nature climate change

SUPPLEMENTARY INFORMATION

DOI: 10.1038/NCLIMATE1989

The impact of temperature on marine phytoplankton resource allocation and metabolism

Andrew Toseland[1], Stuart Daines[2], James Clark[2], Amy Kirkham[3], Jan Strauss[3], Christiane Uhlig[4], Timothy M. Lenton[2], Klaus Valentin[4], Gareth Pearson[5], Vincent Moulton[1], Thomas Mock[3]

Author affiliations:

[1] School of Computing Sciences, University of East Anglia, Norwich Research Park, Norwich, UK

[2] College of Life and Environmental Sciences, University of Exeter, UK

[3] School of Environmental Sciences, University of East Anglia, Norwich Research Park, Norwich, UK

[4] Alfred-Wegener Institute for Polar and Marine Research, Bremerhaven, Germany

[5] Centre of Marine Sciences, University of the Algarve, Portugal

DOI: 10.1038/NCLIMATE1989

Contents

*Tables and figures for the modelling section of supporting information are embedded within the main text.

Supplementary Materials and Methods (Sampling, Molecular Biology & Bioinformatics)

1 Sampling

1.1 EPAC (Equatorial Pacific)

Samples were taken on two stations (EPAC1: 0°, 155°W; EPAC2: 0°, 140°W) (Fig. S1) during a cruise to the equatorial Pacific Ocean from 15th to 2nd of October 2006 onboard the RV 'Kilo Moana'.

1.2 NPAC (North-East Pacific, Puget Sound)

Samples were taken on one station (NPAC: 47°55.19 N; 122°20'38 W) (Fig. S1) during a Puget Sound cruise on the 15th of August 2007 onboard the 'Sorcerer' (Craig Venter Institute, US). Water for RNA samples was pumped from about 8m depth onboard with a hose and peristaltic pump (Table S1). Cells were immediately filtered onto autoclaved Nucleopore filters (25mm) with a pore size of 2µm. Not more than 500ml were filtered at a time in order to keep the filtration time <5 minutes per filter. Filters were subsequently flash frozen in liquid nitrogen and stored in the laboratory at -80°C. Phytoplankton were collected and concentrated by net tows from about 10m depth to the surface, using 0.25 m diameter nets with a mesh size of 10 µm (Research Nets Inc. Redmond, WA, USA).

1.3 ANT (Southern Ocean, Weddell Sea)

Samples were taken on two stations (ANT1: 65°06.11 S, 57°23.55 W; ANT2: 60°07.11 S, 47°54.55 W) (Fig. S1) during the WWOS (Winter Weddell Outflow Study) cruise in Austral summer 2006 with the German Icebreaker 'Polarstern'. Samples on ANT1 were obtained by icecore drilling and collecting microorganism communities from the lowermost cm (ice-waterinterface) of the ice core (Table S1). For RNA extraction ice samples were melted in or washed with prefiltered (0.2µm) sea

DOI: 10.1038/NCLIMATE1989

water or brine and cells were subsequently filtered onto Isopore filters (Millipore) (25mm) with a pore size of 1.2µm. Filters were subsequently flash frozen and stored in liquid nitrogen. Samples on ANT2 were obtained by fishing ice floes and collecting microorganisms from the ice -water interface as done on ANT1.

1.4 ARC (Arctic) & NATL (North Atlantic)

Phytoplankton community samples were taken in June 2009 on board the RV "Jan Mayan". Water samples at the DCM were taken directly from the CTD rosette (12.5 L Niskin bottles) in waters characterised as Arctic (June 20 SW Spitsbergen at 76' 36"N; 18' 11"E, temperature -1ºC at 35 m) and Atlantic influenced (June 16 at the Polar front south of Bear Island at 73' 55"N; 18' 46"E, temperature +2.1ºC at 50 m). Cells were collected by filtration on 5 µm pore-etched polycarbonate filters, flash–frozen in liquid nitrogen, and stored in a cryoshipper for transport to the laboratory.

2 Temperature, nutrients, chlorophyll a

2.1 EPAC (Equatorial Pacific)

Physical properties (e.g. temperature) were measured with a Seabird 911+ conductivity, temperature, and depth (CTD) profiler (1). Nutrients samples were frozen at -20°C until onshore analysis. Within 2 months after the cruise, the dissolved inorganic N was determined using an Astoria Autoanalyzer. [Si(OH)4] in ambient sea water was measured on board spectrophotometrically (2). Chl a was determined onboard ship by extraction with 90% acetone at -20°C for 24h and measured by in vitro fluorometry on a Turner Designs Trilogy fluorometer using the acidification method (3).

2.2 NPAC (North-East Pacific, Puget Sound)

Temperature of freshly collected water was measured onboard using a mercury thermometer. Water

for nutrient measurements was collected in sterile 15mL Falcon tubes and subsequently frozen at -20°C. Nutrient analysis ([Si(OH)4], NO3, PO4) was conducted using a Technicon Autoanalyzer Model AAII within 3 months after sampling (Table S4). The Hansville buoy (ORCA) at 47°54.44”N and 122°37.62W was used to obtain oceanographic profiles of T, S, density, O2, and in situ fluorescence close to the sampling site (Fig. S1) We retrieved data for 7 profiles from the surface to 20m depth on the 15th of August 2007 (Fig. S8). Four profiles were measured before sampling, one at the time of sampling and 2 afterwards. These data allowed to reconstructing the development of a phytoplankton bloom close to the sampling site. We acknowledge Al Devol and Wendi Ruef for making these data available to us.

