Genomes of nitrogen-fixing eukaryotes reveal a non-canonical model of organellogenesis

Sarah Frail[1, 6], Melissa Steele-Ogus[2, 6], Jon Doenier[1], Solène L.Y. Moulin[2], Tom Braukmann[1,2], Shouling Xu[3], Ellen Yeh[2,4,5,7,8,*]

[1]Department of Biochemistry, Stanford School of Medicine, Stanford, CA 94305, USA
[2]Department of Pathology, Stanford School of Medicine, Stanford, CA 94305, USA
[3]Department of Plant Biology, Carnegie Institution, Stanford, CA 94305, USA
[4]Department of Microbiology & Immunology, Stanford School of Medicine, Stanford, CA 94305, USA
[5]Chan Zuckerberg Biohub – San Francisco, San Francisco, California 94158, USA
[6]These authors contributed equally
[7]Lead contact
[8]Senior author
*Correspondence: ellenyeh@stanford.edu

SUMMARY

Endosymbiotic gene transfer and import of host-encoded proteins are considered hallmarks of organelles necessary for stable integration of two cells. However, newer endosymbiotic models have challenged the origin and timing of such genetic integration during organellogenesis. *Epithemia* diatoms contain diazoplasts, obligate endosymbionts that are closely related to recently-described nitrogen-fixing organelles and share similar function as integral cell compartments. We report genomic analyses of two species which are highly divergent but share a common ancestor at the origin of the endosymbiosis. We found minimal evidence of genetic integration in *E.clementina*: nonfunctional diazoplast-to-nucleus DNA transfers and 6 host-encoded proteins of unknown function in the diazoplast proteome, far fewer than detected in other recently-acquired endosymbionts designated organelles. *Epithemia* diazoplasts are a valuable counterpoint to existing organellogenesis models, demonstrating that endosymbionts can function as integral compartments absent significant genetic integration. The minimal genetic integration makes diazoplasts valuable blueprints for bioengineering endosymbiotic compartments *de novo.*

KEYWORDS

Cyanobacteria, diatom, endosymbiosis, evolution, genome, genomics, horizontal gene transfer, nitrogen fixation, organelle, organellogenesis, symbiosis

INTRODUCTION

Endosymbiotic organelles are uniquely eukaryotic innovations for the acquisition of complex cellular functions, including aerobic respiration, photosynthesis, and nitrogen fixation. Endosymbioses contributed to expansive eukaryotic diversity[1]. An important question in cell evolution and engineering is: how do intermittent, facultative endosymbioses evolve into permanent integral cell compartments, i.e. organelles? With the recognition of the bacterial origin of mitochondria and chloroplasts in the 1980s, Cavalier-Smith & Lee proposed that the key distinction between a transient endosymbiont and an organelle was that organelles do not synthesize all their own proteins[2,3]. Instead, in organelles, some genes are transferred from the endosymbiont genome to the eukaryotic nucleus in a process called endosymbiotic gene transfer (EGT). These and other gene products, now under the control of host gene expression, are imported back into the endosymbiotic compartment to regulate endosymbiont growth and division. This definition has since been commonly applied[4,5]. However, the underlying hypothesis for organelle evolution — that genetic integration resulting from EGT and/or import of host-encoded gene products is essential for maintaining the endosymbiont as an integral cellular compartment — has not been rigorously tested.

In the decades since, increased sampling of eukaryotic diversity has uncovered evidence that, amongst microbes, endosymbioses are a common strategy for acquisition of new functions. Based on observations of EGT and host protein import, new organelles have been recognized: the chromatophore in *Paulinella chromatophora*[6–9] and UCYN-A in *Braarudosphaera bigelowii*[10]. EGT and host protein import have also been observed in obligate, vertically-inherited nutritional endosymbionts of the parasite *Angomonas deanei* and insects, which are not formally recognized as organelles[11,12]. With the benefit of these newer models, our understanding of genetic integration has become more nuanced[13,14]. For

example, the majority of host proteins imported into the *Paulinella* chromatophore do not originate from EGT but rather horizontal gene transfer (HGT) from other bacteria or eukaryotic genes[15], showing that a host's repertoire of pre-existing genes may play an outsized role in facilitating genetic integration[16]. UCYN-A was initially described as having an unstable relationship with its host as it is often lost from host cells during isolation and in culture, suggesting environmental conditions can affect endosymbiont stability despite genetic integration[17]. There have been bigger surprises: Some organisms temporarily acquire plastids from partially-digested prey algae[18,19]. The retained chloroplasts, called kleptoplasts, perform photosynthesis and, in several species, depend on imported host proteins to fill gaps in their metabolic pathways. Despite their genetic integration, these kleptoplasts cannot replicate in the host cell and are not required for host cell survival, indicating that genetic integration is not sufficient to achieve stable integration of the endosymbiont[20–22]. These findings highlight the importance of studying biodiverse organisms to inform new hypotheses for endosymbiotic evolution.

Amongst new model systems, *Epithemia* spp. diatoms offer a unique perspective on organellogenesis. These photosynthetic microalgae contain diazotroph endosymbionts (designated diazoplasts) that perform nitrogen fixation, a biological reaction that converts inert atmospheric nitrogen to bioavailable ammonia[23–28]. The ability to fix both carbon and nitrogen fulfills a unique niche in ecosystems. Numerous *Epithemia* species are globally widespread in freshwater habitats and have recently been isolated from marine environments[29–31]. The *Epithemia* endosymbiosis is very young relative to mitochondria and chloroplasts, having originated ~35 Mya, based on fossil records[32]. Nonetheless, diazoplasts are obligate endosymbionts which are coordinately inherited during host cell division and present in all *Epithemia* species described so far, indicating co-evolution of diazoplasts and their host algae. Finally, *Epithemia* diazoplasts are closely related to UCYN-A, the diazotroph endosymbiont of *B. bigelowii* which was recently designated the first nitrogen-fixing organelle, or nitroplast[10,28,33]. Both *Epithemia* diazoplasts and UCYN-A evolved from free-living *Crocosphaera* cyanobacteria that have engaged in endosymbioses with several host microalgae. The independent evolution of free-living *Crocosphaera* into diazotroph endosymbionts in multiple host lineages enables comparisons that can lead to powerful insights.

If the significance of organelles lies in their function as integral cellular compartments, then metabolic and cellular integration with the host cell are paramount[34]. By these criteria, diazoplasts show a level of host-symbiont integration comparable to UCYN-A. Nitrogen fixation requires large amounts of ATP and reducing power, energy that can be supplied by photosynthesis. Yet nitrogenase, the enzyme that catalyzes nitrogen fixation, is exquisitely sensitive to oxygen produced during oxygenic photosynthesis. In free-living *Crocosphaera*, photosynthesis and nitrogen fixation is temporally separated such that fixed carbon from daytime photosynthesis is stored as glycogen to fuel exclusively nighttime nitrogen fixation. Diazoplasts have lost all photosystem genes and depend entirely on host photosynthesis for fixed carbon[27,35]. Recently, we showed that host and diazoplast metabolism are tightly coupled to support continuous nitrogenase activity throughout the day-night cycle in *E. clementina*: Diatom photosynthesis is required for daytime nitrogenase activity in the diazoplast, while nighttime nitrogenase activity also depends on diatom, rather than diazoplast, carbon stores[28]. In comparison, UCYN-A has lost only photosystem II and is dependent on both host photosynthesis and, likely, its own photosystem I, restricting it to daytime nitrogen fixation[36,37]. *Epithemia* spp. described so far typically contain 1-2 diazoplasts per cell that are vertically inherited during asexual cell division[25,28,30]. Diazoplasts have further been shown to be uniparentally inherited during sexual reproduction, similar to mitochondria and chloroplasts[38]. Coordinated replication of UCYN-A with host cell division has been observed to maintain a single endosymbiont per cell[10]. Similar mechanisms are likely in place to coordinate diazoplast inheritance with diatom division. In fact, the presence of diazoplasts in diverse *Epithemia* species globally widespread in freshwater and marine ecosystems demonstrates that the mechanisms of inheritance are robust through speciation events. Diazoplasts effectively serve as dedicated nitrogen-fixing compartments in *Epithemia*, whether or not genetic integration has occurred.

An important question emerges from these observations: is EGT and/or host protein import required to achieve the level of host-symbiont integration observed between diazoplasts and host *Epithemia*? Based on the similarity of diazoplasts to UCYN-A, the assumption is yes. However, there is evidence that metabolite exchange via endosymbiont-encoded transporters[39] and division coordinated by host proteins outside the endosymbiotic compartment[11,40,41] could form a stable compartment without genetic integration. We previously established freshwater *E. clementina* as a laboratory model for functional studies and herein performed *de novo* assembly and annotation of its genome. The genome

sequence for *E. pelagica*, a recently-discovered marine species, was publicly released by the Wellcome Sanger Institute[30,42]. To facilitate comparison between these species, we also performed *de novo* genome annotation of *E. pelagica*. Notably, no genomes of *B. bigelowii* (which hosts UCYN-A) nor the eukaryotic host in any other diazotroph endosymbiosis have been available. We report genome and transcriptome analyses of these two *Epithemia* species as well as proteome analyses of *E. clementina* with the goals of 1) providing a necessary resource to accelerate investigation of this model and 2) elucidating the role of genetic integration in this very young, stably integrated endosymbiont.

RESULTS

The genome of *E. clementina* is larger and more repetitive than *E. pelagica*

Epithemia spp. are raphid, pennate diatoms composed of at least 50 freshwater species and 2 reported marine species[29,30]. Isolation and characterization of freshwater *E. clementina* was previously reported[28] (Figure S1A). We isolated high molecular weight DNA from axenic *E. clementina* cultures,performed sequencing by long-read Nanopore and short-read Illumina and assembled a 418 Mbp haploid assembly with a high level of heterozygosity of 1.48% (Figure S1B). The final reported haploid assembly is complete, contiguous, and of high sequence quality (Figure S1C, Table 1). A chromosome-level 60 Mbp genome assembly (GCA_946965045) of *E. pelagica*, a marine species, was reported by the Sanger Institute[42]. Whole-genome alignments of *E. clementina* and *E. pelagica* did not show significant syntenic blocks in their nuclear genomes (Figure S1E). In contrast, their diazoplast genomes showed 5 major and 2 minor syntenic blocks (Figure S1F), similar to the synteny reported between diazoplast genomes of other *Epithemia* species[33,35].

Both *E. clementina* and *E. pelagica* genomes were annotated using evidence from protein orthology and transcriptome profiling. The nuclear genomes were predicted to contain 20,203 genes in *E. pelagica* and 26,453 genes in *E. clementina* (Figure 1A). The completeness of their predicted proteomes was assessed based on the presence of known single-copy orthologs in stramenopiles, yielding $BUSCO_{protein}$ scores of 99% for *E. clementina* and 94% for *E. pelagica* (Table 1, Figure S1D). While the gene numbers between *E. clementina* and *E. pelagica* are similar and typical for diatoms, the genome of *E. clementina* is 7 times larger (Figures 1A and 1B, Table 1). The increased genome size is due to a substantial repeat expansion unique to *E. clementina* (Figure 1B and 1C). Notably, the differences in genome size observed amongst diatoms is largely due to repeat content (Figure 1B). Multiple LTR families and DNA transposons show expansions that contribute to the high repeat percent in *E. clementina* (Figure 1C, Table S2). This cross-family expansion may indicate a history of relaxed selection on the repeatome of *E. clementina*, perhaps related to a species bottleneck at the transition to freshwater habitats.

