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.

81 # Results

82

83 ## Comparative genomics and proteomics

84

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

114

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

126 PSI and II and those representing the BUSCO data set (Fig. 2C), indicative of an
127 accelerated rate of evolution. Interestingly, one site in the motifs[24] responsible for zinc-
128 ion binding appear to be under significant positive selection (Ka/Ks > 1) (Fig. 2D and
129 Supplementary Table S 5), which is consistent with adaptive evolution.
130

131 A broader comparative approach with distantly related cold-adapted polar algae such as
132 two strains of *Polarella glacialis* (Dinoflagellates)[25] and their non-polar relatives[26-29]
133 provided evidence for the commonality of the expansion of specifically zinc-finger domain
134 containing gene families in polar microalgae despite species-specific differences in their
135 diversity (Fig. 2E). Interestingly, a comparative genome analysis of gene families that co-
136 expanded together with the zinc-finger domain containing genes identified photosynthesis
137 genes, such as genes involved in light-harvesting, electron generation and transport, and
138 inorganic carbon acquisition (Fig. 2E). Interestingly, many of the proteins that are co-
139 expanded and involved in photosynthesis bind zinc as co-factor such as sedoheptulose-
140 1,7-biphosphatase, fructose-bisphosphate aldolase, and specific carbonic anhydrases.
141 Although there were species-specific differences in the diversity of photosynthesis genes,
142 the Pfam-domain count in % total revealed strong enrichment in polar algae similar to the
143 enrichment of zinc-finger domain containing gene families. The lack of expanded zinc-
144 finger domains in polar heterotropic bacteria and cold-adapted fish (Extended Data Fig.
145 3) suggests that zinc and its binding proteins contribute to regulating photosynthesis and
146 carbon acquisition in polar eukaryotic phytoplankton.
147

148 These comparative genomics data suggest that the demand for zinc might be higher in
149 polar microalgae if the expansion of zinc-finger proteins and zinc-binding proteins
150 involved in photosynthesis causes elevated cellular zinc concentrations. To test this
151 hypothesis, we combined global domain searches for all known zinc-binding proteins in
152 sequenced polar and non-polar green algae and diatoms complemented by quantitative
153 proteomics using *Microglena* sp. and the mesophilic counterpart *Chlamydomonas*
154 *reinhardtii* as model species. The relative contribution of genes encoding zinc-binding
155 proteins in the genome of *Microglena* sp. was estimated to be 11.54% which was
156 significantly higher (p-value = 0.03) compared to other green algae except a mesophilic
157 strain of *Micromonas pusilla* (11.94%; p-value = 0.005) (Supplementary Table S6).
158 However, the species *M. pusilla* is well known to have strains with frequent occurrence
159 in the Arctic Ocean[30]. The genome of the polar diatom *Fragilariopsis cylindrus* was
160 estimated to encode 10.75% zinc-binding proteins, which was significantly higher (p-
161 value = 0.03) compared to the non-polar diatoms *Phaeodactylum tricornutum* and
162 *Thalassiosira pseudonana* (Supplementary Table S6). Quantitative lable-free mass
163 spectrometry was performed with *Microglena* sp. and *C. reinhardtii* under zinc-replete
164 growth conditions to complement these *in-silico* estimates (Supplementary Data 2 and
165 3). A total of 396 and 384 zinc-binding proteins were identified in *Microglena* sp. and *C.*
166 *reinhardtii* protein extracts, which converts to their estimated total copy number of 4.64
167 $\pm 0.22 \times 10^8$ and $2.61 \pm 0.22 \times 10^8$, 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

252 10 nM in both algal species, but *P. subcordiformis* showed a much stronger decline at
253 higher concentrations. Under zinc-replete growth conditions, oxygen evolution
254 experiments with *Microgelana* sp. under white, blue and red light revealed a LCP between
255 2~8 μmol photons m^{-2} s^{-1} depending on the light spectrum applied (Fig. 5E).

256

257 **Zinc-binding genes in pole-to-pole metatranscriptomes**

258

259 To test whether the elevated level of zinc-binding proteins and therefore the cellular zinc
260 quota in polar model algae is representative for natural polar microalgal communities, we
261 correlated the normalized abundance of transcripts encoding zinc-binding proteins from
262 pole-to-pole eukaryotic metatranscriptomes with latitude[34]. Samples were obtained from
263 chlorophyll *a* – maximum layers as part of the project "Sea of Change: Eukaryotic
264 Phytoplankton Communities in the Arctic Ocean" (DOI: 10.25585/1488054). For our
265 correlations, we used transcripts from 346 domains known to bind zinc. The distribution
266 of correlation coefficients (R) between the number of reads of zinc-binding domain-
267 containing genes increase with latitude, as indicated by the mean R > 0 (One sample T-
268 test for the North: T = 9.6421, df = 300, p < 2.2e-16; South: T = 18.549, df = 305, p < 2.2e-
269 16) (Fig. 6A). However, the positive correlation between latitude and the number of reads
270 of zinc-binding domain containing genes is significantly stronger for the Southern (mean
271 (±StDev) R=0.44 (±0.42) compared to the North hemisphere (mean (±StDev) = 0.16
272 (±0.28) (Two sample T-test: T = 9.92, df = 535, p < 0.00001). The same trends with latitude
273 were observed for estimated surface concentrations of dissolved zinc (Fig. 6B) with a
274 weaker trend for the Northern hemisphere (mean R = 0.05; p-value = 1.36e-07) and a
275 stronger trend for the Southern hemisphere (mean R = 0.35; p-value < 2.2e-16). However,
276 as we have yet to build a comprehensive map of the surface ocean zinc inventory,
277 concentrations of dissolved zinc for this study were estimated based on integrating
278 publically available zinc data, new measurements and models (Extended Data Fig. 6).
279 The application of a generalized linear model indeed revealed that temperature
280 explaines most of the variance with zinc as the second most important variable after
281 temperature and before other nutrients tested (Supplementary Table 7). Consequently,
282 our results indicate that polar eukaryotic phytoplankton communities have more zinc-
283 binding proteins caused by both, an elevated copy number of transcripts encoding zinc-
284 binding proteins and an expansion especially of regulatory zinc-finger protein families.
285 These large-scale comparative metatranscriptomics data in combination with omics
286 studies using model algal species provide strong support for the previously identified
287 positive correlation between dissolved zinc concentrations and the modelled latitudinal
288 gradient of zinc-uptake ratios of phytoplankton[16]. Thus, the increased demand for zinc in
289 high-latitude phytoplankton appears to have resulted in elevated zinc-uptake ratios.