2.3 ANT (Southern Ocean, Weddell Sea)

Temperature at the sea-ice-water interphase was measured with a Testo 720 thermometer with PT-100 sensor. Nutrients ([Si(OH)4], NO3, PO4) were measure on melted sea ice onboard ‘Polarstern’ using an Autoanalyser. Chl a was determined onboard ship by extraction with 90% acetone at -20°C for 24h and measured by in vitro fluorometry on a Turner Designs Trilogy fluorometer using the acidification method (3) (Table S4).

3 Independent taxonomic identification of dominant eukaryotic phytoplankton

3.1 NPAC (North-East Pacific, Puget Sound)

Cells from net tow samples were fixed with 1% Lugol (final concentration) and counted with an Olympus BX43 microscope with DIC optics at ×100 and x400 magnification. A magnification of x1000 (oil immersion) was used for species identification (Table S5).

4 Molecular biology

DOI: 10.1038/NCLIMATE1989

4.1 Metatranscriptome (RNA, cDNA, sequencing)

4.1.1 ANT (Antarctic), EPAC (Equatorial Pacific) & NPAC (North Pacific)

Several samples per station and ecosystem (EPAC, NPAC, ANT) were filtered onto 2µ pore size filters and subsequently flush frozen in liquid N and stored at -80˚C. RNA extraction was performed with the ToTally RNA extraction kit (Ambion) according to the manufacturers recommendation. Eukaryotic mRNA was extracted with the Oligotex mRNA purification kit (Qiagen). The same kit was used to do an additional purification with the purified mRNA from the first time to reduce the contamination by rRNA and bacterial mRNA. Due to a very limited amount of double purified eukaryotic mRNA, we pooled all samples from each ecosystem (EPAC, NPAC, ANT). CDNA synthesis on the pooled samples was conducted with the SuperSmart PCR cDNA kit (Clontech) according to manufacturers recommendations. Libraries for next generation sequencing were constructed according to protocols for Roche 454 GS-FLX and GS-Titanium sequencing. GS-FLX sequencing was done at the NERC sequencing facility in Liverpool (UK) and GS-Titanium sequencing was done at Roche 454 (US).

4.1.2 ARC (Arctic) & NATL (North Atlantic)

Extraction of total RNA from replicate filters was performed following standard protocols and a commercial kit (RNAeasy, Qiagen). Synthesis of full-length double-stranded cDNA (ds-cDNA) was performed from 250 ng of total RNA of each sample (SMARTer PCR cDNA Synthesis Kit; Clontech) according to the manufacturer's instructions, allowing synthesis of full-length transcripts while maintaining the gene representation of unamplified samples. Full-length single-stranded DNA templates were then amplified by long-distance PCR using the Advantage 2 PCR Kit (Clontech). Replicate PCR reactions were performed for each library in order to obtain the amount required for sequencing (3 – 5 µg), and subsequently pooled and purified using the MiniElute PCR Purification kit (Qiagen). The cDNA libraries were quantified using NanoDrop (ThermoScientific), and the

quality of final samples was verified using agarose gel electrophoresis. Libraries were sequenced by a commercial service provider (BioCant, Portugal) using 454 FLX Titanium chemistry.

4.2 Fragilariopsis cylindrus culture experiments and ribosomal gene expression

4.2.1 Culture conditions

Fragilariopsis cylindrus (Grunow) Krieger CCMP1102 was obtained from the Provasoli-Guillard National Centre for Marine Algae and Microbiota (NCMA, https://ncma.bigelow.org/, West Boothbay Harbor, USA, formerly CCMP). All cultures were grown and maintained in filtersterilised (0.2 µm pore size) Aquil medium (4) at 4°C under continuous illumination at a photon flux density of 35 µmol photons m-2 s-1. Cultures of *F. cylindrus* were handled under strict sterile conditions and potential bacterial contamination was eliminated as stock cultures were subjected to a multi-antibiotic treatment with Ampicillin (50 µg mL-1), Gentamycin (1 µg mL-1), Streptomycin (25 µg mL-1), Chloramphenicol (1 µg mL-1) and Ciprofloxacin (10 µg mL-1) (5). Epifluorescence microscopy was used to confirm axenic cultures using 4',6-diamidino-2-phenylindole (DAPI) fluorescent nucleic acid staining before the beginning of the experiment. During exponential growth, stock cultures were used to inoculate triplicates of 2L experimental batch cultures for optimal (+4°C), high (+10°C) and low (-2°C) temperature treatments. Bubbling with sterile filtered air and shaking of the culture bottles ensured sufficient CO2 supply and mixing during experimental treatments. Experimental cultures were grown to mid-exponential phase (approximately 500,000 cells mL-1) at +4°C before temperatures were amended to the final experimental temperature (+10°C to -2°C). Subsamples were taken on a daily basis throughout the experiment to determine the maximum quantum yield of photosystem II (Fv/Fm) by pulse-amplitude-modulated fluorometry (Phyto-PAM fluorometer, Walz GmbH, Effeltrich, Germany) and cell counts (Multisizer 3 particle counter, Beckman Coulter, Brea, USA).

