data S2.** Alignment of 87 full-length PDAT nucleotide sequences.

**Supplemental Data S3.** Alignment of 71 full-length nucleotide sequences of core eudicot PDATs and ScPDAT.

Received November 13, 2014; accepted January 10, 2015; published January 13, 2015.

**LITERATURE CITED**

Abascal F, Zarza R, Telford MJ (2010) TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. Nucleic Acids Res 38: W7–W11

Akaike H (1974) New look at the statistical model identification. IEEE Trans Automat Contr 19: 716–723

Alendé N, Nielsen JE, Shields DC, Kahl D (2011) Evolution of the isolectric point of mammalian proteins as a consequence of indels and adaptive evolution. Proteins 79: 1635–1648

Anisimova M, Kosiol C (2009) Investigating protein-coding sequence evolution with probabilistic codon substitution models. Mol Biol Evol 26: 255–271

Arabolaza A, Rodríguez E, Altabe S, Alvarez H, Gramajo H (2008) Multiple pathways for triacylglycerol biosynthesis in Streptomyces coelicolor. Appl Environ Microbiol 74: 2573–2582

Bailey TL, Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol 2: 28–36

Banana A, Carlsso AS, Huang B, Lenman M, Banas W, Lee M, Noiriel A, Benveniste P, Schaller H, Boivier-Navé P, et al. (2005) Cellular sterol ester synthesis in plants is performed by an enzyme (phospholipid:sterol acyl transferase) different from the yeast and mammalian acyl-CoA:sterol acyltransferases. J Biol Chem 280: 34626–34634

Bielański JP, Yang Z (2004) A maximum likelihood method for detecting functional divergence at individual codon sites, with application to gene family evolution. J Mol Evol 59: 121–132

Cao H (2011) Structure-function analysis of diacylglycerol acyltransferase sequences from 70 organisms. BMC Res Notes 4: 249

Cases S, Smith SJ, Zheng YW, Myers HM, Lear SR, Sande E, Novak S, Collins C, Welch CB, Luisi AJ, et al. (1998) Identification of a gene encoding an acyl CoA:diacylglycerol acyltransferase, a key enzyme in triacylglycerol synthesis. Proc Natl Acad Sci USA 95: 13018–13023

Cheng D, Meegalla RL, He B, Cromley DA, Bilheimer JT, Young PR (2001) Human acyl-CoA:diacylglycerol acyltransferase is a tetrameric protein. Biochem J 359: 707–714

Dahlqvist A, Stålh U, Lemman M, Banas A, Lee M, Sandager L, Ronne H, Stymne S (2000) Phospholipid:diacylglycerol acyltransferase: an enzyme that catalyzes the acyl-CoA-independent formation of triacylglycerol in yeast and plants. Proc Natl Acad Sci USA 97: 6487–6492

Drummond A, Ashton B, Cheung M, Heled J, Keane M, Moir R, Stones-Havas S, Thierer T, Wilson A (2013) Geneious v5.3.6. http://www.geneious.com/ (November 29, 2013)

Durrett TP, McClosky DD, Tumanewk AW, Elzinga DA, Ohrbroge J, Pollard M (2010) A distinct DGAT with ser-3 acetyltransferase activity that synthesizes unusual, reduced-venosity oils in E. grandiflorus and transgenic seeds. Proc Natl Acad Sci USA 107: 9464–9469

Fan J, Yan C, Rosston R, Shanklin J, Xu C (2014) Arachidonic lipids, PDAT1 acyltransferase, and SDP1: Three glycerolipid lipases synergistically direct fatty acids toward β-oxidation, thereby maintaining membrane lipid homeostasis. Plant Cell 26: 4119–4134

Fan J, Yan C, Zhang X, Xu C (2013) Dual role for phospholipid:diacylglycerol acyltransferase: enhancing fatty acid synthesis and diverting fatty acids from membrane lipids to triacylglycerol in Arabidopsis leaves. Plant Cell 25: 3510–3528

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlewait J (1999) Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151: 1531–1545

Ghosal A, Banas A, Stålh U, Dahlqvist A, Lindqvist Y, Stymne S (2007) Saccharomyces cerevisiae phospholipid:diacylglycerol acyl transferase

902

Pan et al.

Supplemental Data S1. Alignment of 128 full-length LCAT-like nucleotide sequences.

Supplemental Data S2. Alignment of 87 full-length PDAT nucleotide sequences.

