Running Head: Evolutionary Relationships of the *PDAT* Genes in Plants

To whom correspondence should be addressed: Randall J. Weselake, Department of Agricultural, Food and Nutritional Science, University of Alberta, Edmonton, Alberta, Canada T6G 2P5, Tel.: (780) 492-4401; E-mail: (randall.weselake@ualberta.ca)

Research Area: Genes, Development and Evolution
Genome-wide Analysis of PHOSPHOLIPID:DIACYLGLYCEROL
ACYLTRANSFERASE 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, Randall J. Weselake

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

Summary: Ancient gene duplication may have led to the diversification of a key acyltransferase of plant triacylglycerol synthesis in the core eudicots.
Footnotes:

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantphysiol.org) is: Randall J. Weselake (randall.weselake@ualberta.ca).

Address correspondence to email: randall.weselake@ualberta.ca;
Abstract

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 (TAG) and a lyso-phospholipid. Although the importance of PDAT in TAG 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 show that different selection constraints have acted on three core eudicot clades, which might enable paleo-duplicated PDAT paralogs to either become non-functionalized or develop divergent expression pattern 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.
Introduction

Plant oils with a broad variety of food and industrial applications 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 research community. In 1960, using chicken (Gallus gallus) liver as an enzyme source, Eugene P. Kennedy and his graduate student Samuel B. Weiss demonstrated the use of diacylglycerol (DAG) and fatty acyl-CoAs as substrates for TAG biosynthesis (Weiss et al., 1960). The enzyme catalyzing this reaction is named acyl-CoA:diacylglycerol acyltransferase (DGAT, EC 2.3.1.20). 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 and colleagues isolated the first DGAT gene in mouse (Mus musculus) by homology searches of the expressed sequence tag (EST) databases using the coding sequences from acyl-CoA:cholesterol acyltransferase (ACAT, EC 2.3.1.26), an enzyme that catalyzes the synthesis of cholesteryl esters (Cases et al., 1998). At least six types of DGAT genes (DGAT1, DGAT2, DGAT3, DCR, DAcT and a bifunctional DGAT/wax synthase) have been identified in plants (Saha et al., 2006; Hobbs et al., 1999; Lardizabal et al., 2001; Durrett et al., 2010; Rani et al., 2010; Li et al., 2008). The biochemistry, molecular biology, evolutionary relationship, structure-function relationship, and biotechnological applications of DGAT have been discussed in some recent publications (Liu et al., 2012; Cao, 2011; Turchetto-Zolet et al., 2011).

Recently, it has become obvious that TAG synthesis can also 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 (Rincinus communis), and Crepsis palaestina by Stymne and co-workers during their observation of the use of phospholipids as acyl donor and DAG as acceptor for TAG biosynthesis (Dahlqvist et al., 2000). They further found that PDAT activity is also present in yeast (Saccharomyces cerevisiae) and identified the first PDAT gene (YNR008w, LRO1) from yeast as a homolog of human lecithin:cholesterol 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 (RNAi)-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 (van Erp et al., 2011; Kim et al., 2011). One particular PDAT, *RcoPDAT1A*, appears to be ricinoleic-specific; seed-specific overexpressing 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/LusPDAT5 and LusPDAT2/LusPDAT4) have the unique ability to preferentially channel α-linolenic acid (ALA) 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 has also been detected in bacteria, *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 enzyme 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 PDAT1 (AthPDAT1) and PDAT2 (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 of land plant genomes, with the exception of Brachypodium distachyon. Only one hit was identified in each of algal genomes. In total, 139 sequences were identified and sequence information is provided in Supplemental Table 1. Among the 139 sequences, six from five species (Solanum tuberosum, Populus trichocarpa, Medicago truncatula, Malus domestica and Arabidopsis lyrata) encode less than 200 amino acid residues, which is most likely due to genome annotation errors (Supplemental Table 2). These short sequences were eliminated from further analysis. In addition, the predicted transcripts from Malus domestica genome have multiple stop codons and the predicted transcript from 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 3), 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) trees (Figure 1 and Supplemental Figure 1) 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, 3,
4 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 are previously experimentally characterized as PDAT, including Arabidopsis, flax and castor bean PDATs (Ståhl et al., 2004; Zhang et al., 2009; van Erp et al., 2011; Kim et al., 2011; Pan et al., 2013). The AthLCAT-like2 (AT1G04010) (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 previously predicted to be PLA while the AthLCAT-like1 sequence (AT1G27480) from Group D was classified as LCAT (Yoon et al. 2012).

Next, all protein sequences were classified using the InterPro and PANTHER (Protein Analysis Through Evolutionary Relationships) classification system. The results (Supplemental Table 3) 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 Group B and Group C are classified into PSAT subfamily (PTHR:11440:SF7) and LCAT-like 4-related subfamily (PTHR:11440:SF3), respectively. The sequences in Group D have no PANTHER subfamily classification.

The phylogeny result combined with the PANTHER classification suggests 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 as PDAT while the sequences from Groups B, C and D are named as 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 (*Solanum tuberosum, Populus trichocarpa, Medicago truncatula* and *Arabidopsis lyrata*) with short candidate PDATs 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 of closely related homologs (Supplemental Table 2).
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 *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 alga *C. reinhardtii* identified in this study is consistent with previous studies (van Erp et al., 2011; Pan et al., 2013; Yoon et al., 2012; Ståhl et al., 2004).

**Phylogenetic analysis divides plant PDATs into seven major clades**

To explore the evolutionary relationship of the plant PDAT gene family, we further constructed ML trees using 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 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%), the PDAT gene family can be classified into seven clades designated as Clade I to Clade VII (Figure 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 Clade II to Clade IV. The eudicots can be divided into the basal and the core eudicots (Worberg et al., 2007). As shown in the species tree (Table 1), the basal eudicot (represented as *Aquilegia coerulea*) forms a paraphyly at the base of 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 Table 1.

It is important to point out that taxa in the phylogenetic tree are very unevenly distributed among the clades, ranging from 1 to 28 sequences (Figure 2), which may have negative impact on phylogenetic accuracy (Heath et al., 2008). To exam the phylogenetic accuracy, we first
compared the phylogenetic tree (Figure 2) with the species tree shown in Phytozome (Table 1). As shown, 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 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 dataset and reconstructed the phylogenetic tree with the sequences from Clades V, VI and VII, which have more balanced taxa sampling. The trees generated from pruned and complete datasets are topologically identical (Supplemental Figure 2 and Figure 2), 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 multi-copy 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 (Page and Charleston, 1997; Maddison, 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 species tree as well as consistent phylogenetic topologies inferred from complete and pruned datasets.

