AN ANCIENT ENZYME FINDS A NEW HOME: PREVALENCE AND NEOFUNCTIONALIZATION OF TRYPsin IN MARINE PHYTOPLANKTON1

Yanchun You, Xueqiong Sun
State Key Laboratory of Marine Environmental Science, College of Ocean and Earth Sciences, Xiamen University, Xiamen 361005, China

and Senjie Lin2
State Key Laboratory of Marine Environmental Science, College of Ocean and Earth Sciences, Xiamen University, Xiamen 361005, China
Laboratory of Marine Biology and Biotechnology, Qingdao National Laboratory of Marine Science and Technology, Qingdao 266237, China
Department of Marine Sciences, University of Connecticut, Groton, Connecticut 06340-6048, USA

Trypsin is an ancient protease best known as a digestive enzyme in animals, and traditionally believed to be absent in plants and protists. However, our recent studies have revealed its wide presence and important roles in marine phytoplankton. Here, to gain a better understanding on the importance of trypsin in phytoplankton, we further surveyed the distribution, diversity, evolution and potential ecological roles of trypsin in global ocean phytoplankton. Our analysis indicated that trypsin is widely distributed both taxonomically and geographically in marine phytoplankton. Furthermore, by systematic comparative analyses we found that algal trypsin could be classified into two subfamilies (trypsin I and trypsin II) and exhibited highly duplicated and diversified during evolution. We also observed markedly different domain sequences and organizations between and within the subfamilies, suggesting potential neofunctionalization. Diatoms contain both subfamilies of trypsin, with higher numbers of genes and more environment-responsive expression of trypsin than other lineages. The duplication and subsequent neofunctionalization of the trypsin family may be important in diatoms for adapting to dynamical environmental conditions, contributing to diatoms’ dominance in the coastal oceans. This work advances our knowledge on the distribution and neofunctionalization of this ancient enzyme and creates a new window of research on phytoplankton biology.

Key index words: adaptation; diatom; environmental stimuli; evolution; phylogenetics; tandem duplication; trypsin

Abbreviations: CDD, conserved domain database; Cys, cystine; DCM, deep chlorophyll maximum; Deg, degradation of periplasmic protein; ESTs, expressed sequence tags; HMM, hidden markov models; HtrA, high temperature requirement A; Ka/Ks, the nonsynonymous/synonymous substitution ratio; MATOUv1 + T, marine atlas of tara oceans unigenes + metaT(eukaryotes); MES, mesopelagic zone; MIX, mixed layer; OGA, ocean gene atlas; pI, isoelectric points; SAGER, symbiodiniceae and algal genomic resource; SMART, simple modular architecture research tool; SRF, surface; TL, temperature-limited; TPM, transcript per million; Tryp, trypsin

Trypsin (EC 3.4.21.4) is known as pancreatic serine proteolytic enzyme, which specifically cleaves the carboxyl end of the lysine and arginine residues in polypeptides. First discovered almost 150 years ago (Kühne 1867), trypsin is arguably the first enzyme known to science and the best studied protein, and probably is the best exploited enzyme for protein biotechnology. However, the occurrence and function across the tree of life is poorly understood and underexplored.

Although trypsin represents a conserved family of enzymes occurring in organisms ranging from bacteria to mammals (Rungruangsak-Torrissen and Male 2000, Chen et al. 2003), it has been believed to be absent in plants and protists (Rojas and Doolittle 2002, Querino Lima Afonso et al. 2020). According to the MEROPS nomenclature, there are 109 SIA serine proteases with species identification and annotated as trypsin enzymes, all of which are exclusively from animals. There is a dearth of information on plant/protist trypsins in the PubMed database. In a recent harmful algal bloom metatranscriptomic study, diatom trypsin genes were found to be highly expressed, accounting for 1% of the total diatom transcriptome, when diatoms were dominant and the community was evolving into a dinoflagellate bloom while undergoing a rapid
ambient phosphate decline (Zhang et al. 2019). A follow-up study revealed that a trypsin in the model diatom *Phaeodactylum tricornutum* played an important role in regulating the homeostasis of N:P nutrient stoichiometry (You et al. 2022). These raise questions as to how broadly trypsin occurs, how it has evolved, and whether there is functional differentiation among lineages, in phytoplankton.

In this study, we analyzed the *Tara* Oceans metatranscriptomic dataset to detect occurrence and quantify expression of trypsin family in phytoplankton in the global ocean and their relationship with environmental factors. We further documented sequence diversity, lineage-specific expansion and phylogenetic relationships of the trypsin family by mining existing genome data for nine species that represent the major phyla of algae. We also analyzed trypsin gene structure, evolutionary characteristics, and expression patterns in these lineages. Results indicate taxonomically and geographically wide distribution of trypsin and remarkable diversification in sequence and organization that group trypsin sequences into two subfamilies with potential functional innovation.

**MATERIALS AND METHODS**

Detection of trypsin family in the *Tara* Oceans datasets. Extensive search for trypsins was performed in both the *Tara* Oceans eukaryote unigene catalog and metatranscriptomes (MATOUv1 + metaT; Carradee et al. 2018, Villar et al. 2018), using the trypsin (Pfam ID: PF00089) and trypsin_2 (Pfam ID: PF13365, also described as trypsin-like peptidase) domain based on profile Hidden Markov Models (HMM), with an E-value ≤1.0 E^-10. MATOU is a catalog of 116 million unigenes obtained from poly-A cDNA sequencing for samples of different size fractions and different water layers, available at the OGA website (http://tara-oceans.mio.osupytheas.fr/oceangeneatlas/).