290

291 **Comparative evolutionary genomics using metagenomes**

292

293 Our comparative genomics, proteomics, and metatranscriptomics data suggest a
294 common link between the evolution of polar microalgae from diverse lineages and zinc-
295 based geochemistry. Thus, this link might therefore represent a unifying framework for
296 how the evolution and biology of primary producers in polar oceans feedback onto global

297 cycles of zinc and elements that co-vary such as silicon and carbon. Evidence for parallel
298 evolution in distantly related algal lineages (e.g. diatoms, dinoflagellates, chlorophytes)
299 that have converged in their adaptation under polar conditions including a response to
300 elevated levels of zinc have already provided some support for the existence of this
301 unifying framework (Figs. 1-4). However, for revealing if this framework is univerisal, we
302 need to go beyond individual species and therefore have selected natural phytoplankton
303 communities from across a latitudinal gradient to test if homologs in their polar
304 communities undergo similar adaptive evolution as identified in our polar model species.

305 Thus, we have retrieved homologs from six Arctic phytoplankton metagenomes and
306 compared them to five of their warm-water counterparts (Fig. 6C). All eleven
307 metagenomes were collected on two RV Polarstern (Alfred-Wegener Institute for Polar
308 and Marine Research, Bremerhaven, Germany) expeditions described by Martin et al.[34]
309 and Duncan et al.[35]. First, we tested if the density of zinc-finger domains, for which we
310 have unquivocal evidence of their adaptive evolution in polar model algae, is higher in
311 polar metagenomes. Indeed, the majority of these domains increase in density in polar
312 metagenomes (116 out of 138 pairwise comparisons, 84.0%), (one-sample T-test: N=138,
313 T=3.98, p=0.0001) (Fig. 6D), corroborating the hypothesis that the expansion of the
314 MYND zinc-finger family in the genome of *F. cylindrus* and the C3HC4 zinc-domain
315 containing genes in *Microglena sp*. are not isolated events. Second, we calculated the
316 ratio of non-synonomous to synonomous nucleotide substitutions (dN/dS) in zinc-finger
317 domain containing genes in comparsion to homologs from non-polar counterparts (Fig.
318 6E). Based on more than 10,000 sequences from each environment, the median of dN/dS
319 of these genes in polar phytoplankton metagenomes was significantly higher (Mann-
320 Whitney p-value 2.2×10^{-16}) compared to their non-polar counterparts. These data
321 suggest that purifying selection is more relaxed and/or that natural selection favours
322 diversification in the polar zinc-finger domain containing genes, allowing them to evolve
323 faster in polar phytoplankton. Interestingly, individual zinc-finger domain containing genes
324 families such as CCHC, C2H2, CCCH, MYND, and *met* had the largest and most
325 significant differences in dN/dS compared to their non-polar counterparts (Fig. 6F). This
326 is consistent with the observation that genes encoding for protein repeats like Zn-fingers,
327 new copies tend to experience rapid functional divergence due to the combined effects
328 of relaxed purifying selection and positive selection[36]. It furthermore suggests that the
329 expansion of specific zinc-finger domain containing gene families in natural polar
330 phytoplankton communities facilitated their rapid evolution and possible acquisition of
331 modified or new functions, enabling adaptations to the novel challenging conditions of the
332 surface polar ocean.This expansion not likely to have been driven by any substantial
333 changes in zinc availability[17], but may be more broadly linked to the emergence of polar
334 conditions of which zinc in combination with low temperatures facilitated the adaptive
335 evolution of algae to polar oceans.

336

337 ## Discussion

338 This study shows that elevated zinc concentrations in surface polar oceans have
339 facilitated the adaptive evolution of algae to conditions of these permanently low-
340 temperature environments. Zinc therefore has enabled the formation of some of the most
341 productive food webs on Earth. We conclude this based on a set of integrative and

342 comparative approaches that combine analyses of evolutionary genomics, quantitative
343 proteomics, co-expression analyses, zinc-dependent cellular physiology, pole-to-pole
344 eukaryotic phytoplankton metatranscriptomes, and metagenomes. Thus, the emerging
345 view that the particularly high zinc demands of polar diatoms regulate a large part of the
346 global zinc distribution**[14,15]** can be extended to include polar green algae and
347 dinoflagellates. Thus, massive parallel evolution in these distantly related taxa appears to
348 have enabled these primary producers to cope with the challenging conditions of polar
349 oceans.
350
351 In addition to the expansion and accelerated evolution of specific zinc-binding domain
352 containing protein families in three different algal classes, their co-expression networks
353 reveal that polar microalgae might require an increased diversity of zinc-binding domains
354 to regulate light harvesting, photosynthetic electron generation and transport, inorganic
355 carbon acquisition and other forms of primary metabolism such as nitrogen and fatty acid
356 metabolism. Interestingly, zinc-binding domains were also co-expressed with ice-binding
357 proteins (Fig. 4A), suggesting their involvement in coping with freezing conditions.
358 However, the specific zinc-dependent regulatory mechanisms facilitating primary
359 metabolism under harsh polar conditions remain to be identified in polar algae.
360 Photosynthetic processes would appear to be an appropriate target for their identification
361 because of the co-expression networks and a lack of expanded zinc-domains in non-
362 photosynthetic polar organisms.
363
364 As photosynthesis in polar ecosystems requires regulation under extreme seasonality of
365 24 hours light in summer and long periods of darkness in winter, algae not only need to
366 effectively compensate for over excitation of the photosynthetic electron transport chain
367 in summer, they also require efficient mechanisms to photosynthesis under extremely low
368 irradiance levels. The former has been realized in different polar algae by expanding high-
369 light inducible protein families, which contribute to the dissipation of energy and therefore
370 reduce effects of over excitation (e.g. production of radical oxygen species). For coping
371 with the dark end of the light spectrum, cold-adapted algae have evolved mechanisms to
372 significantly reduce their light compensation point (LCP), which is the light intensity at
373 which oxygen production equals consumption (Fig. 5E). For instance, the LCP of
374 *Microglena* sp., was estimated to be between 2~8 μmol photons m^{-2} s^{-1} (Fig. 5E), which
375 is similar to LCPs measured in polar sea-ice diatoms**[37]**, but they are significantly lower
376 than that of temperate algae such as *C. reinhardtii* (40 μmol photons m^{-2} s^{-1})**[38]**. A low LCP
377 can be achieved by efficient energy conversion and carbon fixation under low light and
378 low temperatures. Hence, it is likely that the expansion of LHC gene families in *Microglena*
379 sp., and the duplication of genes encoding for proteins involved in linear electron transport
380 underpins the reduction of the LCP in this species.
381
382 The co-expression with zinc-binding domain containing genes suggests that these
383 photosynthetic processes require regulatory support to function at polar temperatures. If
384 the regulation of these photosynthetic processes takes place at the level of transcription,
385 translation or even at the level of regulating protein activity remains to be seen. The costs
386 for the adaptation of photosynthetic processes (e.g. Fv/Fm, ETR) to polar environmental