*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 shown in Figure 3 and details
shown in Supplemental Figure 3). Diverse gene structure has been found in the algal PDATs (Clade I): MpuCMPPDAT has no intron, while the rest of the PDAT genes have 9 to 14 introns. By contrast, land plant PDATs (Clades II to 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, PvuPDAT1, GmaPDAT5 and GraPDAT1 with seven exons and six introns, and LusPDAT3 and 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; 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 (Figure 3 and Supplemental Figure 3).

**Evaluation of PDAT protein properties reveals that PDATs that belong to the core eudicot Clade VI had a tendency to maintain acidic isoelectric points (pI values) during evolution**

After evaluation of gene structure, we continued our analysis with a focus on protein properties of 86 PDATs, including protein length, molecular mass and pI values. According to our analyses (Table 2 and Supplemental Table 4), the length and molecular mass of PDATs from Clade I (algae clade) varied substantially. CrePDAT with 1041 amino acid residues and 104.5 kDa is the longest and largest PDAT, while CsuPDAT with 509 amino acid residues and 56.9 kDa 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 kDa to 80.3 kDa, with a mean of 671 amino acid residues and 74.7 kDa. For the pI values, MpuCMPPPDAT has the highest value of 9.53. Except MpuCMPPPDAT, 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 (>7) have been observed in 31 out of 42 PDATs belongs to Clades V and VII.

**The membrane topology of the PDAT proteins is well conserved among most land plants**
We next studied the membrane topology of plant PDATs. The putative transmembrane domains (TMDs) of 86 PDATs were predicted using the TMHMM program. To provide a better comparison of the TMDs among PDATs, the polypeptides with the annotated TMD regions were aligned using ClustalW. The results (examples shown in Figure 4A and details shown Supplemental Figure 4) show that two out of four algal PDATs (Clade I) have one putative TMD (CrePDAT and MpuCMPPDAT), while the other two contain no TMD. CrePDAT was previously predicted to be localized in chloroplasts (Yoon et al., 2012); therefore, MpuCMPPDAT might also be a chloroplast-localized protein. Because the endoplasmic reticulum (ER) is the major site for TAG biosynthesis in plants (Lung and Weselake, 2006), we assume that land plant PDATs are inserted into the ER and interpret the topology results based on the ER structure. The results (examples shown in Figure 4A and details shown Supplemental Figure 4) show that 73 out of 82 land plant PDATs (except BraPDAT2, CruPDAT2, GraPDAT2, GraPDAT3, LusPDAT3, LusPDAT6, RcoPDAT2, TcaPDAT3 and VviPDAT3) have a single putative TMD, with the short N-terminus facing the cytosol and the bulk of the C-terminus residing in the ER lumen. This result is consistent with the topology reported for yeast and Arabidopsis PDATs (Ghosal et al., 2007; Yoon et al., 2012). Our alignment results further indicate that the position of the TMD is highly preserved among land plant PDATs (Figure 4A and Supplemental Figure 4). Mapping the TMD region onto the genomic sequence reveals that the position of the TMD corresponds to the region within the first exon in 71 out of 73 TMD-containing PDATs (examples shown in Figure 3). The alignment of land plant PDATs only (Figure 4B) also shows that the hydrophilic N-terminal region preceding the TMD appears to be the most divergent region, which carries the only common feature: a cluster of consecutive arginine residues. Interestingly, the N-termini of DGAT1 is also the most variable region and carries the arginine cluster (Liu et al., 2012). The role of these conserved arginine residues remains unclear, but it has been speculated that they are potentially an ER-localization signal (Liu et al., 2012).

**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 conserved amino acid residues. The alignment shows that besides the initial methionine 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 C-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 the homology to human LCAT, which is a soluble protein with no TMD. It was previously reported that human LCAT contains several structurally conserved elements (Peelman et al., 1998; Peelman et al., 1999), including a catalytic triad of Ser181-His377-Asp345, a salt bridge between Asp145 and Arg147, and a so-called “lid region”. The Trp61 within the lid region was proposed to play an important role in binding the cleaved fatty acid into the active site for optimal acylation process. Aligning plant PDATs with human LCAT reveals (examples shown in Figure 5 and details shown Supplemental Figure 5) that the Trp61, Asp145, Arg147, Ser181, and Asp345 of LCAT are completely conserved in 86 PDATs, while His377 is conserved in 83 out of 86 PDATs, with the exception of GmaPDAT5, GmaPDAT6 and PvuPDAT1. Since the available EST sequences of GmaPDAT5, GmaPDAT6 and PvuPDAT1 were not long enough to cover the coding regions for His377, 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 alpha/beta 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 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 6. More detailed information is provided in Supplemental Figures 6 and 7. 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 first three sequentially arranged and the fourth one separated. Six major LCAT motif patterns (Figure 6 and Supplemental Figure 7) have been identified based on the MEME combined block diagram. These patterns include: 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, 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 (Figure 6): 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 the 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 the motif 38 in Clade IV and the motif 27 in Clade VII, were found. It is worth mentioning that individual MEME motifs are gapless (no insertions or deletions), which means that the motifs containing gaps can still 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 cladding pattern in the phylogenetic tree.

Eudicot-wide PDAT gene expansion arose 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 (Figure 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 as “paleo-duplicated genes”. Due to the fact that the ancient gene duplication was followed by specie-specific gene duplication, gene loss and chromosome rearrangements, only six species (Vitis vinifera, Theobroma cacao, R. communis, Carica papaya, Eucalyptus grandis, and Phaseolus vulgaris) have maintained the triplicated “paleologous” PDAT genes, each of which is present in one of the three core eudicot clades (Clades V to VII) (Figure 2 and Table 1).

Besides paleo-duplicated 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, Glycine max and Gossypium raimondii, two in Solanum lycopersicum and Manihot esculenta, one in Mimulus guttatus and Brassica rapa (Figure 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 genome has a unique evolutionary history. The lineage-specific whole genome multiplication event(s) has occurred approximately 60 million years ago in Gossypium genomes (Paterson et al., 2012).

The within-clade gene pairs tend to have higher sequence identity than the between-clade gene pairs (Tables 3 and 4). 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 (Ks) as a proxy for time to compare the date of gene duplications. Judging from the Ks values, the within-clade gene pairs have a much lower Ks than between-clade gene pairs,
suggesting more recent duplications (Tables 3 and 4). Because most of the Ks values for
between-clade gene pairs are saturated (>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 200kb can be viewed as a gene cluster (Houb, 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 species of Citrus clementine (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 specie-specific segmental duplication
primarily contributes to the expansion of the PDAT gene family in eudicots.

