Comparative gene analysis focused on silica cell wall formation: Identification of diatom-specific SET domain protein methyltransferases

Michiko Nemoto[1*], Sayako Iwaki[1], Hisao Moriya[1], Yuki Monden[1], Takashi Tamura[1], Kenji Inagaki[1], Shigeki Mayama[2], and Kiori Obuse[1]

[1] Graduate School of Environmental and Life Science, Okayama University, Okayama 700-8530, Japan

[2] Department of Biology, Tokyo Gakugei University, Tokyo 184-8511, Japan

[*]Correspondence: mnemoto@okayama-u.ac.jp

Acknowledgements

Computations were partially performed on the NIG supercomputer at ROIS National Institute of Genetics. This work was supported by the Program to Disseminate Tenure Tracking System from the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan, Grants-in-Aid for Scientific Research (C) (No. 18K05818), Asahi Glass Foundation to MN.

Abstract

Silica cell walls of diatoms have attracted attention as a source of nanostructured functional materials and have immense potential for a variety of applications. Previous studies of silica cell wall formation have identified numerous involved proteins, but most of these proteins are species-specific and are not conserved among diatoms. However, because the basic process of diatom cell wall formation is common to all diatom species, ubiquitous proteins and molecules will reveal the mechanisms of cell wall formation. In this study, we assembled *de novo* transcriptomes of three diatom species, *Nitzschia palea*, *Achnanthes kuwaitensis*, and *Pseudoleyanella lunata*, and compared protein-coding genes of five genome-sequenced diatom species. These analyses revealed a number of diatom-specific genes that encode putative endoplasmic reticulum-targeting proteins. Significant numbers of these proteins showed homology to silicanin-1, which is a conserved diatom protein that reportedly contributes to cell wall formation. These proteins also included a previously unrecognized SET domain protein methyltransferase family that may regulate functions of cell wall formation-related proteins and long-chain polyamines. Proteomic analysis of cell wall-associated proteins in *N. palea* identified a protein that is also encoded by one of the diatom-specific genes. Expression analysis showed that candidate genes were upregulated in response to silicon, suggesting that these genes play roles in silica cell wall formation. These candidate genes can facilitate further investigations of silica cell wall formation in diatoms.

Keywords

Biomineralization, diatom, silica, transcriptome, proteome

Introduction

Diatoms synthesize silica cell walls that have highly complex and elaborate nano-architectures. More than 100,000 diatom species with differing species-specific morphologies have been identified (Round et al. 1990). The unique porous architectures of silica cell walls offer promise for the production of next-generation materials for various applications (Terracciano et al. 2018; Kroger and Brunner 2014). However, to exploit biomimetic silica materials obtained from diatoms, the molecular mechanisms of silica cell wall formation need to be better characterized.

In recent decades, the understanding of proteins involved in silica cell wall formation in diatoms has enormously improved. Among recent insights, long-chain polyamines (LCPAs) and polysaccharides have been characterized as important components of silica cell walls (Brunner et al. 2009; Kroger et al. 2000; Sumper and Lehmann 2006). Moreover, seminal studies of *Cylindrotheca fusiformis* identified frustulins, pleuralins, and silaffins as cell wall-associated proteins, although frustulins extracted from the cell wall with EDTA were shown to be unrelated to silica formation (Kroger et al. 1994; Kroger et al. 1996; van de Poll et al. 1999). Pleuralins and silaffins have been extracted from the cell wall using anhydrous hydrogen fluoride (Kroger et al. 1997; Kroger et al. 1999). It was proposed that pleuralins are not involved in silica formation but rather connect the interface between hypotheca and epitheca (De Sanctis et al. 2016). In other studies, native forms of silaffins with post-translational modifications formed silica from silicic acid solutions (Kroger et al. 2002), and numerous cell wall-associated proteins were later identified from *Thalassiosira pseudonana*. Among these, tpSils, cingulins, silacidin, TpSAPs, and p150 were analyzed for their functions in silica formation (Poulsen and Kroger 2004; Scheffel et al. 2011; Kirkham et al. 2017; Wenzl et al. 2008; Davis et al. 2005; Tesson et al. 2017). G7408 from *Fistulifera solaris* was also identified as a cell wall-associated protein, although its function was not determined (Nemoto et al. 2014).

Apart from frustulins, which are common among diatoms (Armbrust et al. 2004; Bowler et al. 2008; Nemoto et al. 2014), most currently known silica cell wall-associated proteins are species-specific.

Silicanin-1 was the first protein to be identified as conserved among diatoms (Kotzsch et al. 2017) and is localized to silica deposition vesicles (SDVs), which are organelles with specialized roles in silica cell wall formation. When combined with LCPAs, recombinant silicanin-1 promoted silica formation from silicic acid solution. Recently, a silicanin-1 knockout mutant exhibited decreased silica levels and altered pore patterns in valves (Gorlich et al. 2019). These results demonstrate the importance of silicanin-1 in cell wall formation in diatoms. However, silica cell wall formation was not stopped in silicanin-1 knockout mutants, indicating the presence of additional universal proteins that are essential for cell wall formation in diatoms.

During recent years, genome information of several diatoms has become available. In addition to the genome of *T. pseudonana* (Armbrust et al. 2004), which was the first to be sequenced and annotated, the genomes of *Phaeodactylum tricornutum* (Bowler et al. 2008), *F. solaris* (Tanaka et al. 2015), *Thalassiosira oceanica* (Lommer et al. 2012), and *Fragilariopsis cylindrus* (Mock et al. 2017) are now available in database. Furthermore, developments in RNA sequencing technologies have produced more sequence information from several diatom species, whose genomes have not yet been sequenced.

In this study, we performed transcriptome analyses of the diatom species *Nitzschia palea*, *Achnanthes kuwaitensis,* and *Pseudoleyanella lunata* to investigate silica cell wall formation in diatoms. To identify diatom-specific genes, three transcriptome datasets were developed and protein-coding gene sets of five sequenced diatoms were compared. We then used transcriptome data in proteomic analyses and identified silica cell wall-associated proteins in *N. palea*.

Materials and Methods

Strains and culture conditions

N. palea (NIES-487) and *A. kuwaitensis* (NIES-1349) strains were purchased from the National Institute for Environmental Studies, Tsukuba, Japan (NIES). *P. lunata* was kindly provided by Noriaki Nakamura (Fukui Prefectural University). *N. palea* was cultured in CSi medium (Watanabe et al. 1988); *A. kuwaitensis* and *P. lunata* were cultured in f/2 medium (Guillard and Ryther 1962). All cultures were cultured at 20°C with continuous illumination at 15 μmol/m^2/s.

RNA isolation and cDNA library construction

During the mid-logarithmic phase, diatom cells were collected by centrifugation, frozen immediately in liquid nitrogen, and stored at −80°C until use. Total RNA was extracted from diatom samples using TRI reagent (Molecular Research Center, Cincinnati, OH) and was purified using RNeasy mini kits (QIAGEN, Hilden, Germany) according to the manufacturer's instructions. The integrity and quantity of extracted RNA was assessed using an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA). Subsequently, cDNA libraries were prepared using TruSeq Stranded mRNA Sample Prep Kits (Illumina, San Diego, CA).

RNA sequencing and *de novo* transcriptome assembly

cDNA libraries were sequenced on an Illumina Hiseq 2500 platform to generate paired-end reads. Adapter sequences were removed from raw reads using Cutadapt (Martin 2011). Reads were then quality trimmed using Trimmomatic (Bolger et

al. 2014), and the trimmed reads were assembled *de novo* into a transcriptome reference set using Trinity (Haas et al. 2013).

Bioinformatics analysis

Comparative gene analyses were conducted using transcriptome data of *N. palea*, *A. kuwaitensis*, and *P. lunata* and open reading frames (ORF) in genomes of *F. cylindrus* (18,111 sequences), *T. pseudonana* (11,673 sequences), *T. oceanica* (34,642 sequences), *P. tricornutum* (10,408 sequences), and *F. solaris* (20,429 sequences) obtained from the GenBank database of NCBI. OrthoFinder (Emms and Kelly 2019) was used to identify the homologous sequences of known silica cell wall-associated proteins in eight diatom species. OrthoFinder is an alignment-based method that identifies homologous relationships among several sequences. To identify diatom-specific genes, ORFs that were shared by all diatom species were further searched using the local BLAST algorithm with transcriptome and genome data from eight diatom species. BLAST searches were performed using E-values of 1e-4 as cutoffs (Altschul et al. 1997). ORFs that were shared only by diatom species were further searched in the NCBI protein database using BLAST. ORFs that were not homologous with any sequences of non-diatom organisms were selected as diatom-specific genes. This screening process was conducted using an in-house program. Screening for endoplasmic reticulum (ER)-targeting signal sequences was performed using SignalP 4.1 (Petersen et al. 2011) and HECTAR (Gschloessl et al. 2008). To evaluate the expression levels of unigenes in *N. palea*, *A. kuwaitensis*, and *P. lunata*, trimmed reads were obtained from each sample and were aligned to the constructed transcriptome reference using Bowtie (Langmead et al. 2009). Based on these mapping analyses, we estimated expression levels as transcripts per million (TPM) using the RSEM package (Li and Dewey 2011). Conserved domains of protein sequences were then searched using SMART (Schultz et al. 1998), and intrinsically disordered protein (IDP) regions in

protein sequences were predicted using ESpritz (Walsh et al. 2012). Amino acid composition was analyzed using the COPid web server (http://www.imtech.res.in/raghava/copid).

