The diversity of algal phospholipase D homologs revealed by biocomputational analysis

Beligni, M.V.; Bagnato, C.; Prados, M.B.; Bondino, H.; Laxalt, A.M.; Munnik, T.; Ten Have, A.

DOI
10.1111/jpy.12334

Publication date
2015

Document Version
Final published version

Published in
Journal of phycology

J. Phycol. **51,** 943–962 (2015)

DOI: 10.1111/jpy.12334

THE DIVERSITY OF ALGAL PHOSPHOLIPASE D HOMOLOGS REVEALED BY BIOCOMPUTATIONAL ANALYSIS[1]

María Verónica Beligni[3]

Instituto de Investigaciones Biológicas (IIB-CONICET-UNMdP), Facultad de Ciencias Exactas y Naturales, Universidad Nacional de Mar del Plata, CC 1245, Mar del Plata 7600, Argentina

Carolina Bagnato[#]

Consejo Nacional de Investigaciones Científicas y Técnicas, Universidad Nacional de Río Negro, Mitre 630. S. C. de Bariloche 8400, Río Negro, Argentina

María Belén Prados[#]

Instituto de Energía y Desarrollo Sustentable - Comisión Nacional de Energía Atómica, Centro Atómico Bariloche, Av. Bustillo 9500, S. C. de Bariloche 8400, Río Negro, Argentina

Hernán Bondino,[2] *Ana María Laxalt*

Instituto de Investigaciones Biológicas (IIB-CONICET-UNMdP), Facultad de Ciencias Exactas y Naturales, Universidad Nacional de Mar del Plata, CC 1245, Mar del Plata 7600, Argentina

Teun Munnik

Department of Plant Physiology, Swammerdam Institute for Life Sciences, University of Amsterdam, Science Park 904, Amsterdam NL-1098 XH, the Netherlands

and Arjen Ten Have

Instituto de Investigaciones Biológicas (IIB-CONICET-UNMdP), Facultad de Ciencias Exactas y Naturales, Universidad Nacional de Mar del Plata, CC 1245, Mar del Plata 7600, Argentina

Phospholipase D (PLD) participates in the formation of phosphatidic acid, a precursor in glycerolipid biosynthesis and a second messenger. PLDs are part of a superfamily of proteins that hydrolyze phosphodiesters and share a catalytic motif, HxKxxxxD, and hence a mechanism of action. Although HKD-PLDs have been thoroughly characterized in plants, animals and bacteria, very little is known about these enzymes in algae. To fill this gap in knowledge, we performed a biocomputational analysis by means of HMMER iterative profiling, using most eukaryotic algae genomes available. Phylogenetic analysis revealed that algae exhibit very few eukaryotic-type PLDs but possess, instead, many bacteria-like PLDs. Among algae eukaryotic-type PLDs, we identified C2-PLDs and PXPH-like PLDs. In addition, the dinoflagellate *Alexandrium tamarense* features several proteins phylogenetically related to oomycete PLDs. Our phylogenetic analysis also showed that algae bacteria-like PLDs (proteins with putative PLD activity) fall into five clades, three of which are novel lineages in eukaryotes, composed almost entirely of algae. Specifically, Clade II is almost exclusive to diatoms, whereas Clade I and IV are mainly represented by proteins from prasinophytes. The other two clades are composed of mitochondrial PLDs (Clade V or Mito-PLDs), previously found in mammals, and a subfamily of potentially secreted proteins (Clade III or SP-PLDs), which includes a homolog formerly characterized in rice. In addition, our phylogenetic analysis shows that algae have non-PLD members within the bacteria-like HKD superfamily with putative cardiolipin synthase and phosphatidylserine/phosphatidylglycerophosphate synthase activities. Altogether, our results show that eukaryotic algae possess a moderate number of PLDs that belong to very diverse phylogenetic groups.

Key index words: **algae; glycerolipid metabolism; HKD motif; HMMER profiling; lipid signaling; phosphodiester hydrolysis; phospholipase D; protein phylogeny**

***Abbreviations*: CS, cardiolipin synthase; ML, maximum likelihood; MSA, multiple sequence alignment;**

[1]Received 17 April 2014. Accepted 9 July 2015.
[2]Present address: Biotechnology Research Center, Advanta Semillas SAIC, Balcarce, Argentina.
[3]Author for correspondence: e-mail mvbeligni@mdp.edu.ar.
[#]Both authors contributed equally to this work.
Editorial Responsibility: C. Lane (Associate Editor)

PA, phosphatidic acid; PC, phosphatidylcholine; PE, phosphatidylethanolamine; PG, phosphatidylglycerol; PGPS, phosphatidylglycerophosphate synthase; PH, Pleckstrin homology; PLD, phospholipase D; PPK, polyphosphate kinase; PSS, phosphatidylserine synthase; PX, phox homology

Phosphatidic acid (PA) is an important precursor in glycerolipid biosynthesis that also plays a role as a second messenger in eukaryotic cells (Kolesnikov et al. 2012, Pleskot et al. 2013, Kukkonen 2014). One of the mechanisms for PA generation is dependent on the action of phospholipases D (PLDs), enzymes that catalyze the hydrolysis of structural phospholipids at their terminal phosphoester bond. The preferred substrates of most PLDs are phosphatidylcholine (PC), phosphatidylethanolamine (PE) and phosphatidylglycerol (PG) (Munnik et al. 1998, Wang 2002). PLDs are part of a larger superfamily of proteins that share a similar structural core and similar reaction mechanism for a range of substrates, including neutral lipids and even polynucleotide backbones (Pointing and Kerr 1996, Wakelam et al. 1997b, Wang 2000, Bargmann and Munnik 2006, Jang et al. 2012). Pfam refers to this superfamily as "Clan CL0479" or the "PLD superfamily," which is described as a superfamily of proteins related to PLDs that contains eight families (Pfam IDs: DUF1669, PLDc, PLDc_2, PLDc_3, PP_kinase_C, RE_NgoFVII, Regulator_TrmB and Tyr-DNA_phospho). Some of the best characterized members include proteins such as bacterial cardiolipin synthase (CS), phosphatidylserine and phosphatidylglycerophosphate synthase (PSS and PGPS, respectively) and endonuclease Nuc, as well as Pox virus envelope protein (Pointing and Kerr 1996) and even human tyrosyl-DNA phosphodiesterase Tdp1 (Interthal et al. 2001). A large number of enzymes within this superfamily share a conserved $HxKx_4Dx_6GSxN$ motif (HKD) that is responsible for catalytic activity, but some degeneration appears to be allowed in this motif, particularly in the non-PLD members of the superfamily (Lee et al. 1989, Sung et al. 1997, Lerchner et al. 2005, 2006). Canonical HKD proteins with PLD enzymatic activity possess two PLD domains with one perfect HKD motif each, which together form the active site, as concluded from the crystal structure of PLD family members (Stuckey and Dixon 1999, Leiros et al. 2000). Mitochondrial PLD (Mito-PLD) is an exceptional and ancestral PLD homolog with a single HKD catalytic site that homo-dimerizes to generate PA from cardiolipin hydrolysis (Choi et al. 2006, Huang and Frohman 2009).

HKD proteins are not the only proteins exhibiting PLD activity. Several non-HKD, nonhomologous enzymes have been described, and these enzymes have diverse structures and catalytic mechanisms (Zambonelli and Roberts 2005, Selvy et al. 2011). Non-HKD PLDs include human autotaxin (Tokumura et al. 2002), *Streptomyces chromofuscus* PLD (Zambonelli and Roberts 2003) and glycosylphosphatidylinositol PLD (Jones et al. 1997), among others.

The action of HKD PLDs (henceforth PLDs) was first described in plants, where the activity of some of the numerous isoenzymes has been studied and seems to respond to a variety of abiotic stress conditions (Testerink and Munnik 2011), such as drought, hyperosmotic stress, temperature stress and wounding, as well as biotic stress, including pathogenic (Laxalt et al. 2001) and symbiotic interactions (Smékalová et al. 2014). In animals, PLDs play a critical role in cell signaling, modulating cell growth, proliferation, survival, and migration (Wakelam et al. 1997a, Selvy et al. 2011).

Eukaryotic-type PLDs are characterized by the presence of regulatory domains located at the N-terminal portion of the proteins. The Phox- (PX) and Pleckstrin homology (PH) domains are not only characteristic of PLDs in mammals and yeast (Hodgkin et al. 2000) but are also present in other eukaryotes, including plants (Wang 2000, Qin and Wang 2002, Bargmann and Munnik 2006). The PX and PH domains are involved in binding of the enzyme to inositol lipids (Powner and Wakelam 2002). Most plant PLDs, however, belong to the C2-type (Eliás et al. 2002, Liu et al. 2010, Singh et al. 2012), due to the presence of an N-terminal C2 domain, involved in Ca^{2+}-dependent lipid binding. Although most eukaryotic PLDs characterized to date belong to either the PXPH- or the C2-type, some homologs lacking these regulatory domains have recently been identified (Meijer et al. 2005, 2011, Li et al. 2007). These homologs resemble bacterial PLDs, which lack the canonical regulatory domains, are often extracellular and bear N-terminal signal peptides (SP; Zhao et al. 1997, Selvy et al. 2011).

Although the potential of eukaryotic algae (hereafter algae) as a biofuel feedstock has catalyzed a research boom in these organisms, much of their lipid metabolism remains unknown. As an example, no algal PLDs have been thoroughly characterized to date. In *Chlamydomonas moewusii*, a chlorophycean alga related to higher plants, both hyperosmotic stress (Arisz and Munnik 2011) and membrane depolarization (Meijer et al. 2002) proved to stimulate PA formation via the PLD pathway, with PE as the preferred substrate (Munnik et al. 1995, Arisz et al. 2003). In *Silvettia compressa*, a fucoid brown alga, the PLD pathway seems to be important for the formation of microtubule arrays during zygote cell division and cytokinesis (Peters et al. 2007, 2008). In both cases, PLD activity was not ascribed to a particular polypeptide or encoding gene.

Experimental analyses of algae PLD function and lipid metabolism in general will benefit from a comprehensive knowledge of the number and sequence characteristics of PLDs in algae, which is made

possible by the many genome sequences currently available. For identifying all algae enzymes with PLD function that belong to the HKD superfamily, we performed an in-depth PLD sequence data mining, including 27 algae and 53 related and nonrelated eukaryotic species. In addition, we included curated prokaryotic PLD homologs in the assay. Phylogeny and standard biocomputational analyses were performed to make testable predictions of the characteristics and/or functions of the encoded enzymes.

MATERIALS AND METHODS

Proteomes and databases used for the identification of PLD homologs. Protein models from fully annotated genomes (hereafter called complete proteomes, according to http://www.uniprot.org/keywords/KW-0181; UniProt Consortium 2012) of representative eukaryotic algae groups, supplemented with the complete proteomes of a set of nonalgal eukaryotic organisms from different supergroups were used for sequence mining (all available by June 2014). Table S1 in the Supporting Information lists the eukaryotic species analyzed, their internal codes, the source of the proteome sequences and the corresponding references or sequencing projects. We followed taxonomic relationships between taxa according to Adl et al. (2012). Prokaryotic PLD homologs were extracted from Swiss-Prot, the manually annotated and reviewed section of UniProt, and from Reference Proteomes (EMBL-EBI, http://www.ebi.ac.uk/).

Identification of algae PLD homologs. PLD homologs were identified by HMMER version 3 (Eddy 1998) iterative profiling starting from a seed alignment of 48 proteins within the Pfam PLDc family (PF00614) (http://pfam.sanger.ac.uk/family/PF00614#tabview=tab3) (Sonnhammer et al. 1997, 1998b). The multiple sequence alignment (MSA) was visualized and manually corrected using SeaView (Gouy et al. 2010) and GeneDoc (Nicholas et al. 1997) and subsequently used to generate a position-specific scoring table (hidden Markov model, hmm) using the hmmbuild tool from the HMMER suite (Eddy 1998, Johnson et al. 2010). This model was used to search a compiled fasta file containing all the complete proteomes of the selected eukaryotic species, using the hmmsearch tool. Sequences above the default inclusion threshold (E-value = 0.01) were retrieved and redundant sequences showing 100% identity were eliminated using CD-HIT (Li and Godzik 2006). The remaining sequences were aligned using MAFFT at http://mafft.cbrc.jp/alignment/server/index.html (Katoh et al. 2002). The resulting MSA was manually corrected and used for the generation of a new hmm file, restarting the whole cycle. This process was repeated until convergence, at which point no new information was obtained when a new data-mining cycle was done. For dinoflagellates, euglenids and red algae, complete proteomes are scarce or unavailable; hence, hmm files were used to search the nonredundant database of the HMMER website (http://hmmer.janelia.org/, hmmsearch tool). In addition, to retrieve EST information that could be valuable for the mentioned taxa, two PLDs from each resulting phylogenetic group were used as queries for tblastn search of the EST NCBI database. (http://blast.ncbi.nlm.nih.gov/Blast.cgi). To ensure that we obtained the most complete data set of eukaryotic PLDs possible, the sequences from each phylogenetic group were retrieved and used in group-specific data mining following the same procedure as the one described above. To analyze algae PLD homologs in a true phylogenetic context, both the seed and the final hmm files were used to search for prokaryotic homologs. This was done by searching (using hmmsearch) the Swiss-Prot and Reference Proteomes databases at the HMMER website, restricted to Bacteria and Archaea. All nonidentical prokaryotic sequences below a established inclusion threshold (E-value = 10^{-8}) were added to the analysis.

