An Ancient Enzyme Finds a New Home: Prevalence and Neofunctionalization of Trypsin in Marine Phytoplankton

Yanchun You  
College of Ocean and Earth Sciences, Xiamen University,

Xueqiong Sun  
College of Ocean and Earth Sciences, Xiamen University,

Senjie Lin (senjie.lin@uconn.edu)  
University of Connecticut  https://orcid.org/0000-0001-8831-6111

Article

Keywords: trypsin, phylogenetics, tandem duplication, evolution, environmental stimuli

Posted Date: November 16th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1008417/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
An ancient enzyme finds a new home: prevalence and neofunctionalization of trypsin in marine phytoplankton

Yanchun You¹, Xueqiong Sun¹, Senjie Lin¹²³*

Yanchun You: fjyhy@163.com

Xueqiong Sun: 22320201151289@stu.xmu.edu.cn

Senjie Lin: senjie.lin@uconn.edu

¹ State Key Laboratory of Marine Environmental Science, College of Ocean and Earth Sciences, Xiamen University, Xiamen 361005, China.

² Laboratory of Marine Biology and Biotechnology, Qingdao National Laboratory of Marine Science and Technology, Qingdao 266237, China.

³ Department of Marine Sciences, University of Connecticut, Groton, CT, United States.

*Correspondence: Senjie Lin: senjie.lin@uconn.edu

Abstract

Trypsin is an ancient protease best known as a digestive enzyme in animals, and traditionally believed to be absent in plants and protists. Here, we surveyed the distribution, diversity, evolution and potential functions of trypsin genes in global ocean phytoplankton, the major primary producers in the aquatic ecosystem. Our analysis indicates that trypsin genes are widely distributed both taxonomically and geographically in marine phytoplankton. Furthermore, by systematic comparative analyses we documented lineage-specific diversity and expansion of trypsin genes in the evolution of marine phytoplankton. Genome-wide analyses revealed that trypsin genes were more prevalent in diatoms than in other lineages. Moreover, the expression
of trypsin genes in diatom tended to be more responsive to environmental stimuli. The
duplication and neofunctionalization of trypsin genes may be important in diatoms to
adapt to dynamical environmental conditions, contributing to diatoms’ dominance in
the coastal oceans. This work advances our knowledge on the distributions and
neofunctionalizations of this ancient enzyme and creates a new research direction in the
phytoplankton biology.

Keywords: trypsin, phylogenetics, tandem duplication, evolution, environmental
stimuli
Introduction

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\(^1\), trypsin is arguably the first enzyme known to science and the best studied protein, and probably is the best exploited enzyme for protein biotechnology. Yet beyond the knowledge of it as a digestive enzyme, understanding of its occurrence and function across the tree of life is very limited.

Trypsin represents a conserved family of enzymes occurring in organisms ranging from bacteria to mammals\(^2,3\). It is believed to be absent in plants and protists\(^4,5\). Indeed, many animal trypsin genes have extremely well-understood physiological roles, to digest protein food and to activate zymogen, an activator of other proteases\(^6,7\), etc.. In the case of fish, the expression of trypsin is affected by influences protein digestion rate, amino acid absorption and transport, protein synthesis, and growth rate\(^2,8,9\). However, the gene characteristics, phylogenetic perspectives and function of trypsin in phytoplankton are unexplored and enigmatic.

In the current MEROPS database, there are 109 trypsin enzymes, all of which are exclusively from animals. In a 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\(^10\). This raises questions as to whether trypsin occurs broadly, how it has evolved, and what roles it plays, in phytoplankton, which represent the majority of plantae diversity and 50% of global carbon fixation and oxygen production. Here, we analyzed the *Tara* Oceans metatranscriptomic dataset to detect occurrence and quantify expression of trypsin genes in phytoplankton in the global ocean and their relationship with environmental factors. We further documented sequence diversity, lineage-specific expansion and phylogenetic perspectives of trypsin genes by mining existing algal genome data and analyzed gene structure, evolutionary characteristics, and expression patterns from nine species that represent major phyla of algae.