Phylogenetic analysis

18S rRNA genes from *A. kuwaitensis* (NIES-1349) and *C. fusiformis* (NIES-2351) were amplified and sequenced using the primers 5′-GGT GAT CCT GCC AGT AGT CAT ATG CTT G-3′ (ss5) and 5′-GAT CCT TCC GCA GGT TCA CCT ACGGAA ACC-3′ (ss3). 18S rRNA sequences of representative taxa (Theriot et al. 2010) and the diatom species *N. palea* (LC037443.1), *P. lunata* (LC164794.1), *F. cylindrus* (LC189084.1), *F. solaris* (AB769957.1), and *T. oceanica* (HM991696.1) were also included in the analyses. All 18S rRNA sequences were aligned using MEGA, version 6.0 (Tamura et al. 2013). Phylogenetic analyses of 18S rRNA were then performed using RAxML, version 8.2.12 (Stamatakis 2014). These analyses were performed using the GTR+G+I model with 1,000 bootstrap replicates.

Proteome analyses

Proteins were extracted from silica cell walls as described previously with slight modifications (Kroger et al. 1997; Kroger et al. 2000). Briefly, cells were harvested from 1.5 L cultures and were stored at −20°C until use. Harvested cells were then boiled twice in 2% sodium dodecyl sulfate (SDS)/100 mM EDTA for 30 min. The resulting cell wall samples were repeatedly washed with acetone until colorless. After washing with MilliQ water, cell wall samples were lyophilized and mixed with hydrofluoric acid on ice. After 30 min, mixtures were centrifuged and supernatants were collected and dried *in vacuo*. Residues were finally dissolved in 100 mM Tris (pH 8.8) and were neutralized with 1 M Tris. In some cases, protein

extracts were dialyzed against MilliQ water using a dialysis membrane with a molecular weight cutoff (MWCO) of 100 to 500 Da at 4°C for at least 50 h. Protein solutions were finally condensed by lyophilization.

Extracted proteins were subjected to tricine SDS-polyacrylamide gel electrophoresis, and gels were stained using the Silver Stain MS Kit (Fujifilm Wako Pure Chemicals, Osaka, Japan). Protein bands were then excised from gels and were digested with proteases as described previously (Shevchenko et al. 1996). In-gel digests of individual protein bands were analyzed using high-performance liquid chromatography (HPLC)-Chip and quadrupole time-of-flight mass spectroscopy (QTOF-MS; G6520 and G4240, Agilent Technologies, Santa Clara, CA). The translated databases were constructed from the present transcriptome data. Tandem mass spectra (MS/MS) data of cell wall-associated proteins were then searched against a translated database using the MASCOT algorithm for protein identification. The parameters used for MASCOT were used as previously described (Nemoto et al. 2012).

Quantitative reverse transcriptase-polymerase chain reaction (qRT-PCR) expression analysis of unigenes in *N. palea* during cell wall formation

N. palea cells were grown until the early-logarithmic phase in 1.5 L of pre-cultures. Cells were harvested and washed in Si-free CSi medium and then resuspended in 150 mL of Si-free CSi medium and equally distributed into three polycarbonate flasks. After 24 h of incubation, sodium silicate was added to cultures at a concentration of 352 μM. Cultures were sampled prior to the addition of sodium silicate and at 5 min, 5 h, 11 h, 14 h, 17 h, and 20 h after silicate addition. Si contents and cell numbers were then determined, and RNA was isolated. To analyze Si concentrations, culture media samples were centrifuged at 14,000 × g for 5 min and the supernatants were collected and stored at −20°C until analysis.

Silicate concentrations were determined using the molybdate method as described previously (Strickland and Parsons 1968; Hildebrand et al. 2007).

Total RNA was extracted for qRT-PCR analysis as described above and then converted into first strand cDNA using a LunaScript RT SuperMix Kit (New England Biolabs, Ipswich, MA). Subsequent qRT-PCR analyses were performed using a Mx3000P QPCR System (Agilent Technologies, Santa Clara, CA) with Luna Universal qPCR Master Mix (New England BioLabs, Ipswich, MA). Primer specificity was confirmed in melting curve analyses, and target mRNA expression levels were normalized to that of the TATA box-binding protein (Np272). Relative expression levels were calculated using the ΔΔCt method. All primers used for PCR are listed in Table S1.

Data availability

Reads from Illumina Hiseq analyses were deposited into the DDBJ Sequence Read Archive with accession number DRA009136.

Results

RNA sequencing and *de novo* transcriptome assembly

N. palea, A. kuwaitensis, and *P. lunata* cells were grown to the logarithmic phase and were then harvested for RNA extraction. RNA sequences were determined using an Illumina Hiseq platform followed by quality read trimming. Trimmed reads were then assembled *de novo* into transcripts. Results for 40,498, 71,930, and 45,400 contigs were obtained for *N. palea*, *A. kuwaitensis*, and *P. lunata*, respectively (Table S2). Related contigs were clustered into unique gene sequences,

and 31,946, 60,767 and 38,314 unigenes were generated for *N. palea*, *A. kuwaitensis*, and *P. lunata*. Among these, 19,714 (62%), 34,888 (57%), and 27,098 unigenes (70%) were covered by at least 10 high-quality reads. Numbers of unigenes in transcriptome data were much larger than numbers of predicted ORFs in the diatom genome.

Identification of homologous sequences of known silica cell wall-associated proteins

The genes showing homology with genes coding for known silica cell wall-associated proteins were screened from transcriptome data for the present diatom species and from genome data for five genome-sequenced diatoms using OrthoFinder software (Emms and Kelly 2019). Homologs of frustulins and silicanin-1 were identified from all diatom species (Table 1 and Table S3). Frustulins are predominantly localized in organic casings of the silica cell walls and may not be directly involved in silica formation (van de Poll et al. 1999). Silicanin-1 is a recently reported membrane protein that is localized in SDVs and is highly conserved among all diatoms (Kotzsch et al. 2017). Recombinantly expressed silicanin-1 fragments showed silica precipitation in the presence of LCPAs. Homologs of p150 were identified from all diatoms, except from *F. solaris*. The acidic protein p150 is associated with the girdle band region, and its role in silicification is not well understood (Davis et al. 2005). Sequence alignments revealed homologs with similarity to the C-terminal region of p150, including the chitin-binding domain (Fig. S1). Homologs of SiMat3 and SiMat5 were identified from *P. lunata* and *T. oceanica*, respectively. SiMats, standing for silica matrices, are proteins composed of insoluble organic matrices of silica cell wall (Kotzsch et al. 2016). Homologs of TpSAP1 and TpSAP3 were identified from *P. lunata*, *F. cylindrus*, and *T. oceanica*. SAPs, stand for silicalemma-associated proteins, are a family of proteins associated with the SDV membrane (silicalemma) (Tesson et al. 2017). The knockdown lines of TpSAP1 and TpSAP3 in *T. pseudonana*

displayed altered silica morphologies. The sequence alignments of SiMat3, SiMat5, TpSAP1, and TpSAP3 with their homologs are shown in Fig. S2 to Fig. S5. The cell wall-associated proteins silaffins, tpSils, cingulins, silacidin, and G7408 were not detected in genome and transcriptome data of all diatoms. Homologs of pleuralin-1 were identified only in *N. palea* (Table 1 and Table S3). Because homology searches of pleuralin-1 in the NCBI database of genome-sequenced diatoms retrieved no hits, we concluded that these homologs in *N. palea* are the only proteins with homology to pleuralin-1. Pleuralin-1 was previously extracted from cell walls of *C. fusiformis* using hydrogen fluoride (Kroger et al. 1997) and was reportedly localized to pleural bands of the epitheca (Kroger and Wetherbee 2000). Pleuralin-1 has a unique sequence comprising an N-terminal proline-rich region, followed by proline, serine, cysteine, and aspartate-rich repeat sequences (PSCD domains) and a C-terminal region. Among the homologous sequences, the Np5566 sequence, which was found to be the best hit sequence by the BLAST search, was aligned with the Pleuralin-1 sequence (Fig. S6). Although the 5′ terminus of the Np5566 sequence was truncated, similarities with N-terminal proline-rich and C-terminal regions of pleuralin-1 were observed. The positions of cysteine and acidic amino acid residues in the C-terminal region are particularly well conserved between pleuralin-1 and Np5566.

