the plant journal

S E B SOCIETY FOR EXPERIMENTAL BIOLOGY

The Plant Journal (2022) **110,** 1700–1716 doi: 10.1111/tpj.15765

Shedding light on silica biomineralization by comparative analysis of the silica-associated proteomes from three diatom species

Alastair W. Skeffington[1,2,*], Marc Gentzel[3], Andre Ohara[2], Alexander Milentyev[4], Christoph Heintze[2], Lorenz Böttcher[2], Stefan Görlich[2], Andrej Shevchenko[4], Nicole Poulsen[2] and Nils Kröger[2,5,6,*]

[1]Max-Planck-Institute of Molecular Plant Physiology, 14476, Potsdam, Germany,
[2]B CUBE Center for Molecular Bioengineering, TU Dresden, 01307, Dresden, Germany,
[3]Center for Cellular and Molecular Bioengineering, TU Dresden, 01307, Dresden, Germany,
[4]Max-Planck-Institute of Molecular Cell Biology and Genetics, 01307, Dresden, Germany,
[5]Cluster of Excellence Physics of Life, TU Dresden, 01062, Dresden, Germany, and
[6]Faculty of Chemistry and Food Chemistry, TU Dresden, 01062, Dresden, Germany

Received 29 November 2021; revised 17 March 2022; accepted 3 April 2022.; published online 10 April 2022.

*For correspondence (e-mail a.skeffington@exeter.ac.uk; nils.kroeger@tu-dresden.de).

SUMMARY

Morphogenesis of the intricate patterns of diatom silica cell walls is a protein-guided process, yet to date only very few such silica biomineralization proteins have been identified. Therefore, it is currently unknown whether all diatoms share conserved proteins of a basal silica forming machinery, and whether unique proteins are responsible for the morphogenesis of species-specific silica patterns. To answer these questions, we extracted proteins from the silica of three diatom species (*Thalassiosira pseudonana, Thalassiosira oceanica*, and *Cyclotella cryptica*) by complete demineralization of the cell walls. Liquid chromatography coupled with tandem mass spectrometry (LC-MS/MS) analysis of the extracts identified 92 proteins that we name 'soluble silicome proteins' (SSPs). Surprisingly, no SSPs are common to all three species, and most SSPs showed very low similarity to one another in sequence alignments. In-depth bioinformatics analyses revealed that SSPs could be grouped into distinct classes based on short unconventional sequence motifs whose functions are yet unknown. The results from the *in vivo* localization of selected SSPs indicates that proteins, which lack sequence homology but share unconventional sequence motifs may exert similar functions in the morphogenesis of the diatom silica cell wall.

Keywords: biosilica, silica morphogenesis, intrinsically disordered proteins, silaffins, GFP-tagging, frustule, fultoportula, *Thalassiosira pseudonana*, *Thalassiosira oceanica* and *Cyclotella cryptica*.

INTRODUCTION

Silica is the hydrated, amorphous oxide of the element silicon ($SiO_2{\cdot}nH_2O$) and the second most abundant biologically formed mineral (biomineral) after calcium carbonate (Lowenstam & Weiner, 1989). It occurs in all eukaryotic supergroups (Marron et al., 2016), but the biological production of silica is dominated by a group of microalgae called diatoms (Nelson et al., 1995). Oceanic diatoms alone account for about 20% of global primary biological production (Benoiston et al., 2017), and therefore closely link the biogeochemical cycles of silicon and carbon. Diatoms use silica as their cell wall material, which is believed to have been a key factor in their ecological success, functioning as an armour against predators (Pančić et al., 2019), improving nutrient uptake (Mitchell et al., 2013), providing photoprotection (Goessling et al., 2022), enhancing light harvesting for photosynthesis (Goessling et al., 2022), and as a means to regulate buoyancy (Raven & Waite, 2004). Furthermore, biogenesis of the intricately nano- and micropatterned silica cell walls of diatoms has attracted great interest from materials scientist as paradigms for the bottom-up production of 3D, hierarchically porous materials (Nassif & Livage, 2011).

Silica formation in diatoms occurs in specialized, lipid bilayer-bound, intracellular compartments called silica deposition vesicles (SDVs). There are two different types of SDVs. One produces plate- or dome-shaped silica structures called valves, and the other produces biosilica rings, called girdle bands. During cell division, each daughter cell produces one valve, while girdle bands are produced during interphase (Hildebrand et al., 2007). When morphogenesis of a valve or girdle band inside a SDVs has been completed, the silica structure is exocytosed and assembled into a cell wall that completely encases the cell.

It has been proposed that the lumen of each SDV possesses a silica-forming matrix composed of organic macromolecules (Hecky et al., 1973; Kröger & Sumper, 2004; Volcani, 1981). In this model, the species-specific silica nano- and micropatterns would result from differences in the composition of the organic matrix components. Currently, the biomolecular composition of SDVs is unknown, because a method for their isolation has not yet been established. However, it is assumed that some of the organic molecules associated with the mature cell wall originate from the silicification machinery of the SDV (Hecky et al., 1973). Complete demineralization of diatom cell walls using an acidic ammonium fluoride solution, solubilized phosphoproteins and long-chain polyamines (Kröger & Sumper, 2004), while organic matrices composed of proteins and polysaccharides remained insoluble (Brunner et al., 2009; Buhmann et al., 2014; Kotzsch et al., 2017; Pawolski et al., 2018; Scheffel et al., 2011; Tesson & Hildebrand, 2013). *In vitro* silica formation experiments using mixtures of the phosphoproteins and long-chain polyamines, both in the presence and absence of insoluble organic matrices yielded porous silica patterns that, in some cases, mimicked native silica morphologies, suggesting that the biosilica associated organic components might play a role in silica morphogenesis *in vivo* (Pawolski et al., 2018; Poulsen et al., 2003; Poulsen & Kröger, 2004; Scheffel et al., 2011).

To date, diatom biosilica associated proteins have been biochemically characterized from only two diatom species, *Cylindrotheca fusiformis* (Kröger et al., 2001, 2002; Poulsen et al., 2003) and *Thalassiosira pseudonana* (Kotzsch et al., 2016; Poulsen & Kröger, 2004; Scheffel et al., 2011). The most prominent groups of proteins, called silaffins, exhibit no significant sequence similarity to each other or to any other proteins in the NCBI database, but they share a very high degree of phosphorylation of serine and threonine residues and polyamine modifications of lysine residues (Poulsen & Kröger, 2004; Scheffel et al., 2011). A large number of genes have been identified that are responsive to silicic acid content in the culture medium (Mock et al., 2008; Sapriel et al., 2009) or upregulated during valve formation (Brembu et al., 2017; Shrestha et al., 2012). However, for almost all of these genes it is unknown whether the encoded proteins are associated with the biosilica and/or involved in silica morphogenesis.

In the present study, we aimed to expand our knowledge of the silica-associated proteome of diatoms substantially, and to identify proteins that might be responsible for differences in silica morphology between species. To achieve this, we chose three closely related species of centric diatoms: *T. pseudonana*, *Thalassiosira oceanica*, and *Cyclotella cryptica*. It should be noted that the genus *Thalassiosira* is paraphyletic (Kaczmarska et al., 2005), and *T. pseudonana* is more closely related to *C. cryptica* than it is to *T. oceanica* (Alverson et al., 2011). These three species have a conserved architecture of cylindrically shaped cell walls, with a complex silica network structure in the valves. The network consists of radially oriented ribs forming a branched pattern from the centre of the valve to its rim (Figure 1a,g,b,h yellow and green lines), which are more pronounced in *C. cryptica* than in *T. pseudonana*. *Thalassiosira oceanica* differs in that the ribs are obscured by a covering layer of porous silica (Figure 1d,e). Neighbouring ribs are connected by seemingly irregularly spaced short silica bridges in *C. cryptica* and *T. pseudonana*. The gaps in the meshwork of interconnected ribs and silica bridges contain silica that is perforated with 20–30 nm sized circular pores (known as cribrum pores; Figure 1b,h). Multiple tube-like structures, termed fultoportulae, are regularly spaced near the rim of the valves in all three species (Figure 1a,d,g,b,e,h, red circles). About half of the valves also contain one or two fultoportulae near the centre (Figure 1a,d, purple circle). Fultoportulae are involved in the secretion of chitin fibres (Herth, 1979) and have a restricted phylogenetic distribution within the centric diatoms (Kaczmarska et al., 2005). The girdle band regions of the cell walls are less complex than the valves, exhibiting a pattern of alternating porous and non-porous regions (Figure 1c,f,i).

In this work, we identify soluble silica-associated proteins from these three species, thereby shedding light on the machinery that might be responsible for the commonalities and differences in silica structure, as well as identifying more conserved proteins that might have a more basal role in silica biogenesis.

RESULTS

Extraction of silica-associated proteins and protein identification

Biosilica was isolated from *T. pseudonana*, *T. oceanica*, and *C. cryptica* cultures using a previously established method for diatoms (Kröger et al., 2000), followed by complete demineralization with ammonium fluoride at pH 4.5 (Kotzsch et al., 2016). The soluble extracts were separated by sodium dodecyl sulphate–polyacrylamide gel electrophoresis (SDS-PAGE) and stained with ‘Stains-All’, which revealed multiple bands in the extract of *T. pseudonana* and *C. cryptica* with apparent molecular masses ranging from approximately 15 kDa to >170 kDa (Figure 2). In contrast, the *T. oceanica* extract was dominated by a broad band ranging from approximately 55 kDa to approximately 130 kDa. Two replicate ammonium fluoride extracts for each of the three species were subjected to proteolytic

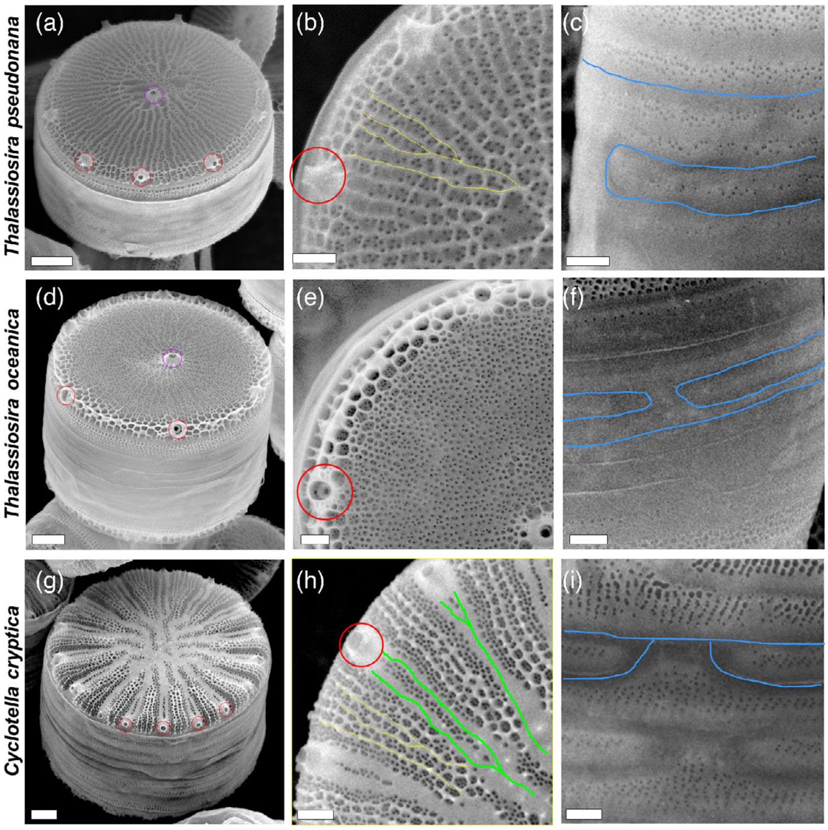


Figure 1. Diatom silica structure. Scanning electron microscopy images of isolated silica cell walls from (a–c) *Thalassiosira pseudonana*, (d–f) *Thalassiosira oceanica*, and (g–i) *Cyclotella cryptica*. For each species, the cell wall is presented in (a,d,g) oblique view, and details from the (b,e,h) valve region and (c,f,i) girdle band region are shown. Yellow and green lines indicate radial ribs and wide radial ribs, respectively. Blue lines designate non-porous regions in girdle bands. Examples of fultoportulae on the rim are labelled with red circles, and central fultoportulae with purple circles. Scale bars: (a,d,g) 1 μm and (b,c,e,f,h,i) 400 nm.

digestion and subsequently analysed by high pressure LC-MS/MS (Methods S1). After protein identification and inference (see Methods S1), we discarded seven proteins as potential contaminants. Four of these were proteins with a predicted chloroplast transit peptide ('high confidence' ASAfind predictions; Gruber et al., 2015), two were identified as histones, and one as a conserved small nucleolar RNA-binding protein. The remaining 92 proteins were taken forward for downstream analyses (Table S1): 29 from *C. cryptica*, 55 from *T. pseudonana*, and eight from *T. oceanica*. These proteins were collectively termed 'soluble silicome proteins' (SSPs).