Identification of trypsins from sequenced algal genomes. Whole genome sequences were downloaded from the Ensembl Protists Database (http://protists.ensembl.org) and the NCBI Genome Database (https://www.ncbi.nlm.nih.gov/genome/).

To identify trypsins, the hmmsearch analysis was conducted. We downloaded the HMM profile of trypsin and trypsin_2 (PF00089 and PF13365) from Pfam protein family database (http://pfam.xfam.org/) and used it as the query (P-value <0.001) by hmmsearch from the whole genome protein sequences. Top hits of the search were selected to build the species-specific HMM for the 2nd round of hmmsearch to yield all trypsin genes from the selected genome. To avoid missing probable trypsin members because of incomplete trypsin domains, a BLASTP-algorithm based search was conducted using trypsin amino acid sequences from NCBI and UniProt database as queries with an e-value ≤1 e-5 as the threshold. After removing redundant sequences, the identified putative trypsin protein sequences were submitted to CDD (https://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml) and SMART (http://smart.embl-heidelberg.de/) to confirm the conserved trypsin domain. All the non-redundant and high-confidence genes were assigned as trypsin, named with abbreviated species names, *Cyndioschyzon merolae* (Cm), *Thalassiosira pseudonana* (Tp), *Phaeodactylum tricornutum* (Pt), *Thalassiosira oceanica* (To), *Fistulifera solaris* (Fs), *Fragilariaopsis cylindrus* (Fc), *Pseudo-nitzschia multiseries* (Pm), *Fugacium kawaguiti* (Fk), *Chlamydomonas reinhardtii* (Cr), followed by Tryp standing for trypsin, then the order of their position on the chromosome/scaffold/contig.

Structural analysis of identified trypsins. All of the high-confidence trypsin sequences we obtained were submitted to ExPASy (http://web.expasy.org/protparam/) to calculate the number of amino acids, molecular weights and theoretical isoelectric points (pI). The MEME program (version 4.11.2, http://alternate.meme-suite.org/tools/meme) was used to identify the conserved motifs in the trypsin sequences, with the following parameters: any number of repetitions, minimum of 10 misfits and an optimum motif width of 6–200 amino acid residues. The chromosomal positions of the trypsin genes were acquired from the genome datasets of the species. MapChart software (Voorrips 2002) was used for the mapping of trypsin genes’ chromosomal positions and relative distances. The Plant-mPloc (Chou and Shen 2010) and SignalP-6.0 (Teufel et al. 2022) were used to predict the subcellular localization and signal peptides of identified algal trypsin.

Phylogenetic analysis of identified trypsins. OrthoFinder (Emms and Kelly 2015, 2019) was used to identify the orthologous and paralogous groups among 9 marine phytoplankton species and the plant model *Arabidopsis thaliana*. All-versus-all BLASTP with an E-value cutoff of 1 e-05 were performed, orthologous and paralogous genes were clustered using OrthoFinder. Single-copy orthologous genes were extracted from the clustering results.

To examine the grouping of the identified trypsins, the deduced amino acid sequences of these genes from the nine species of algae with sequenced genomes were subjected to phylogenetic analysis. The alignment of the sequences was carried out using ClustalW (Thompson et al. 1994, Larkin et al. 2007) and inspected manually for necessary correction. An unrooted Maximum Likelihood phylogenetic tree was constructed using MEGA X (Kumar et al. 2018, Stecher et al. 2007) and inspected manually for necessary correction. The resulting tree file was visualized with iTol (https://itol.embl.de/).

Analyses of trypsin gene expression and correlation with environmental factors in the global ocean. The ambient nutrient conditions corresponding to trypsin expression data in the global ocean were extracted from PANGAEA (doi:10.1594/PANGAEA.836319; Pesant et al. 2015). The following nine environmental nutrient parameters were chosen for correlation analysis: iron_5m (Fe, μmol L^-1), ammonium_5m (NH_4^+, μmol L^-1), NO_2-NO_3 (μmol L^-1), NO_3^- (μmol L^-1), PO_4^3- (μmol L^-1), Si (μmol L^-1), NO_2^-5m (μmol L^-1), NO_3^-5m (μmol L^-1). The correlations of these environmental parameters with trypsin mRNA profiles were investigated using pairwise Spearman correlation analysis, with a “fdr”-adjusted P value. Correlations were considered significant with a P value <0.05 and an absolute coefficient >0.3.

Analysis of trypsin gene expression in diatoms and dinoflagellates in existing databases. The gene expression data of diatom trypsin genes (PtTryp for *Phaeodactylum tricornutum* and TpTryp for *Thalassiosira pseudonana* trypsin) and dinoflagellate trypsin genes (FKTryp for *Fugacium kawaguiti* trypsin) were downloaded from the Diatom EST Database (http://www.diatomics.biologie.envs.fr/EST3/), and SAGER Database (http://samgr.org.cn/index.php), respectively. The gene expression data included a range of environmental conditions, as detailed in websites. The counts of ESTs (Expressed Sequence Tags) for PtTryp and TpTryp and TPM (Transcript}
per million) value for FkTryp were used to analyze the expression pattern. The results visualized using TBtools (Chen et al. 2018).