Identification of diatom-specific genes

We hypothesized that some core genes involved in silica cell wall formation are conserved and exclusively present in diatoms. To identify diatom-specific core genes, transcriptome sequences of *N. palea*, *A. kuwaitensis*, and *P. lunata* and all predicted genes in the genomes of *F. cylindrus*, *T. oceanica*, *T. pseudonana*, *F. solaris*, and *P. tricornutum* were compared using BLAST searches (Fig. 1). A total of 6,841 to 15,654 genes were shared between the eight diatom species

(Fig. 2A). Among them, 590 to 1,830 did not share homology with genes of non-diatom organisms (Fig. 2A). All previously reported silica cell wall-associated proteins contain N-terminal ER signal peptides. Therefore, we predicted the presence of ER signal peptides in deduced amino acid sequences of these diatom-specific genes. Because transcriptome data contain 5′ truncated sequences, which leads to difficulties in predicting signal peptides, sequences of 845 diatom-specific genes in the *F. cylindrus* genome were further used for the analyses. Analyses of predicted protein sequences using SignalP and HECTAR retrieved 73 genes that encode the putative ER-targeting proteins listed in Table S4. Genes of other diatoms that best matched sequences of *F. cylindrus* are listed in Table S4. Expression levels of corresponding unigenes in *N. palea*, *A. kuwaitensis*, and *P. lunata* were estimated using an alignment-based method (Li and Dewey 2011) (Table S4).

Initially, we used these 73 proteins in BLAST searches for genes encoding known silica cell wall-associated proteins. Twelve of these proteins shared homology with silicanin-1 (Table S4), and with the exception of few truncated unigenes, silicanin-1 and silicanin-1-like proteins possessed C-terminal transmembrane domains. Moreover, NQ-rich domains that are characteristic of silicanin-1 were present in these silicanin-1 like proteins (Kotzsch et al. 2017).

In further analyses, the remaining 61 proteins were analyzed using SMART software and domain structures were compared. Seven proteins contained SET domains (Fig. 2B and Table S4), and five of these proteins were homologous to each other. Several SET domain proteins have been shown to methylate lysine residues of substrate proteins (Herz et al. 2013). Among them, histone methyltransferases (HMTs) and Rubisco large subunit methyltransferases (LSMTs) are well characterized, whereas substrates of most other SET domain proteins remain unknown (Trievel et al. 2002). Diatom-specific SET domain proteins contain 1 to 4 SET domains (Fig. 3A) and were named Bacillariophyceae-specific SET domain (BacSET) proteins. BacSET1 proteins of *N. palea*, *P. tricornutum*, and *F. solaris* contain spermine/spermidine

synthase domains, through which they transfer aminopropyl groups from decarboxylated *S*-adenosyl-L-methionine (AdoMet) to amine for spermine/spermidine biosynthesis (Wu et al. 2007). Fig. 3B shows a comparison of SET domains from diatoms with those from *Arabidopsis thaliana* LSMT (AtLSMT) and *Saccharomyces cerevisiae* HMT (ScHMT). In these analyses, positions of putative catalytic Tyr residues and other residues that are involved in AdoMet binding were well conserved, whereas residues involved in substrate lysine binding were not. In homology searches, we identified sets of putative LSMTs and HMTs that share high sequence homology with AtLSMT and ScHMT genes from diatoms (data not shown). Subsequent protein domain analyses further identified three proteins with protein binding domains, such as PDZ and WW domains (Table S4).

It has been reported that a number of biomineral-associated matrix proteins contain IDP regions and often comprise highly repetitive and biased amino acid sequences (Magdalena et al. 2012). Among the 73 diatom-specific proteins identified herein, six contained predicted long IDP regions of ≥30 residues (Table S4). These proteins have unique repetitive sequences that are enriched with specific amino acid residues, as commonly observed in IDPs. In similar predictions, previously reported silica cell wall-associated proteins were demonstrated to contain putative long IDP regions (Kotzsch et al. 2016; Kroger et al. 1997; Kroger et al. 1999; Poulsen and Kroger 2004; Scheffel et al. 2011; Wenzl et al. 2008). The remaining proteins contained unique sequences with biased amino acid compositions, including an AK-rich protein, several acidic amino acid-rich proteins, an S-rich protein, and a GS-rich protein (Table S4).

Proteomic analyses of cell wall-associated proteins in *N. palea*

In previous studies, numerous silica cell wall-associated proteins have been identified in diatoms. Hence, to identify cell

wall-associated proteins in *N. palea*, *A. kuwaitensis*, and *P. lunata*, proteins were extracted from the three diatom species using hydrofluoric acid. This protocol for extracting protein from cell walls slightly differs from the previously reported method (Kroger et al. 1997), which used anhydrous hydrogen fluoride for protein extraction. Under these conditions, no clear protein bands were observed in cell wall extracts from *A. kuwaitensis* and *P. lunata*, but several protein bands of approximately 10 and 15 kDa were observed in *N. palea* samples (Fig. 4A). The protein content of the detergent-purified silica cell walls of *N. palea* was 4.5% ± 0.7% of dry weight (n = 5). Peptides were extracted from the protein bands (Fig. 4A) and were analyzed using liquid chromatography-MS/MS. We then generated a translated database from the assembled transcripts of *N. palea* and analyzed peptides using an MS/MS ion search with the translated database. As a result, the identified peptide sequences were assigned to 14 unigenes (Table S5 and Fig. 4A). Among these, eight are photosynthesis-related proteins, such as fucoxanthin chlorophyll a/c-binding proteins, which are known to be abundant in diatoms. These photosynthesis-related proteins and the other four proteins (elongation factor-like protein, ATP synthase, mitochondrial ATPase, and 14-3-3-like protein) were unrelated to silica cell wall formation and were likely contaminants in our procedures. The remaining two proteins include a novel protein (Np23207) that does not share homology with any known protein. The function of the other protein Np12828 is also unknown, but it was identified as a diatom-specific protein (Table S4). These proteins are termed silica matrix proteins (SMPs). NpSMP1 (Np12828) contains an N-terminal signal peptide and characteristic repetitive sequences (Fig. 4B). One of the diatom-specific genes, Np6444, shares homology with NpSMP1.

Expression patterns of unigenes during cell wall formation

In functional assessments of the unigenes identified above, expression patterns during cell wall formation in *N. palea* were analyzed. We cultured cells in Si-free medium for 24 h as described previously (Hildebrand et al. 2007). After adding silicate, cell growth and media Si concentrations were monitored at each sampling time point (Fig. 5A). In culture media, Si concentrations were decreased to 0.2 μM after 20 h. The cell numbers gradually increased, and doubling was observed after 14 h (Fig. 5A).

Expression patterns of unigenes during cell wall formation were determined using qRT-PCR analyses of total RNA from cells collected at each time point. Based on these expression patterns, unigenes were classified into four groups. Specifically, silicanin-1 (Np16494), NpSMP2 (Np23207), NpBacSET6 (Np5036), and NpBacSET7 (Np13862) were significantly induced at 5 h after Si replenishment (Fig. 5B, C), coinciding with the initial absorption of Si (Fig. 5A). NpSMP1 (Np12828), NpBacSET5 (Np8261), and NpBacSET1 (Np20729) expression levels increased with the progression of cell growth and decreased after cell growth had ceased. In contrast, NpBacSET3 (Np8008), NpBacSET4 (Np11017), and NpBacSET2 (Np10391) were consistently expressed over 5 to 20 h time points. Finally, pleuralin-1 like protein (Np5566) exhibited a unique expression pattern, with dramatic upregulation at 14 h that coincided with cell division (Fig. 5B).

Discussion

In this study, we conducted the first extensive comparative analysis of genes from five genome-sequenced diatoms and the transcriptomes of *N. palea*, *A. kuwaitensis*, and *P. lunata*, with a focus on genes involved in silica cell wall formation.

Unigenes were more numerous in transcriptome data than ORFs in the diatom genome, likely reflecting

fragmented short read assemblies. To evaluate the qualities of transcriptome assemblies, numbers of genes that were conserved between the genomes of five diatoms and the transcriptomes of three diatoms were compared, revealing 7,157 genes that were conserved between the five diatom genomes. We compared these genes with transcriptome sequences of three diatom species and revealed 6,841 common genes (96%). These analyses support the completeness of our *de novo* transcriptome assemblies.

Homologs of the known silica cell wall-associated proteins frustulins and silicanin-1 were identified in all diatom species. These proteins are reportedly conserved among diatom species (Nemoto et al. 2014; Bowler et al. 2008; Kotzsch et al. 2017; Armbrust et al. 2004), whereas homologs of other proteins, including tpSils and cingulins, were not identified even in *T. oceanica*, which is phylogenetically closest to *T. pseudonana* (Table 1, Table S3, Fig. S7). We contend, therefore, that silica cell wall-associated proteins include essential proteins that are highly conserved among diatoms, in addition to proteins with high sequence heterogeneity.

The silica cell wall-associated protein pleuralin-1 was originally identified from *C. fusiformis* but was apparently absent in other diatom genomes (Kroger and Poulsen 2008). In agreement, our homology searches for pleuralin-1 in the NCBI nonredundant database retrieved no hits, although a number of pleuralin-1-like proteins were found to be present in *N. palea* (Table 1 and Table S3). Pleuralin-1 was previously isolated from cell walls using anhydrous hydrogen fluoride (Kroger et al. 1997). Although its specific roles in silica cell wall formation are unknown, pleuralin-1 has been found to be attached to newly deposited pleural bands of the hypotheca (Kroger and Wetherbee 2000; Kroger and Poulsen 2008). In the diatom cell cycle, pleural bands of the hypotheca are synthesized after cell separation. The pleuralin-1-like protein Np5566 was most highly expressed in *N. palea* at 14 h, coinciding with the end of cell division (Fig. 5). These results are

consistent with previously reported localization dynamics of pleuralin-1 (Kroger and Wetherbee 2000).