Phylogenetic analysis of PLD homologs. The identified sequences were aligned to the hmm of the last mining cycle using hmmalign (HMMER) and the MSAs manually corrected in GeneDoc (Nicholas et al. 1997). To use only phylogenetically informative regions for the reconstruction of phylogenetic trees, the MSAs were trimmed using BMGE (Block Mapping and Gathering with Entropy) (Criscuolo and Gribaldo 2010). BMGE optional arguments were determined based on the conservation of secondary structure elements and were: -m BLOSUM30 -h 0.7 -g 0.65:0.3 -b 3. This procedure eliminated not only alignment columns according to entropy and gap content but also eliminated sequences that contained a gap proportion higher than 65%. All the putative algae PLDs that were eliminated using this criteria were confronted with the group-specific hmm profiles mentioned in the previous section, and the closest phylogenetic group was determined by the highest HMMER scores. The eliminated sequences were included in the Tables, but not in the final trimmed alignment used for phylogenetic tree reconstruction. Maximum likelihood (ML) and Bayesian methods were used for phylogenetic tree reconstruction. For ML, PhyML 3.0 was used with default settings and 600 bootstraps (Guindon et al. 2010), with the LG model. For Bayesian analysis, we used Mr. Bayes 3.2.3 (Ronquist et al. 2012) with the WAG model, with its own default priors and invariant gamma distribution approximated with four substitutions. Two Metropolis-coupled Markov Chain Monte Carlo runs with four chains each and a temperature of T = 0.05 were done. Convergence of the chains in the Bayesian analysis was assessed by monitoring that the standard deviation of split frequencies was < 0.01 and with AWTY online, a system for the graphical exploration of MCMC convergence in Bayesian phylogenetic inference (http://king2.scs.fsu.edu/CEBProjects/awty/awty_start.php) (Nylander et al. 2008). Trees were visualized with Dendroscope v3.2.10 (Huson and Scornavacca 2012) and final tree editing was done using iTOL (Letunic and Bork 2011).

Protein localization predictions. Subcellular localization of the selected proteins was predicted using both manual and automated methods. Four neural network-based tools optimized for different taxa were used: PredAlgo (https://giavapgenomes.ibpc.fr/cgi-bin/predalgodb.perl?page=main) was used for the prediction of transit peptides in Archaeplastida algae (Tardif et al. 2012), Hectar (http://www.sb-roscoff.fr/hectar/) was used for Heterokonts (Gschloessl et al. 2008), PATS (http://gecco.org.chemie.uni-frankfurt.de/pats/pats-index.php) was used for Apicomplexans (Zuegge et al. 2001) and TargetP (http://www.cbs.dtu.dk/services/TargetP/) for the rest (Emanuelsson et al. 2007). In addition, a manual method was used for organisms containing secondary plastids, as described elsewhere (Jiroutová et al. 2007). Briefly, sequences were first analyzed using SignalP (http://www.cbs.dtu.dk/services/SignalP/) (Petersen et al. 2011). For those sequences with positive prediction of endoplasmic reticulum trafficking, the SPs were manually removed and the resulting sequences were further analyzed using TargetP to test for the presence of plastid targeting peptides. For TargetP, considering that the analyzed taxa constitute a diverse group, no cutoffs were established and the default "winner-takes-all" criterion was used for determining the potential subcellular localizations. The prediction was only done for complete sequences, therefore partial sequences were annotated as "NGE" (not good enough evidence for prediction). The presence of transmembrane (TM) helices was predicted with TMHMM (http://www.cbs.dtu.dk/services/TMHMM/)

(Sonnhammer et al. 1998a). For prokaryotic sequences, subcellular localization was assigned according to UniProt annotation information combined with SignalP and TMHMM results.

Functional annotations. The level of evidence for the existence of the selected proteins was retrieved analyzing published data available at NCBI (http://www.ncbi.nlm.nih.gov/). Alternatively, tblastn analyses against EST collections were performed at the corresponding genome or database websites (e.g., JGI Genomes or Diatom Database). The results were classified according to the UniProt nomenclature as follows: "protein" indicates that there is clear experimental evidence for the existence of the protein (partial or complete Edman sequencing, clear identification by mass spectrometry, X-ray or NMR structure, identification of protein–protein interaction or detection of the protein using antibodies); "transcript" indicates the existence of a transcript determined by expression data (cDNAs, RT-PCR, Northern blots or ESTs); "predicted" is used for entries without evidence at protein or transcript levels, the existence of the protein is therefore inferred from gene models. The evidence level for all algae PLD homologs identified is summarized on Table S2 in the Supporting Information.

Domain analysis was performed using Pfam search mode (Sonnhammer et al. 1997, 1998b). Sequence logos for each group were generated using WebLogo (http://weblogo.berkeley.edu/) with default settings, including sample size correction (Crooks et al. 2004), starting from clade-specific subalignments constructed with MAFFT (Katoh et al. 2002) at default settings and trimmed with BMGE using the following optional arguments: -m BLOSUM62 -h 0.5 -g 0.7: 0.3 -b 3.

Modeling of PLD homologs. Three-dimensional (3D) modeling was done using I-Tasser online (http://zhanglab.ccmb.med.umich.edu/I-TASSER/; Zhang et al. 2005) with default settings. The prediction of homo-dimers was performed using COTH (http://zhanglab.ccmb.med.umich.edu/COTH/; Mukherjee and Zhang 2011). Models were visualized and edited with Jmol (http://www.jmol.org/; Jmol Team 2014) and VMD v1.9.1 (Humphrey et al. 1996). The structural aligments were made using the STAMP structural alignment module, part of VMD.

RESULTS

The HKD superfamily. To identify all possible HKD-PLD homologs in algae, we performed an iterative sequence mining of the complete proteomes of the eukaryotic species shown in Figure 1A. As outlined in Figure 1B, a seed alignment (MSA) of PLDs from reference organisms was used to generate a hmm that was then used to search the mentioned complete proteomes. The identified sequences were aligned and the MSA was manually corrected. The procedure was iterated until convergence occurred. Since for some groups of algae the number of fully annotated genomes is scarce or not available, we also searched the nonredundant database at the HMMER website for potential PLDs using the hmm file generated during the last cycle of data mining and the EST database at the NCBI website using tblastn. To increase the phylogenetic context of this analysis, the hmm files were also used to search for well-annotated prokaryotic sequences using the Swiss-Prot and Reference Proteomes databases at the HMMER website, and restricting the search to Bacteria and Archaea. Using this criteria, we identified 188 sequences from Bacteria and 15 from Archaea within the established cut-off limits. The complete set of sequences obtained in this manner was aligned, trimmed using BMGE (see Appendix S1 in the Supporting Information for a trimmed alignment) and used for phylogenetic analyses. As a result of trimming, all the sequences containing a gap proportion higher than 65% were eliminated and not used for tree reconstruction.

As mentioned earlier, one challenge of working with the PLD family is the existence of other proteins that contain HKD motifs (and hence are part of the HKD superfamily) but possess other phosphodiesterase-related activities. Bacterial CSs, PSSs, PGPSs, and endonucleases are some of the best characterized homologs (Pointing and Kerr 1996). Although most of these proteins are bacterial, a few examples have been found in eukaryotes (Schäfer et al. 2007, Yoshikawa et al. 2010). Here, phylogenetic analysis was used to locate HKD superfamily members within the different families. The resulting tree, depicted in Figure 2, shows a clear hierarchical clustering (also seen in the radial phylogram inset). Bayesian and ML topologies were identical for all major clades, showing only small differences within clades (see Appendices S2 and S3 in the Supporting Information for Bayesian and ML-bootstrapped tree files). There appear to be two, well-supported, major clades. The first one consists of the eukaryotic-type PLD homologs (Fig. 2 and Appendices S2 and S3, Bayesian posterior probability = 0.83, ML Bootstrap support = 87). This clade is, as expected, mostly composed of proteins with two HKD catalytic domains and amino-terminal regulatory domains. The second major clade contains all the prokaryotic PLD homologs included in the analysis, as well as many sequences from eukaryotes with similarity to bacterial PLDs and to other prokaryotic HKD enzymes. Here, we refer to the members of this clade collectively as the bacteria-like HKD superfamily (Fig. 2 and Appendices S2 and S3, Bayesian posterior probability = 0.996, ML Bootstrap support = 98). This clade contains proteins that lack regulatory domains and many members with single catalytic domains. Interestingly, while all the members of the eukaryotic-type PLD clade come from the eukaryotic domain, the bacteria-like HKD superfamily is composed of proteins from both prokaryotes and eukaryotes.

Phylogenetic analysis resulted more challenging for the bacteria-like HKD superfamily than for the clade of eukaryotic-type PLDs. First, for bayesian analysis, a higher amount of generations were needed for some of the splits of the bacteria-like clades to reach convergence between runs (~20 millions). When analyzing many compatible trees from different runs and even from different analyses, we

A

Classification at Higher Ranks		Lower Ranks
ARCHAEPLASTIDA	**Glaucophyta**	*Cyanophora paradoxa*
	Rhodophyta (Red Algae)	*Cyanidioschyzon merolae*
		Chondrus crispus
		Porphyridium purpureum
	Chloroplastida	**Chlorophyta (Green Algae)**
		Coccomyxa subellipsoidea
		Chlorella variabilis
		Chlamydomonas reinhardtii
		Volvox carteri
		Bathycoccus prasinos
		Micromonas
		Ostreococcus
		Streptophyta
		Physcomitrella patens
		Selaginella moellendorffii
		Oryza sativa
		Zea mays
		Arabidopsis thaliana
		Glycine max
		Solanum lycopersicum
		Vitis vinifera
SAR	**Rhizaria**	*Bigelowiella natans*
	Alveolata	**Protalveolata**
		Perkinsus marinus
		Dinoflagellata (Dinoflagellates)
		Alexandrium tamarense
		Symbiodinium minutum
		Apicomplexa
		Babesia bovis
		Plasmodium
		Theileria
		Cryptosporidium parvum
		Eimeria tenella
		Toxoplasma gondii
		Ciliophora
		Ichthyophthirius multifiliis
		Tetrahymena thermophila
		Paramecium tetraurelia
	Stramenopiles	**Labyrinthulomycetes**
		Aplanochytrium kerguelense
		Aurantiochytrium limacinum
		Schizochytrium aggregatum
		Peronosporomycetes
		Albugo laibachii
		Phytophthora infestans
		Pythium irregulare
		Eustigmatophyta
		Nannochloropsis
		Pelagophyta
		Aureococcus anophagefferens
		Phaeophyta (Brown Algae)
		Ectocarpus siliculosus
		Diatomea (Diatoms)
		Fragilariopsis cylindrus
		Pseudo-nitzschia multiseries
		Phaeodactylum tricornutum
		Thalassiosira pseudonana
	Haptophyta	*Emiliania huxleyi*
	Cryptophyceae	*Guillardia theta*
EXCAVATA	**Discoba**	**Euglenozoa**
		Euglena gracilis
		Leshmania major
		Trypanosoma
		Heterolobosea
		Naegleria gruberi
	Metamonada	**Fornicata**
		Giardia intestinalis
		Parabasalia
		Trichomonas vaginalis
AMOEBOZOA		*Dictyostelium purpureum*
OPISTHOKONTA	**Holozoa**	**Metozoa**
		Nematostella vectensis
		Daphnia pulex
		Branchiostoma floridae
		Danio rerio
		Xenopus tropicalis
		Bos taurus
		Homo sapiens
	Nucletmycea	**Fungi**
		Aspergillus niger
		Neurospora tetrasperma
		Saccharomyces cerevisiae
		Trichoderma longibrachiatum
		Penicillium canescens
		Rhizopus microsporus

B

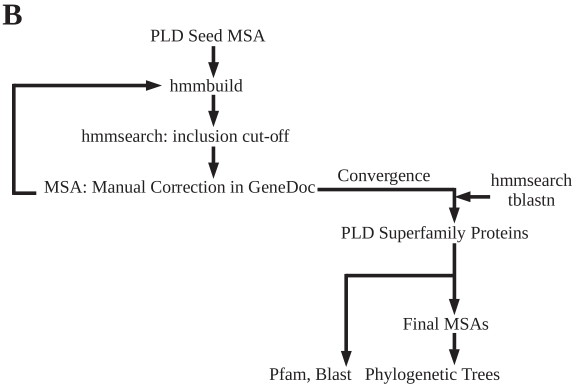


FIG. 1. Identification of PLD homologs in algae by biocomputational analysis. (A) Table showing the species used for biocomputational analysis. Representative algae with complete proteomes, as well as selected organisms from Archaeplastida, SAR, Haptophyta, Cryptophyceae, Excavata, Opisthokonta, and Amebozoa were selected for HMMER iterative data mining. Algae species are highlighted as separate boxes. Where only the genus name is provided, more than one species were analyzed. PLD homologs from *Euglena gracilis*, *Chondrus crispus* and *Alexandrium tamarense*, as well as prokaryotic PLD homologs, were obtained searching several databases within the HMMER website in a noniterative manner. (B) Flowchart of PLD sequence identification. A multiple sequence alignment (MSA) of characterized PLDs from diverse organisms was used to build a HMMER profile using the hmmbuild tool, which was subsequently used to search the proteomes shown in A) (with hmmsearch). Sequences positively identified were aligned into a new MSA and further scrutinized for the presence of the hallmark catalytic motif. This MSA was used to create a new hmm for iterative sequence identification until convergence. After the addition of proteins identified through the databases in the HMMER website, the final set of HKD superfamily proteins was used to reconstruct phylogenetic trees (after alignment and trimming) and to perform Pfam and Blast analyses for functional annotation of proteins and domains.