387 conditions in marine microalgae appear to be an increased demand for the trace metal
388 zinc as indicated by increased copy numbers of zinc-binding proteins compared to non-
389 polar relatives. Field studies on natural phytoplankton communities from non-polar
390 regions (e.g. subarctic Pacific), which corroborate our laboratory-based growth
391 experiments, confirm the significantly lower requirement of zinc for sustaining maximum
392 growth rates in non-polar microalgae[39]. More broadly, the particularly high zinc demands
393 of polar phytoplankton and the role played in zinc biogeochemical cycling, indicates that
394 changes in their abundance and biodiversity due to a changing climate will directly
395 modulate zinc cycling throughout the global ocean, potentially affecting the regional
396 emergence of zinc limitation in low latitude oceans.

397

398 **Methods**

399 Microalgal materials and growth conditions

400 The polar species *Microglena* sp. YARC (YSFRI-03-00001) and the non-polar
401 species *Platymonas subcordiformis* (YSFRI-03-00011), *Chlorella* sp. (YSFRI-03-00010),
402 *Chlamydomonas reinhardtii* (YSFRI-03-00002), and *Chlamydomonas euryale* (YSFRI-
403 03-00058), were all obtained from the Algal genetic resources center (Algae Culture
404 Collection) at the Institute of Yellow Sea Fisheries Research Institute (YSFRI), Chinese
405 Academy of Fishery Sciences. Cultures were single cell sorted using monoclonal
406 screening on solid medium and bacterial contaminants were removed by treatment with
407 ampicillin (50 μg ml^{-1}) and kanamycin (50 μg ml^{-1}). Batch cultures of all species were
408 grown in nutrient replete Provasoli seawater medium[40] with 6°C + 40 μmol photons m^{-2}
409 s^{-1} for polar species and 25°C + 100 μmol photons m^{-2} s^{-1} for non-polar species
410 considering their inherent differences in temperature limits and light conditions required
411 for optimal growth.

412 Genome sequencing

413 High molecular weight DNA for whole-genome sequencing was extracted from a
414 monoclonal culture using Plant DNA Isolation Reagent (Takara Biomedical Technology
415 (Beijing) Co., Ltd.). Genomic DNA was sheared using a sonication device for short-insert
416 paired-end (PE) library construction. Short-insert libraries with a size of 180, 200, 300,
417 500 bp were constructed according to the instructions described in the Illumina library
418 preparation kit (Illumina Co., U.S.A.). All libraries were sequenced on an Illumina HiSeq
419 2500 sequencer. The raw reads were subsequently trimmed for quality using
420 Trimmomatic[42] (v.0.35). Illumina sequence adaptors were removed, low quality bases
421 from the start or end of raw reads were trimmed, and reads were scanned using a 4-bp
422 sliding window and trimmed when the average quality per base dropped below 15. The
423 clean data obtained from this process were used for subsequent analyses. For PacBio
424 library construction, the genomic DNA was sheared to ~10 kb, and short fragments below
425 the size of 7 kb were filtered using BluePipin (Sage Science Co., U.S.A). Filtered DNA
426 was then converted into the proprietary SMRTbell library using the PacBio DNA Template
427 Preparation Kit (Pacific Biosciences, U.S.A).

428 Genome assembly

429 Five different assembly methods, including DBG2OLC (hybrid)[43], Falcon (Pacbio
430 only) (https://github.com/PacificBiosciences/FALCON/)[44], SmartDenovo (Pacbio only),

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

440 ## Repeat annotation and mask

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

452 ## Gene prediction and annotation

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

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

474 ## 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 $T = d/2\mu$.

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: $\mu = d/2T$, 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 ($2\Delta \ln L$), and used a chi-square test, and compared it to a χ^2 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

519 31‰ salinity), low temperatures (LT, −2°C for 5 days), high temperatures (HT, +12°C for
520 5 days), high salinity (HS, 96‰ for 5 days), lessened salinity (LS, 16‰ for 5 days), and
521 ultraviolet radiation (UV, 60 μw cm^{-2} UV-B irradiance treatment for 4 hours). For different
522 illumination intensity treatment, *Microglena* were firstly darkened for 36 hours at their
523 suitable conditions, and then exposed to illumination intensity of 3, 40 and 200 μmol
524 photons m^{-2} s^{-1} for 1 hour.

525 Total RNA was extracted from 0.5 g tissue using an E.Z.N.A. Total RNA Kit (OMEGA,
526 America) according to the manufacturer's protocol. After total RNA was extracted, mRNA
527 were enriched using oligo (dT) magnetic beads. Then the strand-specific RNA-seq
528 libraries were constructed using NEBNext® Ultra™ II Directional RNA Library Prep Kit for
529 Illumina (NEB, Ipswich, MA, USA) in accordance with the manufacturer's instruction. The
530 quality of RNA-seq libraries were assessed by using a Fragment Analyzer (Advanced
531 Analytical, IA, USA), and the resulting libraries were sequenced on an Illumina HiSeq
532 2500 instrument producing pair-end reads of 150 nucleotides. The clean paired-end reads
533 were mapped to the *Microglena* sp. genome using TopHat v2.0.12[53]. Then, the FPKM
534 (fragments per kilobase of transcript sequence per million base pairs sequenced) value
535 of each gene was calculated to estimate gene expression levels using Cufflinks[54] v2.2.1
536 (http://cole-trapnell-lab.github.io/cufflinks/). Heatmaps of expression patterns were
537 generated in R using the pheatmap.