Highly divergent *E. clementina and E. pelagica* genomes share many unique gene families

As a measure of divergence at a functional level, we compared the amino acid identity between orthologs across proteomes of several pairs of representative diatom and metazoan species (Figure 1D). Despite their estimated 35 Mya of speciation, *E. pelagica* and *E. clementina* showed a similar distribution of identity across protein orthologs as humans and pufferfish (*Homo sapiens* and *Takifugu rubripes*), which are estimated to have shared a common ancestor 429 Mya[43]. This rapid divergence, relative to age, is also observed in other diatom species, for example, *T. pseudonada*/ *T. oceanica* (70 Mya) and *P. multistriata*/ *P. multiseries* (6.3 Mya)[43] (Figure 1D and Figure S2A). The loss of synteny, substantial differential repeat expansion, and low protein ortholog identity suggest that *E. pelagica* and *E. clementina* have diverged substantially during speciation, reflecting the rapid evolution rates of diatoms[44,45] (Figure 1D and Figure S2A).

Because rapid divergence is common across diatoms, we evaluated the gene content of *E. pelagica* and *E. clementina* in comparison with other diatom species with complete genomes available. Gene families, defined by orthogroups, were identified for each species. We then quantified the number of uniquely shared gene families between subsets of diatom species, i.e. gene families shared between that group of species and not found in any other diatoms. Of 10,740 and 10,612 gene families identified in *E. clementina* and *E. pelagica* respectively, they share 8,942 gene families, a greater overlap than is observed between any other pair of diatom species (Figures S2A). Of these, 1,109 gene families are uniquely shared between *E. clementina* and *E. pelagica*, more than any other species grouping including the more recently speciated *Pseudo-nitzchia* species (Figure 1E). This *Epithemia*-specific gene set is

169 significantly enriched for functions relating to carbohydrate transport and membrane biogenesis, which
170 may have been important for adaptation to the endosymbiont (Figure S2B).
171 Because HGT is known to be a source of genes for endosymbiont functions expressed by the
172 host[9] and 3-5% of diatom proteomes have been attributed to bacterial HGT[46], a significantly greater
173 proportion than detected in other eukaryotic proteomes[47], we identified candidate HGTs in the *Epithemia*
174 genomes. A total of 118 and 97 candidate HGTs were identified in *E. clementina* and *E. pelagica*
175 respectively, of which 51 are only detected in *Epithemia* (Figures S2C, S2D and S2E, Table S3). Notably,
176 the two *Epithemia* species share a greater number of predicted horizontally-acquired genes than other
177 diatom species and more than expected by gene family overlap alone. However, we were unable to
178 identify enriched functions in this relatively small gene set that were informative. Overall, the uniquely
179 shared features of the divergent nuclear genomes of *Epithemia* genus are valuable for identifying
180 potential signatures of endosymbiotic evolution.
181
182 **Diazoplast-to-nucleus transfer of DNA is actively occurring in *E. clementina***
183 Having broadly compared the *Epithemia* genomes for shared features, we turned to specifically
184 interrogate genetic integration between *Epithemia* and their diazoplasts. We specifically distinguished
185 EGT, which we defined as the transfer of functional genes from the endosymbiont to the host nucleus,
186 from endosymbiont-to-nucleus transfers of DNA, which is believed to be more frequent and often
187 nonfunctional. Indeed, it has been shown that nuclear integrations of organellar DNA originating from
188 mitochondria (designated NUMT) and plastids (NUPT) still occur[48,49]. Given the significantly younger age
189 of the diazoplast, it is not clear whether nuclear integrations of diazoplast DNA (which we will refer to as
190 NUDT) and/or functional transfers of genes (EGT) have occurred. To identify transfers of endosymbiont
191 DNA to the host nucleus, we performed homology searches against the nuclear assemblies of *E.*
192 *clementina* and *E. pelagica.* As queries, we used the diazoplast genomes of 4 *Epithemia* species
193 (including *E. clementina* and *E. pelagica*) and the 5 most closely-related free-living cyanobacteria species
194 for which whole genomes were available (Table S1). To prevent spurious identifications, alignments were
195 excluded if they were <500 contiguous base pairs in length. In *E. clementina*, we identified seven
196 segments, ranging from 1700-6400bp, with homology to the *E. clementina* diazoplast (Figures 2A and 2B,
197 Data S1A-S1G). No homology to free-living cyanobacteria genomes was detected. The *E. coli* genome
198 and a reversed sequence of the *E. pelagica* diazoplast were used as negative control queries and yielded
199 no alignments. Finally, no regions of homology to any of the queries were detected in the nuclear genome
200 of *E. pelagica*.
201 NUDTs showed features suggesting they were distinct from diazoplast genomic sequences and
202 unlikely to be assembly errors. First, 4 of the 7 NUDTs were supported by long reads equivalent to 1x
203 coverage of the genome indicating the insertions were homozygous. 3 NUDTs contained on ctg003410,
204 ctg001640, ctg005680 showed the equivalent of 0.5x genome coverage, consistent with a heterozygous
205 insertion in the diploid eukaryotic genome (Figure 2E, Data S1A, S1B, and S1D). Second, NUDTs had
206 low GC content similar to that of the diazoplast but contain many single nucleotide variants (SNVs) with
207 mean identity of 98.4% to their source sequences, indicative of either neutral or relaxed selection (Figures
208 2F and 3B). Finally, each NUDT was composed of multiple fragments corresponding to distal regions in
209 the endosymbiont genome, ranging from as few as 8 distal fragments composing the NUDT on
210 ctg002090 to as many as 42 on ctg003780 (Figures 2C, 2D, and S3A-S3E). This composition of NUDTs
211 indicates either that fragmentation and rearrangement of the diazoplast genome occurred prior to
212 insertion into the eukaryotic genome or that NUDTs were initially large insertions that then underwent
213 deletion and recombination. Overall, the detection of NUDTs shows that diazoplast-to-nucleus DNA
214 transfer is occurring in this very young endosymbiosis.
215
216 **Most NUDTs are decaying and non-functional**
217 To determine whether any of the identified NUDTs have resulted in EGT, we identified diazoplast
218 genes present in NUDTs and evaluated their potential for function. A total of 124 diazoplast genes and
219 gene fragments were carried over into the NUDTs (Figure 3A). (A few of these diazoplast genes have
220 conserved eukaryotic homologs and were also predicted as eukaryotic genes in the *E. clementina*
221 genome annotation (Data S1A-S1G).) 121 diazoplast genes detected in NUDTs are truncated >30%
222 compared to the full-length diazoplast gene (Figure 3A). Of the three remaining, two genes contained on
223 ctg002090 showed accumulation of SNVs that resulted in a premature stop codon and a nonstop
224 mutation (Figures 3B, S3F and S3G). We performed transcriptomics to assess the expression from

NUDTs. Neither of the two genes on ctg002090 showed appreciable expression. All except one NUDT showed <2 bins per million mapped reads (BPM), equivalent to background transcription levels within the region (Figures 3C and Data S1A-S1G). The truncation, mutation accumulation, and lack of appreciable expression of diazoplast genes encoded in NUDTs suggest that most are nonfunctional.

Only a single EGT candidate was detected contained on ctg005680: an intact sulfotransferase (*tusA*) gene that is 100% identical to the diazoplast-encoded gene (Figures 3A and 3B, Data S1A). The NUDT that contains this candidate appears to be very recent as it is heterozygous and shows 99.7% identity to the source diazoplast sequence (Figure 3B). Interestingly, *tusA* is implicated in Fe-S cluster regulation that could be relevant for nitrogenase function. Due to the high sequence identity, it is not possible to distinguish transferred *tusA* from that of *tusA* encoded in the diazoplast genome by sequence alone. However, transcript abundance above background levels was only detected in rRNA depleted samples that contain diazoplast transcripts and not in polyA-selected samples that remove diazoplast transcripts, indicating that the observed expression is largely attributed to diazoplast-encoded *tusA* (Figure 3C). Moreover, host proteins imported to endosymbiotic compartments often use *N*-terminal (occasionally *C*-terminal) targeting sequences[10,50,51]. We were unable to identify any added sequences in the transferred *tusA* indicative of a targeting sequence; the sequence immediately surrounding consisted only of native diazoplast sequence carried over with the larger fragment (Data S1A). Though there is no evidence for gene function, the transfer of this intact gene indicates that the conditions for EGT are present in *E. clementina*[4].

Few host-encoded proteins are detected in the diazoplast proteome

The critical step in achieving genetic integration is evolution of pathways for importing host proteins into the endosymbiont. While EGT and HGT from other bacteria can expand the host's genetic repertoire, neither transferred genes nor native eukaryotic genes can substitute for or regulate endosymbiont functions unless the gene products are targeted to the endosymbiotic compartment. Abundant host-encoded proteins were detected in the proteomes of recently-acquired endosymbionts that have been designated organelles: 450 in the chromatophore of *Paulinella*[9] and 368 in UCYN-A[10]. In both the chromatophore and UCYN-A, several host-encoded proteins detected in the endosymbiont fulfill missing functions that complete endosymbiont metabolic pathways, providing further support for the import of host-encoded proteins.

To determine whether host protein import is occurring in the diazoplast, we identified the proteome of the *E. clementina* diazoplast. We were unable to maintain long-term *E. pelagica* cultures to perform proteomics for comparison. Diazoplasts were isolated from *E. clementina* cells by density gradient centrifugation (Figure 4A). The purity of isolated diazoplasts was evaluated by light microscopy. The protein content of isolated diazoplasts and whole *E. clementina* cells containing diazoplasts were determined by LC-MS/MS. A total of 2481 proteins were identified with ≥2 unique peptides: 754 proteins were encoded by the diazoplast genome (detected/total protein coding = 43% coverage) and 1727 proteins encoded by the nuclear genome (6.5% coverage) (Figure 4B, Table S4). Of note, TusA, the only EGT candidate identified, was not detected in either proteome. To identify proteins enriched in either the diazoplast or host compartments, we compared protein abundance in isolated diazoplasts and whole cell samples across 3 biological replicates (Figure 4C). 492 diazoplast-encoded proteins were significantly enriched in the diazoplast and none were enriched in whole cell samples, supporting the purity of the isolated diazoplast sample. Similarly, most host-encoded proteins (1281) were significantly enriched in whole cell samples, indicating localization in host compartments. Six unique host-encoded proteins were significantly enriched in diazoplast samples, suggesting possible localization to the diazoplast. Five were encoded by Ec_g00815, Ec_g12982, Ec_g13000, Ec_g13118, and Ec_g25610. The sixth protein was encoded by two identical genes, Ec_g24166 and Ec_g03819, resulting from an apparent short duplication of it and two neighboring genes. Because the duplication makes Ec_g24166 and Ec_g03819 indistinguishable by amino acid sequence, we considered them one import candidate. Of the 6 host protein import candidates, Ec_g25610 and Ec_g13000 were detected only in the diazoplast sample, while the rest were identified in both diazoplast and whole cell samples. We are unable to rule out the possibility of nonspecific enrichment since neither genetics nor immunofluorescence are available in *E. clementina* to further validate their protein localization.