RESULTS

Taxonomic and geographic distribution of trypsins in natural assemblages of marine plankton. To understand how broadly trypsin occurs in marine plankton, we mined the Tara Oceans datasets by using the Hidden Markov Models of trypsin and trypsin_2 based on hmmr search. For convenience, the trypsin and trypsin_2 domain-containing genes are named henceforth as trypsin I and trypsin II, respectively, and both combined are regarded as the trypsin family. The analysis of the Tara Oceans data yielded 129,512 and 6163 identified hits of Trypsin I and II, respectively, and showed that the trypsin family occurred in all global sampling sites across different sample depths, including surface (SRF), deep chlorophyll maximum (DCM), mesopelagic zone (MES), mixed layer (MIX). Trypsin I were more prevalent in SRF and DCM, accounting for 57.83% and 38.08% of the total identified trypsin I hits, respectively, and tended to be predominant in the 20–2000 µm size fraction of plankton (accounting for 60.91% and 62.06% of the total identified trypsin I hits in SRF and DCM, respectively) than in small size fractions (Fig. 1a). In contrast, trypsin II also showed SRF and DCM biased distributions, accounting for 63.26% and 31.41% of the total identified trypsin II hits, respectively, but were predominantly in the smaller sized organisms (0.8–20 µm, accounting for 55.47% and 64.30% of the total identified trypsin II hits in SRF and DCM, respectively; Fig. 1b). Taken together, trypsin I and trypsin II exhibited similar latitudinal and depth distribution but different size fractionwise distribution.

Taxonomically, 96% of the trypsin I transcripts detected was originated from metazoans, and a small percentage from dinoflagellates (2.0%), haptophytes (0.2%), diatoms (0.1%), and cryptophytes (0.1%; Fig. 1c). For trypsin II, in contrast, 41% of the hits were from haptophytes, 20% from stramenopiles (dominantly Pelagophyceae), 12% from dinoflagellates, 8% from chlorophytes, and only 6% from metazoans (Fig. 1d). Taken together, trypsin I and trypsin II are widely distributed in marine plankton both taxonomically and geographically, but are dominantly in Arthropoda and phytoplankton, respectively.

Abundance and evolutionary dynamics of trypsins in algal genomes. To better understand the lineage-specific diversity and expansion of the trypsin family during phytoplankton evolution, we searched for trypsins from available sequenced genomes across four major phyla of algae, including Rhodophyta (Cyanidioschyzon merolae), Chlorophyta (Chlamydomonas reinhardtii), Dinophyta (Fugacium kawagutii) and Bacillariophyta (Thalassiosira pseudonana, Phaeodactylum tricornutum, Thalassiosira oceanica, Fistulifera solaris, Fragilariopsis cylindrus and Pseudo-nitzschia multiseries). For comparison of trypsin genes between algae and land plants, the genome of the land plant model Arabidopsis thaliana was analyzed. A total of 291 algal trypsin genes and 16 A. thaliana trypsin genes were identified and named based on their species name abbreviations and chromosomal locations (Table S1 in the Supporting Information). The number of trypsin genes varies from 5 to 65 in the examined algal genomes (Fig. 2), accounting for 0.95% to 5.68% of their predicted proteomes.

To explore the evolutionary trend of trypsin, we inferred a species phylogenetic tree based on the NCBI Taxonomy Common Tree (https://www.ncbi.nlm.nih.gov/Taxonomy/CommonTree/wwwcmt.cgi) database and mapped the genome size, coding gene number, and the number of identified trypsin genes on the tree (Fig. 2). The strongly supported tree suggest that ancestral algae, dated prior to the Cyanidioschyzon merolae, likely already contained a repertoire of trypsin genes (5 genes). Based on the tree, it appears that diatoms have undergone a greater trypsin gene family expansion than dinoflagellates, which is disproportionate to their genome sizes (Fig. 2). However, the numbers of trypsin genes vary considerably, from 5 to 65, across alga lineages, even between diatom species, showing no phylogenetic trends. For instance, Phaeodactylum tricornutum and Fistulifera solaris are phylogenetically closely related, both belonging to the order of Naviculales, similar in genome size, but are very different in trypsin gene copy number (10 and 65 copies, respectively), indicating a trypsin gene expansion hotspot in F. solaris.

The trypsin gene also varies significantly between lineages in chromosomal locations, open reading frame lengths, amino acid numbers, molecular weights and isoelectric points (pI) (Table S1). Large differences can even be found between different members of the same lineage. For example, for
PtTryp alone, the length ranges from 366 (PtTryp3) to 726 amino acids (PtTryp8), and the molecular weights from 39.66 kDa (PtTryp5) to 76.63 kDa (PtTryp8).

Classification of trypsin family into two classic subfamilies. To analyze the evolutionary relationships of trypsins among the nine selected algae, an unrooted phylogenetic tree was constructed using their conserved amino acid sequences (Fig. 3). Consistently, the 291 algal trypsin genes were well separated into two distinct clades, which corresponded to the subfamilies defined above: clade I represents trypsin I (162 genes; Fig. 3) and clade II represents trypsin II (129 genes; Fig. 3). Most of the trypsin I homologs (97.53%) are originated from diatoms and dinoflagellates. By contrast, trypsin II homologs were present in all the selected nine species of algae spanning four phyla.