Different selection pressures have acted on the paleo-duplicated PDAT paralogs

It has long been thought that gene duplication plays a crucial role in the evolution of gene
diversity (Ohno, 1970; Roth et al., 2007; Hughes, 1994). To explore the evolutionary fate of
paleo-duplicated eudicot PDATs, we performed selection pressure analyses. 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 SmoPDAT1 as an outgroup (Supplemental Figure 8).

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 (Figure2 and Supplemental Figure 8), 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, Clade VI along with the outgroup comprised the background
partition. We call this analysis as “CmC VII”. The CmC model 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 5). Parameter estimates indicate that a larger set of sites (~56%) evolving under stronger purifying selection (ω₀ =0.02764), and a smaller set of sites (~43%) evolving 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 significant better fit than the null model (Table 5). 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 ω₃=0.14847; Clade VII ω₃=0.23272) is close to the ω ratio for Clade V (ω₃=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: Clade V, Clade VI and Clade VII. We call this analysis as “Ex-CmC”. The null hypothesis is that selection pressure is the same for Clade V and Clade 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 as “Ex-Null”. The LRT result (Table 5) indicates that the null hypothesis is rejected, supporting that different selection pressures have acted on Clade V and Clade VI. To further confirm our result, we excluded the Clade VII from the analysis and only compared the functional divergence between Clade V and Clade VI. Using this dataset, we found that setting either Clade V or Clade VI as the foreground partition yields a significant LRT result (P<0.001, Table 6), indicating that selective constraint indeed differs between Clade V and Clade VI.

Taken together, these results indicate that three core eudicot clades have evolved under divergent selection pressures and PDATs in Clade VII experienced the lowest selection constraint compared to PDATs from the other two clades.

Discussion

Despite the fact that many studies have revealed the crucial role of PDATs in TAG biosynthesis, our knowledge on 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 have surveyed 40 different plant species and identified 139 LCAT-like sequences. The results of phylogeny (Figure 1) and PANTHER classification analyses (Supplemental Table 3), along with the previous findings indicated that only the sequences with E-value $< 1e-15$ from land plants and the 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 (Figure 2).

Four algal *PDAT*s form a separate well-supported clade (Clade I) from land plant *PDAT*s. This phylogenetic separation is supported by their different gene structure (Figure 3), protein properties (Table 2) and motif composition (Figure 6). The observed differences between algal and land pant 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, which has not only PDAT and DAG:DAG acyltransferase function, but also galactolipid:DAG acyltransferase 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 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 (Figure 3 and Supplemental Figure 3), 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 (Figure 4 and Supplemental Figure 4). Third, all LCAT-like motifs (Figure 6 and Supplemental Figure 7) and LCAT-conserved amino acid residues (Figure 5 and Supplemental Figure 5) are located at the C-terminal end of the TMD, suggesting that the active and/or binding sites of land plant PDATs are possibly facing the luminal side of the ER. Fourthly, the C-terminal portion and the region between the TMD and the first LCAT-like motif are highly conserved (Figure 6 and Supplemental Figure 7). 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-terminal 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 (Figure 4B). The motif occurrences in the N-terminus are quite unique for PDATs from each clade (Figure 6 and Supplemental Figure 7). 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 (McFie et al., 2010; Weselake et al., 2006; Pelham, 2000; Cheng et al., 2001).

Our analysis also reveals a eudicot-wide PDAT gene expansion. Combined with evidence from the phylogenetic (Figure 2) and Ks analyses (Tables 3 and 4), 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, VI and 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 (Figure 6 and Supplemental Figure 7). A tendency for the conservation of acidic pI values in Clade VI (Table 2 and Supplemental Table 4) 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 paleo-duplicated 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.

*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, which show evidence of diverging TAG-synthesizing function. Our study showed that there is a eudicot-wide PDAT gene expansion, but the questions remain: “why do eudicots contain multiple copies of PDATs in their genome?” and “Whether gaining functional divergence among PDAT paralogs is a general trend in the evolution of the eudicot PDATs and how this may happen?” Now, we may be able answer these questions from an evolutionary perspective.

Gene duplication is believed to be one of the major driving forces for the evolutionary novelties, including neo-functionalization (Ohno, 1970; Force et al., 1999; Roth et al., 2007) and sub-functionalization (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, paleo-duplicated PDAT genes have been subjected to different selective constraints (Tables 5 and 6).