Results
Taxonomic and geographic distribution of trypsin in natural assemblages of marine plankton

To understand how broadly trypsin occurs in marine plankton, we mined existing data. Using Hidden Markov Models, we aligned trypsin (PF00089) and trypsin-like (PF13365) conserved domains against the MATOU databases. Trypsin and trypsin-like homologous genes were presented in all global sampling sites in the Tara Oceans expedition with 129,512 and 6,167 hits, respectively. Interestingly, trypsin sequences were found predominantly in the surface water layer (SRF) and, less abundantly in the deep chlorophyll maximum layer (DCM), and both were more abundant in the larger sized organisms (180-2000 µm) (Fig. 1a). The trypsin-like sequences show similar distribution with trypsin with respect to latitudes and the sampled depths, but were more abundant in the smaller sized organisms (0.8-5 µm) (Fig. 1b).

Taxonomically, 96% of the trypsin hits 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-like, 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). Therefore, trypsin and trypsin-like genes are widely distributed in marine plankton both taxonomically and geographically, but are dominantly in Arthropoda and phytoplankton, respectively.

Abundance and evolutionary dynamics of trypsin genes in algal genomes

To better understand the lineage-specific diversity and expansion of trypsin during phytoplankton evolution, we searched for trypsin sequences from available sequenced genomes across four major phyla of algae, including Rhodophyta (C. merolae), Chlorophyta (C. reinhardtii), Dinophyta (F. kawagutii) and Bacillariophyta (T. pseudonana, P. tricornutum, T. oceanica, F. solaris, F. cylindrus and P. multiseries). For comparison, the genome of the land plant model A. thaliana was also 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 (supplementary table 1). The number of trypsin genes varies in
examined algal genome from 5 to 65 (Fig. 2), accounting for 0.93% 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 *C. merolae*, likely already contained a repertoire of trypsin 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 across alga lineages, even between diatom species, showing no phylogenetic trends. For instance, *P. tricornutum* and *F. 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 (pIs) (supplementary table 1). 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 genes into two classic subfamilies**

To analyze the evolutionary relationships of trypsin genes among the nine selected algae, an unrooted phylogenetic tree was constructed using their conserved amino acid sequences (Fig. 3). The 291 trypsin genes were well separated into two distinct subfamilies: trypsin (162 genes, Fig. 3, clade I) and trypsin-like (129 genes, Fig. 3, clade II). Most of the trypsin genes are
originated from diatoms and dinoflagellates. By contrast, trypsin-like genes are more diverse, including sequences from all the selected nine species of algae.

Remarkably, some trypsin members were more closely related to those in the same subfamily from different species than to the other trypsin genes from the same species, indicative of ancestral gene duplication and subsequent divergence. 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 recently gene duplication events within species. Within a subfamily, most trypsin genes were clustered by phylum, suggesting a lineage-specific expansion within the trypsin or trypsin-like subfamily in algae. In C. merolae, an ancient taxon of rhodophytes, the phylum from which all non-green algae are believed to have emerged, all of its trypsin genes appeared to belong to the trypsin-like subfamily. These results indicate that trypsin genes might have arisen from trypsin-like genes.

Trypsin gene clusters and evidence of duplications and differential losses

We found that trypsin genes are not randomly distributed on different chromosomes, scaffolds or contigs, and some algal trypsin genes exist as gene clusters (Extended Data Fig. 1). Based on current genome assembly level, four of the nine alga genomes we examined contain trypsin genes organized in clusters, as shown in extended data figure 1, C. reinhardtii with 5 clusters, F. cylindrus with 4 clusters, F. solaris with 5 clusters, and T. pseudona with 4 clusters, respectively. In sum, the algae with gene clusters appear to host more trypsin genes, which might have resulted from gene duplication. Considering that the genome assembly of some algae were only at contig level, thus, some trypsin gene clusters may be missing, further gene duplication events were analyzed. For two proteins to belong to the duplicated gene pairs, two criteria must be met: (1) length of alignable sequence covers >75% of the longer gene, and (2) the similarity of aligned regions is >75% \(^{11}\). Based on these criteria, three species, C. reinhardtii, T. oceanica and F. solaris, had three genes in three duplicated gene pairs out the genome collinearity blocks, 10 genes in 6 duplicated gene pairs out the genome collinearity blocks, and 34 genes in 24 duplicated gene pairs in collinearity blocks, respectively (Fig. 4; Extended Data Table 1). To examine the evolutionary constraints acting on 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 (Extended Data Table 1), indicative of purifying selection after duplication. These results indicated that some trypsin genes might have arisen by segmental chromosome duplication.

