Figure #	Figure title One sentence only	Filename This should be the name the file is saved as when it is uploaded to our system. Please include the file extension. i.e.: *Smith_ED_Fig1.jpg*	Figure Legend If you are citing a reference for the first time in these legends, please include all new references in the main text Methods References section, and carry on the numbering from the main References section of the paper. If your paper does not have a Methods section, include all new references at the end of the main Reference list.
Extended Data Fig. 1	**Morphology and phylogenetic relationship of *Microglena* sp. YARC.**	Extended_Data_Fig1.pdf	**a, b,** Optical microscope (**a**) and transmission electron microscopy (**b**) of *Microglena* sp. YARC. **c,** A maximum likelihood tree of green algae based on the five chloroplast protein-coding genes (*atp*B, *psa*A, *psa*B, *psb*C and *rbc*L) using maximum likelihood method. The tree with branches showing bootstrap support > 50.
Extended Data Fig. 2	**Genome characters of *Microglena* sp. YARC.**	Extended_Data_Fig2.pdf	**a,** Estimation of the genome size by flow cytometry. The maize diploid cells (genome size 2.3 Gb) were used as reference. The relative fluorescence values of maize diploid cells were 55.15 ± 2.43, and the relative fluorescence value of *Microglena* haploid cells were 15.11 ± 0.73. Based on the ratio of fluorescence values, the genome size of *Microglena* was estimated to be about 1.26 ± 0.02 Gb. **b,** *K*-mer estimation of the genome size of *Microglena* sp. YARC. **c,** Cytogenetic studies showed that *Microglena* sp. YARC has n = 6 chromosomes. **d,** The Hi-C assisted assembly of

			Microglena sp. pseudomolecules. Heatmap showing Hi-C interactions under the resolution of 200 kb, and the antidiagonal pattern for the intrachromosomal interactions may reflect the Rabl configuration of chromatins. **e,f,** Ks (**e**) and 4dtv (**f**) distribution of the whole *Microglena* sp. genome.
Extended Data Fig. 3	**Comparison of total zinc finger domains in bacteria and fish genomes.**	Extended_Data_Fig3.pdf	**a,** Total zinc finger domains comparison between psychrophillic (blue) and mesophilic bacteria (tawny). **b,** Zinc finger comparison between two Antarctic fishes and four Mesophilic fishes. Statics analysis of two-sided Duncan's test showed no significant difference between each species. For all boxplots, box bounds represent the first and third quartiles and whiskers 1.5× the interquartile range; the center line represents the median.
Extended Data Fig. 4	**Phylogenetic and expansion analysis of carbonic anhydrase encoding genes in polar green alga *Microglena* and diatom *F. cylindrus*.**	Extended_Data_Fig4.pdf	
Extended Data Fig. 5	**Expansion of light harvesting proteins in *Microglena* sp. YARC.**	Extended_Data_Fig5.pdf	**a,** Expansion of chlorophyll A-B binding protein domains as a function of total annotated domains for green algal, diatom and dinoflagellate genomes.

			Solid for polar algae and hollow for non-polar algae. **b,** Unrooted genealogy of LHC genes in *Microglena* (Red), *Chlamydomonas reinhardtii* (blue), *Chlamydomonas eustigma* (green), *Volvox carteri* (yellow) and *Gonium pectorale* (purple).
Extended Data Fig. 6	**Upper 100m annually averaged dissolved zinc (nmol L^{-1}) from the PISCES model with observations taken within the upper 100m overlain as colored circles.**	Extended_Data_Fig6.pdf	Sampling locations for this study are indicated with red crosses.

1

Item	**Present?**	**Filename** **This should be the name the file is saved as when it is uploaded to our system, and should include the file extension. The extension must be .pdf**	**A brief, numerical description of file contents.** **i.e.:** ***Supplementary Figures 1-4, Supplementary Discussion, and Supplementary Tables 1-4.***
Supplementary Information	Yes	Supplementary_file.pdf	***Supplementary Methods, Supplementary Tables 1-7.***
Reporting Summary	Yes	nr-reporting-summary_NATECOLEVOL_210212798.pdf	
Peer Review Information	Yes	*Mock_PRfile.pdf*	

2

Type	**Number** **If there are multiple files of the same type this should be**	**Filename** **This should be the name the file is saved as when it is uploaded to our system, and should include the file**	**Legend or Descriptive Caption** **Describe the contents of**

	the numerical indicator. i.e. "1" for Video 1, "2" for Video 2, etc.	extension. i.e.: *Smith_ Supplementary_Video_1.mov*	the file
Supplementary Data	6	Supplementary_Data	Supplementary Data 1 PFAM and Interproscan annotation of zinc finger containing LTRs. Supplementary Data 2 Protein copy number of Microglena sp. YARC. Supplementary Data 3 Protein copy number of Chlamydomonas reinhardtii. Supplementary Data 4 Copy number of zinc binding proteins of Microglena sp. YARC. Supplementary Data 5 Copy number of zinc binding proteins of Chlamydomonas reinhardtii. Supplementary Data 6 Co-expression gene pairs of Microglena sp. YARC.

3

Parent Figure or Table	**Filename** **This should be the name the file is saved as when it is uploaded to our system, and should include the file extension. i.e.: *Smith_SourceData_Fig1.xls,* or *Smith_ Unmodified_Gels_Fig1.pdf***	**Data description** **i.e.: Unprocessed Western Blots and/or gels, Statistical Source Data, etc.**
Source Data Fig. 2	source_data_figure2.xlsx	Original source data
Source Data Fig. 3	source_data_figure3.xlsx	Original source data
Source Data Fig. 4	source_data_figure4.xlsx	Original source data
Source Data Fig. 5	source_data_figure5.xlsx	Original source data

The role of zinc in the adaptive evolution of polar phytoplankton

Naihao Ye[1,2]*†, Wentao Han[1]†, Andrew Toseland[3], Yitao Wang[1], Xiao Fan[1], Dong Xu[1], Cock van Oosterhout[3], **Sea of Change Consortium**[4], Igor V. Grigoriev[5], Alessandro Tagliabue[6], Jian Zhang[1], Yan Zhang[1], Jian Ma[1], Huan Qiu[7], Youxun Li[8], Xiaowen Zhang[1,2]*, Thomas Mock[3]*

[1]Yellow Sea Fisheries Research Institute, Chinese Academy of Fishery Sciences, Qingdao 266071, China. [2]Function Laboratory for Marine Fisheries Science and Food Production Processes, Qingdao National Laboratory for Marine Science and Technology, Qingdao 266200, China.[3]School of Environmental Sciences, University of East Anglia, Norwich Research Park, Norwich, NR4 7TJ, U.K. [4]A list of all participants and their affiliations appear at the end of the paper. [5]DOE-Joint Genome Institute, 1 Cyclotron Road, Berkeley, CA 94720, U.S.A. [6]School of Environmental Sciences, University of Liverpool, Liverpool, L69 3GP, U.K. [7]Independent scholar, Gaithersburg, Maryland 20878, U.S.A. [8]Marine Science Research Institute of Shandong Province, Qingdao 266104, China.

*__Correspondence__: yenh@ysfri.ac.cn (N.Y.); zhangxw@ysfri.ac.cn (X.Z.); t.mock@uea.ac.uk

†These authors contributed equally to this work.

Abstract

Zinc is an essential trace metal for oceanic primary producers with the highest concentrations in polar oceans. However, its role in the biological functioning and adaptive evolution of polar phytoplankton remains enigmatic. Here, we have applied a combination of evolutionary genomics, quantitative proteomics, co-expression analyses, and cellular physiology to suggest that model polar phytoplankton species have a higher demand for zinc because of elevated cellular levels of zinc-binding proteins. We propose that adaptive expansion of regulatory zinc-finger protein families, co-expanded and co-expressed zinc-binding proteins families involved in photosynthesis and growth in these microalgal species and their natural communities were identified to be responsible for the higher zinc demand. The expression of their encoding genes in eukaryotic phytoplankton metatranscriptomes from pole to pole was identified to correlate not only with dissolved zinc concentrations in the upper ocean but also with temperature, suggesting that environmental conditions of polar oceans are responsible for an increased demand of zinc. These results suggest that zinc plays an important role in supporting photosynthetic growth in eukaryotic polar phytoplankton, and that this has been critical for algal colonization of low temperature polar oceans.

Main

Oceanic phytoplankton contribute ca. 50% of annual primary productivity[1], and their biology and evolution is interlinked with ocean geochemistry throughout Earth history[2,3]. Biologically essential trace metals play an important role in this regard as any trace metal limitation can feedback onto the global nitrogen and carbon cycle. Although iron has received a lot of attention due to its role in carbon and nitrogen assimilation, zinc also supports a number of fundamental biological processes such as DNA/RNA replication and regulation, photosynthesis and carbon fixation[4,5]. Indeed, due to these requirements, zinc is one of the most abundant trace metals in the phytoplankton cell[6,7] . Oceanic surface waters display marked variability in their dissolved zinc concentrations, ranging from several nanomolar in the Southern Ocean[8,9] to vanishingly low levels in the tropical oceans[10,11]. Laboratory experiments have shown that variations in the availability of zinc in seawater are directly linked to cellular zinc levels[12,13]. For polar diatoms in particular, elevated demands for zinc[6] have been found to be a primary driver of the overall zinc distribution throughout the global ocean in several modelling studies[14,15,16]. Over geologic timescales, the availability of many trace metals has been thought to be affected by periods of anoxia and euxinia[3], but reconstructions of past zinc levels estimate broadly constant zinc concentrations through time[17].

The reasons behind the enhanced requirement of zinc by natural polar phytoplankton communities, especially in the Southern Ocean, remains enigmatic, but it implies that polar microalgae have an intrinsically higher zinc demand. Preliminary evidence for their high zinc demand was provided by the first genome sequence of a cold-adapted microalga, the diatom *Fragilariopsis cylindrus* from the Southern Ocean[18]. Unlike microalgae from temperate oceans, the genome of *F. cylindrus* was characterized by adaptive expansion of MYND zinc-finger proteins[18]. Even though the zinc requirement of one expanded zinc-binding protein family (e.g. MYND) is likely to be much lower than the external supply from the environment, the signature of the expansion suggests this may constitude a selective advantage. Furthermore, their expansion was estimated to have taken place within the last 30 million years, which coincides with the formation of the Southern Ocean and therefore glaciation of the Antarctic continent[18]. Thus, these data suggest that elevated concentrations of the trace metal zinc in the Southern Ocean may have contributed to diatom colonization of this polar marine ecosystem, and here we critically examine this hypothesis. Our study applies an integrative approach that includes quantitative proteomics with polar and non-polar model algae to test whether the former possess a higher zinc demand overall based on all proteins that contain zinc as co-factor. Complementary transcriptome and physiological measurements together with metagenome and metatranscriptome data from natural pole-to-pole algal communities are providing additional evidence that zinc plays an important role in supporting photosynthetic growth in eukaryotic polar phytoplankton.

Results

Comparative genomics and proteomics

To address parallel evolution in distantly related polar algal species, the genome of the green alga *Microglena* sp. YARC was sequenced (Extended Data Fig. 1) and compared with *F. cylindrus*, as well as other recently sequenced polar algae and their close relatives from non-polar ecosystems serving as controls. *Microglena* sp. was isolated from the Southern Ocean and sequenced using a combination of Illumina and PacBio RSII platforms based on Hi-C libraries to improve long-range contiguity. Although our *k*-mer analysis revealed a haploid genome, the estimated size of ca. 950 Mbps was unexpected (Extended Data Fig. 2a and b) as all previously sequenced green algal genomes are smaller in size[19] (range: 12 - 540 Mb), including the recently sequenced Antarctic green alga *Chlamydomonas* sp. ICE-L[20]. The size expansion of the *Microglena* sp. genome is the result of repeats, which contribute 79% (Extended Data Fig. 2b and Supplementary Table 1). Our current assembly (91% complete based on BUSCO) captures ca. 60% of the estimated genome size (Supplementary Table 2), and Hi-C data enabled us to combine the scaffolds into 6 chromosomes, which is in agreement with the estimated number of chromosomes based on karyography measurements (Fig. 1A; Extended Data Fig. 2c and d). We annotated 19,596 protein encoding genes (Supplementary Table 2) based on transcriptome sequencing under different stress conditions. Our synteny and homology analysis revealed no evidence for whole-genome duplication (Fig. 2E and F), and we identified that only ~0.25% of its total gene inventory potentially was acquired via recent horizontal gene transfer (Supplementary Table 3). The unprecedented repeat content in the genome of *Microglena* sp. is likely the result of transposon activity and their expansion. For instance, we found that the transposon family of long terminal repeats (LTRs) was expanded particularly over the past 40 million years (Fig. 1B), which is in accordance with the formation of the Antarctic circumpolar current[21]. Interestingly, we found that over 17% of the intact LTRs contain zinc-knuckle domains (CX2CX3GHX4C), with a peak expansion at approximately 20 Mya ago (Fig. 1B and Supplementary Data 1). As zinc-knuckle domains are involved in the regulation of mRNA metabolism[22,23], and as they are expressed under polar conditions in *Microglena* sp. (Fig. 1C), it suggests that they were required for the regulation of the LTRs.

Similar to our findings for the the cold-adapted diatom *F. cylindrus*[18], specific regulatory zinc-domain containing gene families were expanded in the genome of *Microglena* sp. such as the C3HC4 family (Fig. 2A and Supplementary Table 4). However, the expansion of the C3HC4 family in *Microglena* sp. was likely driven by long interspersed nuclear elements (LINEs), given that the peak of their insertion time coincides with the peak of the expansion of the C3HC4 family, and because they have accumulated in the flanking regions (≤ 5 kbp) of the C3HC4-containing genes (Fig. 2B). This close association of zinc-binding domains and transposable elements has not been identified in any of the other algal genomes to date. Similar to what we have seen in specific zinc-binding domains from *F. cylindrus*, *Microglena* sp. has elevated ratios of non-

synonymous to synonymous substitutions (Ka/Ks) of C3HC4-containing genes in comparison to control genes such as PSI and II and those representing the BUSCO data set (Fig. 2C), indicative of an accelerated rate of evolution. Interestingly, one site in the motifs[24] responsible for zinc-ion binding appear to be under significant positive selection (Ka/Ks > 1) (Fig. 2D and Supplementary Table S 5), which is consistent with adaptive evolution.

A broader comparative approach with distantly related cold-adapted polar algae such as two strains of *Polarella glacialis* (Dinoflagellates)[25] and their non-polar relatives[26-29] provided evidence for the commonality of the expansion of specifically zinc-finger domain containing gene families in polar microalgae despite species-specific differences in their diversity (Fig. 2E). Interestingly, a comparative genome analysis of gene families that co-expanded together with the zinc-finger domain containing genes identified photosynthesis genes, such as genes involved in light-harvesting, electron generation and transport, and inorganic carbon acquisition (Fig. 2E). Interestingly, many of the proteins that are co-expanded and involved in photosynthesis bind zinc as co-factor such as sedoheptulose-1,7-biphosphatase, fructose-bisphosphate aldolase, and specific carbonic anhydrases. Although there were species-specific differences in the diversity of photosynthesis genes, the Pfam-domain count in % total revealed strong enrichment in polar algae similar to the enrichment of zinc-finger domain containing gene families. The lack of expanded zinc-finger domains in polar heterotropic bacteria and cold-adapted fish (Extended Data Fig. 3) suggests that zinc and its binding proteins contribute to regulating photosynthesis and carbon acquisition in polar eukaryotic phytoplankton.