We considered the possibility that the lower complexity of the *T. oceanica* samples could be an artefact due to a poor *T. oceanica* database, resulting in proteins not being identified that might be present not only in *T. oceanica* but also in at least one of the other two species. To examine this possibility, we performed searches of the *T. oceanica* derived MS/MS spectra against the *T. pseudonana* and *C. cryptica* databases. These searches did not result in any hits, supporting the assumption that the lower complexity of SSPs might be a genuine feature of the *T. oceanica* samples.

Previously characterized proteins and known domains

Seven of the 92 SSPs, all from *T. pseudonana*, had previously been identified at the protein level: Tp_p150 (Davis et al., 2005), TpSil1 and TpSil3 (Poulsen & Kröger, 2004), TpSil4 (Sumper & Brunner, 2008), and SiMat3, SiMat4, and SiMat7 (Silacanin-1, Sin1) (Kotzsch et al., 2016). The 85 novel proteins were distinguished by an acronym for the species of origin and a numeric identifier (e.g. TpSSP1). A role of the SSPs in silica formation was supported by the fact that 35 of the *T. pseudonana* SSPs had been previously identified in transcriptomic experiments relating to silica metabolism (Table S2) (Brembu et al., 2017; Mock et al., 2008; Shrestha et al., 2012), and 13 were identified in more than one of the previous studies.

Twenty-nine of the SSPs contained known protein domains, including three with chitin-binding domains and eight with transmembrane domains. It has been previously shown that silaffins are extensively post-translationally processed (Kröger et al., 1999; Poulsen et al., 2003; Poulsen & Kröger, 2004), and several proteins were identified with domains that indicate they could be involved in these processes (Appendix S1).

Sequence characteristics of SSPs

Global sequence analysis of the SSPs revealed that they are significantly lower in sequence complexity (Figure 3a) than other proteins from the respective proteomes of the three species (hereafter termed the 'background proteomes': predicted proteome minus the SSPs). They are

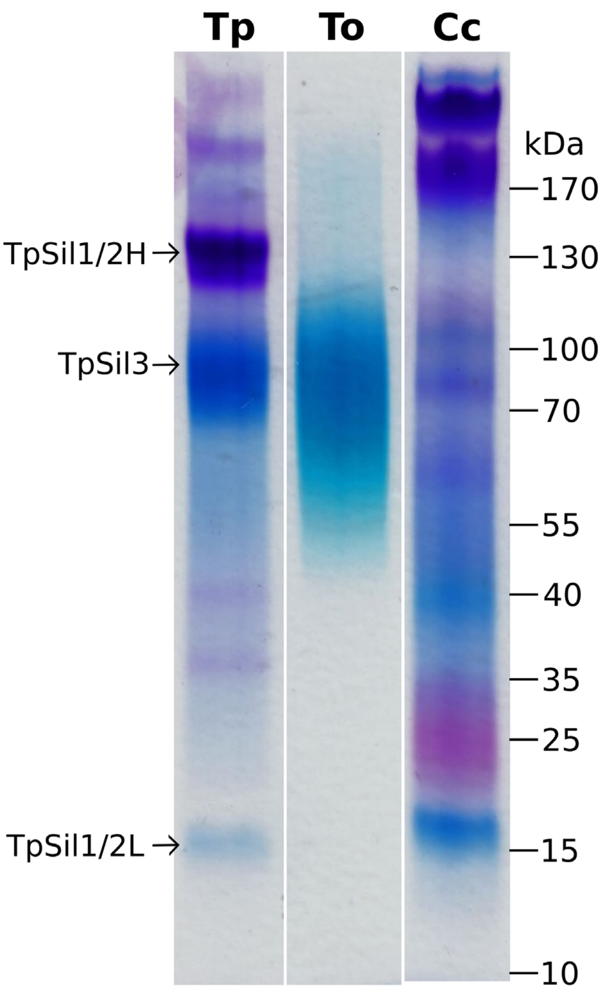


Figure 2. Sodium dodecyl sulphate–polyacrylamide gel electrophoresis analysis of the ammonium fluoride soluble extracts. Samples were run on the same gel and were stained with the dye 'Stains-All'. Cc, *Cyclotella cryptica*; To, *Thalassiosira oceanica*; Tp, *Thalassiosira pseudonana*. Arrows indicate previously characterized *T. pseudonana* silaffins (Poulsen & Kröger, 2004). Different aliquots of the extracts were loaded as described in Experimental procedures section. Note that the colour and intensity of bands in 'Stains-All'-stained sodium dodecyl sulphate–polyacrylamide gel electrophoresis are strongly influenced by the composition of the glycan moieties in glycoproteins. Differences in band patterns may thus partly reflect differences in the protein glycosylation machineries rather than the polypeptide backbones.

also predicted to be disordered over a significantly greater proportion of their length than the background proteomes (Figure 3b), and display substantial biases in amino acid composition (Figure S1). The probability of achieving bias at least this extreme through random sampling of the proteome is <0.001 for *T. pseudonana* and *C. cryptica*, and <0.01 for *T. oceanica*.

Sequence similarity among SSPs is generally low, and no protein has putative homologues in the SSP set of all three species. In the whole dataset, there are only 10 pairs of proteins with an identity >35% in global sequence alignments (Table S3). Several groups of these proteins included both *T. pseudonana* and *C. cryptica* proteins, suggesting that there is some conservation in the silica-associated proteomes of these species. For example, *C. cryptica* homologues of TpSil1 and Tp_p150 were identified as CcSSP6 and CcSSP10 respectively. TpSSP2 and CcSSP2 are also noteworthy as they are 44% identical and were both identified with high (normalized) spectral counts.

There are no proteins with good similarity (BLAST e-value <1e-30 and query coverage >50%) to the *T. oceanica* SSPs represented in either the proteomes or the genomes of the other two species. However, 30 (55%) of the *T. pseudonana* SSPs and 10 (34%) of the *C. cryptica* SSPs show similarity to proteins in the predicted proteome of *T. oceanica*, yet none of these were found among the *T. oceanica* SSPs. Where *T. pseudonana* lacked a similar SSP to a *C. cryptica* SSP (or vice versa, based on the >35% identity in global sequence alignments, 66 proteins in total) the SSPs were not present in the proteome of the other species for two different reasons: In 52% of these cases (34 proteins) there is no similar protein in the other predicted proteome (BLAST e-value <1e-30 and qcov >50%), indicating the protein is either unique to one of the species or that the predicted proteome is incomplete in the other species. In the other 48% of cases (32 proteins), a similar protein is represented in the other predicted proteome, so the protein could be genuinely absent in the silica extract or simply not detected by proteomics due to low abundance or differing modifications.

Despite the low level of sequence similarity among the SSPs, at the level of whole protein sequence alignments, manual inspection of the sequences revealed that short sequence motifs were often shared between subsets of the proteins. To understand the nature of such smaller scale sequence similarities, we searched for sequence motifs that were overrepresented in the SSPs relative to the background proteomes. Enriched motifs were identified using the motif-x software (Chou & Schwartz, 2011) via the ProminTools package (Skeffington & Donath, 2020). In total, 113 motifs were found to be enriched in the SSPs relative to the background proteome (motif-x *P*-value cut-off = 10^{-6}), reflecting the highly biased composition of the sequences (Table S4). None of the enriched motifs are common to all SSPs.

Previously, K..K motifs (period = any amino acid) have been identified in many silica-associated proteins, and experimentally shown to be involved in targeting these proteins to the cell wall (Poulsen et al., 2013), and to promote silica formation *in vitro* (Wieneke et al., 2011). The motif K..K is common in all three background proteomes and is only three-fold enriched in the SSPs. However, specific variants of K..K with restricted identities for both the central and the flanking amino acids were among the

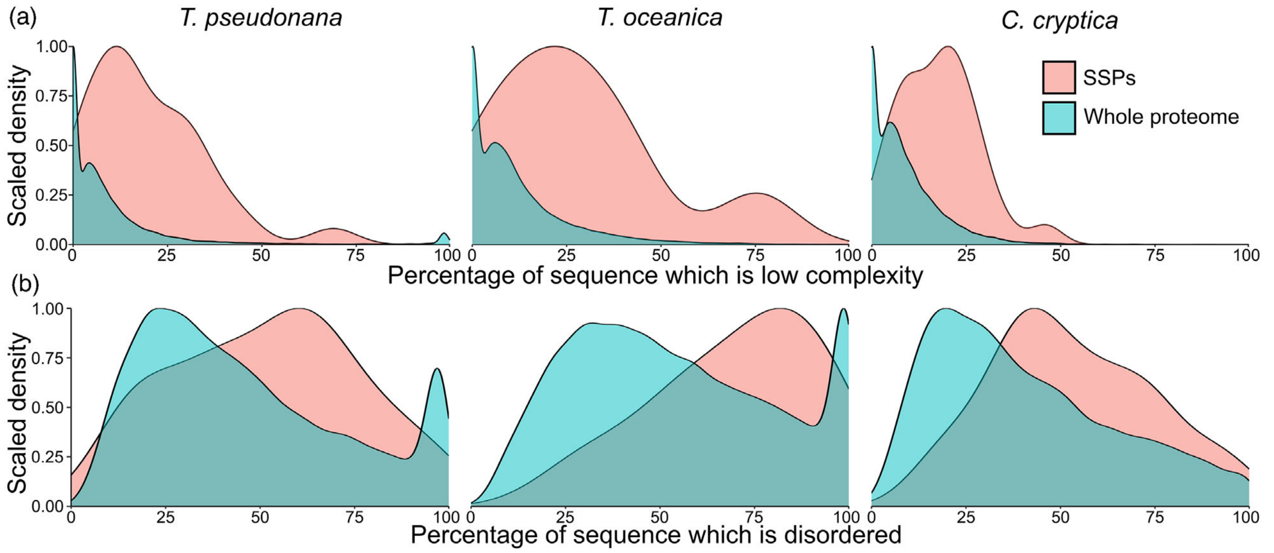


Figure 3. Global properties of the soluble silicome proteins (SSPs) from *Cyclotella cryptica*, *Thalassiosira oceanica*, and *Thalassiosira pseudonana*. Density plots of the percentage of (a) low complexity sequence in individual proteins and (b) predicted disordered sequence in individual proteins. In each case, the SSPs and background proteome distributions are significantly different (Wilcoxon rank sum test with continuity correction, $P < 10^{-6}$).

most prominent motifs in the data set. In particular, PKA.K, KS.KS, GKS.K, SKS.K, SKA.K, AKS.K, KS.KA, and GKA.K were all more than 20-fold enriched. This included DAKA.K, which is 124-fold enriched and found in eight SSPs, KSSKA is 77-fold enriched and present in 11 SSPs, and KAEK is 33-fold enriched and found in nine SSPs.