The observed heterogeneity in selection pressure among the core eudicot clades might enable the changes in the genes’ functions and/or the development of expression-level divergence among
duplicated genes. Consistent with this hypothesis, the previous studies (Ståhl et al., 2004; Zhang et al., 2009; van Erp et al., 2011; Kim et al., 2011; Pan et al., 2013) showed that PDATs, including Arabidopsis, flax and castor bean PDATs, from Clade VII (AthPDAT2, LusPDAT3, LusPDAT6, RcoPDAT2) do not have an apparent function in TAG biosynthesis. Clade VII under the weakest selection constraint (Table 5) 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 non-functionalization of Clade VII, previous studies (Ståhl et al., 2004; Zhang et al., 2009; van Erp et al., 2011; Kim 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 (Clade V and Clade VI). More interestingly, many PDATs in Clade VI appear to have been sub-functionalized at the expression level into the non-seed tissues. Studies on flax (Pan et al., 2013) and castor bean PDATs (Kim et al., 2011) revealed that the PDAT paralogs sitting on Clade VI (LusPDAT2, LusPDAT4 and RcoPDAT1B) are expressed at a high level in non-seed tissues and very low expression levels are detected in seeds. In this study, we also used the RNA-Seq data to examine the expression profile of PDATs in soybean (G. max) and common bean (P. vulgaris). Consistent with the hypothesis, soybean and common bean PDAT paralogs in Clade VI (GmaPDAT1, GmaPDAT2 and PvuPDAT3) have significantly higher expression in leaves, flowers and roots than in developing seeds (Supplemental Figure 9 and 10). Although similar expression levels of Arabidopsis PDAT1 (from Clade VI) in leaves, roots, flowers and developing seeds were reported, a semi-quantitative RT-PCR approach with seeds at single developmental stage (mid stage) was used in the study (Ståhl et al., 2004). To obtain a more detailed expression data, we extracted the microarray expression data for AthPDAT1 from the AtGenExpress database (http://jsp.weigelworld.org/expviz/expviz.jsp). The result (Supplemental Figure 11) 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 only been detected in rapidly developing leaves rather than in seeds under both over-expression and RNAi approaches (Mhaske et al., 2005; Fan et al., 2013). A fairly recent study (Fan et al., 2014) indicated that the AthPDAT1-mediated TAG synthesis is involved in the process of diverting fatty acids from membrane lipids towards peroxisomal beta-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, the studies on mammalian proteins suggest that the shifts in the pI values may be due to the functional divergence of proteins (Khaldi 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 the tissue-subfunctionalization. At the same time, it must be noted that three PDATs from the third core eudicot clade (Clade V) characterized to date, including LusPDAT1, LusPDAT5 and RcoPDAT1A, have seed-specific expression patterns and the encoded enzymes have unique substrate selectivity properties (Pan et al., 2013; Kim et al., 2011). The identified ALA-preferring flax PDATs (Pan et al., 2013) and hydroxy fatty acid-selective castor PDAT (van Erp et al., 2011) support the speculation that: 1) the substrate selectivity of PDAT has co-evolved 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, lycophyte and monocots form monophyletic clades. Two PDAT paralogs from maize (Z. mays) have similar expression pattern (Supplemental Figure 12). The PDAT expression profile in rice (Oryza sativa) did not show any significantly differential expression among tissues (Supplemental Figure 12). 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 paleo-duplicated PDAT paralogs to become non-functionalized and another two paralogs to develop divergent expression pattern. Since it has been well recognized that gain of functional diversification and expression-level divergence is a key process in promoting retention of duplicated genes in the genome (Lynch and Force, 2000; Li et al., 2005; Torgerson and Singh, 2004; Roth et al., 2007), non-functionalization 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 paleo-duplicated 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 non-functionalization and expression divergence of paleo-duplicated 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 (van Erp et al., 2011; Pan et al., 2013; Yoon et al., 2012; Ståhl et al., 2004). 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 the TBLASTN search using the Arabidopsis PDAT1 (AthPDAT1) and PDAT2 (AthPDAT2) protein sequences as queries against the Phytozome databases (http://www.phytozome.net/). All 39 listed plant species, including algae (Chlamydomonas reinhardtii, Volvox carteri, Coccomyxa subellipsoidea C-169, Micromonas pusilla CCMP1545 and Ostreococcus lucimarinus), a lycophyte Selaginella moellendorffii, a moss Physcomitrella patens, monocots (Sorghum bicolor v1.4, Zea mays, Setaria italica, Panicum virgatum, Oryza sativa and Brachypodium distachyon), and eudicots (Manihot esculenta, Ricinus communis, Linum usitatissimum, Populus trichocarpa, Medicago truncatula, Phaseolus vulgaris, Glycine max, Cucumis sativus, Prunus persica, Malus domestica, Fragaria vesca, Arabidopsis lyrata, Capsella rubella, Brassica rapa, Thellungiella halophila, Carica papaya, Gossypium raimondii, Theobroma cacao, Citrus sinensis, Citrus clementina, Eucalyptus grandis, Vitis vinifera, Solanum tuberosum, Solanum lycopersicum, Mimulus guttatus
v1.1 and Aquilegia coerulea Goldsmith), were included in the analysis. Because the predicted transcripts for Medicago truncatula genome are not available in the Phytozome database, the LegumeIP database was used instead for this species (http://plantgrn.noble.org/LegumeIP/blast.do). The cDNA, genomic DNA and amino acid sequences corresponding to each PDAT or putative PDAT were downloaded from the Phytozome database. The theoretical molecular mass \( (m) \) and isoelectric point \( (pI) \) 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 Pfam (Punta et al., 2012; http://pfam.sanger.ac.uk/search) and PANTHER classification systems (Mi et al., 2013; http://www.pantherdb.org/). 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 the confusion, the names for the previously reported Arabidopsis (A. thaliana), castor bean (R. communis), flax (L. usitatissimum) and alga (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 TranslatorX server (Abascal et al., 2010; http://translatorx.co.uk/). Then, jModelTest 0.1.1 analysis (Posada, 2008) was carried out to select the best-fit model under the Akaike Information Criterion (AIC) framework (Akaike, 1974). The result of jModelTest indicates that the best-fit substitution model to determine the evolution for all data sets is GTR with the shape of the gamma distribution plus the proportion of invariable sites (G+I). According to the best-fit model, maximum likelihood (ML) phylogenetic analysis was constructed via the CIPRES Web Portal http://www.phylo.org/portal/Home using MrBayes 3.2.2 (Huelsenbeck and Ronquist, 2001) with 1,000,000 generations, four Markov chains and two runs. The first 25% of tree from all runs were discarded as burn-in. To verify the reliability of phylogenetic analysis, a
maximum likelihood tree was also performed using online program RAxML (Randomized Axelerated Maximum Likelihood) (Stamatakis, 2006; http://www.trex.uqam.ca/index.php?action=raxml&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 Datasets 1-3.

**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 (Multiple Expectation Maximization for Motif Elicitation) program (Bailey and Elkan, 1994; http://meme.nbcr.net/meme/cgi-bin/meme.cgi) with the following parameters: the distribution of motifs = any number of repetitions, maximum number of motifs = 100, and optimum motif width = 3 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 200kb region of a chromosome (Houb, 2001). The chromosomal locations of PDAT genes were determined using the Phytozome’s GBrowse genome browser.

For calculating the synonymous substitution rates ($K_s$), 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 codon alignments via PAL2NAL (Suyama et al., 2006; http://www.bork.embl.de/pal2nal/). The resulting codon
alignments were imported into the codon substitution model (CodeML) implemented in the PAML v4.4c software package (Yang, 1997) for $K_s$ calculation. The Goldman and Yang maximum likelihood method and the F3x4 model were used in analyses.

The selective pressure operating on the core eudicot $PDATs$ was estimated using the ratio ($\omega$) of the nonsynonymous substitution rate ($K_\text{N}$) versus the synonymous substitution rate ($K_\text{S}$) as an indicator (Anisimova and Kosiol, 2009; Yang and Bielawski, 2000): $0 < \omega < 1$ corresponds to purifying selection, $\omega = 1$ indicates neutral selection, and $\omega > 1$ suggests positive selection. The estimation of the $\omega$ ratio was performed using the CodeML program within the PAML package.

