New Phytologist Supporting Information

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

Authors: Jirina Zackova Suchanova, Gust Bilcke, Beata Romanowska, Ali Fatlawi, Martin Pippel, Alastair Skeffington, Michael Schroeder, Wim Vyverman, Klaas Vandepoele, Nils Kröger, Nicole Poulsen

Article acceptance date: 2 July 2023

Supplementary Material & Methods

Gel filtration chromatography

The adhesive material (~5.0 mg) was solubilized in 1 mL of 0.5 M borate buffer (pH 8.5) containing 2 M hydroxylamine (Sigma Aldrich) and incubated for 3 h at 45°C. The solubilized material was centrifuged for 10 min at 20,000 xg and then loaded on a HiLoad 16/600 Superdex 200p column (GE Healthcare) that had been pre-equilibrated in 500 mM NaCl/50 mM Tris-HCl pH 7.5 at a flow rate of 1 mL/min. The elution of the adhesive components was monitored at 226 and 280 nm and 2.0 mL fractions were collected. The resulting fractions were pooled and concentrated 8-fold using an Amicon Ultra 4 Centrifugal Filter Unit and then 20 µL of each pooled fraction was analyzed on a 6% Schägger Gel and stained with Stains All. A size calibration was performed under the same conditions using protein molecular weight standards (Serva, Germany, cat. No. 39064).

Anhydrous hydrofluoric acid treatment

To deglycosylate the AM, the freeze-dried hydroxylamine solubilized material was resuspended in 1 mL freshly condensed anhydrous hydrofluoric acid (HF) and incubated on ice for 1 h. The HF was evaporated using a gentle nitrogen stream followed by a final drying step in a UNIVAPO vacuum concentrator centrifuge (25°C, -0.8 bar, 1 h). The dry samples were immediately dissolved in 50 mM ammonium acetate and when necessary, the pH was neutralized with ammonium hydroxide.

LC-MS/MS proteomics analysis

The desalted adhesive material was submitted to Proteomics Facility at Max Planck Institute of Molecular Cell Biology and Genetics in Dresden for reversed phase liquid chromatography nanospray tandem mass spectrometry (LC-MS/MS). The desalted AM was submitted as two samples: HF and non-HF treated. The recovered peptides were dissolved in 20 µL of 5% aqueous formic acid and a 5 µL aliquot was analyzed by GeLC–MS/MS on an Ultimate3000 nanoLC system (Dionex, Amsterdam, The Netherlands) interfaced on-line to a LTQ Orbitrap Velos hybrid tandem mass spectrometer (Thermo Fisher, Bremen, Germany) applying a 180 min elution program. Acquired spectra were searched against 3'-frame translated *C. australis* transcriptome (18699 entries) (Poulsen, 2023) using Mascot software (Matrix Science, London, UK, v.2.2.04). The hits were evaluated by Scaffold program v.4.3.4 (Proteome Software Inc., Portland, OR) and identifications were accepted if two of more peptides could

be established at greater probability than 95% (Peptide Prophet Algorithm) and 99% (Protein Prophet Algorithm) for peptides and proteins, respectively. False Discovery Rate (FDR) was below 1.5% as calculated by Scaffold.

The hydroxylamine-solubilized only AM was submitted to the Mass Spectrometry Facility at Joint Technology Platform of Center for Molecular and Cellular Bioengineering in Dresden. Collected protein material was separated by SDS gel electrophoresis and proteins were digested in gel either with trypsin (Promega, Mannheim, Germany) according (Shevchenko *et al.*, 2006). Digestion was performed in 10 mM ammonium bicarbonate buffer and enzyme concentration was 10 ng/µL. Extracted peptides were dried in vacuum and stored until analysis at -20°C. For analysis by HPLC-MS/MS protein digests were recovered in 3 µl 30% formic acid containing retention time peptide standard (25 fmol/µL, Thermo Pierce) and diluted by addition of 20 µL H_2O. For analysis by HPLC-MS/MS 5 µl were injected into Ultimate3000 nanoLC system (Dionex, Amsterdam, The Netherlands) system equipped with a 75 µm x 2 cm trap column and 75 µm x 15 cm separation column (Acclaim C18 PepMap, Thermo Fisher, Bremen, Germany). Samples were separated in a linear gradient of solvent A (0.1% formic acid) and solvent B (60% acetonitrile, 0.1% formic acid) from 0% B to 55% B in 90 min. The flow rates for sample loading on the trap column was 3 µL/min while the flow rate for separation and analysis was 200 nl/min. The mass spectrometer (LTQ Orbitrap XL ETD, Thermo Fisher, Bremen, Germany) was operated in data dependent analysis (DDA) mode with selection of the 8 most intense ion signals of charge states +2,+3 and +4. MS1 spectra were acquired in the Orbitrap at a nominal resolution of 60000 at 400 Th and MS/MS in the ion trap using CID and a normalized collision energy (NCE) of 35. Acquired spectra were searched against 3'-frame translated *C. australis* transcriptome by Mascot software (Matrix Science, London, UK, v.2.6.0). The hits were evaluated by Scaffold program v.4.8.2 (Proteome Software Inc 2017; Craig & Beavis 2003; Nesvizhskii & Keller 2003) and accepted if two of more peptides were matched under 95% and 99% probability thresholds for peptides and proteins respectively. False Discovery Rate (FDR) was below 1.5% as calculated by Scaffold.

***C. australis* genomic DNA isolation**

HMW genomic DNA (gDNA) isolation, $4x10^8$ cells were resuspended in CTAB extraction solution (AppliChem) containing 2% β-mercaptoethanol (Merck) and incubated for 1 hour at 65°C. The cell debris was removed by centrifugation (3,200 g, 30 min) and the resulting supernatant was treated with RNase A (50 µg/mL, Fermentas) for 1 hour at 37°C and then extracted six times with phenol:chloroform:isoamyl alcohol (25:24:1) and twice with

chloroform:isoamyl alcohol (24:1). To precipitate the DNA, 2.5 volumes of ice-cold ethanol (100%, Merck) and 0.1 volumes 1 M NaCl (Sigma Aldrich) were added and the solution was incubated overnight on ice. The precipitated gDNA was collected by centrifugation (5,000 xg, 10 min, 4°C) and washed twice with 70% ethanol. After drying, the gDNA was dissolved in TE buffer (AppliChem).

C. australis PacBio Genome Assembly

De novo genome assembly was performed with DAmar (https://github.com/MartinPippel/DAmar). This assembler is based on an improved MARVEL assembler (https://github.com/schloi/MARVEL, commit ID: 5e17326) [1] and the integration of parts from the DAZZLER, DALIGNER, DAMASKER, DASCRUBBER, DAZZ_DB, (https://github.com/thegenemyers; [2]) and the DACCORD (version: 0.0.635, [3]) code base. To assemble the genome, we performed the following steps: setup, PacBio read patching, assembly and error-polishing.

a) Setup Phase

PacBio reads were filtered by choosing only the longest read of each zero-mode waveguide (ZMW) and requiring subsequently a minimum read length of 4 kb. The resulting 333 thousand reads (69X coverage) were stored in a DAZZLER database. As the chloroplast and mitochondria reads were strongly overrepresented in the database both organelles were first fully assembled with the DAmar organelle pipeline (see 1e). Afterwards any PacBio read that aligns with the assemblies were filtered out. The reduced dazzler database consists of 263 thousand reads (54X coverage).

b) Read Patching

The patch phase detects and corrects read artefacts including missed adapters, polymerase strand jumps, chimeric reads, and long low-quality read segments that are the primary impediments to long contiguous assemblies. To this end, we first computed local alignments of all raw reads. Because local alignment computation is by far the most time- and storage-consuming part of the pipeline, we reduced runtime and storage by masking repeats in the reads as follows. First, low-complexity intervals, such as microsatellites or homopolymers, were masked with DBdust (https://github.com/thegenemyers/DAZZ_DB). Second, tandem repeats were masked by using datander and TANmask (https://github.com/thegenemyers/DAMASKER). Third, we used a read alignment step to detect repeats. To this end, we first split all reads into groups representing 1× read coverage. For each group, we then aligned all reads against all others in the same group with daligner (https://github.com/thegenemyers/DALIGNER) and masked all local regions in each read

where ≥50 other reads aligned. The repeat masks were subsequently used to prevent k-mer seeding in repetitive regions when computing all local alignments between all reads. Then we applied LAfix to detect and correct read artefacts.

c) *De novo* assembly

In the assembly phase, we first calculated all overlaps between patched reads using the same alignment strategy of the patch phase. The subsequent steps of (i) computing a quality track for all reads, (ii) computing a detailed repeat mask, (iii) filtering overlap piles, (iv) computing the overlap graph, (v) touring the overlap graph to obtain primary contigs, and (vi) base error correction to create consensus sequences for the primary contigs follow the steps of the original MARVEL assembly pipeline [1], [4].

d) Error Polishing

The assembly and the organelle contigs were further polished by using the raw PacBio reads and applying two rounds of gcpp (https://github.com/PacificBiosciences/gcpp) polishing.

To further correct base errors and reduce remaining length errors in homopolymer regions we used merfin (https://github.com/arangrhie/merfin; [5]) based on an Illumina read set. Therefore, we ran gcpp a third time to create a variant (vcf) file. Only remaining homozygous variants (QUAL>1 && (GT="AA" || GT="Aa")), were used for the merfin polishing.

e) Mitochondrion and Chloroplast assemblies

The mitochondrion and chloroplast genomes were assembled with the DAmar organelle assembly pipeline based on the closely related references from *Durinskia baltica* (NCBI accession: NC_014287, ENA accession JN378735). The final circular mitochondrial genome has length of 66,255 bp and the chloroplast genome has a length of 133,148 bp.

The final *Craspedostauros australis* genome assembly was checked for contaminations with blobtoolkit (version 1.1) and an in-house pipeline which screens several BLAST databases. The base accuracy was estimated with merqury (version 1.3) [6] which reported a QV score of 38.7 and a k-mer completeness of 99.2. BUSCO scores were created based on the stramenopiles_odb10 lineage data set (version 5.2.2) [7].

Assembly sequence analysis of region CaTrailin_4 and Ca5609

a) CaTrailin_4

The genomic sequence of protein CaTrailin4 and the raw PacBio reads were aligned with minimap2 (version: 2.24-r1122) to the genome assembly. Tandem repeats were identified with the program Tandem Repeat Finder (TRF, version 4.09) and converted into a bed file format. CaTrailin4 unambiguously mapped to contig 00004_0 (uCraAus1_00004_0_1:344.489-373.438 (-) = 28.950bp). Alignments and repeat annotations were loaded into the Integrative Genomics Viewer (IGV, version 2.12.3) for a further manual inspection.

The mapping location of CaTrailin4 has a complex tandem repeat structure (Figure: CA195_Alignment – TandemRepeatTrack). When looking at the coverage profile (Figure: CA195_Alignment - PacBioCoverage), the alignments show a slight decrease in read coverage - 38.6X over the full contig versus 27.0X over the CaTrailin_4 region. But this is not unexpected for such a complicated repeat region. This is also reflected by a dot plot of the CaTrailin_4 self-alignment (Figure Dotplot).

There is no position in CaTrailin4 that is not supported by PacBio reads, in that case all alignments would break. Furthermore, a single PacBio read (highlighted in blue) fully spans the CaTrailin4 region (Figure: CaTrailin4_Alignment – PacBioAlignments). The long PacBio reads (>31Kb) highlighted in red and green, that could be anchored at both non-repetitive sites adjacent to the CaTrailin4 region, also support the different tandem repeat modules i.e. they span multiple different tandem repeat units (363Kb – 372Kb).

b) Ca5609

The gene Ca5609 is split on two different contigs as not a single PacBio read fully spans the location. All the reads that map to contig uCraAus1_00018_0_1 that reach a bit further into contig uCraAus1_00034_0_1 are fully trapped in the repetitive part of the sequence and cannot be anchored outside the repetitive region.

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

Gene	Primer Name	Sequence
Ca807		
3'RACE	Ca807_1st PCR	GGACGCCAACGGCCTCCCCTC
	Ca807_2nd PCR	GAGGACCTCAGCAAGCAGAAC
Full-length gene model	Ca807_FULL_1f	ACGAACCCCCACTTCCACTTC
	Ca807_FULL_2r	CTCCTCTGGGGAGCATGCTGG
Ca1637		
3'RACE	Ca1637_1st PCR	ATGCAGGTCGGCAACACCACC
	Ca1637_2nd PCR	CGGAGGCCACCTTCACCTAC
Full-length gene model	Ca1637_FULL_1f	ATTCGGCCAACCGCAACCCAC
	Ca1637_FULL_2r	CAGCCGTAGACACTGCACTTG
Ca5255		
3'RACE	Ca5255_1st PCR	GCTCTCACAAAGCAAGCGC
	Ca5255_2nd PCR	CGAACATGCCATGGACGGC
Full-length gene model	Ca5255_FULL_1f	CAAAGCAAGCGCAAATCAAACAC
	Ca5255_FULL_2r	GCATTCTACGTACTCTTCATGG
Ca949		
Full-length gene model	Ca949_FULL_1f	GATCAAAGCCACGCACACACC
	Ca949_FULL_2r	ACGCAATAGTACTACGATATC
Ca11384		
3'RACE	Ca11384_3RACE_1f	GCACGCAAGCTCAGCGACAAG
	Ca11384_3RACE_2f	ATGGTGAACGTGTGCCATTAC
Full-length gene model	Ca11384_FULL_4r	GAATGGGTTCGGAAACAATGC
	Ca11384_FULL_3f	GAGAGAATAAGTAGGCACACAC
Ca1945		
3'RACE	Ca1945_1st PCR	GACTATTGTCGATCTGTGGTTC
	Ca1945_2nd PCR	CCCACCGTCCAGTACACATTC

Construction of genes encoding GFP-tagged CaTrailin2 and CaTrailin3

The CaTrailin2 gene was PCR amplified from genomic DNA using the oligonucleotides 5'-GATC TAC GTA ATG ATC ATG AAG ATG CTG TCG-3' and 5'-AGGT TCT AGA TCC CTC GAC AGG TTC CGC ATC-3' that introduced a SnaBI (bold) and XbaI (italics) restriction sites. The resulting 2603-bp PCR product was digested with SnaBI and XbaI, ligated into the SnaBI and XbaI sites of pCafcp_GFP (Poulsen, 2023) and transformed into Dh5α competent *E. coli*. The sequence of the resulting plasmid pCafcp_CaTrailin_2_GFP was confirmed by DNA sequencing (Eurofins Genomics).