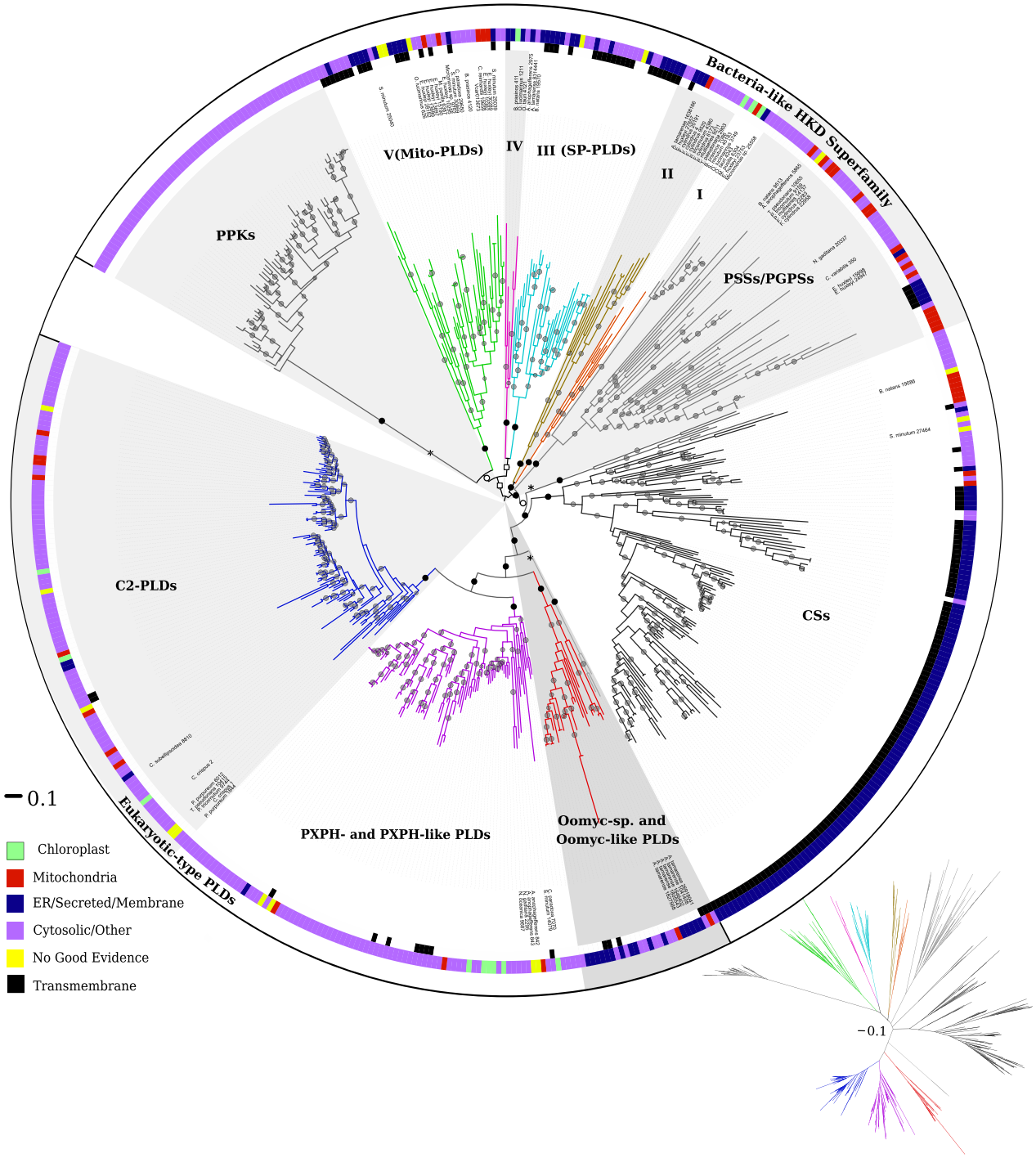


FIG. 2. Phylogenetic relationships of algae HKD superfamily members. Rooted circular phylogram representation of the tree generated by the maximum likelihood (ML) and Bayesian methods on the conserved regions of proteins from the HKD superfamily. The main groups are: C2-PLDs, PXPH-PLDs (including PXPH-like PLDs), oomycete-specific PLDs and oomycete-like PLDs (Oomyc-sp. and Oomyc-like PLDs) and bacteria-like HKD superfamily members. From this last group, cardiolipin synthases (CSs), phosphatidyl synthases (including both PSSs and PGPSs), polyphosphate kinases (PPKs) and the five clades of bacteria-like PLDs are shown. All the groups, with the exception of the clades indicated with asterisks, contain algae members. ML bootstrap support values higher than 70% are shown with gray circles. Only for the major clades, black circles indicate support values higher than 0.7 (or 70%) for both Bayesian and ML analyses, while open squares indicate that branch support is provided only by bootstrap, but not Bayesian, analysis. The open white circles indicate the only two major clades with support values between 0.50 and 0.70. Abbreviated species names and internal number codes are shown only for leaves corresponding to algae proteins. The scale bar represents 0.1 amino acid substitution per site. The inset below shows the same tree as an unrooted radial phylogram. To view the trimmed MSA used to reconstruct the trees, see Appendix S1; to see the raw trees in Nexus format, see Appendices S2 and S3. The two strips correspond to predicted subcellular localization (outer) and prediction of transmembrane domains (inner).

evidenced quite some heterogeneity in the organization of some of the subclades within the bacteria-like HKD superfamily (data not shown). This heterogeneity was the cause of a few low to moderate support values within the bacteria-like HKD superfamily when using all compatible trees to build a consensus (Figs. 2 and 3), and it also generated a few politomies when using the 50% majority rule (data not shown). In contrast to the bacteria-like HKD superfamily, the topologies of the eukaryotic-type subclades were much more homogeneous between samples, producing higher support values and an earlier reach of convergence (data not shown).

An analysis of eukaryotic-type PLDs reveals that the four major subclades consist of proteins all likely to have PLD activity, as previously reported for many of the homologs (Selvy et al. 2011). These clades include the canonical C2- and PXPH-PLDs (including a few PXPH-like proteins) and two clades containing the majority of known oomycete PLD sequences (Meijer et al. 2011), as well as a number of related algae sequences from the dinoflagellate *Alexandrium tamarense*, referred to as oomycete-like PLDs (Fig. 2). On the contrary, the bacteria-like HKD superfamily is composed of several clades of proteins with distinct enzymatic activities, as evidenced by the many well-characterized homologs annotated on Swiss-Prot. The first major clade includes both prokaryotic and eukaryotic CSs (Bayesian posterior probability = 1.00, ML Bootstrap support = 100). Although CSs include many well-characterized proteins from bacteria and a number of homologs from protists, we identified only two sequences from algae (one from the chlorarachniophyte *Bigelowiella natans* and one from the dinoflagellate *Symbiodinium minutum*; Table 1). The second clade is composed of PSSs (Bayesian posterior probability = 1.00, ML Bootstrap support = 100). PSSs are more widespread among algae than CSs, and include members from the glaucophyte *Cyanophora paradoxa*, the chlorophyte *Chlorella variabilis*, the eustigmatophyte *Nannochloropsis gaditana*, the pelagophyte *Aureococcus anophagefferens*, *B. natans* and diatoms (Table 1). Then there are five clades (numbered I–V) of putative PLDs containing proteins from algae, which will be discussed in further detail. Last, there is a clade composed solely of proteins from prokaryotes that corresponds to bacterial polyphosphate kinases (PPKs), non-PLD members of the HKD clan (see Introduction) required for the synthesis of inorganic polyphosphate (polyP) from ATP (Tzeng and Kornberg 2000).

Most algae have few putative HKD PLDs. Algae in general possess considerably fewer PLDs than higher plants. This is most remarkable for green and red algae, given their monophyletic origin with plants. As an example, we identified 12 PLD sequences in the annotated genome of *Arabidopsis thaliana*, 16 in *Oryza sativa*, 18 in *Solanum lycopersicum* and 23 in *Glycine max*, whereas glaucophytes, green and red algae appear to have 3 PLDs at most (Table 1). In some striking cases, such as *C. variabilis* (green alga) and *Cyanidioschyzon merolae* (red alga), no sequences encoding a protein with putative PLD activity were identified at all. The algae with the highest number of PLDs were the dinoflagellates *S. minutum* and *A. tamarense* and the coccolithophore haptophyte *Emiliania huxleyi*, which appear to have 9, 7, and 10 PLDs, respectively. All diatoms analyzed have at least one PLD, with as high as four (*Fragilariopsis cylindrus*; Table 1). *Guillardia theta*, a cryptophyte and *Ectocarpus siliculosus*, a brown alga, do not appear to have any HKD-containing proteins with putative PLD activity. The pelagophyte *A. anohagefferens* features two PLDs, whereas *B. natans* and the eustigmatophytes *N. gaditana* and *Nannochloropsis oceanica* appear to have single proteins with putative PLD activity (Table 1).

Algae PLDs belong to diverse phylogenetic groups. Analysis of the tree in Figure 2 reveals that algae PLDs belong to a wide number of distinct classes. Table 2 and Table S2 show a list of all the sequences with putative PLD activity identified in algae, together with a set of features that characterize each one of them. As already outlined for the case of the whole HKD superfamily, we found four main groups of putative PLDs: C2-PLDs, PXPH-PLDs (including PXPH-like PLDs), oomycete-specific PLDs (and related oomycete-like PLDs) and bacteria-like PLDs. The catalytic motifs of each of the groups have unique distinguishing features, as shown in the sequence logos in Figure 4.

Canonical eukaryotic PLDs: C2- and PXPH-PLDs. Within the clade of eukaryotic-type PLDs, C2- and PXPH-PLDs appear to have evolved from a common ancestor (Fig. 2, Bayesian posterior probability = 0.998, Bootstrap support = 89). C2-PLDs include most of the land plant PLDs and very few proteins from green (*Coccomyxa subellipsoidea*, one polypeptide) and red algae (*Chondrus crispus*, two PLDs; *Porphyridium purpureum*, two PLDs; Fig. 2 and Table 2). This group also includes single members from the diatoms *Thalassiosira pseudonana* and *Phaeodactylum tricornutum* and several proteins from nonalgae protists (Fig. 2). Remarkably, two of the algae C2-PLDs have imperfect HKD1 motifs: in one of the two *C. crispus* C2-PLDs, HKD1 is replaced by HKT; in *C. subellipsoidea*, HKD1 is modified to HKS (Table 2).

PXPH-PLDs include most PLDs from amoebozoans and opisthokonts, some land plant PLDs and a few members from nonphotosynthetic protists, including the oomycete *Phytophthora infestans*. We could not identify any true PXPH-PLDs in any the algae species analyzed (Fig. 2). However, we did identify several algal proteins with their catalytic domains most closely related to those of PXPH-PLDs, but that lack the PX or PH domains, or both. For this reason, this subclade was named PXPH-like PLDs. A few nonphotosynthetic protists also seem to

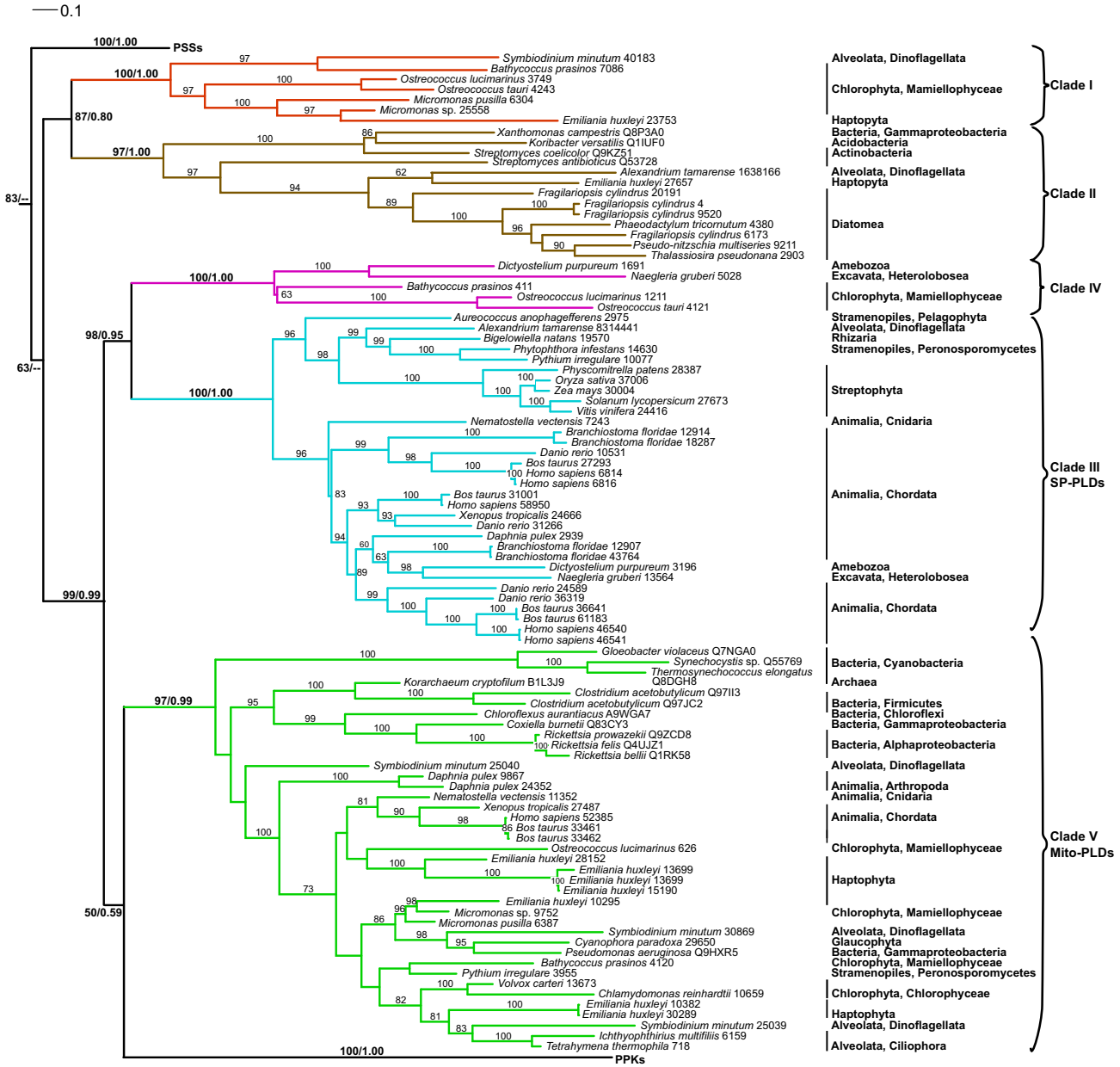


FIG. 3. Composition and phylogenetic relationships of bacteria-like PLDs. Detailed phylogram showing the full composition of Clades I–V of bacteria-like PLDs and collapsed clades for PPKs and PSSs. Full species names and internal codes are shown for eukaryotic sequences, full species names and Uniprot IDs are shown for prokaryotic sequences. Each clade is delimited with brackets. Taxonomic classification at the higher ranks is given for all the species involved in the phylogram. ML Bootstrap support is shown for all splits. For the major splits (in bold), Bayesian posterior probabilities are shown after the slash only when the posterior probability is >0.50. The scale bar represents 0.1 aminoacid substitution per site.

have this type of PXPH-like PLDs (see trees in Appendices S2 and S3).