Comparisons of genes between eight diatom species identified 590 to 1,830 diatom-specific genes. Further analysis revealed that among 845 diatom-specific genes in *F. cylindrus*, 73 encoded putative ER-targeting proteins. In agreement with a previous report (Kotzsch et al. 2017) in which silicanin-1 was established as a conserved diatom protein, a silicanin-1 gene was found among the diatom-specific genes (Table S4). Furthermore, 11 proteins were homologous to silicanin-1, suggesting that our method successfully identifies diatom-specific proteins with potential roles in cell wall formation. Also among the diatom-specific genes, we characterized seven as encoding BacSET proteins (Fig. 2B and Table S4). SET domain proteins have been shown to mono-, di-, and tri-methylate lysine residues of substrate proteins (Dillon et al. 2005). Silaffins, which are silica cell wall-associated proteins that regulate silica polymerization in *C. fusiform*, have dimethylated or trimethylated lysine residues (Kroger et al. 1999; Kroger et al. 2002). The di- and tri-methylated lysine residues have also been detected in silica cell wall-associated proteins from other diatom species (Poulsen and Kroger 2004; Wenzl et al. 2004). In addition, LCPAs from silica cell walls have been shown to be *N*-methylated (Sumper and Brunner 2006). Accordingly, detailed analyses of synthetic polyamines and silaffin peptides revealed that silica polymerization activity and the resulting silica morphologies are dependent on methylation status (Bernecker et al. 2010; Lechner and Becker 2012). Quaternary ammonium groups of these molecules have been shown to favor the formation of oligosilicate anions that affect silica polymerization (Hoebbel et al. 1980). By comparing sequences of our BacSET proteins with those of HMT and LSMT, we revealed that the residues involved in catalysis and AdoMet binding are well conserved, whereas those involved in substrate lysine binding are less conserved in the BacSET proteins (Fig. 3B). The present BacSET protein family might be a novel, diatom-specific family of methyltransferases that target unique substrates, such

as cell wall formation-related proteins and LCPAs. In particular, NpBacSET1, PtBacSET1, and FsBAcSET1 contain spermine/spermidine synthase domains that are involved in aminopropyl transfer. Given the presence of multiple catalytic domains, these proteins may be involved in the biosynthesis of diatom-specific LCPAs or aminopropyl group transfer to cell wall formation-related proteins, such as silaffins that contain lysine modified with 6–11 *N*-methyl propylamine repeats (Kroger et al. 1999). In this study, we present gene expression analyses showing diverse patterns of SET domain proteins (Fig. 5). During cell wall formation in diatoms, valve parts are initially formed in valve SDVs, and after completion of valve formation, girdle bands are formed in separate girdle band SDVs. Two SET domain proteins (NpBacSET6 and NpBacSET7) were induced during the early stages of cell wall formation, suggesting roles in methylation of molecules that are related to valve formation (Fig. 5). Other SET domain proteins (NpBacSET1 and NpBacSET5) had increased expression levels during later stages of cell wall formation. Perhaps these proteins are involved in girdle band biogenesis (Fig. 5).

Numerous IDPs have been isolated from biominerals and characterized for their involvement in biomineralization (He et al. 2003; Sarem et al. 2017; Shen et al. 1997). Silica cell wall-associated proteins, such as silaffins, cingulins, and silacidins, were also predicted to be IDPs. Among our diatom-specific genes, six encode putative long IDP regions (Table S4). IDPs have been shown to exhibit conformational flexibility and structural dynamics (Uversky 2019), allowing control of biomineralization by conformationally adapting to minerals across several mineral phases.

Several proteins have sequences with biased amino acid compositions, such as high ratios of Lys, Ala, Asp, Glu, Gly, or Ser (Table S4). Similar characteristics have been observed in silica cell wall-associated proteins. In particular, silaffins and other proteins are known as GS-rich proteins (Kroger et al. 1999; Scheffel et al. 2011; Nemoto et al. 2014),

silacidin is an acidic amino acid-rich protein, and tpSil3p from *T. pseudonana* is a KA-rich protein (Wenzl et al. 2008; Poulsen and Kroger 2004). We believe that the amino acids enriched in these proteins have specific functions in silica cell wall formation.

In contrast with previous studies (Kroger et al. 1997; Kroger et al. 1999), we extracted silica cell wall-associated proteins with hydrofluoric acid instead of anhydrous hydrogen fluoride. The resulting protein contents of silica cell walls from *N. palea* were less than one-tenth of those in *T. pseudonana* (Kroger et al. 1999), suggesting partial degradation of proteins during hydrofluoric acid treatments. Moreover, our proteomic analyses of cell wall-associated proteins in *N. palea* identified 14 proteins, including two with unknown function. The remaining 12 proteins were associated with photosynthesis or other functions that are unrelated to silica cell wall formation (Fig. 4A and Table S5). Hence, despite repeated washing of cell wall fractions with SDS and acetone, many contaminants may remain in our samples, as previously reported (Kotzsch et al. 2016). Nonetheless, the gene Np12828, which encodes one of two NpSMPs, was identified as a diatom-specific gene in our comparative analyses (Table S4). We show that NpSMP1 expression was significantly induced at 11 h and decreased at 20 h, whereas that of the other cell wall-associated protein NpSMP2 was induced at 5 h and started to decrease at 17 h (Fig. 5), suggesting roles of NpSMP2 and NpSMP1 in valve and girdle band formation, respectively.

In conclusion, we constructed *de novo* transcriptome assemblies of the diatom species *N. palea*, *A. kuwaitensis*, and *P. lunata*. These transcriptomes offer new resources for understanding silica biomineralization and other physiological machinery in diatoms, which are ecologically important organisms. Comparisons of these transcriptomes with known silica cell wall-associated proteins revealed that *N. palea* carries homologs of pleuralin-1 that had only been identified previously in *C. fusiformis.* Moreover, our comprehensive comparative analysis of eight diatom species identified a number of diatom-

specific genes that have likely associations with silica cell wall formation, including a previously unrecognized family of SET domain proteins. Subsequent transcriptome and proteomic analyses revealed the presence of specific silica cell wall-associated proteins in *N. palea*. Further comparisons of diatom genes with genes from other organisms producing biosilica structures, such as parmales and siliceous sponges, will likely reveal genes with general roles in biosilica formation. Moreover, continued investigations of the genes identified in this study are required to determine specific substrates of diatom-specific SET domain proteins and characterize the subcellular localizations of candidate proteins. Such studies will greatly extend the knowledge of silica cell wall formation in diatoms, with likely insights into exploitable mechanisms of biosilica formation.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W, Lipman, DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25**:** 3389-3402

Armbrust EV, Berges JA, Bowler C, Green BR, Martinez D, Putnam NH, Zhou SG, Allen AE, Apt KE, Bechner M, Brzezinski MA, Chaal BK, Chiovitti A, Davis AK, Demarest MS, Detter JC, Glavina T, Goodstein D, Hadi MZ, Hellsten U, Hildebrand M, Jenkins BD, Jurka J, Kapitonov VV, Kroger N, Lau WWY, Lane TW, Larimer FW, Lippmeier JC, Lucas S, Medina M, Montsant A, Obornik M, Parker MS, Palenik B, Pazour GJ, Richardson PM,

Rynearson TA, Saito MA, Schwartz DC, Thamatrakoln K, Valentin K, Vardi A, Wilkerson FP, Rokhsar DS (2004) The genome of the diatom *Thalassiosira pseudonana*: Ecology, evolution, and metabolism. Science 306**:** 79-86

Bernecker A, Wieneke R, Riedel R, Seibt M, Geyer A Steinem C (2010) Tailored Synthetic Polyamines for Controlled Biomimetic Silica Formation. J Am Chem Soc 132**:** 1023-1031

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30**:** 2114-2120

Bowler C, Allen AE, Badger JH, Grimwood J, Jabbari K, Kuo A, Maheswari U, Martens C, Maumus F, Otillar RP, Rayko E, Salamov A, Vandepoele K, Beszteri B, Gruber A, Heijde M, Katinka M, Mock T, Valentin K, Verret F, Berges JA, Brownlee C, Cadoret JP, Chiovitti A, Choi CJ, Coesel S, De Martino A, Detter JC, Durkin C, Falciatore A, Fournet J, Haruta M, Huysman MJJ, Jenkins BD, Jiroutova K, Jorgensen RE, Joubert Y, Kaplan A, Kroger N, Kroth PG, La Roche J, Lindquist E, Lommer M, Martin-Jezequel V, Lopez PJ, Lucas S, Mangogna M, Mcginnis K, Medlin LK, Montsant A, Oudot-Le Secq MP, Napoli C, Obornik M, Parker MS, Petit JL, Porcel BM, Poulsen N, Robison M, Rychlewski L, Rynearson TA, Schmutz J, Shapiro H, Siaut M, Stanley M, Sussman MR, Taylor AR, Vardi A, Von Dassow P, Vyverman W, Willis A, Wyrwicz LS, Rokhsar DS, Weissenbach J, Armbrust EV, Green BR, Van De Peer Y, Grigoriev IV (2008) The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes. Nature 456**:** 239-244