The CaTrailin3 gene was PCR amplified from genomic DNA using the oligonucleotides 5'-GATC TAC GTA ATG TTC AAG CAC ATC ACC GTC-3' and 5'-GAAT TCT AGA TGT GGA GGT TTG ACC CGG CGG-3' that introduced a SnaBI (bold) and XbaI (italics) restriction sites. The resulting 4,421-bp PCR product was cloned into pJet1.2 using SURE2 competent *E. coli* cells (Agilent) and sequenced. The resulting plasmid, pJet1.2/CaTrailin_3, was digested with SnaBI and XbaI and introduced into the SnaBI and XbaI sites of pCafcp_GFP (Poulsen, 2023), generating the final plasmid pCafcp_CaTrailin_3_GFP.

Bioinformatics analysis of AM proteins.

Protein domains were identified using ScanProsite (de Castro *et al.*, 2006), Simple Modular Architecture Research Tool (Letunic & Bork, 2018), Pfam (Finn *et al.*, 2016), and InterPro (Finn *et al.*, 2017). Signal peptides were identified by the SignalP-5.0 program (Armenteros *et al.*, 2019). Intrinsically disordered regions were identified using IUPred3 (Erdos *et al.*, 2021). Homologs with other eukaryotic species were identified by performing a search against the MMETSP dataset (Keeling *et al.*, 2014), PLAZA Diatoms 1.0 (Osuna-Cruz *et al.*, 2020) and Phycocosm (Grigoriev *et al.*, 2021). The minimal e-value for homologs was set to 10^{-20}. BLAST hits to *Tiarina fusa* (ciliophora) or the dinoflagellate *Karenia brevis* (myzozoa) are possibly due to (cross-)contamination of the transcriptomes and have therefore been excluded from further analysis (Marron *et al.*, 2016; Van Vlierberghe *et al.*, 2021).

Phylogenetic analysis of diatom CAA domains

A seed alignment was created using the amino acid sequences of the four *Craspedostauros australis* and three *Amphora coffeaeformis* proteins that contain a CAA-domain (see Table S2). After alignment with MAFFT v7.453 (Katoh & Standley, 2013), manual trimming was performed with MEGA X (Kumar *et al.*, 2018) to retain only the CAA domain, and a HMM

profile was created with hmmbuild in HMMER v3.1b2. Next, hmmsearch was used for an initial homolog search against the full PLAZA Diatoms protein database, which contains 666,550 protein sequences of 10 diatom species and 16 other eukaryotic species (Osuna-Cruz *et al.*, 2020). An updated HMM profile including the six diatom hits with an E-value below 1e-20 was created by aligning the new protein sequences with the original seven CAA-domains with MAFFT, followed by manual trimming. Two extensive reference datasets were selected for further homology searches. Firstly, the NCBI non-redundant (nr) dataset v2022-02-05 was retrieved from ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nr.gz. Secondly, 224 "clean" transcriptome samples were retrieved from the decontaminated Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP) (Keeling *et al.*, 2014; Van Vlierberghe *et al.*, 2021). To translate the open-reading frames (ORF) of these nucleotide sequence datasets to amino acid sequences, Transdecoder v5.0.2 (available at github.com/TransDecoder) was used. To allow detection of the correct ORF based on homology, all possible translated ORFs were queried against the merged SwissProt v2022-02-09 and PLAZA Diatoms 1.0 proteome with Diamond v2.0.14 (Buchfink *et al.*, 2015) in --ultra-sensitive mode. Next, the hmmsearch function from HMMER v3.1b2 was used to search homologs, and only hits with an E-value below 1e-20 were retained. To reduce the number of bacterial sequences CD-HIT v 4.8.1 (Fu *et al.*, 2012) was used to select representative sequences, specifying a global sequence identity cutoff (-c) of 0.7. The NCBI taxonomy (downloaded on 2022-02-11) of the final hits was retrieved using Taxonkit v0.10.1 (Shen & Ren, 2021). Homologous CAA domains from all sources (proteomics, NCBI nr, PLAZA Diatoms and MMETSP) were combined and aligned using MAFFT v7.453, followed by automatic gap trimming with trimal v 1.4.1 (gap threshold 0.8). A maximum-likelihood phylogenetic tree was inferred using IQ-tree v 2.1.3 (Nguyen *et al.*, 2015) with automatic model selection from the following set of substitution models: JTT,LG,WAG,Blosum62,VT and Dayhoff, using empirical amino acid frequencies (-mfreq = F) and the FreeRate rate heterogeneity model (-mrate R) for model selection. Finally, phylogenetic trees were plotted with ggtree and ggtreeExtra in R (Yu *et al.*, 2017; Xu *et al.*, 2021). To construct the species tree in Fig. 2B, a maximum-likelihood 11-gene phylogeny consisting of 1151 taxa was retrieved from (Nakov *et al.*, 2018). The topology of the selected species was visualized with ggtree after pruning the tree using the drop.tip command from Ape v5.6-1 for R (Paradis & Schliep, 2019).

Immunolabelling of *C.australis* whole cells and biofilms

Immunolabelling of *C.australis* cells was performed according to previously established method (Aumeier, 2015). If not stated otherwise, all procedures were performed at RT under constant rocking. Following cell counting, the cells were transferred to a 1.5 mL Eppendorf tube ($1.8x10^6$ cells/tube) and fixed with 1% glutaraldehyde + 1.5% paraformaldehyde in Actin Stabilizing Buffer (ASB; 50 mM Pipes, 10 mM EGTA, 5 mM $MgCl_2$, 50 mM KCl, 1% DMSO, pH 7.1) for 15 min. The cells were then rinsed twice with ASB (5 min each), treated with fresh sodium borohydride (2x 15 min) and washed again with ASB (3x 5 min). The non-specific antibody interactions were blocked by incubation with 0.1% Fish gelatin (Sigma Aldrich) and 1% BSA in PBS for 1 h. Following blocking, the cells were incubated with the primary polyclonal antibodies (5 µg/mL; GenScript) diluted in 1% BSA in PBS overnight at 4°C under constant rocking. The cells were then washed with PBS (2x 5 min), incubated for 1 h with the secondary antibody (Goat-anti-rabbit conjugated with AlexaFluor488, 1:3000, Life Technologies) diluted in 1% BSA in PBS and washed again with PBS (2x 5 min). Finally, the cells were resuspended in 200 µl of PBS and 50 µL ibidi mounting medium (Ibidi), mixed with 250 µl of 3% agarose (Biozym) and distributed on a glass bottom dish (µ-Dish 35 mm, ibidi).

To visualize biofilm formation, *C. australis* cells were incubated in a glass bottom chamber slide ($2.5x10^3$ cells/well, µ-slide 8 well chamber, ibidi) with a 12 h light/12 h dark cycle at 18°C for 1, 4, 8 or 12 days. All steps were performed at RT. The culture medium was aspirated and the cells incubated in 2% BSA in PBS (BS) for 1 hour. The samples were then incubated for 2 hours with primary antibodies or pre-immune serum (5 $\mu g \cdot mL^{-1}$) diluted in BS + 0.05% Tween 20. The samples were washed (1x 2 sec, 1x 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 hour 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 63x (1.4) Oil DIC M27 objective or C-Apochromat 40x/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).

Conjugation of fluorescent dyes to lectin

Both lectins were labelled with fluorescent dyes via N-hydroxysuccinimide (NHS)-activated crosslinker. *Aleuria aurantia* lectin (AAL, Vector Laboratories) was dissolved in sodium bicarbonate buffer (0.2 M, pH 8.3) into final concentration 1 mg/ml and 7 nmol of AAL was used for conjugation to either Atto 488 NHS ester (Sigma Aldrich) (Thermofisher Scientific). Three times molar access of the dye was added into the solution and the sample was incubated for 3 hours at RT in the dark with gentle rotation. Then 100 µl of Tris-HCl was added (1 M, pH 8.5) to block the excess activated NHS groups and the unconjugated dye was removed by separation of the sample on the PD MiniTrap G-10 desalting column (GE Healthcare). Gravity protocol was used for removal of the free dye and washing steps were performed according to manufacturer's protocol. Briefly, PD MiniTrap G-10 column was prepared and equilibrated by washing with PBS (repeated 3 times) and then 0.5 ml of the sample was loaded onto the column. PBS was added to fill the column to 0.7 ml and the flow through was discarded. Finally, the sample was eluted with 0.2 ml PBS. The elution was repeated 3 times and eluates were kept separate. Eluates were tested on fluorescence microscope and the most effective one (eluate No.2) was aliquoted and frozen in -20°C.

Lectin labelling

The procedure for lectin labelling was similar to the immunolabelling. However, instead of antibodies the samples were incubated for 1 h with lectin from *Aleuria aurantia* (AAL: Vector Laboratories) conjugated with Atto488 (4 $\mu g \cdot mL^{-1}$). Confocal microscopy images were taken using a Zeiss LSM780 inverted microscope equipped with a Zeiss Plan Apochromat 63x (1.4) Oil DIC M27 objective or C-Apochromat 40x/1.2 W Corr M27 objective. Two channels were used to separately monitor chloroplast together with AlexaFluor647 fluorescence (emission 654-693 nm) and Atto488 channel (emission at 480-515 nm). Images were analyzed using the ZEN 3.1 software (Zeiss). The control experiments were performed at the same conditions without the labelling with lectins.

Dot immunoblot analysis of the hydroxylamine solubilized adhesive material

As attempts to transfer the HMW solubilized adhesive material for a western blot failed we have performed a dot immunoblot of the adhesive material separated by gel filtration chromatography. For the dot immunoblot assay, 4 µL of each pooled fraction was spotted onto a nitrocellulose membrane and allowed to air dry. The membrane was blocked for 90 min in Rotiblock (Carl Roth, Germany) and then incubated with the primary antibody (5 µg/mL)

diluted in Rotiblock overnight at 4°C. The membrane was then washed three times for 10 min in PBS containing 0.05% Tween (PBST), and then incubated with the secondary antibody (anti-rabbit IgG-HRP, 1:10,000, Sigma Aldrich) diluted in Rotiblock for 1 h at RT. Finally, the membrane was washed three times for 10 min with PBST, incubated with Pierce SuperSignal West Pico chemiluminescent reagent (Thermo Scientific, Germany) for 5 min and then the signal was detected on a Chemidoc Imaging System (Bio-Rad, Germany).

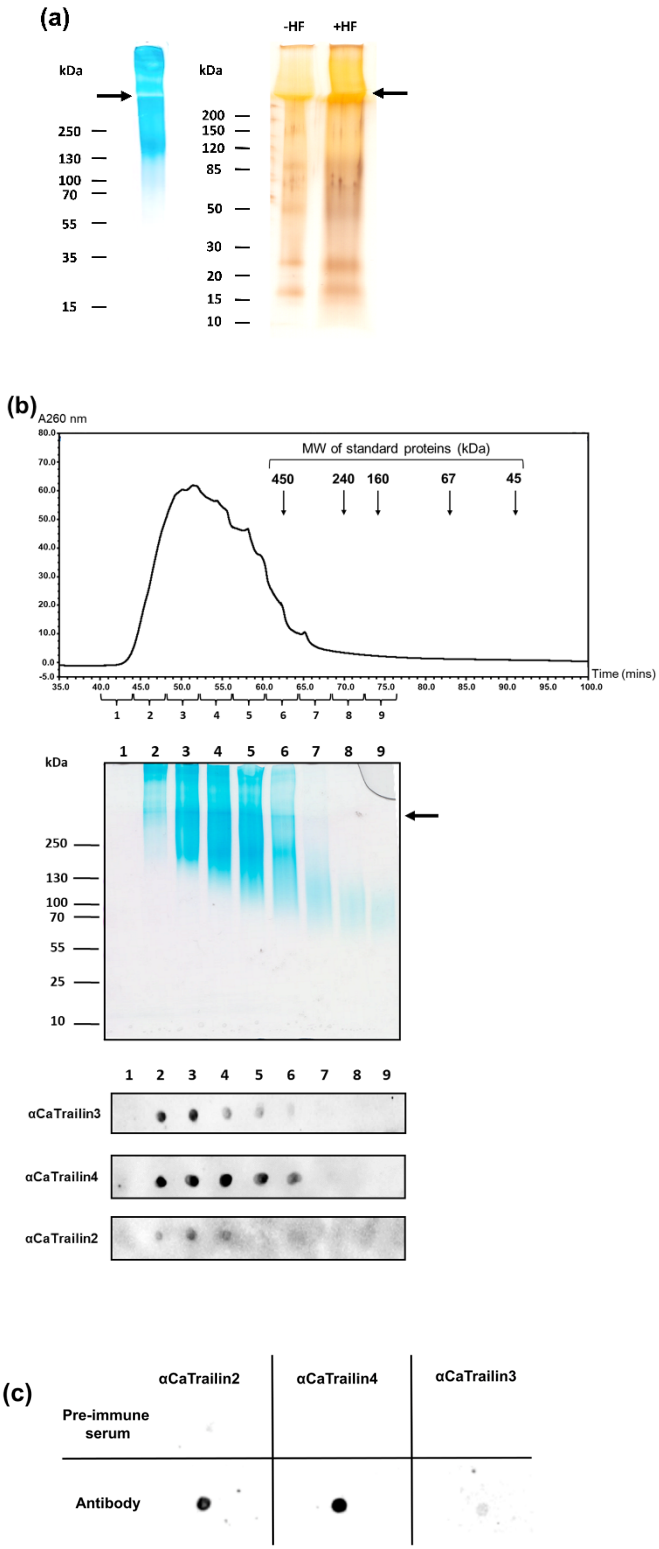


Figure S1.
Biochemical characterization of the hydroxylamine solubilized adhesive trails from *C. australis*. (**a**) SDS-PAGE of adhesive material stained with Stains All (left) and Silver Stain (right). The interface between the stacking gel and resolving gel is highlighted with an arrow to indicate that much of the adhesive material is very high molecular weight and does not migrate into the resolving gel. **(b)** Gel filtration chromatography separation of hydroxylamine solubilized *C. australis* adhesive material (AM) (upper panel). The arrows refer to the molecular weight of standard proteins (in kDa) using Ferritin (450 kDa), Catalase (240 kDa), Aldolase (160 kDa), Albumin bovine (6kDa), Albumin egg (45 kDa), Chymotrypsinogen A (17.8 kDa). The pooled fractions were separated on 16% Schägger gel with subsequent Stains All staining (middle panel). Each fractioned pool was spotted onto nitrocellulose membrane, air dried and then probed with the antibodies αCaTrailin3, αCaTrailin4 and αCaTrailin2. **(c)** Control dot blot probing the crude, adhesive material (9 nmol carbohydrates) using both the pre-immune serum and the polyclonal antibodies against CaTrailin2, CaTrailin3 and CaTrailin4.

