Diatom adhesive trail proteins acquired by horizontal gene transfer from bacteria serve as primers for marine biofilm formation

Jirina Zackova Suchanova[1], Gust Bilcke[2,3,4], Beata Romanowska[1], Ali Fatlawi[5,6], Martin Pippel[7], Alastair Skeffington[8], Michael Schroeder[5,6], Wim Vyverman[2], Klaas Vandepoele[3,4], Nils Kröger[1,9,10] and Nicole Poulsen[1]

[1]B CUBE Center for Molecular Bioengineering, Technische Universität Dresden, Dresden, 01307, Germany; [2]Department of Biology, Protistology and Aquatic Ecology, Ghent University, Ghent, 9000, Belgium; [3]Department of Plant Biotechnology and Bioinformatics, Ghent University, Technologiepark 71, Ghent, 9052, Belgium; [4]VIB Center for Plant Systems Biology, Technologiepark 71, Ghent, 9052, Belgium; [5]Biotechnology Center (BIOTEC), Technische Universität Dresden, Tatzberg 47-49, Dresden, 01307, Germany; [6]Centre for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI), Chemnitzer Str. 46b, Dresden, 01187, Germany; [7]Max Planck Institute of Molecular Cell Biology and Genetics, Germany Center for Systems Biology, Pfotenhauerstraße 108, Dresden, 01307, Germany; [8]Biological and Environmental Sciences, Faculty of Natural Sciences, University of Stirling, Stirling, FK9 4LA, UK; [9]Cluster of Excellence Physics of Life, Technische Universität Dresden, Dresden, 01062, Germany; [10]Faculty of Chemistry and Food Chemistry, Technische Universität Dresden, Dresden, 01062, Germany

Authors for correspondence:
Nicole Poulsen
Email: nicole.poulsen@tu-dresden.de

Nils Kröger
Email: nils.kroeger@tu-dresden.de

Received: *7 March 2023*
Accepted: *2 July 2023*

New Phytologist (2023) **240:** 770–783
doi: 10.1111/nph.19145

Key words: bioadhesion, biofilm, diatom, gliding motility, horizontal gene transfer.

Summary

- Biofilm-forming benthic diatoms are key primary producers in coastal habitats, where they frequently dominate sunlit intertidal substrata. The development of gliding motility in raphid diatoms was a key molecular adaptation that contributed to their evolutionary success. However, the structure–function correlation between diatom adhesives utilized for gliding and their relationship to the extracellular matrix that constitutes the diatom biofilm is unknown.
- Here, we have used proteomics, immunolocalization, comparative genomics, phylogenetics and structural homology analysis to investigate the evolutionary history and function of diatom adhesive proteins.
- Our study identified eight proteins from the adhesive trails of *Craspedostauros australis*, of which four form a new protein family called Trailins that contain an enigmatic Choice-of-Anchor A (CAA) domain, which was acquired through horizontal gene transfer from bacteria. Notably, the CAA-domain shares a striking structural similarity with one of the most widespread domains found in ice-binding proteins (IPR021884).
- Our work offers new insights into the molecular basis for diatom biofilm formation, shedding light on the function and evolution of diatom adhesive proteins. This discovery suggests that there is a transition in the composition of biomolecules required for initial surface colonization and those utilized for 3D biofilm matrix formation.

Introduction

Many marine microorganisms are capable of adhering to surfaces underwater where they form taxonomically diverse, often photosynthetically active biofilm communities embedded in a three-dimensional (3D) matrix of self-produced extracellular polymeric substances (EPS). In fact biofilms are the most abundant microbial lifestyle on the planet, accounting for *c.* 80% of all bacterial cells (Flemming & Wuertz, 2019). In shallow coastal habitats, diatoms and cyanobacteria-rich biofilms contribute immensely to global primary productivity, being responsible for *c.* 20% of marine carbon fixation even though they only occupy *c.* 0.03% of the ocean's surface area (Pinckney, 2018). The carbohydrate-rich EPS provides numerous advantages for the encapsulated cells, including protection against desiccation and mechanical stresses, thereby enhancing their ability to thrive in highly dynamic environments (Flemming & Wingender, 2010; Steele *et al.*, 2014).

The evolution of a raphe system in pennate diatoms was a significant milestone that enabled the development of a unique mode of active gliding motility and the production of copious amounts of EPS, which facilitated the colonization of a wide range of benthic habitats (Ashworth, 2013; Nakov *et al.*, 2018). Furthermore, the gliding of diatoms allows cells to actively optimize their position in the environment allowing them to move towards favourable light and nutrient conditions, while avoiding desiccation and toxic compounds. Diatom motility is critically dependent on the presence of a specialized slit in the silica cell wall, termed the raphe. It is hypothesized that the secretion of EPS strands through the raphe slit provides a continuous physical link between the plasma membrane and the substratum (Edgar

& Pickett-Heaps, 1984). The EPS strands have remarkable properties: they form intertwined long tethers that can extend a distance of up to 30 µm away from the cell body; exhibit high adhesive, elastic and tensile strength; and are adhesive underwater on a wide range of materials (Higgins *et al.*, 2003; Holland *et al.*, 2004; Dugdale *et al.*, 2006; Gutierrez-Medina *et al.*, 2022). One of the proposed models for gliding is based on the observation that two actin bundles are positioned immediately below the plasma membrane along the entire length of the raphe (Edgar & Zavortink, 1983; Edgar & Pickett-Heaps, 1984). It is hypothesized that the EPS strands are connected to the actomyosin motor system via a continuum of biomolecules that span the plasma membrane, which is referred to as the adhesion motility complex (AMC). According to this model, the actomyosin motor system translocates the EPS strands that are anchored to the substratum in a rearward direction, thus propelling the cell forwards (Edgar, 1983; Edgar & Zavortink, 1983; Edgar & Pickett-Heaps, 1984). This model was supported by inhibition studies with drugs specific for actin and myosin (Poulsen *et al.*, 1999), but the EPS composition and the other components of the AMC remained unknown.

To investigate the AMC-based mechanism for diatom gliding, we have embarked on identifying the components of the machinery. As diatoms glide across a surface, they often deposit behind them a trail of the EPS strands that are composed of acidic polysaccharides and proteins (Lind *et al.*, 1997; Higgins *et al.*, 2000; Poulsen *et al.*, 2014). We have previously performed a proteomics study and identified 21 putative adhesion proteins from *Amphora coffeaeformis* (Lachnit *et al.*, 2019). These proteins contain several features that also occur in adhesive proteins from other organisms and a newly described, diatom-specific GDPH-domain. Immunolocalization of one GDPH-domain-containing protein, Ac629, confirmed its presence in the raphe and EPS trails, which is consistent with a role in diatom adhesion (Lachnit *et al.*, 2019). However, bioinformatics analysis revealed that not all motile diatoms possess proteins with a GDPH-domain (Lachnit *et al.*, 2019), indicating that other protein domains must also be able to mediate diatom adhesion. To address this question, we have pursued a proteomics analysis of the adhesive trails from *Craspedostauros australis*, which is a raphid diatom species lacking GDPH-domain-bearing proteins. Through long-read genome sequencing and assembly, comparative genomics, phylogenetics, structural homology and immunolocalization studies, we aimed to investigate the evolutionary history of diatom adhesive proteins and their role in the formation of diatom trails and biofilms.

Materials and Methods

Cell cultures

Craspedostauros australis (Cox, CCMP 3328) was grown in artificial seawater medium (ESAW; Harrison *et al.*, 1980) at 18°C and 12 h : 12 h, light : dark cycle with an intensity of 100 µmol photons m^{-2} s^{-1}, using cool-white lamps. Antibiotic treatment with penicillin and streptomycin (both 100 µg ml^{-1}) was performed periodically to ensure cultures remained axenic.

Isolation of diatom adhesive material

The *C. australis* adhesive material (AM) was purified according to a previously published procedure (Poulsen *et al.*, 2014) and stored at −20°C. The freeze-dried AM was solubilized in 0.5 M borate buffer (pH 8.5) containing 2 M hydroxylamine (Sigma Aldrich) for 2 h at 45°C and then desalted against 50 mM ammonium acetate using a PD MidiTrap G-10 column (size exclusion limit: 700 Da; GE Healthcare, Uppsala, Sweden) according to the manufacturer's instructions. Detailed methods for the anhydrous hydrofluoric acid treatment and gel filtration chromatography are in the Methods S1.

LC–MS/MS proteomics analysis

The solubilized AM was subjected to SDS-PAGE and after a short separation (*c*. 4 cm migration into the separating gel) visualized by Coomassie Blue staining. An entire gel lane was cut into four slabs, and each slab was subjected to in-gel proteolytic digestion with trypsin (Promega, Mannheim, Germany) overnight as described in Shevchenko *et al.* (2006). Detailed LC–MS methods are provided in the Methods S1.

PacBio genomic DNA sequencing

HMW genomic DNA (gDNA) was isolated using a CTAB and phenol-chloroform method. The Dresden-concept genome centre (https://dresden-concept.de/genome-center/?lang=en) prepared the libraries for long-read PacBio sequencing. Detailed methods are described in the Methods S1.

Confirmation of gene models

RACE and RT-PCR were used to determine the full-length gene models of the *C. australis* adhesive trail proteins. *C. australis* cDNA and gDNA were prepared as described previously (Poulsen *et al.*, 2023). Two nested RACE PCRs were performed as described previously (Poulsen *et al.*, 2023) using cDNA attached to the Oligo $(dT)_{25}$ Dynabeads and primer combinations listed in Supporting Information Table S1 using DreamTaq (Thermo Fisher Scientific, Dreieich, Germany). The resulting PCR products were cloned into pJet1.2 (Thermo Fisher Scientific) and transformed into the DH5α *E. coli* strain and subsequently sequenced by Eurofins Genomics (Ebersberg, Germany). To confirm the complete gene model and intron-exon boundaries, a PCR was performed over the entire predicted coding region using cDNA and gDNA as a template. The PCR was performed using Q5 DNA polymerase (NEB) according to the manufacturer's instructions and using the GC enhancer buffer. The resulting PCR products were directly sequenced by Microsynth Seqlab (Göttingen, Germany). Primer sequences are provided in the Table S1.

Functional annotation and homology searches

Proteins identified by mass spectrometry were manually annotated according to transcriptomic and genomic data (Poulsen *et al*., 2023) and the PacBio genome assembly generated in this study. Protein sequence homology analysis was performed using a BLASTX search against the nonredundant protein sequences in National Center for Biotechnology Information (NCBI) database with an E-value cut-off of 10^{-10}. Parameters used for further bioinformatics searches are found in the Methods S1. The evolutionary origin of diatom CAA-domains was assessed by creating Hidden Markov Model (HMM) profiles which were used to search for homologous protein domains in the PLAZA Diatoms 1.0, MMETSP and the NCBI nr datasets (Keeling *et al*., 2014; Osuna-Cruz *et al*., 2020), as described in Methods S1.