Instead we sought additional evidence to support the import of these host proteins by evaluating their potential functions[52]. No domains, GO terms, or BLAST hits (other than to hypothetical proteins found in other diatoms) were identified for any of the candidates except for Ec_g13118 which is annotated

281 as an E3 ubiquitin ligase. In contrast to the unclear functions of these candidates, several host proteins
282 detected in the chromatophore and UCYN-A proteome were assigned to conserved cyanobacterial
283 growth, division, or metabolic pathways in these organelles. Moreover, none of the candidates for import
284 into the diazoplast have apparent homology to proteins encoded in diazoplast or free-living *Crocosphaera*
285 genomes to suggest they might fulfill unidentified cyanobacterial functions. Instead, all candidates are
286 diatom-specific proteins: 3 candidates (Ec_g24166/Ec_g03819, Ec_g12982, and Ec_g13000) belonged to
287 orthogroup OG0000250 which is uniquely shared with *E. pelagica* but no other diatoms (Figure 1E). The
288 remaining 3 belonged to separate orthogroups (OG0001966, OG0004498, and OG0009247) which are
289 shared broadly among diatoms including *E. pelagica*. Alignments of these candidates with their homologs
290 did not show *N*- or *C*-terminal extensions consistent with possible targeting sequences for endosymbiont
291 localization. Even if targeted to the diazoplast, we hypothesize that conserved eukaryotic proteins,
292 especially diatom-specific proteins, are less likely to substitute for endosymbiont functions than bacterial
293 proteins. Our functional analysis suggests these import candidates are unlikely to have critical functions in
294 cyanobacterial pathways.
295 Finally, the detection of ~100-fold fewer import candidates in the diazoplast indicate that host
296 protein import, if occurring, is far less extensive than in the chromatophore or UCYN-A. Since the
297 sensitivity of proteomics is highly dependent on biomass, we estimated the coverage of the diazoplast
298 proteome based on the ratio of diazoplast-encoded proteins detected (754) compared to the total
299 diazoplast protein-coding genes (1585). The coverage of the diazoplast proteome (48%) was comparable
300 to the coverage of the published chromatophore proteome (422/867= 49%) and that of UCYN-A
301 (609/1186= 51%) and therefore does not account for the low number of import candidates[9,10]. Overall, the
302 number of host proteins detected in the diazoplast was significantly fewer and their functional significance
303 unclear compared to host proteins detected in the chromatophore and UCYN-A.
304
305 **DISCUSSION**
306 EGT and host protein import have been held as a necessary to achieve the "permanent"
307 integration of organelles[2,5]. But this hypothesis for organelle evolution has been challenged by findings in
308 young endosymbionts from diverse organisms[13,14,34]. We report analysis of two genomes of *Epithemia*
309 diatoms and evaluate the extent of their genetic integration with their nitrogen-fixing endosymbionts
310 (diazoplasts), thereby adding this very young endosymbiosis to existing model systems that can elucidate
311 the integration of two cells into one.
312 *A window into the early dynamics of nuclear gene transfers.* Our first significant finding was the
313 detection of active diazoplast-to-nucleus DNA transfers but, as yet, no functional EGT in *Epithemia*. Our
314 observations support findings in the chromatophore and UCYN-A that EGT is not necessary for genetic
315 integration[10,15,53]. Given that EGT does not necessarily precede evolution of host protein import pathways,
316 it may be a suboptimal solution for the inevitable genome decay in small asexual endosymbiont
317 populations as a consequence of Muller's ratchet[54,55]. Instead, the decayed nature of the NUDTs we
318 detected in *E. clementina* is consistent with stochastic, transient, ongoing DNA transfer. Nonfunctional
319 DNA transfers were previously only described from mitochondria or plastids with far more reduced
320 genomes. The status of nuclear transfers from more recently-acquired organelles is unknown, as only
321 protein-coding regions were used as queries to identify chromatophore transfers in *Paulinella* and only a
322 transcriptome is available for the UCYN-A host, *B. bigelowii*[10,15,56]. NUDTs in *Epithemia* genomes
323 therefore provide a rare window into the early dynamics of DNA transfer. For example, using the same
324 homology criteria, we identified 5 NUMTs but no NUPTs in *E. clementina*. The NUMTs were significantly
325 shorter than NUDTs and did not show rearrangement, which may suggest different mechanisms of
326 transfer for NUDTs, NUMTs, and NUPTs in the same host nucleus. In addition, between-species
327 differences may identify factors that affect transfer rates. The lack of observed NUDTs in *E. pelagica*
328 suggest constraints on diazoplast-to-nucleus transfers. Previous observations in plant chloroplasts
329 supported the limited transfer window hypothesis, which proposes the mechanism of gene transfer
330 requires endosymbiont lysis and therefore the frequency of gene transfers correlates with the number of
331 endosymbionts per cell[57–59]. However since *E. pelagica* and *E. clementina* contain similar numbers of
332 diazoplasts per cell (1-2)[28,30], the limited transfer window hypothesis does not explain the observed
333 differences. Instead, there may be additional constraints imposed such as lower tolerance to DNA
334 insertions in the comparatively smaller, non-repetitive genome of *E. pelagica*. Finally, the lack of NUDT
335 gene expression, even with transfer of a full-length unmutated *tusA* gene, points to barriers to achieving
336 eukaryotic expression from bacterial gene sequence. *Epithemia* is an apt model system to interrogate

how horizontal gene transfer impacts eukaryotic genome evolution with at least 20 species easily obtained from freshwater globally and consistently adaptable to laboratory cultures[28–30,33,35].

Epithemia diazoplasts as a counterpoint to existing models of organellogenesis. A second unexpected finding was the detection of only 6 host protein candidates in the diazoplast proteome, much fewer and with less clear functional significance than in comparable endosymbionts that have been designated organelles. Methods for validating the localization of these import candidates are unavailable in *Epithemia.* Even if confirmed to target to the diazoplast, the candidates lack conserved domains or homology with cyanobacterial proteins to indicate they replace or supplement diazoplast metabolic function, growth, or division. Our findings are not explained by current models of organellogenesis that propose import of host proteins as a necessary step to establish an integral endosymbiotic compartment. In the traditional model described in the introduction, host protein targeting is a “late” bottleneck step required for the regulation of the endosymbiont growth and division. More recently, “targeting-early” has been proposed to account for establishment of protein import pathways prior to cellular integration as observed in kleptoplasts[19,20]. In this model, protein import is selected over successive transient endosymbioses, possibly driven by the host’s need to export metabolites from the endosymbiont via transporters or related mechanisms[60]. The establishment of protein import pathways then facilitates endosymbiont gene loss with metabolic functions fulfilled by host proteins leading to endosymbiont fixation. Contradicting both models, we observed minimal evidence for genetic integration despite millions of years of co-evolution resulting in diverse *Epithemia* species retaining diazoplasts, indicating that genetic integration is not necessary for its stable maintenance. At a minimum, the unclear functions of the few host proteins identified in the diazoplast proteome, if imported, suggest that the genesis of host protein import in *Epithemia* is very different than would be predicted by current models.

Diazotroph endosymbioses are fundamentally different from photosynthetic endosymbioses that are the basis for current organellogenesis models. First, the diazoplast is derived from a cyanobacterium that became heterotrophic by way of losing its photosynthetic apparatus. Regulation of endosymbiont growth and division by the availability of host sugars (without requiring an additional layer of regulation via import of host metabolic enzymes) may be more facile with heterotrophic endosymbionts maintained for a nonphotosynthetic function compared to autotrophic endosymbionts. It will be interesting to see how integration of the diazoplast differs from the endosymbiont of *Climacodium freunfeldianum*, another diazotrophic endosymbiont descended from *Crocosphaera* that likely retains photosynthesis[61,62]. Second, ammonia, the host-beneficial metabolite in diazotroph endosymbioses, can diffuse through membranes in its neutral form and does not require host transporters for efficient trafficking[63]. Previously, we observed efficient distribution of fixed nitrogen from diazoplasts into host compartments following $^{15}N_2$ labeling[28]. Ammonia diffusion may have reduced early selection pressure for host protein import as posited by the targeting-early model. Finally, the eukaryotic hosts in most diazotroph endosymbioses are already photosynthetic, in contrast to largely heterotrophic hosts that acquired photosynthesis by endosymbiosis. For instance, cellular processes that enabled intracellular bacteria to take up residence in the ancestor of *Epithemia* spp. were likely different than those of the bacterivore amoeba ancestor of *Paulinella chromatophora*. Autotrophy and lack of digestive pathways would reduce the frequency by which bacteria might gain access to the host cell, such that the selection of host protein import pathways over successive transient interactions would be less effective. Overall, a universal model of organellogenesis is premature given the limited types of interaction that have been investigated in depth, highlighting instead the importance of increasing the diversity of systems studied.

Are diazoplasts “organelles”? As detailed in the introduction, diazoplasts show metabolic and cellular integration with their host alga comparable to that of UCYN-A, the first documented nitrogen-fixing organelle[10,17]. However, while hundreds of host proteins were detected in the UCYN-A proteome, including many likely to fill gaps in its metabolic pathways, a handful of host proteins with unknown function were detected in the diazoplast proteome. Based on the conventional definition which specifies genetic integration as the dividing line between endosymbionts and organelles, diazoplasts would not qualify[2,5]. However, over a decade ago, Keeling and Archibald[34] suggested that “if we use genetic integration as the defining feature of an organelle, we will never be able to compare different routes to organellogenesis because we have artificially predefined a single route.” They further hypothesized that if an endosymbiont became fixed in its host absent genetic integration, “it might prove to be even more interesting… by focusing on how it did integrate, perhaps we will find a truly parallel pathway for the integration of two cells.” The diazoplast appears to be such a parallel case in which non-genetic interactions were sufficient to integrate two cells.

If not gene transfer and host protein import, then what is the "glue" that holds this endosymbiosis together? The loss of cyanobacterial photosystems and dependence on host photosynthesis indicate that diazoplasts acquired new transporters for host sugars[28]. There are several examples of cyanobacteria that express genes for sugar transport[39,64,65]. Therefore, acquisition via horizontal gene transfer from another bacteria to the diazoplast ancestor, prior to the endosymbiosis, rather than targeting of a eukaryotic transporter post-endosymbiosis, seems more likely. Consistent with this hypothesis, potential transporters were not detected amongst host protein import candidates detected in UCYN-A[10]. In fact, mixotrophy may have facilitated the adoption of an endosymbiotic lifestyle by the free-living cyanobacterium. Regardless of the timing of acquisition, sugar transport function could allow diazoplast growth to be regulated by nutrient availability from the host. Similarly, cytosolic host proteins may coordinate diazoplast division from outside the endosymbiotic compartment[11,41]. Eukaryotic dynamins required for mitochondria and chloroplast fission localize to the surface of the organellar outer membrane, acting coordinately with bacterial fission factors located in the organelle[40,66]. Diazoplasts appear to be surrounded by a host-derived membrane[26,67] (which may be lost during diazoplast purification) (Figure 4A). In analogy to dynamins, host proteins localized to this outer membrane may mediate diazoplast fission without requiring protein import pathways. Finally, cell density[68] and mechanical confinement[69] have been demonstrated to limit the growth of cyanobacteria, suggesting that host regulation of the volume of the endosymbiotic compartment could also be an effective mechanism. The mechanisms for the robust metabolic exchange and coordinated division observed in diazoplasts will be the focus of future studies.