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

540

541 ## Total protein extraction and digestion

542 *Microglena* sp. and *C. reinhardtii* were grown under optimal growth conditions until
543 the middle of their exponential growth phase where cell were harvested. Then the algal
544 samples were grinded with liquid nitrogen, then BPP buffer were added in the ratio of
545 1:10. The solution were centrifuged at 12000 × g for 20 min at 4℃, and supernatants were
546 collected. The equal volume of Tris-saturated phenol were added and vortexed for 10 min
547 at 4°C. The solution were centrifuged at 12000g for 20 min at 4℃ and the phenol phase
548 were collected. The equal volume of BPP were added and vortexed for 10 min at 4°C.
549 The solution were centrifuged at 12000 × g for 20 min at 4℃ and the phenol phase were
550 collected. Five volume of pre-cooled 0.1M ammonium acetate in methanol were added
551 and precipitated protein at -20°C overnight. The supernatant was discarded by
552 centrifugation, and the precipitate was washed twice with 90% acetone. Discard the
553 supernatant by centrifugation and air dry the precipitate. The precipitate was re-
554 suspended with lysis buffer (1% SDS, 8 M urea, cocktail), then sonicate for 3 min on ice.
555 The lysates were centrifuged, and supernatants were collected to test the concentration
556 of protein in all samples. Protein concentrations were determined by Bicinchoninic acid
557 (BCA) method by BCA Protein Assay Kit (Beyotime biotechnology). Protein quantification
558 was performed according to the kit protocol.

559 Protein digestion was performed according to the standard procedure. Briefly, for
560 each sample tube containing 100 μg protein, appropriate TCEP was added to the final
561 concentration of 10mM and the tubes were incubated at 37 °C for 60 min. Appropriate
562 IAM was added to the final concentration of 40mM and reaction for 40 min in dark. Add
563 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

608 were rinsed three times with ultrapure water (Merck Millipore Corporation, Darmstadt,
609 Germany). To precisely manipulate the zinc concentration, we used the artificial seawater
610 (ATCC medium 1661 with minor modification) to prepare the zinc free Provasoli seawater
611 medium. The artificial seawater and Provasoli seawater medium was prepared using
612 ultrapure water. The zinc concentration was adjusted by $ZnSO_4$ to the final concentration
613 of 0, 10, 20 and 50 nM.

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

618 The specific growth rate (SGR) was calculated using the equation: SGR (increase
619 day^{-1}) = (lnN2-lnN1)/(T2-T1) where N1=cell number at time T1, N2=cell number at time
620 T2. Chlorophyll *a* and carotenoid contents were extracted by 95% ethanol. The contents
621 of chlorophyll *a* was determined spectrophotometrically as follows: $Chl_a = 13.36 \times A_{664.2} -$
622 $5.19 \times A_{648.6}$.

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

634

635 Relative abundance of zinc-binding domains in marine metatranscriptomes

636 The relative abundance of zinc-binding domains in the oceans was assessed using
637 the metatranscriptome data from the "Sea of Change: Eukaryotic Phytoplankton
638 Communities in the Arctic Ocean" project (DOI: 10.25585/1488054) hosted at JGI. This
639 dataset consists of sequence data from four separate cruises: ARK-XXVII/1 (PS80) - 17th
640 June to 9th July 2012; Stratiphyt-II - 1st May to 30th April 2011; ANT-XXIX/1 (PS81) - 1st to
641 24th November 2012 and ANT-XXXII/2 (PS103) - 20th December 2016 to 26th January
642 2017 and covers a transect of the Atlantic Ocean from Greenland to the Weddell Sea
643 (71.36°S to 79.09°N), (https://www.pangaea.de/expeditions/cr.php/Polarstern)[34]. Each
644 metatranscriptome dataset had been assembled and annotated with the JGI/IMG
645 pipeline[**64**]. We downloaded Pfam annotations and self-mapping files (alignments of the
646 raw reads to assembled contigs) for 72 metatranscriptomes. Using a custom Perl script,
647 we identified all unique contigs containing zinc-binding Pfam domains (from a filtered list
648 of 346) and converted the number of reads mapping to each contig to a percentage of
649 total mapped reads for each sample ((#reads mapped / #total reads mapped) * 100). We
650 then calculated the overall total for each domain for each sample with known latitude. To
651 assess the correlation between the normalised abundance of each zinc-binding domain

652 and latitude, we used the R function cor.test (Pearson's correlation coefficient, R) for each
653 of the N=301 domains in North, and N=306 domains in South. Finally, we plotted the
654 distributions of these 301 and 306 R coefficients of the Northern and Southern
655 hemisphere samples, respectively, and we tested whether the means of these
656 distributions were significantly greater than zero using a One Sample T-test. We thus
657 tested whether there was an overall positive correlation between the relative number of
658 reads of zinc-binding Pfam domains and latitude across more than 300 domains. We also
659 tested whether the distributions of R between both hemispheres differed from one another,
660 using a Two Sample T-test.