Structural analysis of diatom Choice-of-Anchor A (CAA)-domains

ALPHAFOLD (v.2.0.1) was used to predict the structure of the CAA-domain and compared with experimentally determined 3D protein structures in the PDB database (https://www.rcsb.org/). The amino acid sequence of CaTrailin4 was submitted to ALPHAFOLD (v.2.0.1 with the full_dbs option) to predict its 3D structure (Jumper *et al*., 2021). Reported quality scores are encoded in the predicted structures b-factors. To identify structural homologues of the CaTrailin_4 CAA-domain, we carried out a dedicated computational screen, as there are currently no tools to compare a predicted structure to all known structures. First, the Uniprot database was downloaded on 22.11.2021 and 170 041 PDB IDs indexed in Uniprot were extracted. If a Uniprot entry had multiple PDB IDs associated with it, then a representative was selected. Criteria for selection were number of residues covered, resolution and then method. After selecting representatives, 54 943 PDB IDs were remaining. The PDB IDs were then compared with the predicted structure of the CaTrailin_4 CAA-domain using structural alignment (CE align) in Pymol. RMSD and percentage of aligned residues were recorded for each comparison. Overall, there were 223 structures with RMSD < 10 Å (see Table S7 for the top eight hits). Next, we sorted these structures by percentage-aligned residues. The top eight PDB IDs were 4NU2, 6QVI, 3UYU, 5B5H, 3WP9, 3VN3, 6A8K and 7BWX. All but 6QVI are ice-binding proteins. We aligned CaTrailin_4 CAA-domain with 4NU2 using Pymol's CE align. We exported the sequence representation of the structural alignment and manually added secondary structure. To assess ice-binding, we used the tool AFPredictor (Doxey *et al*., 2006).

Genetic transformation of *Craspedostauros australis*

Craspedostauros australis was genetically transformed using a previously developed protocol (Poulsen *et al*., 2023). In brief, 10^8 cells were plated on an ESAW agar plate and 5 μg of plasmid DNA coated on W-microparticles (M17; Bio-Rad) was delivered into the cells using the Bio-Rad Biolistic Particle Delivery System (1550 psi, 28 mmHg vacuum). The cells were allowed to recover for 24 h, and then, 5×10^6 cells were plated on ESAW agar plates containing 450 μg ml^{-1} nourseothricin (Jena Bioscience, Jena, Germany) and incubated in constant light at 18°C.

Production of polyclonal antibodies

Custom made rabbit polyclonal antibodies were obtained from GenScript (Piscataway, NJ, USA). Epitopes were chosen based on the manufacturer's prediction algorithm and regions that were covered by the protein sequencing: CaTrailin3 DEDLSKQNTGKTIN; CaTrailin2 EDNLDQIRIITESN; CaTrailin4 DDNVPYEETQRHTA. The antibodies were raised against selected protein sequences and purified by affinity chromatography using the antigen as a ligand. To obtain the IgG fraction, the antibodies were further affinity purified using Protein G mag sepharose (GE Healthcare).

Immunofluorescence and lectin labelling

For immunolabelling of adhesive trails, 5×10^3 *C. australis* cells were incubated for 16 h on a glass bottom chamber slide (μ-slide 8-well chamber, ibidi, Gräfelfing, Germany). All steps were performed at RT. The culture medium was aspirated and the cells incubated in 2% BSA in PBS (BS) for 1 h. The samples were then incubated for 2 h with primary antibodies or preimmune serum (5 μg ml^{-1}) diluted in BS + 0.05% Tween 20. The samples were washed (1× 2 s, 1× 5 min) with PBS + 0.05% Tween 20 (PBST) and then incubated in secondary antibody (Goat-anti-rabbit conjugated with AlexaFluor488, 1 : 3000 diluted in PBST; Life Technologies) for 1 h in the dark. After washing twice with PBST and overlaying with ibidi mounting medium (Ibidi), confocal microscopy was performed using a Zeiss LSM780 inverted microscope equipped with a Zeiss Plan Apochromat 63× (1.4) Oil DIC M27 objective or C-Apochromat 40×/1.2 W Corr M27 objective. Two channels were used to separately monitor chloroplast fluorescence (emission at 654–693 nm) and AlexaFluor488 or Atto488 channel (emission at 480–515 nm). Images were analysed using the ZEN 3.1 software (Zeiss). Detailed methods for the lectin labelling and whole cell and biofilm immunolabelling are provided in the Methods S1.

Results

Identification of *Craspedostauros australis* adhesive trail proteins

The biochemical characterization of biological adhesives is hampered by their insolubility in reagents typically used in biochemical studies. Recently, we demonstrated that hydroxylamine completely solubilizes the *A. coffeaeformis* AM, which enabled identification of the first diatom adhesive trail proteins (Lachnit *et al*., 2019). Following the same approach, we used hydroxylamine to solubilize the *C. australis* AM. SDS-PAGE in combination with Stains-All staining revealed that the *C. australis* AM, like *A. coffeaeformis*, is dominated by acidic, high-molecular-weight (HMW) components (> 250 kDa) and a minor fraction of low molecular weight (LMW) components (< 30 kDa; Fig. S1a). In contrast to the *A. coffeaeformis* LMW components, which can be stained with Coomassie blue, the *C. australis* LMW components can only be stained with silver, indicating clear

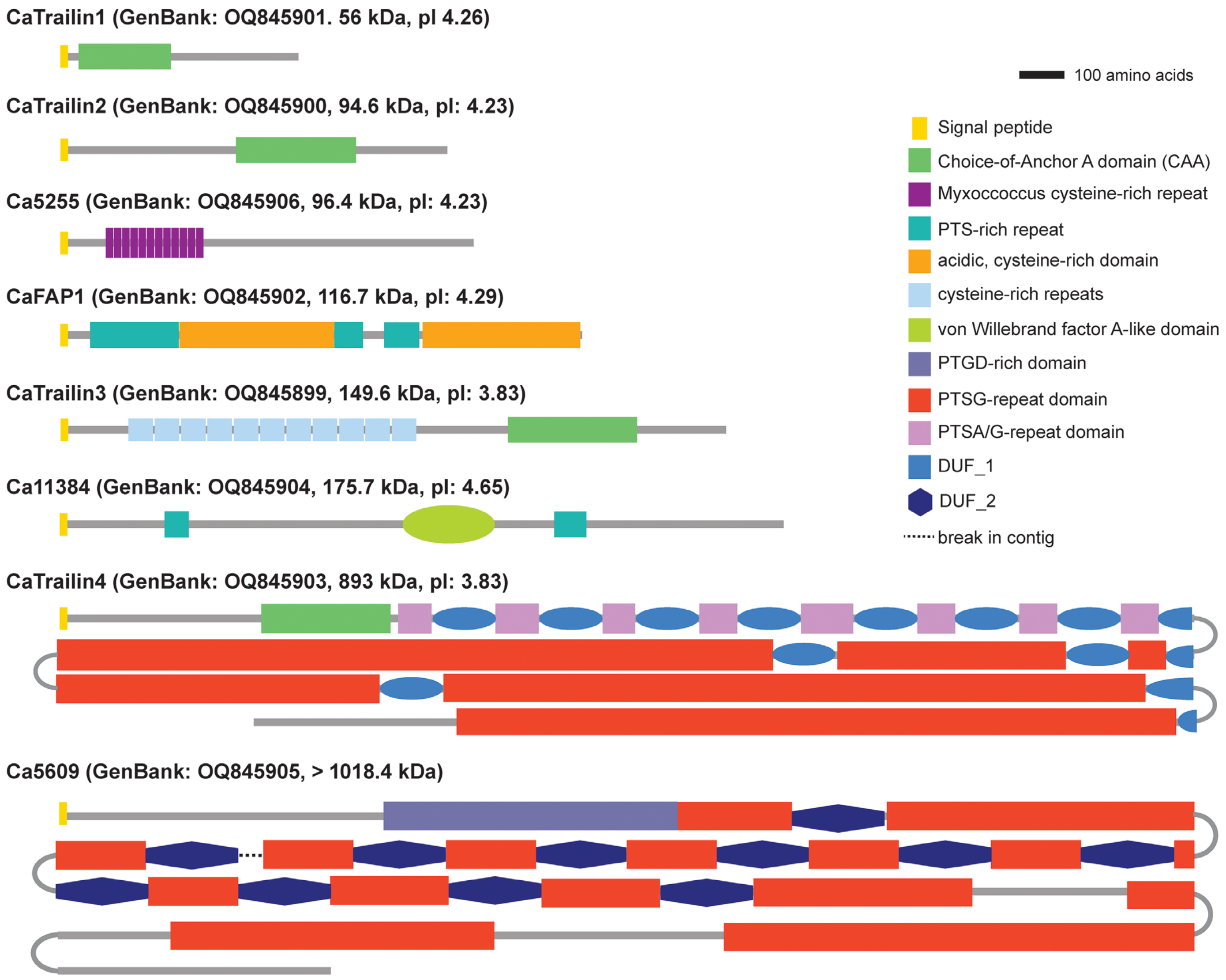
CaTrailin1 (GenBank: OQ845901. 56 kDa, pI 4.26)
CaTrailin2 (GenBank: OQ845900, 94.6 kDa, pI: 4.23)
Ca5255 (GenBank: OQ845906, 96.4 kDa, pI: 4.23)
CaFAP1 (GenBank: OQ845902, 116.7 kDa, pI: 4.29)
CaTrailin3 (GenBank: OQ845899, 149.6 kDa, pI: 3.83)
Ca11384 (GenBank: OQ845904, 175.7 kDa, pI: 4.65)
CaTrailin4 (GenBank: OQ845903, 893 kDa, pI: 3.83)
Ca5609 (GenBank: OQ845905, > 1018.4 kDa)
100 amino acids
Signal peptide
Choice-of-Anchor A domain (CAA)
Myxococcus cysteine-rich repeat
PTS-rich repeat
acidic, cysteine-rich domain
cysteine-rich repeats
von Willebrand factor A-like domain
PTGD-rich domain
PTSG-repeat domain
PTSA/G-repeat domain
DUF_1
DUF_2
break in contig

Fig. 1 Schematic primary structures of the proteins identified in the *Craspedostauros australis* adhesive trails. For reasons explained in the text, the gene model of Ca5609 is tentative. The molecular mass (in kDa) and isoelectric point (pI) of each protein were calculated from the polypeptide sequence lacking the predicted signal peptide. CAA-domain, Choice-of-Anchor A-domain; DUF, domain of unknown function; PTSA/G-rich, proline, threonine, serine, alanine/glycine-rich; PTSGD-rich, proline, threonine, serine, glycine and aspartic acid-rich; PTS-rich, proline, threonine, serine-rich.

differences in the chemical composition of the LMW components from both diatom species (Fig. S1a).