Brunner E, Richthammer P, Ehrlich H, Paasch S, Simon P, Ueberlein S, Van Pee KH (2009) Chitin-Based Organic Networks: An Integral Part of Cell Wall Biosilica in the Diatom *Thalassiosira pseudonana* Angew Chem Int

Edit 48**:** 9724-9727

Davis AK, Hildebrand M, Palenik B (2005) A stress-induced protein associated with the girdle band region of the diatom *Thalassiosira pseudonana* (Bacillariophyta). J Phycol 41**:** 577-589

De Sanctis S, Wenzler M, Kroger N, Malloni WM, Sumper M, Deutzmann R, Zadravec P, Brunner E, Kremer W, Kalbitzer HR (2016) PSCD Domains of Pleuralin-1 from the Diatom *Cylindrotheca fusiformis*: NMR Structures and Interactions with Other Biosilica-Associated Proteins. Structure 24**:** 1178-1191

Dillon SC, Zhang X, Trievel RC, Cheng XD (2005) The SET-domain protein superfamily: protein lysine methyltransferases. Genome Biol. https://doi.org/ 10.1186/gb-2005-6-8-227

Emms DM, Kelly S (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. https://doi.org/10.1186/s13059-019-1832-y

Gorlich S, Pawolski D, Zlotnikov I, Kroger N (2019) Control of biosilica morphology and mechanical performance by the conserved diatom gene Silicanin-1. Communications Biology. https://doi.org/10.1038/s42003-019-0436-0

Gschloessl B, Guermeur Y, Cock JM (2008) HECTAR: A method to predict subcellular targeting in heterokonts. BMC Bioinformatics. https://doi.org/10.1186/1471-2105-9-393

Guillard R, Ryther J (1962) Studies of marine planktonic diatoms.1. *Cyclotella nana* Hustedt, and *Detonula confervacea*(cleve) gran Can J Microbiol 8**:** 229-239

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, Macmanes MD, Ott M, Orvis J, Pochet N, Strozzi F, Weeks N, Westerman R, William T, Dewey CN, Henschel R, Leduc RD, Friedman N, Regev A (2013) De novo transcript sequence reconstruction from RNA-seq using

the Trinity platform for reference generation and analysis. Nat Protoc 8**:** 1494-1512

He G, Dahl T, Veis A, George A (2003) Nucleation of apatite crystals in vitro by self-assembled dentin matrix protein 1. Nat Mater 2**:** 552-558

Herz HM, Garruss A, Shilatifard A (2013) SET for life biochemical activities and biological functions of SET domain-containing proteins. Trends Biochem Sci 38**:** 621-639

Hildebrand M, Frigeri LG, Davis AK (2007) Synchronized growth of *Thalassiosira pseudonana* (Bacillariophyceae) provides novel insights into cell-wall synthesis processes in relation to the cell cycle. J Phycol 43**:** 730-740

Hoebbel D, Garzo G, Engelhardt G, Ebert R, Lippmaa E, Alla M (1980) On The Constitution Of Silicate Anions In Tetraethylammonium Silicates And Their Aqueous-Solutions. Z Anorg Allg Chem 465**:** 15-33

Kirkham AR, Richthammer P, Schmidt K, Wustmann M, Maeda Y, Hedrich R, Brunner E, Tanaka T, Van Pee KH, Falciatore A, Mock T (2017) A role for the cell-wall protein silacidin in cell size of the diatom *Thalassiosira pseudonana*. ISME J 11**:** 2452-2464

Kotzsch A, Groger P, Pawolski D, Bomans PHH, Sommerdijk N, Schlierf M, Kroger N (2017) Silicanin-1 is a conserved diatom membrane protein involved in silica biomineralization. BMC biol. https://doi.org/10.1186/s12915-017-0400-8

Kotzsch A, Pawolski D, Milentyev A, Shevchenko A, Scheffel A, Poulsen N, Kroger N (2016) Biochemical Composition and Assembly of Biosilica-associated Insoluble Organic Matrices from the Diatom *Thalassiosira pseudonana*. J Biol Chem 291**:** 4982-4997

Kroger N, Bergsdorf C, Sumper M (1994) A new calcium-binding glycoprotein family constitutes a major diatom cell-

wall component. EMBO J 13**:** 4676-4683

Kroger N, Bergsdorf C, Sumper M (1996) Frustulins: Domain conservation in a protein family associated with diatom cell walls. Eur J Biochem 239**:** 259-264

Kroger N, Brunner E (2014) Complex-shaped microbial biominerals for nanotechnology. Wires Nanomed Nanobi 6**:** 615-627

Kroger N, Deutzmann R, Bergsdorf C, Sumper M (2000) Species-specific polyamines from diatoms control silica morphology. P Natl Acad Sci Usa 97**:** 14133-14138

Kroger N, Deutzmann R, Sumper M (1999) Polycationic peptides from diatom biosilica that direct silica nanosphere formation. Science, 286**:** 1129-1132

Kroger N, Lehmann G, Rachel R, Sumper M (1997) Characterization of a 200-kDa diatom protein that is specifically associated with a silica-based substructure of the cell wall. Eur J Biochem 250**:** 99-105

Kroger N, Lorenz S, Brunner E, Sumper M (2002) Self-assembly of highly phosphorylated silaffins and their function in biosilica morphogenesis. Science 298**:** 584-586

Kroger N, Poulsen N (2008) Diatoms-From Cell Wall Biogenesis to Nanotechnology. Annu Rev Genet 42**:** 83-107

Kroger N, Wetherbee R (2000) Pleuralins are involved in theca differentiation in the diatom *Cylindrotheca fusiformis.* Protist 151**:** 263-273

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. https://doi.org/10.1186/gb-2009-10-3-r25

Lechner CC, Becker CFW (2012) Exploring the effect of native and artificial peptide modifications on silaffin induced

silica precipitation. Chem Sci 3: 3500-3504

Li B, Dewey CN (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. https://doi.org/10.1186/1471-2105-12-323

Lommer M, Specht M, Roy AS, Kraemer L, Andreson R, Gutowska MA, Wolf J, Bergner SV, Schilhabel MB, Klostermeier UC, Beiko RG, Rosenstiel P, Hippler M, Laroche J (2012) Genome and low-iron response of an oceanic diatom adapted to chronic iron limitation. Genome Biol. https://doi.org/10.1186/gb-2012-13-7-r66

Magdalena W, Piotr D, Andrzej O (2012) Intrinsically Disordered Proteins in Biomineralization. In: Seto J (ed) Advanced Topics in Biomineralization. IntechOpen Limited, London

Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetjournal 17: 10-12

Mock T, Otillar RP, Strauss J, Mcmullan M, Paajanen P, Schmutz J, Salamov A, Sanges R, Toseland A, Ward BJ, Allen AE, Dupont CL, Frickenhaus S, Maumus F, Veluchamy A, Wu TY, Barry KW, Falciatore A, Ferrante MI, Fortunato AE, Glockner G, Gruber A, Hipkin R, Janech MG, Kroth PG, Leese F, Lindquist EA, Lyon BR, Martin J, Mayer C, Parker M, Quesneville H, Raymond JA, Uhlig C, Valas RE, Valentin KU, Worden AZ, Armbrust EV, Clark MD, Bowler C, Green BR, Moulton V, Van Oosterhout C, Grigoriev IV (2017) Evolutionary genomics of the cold-adapted diatom *Fragilariopsis cylindrus*. Nature 541: 536-540

Nemoto M, Maeda Y, Muto M, Tanaka M, Yoshino T, Mayama S, Tanaka, T (2014) Identification of a frustule-associated protein of the marine pennate diatom *Fistulifera* sp strain JPCC DA0580. Mar Genom 16: 39-44

Nemoto M, Wang QQ, Li DS, Pan SQ, Matsunaga T, Kisailus D (2012) Proteomic analysis from the mineralized radular teeth of the giant Pacific chiton, *Cryptochiton stelleri* (Mollusca). Proteomics 12: 2890-2894

Petersen TN, Brunak S, Von Heijne G, Nielsen H (2011) SignalP 40: discriminating signal peptides from transmembrane regions. Nat Methods 8**:** 785-786

Poulsen N, Kroger N (2004) Silica morphogenesis by alternative processing of silaffins in the diatom *Thalassiosira pseudonana*. J Biol Chem 279**:** 42993-42999

Round FE, Crawford RM, Mann DG (1990) The Diatoms : biology & morphology of the genera. Cambridge University Press, Cambridge

Sarem M, Ludeke S, Thomann R, Salavei P, Zou ZY, Habraken W, Masic A, Shastri VP (2017) Disordered Conformation with Low Pii Helix in Phosphoproteins Orchestrates Biomimetic Apatite Formation. Adv Mater. https://doi.org/10.1002/adma201701629