Application of cell evolution models to bioengineering. Diazoplasts provide another example that the current organelle definition does not account for observations in many biological systems and may be overdue for revision to reflect biological significance in the spectrum of endosymbiotic interactions. At a minimum, it is time to disentangle the current definition of an organelle from models that elucidate the formation of integral cellular compartments. Identifying mechanisms to integrate cells is more than an academic exercise. The ability to engineer bacteria as membrane compartments to introduce new metabolic functions in eukaryotes would be transformative[70,71]. For example, nitrogen-fixing crop plants that could replace nitrogen fertilizers is a major goal for sustainable agriculture. Efforts to transfer the genes for nitrogen fixation to plant cells have been slow, hampered by the many genes required as well as the complex assembly, high energy requirements, and oxygen sensitivity of the reaction. We previously proposed an alternative strategy inspired by diazotroph endosymbioses: introducing nitrogen-fixing bacteria into plant cells as an integral organelle-like compartment[28]. This approach has the advantage that diazotrophs express all required genes with intact regulation, coupled to respiration, and in a protected compartment. Diazoplasts, which achieve stable integration without significant genetic integration, are an important alternative to UCYN-A and other organelles, which are defined by their genetic integration, to inform this strategy. Identifying the nongenetic interactions that facilitated diazoplast integration with *Epithemia* will be critical for guiding bioengineering.

Ongoing genome reduction may drive genetic integration in diazotroph endosymbioses. The fewer number of host protein candidates and their lack of clear function in diazoplasts versus UCYN-A is not associated with differences in their function as nitrogen-fixing cellular compartments. Rather, an alternative explanation points towards differences in the extent of genome reduction in diazoplasts, which encode 1585-1848 protein-coding genes, compared to UCYN-A, which encode 1200-1246 protein-coding genes[72]. Among the genes missing from the UCYN-A genome are cyanobacterial IspD, ThrC, PGLS, and PyrE; for each, an imported host protein was identified that could substitute for the missing function[10]. In contrast, these genes are retained in diazoplast genomes, including those of *E. clementina* and *E. pelagica*, obviating the need for host proteins to fulfill their functions (Figure S4). Consistent with diazoplasts and UCYN-A being at different stages of genome reduction, diazoplast genomes contain >150 pseudogenes compared to 57 detected in the UCYN-A genome, suggesting diazoplasts are in a more active stage of genome reduction[27,33,35]. Interestingly, even genes retained in UCYN-A, namely PyrC and HemE, have imported host-encoded counterparts[10]. The endosymbiont copies may have acquired mutations resulting in reduced function, necessitating import of host proteins to compensate. Alternatively, once efficient host protein import pathways were established, import of redundant host proteins may render endosymbiont genes obsolete, further accelerating genome reduction. Genetic integration may in fact be destabilizing for an otherwise stably integrated endosymbiont, at least initially, as it substitutes essential endosymbiont genes with host-encoded proteins that may not be functionally equivalent and require energy-dependent import pathways. Comparing these related but independent

diazotroph endosymbioses yields valuable insight, which otherwise would not be apparent. Diazoplasts at 35 Mya may represent an earlier stage of the same evolutionary path as ~140 Mya UCYN-A, in which continued genome reduction will eventually select for protein import pathways. Alternatively, diazoplasts may have evolved unique solutions to combat destabilizing genome decay, for example through the early loss of mobile elements.[27,33,73] Whether they represent an early intermediate destined for genetic integration or an alternative path, diazoplasts provide a valuable new perspective on endosymbiotic evolution.

Limitations of the study

- Accurate gene family analysis is dependent on species sampling. While we sought to sample species representative of diatom diversity, some of our reported *Epithemia*-specific gene families may be shared by non-*Epithemia* species not present in the data set.
- While we included free-living *Crocosphaera* relatives for our homology search for NUDTs, we cannot eliminate the possibility that there are NUDTs or EGTs derived from sequences that were once in diazoplast genomes but have been lost.
- We did not observe expression associated with genes contained in NUDTs, however, we cannot exclude that they may be expressed under untested conditions.
- While differential centrifugation was an effective means of diazoplast enrichment, there may still be contamination by other cellular compartments.
- Low-abundance proteins may fall below the detection threshold in label-free proteomics. While our proteome coverage was comparable to previously reported studies, we cannot eliminate the possibility that there are undetected host proteins enriched in the diazoplast fraction.
- Tools such as immunofluorescence and genetics have not yet been established in *Epithemia* diatoms. We are therefore unable to confirm the localization or function of any *Epithemia* proteins.

RESOURCE AVAILABILITY

Lead contact

Requests for further information and resources should be directed to and will be fulfilled by the lead contact, Ellen Yeh (ellenyeh@stanford.edu)

Materials availability

Cultures and reagents used in this study are available upon request from lead contact.

Data and code availability

Data supporting the findings of this work will be made available upon publication on Mendeley at DOI:10.17632/rr9t3ccbc5.1. Sequencing data generated for this study will be made available at NCBI BioProject accession PRJNA1147773. The genome and annotation of *E. clementina* have been deposited and will be made available at NCBI accession [GCA ID TBD], TaxID 3042617.

The genome assembly pipeline is published at https://github.com/doenjon/Epithemia_assembly
Any additional information required to re-analyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGEMENTS

We thank Chriz Schvarcz and Kelsey McBeain for the generous sharing of *E. pelagica* cultures. We thank Dr. Devaki Bhaya and Dr. Arthur Grossman and members of their labs for support and feedback on the project. We are grateful to Scott Miller and Heidi Abresch for their *Epithemia* expertise and discussion. Our thanks to Andres Reyes for mass spectrometry technical training. We thank Daniel S. Rokhsar, Jonathan Zehr, Kendra Turk-Kubo, Andy Alverson, Elizabeth Ruck, Paolo Carnevali, Dmitri Petrov, and Cedric Feschotte for advice during the project. Anti-NifDK polyclonal antibodies were kindly provided by Dennis Dean. E.Y. is a Chan-Zuckerberg Biohub – San Francisco Investigator and supported by Burroughs Wellcome Fund. S.F. was partially funded by NIH training grant (T32GM007276). M.S.O. was partially funded by NIH training grant (5T32AI007328-32). The contents of this manuscript are solely the responsibility of the authors and do not represent the official views of the NCRR or the National Institutes

of Health. S.-L.X is funded by the NIH grant S10OD030441 and the Carnegie Endowment Fund to the Carnegie Mass Spectrometry Facility.

AUTHOR CONTRIBUTIONS

S.F., M.S.O., and E.Y, wrote the manuscript with input from all authors. S.F. and E.Y. conceptualized project. S.F., J.D., M.S.O., T.B., and E.Y. contributed to project design and strategy. Initial DNA extraction and preliminary analysis was performed by S.F., J.D., and T.B. Axenic cultures were isolated by S.L.Y.M. Final DNA isolation and sequencing experiments were performed by S.F. and S.L.Y.M. S.F. isolated and performed RNA sequencing experiments. Final genome assembly was performed by J.D. Genome annotation and analysis were performed by S.F. S.-L.X. provided proteomics resources and expertise. M.S.O. performed and analyzed proteomics. All authors read and approved the final paper.

DECLARATION OF INTERESTS

The authors declare no competing interest.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES

During the preparation of this work, the author(s) used GPT-4o mini and Claude 3.5 Sonnet for assistance with minor bioinformatic troubleshooting. The authors reviewed and edited all AI outputs and take full responsibility for the content of this publication.

SUPPLEMENTAL INFORMATION

Supplemental Figures S1-S4

Table S1. NCBI Accessions and access links for all species used in this study, related to Figures 1, 2, S1 and S2.

Table S2. RepeatMasker output tables for *Epithemia*, related to Figures 1B and 1C

Table S3. Gene codes, putative identity, and gene tree of *Epithemia*-specific horizontally transferred genes, related to Figures 1D, 1E, and S2C-E.

Table S4. Processed and unprocessed proteomics data, related to Figure 4.

Data S1. Integrative genomics viewer screenshots of genomic context, expression, and read support of all NUDTs, related to Figures 1 and 2.

TABLES

Table 1. *Epithemia* genome assembly statistics

Summary of assembly statistics for *E. clementina* and, where applicable, *E. pelagica*. Quality value (QV) represents a log-scaled estimate of the base accuracy across the genome, where a QV of 40 is 99.99% accurate. N50 and L90 are measures of genome contiguity. N50 represents the contig length (bp) such that 50% of the genome is contained in contigs ≥ N50. L90 represents the minimum number of contigs required to contain 90% of the genome. Finally, BUSCO (Benchmarking of Single Copy Orthologues) is an estimate of completeness of the genome ($BUSCO_{genome}$) and proteome ($BUSCO_{protein}$) of *E. clementina* and *E. pelagica*. '-' indicates no statistic.

MAIN FIGURE TITLES AND LEGENDS

Figure 1. Highly divergent *E. clementina and E. pelagica* genomes share many unique gene families

(A) Genome size and total gene number for published diatom genomes compared with *Epithemia* species (dark blue). (See also Figure S1, Table S1.)

(B) Comparison of repeat content in diatom genomes showing size of the whole genome (grey dots) or the genome excluding repeat elements (orange dots). X-axis is the same as 1C.

(C) Breakdown of repeat types in diatom genomes showing amount in Mbp of the genome occupied by repeat elements of specific class, indicated by color. (See also Table S2.)

(D) Cumulative distribution of amino acid identity between pairwise orthologs from reference species. Estimated divergence time of species pair is indicated (right bar graph). (See also Figure S2.)

(E) UpSet plot depicting the number of uniquely shared orthogroups between all diatom species (first column) or subsets of 2-4 species. Orthogroups shared by *E. pelagica* and *E. clementina* are highlighted in brown. Columns are ranked by the number of uniquely shared orthogroups. (See also Figure S2.)

Figure 2. Detection of nuclear integrations of diazoplast DNA (NUDTs)
(A) A representative, NUDT containing *E. clementina* nuclear genome locus on contig ctg002090. Tracks shown from top to bottom: nuclear sub-region being viewed (red box) within the contig (black rectangle); length of the sub-region, with ticks every 500bp; nanopore sequencing read pileup, showing long read support across the NUDT; location of repeat masked regions (dark grey bars); locations of homology to *E. clementina* diazoplast identified by BLAST, demarcating the NUDT (blue shade); regions of homology to the *E. clementina* diazoplast identified by minimap2 alignment, colors represent SNVs between the diazoplast and nuclear sequence. (See also, Data S1F.)
(B) Same as A, for the NUDT on contig ctg003780. (See also, Data S1G.)
(C) Circlize plot depicting the fragmentation and rearrangement of NUDTs. The diazoplast genome (blue) and the NUDT on contig ctg002090 (brown) with chords connecting source diazoplast regions to their corresponding nuclear region, inversions in red. The length of the NUDT is depicted at 100x true relative length for ease of visualization. (See also, Figure S3A-S3E.)
(D) Same as C, for the NUDT on contig ctg003780.
(E) Ratio of long read depth of NUDT compared to average read depth for the containing contig. Heterozygous insertions (light grey bars) show approximately 0.5x depth; homozygous insertions (black bars) show approximately 1.0x depth.
(F) GC content of NUDTs, compared to mean GC content for 5kb sliding windows of the diazoplast genome (blue dashed line) and the nuclear genome (brown dashed line). Shaded regions represent mean ± 1 SD.