Table S3: *C. australis* PacBio genome assembly statistics:

Coverage	54
Total Length	74,631,628
Number of contigs	88
Contig N50	1,724,159
Contig N90	411,438
GC content	52.93%
Shortest contig	53,658
Longest contig	3,871,151

Table S4: *C. australis* PacBio genome completeness statistics. Values shown are BUSCO (version 5.2.2) scores for stramenopiles ODB10 data set.

Complete BUSCOs	97
Complete and single-copy	94
Complete and duplicated	3
Fragmented BUSCOs	2
Missing BUSCOs	1
Total groups searched	100
% complete	97.0

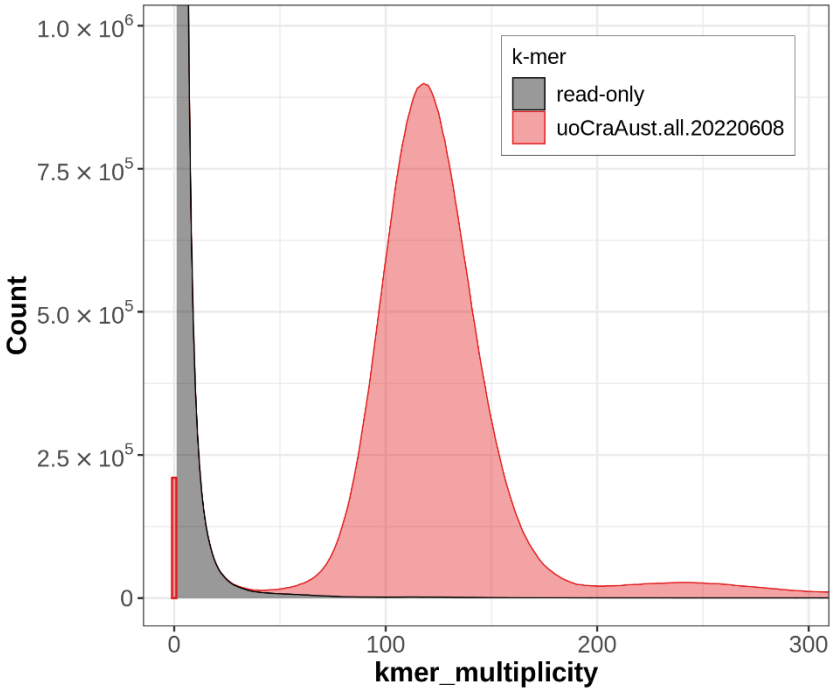


Figure S2: Merqury assembly spectrum plots for evaluating k-mer completeness.

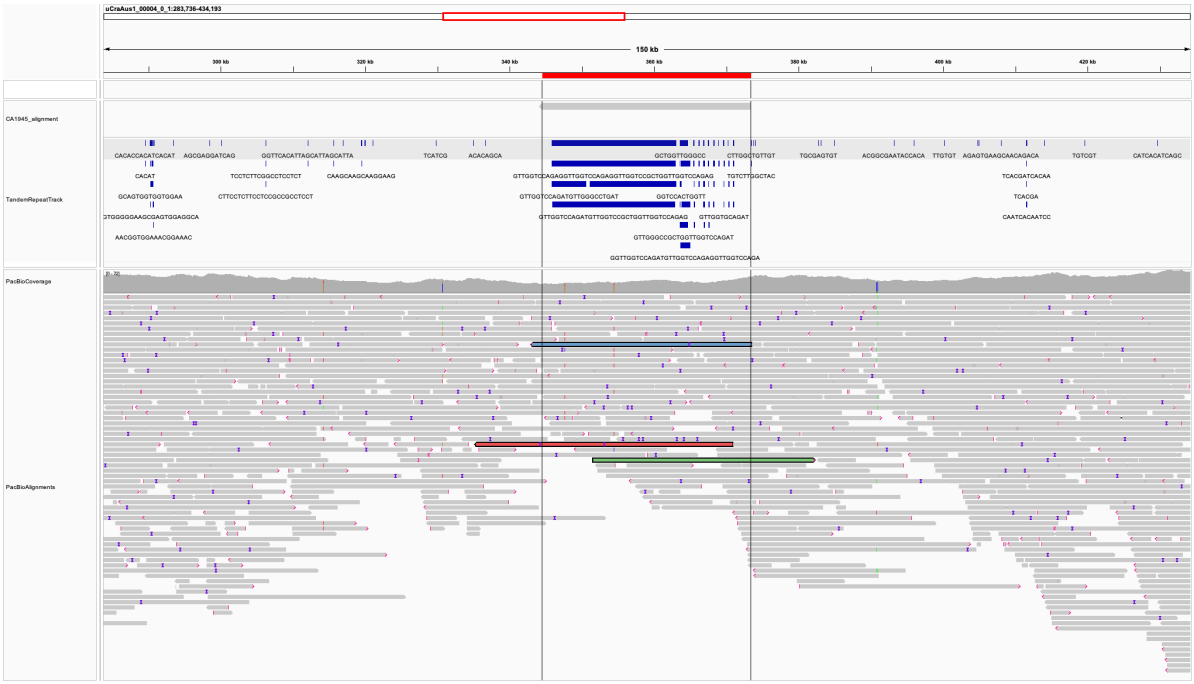


Figure S3. CaTrailin4 PacBio sequence alignment snapshot. Alignment snapshot of contig uCraAus1_00004_0_1 (283Kb – 434Kb) that includes the CaTrailin4 gene region highlighted in red and delimited by vertical black lines. Row CaTrailin4_alignment shows the alignment position of the CaTrailin4 gene sequence. The following row TandemRepeatTrack shows the tandem repeat annotation in blue blocks labelled with the tandem motifs that were reported by TRF. Rows PacBioCoverage and PacBioAlignments show the alignment coverage and the alignments of the raw PacBio reads that were mapped with minimap2.

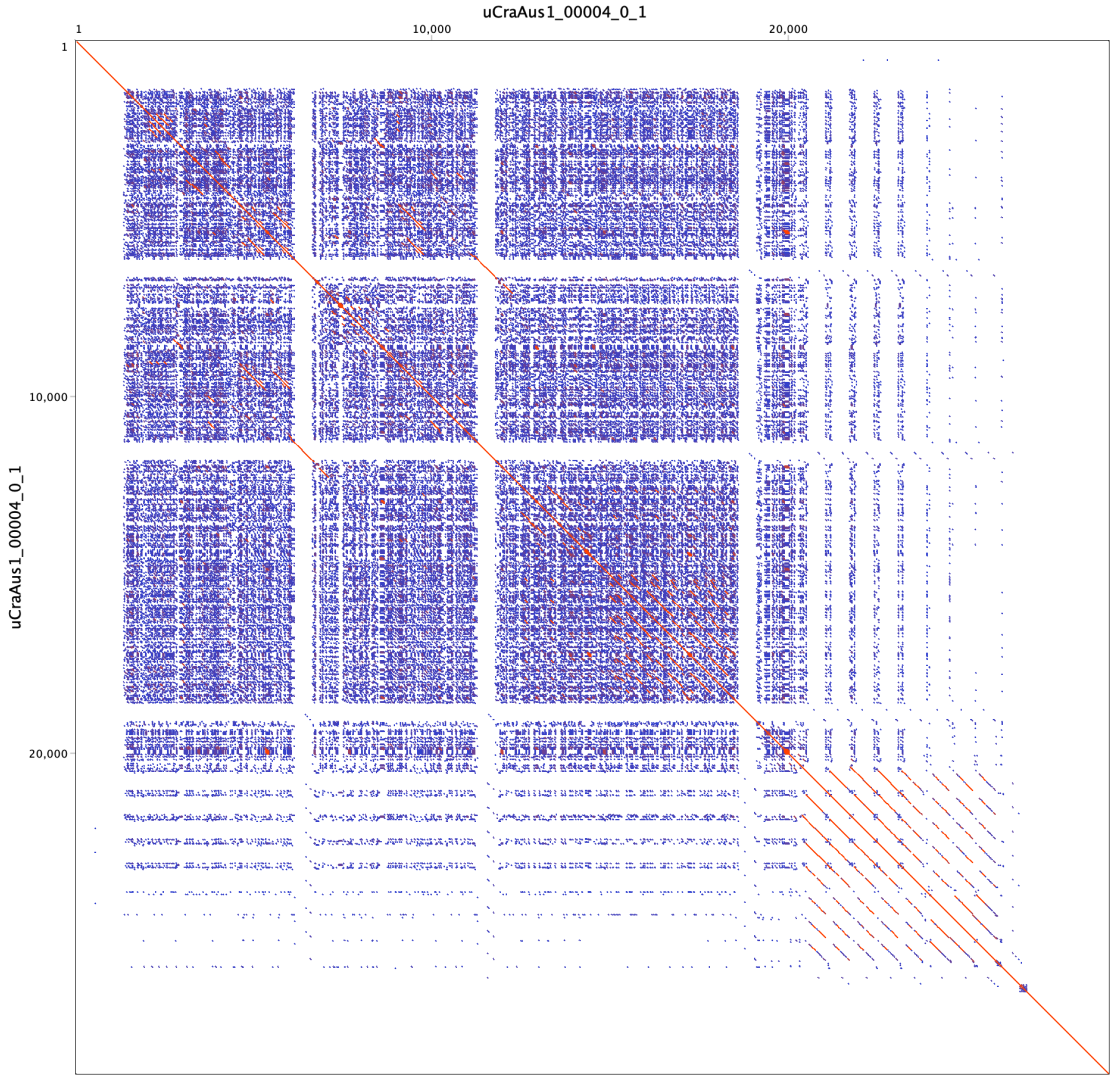


Figure S4. Dotplot: Dot plot of the CaTrailin4 region of contig uCraAus1_00004_0_1 (344kb – 373Kb). Created via Geneious (version 2020.0.5).

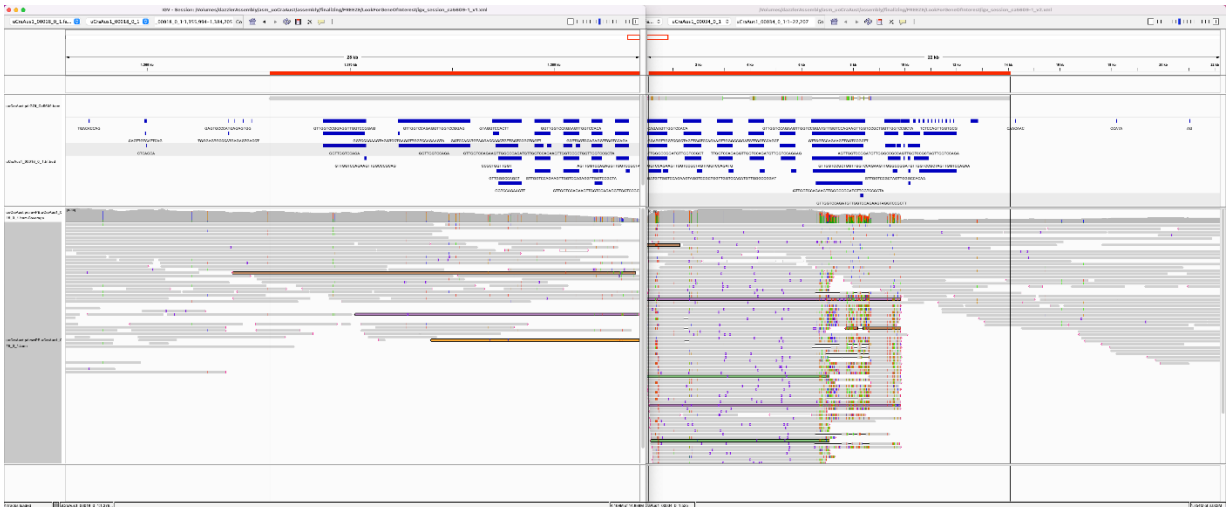
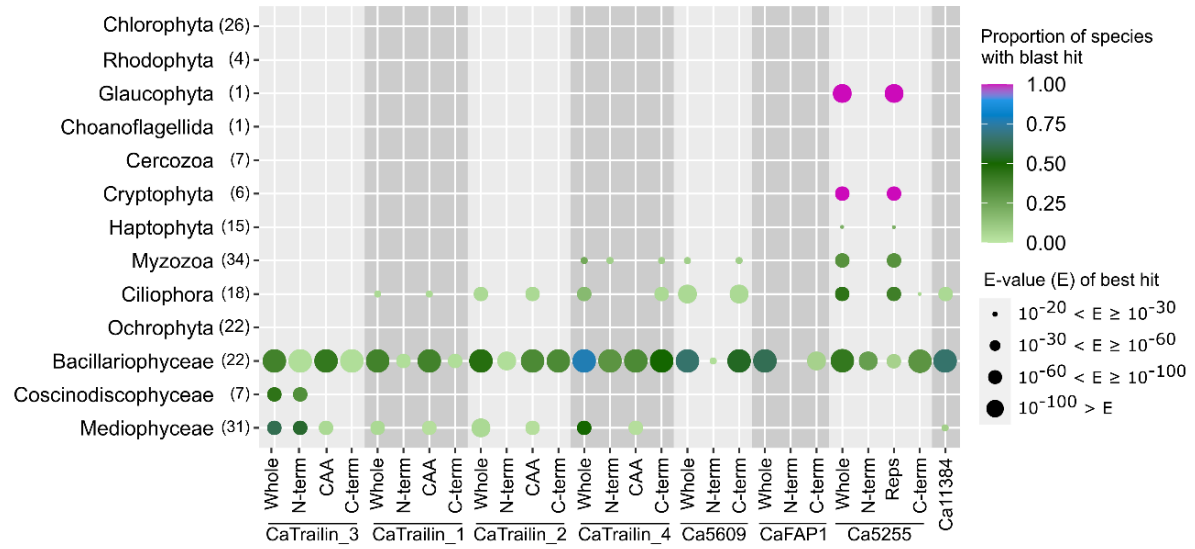


Figure S5. Ca5609 PacBio sequence alignment snapshot. Alignment snapshot of contig uCraAus1_00018_0_1 and uCraAus1_00034_0_1 that includes the Ca5609 gene region highlighted in red and is delimited by vertical black lines. The following row TandemRepeatTrack shows the tandem repeat annotation in blue blocks labelled with the tandem motifs that were reported by TRF. Rows PacBioCoverage and PacBioAlignments show the alignment coverage and the alignments of the raw PacBio reads that were mapped with minimap2.