Given that most SSPs could not be grouped based on full-length sequence comparisons, we investigated whether the proteins could be grouped based on their motif content. Distance correlation coefficients (Székely & Rizzo, 2014) were calculated for all pairs of the 92 SSPs based on their enrichment in the identified motifs (see Experimental procedures, Table S5). The 52 protein sequences with strong motif-based similarity to at least one other protein (dcor >0.65) were clustered based on their enrichment in the 70 most enriched motifs. The remaining motifs tended to be enriched to a low level in a high proportion of the SSPs, and thus did not aid in grouping the sequences. This resulted in eight protein clusters (Figure 4), with distinct patterns of motif enrichment. Cluster 1 (10 proteins) and Cluster 4 (10 proteins) are enriched in different variants of K..K motifs (particularly KSGK and FKS.K in cluster 1, and KSSKA in Cluster 4). Cluster 2 (three proteins) is rich in SMSM, as are a number of proteins in Cluster 5, which are additionally highly enriched in DAKA.K. Cluster 3 (four proteins) is enriched in S..SSKS, while Cluster 6 (six proteins including TpSil1 and TpSiMat4) is enriched in various P, T, and S containing motifs. Finally, Cluster 7 is enriched in a range of C and G containing motifs, and Cluster 8 in various N and Q containing motifs.

Pairwise BLASTp comparisons of the proteins within the clusters (Figure S2) showed that the motif-based clustering revealed similarities between the proteins not captured by BLAST comparisons. Even in highly unselective BLAST comparisons (e-value cut-off of 0.01) in clusters 1–6, only eight of the total 120 pairwise comparisons generated a match.

Phylogenetic distribution of SSPs

Two approaches were used to assess the phylogenetic distribution of the full-length SSP sequences. The first makes use of the precalculated orthogroups available at PLAZA diatoms (Osuna-Cruz et al., 2020), which include sequences from 10 diatom species with sequenced genomes as well as 16 other eukaryotes. Proteins can then be classified as being species-specific or restricted to the centric diatoms, diatoms, Chromista, or as having a broad distribution across the eukaryotes. The second method was to run tBLASTn searches of the SSP sequences against transcriptomes from the Marine Microbial Eukaryote Sequencing Project (Keeling et al., 2014), including 136 diatom transcriptomes and 390 from other taxa. A version of the MMETSP data was used that had been cleaned of contaminants (see Experimental procedures). Of these two methods, orthogroups provide a more rigorous assessment of shared ancestry. Here they are built on genome sequence information, so proteins will not be missing due to low or absent expression, as may be the case in the MMETSP transcriptomes. However, the orthology analysis does not capture the diversity of the diatoms or the eukaryotes, which is much better represented by

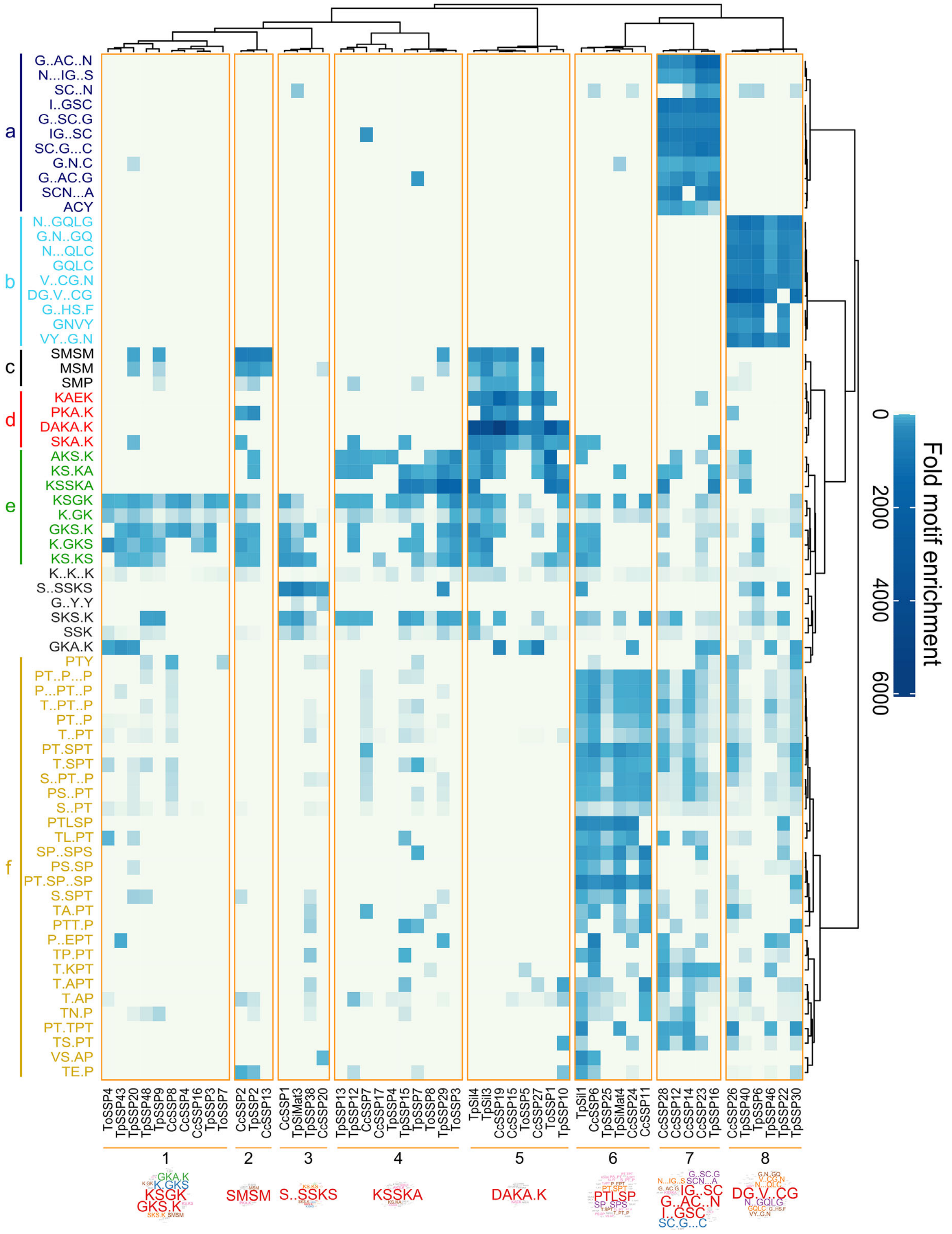


Figure 4. Heatmap displaying motif enrichment with respect to the relevant background proteomes. To be included in this analysis, each protein had to be highly correlated in its motif profile to at least one other protein (see Experimental procedures section). Eight protein clusters were identified and are indicated at the bottom of the heatmap. For each cluster, the motif profile is represented as a word cloud. The height of the letters is proportional to the mean enrichment of the motif in that protein cluster, but the scale differs for each word cloud. (a–e) On the left margin, groups of motifs of interest (selected by co-occurrence in similar proteins and/or by similarity in amino acid composition) are coloured and labelled. These correspond to the motif groups represented in Figure 6.

Figure 5. Phylogenetic distribution of soluble silicome proteins. The central panel summarizes the BLAST analysis of the soluble silicome proteins against the MMETSP transcriptomes (see Experimental procedures section). Only proteins with BLAST hits at an e-value $<10^{-30}$ are included in this analysis. Each column displays data from a different taxonomic group. The number of hits per species is visualized as the colour of circles, while the e-value of the best hit is indicated via the size of the circles. Proteins are classified as specific to the genus *Thalassiosira*, class Mediophyceae, phylum Bacillariophyta (diatoms), kingdom Chromista or as having a cross-kingdom classification. The search results for two highly conserved proteins, α-tubulin and Histone-2A from *Thalassiosira pseudonana*, are displayed for comparison. The phylogenetic distribution of the PLAZA diatom orthogroup to which each protein belongs is noted. TpSil1 was not represented in the PLAZA database (B8LDT2 is mis-annotated as TpSil1 in the database). Chr, Chromista specific; Dia, diatom specific; Euk, eukaryote-wide distribution; Med, Mediophyceae specific (note PLAZA only contains Mediophyceae from the Thalassiosira clade); SpSp, species specific.

the MMETSP transcriptomes. Overall, it is clear that neither method provides a definitive assessment of phylogenetic distribution, but rather provides complementary information to each other. The grouping of proteins by phylogenetic distribution in Figure 5 is based on the MMETSP BLAST searches (e-value cut-off of 10^{-30} used for the group assignment), and the PLAZA classifications are indicated.

Of the 21 proteins found to be species-specific (and therefore not displayed in Figure 5), nine were from *T. oceanica*, six from *C. cryptica* and six from *T. oceanica*. The largest phylogenetic grouping consisted of the 38 proteins (23 from *T. pseudonana*, 14 from *C. cryptica*, one from *T. oceanica*) restricted to the Mediophyceae (Thalassiosirales and bipolar centric diatoms). Among these, nine proteins were particularly well conserved, being found in 54–67% of Mediophyceae species, which makes them good candidates for the Mediophyceae-specific silica biomineralization machinery. These proteins are TpSiMat3, TpSSP46, Tp_p150 and its homologue CcSSP10, TpSSP3 and its homologues CcSSP4 and TpSSP5, TpSSP6 and its homologue TpSSP22.

The Bacillariophyta-specific SSPs comprise 13 proteins (eight from *T. pseudonana* and five from *C. cryptica*), which are candidates for being components of a basal, diatom-specific biomineralization machinery. Of these, five proteins were identified in >80% of diatom species. These were TpSSP41, TpSSP28, and CcSSP20 along with its homologues TpSSP31 and TpSSP36. TpSin1 was classified as having a Chromista-wide distribution by the MMETSP analysis, due to blast hits (best hit e-value = 1.8E-36, 25% query coverage, 65% identity) in the dinoflagellates *Glenodinium foliaceum* and *Kryptoperidinium foliaceum*. These species have tertiary plastids of diatom origin and a residual diatom-derived nucleus (Žerdoner Čalasan et al., 2018), and it is thus conceivable that a diatom TpSin1 is retained in the organism's DNA. Because of this, TpSin1 can probably be considered diatom-specific, and it is the only diatom-specific protein found in the transcriptomes of >90% of the MMETSP diatom species analysed.