These comparative genomics data suggest that the demand for zinc might be higher in polar microalgae if the expansion of zinc-finger proteins and zinc-binding proteins involved in photosynthesis causes elevated cellular zinc concentrations. To test this hypothesis, we combined global domain searches for all known zinc-binding proteins in sequenced polar and non-polar green algae and diatoms complemented by quantitative proteomics using *Microglena* sp. and the mesophilic counterpart *Chlamydomonas reinhardtii* as model species. The relative contribution of genes encoding zinc-binding proteins in the genome of *Microglena* sp. was estimated to be 11.54% which was significantly higher (p-value = 0.03) compared to other green algae except a mesophilic strain of *Micromonas pusilla* (11.94%; p-value = 0.005) (Supplementary Table S6). However, the species *M. pusilla* is well known to have strains with frequent occurrence in the Arctic Ocean[30]. The genome of the polar diatom *Fragilariopsis cylindrus* was estimated to encode 10.75% zinc-binding proteins, which was significantly higher (p-value = 0.03) compared to the non-polar diatoms *Phaeodactylum tricornutum* and *Thalassiosira pseudonana* (Supplementary Table S6). Quantitative lable-free mass spectrometry was performed with *Microglena* sp. and *C. reinhardtii* under zinc-replete growth conditions to complement these *in-silico* estimates (Supplementary Data 2 and 3). A total of 396 and 384 zinc-binding proteins were identified in *Microglena* sp. and *C.*

reinhardtii protein extracts, which converts to their estimated total copy number of 4.64 ± 0.22 x 10^8 and 2.61 ± 0.22 x 10^8, respectively (Supplementary Data 4 and 5). To compare the total copy number of all zinc-binding proteins between both species, we normalized them using the ratio of the copy number of zinc-binding proteins over the copy number of Actin proteins. As Actin is not known to bind zinc and because it has a relatively stable copy number, it serves as an appropriate reference for normalization[31]. This analysis revealed that *Microglena* sp. contains a significantly (p-value < 2e-16; Wilcox test) higher copy number of zinc-binding proteins than *C. reinhardtii* (Fig. 3A). However, their separation into orthologs, paralogs and species-specific proteins revealed that only othologs and species-sepcific proteins were enriched in *Microglena* sp. compared to *C. reinhardtii* (Fig. 3B1-3). A similar ratio was observed for the less abundant group of zinc-finger proteins dominated by orthologs (88.5%) (Fig. 3C). The proportion of zinc-finger proteins to the total copy number of zinc-binding proteins in *Microglena sp.* was estimated to be 1.73% whereas it was only 0.63% in *C. reinhardtii.* Hence, these quantitative lable-free mass spectrometry data corroborate our comparative genome analyses including the evolutionary expansion of regulatory zinc-finger protein families such as MYND in *F. cylindrus*[18] and C3HC4-containing genes in *Microglena* sp.. These genomics and proteomics-based zinc-quota assessments were complemented by direct measurements of zinc ions in *Microglena* sp. and four different non-polar green algae including *C. reinhardtii* (Fig. 3D). *Microglena* sp. was the species with the highest intracellular zinc concentration including *Platymonas subcordiformis*, a non-polar green alga of similar cell size.

Co-expression networks

The quantitative proteomics data with *Microglena* sp. suggest that regulatory zinc-finger proteins such as C3HC4 might be co-regulated with photosynthetic proteins because both groups have a higher copy number compared to *C. reinhardtii.* To test this idea, we conducted co-expression analyses in *Microglena* sp. under diverse polar growth conditions in comparison to other polar and non-polar relatives. The latter serve as controls for revealing polar-specific co-expression networks potentially related to zinc. Thus, to reveal if photosynthetic proteins are likely a target of the regulatory zinc-binding proteins such as C3HC4 and MYND, we identified gene co-expression networks (Fig. 4A). Co-regulation of both groups of genes would suggest they are controlled by similar regulatory programmes and therefore members of the same pathway, implying causality relationships. We conducted extensive transcriptome profiling with *Microglena* sp. under different light, salinity, temperature and nutrient conditions simulating polar-relevant growth. Comparable transcriptome data were obtained from *F. cylindrus*[18,32], and non-polar algae using publically available transcriptomes[33].

More than 3,200 genes (16%) in *Microglena* sp. and more than 5,800 (27%) genes in *F. cylindrus* were significantly (Pearson's r ≥ 0.9; p-value ≤ 0.0001) co-expressed with the

expanded families of zinc-binding domain-containing genes (Fig. 4A and Supplementary Data S6). The same analysis using *M. pusilla*, *C. reinhardtii*, *T. pseudonana* CCMP 1335 and *P. tricornutum* CCMP2561 as a non-polar control species, only resulted in less than 800 (7%) co-expressed genes in each species (Fig. 4B), suggesting that polar conditions have not only caused co-expansion of specific zinc-binding and photosynthesis genes, but that proteins of both groups might interact to facilitate growth under polar-specific environmental conditions. The most enriched KEGG pathways for co-expressed genes were part of primary metabolism such as nitrogen and fatty acid metabolism and photosynthesis, including light-harvesting and inorganic carbon acquisition via carbonic anhydrases (CAs) (Fig. 4C). Interestingly, CA-families such as the α-family in *F. cylindrus* and the β-family in *Microglena* sp. were not only co-expressed with the MYND and C3HC4 zinc-domain containing genes, but they were also expanded, possibly because they require zinc as a co-factor (Extended Data Fig. 4). Another example of parallel evolution in polar algae is the expansion and co-expression of genes encoding high-light inducible, light-harvesting proteins such as *lhc*SRs and *cbr*s in *Microglena* sp. (Extended Data Fig. 5) and genes of the *Lhcx* clade in *F. cylindrus*, which suggests adaptation to cold-induced photoinhibition under conditions of 24 hours light and freezing temperatures (≤ -1.8□)[18]. For *Microglena* sp. for instance, the expansion of the *lhc*SRs and *cbr*s families coincides with the glaciation of the Antarctic continent (Fig. 4D), providing additional support for environment-induced adaptation of photosynthetic processes. In addition, proteins involved in electron transport have been duplicated (Fig. 4E), and some of the duplicated copies were differentially expressed under different light and stress conditions (Fig. 4F). These findings are consistent with sub- or neofunctionalization of duplicated member genes, which is likely to have facilitated adaptations to the extreme polar-specific.

Zinc-dependent growth and photophysiology

A higher demand for the trace element zinc in polar microalgae due to regulatory processes involved in photosynthesis and primary metabolism likely will mean that these algae are more susceptible to zinc limitation compared to their non-polar relatives. To test this hypothesis, we measured zinc-dependent cell-division rates, chlorophyll *a* concentrations, the quantum yield of photosynthesis (Fv/Fm) and relative electron transports rates for *Microglena* sp. and the non-polar green alga *P. subcordiformis* under different zinc concentrations (Fig. 5A-D). To estimate the adaptation to light limitation of *Microglena* sp., we performed photosynthesis-response curves (O_2 evolution) under relevant light spectra and intensities under zinc-replete growth conditions (Fig. 5E). The light compensation point (LCP) of photosynthesis was used as a measure of adaptation to light limitation in *Microglena* sp. Zinc-dependent photophysiology showed that *Microglena* sp. required at least twice as much zinc in the growth medium for a maximum specific growth rate, chlorophyll a concentration, and the

quantum yield of photosynthesis (Fv/Fm) compared to the non-polar green alga *P. subcordiformis* (Fig. 5A-D). However, the response of both microalgae to different zinc concentrations in the growth medium was less diverged for relative electron transport rates (ETRs) (Fig. 5D). They peaked at 10 nM in both algal species, but *P. subcordiformis* showed a much stronger decline at higher concentrations. Under zinc-replete growth conditions, oxygen evolution experiments with *Microgelana* sp. under white, blue and red light revealed a LCP between 2~8 µmol photons m^{-2} s^{-1} depending on the light spectrum applied (Fig. 5E).

Zinc-binding genes in pole-to-pole metatranscriptomes

To test whether the elevated level of zinc-binding proteins and therefore the cellular zinc quota in polar model algae is representative for natural polar microalgal communities, we correlated the normalized abundance of transcripts encoding zinc-binding proteins from pole-to-pole eukaryotic metatranscriptomes with latitude[34]. Samples were obtained from chlorophyll *a* – maximum layers as part of the project "Sea of Change: Eukaryotic Phytoplankton Communities in the Arctic Ocean" (DOI: 10.25585/1488054). For our correlations, we used transcripts from 346 domains known to bind zinc. The distribution of correlation coefficients (R) between the number of reads of zinc-binding domain-containing genes increase with latitude, as indicated by the mean R > 0 (One sample T-test for the North: T = 9.6421, df = 300, p < 2.2e-16; South: T = 18.549, df = 305, p < 2.2e-16) (Fig. 6A). However, the positive correlation between latitude and the number of reads of zinc-binding domain containing genes is significantly stronger for the Southern (mean (±StDev) R=0.44 (±0.42) compared to the North hemisphere (mean (±StDev) = 0.16 (±0.28) (Two sample T-test: T = 9.92, df = 535, p < 0.00001). The same trends with latitude were observed for estimated surface concentrations of dissolved zinc (Fig. 6B) with a weaker trend for the Northern hemisphere (mean R = 0.05; p-value = 1.36e-07) and a stronger trend for the Southern hemisphere (mean R = 0.35; p-value < 2.2e-16). However, as we have yet to build a comprehensive map of the surface ocean zinc inventory, concentrations of dissolved zinc for this study were estimated based on integrating publically available zinc data, new measurements and models (Extended Data Fig. 6).

The application of a generalized linear model indeed revealed that temperature explaines most of the variance with zinc as the second most important variable after temperature and before other nutrients tested (Supplementary Table 7). Consequently, our results indicate that polar eukaryotic phytoplankton communities have more zinc-binding proteins caused by both, an elevated copy number of transcripts encoding zinc-binding proteins and an expansion especially of regulatory zinc-finger protein families. These large-scale comparative metatranscriptomics data in combination with omics studies using model algal species provide strong support for the previously identified positive correlation between dissolved zinc concentrations and the modelled latitudinal gradient of zinc-uptake ratios of phytoplankton[16]. Thus, the increased demand for zinc in high-latitude phytoplankton appears to have resulted in elevated zinc-uptake ratios.

Comparative evolutionary genomics using metagenomes

Our comparative genomics, proteomics, and metatranscriptomics data suggest a common link between the evolution of polar microalgae from diverse lineages and zinc-based geochemistry. Thus, this link might therefore represent a unifying framework for how the evolution and biology of primary producers in polar oceans feedback onto global cycles of zinc and elements that co-vary such as silicon and carbon. Evidence for parallel evolution in distantly related algal lineages (e.g. diatoms, dinoflagellates, chlorophytes) that have converged in their adaptation under polar conditions including a response to elevated levels of zinc have already provided some support for the existence of this unifying framework (Figs. 1-4). However, for revealing if this framework is univerisal, we need to go beyond individual species and therefore have selected natural phytoplankton communities from across a latitudinal gradient to test if homologs in their polar communities undergo similar adaptive evolution as identified in our polar model species.

Thus, we have retrieved homologs from six Arctic phytoplankton metagenomes and compared them to five of their warm-water counterparts (Fig. 6C). All eleven metagenomes were collected on two RV Polarstern (Alfred-Wegener Institute for Polar and Marine Research, Bremerhaven, Germany) expeditions described by Martin et al.[34] and Duncan et al.[35]. First, we tested if the density of zinc-finger domains, for which we have unquivocal evidence of their adaptive evolution in polar model algae, is higher in polar metagenomes. Indeed, the majority of these domains increase in density in polar metagenomes (116 out of 138 pairwise comparisons, 84.0%), (one-sample T-test: N=138, T=3.98, p=0.0001) (Fig. 6D), corroborating the hypothesis that the expansion of the MYND zinc-finger family in the genome of *F. cylindrus* and the C3HC4 zinc-domain containing genes in *Microglena sp*. are not isolated events. Second, we calculated the ratio of non-synonomous to synonomous nucleotide substitutions (dN/dS) in zinc-finger domain containing genes in comparsion to homologs from non-polar counterparts (Fig. 6E). Based on more than 10,000 sequences from each environment, the median of dN/dS of these genes in polar phytoplankton metagenomes was significantly higher (Mann-Whitney p-value 2.2×10^{-16}) compared to their non-polar counterparts. These data suggest that purifying selection is more relaxed and/or that natural selection favours diversification in the polar zinc-finger domain containing genes, allowing them to evolve faster in polar phytoplankton. Interestingly, individual zinc-finger domain containing genes families such as CCHC, C2H2, CCCH, MYND, and *met* had the largest and most significant differences in dN/dS compared to their non-polar counterparts (Fig. 6F). This is consistent with the observation that genes encoding for protein repeats like Zn-fingers, new copies tend to experience rapid functional divergence due to the combined effects of relaxed purifying selection and positive selection[36]. It furthermore suggests that the expansion of specific zinc-finger domain containing gene families in natural polar phytoplankton communities facilitated their rapid evolution and possible acquisition of modified or new functions, enabling adaptations to the novel challenging conditions of the surface polar ocean.This expansion not likely to have been driven by any substantial changes in zinc availability[17], but may be more broadly linked to the emergence of polar conditions of which zinc in combination with low temperatures facilitated the adaptive evolution of algae to polar oceans.

Discussion

This study shows that elevated zinc concentrations in surface polar oceans have facilitated the adaptive evolution of algae to conditions of these permanently low-temperature environments. Zinc therefore has enabled the formation of some of the most productive food webs on Earth. We conclude this based on a set of integrative and comparative approaches that combine analyses of evolutionary genomics, quantitative proteomics, co-expression analyses, zinc-dependent cellular physiology, pole-to-pole eukaryotic phytoplankton metatranscriptomes, and metagenomes. Thus, the emerging view that the particularly high zinc demands of polar diatoms regulate a large part of the global zinc distribution[14,15] can be extended to include polar green algae and dinoflagellates. Thus, massive parallel evolution in these distantly related taxa appears to have enabled these primary producers to cope with the challenging conditions of polar oceans.

In addition to the expansion and accelerated evolution of specific zinc-binding domain containing protein families in three different algal classes, their co-expression networks reveal that polar microalgae might require an increased diversity of zinc-binding domains to regulate light harvesting, photosynthetic electron generation and transport, inorganic carbon acquisition and other forms of primary metabolism such as nitrogen and fatty acid metabolism. Interestingly, zinc-binding domains were also co-expressed with ice-binding proteins (Fig. 4A), suggesting their involvement in coping with freezing conditions. However, the specific zinc-dependent regulatory mechanisms facilitating primary metabolism under harsh polar conditions remain to be identified in polar algae. Photosynthetic processes would appear to be an appropriate target for their identification because of the co-expression networks and a lack of expanded zinc-domains in non-photosynthetic polar organisms.

As photosynthesis in polar ecosystems requires regulation under extreme seasonality of 24 hours light in summer and long periods of darkness in winter, algae not only need to effectively compensate for over excitation of the photosynthetic electron transport chain in summer, they also require efficient mechanisms to photosynthesis under extremely low irradiance levels. The former has been realized in different polar algae by expanding high-light inducible protein families, which contribute to the dissipation of energy and therefore reduce effects of over excitation (e.g. production of radical oxygen species). For coping with the dark end of the light spectrum, cold-adapted algae have evolved mechanisms to significantly reduce their light compensation point (LCP), which is the light intensity at which oxygen production equals consumption (Fig. 5E). For instance, the LCP of *Microglena* sp., was estimated to be between 2~8 µmol photons m^{-2} s^{-1} (Fig. 5E), which is similar to LCPs measured in polar sea-ice diatoms[37], but they are significantly lower than that of temperate algae such as *C. reinhardtii* (40 µmol photons m^{-2} s^{-1})[38]. A low LCP can be achieved by efficient energy conversion and carbon fixation under low light and low temperatures. Hence, it is likely that the expansion of LHC gene families in *Microglena* sp., and the duplication of genes encoding for proteins involved in linear electron transport underpins the reduction of the LCP in this species.