(a)


Chlorophyta (26)
Rhodophyta (4)
Glaucophyta (1)
Choanoflagellida (1)
Cercozoa (7)
Cryptophyta (6)
Haptophyta (15)
Myzozoa (34)
Ciliophora (18)
Ochrophyta (22)
Bacillariophyceae (22)
Coscinodiscophyceae (7)
Mediophyceae (31)
Whole
N-term
CAA
C-term
Reps
Ca11384
CaTrailin_3
CaTrailin_1
CaTrailin_2
CaTrailin_4
Ca5609
CaFAP1
Ca5255
Proportion of species with blast hit
1.00
0.75
0.50
0.25
0.00
E-value (E) of best hit
10-20 < E ≥ 10-30
10-30 < E ≥ 10-60
10-60 < E ≥ 10-100
10-100 > E

(b)

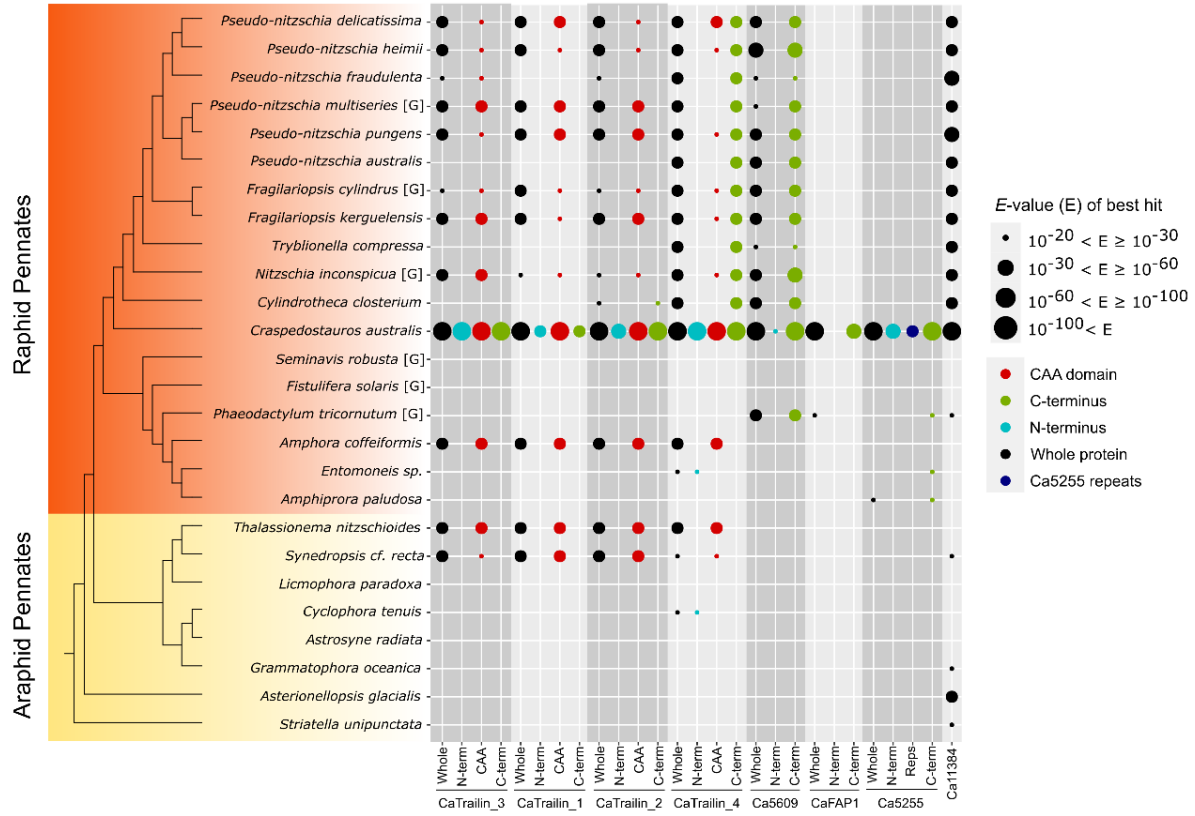

Raphid Pennates
Araphid Pennates
Pseudo-nitzschia delicatissima
Pseudo-nitzschia heimii
Pseudo-nitzschia fraudulenta
Pseudo-nitzschia multiseries [G]
Pseudo-nitzschia pungens
Pseudo-nitzschia australis
Fragilariopsis cylindrus [G]
Fragilariopsis kerguelensis
Tryblionella compressa
Nitzschia inconspicua [G]
Cylindrotheca closterium
Craspedostauros australis
Seminavis robusta [G]
Fistulifera solaris [G]
Phaeodactylum tricornutum [G]
Amphora coffeiformis
Entomoneis sp.
Amphiprora paludosa
Thalassionema nitzschioides
Synedropsis cf. recta
Licmophora paradoxa
Cyclophora tenuis
Astrosyne radiata
Grammatophora oceanica
Asterionellopsis glacialis
Striatella unipunctata
Whole
N-term
CAA
C-term
Reps
Ca11384
CaTrailin_3
CaTrailin_1
CaTrailin_2
CaTrailin_4
Ca5609
CaFAP1
Ca5255
E-value (E) of best hit
10-20 < E ≥ 10-30
10-30 < E ≥ 10-60
10-60 < E ≥ 10-100
10-100< E
CAA domain
C-terminus
N-terminus
Whole protein
Ca5255 repeats

Figure S6 Phylogenetic distribution of *C. australis* adhesive proteins (**a**) Distribution of BLAST hits of *C. australis* adhesive proteins across the Eukaryotic tree of life represented in the Marine Microbial Eukaryotic Transcriptome Sequencing Project (MMETSP) transcriptomes, including pennate diatoms (Class: Bacillariophyceae), radial centric diatoms (Class: Coscindodiscophyceae) and polar centric diatoms (Class: Mediophyceae). Either the entire protein sequence or partial sequence (non-repetitive N-terminal or C-terminal sections) were queried against the MMETSP transcriptome database (all other species) using tBLASTn with low complexity filtering. The size of the circles indicates e-value of the best BLAST hit, while the color indicates the proportion of species in the taxonomic group that contains a hit at a minimum e-value of 1E-20. The number of species in the taxonomic group is indicated next to the group name in brackets. (**b**) Distribution of BLAST hits of *C. australis* adhesive proteins across the Bacillariophyceae (pennate diatoms) phylogenetic tree. Either the entire protein sequence or partial sequence (non-repetitive N-terminal or C-terminal sections) were queried against Phycosm genomic database (species labelled [G]) or the MMETSP transcriptome database (all other species) using tBLASTn with low complexity filtering. The size of the circles indicates e-value of the best BLAST hit.

Table S5. Clustal Omega sequence alignment of **(A)** CaTrailin3 Cysteine rich repeats, **(B)** Ca5255 Cysteine rich repeats, **(C)** CaTrailin4 NCVD repeats and Ca5609 QPTY-repeats. The consensus symbols represent: an * (asterix) fully conserved residue; a : (colon) indicates conservation between groups of strongly similar properties; a . (period) indicates conservation between groups of weakly similar properties.

A. CaTrailin3 Cys repeat

```
R1  CGTGWMDTFNA-CMTSKP-CPGGQNSECADGQSCFGGINCSGNSYGSTDLPYSGPKF
R2  CGSSWTDASNT-CASAAP-CPGGQDNECPDGLKCFSDINCATSTTYTPPVESSSYSGPNF
R3  CGSSWTGAKNG-CATGAP-CPGGQNDECPDGESCFGGITCPTPTTYGAGSTSSFLGDSAGSSSSASSYTGPNF
R4  CGSSWTGANDG-CATGAP-CPGAQDSECPDGQSCFGGITCPTPTTYGSTSTSAATSGYTGPNF
R5  CGIDWMNAKNI-CAGGAP-CPGAQDSECPDGLKCFGGITCETPTTYGSTSTSTATSSYSGPNF
R6  CGSSWTSANDG-CATGAP-CPGAQDSECPDGQSCFGGITCPTPTTYGAGSTSSFLDDTAGSSSSTSSYTGPNF
R7  CGTSWTGAKNG-CATGAP-CPGAQDSECPDGQSCFGGITCPTPTTYGSTSAATSGYTAPNF
R8  CGSSWTDANNN-CATGAP-CPGAQDSECPDGQSCFGGITCPTPTTYGSGSTSSFLGGSAGSSSASGYTAPNF
R9  CGSSWTNANDN-CATGAP-CPGAQDSECPDGQSCFGGITCPTPTTYGAGSTSSFLDNTTPSYTGPSF
R10 CGDSWSLANTN-CATGAP-CPGGTDGECPGGQSCFGGITCPTPTTYGTGAAPVPNLNF
R11 CGPTYDDAEANKCGNGAQACPGGTADECPGSQDCFKVSACPGAAYLATEQESNASVKAVGSSPDASFGNPSSLSASL
    **        *     ***    **    **    *
```

B. Ca5255 Cys repeat

```
R1   CDDGNKINDDSCTNQCTSAK---CGDGIMQSGEE
R2   CDDGNKVDNDGCTNGCKLPK---CGDGFMQAGEE
R3   CDDGNAVNNDGCTNGCKLPT---CGDGIVQNGEE
R4   CDDGNAVNNDGCTNTCKNAK---CGDGIMQAGEE
R5   CDDGNAVNNDGCTNGCKLPT---CGDGIVQNGEE
R6   CDDGNNSNTDSCTNNCKNAK---CGDGIMQAGEE
R7   CDDGNAVNNDGCTNGCKLPT---CGDGIVQNGEE
R8   CDDGNNSNTDSCTNTCKNAK---CGDGFMQAGEE
R9   CDDGNAVNNDGCTNGCKLPT---CGDGIVQNGEE
R10  CDDGNNSNTDSCTNNCKNAK---CGDGFMQAGEE
R1   CDDGNNSNTDGCTNGCKLPK---CGDGIVQNGEE
R12  CDDGNNNNNDNCTNNCEIPVPNTCGNGIVDPGEE
          *****    * *** *       ** *    ***
```