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

662 We downloaded contig sequences, gff files describing the contig coordinates of
663 predicted genes and Pfam annotation tables for 11 metagenomes from the "Sea of
664 Change: Eukaryotic Phytoplankton Communities in the Arctic Ocean" project (DOI:
665 10.25585/1488054) hosted at JGI. Six metagenomes were from polar regions (latitude
666 69.32°N to 79.02°N) and five from non-polar (latitude 34.88°N to 17.28°S) (Fig. 6c). As
667 only amino-acid sequences were available for predicted genes, we first set up a custom
668 Perl/BioPerl script to pull out gene sequences from the contig file for each sample based
669 on the contig name, contig coordinates and strand orientation for each predicted gene
670 from the corresponding gff file. Next, we compiled a list of all unique zinc-finger Pfam
671 domains ('zf-' prefix, 138 unique domains in total) contained in all 11 metagenome
672 samples and for each sample we produced a fasta file of all sequences containing each
673 zinc-finger domain. We then combined fasta files from each domain from all polar/non-
674 polar samples. We calculated the density of zinc-finger containing genes, tallying the
675 number of genes per one Mb CDS. We did this for both the polar and non-polar
676 metagenomes. We then subtracted the two distributions of the polar minus the non-polar
677 environment and repeated this procedure for all 138 zinc-finger domain containing genes.
678 Finally, we calculated the mean and the standard deviation of this difference for each of
679 these 138 genes (see 6d). This shows that the majority (116 out of 138, 84%) of zinc-
680 finger domains increase in density in the polar environment (one-sample T-test: N=138,
681 T=3.98, p=0.0001). The figure displays the absolute differences ranked, with the absolute
682 difference decreasing from the left to right (Fig. 6d).
683

684 The dN/dS analysis pipeline was implemented as following. For each multi-fasta file of
685 sequences representing all genes containing a specific zinc-finger domain from either
686 polar/non-polar samples we first clustered the sequences with cd-hit v4.6.8[65] at 100%
687 global identity to remove identical sequences. The non-redundant nucleotide sequences
688 were then translated into their first reading frame, amino-acid sequences with the BBmap
689 v37.28[66] utility translate6frames.sh and clustered with cd-hit requiring >=40% identity
690 over >=50% coverage of the longer sequence. We then produced a summary of the
691 sequence clustering with the cd-hit utility script clstr2txt.pl and used this with a custom
692 perl script to pull out the original nucleotide sequences for each cluster containing >=5
693 sequences into a separate multi-fasta file.

694 For each cluster we then aligned the gene sequences with Prank v170427[67] in codon
695 mode; removed poorly aligned sequences with trimal v1.2[67] requiring a residue overlap
696 of 0.5 and a sequence overlap of 50 and then removed gappy columns with Gblocks

697 v0.91b[68] in codon mode. We then produced a phylogenetic tree with RAxML v8.2.9
698 (Stamatakis, 2014), using the GTRGAMMA substitution model, with 100 bootstrap
699 replicates. The curated multiple sequence alignment was then converted to Paml format
700 with Prank and the alignment and RAxML tree was used for Paml v4.9[62] Codeml analysis
701 using model M0 (one average dN/dS ratio).

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

713

714 ## Acknowledgments

715 This work was supported by the national key research and development program of China
716 (2018YFD0900305), Special funds of Shandong Province for Pilot National Laboratory
717 for Marine Science and Technology (Qingdao) (2021QNLM050103-1), National Natural
718 Science Foundation of China (41676145, 32000404), Central Public-interest Scientific
719 Institution Basal Research Fund, YSFRI, CAFS (20603022020019, 20603022021019),
720 China Agriculture Research System (CARS-50), Taishan Scholars Funding of Shandong
721 Province, Young Taishan Scholars Program (tsqn202103136). The metatranscriptome
722 sequencing was conducted by the U.S. Department of Energy Joint Genome Institute, a
723 DOE Office of Science User Facility, which is supported by the Office of Science of the
724 U.S. Department of Energy under Contract No. DE-AC02-05CH1123. AT was supported
725 by the European Research Council (ERC) under the European Union's Horizon 2020
726 research and innovation programme (Grant agreement No. 724289). TM acknowledges
727 funding from the U.S. Department of Energy, Joint Genome Institute (Grant 532,
728 Community Science Program) and the Natural Environment Research Council (NERC)
729 (Grants NE/K004530/1; NE/R000883/1). TM and CvO acknowledge partial funding from
730 the School of Environmental Sciences at the University of East Anglia, Norwich Research
731 Park, United Kingdom.

732

733 ## Author contributions

734 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.
735 Q. analyzed the data; Y.W., D.X., J.Z., Y.Z., J.M. and Y.L. conducted the laboratorial
736 experiment; Sea of Change Consortium collected the samples and did DNA and RNA
737 extractions. The consortium also contributed to sequence data analysis; I.V.G.
738 coordinated metagenome and metatranscriptome sequencing. T.M., X.Z., N.Y., A.T. and
739 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 proteome raw data of *Microglena* sp. and *C. reinhardtii* were deposited into the intergrated proteome resources (iProX, under project accession IPX0004190000) 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