The co-expression with zinc-binding domain containing genes suggests that these photosynthetic processes require regulatory support to function at polar temperatures. If the regulation of these photosynthetic processes takes place at the level of transcription, translation or even at the level of regulating protein activity remains to be seen. The costs for the adaptation of photosynthetic processes (e.g. Fv/Fm, ETR) to polar environmental conditions in marine microalgae appear to be an increased demand for the trace metal zinc as indicated by increased copy numbers of zinc-binding proteins compared to non-polar relatives. Field studies on natural phytoplankton communities from non-polar regions (e.g. subarctic Pacific), which corroborate our laboratory-based growth experiments, confirm the significantly lower requirement of zinc for sustaining maximum growth rates in non-polar microalgae[39]. More broadly, the particularly high zinc demands of polar phytoplankton and the role played in zinc biogeochemical cycling, indicates that changes in their abundance and biodiversity due to a changing climate will directly modulate zinc cycling throughout the global ocean, potentially affecting the regional emergence of zinc limitation in low latitude oceans.

Methods

Microalgal materials and growth conditions

The polar species *Microglena* sp. YARC (YSFRI-03-00001) and the non-polar species *Platymonas subcordiformis* (YSFRI-03-00011), *Chlorella* sp. (YSFRI-03-00010), *Chlamydomonas reinhardtii* (YSFRI-03-00002), and *Chlamydomonas euryale* (YSFRI-03-00058), were all obtained from the Algal genetic resources center (Algae Culture Collection) at the Institute of Yellow Sea Fisheries Research Institute (YSFRI), Chinese Academy of Fishery Sciences. Cultures were single cell sorted using monoclonal screening on solid medium and bacterial contaminants were removed by treatment with ampicillin (50 μg ml^{-1}) and kanamycin (50 μg ml^{-1}). Batch cultures of all species were grown in nutrient replete Provasoli seawater medium[40] with 6°C + 40 μmol photons m^{-2} s^{-1} for polar species and 25°C + 100 μmol photons m^{-2} s^{-1} for non-polar species considering their inherent differences in temperature limits and light conditions required for optimal growth.

Genome sequencing

High molecular weight DNA for whole-genome sequencing was extracted from a monoclonal culture using Plant DNA Isolation Reagent (Takara Biomedical Technology (Beijing) Co., Ltd.). Genomic DNA was sheared using a sonication device for short-insert paired-end (PE) library construction. Short-insert libraries with a size of 180, 200, 300, 500 bp were constructed according to the instructions described in the Illumina library preparation kit (Illumina Co., U.S.A.). All libraries were sequenced on an Illumina HiSeq 2500 sequencer. The raw reads were subsequently trimmed for quality using Trimmomatic[42] (v.0.35). Illumina sequence adaptors were removed, low quality bases from the start or end of raw reads were trimmed, and reads were scanned using a 4-bp sliding window and trimmed when the average quality per base dropped below 15. The clean data obtained from this process were used for subsequent analyses. For PacBio library construction, the genomic DNA was sheared to ~10 kb, and short fragments

below the size of 7 kb were filtered using BluePipin (Sage Science Co., U.S.A). Filtered DNA was then converted into the proprietary SMRTbell library using the PacBio DNA Template Preparation Kit (Pacific Biosciences, U.S.A).

Genome assembly

Five different assembly methods, including DBG2OLC (hybrid)[43], Falcon (Pacbio only) (https://github.com/PacificBiosciences/FALCON/)[44], SmartDenovo (Pacbio only), LRscaf (hybrid)[45], and WTDBG2 (Pacbio only) were conducted with the PacBio and Illumina data. Finally, the genome assembly by WTDBG2 approach was adopted. All of the subreads from PacBio sequencing were assembled using WTDBG2[46] software with default values for all parameters (parameters: -L 5000 -p 19 -A -S 2 -e 2) (https://github.com/ruanjue/wtdbg2). The assembly sequence was then polished using Quiver (SMRT Analysis v2.3.0) with default parameters. To increase the accuracy of the assembly genome, six rounds of iterative error correction were performed using the above Illumina clean data. Scaffolding was performed using Hi-C-based proximity-guided assembly using the HiC-Pro, version 2.8.0, pipeline[47].

Repeat annotation and mask

Both RepeatModeler and RepeatMasker (http://www.repeatmasker.org) were used to perform de novo identification and mask of repeats. To make sure the integrity of genes in the subsequent analysis, the low complexity or simply repeats were not masked in this analysis, because some of which could be found in genes. Finally, 42.77% of assembled bases were masked. Candidate LTR-RTs in *Microglena* sp. were identified using LTRharvest[48] (version 1.5.10) and LTR_Finder[49] (version 1.07). LTRharvest with parameters '-similar 90 -vic 10 -seed 20 -seqids yes -minlenltr 100 -maxlenltr 7000 -mintsd 4 -maxtsd 6 -motif TGCA -motifmis 1' and LTR_finder with parameters '-D 15000 -d 1000 -L 7000 -l 100 -p 20 -C -M 0.9'. The identified LTR-RT candidates were filtered with LTR_retriever program[50] with default parameters. The reverse transcriptase paralogs within the intact LTRs were annotated by using Prodigal[51].

Gene prediction and annotation

Protein-coding region identification and gene prediction were performed through a combination of homology-based prediction, ab initio prediction, and transcriptome-based prediction method[52]. Proteins from several species including *C. reinhardtii*, *Chlorella variabilis*, *Coccomyxa subellipsoidea*, *Gonium pectorale*, *Micromonas pusilla* CCMP1545, *Micromonas commoda* RCC299, *Volvox carteri*, and *Ostreococcus lucimarinus*, were downloaded from NCBI (https://www.ncbi.nlm.nih.gov/genome). Protein sequences were mapping against *Microglena* sp. with exonerates. The blast hits were used in predicting the exact gene structure of the corresponding genome regions. The ab initio prediction software AUGUSTUS was used to predict coding-regions in the repeat-masked genome[52]. RNA-seq data were mapped against the assembly using Tophat[53] (version 2.1.1). Cufflinks[54] (version 2.2.1), and then was used to deal with the transcripts from the results of Tophat to obtain gene models. All gene models from the above three methods were integrated by EvidenceModeler (EVM) into a non-redundant gene set[55]. The weighting values for homology-based prediction, ab initio prediction, and transcriptome-based prediction method are 2, 1 and 10, respectively.

Functional annotations of the obtained gene set were conducted using BLASTP with an E-value 1e-5 against the NCBI-NR, SwissProt database, and KOG database. Protein domains were annotated by mapping to the InterPro and Pfam databases using InterProScan and HMMER[56]. The pathways in which genes might be were derived from genes mapping against the KEGG databases. The Gene Ontology (GO) terms for genes were extracted from the corresponding InterProscan or Pfam results.

CalculatingLTR insertion time

Intact LTR-RTs were identified using LTR_retriever program10[50]. We performed the following flow to calculate the insertion time of LTR: (1) calculate DNA substitution rate (μ) of *Microglena*; (2) aligning the two LTRs of each intact LTR-RT using the programme "Stretcher" (EMBOSS package)[57], (3) measuring the nucleotide distance (d) between LTRs using the Kimura two-parameter method (K2P)[58] as implemented in the programme "Distmat" (EMBOSS package) (Rice et al., 2000); and (4) measuring the insertion time of each LTR using the formula of T = d/2μ.

To calculate the DNA substitution rate (μ) of *Microgelna* sp., we firstly do an all-versus-all alignments between *Microglena* and *C. eustigma* by using OrthoFinder[59] (version 2.2.6) to obtain the orthologous pairs. Nucleotide distance (d) between orthologous pairs was estimated using the Kimura two-parameter (K2p) (transition-transversion ratio) criterion[58] as implemented in the program 'Distmat' (EMBOSS package[57]. Substitution rates (μ) were inferred using the formula: μ = d/2T, where T is the divergence time between *Microglena* and *C. eustigma* of about 432MYA. A total of 6413 orthologs were obtained and the median value of d was 0.065.

Calculating gene family duplication time

To estimate the gene duplication time of the annotated genes in the *Microglena* genomes, we first calculate the molecular clock rate (r) of *Microglena*. We obtained the orthologs between *Microglena* and *C. eustigma* by using OrthoFinder[59] (version 2.2.6). For each alignment result, the Ks values were calculated using KaKs Calculator[60] and single-linkage clustering for the Ks values was performed using the hclust function in the R package. A total of 6,413 orthologs were obtained and the median value of *ks* was 2.66. The r was estimated using the formula r = ks/2T with T = 432 Mya according to species differentiation time of *Microglena* and *C. eustigma* and the r was 3.08×10^{-9} substitutions per synonymous site per year. To calculate the gene family duplication time, we firstly performed all-versus-all alignments of the coding sequences within a gene subfamily and then calculate Ks values by KaKs Calculator[61] for each alignment result. The duplication time was estimated using the T=Ks/2r.

Evolutionary history analysis of zinc finger genes

To examine whether these recently duplicated zinc finger gene members diversified under the scrutiny of positive selection, we calculated the ka/ks of BUSCO, C3HC4 Zinc finger genes, expanded photoprotection *Lhc*SR and *CBR* genes, and the doubled PSII+I core encoding genes by KaKs_calculator[60].

To determine whether positive selection had acted at specific sites in the Zinc finger C3HC4 sequences, we compared three models of positive selection in PAML[62], M3-discrete, M2a-positive selection, M8-beta and ω to their null models of neutral evolution (M0-one ratio, M1a-nearly neutral, and M7-beta, respectively). LRT was

performed to test which model fits the data best. We multiplied the log likelihood times two (2ΔlnL), and used a chi-square test, and compared it to a χ2 distribution with 2 degrees of freedom.

Gene expression analysis

Microglena batch cultures for transcriptome sequencing were grown in three biological replicates under normal growth conditions (+6°C, nutrient-replete Provasoli seawater medium[40], 12-h/12-h light/dark photoperiod at 40 μmol photons m^{-2} s^{-1} and 31‰ salinity), low temperatures (LT, −2°C for 5 days), high temperatures (HT, +12°C for 5 days), high salinity (HS, 96‰ for 5 days), lessened salinity (LS, 16‰ for 5 days), and ultraviolet radiation (UV, 60 μw cm^{-2} UV-B irradiance treatment for 4 hours). For different illumination intensity treatment, *Microglena* were firstly darkened for 36 hours at their suitable conditions, and then exposed to illumination intensity of 3, 40 and 200 μmol photons m^{-2} s^{-1} for 1 hour.

Total RNA was extracted from 0.5 g tissue using an E.Z.N.A. Total RNA Kit (OMEGA, America) according to the manufacturer's protocol. After total RNA was extracted, mRNA were enriched using oligo (dT) magnetic beads. Then the strand-specific RNA-seq libraries were constructed using NEBNext® Ultra™ II Directional RNA Library Prep Kit for Illumina (NEB, Ipswich, MA, USA) in accordance with the manufacturer's instruction. The quality of RNA-seq libraries were assessed by using a Fragment Analyzer (Advanced Analytical, IA, USA), and the resulting libraries were sequenced on an Illumina HiSeq 2500 instrument producing pair-end reads of 150 nucleotides. The clean paired-end reads were mapped to the *Microglena* sp. genome using TopHat v2.0.12[53]. Then, the FPKM (fragments per kilobase of transcript sequence per million base pairs sequenced) value of each gene was calculated to estimate gene expression levels using Cufflinks[54] v2.2.1 (http://cole-trapnell-lab.github.io/cufflinks/). Heatmaps of expression patterns were generated in R using the pheatmap.

Gene-to-gene correlations were measured by Pearson's correlation that provides links between genes with similarities in expression pattern across multiple transcriptomes.

Total protein extraction and digestion

Microglena sp. and *C. reinhardtii* were grown under optimal growth conditions until the middle of their exponential growth phase where cell were harvested. Then the algal samples were grinded with liquid nitrogen, then BPP buffer were added in the ratio of 1:10. The solution were centrifuged at 12000 × g for 20 min at 4℃, and supernatants were collected. The equal volume of Tris-saturated phenol were added and vortexed for 10 min at 4°C. The solution were centrifuged at 12000g for 20 min at 4℃ and the phenol phase were collected. The equal volume of BPP were added and vortexed for 10 min at 4°C. The solution were centrifuged at 12000 × g for 20 min at 4℃ and the phenol phase were collected. Five volume of pre-cooled 0.1M ammonium acetate in methanol were added and precipitated protein at -20°C overnight. The supernatant was discarded by centrifugation, and the precipitate was washed twice with 90% acetone. Discard the supernatant by centrifugation and air dry the precipitate. The precipitate was re-suspended with lysis buffer (1% SDS, 8 M urea, cocktail), then sonicate for 3 min on ice.

The lysates were centrifuged, and supernatants were collected to test the concentration of protein in all samples. Protein concentrations were determined by Bicinchoninic acid (BCA) method by BCA Protein Assay Kit (Beyotime biotechnology). Protein quantification was performed according to the kit protocol.

Protein digestion was performed according to the standard procedure. Briefly, for each sample tube containing 100 μg protein, appropriate TCEP was added to the final concentration of 10mM and the tubes were incubated at 37 °C for 60 min. Appropriate IAM was added to the final concentration of 40mM and reaction for 40 min in dark. Add six volumes of cold acetone to the sample tube. Invert the tube three times and incubate the tube at –20 °C until precipitate forms (~4 h).The acetone was removed by centrifugation at 10000g for 20min and precipitated protein was resuspended with150μl 100mM TEAB Buffer. To each sample tube, according to the proportion 1:50 added the trypsin solution and incubate the tubes at 37 °C overnight.

Peptide desalination and quantification

The peptides were vacuum dried, then resusoended with 2% acetonitrile and 0.1% TFA. Samples were desalted with Sep-Pak, and vacuum dried. Peptide concentrations were determined by peptide qutification kit (Thermo, Cat.23275). Loading buffer was added to each tube to prepare samples for mass spectrometry analysis, and the concentration of each samples was 0.5μg/μl .

Mass spectrometry analysis

Mass spectrometry for proteomics analysis was performed on three biological replicates. Experiments were performed on a Q Exactive mass spectrometer that was coupled with Easy-nLC 1200. Each peptide sample was injected for nanoLC-MS/MS analysis. The sample was loaded onto a the C18-reversed phase column(75 μm x 25 cm , Thermo ,USA) in buffer A (2% acetonitrile and 0.1% Formic acid) and separated with a linear gradient of buffer B (80% acetonitrile and 0.1% Formic acid) at a flow rate of 300 nl/min. The electrospray voltage of 1.8 kV versus the inlet of the mass spectrometer was used. Q Exactive HF-X was operated in the data-dependent mode to switch automatically between MS and MS/MS acquisition. Survey full-scan MS spectra (m/z 350-1300) were acquired with a mass resolution of 70K, followed by twenty sequential high energy collisional dissociation (HCD) MS/MS scans with a resolution of 17.5K. In all cases, one microscan was recorded using dynamic exclusion of 18 seconds.

Sequence Database Searching

MS/MS spectra were searched using ProteomeDiscoverer™ Software 2.4 against protein data of *Microglena* and *C. reinhardtii* (assembly v3.0). The highest score for a given peptide mass (best match to that predicted in the database) was used to identify parent proteins. The parameters for protein searching were set as follows: tryptic digestion with up to two missed cleavages, carbamidomethylation of cysteines as fixed modification, and oxidation of methionines and protein N-terminal acetylation as variable modifications. Peptide spectral matches were validated based on q-values at a 1% false discovery rate (FDR).

Protein copy number estimations

Protein copy number calculations were performed in Perseus using the Proteomic Ruler plugin.40 This method uses the peak intensities of histone proteins, which are proportional to DNA content, to estimate protein abundance on a per cell basis[63].