C. CaTrailin4 DUF1

```
NCVD_1       NCVDPNGAKFTGSTDPTFTADPLPIVITFESDDGSTVEVDVVQNILNGEIKHVFIKYRNP   60
NCVD_2       NCVDPNGAKLTKSTDPTYTADPLPIVITFESDDGTSVEFDFVQNLSTDTIPHVFIKYRNQ   60
NCVD_8       NCVDPNGAVFKESTDPTYTADPLPIVLTFESDDGTSVEFDFVQNIINGVIPHVFIKYRNQ   60
NCVD_9       NCVDPNGAVFKESTDPTYTADPLPIVLTFESDDGTSVEFDFVQNIINGVIPHVFIKYRNQ   60
NCVD_6       NCVDPNGAVFKESTDPTYTADPLPIVLTFESDDGTSVEFDFVQNIINGVIPHIFIKYRNQ   60
NCVD_7       NCVDPNGAVFKESTDPTYTADPLPIVLTFESDDGTSVEFDFVQNIINGVIPHVFIKYRNQ   60
NCVD_5       NCVDPNGAIFKESTDPTYTADPLPIVLTFESDDGTSVEFDFVQNIINGVIPHIFIKYRNQ   60
NCVD_3       NCVDPNGAVFKESTDPTYTADPLPIVLTFESDDGTSVEFDFVQNIINGVIPHVFIKYRSQ   60
NCVD_4       NCVDPNGAVFKESTDPTYTADPLPIVLTFESDDGTSVEFDFVQNIINGVIPHVFIKYRSQ   60
NCVD_10      NCVDPDGAKLKGSTDPDYTADPKPIVLTFESDDGTLVEFDFVQNIINGVIPHVFIKYRNA   60
NCVD_11      NCVDPDGAKFVRSTDPSYTADPKPIILTFESDDATLVEFDFIQNIINGVIPHVFIKYRKS   60
NCVD_12      NCVDPDGAKFVKSTDASYTADPKPIILTFESDDATLVEFDFIQNIINGVIPHVFIQYRKS   60
             *****:** :  ***  :**** **::******.: **.*.:**: .. * *:**:**.

NCVD_1       ESGMDDCAKHDDVEFGKKIPITAECYGQITTITIYIYTGEDFPAEECEACEIPD------   114
NCVD_2       ESGDEECAQFDTVTLGTTLPIAAECFDGVTGVTIYIYTGEDFPSEECEACSIPD---EGD   117
NCVD_8       ETGTEDCAQFDTVGLGATYPLTAECFDGVTGVTIYIYTGEDFPSEECEACSIPD--GEGD   118
NCVD_9       ETGTEDCAQFDTVGLGATYPLTAECFDGVTGVTIYIYTGEDFPSEECEACSIPD--GEGD   118
NCVD_6       ETGTEDCAQFDTVGLGATYPLTAECFEGVTSVTIYIYTGEDFPSEECEACSIPD--GEGD   118
NCVD_7       ETGTEDCAQFDTVGLGATYPLTAECFEGVTSVTIYIYTGEDFPSEECEACSIPD--GEGD   118
NCVD_5       ETGTEDCAQFDNVETGATLPFAAECFSGVTGVTIYIYTGEDFPSEECEACSIPD---EGD   117
NCVD_3       ESGDEECAQFDTVTLGTTLPFAAECFDGVTGVTIYIYTGEDFPSEECEACSIPD---EGD   117
NCVD_4       ESGDEECAQFDTVTLGTTLPFAAECFEGVTGVTIYIYTGEDFPSEECEACSIPD---EGD   117
NCVD_10      ETGINECSEIMNAGEDSTFSYAAECIDQTTMVTIYIYTGDDFAGEDCEACSLPTDDGETG   120
NCVD_11      ETGINECAEFANVGENTPISLSAECVDMSTNVVIYIYTGDDFASENCEACSLPSDDGEGG   120
NCVD_12      ETGINECAEFANVGENTPISLSAECVDMSTNVVIYIYTGDDFASENCEACSLPSDDGEGG   120
             *:* ::*::   .  .     :***    * :.******:** .*:****.:*

NCVD_1       -GDPADFEAYSFEIPCDTECGEP       136
NCVD_2       SGNPDNFEAYYFEVPCQTECIDP       140
NCVD_8       SGNPDNFEAYYFEVPCQTECIDP       141
NCVD_9       SGNPDNFEAYYFEVPCQTECIDP       141
NCVD_6       SGNPDNFEAYYFEVPCQTECIDP       141
NCVD_7       SGNPDNFEAYYFEVPCQTECIDP       141
NCVD_5       SGNPDNFEAYYFEVPCQTECIDP       140
NCVD_3       SGNSDNFEAYHFEVPCQTECIDP       140
NCVD_4       SGNTDNFEAYYFEVPCQTECIDP       140
NCVD_10      SGNPDNFEAYYFELPCNTECIEP       143
NCVD_11      SGNPDNFAAYYFDVPCNTECIEP       143
NCVD_12      SGNPDNFAAYYFDVPCNTECIEP       143
              *:  :* ** *::**:*** :*
```

D. Ca5609 DUF2

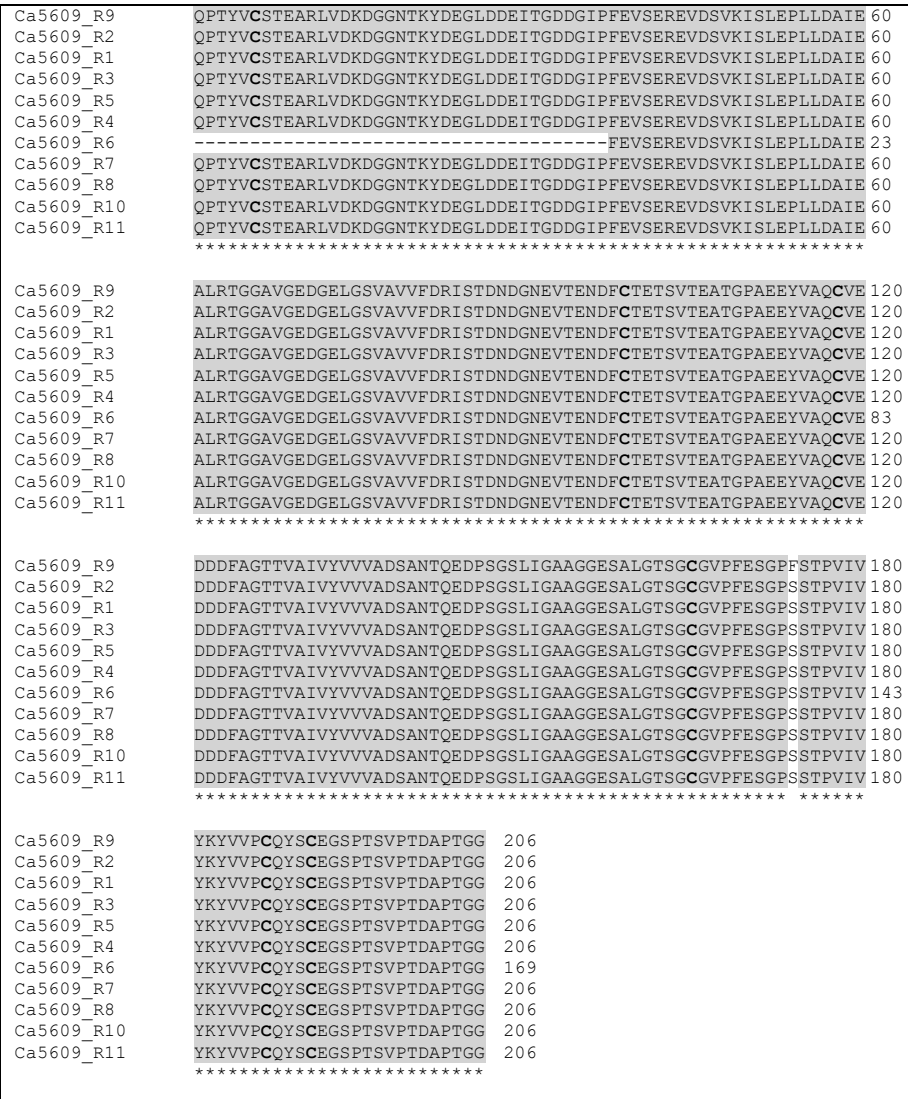


Table S6. BLASTp alignments of *C. australis* adhesive proteins highlighting regions of sequence similarity that are confined only to regions of the proteins that contain known protein domains.

CaTrailin3

>TPX53846.1 chitinase [Chytriomyces confervae]
MLSSRAILVLLALTQILAMPTLPLINARDGASTRCGKSWADANGRCGQACPSGQNSQCSNGETCFKDLAT
ASCGGGSPAPAPAPGNTGGSTRCGKSWADANGKCGQSCPTGQNSQCSNGETCFKDLATASCGGGSPAPAP
APAPAPAPGNTGGSARCGKSWADANGKCGQSCPGGQNSECSNGETCFKDLATAPCGGSSPAPAPAPAPAP
APGNTGGSTRCGKSWADANGKCGQSCPGGQNSECSNGETCFKDLATASCGGGSPAPAPAPAPAPAPGNTG
GSTRCGKSWADANGKCGQSCPGGQNSECSNGETCFKDLATASCGGSSPAPAPAPAPAPGNTGGSIRCGKS
WADANGKCGQSCPGGQNSECSNGETCFKDLATASCGNSPAPGPAPAPGPQPNDPSNPTSIQSLISESKFN
SALQTCGISKSGLYQSLVKGFTAPLANLRELALLVGNTAHESGAFIYVEEIACAGVTSPTGNCPYGLYHG
RGYIQLSWDYNYRAAASALNRPDIFSNPWVVQQDEATNWSTVQWYWTTAVQPALKANGYTLAASVRAING
GLECGGNPIAAKRIQFVHCFEQQFTGTQSGQTSC

Yellow: region with sequence homology to CaTrailin3
Grey: chitinase domain

>WP_017559785.1 choice-of-anchor A family protein [Nocardiopsis baichengensis]
MGTNLDVASDRKRRSTGRRAVVLLGTGAAGAAVLAGALALSSVAATLPGGLGPCVPGDCPDPYPEVDNGP
IAGRDNGINVFVGGDMLVREGAAEAEGKVVVLGDFDMDKREGASQIYNVGVAGVGSRVPPDDGTDFLTAG
GNVTVADGQRVLAEEGSVHGVVRYGGELQGTVVPEEVQDPDAAAPYADLRSELAAASDCYAYPHGEQREA
TGTADNQGHTTVFTGDGTSSLQVFNVDFDLTNTDGGQQGFSFVDIPTDATVLINVTGDRRTINTYMGELP
DGLRERILWNFPDATEVNLKGTGQFQGSVLVGEPGSTTRLSLPGTNGRFFTTGTLSHESDGGSGGQELHA
YPFNGDLPDCRNLGPSPTPTPSPSPSVSPTPSPEPTPSPTLEPSPSPSPEPSPSPEPSPSPEPSPTPTPE
PTDSPGPTPEPTVSPEPTDSPGPTDEPTDGPTPEPTDSPEPTDGPTDGPTPGPDGGGGDGGHGDAGDDAP
APDRGGLPLTGPEIAGLVAAALVLLAFGGLAVYASWRRRADGDGPG

Grey: CAA domain
Underlined: region with sequence homology to CaTrailin3

CaTrailin1

>WP_084461253.1 choice-of-anchor A family protein [Curvibacter gracilis]
MTAAFSHPGFPILLSLSALLSTPSYAGSGLGAETGPCLGPDCPSHYGPPAACRKLGAEGRDEAYSLVVGG
HFHVLQGSEAEGRIFVGGDLVLRTRGYYNLVQVGAGACVVPPDADQASSPPHLVVGGQIQPVDPSAVLAV
GQPNLRSTALIGGARVGAGQVQARGGVRYQTSPTPPPVDFNSLAQRSAYWATLAPTGTVRGQHPMVLQGD
GQSALQVFKLSADLPSGGLRLKGIPPGATVLINHPGTPTVSLQFYDMTDPAGHGGFQFDTELTRRMLWNF
PQASLVKLGGGAQWQGSVLVARGALYNALPGLNGRVLVAGDLTQDSQGSEFHNYDFKGELPDPPTDDPPA
TEPGHDTSPPEPPPEVPPEPPPESPPPGLMQAALNIQVRYDGDALRITDAPLLQIESQCERSGTQSVRIT
PPAAGWIEGLQTDESCLLVAKLERAAQLPRGWQWPSPQPGLWVGDNPRVITPDTNPVELLLQIRPVMNAS
LRVLPRYLGNHRAVQTHPRIELNLLCSHSGEQSLLLQAPQVAQSGLLQDGEACEVLARRITGAVLQEGHT
LSLPDTGWWSPGSTLVAGDTPDPIPLDISILAASPGDESEPPPAPPPVPPSESGPSTLNSIPLLTPWGLL
GLASGLGLLSLGTWMGPAIDRRKRRTASDKVNESRGCAGSSRPGSGPAAH

Grey: CAA domain
Underlined: region with sequence homology to CaTrailin1

CaTrailin2

>WP_084461253.1 choice-of-anchor A family protein [Curvibacter gracilis]
MTAAFSHPGFPILLSLSALLSTPSYAGSGLGAETGPCLGPDCPSHYGPPAACRKLGAEGRDEAYSLVVGG
HFHVLQGSEAEGRIFVGGDLVLRTRGYYNLVQVGAGACVVPPDADQASSPPHLVVGGQIQPVDPSAVLAV
GQPNLRSTALIGGARVGAGQVQARGGVRYQTSPTPPPVDFNSLAQRSAYWATLAPTGTVRGQHPMVLQGD
GQSALQVFKLSADLPSGGLRLKGIPPGATVLINHPGTPTVSLQFYDMTDPAGHGGFQFDTELTRRMLWNF
PQASLVKLGGGAQWQGSVLVARGALYNALPGLNGRVLVAGDLTQDSQGSEFHNYDFKGELPDPPTDDPPA
TEPGHDTSPPEPPPEVPPEPPPESPPPGLMQAALNIQVRYDGDALRITDAPLLQIESQCERSGTQSVRIT
PPAAGWIEGLQTDESCLLVAKLERAAQLPRGWQWPSPQPGLWVGDNPRVITPDTNPVELLLQIRPVMNAS
LRVLPRYLGNHRAVQTHPRIELNLLCSHSGEQSLLLQAPQVAQSGLLQDGEACEVLARRITGAVLQEGHT
LSLPDTGWWSPGSTLVAGDTPDPIPLDISILAASPGDESEPPPAPPPVPPSESGPSTLNSIPLLTPWGLL
GLASGLGLLSLGTWMGPAIDRRKRRTASDKVNESRGCAGSSRPGSGPAAH