Of the 304 trypsin genes found from the ten species, all but 70 genes were clustered in orthologous groups or paralogous groups by OrthoFinder analysis. From the phylogenetic tree of 234 orthologous or paralogous trypsin genes, nine clades emerged (Fig. 5). All the species included in the analysis contained a rich repertoire of trypsin genes of clade I, and its late diverging position indicates that this is the most recent emergent of the trypsin family. The other eight clades, however, are completely absent in *C. 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 clades II through IX, diatoms retained most of the trypsin of clade II, VII, VIII and IX, and the other species are in between the two extremes. The remarkable sequence divergence in the more basal clades (e.g. IX) suggests 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 by MEME and hmmscan. As showed in figure 3, 10 conserved motifs and three conserved domains were identified, and with subclade-regular composition and distribution patterns. Based on the comprehensive analyses from Pfam, CDD and SMART database, motifs 2, 3, 6, 8 and 9 were responsible for the region of the conserved trypsin, and motifs 5 and 10 belong to the trypsin-like serine protease, motifs 1, 4 and 7 were unknown, but they share similar sequence with PPSP repeat (Fig. 3; detail consensus amino acid sequence shown in Extended Data Fig. 2 and Extended Data Table 2). Furthermore, the three most conserved domains were trypsin, trypsin-like and PDZ-like domain, and some trypsin-like domain tend to couple with PDZ-like 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, the two clades of trypsin genes show diverged motif and domain
compositions and distributions, but with similar patterns among the closely branches, indicating that trypsin and trypsin-like genes may play different roles in cells. There are three different domain composition and distribution patterns, that is, trypsin domains were only found in trypsin genes, trypsin-like domains solely or couple with PDZ domains in trypsin-like genes. Furthermore, a close relationship was found between the phylogenetic affiliation, conserved domain and motif composition and distribution patterns, indicating that they were highly conserved, which also supports the close evolutionary relationship of these members.

**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. Interestingly, the majority (74.93%) of trypsin transcripts were found in the larger size fractions (32.31% and 42.62% in 20-180 µm and 180-2000 µm, respectively, 48.89% and 26.04% in SRF and DCM, respectively) (Fig. 6a). By contrast, the majority (62.64%) of trypsin-like transcripts were found in the smaller size fraction (0.8-5 µm, 41.97% and 20.67% in SRF and DCM, respectively) (Fig. 6b). According to Malviya et al. (2016), *Tara Oceans* smaller size fractions (0.8-5 and 5-20 µm) are mostly composed of small diatoms and single cells. Overall, both trypsin subfamilies were more abundantly expressed at surface than that at the DCM. The different size-class distributions of trypsin and trypsin-like suggest that they may have different functions. Furthermore, as showed in figure 6, the trypsin subfamily showed no correlation with ambient nutrient abundances except for SRF_5-20 and SRF_0.8-5 with iron and SRF_0.8-5 with nitrite. In contrast, the trypsin-like subfamily was closely correlated with ambient nutrient conditions, especially for the medium size fractions (5-180 µm) (Fig. 6b). Moreover, the mRNA abundance of trypsin-like were significant positively correlated with ambient nutrient changes, while showed anticorrelated with Fe and Si. With trypsin, only the mRNA abundance of SRF_0.8-5 µm showed anticorrelation with Fe and positive correlation with NO2_5m, while SRF_5-20 µm was positive correlated with Fe. Moreover, NO2, NO3_5m and NO3_NO2 could modulated more size-class of trypsin-like expression.
Conserved and divergent patterns of trypsin gene expression in response to environmental stress in cultured diatoms and dinoflagellates