To identify proteins in the AM, we performed a proteomics analysis of two different samples, hydroxylamine solubilized adhesive material (1) before and (2) after hydrofluoric acid (HF) treatment, which was used to remove glycan moieties that may impede protein sequencing. The proteomics analysis resulted in the identification of eight proteins (Fig. 1; Table S2). Among these was CaFAP1, which was recently shown to be a cell surface glycoprotein that is sloughed off the cell wall during gliding and remains associated with the adhesive trails (Poulsen *et al.*, 2023). The presence of CaFAP1 in the AM was therefore expected and served as a positive control for the identification of adhesive trail proteins. Four of the eight proteins contain 'choice-of-A-anchor' (CAA) domains and were named Trailins, because immunolabelling confirmed their presence in the adhesive trails (see 'Trailin immunolocalization' in the Results section).

Determining the full-length sequences of CaTrailin4 and Ca5609 from the short-read genome assembly (Poulsen *et al.*, 2023) was not possible as the two genes were truncated at their 5′-ends within regions encoding long repetitive stretches. Therefore, we combined Pacific Biosciences (PacBio, Menlo Park, CA, USA) Single Molecule Real-Time long-read sequencing with the Illumina reads to create a new high-quality genome assembly, which spans 88 contigs with an N50 of 1724 159 bp and a BUSCO completeness score of 97% (Fig. S2; Tables S3, S4). A single continuous 31 kb read allowed for the manual prediction of the CaTrailin4 gene model, which is 28 951 bp long and predicted to be devoid of introns (Figs S3, S4). The Ca5609 gene model is split within the tandem repeat region across two different contigs, as not a single PacBio read fully spans the genomic region around this gene (Fig. S5). Therefore, the precise length of the tandem repeat region remains unknown.

To detect individual proteins in the AM, we raised peptide antibodies against CaTrailin-2, -3 and -4. Gel filtration chromatography revealed that the majority of the hydroxylamine solubilized AM exhibited molecular masses > 450 kDa (i.e. the molecular mass of the largest standard protein; Fig. S1b). Dot blot analysis with the antibodies demonstrated that CaTrailins-2, -3 and -4 were each present in these very high-molecular-mass fractions (Fig. S1b,c), yet only CaTrailin4 (894 kDa) was expected to elute in this fraction. This indicates that CaTrailin3 (95 kDa) and CaTrailin4 (163 kDa) contain either extensive post-translational modifications and/or are engaged in supramolecular complexes.

Primary structures of the diatom adhesive trail proteins

Sequence analysis of the eight AM proteins (Fig. 1; Table S2) revealed that each contains a predicted N-terminal signal peptide, as is expected for secreted proteins. None of the proteins contained a GDPH-domain, which was the most abundant diatom-specific sequence feature in the *A. coffeaeformis* adhesive proteins (Lachnit *et al*., 2019). Instead, other protein domains were present as described in the following.

Proline-Threonine-Serine (PTS)-rich domains CaFAP1, CaTrailin4 and Ca5609 contain numerous PTS repeats that constitute *c*. 23%, *c*. 70% and *c*. 50%, respectively, of their sequence. PTS repeats are a hallmark feature of highly glycosylated extracellular proteins, such as mucins and hydroxyproline-rich-glycoproteins (Perez-Vilar & Hill, 1999; Mathieu-Rivet *et al*., 2020). CaFAP1 was shown to be glycosylated (Chiovitti *et al*., 2003). Previously, we demonstrated that the crude AM contains *c*. 70% carbohydrate (Poulsen *et al*., 2014); therefore, we expect CaTrailin4 and Ca5609 to be extensively O-glycosylated within their PTS-rich domains.

Cys-rich repeats Regions rich in cysteine residues are present in three of the AM proteins (CaFAP1, CaTrailin3 and Ca5255). CaFAP1 has a modular structure of alternating PTS-rich and cysteine-rich domains that resembles mucin-like proteins (Poulsen *et al*., 2023). CaTrailin3 contains 11 imperfect cysteine-rich repeats, each with six conserved cysteine residues (Table S5) that do not match any known protein domains, but are found in several other diatoms (including centrics, which are nonmotile) as well as some fungal species (Fig. S6).

Ca5255 contains 12 *Myxococcus* cys-rich repeats/domain of unknown function (DUF4215; IPR011936; Table S5). Myxobacteria are biofilm-forming bacteria that exhibit gliding motility (Nan *et al*., 2010; Faure *et al*., 2016), but it is unknown whether this domain plays a role in surface adhesion.

Other repetitive sequences CaTrailin4 and Ca5609 also contain repetitive sequences that interspace some of the PTS-rich regions, termed domain of unknown function (DUF) 1 and 2, respectively (Fig. 1; Table S5). Both DUFs contain six cysteine residues that might form disulphide bonds to stabilize a particular fold.

Choice-of-Anchor A (CAA) domain/pAdhesive_15 (Interpro: IPR026588; Pfam: PF20597) CAA-domains are present in the four *C. australis* Trailins and were previously identified in three *A. coffeaeformis* AM proteins (Lachnit *et al*., 2019). The term '-choice-of-anchor' refers to the fact that bacterial proteins possessing this domain are surface-exposed proteins, anchored to the cell membrane via one of three conserved transmembrane domain types (LPXTG cell wall anchor, PEP-CTERM domain or type IX secretion system; Xu *et al*., 2004). However, the CAA-domain itself is not the cell membrane anchor. The *C. australis* CAA-domain-bearing proteins do not contain a membrane anchor domain.

Von Willebrand factor A (vWA)-like domain (IPR036465) Ca11384 contains a vWA-like domain. This domain is typically involved in the formation of multiprotein complexes in a wide variety of biological processes, including adhesion and extracellular matrix assembly (Colombatti *et al*., 1993).

Phylogenetic analysis of the adhesive trail proteins

Phylogenetic analysis revealed that outside the known protein domains (vWA-like, CAA, cysteine-rich repeats) and PTS-rich regions, the eight *C. australis* AM proteins are specific to diatoms (Table S6; Fig. S6) and four are found exclusively in *C. australis* (CaFAP1, Ca5255, CaTrailin1 and CaTrailin2). The C-terminal regions of two proteins are specific to raphid pennates (CaTrailin4 and Ca5609) and are conserved among the orders Bacillariales and Mastogloiales but absent from the Naviculales. Ca11384 shows a similar pattern of homology among raphid pennates. The N-terminal domain of CaTrailin3 lacks similarity to pennate sequences but has putative centric homologues. It is notable that none of the proteins share sequence homology along their entire length with proteins from other diatoms or nondiatoms species.

To investigate the evolutionary origin of the diatom CAA-domains, a hidden Markov Model (HMM) profile was constructed from the *C. australis* and *A. coffeaeformis* CAA-domains from the adhesive trail proteins thereby defining a 'diatom CAA-domain' with a length of 292 amino acids. Using this HMM profile, a comprehensive set of homologues from both eukaryotes and prokaryotes was compiled from three protein databases: PLAZA Diatoms 1.0 (Osuna-Cruz *et al*., 2020), NCBI nonredundant database, and the decontaminated MMETSP database (Keeling *et al*., 2014; Van Vlierberghe *et al*., 2021). Combining hits from these sources, we found that CAA-domains are largely restricted to bacteria, rotifers (aquatic invertebrates) and diatoms (Fig. 2a). Among the CAA-containing hits from diatoms, five genes (three from *Fragilariopsis cylindrus*, one from *Pseudo-nitzschia multistriata* and one from *Phaeodactylum tricornutum*) were assessed in a recent study, in which four were shown to be acquired through horizontal gene transfer (HGT) from bacteria (Vancaester *et al*., 2020). Indeed, phylogenetic analysis shows that diatom CAA-domains are nested within bacterial clades, supporting their bacterial origin (Fig. 2a). Notably, diatom