Supplemental Data S3. Alignment of 71 full-length nucleotide sequences of core eudicot PDATs and ScPDAT.
(PDAT) devoid of its membrane anchor region is a soluble and active enzyme retaining its substrate specificities. Biochim Biophys Acta 1771: 1457–1463
Guo AY, Zhu QH, Chen X, Luo JC (2007) [GSDS: a gene structure display server]. Yi Chuan 29: 1023–1026
Heath TA, Hedtke SM, Hillis DM (2008) Taxon sampling and the accuracy of phylogenetic analyses. J Syst Evol 46: 239–257
Hobbs DH, Lu C, Hills MJ (1999) Cloning of a cDNA encoding diacyglycerol acyltransferase from Arabidopsis thaliana and its functional expression. FEBS Lett 452: 145–149
Holub EB (2001) The arms race is ancient history in Arabidopsis, the wildflower. Nat Rev Genet 2: 516–527
Hulsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17: 754–755
Hughes AL (1997) Gene trees in species trees. Syst Biol 46: 290–303
Hughes AL, Verschelde H, Vanloo B, Ampe C, Labeur C, Tavernier J, Vandenckervke J, Rossenve M (1999) Effects of natural mutations in lecithin:cholesterol acyltransferase on the enzyme structure and activity. J Lipid Res 40: 59–69
Peelman F, Vinaintom N, Verhee A, Vanlou B, Verschelde JI, Labeur C, Sungle-Sace M, Duverger N, Hutchinson G, Vandenckervke J, et al (1998) A proposed architecture for lecithin cholesterol acyltransferase (LCAT): identification of the catalytic triad and molecular modeling. Protein Sci 7: 587–599
Pelham HRB (2000) Using sorting signals to retain proteins in endoplasmic reticulum. Methods Enzymol 327: 279–283
Posada D (2008) jModelTest: phylogenetic model averaging. Mol Biol Evol 25: 1253–1256
Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Eddy SR, et al (2012) The Pfam protein families database. Nucleic Acids Res 40: D296–D302
Rani SH, Krishna THA, Saha S, Negi AS, Rajasekharan R (2010) Defective in cuticular ridges (DCR) of Arabidopsis thaliana, a gene associated with surface cutin formation, encodes a soluble diacylglycerol acyltransferase. J Biol Chem 285: 38337–38347
Roth C, Rastogi S, Arvestad L, Dittmar K, Light S, Ekman D, Liberles DA (2007) Evolution after gene duplication: models, mechanisms, systems, and organisms. J Exp Zoolog B Mol Dev Evol 308: 58–73
Saha S, Enugutti B, Rajakumar S, Rajasekharan R (2006) Cytosolic triacylglycerol biosynthetic pathway in oilseeds: molecular cloning and expression of peanut cytosolic diacylglycerol acyltransferase. Plant Physiol 141: 1533–1543
St Dahl U, Carlsson AS, Lenman M, Dahlqvist A, Huang B, Banas W, Banas A, Stymne S (2004) Cloning and functional characterization of a phospholipid:diacylglycerol acyltransferase from Arabidopsis. Plant Physiol 135: 1324–1335
Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22: 2688–2690
Suyama M, Torrents D, Bork P (2006) PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res 34: W609–W612
Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673–4680
Torgerson DG, Singh RS (2004) Rapid evolution through gene duplication and subfunctionalization of the testes-specific a4 proteasome subunits in Drosophila. Genetics 168: 1421–1432
Turchetto-Zolet AC, Maraschin FS, de Morais GL, Caglioni A, Andrade CMB, Margis-Pinheiro M, Margis R (2011) Evolutionary view of acyl-CoA diacylglycerol acyltransferase (DGAT), a key enzyme in neutral lipid biosynthesis. BMC Evol Biol 11: 263
van Erp H, Bates PD, Burgal J, Shockey J, Browse J (2011) Castor phospholipid:diacylglycerol acyltransferase facilitates efficient metabolism of hydroxy fatty acids in transgenic Arabidopsis. Plant Physiol 155: 683–693
Wang Y, Wang X, Paterson AH (2012) Genome and gene duplications and gene expression divergence: a view from plants. Annu N Y Acad Sci 1256: 1–14
Weadick CJ, Chang BSW (2012) An improved likelihood ratio test for detecting site-specific functional divergence among clades of protein-coding genes. Mol Biol Evol 29: 1297–1301
Weiss SB, Kennedy EP, Kitasov JY (1960) The enzymatic synthesis of triglycerides. J Biol Chem 235: 40–44
Weselake RJ, Madhavji M, Szarka SJ, Patterson NA, Wiehler WB, Nyikoforuk CL, Burton TL, Boora PS, Mosimann SC, Foroud NA, et al (2006) Acyl-CoA-binding and self-associating properties of a recombiant 133
kDa N-terminal fragment of diacylglycerol acyltransferase-1 from oilseed rape.

BMC Biochem 7: 24

Worberg A, Quandt D, Barniske A, Löhne C, Hilu KW, Borsch T (2007) Phylogeny of basal eudicots: insights from non-coding and rapidly evolving DNA. Org Divers Evol 7: 55–77

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

Yang Z, Bielawski JP (2000) Statistical methods for detecting molecular adaptation. Trends Ecol Evol 15: 496–503

Yang Z, Wong WSW, Nielsen R (2005) Bayes empirical Bayes inference of amino acid sites under positive selection. Mol Biol Evol 22: 1107–1118

Yoon K, Han D, Li Y, Sommerfeld M, Hu Q (2012) Phospholipid:diacylglycerol acyltransferase is a multifunctional enzyme involved in membrane lipid turnover and degradation while synthesizing triacylglycerol in the unicellular green microalga Chlamydomonas reinhardtii. Plant Cell 24: 3708–3724

Yoshida I, Sugiura W, Shibata J, Ren F, Yang Z, Tanaka H (2011) Change of positive selection pressure on HIV-1 envelope gene inferred by early and recent samples. PLoS ONE 6: e18630

Zhang M, Fan J, Taylor DC, Ohlrogge JB (2009) DGAT1 and PDAT1 acyltransferases have overlapping functions in Arabidopsis triacylglycerol biosynthesis and are essential for normal pollen and seed development. Plant Cell 21: 3885–3901