Figure 3. Most NUDTs are decaying and non-functional
(A) Truncation of diazoplast genes contained within each NUDT relative to the full-length diazoplast gene.
(B) Nucleotide identity of diazoplast genes that are <30% truncated (points) contained within each NUDT compared to identity of the full containing NUDT sequence (bars). (See also, Figures S3F and S3G.)
(C) Normalized expression across each NUDT (blue highlight) +/- 1kb of the genomic region surrounding the NUDT. For each NUDT, a pair of tracks shows RNA-seq reads after polyA enrichment of whole RNA plotted within background signal range, from 0 - 0.1 BPM (top, grey) and RNA-seq reads after rRNA depletion of whole RNA, plotted from 0 - 7 BPM (bottom, black). The region corresponding to the *tusA* gene in ctg005680 is highlighted in dark blue (See also Data S1A-S1G.)

Figure 4. Few host-encoded proteins are detected in the diazoplast proteome
(A) Electron micrographs of (top) *E. clementina* cells with diazoplast (D), chloroplast lobes (C), and lipid bodies (L) indicated and (bottom) diazoplasts following purification with thylakoids (yellow arrow) indicated.
(B) Number of diazoplast-encoded (left) and host-encoded (right) proteins identified by LC-MS/MS. Total number of proteins identified from each respective proteome is shown above each stacked bar. Colored bars and numbers indicate proteins identified in purified diazoplasts only, whole cells only, or both.
(C) Volcano plot showing the enrichment of diazoplast-encoded (blue) and host-encoded (brown) proteins in whole cells or purified diazoplasts, represented by the difference between log2-transformed iBAQ values. Proteins enriched in the diazoplast are on the left side of the graph while those enriched in the host are on the right; the darker shade of each color represents significantly enriched hits. Host-encoded proteins significantly enriched in the diazoplast are shown with larger brown markers.

SUPPLEMENTAL FIGURE TITLES AND LEGENDS

Figure S1. *E. clementina* genome assembly statistics and features
(A) Scanning electron micrographs of *E. clementina,* scale bar 5μm. Top: View looking down on the dorsal girdle band. Middle: View down the apical axis. Bottom: View of the ventral face, lined by prominent fenestral bars regularly spaced between the radial striae. The raphe lies along the strongly curved keel on the ventral margin and pinches slightly towards the dorsal margin.

(B) GenomeScope spectrum of 35-mer multiplicity collected from the Illumina sequencing reads. Peak at 1x coverage (~90) and 2x coverage (~180), consistent with a diploid genome.
(C) Merqury spectrum of k-mer multiplicity collected from the Illumina sequencing reads, stacked lines colored by number of times k-mer is seen in the genome assembly. Few k-mers within the heterozygous and homozygous peaks are read-only (black), suggesting that the assembly is not missing significant sequence present in the reads.
(D) Stramenopile-specific Benchmarking Universal Single-Copy Orthologs (BUSCOs) for *E. pelagica* and *E. clementina* genomes and proteomes. Both genomes contain all stramenopile BUSCOs, however the *E. pelagica* annotation is less complete. The genome and proteome of *E. clementina* show some duplication. This duplication of stramenopile BUSCOs is similar or less than that found in previously published genomes of diatoms, for example, *Nitzschia putrida* (GCA_016586335, 39% duplicate BUSCOs)*, Fragilariopsis cylindrus* (GCA_900095095.1, 40% duplicate BUSCOs*), and Fragilaria radians* (GCA_900642245.1, 19% duplicate BUSCOs).
(E) Whole genome alignment of the *E. clementina* and *E. pelagica* genome assemblies. White indicates no sequence homology, yellow indicates alignments at <25% nucleotide identity. There is only 4.76% sequence homology between the two genomes at the nucleotide level, all at <25% identity.
(F) Genomic synteny between the whole genome alignments of the *E. clementina* and *E. pelagica* diazoplasts, showing 7 syntenic blocks.

Figure S2. Gene family enrichment and HGT statistics
(A) Asymmetrical heatmap of ortholog comparisons between diatom species pairs, showing mean amino acid identity (MAAI) of pairwise orthologs (top) and Jaccard similarity index (JSI) of orthogroups (bottom).
(B) Enrichment analysis of clusters of orthologous group (COG) categories in *E. clementina* and *E. pelagica.* Color indicates difference in proportion of each COG category within genes in *Epithemia* unique gene families and the whole functionally annotated gene set. An asterisk indicates categories that were found to be statistically significantly enriched.
(C) Total count of HGT candidate genes across diatom species, considering only high-confidence candidates.
(D) Proportion of the genes considered for HGT analysis that were identified as high-confidence HGT candidates. Genes were considered for analysis if they had >10 identifiable homologues.
(E) UpSet plot depicting the number of uniquely shared, HGT candidate containing orthogroups between diatom species groupings (including singlets). Columns are ranked by the number of uniquely shared orthogroups.

Figure S3. NUDT fragmentation and gene containing regions
(A-E) Circlize plots depicting the fragmentation and rearrangement of the NUDTs. The diazoplast genome (blue) and the NUDT on labeled contig (brown) with chords connecting source diazoplast regions to their corresponding nuclear region, inversions in red. The length of the NUDT is depicted at 100x true relative length for ease of visualization.
(F) Translation in all potential frames of the gene contained within the NUDT on contig ctg002090 (transcriptional repressor, gene ID: P3f56_RS08570). The copy within the NUDT (bottom) is untruncated (100% of the full-length gene) by nucleotide sequence and is 96% identical to the corresponding diazoplast gene (top). Compared to the diazoplast-encoded gene, the gene contained in the NUDT has a mutation that results in a premature stop codon at amino acid 39 (out of 177). Red highlight indicates a potential translation. 5'3' Frame 1 is the native diazoplast frame.
(G) Translation in all potential frames of the gene contained within the NUDT on contig ctg002090 (low-complexity tail membrane protein, gene ID: P3F56_RS01750). The gene is 9% truncated at the 3' terminus (91% of the full-length gene). Compared to the diazoplast-encoded gene, the gene contained in the NUDT has several non-synonymous mutations and is missing 16 amino acids at the C-terminus. Red highlight indicates a potential translation. 5'3' Frame 1 is the native diazoplast frame.

Figure S4. Comparative pathway analysis of diazoplasts and close relatives
KEGG pathway analysis of *E. clementina, E. turgida, E. gibberula* diazoplasts as well as *C. subtropica* and UCYN-A*,* indicating presence (green circle) or absence (red x) in the genome. Filled green circle indicates evidence for import of a host-encoded protein; filled black circle indicates presence in the

endosymbiont genome and evidence for import of a host-encoded protein. A black dashed box indicates species for which data exists for both endosymbiont and host nuclear genes.

SUPPLEMENTAL DATA TITLES AND LEGENDS

Data S1: Detailed genome tracks across NUDT regions, related to Figures 3 and 4

(**A-G**) For all NUDTs, full context genome tracks from the Integrated Genomics Viewer zoomed in to the NUDT region (left) or zoomed out to a 20kb surrounding region (right). Tracks from top to bottom are:

1) Region file of masked repeat regions;
2) Feature file of *E. clementina* gene models;
3) Region file of *E. clementina* diazoplast homology found by BLAST, demarcates the NUDT;
4-7) Alignment files of homology found by minimap2 when aligning 4) *E. clementina* diazoplast, 5) *E. pelagica* diazoplast, 6) *E. turgida* diazoplast, and 7) *E. gibberula* diazoplast to the *E. clementina* nuclear genome;
8-10) Normalized expression data in BPM of RNA seq from combined replicates of poly-adenylated transcript enriched RNA collected from three treatment conditions.;
11-13) Normalized expression data in BPM of RNA seq from combined replicates across of ribosomal RNA depleted RNA collected from three treatment conditions.;
14) Read pileup of axenic nanopore reads. Colored bars at certain sites indicate proportion of SNVs across the reads deviating from the haplotype reference assembly often resulting from a heterozygous site but sometimes from reads accumulating at undiscernible copies of repeat elements.;
15) Alignment file of axenic, nanopore long reads aligned to the reference *E. clementina* genome. An aligned read identical to the reference sequence is rendered as a single plain grey bar. Colors at sites along the read denote SNVs from the reference assembly. Small indels are denoted by small purple bars. A thin black bar within a read represents a region not present in the read that is present in the haplo-assembly (i.e. larger indels). Very light grey bars are secondary alignments, which accumulate at repeat elements.

STAR★METHODS

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Diatom strains and culture conditions

Wild isolates of *E. clementina* were cultivated in CSi-N media in vented flasks under 10 µmole photon m -2 s-1 of white light at 20°C. Cultivation procedures are also detailed in[28]. *E. pelagica* cells from wild isolates reported in Schvarcz et al.[30] were shipped overnight from University of Hawaii at Manoa, HI in an insulated box. Upon arrival cultures were centrifuged at 1000 x g, media was aspirated, and cells were resuspended in trizol for RNA extraction.

METHOD DETAILS

Generation of axenic culture

Cultures were initially isolated from a single cell of *E. clementina* and were thus clonal but xenic. To generate axenic cultures for sequencing, cells were incubated overnight with lysozyme to disrupt the cell walls of gram-positive bacteria, then treated to a 30-minute pulse of antibiotic cocktail (100 µg/mL Carbenicillin, 25 µg/mL Chloramphenicol, 5 mg/mL Levofloxacin, 50 mg/mL Rifampicin, 50 µg/mL Streptomycin). Immediately following, cultures were spray plated[74]. A small volume of dilute culture suspension was aspirated into a glass Pasteur pipette and held perpendicular to a stream of sterile air. The air atomizes the culture; the small droplets are then captured on a CSi-N agar plate. This process isolates single cells of *E. clementina* and disrupts their associated bacterial community. Cells on the agar plates were allowed to form colonies which were screened for any visual bacterial growth. Only those colonies that lacked bacterial growth were chosen for further cultivation. The resulting strain was expanded and confirmed axenic in subsequent sequencing experiments.