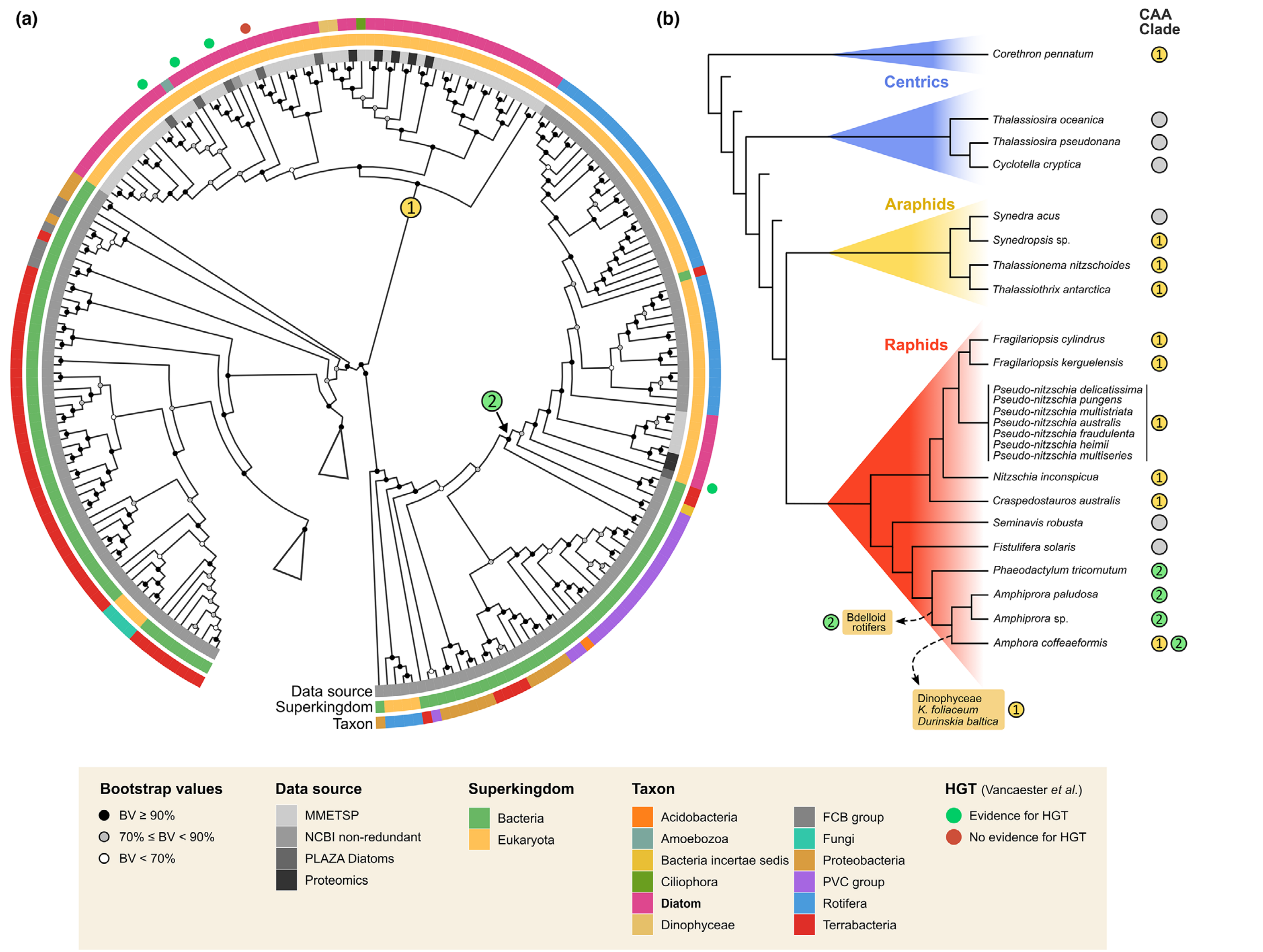

Fig. 2 Evolutionary history of diatom Choice-of-Anchor A (CAA)-domains. (a) Midpoint-rooted maximum-likelihood phylogenetic tree of homologues of the diatom CAA-domain across the tree of life, based on a alignment of the trimmed CAA-domain with a length of 230 amino acids. Node symbols indicate bootstrap support (1000 ultrafast bootstrap repeats). The coloured outer rings indicate the data source and taxonomic binning of each hit. The prediction of possible horizontal gene transfer (HGT) for selected genes from Vancaester *et al*. (2020) is shown by coloured circles outside the phylogeny. Two clades of CAA-domain-containing proteins from bacteria were collapsed for legibility. The two separate clades of diatom CAA-domains are indicated with a coloured circle on the branches. (b) Species tree showing the distribution of diatom CAA-domains in the predicted proteomes of selected species from the major clades of diatoms (indicated with coloured triangles). To the right, the presence of CAA-domains is indicated, specifying the diatom CAA-domain clade (panel a). Dotted arrows show the approximate phylogenetic origin of putative secondary HGT events from diatoms to other taxa (bdelloid rotifers and dinoflagellates). Phylogenetic relationships of diatoms were parsed from Nakov *et al*. (2018). MMETSP, Marine Microbial Eukaryotic Transcriptome Sequencing Project (Keeling *et al*., 2014); NCBI, National Center for Biotechnology Information.

CAA-domains cluster into two distinct clades, which we designated type-1 and type-2 (Figs S7, S8), suggesting two independent HGT events (Fig. 2a). Mapping the occurrence of these two CAA-domains to the diatom species tree (Fig. 2b) reveals that type-1 CAA-domains are distributed across the major clades of diatoms, suggesting that the domain was acquired by a common ancestor of diatoms. Meanwhile, type-2 CAA-domains likely have a more recent origin, as they are restricted to a subclade of raphid diatoms. Interestingly, dinoflagellate CAA-domains (*Kryptoperidinium foliaceum* and *Durinskia baltica*) are nested within the type-1 raphid clade (Fig. 2b). The origin of these genes could be explained by the fact that both species belong to the so-called dinotoms, a select group of dinoflagellates possessing a chloroplast obtained through tertiary endosymbiosis of a diatom (Imanian *et al.*, 2010). Likewise, diatom type-2 CAA-domains were acquired by bdelloid rotifers belonging to the genera *Rotaria*, *Adineta* and *Didymodactylos* (Fig. S8).

3D structure prediction of the diatom CAA-domain

Very recently, a large-scale bioinformatics study identified 24 clusters of bacterial putative adhesive domains (Monzon &

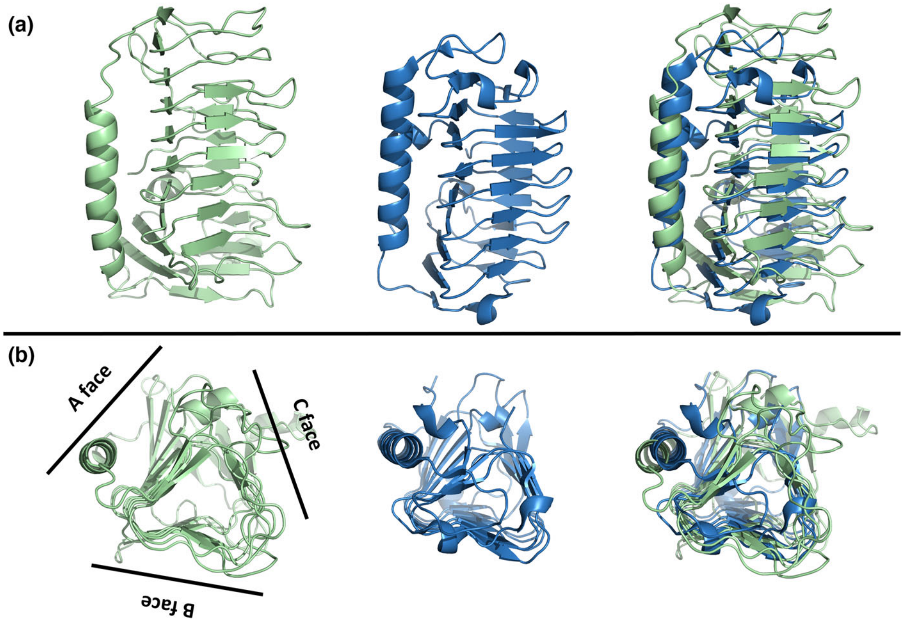

Fig. 3 Relationship between 3D structures of Choice-of-Anchor A (CAA)-domains and ice-binding proteins. (a) Left: predicted structure of the CAA-domain from CaTrailin4; middle: experimentally determined structure of the ice-binding protein *Flavobacterium frigoris* ice-binding protein (FfIBP) (PDB ID 4NU2); right: superposition of the CaTrailin4 CAA-domain (green) and FfIBP (blue). (b) The CaTrailin4 CAA-domain and FfIBP adopt a β-helical fold with three faces.

Bateman, 2022), wherein 'cluster 10' included a representative of a CAA-domain-bearing protein from *Bacillus anthracis* (UniProt: BA_0871; Xu *et al.*, 2004). Using Alphafold, the predicted structure of the CAA-domain (now also termed *pAdhesive_15*) was shown to be similar to an ice-binding protein (mLeIBP) from the fungus *Leucosporidium* sp. (PDB ID 4NUH:A; Monzon & Bateman, 2022). This is a puzzling result in the context of CAA-domain-bearing adhesion proteins from *C. australis*, as this diatom species is not found associated with ice in nature. To further investigate this, we submitted the CAA-domain from CaTrailin4 to Alphafold 2 (Jumper *et al.*, 2021). The predicted structure of the CaTrailin4 CAA-domain is a β-helical fold consisting of two units (see Figs 3, S9) held together by a long alpha helix. Out of the predicted positions of the 4230 atoms, 72% were of high quality and 20% of medium quality, indicating that the predicted structure of the CAA-domain is highly reliable and the predicted topology likely correct (Fig. S9). Only the exact position of a few loops (amino acids 40–50 and 70–75) is uncertain.

The top eight hits from the structural homology search against the experimentally determined 3D structures from the Protein Data Bank (PDB) are listed in Table S7, which align with 70–85% of their residues to the CaTrailin4 CAA with an overall RMSD between 4 and 6 Å. The highest structural similarity is to the ice-binding protein FfIBP from the Antarctic bacterium *Flavobacterium frigoris* (see Fig. 3; Do *et al.*, 2014). Both structures align with 85% of the residues and a root mean square deviation of 5.18 Å. In both, the CaTrailin4 CAA-domain and FfIBP, the β-helical core has three faces and the long alpha helix packs against face A (see Fig. 3b). A sequence representation of the structural alignment highlights the lack of sequence similarity with only 25 residues across the full sequences being identical (see Fig. S10). Yet, 85% of residues are structurally aligned (Fig. S10, shown in red) including the long helix as well as most of the strands making up the three faces. Interestingly, although seven out of these eight structures are ice-binding proteins from bacteria, fungi and diatoms living in cold environments, the one protein that is not ice-binding, ComZ, is from *Thermus thermophilus*, a bacterium with an optimal growth temperature of *c.* 65°C. ComZ is thought to be located on the pilus tip and to play a role in DNA uptake (Salleh *et al.*, 2019).

The strong structural similarity to ice-binding proteins begs the question as whether the diatom CAA-domains are capable of ice-binding. To address this question, we checked the CAA-domain of CaTrailin4 for the presence of ice-binding motifs (IBMs; Do *et al.*, 2014) and ordered surface carbons (OSCs; Doxey *et al.*, 2006), which are predictive of ice-binding regions. The tetrapeptide sequence T-A/G-X-T/N is present at three locations in FfIBP and regarded as an IBM (Do *et al.*, 2014). The three locations align structurally well with the CaTrailin4 CAA-domain, but they are not conserved in sequence, indicating the absence of IBMs. To further corroborate this finding, we screened the CaTrailin4 CAA-domain for OSCs (Doxey *et al.*, 2006), which are specific geometric constellations enabling binding to the highly structured and regular surface of an ice crystal. It was shown that the absolute and relative amount of the protein surface areas covered by OSCs is predictive of ice-binding with values of > 325 Å^2 and > 6%, respectively (Doxey *et al.*, 2006). FfIBP meets these criteria, whereas the CAA-domain of CaTrailin4 does not. Therefore, we assume that CaTraillin4 is highly unlikely to serve the purpose of binding to ice.

Trailin immunolocalization

To gain insight into the Trailins function(s), we analysed their localization on the cell surface, in the adhesive trails, and in the biofilm matrix. Initial attempts to express the Trailins (CaTrailin2 and -3) as GFP-fusion proteins *in vivo* were unsuccessful. The GFP-fusion proteins remained trapped inside the cell (Fig. S11), which is analogous to previous attempts to visualize putative *P. tricornutum* and *A. coffeaeformis* adhesion proteins (Buhmann *et al.*, 2014; Willis *et al.*, 2014; Lachnit *et al.*, 2019).