Scheffel A, Poulsen N, Shian S, Kroger N (2011) Nanopatterned protein microrings from a diatom that direct silica morphogenesis. P Natl Acad Sci Usa 108**:** 3175-3180

Schultz J, Milpetz F, Bork P, Ponting CP (1998) SMART, a simple modular architecture research tool: Identification of signaling domains. P Natl Acad Sci Usa 95**:** 5857-5864

Shen X, Belcher AM, Hansma PK, Stucky GD, Morse DE (1997) Molecular cloning and characterization of lustrin A, a matrix protein from shell and pearl nacre of *Haliotis rufescens*. J Biol Chem 272**:** 32472-81

Shevchenko A, Wilm M, Vorm O, Mann M (1996) Mass spectrometric sequencing of proteins from silver stained polyacrylamide gels. Anal Chem 68**:** 850-858

Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies Bioinformatics 30**:** 1312-1313

Strickland JDH, Parsons TR (1968) A Practical Handbook of Seawater Analysis. Fisheries Research Board of Canada, Ottawa

Sumper M, Brunner E (2006) Learning from diatoms: Nature's tools for the production of nanostructured silica. Adv Funct Mater 16**:** 17-26

Sumper M, Lehmann G (2006) Silica pattern formation in diatoms: Species-specific polyamine biosynthesis Chembiochem 7**:** 1419-1427

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. Mol Biol Evol 30**:** 2725-2729

Tanaka T, Maeda Y, Veluchamy A, Tanaka M, Abida H, Marechal E, Bowler C, Muto M, Sunaga Y, Yoshino T, Taniguchi T, Fukuda Y, Nemoto M, Matsumoto M, Wong PS, Aburatani S, Fujibuchi W (2015) Oil Accumulation by the Oleaginous Diatom *Fistulifera solaris* as Revealed by the Genome and Transcriptome Plant Cell 27**:** 162-176

Terracciano M, De Stefano L, Rea I (2018) Diatoms Green Nanotechnology for Biosilica-Based Drug Delivery Systems Pharmaceutics. https://doi.org/10.3390/pharmaceutics10040242

Tesson B, Lerch SJL, Hildebrand M (2017) Characterization of a New Protein Family Associated With the Silica Deposition Vesicle Membrane Enables Genetic Manipulation of Diatom Silica. Sci Rep-Uk, https://doi.org/10.1038/s41598-017-13613-8

Theriot EC, Ashworth M, Ruck E, Nakov T, Jansen RK (2010) A preliminary multigene phylogeny of the diatoms (Bacillariophyta): challenges for future research. Plant Ecol Evol 143**:** 278-296

Trievel RC, Beach BM, Dirk LMA, Houtz RL, Hurley JH (2002) Structure and catalytic mechanism of a SET domain protein methyltransferase. Cell 111**:** 91-103

Uversky VN (2019) Intrinsically Disordered Proteins and Their "Mysterious" (Meta)Physics. AIP Conf Proc. https://doi.org/10.3389/fphy201900010

Van De Poll WH, Vrieling EG, Gieskes WWC (1999) Location and expression of frustulins in the pennate diatoms *Cylindrotheca fusiformis*, *Navicula pelliculosa,* and *Navicula salinarum* (Bacillariophyceae). J Phycol 35**:** 1044-1053

Walsh I, Martin AJM, Di Domenico T, Tosatto SCE (2012) ESpritz: accurate and fast prediction of protein disorder. Bioinformatics 28**:** 503-509

Watanabe M, Kasai F, Sudo R (1988) NIES Collection list of strains Second Edition Microalgae and protozoa. NIES, Japan

Wenzl S, Deutzmann R, Hett R, Hochmuth E, Sumper M (2004) Quaternary ammonium groups in silica-associated proteins. Angew Chem Int Edit 43**:** 5933-5936

Wenzl S, Hett R, Richthammer P, Sumper M (2008) Silacidins: Highly acidic phosphopeptides from diatom shells assist in silica precipitation in vitro. Angew Chem Int Edit 47**:** 1729-1732

Wu H, Min JR, Ikeguchi Y, Zeng H, Dong AP, Loppnau P, Pegg AE, Plotnikov AN (2007) Structure and mechanism of spermidine synthases. Biochemistry 46**:** 8331-8339

Figure Legends

Fig. 1 Flow chart of the method used to identify the diatom-specific proteins with ER signal sequences.

Fig. 2 Comparison of transcriptomes and genomes of diatoms. (A) Categorization of the predicted protein-coding genes in transcriptome and genome sequences of diatoms. (B) Features of diatom-specific proteins with ER signal sequences in *F. cylindrus*, including 12 proteins showing homologies with silicanin-1, 7 proteins with SET domains, 5 proteins with biased amino acid (AA) compositions, 4 proteins containing predicted long IDP regions, 2 proteins containing predicted long IDP regions and protein binding domains (PDZ domain and WW domain), and 1 protein with a protein binding domain (PDZ domain).

Fig. 3 Analysis of Bacillariophyceae-specific SET domain (BacSET) proteins. (A) schematic of SET domain protein structures in diatoms; (B) representative amino acid sequence alignment of predicted SET domains in diatoms with large subunit methyltransferases (LSMT) and histone methyltransferases (HMT); putative catalytic Tyr residues are indicated by red arrow heads. Blue and black arrow heads indicate residues reported to be involved in AdoMet and substrate lysine binding respectively. A green arrow head indicates residues implicated in both (Trievel et al. 2002). Species abbreviations: *At, Arabidopsis thaliana*; *Sc*, *Saccharomyces cerevisiae*; *Fc*, *Fragilariopsis cylindrus*; *Pt*, *Phaeodactylum tricornutum*; *Fs*, *Fistulifera solaris*; *To*, *Thalassiosira oceanica*; *Pl*, *Pseudoleyanella lunata*; *Np*, *Nitzschia palea*; *Tp*, *Thalassiosira pseudonana*; *Ak*, *Achnanthes kuwaitensis*

Fig. 4 Identification of the proteins associated with the silica cell walls of *N. palea*. (A) tricine SDS-PAGE analyses of proteins from silica cell walls of *N. palea* and summary of identified proteins from each band; (B) full amino acid sequence of silica matrix protein1 (SMP1) and partial amino acid sequence of SMP2 from *N. palea*; the underline indicates a putative signal sequence. Basic and acidic amino acid residues are shown in blue and red letters, respectively. Repeat sequences are highlighted in yellow.

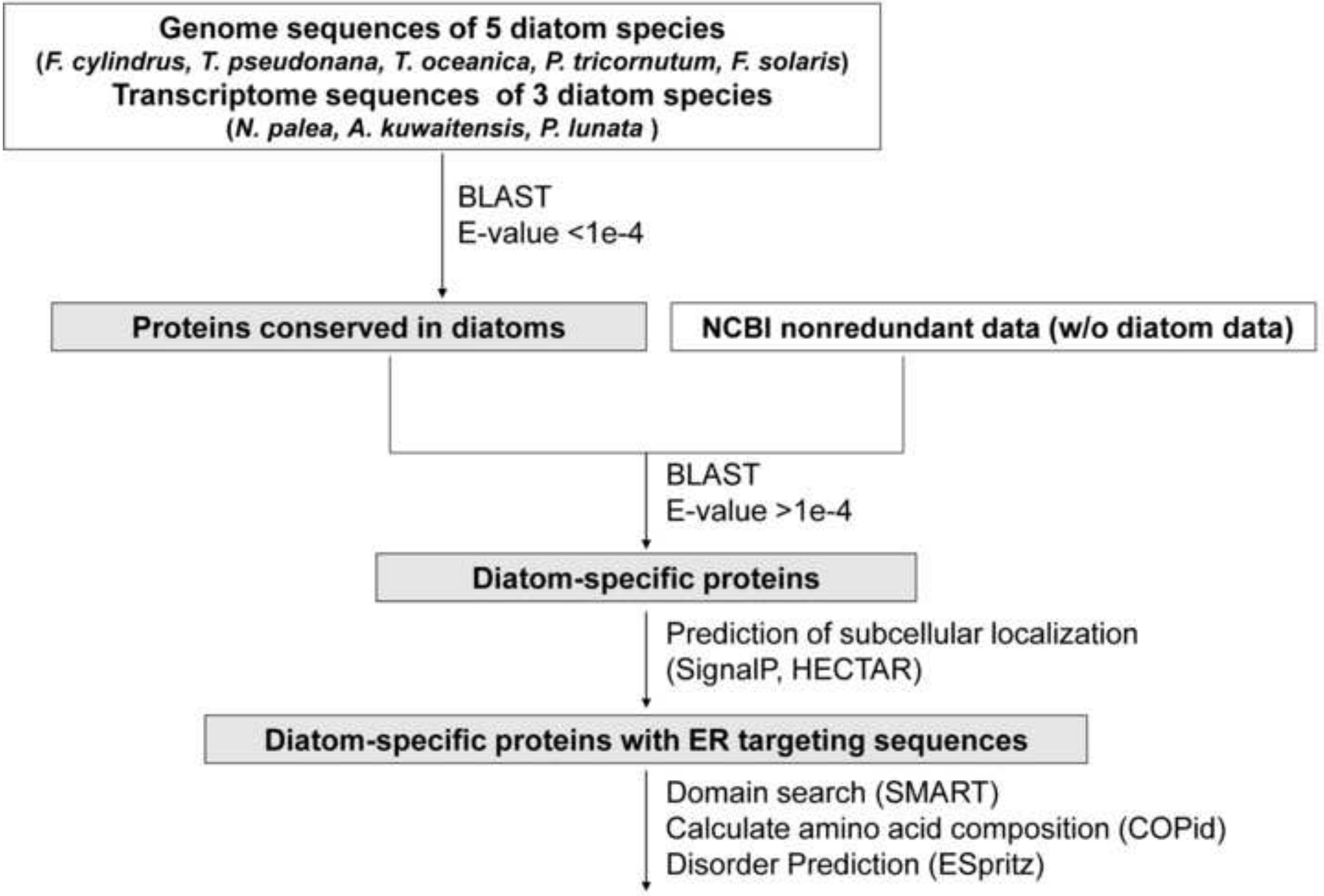
Fig. 5 Expression analysis of unigenes during cell wall formation in *N. palea*. (A) time-course of cell density and Si uptake after replenishment of silicate in Si-starved cultures; expression profiles of unigenes of the SMPs, pleuralin-1-like and silicanin-1 (B) and the SET domain proteins (C) during silicon uptake and cell growth. Data are presented as means ± standard errors of the mean.