DOI: 10.1038/NCLIMATE1989

4.2.2 RNA extraction and purification.

Cells were harvested on the third day after the cultures reached the experimental temperatures using 1.2 µm membrane filters (Isopore Membrane, Millipore, MA, USA). The volume of F. cylindrus culture per filter sample was recorded to calculate the number of cells per filter sample. Total RNA was extracted using TRI Reagent (Sigma-Aldrich, St. Louis, USA) followed by DNase I (Qiagen, Hilden, Germany) treatment (1h, 37°C) and purification using RNeasy MinElute Cleanup Kit (Qiagen, Hilden, Germany). Purity of RNA was checked on a NanoDrop (Thermo Fisher Scientific, Waltham, USA) and integrity using denaturing 2% formaldehyde gels. Concentrations after RNA cleanup were determined in duplicate readings using a NanoDrop.

4.2.3 Total RNA concentration per *Fragilariopsis cylindrus* cell as a function of growth temperature.

The total RNA yield obtained for each filter sample was used to calculate the RNA concentration per cell. Therefore the total RNA yield per filter was divided by the number of cells in each filter sample. The obtained total RNA concentration per cell was plotted as a function of growth temperature and used for a linear regression analysis.

4.2.4 Reverse transcription, primer design, and Q-PCR conditions.

First strand synthesis was performed using Superscript II reverse transcriptase (Invitrogen, Carlsbad, USA) utilising Anchored Oligo(dT)20 Primer (Invitrogen, Carlsbad, USA). Reverse transcription (RT) of 500 ng of total RNA was carried out according to manufacturer's recommendations in 20µL reactions at 42°C for 50 minutes, followed by inactivation at 70°C for 15 minutes. Immediately prior to transcription all reaction mix was spiked with artificial RNA of the major allergen (MA) gene of the butterfly Pieris rapae (cabbage white butterfly, *Lepidoptera: Pieridae*) to verify efficiency of RT reactions and to provide an exogenous control. The MA gene

provides an ideal control, because few insects are present in the marine environment and P. rapae is considered alien to polar or marine diatoms. The control gene MA was constantly detected in all samples at a cycle threshold of 36.14 (±0.17, n=12) indicating consistent efficiency of the RT reaction. As a control for DNA contamination, RNA was pooled from each biological replicate and first strand synthesis reaction mix was added omitting reverse transcriptase. Oligonucleotides (Table S8) were designed towards the 3' end of the gene of interest using the webbased RealTimeDesign Software (available at http://www.biosearchtech.com/realtimedesign, Biosearch Technologies, Novato, USA). BLAST searches of the primer sequences against the *F. cylindrus* genome sequence (http://genome.jgi-psf.org/Fracy1/Fracy1.home.html) were performed to check for target specificity and if necessary primer sequences were modified manually. Oligonucleotides were assessed for melting temperature, hairpins, and primer dimers using the webbased tool OligoAnalyzer 3.1 (available at http://eu.idtdna.com/analyzer/Applications/OligoAnalyzer; Integrated DNA Technologies, Coralville, USA) and synthesised by Eurofins MWG Operon (Ebersberg, Germany). For second strand amplification, 5 µL of a 10-fold diluted RT reaction mix was supplemented with 20 µL 2× SensiMix SYBR Green NoROX Master Mix (Bioline, London, UK). Each primer was added at a concentration of 200 nM. Amplifications were performed in white 96-well plates on a CFX96 Real Time System (Bio-Rad, Hercules, USA) using the following conditions: initial denaturation 95°C, 10 minutes, followed by 40 amplification and quantification cycles of 15 seconds at 95°C, 15 seconds at 59°C, 10 seconds at 72°C. Finally, a melting curve analysis (65°C to 95°C, increments of 0.5 °C, dwelling time 5 seconds) was carried out to check for primer dimers and non-specific amplification.

4.2.5 qPCR data analysis.

The cycle thresholds were automatically determined using the CFX Manager Software Version 1.1 (Bio-Rad, Hercules, USA). The REST-MCS© software (available at http://rest.genequantificationinfo/) was used to test the expression of target genes under both

DOI: 10.1038/NCLIMATE1989

experimental conditions normalised by a reference gene index containing the endogenous and exogenous controls TBP, RNAP II and MA and significances were tested by a Pair Wise Fixed Reallocation Randomisation Test using 2000 iterations.

4.3 Biochemical studies

4.3.1 Diatom Cultures

Diatom cultures (*Thalassiosira pseudonana* CCMP1335 and *Fragilariopsis cylindrus* CCMP1102) were grown in artificial seawater (NEPC) under 24 hours light at 100 µE. Photosynthetic health was estimated using phytoPAM ED (WALZ) spectrometer. Healthy cultures (fv/fm >0.6) were transferred to incubators (Sanyo) at the experimental temperatures and allowed to acclimatise for 24 hours before 25000 cells/ml was diluted into new media that was pre-warmed or cooled to the experimental temperature. The cultures' growth was monitored daily using a coulter counter (Beckman).

4.3.2 Western blots

100 ml of mid-exponential phase culture was pelleted by centrifugation and total proteins were extracted by adding 50 µl of lysis buffer (50 mM Tris pH 6.8, 2% SDS) to the cell pellet. Cell lysates were incubated at room temperature for 30 min before separation from cellular debris by centrifuging at 10,000 g at 4 °C for 30 minutes. 35 µg protein extracts were resolved on 12.5 % SDS-PAGE gels and transferred to nitrocellulose transfer membranes using criterion blotter (BioRad). Loading was checked by incubating the membrane with the protein stain Ponceau S for 20 minutes at room temperature. The S14 ribosomal protein was hybridised with 1:1000 dilution of S14 antibody (AS09 477, Agrisera), for 1 hour at room temperature, followed by a 1:10,000 dilution of horseradish peroxidase (HRP)-conjugated goat anti-rabbit secondary antibody (Promega) in 1×PBS, 1% milk, 0.1% Tween 20. Signals were visualised using the enchanced chemiluminescence

(ECL) kit (Amersham Biosciences) and Lasimager 2000 software (Fuji).