Remarkably, as shown in Figure 3, some trypsin family members were more closely related to counterparts in the same subfamily from different species than to the other trypsin genes from the same species, indicative of ancestral gene duplication. By contrast, some trypsin members were more closely related to those from the same species than to the other trypsin genes from different species, indicating recent gene duplication events within species. Within a subfamily, most trypsin genes were clustered by phylum (Fig. 3), suggesting expansion of trypsin subfamilies within each algal phylum. In Cyanidioschyzon merolae, an early diverging lineage of algae (Fig. 2), all of its trypsin genes appeared to belong to the trypsin II subfamily. In contrast, there were no trypsin I genes identified in bacteria based on our analysis (data not shown). These results suggest that trypsin II is basal and trypsin I might have arisen from trypsin II duplication and subsequent divergence.

Trypsin gene clusters and evidence of duplications and differential losses. We noted that trypsin genes are not randomly distributed on different chromosomes, scaffolds or contigs, and some algal trypsin genes exist as gene clusters (Fig. S1 in the Supporting Information). Based on current genome assembly level, four of the nine alga genomes we examined contain trypsin genes organized in clusters, as shown in Figure S1, Chlamydomonas reinhardtii with 5 clusters, Fragilariaopsis cylindrus with 4 clusters, Fistulifera solaris with 5 clusters, and Thalassiosira pseudonana with 4 clusters, respectively. In sum, trypsin genes occur in higher copy numbers when organized in clusters, evidence of gene duplication. Considering that the genome assembly for some species was only at the contig level, some trypsin gene clusters may be missed by the cluster analysis. To address this issue, further gene duplication events were analyzed. For two proteins to belong to a duplicated gene pair, two criteria must be met: the length of alignable sequence covers >75% of the longer gene, and the similarity of aligned regions is >75% (Gu et al. 2002). Based on these criteria, duplicated trypsin gene clusters were detected in C. reinhardtii, T. oceanica and F. solaris from their genome collinearity analysis (Fig. 4; Table S2 in the Supporting Information). Chlamydomonas reinhardtii has a total of 38 trypsin genes, three of which are organized in three duplicated pairs with tandem localization in a chromosome, while not in the genome synteny blocks (CrTryp12, CrTryp13, CrTryp14; Fig. 4a). In T. oceanica, nine of the 58 trypsin genes are organized in six duplicated pairs with random localization in supercontigs.
located in the core is the phylogenetic tree of 291 trypsin family from all nine algal species, which clearly groups these proteins into two subfamilies (trypsin I and trypsin II). The unrooted maximum likelihood phylogenetic tree was constructed with MEGA7, and the bootstrap test replicate was set as 1000 times. Cm: Cyanidoschyzon merolae, Tp: Thalassiosira pseudonana, Fs: Phaeodactylum tricornutum, To: Thalassiosira oceanica, Fs: Fistulifera solas, Fk: Fragilaropsis cylindrus, Pm: Pseudo-nitzschia multiseries, Fk: Fugacium kawaguti, Cr: Chlamydomonas reinhardtii. Inner cycle depicts trypsin family from different lineages of algae corresponding to branches of the phylogenetic tree, each lineage in a unique color: red, diatoms; purple, dinoflagellate; blue, green algae; yellow, red algae. Middle circle: green, deep pink and yellow rectangles represent trypsin, trypsin_2 and PDZ conserved domains, respectively. Outer circle illustrates different organizations of ten putative motifs, each in a different color of rhombus. UN: unknown, tryp: trypsin. For details of motifs refer to Table S3.

(Fig. 4b). Forty-eight of the 65 trypsin genes in *F. solas*is are organized in 24 duplicated pairs with tandem/random localization in scaffolds that form syntenic blocks (Fig. 4, c, d). These results indicated that some trypsin genes might have arisen by segmental chromosome duplication and tandem duplication. To examine the evolutionary constraints acting on the trypsin gene family, the
nonsynonymous/synonymous substitution ratio ($Ka/Ks$) was calculated for the duplicated trypsin gene pairs. The majority of the duplicated trypsin gene pairs exhibited a $Ka/Ks$ ratio $< 1$ (Table S2), indicative of purifying selection after duplication.

Of the 307 trypsin genes found from the ten species, all but 73 genes were clustered in orthologous groups or paralogous groups as the OrthoFinder analysis revealed. From the phylogenetic tree of 234 orthologous or paralogous trypsin genes, nine groups emerged (Fig. 5). All the species included in the analysis contained a rich repertoire of trypsin genes of group I (52.12%), and its late diverging position indicates that this is the most recent emergent of the trypsin family. The other eight groups, however, are completely absent in *Cyanidioschyzon merolae* and variably occur across different species. These indicate that trypsin has experienced multiple
gene duplication events and differential losses. In the course of evolution, *C. merolae* lost all paralogs from groups II through IX, diatoms retained most of the trypsins in groups II, VII, VII, and IX, and the other species are in between the two extremes. The remarkable sequence divergence across different groups and species suggests that neofunctionalization might have occurred.

**Conserved and diverged trypsin gene structure.** To further explore the potential functions of algal trypsin genes, the conserved motifs and domains of trypsin proteins were identified using MEME and hmm-scan. As showed in Figure 3, ten conserved motifs and three conserved domains were identified, showing subclade-specific composition and distribution patterns. Based on the comprehensive analyses from Pfam, CDD and SMART database, motifs 2, 3, 6, 8, and 9 constituted the conserved region of trypsin I; motifs 5 and 10 belonged to trypsin II; motifs 1, 4 and 7 were unknown, but they were similar to Pro-Pro-Ser-Pro amino acid residue repeat (Fig. 3; detailed consensus amino acid sequence is shown in Fig. S2 and Table S3 in the Supporting Information). Furthermore, according to the Pfam database nomenclature, the three most conserved domains were trypsin, trypsin_2 and PDZ domain. The PDZ domain is believed to target signaling molecules to sub-membranous sites and occurs in diverse signaling proteins.