Scanning electron microscopy

Xenic cultures of *E. clementina* were resuspended, pelleted at 23˚C at 1000 x g, rinsed with CSi-N media, and then resuspended in 250uL of PBS. Cells were transferred in a droplet to poly-L-lysine

726 coated 12mm diameter glass cover slips and left to sit on a flat surface for 5min. The PBS was gently
727 aspirated, and a droplet of 4% paraformaldehyde in PBS was added to coat the entire cover slip surface.
728 Cells were fixed for 10min in the dark and then the cover slip was rinsed twice with PBS. An ethanol
729 dehydration series was performed wherein cover slips were sequentially immersed in 60%, 70%, 80%,
730 90%, and 100% v/v ethanol in PBS. The cover slips were gently dried on a 42˚C heat block. The cover
731 slips were secured to a low-profile pin mount and sputter coated in a Leica ACE600 High Vacuum Sputter
732 Coater with gold to a thickness of 6nm. Samples were imaged on a Zeiss Sigma FE-SEM.
733
734 **High molecular weight DNA extraction**
735 *E. clementina* were grown to a density of approximately 400,000 cells/mL and 20-30 million cells
736 were used as input to HMW DNA extraction. Xenic cultures were first subjected to a round of
737 centrifugation through a discontinuous Percoll gradient to deplete excess bacteria. *E. clementina* cells
738 pellet out of the solution entirely, whereas a portion of their bacterial community stays suspended in
739 various Percoll fractions. Centrifugation steps were performed at 23˚C at 1000 x g. For both xenic and
740 axenic cultures, HMW DNA was isolated using a nuclei extraction method[75]. Cells were suspended in a
741 minimal volume of nuclear isolation buffer (NIB) and the transferred to a mortar where they were flash
742 frozen and then ground with the pestle until a paste formed. This grinding process was repeated a total of
743 three times. Cell homogenate was transferred to a 15mL falcon tube containing NIB, rinsing the mortar
744 with NIB if necessary, and incubated at 4˚C for 15min. No miracloth filtering step was performed. The cell
745 homogenate was spun down at 4˚C and 2900 x g. The resulting nuclei pellet was rinsed with 15mL NIB
746 until the solution was clear of any photosynthetic pigments. The resulting nuclei/cellular compartment mix
747 was used as input for the Nanobind plant nuclei big DNA kit from PacBio. Steps were followed as listed in
748 the kit protocol except for large cell inputs, in which reagent volumes were doubled and the Proteinase K
749 digestion step was extended to 2hrs. The isolated DNA from this protocol was processed with the Short
750 Read Eliminator kit from PacBio to deplete DNA fragments < 25kb in length. The final, HMW DNA was
751 used as input for nanopore library preparation.
752
753 **Nanopore library preparation and sequencing**
754 For all sequencing runs from axenic cultures of *E. clementina*, 1-2µg of HMW DNA was used as
755 input to the Oxford Nanopore Technology sequencing by ligation kit (SQK-LSK112). The nanopore
756 protocol (Version: GDE_9141_v112_revC_01Dec2021) was followed with the following minor
757 modifications: end repair incubation was lengthened to 30min at 20˚C and the adapter ligation incubation
758 was lengthened to 60min at room temperature. Resulting libraries were loaded onto primed, high-
759 accuracy MinION R10.4 flow cells (FLO-MIN112) at a target amount of 9 fmoles of 10kb DNA. In actuality,
760 DNA sizes ranged within samples and between sequencing runs, but 9 fmole maintains a recommended
761 loading amount for the flow cell at a range of fragment sizes. All sequencing of xenic cultures was
762 performed similarly, but with previous iterations of the sequencing kit (SQK-LSK110) and the flow cell
763 MinION R9.4.1 (FLO-MIN111). If pore occupancy dipped below roughly 1/3 of starting occupancy during
764 the sequencing run, the run was paused, and the flow cell was washed with the Flow Cell Wash Kit (EXP-
765 WSH004) from Nanopore according to the associated protocol. The same prepped library was then
766 reloaded onto the flow cell and the sequencing run was restarted. Each run was left to sequence for 3-5
767 days, or until pore occupancy was near zero.
768
769 **Isolation of genomic DNA for Illumina sequencing**
770 DNA was extracted from axenic *E. clementina* cultures following the QIAGEN DNeasy Plant Pro
771 Kit (69206) protocol. For the lysis step, cell suspension was transferred to the kit's tissue-disrupting tubes
772 included along with 100mg 0.5mm autoclaved glass beads added and placed in a bead-beater and
773 shaken for one minute. 300ng of the isolated axenic *E. clementina* DNA was used as input to the
774 NEBNext Ultra II FS DNA Library Kit for Illumina (E7805S). A fragmentation time of 16min was used for a
775 target insert size of 200-450bp. Samples were indexed with NEBNext Multiplex Oligos for Illumina Dual
776 Index Primers Set 1 (E7600S). DNA concentration of resulting libraries was determined with a Qubit
777 dsDNA Quantification Assay High Sensitivity kit (Q32851). Final libraries were checked for quality and
778 size-range using an Agilent Bioanalyzer High-Sensitivity DNA chip. The final mean insert size was 440bp,
779 with a well-formed size distribution around the mean and minimal adapter dimers. The library was
780 sequenced on an Illumina NextSeq 2000 P3 for 2 x 150bp reads. Raw reads were trimmed and paired

781 with fastp (--qualified_quality_phred 20, --unqualified_percent_limit 20) for a final total of in 402 million
782 read pairs from axenic *E. clementina*[76].
783
784 **RNA isolation and sequencing**
785 To capture a wide range of transcripts, axenic cultures of *E. clementina* were exposed to different
786 nitrogen conditions and collected at different times in the day-night cycle. Axenic cultures of *E. clementina*
787 were seeded in 175cm^2 sterile vented flasks at a density of 1.2 million cells per flask. For conditions of
788 nitrogen repletion, media in the flask contained 100µM of ammonium. Cells were kept in -N or +NH_4^+
789 conditions for 72 hours and harvested two hours into the day period. Cells in nitrogen depleted conditions
790 were additionally collected two hours into the night period. All cells were scraped from the flask,
791 centrifuged to concentrate, resuspended in trizol, and flash frozen. Each condition was collected in
792 triplicate for each experiment, and the whole experiment was performed twice. To lyse, the trizol
793 suspended cells were held on ice and a sonicator probe was submerged at the center of the tube.
794 Sonication was performed with a microtip at 50/50 on/off pulses for one minute at an intensity setting of
795 six on a Branson 250 Sonifier (B250S). RNA was isolated using the QIAGEN RNeasy Plus Universal Mini
796 Kit (74134) following the included protocol. 500ng of RNA per condition per replicate collected from the
797 first experiment was used as input to the NEBNext poly(A) mRNA Magnetic Isolation module (E7490L) to
798 enrich for mRNA and the NEBNext Ultra II Directional RNA Library Prep Kit for Illumina (E7760L) and
799 indexes from NEBNext Multiplex Oligos for Illumina Dual Index Primers Set 1 (E7600S) were used for
800 library preparation. 350ng of RNA per condition per replicate collected from the second experiment was
801 used as input to the Zymo-Seq RiboFree Universal cDNA Kit (R3001) and indexed with Zymo-Seq UDI
802 Primer Set (Indexes 1-12) (D3008). For each experiment, libraries were pooled and sequenced on an
803 Illumina NextSeq 2000 P3 for 2 x 150bp reads.
804 *E. pelagica* cultures provided by courtesy of Chris Schvarcz and Kelsey McBeain proved to be
805 unculturable long-term in lab conditions after shipment. Therefore, RNA was extracted upon receipt of
806 overnight shipment from University of Hawaii at Manoa, HI. Otherwise, the same method of poly(A)-
807 enrichment and Illumina sequencing was used as for *E. clementina*.
808
809 **Data filtering and genome assembly of *E. clementina***
810 Initial genome size, ploidy, and repeat content estimates were made by counting k-mers in the
811 axenic Illumina reads with jellyfish v2.2.10 (-C -m -k 35 -s 5G) and plotting with GenomeScope[77,78]. The
812 raw fast5 sequencing files were basecalled with guppy v1.1.alpha13-0-g1ec7786. Reads were filtered
813 based on minimum length 3kb and quality 20 with Nanofilt v2.8.0[79]. Read statistics were calculated with
814 NanoPlot v1.30.1. Basic quality checks were performed with fastqc v0.11.9[80]. Post filtering, 19.5Gb of
815 sequence from axenic cultures of *E. clementina* and 30.2Gb of sequencing data from xenic cultures were
816 used for a two-step assembly process. First, axenic reads were assembled with NextDenovo v2.5.0[81].
817 Then, xenic nanopore sequencing data was aligned to the axenic assembly using minimap2 (-ax map-
818 ont) v2.24-r1122 to identify probable diatom reads in the xenic data[82]. Finally, axenic and diatom-mapped
819 xenic nanopore reads were combined and assembled with NextDenovo (using default or machine-specific
820 options, except read_cutoff=5k, genome_size=350M). Axenic Illumina data was mapped to the assembly
821 with BWA v0.7.17-r1188[83]. Contigs in the assembly were removed if less than 70% of the contig was
822 covered by the axenic Illumina reads or if those reads mapped at significantly lower depth than to the rest
823 of the contigs (< 4% of mean depth). The axenic Illumina reads were then used as input for 3 rounds of
824 polishing with Racon v1.5.0 and one round of Polca (part of MaSuRCA v4.0.5)[84–86]. Further contamination
825 analysis of the assembly and reads was performed with blobtools v1.1.1[87]. Organellar genomes for the
826 diazoplast, chloroplast, and mitochondria were assembled and annotated as previously reported[28]. All
827 contigs in the assembly were aligned to the organellar genomes and to the diazoplast genome to check
828 for remaining organellar contaminants in the assembly. Any remaining organellar contigs contaminating
829 the nuclear assembly were identified and removed if they aligned end-to-end to the already assembled
830 organellar and endosymbiont genomes. Basic assembly statistics were extracted with QUAST v5.2.0[88].
831 Final assembly completeness and consensus quality (QV) was assessed with the k-mer spectra tool
832 Merqury v1.3[89]. The QV of our final assembly was 38.52. BUSCO v5.3.2 in genome mode was also used
833 to estimate completeness at the gene level[90].
834
835 **Repeat masking of *E. clementina* and *E. pelagica***

The final nuclear assembly of *E. clementina* and the publicly released[42] but raw nuclear sequence of *E. pelagica* (uoEpiScrs1.2 GCA_946965045.2) were used as input to the RepeatModeler2 and RepeatMasker pipelines. To identify and classify the repeat elements for both *Epithemia* genomes, the workflow for RepeatModeler v2.0.2 with built-in LTR detection and classification was run[91] (BuildDatabase -engine rmblast, RepeatModeler -LTRStruct, RepeatClassifier -engine rmblast). Since the repeat models for the organisms are *de novo* and repeat data for diatoms in the source databases may be limited, the repeat families classified as 'Unknown' were further interrogated to ensure no protein-coding genes were annotated as repeats. To do this, the Unknown repeat families were used as input to ncbi-blast+ against the NCBI non-redundant (NR) protein database[92,93] (November 3rd 2022). Approximately 8% of Unknown repeats had significant similarity to eukaryotic and Bacillariophyta proteins. Out of caution, these regions annotated as Unknown repeats with protein hits were removed from the repeat database to be kept unmasked. Finally, RepeatMasker v4.1.2-p1 (-engine rmblast, -s no_is -norna -gff -xsmall) was run on the genomes to soft-mask all repeat regions. The ParseRM tool by Aurelie Kapusta was used to extract repeat type and divergence from consensus from the raw .classified and .align output files from RepeatMasker[94].