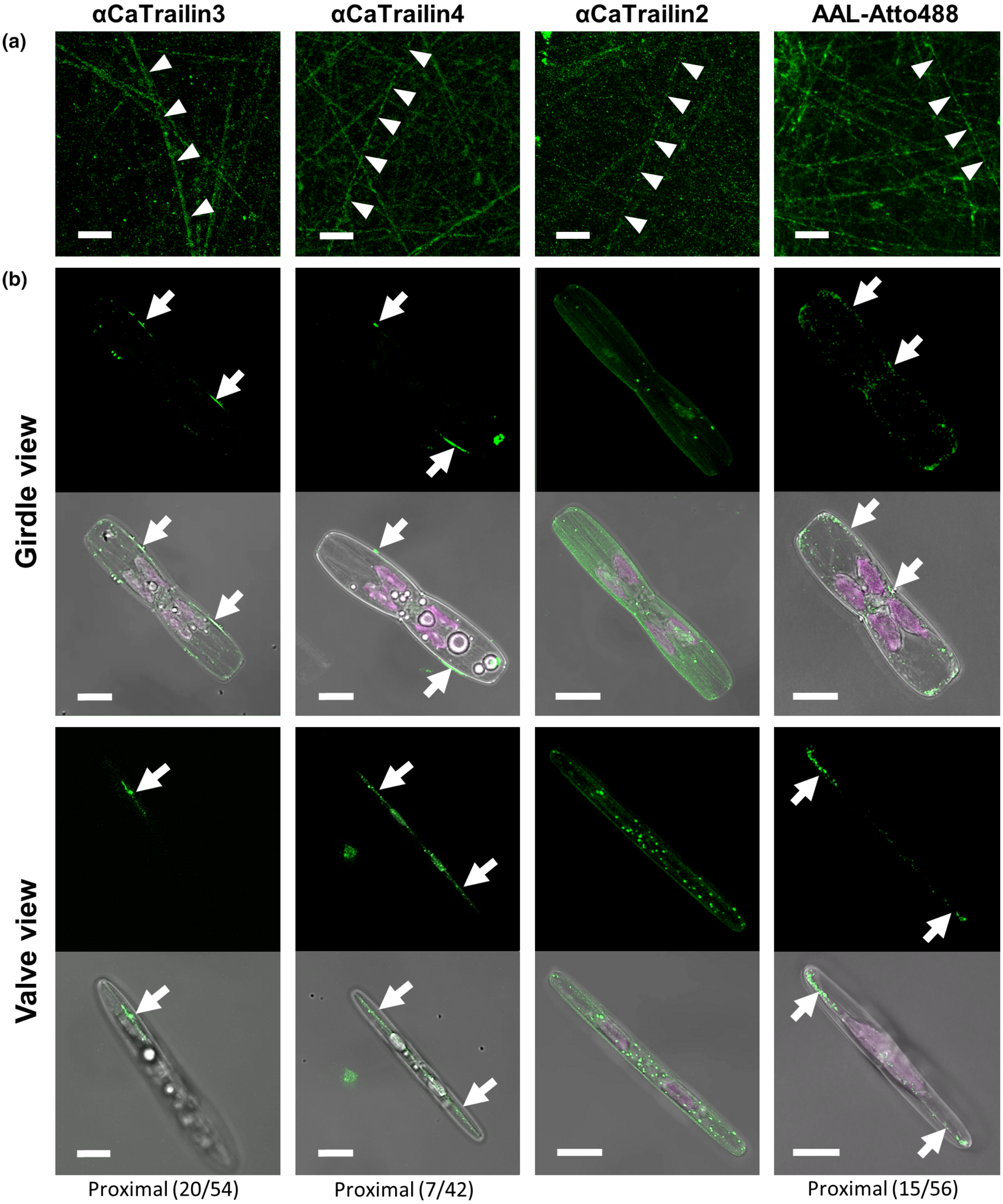

Fig. 4 Localization of adhesive proteins. Immunolabelling of *Craspedostauros australis* (a) adhesive trails and (b) chemically fixed cells with Trailin-specific antibodies αCaTrailin3, αCaTrailin4 and αCaTrailin2, as well as the fucose-specific lectin (*Aleuria aurantia* Lectin (AAL)-Atto488). Arrowheads indicate the position of one continuous adhesive trail and the arrows indicate the position of the raphe. The arrows in both the fluorescence and bright-field images are in the same position. All images are presented as the maximum intensity Z-projection. For valve view, only a part of all Z-stacks was used for maximum intensity projection, and the position and number of *Z*-stacks is indicated below the image. The green colour represents the fluorescence signal from either the Atto488 dye (AAL-Atto488) or the AlexaFluor488 secondary antibody. The magenta colour represents chloroplast autofluorescence. Bars: (a) 20 μm; (b) 10 μm.

Therefore, we pursued immunolocalization studies with peptide antibodies raised against CaTrailin2, -3 and -4. As a control, we have employed the fucose-specific lectin (*Aleuria aurantia*, AAL) that recognizes the *C. australis* adhesive trails (Fig. 4; Neu & Kuhlicke, 2017). All three antibodies (αCaTrailin2, -3 and -4) showed specific labelling of adhesive trails (Fig. 4a). αCaTrailin2

Table 1 Summary of lectin and Trailin immunolabelling during *Craspedostauros australis* biofilm development, which are indicated by their relative abundance (+/++/+++) or absence (–).

		Day 1		Day 4		Day 8		Day 12	
	Raphe localization	Trails	Biofilm	Trails	Biofilm	Trails	Biofilm	Trails	Biofilm
AAL	+	+	–	++	+	+	++	+	+++
CaTrailin2	–	+	–	+	–	+	–	+	–
CaTrailin3	+	+	–	+	–	–	–	–	–
CaTrailin4	+	+	–	++	+	–	–	–	–

and -3 also recognized numerous small particles on the substratum, which may arise from colloidal secreted material that settles from solution (Fig. 4a). The strongest labelling of the adhesive trails was obtained using αCaTrailin4 (Fig. 4a; Table 1). Control experiments performed using preimmune serum and secondary antibody alone showed no labelling of the adhesive trails (Fig. S12).

To determine whether the Trailins are secreted together with the adhesive material through the raphe slit, their location on the cell surface was determined. Like the trail-specific lectin AAL, CaTrailin3 and -4 were located only at the raphe slits, whereas CaTrailin2 was present throughout the cell surface but notably absent from the raphe slits (Fig. 4b). The CaTrailin2 location is similar to the cell surface protein CaFAP1, which becomes secondarily associated with the trails after being secreted from the cell through a raphe-independent mechanism (Poulsen *et al*., 2023).

There is previous circumstantial evidence that the biomolecular composition of the primary adhesive trails and extracellular matrix of biofilms differ (Smith & Underwood, 1998; Wetherbee *et al*., 1998; Underwood *et al*., 2004; Tong & Derek, 2021). However, the direct visualization of proteins in the diatom biofilm matrix and how these change over time has not yet been accomplished. Here, we used immunolabelling with the αCaTrailin2, -3 and -4 antibodies in combination with confocal laser scanning microscopy (CLSM) to analyse their distribution at different stages of biofilm development. To visualize the development of the biofilm, we also employed the fluorescently labelled lectin AAL, which binds to fucose-bearing biopolymers both in the adhesive trails (see Fig. 4a) and the biofilm matrix (Fig. 5). After Day 1, AAL-Atto488 labelling revealed numerous, overlapping but discrete straight and curved lines, which are the primary adhesive trails deposited on the substratum by moving diatoms (Fig. 5a,b). The trails generate a meshwork of *c*. 6 μm thickness with occasional patches of globular EPS material deposited on top of the trails or around the cells (Fig. 5b). After Day 4, the density of the trail meshwork and the globular EPS material had markedly increased generating a *c*. 10 μm thick layer (Fig. 5e,f). In many places, globular EPS material piled up on top of this layer reaching *Z*-heights of up to 20 μm (Fig. 5e,f). After Day 8, the morphology of the biofilm changed and became dominated by a narrow meshwork of rather short fibres while the long adhesive trails seemed to be absent (Fig. 5i,j). Large EPS agglomerates were abundant throughout the biofilm, and most cells are found in the upper biofilm layer (Fig. 5i,j). At Day 12, the biofilm was still composed of a narrow meshwork of short fibres and extensive EPS aggregates up to 30 μm in *Z*-height and tens of micrometre wide in the X-Y direction (Fig. 5m,n). Both cell-free and cell-bearing EPS aggregates were observed (Fig. 5m,n).

The labelling with AAL-Atto488 showed that fucose-bearing biopolymers are seemingly continuously produced at all stages of biofilm development. To investigate the contribution of Trailins to biofilm development, we performed separate immunolabelling experiments with αCaTrailin4 (Fig. 5; Table 1), αCaTrailin2 and αCaTrailin3 (Fig. S12; Table 1). This demonstrated that the primary adhesive trails were recognized by all three antibodies on Days 1 and 4 (Figs 5c,d,g,h, S12a–h). At Day 4, the results with the αCaTrailin4 labelling were very similar to the AAL lectin, showing adhesive trails and globular agglomerates (Fig. 5g,h), whereas the labelling with αCaTrailin2 and αCaTrailin3 showed mainly adhesive trails (Fig. S12e–h). In striking contrast to the AAL-Atto488 labelling, none of the antibodies labelled the biofilm matrix on Days 8 and 12 (Figs 5k,l,o,p, S12i–p). Occasionally, trail-like labelling patterns were observed with αCaTrailin2 on Day 8, but the labelling was much weaker and infrequent when compared to the previous days (Fig. S12k,l). We were concerned that the absence of antibody labelling on Days 8 and 12 might be an artefact caused by the gelatinous nature of the biofilm, slowing down diffusion of the antibodies thereby preventing their binding to the trails on the underlying substratum. To investigate the ability of IgG antibody molecules to diffuse through the biofilm, Protein G beads (*c*. 1 μm) that bind IgG molecules were seeded together with the diatoms on a fresh substratum. After 12 d, the secondary antibody (i.e. labelled IgG molecules) was readily able to diffuse through the entire biofilm and bind to the Protein G beads on the substratum (Fig. S13). This demonstrated that the absence of CaTrailins 2–4 in the Days 8 and 12 biofilms is not caused by their restricted accessibility for the antibody. Instead, we conclude that after Day 4 the production of CaTrailins decreases and ceases by Day 8, and within the same time period, the proteinaceous component of the CaTrailin containing primary trails becomes proteolytically degraded. A final control experiment using the preimmune serum revealed that the strong green fluorescence of cells seen in Days 8 and 12 is a result of cell death and not due to binding of the primary or secondary antibody (Fig. S14). We observed that upon cell death the chloroplast autofluorescence shifts towards the blue spectrum and is visible in the green channel.