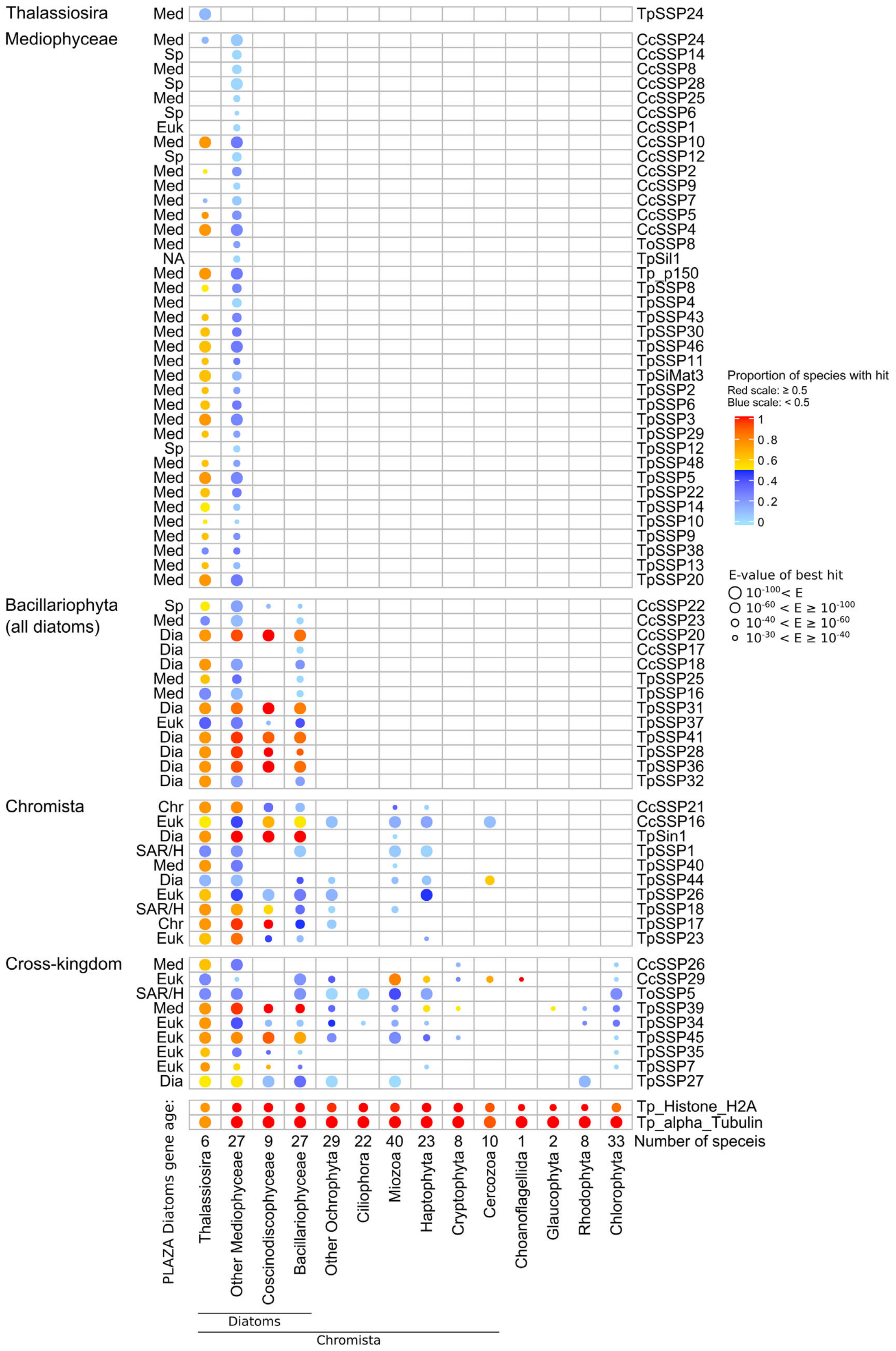
Nineteen proteins were classified as having a broader phylogenetic distribution with putative homologues outside the diatoms (14 from *T. pseudonana*, four from *C. cryptica*, one from *T. oceanica*), often with BLAST hits in Ochrophytes, Miozoa, and Haptophyta. They were classified as either having a Chromista, or Eukaryote-wide distribution. Twelve of these proteins had characterized domains (e.g. sulfatase, phytase, protease), which were then key drivers of BLAST similarity. It is conceivable that these domains may have unique functions when set in the context of a diatom-specific sequence. For example, silica formation in sponges is catalysed by a protein that contains a cysteine protease-like domain (Cha et al., 1999).

Silica-targeting sequences

Although there were no enriched motifs shared by all SSPs, we were interested to see if they shared sequences that might be responsible for 'silica targeting', i.e. directing the proteins into the biosilica cell wall. A TpSil3-derived 13mer peptide containing five lysine residues, was previously shown to be sufficient (along with a N-terminal signal peptide) for silica targeting of green fluorescent protein (GFP) in *T. pseudonana* with high efficiency (Poulsen et al., 2013). Similar so-called penta-lysine clusters matching the regular expression K..K..K.{1,2}K..K are also present in the other previously known silaffins, TpSil1, TpSil2, and TpSil4 (Poulsen et al., 2013). Among the SSPs, 27 (29%) contained at least one penta-lysine cluster, while there were matches in only 1.5% of sequences in the background proteomes, excluding SSPs. This raises the question how the other SSPs become targeted to the biosilica cell wall. Poulsen et al. (2013) showed that sequence regions containing at least two K..K motifs with a maximal spacing of 10 residues could mediate silica targeting to some extent (Poulsen et al., 2013). Such regions are present in an additional 40 SSPs, yet 25 SSPs lacked any of the silica targeting sequences, including 10 that contain no K..K motif at all. SSPs devoid of any known silica targeting sequences might rely either on a yet unknown silica targeting mechanism, might reach the silica cell wall by forming clusters with one of the SSPs that bear a silica targeting sequence, or some could potentially be contaminants not targeted to the silica cell wall.

Overview of SSP characteristics

Key outputs of the bioinformatic analyses of the SSPs are summarized in Figure 6, which provides an integrated overview how the results of the above analyses relate to one another. Species-specific proteins tend to be (i)


Thalassiosira
Med TpSSP24
Mediophyceae
Med CcSSP24
Sp CcSSP14
Med CcSSP8
Sp CcSSP28
Med CcSSP25
Sp CcSSP6
Euk CcSSP1
Med CcSSP10
Sp CcSSP12
Med CcSSP2
Med CcSSP9
Med CcSSP7
Med CcSSP5
Med CcSSP4
Med ToSSP8
NA TpSil1
Med Tp_p150
Med TpSSP8
Med TpSSP4
Med TpSSP43
Med TpSSP30
Med TpSSP46
Med TpSSP11
Med TpSiMat3
Med TpSSP2
Med TpSSP6
Med TpSSP3
Med TpSSP29
Sp TpSSP12
Med TpSSP48
Med TpSSP5
Med TpSSP22
Med TpSSP14
Med TpSSP10
Med TpSSP9
Med TpSSP38
Med TpSSP13
Med TpSSP20
Bacillariophyta (all diatoms)
Sp CcSSP22
Med CcSSP23
Dia CcSSP20
Dia CcSSP17
Dia CcSSP18
Med TpSSP25
Med TpSSP16
Dia TpSSP31
Euk TpSSP37
Dia TpSSP41
Dia TpSSP28
Dia TpSSP36
Dia TpSSP32
Chromista
Chr CcSSP21
Euk CcSSP16
Dia TpSin1
SAR/H TpSSP1
Med TpSSP40
Dia TpSSP44
Euk TpSSP26
SAR/H TpSSP18
Chr TpSSP17
Euk TpSSP23
Cross-kingdom
Med CcSSP26
Euk CcSSP29
SAR/H ToSSP5
Med TpSSP39
Euk TpSSP34
Euk TpSSP45
Euk TpSSP35
Euk TpSSP7
Dia TpSSP27
PLAZA Diatoms gene age:
Tp_Histone_H2A
Tp_alpha_Tubulin
Number of speceis
6 Thalassiosira
27 Other Mediophyceae
9 Coscinodiscophyceae
27 Bacillariophyceae
29 Other Ochrophyta
22 Ciliophora
40 Miozoa
23 Haptophyta
8 Cryptophyta
10 Cercozoa
1 Choanoflagellida
2 Glaucophyta
8 Rhodophyta
33 Chlorophyta
Diatoms
Chromista
Proportion of species with hit
Red scale: ≥ 0.5
Blue scale: < 0.5
1
0.8
0.6
0.4
0.2
0
E-value of best hit
$10^{-100} < E$
$10^{-60} < E \geq 10^{-100}$
$10^{-40} < E \geq 10^{-60}$
$10^{-30} < E \geq 10^{-40}$

disordered over a greater portion of their length compared with other SSPs (mean of 75% versus 45%, Welch two sample *t*-test, $t = 7$, df = 50, $P = 4.6E\text{-}9$), and (ii) of low sequence complexity over a greater portion of their length (mean of 29% versus 15%, $t = 3.8$, df = 34, $P = 0.0006$). Compared with proteins in broader phylogenetic categories, proteins classed as species-specific or Mediophyceae age fall at a higher proportion into one of the eight motif clusters (70% versus 31%; Figure S3). It is notable that there are few long regions of sequence similarity between the SSPs of the Mediophycea or Bacillariophyta age apart from known domains such as RCC1 repeats. Instead, such SSPs seem to share particular combinations of very short, unconventional motifs within intrinsically disordered protein domains of low amino acid complexity. Currently, there is insufficient understanding of the structure–function correlation in such proteins and their evolution, making it impossible to identify orthologues based on sequence analysis. To attempt the identification of putative orthologues, we investigated the location patterns of selected SSPs in the biosilica assuming that orthologous proteins should exhibit highly similar location patterns.

Localization of selected SSPs

A selection of SSPs were expressed as GFP fusion proteins in their species of origin and under their native promoters, to validate their silica association *in vivo*, and to examine their precise positioning with the biosilica. We focused on species-specific, and Mediophyceae-specific proteins, because the former might account for species-specific differences in the biosilica structures (e.g. rib patterns), while the latter might have a role in the biogenesis of structural features common to the three species (e.g. fultoportulae, cribrum pores, and girdle bands) (see Figure 1).

In each case, tight association with the silica was tested by extracting the cells with a hot solution of EDTA and SDS, which removes intracellular material and cell wall proteins that are only loosely associated with the silica (Poulsen et al., 2013). All except one of the proteins that were GFP-tagged in the present study, showed identical location patterns in live cells and the purified silica (see below). The GFP-tagging experiments could only be performed with *T. pseudonana* and *C. cryptica*, because transformation experiments using the protocols for both species (Kumari et al., 2020; Poulsen et al., 2006) did not succeed in obtaining transgenic clones when applied to *T. oceanica*. Schemes showing the position of the GFP tag for each protein are shown in Figure S4.

Given the potential role of silaffins in the morphogenesis of species-specific silica structures (Poulsen et al., 2003; Poulsen & Kröger, 2004), we started by examining the locations of some of the *T. pseudonana* silaffins and compared these with the locations of the most similar SSPs that we had identified in *C. cryptica* silica. The closest homologue to silaffin TpSil1 in the *C. cryptica* SSPs is CcSSP6 (36% sequence identity in global alignment, motif-based correlation 0.60). Both proteins are Mediophyceae-specific, belong to motif Cluster 6: rich in both ‘type f’ motifs (e.g. PTLSP) and ‘type e’ motifs (e.g. KSGK) (see Figure 4). They also both have a helical region at the extreme C-terminus of the sequence, while TpSil1 has a chitin-binding domain which is absent in CcSSP6. TpSil1-GFP and CcSSP6-GFP were both located exclusively in the valve part of the cell wall, and tightly associated with the silica (Figure 7a,b). They were located throughout the valve, but the strongest GFP signal was detected towards the rim. In valve view, CcSSP6-GFP exhibits a sector-type pattern that is consistent with the location between the wide ribs of the valve (compare Figure 7b with Figure 1h). However, this requires further experimental validation.