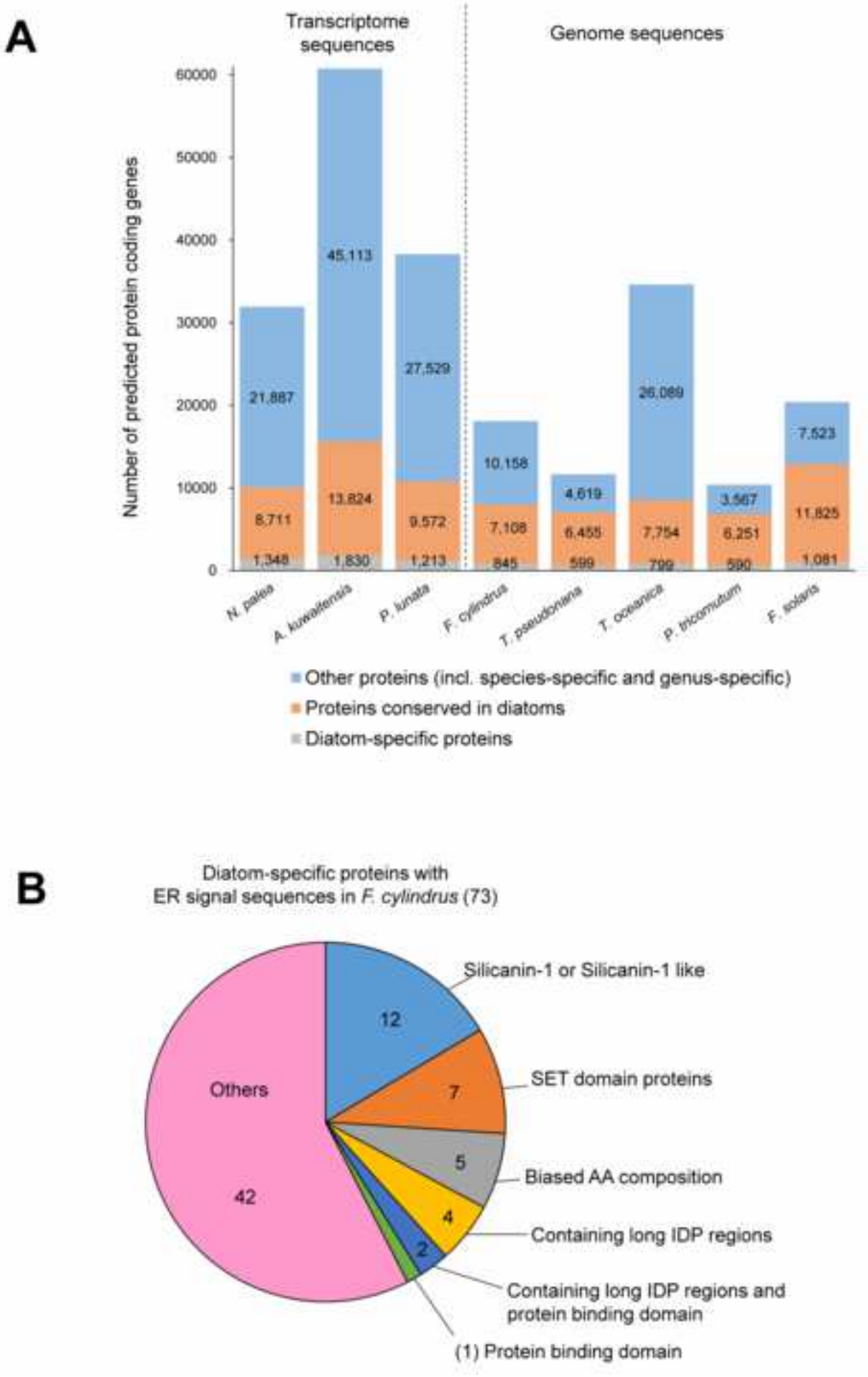
Table 1. Genes of diatoms encoding known silica cell wall-associated proteins

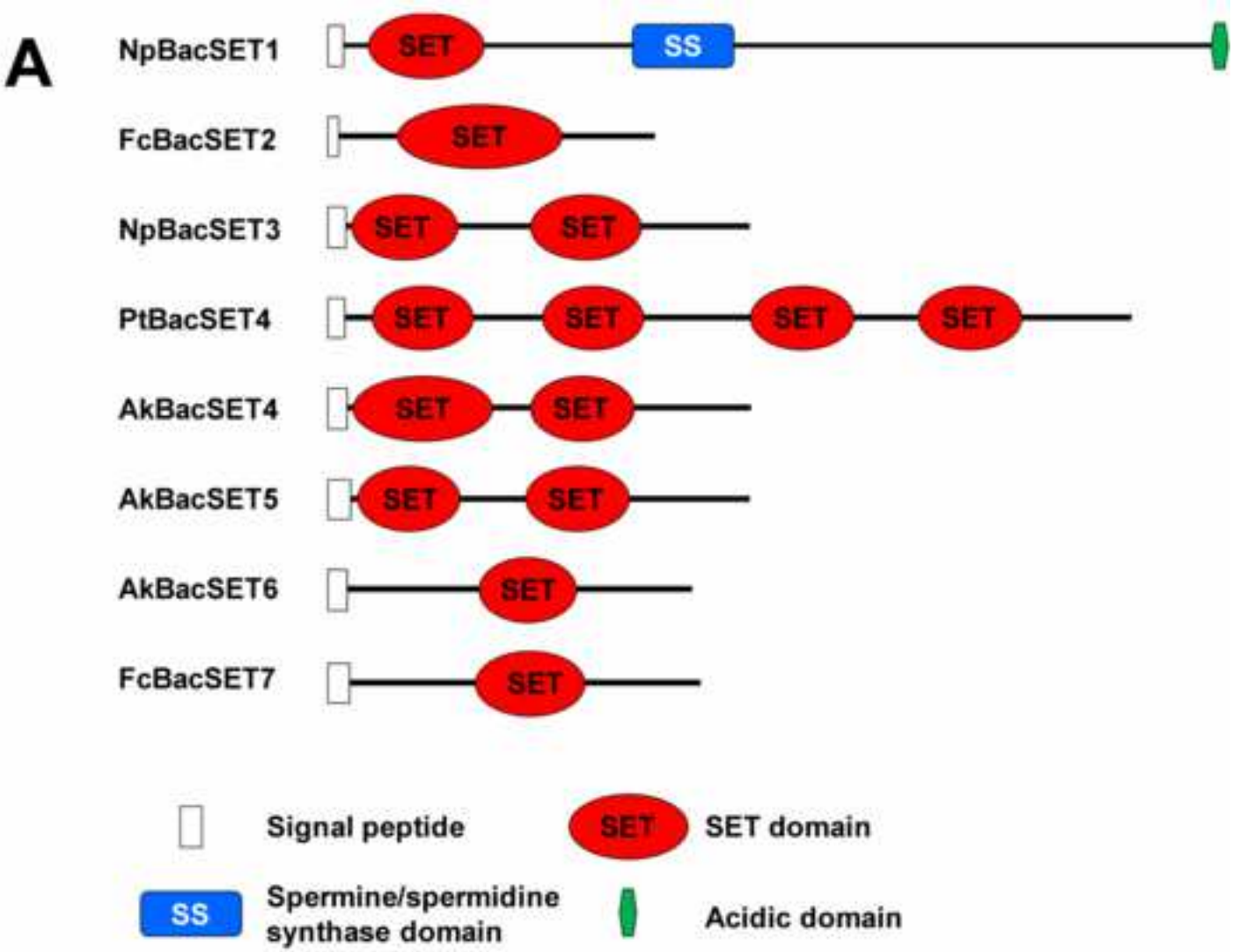
Silica cell wall-associated proteins	Reference	*C. fusiformis*	*N. palea*	*A. kuwaitensis*	*P. lunata*	*F. cylindrus*	*T. pseudonana*	*T. oceanica*	*P. tricornutum*	*F. solaris*
Sil1p	Kröger et al., 1999	○	—	—	—	—	—	—	—	—
pleuralin-1	Kröger et al., 1997	○	✓	—	—	—	—	—	—	—
frustulins	Kröger et al., 1994 Kröger et al., 1996	○	✓	✓	✓	✓	✓	✓	✓	✓
tpSil1p	Poulsen et al., 2004	NT	—	—	—	—	○	—	—	—
tpSil2p	Poulsen et al., 2004	NT	—	—	—	—	○	—	—	—
tpSil3p	Poulsen et al., 2004	NT	—	—	—	—	○	—	—	—
CinY1	Scheffel et al., 2011	NT	—	—	—	—	○	—	—	—
CinY2	Scheffel et al., 2011	NT	—	—	—	—	○	—	—	—
CinY3	Scheffel et al., 2011	NT	—	—	—	—	○	—	—	—
CinY4	Kotzsch et al., 2016	NT	—	—	—	—	○	—	—	—
CinW1	Scheffel et al., 2011	NT	—	—	—	—	○	—	—	—
CinW2	Scheffel et al., 2011	NT	—	—	—	—	○	—	—	—
CinW3	Scheffel et al., 2011	NT	—	—	—	—	○	—	—	—
silicanin-1	Kotzsch et al., 2017	NT	✓	✓	✓	✓	○	✓	✓	✓
SiMat2	Kotzsch et al., 2016	NT	—	—	—	—	○	—	—	—
SiMat3	Kotzsch et al., 2016	NT	—	—	✓	—	○	—	—	—
SiMat4	Kotzsch et al., 2016	NT	—	—	—	—	○	—	—	—