778 Wilcoxon test. For all boxplots, box bounds represent the first and third quartiles and
779 whiskers 1.5× the interquartile range; the center line represents the median. **d**, The
780 conserved motifs in *Microglena* C3HC4 zinc finger domains of clade III and the positive
781 selection site (red asterisk). I (C, C), II (C, H), III (C, C) and IV (C, C) indicates the four
782 pairs of amino acids participated in binding zinc ions. **e,** Heatmap of co-expanded Pfam
783 domains in eight sequenced dinoflagellate, three diatom, and ten green algal genomes.
784 Significantly expanded PFAM domains in polar microalgae (two-sided Fisher's exact
785 test p-value ≤ 0.05) are highlighted with an asterisk. The bar graph indicates the
786 average domain count (in % total) of respective Pfam domains in all 21 polar (yellow)
787 and non-polar (blue) algal species. Pg3, *Polarella glacialis* CCMP1383; Pg8, *Polarella*
788 *glacialis* CCMP2088; Syc, *Symbiodinium* sp. clade C; Sm, *Symbiodinium*
789 *microadriaticum*; Sya, *Symbiodinium* sp. clade A3; Bm, *Breviolum minutum*; Cg,
790 *Cladocopium goreaui*; Fk, *Fugacium kawagutii*; Fc, *Fragilariopsis cylindrus*; Pt,
791 *Phaeodactylum tricornutum*; Tp, *Thalassiosira pseudonana*; Mg, *Microglena* sp. YARC;
792 Ce, *Chlamydomonas eustigma*; Cr, *Chlamydomonas reinhardtii*; Vc, *Volvox carteri*; Gp,
793 *Gonium pectorale*; Cs, *Coccomyxa subellipsoidea*; Cv, *Chlorella variabilis*; Mc,
794 *Micromonas commoda* RCC299; Mp, *Micromonas pusilla*; Ol, *Ostreococcus*
795 *lucimarinus*.
796
797 **Fig. 3. Comparison of protein copy numbers, cellular zinc concentrations and cell**
798 **sizes between polar and non-polar green microalgae. a, b,** Comparison of the copy
799 number of the total zinc-binding proteins (**a**) and their separation into orthologs,
800 paralogs and species-specific proteins (**b**). **c,** Comparison of the zinc-finger proteins of
801 polar *Microglena* (Mg) and non-polar *C. reinhardtii* (Cr) (two-sided wilcoxon test). N = 3
802 biologically independent samples. Box plots show the Q1 and Q3 (the 25th and 75th
803 percentile, or the interquartile range, IQR), with the median in the centre, and the
804 whiskers denoting Q1 – 1.5 * IQR and Q3 + 1.5 * IQR. **d**, Number of intracellular zinc
805 atoms per cell (circles, left y-axis) and cell size (triangles, right y-axis) of polar (red) and
806 non-polar green algae (blue).Two-sided Duncan's multiple range test, $p < 0.05$. N = 3
807 biologically independent samples. Data are presented as mean values ± SEM. Mg,
808 *Microglena* sp.; Ps, *Platymonas subcordiformis*; Cr, *Chlamydomonas reinhardtii*; Ce,
809 *Chlamydomonas euryale,* Cs, *Chlorella* sp..
810
811 **Fig. 4. The co-expression of zinc finger genes and the expansion and expression**
812 **of photosynthesis genes in *Microglena* sp.. a, b,** Co-expression networks of zinc
813 finger genes and their co-regulated genes in (**a**) the polar microalgae *Microglena* sp.
814 (Mg) and *Fragilariopsis cylindrus* (Fc). and in (**b)** their non-polar counterparts
815 *Chlamydomonas reinhardtii* (Cr), *Phaeodactylum tricornutum* (Pt) and *Thalassiosira*
816 *pseudonana* (Tp). **c**, Co-expression ratios in % total for zinc-finger genes, all co-
817 expressed genes and photosynthesis and carbon fixing genes in polar and non-polar
818 microalgae. Photo, Photosynthesis genes; LHCs, Antenna proteins encoding genes;
819 CarF, carbon fixing genes; CAs, carbonic anhydrase genes; IBPs, ice-binding protein

820 encoding genes; NiM, nitrogen metabolism genes; GlylM, glycerolipid metabolism
821 genes; FAs, fatty acid biosynthesis genes; Oxid, oxidative phosphorylation genes. **d**,
822 Distribution of synonymous substitutions (*Ks*) of four LHC subfamily genes. The number
823 on each curve represents the peak of *Ks*. LhcSR and CBR encoding genes were
824 candidates for photoprotection, and Lhcb and Lhca were light harvesting proteins for
825 PSII and PSI, respectively. **e**, Unrooted genealogy of PSI (*psa*D, *psa*E, *psa*F, *psa*K,
826 *psa*L), PSII (*psb*X, *psb*W, *psb*R, *psb*O) and photosynthetic electron transport chain *pet*C
827 genes in 11 green algae: Ms, *Microglena* sp.; Cb, *Chara braunii*; Chl, *Chlorella*
828 *variabilis*; Ce, *Chlamydomonas eustigma*; Cr, *Chlamydomonas reinhardtii*; Gp, *Gonium*
829 *pectorale*; Cs, *Coccomyxa subellipsoidea*; Kf, *Klebsormidium flaccidum*; Mc,
830 *Micromonas* RCC229; Mp, *Micromonas* CCMP1545; Um, *Ulva mutabilis*; Vc, *Volvox*
831 *carteri*; Ts, *Tetrabaena socialis*; Ol, *Ostreococcus lucimarinus*; Cv, *Chlorella variabilis*. **f**,
832 Diverged expression of *psb*O, *psb*W, *psb*R and *pet*C paralogous genes under high salt
833 (HS), at low temperature (LT) and under control growth conditions (C). N = 3 biologically
834 independent samples. Box plots show the Q1 and Q3 (the 25th and 75th percentile, or
835 the interquartile range, IQR), with the median in the centre, and the whiskers denoting
836 Q1 – 1.5 * IQR and Q3 + 1.5 * IQR. Different letters on error bars indicate statistically
837 significant differences (Two-sided Duncan's multiple range test, $p < 0.05$).
838
839 **Fig. 5. Zinc-dependent growth rates and photophysiology of polar vs non-polar**
840 **green algal species. a**, **b, c, d,** Specific growth rate (**a**), chlorophyll *a* concentration (**b**),
841 quantum yield of photosynthesis (**c**) and the relative electron transport rate (**d**) of the
842 polar green alga *Microglena* sp. and the non-polar relative *P. subcordiformis* under
843 different concentrations of zinc in the growth medium. N = 3 biologically independent
844 samples. Different letters on error bars indicate statistically significant differences (Two-
845 sided Duncan's multiple range test, $p < 0.05$). For all boxplots, box bounds represent
846 the first and third quartiles and whiskers 1.5x the interquartile range; the center line
847 represents the median. **e,** Photosynthesis-response curve of *Microglena* sp. under
848 white, blue and red light. Oxygen evolution (μmol O_2 L^{-1} s^{-1} $cell^{-1}$) was measured at 6°C
849 using micro-electrodes. N=3 biologically independent samples. Data are presented as
850 mean values ± SEM.
851
852 **Fig. 6. Eukaryotic phytoplankton metatranscriptomes and -genomes with focus**
853 **on zinc-binding domain containing genes and their substitution rates (dN/dS) in**
854 **relation to latitude and the estimated concentration of dissolved zinc in the**
855 **surface ocean. a,** Distribution of correlation coefficients (R) between the number of
856 reads of zinc-binding domain containing genes and the latitude of the sample (a). Both
857 in the Northern and Southern hemisphere, the relative number of reads of Zn-binding
858 domain containing genes increase at higher latitude, as indicated by the mean R > 0
859 (One sample T-test for the North: T = 9.6421, df = 300, p < 2.2e-16; South: T = 18.549,
860 df = 305, p < 2.2e-16). The positive correlation between latitude and the number of
861 reads of zinc-binding domain containing genes is significantly stronger for the Southern
862 hemisphere (mean (± SD) R = 0.44 (± 0.42) compared to the North (mean (± SD) = 0.16
863 (± 0.28) (Two sample T-test: T = 9.92, df = 535, p < 0.00001). **b,** Estimated
864 concentrations of dissolved zinc (nmol/kg) for stations subjected to metatranscriptome
865 sequencing. **c,** Geographical map showing the sampling locations for metagenomes