TpSil3 is structurally the best characterized silaffin of the three diatom species (Poulsen & Kröger, 2004; Sumper et al., 2007). It is located throughout the valves and in the two girdle bands closest to each valve, but absent from the girdle band in the mid cell region (Figure 7c). These observations are consistent with a previous report (Scheffel et al., 2011). The *C. cryptica* SSP most similar to TpSil3 is CcSSP19. Although they share only 29% sequence identity in global alignments, they have a very high motif-based sequence correlation of 0.89. They both belong to Cluster 5, and are both rich in acidic ‘type d’ motifs such as DAKA.K. Thus, they can be considered putative orthologues, despite being classified as species-specific in the phylogenetic analysis (see Figure 5). Like TpSil3-GFP, CcSSP19-GFP is observed in both the valves and the girdle bands (Figure 7d). However, in contrast to TpSil3-GFP, the CcSSP19-GFP appears to be concentrated in radial structures of the valve, which are reminiscent of the pattern and dimensions of the wide ribs (see Figure 1h). It also exhibits a striped location pattern in the girdle bands, with curved regions, which might correspond to non-porous valve regions (Figure 1i). These data support the idea that both proteins may have similar functions, but with species-specific modulations. Interestingly, it appears that the localizations of CcSSP6 and CcSSP19 may be complementary to one another: the former present predominantly between the wide ribs, and the latter in the wide ribs (compare Figure 7b and Figure 7d bottom panels). The more homogeneous distribution of TpSil3-GFP is consistent with the absence of sector-type silica architecture in the *T. pseudonana* valve.

Motif Cluster 1 (rich in ‘type e’ motifs such as KSGK) is the most common cluster among Mediophyceae-specific proteins (Figure S3), and did not contain any previously characterized proteins. We chose to examine TpSSP3 and CcSSP4 as representatives of this cluster, which shared 44% identity in a full-length sequence alignment and were identified with high spectral counts (Table S1). Closer

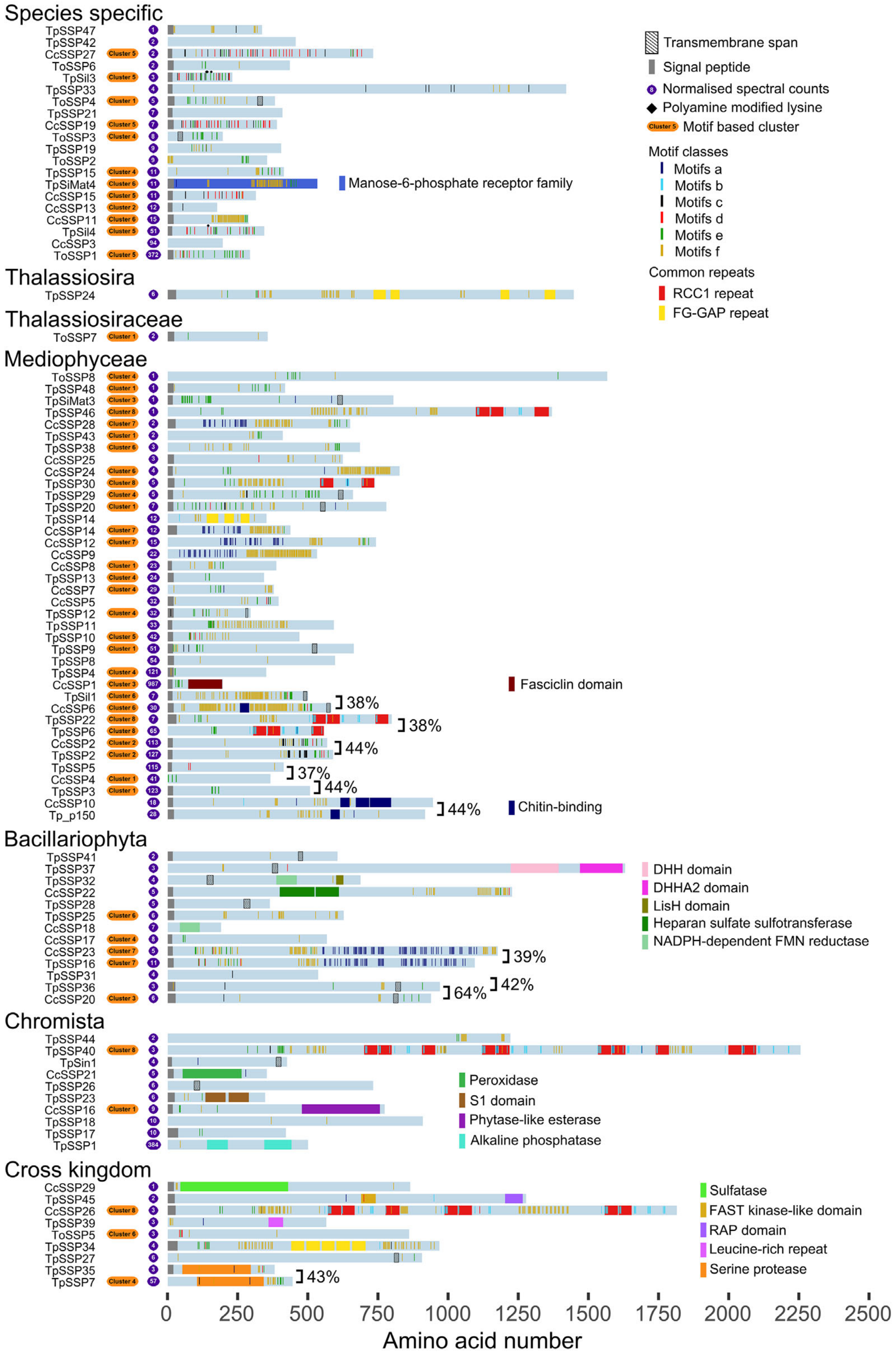


Figure 6. Schematic diagrams summarizing the key features of the soluble silicome proteins, including functional domains, signal peptides, and transmembrane domains. The positions of six classes of motifs are displayed. The classes and colour coding correspond to Figure 4, and cluster membership are indicated. The phylogenetic groupings are derived from the MMETSP data shown in Figure 5. The most similar pairs of proteins according to global pairwise sequence alignments are bracketed and their percentage identity shown. The normalized spectral counts (counts per 100 residues) for each protein across all samples are shown to the left of each protein chain with a purple background.

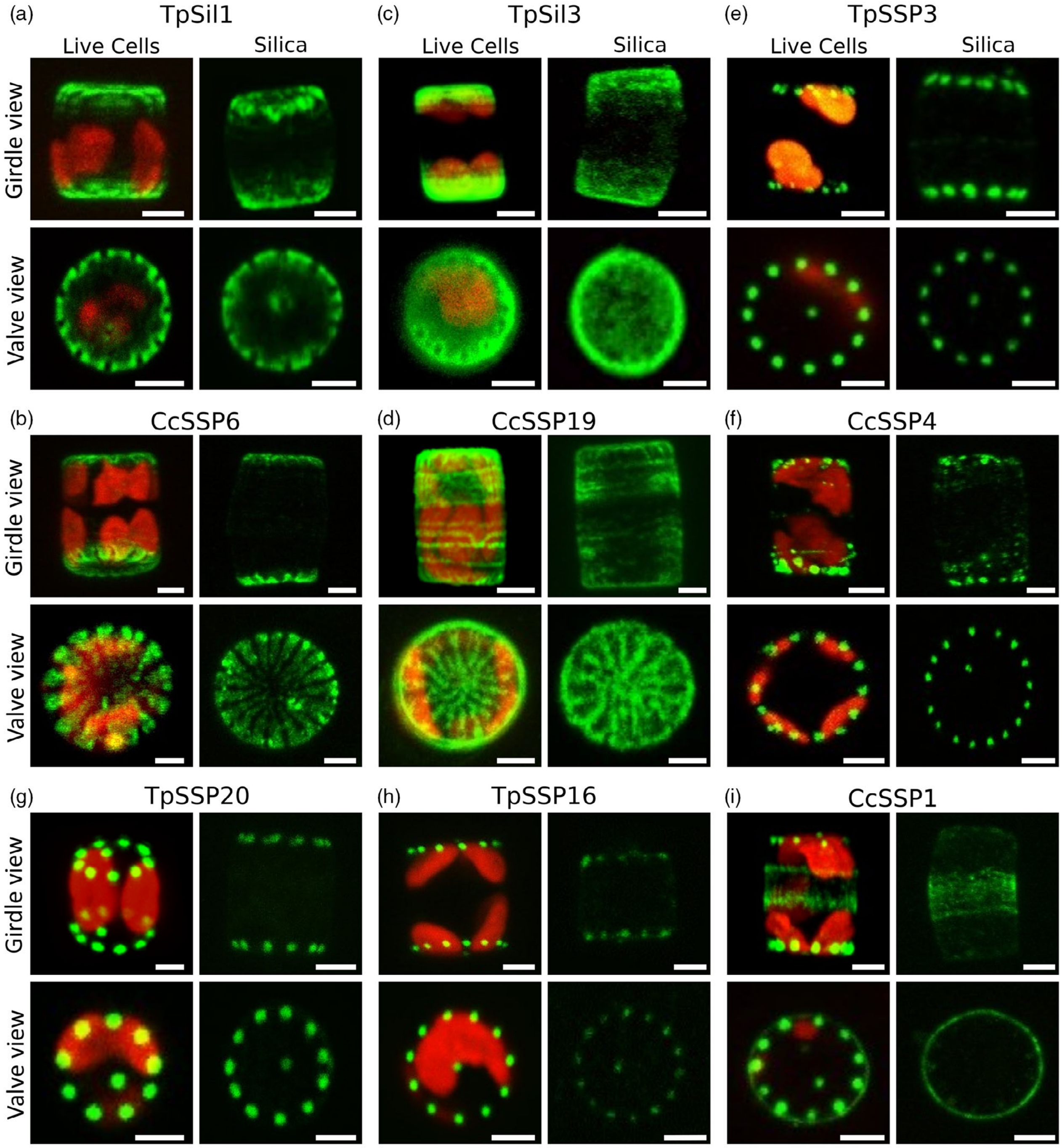


Figure 7. Localization of SS-GFP fusion proteins. (a) TpSil1-GFP^{int}, (b) CcSSP6^{int}-GFP, (c) TpSil3-GFP, (d) CcSSP19-GFP, (e) TpSSP3-GFP, (f) CcSSP4-GFP, (g) TpSSP20-GFP, (h) TpSSP16-GFP, and (i) CcSSP1-GFP fusion proteins in *Thalassiosira pseudonana* and *Cyclotella cryptica*. Each protein is expressed in its species of origin. Confocal fluorescence microscopy images are of live cells and isolated biosilica in valve and girdle band orientation. Green, GFP fluorescence; red, chloroplast autofluorescence. Scale bars: 2 μm.

inspection of the CcSSP4 gene model revealed that it was truncated, so the non-truncated version (52% identity to TpSSP3), which included a signal peptide, was used for GFP tagging (Figure S5). When expressed as GFP fusion proteins, TpSSP3 and CcSSP4 were located exclusively in the valves (Figure 7e,f). Remarkably, they exhibited an evenly spaced punctate location around the circumference of the valve, in a pattern matching the positioning of the fultoportulae (see Figure 1a,g). We employed correlative fluorescence and electron microscopy imaging to