An interesting case is provided by eustigmatophytes. HHMER profiling shows that both *N. gaditana* and *N. oceanica* have a gene that codes for a protein with very high similarity to PXPH-PLDs, but with both catalytic histidines (H) and lysines (K) modified. In *N. gaditana*, HKD1 is replaced by NTD, while in *N. oceanica*, HKD1 is replaced by NSD. In both organisms, HKD2 is modified to NRD (Table 2). Interestingly, both these proteins seem to have a PX-like domain detectable by delta blast, but not Pfam. Modeling of both proteins on known PLDs yielded a 3D structure with a 100% confidence (data not shown).

Oomycete-like PLDs. Figure 2 shows that there are two well-supported clades of PLDs composed largely by proteins from oomycetes, which we refer to as oomycete-specific PLDs. One of them is exclusive to this taxonomic group, and includes the previously

TABLE 1. Number of HKD superfamily homologs identified in the selected algae by HHMer and blast analyses.

Organism	Putative PLDs	Putative cardiolipin synthases	Putative phosphatidylserine/ phosphatidylglycerophosphate synthases
Cyanophora paradoxa	3	–	2 (consensus from Contig 37097-abinit-gene-0.1, consensus from Contig 40239-abinit-gene-0.1)*,†
Cyanidioschyzon merolae	–	–	–
Chondrus crispus‡	2	–	–
Porphyridium purpureum	2	–	–
Chlorella variabilis	–	–	1 (E1Z2U1)§
Coccomyxa subellipsoidea	1	–	–
Chlamydomonas reinhardtii	1	–	–
Volvox carteri	1	–	–
Bathycoccus prasinos	3	–	–
Micromonas sp.	3	–	–
Micromonas pusilla	3	–	–
Ostreococcus lucimarinus	3	–	–
Ostreococcus tauri	3	–	–
Bigelowiella natans	1	1 (jgi|Bigna1|21168|gw1.5.39.1)*	1 (jgi|Bigna1|79212|fgenesh1_pg.60_#_83)*
Symbiodinium minutum	9	1 (symbB1.v1.2.024281.t1)*	–
Alexandrium tamarense‡	7	–	–
Nannochloropsis gaditana	1	–	1 (K8YRD7)§
Nannochloropsis oceanica	1	–	–
Aureococcus anophagefferens	2	–	1 (F0YJZ1)§
Ectocarpus siliculosus	–	–	–
Fragilariopsis cylindrus	4	–	2 (jgi|Fracy1|251375|fgenesh2_pg.35_#_120, jgi|Fracy1|244906|fgenesh2_pg.14_#_204)*
Phaeodactylum tricornutum	2	–	1 (B7G454)§
Pseudo-nitzschia multiseries	1	–	1 (jgi|Psemu1|291269|fgenesh1_pg.662_#_4)*
Thalassiosira pseudonana	3	–	1 (B8C566)§
Emiliania huxleyi	10	–	2 (R1D6M1)§ (R1D3E8)§
Guillardia theta	–	–	–
Euglena gracilis‡	–	–	–

*Protein IDs obtained from the corresponding genome pages detailed in Table S1.
†These sequences were eliminated before tree reconstruction due to their gap content. Their putative group localization was determined according to their hmmserach scores (see Methods).
‡Complete proteomes were unavailable at the time of analysis. Data might be incomplete.
§Uniprot IDs.

characterized Pi-TM-PLD protein from *Phytophthora infestans* (Meijer et al. 2011). The other clade consists of several PLD-like proteins also previously described in *P. infestans* (Meijer et al. 2011), along with orthologs in other oomycetes. As mentioned above, phylogenetic analysis provides strong evidence that this last clade of oomycete-specific PLDs clusters with another subset composed of five PLDs from the distantly related alga *A. tamarense* (oomycete-like PLDs), forming a monophyletic clade (Fig. 2, Bayesian posterior probability = 1.00, Bootstrap Support = 100). Interestingly, all the sequences within this clade also have a modified HKD1, while HKD2 remains canonical (Fig. 4). In *A. tamarense*, the HKD1 is modified to HKR or HKQ (Table 2). Protein subcellular localization predictions indicate that most of the proteins within this clade likely enter the secretory pathway (Fig. 2).

Bacteria-like PLDs: clades I-V. A large number of algae PLDs show significant similarities to prokaryotic PLD homologs. The catalytic motifs of the proteins within this group have unique characteristics. Bacteria-like PLDs show a lower degree of conservation in the catalytic aspartic acid (D), either in HKD1 or HKD2, which is commonly replaced by other amino acids with either similar or different chemical properties. In contrast, modification of this amino acid is rare for C2 or PXPH-PLDs. Both catalytic motifs of bacteria-like PLDs have an invariant asparagine (N) downstream of the HxKxxxxD (mainly as GGxN, GSxN, SSxN, SGxN, TGxN, GTxN, or STxN), whereas C2, PXPH-like and oomycete-like PLDs all have a GGxD in HKD1 and a GSxN in HKD2 (Fig. 4 and Appendix S1). Bacteria-like PLDs also have a greater number of amino acids with bulky hydrophobic side chains (Y, F, W) in both catalytic motifs (Fig. 4). Bacteria-like PLDs identified by our hmmer profiling can be further divided into five main clades. These clades might be paraphyletic with respect to CSs and PSSs/PGPSs, although with a moderate statistical support provided only by ML (Fig. 2, Bootstrap support = 63). In addition, clade I shares a common ancestor with clade II, while clade III, IV, and V form a monophyletic clade together with PPKs. A detailed tree showing the full taxon composition of the five bacteria-like PLD clades, as well as their phylogenetic relationship with PPKs, is shown in Figure 3.

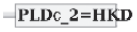
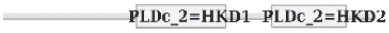
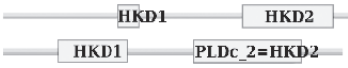
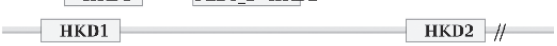
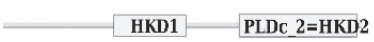
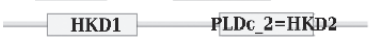
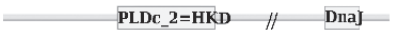
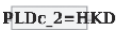
Clade I. This subgroup is composed exclusively of algal proteins, including the prasinophytes *Ostreococ-*

TABLE 2. Main predicted characteristics of algae proteins with putative PLD enzymatic activity.

	Detailed Organism	Internal ID	Uniprot ID	Length	PLD Group/Clade [a]	Catalytic motif		Domain architecture [b]
						HKD1	HKD2	
ARCHAEPLASTIDA	*Chondrus crispus*	Chc0000001	R7QAJ1	898	C2	HKD	HKD	C2 HKD1 HKD2 PLD_C
		Chc0000002	R7QG88	858	C2	HKT	HKD	C2 HKD1 PLDc_2 = HKD2 PLD_C
	Porphyridium purpureum	Ppur001844	N.A.	844	C2	HKD	HKD	PLDc=HKD1 PLDc=HKD2 PLD_C
		Ppur006012 [c]	N.A.	530	C2	HKD	–	C2 PLDc=HKD
		Ppur006013 [c]	N.A.	158	? (C2)	HKD	–	PLDc=HKD
	Volvox carteri f. nagariensis	Vcar013673	D8UHV0	226	V (MITO)	HKD	–	PLDc_2 = HKD
	Chlamydomonas reinhardtii CC-503 mt+	Crei010659	A8IW99	223	V (MITO)	HKD	–	PLDc_2 = HKD
	Bathycoccus prasinos	Bpra000411	K8EYG9	526	IV	HKD	–	PLDc_2=HKD
		Bpra004120	K8EYZ8	264	V (MITO)	HKD	–	PLDc_2=HKD
		Bpra007086	K8ENR5	509	I	HKK	HKK	PLDc_2=HKD1 PLDc_2=HKD2
	Micromonas pusilla CCMP1545 v3.0	Mpus006304	C1N565	489	I	HKH	HKR	HKD1 PLDc_2=HKD2
		Mpus006387	C1MQC2	199	V (MITO)	HKD	–	PLDc_2=HKD
		Mpus006717	C1N7K0	820	? (N.C.)	HKD	HKD	PLDc_2=HKD1 PLDc_2=HKD2
	Micromonas sp. RCC299	Misp009752	C1EET0	196	V (MITO)	HKD	–	PLDc_2 = HKD
		Misp012311	C1FH57	793	? (N.C.)	HKD	HKD	PLDc_2 = HKD1 PLDc_2 = HKD2
		Misp025558	C1EJG8	536	I	HKH	HKE	HKD1 PLDc_2 = HKD2
	Ostreococcus lucimarinus CCE9901	Oluc000626	A4S175	Partial	V (MITO)	HKD	–	PLDc_2 = HKD
		Oluc001211	A4RRH2	440	IV	HKD	HKD	PLDc_2 = HKD1 PLDc_2 = HKD2
		Oluc003749	A4RS98	464	I	HKE	HKE	Lsr2 PLDc_2 = HKD1 PLDc_2 = HKD2
	Ostreococcus tauri iOTH95	Otau003862	N.A.	149	? (PXPH-like)	HKD	–	PLDc_2 = HKD
		Otau004121	Q01G57	396	IV	HKD	HKD	PLDc_2= HKD1 PLDc_2= HKD2
		Otau004243	Q01FD3	386	I	HKQ	HKE	HKD1 PLDc_2 = HKD2
	Coccomyxa subellipsoidea C-169	Csub008810	I0YIM4	766	C2	HKS	HKD	C2 HKD1 PLDc_2 = HKD2 PLD_C
	Cyanophora paradoxa	Cypa007070	N.A.	299	PXPH-like	HKD	–	HKD
		Cypa029650	N.A.	198	V (MITO)	HKD	–	PLDc_2=HKD
		Cypa031663	N.A.	Partial	? (PXPH-like)	HKD	–	HKD
HAPTOPHYTA	*Emiliania huxleyi CCMP1516*	Ehux005194	R1FFU4	110	? (III)	HKD	–	PLDc_2 = HKD
		Ehux010295	R1C5G9	184	V (MITO)	HKD	–	PLDc_2 = HKD
		Ehux010382	R1C2H1	453	V (MITO)	HKD	–	KH-1 PLDc_2 = HKD
		Ehux013697	R1D7T9	385	V (MITO)	HKD	–	PLDc_2 = HKD
		Ehux013699	R1BLR0	863	V (MITO)	HK–	HKD	PLDc_2 = HKD1 PLDc_2 = HKD2
		Ehux015190	R1D9S8	777	V (MITO)	HKD	–	PLDc_2 = HKD
		Ehux023753	R1D2U9	414	I	HKA	–	HKD
		Ehux027657	R1ECX6	705	II	HKD	HKD	HKD1 HKD2
		Ehux028152	R1EL36	274	V (MITO)	HKD	–	PLDc_2 = HKD
		Ehux030289	R1CBL6	453	V (MITO)	HKD	–	KH-1 PLDc_2 = HKD

(continued)

TABLE 2. (continued)

	Detailed Organism	Internal ID	Uniprot ID	Length	PLD group	Catalytic motif		Domain architecture [a]
						HKD1	HKD2	
SAR	*Bigelowiella natans CCMP2755*	Bnat019570	N.A.	795	III (SP)	HKD	HKD	
	Aureococcus anophagefferens CCMP1984	Aano000842 [c]	F0Y0G9	Partial	PXPH-like	HKD	–	
		Aano000843 [c]	F0Y0H0	Partial	PXPH-like	HKD	–	
		Aano002975	F0Y438	505	III (SP)	HKD	HKD	
	Thalassiosira pseudonana CCMP 1335	Tpse002903	B5YN52	756	II	HKD	HKD	
		Tpse006834	B8C630	645	? (Mito)	HKL	HKD	
		Tpse010415	B8BY84	Partial	C2	HKD	HKD	
	Phaeodactylum tricornutum CCAP1055/1	Ptri004380	B7G133	567	II	HKD	HKD	
		Ptri009744	B7FZB6	Partial	C2	HKD	HKD	
	Pseudo-nitzschia multiseries CLN-47 v1.0	Pmul009211	N.A.	574	II	HKD	HKD	
	Fragilariopsis cylindrus CCMP 1102	Fcyl000004	N.A.	592	II	HKD	HKD	
		Fcyl006173	N.A.	688	II	HKD	HKD	
		Fcyl009520	N.A.	592	II	HKD	HKD	
		Fcyl020191	N.A.	580	II	HKD	HKD	
	Nannochloropsis gaditana CCMP526	Ngad002295	W7TJZ4	1422	PXPH-like	NTD	NRD	
	Nannochloropsis oceanica CCMP1779	Noce009687	N.A.	1485	PXPH-like	NSD	NRD	
	Alexandrium tamarense CCMP1598	Ata1638166	N.A.	627	II	HKD	HKD	
		Ata1041408	N.A.	546	Oomycete-like	HKQ	HKD	
		Ata1805543	N.A.	551	Oomycete-like	HKQ	HKD	
		Ata26918041	N.A.	638	Oomycete-like	HKR	HKD	
		Ata1827568	N.A.	544	Oomycete-like	HKQ	HKD	
		Ata1646402	N.A.	546	Oomycete-like	HKQ	HKD	
		Ata8314441	N.A.	434	III (SP)	HKD	HKD	
	Symbiodinium minutum	Smin009945	N.A.	608	? (PXPH-like)	HKD	HKD	
		Smin014079	N.A.	1235	PXPH-like	HKD	HKD	
		Smin018127	N.A.	931	? (Mito)	HKD	–	
		Smin025039 [d]	N.A.	1217	V (MITO)	HND	–	
		Smin025040	N.A.	159	V (MITO)	KKD	–	
		Smin030869	N.A.	286	V (MITO)	HKD	–	
		Smin038356	N.A.	Partial	? (N.C.)	HKR	–	
		Smin040183	N.A.	450	I	HKR	HKE	
		Smin044551	N.A.	Partial	? (N.C.)	HKR	–	

[a] Sequences that were eliminated from the trimmed alignment prior to tree reconstruction are indicated with a question mark. The closest phylogenetic group is indicated between parentheses and was estimated by comparing the scores obtained by each sequence when confronted to group-specific hmm profiles (see Methods). N.C., non conclusive results.