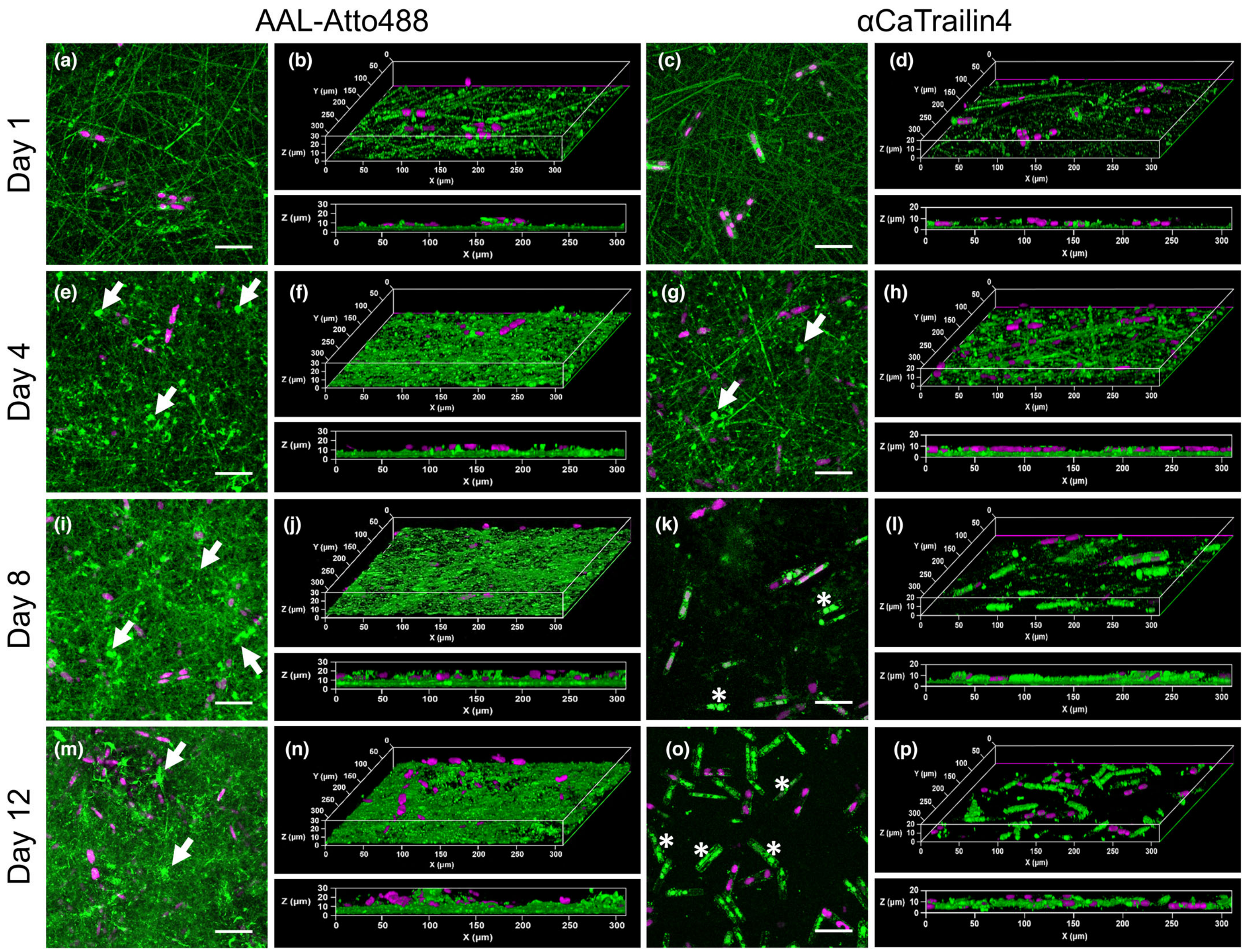

Fig. 5 Development of the *Craspedostauros australis* biofilm matrix. All images are from confocal microscopy and presented as the maximum intensity Z-projection. (a, c, e, g, i, k, m, o) Top view and (b, d, f, h, j, l, n, p) oblique and side-on view of the confocal fluorescence images obtained by probing the biofilm with *Aleuria aurantia* Lectin (AAL)-Atto488 (left) and αTrailin4 antibodies (right). The days indicated on the left margin state the time after seeding the cells on the surface, at which the labelling of the surfaces with the probes indicated on the top margin was performed. White arrows depict some of the agglomerates of adhesive material. Green colour represents the fluorescence signal from AlexaFluor488 secondary antibody or Atto488 conjugate and magenta colour chloroplast autofluorescence. Some dead cells are indicated by an asterisk. Bars: 50 µm, 3D reconstruction: 300 µm (X) × 300 µm (Y) × 20 µm (Z).

Discussion

In this study, we describe the discovery, structure and localization of novel adhesive trail proteins involved in gliding motility of the model adhesion diatom *C. australis*. Combined with the recently described adhesive trail proteins from *A. coffeaeformis* (Lachnit *et al*., 2019), our studies reveal that each species contains a unique set of adhesive trail proteins that are modular and contain PTS-rich regions and the presence of CAA and/or GDPH-domains. Many extracellular matrix (ECM) proteins are composed of multiple protein domains that have evolved through exon/domain-shuffling and resulted in the creation of hundreds of ECM proteins (Engel, 1996; Hynes, 2012). This complexity often makes it challenging to accurately identify homologous proteins when relying on BLAST searches and some caution must be taken when inferring their function. The diatom adhesive proteins described in this study are no exception; although they contain some known protein domains, the regions outside of these domains are either diatom-specific or even species-specific and their exact functions are not yet fully understood.

Using long-read PacBio genome sequencing, we discovered extremely large (> 1 MDa), repetitive, PTS-rich proteins that we suspect to be highly glycosylated as the isolated crude adhesive material contains roughly 70% (w/w) carbohydrates (Poulsen *et al*., 2014). The occurrence of extremely large, repetitive proteins in underwater adhesives is not unique to diatoms. For example, Stewart and co-workers identified a *c*. 650 kDa adhesive silk protein, H-fibroin, from caddisflies (Frandsen *et al*., 2019). The use of long-read sequencing technologies in the latter study and our present work highlights the potential of these technologies to identify and characterize high-molecular-weight, complex adhesive proteins that have evaded detection using traditional sequencing technologies.

Here, we report the finding that diatoms have acquired CAA-domains from bacteria through horizontal gene transfer (HGT). Throughout their evolutionary history, HGT has allowed diatoms to expand their ability to adapt to different ecological niches. (Bowler *et al*., 2008; Vancaester *et al*., 2020; Dorrell *et al*., 2021). For example, the acquisition of bacterial genes and their integration in several metabolic pathways such as the ornithine-urea cycle and vitamin B12 uptake has contributed to their ability to outcompete other phytoplankton (Bowler *et al*., 2008; Vancaester *et al*., 2020; Dorrell *et al*., 2021). Our observation that the CAA-domain was acquired at least twice independently and was widely retained across diatoms suggests these genes are under purifying selection pressure and potentially confer a functional advantage. Coupling the abundance of CAA-domains in the adhesive material of benthic diatoms with their presumed role in adhesion in their bacterial ancestors suggests that the CAA-domain offers novel adhesive functionalities to diatoms.

Recent studies suggest that HGT is continuously occurring in diatoms and that environmental pressures and ecological niche adaption result in a high rate of HGT loss (Dorrell *et al*., 2021, 2023). Of particular note is the recent description of Antarctic vs Arctic clades of diatom IBPs that suggests two independent HGT events for these two Polar Regions (Dorrell *et al*., 2023). This genome plasticity may account for our inability to identify a single 'universal' adhesive protein, as species-specific adhesive proteins can evolve through HGT acquisitions and losses in their local habitat. As the CAA-domain is present in numerous, but not all raphid diatoms, it is not essential for gliding motility. In this context, it is interesting to note that the diatom-specific GDPH-domain, which we previously identified in the *A. coffeaeformis* adhesive trail proteins (Lachnit *et al*., 2019), is present in many but not all raphid diatom species. The GDPH-domain is absent from raphid diatoms that contain proteins with a type-1 CAA-domain, while raphid diatoms that possess GDPH-domain-bearing proteins either lack proteins with CAA-domain or contain type-2 CAA-domain-bearing proteins. This observation might indicate that diatom gliding can be accomplished by three types of protein assemblies that contain: type-2 CAA-domains and GDPH-domains; type-1 CAA-domains; or only GDPH-domains. Furthermore, the presence of the type-1 CAA-domain in a single centric diatom (*Corethron*) and some araphid species suggests that in nonmotile species the CAA-domain may play a role in other adhesive functions such as mucilage pads, chain formation or cell aggregation (Edgar & Pickett-Heaps, 1984; Hoagland *et al*., 1993; Thornton, 2002).

The Alphafold protein structure prediction revealed that while the CAA-domain shares structural similarity with bacterial and eukaryotic ice-binding proteins and the ComZ pilus protein from thermophilic bacteria, there is no sequence similarity. It may seem paradoxical that this protein structure is present in both psychrophilic and thermophilic organisms as well as diatoms from temperate habitats, yet all these organisms share the ability to adhere to surfaces underwater. Notably, the sea-ice diatom *Fragilariopsis cylindrus* possesses both true ice-binding proteins (IBP; > 50 proteins; IPR021884) that protect them from freezing and also serve to attach cells to ice (Krell *et al*., 2008; Raymond & Kim, 2012; Mock *et al*., 2017; Dorrell *et al*., 2023) as well as proteins bearing the type-1 CAA-domain. In contrast, *C. australis* contains no proteins with a 'true' IBP domain, but encodes six CAA-domain-bearing proteins, of which four were identified here in the adhesive material. While the exact ice-binding mechanism of proteins that possess this β-solenoid protein structure is unknown, it has been suggested that they are able to order surface-associated water molecules, which are fine-tuned to merge with the water molecules at the ice-water interface (Yamauchi *et al*., 2020; Khan *et al*., 2021). Although the diatom CAA-domains exhibit an almost identical β-solenoid protein structure to ice-binding proteins, they are not predicted to be ice-binding as they lack the canonical threonine-rich ice-binding sites on the flat surface of one of the β-sheets. Nevertheless, this highly conserved protein structure might endow these diatom proteins with the ability to order water molecules allowing for the alignment of protein surface and the interfacial water molecules on any surface underwater. Future studies using recombinant CAA-domains are needed to test the hypothesis that they play a role in the underwater adhesion of diatoms.

One of the most puzzling aspects of the current model for diatom gliding is that the adhesive material secreted through the raphe slits functions as both an adhesive and a transducer of the actomyosin-generated force from inside the cell to the substratum. To achieve this dual function, the adhesive material needs to reach from the plasma membrane through the raphe slit to the underlying substratum, which equals a distance ≥ 1 μm in *C. australis* and many other diatoms. It was previously suggested that proteins with high tensile strength and elasticity could fulfil this function (Dugdale *et al*., 2006), prompting us to speculate that the extremely large, modular proteins in the *C. australis* adhesive trails, CaTrailin4 and Ca5609, might be such proteins. The repetitive PST-rich domains in these proteins are predicted to be intrinsically disordered (Table S2), and we regard it very likely that they are highly O-glycosylated, like many other PTS-rich extracellular proteins from algae (Hallmann, 2003; Tatli *et al*., 2018; Mathieu-Rivet *et al*., 2020). The glycan moieties may be highly negatively charged due to the presence of uronic acids, which are abundant in the *C. australis* adhesive material (Poulsen *et al*., 2014). A high negative charge density would, due to electrostatic repulsion, stretch out the PTS polypeptide regions, which in the case of CaTrailin4 could span *c*. 2.0 μm (*c*. 6500 aa, 3.5 Å per amino acid (Ainavarapu *et al*., 2007)). This distance is in fact very close to the measured length of EPS strands from *C. australis* (3.5 ± 1.2 μm; Higgins *et al*., 2003) and *Nitzschia communis* (Gutierrez-Medina *et al*., 2022). Recently, a 1.5 MDa adhesive protein (*Mp*IBP) from an arctic bacterium (*Marinomonas primoryensis*) was shown to be 600 nanometres long, with an exceptionally long extender region (RII) composed of *c*. 120 tandem Ig-like domains that serves to project the adhesive regions of the protein into the medium (Guo *et al*., 2017). Furthermore, it has also been demonstrated that the 'elastic reach' of cell adhesion molecules (i.e. the distance between the two contact sites that the protein can be displaced without breaking) can depend on the number of tandem IgG repeats (Carl *et al.*, 2001). The repetitive tandem repeats in CaTrailin4 (DUF1) and Ca5609 (DUF2) might serve an analogous role enabling the reversible mechanical extensibility of these proteins. To investigate the structure–property relationship in CaTrailin4 and Ca5609, the native proteins need to be isolated from the solubilized adhesive material whose preparation was established in the present work.