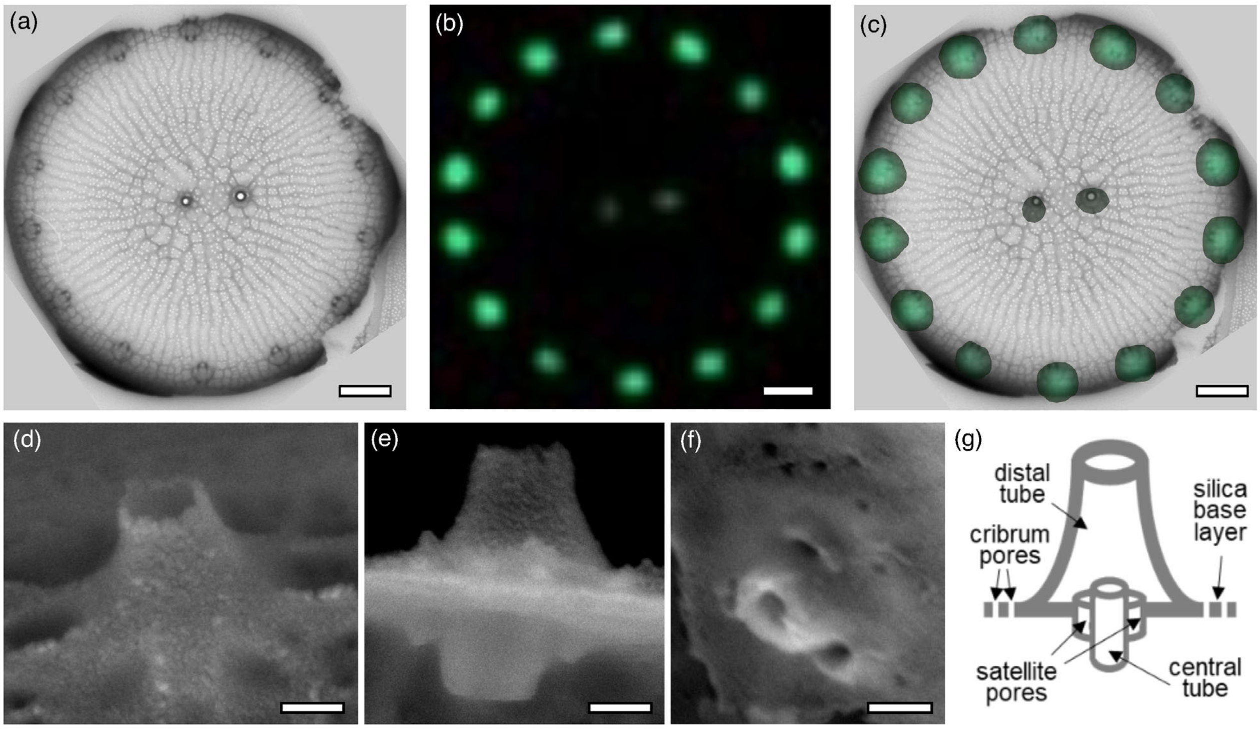


Figure 8. *Thalassiosira pseudonana*fultoportula structure. (a) Transmission electron microscopy image of a *T. pseudonana* valve from a TpSSP3-GFP expressing cell line. (b) GFP fluorescence from the same valve. (c) Overlay of GFP signal and transmission electron microscopy image. Scale bars: (a–c) 500 nm. (d–f) Scanning electron microscopy images from different individual fultoportulae in *T. pseudonana*. (d) Distal tube in oblique view; (e) view at a valve edge showing the distal tube in the upper half and the central tube with satellite pores on the bottom half of the image; (f) view into the opening of the distal tube showing the central tube and the associated satellite pores. (g) Schematic of the fultoportula structure. Scale bars for (d–g), 100 nm.

investigate the location of TpSSP3-GFP relative to the fultoportulae in *T. pseudonana*. The result shows that TpSSP3-GFP is indeed precisely co-located with the fultoportulae (Figure 8a–c).

TpSSP20 belongs to motif Cluster 1 and is well conserved in the Mediophyceae, yet it differs from the Cluster 1 proteins analysed above (CcSSP3 and CcSSP4) in being also enriched in SMSM motifs and containing a single transmembrane domain. Despite these differences, it was also located exclusively in the valves and showed a regular, punctate pattern consistent with fultoportulae location (Figure 7g).

Motif Cluster 2 also contained only uncharacterized proteins, but all its members have two characteristic regions: an N-terminal region rich in 'type f' motifs (Ser, Pro, and Thr rich), and a C-terminal region rich in the Cys-rich 'type a' motifs such as IGxxSC and GxxACxxN. We chose to localize TpSSP16 as a representative of this group of proteins. It shares 29% sequence identity with CcSSP23. Both these sequences are well conserved across the Mediophyceae, but also have some similar sequences within the Bacillariophyceae, but none in the Coscinodiscophyceae. TpSSP16-GFP also exhibited the characteristic fultoportula localization pattern (Figure 7h). No data could be obtained for CcSSP23-GFP, which failed to display GFP fluorescence in transformant clones.

Given that CcSSP1 lacked similar sequences among *T. pseudonana* SSPs, had a fasciclin domain (that is absent from the *T. pseudonana* SSPs) and was identified by more spectral counts than any other *C. cryptica* SSP, we considered CcSSP1 to be a good candidate for a protein, which could be responsible for some structural differences between *T. pseudonana* and *C. cryptica* silica. This protein is found in only 6% of Mediophyceae species in the MMETSP data, and belongs to motif Cluster 3, being rich in SxxSSKS motifs as well as 'type e' motifs such as KSGK. The GFP fusion protein was observed in both girdle bands and valves. In the girdle bands, CcSSP1-GFP was homogeneously distributed in the mid-cell region, and in the valve it exhibited the fultoportula location pattern (Figure 7i). After EDTA/SDS treatment the GFP fluorescence from the fultoportulae was entirely removed, whereas the GFP fluorescence from the girdle bands remained (Figure 7i).

DISCUSSION

The present study vastly increases the number of known silica-associated proteins, providing unique insights into their phylogenetic relationships, sequence characteristics, and possible functions in silica morphogenesis. From the three species, we identified 92 SSPs, 75 of which had not been previously identified at the protein level. Almost 80%

(73 of 92) of the SSPs appear to be diatom specific, and they are often rich in unusual sequence motifs (Figure 4). Sequence similarity among SSPs is generally low: only 10 of 4186 possible pairs of SSPs displayed a global sequence identity above 35%. It has previously been noted that biomineral-associated proteins tend to be biased in amino acid composition and intrinsically disordered (Evans, 2019). However, such qualitative observations have only rarely been quantitatively analysed or put into perspective to the whole proteomes of the biomineralizing organisms (Skeffington & Donath, 2020). Here we show that the SSPs differ as a group from their respective background proteomes in (i) being significantly more biased in amino acid composition, (ii) exhibiting significantly lower sequence complexity, and (iii) being predicted to be intrinsically disordered over a larger proportion of their lengths (see Figure 3). These trends were more pronounced in species-specific SSPs, although this may simply reflect the difficulties of assigning orthology relationships to low complexity, disordered proteins.

It was very striking, from both our proteomic and SDS-PAGE data, that the SSPs of *T. oceanica* are quite different from that of *T. pseudonana* and *C. cryptica*. In fact, none of the SSPs from *T. pseudonana* or *C. cryptica* appeared to have homologues in the *T. oceanica* SSP set, and many proteins were unique to either *T. pseudonana* or *C. cryptica* SSPs. A possible reason could be differences in protein abundance or post-translation modification patterns meaning that proteins are more likely to be detected in one species than another. It is also possible that some of the SSPs are principally components of the ammonium fluoride insoluble organic matrices, and a small proportion of them are solubilized to different extents in the three species. However, only three of the 12 proteins that were previously found in the insoluble organic matrix from *T. pseudonana* were present among the SSPs (TpSil1, TpSiMat3, and TpSiMat4). Consequently, we expect that many of the proteins that constitute the insoluble organic matrices of *C. cryptica* and *T. oceanica* are probably absent from the SSP data set. To identify the full set of biosilica-associated proteins from the three species will require the development of methods to characterize comprehensively the protein complement of the insoluble organic matrices.

Whatever the underlying biochemical reason, our data do suggest strong differences in the SSP complement between species. This could be interpreted as evidence that the soluble silica proteins are unlikely to be important components of the basal silicification machinery of diatoms, and that their role may be more related to the fine-tuning of the silica structures. However, an alternative explanation is that orthologues (or at least functional homologues) of SSPs do exist in all three species, but they are not readily recognized in sequence alignments and BLAST searches due to the low complexity nature of the sequences. In support of this idea, we found that clustering SSPs by motif content resulted in groups of proteins with members derived from all three species. For example, TpSil3 and TpSil4 belonged to the same cluster as CcSSP19 and ToSSP1, leading us to hypothesize that these proteins may have similar roles in silica formation in each species, driven by the particular physio-chemical properties of the amino acid side chains in the motifs. Indeed, GFP-tagging demonstrated that TpSil3 and CcSSP19 have a very similar distribution within the valve silica. To test the functional equivalence of proteins within motif-clusters more formally, knock-out mutants must be generated and characterized. However, because there are often multiple proteins from a species within a cluster indicates that redundancy may make this a challenging approach in practice.

We investigated the intracellular locations of seven newly identified SSPs *in vivo* through GFP-tagging, and each was found to be tightly associated with the biosilica. The result of this spot-check strongly supports the notion that most SSPs are probably genuine silica-associated proteins rather than contaminants.

An unexpected outcome of the GFP-tagging studies was the identification of five proteins with a fultoportula-specific location, which has never been observed before. Fultoportulae have a much more complex morphology than any of the other structural elements were in the cell walls of *Thalassiosira* and *Cyclotella* species. A fultoportula is composed of a central silica tube and two to four satellite pores, which each penetrate the silica base layer of the valve (Figure 8d–g). At the distal side, the pores are covered by a funnel-shaped tube. It is conceivable that the biosynthesis of such an elaborate structure will require a larger number of morphogenic proteins compared with the comparatively simple patterns of branched, cross-linked ribs and cribrum pores. One of fultoportulae-located proteins, TpSSP20, was previously suggested to be involved in cribrum pore formation based on RNAi experiments targeting the *tpssp20* gene (Trofimov et al., 2019). The resulting mutant phenotype exhibited alterations in the area and density of the cribrum pores in the valve, but no effect on the positioning or morphology of the fultoportulae was reported. Considering the location of TpSSP20, the result from the RNAi work is rather surprising, yet it is possible that the concentration of the TpSSP20-GFP fusion protein in the cribrum pores was is too low to be detectable by fluorescence microscopy. In the RNAi knock-down study there was also no analysis of the degree of TpSSP20 gene silencing. Given the rather mild effect on cribrum pores in the RNAi mutants, it is conceivable that the downregulation of TpSSP20 achieved was insufficient to affect fultoportula morphogenesis. The generation of knock-out mutants entirely lacking TpSSP20 and the other fultoportula located proteins would provide more definite insight into their role in silica morphogenesis. It would also be interesting to modulate the expression of CcSSP19 and CcSSP6, as their localization suggests that they could have roles in determining the width and spacing of the wide ribs.