As illustrated in Figure 3, there are three different domain composition and distribution patterns. Trypsin I subfamily contains only the trypsin domain; Trypsin II subfamily either contains the trypsin_2 domain only (38.93%) or domain trypsin_2 coupled with the PDZ domain (61.07%). Further analysis on the conserved residues in the model diatom *Phaeodactylum tricornutum* showed that the known conserved residues (e.g., the trypsin catalytic triad, substrate-binding site, and the Cys residues at the conserved disulfide bridges) of the trypsin family had different degrees of divergence between PtTryp members, especially between trypsin I and trypsin II members (Fig. S3 in the Supporting Information). Together, the two subfamilies of trypsin showed disparate compositions and distributions of important residues, motif and domain, but with similar patterns among the closely branches. The more remarkable divergence in the distant branches might have rendered the genes for new functions.

**Trypsin expression profiles in different plankton groups and nutrient conditions in the global ocean.** To further understand the potential ecological functions of the
ubiquitous plankton trypsin genes, we investigated the relationship between trypsin gene expressions and environmental factors based on the Tara Oceans database. As shown in Figure 6, the majority of trypsin I transcripts were found in the larger size fractions (74.93% in 20–2000 μm), whereas the majority of trypsin II transcripts were found in the smaller size fraction (76.88% in 0.8–20 μm). Furthermore, the trypsin I transcript abundance showed little correlation with the ambient nutrient condition, whereas the trypsin II transcript abundance was closely correlated with the ambient nutrient condition (Fig. 6b). Trypsin I mRNA abundance of SRF_0.8–5 μm showed negative correlation with Fe and positive correlation with NO2−5m, whereas that of SRF_5–20 μm was positively correlated with Fe (Fig. 6a). In contrast, the mRNA abundance of trypsin II was positively correlated with most ambient nutrients examined, except for Fe and Si that showed negative correlations with trypsin II transcript abundance. The different size-fraction organism-biased distributions and different ambient nutrient-triggered responses of trypsin I and trypsin II mRNA abundance suggest that these two subfamilies of trypsin have diverged in function during evolution.

Conserved and divergent patterns of trypsin gene expression in response to stress in cultured diatoms and dinoflagellates. To gain a deeper insight into the functions of algal trypsin genes analyzed, we took advantage of currently available diatom EST libraries and Symbiodiniceae and Algal Genomic Resource (SAGER), which included 16 transcriptomic libraries of Phaeodactylum tricornutum, seven of Thalassiosira pseudonana, and six of Fugacium kawagutii, each derived from cells grown under different conditions (Maheswari et al. 2009, Li et al. 2020). After normalizing trypsin gene expression, a hierarchical clustering (Eisen et al. 1998) of the P. tricornutum, T. pseudonana and F. kawagutii trypsin genes was made to identify groups of genes with similar expression patterns and the libraries with similar gene expression profiles. The expression patterns of dinoflagellate trypsin genes exhibited constitutive expression under different growth conditions, while the expression of diatom trypsin genes tended to be more dynamic, and exhibited gene-differential and growth condition-specific patterns (Fig. 7). For example, trypsin expression in P. tricornutum exhibited three distinct patterns (clusters I, II, and III), genes from cluster II responded to fewer growth conditions than those from cluster III. Notably, PtTryp2 (cluster I) was the only gene showing differential expression across all the 16 different growth conditions examined (Fig. 7a), indicating that the gene might play important roles in responding to various environmental stressors.

In Thalassiosira pseudonana, only half of the 36 identified trypsin genes were expressed under the seven different conditions, and most of the other 18 genes are organized in tandem repeats, which are likely to be pseudogenes. Moreover, the T. pseudonana trypsin genes displayed three different condition-specific expression patterns (clusters I, II, and III), and genes of clusters I and II were less responsive than genes of cluster III (Fig. 7b). TpTryp14 was significantly more highly expressed in all T. pseudonana libraries than other trypsin genes in the species but in the “temperature-limited” (TL) library in which the gene’s expression was undetectable. Overall, T. pseudonana trypsin genes appeared to be most strongly responsive to the nitrate plus among the conditions examined, while least responsive to the temperature-limited condition.

In the dinoflagellate Fugacium kawagutii, in sharp contrast to diatoms, most of the trypsin genes, except FkTryp14 and FkTryp13, were found to be stably expressed across different growth conditions (Fig. 7c). These genes may be involved in basic physiological processes and hence behave like housekeeping genes. All these results showed that algal trypsin genes exhibited different expression patterns in response to environmental conditions, depending on the species and the specific trypsin gene, suggesting rampant structural and functional differentiation.