To gain a deeper insight into the functions of the trypsin genes analyzed, we took advantage of currently available diatom EST libraries and Symbiodiniceae and Algal Genomic Resource (SAGER), which included 16 transcriptomic libraries in *P. tricornutum*, seven in *T. pseudonana*, and six in *F. kawagutii*, each derived from cells grown under different conditions. After normalizing trypsin gene expression for the diatoms and dinoflagellates, a hierarchical clustering 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). In diatoms, the expression of trypsin genes showed responses to growth conditions (Fig. 7a), with more genes responding to more specific conditions. For example, in *P. tricornutum*, trypsin genes from clade I responded to fewer growth conditions than those from clade II. Notably, *PtTryp2* was the only gene showing differential expression across all these 16 different growth conditions with a higher expression (Fig. 7a), indicating that the gene might play important roles in responding to various environmental stresses. In *T. pseudonana*, only half of the 36 identified trypsin genes were expressed under the seven different conditions, most of the other 18 genes are organized in tandem repeats, which are likely to be pseudogenes. Similarly, in *T. pseudonana* clusters of trypsin genes represent more condition-specific expression pattern, genes of clade I and II were less responsive, while genes of clade III were more responsive (Fig. 7b). *TpTryp14* was significantly more highly expressed in all *T. pseudonana* libraries but in the ‘temperature limited’ (TL) library. *T. pseudonana* trypsin genes appeared to be more strongly responsive to the nitrate plus, while in temperature limited was just the opposite. Most *F. kawagutii* trypsin genes, except *FkTryp14* and *FkTryp13*, were found to be consistently expressed throughout different growth conditions (Fig. 7c), those highly and moderately expressed genes (clade I and II) may function as housekeeping genes involved in basic physiological processes. All these results showed that algal trypsin genes with
diverse and complicated expression patterns, suggesting their diversified functions and may function as biological interactions or cross-talk between in response to particular stimuli.

Discussions

Widespread occurrence of trypsin in marine phytoplankton

In many eukaryotic taxonomic groups, especially in animals, trypsin gene family has undergone repeated cycles of duplication and divergence to perform a wide spectrum of physiological activities\textsuperscript{15-21}, but were previously believed to be absent in plants and protists\textsuperscript{4}. With the advances of sequencing technology, increasing plant and protist genomes have been sequenced, facilitating detection of trypsin genes. However, most of the sequences from plants or protists that contain the trypsin or trypsin-like domains are usually recognized as unknown function proteins, and much less data is available for marine phytoplankton trypsin. Our report here is a systematic documentation of trypsin and trypsin-like genes in marine phytoplankton. We identified 291 trypsin genes across four major phyla of algae that had sequenced genomes (Bacillariophyta, Dinophyta, Cyanophyta, and Chlorophyta). In addition, we identified 129,512 and 6,167 putative trypsin and trypsin-like genes from Tara Oceans metatranscriptomic datasets, indicating wide taxonomic and geographic distribution of trypsin and trypsin-like genes in the global ocean.

Extensive duplication and evolutionarily conserved structure of trypsin in marine phytoplankton

Expansion of gene families tend to emerge new biological functions, which with the relatively high rate of being duplicated or retained\textsuperscript{22}. Our analysis results show high copy numbers of trypsin and trypsin-like genes occur in some algal lineages, suggesting the functional importance of these genes. Their high copy numbers indicate that rampant gene duplication might have occurred 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, repetitive 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 sequences17,23-25. These examples suggest that local duplication events by unequal crossing-over as well as segmental duplication have contributed to the expansion of this 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 occurred to trypsin genes during the evolution of marine phytoplankton, creating the highly variable gene copy numbers between algal species. The expansion of trypsin gene families 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, indicating that they were highly conserved (Fig. 3). 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 function26. Moreover, the two clades of trypsin genes showed diverging motif and domain compositions and distributions. The clade I trypsin genes tend to only contain the N-terminus trypsin domain, and mainly consist of trypsin motifs, while some of the clade II trypsin-like genes only have N-terminus trypsin-like domain that consist of trypsin and trypsin-like motifs, and some genes are additionally coupled with PDZ domain.