Finally, we were interested to see whether any of the SSPs might be part of basal silicification machinery common to all diatoms. Four proteins [TpSSP41, TpSSP28, CcSSP20 (with homologues TpSSP31 and TpSSP36), and Sin1] had homologues in >80% of diatom species represented in the MMETSP transcriptomes. Given that some of these transcriptomes are probably incomplete, and that they are very unlikely to represent the full expression potential of the organism, it is currently reasonable to suppose that these four proteins are conserved throughout the diatoms. The only one of these proteins to have been previously studied is TpSin1, a knock-out of which displayed changes in silica architecture and reduced the strength of the silica cell wall (Görlich et al., 2019). In the future it will be interesting to see if these proteins can be detected in the silica of diatoms from diverse lineages, and to explore their functional roles through reverse genetics.

EXPERIMENTAL PROCEDURES

Culture conditions

Thalassiosira pseudonana (Hustedt) Hasle et Heimdal clone CCMP1335, *C. cryptica* Reimann, Lewin et Guillard strain CCMP332, and *T. oceanica* (Hustedt) Hasle et Heimdal CCMP1005 were grown in an enriched artificial seawater medium (EASW) (Harrison et al., 1980) at 18°C under constant light at 5000–10 000 lux.

Identification of silica cell wall-associated proteins

Isolation of diatom silica from approximately 50 g wet cell pellet of a stationary cell culture and subsequent ammonium fluoride extraction was performed as described for *T. pseudonana* (Kotzsch et al., 2016). The ammonium fluoride extracts were centrifuged at 3200 ***g*** for 30 min and the supernatants were desalted by three rounds of ultrafiltration (molecular weight cut-off: 3–6 kDa), each resulting in a 10-fold dilution with 10 mM ammonium acetate. Extracts were analysed by SDS-PAGE using 6% Tris-Tricine gels (Schägger & von Jagow, 1987) stained using 'Stains-All' (Campbell et al., 1983), where 0.2% (*T. pseudonana*), 0.5% (*T. oceanica*), and 5% (*C. cryptica*) of the total ammonium fluoride soluble extract was loaded (Figure 2). In solution and in-gel digests were analysed by LC-MS/MS and database searches carried out using Mascot (Methods S1) (Matrixscience, London, UK). Protein inference was carried out using Scaffold (ProteomeSoftware, Portland, OR, USA) (protein probability >0.99, at least two unique peptides, found in two biological replicates: see Methods S1). For MIAPI compliance, raw data have been deposited in the PRIDE database (Chambers et al., 2012) (part of the ProteomeXchange consortium; Deutsch et al., 2020), with accession no. PXD026496. Scaffold files deposited with the PRIDE submission provide full details of identification.

Bioinformatic methods

The properties of the protein sequences were investigated using ProminTools (Skeffington & Donath, 2020), which relies on various software tools: low complexity regions were identified using Seg (Wootton & Federhen, 1993) with default parameters; predicted intrinsic disorder was calculated using VSL2 (Peng et al., 2006); biases in amino acid content were identified using the FLPS software (Harrison, 2017) with a *P*-value cut-off of 10^{-6} and bias quantified as previously described (Skeffington & Donath, 2020). Motif finding was carried out using Motif-x (Chou & Schwartz, 2011) via ProteinMotifFinder (Skeffington & Donath, 2020). Clustering based on motif content relied on the Ward.D method, and a distance matrix based on the Distance Correlation (Székely & Rizzo, 2014) measure (further details in Methods S2). Pairwise global sequence alignments were calculated using EMBOSS Needle (Needleman & Wunsch, 1970) with default parameters. Cellular locations were predicted using Hectare (Gschloessl et al., 2008) and ASAfind (Gruber et al., 2015), while transmembrane prediction was carried out using PureseqTM (Qing et al., 2019) and MEMSAT-SVM (Nugent & Jones, 2009). Known domains were discovered using INTERPRO v. 84.0 (Jones et al., 2014).

The phylogenetic distribution of proteins was carried out using the MMETSP transcriptome data (Keeling et al., 2014) in particular the '_clean.fasta' files from which contaminants have been removed (software: https://github.com/kolecko007/mmetsp_cleanup, data: www.imicrobe.us/#/projects/104) (Marron et al., 2016). tBLASTn searches were carried out with Seg-based filtering of low complexity regions turned on, with default parameters. For further computational details see Methods S3.

Generation and characterization of transgenic cell lines

The construction for gfp fusions with genes of interest are described in detail in Methods S4. Introduction of fusion genes into *T. pseudonana* and *C. cryptica* and selection of transformants were performed as described previously (Kumari et al., 2020; Poulsen et al., 2006). Cells originating from at least 10 independent colonies were examined by epifluorescence microscopy to check for a consistent location phenotype for each GFP fusion protein. Confocal microscopy was performed with live cells and extracted silica from representative cell lines (see Methods S5).

ACKNOWLEDGEMENTS

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) through Research Unit 2038 'Nanomee' (grants KR 1852/8-2 to NK, and SH94/4-2 to ASh) and by a fellowship from the Alexander von Humboldt Stiftung (to AWS). The facility 'Molecular Analysis - Mass Spectrometry' at the CMCB (TU Dresden) was generously supported by grants of the European Regional Development Fund (ERDF/EFRE) (contract no. 100232736), the German Federal Ministry of Education and Research (BMBF, 03Z22EB1), and the German Research Foundation (DFG, INST 269/731-1 FUGG). We thank Damian Pawolski for providing the scanning electron microscopy images in Figure 1. We also thank Anke Bormmann, Jennifer Klemm, Jannes Haupstein, and Alexander Kotzsch for assistance with cloning of the GFP constructs and diatom transformations. Open Access funding enabled and organized by Projekt DEAL.

AUTHOR CONTRIBUTIONS

AWS analysed the data, developed bioinformatic methods, and wrote the paper. NK and NP conceived the project, designed experiments, analysed data, performed bioinformatics analyses, and wrote the paper. CH, AO, LB, and SG designed and performed experiments and analysed data. MG designed proteomic experiments and analysed data.

CONFLICT OF INTERESTS

The authors declare that they have no competing interests.

DATA AVAILABILITY STATEMENT

Raw proteomics data and Scaffold file containing all peptide information and spectra are available from the PRIDE repository with accession no. PXD026496.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Bias in amino acid frequencies for the SSPs of the three species compared with the respective background proteomes. Values above zero indicate enrichment, values below zero indicate depletion.

Figure S2. Results of all versus all pairwise blastp comparisons for the protein clusters identified in Figure 4. The percentage identity of the top HSP for each comparison is indicated by the colour scale and the size of the squares. Query sequences are rows, subject sequences are columns. Note that the nature of the blast comparison means the results are not always exactly symmetrical.

Figure S3. Heatmap showing the distribution of protein motif-cluster membership among the phylogenetic categories.

Figure S4. Schematic structures of GFP fusion proteins. The letters flanking the eGFP bar represent the SSP amino acid sequence in the immediate vicinity of the eGFP tag.

Figure S5. The revised gene model of CcSSP4. (a) Blast of the CcSSP4 protein sequence (black) against *C. cryptica* genomic contigs revealed two overlapping contigs (green). (b) Predicted cDNA sequence. (c) Predicted protein sequence. The predicted signal peptide is shown in bold.

Table S2. *Thalassiosira pseudonana* SSP genes identified in transcriptomics studies of diatom responses to silicon availability. Study 1 (Mock et al., 2008) identified genes with altered expression in cells under Si and Fe limitation. Study 2 (Shrestha et al., 2012) classified genes as 'silaffin-like response genes' (SLRGs) if they matched the expression profile of TpSil3, a cell culture that was synchronized using the silicon starvation-replenishment procedure. Study 3 (Brembu et al., 2017) was also performed with synchronized cells. The TpSil2 cluster contained genes that strongly decreased in expression under Si starvation and showed a strong induction after Si replenishment. The CinY1 cluster showed a similar regulation, but of smaller magnitude, and the SiMat7-like cluster showed an expression peak 2–4 h after Si replenishment and a subsequent decrease in expression. Study 4 (Scheffel et al., 2011) used a bioinformatics analysis to identify silaffin-like proteins in the *T. pseudonana* genome.

Methods S1. Proteomics analysis.

Methods S2. Clustering proteins based on motif content.

Methods S3. MMETSP blast analysis.

Methods S4. Generating SSP-GFP fusion constructs.

Methods S5. Confocal fluorescence microscopy imaging.

Appendix S1. Known domains found in SSPs.

Table S1. Summary of the characteristics of the SSPs.

Table S3. Results of all versus all global sequence alignments with EMBOSS Needle. Values are percentage identity.

Table S4. Motif enrichment values for all SSPs and enriched motifs. Values are fold enrichment relative to the background proteome.

Table S5. Motif-based protein correlation matrix.

Table S6. MMETSP transcriptomes used in analyses and taxonomic assignment.

REFERENCES

Alverson, A.J., Beszteri, B., Julius, M.L. & Theriot, E.C. (2011) The model marine diatom *Thalassiosira pseudonana* likely descended from a freshwater ancestor in the genus Cyclotella. *BMC Evolutionary Biology*, **11**, 125.

Benoiston, A.-S., Ibarbalz, F.M., Bittner, L., Guidi, L., Jahn, O., Dutkiewicz, S. et al. (2017) The evolution of diatoms and their biogeochemical functions. *Proceedings of the Royal Society B: Biological Sciences*, **372**, 20160397.

Brembu, T., Chauton, M.S., Winge, P., Bones, A.M. & Vadstein, O. (2017) Dynamic responses to silicon in *Thalasiossira pseudonana* - identification, characterisation and classification of signature genes and their corresponding protein motifs. *Scientific Reports*, **7** Article number: 4865.

Brunner, E., Richthammer, P., Ehrlich, H., Paasch, S., Simon, P., Ueberlein, S. et al. (2009) Chitin-based organic networks: an integral part of cell wall biosilica in the diatom thalassiosira pseudonana. *Angewandte Chemie, International Edition*, **48**, 9724–9727.

Buhmann, M.T., Poulsen, N., Klemm, J., Kennedy, M.R., Sherrill, C.D. & Kröger, N. (2014) A tyrosine-rich cell surface protein in the diatom Amphora coffeaeformis identified through transcriptome analysis and genetic transformation. *PLoS One*, **9**, e110369.

Campbell, K.P., MacLennan, D.H. & Jorgensen, A.O. (1983) Staining of the Ca^{2+}-binding proteins, calsequestrin, calmodulin, troponin C, and S-100, with the cationic carbocyanine dye 'stains-all'. *Journal of Biological Chemistry*, **258**, 11267–11273.

Cha, J.N., Shimizu, K., Zhou, Y., Christiansen, S.C., Chmelka, B.F., Stucky, G.D. et al. (1999) Silicatein filaments and subunits from a marine sponge direct the polymerization of silica and silicones in vitro. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 361–365.

Chambers, M.C., MacLean, B., Burke, R., Amodei, D., Ruderman, D.L., Neumann, S. et al. (2012) A cross-platform toolkit for mass spectrometry and proteomics. *Nature Biotechnology*, **30**, 918–920.

Chou, M.F. & Schwartz, D. (2011) Biological sequence motif discovery using motif-x. *Current Protocols in Bioinformatics*, **13**, 15–24.

Davis, A.K., Hildebrand, M. & Palenik, B. (2005) A stress-induced protein associated with the girdle band region of the diatom *Thalassiosira pseudonana* (Bacillariophyta). *Journal of Phycology*, **41**, 577–589.