From the moment a surface is submerged in an aquatic habitat, biofilm-forming organisms will attach and commence the production of a 3D extracellular matrix (Callow & Callow, 2011; de Carvalho, 2018). While such biofilms have important roles in biogeochemical cycling and ecological functions of benthic communities, the accumulation of microbial biofilms (termed biofouling) on manufactured structures (e.g. ship hulls, piping and aquaculture nets) continues to be a very costly problem. A 'heavy' biofilm, which describes the condition where the underlying paint colour is difficult or impossible to determine, increases the drag forces on a ship's hull up to 18%, resulting in enhanced fuel consumption (Schultz *et al.*, 2011). Mitigating the detrimental effects (e.g. corrosion and increased CO_2 emissions) and costs associated with biofouling requires an understanding of the adhesive biomolecules to develop a targeted approach to prevent initial adhesion. In this study, we have demonstrated that there appears to be a transition in the composition of the biomolecules required for initial adhesion to those present in the mature biofilm matrix. The components of the primary adhesives, particularly the widespread CAA-domain, might be suitable targets for the development of antifouling coatings that prevent the adhesion of a wide variety of diatoms and possibly also bacteria. The discoveries described here are the fundament to achieve a detailed mechanistic understanding of how diatoms accomplish their remarkable underwater adhesion and establish a biofilm community.

Acknowledgements

We thank Jennifer Klemm for technical support, Stefan Görlich for his initial assistance with the structural protein analysis, Stefan Golfier for critically reading the manuscript, and the CMCB Technology Platform at TU Dresden (Light Microscopy Facility, Molecular Analysis/Mass Spectrometry Facility) and the Dresden-concept Genome Center. This work was supported by grants from the Deutsche Forschungsgemeinschaft (PO 2256/1-1) to NP, (KR 1853/9-1) to NK. GB is a postdoctoral fellow supported by Fonds Wetenschappelijk Onderzoek (FWO, 1228423 N). KV and WV want to acknowledge the funding obtained by the BOF project GOA01G01323. The Molecular Analysis/Mass Spectrometry Facility was supported by grants from the Deutsche Forschungsgemeinschaft (INST 269/731-1 FUGG), from the German Federal Ministry of Education and Research (#03Z22EB1), and the European Regional Development Fund (ERDF/EFRE, Contract #100232736). Open Access funding enabled and organized by Projekt DEAL.

Competing interests

None declared.

Author contributions

NP and NK conceived the project. NP, NK and JZS designed research. NP, BR, JZS, GB, AF, AS and MP performed experiments. NP, JZS, BR, GB, AF, MS, MP, AS, WV, KV and NK analysed the data. NP, NK, JZS, GB and MS wrote the manuscript. All authors contributed to the corrections of the manuscript.

ORCID

Gust Bilcke https://orcid.org/0000-0002-9499-2295
Ali Fatlawi https://orcid.org/0000-0002-0788-2363
Nils Kröger https://orcid.org/0000-0002-8115-4129
Martin Pippel https://orcid.org/0000-0002-8134-5929
Nicole Poulsen https://orcid.org/0000-0002-4533-8860
Beata Romanowska https://orcid.org/0000-0002-1372-1101
Alastair Skeffington https://orcid.org/0000-0002-5645-8120
Klaas Vandepoele https://orcid.org/0000-0003-4790-2725
Wim Vyverman https://orcid.org/0000-0003-0850-2569
Jirina Zackova Suchanova https://orcid.org/0000-0003-1484-0599

Data availability

The data that support the finding of this study are available in the Methods S1 of this article. The proteomics data are deposited on PRIDE (accession: PXD039465). The sequencing data of this study are available through NCBI (Bioproject: PRJNA850956). The genomic DNA and protein sequences have been deposited at GenBank: CaTrailin1: OQ845901; CaTrailin2: OQ845900; CaTrailin3: OQ845899; CaTrailin4: OQ845903; CaFAP1: OQ845902; Ca11384: OQ845904; Ca5609: OQ845905; and Ca5255: OQ845906.

References

Ainavarapu RK, Brujic J, Huang HH, Wiita AP, Lu H, Li LW, Walther KA, Carrion-Vazquez M, Li HB, Fernandez JM. 2007. Contour length and refolding rate of a small protein controlled by engineered disulfide bonds. *Biophysical Journal* 92: 225–233.

Ashworth MP. 2013. *Rock snot in the age of transcriptomes: using a phylogenetic framework to identify genes involved in diatom extracellular polymeric substance-secretion pathways.* PhD thesis. Austin, TX, USA: University of Texas at Austin.

Bowler C, Allen AE, Badger JH, Grimwood J, Jabbari K, Kuo A, Maheswari U, Martens C, Maumus F, Otillar RP *et al.* 2008. The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes. *Nature* 456: 239–244.

Buhmann MT, Poulsen N, Klemm J, Kennedy MR, Sherrill CD, 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.

Callow JA, Callow ME. 2011. Trends in the development of environmentally friendly fouling-resistant marine coatings. *Nature Communications* 2: 244.

Carl P, Kwok CH, Manderson G, Speicher DW, Discher DE. 2001. Forced unfolding modulated by disulfide bonds in the Ig domains of a cell adhesion molecule. *Proceedings of the National Academy of Sciences, USA* **98**: 1565–1570.

de Carvalho CCCR. 2018. Marine biofilms: a successful microbial strategy with economic implications. *Frontiers in Marine Science* 5: 126.

Chiovitti A, Bacic A, Burke J, Wetherbee R. 2003. Heterogeneous xylose-rich glycans are associated with extracellular glycoproteins from the biofouling diatom *Craspedostauros australis* (Bacillariophyceae). *European Journal of Phycology* **38**: 351–360.

Colombatti A, Bonaldo P, Doliana R. 1993. Type A modules: interacting domains found in several non-fibrillar collagens and in other extracellular matrix proteins. *Matrix* **13**: 297–306.

Do H, Kim SJ, Kim HJ, Lee JH. 2014. Structure-based characterization and antifreeze properties of a hyperactive ice-binding protein from the Antarctic bacterium *Flavobacterium frigoris* PS1. *Acta Crystallographica Section D: Structural Biology* **70**: 1061–1073.

Dorrell RG, Kuo A, Fussy Z, Richardson EH, Salamov A, Zarevski N, Freyria NJ, Ibarbalz FM, Jenkins J, Pierella Karlusich JJ *et al.* 2023. Convergent evolution and horizontal gene transfer in Arctic Ocean microalgae. *Life Science Alliance* **6**: e202201833.

Dorrell RG, Villain A, Perez-Lamarque B, Audren de Kerdrel G, McCallum G, Watson AK, Ait-Mohamed O, Alberti A, Corre E, Frischkorn KR *et al.* 2021. Phylogenomic fingerprinting of tempo and functions of horizontal gene transfer within ochrophytes. *Proceedings of the National Academy of Sciences, USA* **118**: e2009974118.

Doxey AC, Yaish MW, Griffith M, McConkey BJ. 2006. Ordered surface carbons distinguish antifreeze proteins and their ice-binding regions. *Nature Biotechnology* **24**: 852–855.

Dugdale TM, Willis A, Wetherbee R. 2006. Adhesive modular proteins occur in the extracellular mucilage of the motile, pennate diatom *Phaeodactylum tricornutum. Biophysical Journal* **90**: L58–L60.

Edgar LA. 1983. Mucilage secretions of moving diatoms. *Protoplasma* **118**: 44–48.

Edgar LA, Pickett-Heaps J. 1984. Diatom locomotion. In: Round FE, Chapman DJ, eds. *Progress in phycological research.* Bristol, UK: Biopress Ltd, 47–88.

Edgar LA, Zavortink M. 1983. The mechanism of diatom locomotion. II: identification of Actin. *Proceedings of the Royal Society of London. Series B, Biological Sciences* **218**: 345–348.

Engel J. 1996. Domain organizations of modular extracellular matrix proteins and their evolution. *Matrix Biololgy* **15**: 295–299.

Faure LM, Fiche JB, Espinosa L, Ducret A, Anantharaman V, Luciano J, Lhospice S, Islam ST, Treguier J, Sotes M *et al.* 2016. The mechanism of force transmission at bacterial focal adhesion complexes. *Nature* **539**: 530–535.

Flemming HC, Wingender J. 2010. The biofilm matrix. *Nature Reviews Microbiology* **8**: 623–633.

Flemming HC, Wuertz S. 2019. Bacteria and archaea on Earth and their abundance in biofilms. *Nature Reviews Microbiology* **17**: 247–260.

Frandsen PB, Bursell MG, Taylor AM, Wilson SB, Steeneck A, Stewart RJ. 2019. Exploring the underwater silken architectures of caddisworms: comparative silkomics across two caddisfly suborders. *Philosophical Transactions of the Royal Society B: Biological Sciences* **374**: 20190206.

Guo S, Stevens CA, Vance TDR, Olijve LLC, Graham LA, Campbell RL, Yazdi SR, Escobedo C, Bar-Dolev M, Yashunsky V *et al.* 2017. Structure of a 1.5-MDa adhesin that binds its Antarctic bacterium to diatoms and ice. *Science Advances* **3**: e1701440.

Gutierrez-Medina B, Pena Maldonado AI, Garcia-Meza JV. 2022. Mechanical testing of particle streaming and intact extracellular mucilage nanofibers reveal a role of elastic force in diatom motility. *Physical Biology* **19**: 56002.

Hallmann A. 2003. Extracellular matrix and sex-inducing pheromone in Volvox. *International Review of Cytology* **227**: 131–182.

Harrison PJ, Waters RE, Taylor FJR. 1980. A broad-spectrum artificial seawater medium for coastal and open ocean phytoplankton. *Journal of Phycology* **16**: 28–35.

Higgins MJ, Crawford SA, Mulvaney P, Wetherbee R. 2000. The topography of soft, adhesive diatom 'trails' as observed by atomic force microscopy. *Biofouling* **16**: 133–139.