Physiological responses to different zinc concentrations

In order to minimize contamination, the polycarbonate (PC) bottles (Thermo Scientific™ Nalgene™ Products, USA) were soaked for 1 week in 1 mol L^{-1} hydrochloric acid ("HPLC" grade, China National Pharmaceutical Group Corporation, China) and then were rinsed three times with ultrapure water (Merck Millipore Corporation, Darmstadt, Germany). To precisely manipulate the zinc concentration, we used the artificial seawater (ATCC medium 1661 with minor modification) to prepare the zinc free Provasoli seawater medium. The artificial seawater and Provasoli seawater medium was prepared using ultrapure water. The zinc concentration was adjusted by $ZnSO_4$ to the final concentration of 0, 10, 20 and 50 nM.

The algal samples were pre-cultivated in non zinc seawater medium for one month, and then re-inoculated to new medium with different zinc concentration. Three biological repetitions were used for each concentration. Cell numbers were calculated by Hemacytometers (Thermo Fisher Scientific, USA).

The specific growth rate (SGR) was calculated using the equation: SGR (increase day^{-1}) = (lnN2-lnN1)/(T2-T1) where N1=cell number at time T1, N2=cell number at time T2. Chlorophyll *a* and carotenoid contents were extracted by 95% ethanol. The contents of chlorophyll *a* was determined spectrophotometrically as follows: $Chl_a = 13.36 \times A_{664.2} - 5.19 \times A_{648.6}$.

To determine the intracellular zinc content, all algal samples were cultrued in normal Provasoli seawater medium prepared by using artificial seawater under their optimum growth condition. A sample of 0.05-0.1g freeze drying microalgae was placed in the digestion vessel and 5 mL of concentrated nitric acid was added. The vessels were capped and placed in a microwave pressure digestor Speedwave WX-8000 (Preekem) and subjected to microwave-mediated heating according to the following program: 100ºC for 3 min, 140ºC for 3 min, 160ºC for 3 min, 180ºC for 3 min, 190ºC for 15 min. After cooling, acid digests were made up to 50 mL with Milli-Q water. The zinc content was determined by inductively coupled plasma-atomic emission spectrometry (ICP-OES) (Thermo Scientific iCAP 7200, USA). Intracellular zinc contents were then normalized per cell.

Relative abundance of zinc-binding domains in marine metatranscriptomes

The relative abundance of zinc-binding domains in the oceans was assessed using the metatranscriptome data from the "Sea of Change: Eukaryotic Phytoplankton Communities in the Arctic Ocean" project (DOI: 10.25585/1488054) hosted at JGI. This dataset consists of sequence data from four separate cruises: ARK-XXVII/1 (PS80) - 17th June to 9th July 2012; Stratiphyt-II - 1st May to 30th April 2011; ANT-XXIX/1 (PS81) - 1st to 24th November 2012 and ANT-XXXII/2 (PS103) - 20th December 2016 to 26th

January 2017 and covers a transect of the Atlantic Ocean from Greenland to the Weddell Sea (71.36°S to 79.09°N), (https://www.pangaea.de/expeditions/cr.php/Polarstern)[34]. Each metatranscriptome dataset had been assembled and annotated with the JGI/IMG pipeline[64]. We downloaded Pfam annotations and self-mapping files (alignments of the raw reads to assembled contigs) for 72 metatranscriptomes. Using a custom Perl script, we identified all unique contigs containing zinc-binding Pfam domains (from a filtered list of 346) and converted the number of reads mapping to each contig to a percentage of total mapped reads for each sample ((#reads mapped / #total reads mapped) * 100). We then calculated the overall total for each domain for each sample with known latitude. To assess the correlation between the normalised abundance of each zinc-binding domain and latitude, we used the R function cor.test (Pearson's correlation coefficient, R) for each of the N=301 domains in North, and N=306 domains in South. Finally, we plotted the distributions of these 301 and 306 R coefficients of the Northern and Southern hemisphere samples, respectively, and we tested whether the means of these distributions were significantly greater than zero using a One Sample T-test. We thus tested whether there was an overall positive correlation between the relative number of reads of zinc-binding Pfam domains and latitude across more than 300 domains. We also tested whether the distributions of R between both hemispheres differed from one another, using a Two Sample T-test.

Density and dN/dS analyses of zinc-fingers from eukaryotic metagenomes

We downloaded contig sequences, gff files describing the contig coordinates of predicted genes and Pfam annotation tables for 11 metagenomes from the "Sea of Change: Eukaryotic Phytoplankton Communities in the Arctic Ocean" project (DOI: 10.25585/1488054) hosted at JGI. Six metagenomes were from polar regions (latitude 69.32°N to 79.02°N) and five from non-polar (latitude 34.88°N to 17.28°S) (Fig. 6c). As only amino-acid sequences were available for predicted genes, we first set up a custom Perl/BioPerl script to pull out gene sequences from the contig file for each sample based on the contig name, contig coordinates and strand orientation for each predicted gene from the corresponding gff file. Next, we compiled a list of all unique zinc-finger Pfam domains ('zf-' prefix, 138 unique domains in total) contained in all 11 metagenome samples and for each sample we produced a fasta file of all sequences containing each zinc-finger domain. We then combined fasta files from each domain from all polar/non-polar samples. We calculated the density of zinc-finger containing genes, tallying the number of genes per one Mb CDS. We did this for both the polar and non-polar metagenomes. We then subtracted the two distributions of the polar minus the non-polar environment and repeated this procedure for all 138 zinc-finger domain containing genes. Finally, we calculated the mean and the standard deviation of this difference for each of these 138 genes (see 6d). This shows that the majority (116 out of 138, 84%) of zinc-finger domains increase in density in the polar environment (one-sample T-test: N=138, T=3.98, p=0.0001). The figure displays the absolute differences ranked, with the absolute difference decreasing from the left to right (Fig. 6d).

The dN/dS analysis pipeline was implemented as following. For each multi-fasta file of sequences representing all genes containing a specific zinc-finger domain from either

polar/non-polar samples we first clustered the sequences with cd-hit v4.6.8[65] at 100% global identity to remove identical sequences. The non-redundant nucleotide sequences were then translated into their first reading frame, amino-acid sequences with the BBmap v37.28[66] utility translate6frames.sh and clustered with cd-hit requiring >=40% identity over >=50% coverage of the longer sequence. We then produced a summary of the sequence clustering with the cd-hit utility script clstr2txt.pl and used this with a custom perl script to pull out the original nucleotide sequences for each cluster containing >=5 sequences into a separate multi-fasta file.

For each cluster we then aligned the gene sequences with Prank v170427[67] in codon mode; removed poorly aligned sequences with trimal v1.2[67] requiring a residue overlap of 0.5 and a sequence overlap of 50 and then removed gappy columns with Gblocks v0.91b[68] in codon mode. We then produced a phylogenetic tree with RAxML v8.2.9 (Stamatakis, 2014), using the GTRGAMMA substitution model, with 100 bootstrap replicates. The curated multiple sequence alignment was then converted to Paml format with Prank and the alignment and RAxML tree was used for Paml v4.9[62] Codeml analysis using model M0 (one average dN/dS ratio).

The omega (dN/dS) ratio results were parsed from each successful Codeml run and added to a summary file along with the domain name and environment type (polar/non-polar). In total, the analysis yielded 1,977 clusters containing 23,792 sequences from non-polar samples and 1,310 clusters containing 14,662 sequences from polar samples. A total of 95 unique zinc-finger domains produced dN/dS results. The results summary were imported into R and we performed a Mann-Whitney test testing the hypothesis that the distribution of dN/dS from polar samples was higher than non-polar samples. The same test was carried out for each individual zinc-finger domain and those domains producing a p-value of <= 0.001 were retained for plotting. Next we calculated the number of unique genes containing each zinc-finger domain from each sample. The raw counts were normalized to number of genes per megabase of coding sequence and imported into R.

Acknowledgments

This work was supported by the national key research and development program of China (2018YFD0900305), Special funds of Shandong Province for Pilot National Laboratory for Marine Science and Technology (Qingdao) (2021QNLM050103-1), National Natural Science Foundation of China (41676145, 32000404), Central Public-interest Scientific Institution Basal Research Fund, YSFRI, CAFS (20603022020019, 20603022021019), China Agriculture Research System (CARS-50), Taishan Scholars Funding of Shandong Province, Young Taishan Scholars Program (tsqn202103136). The metatranscriptome sequencing was conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH1123. AT was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 724289). TM acknowledges funding from the U.S. Department of Energy, Joint Genome Institute (Grant 532, Community Science Program) and the Natural Environment Research Council (NERC) (Grants NE/K004530/1; NE/R000883/1). TM

and CvO acknowledge partial funding from the School of Environmental Sciences at the University of East Anglia, Norwich Research Park, United Kingdom.

Author contributions

N.Y., T.M. and X.Z. designed the study; X.Z., W.H., T.M., C.v.O., X.F., N.Y., A.T. and H. Q. analyzed the data; Y.W., D.X., J.Z., Y.Z., J.M. and Y.L. conducted the laboratorial experiment; Sea of Change Consortium collected the samples and did DNA and RNA extractions. The consortium also contributed to sequence data analysis; I.V.G. coordinated metagenome and metatranscriptome sequencing. T.M., X.Z., N.Y., A.T. and C.v.O. co-wrote the manuscript.

Competing interests statement

The authors declare no competing interests.

Data availability

The *Microglena* sp. genome assembly data were deposited in NCBI GenBank (under BioProject accession PRJNA787402 and Genome accession JAJSRW000000000). All raw transcriptome sequencing data of *Microglena* sp. were deposited into the Sequence Read Archive (under BioProject accession PRJNA814737). The mass spectrometry proteomics data of *Microglena* sp. and *C. reinhardtii* have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the iProX partner repository with the dataset identifier PXD032702. Source data are provided with this paper.

Figure legends

Fig. 1. The expansion and expression of transposable elements in *Microglena* sp. **a,** A schematic representation of assembled *Microglena* sp. genomic characteristics. Tracks from outer to inner: Track 1, chromosome length; Track 2, Protein-coding genes present in the scaffolds, red represents genes on forward strand and green for genes on reverse strand; Track 3, Distribution of gene density with sliding windows of 1 Mb; Track 4, Distribution of repeat element density with sliding windows of 1 Mb; Track 5: Mapping depths of the whole genome; Track 6, Mapping depths on LTR transposon elements; Track 7, Mapping depths on LINE transposon elements; Track 8, Mapping depths on DNA repeat elements; Track 9, Mapping depths on intron regions; Track10, Mapping depths on exon regions; Track 11, Paralogous in *Microglena*. **b**, Distributions of insertion times calculated for intact LTRs and zinc knuckle domain containing intact LTRs. **c**, Heat map of expressed LTRs, LINEs and zinc finger containing LTRs (ZF-

LTRs) under different conditions. LT, low temperature (-2℃for 5 days); HS, high salt (99‰ for 5 days) ; LS, low salt (16‰ for 5 days), UV (60μw cm^{-2} UVB for 4 hours), HT (15℃for 5 days); C (control).

Fig. 2. Expansion and evolution of zinc-finger proteins in polar microalgae. a, Expansion of C3HC4 zinc-binding domains as a function of total annotated domains for selected green algal genomes. **b**, Insertion time compared between zinc finger (C3HC4 type, PF13920) domains and their flanking LTRs and LINEs. Red line, LTRs; blue line, zinc finger domains; green line, flanking LINEs. **c**, Comparison of the ratio of the non-synonymous over the synonymous substitutions (Ka/Ks) between BUSCO, C3HC4 zinc finger genes and photosynthesis genes. The *p*-value are calculated using a two-sided Wilcoxon test. For all boxplots, box bounds represent the first and third quartiles and whiskers 1.5× the interquartile range; the center line represents the median. **d**, The conserved motifs in *Microglena* C3HC4 zinc finger domains of clade III and the positive selection site (red asterisk). I (C, C), II (C, H), III (C, C) and IV (C, C) indicates the four pairs of amino acids participated in binding zinc ions. **e,** Heatmap of co-expanded Pfam domains in eight sequenced dinoflagellate, three diatom, and ten green algal genomes. Significantly expanded PFAM domains in polar microalgae (two-sided Fisher's exact test p-value ≤ 0.05) are highlighted with an asterisk. The bar graph indicates the average domain count (in % total) of respective Pfam domains in all 21 polar (yellow) and non-polar (blue) algal species. Pg3, *Polarella glacialis* CCMP1383; Pg8, *Polarella glacialis* CCMP2088; Syc, *Symbiodinium* sp. clade C; Sm, *Symbiodinium microadriaticum*; Sya, *Symbiodinium* sp. clade A3; Bm, *Breviolum minutum*; Cg, *Cladocopium goreaui*; Fk, *Fugacium kawagutii*; Fc, *Fragilariopsis cylindrus*; Pt, *Phaeodactylum tricornutum*; Tp, *Thalassiosira pseudonana*; Mg, *Microglena* sp. YARC; Ce, *Chlamydomonas eustigma*; Cr, *Chlamydomonas reinhardtii*; Vc, *Volvox carteri*; Gp, *Gonium pectorale*; Cs, *Coccomyxa subellipsoidea*; Cv, *Chlorella variabilis*; Mc, *Micromonas commoda* RCC299; Mp, *Micromonas pusilla*; Ol, *Ostreococcus lucimarinus*.

Fig. 3. Comparison of protein copy numbers, cellular zinc concentrations and cell sizes between polar and non-polar green microalgae. a, b, Comparison of the copy number of the total zinc-binding proteins (**a**) and their separation into orthologs, paralogs and species-specific proteins (**b**). **c,** Comparison of the zinc-finger proteins of polar *Microglena* (Mg) and non-polar *C. reinhardtii* (Cr) (two-sided wilcoxon test). N = 3 biologically independent samples. Box plots show the Q1 and Q3 (the 25^{th} and 75^{th} percentile, or the interquartile range, IQR), with the median in the centre, and the whiskers denoting Q1 – 1.5 * IQR and Q3 + 1.5 * IQR. **d**, Number of intracellular zinc atoms per cell (circles, left y-axis) and cell size (triangles, right y-axis) of polar (red) and non-polar green algae (blue).Two-sided Duncan's multiple range test, $p < 0.05$. N = 3 biologically independent samples. Data are presented as mean values ± SEM. Mg,

Microglena sp.; Ps, *Platymonas subcordiformis*; Cr, *Chlamydomonas reinhardtii*; Ce, *Chlamydomonas euryale,* Cs, *Chlorella* sp..

Fig. 4. The co-expression of zinc finger genes and the expansion and expression of photosynthesis genes in *Microglena* sp.. a, b, Co-expression networks of zinc finger genes and their co-regulated genes in (**a**) the polar microalgae *Microglena* sp. (Mg) and *Fragilariopsis cylindrus* (Fc). and in (**b)** their non-polar counterparts *Chlamydomonas reinhardtii* (Cr), *Phaeodactylum tricornutum* (Pt) and *Thalassiosira pseudonana* (Tp). **c**, Co-expression ratios in % total for zinc-finger genes, all co-expressed genes and photosynthesis and carbon fixing genes in polar and non-polar microalgae. Photo, Photosynthesis genes; LHCs, Antenna proteins encoding genes; CarF, carbon fixing genes; CAs, carbonic anhydrase genes; IBPs, ice-binding protein encoding genes; NiM, nitrogen metabolism genes; GlylM, glycerolipid metabolism genes; FAs, fatty acid biosynthesis genes; Oxid, oxidative phosphorylation genes. **d**, Distribution of synonymous substitutions (*Ks*) of four LHC subfamily genes. The number on each curve represents the peak of *Ks*. LhcSR and CBR encoding genes were candidates for photoprotection, and Lhcb and Lhca were light harvesting proteins for PSII and PSI, respectively. **e**, Unrooted genealogy of PSI (*psa*D, *psa*E, *psa*F, *psa*K, *psa*L), PSII (*psb*X, *psb*W, *psb*R, *psb*O) and photosynthetic electron transport chain *pet*C genes in 11 green algae: Ms, *Microglena* sp.; Cb, *Chara braunii*; Chl, *Chlorella variabilis*; Ce, *Chlamydomonas eustigma*; Cr, *Chlamydomonas reinhardtii*; Gp, *Gonium pectorale*; Cs, *Coccomyxa subellipsoidea*; Kf, *Klebsormidium flaccidum*; Mc, *Micromonas* RCC229; Mp, *Micromonas* CCMP1545; Um, *Ulva mutabilis*; Vc, *Volvox carteri*; Ts, *Tetrabaena socialis*; Ol, *Ostreococcus lucimarinus*; Cv, *Chlorella variabilis*. **f**, Diverged expression of *psb*O, *psb*W, *psb*R and *pet*C paralogous genes under high salt (HS), at low temperature (LT) and under control growth conditions (C). N = 3 biologically independent samples. Box plots show the Q1 and Q3 (the 25^{th} and 75^{th} percentile, or the interquartile range, IQR), with the median in the centre, and the whiskers denoting Q1 – 1.5 * IQR and Q3 + 1.5 * IQR. Different letters on error bars indicate statistically significant differences (Two-sided Duncan's multiple range test, $p < 0.05$).