Nucleotide alignments were generated using the 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 ($M_0$, model = 0 and NS site = 0) analyses. The trees with branch lengths generated by $M_0$ analyses were further used to investigate functional divergence. To test divergent selective pressures among the core eudicot clades, we used the Clade model C ($CmC$, model = 3 and NS site = 2) of Bielawski and Yang (Bielawski and Yang, 2004) as modified by Yang, Wong, and Nielsen (Yang et al., 2005). The $CmC$ assumes that the phylogeny can be divided into the 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 ($0 < \omega_0 < 1$); site class 1 is under neutral selection ($\omega_1 = 1$); site class 2 is the divergent site class where independent $\omega$ is estimated to the background ($\omega_2 > 0$) and foreground partitions ($\omega_3 > 0$). The null model $M_2a_{\text{rel}}$ hypothesizing the same $\omega$ between the foreground and background partitions also has three site classes. The first two site classes are the same as the ones in $CmC$, while a third site class is represented by a single $\omega$ ratio for all branches across the phylogeny ($\omega_2 > 0$). Likelihood ratio tests (LRTs) were used to compare the fit of the $CmC$ against the null model $M_2a_{\text{rel}}$ (Weadick and Chang, 2012). LRTs were performed by comparing twice the difference in $ln$ likelihood scores of $CmC$ and $M_2a_{\text{rel}}$ against a $\chi^2$ 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 $\omega$ values to avoid local optima. Like $CmC$, extended clade model (Yoshida et al., 2011) also assumes three site classes. The first two site classes are
the same as CmC model. The final class (site class 2) allows to model divergent selection for
more than two phylogeny partitions, each with a separately estimated ratio. We specified three
partitions in our analysis (Clade V, Clade VI and Clade VII) and three separate $\omega$ ratios were
obtained for the three tree partitions ($\omega_2$ for Clade V, $\omega_3$ for Clade VI, and $\omega_4$ for Clade VII). The
null model has only two partitions, with $\omega_2$ for Clade VII and $\omega_3$ for both Clade V and Clade VI.
LRTs were used to compare the fit of the extended clade model against the null model.

Acknowledgments

RJW gratefully acknowledges the support provided by Genome Alberta/Genome Canada, Genome
Prairie, the Canada Research Chairs Program, the Canada Foundation for Innovation, the Research
Capacity Program of Alberta Enterprise and Advanced Education, AVAC Ltd., Alberta Innovates Bio
Solutions and the Natural Sciences and Engineering Research Council of Canada.

References


AC-19: 716-723

Alendé N, Nielsen JE, Shields DC, Khaldi N (2011) Evolution of the isoelectric point of
mammalian proteins as a consequence of indels and adaptive evolution. Proteins: Struct, Funct,
Bioinf 79: 1635-1648

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


Bailey TL, Elkan C (1994) Fitting a mixture model by expectation maximization to discover
motifs in biopolymers. Proceedings /International Conference on Intelligent Systems for
Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular
Biology 2: 28-36


region is a soluble and active enzyme retaining its substrate specificities. Biochim Biophys Acta 1771: 1457-1463


Khaldi N, Shields DC (2011) Shift in the isoelectric-point of milk proteins as a consequence of adaptive divergence between the milks of mammalian species. Biology Direct 6: 40-40


McCartney AW, Dyer JM, Dhanoa PK, Kim PK, Andrews DW, McNew JA, Mullen RT (2004) Membrane-bound fatty acid desaturases are inserted co-translationally into the ER and contain different ER retrieval motifs at their carboxy termini. Plant J 37: 156-173


(Linum usitatissimum L.) seed catalyzing the selective production of trilinolenin. J Biol Chem 288: 24173-24188


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
**Figure Legends**

Figure 1. Cladogram of 128 LCAT-like sequences from 38 plant species. The maximum likelihood tree was generated using the MrBayes program, and the support values along branches are Bayesian posterior probabilities (shown as percentage). The land plant LCAT-like sequences are grouped into four major clades, designated as Groups A, B, C and D. The scale bar denotes the number of nucleotide replacements per site. More detailed data are presented in Supplemental Figure 1.

Figure 2. Phylogenetic relationship of the PDAT gene family. The maximum likelihood tree was generated using the MrBayes program. PDATs are grouped into 7 distinct clades (I to 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.

Figure 3. Schematic diagram of gene structures of representative PDATs. The thin lines represent introns and thick bars represent exons. The numbers above the gene structure indicate intron phases. A scale bar with a unit of base pair (bp) is graphed on the top. TMD denotes transmembrane domain. Gene identifiers and abbreviations of listed PDATs can be found in the Supplemental Table 1.

Figure 4. (A) Alignment of transmembrane domains (TMDs) in representative plant PDATs. The putative TMDs are annotated as red arrows. The alignment was generated using ClustalW implemented in Geneious software and represented as thick lines (aligned characters) and thin lines (gaps). Overall alignment identity of 86 final selected plant PDATs and a scale bar indicating the numbers of amino acid residues are graphed on the top. (B) Overall alignment identity of 80 land plant PDATs. Gene identifiers and abbreviations for listed PDATs can be found in the Supplemental Table 1.

Figure 5. Alignment of representative PDAT polypeptides with the human lecithin:cholesterol acyltransferase (HsLCAT; Genbank Accession No. AAB34898.1). The thick line represents the LCAT polypeptide. The positions of the structurally conserved elements in LCAT, including a catalytic triad of Ser181-His377-Asp345, a salt bridge between Asp145 and Arg147, and a so-called “lid region” (annotated as blue arrow) containing Trp 61, are indicated above. Stars (★) indicate the positions of the conserved amino acids of LCAT in PDAT polypeptides. Gene identifiers and abbreviations for the listed PDATs can be found in the Supplemental Table 1.

Figure 6. Motif patterns in representative PDATs. Motif occurrences were predicted using MEME program and the polypeptides were aligned using ClustalW implemented in Geneious software. Thick lines represent aligned characters and thin lines represent gaps. Different colored and numbered boxes represent separate and distinct motifs. The size of the box does not correspond to the size of the motif due to the alignment. The putative transmembrane domains
(TMDs) are annotated as red arrows. Stars (★) indicate the LCAT-like motifs. Based on the position of the TMD and LCAT-like motifs, the PDAT polypeptide was further divided into four regions. Gene identifiers and abbreviations for PDATs can be found in the Supplemental Table 1.
Table 1. 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 (*Aquilegia coerulea*) and the core eudicots. The tree is not scaled.