Grey: CAA domain
Underlined: region with sequence homology to CaTrailin2

Ca5255

>KAG7347470.1 cold-active serine alkaline protease [Nitzschia inconspicua]
*MRHRSFFPFWLLSLHSSRYLVVHG*EKSRKSNLRNSKHSSRIPATNGVVNFLPFLVENAGRTLGNGRKLHK
NDSDVGTEEQKGKRMSTQARIGNGFERGHNRSIQQEKPLLNKRKVEVRNLTFQRCRNGGTQFVVQCTTGS
EEHCYRELARANAVIVNELPNSDFFVVCVHTAEEKKLLNELTDVVDMEEDCIRTLSYLPELTKPVDRREL
QSGQQIPYGATMVNAPQFWQTKGDKGGSAKVCIIDTGLNIGHEDIQGAVFSGSTDSSVVADWDSDDAGHG
THVAGTIAAVDNNIGVVGVAPEAELLIVKVFEGANSQFTASSLVSALEECRKGGANIINMSLGGPSSSVV
ERNKVNQLANQGIQLIAASGNSGDTSNPIEYPASYDKVISVAAVDEDRHIAIFSTHNNEVDVAAPGVDIL
SLTNACSTCYGLYSGTSMATPHVAGVFALLMSKYPTKSISQIREAIQESASDSGACGIDRMFGHGIVDVM
AAAVYLESGVSASEQNNCINTKITVTTDKWGSETSYVIRNSDGEVVYKNGPYSNQIKTYTDVFQLPDDCY
EFELLDSYGDGICCEEGSGSFKVEYDGVEEVYNDSFLNGNSVKASFGCGSGGDVGSGSPPSCGNGILEDI
EECDDGNQNNNDSCTNECKIAVCGDGIVGPGEQCDDGNNVNDDSCRNDCTSALCGDGVVQSGEECDDGNN
SNNDTCTNACKNARCGDGFLQPGEECDDGNGDNGDGCSSDCKVETISTSCQAGEAEVELLFQSDAYSYNE
NELYFYQSTDQTSVGDIDFIWIGTKQGIESNKKYEMSACVDETKCYKFFFFDSYGDGLWGNDGLRLSWNG
VEVLSVAPYEVGPKWDGGPTIYWTEDLGACS

italics: signal peptide
magenta: peptidase domain
Green: Myxococcus cysteine-rich repeat
Underlined: region with sequence homology to Ca5255

>MCB0319749.1 DUF4215 domain-containing protein [Bdellovibrionales bacterium]
*MMKNAFSRLSVRLVAFFCLGLICTSSA*RAQTYPPECQPLLDGSVLFVAAECLNDSSALISDSTTSDSNKW
QYSFDSNSDGVNGSNVGGTAYEIYGIAIRELNTEVWVVINSNIPVTGVPSASAAGGSVSWGDLFWNFTGN
DFEAADTTSNHFAIRFVTSNESGVPSLGVYREVSAKSVTAQNVGYASWDVYAAFVASHGQDASLGDFGLN
QTYYGGHLSLNVIDIGTFVGPITFLSPAELLSAGFDDSLFDGTTTIAFRFPKSYLAQYPDTDGDGQDDCI
DPCGNGKIESPEECDDGNFVNTDECSNLCTLPRCGDAIVQSAKGEECDDGNQVNDDSCSNACTNPKCGDG
ILQSGEQCDDGNTNNTDSCSNGCTTAVCGDGIVQTDEQCDDGNQIDSDECTNLCTTPKCGDGILQAGEEC
DDGNTSDSDSCTSTCTNARCGDGILQDGEECDDGNTSNSDSCSNTCTTPVCGDEIIQSGEECDDGNSVDD
DGCTNSCALPKCGDGILQGGEQCDDGNSNNSDSCTNACTTPKCGDGVVQAGESCDDGNGNNSDACTTSCE
VAACGDGFVQPGELCDDGNTIIDDECTNQCTLPNCGDGIVQSGEECDDGNSVNSDTCTNVCTVANCGDGI
KADSEQCDDGNQINTDACSNDCTLPVCGDGIIQGGELCDDGNLVDDDDCTSFCTLPTCGDGIVQAGEECD
DGNSVNNDSCTDQCTSARCGDGIVQAPEECDDGNEVDNDSCSVSCKVIDVVPEDCKEYNIEPSQFILDQG
VKLQEAFINLQILKKYVRLGGSKKYVADARAEAHTLQTEGWVLSWSVASQGVICTSSTENCVTSTANGDM
LAQYTERTEGLFDLTKQALRKLRKVGGSKKLIRTLRKRALKLLNRNLSERDKIPTNSTECTTS

italics: signal peptide
Green: Myxococcus cysteine-rich repeat
Underlined: region with sequence homology to Ca5255

Ca11384

>MMETSP1070 (*Minutocellus polymorphus)*
VTFEYIVADGNGATDSGIVTLTLFGQDTFAPSSSPTELFVCPDQTVNAANQGTISGSNVSGGTTSQTLQFDLF
ASSFTSNGMACVDGLIGIKPPYNILFVIDVSGSTRNNFSGTAVGDVNGDGRSNTILDAQIDTVLKAIEAIINT
PSLTNDNVNIGIVTFSTSGNYVGNWSPTDENNPDAIDPALETTLKSLRSGGYTNFDDALDKAIVYFEGSDSSK
GGAPDVSSRTNRMYFLSDGLPNQCGDGDPNTAEDYCGEDEIYNDAAGATVFASELQSLELYSVAIYAIGVGEA
SDVSAGSGLDKIDNTENPFTGEKAVQVTTTDALTSNILESPVYAEVFDFKLKVNGAIVPGVDETSLVTDISGY
VLGVNEISGLDPTNGASNTIIATVLLDFDGDSATPDDQLEISTIVSIIGAAGLPV*

Teal: von Willebrand factor type A (vWFA) domain
Underlined: region with sequence homology to Ca11384

>VEU43960.1 unnamed protein product [Pseudo-nitzschia multistriata]
MNRFFLSIAAASLALVSADVNFKEMHERKLAMEEDHRVMEEYLSTKSGTKSGTKSDTKSGRQLDASFFQT
PSPSASPSKVPTQTPSASPTTSEPSKQPTTSPSAEPSSIPSRLSEEYGGLPWPSASPSNFYEPAPKCGTI
ETKDGSERLLGPLPDENYYCNGDSYAWYRRTFTHYDYEFELSQYCEYKFSSPETGSSVIMRDVNHWFEQA
ATSLKGWDEVGYKASPEEYVFSGYAVPEFDDIEYESDIADAECPLNYCIRITGPSVGLNTNEELSDGSAE
PYLLYAFSSWEGYSTFEGFLSSTMYSWNAHHIFMTNHYLKCSCPSFCIDEEFYVQDFKPLATCYVEDICG
VKTFQIAFSDDYPETQRVSWENTGTTHVMEWFSECANVMHGCPQLASDFSPARFSGVGHFSSIRGPDVYT
YDEYESATNVYGFGGWRGFGWDYYSQPALLWECPSPEHLERLVEFGDLISDYSFFRHDSIDTSVPETCPE
PPLIEPEDPIVFDPEGPPCTEVENEGLFCFELKHATCYIRDTHECSEFEFVIEGVLDGAEDMWVTTDVEV
WDSWCHFDGIGCGIPSYWDEQPPYTYENSPWAFVSWSDYQHQWLEFQSGSSGRPSAIGGGNSYRLWNCDD
ATQAYLARDFTDAVAASSSGDLYQVPPLDCTESPTASPSMGPTSSIQPSNSPTMSPTICGESYLNSESLN
IDLAIDLSYSTYLYLFSSEVDIGDVNGDGKANTILDAQVQAIEELLVAILESETLGNRNCEINLISFHTD
ATNHGTFLPLSEDESSINTRLMHYIKTELRAPTSDLEVQLTNNGFTNFDAALDLAGDYFEYEATPDRLNL
LVFLSDGEPNVRGDGDDEGYCADTAGVWNSNPTNGFPAEVQCADLDLEAGVRHTFCRADDPECVPRNPYQ
ECVRGMTKCINSPAVTQYESEINRLTELRVERLAIGVGDASNVAEGSALWMIDNNPGKDLGVLPIQALDL
EALSNALKSLCILNTDPPTAEPSHQPSDQPTSSTAPSVIPTDVPSSEPTDVPSSSPTEEPEFTIETPSPT
KSPTQTPTKTPTKTPTASPSQEPSTSEPTGSPTAFPTRLPSASPTSEPTTAEPSAAPSQYPSGSFYPSSA
PTESPTESPAPSQSPTDAPTTSPTVSPTVSPTLSPTASPTKSPTASPTSSPTNKPTEIPTTSPTDSPTSS
VAPTFLLPECYDYPKLVKKDSSDTGICFFSEDMIFIEQQNTTEVGLRINNVWSKTLPQDVILFSHSNGVN
SVRGGNGFECDSEDGNAIDLAGSDEIMVQCHSEGDGEPYLAVVDLMIIDSNIPVNDVNHPCNPNEVIPNA
CSWRMVIPCEEDVMCTPEPTDSPSVYPTYVPTFGESTESPTTLTTFHSESPTTGTTFHEKTESPTHELDR
HTNDDESNDIVYESECPEDIVLLEQEGVTEFPEGGLYVVGREDGTVTVKLTQTFTDATVDSLFYQYQIGR
FSNKCFEDKDVPVTDSVDITILCTEHSKIALLELWVVDALEHNLLSESDNSVVPECCHPTHPEGTPATKY
LFEIKCLTVCADAVE

Teal: von Willebrand factor type A (vWFA) domain
Underlined: region with sequence homology to Ca11384

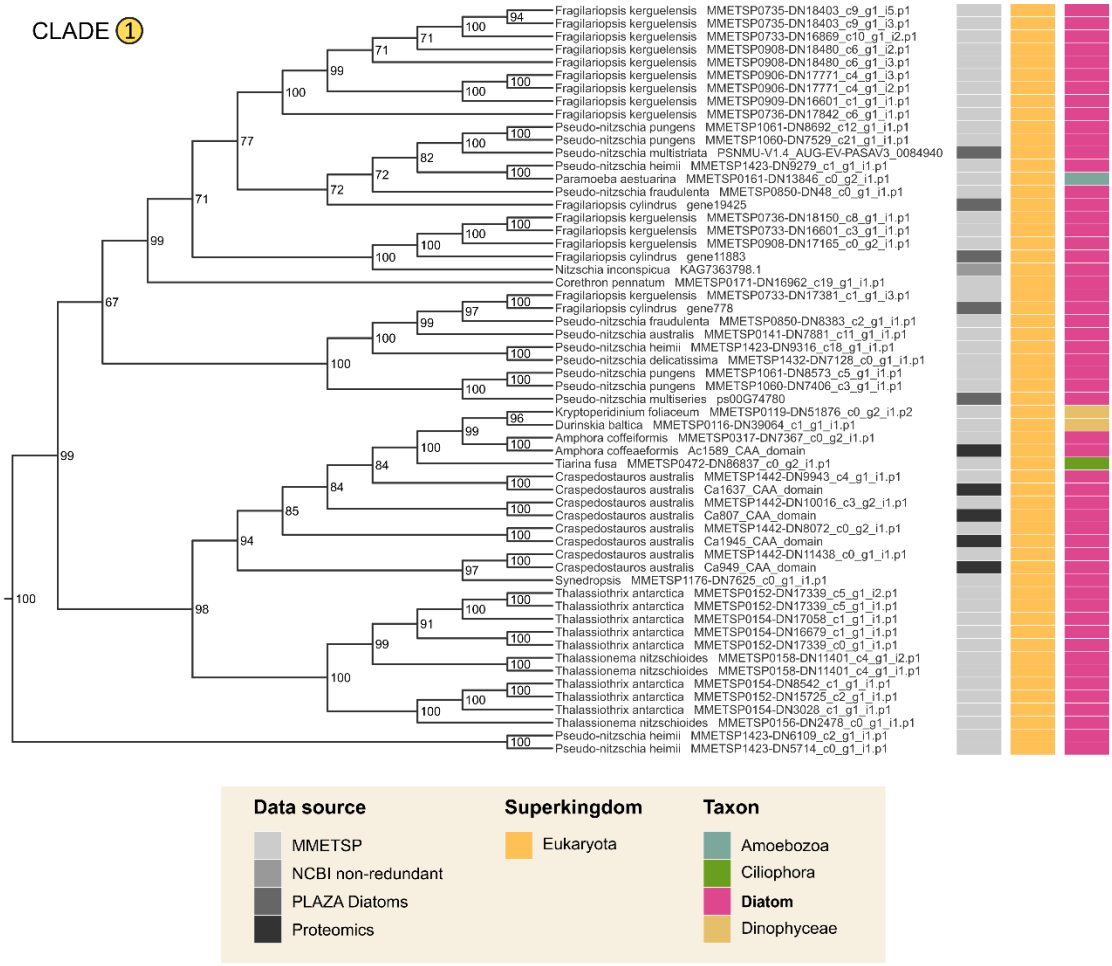


Figure S7. Midpoint-rooted maximum-likelihood phylogenetic tree of homologs of the Diatom Choice-of-Anchor A (CAA)-like domain, showing only proteins belonging to Clade 1 (subset of the tree shown in Fig. 2a). Node labels indicate bootstrap support (1000 ultrafast bootstrap repeats). For each hit, the species name and gene identifier is shown, as well as coloured boxes indicating the data source and taxonomic position.