[b] Catalytic domains are shown as follows: PLDc = HKD, indicates Pfam significant hit to PLD catalytic domain (Pfam domain PF00614) containing an HKD motif ; PLDc_2 = HKD, Pfam significant hit to PLD-like domain 2 (PF13091) containing an HKD motif; PLDc_ 3, Pfam significant hit to PLD-like domain 3 (PF13918) *not* containing an HKD motif; Where only HKD1 and HKD2 are written, prediction of the catalytic motif was positive according to our HMMER profiling, but insignificant to Pfam. Other domains are: KH-1, K homology domain (PF00013); C2, C2 domain (PF00168); PLD_C, PLD C terminal domain (PF12357); Lsr2, Mycobacterium DNA-bridging domain (PF11774); zf-GRF, GRF zinc finger (PF06839); DnaJ, DnaJ domain (PF00226)

[c] The pairs Aano000842-Aano000843 and Ppur006012-Ppur006013 most likely correspond to single polypeptides.

[d] Smin0025039 most likely corresponds to two separate proteins.

cus tauri, *Ostreococcus lucimarinus*, *Bathycoccus prasinos*, *Micromonas pusilla*, and *Micromonas* sp., in addition to single members from *S. minutum* and *E. huxleyi*. The members of this clade have Pfam PLDc_2 catalytic domains, mentioned in the Introduction section, as either significant or insignificant hits

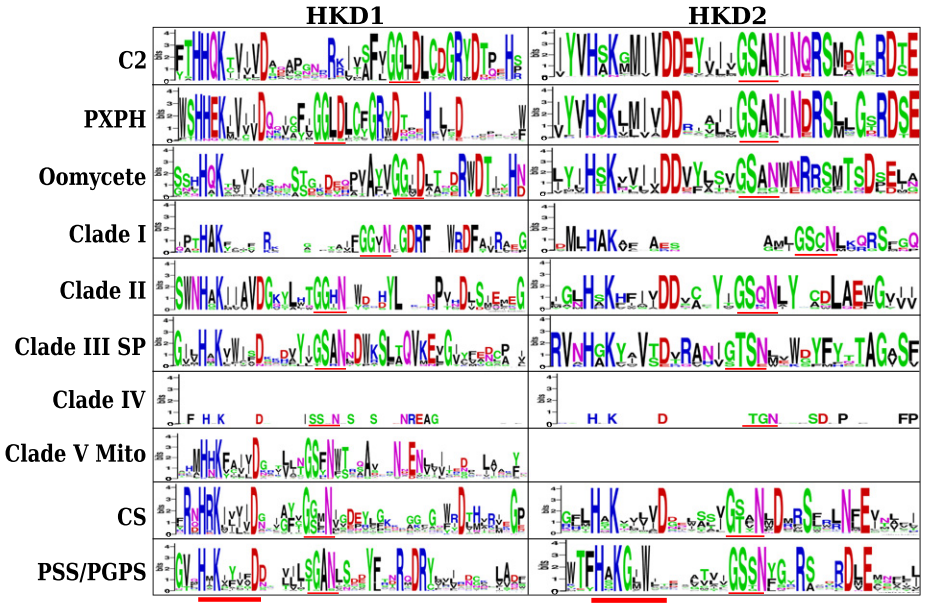


FIG. 4. Sequence logos of the HKD motifs (HKD1 and HKD2) of each group within the HKD superfamily. The logos were organized to align the HKD motifs. The second region of high conservation, the GGxN motif (and its variations), is shown with bars at the bottom of the Figure to guide logo comparison. The only clade in which all the polypeptides have a single full HKD motif is Bacteria-like Clade V (Mito-PLDs). Sample size correction was applied. Bit scores are shown on the *Y*-axes.

(Table 2). Four of the proteins within this clade have D-to-E replacements in the catalytic aspartic acid (D) in HKD2, while there is no conservation in this position in HKD1, where we observed replacements to H, Q, K, R, A, or E (Table 2, Fig. 4 and Appendix S1).

Several members within this clade have features that question what the functions of these proteins really are. For example, the *O. lucimarinus* protein has, toward the amino terminus, a domain similar to Lsr2, a small, basic DNA-bridging protein present in *Mycobacterium* and related actinomycetes. Lsr2 is a functional homolog of the H-NS-like proteins from gram-negative bacteria, which play a role in nucleoid organization and also function as pleiotropic regulators of gene expression (Gordon et al. 2008).

Although the *O. tauri* protein within this clade appears as a PLDc_2 member and Pfam does not detect it as homolog of the NgoFVII subfamily within clan CL0479, its HKD motifs resemble those of the catalytic domain of type II restriction endonucleases BfiI and NgoFVII (Zaremba et al. 2004, Rao et al. 2014).

The gene models of the two *Micromonas* sp. proteins within this clade have additional domains as well. For example, one of them has, toward the carboxy end, a domain with high similarity to lipoate-protein ligase B (BPL_LplA_LipB in Pfam), a group of proteins that ligate cofactors, such as biotin and lipoate, to acyl carrier proteins (Christensen and Cronan 2009). However, these two protein sequences have been subsequently modified in Uniprot and NCBI (as shown in Table 2), indicating that these additional domains might in fact correspond to different proteins. Further refinement of gene models and EST sequencing will certainly allow to determine the precise nature of these proteins.

Clade II. The eukaryotic members of this subgroup are exclusively from algae: several PLDs from diatoms, *E. huxleyi* and *A. tamarense* (Fig. 2 and Table 2). In addition, this clade includes several bacterial PLDs, including a member from the Gammaproteobacteria *Xanthomonas campestris* and secreted proteins from the Acidobacteria *Koribacter versatilis* and from two proteins from the *Streptomyces* genus (Actinobacteria; Fig. 3). The catalytic domains of all the members of this clade have invariant HKD motifs (Fig. 4 and Appendix S1) and belong to Pfam PLD-like domain 2 (PLDc_2), although in some cases the hits are insignificant according to the Pfam domain gathering threshold (Table 2). Protein localization predictions suggest that the algae proteins within this clade are soluble and, with the exception of the *A. tamarense* and the *E. huxleyi* proteins, most likely cytosolic (Fig. 2).

Clade III: SP-PLDs. This subgroup consists of PLDs from several taxonomically diverse eukaryotic species, which include a few land plants, opisthokonts and amebozoans, a few nonphotosynthetic protists and single PLDs from the algae *A. anophagefferens*, *B. natans*, and *A. tamarense* (Fig. 3). This group contains the previously described SP-PLD from rice (*O. sativa*; Li et al. 2007). The N-terminal catalytic domain of all the proteins within this clade belongs to Pfam PLD-like domain 2 (Table 2, PLDc_2 = HKD1), whereas the C-terminal catalytic domain belongs to PLD-like domain 3 (Table 2, PLDc_3). In addition, the HKD2 of the algae homologs within this clade is, similar to the rice counterpart, a canonical HKD, unlike the mammalian SP-PLDs, in which there is a D-to-E replacement (Li et al. 2007, Selvy et al. 2011). In this analysis, only a protein from the nonphotosynthetic

protist *Naegleria gruberi* showed such replacement (Appendix S1).

In addition to the sequence similarity between the various members of this group, protein localization prediction shows that, besides the rice protein, many other counterparts might enter the secretory pathway as well (Fig. 2). Although many of the proteins within this clade have TM domains, this does not appear to be the case for the algae members (Fig. 2).

Clade IV. This subgroup is composed of single sequences from eukaryotes, including the prasinophytes *O. tauri*, *O. lucimarinus* and *B. prasinos*, in addition to single members from the amoebozoan *Dictyostelium purpureum* and the protist *N. gruberi* (Fig. 3). Although the catalytic domains of these proteins belong to Pfam PLDc-like domain 2 (Table 2, PLDc_2 = HKD), the surroundings of the HKD motifs are very different from the rest of the clades (Fig. 4 and Appendix S1). A remarkable example is constituted by a stretch of prolines very close to each other (e.g., PxxxPxPxxPxP or PxxxPxPxxPxPP) downstream of HKD2.

Clade V: Mito-PLDs. This clade is one of the most diverse within the bacteria-like HKD superfamily. It is composed of members from both the eukaryotic and prokaryotic domains. Within the prokaryotic domain, there are members from Archaea, Cyanobacteria, Firmicutes, Chloroflexi, and both Alpha- and Gammaproteobacteria (Fig. 3). The eukaryotic proteins within this clade include single members from chlorophytes, including the only HKD superfamily members identified in the volvocales *Chlamydomonas reinhardtii* and *Volvox carteri*. This group also contains seven quite similar sequences from *E. huxleyi*, three from *S. minutum* and single proteins from both *Micromonas* species, *O. lucimarinus*, *B. prasinos* and the glaucophyte *C. paradoxa*, along with a few sequences from ciliate protists and several from opisthokonts and the amoebozoan *D. purpureum* (Fig. 2, see also Table 2 and Fig. 3). This clade includes mammalian Mito-PLD (also known as Phospholipase D6 or cardiolipin hydrolase), a protein proposed to generate PA from cardiolipin hydrolysis on the mitochondrial surface (Choi et al. 2006).

Most of the algae proteins within clade V have a single PLD catalytic domain; most of them have between ~200 and 450 amino acids and clearly constitute half a full-length PLD (Table 2). However, two of the *E. huxleyi* proteins are much larger, similar to full PLDs (Table 2, 863 and 777 amino acids), and possess parts of a second PLD-like domain, albeit that the catalytic HKD motifs are absent or compromised (Table 2 and Appendix S1). As an example, one of them (Uniprot ID: R1BLR0) has a deletion of the catalytic D and the minimum amount of 6 amino acids that seems to be necessary between the catalytic HxK and the downstream GSxN, which argues that this motif might not be functional. These two proteins possess a single TM domain at their amino-terminal portion (Fig. 2), indicating they could be anchored to a membrane with both ends facing opposite sides. According to their predicted subcellular localization, one of these proteins (Eh00013699, Uniprot ID: R1BLR0) is likely to be located on the plasma membrane, with the catalytic domain in the extracellular medium. The other protein (Eh00015190, Uniprot ID: R1D9S8), could be anchored to the outer chloroplast envelope, with the enzymatic activity facing the cytosol (Fig. 2). Two of the Mito-PLDs from *S. minutum* also show TM domains. One of them, Smin030869 (gene ID symbB1.v1.2.027300.t1) is predicted to be secreted, having one N-terminal TM domain (Fig. 2) and, hence, both ends of the protein facing opposite sides of the plasma membrane. The other protein from *S. minutum*, Smin025039 (gene ID symbB1.v1.2.022118.t1) is also predicted to enter the secretory pathway and to have 5 TM domains. However, the presence of a DNaJ domain at the C terminus of this sequence and its unusual length (Table 2) suggest that the annotated gene model for this protein might be wrong and the PLDc_2 and DnaJ domains actually belong to two separate proteins. Since the TM domains are located in the middle portion of the protein (which has no similarity to any reported domain according to Pfam), it remains to be determined whether any of them is present in this PLD.

The algae proteins within clade V are phylogenetically related to Mito-PLDs, but their predicted subcellular localization seems to vary among the different members (Fig. 2). Five of the proteins within this clade are predicted to be targeted to mitochondria (Fig. 2).

The *C. reinhardtii* Mito-PLD has an overall sequence identity of 40% and a similarity of 58% when compared with its mouse counterpart (by blastp). Modeling of the *C. reinhardtii* protein on mouse Mito-PLD yielded positive results with 100% confidence and showed that the two proteins correspond very closely (Fig. 5A). The formation of homo-dimers in the *C. reinhardtii* homolog could be positively identified and predicted with high confidence when compared with the homo-dimer of *Salmonella typhimurium* PLD-like protein Nuc (Fig. 5, B and C), suggesting that *C. reinhardtii* Mito-PLD also homo-dimerizes for catalytic activity.

DISCUSSION

By computational analysis, we identified a set of algae proteins from the HKD superfamily, including polypeptides with putative PLD activity. Our iterative HHMER profiling identified all the PLD family members previously reported for studied plants, such as rice, as well as other organisms, such as the oomycete *P. infestans*. This indicates that all the HKD superfamily members were most likely identi-

fied from the algae data set, provided that a translated gene model exists for that polypeptide.