<table>
<thead>
<tr>
<th>Species</th>
<th>Family</th>
<th>Clade I</th>
<th>Clade II</th>
<th>Clade III</th>
<th>Clade IV</th>
<th>Clade V</th>
<th>Clade VI</th>
<th>Clade VII</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Manihot esculenta</em></td>
<td>Euphorbiaceae</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td><em>Ricinus communis</em></td>
<td>Euphorbiaceae</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td><em>Linum usitatissimum</em></td>
<td>Linaceae</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td><em>Phaseolus vulgaris</em></td>
<td>Fabaceae</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td><em>Glycine max</em></td>
<td>Fabaceae</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td><em>Cucumis sativus</em></td>
<td>Cucurbitaceae</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td><em>Prunus persica</em></td>
<td>Rosaceae</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td><em>Fragaria vesca</em></td>
<td>Rosaceae</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td><em>Arabidopsis thaliana</em></td>
<td>Brassicaceae</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td><em>Cappella rubella</em></td>
<td>Brassicaceae</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td><em>Brassica rapa</em></td>
<td>Brassicaceae</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td><em>Theobroma cacao</em></td>
<td>Malvaceae</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td><em>Carica papaya</em></td>
<td>Caricaceae</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td><em>Gossypium hirsutum</em></td>
<td>Malvaceae</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td><em>Citrus sinensis</em></td>
<td>Rutaceae</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td><em>Citrus clementine</em></td>
<td>Rutaceae</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td><em>Eucalyptus grandis</em></td>
<td>Myrtaceae</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td><em>Vitis vinifera</em></td>
<td>Vitaceae</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td><em>Solanum lycopersicum</em></td>
<td>Solanaceae</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td><em>Mimulus guttatus</em></td>
<td>Scrophulariaceae</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td><em>Aquilegia coerulea</em></td>
<td>Ranunculaceae</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><em>Sorghum bicolor</em></td>
<td>Poaceae</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><em>Zea mays</em></td>
<td>Poaceae</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td><em>Setaria italica</em></td>
<td>Poaceae</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><em>Panica virgata</em></td>
<td>Poaceae</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><em>Oryza sativa</em></td>
<td>Poaceae</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><em>Brachypodium distachyon</em></td>
<td>Poaceae</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><em>Selaginella moellendorffii</em></td>
<td>Selaginellaceae</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><em>Physcomitrella patens</em></td>
<td>Funariaceae</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><em>Chlamydomonas reinhardtii</em></td>
<td>Chlamydomonadaceae</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><em>Volvocales</em></td>
<td>Volvaceae</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><em>Coccomyxa subellipsoidea</em></td>
<td>Coccomyxaaceae</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><em>Micromonas pusilla</em></td>
<td>Monerolaeaceae</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Table 2. Summary of PDAT protein properties

<table>
<thead>
<tr>
<th>Clade</th>
<th>Total No. of PDATs</th>
<th>Protein length (amino acid residues)</th>
<th>Molecular mass (kDa)</th>
<th>Isoelectric point (pI)</th>
<th>No. of PDATs with pI&gt;7.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>4</td>
<td>757±223</td>
<td>79.6±19.6</td>
<td>6.90±1.76</td>
<td>0</td>
</tr>
<tr>
<td>II</td>
<td>1</td>
<td>693</td>
<td>76.9</td>
<td>6.50</td>
<td>0</td>
</tr>
<tr>
<td>III</td>
<td>1</td>
<td>647</td>
<td>71.8</td>
<td>6.21</td>
<td>0</td>
</tr>
<tr>
<td>IV</td>
<td>8</td>
<td>682±4.87</td>
<td>75.4±0.4</td>
<td>6.25±0.13</td>
<td>0</td>
</tr>
<tr>
<td>V</td>
<td>17</td>
<td>670±21.96</td>
<td>74.5±2.53</td>
<td>7.65±0.94</td>
<td>12</td>
</tr>
<tr>
<td>VI</td>
<td>28</td>
<td>676±21.66</td>
<td>75.1±2.52</td>
<td>6.34±0.61</td>
<td>1</td>
</tr>
<tr>
<td>VII</td>
<td>25</td>
<td>663±40.7</td>
<td>74.1±4.76</td>
<td>7.94±0.97</td>
<td>19</td>
</tr>
<tr>
<td>Min</td>
<td>-</td>
<td>509</td>
<td>56.9</td>
<td>5.27</td>
<td></td>
</tr>
<tr>
<td>Max</td>
<td>-</td>
<td>1041</td>
<td>104.5</td>
<td>9.53</td>
<td></td>
</tr>
</tbody>
</table>
Table 3. Sequence identity and synonymous substitution rates (Ks values) of within-clade duplicated PDAT gene pairs

<table>
<thead>
<tr>
<th>Species</th>
<th>Gene Pair</th>
<th>Clade</th>
<th>Sequence Identity (%)</th>
<th>Ks</th>
<th>Mean Ks</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Glycine max</em></td>
<td>GmaPDAT3 vs GmaPDAT4</td>
<td>V</td>
<td>96.1</td>
<td>0.08996 ± 0.01655</td>
<td>0.125</td>
</tr>
<tr>
<td></td>
<td>GmaPDAT1 vs GmaPDAT2</td>
<td>VI</td>
<td>95.9</td>
<td>0.12586 ± 0.01946</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GmaPDAT5 vs GmaPDAT6</td>
<td>VII</td>
<td>92.0</td>
<td>0.15779 ± 0.02457</td>
<td></td>
</tr>
<tr>
<td><em>Gossypium raimondii</em></td>
<td>GraPDAT4 vs GraPDAT6</td>
<td>V</td>
<td>79.1</td>
<td>0.70247 ± 0.06247</td>
<td>0.794</td>
</tr>
<tr>
<td></td>
<td>GraPDAT1 vs GraPDAT5</td>
<td>VI</td>
<td>76.4</td>
<td>0.80740 ± 0.07259</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GraPDAT2 vs GraPDAT3</td>
<td>VII</td>
<td>76.9</td>
<td>0.87113 ± 0.07781</td>
<td></td>
</tr>
<tr>
<td><em>Solanum lycopersicum</em></td>
<td>SlyPDAT3 vs SlyPDAT4</td>
<td>VI</td>
<td>81.4</td>
<td>0.71665 ± 0.06707</td>
<td>0.755</td>
</tr>
<tr>
<td></td>
<td>SlyPDAT1 vs SlyPDAT5</td>
<td>VII</td>
<td>79.7</td>
<td>0.79243 ± 0.06856</td>
<td></td>
</tr>
<tr>
<td><em>Manihot esculenta</em></td>
<td>MesPDAT3 vs MesPDAT5</td>
<td>VI</td>
<td>91.0</td>
<td>0.29142 ± 0.03067</td>
<td>0.316</td>
</tr>
<tr>
<td></td>
<td>MesPDAT2 vs MesPDAT4</td>
<td>VII</td>
<td>87.6</td>
<td>0.34022 ± 0.03314</td>
<td></td>
</tr>
<tr>
<td><em>Linum usitatissimum</em></td>
<td>LusPDAT1 vs LusPDAT5</td>
<td>V</td>
<td>97.1</td>
<td>0.08181 ± 0.01528</td>
<td>0.100</td>
</tr>
<tr>
<td></td>
<td>LusPDAT2 vs LusPDAT4</td>
<td>VI</td>
<td>95.8</td>
<td>0.12603 ± 0.01866</td>
<td></td>
</tr>
<tr>
<td></td>
<td>LusPDAT3 vs LusPDAT6</td>
<td>VII</td>
<td>96.2</td>
<td>0.09326 ± 0.01564</td>
<td></td>
</tr>
<tr>
<td><em>Mimulus guttatus</em></td>
<td>MguPDAT2 vs MguPDAT3</td>
<td>VI</td>
<td>76.7</td>
<td>1.35287 ± 0.14409</td>
<td>1.353</td>
</tr>
<tr>
<td><em>Brassica rapa</em></td>
<td>BraPDAT1 vs BraPDAT2</td>
<td>VI</td>
<td>91.5</td>
<td>0.3522</td>
<td>0.352</td>
</tr>
</tbody>
</table>
Table 4. Sequence identity and synonymous substitution rates (Ks values) of between-clade duplicated PDAT gene pairs. The Ks values for individual gene pair can be found in Supplemental Table 4.