**DISCUSSION**

Widespread occurrence of trypsin in marine phytoplankton. The trypsin gene family is common in animals in which the genes have undergone repeated cycles of duplication and divergence to perform a wide spectrum of physiological activities (Neurath 1984, Davis et al. 1985, Stevenson et al. 1986, Hoo- per 1990, Muller et al. 1993, Roach et al. 1997, Arenas et al. 2010), but were previously reported to be absent in plants and protists (Rojas and Doolittle 2002, Querino Lima Afonso et al. 2020). However, our recent metatranscriptomic study on a phytoplankton community, diatom trypsin genes were found to be highly expressed, accounting for 1% of the total diatom transcriptome when diatoms were dominant (Zhang et al. 2019). In addition, our further study revealed that a trypsin of Phaeodactylum tricornutum played an important role in regulating the homeostasis of N:P stoichiometry (You et al. 2022). Together, these studies indicated that trypsin is not only present in phytoplankton, but also plays an important role and, therefore, deserves more attention and in-depth study. However, although in recent years, along with genome sequencing, an increasing number of trypsin domain-containing genes have been identified, no standardized nomenclature has been employed (e.g., trypsin, trypsin-like, HtrA, Deg.), resulting in incorrect and confusing inference of trypsin distribution and function across different species. Hence, a genome-wide phylogeny- and domain-based
Identification and nomenclature is required for marine phytoplankton trypsin, so as to simplify scientific communication when comparing different phytoplankton species and in making inferences about the functions of trypsin more reliable. Here, our report is a systematic documentation of the trypsin family in marine phytoplankton. We identified 291 trypsin genes across four major phyla of algae that had sequenced genomes (Bacillariophyta, Dinophyta, Cyanophyta, and Chlorophyta), and identified 129,512 and 6163 putative trypsin I and trypsin II genes from the global ocean expedition data, the

Fig. 6. Sample distribution of identified trypsin transcripts and pairwise Spearman correlation between trypsin subfamily mRNA level and environmental parameters. (a) The trypsin I genes that contain trypsin domain (Pfam ID PF00089) showing little correlation with environmental factors. (b) The trypsin II genes that contain trypsin_2 domain (Pfam ID PF13365) showing extensive correlations with environmental factors. The correlations were run for each size fraction class and P values were adjusted based false discovery rate (FDR). Only significant correlations ($P$ value $< 0.05$) are shown. Data of mRNA level and environmental parameters were from the 65 $Tara$ Oceans sampling stations at which seawater samples and environmental data were collected from surface (SRF) and deep chlorophyll maximum depths (DCM). The width of line that links from gene to sample is proportional to the ratio of trypsin mRNA abundance distribution in different samples. The length of factor bar (on the right) is proportional to the effect of environmental factors. The $+$/$-$ sign represents positive/negative correlation. [Color figure can be viewed at wileyonlinelibrary.com]
Fig. 7. Expression profiles of the trypsin family of *Phaeodactylum tricornutum*, *Thalassiosira pseudonana* and *Fugacium kawagutii* across different conditions. (a) Hierarchical clustering of expression patterns of *P. tricornutum* trypsin family across 16 different growth conditions. (b) Hierarchical clustering of expression patterns of *T. pseudonana* trypsin family in seven growth conditions. (c) Hierarchical clustering of expression patterns of *F. kawagutii* trypsin family under six growth conditions. The gene name of trypsin I and II was represented as red and blue, respectively. The subcellular localization was predicted by Plant-mPLOC. The heatmaps in (a) and (b) were based on the log2 transformed ESTs value derived from public Diatom EST Database whereas that in (c) was based on the log2 transformed TPM (transcript per million) value derived from public SANGER Database. Abbreviations of library names, Pt2DB: OS, original 12,000 standard; AA, ammonium adapted; FL, iron limited; SP, silica plus; SM, silica minus; NR, nitrate replete; NS, nitrate starved; UA, urea adapted; LD, low decadienal treated; HD, high decadienal treated; OM, oval morphotype; TM, triradiate morphotype; TA, tropical accessions; BL, blue light; C1, high CO2 for 1 d; C4, high CO2 for 4 d. Tp2DB: FL, iron limited cells (-Fe); NL, nitrate limited cells (-NO3); NP, nitrate plus; TL, temperature-limited; CL, carbon dioxide limited cells (-CO2); OL, old library; SL, silicate limited cells (Sunagawa et al. 2015). SANGER: -Cu, without Cu; -Mn, without Mn; -Ni, without Ni; +1/5 Zn: 2 nM, 1/5 of its normal concentration; +1/5 Fe, 50 nM, 1/5 of its normal concentration; NR, nitrate replete. The gradient shows the level of expression, with the darker shades representing higher expression. [Color figure can be viewed at wileyonlinelibrary.com]
Tara Oceans metatranscriptomic datasets. We find that the trypsin family is widely present across the phytoplankton taxonomic spectrum and the global ocean geographic range, advancing our knowledge on the distributions and diversity of this ancient enzyme.

**Extensive duplication and evolutionarily conserved structure of trypsin in marine phytoplankton.** Expansion of gene families, driven by gene duplication and retention, have the tendency to result in gene divergence between the family members and acquisition of novel functions in some of the members (Lespinet et al. 2002). Our analysis results show high copy numbers of trypsin genes in some algal lineages, suggesting the functional importance of these genes. Their high copy numbers signal rampant gene duplication during evolution. This is supported by multiple lines of evidence. First, compared with known species, the number of trypsin genes in phytoplankton tends to account for a larger proportion of the total number of genes in the whole genome, especially for diatoms. Second, the topology of phylogenetic tree suggests multiple duplication events during algal evolutions, especially in diatoms (Fig. 3). Third, trypsin genes cluster in tandem repeats or otherwise in pairs or in apparently duplicated segments, indicating that trypsin gene duplication is common in marine phytoplankton (Fig. 4). Last, a mass of identified trypsin genes (76.97%) from ten sequenced algal genomes were clustered in orthologous groups or paralogous groups (Fig. 5). Similarly, trypsin gene duplications have occurred in human, mosquito, Drosophila melanogaster and Plutella xylostella genome sequences (Davis et al. 1985, Rowen et al. 1996, Wu et al. 2009, Lin et al. 2015). These examples suggest that local duplication events by unequal crossing-over as well as segmental duplication have contributed to the expansion of the trypsin gene family in marine genomes. Furthermore, the orthologous and paralogous gene analysis, combined with phylogenetic trees, indicate that complex pattern of differential losses and duplications have occurred to trypsin genes during the evolution of marine phytoplankton, creating the highly variable gene copy numbers between algal species. The expansion of the trypsin family in some diatoms may confer adaptive advantage, and in those cases, neofunctionalization might be involved.