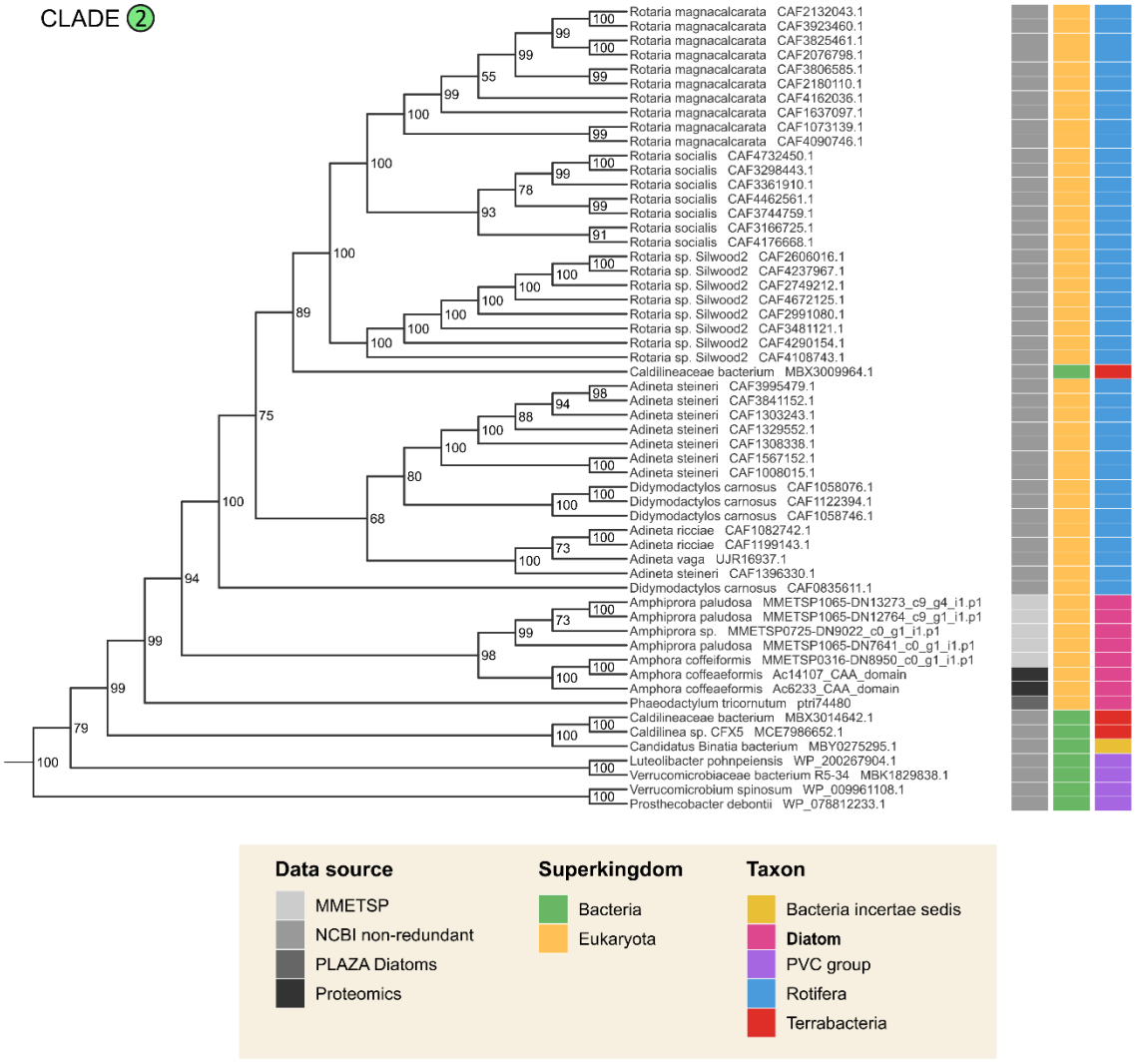


Figure S8. Midpoint-rooted maximum-likelihood phylogenetic tree of homologs of the Diatom Choice-of-Anchor A (CAA)-like domain, showing only proteins belonging to Clade 2 (subset of the tree shown in Fig. 2a). Node labels indicate bootstrap support (1000 ultrafast bootstrap repeats). For each hit, the species name and gene identifier is shown, as well as colored boxes indicating the data source and taxonomic position.

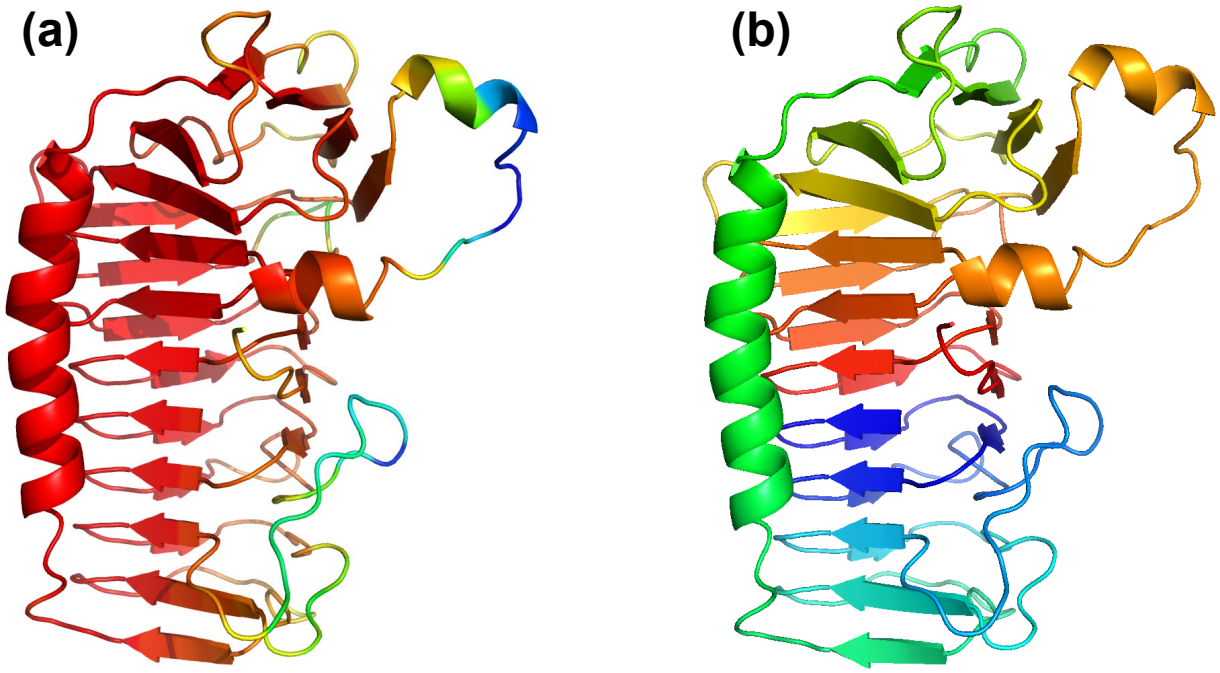


Figure S9. Alphafold structural prediction of the CaTrailin4 Choice-of-Anchor A (CAA)domain. **(a)** Quality of CaTrailin4 model. Red indicates high quality and blue low quality. **(b)** CaTrailin4 CAA domain model consists of two beta helical cores (red/orange and blue, respectively) held together by an alpha helix (green).

Table S7. The top eight predictions from the structural homology search

Name	Species		Uniprot ID	Uniprot Accession	PDB ID	Chain	Method	Resolution	RMSD	Aligned residues	Ice-binding
FfIBP	*Flavobacterium frigoris*	Bacterium	IBP_FLAFP	H7FWB6, I3QNZ8	4NU2	A	X-ray	2.10A	5,18	85%	Yes
ComZ	*Thermus thermophilus*	Bacterium	Q72JC1_THET2	Q72JC1	6QVI	A	X-ray	2.72A	5,49	80%	No
AFP	*Leucosporidium* sp. (Arctic yeast)	Fungus	IBP_LEUSY	C7F6X3	3UYU	A	X-ray	1.57A	4,73	80%	Yes
K1-A	*Typhula ishikariensis* (Gray snow mold fungus)	Fungus	IBPKA_TYPIS	Q76CE8	5B5H	A	X-ray	1.00A	5,06	75%	Yes
ColAFP	*Colwellia* sp.	Bacteria	IBP_COLSX	A5XB26	3WP9	A	X-ray	1.60A	5,82	73%	Yes
K3-B1	*Typhula ishikariensis* (Gray snow mold fungus)	Fungus	IBPKB_TYPIS	Q76CE6	3VN3	A	X-ray	0.95A	4,97	72%	Yes
fcIBP	*Fragilariopsis cylindrus*	Diatom	D0FHA3_9STRA	D0FHA3	6A8K	A	X-ray	1.40A	5,27	72%	Yes
Anp	*Antarctomyces psychrotrophicus*	Fungus	A0A2Z6DSM4_ANTPS	A0A2Z6DSM4	7BWX	A	X-ray	1.90A	4,21	70%	Yes

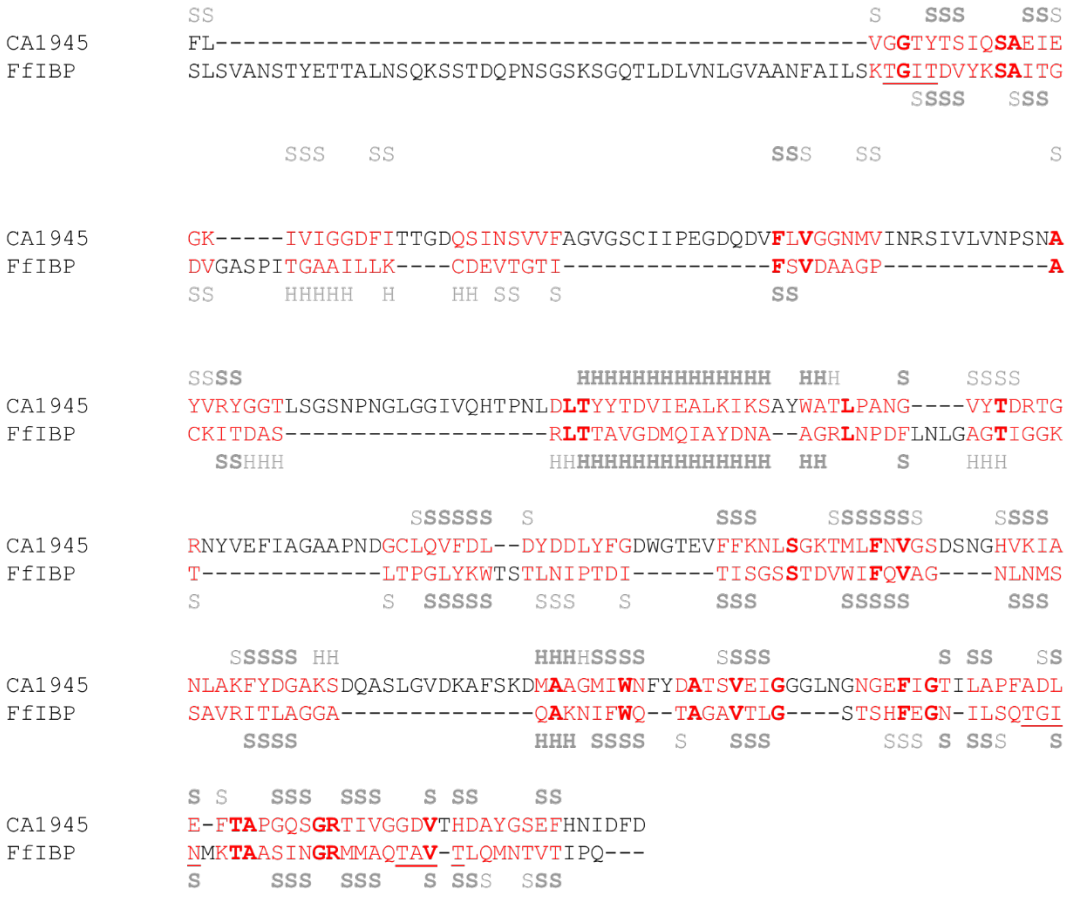


Figure S10. Sequence representation of structural alignment of the CaTrailin4 Choice-of-Anchor A (CAA)-domain and FfIBP. Structurally aligned residues are colored in red, sequence identity is highlighted in bold. Identical secondary structure (S for strand and H for helix) is emphasized in bold. Three ice-binding motifs (T-A/G-X-T/N) are underlined in the sequence of FfIBP. (FfIBP: *Flavobacterium frigoris* ice-binding protein).

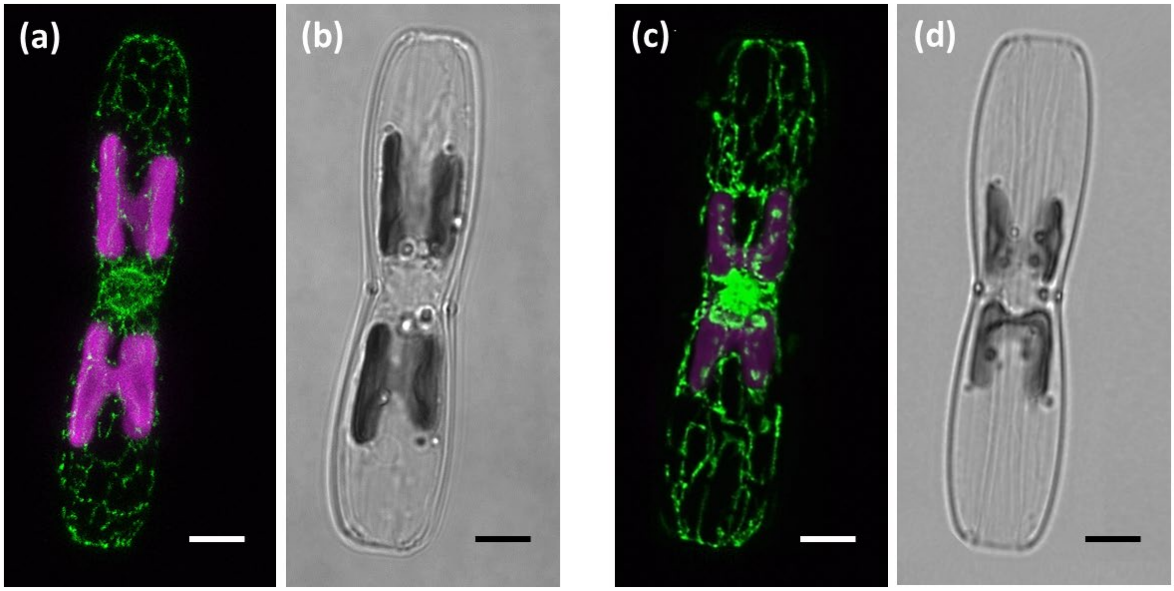


Figure S11. Confocal microscopy (z-projection) images of cell lines expressing the adhesive material (AM) proteins CaTrailin2-GFP and CaTrailin3-GFP transformants. Localization of CaTrailin2-GFP **(a)** and CaTrailin3-GFP **(c)** is shown in green, whereas the magenta color depicts chloroplast autofluorescence. **(b, d)** The corresponding light micrographs. Scale bar: 5 µm.