Fig. 5. Zinc-dependent growth rates and photophysiology of polar vs non-polar green algal species. a, **b, c, d,** Specific growth rate (**a**), chlorophyll *a* concentration (**b**), quantum yield of photosynthesis (**c**) and the relative electron transport rate (**d**) of the polar green alga *Microglena* sp. and the non-polar relative *P. subcordiformis* under different concentrations of zinc in the growth medium. N = 3 biologically independent samples. Different letters on error bars indicate statistically significant differences (Two-sided Duncan's multiple range test, $p < 0.05$). For all boxplots, box bounds represent the first and third quartiles and whiskers 1.5x the interquartile range; the center line represents the median. **e,** Photosynthesis-response curve of *Microglena* sp. under white, blue and red light. Oxygen evolution (µmol O_2 L^{-1} s^{-1} $cell^{-1}$) was measured at 6°C using micro-electrodes. N=3 biologically independent samples. Data are presented as mean values ± SEM.

Fig. 6. Eukaryotic phytoplankton metatranscriptomes and -genomes with focus on zinc-binding domain containing genes and their substitution rates (dN/dS) in

relation to latitude and the estimated concentration of dissolved zinc in the surface ocean. a, Distribution of correlation coefficients (R) between the number of reads of zinc-binding domain containing genes and the latitude of the sample (a). Both in the Northern and Southern hemisphere, the relative number of reads of Zn-binding domain containing genes increase at higher latitude, as indicated by the mean R > 0 (One sample T-test for the North: T = 9.6421, df = 300, p < 2.2e-16; South: T = 18.549, df = 305, p < 2.2e-16). The positive correlation between latitude and the number of reads of zinc-binding domain containing genes is significantly stronger for the Southern hemisphere (mean (± SD) R = 0.44 (± 0.42) compared to the North (mean (± SD) = 0.16 (± 0.28) (Two sample T-test: T = 9.92, df = 535, p < 0.00001). **b,** Estimated concentrations of dissolved zinc (nmol/kg) for stations subjected to metatranscriptome sequencing. **c,** Geographical map showing the sampling locations for metagenomes (Chlorophyll *a* maximum layer). Blue colour indicates sampling stations north (P = polar ≥ 66.3°, n=6) and yellow colour south (NP = non-polar ≤ 66.3°, n=5) of the Arctic Circle. **d,** Difference in density of zinc-finger containing genes (mean (± SD) number of genes per Mb CDS) between polar and non-polar metagenomes. The majority (116 out of 138) of zinc-finger domains increase in density in the polar environment (one-sample T-test (two-tailed): N=138, T=3.98, p=0.0001). Figure shows the mean difference and SD of all 138 zinc-finger domains. The absolute differences are placed in rank order so that they decrease from left to right. **e,** Box plot of dN/dS from all identified zinc-finger domain containing genes. Number indicates Mann-Whitney p-value**. f,** Box plot of dN/dS from zinc-finger domain containing genes deemed to be significantly (Mann-Whitney p-value ≤ 0.001) higher in genes from polar compared to non-polar metagenomes. Box plots show the Q1 and Q3 (the 25^{th} and 75^{th} percentile, or the interquartile range, IQR), with the median in the centre, and the whiskers denoting Q1 – 1.5 * IQR and Q3 + 1.5 * IQR. Raw data in panel E are shown as dots.

References

1. Field, C. B., Behrenfeld, M. J., Randerson, J. T. & Falkowski, P. Primary Production of the Biosphere: Integrating Terrestrial and Oceanic Components. *Science* **281**, 237 (1998).

2. Anbar, A. D. & Knoll, A. H. Proterozoic ocean chemistry and evolution: a bioinorganic bridge? *science* **297**, 1137-1142 (2002).

3. Saito M. A., Sigman D. M. & Morel, F. M. M. The bioinorganic chemistry of the ancient ocean: the co-evolution of cyanobacterial metal requirements and biogeochemical cycles at the Archean–Proterozoic boundary? *Inorg Chim Acta* **356**, 308-318 (2003).

4. Morel, F. M. M., Lam, P. J. & Saito, M.A. Trace Metal Substitution in Marine Phytoplankton. *Annu Rev Earth Pl Sc* **48**, 491-517 (2020).

5. Morel, F. M. & Price, N. M. The biogeochemical cycles of trace metals in the oceans. *Science* **300**, 944-947 (2003).

913 6. Twining, B. S. & Baines, S. B. The Trace Metal Composition of Marine
914 Phytoplankton. *Annu Rev Mar Sci* **5**, 191-215 (2013).

915 7. Ho, T.-Y. et al. The Elemental Composition of Some Marine Phytoplankton. *J Phycol*
916 **39**, 1145-1159 (2003).

917 8. Ellwood, M.J. Wintertime trace metal (Zn, Cu, Ni, Cd, Pb and Co) and nutrient
918 distributions in the Subantarctic Zone between 40–52°S; 155–160°E. *Mar Chem* **112**,
919 107-117 (2008).

920 9. Zhao, Y., Vance, D., Abouchami, W. & de Baar, H. J. W. Biogeochemical cycling of
921 zinc and its isotopes in the Southern Ocean. *Geochim Cosmochim Ac* **125**, 653-67
922 (2014).

923 10. John, S. G., Helgoe, J. & Townsend, E. Biogeochemical cycling of Zn and Cd and
924 their stable isotopes in the Eastern Tropical South Pacific. *Mar Chem* **201**, 256–262
925 (2018).

926 11. Middag, R., de Baar, H. J. W. & Bruland, K.W. The Relationships Between
927 Dissolved Zinc and Major Nutrients Phosphate and Silicate Along the GEOTRACES
928 GA02 Transect in the West Atlantic Ocean. *Global Biogeochem Cy* **33**, 63–84 (2019).

929 12. Sunda, W. G. & Huntsman, S. A. Feedback interactions between zinc and
930 phytoplankton in seawater. *Limnol Oceanogr* **37**, 25-40 (1992).

931 13. Sunda, W. G. & Huntsman, S. A. Cobalt and zinc interreplacement in marine
932 phytoplankton: Biological and geochemical implications. *Limnol Oceanogr* **40**, 1404-
933 1417 (1995).

934 14. Vance, D. et al. Silicon and zinc biogeochemical cycles coupled through the
935 Southern Ocean. *Nat Geosci* **10**, 202 (2017).

936 15. Weber, T., John, S., Tagliabue, A. & DeVries, T. Biological uptake and reversible
937 scavenging of zinc in the global ocean. *Science* **361**, 72 (2018).

938 16. Roshan, S., DeVries, T., Wu, J. & Chen, G. The Internal Cycling of Zinc in the
939 Ocean. *Global Biogeochem Cy* **32**, 1833-1849 (2018)..

940 17. Scott, C. et al. Bioavailability of zinc in marine systems through time. *Nat Geosci* **6**,
941 125-128 (2012).

942 18. Mock, T. et al. Evolutionary genomics of the cold-adapted diatom *Fragilariopsis*
943 *cylindrus*. *Nature* **541**, 536-540 (2017).

944 19. Blaby-Haas, C. E. & Merchant, S. S. Comparative and Functional Algal Genomics.
945 *Annu Rev Plant Biol* **70**, 605-638 (2019).

946 20. Zhang, Z. H. et al. Adaptation to Extreme Antarctic Environments Revealed by the

947 Genome of a Sea Ice Green Alga. *Curr Biol* **30**,1-12 (2020).

948 21. Clarke, A. et al. The Southern Ocean benthic fauna and climate change: a historical
949 perspective. *Phil Trans R Soc Lond B* **338**, 299-309 (1992).

950 22. Klug, A. The Discovery of Zinc Fingers and Their Applications in Gene Regulation
951 and Genome Manipulation. Annu Rev Biochem **79**, 213-231 (2010).

952 23. Krishna, S. S., Majumdar, I. & Grishin, N. V. Structural classification of zinc fingers:
953 SURVEY AND SUMMARY. *Nucleic Acids Res* **31**, 532-550 (2003).

954 24. Barlow, P. N. et al. Structure of the C3HC4 domain by 1H-nuclear magnetic
955 resonance spectroscopy.: A new structural class of zinc-finger. *Journal of molecular*
956 *biology* **237**, 201-211 (1994).

957 25. Stephens, T. G. et al. Genomes of the dinoflagellate Polarella glacialis encode
958 tandemly repeated single-exon genes with adaptive functions. *BMC Biology* **18**, 1-21
959 (2020).

960 26. Aranda M. et al. Genomes of coral dinoflagellate symbionts highlight evolutionary
961 adaptations conducive to a symbiotic lifestyle. *Sci Rep.* **6**, 39734 (2016).

962 27. Liu H. et al. Symbiodinium genomes reveal adaptive evolution of functions related to
963 coral-dinoflagellate symbiosis. *Commun Biol.* **1**, 95 (2018).

964 28. Shoguchi E. et al. Draft assembly of the Symbiodinium minutum nuclear genome
965 reveals dinoflagellate gene structure. *Curr Biol.* **23**, 1399–408 (2013).

966 29. Shoguchi, E.et al. Two divergent Symbiodinium genomes reveal conservation of a
967 gene cluster for sunscreen biosynthesis and recently lost genes. *BMC Genomics.* **19**,
968 458 (2018).

969 30. Hoppe, C. J. M., Flintrop, C. M. & Rost, B. The Arctic picoeukaryote Micromonas
970 pusilla benefits synergistically from warming and ocean acidification. *Biogeosciences*
971 **15**, 4353-4365 (2018).

972 31. Ferguson, R. E.et al. Housekeeping proteins: A preliminary study illustrating some
973 limitations as useful references in protein expression studies. *Proteomics* **5**, 566-571
974 (2005).

975 32. Aslam, S.N. et al. Identifying metabolic pathways for production of extracellular
976 polymeric substances by the diatom Fragilariopsis cylindrus inhabiting sea ice. *ISME J*
977 **12**, 1237-1251 (2018).

978 33. Valenzuela, J. J. et al. Ocean acidification conditions increase resilience of marine
979 diatoms. *Nat Commun* **9**, 2328 (2018).

980 34. Martin, K. et al. The biogeographic differentiation of algal microbiomes in the upper

ocean from pole to pole. *Nat Commun.* **12**, 5483 (2021).

35. Duncan, A. et al. Metagenome-assembled genomes of phytoplankton communities across the Arctic Circle. *bioRxiv* (2020).

36. Persi, E., Wolf, Y. I. & Koonin, E. V. Positive and strongly relaxed purifying selection drive the evolution of repeats in proteins. *Nat Commun* **7**, 13570 (2016).

37. Mock, T. & Gradinger, R. Determination of Arctic ice algal production with a new in situ incubation technique. *Mar Ecol Prog Ser* **177**, 15-26 (1999).

38. Rühle, T., Hemschemeier, A., Melis, A. & Happe, T. A novel screening protocol for the isolation of hydrogen producing Chlamydomonas reinhardtiistrains. *BMC Plant Biol* **8**, 107 (2008).

39. Crawford, D. W. et al. Influence of zinc and iron enrichments on phytoplankton growth in the northeastern subarctic Pacific. *Limnol Oceanogr* **48**, 1583-1600 (2003).

40. Provasoli, L. *Media and prospects for the cultivation of marine algae: Cultures and Collections of Algae* (Jap. Soc. Plant Physiol, 1968).

41. Ranallo-Benavidez, T. R., K. S. J. & Michael C. S. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat Commun* **11**, 1-10 (2020).

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

43. Ye, C. X. et al. DBG2OLC: Efficient Assembly of Large Genomes Using Long Erroneous Reads of the Third Generation Sequencing Technologies. *Sci Rep* **6**, 31900 (2016).

44. Chin, C. S., et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat Methods* **13**, 1050-1054 (2016).

45. Qin, M. et al. LRScaf: improving draft genomes using long noisy reads. *BMC Genomics* **20**, 955 (2019).

46. Ruan, J. & Li, H. Fast and accurate long-read assembly with wtdbg2. *Nat Methods* **17**, 155-158 (2020).

47. Servant, N.et al. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol* **16**, 259 (2015).

48. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* **9**, 18 (2008).

49. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR

retrotransposons. *Nucleic Acids Res* **35**, W265-W268 (2007).

50. Ou, S. J. & Jiang, N. LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons. *Plant Physiol* **176**, 1410-1422 (2018).

51. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).

52. Stanke, M., Schöffmann, O., Morgenstern, B. & Waack, S. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics* **7**, 62 (2006).

53. Kim, D.et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14**, R36 (2013).

54. Trapnell, C. et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* **7**, 562-578 (2012).

55. Haas, B.J. et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol* **9**, R7 (2008).

56. Mitchell, A. L. et al. InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Res* **47**, D351-D360 (2018).

57. Rice, P., Longden, I. & Bleasby, A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet* **16**, 276-277 (2000).

58. Kimura, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol* **16**, 111-120 (1980)..

59. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* **20**, 238 (2019).

60. Zhang, Z. et al. KaKs_Calculator: Calculating Ka and Ks Through Model Selection and Model Averaging. *Genom Proteom Bioinf* **4**, 259-263 (2006).

61. Kumar, S., Stecher, G., Suleski, M. & Hedges, S. B. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Mol Biol Evol* **34**,1812-1819 (2017).

62. Yang, Z. H. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**, 1586-1591 (2007).

63. Wiśniewski, J. R., Hein, M. Y., Cox, J. & Mann, M. A ‘proteomic ruler’ for protein copy number and concentration estimation without spike-in standards. *Mol Cell Proteomics* **13**, 3497-3506 (2014).

64. Huntemann, M. et al. The standard operating procedure of the DOE-JGI Metagenome Annotation Pipeline (MGAP v.4). *Stand Genomic Sci* **10**, 86 (2016).