4.3.3 Translation efficiency experiment

A transgenic *T. pseudonana* strain expressing a novel gene fused to eGFP on the inducible nitrate reductase (NR) promotor was grown in NEPC containing 550 µM NH4Cl as the sole nitrogen source, repressing the expression of the transgene. Mid-exponential phase cells were transferred to nitrogen-free NEPC for one generation time (approx. 12h at 20°C, 24h at 11°C and 48h at 4°C) before 550 µM NaNO3 was added to activate the NR cassette. Triplicate cultures were monitored using a FACSCalibur (Beckton Dickinson) Flow Cytometer. Cells were discriminated by plastid red autofluorescence versus the eGFP green fluorescence arising when the eGFP was translated. Populations were gated and percentage of total population calculated. The lag phase was calculated as the first time point at which the percentage of cells in the eGFP gate was found to be significantly higher than the T0 measurement (t-test, $p<0.05$). Translation efficiency (m) was calculated as the slope of the curve after the lag phase.

5 Bioinformatics

5.1 Quality filtering

Quality clipping was performed as in Marchetti et al. (6) using a single base sliding window each sequence was trimmed from 3' to 5' until a base is reached with a Phred quality score of >=14 is met. To identify potential sequencing artifacts, all sequences were clustered with CD-HIT-est (7) at 100% requiring 100% coverage of both sequences. Only the cluster representatives were retained, cluster members (exact duplicates) were deemed potential artifacts omitted and from further processing. The 5' primer (AAGCAGTGGTATCAACGCAGAGT) was detected using PatMan (8) allowing up to 4 mismatches and 2 gaps, match coordinates from PatMan were used to trim primer regions using a custom BioPerl script. The 3', 17 base oligo-dt primer was identified using Dust

DOI: 10.1038/NCLIMATE1989

(word size 2, complexity value of 50) to get the coordinates of low complexity regions. Each identified region was examined and, if it was of an appropriate length (>= 15 bases) and composed of >=75% adenine or >=75% thymine the region was trimmed out. Low complexity sequences were identified using Dust. Using default parameters, sequences were run through Dust and low complexity regions masked with Xs. The proportion of masked bases for each sequence was calculated and sequences comprising of >= 70% low complexity region were filtered out. Finally any sequences less than 50bp in length were removed.

Despite specifically targeting eukaryotic mRNA by attaching oligo-dt primers to the poly-A region of transcripts, some non mRNA may be present in the samples. John et al. (9) reported that ~2% of sequences of a small scale eukaryotic metatranscriptome held significant similarity to ribosomal rna (rRNA). In order to detect putative rRNA sequences we performed blastn (10) searches (Default settings, no complexity filtering) against both the large and small subunit databases of the Silva ribosomal database (11). Sequences returning hits with bit scores ⩾50 were deemed putative rRNA and excluded from further analysis. The final processing stage was to cluster sequence sets to remove redundancy and speed up homology searches. Each sequence set was clustered using CD-HI-est at >=95% overall identity and requiring >=50% coverage of the representative sequence. A lookup table of cluster details (Cluster representative ID, Cluster size, Cluster member Ids) was created in order to scale the annotation results of cluster representatives accordingly.

5.2 Exploring data set composition through sequence clustering

Sequences from all 5 data sets had an environment specific prefix added to their accession and were pooled together. All sequences were then translated into their longest open reading frames (minimum length >= 10 amino acids) and clustered with CD-HIT (>=60% overall identity, >=50% coverage of the representative sequence). Using a custom Perl script, the resulting clusters were examined individually and the sequence ids of all cluster members were appended to a list for each environment involved in that cluster. The resulting lists were used to create the 5 group venn

diagram in R using the 'venn' function of the gplots package. Lists of sequence ids for each of the 31 sections of the venn diagram were produced using R set operators for downstream analysis.

5.3 PhymmBL taxonomic affiliations

The taxonomic composition of the samples were determined using PhymmBL (12), a hybrid classifier which combines BLAST alignments with nucleotide composition based interpolated markov models. By default, PhymmBL uses bacterial and archaeal genomes from NCBI GenBank (13) as a reference. It is however, extensible and has been successfully applied to eukaryotic data (14). We created a representative set of 44 eukaryote organisms using genomes and ESTs covering the major eukaryote groups but with a focus on algal species for this analysis (See Table S3 for list of organisms used and taxonomic labels). Genome sequences were downloaded from NCBI GenBank and JGI (with 4 exceptions: *Cyanidioschyzon merolae* from Cyanidioschyzon merolae Genome Project http://merolae.biol.s.u-tokyo.ac.jp/download; *Strongylocentrotus purpuratus* from Sea Urchin Genome Project http://www.hgsc.bcm.tmc.edu/projectspecies-o- Strongylocentrotus; *Danio rerio* from UCSC http://genome.ucsc.edu/cgi-bin/hgGateway?db=danRer5; and *Homo sapiens* from Genome Reference Consortium http://hgdownload.cse.ucsc.edu/goldenPath/hg19/chromosomes/). EST sequences were downloaded from NCBI-dbEST and clustered with CD-HIT-est at 95% similarity to ensure non-redundancy of sequences. Taxonomic classifications for the PhymmBL configuration file were taken from the NCBI taxonomy (15) and AlgaeBase (16). The sequence files and taxonomic details were added to PhymmBL in batch mode and IMMs created for each new organism. PhymmBL results were filtered with a confidence score cutoff of $\geqslant 0.9$ at the phylum level**.** Subsets of sequences matching to the phyla Bacillariophyta and Dinoflagellata were extracted for further analysis.