Higgins MJ, Molino P, Mulvaney P, Wetherbee R. 2003. The structure and nanomechanical properties of the adhesive mucilage that mediates diatom-substratum adhesion and motility. *Journal of Phycology* **39**: 1181–1193.

Hoagland KD, Rosowski JR, Gretz MR, Roemer SC. 1993. Diatom extracellular polymeric substances – function, fine-structure, chemistry, and physiology. *Journal of Phycology* **29**: 537–566.

Holland R, Dugdale TM, Wetherbee R, Brennan AB, Finlay JA, Callow JA, Callow ME. 2004. Adhesion and motility of fouling diatoms on a silicone elastomer. *Biofouling* **20**: 323–329.

Hynes R. 2012. The evolution of metazoan extracellular matrix. *Journal of Cell Biology* **196**: 671–679.

Imanian B, Pombert JF, Keeling PJ. 2010. The complete plastid genomes of the two 'Dinotoms' *Durinskia baltica* and *Kryptoperidinium foliaceum. PLoS ONE* **5**: e10711.

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Zidek A, Potapenko A *et al.* 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**: 583–589.

Keeling PJ, Burki F, Wilcox HM, Allam B, Allen EE, Amaral-Zettler LA, Armbrust EV, Archibald JM, Bharti AK, Bell CJ *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.

Khan NMU, Arai T, Tsuda S, Kondo H. 2021. Characterization of microbial antifreeze protein with intermediate activity suggests that a bound-water network is essential for hyperactivity. *Scientific Reports* **11**: 5971.

Krell A, Beszteri B, Dieckmann G, Glockner G, Valentin K, Mock T. 2008. A new class of ice-binding proteins discovered in a salt-stress-induced cDNA library of the psychrophilic diatom *Fragilariopsis cylindrus* (Bacillariophyceae). *European Journal of Phycology* **43**: 423–433.

Lachnit M, Buhmann MT, Klemm J, Kröger N, Poulsen N. 2019. Identification of proteins in the adhesive trails of the diatom *Amphora coffeaeformis. Philosophical Transactions of the Royal Society B: Biological Sciences* **374**: 20190196.

Lind JL, Heimann K, Miller EA, vanVliet C, Hoogenraad NJ, Wetherbee R. 1997. Substratum adhesion and gliding in a diatom are mediated by extracellular proteoglycans. *Planta* **203**: 213–221.

Mathieu-Rivet E, Mati-Baouche N, Walet-Balieu ML, Lerouge P, Bardor M. 2020. N- and O-glycosylation pathways in the microalgae Polyphyletic Group. *Frontiers in Plant Science* **11**: 609993.

Mock T, Otillar RP, Strauss J, McMullan M, Paajanen P, Schmutz J, Salamov A, Sanges R, Toseland A, Ward BJ *et al.* 2017. Evolutionary genomics of the cold-adapted diatom *Fragilariopsis cylindrus. Nature* **541**: 536–540.

Monzon V, Bateman A. 2022. Large-scale discovery of microbial fibrillar adhesins and identification of novel members of adhesive domain families. *Journal of Bacteriology* **204**: e0010722.

Nakov T, Beaulieu JM, Alverson AJ. 2018. Accelerated diversification is related to life history and locomotion in a hyperdiverse lineage of microbial eukaryotes (Diatoms, Bacillariophyta). *New Phytologist* **219**: 462–473.

Nan B, Mauriello EM, Sun IH, Wong A, Zusman DR. 2010. A multi-protein complex from *Myxococcus xanthus* required for bacterial gliding motility. *Molecular Microbiology* **76**: 1539–1554.

Neu TR, Kuhlicke U. 2017. Fluorescence lectin bar-coding of glycoconjugates in the extracellular matrix of biofilm and bioaggregate forming microorganisms. *Microorganisms* **5**: 5.

Osuna-Cruz CM, Bilcke G, Vancaester E, De Decker S, Bones AM, Winge P, Poulsen N, Bulankova P, Verhelst B, Audoor S *et al.* 2020. The *Seminavis robusta* genome provides insights into the evolutionary adaptations of benthic diatoms. *Nature Communications* **11**: 3320.

Perez-Vilar J, Hill RL. 1999. The structure and assembly of secreted mucins. *Journal of Biological Chemistry* **274**: 31751–31754.

Pinckney JL. 2018. A mini-review of the contribution of benthic microalgae to the ecology of the continental shelf in the south atlantic bight. *Estuaries and Coasts* **41**: 2070–2078.

Poulsen N, Hennig H, Geyer VF, Diez S, Wetherbee R, Fitz-Gibbon S, Pellegrini M, Kröger N. 2023. On the role of cell surface associated, mucin-like glycoproteins in the pennate diatom *Craspedostauros australis* (Bacillariophyceae). *Journal of Phycology* **59**: 54–69.

Poulsen N, Kröger N, Harrington MJ, Brunner E, Paasch S, Buhmann MT. 2014. Isolation and biochemical characterization of underwater adhesives from diatoms. *Biofouling* **30**: 513–523.

Poulsen NC, Spector I, Spurck TP, Schultz TF, Wetherbee R. 1999. Diatom gliding is the result of an Actin-myosin motility system. *Cell Motility and the Cytoskeleton* **44**: 23–33.

Raymond JA, Kim HJ. 2012. Possible role of horizontal gene transfer in the colonization of sea ice by algae. *PLoS ONE* 7: e35968.

Salleh MZ, Karuppiah V, Snee M, Thistlethwaite A, Levy CW, Knight D, Derrick JP. 2019. Structure and properties of a natural competence-associated pilin suggest a unique pilus tip-associated DNA receptor. *mBio* **10**: e00614-19.

Schultz MP, Bendick JA, Holm ER, Hertel WM. 2011. Economic impact of biofouling on a naval surface ship. *Biofouling* 27: 87–98.

Shevchenko A, Tomas H, Havlis J, Olsen JV, Mann M. 2006. In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nature Protocols* 1: 2856–2860.

Smith DJ, Underwood GJC. 1998. Exopolymer production by intertidal epipelic diatoms. *Limnology and Oceanography* 43: 1578–1591.

Steele DJ, Franklin DJ, Underwood GJC. 2014. Protection of cells from salinity stress by extracellular polymeric substances in diatom biofilms. *Biofouling* **30**: 987–998.

Tatli M, Ishihara M, Heiss C, Browne DR, Dangott LJ, Vitha S, Azadi P, Devarenne TP. 2018. Polysaccharide associated protein (PSAP) from the green microalga *Botryococcus braunii* is a unique extracellular matrix hydroxyproline-rich glycoprotein. *Algal Research–Biomass Biofuels and Bioproducts* **29**: 92–103.

Thornton DCO. 2002. Diatom aggregation in the sea: mechanisms and ecological implications. *European Journal of Phycology* **37**: 149–161.

Tong CY, Derek CJC. 2021. The role of substrates towards marine diatom *Cylindrotheca fusiformis* adhesion and biofilm development. *Journal of Applied Phycology* **33**: 2845–2862.

Underwood GJC, Boulcott M, Raines CA, Waldron K. 2004. Environmental effects on exopolymer production by marine benthic diatoms: dynamics, changes in composition, and pathways of production. *Journal of Phycology* **40**: 293–304.

Van Vlierberghe M, Di Franco A, Philippe H, Baurain D. 2021. Decontamination, pooling and dereplication of the 678 samples of the Marine Microbial Eukaryote Transcriptome Sequencing Project. *BMC Research Notes* **14**: 306.

Vancaester E, Depuydt T, Osuna-Cruz CM, Vandepoele K. 2020. Comprehensive and functional analysis of horizontal gene transfer events in diatoms. *Molecular Biology and Evolution* **37**: 3243–3257.

Wetherbee R, Lind JL, Burke J, Quatrano RS. 1998. The first kiss: establishment and control of initial adhesion by raphid diatoms. *Journal of Phycology* **34**: 9–15.

Willis A, Eason-Hubbard M, Hodson O, Maheswari U, Bowler C, Wetherbee R. 2014. Adhesion molecules from the diatom *Phaeodactylum tricornutum* (Bacillariophyceae): genomic identification by amino-acid profiling and *in vivo* analysis. *Journal of Phycology* **50**: 837–849.

Xu Y, Liang XW, Chen YH, Koehler TM, Hook M. 2004. Identification and biochemical characterization of two novel collagen binding MSCRAMMs of *Bacillus anthracis*. *Journal of Biological Chemistry* **279**: 51760–51768.

Yamauchi A, Arai T, Kondo H, Sasaki YC, Tsuda S. 2020. An ice-binding protein from an Antarctic ascomycete is fine-tuned to bind to specific water molecules located in the ice prism planes. *Biomolecules* **10**: 759.

Supporting Information

Additional Supporting Information may be found online in the Supporting Information section at the end of the article.

Fig. S1 Biochemical characterization of the hydroxylamine solubilized adhesive trails from *Craspedostauros australis*.

Fig. S2 Merqury assembly spectrum plots for evaluating k-mer completeness.

Fig. S3 CaTrailin4 PacBio sequence alignment snapshot.

Fig. S4 Dotplot: Dot plot of the CaTrailin4 region of contig uCraAus1_00004_0_1 (344–373 Kb).

Fig. S5 Ca5609 PacBio sequence alignment snapshot.

Fig. S6 Distribution of Blast hits of *Craspedostauros australis* adhesive proteins across the Eukaryotic tree of life.

Fig. S7 Midpoint-rooted maximum-likelihood phylogenetic tree of homologues of the Diatom CAA-like domains belonging to Clade 1.

Fig. S8 Midpoint-rooted maximum-likelihood phylogenetic tree of homologues of the Diatom CAA-like domains belonging to Clade 2.

Fig. S9 Alphafold structural prediction of the CaTrailin4 CAA-domain.

Fig. S10 Sequence representation of structural alignment of the CaTrailin4 CAA-domain and FfIBP.

Fig. S11 Confocal microscopy images of cell lines expressing the AM proteins CaTrailin2-GFP and CaTrailin3-GFP.

Fig. S12 Immunofluorescence control experiments.

Fig. S13 Monitoring the presence of adhesive proteins within the biofilm by immunolabelling.

Fig. S14 Control live cell imaging experiments.

Methods S1 Supplementary Materials and Methods.

Table S1 Primers used for determining the full-length gene models.

Table S2 Tabular overview of identified protein hits.

Table S3 *Craspedostauros australis* PacBio genome assembly statistics.

Table S4 *Craspedostauros australis* PacBio genome completeness statistics.

Table S5 Clustal Omega sequence alignments.

Table S6 Blastp alignments of *Craspedostauros australis* adhesive proteins.

Table S7 Top eight predictions from the structural homology search.

Please note: Wiley is not responsible for the content or functionality of any Supporting Information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.