A close relationship was found between the phylogenetic affiliation, conserved domain and motif composition and distribution patterns of trypsin genes, indicating that the gene family is highly conserved within species but has evolved with speciation (Fig. 3). Based on the phylogenetic analysis and conserved domain analysis, the phytoplankton trypsin family was well separated into two distinct subfamilies: trypsin I (Fig. 3, clade I) and trypsin II (Fig. 3, clade II). The conserved motif and domain structure of trypsin proteins, especially those that are responsible for catalytic triad and substrate-binding pocket, is crucial for their function (Zou et al. 2006). Moreover, the two subfamilies of trypsin showed divergent motif and domain compositions and distributions. The trypsin I genes tend to only contain the N-terminus trypsin domain, and mainly consist of trypsin motifs, while some of the trypsin II only have N-terminus trypsin_2 domain that consist of trypsin and trypsin_2 motifs, and some of these genes are additionally coupled with PDZ domain. Furthermore, according to the MEROPS nomenclature, the trypsin I and trypsin II were found belonging to S1 proteases, but corresponding to the S1A clan (chymotrypsin) and S1C clan (DegP peptidase), respectively. Taken together, functional differentiation of algal trypsins with adaptive innovations is likely, but the exact functions of these different types of trypsins remain to be studied in the future.

**Potential roles of phytoplankton trypsin in response to environmental stress.** Trypsin is known to function in peptide hydrolysis, and protein cleavage and degradation in various physiological processes in both vertebrates and invertebrates, for example, coagulation, clot resolution, digestion, fertilization, blood pressure regulation, tissue development and homeostasis, and immunity (Hellman and Thorpe 2014). However, little is known about the function of trypsin family in phytoplankton. Based on further literature review, we found that 16 trypsin domain-containing genes have been reported in Arabidopsis thaliana but were annotated as HtrA (High Temperature Requirement A) / Deg (Degradation of periplasmic protein) genes, which have been shown to function in maintaining protein homeostasis and involved in protein processing (Schuhmann and Adamska 2012, Schuhmann et al. 2012). The HtrA/Deg were initially identified and named according to function in Escherichia coli mutants, which exhibited growth repression at elevated temperatures and failed to digest misfolded protein in the periplasm, respectively (Lipinska et al. 1988, Strauch and Beckwith 1988). The phytoplankton trypsin II genes reported in the present study feature a trypsin_2 domain and 1–2 PDZ domains and share the same structure architecture of HtrA/Deg previously reported (Claussen et al. 2011, Sulskis et al. 2021), which were considered homologs of plant DegP/HtrA. Hence, the large amount of phytoplankton trypsin gene duplication with trypsin I and trypsin II differentiation and specific characteristics documented in the present study suggests that they may have evolved different important functions. This is consistent with the strikingly high expression of diatom trypsins in a diatom-dominated natural plankton assemblage (Zhang et al. 2019) and supported by the recent discovery of a trypsin’s function in regulating the homeostasis of N:P stoichiometry (You et al. 2022).
Trypsin is associated with many intracellular and extracellular events. The classic trypsin (trypsin I in this study) in fish and other animals is involved in important cell processes and responses to various environmental factors (Rungruangsak-Torissen and Male 2000, Rungruangsak-Torissen et al. 2006, Muhlía-Almazán et al. 2008). The family of Deg/HtrA proteases (trypsin II in this study) plays an important role in quality control of cellular proteins (e.g., the Deg proteases assisting in degradation of damaged photosynthetic proteins in photosynthetic organisms; Sun et al. 2010, Schuhmann and Adamska 2012, Tam et al. 2015). Hence, we hypothesize that both of the phytoplankton trypsin I and II may be involved in important cell processes and responses to various environmental factors in phytoplankton. Based on this hypothesis, it is expected that trypsin expression would be modulated by environmental factors. Our in-depth analysis on Tara Oceans global eukaryotic metatranscriptome (Carradec et al. 2018), diatom EST Database (Maheswari et al. 2009) and SAGER Database (Yu et al. 2020) indicates significant correlations of trypsin I and trypsin II expressions with different environmental factors. With the support from the various existing datasets and our recent findings, our hypothesis about the importance of the ancient enzyme in phytoplankton stands for rigorous examination in future research.