5.4 Pfam

Cluster representative sequences were translated into all 6 reading frames (Min length 10aa) and

DOI: 10.1038/NCLIMATE1989

homology searches against the Pfam protein database (17) performed using pfam_scan.pl (Pfam-A only, default gathering thresholds used). Results filtered using custom Perl script to detect best match(es), remove conflicting matches across different reading frames and scale results to cluster size.

5.5 GO Term Enrichment & Term Clouds

All detected Pfam domains were mapped to their corresponding GO term(s) (18) using a custom Perl script and the mapping file Pfam2Go (http://www.geneontology.org/external2go/pfam2go). Then, for each possible pair of environments we performed a Fisher's exact test on each GO term. Enriched GO terms were identified using a Bonferroni corrected p-value <0.001 and used to create term clouds (One for each environment in the pairwise comparison). Lists of enriched GO terms were created - one for each environment, with the frequency of a GO term in the list determined by the absolute difference in the normalised abundance of the term between the two environments. The term clouds were created with Worditout.com using direct colour blending from blue (low frequency) to red (high frequency). See Fig. S12 for example term cloud.

5.6 KEGG

KEGG pathways (19) for cluster representative sequences were identified using the KEGG/KAAS web-server (20) (Using single-directional best hit EST mode against a eukaryote representative gene set, bit-score cut-off >=40). The resulting KO (Kegg Orthology) list were scaled by cluster size and filtered using MinPath (21) to get a minimal set of pathways. Hits for KEGG pathways K000230: Purine metabolism and K000240: Pyrimidine metabolism were summed and plotted against temperature for each environment

5.7 CCA

Canonical Correspondance Analysis was performed using the VEGAN package in R. We treated the

transpose of the normalised Pfam count tables as our species data and created a second table of environmental factors: temperature, salinity, latitude, longitude, nutrient levels etc. Where environmental data was unavailable we used the World Ocean Atlas (http://www.nodc.noaa.gov/OC5/SELECT/woaselect/woaselect.html for nutrient levels, taking the annual surface mean measurements). For light levels we used the Pangaea information system website (htpp://www.pangaea.de) to find in-situ PAR readings over a depth gradient for environments analogous to our samples. By plotting PAR against depth and fitting an exponential regression line we could extract the equation for the PAR – depth relationship and plug in our depth measurement to get an estimated PAR. The data sets used were:

ANT: Nicolaus, M et al. (2012): Downward spectral solar irradiance as measured in different depths under sea ice (transmitted irradiance) at sea ice station PS78/267-1. doi:10.1594/PANGAEA.786857,

EPAC: Eldin, Gerard; Rodier, Martine; Dupouy, D (2004): Physical oceanography at CTD station FLUPAC_119. doi:10.1594/PANGAEA.186766

NATL & ARC: Fosså, Jan Helge; Kutti, Tina; Bergstad, Odd Aksel; Knutsen, Tor; Svellingen, Ingvald; Wangensten, Jarle; Johannessen, Reidar; Steinsland, Asgeir (2011): Physical oceanography during R/V H. Mosby cruise IMR-2009615. Institute of Marine Research, Bergen, doi:10.1594/PANGAEA.756308

NPAC: Whitney, Frank (2002): Physical oceanography at station IOS_97-11_CTD045. doi:10.1594/PANGAEA.79563

In the case of ANT and EPAC where there were multiple samples we took the mean values. All environmental data was log2 transformed and an offset added to temperature values to make them positive. To highlight specific proteins for nitrate reductases, fucoxanthin chlorophyll binding proteins (FCPs), ribosomal proteins and silicon transporters we took one gene of each type from 3

DOI: 10.1038/NCLIMATE1989

diatoms: *Thalassiosira pseudonana*, *Phaeodactylum tricornutum* and *Cylindrotheca fusiformis* (From NCBI RefSeq/GenBank see Table S9). Each gene was compared to Pfam-A (gathering threshold cutoff) and the detected domains used to represent that gene. One hundred percent of the total inertia (1.563) was explained by the set of environmental constraints (temperature, light, nitrate and phosphate). The four CCA dimensions accounted for 0.58310 (37.31%), 0.49761 (31.84%), 0.41433 (26.50864) and 0.06816 (4.36%) respectively.

Correlations between environmental factors and the normalised abundance of hits to the GO term for translation was performed using the Pairs function in R.

5.8 Heatmaps

All heatmaps were created using the Heatmap.2 function in R. For the taxonomy heatmap, only phymmBL classified algal groups were used. The percentage of hits to each phyla were read in as a table and used to create two correlation matrices (One for the table and one for it's transpose). Distances were measured as 1 – Pearson correlation coefficient between rows/columns and the matrices were used to create row and column dendrograms (complete linkage clustering). The abundance data was finally scaled and centred by column. For GO terms the heatmap was created as above, but only biological process GO terms over an abundance cutoff (>=0.5% of hits in at least one data set).