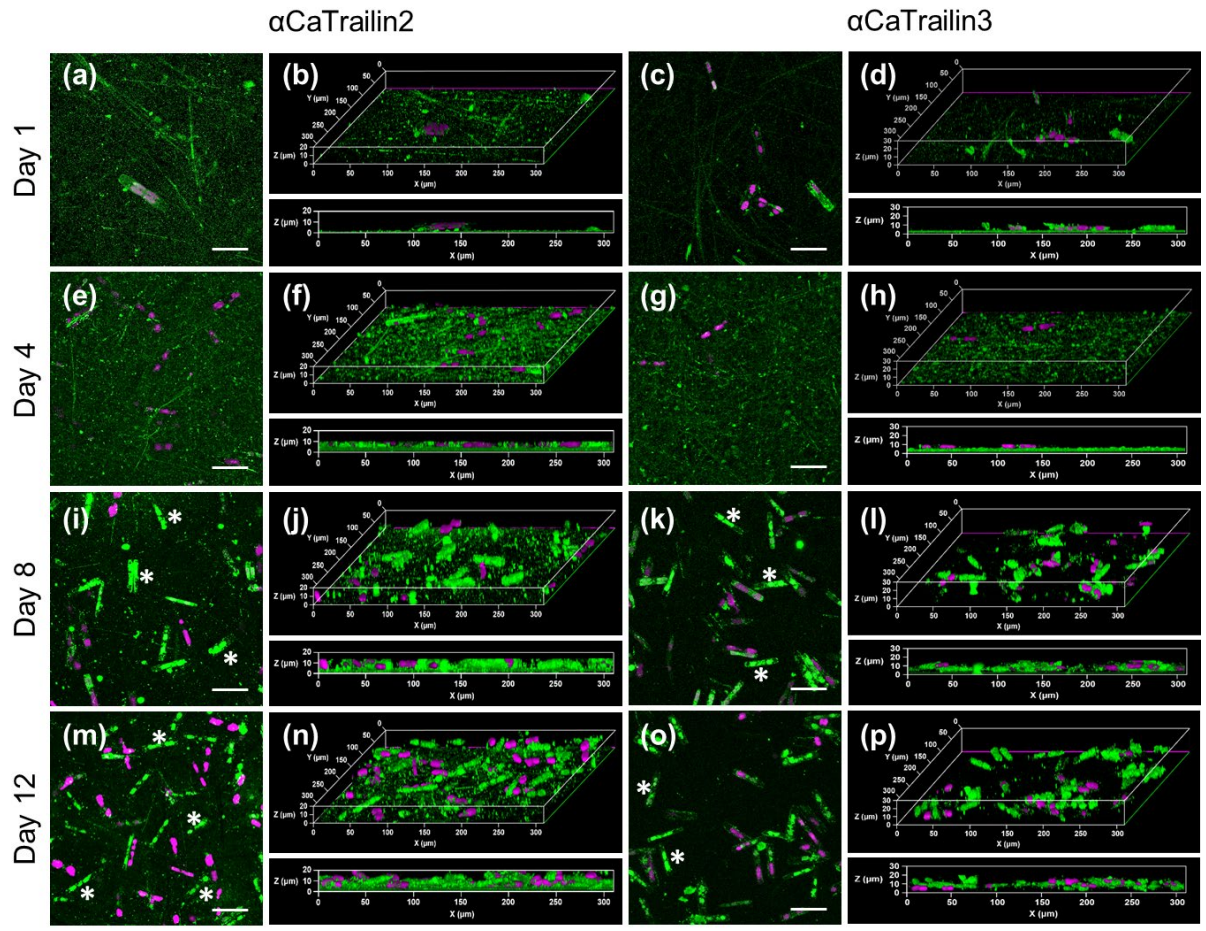


Figure S12. Monitoring the presence of adhesive proteins within biofilm by immunolabelling. Immunolocalization of adhesive proteins CaTrailin2 **(left)** and CaTrailin3 **(right)** in *C. australis* adhesive material after 1, 4, 8 and 12 days. The 3D reconstruction of the confocal images clearly shows the increase in height of adhesive material for both CaTrailin2 and CaTrailin3. All confocal images are presented as the maximum intensity Z-projection. The green color represents the fluorescence signal from AlexaFluor488 secondary antibody. The magenta color represents chloroplast autofluorescence. Some of the dead cells in the biofilm are indicated with an asterisk. Scale bar: 50 µm, 3D reconstruction: 300 µm (X) x 300 µm (Y) x 30 µm (Z).

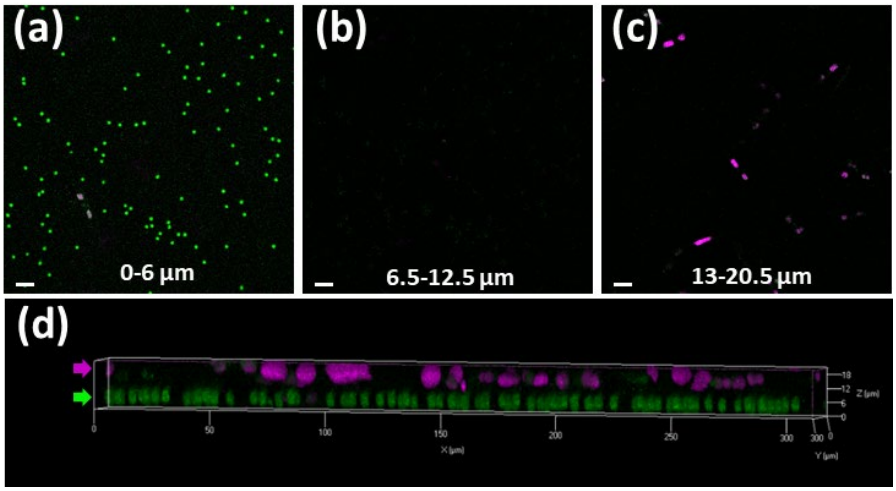


Figure S13. Control live cell imaging experiments. Confocal microscopy images of *C. australis* cells and Protein G beads after 12 days. Cells were visualized without any labelling, fluorescence signal is given only from autofluorescence of chloroplasts. **(a-d)** Green color shows the Protein G beads labelled with the secondary antibody, the magenta color represents the autofluorescence **(a)** maximum intensity Z-projection from the surface (0 µm) to 6 µm **(b)** maximum intensity Z-projection from 6.5 to 12.5 µm **(c)** maximum intensity Z-projection from 13µm to the top of the biofilm (20.5 µm) **(d)** The 3D reconstruction of the confocal images depics the location of the Protein G beads located underneath the biofilm. 3D reconstruction: **45 µm (X) x 45 µm (Y) x 10 µm (Z)**. Scale bar: 10 µm

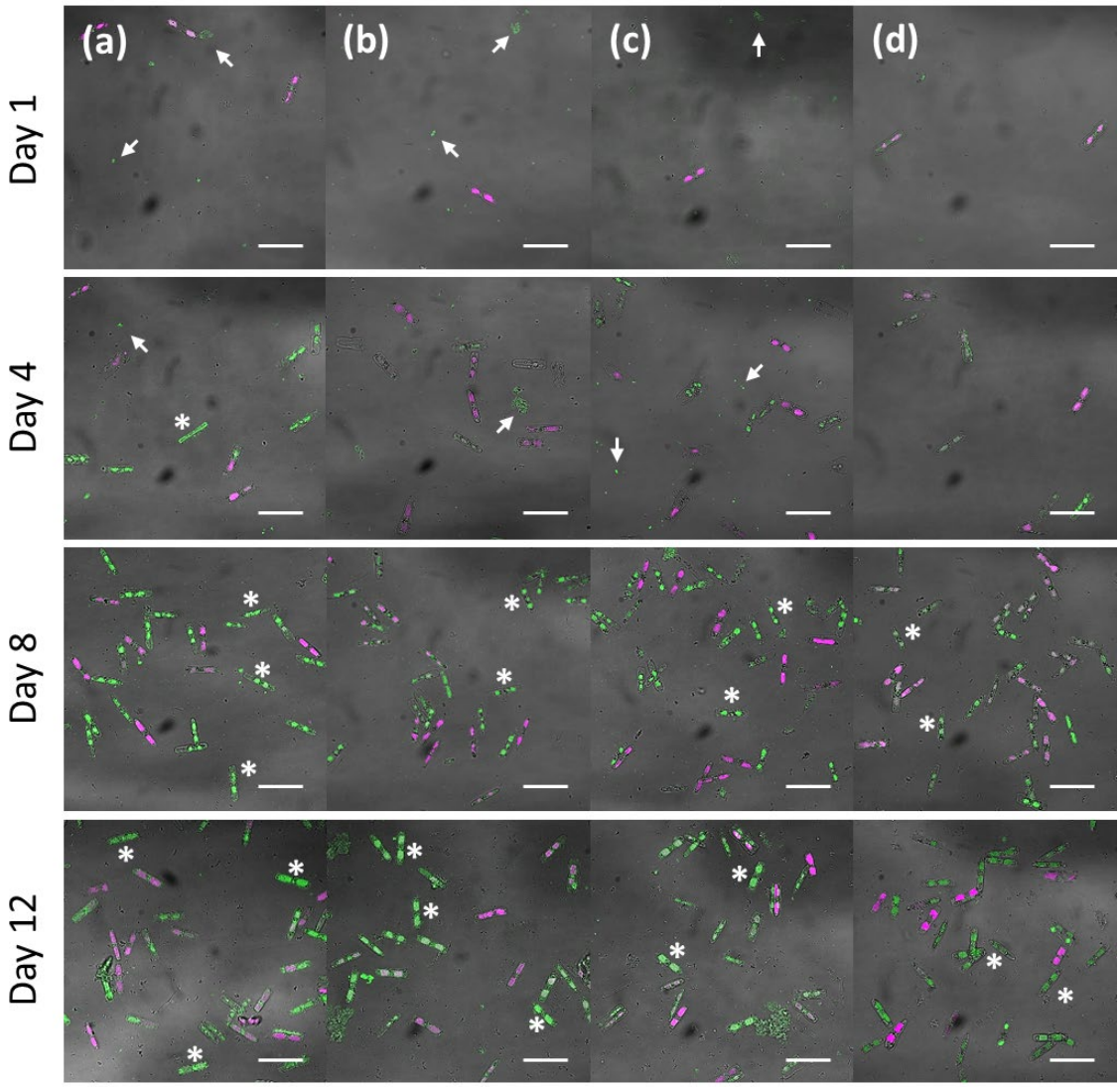


Figure S14. Immunofluorescence control experiments. Confocal microscopy images of *C. australis* adhesive material after 1, 4, 8 and 12 days. Samples were incubated with IgG purified pre-immune rabbit serum substituted for primary antibody αCaTrailin3 (**a**), αCaTrailin4 (**b**) and αCaTrailin2 (**c**). As a negative control, the cells were incubated only in blocking solution without the presence of primary antibody (**d**). All confocal images are presented as a merge of green, magenta and brightfield channel after the maximum intensity Z-projection. The red color represents chloroplast autofluorescence. Some of the dead cells in the biofilm are indicated with an asterisk and some non-specific clumps are depicted with arrows. Scale bar: 50 μm.

References

Armenteros JJA, Tsirigos KD, Sønderby CK, Petersen TN, Winther O, Brunak S, von Heijne G, Nielsen H. 2019. SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nature Biotechnology* **37**: 420-423.

Aumeier C. 2015. *The Cytoskeleton of Diatoms : Structural and Genomic Analysis.* PhD thesis, Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn, Germany.

Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. *Nature Methods* **12**: 59-60.

de Castro E, Sigrist CJ, Gattiker A, Bulliard V, Langendijk-Genevaux PS, Gasteiger E, Bairoch A, Hulo N. 2006. ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Research* **34**: W362-365.

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.

Erdos G, Pajkos M, Dosztanyi Z. 2021. IUPred3: prediction of protein disorder enhanced with unambiguous experimental annotation and visualization of evolutionary conservation. *Nucleic Acids Research* **49**: W297-W303.

Finn RD, Attwood TK, Babbitt PC, Bateman A, Bork P, Bridge AJ, Chang HY, Dosztanyi Z, El-Gebali S, Fraser M, et al. 2017. InterPro in 2017-beyond protein family and domain annotations. *Nucleic Acids Research* **45**: D190-D199.

Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, et al. 2016. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Research* **44**: D279-285.

Fu LM, Niu BF, Zhu ZW, Wu ST, Li WZ. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**: 3150-3152.

Grigoriev IV, Hayes RD, Calhoun S, Kamel B, Wang A, Ahrendt S, Dusheyko S, Nikitin R, Mondo SJ, Salamov A, et al. 2021. PhycoCosm, a comparative algal genomics resource. *Nucleic Acids Research* **49**: D1004-D1011.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* **30**: 772-780.

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.

Kumar S, Stecher G, Li M, Knyaz C, Tamura K. 2018. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Molecular Biology and Evolution* **35**: 1547-1549.

Letunic I, Bork P. 2018. 20 years of the SMART protein domain annotation resource. *Nucleic Acids Research* **46**: D493-D496.

Marron AO, Ratcliffe S, Wheeler GL, Goldstein RE, King N, Not F, de Vargas C, Richter DJ. 2016. The evolution of silicon transport in eukaryotes. *Molecular Biology and Evolution* **33**: 3226-3248.

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.

Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution* **32**: 268-274.

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.

Paradis E, Schliep K. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**: 526-528.

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. 2004. Silica morphogenesis by alternative processing of silaffins in the diatom *Thalassiosira pseudonana*. *Journal of Biological Chemistry* **279**: 42993-42999.

Shen W, Ren H. 2021. TaxonKit: A practical and efficient NCBI taxonomy toolkit. *Journal of Genetics and Genomics* **48**: 844-850.

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.

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.

Xu SB, Dai ZH, Guo PF, Fu XC, Liu SS, Zhou L, Tang WL, Feng TZ, Chen MJ, Zhan L, et al. 2021. ggtreeExtra: Compact visualization of richly annotated phylogenetic data. *Molecular Biology and Evolution* **38**: 4039-4042.

Yu GC, Smith DK, Zhu HC, Guan Y, Lam TTY. 2017. GGTREE: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution* **8**: 28-36.