Gene annotation of *E. clementina* and *E. pelagica*

For both *E. pelagica* and *E. clementina,* the masked nuclear genome of each organism was annotated in two independent runs of BRAKER2 v2.1.6, which applied installs of GeneMark-ETP v1.0, AUGUSTUS v3.4.0, and ProtHint v2.6.0. First, the BRAKER2 pipeline was given extrinsic protein evidence as input. Protein sequences were sourced from the orthoDB v10 protozoa database which was manually edited to include diatom proteins from recent annotations. Second, the BRAKER2 pipeline was given transcriptomic evidence from the source organism[95–104]. To produce the aligned RNA-seq evidence, the RNAseq reads were quality filtered, trimmed and paired with fastp[76] v0.22.0 (--qualified_quality_phred 20, --unqualified_percent_limit 20), and then aligned to the source genome with hisat2 v2.1.0[105] (--rna-strandness RF). Alignment files were sorted and converted to binary alignment files with samtools v1.16.1[106]. For *E. pelagica*, a single 280 million read Illumina run from polyA-enriched RNA was used as input. For *E. clementina*, actively maintained lab cultures enabled more extensive sequencing of the transcriptome. RNA from 30 samples and five different conditions using both polyA-enrichment and rRNA-depletion methods of isolation were used as input. The outputs of these two independent protein-based and transcriptome-based annotations were merged using TSEBRA v1.0.3 into a single annotation[98]. Both the input and output general transfer format (GTF) file were fixed with the fix_gtf_ids.py script included with TSEBRA. The output GTF files were converted to multi-isoform fasta files, removing any pseudo genes or genes interrupted by stop codons using gffread v0.12.7[107] (-J --no-pseudo -y). Completeness of the final annotation was assessed with BUSCO v5.3.2 in proteins mode. To inspect isoforms, the AGAT v1.0.0 agat_sp_keep_longest_isoform.pl tool was used[108]. Functional annotation was performed with eggNOG-mapper[109,110]. For each species, genes in shared, *Epithemia*-specific orthogroups were used for tests of enrichment against a gene universe of functionally annotated genes with p-value cutoff 0.1. clusterProfiler[111] was used for tests of significance. Proportion of each COG term assigned within each set of functionally annotated genes was calculated and used to calculate difference in proportion.

Orthologue analysis

Curated species proteomes and genomes were downloaded from NCBI or associated online repositories[42,45,112–120] (Table S1). The agat package was used to remove short isoforms (agat_sp_keep_longest_isoforms.pl). Where necessary, gene feature files were reformatted[108] (agat_sp_manage_attributes.pl -p gene -att transcript_id). Finally, longest isoform proteomes were produced from the gene feature files and the corresponding species genome with gffread, removing genes without a complete, valid coding sequence and removing pseudo-genes[107] (gffread -J –no-pseudo -y). The resulting proteomes were used as input for Orthologue analysis.

Orthogroups were identified with orthofinder v2.5.4 (-M msa -T iqtree) and orthogroup overlaps between species were extracted from Orthogroups_SpeciesOverlaps[121]. To quantify shared orthologues between species without biasing for total proteome size differences, the Jaccard similarity coefficient for each species pair was calculated according to the standard Jaccard index formula where A and B are the total number of self-orthologues identified for each organism and $A \cap B$ is the number of orthologues identified between the organisms as contained in the OrthologuesStats_one-to-one file. To identify uniquely overlapping orthogroups (e.g. orthogroups shared between two species and not by any other

species), orthogroup sets from Orthogroups.GeneCount.tsv were parsed and plotted with UpSetR[122]. To quantify sequence similarity, orthologue pairs were identified by reciprocal best BLAST between organism pairs and the full-length percent amino acid identity was calculated from the BLAST outputs, similar to the method used in[123].

NUDT homology search

Whole genomes of free-living cyanobacterial relatives of the endosymbiont were curated along with available whole endosymbiont genomes (Table S1). These sequences were used as query for homology searches against the nuclear genomes of *E. pelagica* and *E. clementina*. Command line BLASTN with defaults, BLASTN using the custom settings previously validated for NUMT search[92,124] (-reward 1 -penalty -1 -gapopen 7 -gapextend 2), minimap2 (-ax asm5 and -HK19 modes), and nucmer v4.0.0rc1 were all used to perform these homology searches[82,125]. As negative controls, the reversed sequence of the *E. pelagica* mitochondria and the *E. coli* genome were used. For all cyanobacterial and endosymbiont queries, BLASTN was the most sensitive and least stringent, identifying all homology regions identified by other programs. Contiguous regions of homology < 500bp in length were not considered, though most short alignments were < 100bp. The resulting > 500bp contiguous regions of homology were considered candidate NUDTs. Seven regions in total for *E. clementina* and none in *E. pelagica.* To verify that these alignments were not a result of misassembly, long-reads from nanopore sequencing of axenic cultures were aligned (minimap -ax map-ont) and the reads spanning the border of the insertion were counted and the depth compared to that of the contig. The read depth for the contig was calculated with samtools depth (considering only primary alignments to minimize skews from repetitive regions). To check for expression within the NUDTs, RNA-seq data from both polyA enrichment and rRNA depletion experiments was mapped as previously described and normalized with deeptools v3.3.1 bamCoverage (--normalizeUsing BPM -p max -bs 100)[126]. Corresponding source regions from the endosymbiont and percent identities were pulled from the blast results. Using bedtools v2.30.0 intersect, the source regions were overlapped with endosymbiont gene regions[127]. These coordinates were then mapped back to the nuclear region. Nuclear and diazoplast sequences corresponding to these identified gene fragment containing regions were aligned using EMBOSS Needle v6.6.0.0, which calculates percent identity[128]. The truncation was calculated by dividing the length of the gene fragment by the total length of the corresponding diazoplast gene. For both the nuclear and diazoplast genomes of *E. clementina*, GC content variation was analyzed in sliding windows of 5000bp with a step size of 1000bp using bedtools makewindows and bedtools nuc. All alignments were visualized with the Integrative Genomics Viewer[129] (IGV) and plotted with circlize[130].

Identification of Horizontal Gene Transfers

Diatom proteomes (Table S1) including the *de novo* predicted *E. pelagica* and *E. clementina* were used as input to a custom HGT pipeline adapted from[56]. In brief, the program uses diamond v2.0.14 to collect homologues from the NCBI NR database for each gene in an organism[131]. To best ensure representation of genes from a diverse range of taxa, three diamond runs were performed against different subsections of NR: Bacteria, the SAR supergroup, and the remainder of the database. These results are parsed so that, where possible, the final list of homologues for each gene consists of no more than 70% of any one kingdom and does not contain any hits to self (relevant for diatom proteins already in the NCBI NR). Proteins with under 10 identifiable homologues were excluded from further analysis. Proteins were aligned with mafft v7.525 (--auto) and poorly aligned regions trimmed with trimAl v1.4.rev15 (-automated1)[132,133]. The L-INS-i method in mafft was selected for most alignments. These alignments were used as inputs for generation of phylogenetic trees. FastTree v2.1.1 was used to construct trees[134]. The topology of these trees was parsed by PhySortR v1.0.8 (min.support = 0.7, min.prop.target = 0.7, clade.exclusivity = 0.9) to identify trees in which the diatom gene of interest is more closely related to bacterial homologues than to eukaryotic ones[135]. The results were parsed using custom scripts to remove genes with fewer than five bacterial taxa in the tree. PhySortR designates genes as All Exclusive, Exclusive, Non-Exclusive, or Negative based on the tree topology. We treated All Exclusive and Exclusive results as high confidence and Non-Exclusive as low confidence. In reality, HGT candidates with Non-Exclusive tree topology are a mix of ambiguous topology as well as likely real HGTs shared between the diatoms or other eukaryotes. The Non-Exclusive HGT candidates were further filtered based on Alienness score (AI)[136]. The alienness score was calculated with both the best prokaryote Evalue and with the best prokaryote Evalue after the first group of *Bacillariophyta* results, to account for HGTs that may be shared

948 between diatom species. HGT candidates with positive AI scores were kept for subsequent analysis.
949 Species of origin for HGT candidates was inferred using the taxonomic breakdown of the top blast result.
950
951 **Diazoplast Isolation**
952 *E. clementina* cells were harvested by scraping, then washed twice in CSI-N growth medium by
953 centrifugation at 2,000xg, and re-suspensed in spheroid body isolation buffer (50 mM HEPES pH 8.0, 330
954 mM D-sorbitol, 2 mM EDTA NaOH pH 8.0, 1 mM $MgCl_2$). Cells were then placed in a bath sonicator for
955 10 minutes followed by 3 low pressure cycles (500 psi) and by 5 high pressure cycles (2,000 psi) in an
956 EmulsiFlex-C5 Homogenizer (Avestin) or until most cells appeared lysed under a microscope. After a 1-
957 minute spin at 100xg to pellet the unbroken cells and broken frustules, the supernatant was collected and
958 centrifuged at 3,000xg for 5 minutes to concentrate the diazoplasts and other organelles to a volume of 3-
959 4 mL. This fraction was then split equally, and each half was laid on a discontinuous Percoll gradient.
960 89% Percoll, 10% 10xPBS, and 1% 1M HEPES pH 8.0 was diluted with SIB to generate the gradient,
961 which consisted of 2 mL 90%, 3 mL 70%, 3 mL 60%, 3 mL 50%. The gradient was centrifuged for 20
962 minutes at 12,000xg, 4º using a Beckman Optima L-90K ultracentrifuge with SW-41 rotor.
963 The boundaries between the 60% and 70% layers and the 70% and 90% layers were collected,
964 counted, and checked for purity via light microscopy. They were then diluted 1:6 in SIB Buffer and
965 centrifuged at 2,000xg for 2-3 minutes to collect diazoplasts, which were resuspended in 200 µL
966 Extraction Buffer (100mM Tris-HCl, pH8.0, 2% (wt/vol) SDS, 5mM EGTA, 10mM EDTA, 1mM PMSF, 2x
967 protease inhibitor (1 tablet each of cOmplete™ Protease Inhibitor Cocktail, catalog number 4693116001
968 and Pierce™ Protease Inhibitor tablet, EDTA free, catalog number A32965)). During optimization,
969 enrichment was assessed by Western blot for NifDK on both the diazoplast and whole cell extracts.
970
971 **Protein Extraction, Preparation, and LCMS/MS**
972 We generated whole cell lysate by homogenizing with a bead beater at 3000 strokes per minute
973 for 3 minutes with 1 mm glass beads (BioSpec Products catalog number 11079110) or until most cells
974 appeared lysed under a microscope. Diazoplasts were lysed similarly using 0.5 mm beads (BioSpec
975 Products catalog number 11079105). Beads were pelleted at 100xg for 1 minute and the supernatant was
976 removed; the beads were washed twice with 50 µL extraction buffer each by vortexing and spinning.
977 These fractions were then added to the supernatant for a total of 300 µL, followed by an equal volume of
978 cold Tris-buffered phenol (pH 7.5-7.9). This solution was vortexed for 1 minute, centrifuged at 18,000 x g
979 for 15 minutes at 4º C. The upper phase was discarded, then extracted with an equal volume of cold
980 50mM Tris-HCl, pH8.0. The phenol phase was extracted with Tris-HCl a total of three times, followed by
981 addition of 0.1 M ammonium acetate in methanol and overnight incubation at -80º C. Samples were then
982 transferred to new tubes and centrifuged at 18,000 x g for 20 minutes at 4º C. The supernatant was
983 discarded and the pellet was washed once in 0.1 M ammonium acetate in methanol and twice in 1 mL
984 cold methanol by centrifugation for 5 minutes at 18,000 x g at 4º C, followed by a final short spin and
985 removal of trace methanol. The pellet was then resuspended in 150 µL resuspension buffer (6M
986 Guandine-HCl in 25mM NH_4HCO_3 pH8.0). Each sample was then reduced with TCEP at a final
987 concentration of 2 µM (Thermo Scientific catalog number 20490) for 1 hour at 56º C, alkylated with
988 iodoacetamide (Thermo Scientific catalog number 90034) at a final concentration of 10 mM for 1 hour at
989 ambient temperature, and then diluted with 3 volumes of 25mM NH4HCO3. Sequencing grade modified
990 trypsin (Promega catalog number V5111) was added at a ratio of 1:50 followed by overnight incubation at
991 37º C, then repeated the next morning, followed by quenching the reaction by adding formic acid to a final
992 concentration of 1%. Each sample was then loaded onto a C18 cartridge (Sep-pak waters catalog
993 number WAT054960), activated with 80% acetonitrile and 0.1% formic acid. The flow-through was loaded
994 a total of three times, followed by five washes with 1 mL 0.1% formic acid. The samples were then eluted
995 with 200 µL of 80% acetonitrile 1% formic acid and the flow-through re-loaded a total of three times.
996 Peptide concentration was determined using Pierce™ Quantitative Colorimetric Peptide Assay
997 (Thermo Fisher catalog number 23275). 1 µg of peptides from each sample was loaded on either on a Q-
998 Exactive HF hybrid quadrupole-Orbitrap mass spectrometer (Thermo Fisher) (1 replicate) or an Eclipse
999 Tribrid mass spectrometer (Thermo Fisher) (2 replicates), equipped with an Easy LC 1200 UPLC liquid
1000 chromatography system (Thermo Fisher). Peptides were first trapped using a trapping column (Acclaim
1001 PepMap 100 C18 HPLC, 75 µm particle size, 2 cm bed length), then separated using analytical column
1002 AUR3-25075C18, 25CM Aurora Series Emitter Column (25 cm x 75 µm, 1.7µm C18) (IonOpticks). The
1003 flow rate was 300 nL/min, and a 120-min gradient was used. Peptides were eluted by a gradient from 3 to