866 (Chlorophyll *a* maximum layer). Blue colour indicates sampling stations north (P = polar
867 ≥ 66.3°, n=6) and yellow colour south (NP = non-polar ≤ 66.3°, n=5) of the Arctic
868 Circle. **d,** Difference in density of zinc-finger containing genes (mean (± SD) number of
869 genes per Mb CDS) between polar and non-polar metagenomes. The majority (116 out
870 of 138) of zinc-finger domains increase in density in the polar environment (one-sample
871 T-test (two-tailed): N=138, T=3.98, p=0.0001). Figure shows the mean difference and
872 SD of all 138 zinc-finger domains. The absolute differences are placed in rank order so
873 that they decrease from left to right. **e,** Box plot of dN/dS from all identified zinc-finger
874 domain containing genes. Number indicates Mann-Whitney p-value**. f,** Box plot of dN/dS
875 from zinc-finger domain containing genes deemed to be significantly (Mann-Whitney p-
876 value ≤ 0.001) higher in genes from polar compared to non-polar metagenomes. Box
877 plots show the Q1 and Q3 (the 25th and 75th percentile, or the interquartile range, IQR),
878 with the median in the centre, and the whiskers denoting Q1 – 1.5 * IQR and Q3 + 1.5 *
879 IQR. Raw data in panel E are shown as dots.
880
881

882 ## References

883 1. Field, C. B., Behrenfeld, M. J., Randerson, J. T. & Falkowski, P. Primary Production
884 of the Biosphere: Integrating Terrestrial and Oceanic Components. *Science* **281**, 237
885 (1998).

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

888 3. Saito M. A., Sigman D. M. & Morel, F. M. M. The bioinorganic chemistry of the
889 ancient ocean: the co-evolution of cyanobacterial metal requirements and
890 biogeochemical cycles at the Archean–Proterozoic boundary? *Inorg Chim Acta* **356**,
891 308-318 (2003).

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

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

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

898 7. Ho, T.-Y. et al. The Elemental Composition of Some Marine Phytoplankton. *J Phycol*
899 **39**, 1145-1159 (2003).

900 8. Ellwood, M.J. Wintertime trace metal (Zn, Cu, Ni, Cd, Pb and Co) and nutrient
901 distributions in the Subantarctic Zone between 40–52°S; 155–160°E. *Mar Chem* **112**,
902 107-117 (2008).

903 9. Zhao, Y., Vance, D., Abouchami, W. & de Baar, H. J. W. Biogeochemical cycling of
904 zinc and its isotopes in the Southern Ocean. *Geochim Cosmochim Ac* **125**, 653-67
905 (2014).

906 10. John, S. G., Helgoe, J. & Townsend, E. Biogeochemical cycling of Zn and Cd and
907 their stable isotopes in the Eastern Tropical South Pacific. *Mar Chem* **201**, 256–262
908 (2018).

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

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

914 13. Sunda, W. G. & Huntsman, S. A. Cobalt and zinc interreplacement in marine
915 phytoplankton: Biological and geochemical implications. *Limnol Oceanogr* **40**, 1404-
916 1417 (1995).

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

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

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

923 17. Scott, C. et al. Bioavailability of zinc in marine systems through time. *Nat Geosci* **6**,
924 125-128 (2012).

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

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

929 20. Zhang, Z. H. et al. Adaptation to Extreme Antarctic Environments Revealed by the
930 Genome of a Sea Ice Green Alga. *Curr Biol* **30**,1-12 (2020).

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

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

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

937 24. Barlow, P. N. et al. Structure of the C3HC4 domain by 1H-nuclear magnetic resonance
938 spectroscopy.: A new structural class of zinc-finger. *Journal of molecular biology* **237**, 201-
939 211 (1994).

940 25. Stephens, T. G. et al. Genomes of the dinoflagellate Polarella glacialis encode
941 tandemly repeated single-exon genes with adaptive functions. *BMC Biology* **18**, 1-21
942 (2020).

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

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

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

949 29. Shoguchi, E.et al. Two divergent Symbiodinium genomes reveal conservation of a
950 gene cluster for sunscreen biosynthesis and recently lost genes. *BMC Genomics.* **19**, 458
951 (2018).

952 30. Hoppe, C. J. M., Flintrop, C. M. & Rost, B. The Arctic picoeukaryote Micromonas
953 pusilla benefits synergistically from warming and ocean acidification. *Biogeosciences* **15**,
954 4353-4365 (2018).

955 31. Ferguson, R. E.et al. Housekeeping proteins: A preliminary study illustrating some
956 limitations as useful references in protein expression studies. *Proteomics* **5**, 566-571
957 (2005).

958 32. Aslam, S.N. et al. Identifying metabolic pathways for production of extracellular
959 polymeric substances by the diatom Fragilariopsis cylindrus inhabiting sea ice. *ISME J*
960 **12**, 1237-1251 (2018).

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

963 34. Martin, K. et al. The biogeographic differentiation of algal microbiomes in the upper
964 ocean from pole to pole. *Nat Commun.* **12**, 5483 (2021).

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

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

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

971 38. Rühle, T., Hemschemeier, A., Melis, A. & Happe, T. A novel screening protocol for the
972 isolation of hydrogen producing Chlamydomonas reinhardtiistrains. *BMC Plant Biol* **8**, 107
973 (2008).

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

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

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

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

982 43. Ye, C. X. et al. DBG2OLC: Efficient Assembly of Large Genomes Using Long
983 Erroneous Reads of the Third Generation Sequencing Technologies. *Sci Rep* **6**, 31900
984 (2016).