**Potential roles of phytoplankton trypsin in response to environmental stress**

Trypsin is known to function in peptide hydrolysis, protein cleavage, and degradation in various physiological processes in both vertebrates and invertebrates, e.g. coagulation, clot resolution, digestion, fertilization, blood pressure regulation, tissue development and homeostasis, and immunity27. Moreover, in A. thaliana, we found that 16 Deg/HtrA protease family genes, which have been shown to function in maintaining protein homeostasis and protein processing28,29,
contain trypsin or trypsin-like domains. Nothing is known about the function of trypsin and
trypsin-like proteins in phytoplankton, but the large amount of gene duplication documented in
the present study suggests that they may have an important role in phytoplankton. This is further
supported by the strikingly high expression of diatom trypsins in a diatom-dominated natural
plankton assemblage\textsuperscript{30}.

We hypothesize that like trypsin in fish and other animals\textsuperscript{2,8,31}, phytoplankton trypsin may be
involved in important cell processes and responses to various environmental factors. Based on
this hypothesis, it is expected that trypsin expression would be modulated by environmental
factors. Our in-depth analysis on \textit{Tara Oceans} global eukaryotic metatranscriptome\textsuperscript{32}, diatom
EST Database\textsuperscript{12} and SAGER Database\textsuperscript{33} indicates significant correlations of trypsin and
trypsin-like expressions with environmental factors. With the support from the various existing
datasets, our hypothesis about the importance of the ancient enzyme in phytoplankton stands
for rigorous and specific examination in future research.

\textbf{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. In this study, a wide
taxonomic and geographic distribution of trypsin and trypsin-like genes was found, illuminating
the potential importance of the trypsin-domain function in phytoplankton in the global ocean.

We conducted integrative analyses on the trypsin gene structure, lineage-specific duplication,
phylogenetic perspectives and expression modulation of trypsin family in algae. The
differences in diverse duplication and expression of the trypsin gene family across different
species hint at a key role. Our results suggest that major steps in the evolution of the gene family
reflected key events triggering diatom radiation and diversification. 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. The more gene expansion, diversified
features and expression patterns of the trypsin across different diatom species are inferred to be
associated with the capacity of diatom to out-competing other phytoplankton. The conservation
of subclades, lineage-specific duplication, the high homolog retention rate and the conservation
of expression patterns among different algae underlie the high biological importance of the trypsin gene family in general and of distinct subclades in particular across marine phytoplankton. This study represents a systematic and comprehensive study on genome-wide identification, lineage-specific duplication, phylogenetic perspectives, gene characterization and expression profiling of the trypsin family in phytoplankton. Gene duplication and sequence divergence are helpful for in-depth exploration of the evolution of novel gene functions, and our findings provide a base for functional research on specific trypsin genes, a better understanding of the evolution of phytoplankton trypsin.

**Materials and Methods**

**Detection of trypsin genes in the Tara Oceans datasets**

Extensive search for putative trypsin and trypsin-like genes was performed in both the Tara Oceans eukaryote unigene catalog and metatranscriptomes (MATOUv1+metaT)\(^{32,34}\), using the trypsin (Pfam ID: PF00089) and trypsin-like (Pfam ID: PF13365) domain based on profile Hidden Markov Models (HMM), with an E-value ≤ 1.0 \(\times 10^{-10}\). 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/ocean-gene-atlas/).

**Identification of trypsin genes 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 trypsin genes, the hmmsearch analysis was conducted. We downloaded the HMM profile of trypsin and trypsin-like (PF00089 and PF13365) from Pfam protein family database (http://pfam.xfam.org/) and used it as the query (P < 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 ≤1e-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/bwrpsb/bwrpsb.cgi), Pfam 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, *Cyanidioschyzon merolae* (*Cm*), *Thalassiosira pseudonana* (*Tp*), *Phaeodactylum tricornutum* (*Pt*), *Thalassiosira oceanica* (*To*), *Fistulifera solaris* (*Fs*), *Fragilariopsis cylindrus* (*Fc*), *Pseudo-nitzschia multiseries* (*Pm*), *Fugacium kawagutii* (*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 trypsin genes**

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, maximum 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 was used for the mapping of trypsin genes’ chromosomal positions and relative distances.