65. Fu, L. et al. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150-3152 (2012).

66. Bushnell, B. *BBMap: a fast, accurate, splice-aware aligner* (Lawrence Berkeley National Lab., 2014).

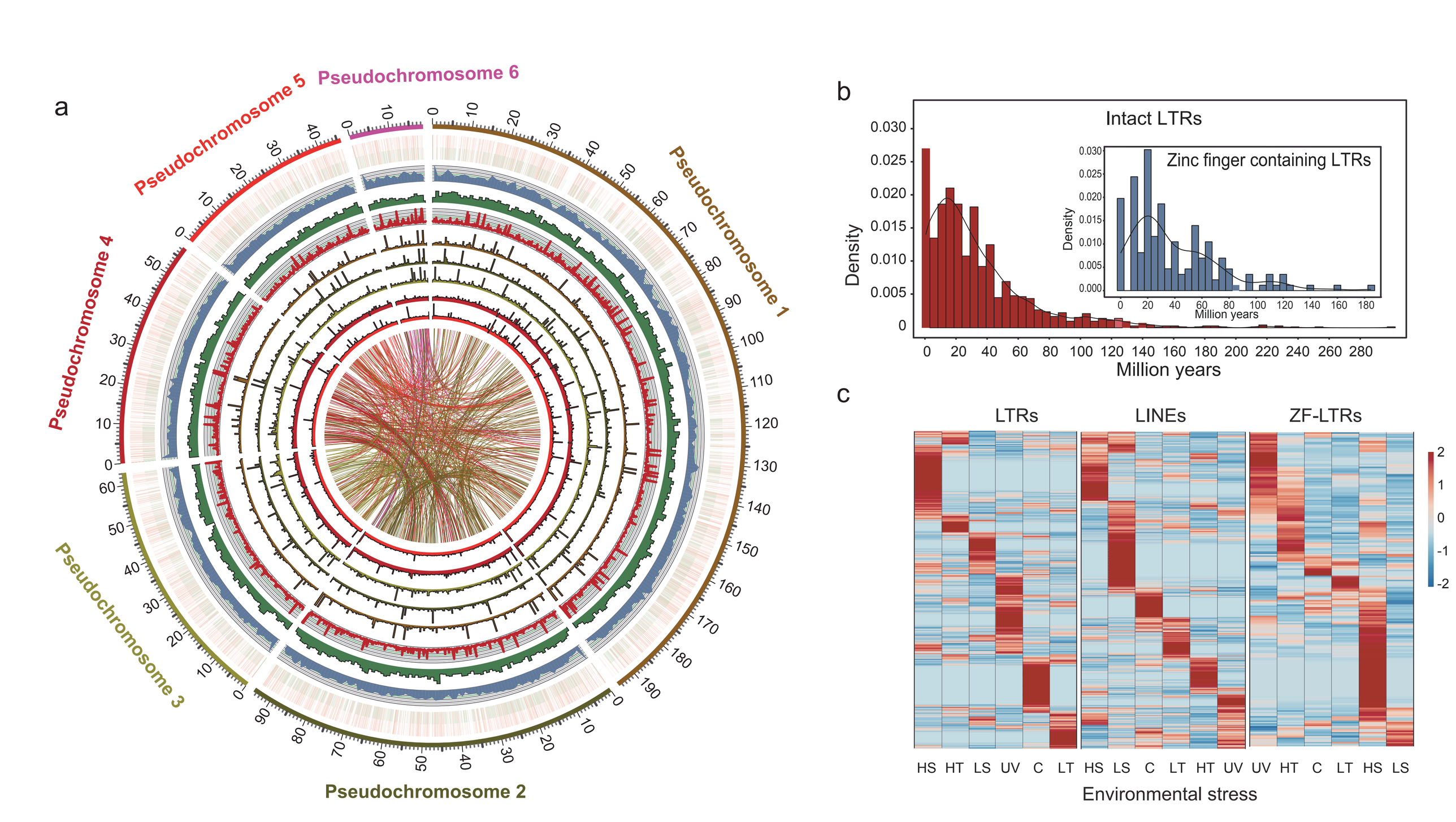
67. Löytynoja, A. *Phylogeny-aware alignment with PRANK: Multiple sequence alignment methods* (Humana Press, 2014).

68. Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* **17**, 540-552 (2000).

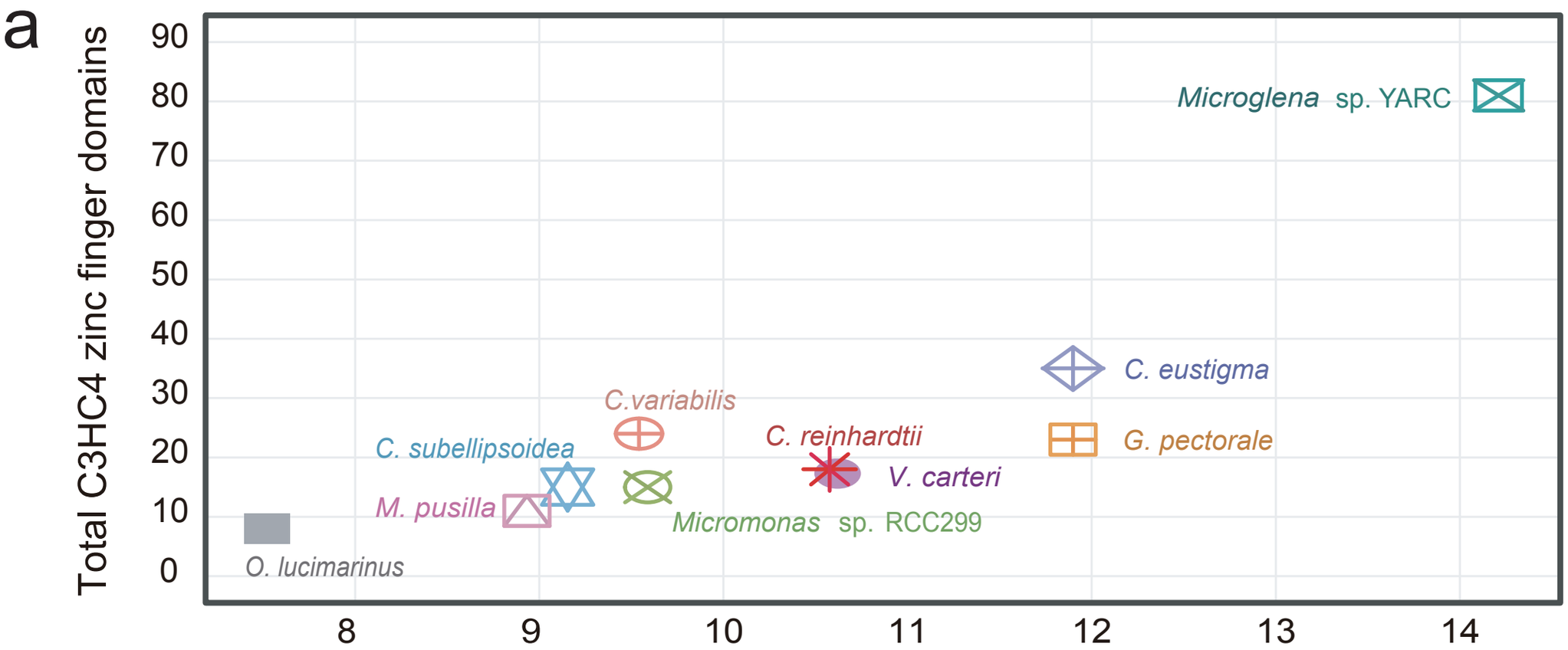
Sea of Change Consortium

Shazia N. Aslam[9], Kerrie Barry[10], Bank Beszteri[11], Corina Brussaard[12], Alicia Clum[10], Alex Copeland[10], Chris Daum[10], Anthony Duncan[13] Emiley Eloe-Fadrosh[10], Allison Fong[14], Brian Foster[10], Bryce Foster[10], Michael Ginzburg[14], Marcel Huntemann[10], Natalia N. Ivanova[10], Nikos C. Kyrpides[10], Kara Martin[12,15], Vincent Moulton[13], Supratim Mukherjee[10], Krishnaveni Palaniappan[10], T.B.K. Reddy[10], Simon Roux[10], Katrin Schmidt[16], Jan Strauss[17], Klaas Timmermans[12], Susannah G. Tringe[10], Graham J.C. Underwood[18], Klaus U. Valentin[14], Willem H. van de Poll[19], Neha Varghese[10]

[9]Department of Chemistry, Norwegian University of Science and Technology, Trondheim, 7491, Norway. [10] DOE-Joint Genome Institute, 1 Cyclotron Road, Berkeley, CA 94720, U.S.A. [11]Department of Biology, University of Duisburg-Essen, Essen, Universitaetsstrasse 2, 45141 Essen, Germany. [12]Royal Netherlands Institute for Sea Research, Texel, The Netherlands, [13]School of Computing Sciences, University of East Anglia, Norwich Research Park, NR47TJ, Norwich, U.K. [14]Alfred-Wegener Institute for Polar and Marine Research, Am Handelshafen 12, 27570 Bremerhaven, Germany. [15]Earlham Institute, Norwich Research Park, Norwich, NR4 7UG, United Kingdom. [16]School of Environmental Sciences, University of East Anglia, Norwich Research Park, NR47TJ, Norwich, U.K. [17]Ocean EcoSystems Biology Unit, GEOMAR Helmholtz Centre for Ocean Research Kiel, Düsternbrooker Weg 20, D-24105 Kiel, Germany. [18]School of Life Sciences, University of Essex, CO4 3SQ, Colchester, U.K. [19]Department of Ocean Ecosystems, Energy and Sustainability Research Institute Groningen, University of Groningen, Nijenborgh 7, 9747 AG Groningen, The Netherlands

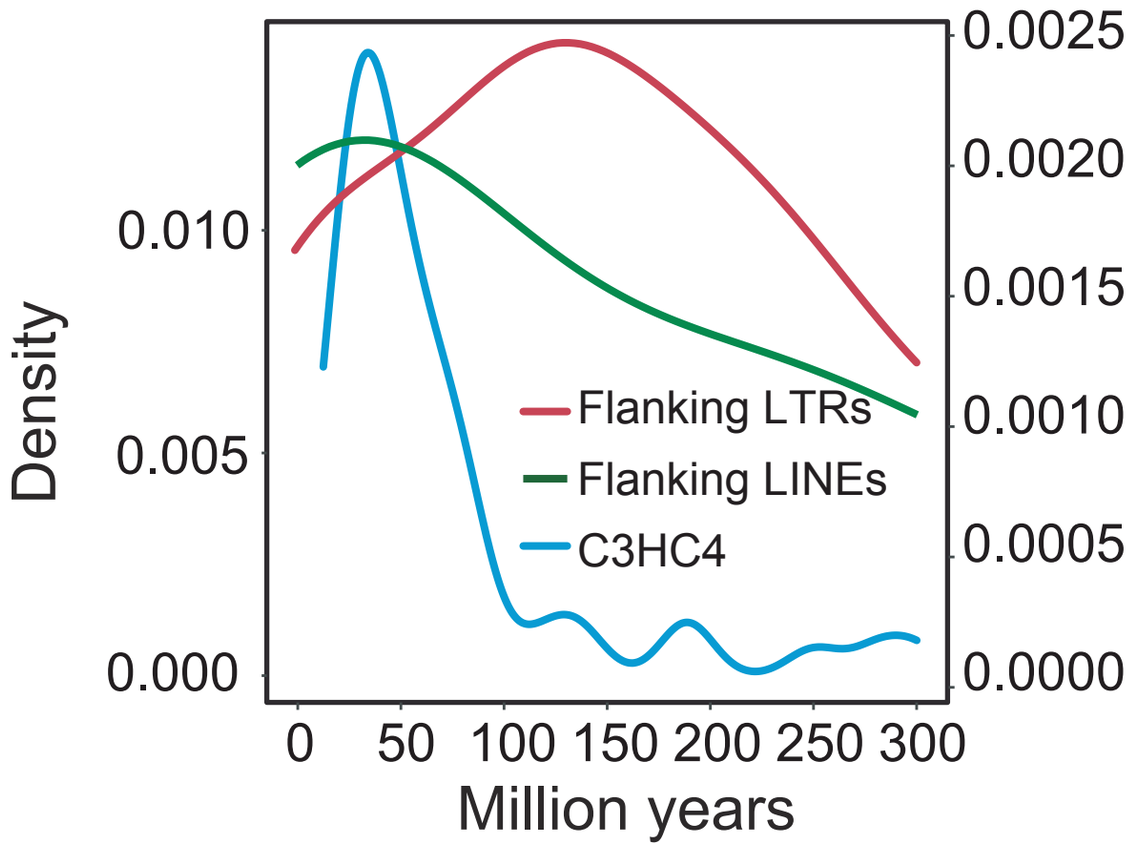

a
Pseudochromosome 6
Pseudochromosome 5
Pseudochromosome 1
Pseudochromosome 4
Pseudochromosome 3
Pseudochromosome 2
b
Intact LTRs
Zinc finger containing LTRs
Density
Million years
c
LTRs
LINEs
ZF-LTRs
HS HT LS UV C LT HS LS C LT HT UV UV HT C LT HS LS
Environmental stress

a

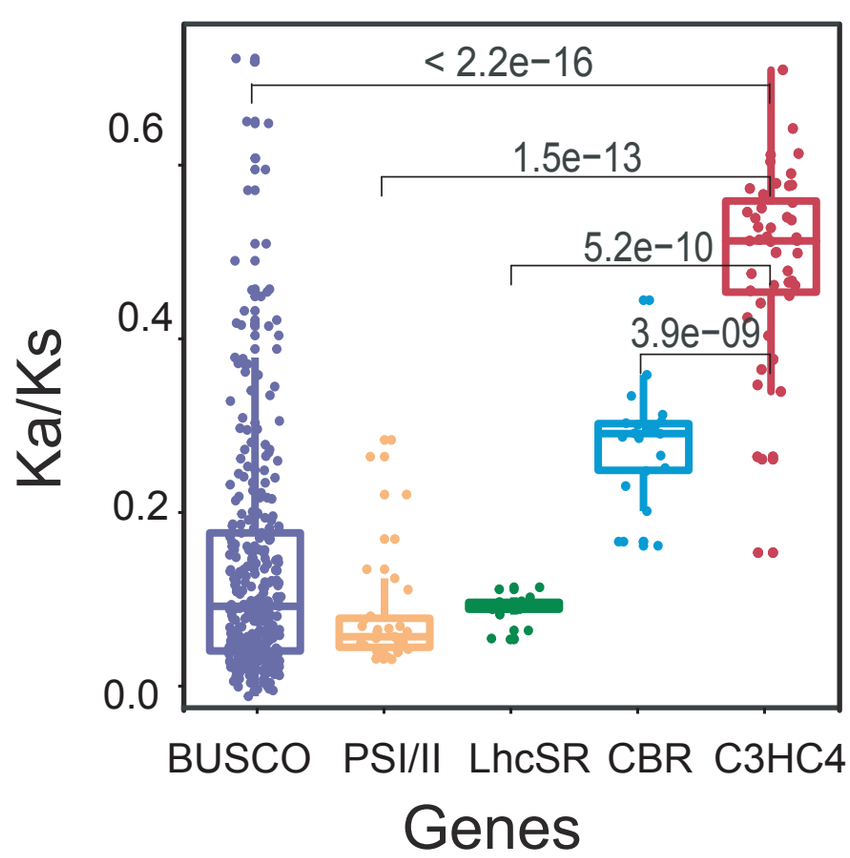


Total annotated domains (x 1, 000)