5.9 Rarefaction

Rarefraction curves were produced with the online Rarefaction tool (http://www.biology.ualberta.ca/jbrzusto/rarefact.php#Calculator) using Chao's estimator for species richness. A list of raw totals for each detected Pfam domain were entered (plus the number of sequences providing no hits) and sampled at 50,000 sequence intervals..

5.10 *Thalassiosira pseudonana* transcriptome

As ~60% of sequences in the NPAC sample had taxonomic affiliations with *T. pseudonana* we chose this data set to perform a comparison with expression data from a *T. pseudonana* genomewide microarray experiment (22). First a spreadsheeet was compiled of differentially expressed (log2 fold change >= 1, p-value <0.05) *T. pseudonana* genes (Table S6) and expression values under low temperature (4°C), and silicate, nitrate, iron and CO2 limitation. Columns were added to each gene for GO, KEGG, KOG (annotations taken from JGI http://genome.jgipsf.org/Thaps3/Thaps3.download.ftp.html)) and Pfam annotations (performed ourselves, search against Pfam-A, gathering threshold cutoff). Sequences from the NPAC sample classified as Bacillariophyta (PhymmBL phylum confidence score >= 0.9) were extracted and BLASTed against the JGI T. pseudonana gene models (BLASTx, e-value <= 1e-5, using soft masking, requiring >=50% coverage of the query and >= 75% overall identity and taking the single best hit). Total matches to each differentially expressed gene were added to the table.

Supplementary References

1. Marchetti, A. et al. Iron and silicic acid effects on phytoplankton productivity, diversity, and chemical composition in the central equatorial Pacific Ocean. *Limnology and Oceanography*, ***55*(1)**, 11 (2010).

2. Strickland, J.D.H. Parsons, A practical handbook of seawater analysis. *Fisheries Research Board of Canada*. (1972).

3. Parsons, T.R., Maita Y., Lalli, C.M. *Manual of chemical and biological methods for seawater analysis*. Pergamon (1984).

4. Morel, F.M., Rueter, J.G., Anderson, D.M., Guillard, R.R.L. AQUIL: A CHEMICALLY DEFINED PHYTOPLANKTON CULTURE MEDIUM FOR TRACE METAL STUDIES12. *Journal of Phycology*, ***15*(2)**, 135-141 (1979).

5. Jaeckisch, N. et al. Comparative genomic and transcriptomic characterization of the toxigenic marine dinoflagellate Alexandrium ostenfeldii. *PloS one*, ***6*(12)**, e28012 (2011).

6. Marchetti, A. et al. Comparative metatranscriptomics identifies molecular bases for the physiological responses of phytoplankton to varying iron availability. *Proceedings of the National Academy of Sciences*, ***109*(6)**, E317-E325 (2012).

7. Li, W., Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, ***22*(13)**, 1658-1659 (2006).

8. Prüfer, K. et al. PatMaN: rapid alignment of short sequences to large databases. *Bioinformatics*, ***24*(13)**, 1530-1531 (2008).

9. John, D.E., Zielinski, B.L., Paul, J.H. Creation of a pilot metatranscriptome library from eukaryotic plankton of a eutrophic bay(Tampa Bay, Florida). *Limnology and Oceanography: Methods*, *7*, 249-259 (2009).

10. Altschul, S. F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, ***25*(17)**, 3389-3402 (1997).

11. Pruesse, E. et al. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic acids research*, ***35*(21)**, 7188-7196 (2007).

12. Brady, A., Salzberg, S.L. Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nature methods*, ***6*(9)**, 673-676 (2009).

13. Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Wheeler, D.L. GenBank. *Nucleic acids research*, ***34***(suppl 1), D16-D20 (2006).

14. Brady, A., Salzberg, S. PhymmBL expanded: confidence scores, custom databases, parallelization and more. *Nature methods*, ***8*(5)**, 367-367 (2011).

15. Ncbi taxonomy. http://www.ncbi.nlm.nih.gov/Taxonomy/

16. Guiry, M.D., Guiry, G.M. AlgaeBase. *AlgaeBase (2008)*.

17. Finn, R.D., et al. The Pfam protein families database. *Nucleic acids research*, ***38***(suppl 1),

D211-D222 (2010).

18. Ashburner, M., et al. Gene Ontology: tool for the unification of biology. *Nature genetics*, ***25*(1)**, 25 (2000).

19. Ogata, H., et al. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic acids research*, ***27*(1)**, 29-34 (1999).

20. Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A.C., Kanehisa, M. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Research*, ***35***(suppl 2), W182-W185 (2007).

21. Ye, Y., Doak, T.G. A parsimony approach to biological pathway reconstruction/inference for genomes and metagenomes. *PLoS computational biology*, ***5*(8)**, e1000465 (2009).

22. Mock, T., et al. Whole-genome expression profiling of the marine diatom Thalassiosira pseudonana identifies genes involved in silicon bioprocesses. *Proceedings of the National Academy of Sciences*, ***105*(5)**, 1579-1584 (2008).

DOI: 10.1038/NCLIMATE1989

Supplementary Tables

Table S1: Geographical locations, method of sampling and water depth. ANT 1, 2: Southern Ocean; NPAC: North-East Pacific; EPAC 1, 2: Equatorial Pacific. NATL North Atlantic and ARC Arctic.