**Phylogenetic analysis of identified trypsin genes**

OrthoFinder 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 1e−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 trypsin and trypsin-like genes, 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\textsuperscript{38,39} and inspected manually for necessary correction. An unrooted Maximum Likelihood phylogenetic tree was constructed using MEGA X\textsuperscript{40,41} software with bootstrap test of 1000 times, based on a discrete Gamma distribution of evolutionary rate variations, which was recommended by the results of Poisson correction model. 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)\textsuperscript{42}. The following nine environmental nutrient parameters were chosen for correlation analysis: iron_5m (Fe, \(\mu\text{mol/l}\)), ammonium_5m (\(\text{NH}_4^+\), \(\mu\text{mol/l}\)), NO\textsubscript{2}-NO\textsubscript{3} (\(\mu\text{mol/l}\)), NO\textsubscript{2} (\(\mu\text{mol/l}\)), NO\textsubscript{3} (\(\mu\text{mol/l}\)), PO\textsubscript{4} (\(\mu\text{mol/l}\)), Si (\(\mu\text{mol/l}\)), NO\textsubscript{2}_5m (\(\mu\text{mol/l}\)), NO\textsubscript{3}_5m (\(\mu\text{mol/l}\)). 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 > 30.

**Analysis of trypsin gene expression in diatoms and dinoflagellates in existing databases**

The gene expression data of diatom trypsin genes (\textit{PtTryp} for \textit{P. tricornutum} and \textit{TpTryp} for \textit{T. pseudonana} trypsins) and dinoflagellate trypsin genes (\textit{FkTryp} for \textit{F. kawagutii} trypsin) were downloaded from the Diatom EST Database (http://www.diatomics.biologie.ens.fr/EST3/) and SAGER Database (http://sampgr.org.cn/index.php), respectively. The gene expression data included a range of environmental conditions, as detailed in websites. The counts of ESTs for \textit{PtTryp} and \textit{TpTryp} and TPM value for \textit{FkTryp} were used to analyze the expression pattern. The results visualized using TBtools\textsuperscript{43}.

**Acknowledgements**
We wish to thank Xin Lin, Ling Li, Chentao Guo 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 KLMEES202006 to YY), the Marine S&T 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 202105 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).

Author contributions

YY and SL conceived and designed the research. YY and XS performed the experiments and data analysis. YY and SL wrote the manuscript.

References

5. Querino Lima Afonso, M., da Fonseca, N. J., Jr., de Oliveira, L. C., Lobo, F. P. & Bleicher, L. Coevolved positions represent key functional properties in the trypsin-like serine...


Figure legends

**Fig. 1** Geographic and taxonomic distribution of trypsin and trypsin-like homologs detected by mining the *Tara* Oceans eukaryote unigene catalog and metatranscriptomes (MATOUv1+metaT) datasets in the OGA webserver. **a**, Geographic distribution and abundance of trypsin homologs in eukaryotes in surface layer (SRF) and deep chlorophyll maximum layer (DCM). **b**, Geographic distribution and abundance of trypsin-like homologs in eukaryotes in SRF and DCM. **c**, Taxonomic distribution of all the detected trypsin homologs in eukaryotes. **d**, Taxonomic distribution of all the detected trypsin-like homologs in eukaryotes. The size of the filled circles shown in (a, b) are proportional to the mRNA abundance of the hits in one location compared to the total number of hits integrated from the different sampled depths, and colors represent the size fractionation range.

**Fig. 2** Evolutionary dynamics of genome size, abundance of coding genes, and abundance of trypsin genes among selected algae and the land plant model *A. thaliana*. The phylogenetic tree was constructed based on NCBI taxonomy and TIMETREE database. Lineage divergence time is indicated at each branch point; Mya, million years ago. The number in trypsin column represents the number of identified trypsin genes, and the number in parenthesis indicates the number of duplication gene pairs.