Deutsch, E.W., Bandeira, N., Sharma, V., Perez-Riverol, Y., Carver, J.J., Kundu, D.J. et al. (2020) The ProteomeXchange consortium in 2020: enabling 'big data' approaches in proteomics. *Nucleic Acids Research*, **48**, D1145–D1152.

Evans, J.S. (2019) The biomineralization proteome: protein complexity for a complex bioceramic assembly process. *Proteomics*, **19**, 1900036.

Goessling, J.W., Sue, Y., Kühl, M. & Ellegaard, M. (2022) Frustule photonics and light harvesting strategies in diatoms. In: Annenkov, V., Seckbach, J. & Gordon, R. (Eds.) *Diatom morphogenesis*. Beverly, MA: Scrivener Publishing LLC.

Görlich, S., Pawolski, D., Zlotnikov, I. & Kröger, N. (2019) Control of biosilica morphology and mechanical performance by the conserved diatom gene Silicanin-1. *Communications Biology*, **2**, 245.

Gruber, A., Rocap, G., Kroth, P.G., Armbrust, E.V. & Mock, T. (2015) Plastid proteome prediction for diatoms and other algae with secondary plastids of the red lineage. *Plant Journal*, **81**, 519–528.

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

Harrison, P.J., Waters, R.E. & Taylor, F.J.R. (1980) A broad spectrum artificial sea water medium for coastal and open ocean phytoplankton 1. *Journal of Phycology*, **16**, 28–35.

Harrison, P.M. (2017) fLPS: fast discovery of compositional biases for the protein universe. *BMC Bioinformatics*, **18**, 476.

Hecky, R.E., Mopper, K., Kilham, P. & Degens, E.T. (1973) The amino acid and sugar composition of diatom cell-walls. *Marine Biology*, **19**, 323–331.

Herth, W. (1979) The site of beta-chitin fibril formation in centric diatoms. II. The chitin-forming cytoplasmic structures. *Journal of Ultrastructure Research*, **68**, 16–27.

Hildebrand, M., Frigeri, L.G. & Davis, A.K. (2007) Synchronized growth of *Thalassiosira pseudonana* (Bacillariophyceae) provides novel insights into cell-wall synthesis processes in relation to the cell cycle. *Journal of Phycology*, **43**, 730–740.

Jones, P., Binns, D., Chang, H.Y., Fraser, M., Li, W., McAnulla, C. et al. (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics*, **30**, 1236–1240.

Kaczmarska, I., Beaton, M., Benoit, A.C. & Medlin, L.K. (2005) Molecular phylogeny of selected members of the order Thalassiosirales (Bacillariophyta) and evolution of the fultoportua. *Journal of Phycology*, **42**, 121–138.

Keeling, P.J., Burki, F., Wilcox, H.M., Allam, B., Allen, E.E., Amaral-Zettler, L.A. et al. (2014) The marine microbial eukaryote Transcriptome sequencing project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. *PLoS Biology*, **12**, e1001889.

Kotzsch, A., Gröger, P., Pawolski, D., Bomans, P.H.H., Sommerdijk, N.A.J.M., Schlierf, M. et al. (2017) Silicanin-1 is a conserved diatom membrane protein involved in silica biomineralization. *BMC Biology*, **15**, 65.

Kotzsch, A., Pawolski, D., Milentyev, A., Shevchenko, A., Scheffel, A., Poulsen, N. et al. (2016) Biochemical composition and assembly of biosilica-associated insoluble organic matrices from the diatom *Thalassiosira pseudonana*. *Journal of Biological Chemistry*, **291**, 4982–4997.

Kröger, N., Deutzmann, R., Bergsdorf, C. & Sumper, M. (2000) Species-specific polyamines from diatoms control silica morphology. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 14133.

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

Kröger, N., Deutzmann, R. & Sumper, M. (2001) Silica-precipitating peptides from diatoms: the chemical structure of silaffin-1A from *Cylindrotheca fusiformis*. *Journal of Biological Chemistry*, **276**, 26066–26070.

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

Kröger, N. & Sumper, M. (2004) The molecular basis of diatom biosilica formation. In: Bäuerlein, E. (Ed.) *Biomineralization: progress in biology, molecular biology and application*. Wenheim, Germany: Wiley Blackwell, pp. 137–158.

Kumari, E., Görlich, S., Poulsen, N. & Kröger, N. (2020) Genetically programmed regioselective immobilization of enzymes in biosilica microparticles. *Advanced Functional Materials*, **30**, 2000442.

Lowenstam, H.A. & Weiner, S. (1989) *On biomineralization*. New York and London: Oxford University Press.

Marron, A.O., Ratcliffe, S., Wheeler, G.L., Goldstein, R.E., King, N., Not, F. et al. (2016) The evolution of silicon transport in eukaryotes. *Molecular Biology and Evolution*, **33**, 3226–3248.

Mitchell, J.G., Seuront, L., Doubell, M.J., Losic, D., Voelcker, N.H., Seymour, J. et al. (2013) The role of diatom nanostructures in biasing diffusion to improve uptake in a patchy nutrient environment (S Humphries, Ed.). *PLoS One*, **8**, e59548.

Mock, T., Samanta, M.P., Iverson, V., Berthiaume, C., Robison, M., Holtermann, K. et al. (2008) Whole-genome expression profiling of the marine diatom *Thalassiosira pseudonana* identifies genes involved in silicon bioprocesses. *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 1579–1584.

Nassif, N. & Livage, J. (2011) From diatoms to silica-based biohybrids. *Chemical Society Reviews*, **40**, 849–859.

Needleman, S.B. & Wunsch, C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, **48**, 443–453.

Nelson, D.M., Tréguer, P., Brzezinski, M.A., Leynaert, A. & Quéguiner, B. (1995) Production and dissolution of biogenic silica in the ocean: revised global estimates, comparison with regional data and relationship to biogenic sedimentation. *Global Biogeochemical Cycles*, **9**, 359–372.

Nugent, T. & Jones, D.T. (2009) Transmembrane protein topology prediction using support vector machines. *BMC Bioinformatics*, **10** Article number: 159.

Osuna-Cruz, C.M., Bilcke, G., Vancaester, E., De Decker, S., Bones, A.M., Winge, P. et al. (2020) The *Seminavis robusta* genome provides insights into the evolutionary adaptations of benthic diatoms. *Nature Communications*, **11**, 3320.

Pančić, M., Torres, R.R., Almeda, R. & Kiørboe, T. (2019) Silicified cell walls as a defensive trait in diatoms. *Proceedings of the Royal Society B: Biological Sciences*, **286**, 20190184.

Pawolski, D., Heintze, C., Mey, I., Steinem, C. & Kröger, N. (2018) Reconstituting the formation of hierarchically porous silica patterns using diatom biomolecules. *Journal of Structural Biology*, **204**, 64–74.

Peng, K., Radivojac, P., Vucetic, S., Dunker, A.K. & Obradovic, Z. (2006) Length-dependent prediction of protein in intrinsic disorder. *BMC Bioinformatics*, **7**, Article number: 208.

Poulsen, N., Chesley, P.M. & Kröger, N. (2006) Molecular genetic manipulation of the diatom *Thalassiosira pseudonana* (Bacillariophyceae). *Journal of Phycology*, **42**, 1059–1065.

Poulsen, N. & Kröger, N. (2004) Silica morphogenesis by alternative processing of silaffins in the diatom *Thalassiosira pseudonana*. *Journal of Biological Chemistry*, **279**(42993–42), 999.

Poulsen, N., Scheffel, A., Sheppard, V.C., Chesley, P.M. & Kröger, N. (2013) Pentalysine clusters mediate silica targeting of silaffins in *Thalassiosira pseudonana*. *Journal of Biological Chemistry*, **288**, 20100–20109.

Poulsen, N., Sumper, M. & Kröger, N. (2003) Biosilica formation in diatoms: characterization of native silaffin-2 and its role in silica morphogenesis. *Proceedings of the National Academy of Sciences of the United States of America*, **21**, 12075–12080.

Qing, W., Ni, C., Li, Z., Li, X., Han, R., Zhao, F. et al. (2019) PureseqTM: efficient and accurate prediction of transmembrane topology from amino acid sequence only. *BioRxiv*, 627307. https://doi.org/10.1101/627307.

Raven, J.A. & Waite, A.M. (2004) The evolution of silicification in diatoms: inescapable sinking and sinking as escape? *New Phytologist*, **162**, 45–61.

Sapriel, G., Quinet, M., Heijde, M., Jourdren, L., Tanty, V., Luo, G. et al. (2009) Genome-wide transcriptome analyses of silicon metabolism in *Phaeodactylum tricornutum* reveal the multilevel regulation of silicic acid transporters. *PLoS One*, **4**, e7458.

Schägger, H. & von Jagow, G. (1987) Tricine-sodium dodecyl sulfate-polyacrylamide gel electrophoresis for the separation of proteins in the range from 1 to 100 kDa. *Analytical Biochemistry*, **166**, 368–379.

Scheffel, A., Poulsen, N., Shian, S. & Kröger, N. (2011) Nanopatterned protein microrings from a diatom that direct silica morphogenesis. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 3175–3180.

Shrestha, R.P., Tesson, B., Norden-Krichmar, T., Federowicz, S., Hildebrand, M. & Allen, A.E. (2012) Whole transcriptome analysis of the silicon response of the diatom *Thalassiosira pseudonana*. *BMC Genomics*, **13**, 499.

Skeffington, A.W. & Donath, A. (2020) ProminTools: shedding light on proteins of unknown function in biomineralization with user friendly tools illustrated using mollusc shell matrix protein sequences. *PeerJ*, **8**, e9852.

Sumper, M. & Brunner, E. (2008) Silica biomineralisation in diatoms: the model organism *Thalassiosira pseudonana*. *ChemBioChem*, **9**, 1187–1194.

Sumper, M., Hett, R., Lehmann, G. & Wenzl, S. (2007) A code for lysine modifications of a silica biomineralizing silaffin protein. *Angewandte Chemie International Edition*, **46**, 8405–8408.

Székely, G.J. & Rizzo, M.L. (2014) Partial distance correlation with methods for dissimilarities. *Annals of Statistics*, **42**, 2382–2412.

Tesson, B. & Hildebrand, M. (2013) Characterization and localization of insoluble organic matrices associated with diatom cell walls: insight into their roles during cell wall formation. *PLoS One*, **8**(4), e61675.

Trofimov, A.A., Pawlicki, A.A., Borodinov, N., Mandal, S., Mathews, T.J., Hildebrand, M. et al. (2019) Deep data analytics for genetic engineering of diatoms linking genotype to phenotype via machine learning. *npj Computational Materials*, **5**, 67.

Volcani, B.E. (1981) Cell wall formation in diatoms: Morphogenesis and biochemistry. In: Simpson, T.L. & Volcani, B.E. (Eds.) *Silicon and siliceous structures in biological systems*. New York, NY: Springer New York.

Wieneke, R., Bernecker, A., Riedel, R., Sumper, M., Steinem, C. & Geyer, A. (2011) Silica precipitation with synthetic silaffin peptides. *Organic and Biomolecular Chemistry*, **9**, 5482–5486.

Wootton, J.C. & Federhen, S. (1993) Statistics of local complexity in amino acid sequences and sequence databases. *Computers and Chemistry*, **17**, 149–163.

Žerdoner Čalasan, A., Kretschmann, J. & Gottschling, M. (2018) Absence of co-phylogeny indicates repeated diatom capture in dinophytes hosting a tertiary endosymbiont. *Organisms, Diversity and Evolution*, **18**, 29–38.