Location	Date	Latitude	Longitude	Sampling method	Water depth
ANT1	08/10/06	65°06.11 S	57°23.55 W	Ice core drilling	Ice-water interface
ANT2	23/09/06	60°07.11 S	47°54.55 W	Fishing ice floes	Ice- water interface
ARC	20/06/09	76°36N	18° 11E	Rosette	35m
EPAC1	27/09/06	0°	155° W	Rosette	10m
EPAC2	25/08/06	0°	140° W	Rosette	40m
NATL	16/06/09	73°55N	18°46E	Rosette	50m
NPAC	15/08/07	47°55.19 N	122°20.38 W	Membrane pump	8m

Table S2: Summary of 454 sequence data for Antarctic (ANT), Arctic (ARC), Equatorial Pacific (EPAC), North Atlantic (NATL), and North Pacific (NPAC) metatranscriptomes. [1]Only exact duplicates were removed: CD-HIT-est clustering at 100% identity requiring 100% coverage of both sequences. [2]Blastn against Silva SSU & LSU database – Best hit, no complexity filtering, bit-score cutoff ≥50. [3]CD-HIT-est clustering ≥95% overall identity, requiring ≥50% coverage cluster representative.

	ANT	ARC	EPAC	NATL	NPAC
# Raw Reads	391,614	514,223	342,252	513,985	313,910
Avg Length (bp)	168.1	278.1	158.9	310.7	258.0
Total Size (Mb)	65.83	143.03	54.4	159.67	81
Potential Artifacts[1]	49,093	3,175	21,942	5,172	14,172
Putative rRNA (%)[2]	3,595 (0.92%)	38,651 (7.52%)	1,324 (0.39%)	68,009 (13.23%)	1,254 (0.4%)
# Filtered Reads	220,844	421,107	246,534	394,187	250,841
Avg Length (bp)	209.3	252.1	161.0	285.6	268.2
Total Size (Mb)	46.22	106.18	39.69	112.58	67.26
GC%	43.43	43.82	47.30	43.99	44.44
# Clusters[3]	29,840	254,423	119,783	252,031	76,564

DOI: 10.1038/NCLIMATE1989

Table S3a: List of eukaryotic genomes and their taxonomic classifications added to PhymmBL reference database.

phylum	class	order	family	genus	species	strain
Bacillariophyta	Coscinodiscophyceae	Thalassiosirales	Thalassiosiraceae	Thalassiosira	pseudonana	CCMP1335
Bacillariophyta	Bacillariophyceae	Naviculales	Phaeodactylaceae	Phaeodactylum	tricornutum	CCAP1055/1
Bacillariophyta	Bacillariophyceae	Bacillariales	Bacillariaceae	Fragilariopsis	cylindrus	NO_VALUE
Ciliophora	Oligohymenophorea	Peniculida	Parameciidae	Paramecium	tetraurelia	sd4-2
Ciliophora	Oligohymenophorea	Hymenostomatida	Tetrahymenidae	Tetrahymena	thermophila	SB210
Apicomplexa	Coccidia	Eucoccidiorida	Cryptosporidiidae	Cryptosporidium	parvum	IowaII
Apicomplexa	Aconoidasida	NO_VALUE	Theileriidae	Theileria	annulata	Ankara
Apicomplexa	Aconoidasida	Haemosporida	NO_VALUE	Plasmodium	yoelii	17XNL
NO_VALUE	Lobosa	Amoebida	Entamoebidae	Entamoeba	histolytica	HM-1:IMSS
Mycetozoa	Dictyostelia	Dictyosteliida	NO_VALUE	Dictyostelium	discoideum	AX4
NO_VALUE	NO_VALUE	Choanoflagellida	Codonosigidae	Monosiga	brevicollis	MX1
Microsporidia	NO_VALUE	NO_VALUE	Unikaryonidae	Encephalitozoon	cuniculi	GB-M1
Basidiomycota	Tremellomycetes	Tremellales	Tremellaceae	Cryptococcus	neoformans	JEC21
Ascomycota	Schizosaccharomycetes	Schizosaccharomycetales	Schizosaccharomycetaceae	Schizosaccharomyces	pombe	972h-
Ascomycota	Pezizomycetes	Pezizales	Tuberaceae	Tuber	melanosporum	Mel28
Ascomycota	Dothideomycetes	Pleosporales	Phaeosphaeriaceae	Phaeosphaeria	nodorum	SN15
Ascomycota	Eurotiomycetes	Eurotiales	Trichocomaceae	Aspergillus	nidulans	FGSC_A4
Ascomycota	LeotiomyceteS	Helotiales	Sclerotiniaceae	Sclerotinia	sclerotiorum	1980_UF-70
Ascomycota	Sordariomycetes	Sordariales	Sordariaceae	Neurospora	crassa	OR74A
Ascomycota	Saccharomycetes	Saccharomycetales	Saccharomycetaceae	Saccharomyces	cerevisiae	S288c
Rhodophyta	Cyanidiophyceae	Cyanidiales	Cyanidiaceae	Cyanidioschyzon	merolae	10D