<table>
<thead>
<tr>
<th>Species</th>
<th>Clades</th>
<th>Mean Sequence identity (%)</th>
<th>Mean Ks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glycine max</td>
<td>V and VI</td>
<td>74.750±0.5</td>
<td>1.686</td>
</tr>
<tr>
<td></td>
<td>V and VII</td>
<td>61.825±0.443</td>
<td>6.944</td>
</tr>
<tr>
<td></td>
<td>VI and VII</td>
<td>61.375±0.506</td>
<td>69.654</td>
</tr>
<tr>
<td>Gossypium raimondii</td>
<td>V and VI</td>
<td>67.825±1.9</td>
<td>2.521</td>
</tr>
<tr>
<td></td>
<td>V and VII</td>
<td>59.575±0.78</td>
<td>8.845</td>
</tr>
<tr>
<td></td>
<td>VI and VII</td>
<td>58.675±1.721</td>
<td>6.686</td>
</tr>
<tr>
<td>Solanum lycopersicum</td>
<td>V and VI</td>
<td>72.75 ±0.495</td>
<td>2.542</td>
</tr>
<tr>
<td></td>
<td>V and VII</td>
<td>62±0.566</td>
<td>11.594</td>
</tr>
<tr>
<td></td>
<td>VI and VII</td>
<td>61.775±0.189</td>
<td>17.506</td>
</tr>
<tr>
<td>Manihot esculenta</td>
<td>V and VI</td>
<td>75.4±0.424</td>
<td>1.551</td>
</tr>
<tr>
<td></td>
<td>V and VII</td>
<td>61.6±0.424</td>
<td>7.608</td>
</tr>
<tr>
<td></td>
<td>VI and VII</td>
<td>60.95±0.252</td>
<td>6.6</td>
</tr>
<tr>
<td>Linum usitatissimum</td>
<td>V and VI</td>
<td>69.4±0.294</td>
<td>15.008</td>
</tr>
<tr>
<td></td>
<td>V and VII</td>
<td>62.675±0.608</td>
<td>63.257</td>
</tr>
<tr>
<td></td>
<td>VI and VII</td>
<td>63.325±0.377</td>
<td>15.013</td>
</tr>
<tr>
<td>Mimulus guttatus</td>
<td>V and VI</td>
<td>68.55±1.485</td>
<td>36.623</td>
</tr>
<tr>
<td></td>
<td>V and VII</td>
<td>60.3</td>
<td>5.655</td>
</tr>
<tr>
<td></td>
<td>VI and VII</td>
<td>61.3±0.283</td>
<td>41.615</td>
</tr>
</tbody>
</table>
Table 5. Parameter estimates, likelihood values and likelihood ratio test (LRT) P values obtained from Clade model (CmC) and extended-Clade model (Ex-CmC) analyses of the 70 core eudicot PDATs data set

<table>
<thead>
<tr>
<th>Model</th>
<th>lnL</th>
<th>0: Purifying</th>
<th>1: Neutral</th>
<th>2: Divergent</th>
<th>2∆ℓ</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ω₀, p₀</td>
<td>ω₁, p₁</td>
<td>ω₂, ω₃ (ω₄)</td>
<td>ω₂, p₂</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M2a_ref</td>
<td>-52695.098844</td>
<td>0.02847</td>
<td>0.57261</td>
<td>1.0000</td>
<td>0.01066</td>
<td>0.18723</td>
</tr>
<tr>
<td>CmC V</td>
<td>-52695.097019</td>
<td>0.02846</td>
<td>0.57248</td>
<td>1.0000</td>
<td>0.01068</td>
<td>0.18733</td>
</tr>
<tr>
<td>CmC VI</td>
<td>-52669.267166</td>
<td>0.02896</td>
<td>0.57538</td>
<td>1.0000</td>
<td>0.00941</td>
<td>0.21333</td>
</tr>
<tr>
<td>CmC VII</td>
<td>-52662.297402</td>
<td>0.02764</td>
<td>0.56217</td>
<td>1.0000</td>
<td>0.01026</td>
<td>0.15745</td>
</tr>
<tr>
<td>Ex-Null</td>
<td>-52659.898615</td>
<td>0.02687</td>
<td>0.55770</td>
<td>1.0000</td>
<td>0.01064</td>
<td>0.23108</td>
</tr>
<tr>
<td>Ex-CmC</td>
<td>-52653.785928</td>
<td>0.02732</td>
<td>0.56258</td>
<td>1.0000</td>
<td>0.00998</td>
<td>0.18319</td>
</tr>
</tbody>
</table>

Clade model: ω₀ is the estimated ω value for site class 0, p₀ is the estimated proportion of sites in site class 0 (purifying selection); ω₁ is the estimated ω value for site class 1, p₁ is the estimated proportion of sites in site class 1 (neutral selection); ω₂ is the estimated ω value for divergent sites on the background partitions, ω₃ is the estimated ω value for divergent sites on the foreground partitions, p₂ is the estimated proportion of sites in site class 2 (divergent selection); lnL is the log likelihood value; 2Δℓ is the likelihood ratio test (LRT) statistic for comparing the CmC and M2a_ref (null) models; p is the p-value of the LRT.