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

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

989 46. Ruan, J. & Li, H. Fast and accurate long-read assembly with wtdbg2. *Nat Methods*
990 **17**, 155-158 (2020).

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

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

995 49. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR
996 retrotransposons. *Nucleic Acids Res* **35**, W265-W268 (2007).

997 50. Ou, S. J. & Jiang, N. LTR_retriever: A Highly Accurate and Sensitive Program for
998 Identification of Long Terminal Repeat Retrotransposons. *Plant Physiol* **176**, 1410-1422
999 (2018).

1000 51. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site
1001 identification. *BMC Bioinformatics* **11**, 119 (2010).

1002 52. Stanke, M., Schöffmann, O., Morgenstern, B. & Waack, S. Gene prediction in
1003 eukaryotes with a generalized hidden Markov model that uses hints from external
1004 sources. *BMC Bioinformatics* **7**, 62 (2006).

1005 53. Kim, D.et al. TopHat2: accurate alignment of transcriptomes in the presence of
1006 insertions, deletions and gene fusions. *Genome Biol* **14**, R36 (2013).

1007 54. Trapnell, C. et al. Differential gene and transcript expression analysis of RNA-seq
1008 experiments with TopHat and Cufflinks. *Nat Protoc* **7**, 562-578 (2012).

1009 55. Haas, B.J. et al. Automated eukaryotic gene structure annotation using
1010 EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol* **9**, R7
1011 (2008).

1012 56. Mitchell, A. L. et al. InterPro in 2019: improving coverage, classification and access
1013 to protein sequence annotations. *Nucleic Acids Res* **47**, D351-D360 (2018).

1014 57. Rice, P., Longden, I. & Bleasby, A. EMBOSS: The European Molecular Biology Open
1015 Software Suite. *Trends Genet* **16**, 276-277 (2000).

1016 58. Kimura, M. A simple method for estimating evolutionary rates of base substitutions
1017 through comparative studies of nucleotide sequences. *J Mol Evol* **16**, 111-120 (1980)..

1018 59. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for
1019 comparative genomics. *Genome Biol* **20**, 238 (2019).

1020 60. Zhang, Z. et al. KaKs_Calculator: Calculating Ka and Ks Through Model Selection
1021 and Model Averaging. *Genom Proteom Bioinf* **4**, 259-263 (2006).

1022 61. Kumar, S., Stecher, G., Suleski, M. & Hedges, S. B. TimeTree: A Resource for
1023 Timelines, Timetrees, and Divergence Times. *Mol Biol Evol* **34**,1812-1819 (2017).

1024 62. Yang, Z. H. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**,
1025 1586-1591 (2007).

1026 63. Wiśniewski, J. R., Hein, M. Y., Cox, J. & Mann, M. A ‘proteomic ruler’ for protein copy
1027 number and concentration estimation without spike-in standards. *Mol Cell Proteomics* **13**,
1028 3497-3506 (2014).

1029 64. Huntemann, M. et al. The standard operating procedure of the DOE-JGI Metagenome
1030 Annotation Pipeline (MGAP v.4). *Stand Genomic Sci* **10**, 86 (2016).

1031 65. Fu, L. et al. CD-HIT: accelerated for clustering the next-generation sequencing data.
1032 *Bioinformatics* **28**, 3150-3152 (2012).

1033 66. Bushnell, B. *BBMap: a fast, accurate, splice-aware aligner* (Lawrence Berkeley
1034 National Lab., 2014).

1035 67. Löytynoja, A. *Phylogeny-aware alignment with PRANK: Multiple sequence alignment*
1036 *methods* (Humana Press, 2014).

1037 68. Castresana, J. Selection of conserved blocks from multiple alignments for their use in

1038 phylogenetic analysis. *Mol Biol Evol* **17**, 540-552 (2000).

1039

1040 **Sea of Change Consortium**

1041 Shazia N. Aslam[9], Kerrie Barry[10], Bank Beszteri[11], Corina Brussaard[12], Alicia Clum[10],
1042 Alex Copeland[10], Chris Daum[10], Anthony Duncan[13] Emiley Eloe-Fadrosh[10], Allison
1043 Fong[14], Brian Foster[10], Bryce Foster[10], Michael Ginzburg[14], Marcel Huntemann[10],
1044 Natalia N. Ivanova[10], Nikos C. Kyrpides[10], Kara Martin[12,15], Vincent Moulton[13], Supratim
1045 Mukherjee[10], Krishnaveni Palaniappan[10], T.B.K. Reddy[10], Simon Roux[10], Katrin
1046 Schmidt[16], Jan Strauss[17], Klaas Timmermans[12], Susannah G. Tringe[10], Graham J.C.
1047 Underwood[18], Klaus U. Valentin[14], Willem H. van de Poll[19], Neha Varghese[10]
1048
1049 [9]Department of Chemistry, Norwegian University of Science and Technology, Trondheim, 7491,
1050 Norway. [10] DOE-Joint Genome Institute, 1 Cyclotron Road, Berkeley, CA 94720, U.S.A.
1051 [11]Department of Biology, University of Duisburg-Essen, Essen, Universitaetsstrasse 2, 45141 Essen,
1052 Germany. [12]Royal Netherlands Institute for Sea Research, Texel, The Netherlands, [13]School of
1053 Computing Sciences, University of East Anglia, Norwich Research Park, NR47TJ, Norwich, U.K.
1054 [14]Alfred-Wegener Institute for Polar and Marine Research, Am Handelshafen 12, 27570
1055 Bremerhaven, Germany. [15]Earlham Institute, Norwich Research Park, Norwich, NR4 7UG, United
1056 Kingdom. [16]School of Environmental Sciences, University of East Anglia, Norwich Research Park,
1057 NR47TJ, Norwich, U.K. [17]Ocean EcoSystems Biology Unit, GEOMAR Helmholtz Centre for Ocean
1058 Research Kiel, Düsternbrooker Weg 20, D-24105 Kiel, Germany. [18]School of Life Sciences,
1059 University of Essex, CO4 3SQ, Colchester, U.K. [19]Department of Ocean Ecosystems, Energy and
1060 Sustainability Research Institute Groningen, University of Groningen, Nijenborgh 7, 9747 AG
1061 Groningen, The Netherlands
1062
1063

1064