Chlorophyta	Mamiellophyceae	Mamiellales	Bathycoccaceae	Ostreococcus	lucimarinus	CCE9901
Chlorophyta	Chlorophyceae	Chlamydomonadales	Volvocaceae	Volvox	carteri	f_nagariensis
Chlorophyta	Chlorophyceae	Chlamydomonadales	Chlamydomonadaceae	Chlamydomonas	reinhardtii	NO_VALUE
Streptophyta	Liliopsida	Poales	Poaceae	Oryza	sativa	japonica
Streptophyta	NO_VALUE	Brassicales	Brassicaceae	Arabidopsis	thaliana	NO_VALUE
Nematoda	Chromadorea	Rhabditida	Rhabditidae	Caenorhabditis	elegans	NO_VALUE
Arthropoda	Branchiopoda	Diplostraca	Daphniidae	Daphnia	pulex	NO_VALUE
Arthropoda	Insecta	Diptera	Drosophilidae	Drosophila	melanogaster	NO_VALUE
Echinodermata	Echinoidea	Echinoida	Strongylocentrotidae	Strongylocentrotus	purpuratus	NO_VALUE
Chordata	Actinopterygii	Cypriniformes	Cyprinidae	Danio	rerio	NO_VALUE
Chordata	Mammalia	Primates	Hominidae	Homo	sapiens	GRCh37
Bacillariophyta	Pelagophyceae	Pelagomonadales	Pelagomonadaceae	Aureococcus	anophagefferens	NO_VALUE
Chlorophyta	Mamiellophyceae	Mamiellales	Mamiellaceae	Micromonas	pusilla	CCMP1545

DOI: 10.1038/NCLIMATE1989

Table S3b: List of eukaryotic EST libraries added to PhymmBL reference database. Sequences downloaded from NCBI DB-EST and clustered with CD-HIT to ensure non-redundancy.

phylum	class	order	family	genus	species	strain
Dinoflagellata	Dinophyceae	Gonyaulacales	Gonyaulacaceae	Alexandrium	catenella	NO_VALUE
Bacillariophyta	Coscinodiscophyceae	Chaetocerotales	Chaetocerotaceae	Chaetoceros	neogracile	NO_VALUE
Glaucophyta	Glaucophyceae	Glaucocystales	Glaucocystaceae	Cyanophora	paradoxa	NO_VALUE
Ochrophyta	Phaeophyceae	Ectocarpales	Ectocarpaceae	Ectocarpus	siliculosus	NO_VALUE
Haptophyta	Prymnesiophyceae	Isochrysidales	Noelaerhabdaceae	Emiliania	huxleyi	NO_VALUE
Cryptophyta	Cryptophyceae	Pyrenomonadales	Geminigeraceae	Guillardia	theta	NO_VALUE
Dinoflagellata	Dinophyceae	Peridiniales	Heterocapsaceae	Heterocapsa	triquetra	NO_VALUE
Dinoflagellata	Dinophyceae	Gymnodiniales	Gymnodiniaceae	Karenia	brevis	NO_VALUE
Dinoflagellata	Dinophyceae	Gymnodiniales	Gymnodiniaceae	Karlodinium	micrum	NO_VALUE
Haptophyta	Pavlovophyceae	Pavlovales	Pavlovaceae	Pavlova	lutheri	NO_VALUE

Table S4: Temperature (T), salinity (in PSU), photosynthetically active radiation (PAR) and major nutrients given in µmol/L. Chlorophyll a (Chl a) given in µg/L or expressed as in-situ fluorescence and light, day length. ANT 1, 2: Antarctic, ARC: Arctic, EPAC 1, 2: Equatorial Pacific, NATL: North Atlantic, NPAC: North-East Pacific. N.d. = no data available. S = Taken from "SPINDLER, Michael. NEOGLOBOQUADRINA PACHYDERMA FROM ANTARCTIC SEA ICE. Proc. NIPR Symp. Polar Biol. Vol. 9. (1996)." P = Data derived from depth-corrected, in-situ measurements from analogous environments from Pangea information system (See supplementary materials and methods for full details). W = Data derived from World Ocean Atlas (Annual mean surface measurements). L= Day lengths calculated using formula from (http://ocean.stanford.edu/courses/EESS151/).

Location	T [C°]	NO3	Si(OH)4	PO4	Salinity (PSU)	PAR (W/m2)[P]	Chl a (ug/L) or *in situ* fluorescence	Day Length (Hours)[L]
ANT1	-2	7.8	6.1	2.2	39.3[S]	0.02	93ug/L	13.73
ANT2	-2	n.d.	n.d.	n.d.	n.d.	0.02	n.d.	12.01
ARC	-1.1	5[W]	2.5[W]	0.5[W]	34.2	2.19	n.d.	24
EPAC1	27	4.72	2.42	0.5	35.3[W]	279.62	0.26 ug /L	11.97
EPAC2	27	4.4	1.88	0.5	35.3[W]	60.02	0.29 ug/L	11.97
NATL	2.1	5[W]	2.5[W]	0.5[W]	34.9	0.32	n.d.	24
NPAC	12	12.47	30.02	1.71	30	324.37	6.98 *in situ* fluorescence	14.1

DOI: 10.1038/NCLIMATE1989

Table S5: Taxonomic composition of major eukaryotic phytoplankton species in NPAC (North-East Pacific (Puget Sound)) on 15th of August 2006. N=3.

Dominant phytoplankton	**Cells / L**
Coscinodiscus walesii	5472 ±894
Chaetoceros spp. single cells	13105 ± 1983
Chaetoceros spp. chains	6280 ± 453
Thalassiosira spp.	91129 ± 7998
Thalassiosira nitzschioides	<1000
Pennate diatoms	<100
Dinoflagellates	<100
Unidentified flagellates	<1000