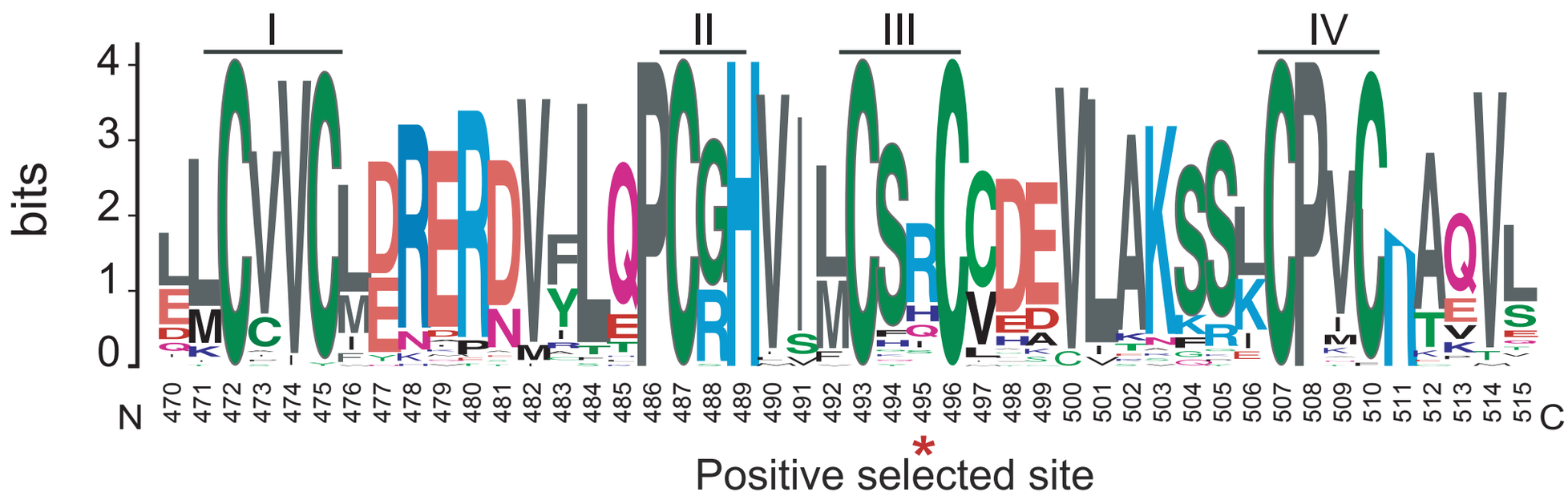
b



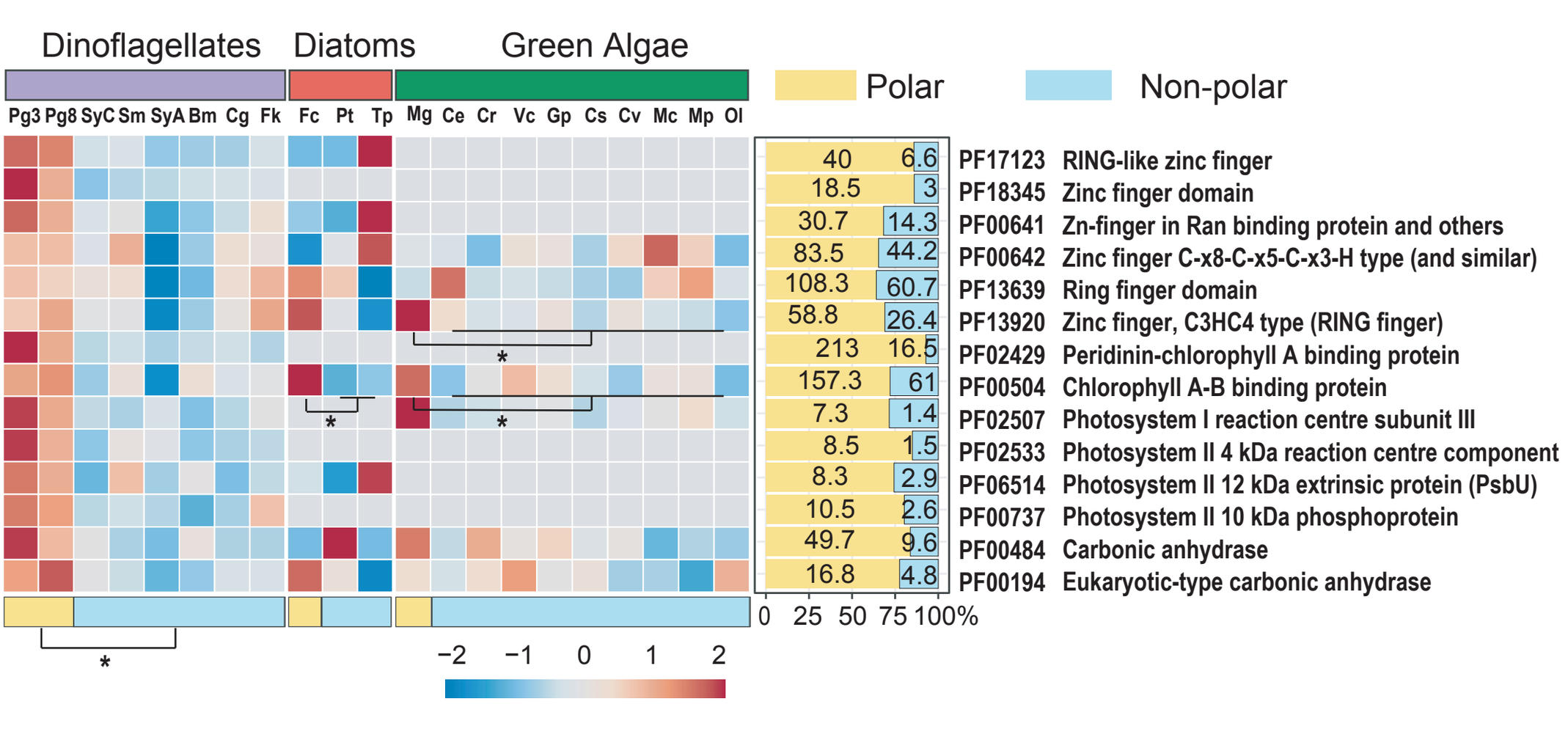
c

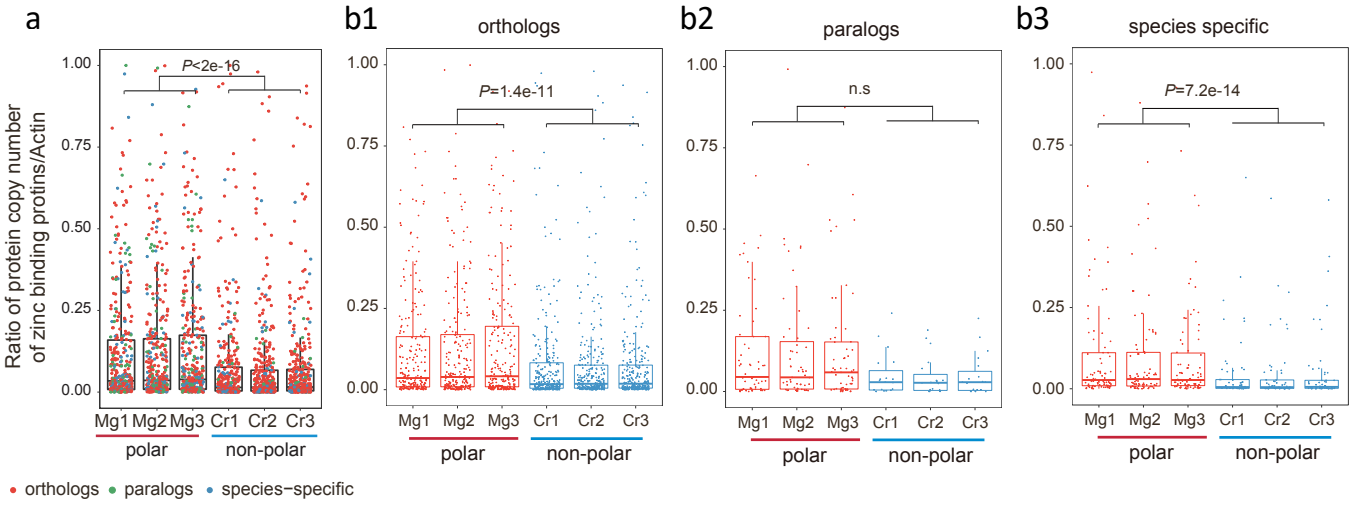


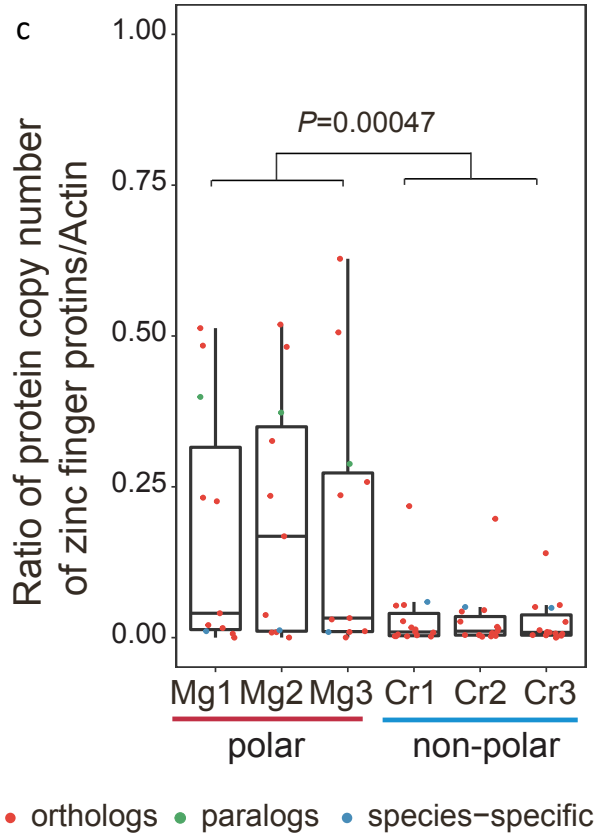
d

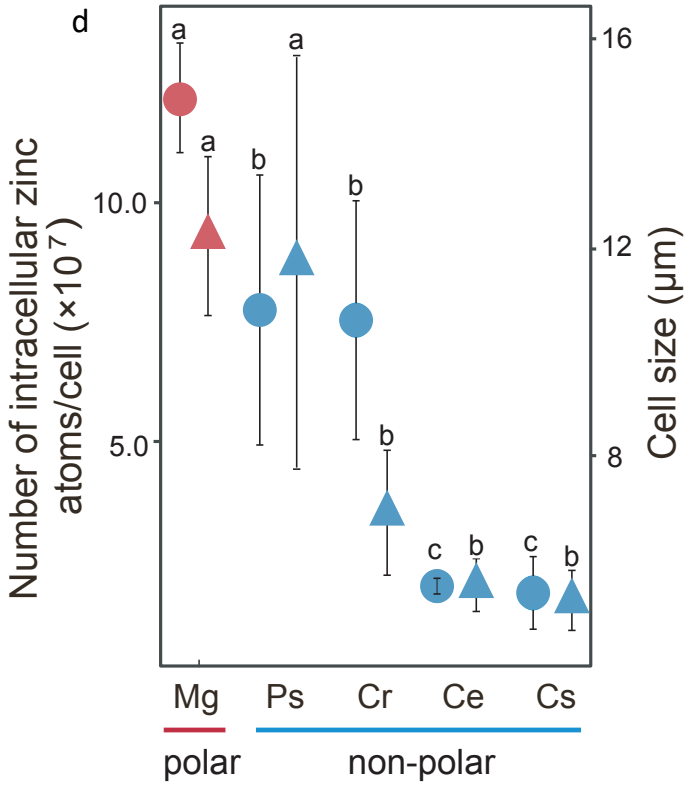


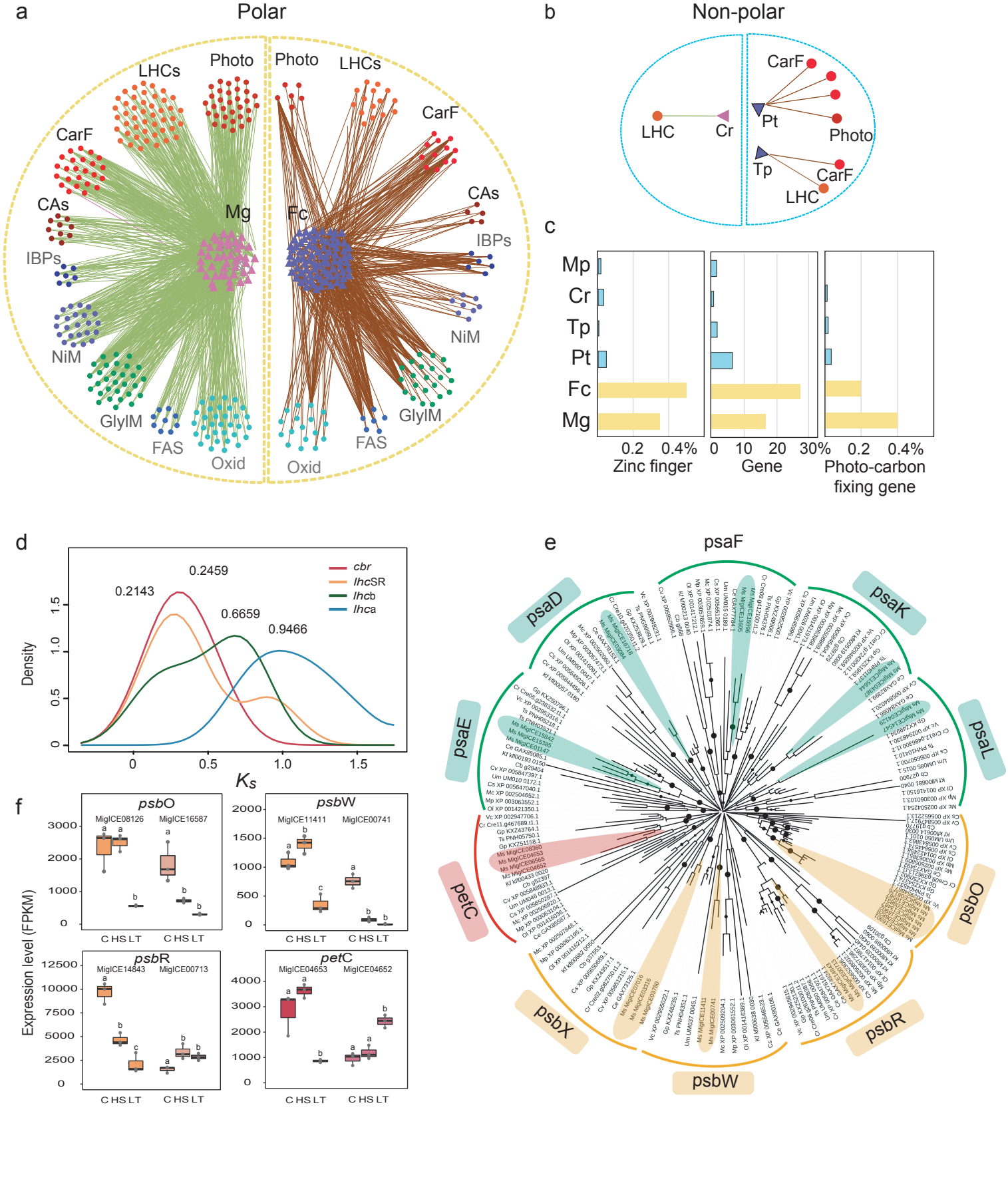
e



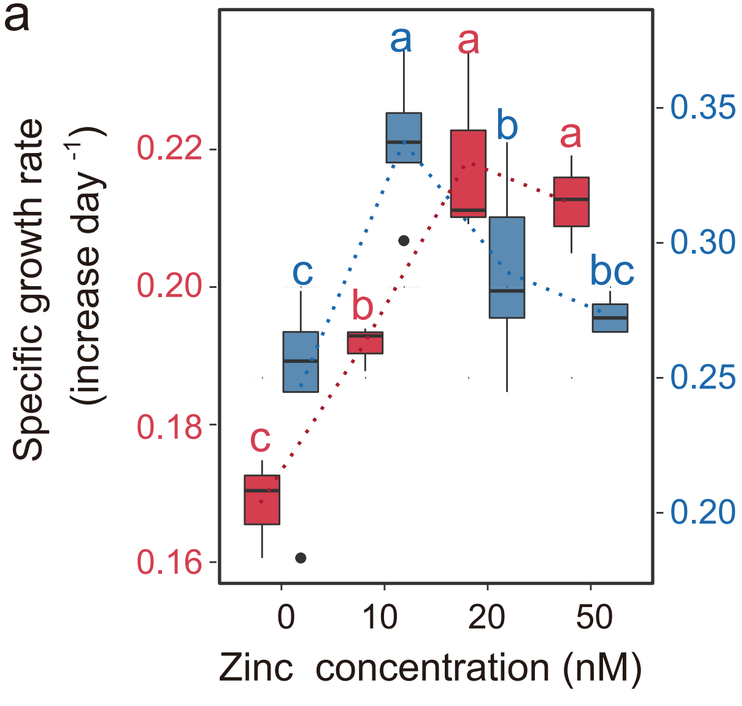

a
b1
orthologs
b2
paralogs
b3
species specific
Ratio of protein copy number of zinc binding protins/Actin
P<2e-16
P=1.4e-11
n.s
P=7.2e-14
1.00
0.75
0.50
0.25
0.00
Mg1 Mg2 Mg3 Cr1 Cr2 Cr3
polar
non-polar
orthologs paralogs species−specific


c
Ratio of protein copy number of zinc finger protins/Actin
P=0.00047
1.00
0.75
0.50
0.25
0.00
Mg1 Mg2 Mg3 Cr1 Cr2 Cr3
polar
non-polar
orthologs paralogs species−specific