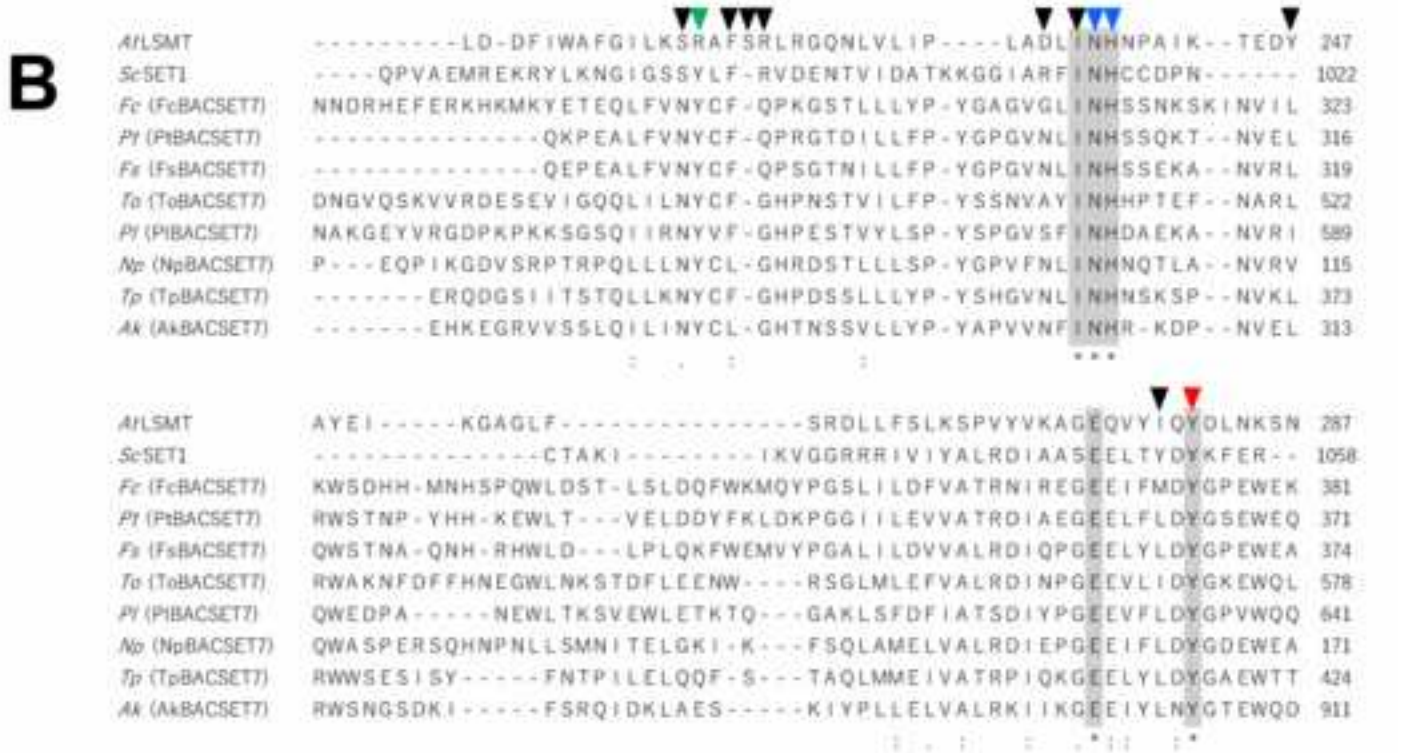
SiMat5	Kotzsch et al., 2016	NT	—	—	—	—	○	✓	—	—
SiMat6	Kotzsch et al., 2016	NT	—	—	—	—	○	—	—	—
p150	Davis et al., 2005	NT	✓	✓	✓	✓	○	✓	✓	—
silacidin	Wenzl et al., 2008	NT	—	—	—	—	○	—	—	—
TpSAP1	Tesson et al., 2017	NT	—	—	✓	—	○	✓	—	—
TpSAP3	Tesson et al., 2017	NT	—	—	—	✓	○	✓	—	—
G7408	Nemoto et al., 2014	NT	—	—	—	—	—	—	—	○

○, protein expression was confirmed; ✓, homologous sequences were found; —, no homologous sequences were found; NT, not tested

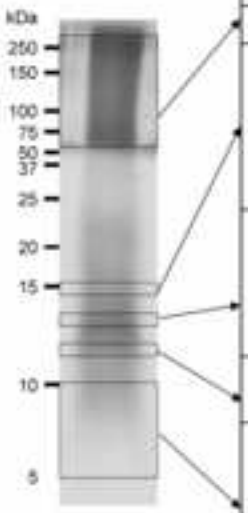








A



Band no.	Peptides	Mascot scores	Unigene ID	Annotation*
1	DYGVEAAVRPTT	52	Np12828	hypothetical protein FRACYDRAFT_268179 (NpSMP1)
2	DELGAQAPLGFF	154	Np18242	fucoxanthin chlorophyll a/c-binding protein precursor
	DALNNFITPPQREP	87	Np2506	elongation factor-like protein
	DVSFESLSALKPF	60	Np7756	chlorophyll a/b-binding protein
	DPLSLAKLGTGEGF	57	Np21947	fucoxanthin chlorophyll a/c protein
	DPLGWLNPAK	43	Np23283	putative plastid light harvesting protein isoform 32
3	DVPVGPELLGRVV	62	Np19225	mitochondrial ATPase, inner membrane
	DEAIAKAETLK	57	Np21788	ATP synthase CF1 beta subunit (chloroplast)
	DPLGLVTDGNQRNF	55	Np410	RecName: Full=Fucoxanthin-chlorophyll a-c binding protein E, chloroplastic; Flags: Precursor
4	DFVSQELRAAE	47	Np7492	photosystem II protein D2 (chloroplast)
	DREPGNFGF	42	Np18071	fucoxanthin chlorophyll a/c protein
5	DPLGLLNGADEERF DIEGTGNEFVGDFRNGYI	82	Np19782	fucoxanthin chlorophyll a/c binding protein
	DESEES	47	Np23207	No hits found (NpSMP2)
	DTLNEESYKDS	43	Np2183	14-3-3-like protein, partial

* Annotation is based on the results of BlastP analysis

B

NpSMP1 (Np12828)

MKFSLLALAAIVGSAAAGRPQLSISVRDGNFDGLDGLNPTLNWEGSAKSGDLNL

DYGVEAAVRPTTDIASLPRNIWGKASTNVAGWGVSARADVDGQDRSSAALEIDA

DNQDAD

LSLRLTASAGGGFNVRQVEATKGLDLNGARLTINPRYNVESEEADVVLGYDNGS

TNVRLTASADSQEVNVKHSLDN

TKIELTASADNQEISIDHQMDN

TNIRLTASADNQEVTISQQVDANNRVAPTINNRGDISVEWERALGDDSSLTARL

KPNDSLNVEWRDNDWTANVDMPLNGANINGANVSIKRDISF

NpSMP2 (Np23207)

PARHTRSKTRKNRMSHEIDDDRHVKHRRTSYDRLLDVSTVEGGSHMQDTSHSED

ESEESDRDILEHDNSTTPNDEGDTASLGDLLSVFGDRATTEDREEMQ