**Fig. 3** Phylogenetic relationship and architectural profiles of conserved domain and conserved motif of trypsin genes in algae. Located in the core is the phylogenetic tree of 291 trypsin and trypsin-like proteins from all nine algal species, which groups these proteins into two clusters (I, trypsin; II, trypsin-like). The unrooted maximum likelihood phylogenetic tree was constructed with MEGA7, and the bootstrap test replicate was set as 1000 times. *Cm*: *Cyanidioschyzon merolae*, *Tp*: *Thalassiosira pseudonana*, *Pt*: *Phaeodactylum tricornutum*, *To*: *Thalassiosira oceanica*, *Fs*: *Fistulifera solaris*, *Fc*: *Fragilariopsis cylindrus*, *Pm*: *Pseudo-nitzschia multiseries*, *Fk*: *Fugacium kawagutii*, *Cr*: *Chlamydomonas reinhardtii*. Inner cycle depicts trypsin genes from different lineages of algae corresponding to branches of the phylogenetic tree, each lineage in a unique color: red, diatoms; purple, dinoflagellate; blue, green alga; yellow, red alga. Middle circle: green, deep pink and yellow rectangles represent
trypsin, trypsin_like 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 Extended Data Table 2.

**Fig. 4 Number and location of duplicated trypsin gene pairs.**

a, Distribution of tandem duplicated trypsin genes on *C. reinhardtii* genome. b, Duplicated trypsin gene pairs in *T.
oceanica*. c, Duplicated trypsin gene pairs in *F. solaris*. d, Distribution of segmentally duplicated trypsin genes on *F. solaris* genome. Duplicated trypsin genes were mapped to their respective locus in the genome in a circular diagram using Circos. Chromosomal or scaffold segments are indicated by different colors (outer track). Duplicated trypsin gene pairs linked with green or red lines (inside of circle). Grey lines indicate collinear blocks in whole genome.

**Fig. 5 Phylogenetic relationships of paralogs and orthologs trypsin genes between different marine phytoplankton.** The phylogenetic tree was constructed on 234 trypsin genes that identified as paralogs and orthologs groups from 307 trypsin genes by OrthoFinder. Based on the paralogs and orthologs groups, the phylogenetic tree was manually defined into 9 clades. All the branches within clades have been collapsed for simplicity and the number of sequences corresponding to each selected species. Sequences from the same species within the same clade were considered as paralogs groups, and sequences from different species were considered as orthologs groups. Combining trypsin phylogenetic tree and the simplified phylogenies of selected species (upper tree), we observe a variable gene number in each clade among the species, indicating an evolutionary scenario characterized by several gene duplication and loss events.

**Fig. 6 Sample distribution of identified trypsin transcripts and pairwise Spearman correlation between trypsin subfamily mRNA level and environmental parameters.**

a, The trypsins that contain trypsin domain (Pfam ID PF00089) showing little correlation with environmental factors. b, The trypsin-like genes that contain trypsin-like peptidase domain (Pfam ID PF13365) showing extensive correlations with environmental factors. The correlations were run for each size 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 ration 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.

**Fig. 7 Expression profiles of trypsin genes of P. tricornutum, T. pseudonana and F. kawagutii across different conditions.** ESTs for *P. tricornutum* and *T. pseudonana* trypsin genes, TPM values for *F. kawagutii* trypsin genes. ESTs and TPM value of trypsin genes were transformed by log_{2} and the heatmap was constructed by Tbtools software. a, Hierarchical clustering of expression patterns of *P. tricornutum* trypsin genes in 16 different growth conditions. b, Hierarchical clustering of expression patterns of *T. pseudonana* trypsin genes in seven growth conditions. c, Expression patterns of *F. kawagutii* trypsin genes under six growth conditions. Abbreviations of library names, Pt2DB: OS, original 12000 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 (-Si). SAGER: -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 of color shows the level of expression, with the darker colors representing higher expression.

**Extended Data**

Extended Data Table 1 Ka/Ks values of duplicated trypsin gene pairs.

Extended Data Table 2 List of the putative motifs of 291 alga trypsin proteins.

Extended Data Fig. 1 Chromosome/scaffold location of trypsin genes. a, *C. reinhardtii*; b, *F. cylindrus*; c, *F. solaris*; d, *T. 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.

Extended Data Fig. 2 WebLogo consensus sequences for 10 most conserved motifs of algal trypsins.

Supplementary File

Supplementary Table 1 Gene characteristics of algal trypsin genes.


**Supplementary Files**

This is a list of supplementary files associated with this preprint. Click to download.

- ExtendedDataCB.pdf
- SupplementaryFile.xlsx