Our results show that PLD homologs have evolved into two major groups, here named eukaryotic-type PLDs and bacteria-like PLD homologs. The first consists of eukaryotic sequences only, whereas the latter includes both eukaryotic and prokaryotic homologs. Several pieces of evidence suggest that the evolution of the bacteria-like HKD superfamily has been more complex than the evolution of eukaryotic-type PLDs. First, the number of generations required to achieve the convergence of splits between runs in bayesian analysis was higher for some of the bacteria-like PLD subclades, and there was a higher heterogeneity between compatible trees for some of these subclades, hence producing lower support values in the consensus. Second, all of the eukaryotic-type homologs characterized biochemically are known to act as PLDs, whereas bacteria-type homologs have several distinct enzymatic activities, being functionally much more diverse as a whole. This is demonstrated by the higher degree of degeneration that appears to be allowed in the HKD motif for the non-PLD members of the bacteria-like HKD superfamily. For instance, Pfam description of PLDc_3 (in terms of logo and HMMER profile) corresponds to a region that does not contain the HKD motif, since many sequences within that subfamily have $QxSx_4D$ or $LxCx_4D$ in HKD1 and $NxKx_4D$ in HKD2. However, many other PLDc_3 homologs do have two perfect HKD motifs. It can be envisaged that during the evolution of the bacteria-like HKD superfamily, various homologs have retained the PLD function, likely related to at least one perfect HKD motif, while others might have acquired new functions. In this work, sequence data mining was performed on the basis of general similarity and the presence of at least one catalytic motif, without discriminating whether it consisted of a perfect or imperfect HKD. Subsequent phylogenetic clustering and biocomputational analysis was used to shed light onto which sequences likely correspond to PLD enzymes and which likely correspond to other, biochemically related, functions. Although this was fairly straightforward for the algae PLD homologs that clustered with well-characterized proteins (e.g., CSs, PSSs/PGPSs, or C2-, SP- and Mito-PLDs), it was not the case for all clades. A remarkable example is provided by bacteria-like Clade I. The presence in some of the members of this clade of domains with similarity to DNA-modifying enzymes opens the question of whether the proteins within clade I function as PLDs or as bacteria-type endonucleases. Another example is the phylogenetic relationship between Clades III, IV, and V of bacteria-like PLDs with PPKs. Although they

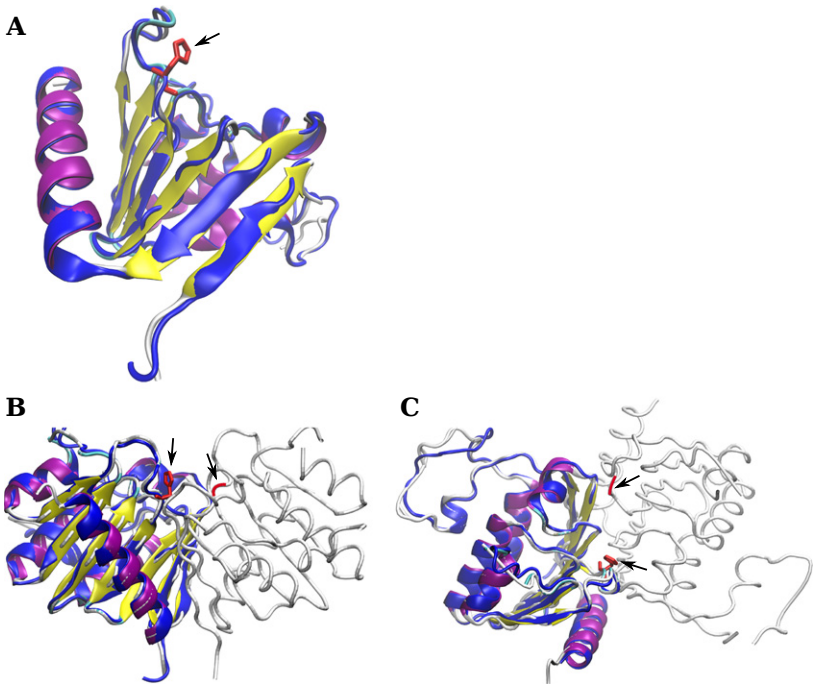


FIG. 5. Prediction of *Chlamydomonas reinhardtii* Mito-PLD tertiary structure. (A) Cartoon representation of the model of the *C. reinhardtii* Mito-PLD homolog (darker cartoon) structurally aligned with the mouse Mito-PLD (lighter cartoon) used for modeling. The side chain of the catalytic histidine of the model is shown as licorice and indicated wth an arrow, β-sheets and α-helices are specifically shown as cartoon, the rest of the model is shown as a tube. (B) Side view of the *C. reinhardtii* Mito-PLD homo-dimer modeled on *Salmonella typhimurium* HKD superfamily member endonuclease Nuc. The monomer model of (A) (shown as cartoon) is structurally aligned to the dimer model (shown as tube). β-sheets, α-helices and catalytic histidines are as indicated as in A. (C) Top view of the homo-dimer, structurally aligned with the monomer, showing the catalytic pocket between the two histidines. β-sheets, α-helices and catalytic histidines are as indicated in A.

appear phylogenetically related, the PPK clade seems to have distanced itself considerably from the last shared ancestor. According to Pfam, the members of Clade III, IV, and V all have PLD domains with at least one intact HKD motif, whereas the members of the PPK clade only have PPK domains, which usually are HRX. It is therefore unlikely that Clades III, IV, and V have PPK activity or viceversa, since many PPKs and Mito-PLDs have been thoroughly characterized biochemically. More likely, this case reaffirms the idea that the bacteria-like HKD superfamily has evolved in a very complex and extensive manner, resulting in a higher functional diversification than eukaryotic-type PLDs.

Analysis of the algae PLD family reveals that most algae PLDs have features typical of bacteria. First, with the exception of the C2-PLDs and the two *Nannochloropsis* PX-like-containing proteins identified, all algae PLDs lack the canonical regulatory domains. Second, many algae PLDs possess single HKDs, similar to Nuc-type bacterial endonucleases and Mito-PLDs. Single HKD PLDs are considered ancestral forms of the enzyme, originated prior to the events of gene duplication and fusion proposed to have generated the more complex plant and animal homologs (Pointing and Kerr 1996).

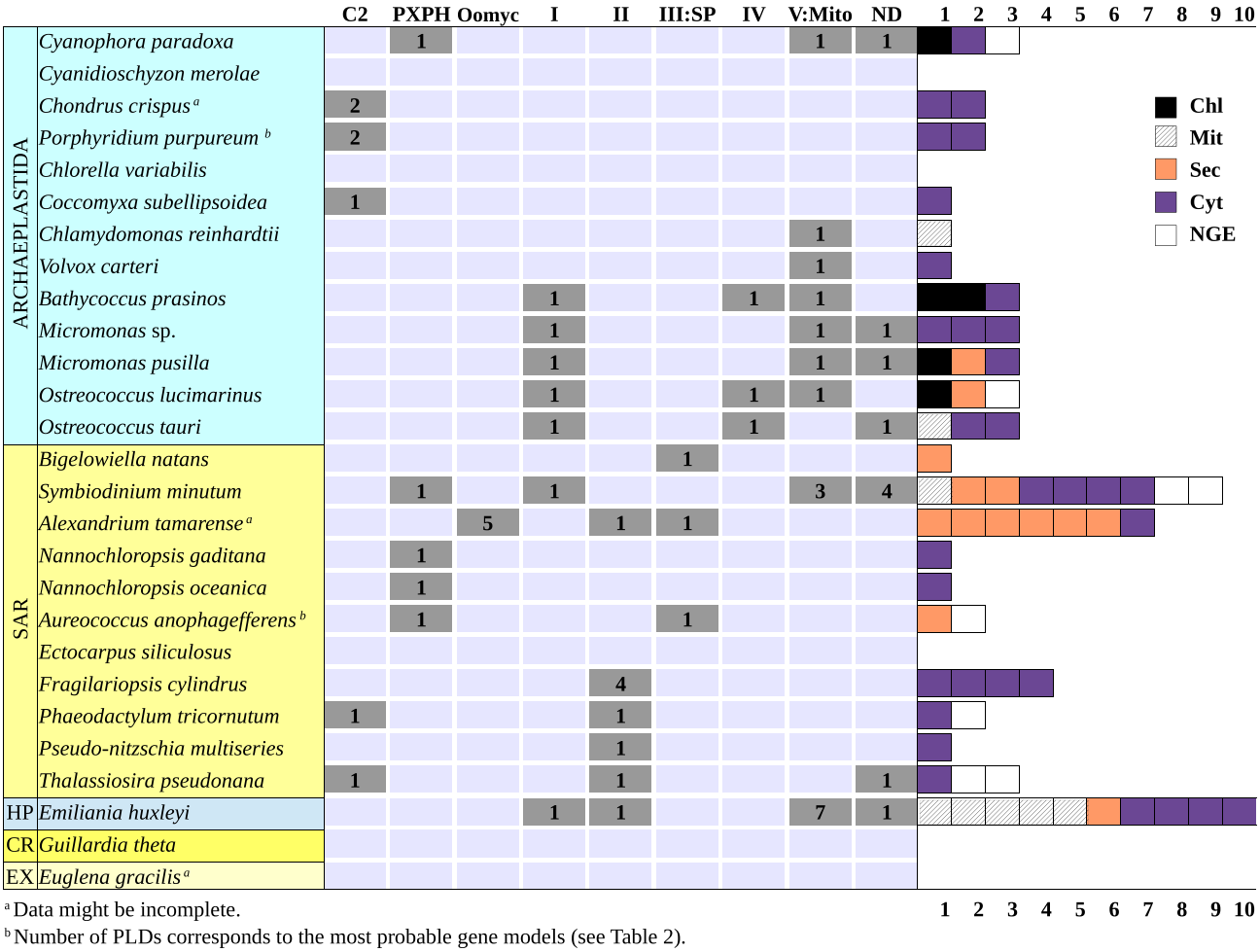
A striking fact is that some algae seem to lack HKD PLDs. There are at least two possibilities with these algae regarding PA signaling and metabolism: the first one is that they might generate PA through the action of non-HKD PLDs. The second is that they might produce PA through a PLD-independent pathway, which involves diacylglycerol kinase (Arisz et al. 2009, Testerink and Munnik 2011). In fact, algae that do have HKD PLDs might also generate PA via the mentioned alternatives, since the three pathways are by no means mutually exclusive.

Algae in the Archaeplastida supergroup have very few PLD isoforms compared with plants, but they are much more diverse. As mentioned above, most PLDs in chlorophytes resemble ancestral PLDs from bacteria, and very few seem to be related to the more complex plant or animal homologs. Indeed, all plants analyzed exhibit many C2-PLDs that are very similar to one another, whereas only one chlorophyte and two red algae seem to have one or two members of this type (Fig. 6). From a functional point of view, this provides insights into the possible differences between algae and plant PLDs. In plants, the existence of a plethora of similar PLD isoforms accounts for the high complexity typical of pluricellular organisms, having the large repertoire of enzymes similar metabolic and signaling mechanisms, with small differences related to tissue specificity, metabolic state (e.g., biotic or abiotic stress) or the developmental stage, among others. In algae, the modest number of very diverse PLDs is likely a reflection of very different roles within a few signaling and metabolic pathways.

Algae outside of Archaeplastida (SAR and Haptophyta) have, in general, many more bacteria-like superfamily members with putative PLD activity than most nonphotosynthetic protists, with the exception of oomycetes. In contrast, bacterial CSs and phosphatidyl synthases seem to be more widespread among non photosynthetic protists, such as apicomplexan parasites and ciliates. There are also some interesting differences in the composition of PLDs between algae in the different supergroups. Most PLDs from Archaeplastida are cytosolic, and only two are predicted to enter the secretory pathway. In contrast, five algae species within SAR have PLDs that seem to be secreted (Fig. 6). SAR and Haptophyta do not possess enzymes with chloroplast target peptides, whereas there are five sequences within Archaeplastida that could be targeted to plastids. The distribution of PLD classes also differs between the supergroups: clades II and III of bacteria-like PLDs are exclusive of algae from SAR (and one from Haptophyta), while clades I and IV are, with only two exceptions, exclusive of Archaeplastida (specifically prasinophytes). Mito-PLDs are not widespread in SAR, considering that only the dinoflagellate *S. minutum* seems to have this type of protein within that supergroup.

In this work, a group of PLDs that seems to be specific to oomycetes showed homologs in the dinoflagellate *A. tamarense*. This group of PLDs is not present in either other algae or the rest of the eukaryotes used for this analysis. Furthermore, oomycete-like PLDs are not present in the genome of the only other dinoflagellate analyzed, *S. minutum*, which has a complete genome sequence draft (Shoguchi et al. 2013). This provides evidence that the presence of this type of protein might not be a general rule in dinoflagellates.

There are three salient exceptions to the reduced number of PLDs in algae, not surprisingly the two dinoflagellates *A. tamarense* and *S. minutum* and the haptophyte *E. huxleyi*. The genomes of dinoflagellates are notoriously large, ranging from 1.5 to 200 Gb (McEwan et al. 2008, Hou and Lin 2009, Shoguchi et al. 2013). *A. tamarense* cells contain ~143 chromosomes and have a genome size of 200 pg/cell (~200,000 Mb), over 60 times the size of the human genome (Hackett and Bhattacharya 2008). In addition, dinoflagellate genomes are extremely complex with genes arranged in tandem arrays, spliced leader (SL) trans-splicing and mRNA transposition mechanisms (Hackett et al. 2005). Analysis of SL sequences in *Alexandrium ostenfeldii* provided evidence for frequent retro-transposition of mRNA species, which probably contributes to overall genome complexity by generating additional gene copies (Jaeckisch et al. 2011). Sequencing of the *S. minutum* genome also revealed a high level of gene duplication (Shoguchi et al. 2013). In *Symdiodinium kawagutii*, more than 20 highly similar SL sequences have been identified, and these were



[a] Data might be incomplete.