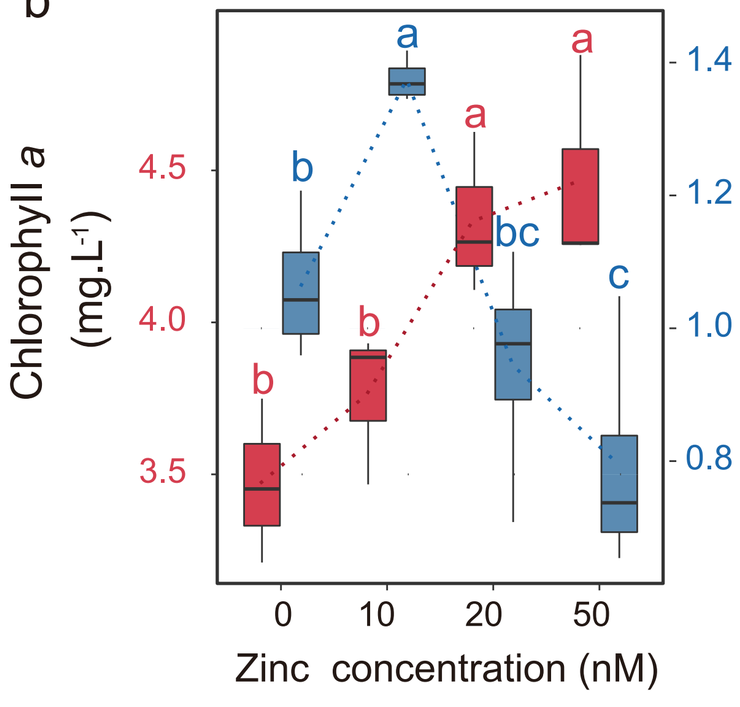

d
Number of intracellular zinc atoms/cell (×10 7)
Cell size (μm)
10.0
5.0
16
12
8
a a b b b c b c b
Mg Ps Cr Ce Cs
polar
non-polar


a
Polar
Photo
LHCs
CarF
CAs
IBPs
NiM
GlylM
FAS
Oxid
Mg
Fc
b
Non-polar
LHC
Cr
Pt
Tp
c
Mp
Cr
Tp
Pt
Fc
Mg
Zinc finger
Gene
Photo-carbon fixing gene
d
Density
Ks
cbr
lhcSR
lhcb
lhca
0.2143
0.2459
0.6659
0.9466
e
psaF
psaD
psaK
psaL
psaE
petC
psbO
psbR
psbW
psbX
f
psbO
psbW
psbR
petC
MigICE08126
MigICE16587
MigICE11411
MigICE00741
MigICE14843
MigICE00713
MigICE04653
MigICE04652
Expression level (FPKM)
C H S L T

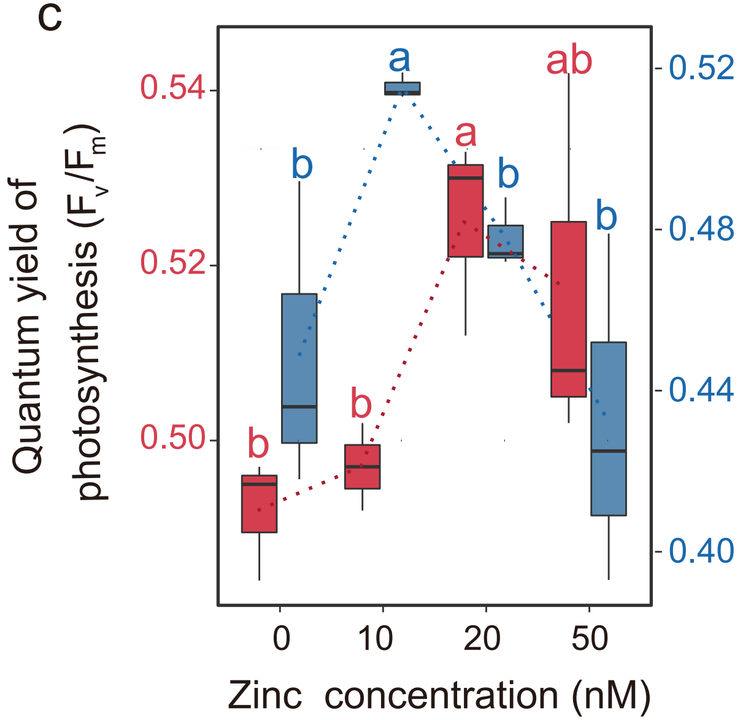
a


Specific growth rate
(increase day -1)
0.22
0.20
0.18
0.16
0.35
0.30
0.25
0.20
a
a
b
a
c
c
b
bc
0
10
20
50
Zinc concentration (nM)

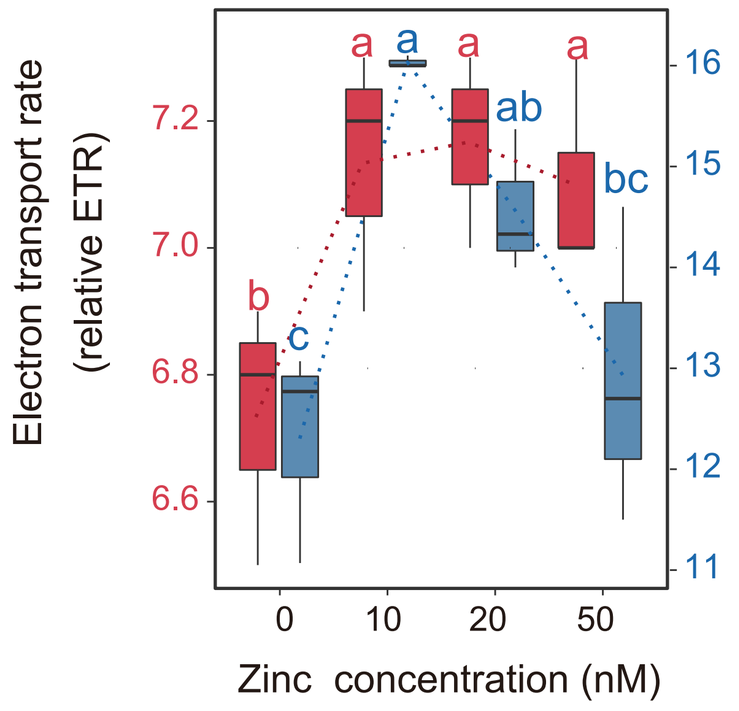
b


Chlorophyll a
(mg.L-1)
4.5
4.0
3.5
1.4
1.2
1.0
0.8
a
a
a
b
b
bc
b
c
0
10
20
50
Zinc concentration (nM)

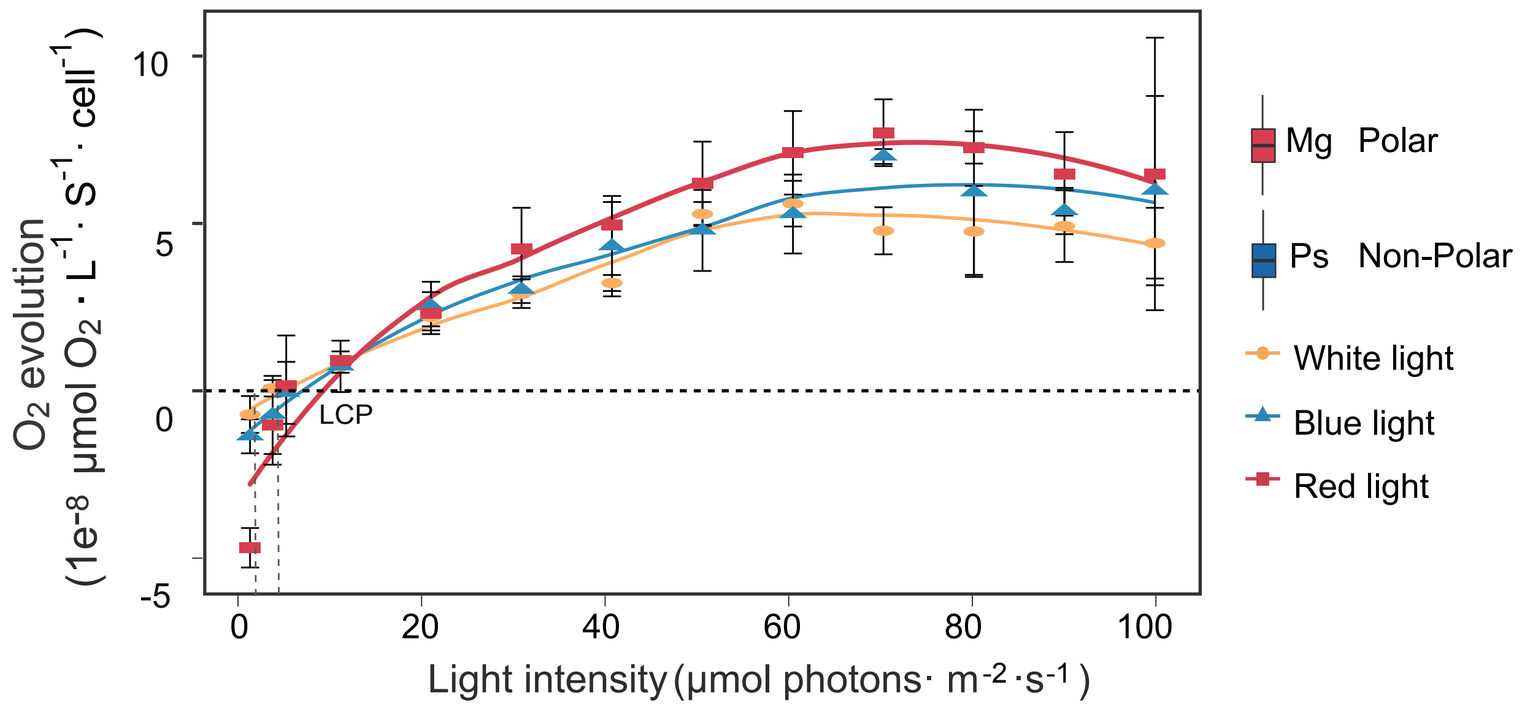
c

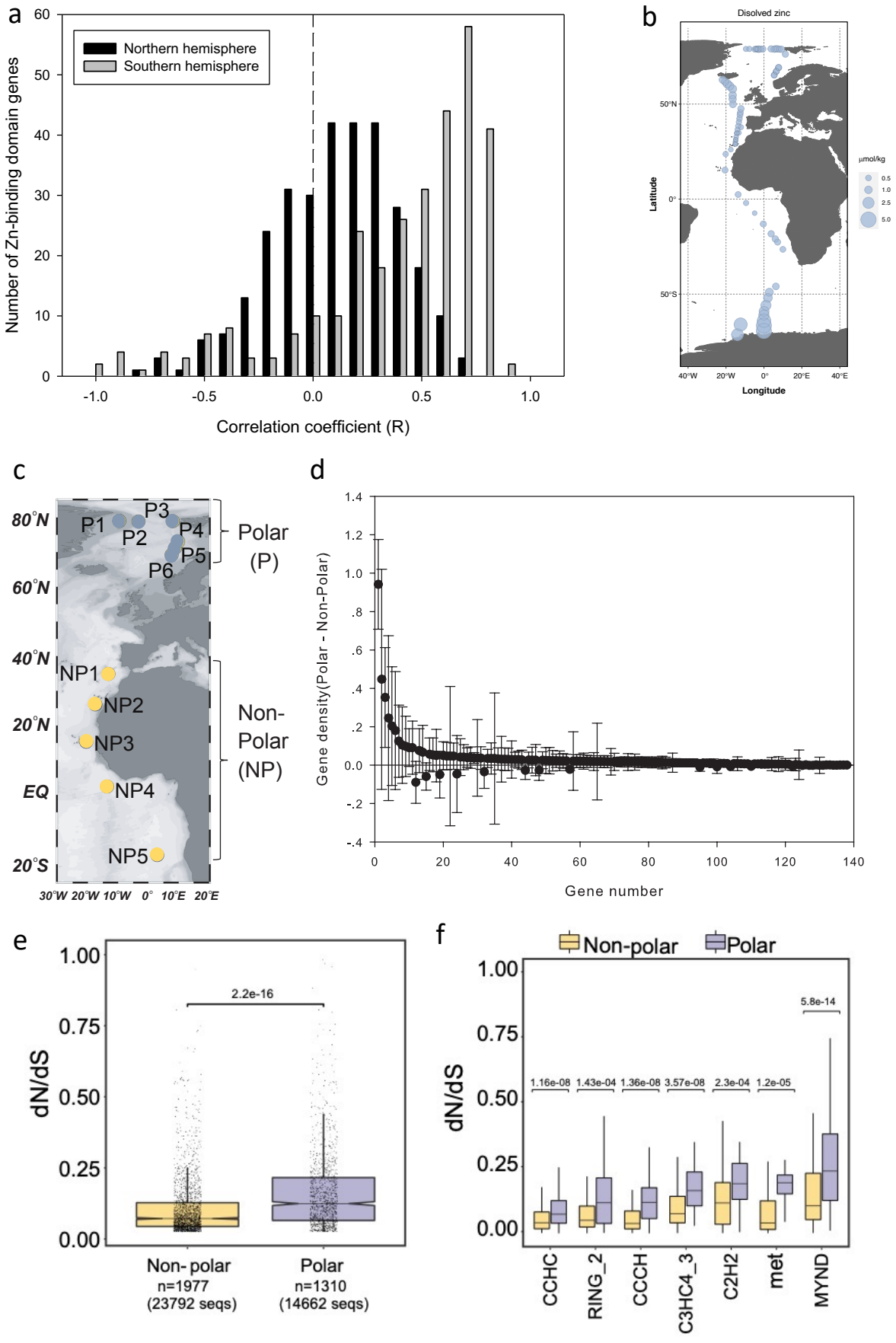

Quantum yield of
photosynthesis (Fv/Fm)
0.54
0.52
0.50
0.52
0.48
0.44
0.40
a
ab
b
a
b
b
b
b
0
10
20
50
Zinc concentration (nM)

d


Electron transport rate
(relative ETR)
7.2
7.0
6.8
6.6
16
15
14
13
12
11
a
a
a
a
ab
bc
b
c
0
10
20
50
Zinc concentration (nM)

e


O2 evolution
(1e-8 μmol O2 · L-1. s-1. cell-1)
10
5
0
-5
LCP
0
20
40
60
80
100
Light intensity (μmol photons· m-2 ·s-1)
Mg Polar
Ps Non-Polar
White light
Blue light
Red light


a
Number of Zn-binding domain genes
Northern hemisphere
Southern hemisphere
Correlation coefficient (R)
b
Disolved zinc
Latitude
Longitude
µmol/kg
c
P1
P2
P3
P4
P5
P6
Polar (P)
NP1
NP2
NP3
NP4
NP5
Non-Polar (NP)
d
Gene density(Polar - Non-Polar)
Gene number
e
dN/dS
2.2e-16
Non- polar n=1977 (23792 seqs)
Polar n=1310 (14662 seqs)
f
Non-polar
Polar
dN/dS
1.16e-08
1.43e-04
1.36e-08
3.57e-08
2.3e-04
1.2e-05
5.8e-14
CCHC
RING_2
CCCH
C3HC4_3
C2H2
met
MYND