CONCLUSIONS

Despite being a classic textbook example and a well-described group of enzymes, trypsin in marine phytoplankton is poorly understood and has been hardly explored. Moreover, there is not standardized nomenclature for trypsin family, leading to a confusing inference of trypsin distribution and function across different species. This study represents a systematic and comprehensive study on trypsin in phytoplankton, including genome-wide identification, characterization of lineage-specific duplication, analysis of structural and potential functional evolution, and profiling of expression patterns in relation to environmental conditions. A wide taxonomic and geographic distribution of the trypsin family was found, illuminating the potential importance of the gene family in phytoplankton in the global ocean. We observed subclade conservation, lineage-specific duplication, the high homolog retention rate, and lineage-specific expression patterns among different algae, implying the high biological importance of the trypsin gene family in marine phytoplankton. Diatoms are one of the most abundant and diverse groups of marine phytoplankton, respond rapidly to the supply of new nutrients, often out-competing other phytoplankton. Our results suggest that major steps in the evolution of the gene family reflect key events triggering diatom radiation and diversification. The more gene expansion, diversified features and expression patterns of the trypsin across different diatom species than other lineages of phytoplankton might be associated with their competitive advantages. The findings of trypsin gene duplication and sequence divergence will be helpful for in-depth exploration of the evolution and functional innovation of trypsin to gain a better understanding on the evolution of phytoplankton trypsin family.

We wish to thank Ling Li, Chentao Guo, and Yujie Wang of Marine EcoGenomics Laboratory (Xiamen University, China) for technical advice and assistance. We also thank Jianwei Chen (BGI Genomics Co., Ltd., China) for his assistance with trypsin orthologous and paralogous analysis. This work was supported by the National Natural Sciences Foundation of China (grant number 41906123 to YY), the Open Fund of CAS Key Laboratory of Marine Ecology and Environmental Sciences, Institute of Oceanology, Chinese Academy of Sciences (grant number KLMES202006 to YY), the Marine S&F Fund of Shandong Province for Pilot National Laboratory for Marine Science and Technology (Qingdao) (grant number 2018SDKJ0406-3 to SL), the MEL Internal Programs of the State Key Laboratory of Marine Environmental Science, Xiamen University (grant number MELR2105 to SL). We wish to thank the program of Excellent Postdoctoral Allowance for Marine Science from the College of Ocean and Earth Sciences (Xiamen University, China).

DATA AVAILABILITY STATEMENT

The data, including the source data that supports the finding of this study, are available within the paper, its supplementary information files, or publicly available datasets. Publicly available trypsin sequences were obtained from Pfam, NCBI and Ensemble databases. Publicly available worldwide distributions and abundances of trypsin genes were acquired from the OGA datasets (https://tara-oceans.mio.osupytheas.fr/ocean-gene-atlas/).

Arenas, A. F., Osorio-Mendez, J. F., Gutierrez, A. J. & Gomez-Marin, J. E. 2010. Genome-wide survey and evolutionary analysis of trypsin proteases in apiocomplexan parasites. Genom. Proteom. Bioinf. 8:103–12.

Carradec, Q., Pelletier, E., Da Silva, C., Alberti, A., Seeleuthner, Y., Blane-Mathieu, R., Lima-Mendez, G. et al. 2018. A global ocean atlas of eukaryotic genes. Nat. Commun. 9:373.

Chen, C., Xia, R., Chen, H. & He, Y. 2018. TBtools, a Toolkit for Biologists integrating various HTS-data handling tools with a user-friendly interface. bioRxiv 289960.

Chen, J. M., Kukor, Z., Le Marechal, U., Toth, M., Tsakiris, L., Raguenes, O., Ferec, C. & Sahin-Toth, M. 2003. Evolution of trypsinogen activation peptides. Mol. Biol. Evol. 20:1767–77.

Chou, K. C. & Shen, H. B. 2010. Plant-nPloC: a top-down strategy to augment the power for predicting plant protein sub-cellular localization. PLoS ONE 5:e11395.

Clausen, T., Kaiser, M., Huber, R. & Ehrmann, M. 2011. HTRA proteases: regulated proteolysis in protein quality control. Nat. Res. Mol. Cell Biol. 12:152–62.

Davis, C. A., Riddell, D. G., Higgins, M. J., Holden, J. J. & White, B. N. 1985. A gene family in Drosophila melanogaster coding for trypsin-like enzymes. Nucleic Acids Res. 13:6605–19.

Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. 1998. Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA 95:14863–8.
Supporting Information

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

Figure S1. Chromosome/scaffold location of trypsin genes. (a) Chlamydomonas reinhardtii; (b) Fragilariopsis cylindrus; (c) Fistulifera solaris; (d) Thalassiosira pseudonana; Chr: chromosome; The bar length represents the length of chromosome or scaffold. The trypsin location of other species that no showed here, due to their absence of gene clusters or their genome assembly at contig level.

Figure S2. WebLogo consensus sequences for 10 most conserved motifs of algal trypsins.

Figure S3. Alignment of Phaeodactylum tricornutum trypsin amino acid sequences. Ten P. tricornutum trypsin sequences were identified in this study, which were aligned using CLUSTAL W. Dashes represent gaps introduced for the alignment. Identical residues are depicted by same background colors. The trypsin domain is indicated by red line. The red letter represents the reported important conserved residues of the trypsin family (Rypniewski et al. 1994). These include the trypsin catalytic triad (Ser-His-Asp), the substrate-binding site, the Cys residues at the conserved disulfide bridges, and the amino terminus of the mature trypsins (peptide IVGG).

Table S1. Gene characteristics of algal trypsin genes.

Table S2. Ka/Ks values of duplicated trypsin gene pairs.

Table S3. List of the putative motifs of 291 alga trypsin proteins.