Extended-Clade model: The first two site classes (class 0 and class 1) are the same as CmC model. The final class (class 2) modeling divergent selection among three partitions, each with a separately estimated ω ratios (ω₂ for Clade V, ω₃ for Clade VI, and ω₄ for Clade VII). The Null-model has only two partitions and two estimated ω ratios (ω₂ for Clade VII and ω₃ for both Clade V and Clade VI). lnL is the log likelihood value; 2Δℓ is the likelihood ratio test (LRT) statistic for comparing the Ex-CmC and Ex-Null models; p is the p-value of the LRT.
Table 6. Parameter estimates, likelihood values, and likelihood ratio test (LRT) P values obtained from Clade model (CmC) analyses of 45 PDATs from the core eudicot Clades V and VI

<table>
<thead>
<tr>
<th>Model</th>
<th>lnL</th>
<th>0: Purifying</th>
<th>1: Neutral</th>
<th>2: Divergent</th>
<th>2Δℓ</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ω0</td>
<td>p0</td>
<td>ω1</td>
<td>p1</td>
<td>ω2, ω3</td>
<td>p2</td>
</tr>
<tr>
<td>M2a_ref</td>
<td>-33307.980006</td>
<td>0.02467</td>
<td>0.6711</td>
<td>1.0000</td>
<td>0.02888</td>
<td>ω2,0.20335</td>
</tr>
<tr>
<td>CmC V</td>
<td>-33298.090883</td>
<td>0.02432</td>
<td>0.66957</td>
<td>1.0000</td>
<td>0.02792</td>
<td>ω2,0.18046,ω3,0.24610</td>
</tr>
<tr>
<td>CmC VI</td>
<td>-33301.896393</td>
<td>0.02506</td>
<td>0.67361</td>
<td>1.0000</td>
<td>0.02779</td>
<td>ω2,0.23597,ω3,0.18545</td>
</tr>
</tbody>
</table>

ω0 is the estimated ω value for site class 0, p0 is the estimated proportion of sites in site class 0 (purifying selection); ω1 is the estimated ω value for site class 1, p1 is the estimated proportion of sites in site class 1 (neutral selection); ω2 is the estimated ω value for divergent sites on the background partitions, ω3 is the estimated ω value for divergent sites on the foreground partitions, p2 is the estimated proportion of sites in site class 2 (divergent selection); lnL is the log likelihood value; 2Δℓ is the likelihood ratio test (LRT) statistic for comparing the CmC and M2a_ref (null) models; p is the p-value of the LRT.
Supplemental Dataset 1: Alignment of 128 full-length LCAT-like nucleotide sequences used in Figure 1 and Supplemental Figure 1. Gene identifiers and abbreviations for listed PDATs can be found in the Supplemental Table 1.

Supplemental Dataset 2: Alignment of 87 full-length PDAT nucleotide sequences used in Figure 2. Gene identifiers and abbreviations for listed PDATs can be found in the Supplemental Table 1.

Supplemental Dataset 3: Alignment of 71 full-length nucleotide sequences of core eudicot PDATs and ScePDAT used in Supplemental Figure 2. Gene identifiers and abbreviations for listed PDATs can be found in the Supplemental Table 1.

Supplemental Figures (Figures 1 to 10) and Tables (Table 1, 2, 4 and 5)

Supplemental Table 3: InterPro domain and PANTHER (Protein Analysis Through Evolutionary Relationships) classification analysis of 128 LCAT-like protein sequences. Gene identifiers and abbreviations for listed PDATs can be found in the Supplemental Table 1.
Figure 1. Cladogram of 128 LCAT-like sequences from 38 plant species. The maximum likelihood tree was generated using the MrBayes program, and the support values along branches are Bayesian posterior probabilities (shown as percentage). The land plant LCAT-like sequences are branched into four major clades, designated as Groups A, B, C and D, whereas all algal candidates are grouped into a single algal clade. The scale bar denotes the number of nucleotide replacements per site. More detailed data are presented in Supplemental Figure 1.
Figure 2. Phylogenetic relationship of the PDAT gene family. The maximum likelihood tree was generated using the MrBayes program. PDATs are grouped into 7 distinct clades (I to 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.
Figure 3. Schematic diagram of gene structures of representative PDATs. The thin lines represent introns and thick bars represent exons. The numbers above the gene structure indicate intron phases. A scale bar with a unit of base pair (bp) is graphed on the top. TMD denotes transmembrane domain. Gene identifiers and abbreviations of listed PDATs can be found in the Supplemental Table 1.
Figure 4. (A) Alignment of transmembrane domains (TMDs) in representative plant PDATs. The putative TMDs are annotated as red arrows. The alignment was generated using ClustalW implemented in Geneious software and represented as thick lines (aligned characters) and thin lines (gaps). Overall alignment identity of 86 plant PDATs and a scale bar indicating the numbers of amino acid residues are graphed on the top. (B) Overall alignment identity of 82 land plant PDATs. Gene identities corresponding to accession information for listed PDATs can be found in the Supplemental Table 1.
Figure 5. Alignment of representative PDAT polypeptides with the human lecithin:cholesterol acyltransferase (HsLCAT; Genbank Accession No. AAB34898.1). The thick line represents the LCAT polypeptide. The positions of the structurally conserved elements in LCAT, including a catalytic triad of Ser181-His377-Asp345, a salt bridge between Asp145 and Arg147, and a so-called "lid region" (annotated as blue arrow) containing Trp61, are indicated above. Stars (*) indicate the positions of the conserved amino acids of LCAT in PDAT polypeptides. Genes identifiers and abbreviations for the listed PDATs can be found in the Supplemental Table 1.
Figure 6. Motif patterns in representative PDATs. Motif occurrences were predicted using MEME program and the polypeptides were aligned using ClustalW implemented in Geneious software. Thick lines represent aligned characters and thin lines represent gaps. Different colored and numbered boxes represent separate and distinct motifs. The size of the box does not correspond to the size of the motif due to the alignment. The putative transmembrane domains (TMDs) are annotated as red arrows. Stars (★) indicate the LCAT-like motifs. Based on the position of the TMD and LCAT-like motifs, the PDAT polypeptide was further divided into four regions. A one-standar and abbreviations for PDATs can be found in the Supplemental Table 1.