[b] Number of PLDs corresponds to the most probable gene models (see Table 2).

FIG. 6. Summary of PLDs identified in algae, classified into their respective groups. A list of algal species is shown on the left. In the central panel, the number of PLD paralogs present in each species is indicated with dark gray boxes. C2, PXPH-like (PXPH), Oomycete-like (Oomyc) and Clades I, II, III (III: SP), IV, and V (V: Mito) of Bacteria-like PLDs are shown. ND indicates the sequences that were not included in the tree and, hence, could not be reliably classified. For *Alexandrium tamarense*, *Euglena gracilis* and *Chondrus crispus*, data might be incomplete. The right panel shows the distribution of predicted protein localization in each species. HP, Haptophyta; CR, Cryptophyta; EX, Excavata; Chl, chloroplast; Mit, mitochondria; Sec; secretory pathway; Cyt, cytosol/other; NGE, not good enough evidence for prediction, sequences do not have starting methionines.

found in essentially all the full-length cDNAs within a 5′-cap selective library (Zhang et al. 2013). These kinds of mechanisms could be operating in dinoflagellates for the generation of multiple copies of genes coding for the several PLDs identified. Unfortunately, *A. tamarense* PLD RNAs extracted from NCBI are not full length, making it impossible to identify the trans-splicing SL consensus sequence most typical of dinaflagellates, "tccgtagccattttggctcaag" (Jaeckisch et al. 2011) or its variations.

Although the estimated genome size of *E. huxleyi* is much smaller (0.14 Gb) than that of dinoflagellates, it is still significantly larger than those of other algae. An analysis of the *E. huxleyi* genome revealed substantial gene expansions in many families, for example specific to iron/macromolecular transport, posttranslational modification, cytoskeletal development and signal transduction, relative to other sequenced eukaryotic algae (Read et al. 2013). The average gene family size in this haptophyte is not only larger than in many other algae to which it was compared with but also similar to other organisms known to code for many proteins in multiple gene copies, such as *Caenorhabditis elegans* or *Phytophthora sojae* (Read et al. 2013).

Mito-PLDs and the Chlamydomonas *case.* Cardiolipin hydrolase has been named Mito-PLD because it has been found associated to the outer mitochondrial membrane of mammals, where it forms homodimers (Huang and Frohman 2009). This enzyme regulates mitochondrial shape through facilitating mitochondrial fusion and has been shown to hydrolyze cardiolipin to generate PA at the mitochondrial surface (Choi et al. 2006, Watanabe et al. 2011, Gao and Frohman 2012). PA would then act as a second

messenger mediating mitochondrial fusion. However, there is some controversy around Mito-PLD activity. Cardiolipin is mostly found in the inner mitochondrial membrane and modeling of the resulting N-terminally tethered Mito-PLD dimer on bacterial Nuc and PLD suggested that the lipid-hydrolyzing catalytic pocket should paradoxically be oriented away from the mitochondrial outer membrane surface (Choi et al. 2006). In addition, during spermatogenesis, Mito-PLD has also been shown to play a critical role in Piwi-interacting RNA (piRNA) biogenesis. PiRNAs constitute a large class of small noncoding RNA molecules expressed in animal cells that provide essential protection against the activity of mobile genetic elements. In this case, Mito-PLD acts as a nonspecific nuclease, cleaving either RNA or DNA single strand substrates with similar affinity (Watanabe et al. 2011, Nishimasu et al. 2012). Taking this into account, whether algae Mito-PLD homologs act as PLDs, endonucleases or both will have to be determined experimentally.

PLD activity was previously measured in the alga *C. moewusii* upon hyperosmotic stress, membrane depolarization and G-protein activation (Munnik et al. 1995, Meijer et al. 2002, Arisz et al. 2003, Arisz and Munnik 2011). In these situations, PE was shown to be the preferred substrate for PLD-dependent PA formation (Arisz et al. 2003). However, the described enzymatic activity was measured in vivo and not attributed to a particular polypeptide. In this analysis, the only PLD-like polypeptide found by HHMER profiling consisted of a cardiolipin hydrolase homolog. Although this raises the question of whether this protein could be responsible for the PLD activity reported in *C. moewusii*, it is not surprising that a single PLD polypeptide could hydrolyze more than one phospholipid. Mito-PLD, whose substrate preference was only analyzed in mammals, also showed some activity on PE (Choi et al. 2006). In addition, since cardiolipin is not a substrate of canonical PLDs, a possible contribution of this lipid to PLD-mediated PA production in *C. moewusii* was not evaluated in the work by Arisz et al. (2003).

Another possibility is that PLD activity in *Chlamydomonas* comes from a non-HKD PLD. Since non-HKD PLDs are not related to one another, they are not suitable for alignment-based similarity identification. Most non-HKD PLDs share a dependence on divalent cations and a poor catalysis of the transphosphatidylation reaction (TPR, in which primary alcohols act as phosphatidyl acceptors instead of water). There are very few non-HKD PLDs that can perform TPRs efficiently, such as human autotaxin (Tania et al. 2010). PLD activity in *C. moewusii* was measured via the transfer of phosphatidyl groups to 1-butanol (Munnik et al. 1995, Munnik 2001). In this scenario, did *C. moewusii* PLD enzymatic activity come from a non-HKD PLD, it would have to be one that can carry out TPRs, of either a known or a novel type.

The Nannochloropsis *case.* In the *Nannochloropsis* genus, HHMER profiling identified a putative PSS/PGPS and two PXPH-like PLDs with catalytic motifs in which the histidine (H) was replaced with an asparagine (N). Although it has been shown that mutation or chemical modification of the catalytic H in several PLDs to N renders the proteins inactive (Gottlin et al. 1998, Rudolph et al. 1999), there are several HKD superfamily members that naturally have an N instead of the catalytic H. The protein p37K from the Poxviridae viral family, for example, is a HKD superfamily member with HKD1 = LCD and HKD2 = NKD. P37K gene deletion prevents formation of the extracellular enveloped virus form and cell to cell viral transmission (Blasco and Moss 1991). The finding that this protein is a PLD homolog has led to the proposal that p37K is involved in processing of the phospholipids of the extracellular enveloped virus envelope acquired from the host's trans-Golgi network, which could be necessary for viral envelope development (Pointing and Kerr 1996, Selvy et al. 2011). For the case of *Nannochloropsis*, it remains to be determined whether the corresponding mRNAs are translated and whether the resulting proteins have PLD activity (Blasco and Moss 1991).

We thank Dr. Mauricio Sica for critical reading of the manuscript. This study was supported by funds to M.V.B. from the Argentinean Consejo Nacional de Investigaciones Científicas y Tecnológicas (CONICET-PIP 11420110100090). MVB, CB, AML, and ATH are CONICET Researchers. MBP and HB were funded with a CONICET Postdoctoral Fellowship.

Adl, S. M., Simpson, A. G., Lane, C. E., Lukes, J., Bass, D., Bowser, S. S., Brown, M. W. et al. 2012. The revised classification of eukaryotes. *J. Eukaryot. Microbiol.* 59:429–93.

Arisz, S. A. & Munnik, T. 2011. The salt stress-induced LPA response in *Chlamydomonas* is produced via PLA_2 hydrolysis of DGK-generated phosphatidic acid. *J. Lipid Res.* 52:2012–20.

Arisz, S. A., Testerink, C. & Munnik, T. 2009. Plant PA signaling via diacylglycerol kinase. *Biochim. Biophys. Acta* 1791:869–75.

Arisz, S. A., Valianpour, F., van Gennip, A. H. & Munnik, T. 2003. Substrate preference of stress-activated phospholipase D in *Chlamydomonas* and its contribution to PA formation. *Plant J.* 34:595–604.

Bargmann, B. O. & Munnik, T. 2006. The role of phospholipase D in plant stress responses. *Curr. Opin. Plant Biol.* 9:515–22.

Blasco, R. & Moss, B. 1991. Extracellular vaccinia virus formation and cell-to-cell virus transmission are prevented by deletion of the gene encoding the 37,000-Dalton outer envelope protein. *J. Virol.* 65:5910–20.

Choi, S., Huang, P., Jenkins, G. M., Chan, D. C., Schiller, J. & Frohman, M. A. 2006. A common lipid links Mfn-mediated mitochondrial fusion and SNARE-regulated exocytosis. *Nat. Cell Biol.* 8:1255–62.

Christensen, Q. H. & Cronan, J. E. 2009. The *Thermoplasma acidophilum* LplA-LplB complex defines a new class of bipartite lipoate-protein ligases. *J. Biol. Chem.* 284:21317–26.

Criscuolo, A. & Gribaldo, S. 2010. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol. Biol.* 10:210.

Crooks, G. E., Hon, G., Chandonia, J. & Brenner, S. E. 2004. WebLogo: a sequence logo generator. *Genome Res.* 14:1188–90.

Eddy, S. R. 1998. Profile hidden Markov models. *Bioinformatics* 14:755–63.

Eliás, M., Potocký, M., Cvrcková, F. & Zárský, V. 2002. Molecular diversity of phospholipase D in angiosperms. *BMC Genom.* 3:2.

Emanuelsson, O., Brunak, S., von Heijne, G. & Nielsen, H. 2007. Locating proteins in the cell using TargetP, SignalP and related tools. *Nat. Protoc.* 2:953–71.

Gao, Q. & Frohman, M. A. 2012. Roles for the lipid-signaling enzyme MitoPLD in mitochondrial dynamics, piRNA biogenesis, and spermatogenesis. *BMB Rep.* 45:7–13.

Gordon, B. R. G., Imperial, R., Wang, L., Navarre, W. W. & Liu, J. 2008. Lsr2 of Mycobacterium represents a novel class of H-NS-like proteins. *J. Bacteriol.* 190:7052–9.

Gottlin, E. B., Rudolph, A. E., Zhao, Y., Matthews, H. R. & Dixon, J. E. 1998. Catalytic mechanism of the phospholipase D superfamily proceeds via a covalent phosphohistidine intermediate. *Proc. Natl. Acad. Sci. USA.* 95:9202–7.

Gouy, M., Guindon, S. & Gascuel, O. 2010. SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol. Biol. Evol.* 27:221–4.

Gschloessl, B., Guermeur, Y. & Cock, J. M. 2008. HECTAR: a method to predict subcellular targeting in heterokonts. *BMC Bioinformatics* 9:393.

Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W. & Gascuel, O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* 59:307–21.

Hackett, J. D. & Bhattacharya, D. 2008. The genomes of dinoflagellates. *In* Katz, L. A. & Bhattacharya, D. [Eds.] *Genomics and Evolution of Microbial Eukaryotes.* Oxford University Press, New York, pp. 48–63.

Hackett, J. D., Scheetz, T. E., Yoon, H. S., Soares, M. B., Bonaldo, M. F., Casavant, T. L. & Bhattacharya, D. 2005. Insights into a dinoflagellate genome through expressed sequence tag analysis. *BMC Genom.* 6:80.

Hodgkin, M. N., Masson, M. R., Powner, D., Saqib, K. M., Ponting, C. P. & Wakelam, M. J. 2000. Phospholipase D regulation and localisation is dependent upon a phosphatidylinositol 4,5-biphosphate-specific PH domain. *Curr. Biol.* 10:43–6.

Hou, Y. & Lin, S. 2009. Distinct gene number-genome size relationships for eukaryotes and non-eukaryotes: gene content estimation for dinoflagellate genomes. *PLoS ONE* 4:e6978.

Huang, H. & Frohman, M. A. 2009. Lipid signaling on the mitochondrial surface. *Biochim. Biophys. Acta* 1791:839–44.

Humphrey, W., Dalke, A. & Schulten, K. 1996. VMD - visual molecular dynamics. *J. Molec. Graphics* 14:33–8.

Huson, D. H. & Scornavacca, C. 2012. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. *Syst. Biol.* 61:1061–7.

Interthal, H., Pouliot, J. J. & Champoux, J. J. 2001. The tyrosyl-DNA phosphodiesterase Tdp1 is a member of the phospholipase D superfamily. *Proc. Natl. Acad. Sci. USA* 98:12009–14.

Jaeckisch, N., Yang, I., Wohlrab, S., Glöckner, G., Kroymann, J., Vogel, H., Cembella, A. & John, U. 2011. Comparative genomic and transcriptomic characterization of the toxigenic marine dinoflagellate *Alexandrium ostenfeldii*. *PLoS ONE* 6:e28012.

Jang, J., Lee, C. S., Hwang, D. & Ryu, S. H. 2012. Understanding of the roles of phospholipase D and phosphatidic acid through their binding partners. *Prog. Lipid Res.* 51:71–81.

Jiroutová, K., Horák, A., Bowler, C. & Oborník, M. 2007. Tryptophan biosynthesis in stramenopiles: eukaryotic winners in the diatom complex chloroplast. *J. Mol. Evol.* 65:496–511.

Jmol Team. 2014. Jmol: an open-source Java viewer for chemical structures in 3D. Available at: http://www.jmol.org/ (last accessed 3 January 2014).

Johnson, L. S., Eddy, S. R. & Portugaly, E. 2010. Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics* 11:431.

Jones, D. R., Avila, M. A., Sanz, C. & Varela-Nieto, I. 1997. Glycosyl-phosphatidylinositol-phospholipase type D: a possible candidate for the generation of second messengers. *Biochem. Biophys. Res. Commun.* 233:432–7.

Katoh, K., Misawa, K., Kuma, K. & Miyata, T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30:3059–66.

Kolesnikov, Y. S., Nokhrina, K. P., Kretynin, S. V., Volotovski, I. D., Martinec, J., Romanov, G. A. & Kravets, V. S. 2012. Molecular structure of phospholipase D and regulatory mechanisms of its activity in plant and animal cells. *Biochemistry Mosc.* 77:1–14.

Kukkonen, J. P. 2014. Lipid signaling cascades of orexin/hypocretin receptors. *Biochimie* 96:158–65.

Laxalt, A. M., ter Riet, B., Verdonk, J. C., Parigi, L., Tameling, W. I., Vossen, J., Haring, M., Musgrave, A. & Munnik, T. 2001. Characterization of five tomato phospholipase D cDNAs: rapid and specific expression of LePLDbeta1 on elicitation with xylanase. *Plant J.* 26:237–47.

Lee, H. H., Choi, M. & Koh, E. 1989. Purification and characterization of the active site of phospholipase D. *Korean Biochem J.* 22:487–93.

Leiros, I., Secundo, F., Zambonelli, C., Servi, S. & Hough, E. 2000. The first crystal structure of a phospholipase D. *Structure* 8:655–67.

Lerchner, A., Mansfeld, J., Kuppe, K. & Ulbrich-Hofmann, R. 2006. Probing conserved amino acids in phospholipase D (*Brassica oleracea* var. *capitata*) for their importance in hydrolysis and transphosphatidylation activity. *Protein Eng. Des. Sel.* 19:443–52.

Lerchner, A., Mansfeld, J., Schäffner, I., Schöps, R., Beer, H. K. & Ulbrich-Hofmann, R. 2005. Two highly homologous phospholipase D isoenzymes from *Papaver somniferum* L. with different transphosphatidylation potential. *Biochim. Biophys. Acta* 1737:94–101.

Letunic, I. & Bork, P. 2011. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res.* 39:W475–8.

Li, W. & Godzik, A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22:1658–9.

Li, G., Lin, F. & Xue, H. 2007. Genome-wide analysis of the phospholipase D family in *Oryza sativa* and functional characterization of PLD beta 1 in seed germination. *Cell Res.* 17:881–94.

Liu, Q., Zhang, C., Yang, Y. & Hu, X. 2010. Genome-wide and molecular evolution analyses of the phospholipase D gene family in Poplar and Grape. *BMC Plant Biol.* 10:117.

McEwan, M., Humayun, R., Slamovits, C. H. & Keeling, P. J. 2008. Nuclear genome sequence survey of the Dinoflagellate Heterocapsa triquetra. *J. Eukaryot. Microbiol.* 55:530–5.

Meijer, H. J. G., Hassen, H. H. & Govers, F. 2011. *Phytophthora infestans* has a plethora of phospholipase D enzymes including a subclass that has extracellular activity. *PLoS ONE* 6: e17767.

Meijer, H. J. G., Latijnhouwers, M., Ligterink, W. & Govers, F. 2005. A transmembrane phospholipase D in Phytophthora; a novel PLD subfamily. *Gene* 350:173–82.

Meijer, H. J. G., ter Riet, B., van Himbergen, J. A. J., Musgrave, A. & Munnik, T. 2002. KCl activates phospholipase D at two different concentration ranges: distinguishing between hyperosmotic stress and membrane depolarization. *Plant J.* 31:51–9.

Mukherjee, S. & Zhang, Y. 2011. Protein-protein complex structure predictions by multimeric threading and template recombination. *Structure* 19:955–66.

Munnik, T. 2001. Phosphatidic acid: an emerging plant lipid second messenger. *Trends Plant Sci.* 6:227–33.

Munnik, T., Arisz, S. A., De Vrije, T. & Musgrave, A. 1995. G protein activation stimulates phospholipase D signaling in plants. *Plant Cell.* 7:2197–210.

Munnik, T., Irvine, R. F. & Musgrave, A. 1998. Phospholipid signalling in plants. *Biochim. Biophys. Acta* 1389:222–72.

Nicholas, K. B., Nicholas, H. B. & Deerfield, D. W. I. 1997. GeneDoc: analysis and visualization of genetic variation. *EMBnet News* 4:14.

Nishimasu, H., Ishizu, H., Saito, K., Fukuhara, S., Kamatani, M. K., Bonnefond, L., Matsumoto, N. et al. 2012. Structure and function of Zucchini endoribonuclease in piRNA biogenesis. *Nature* 491:284–7.

Nylander, J. A., Wilgenbusch, J. C., Warren, D. L. & Swofford, D. L. 2008. AWTY (are we there yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics. *Bioinformatics* 24:581–3.

Peters, N. T., Logan, K. O., Miller, A. C. & Kropf, D. L. 2007. Phospholipase D signaling regulates microtubule organization in the fucoid alga *Silvetia compressa*. *Plant Cell Physiol.* 48:1764–74.

Peters, N. T., Pol, S. U. & Kropf, D. L. 2008. Phospholipid signaling during stramenopile development. *Plant Signal Behav.* 3:398–400.

Petersen, T. N., Brunak, S., von Heijne, G. & Nielsen, H. 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods* 8:785–6.

Pleskot, R., Li, J., Zárský, V., Potocký, M. & Staiger, C. J. 2013. Regulation of cytoskeletal dynamics by phospholipase D and phosphatidic acid. *Trends Plant Sci.* 18:496–504.

Pointing, C. P. & Kerr, I. D. 1996. A novel family of phospholipase D homologues that includes phospholipid synthases and putative endonucleases: identification of duplicated repeats and potential active site residues. *Protein Sci.* 5:914–22.

Powner, D. J. & Wakelam, M. J. O. 2002. The regulation of phospholipase D by inositol phospholipids and small GTPases. *FEBS Lett.* 531:62–4.

Qin, C. & Wang, X. 2002. The Arabidopsis phospholipase D family. Characterization of a calcium-independent and phosphatidylcholine-selective PLD zeta 1 with distinct regulatory domains. *Plant Physiol.* 128:1057–68.

Rao, D. N., Dryden, D. T. F. & Bheemanaik, S. 2014. Type III restriction-modification enzymes: a historical perspective. *Nucleic Acids Res.* 42:45–55.

Read, B. A., Kegel, J., Klute, M. J., Kuo, A., Lefebvre, S. C., Maumus, F., Mayer, C. et al. 2013. Pan genome of the phytoplankton *Emiliania* underpins its global distribution. *Nature* 499:209–13.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A. & Huelsenbeck, J. P. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* 61:539–42.

Rudolph, A. E., Stuckey, J. A., Zhao, Y., Matthews, H. R., Patton, W. A., Moss, J. & Dixon, J. E. 1999. Expression, characterization, and mutagenesis of the Yersinia pestis murine toxin, a phospholipase D superfamily member. *J. Biol. Chem.* 274:11824–31.

Schäfer, P., Cymerman, I. A., Bujnicki, J. M. & Meiss, G. 2007. Human lysosomal DNase IIalpha contains two requisite PLD-signature (HxK) motifs: evidence for a pseudodimeric structure of the active enzyme species. *Protein Sci.* 16:82–91.

Selvy, P. E., Lavieri, R. R., Lindsley, C. W. & Brown, H. A. 2011. Phospholipase D: enzymology, functionality, and chemical modulation. *Chem. Rev.* 111:6064–119.

Shoguchi, E., Shinzato, C., Kawashima, T., Gyoja, F., Mungpakdee, S., Koyanagi, R., Takeuchi, T. et al. 2013. Draft assembly of the *Symbiodinium minutum* nuclear genome reveals dinoflagellate gene structure. *Curr. Biol.* 23:1399–408.

Singh, A., Pandey, A., Baranwal, V., Kapoor, S. & Pandey, G. K. 2012. Comprehensive expression analysis of rice phospholipase D gene family during abiotic stresses and development. *Plant Signal Behav.* 7:847–55.

Smékalová, V., Doskočilová, A., Komis, G. & Samaj, J. 2014. Crosstalk between secondary messengers, hormones and MAPK modules during abiotic stress signalling in plants. *Biotechnol. Adv.* 32:2–11.

Sonnhammer, E. L., Eddy, S. R., Birney, E., Bateman, A. & Durbin, R. 1998b. Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res.* 26:320–2.

Sonnhammer, E. L., Eddy, S. R. & Durbin, R. 1997. Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins* 28:405–20.

Sonnhammer, E. L., von Heijne, G. & Krogh, A. 1998a. A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 6:175–82.

Stuckey, J. A. & Dixon, J. E. 1999. Crystal structure of a phospholipase D family member. *Nat. Struct. Biol.* 6:278–84.

Sung, T. C., Roper, R. L., Zhang, Y., Rudge, S. A., Temel, R., Hammond, S. M., Morris, A. J., Moss, B., Engebrecht, J. & Frohman, M. A. 1997. Mutagenesis of phospholipase D defines a superfamily including a trans-Golgi viral protein required for poxvirus pathogenicity. *EMBO J.* 16:4519–30.

Tania, M., Khan, M. A., Zhang, H., Li, J. & Song, Y. 2010. Autotaxin: a protein with two faces. *Biochem. Biophys. Res. Commun.* 401:493–7.

Tardif, M., Atteia, A., Specht, M., Cogne, G., Rolland, N., Brugière, S., Hippler, M. et al. 2012. PredAlgo: a new subcellular localization prediction tool dedicated to green algae. *Mol. Biol. Evol.* 29:3625–39.

Testerink, C. & Munnik, T. 2011. Molecular, cellular, and physiological responses to phosphatidic acid formation in plants. *J. Exp. Bot.* 62:2349–61.

Tokumura, A., Majima, E., Kariya, Y., Tominaga, K., Kogure, K., Yasuda, K. & Fukuzawa, K. 2002. Identification of human plasma lysophospholipase D, a lysophosphatidic acid-producing enzyme, as autotaxin, a multifunctional phosphodiesterase. *J. Biol. Chem.* 277:39436–42.

Tzeng, C. M. & Kornberg, A. 2000. The multiple activities of polyphosphate kinase of *Escherichia coli* and their subunit structure determined by radiation target analysis. *J. Biol. Chem.* 275:3977–83.

UniProt Consortium 2012. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.* 40: D71–5.

Wakelam, M., Hodgkin, M., Martin, A. & Saqib, K. 1997b. Phospholipase D. *Semin. Cell Dev. Biol.* 8:305–10.

Wakelam, M. J., Martin, A., Hodgkin, M. N., Brown, F., Pettitt, T. R., Cross, M. J., De Takats, P. G. & Reynolds, J. L. 1997a. Role and regulation of phospholipase D activity in normal and cancer cells. *Adv. Enzyme Regul.* 37:29–34.

Wang, X. 2000. Multiple forms of phospholipase D in plants: the gene family, catalytic and regulatory properties, and cellular functions. *Prog. Lipid Res.* 39:109–49.

Wang, X. 2002. Phospholipase D in hormonal and stress signaling. *Curr. Opin. Plant Biol.* 5:408–14.

Watanabe, T., Chuma, S., Yamamoto, Y., Kuramochi-Miyagawa, S., Totoki, Y., Toyoda, A., Hoki, Y. et al. 2011. MITOPLD is a mitochondrial protein essential for nuage formation and piRNA biogenesis in the mouse germline. *Dev. Cell* 20:364–75.

Yoshikawa, F., Banno, Y., Otani, Y., Yamaguchi, Y., Nagakura-Takagi, Y., Morita, N., Sato, Y. et al. 2010. Phospholipase D family member 4, a transmembrane glycoprotein with no phospholipase D activity, expression in spleen and early postnatal microglia. *PLoS ONE* 5:e13932.

Zambonelli, C. & Roberts, M. F. 2003. An iron-dependent bacterial phospholipase D reminiscent of purple acid phosphatases. *J. Biol. Chem.* 278:13706–11.

Zambonelli, C. & Roberts, M. F. 2005. Non-HKD phospholipase D enzymes: new players in phosphatidic acid signaling? *Prog. Nucleic Acid Res. Mol. Biol.* 79:133–81.

Zaremba, M., Urbanke, C., Halford, S. E. & Siksnys, V. 2004. Generation of the BfiI restriction endonuclease from the fusion of a DNA recognition domain to a non-specific nuclease from the phospholipase D superfamily. *J. Mol. Biol.* 336:81–92.

Zhang, Y., Arakaki, A. K. & Skolnick, J. 2005. TASSER: an automated method for the prediction of protein tertiary structures in CASP6. *Proteins.* 61(Suppl. 7):91–8.

Zhang, H., Zhuang, Y., Gill, J. & Lin, S. 2013. Proof that dinoflagellate spliced leader (DinoSL) is a useful hook for fishing dinoflagellate transcripts from mixed microbial samples: *Symbiodinium kawagutii* as a case study. *Protist* 164:510–27.

Zhao, Y., Stuckey, J. A., Lohse, D. L. & Dixon, J. E. 1997. Expression, characterization, and crystallization of a member of the novel phospholipase D family of phosphodiesterases. *Protein Sci.* 6:2655–8.

Zuegge, J., Ralph, S., Schmuker, M., McFadden, G. I. & Schneider, G. 2001. Deciphering apicoplast targeting signals - feature extraction from nuclear-encoded precursors of *Plasmodium falciparum* apicoplast proteins. *Gene* 28:19–26.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's web site:

Table S1. Sources of the proteome files used for PLD biocomputational HHMER profiling.

Table S2. Detailed IDs and evidence of protein existence of algae PLDs identified by HMMER iterative profiling.

Appendix S1. Multiple sequence alignment of PLD homologs from selected algae and other representative taxa.

Appendix S2. Bayesian tree of PLD homologs from algae and other representative taxa reconstructed from the multiple sequence alignment shown in Appendix S1.

Appendix S3. Maximum likelihood tree of PLD homologs reconstructed from the multiple sequence alignment shown in Appendix S1.