28 % solvent B (80 % acetonitrile, 0.1 % formic acid) over 106 min and from 28 to 44 % solvent B over 15 min, followed by a short wash (9 min) at 90 % solvent B. The Q-Exactive HF hybrid quadrupole-Orbitrap mass spectrometer was configured as follows: Precursor scan was from mass-to-charge ratio (m/z) 375 to 1600 (resolution 120,000; AGC 3.0E6, maximum injection time 100ms) and top 20 most intense multiply charged precursors were selected for fragmentation (resolution 15,000, AGC 5E4, maximum injection time 60ms, isolation window 1.0 m/z, minimum AGC target 1.2e3, intensity threshold 2.0 e4, include charge state =2-8). Peptides were fragmented with higher-energy collision dissociation (HCD) with normalized collision energy (NCE) 27. Dynamic exclusion was enabled for 24s. The Orbitrap Eclipse Tribrid mass spectrometer was configured as follows: Precursor scan was from mass-to-charge ratio (m/z) 375 to 1600 (resolution 120,000; AGC 200,000, maximum injection time 50ms, Normalized AGC target 50%, RF lens(%) 30) and the most intense multiply charged precursors were selected for fragmentation (resolution 15,000, AGC 5E4, maximum injection time 22ms, isolation window 1.4 m/z, normalized AGC target 100%, include charge state=2-8, cycle time 3 s). Peptides were fragmented with higher-energy collision dissociation (HCD) with normalized collision energy (NCE) 27. Dynamic exclusion was enabled for 30s.

Proteomics Data Analysis

Maxquant version 2.5.0 was used for proteomics database searches, using default parameters with the following changes: label-free and iBAQ quantification, matched between runs were enabled[137]. For identifications, peptides were searched against the *Epithemia clementina* reference host and diazoplast proteomes. The proteingroups.txt file output from MaxQuant was analyzed using Perseus version 2.0.10.0[138]. iBAQ values were imported and filtered to remove potential contaminants, reverse hits, and those only identified by site. Only proteins identified by two more unique peptides and with a minimum of 5% sequence coverage were included in further analysis. The iBAQ values were then log(2) transformed for normality, proteins with two or more non-valid values were removed, and missing values were imputed from a downshifted normal distribution of the total matrix (width 0.3 standard deviations, down shift 1.8 standard deviations). A two-sided students T-test using the significance analysis of microarrays method (s0=0.1, false discovery rate 0.05, 250 randomizations) was used to determine the enrichment of host-encoded proteins in the diazoplast.

Immunoblot

Whole cell and isolated diazoplast lysates were prepared as described above. Protein concentration was determined using Pierce™ BCA Protein Assay Kit (Thermo Fisher catalog number 23227). 0.5 µg of protein from each sample was diluted in lithium dodecyl sulfate buffer with 100 mM DTT and loaded onto a NuPage™ Bis-Tris gels 4-12% acrylamide (Thermo Fisher catalog number NP0321BOX) in MES Buffer, and separated by electrophoresis, using Precision Plus Protein™ All Blue Standards (BioRad catalog number 1610373). Proteins were then transferred into a nitrocellulose membrane using Bio-rad Transblot Turbo, followed by blocking in LiCOR blocking buffer 1 hour at room-temperature. The membrane was then incubated for two hours at room temperature with primary antibodies anti-NifDK (polyclonal goat at 1:5000 dilution, kindly provided by Dr. Dennis Dean from Virginia Tech, US) to detect nitrogenase and anti-PsbA (1:10,000 dilution rabbit from AgriSera AB, Vanas, Sweden) as an internal loading control. Antibodies were diluted in a solution of 50% TBST and 50% LiCOR blocking buffer. The membrane was then washed three times with TBST and incubated with LiCOR secondary antibodies (IRDye 800CW) for 1 hour (goat α-rabbit for PsbA and donkey α-goat for NifDK). After two rinses with TBST and one with PBS, the blot was imaged using an infra-red LiCOR imager. Intensity of the signal was quantified using Image Studio Lite software v5.2.

QUANTIFICATION AND STATISTICAL ANALYSIS

In all cases, statistical analyses were performed using pre-existing software. The specific statistical tests, settings, and cutoffs used are listed with their accompanying analysis in the preceding method details.

	E. clementina	*E. pelagica*
Genome size (bp)	418,007,894	60,195,788
GC	44.3%	48.19%
QV	38.52	-
Contig/chromosome #	642	15
N50	1,108,441	-
L90	412	-
Gene #	26,453	20,203
Repeat %	80%	27.36%
$BUSCO_{genome}$	100%	100%
$BUSCO_{protein}$	99%	94%
Diazoplast genome size (bp)	3,072,807	2,483,960
Diazoplast gene #	1,910	1,679

1055
1056 **Table 1. *Epithemia* genome assembly statistics**
1057 Summary of assembly statistics for *E. clementina* and, where applicable, *E. pelagica*. Quality value (QV)
1058 represents a log-scaled estimate of the base accuracy across the genome, where a QV of 40 is 99.99%
1059 accurate. N50 and L90 are measures of genome contiguity. N50 represents the contig length (bp) such
1060 that 50% of the genome is contained in contigs ≥ N50. L90 represents the minimum number of contigs
1061 required to contain 90% of the genome. Finally, BUSCO (Benchmarking of Single Copy Orthologues) is
1062 an estimate of completeness of the genome ($BUSCO_{genome}$) and proteome ($BUSCO_{protein}$) of *E. clementina*
1063 and *E. pelagica*. '-' indicates no statistic.
1064

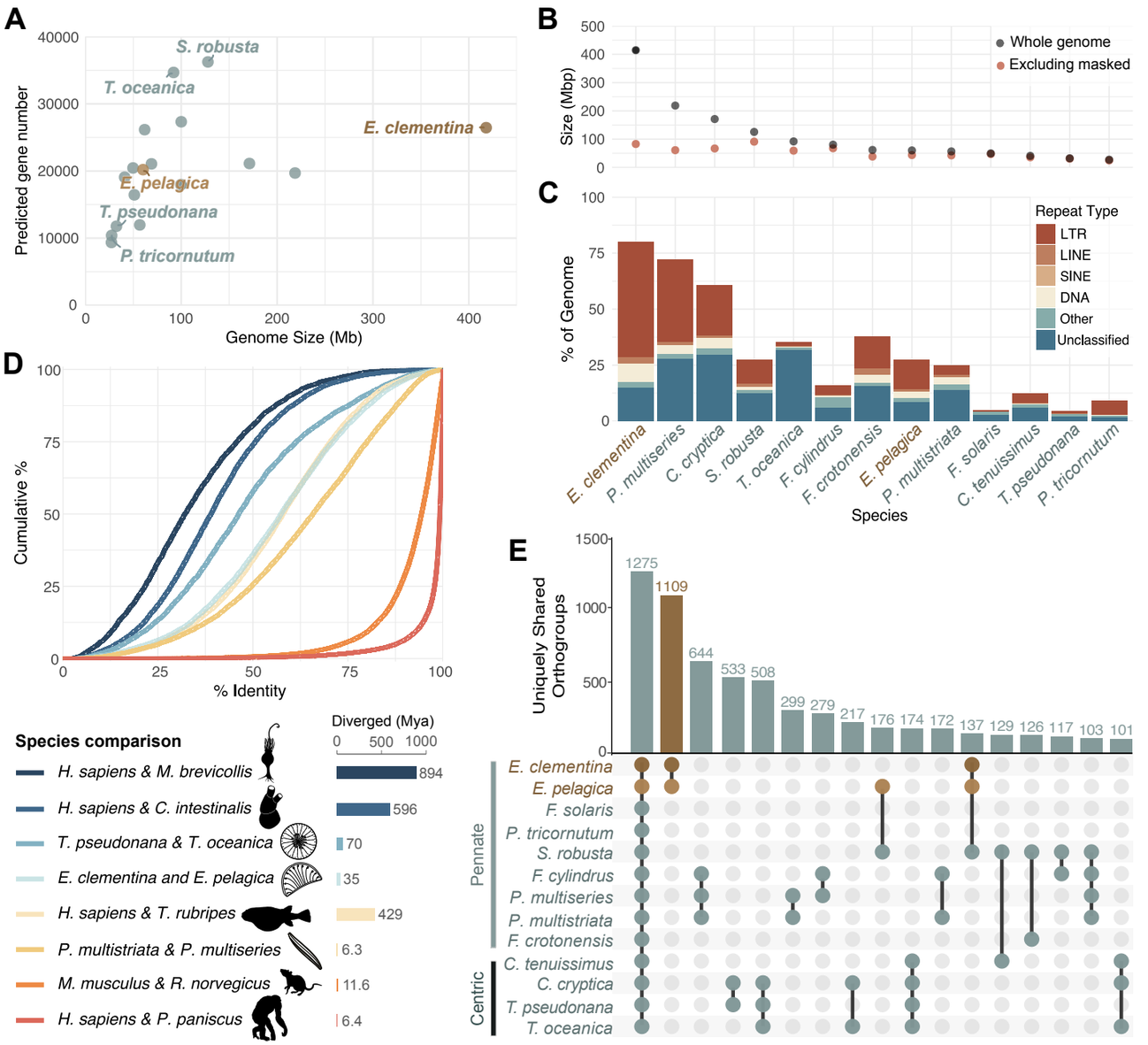

Figure 1. Highly divergent *E. clementina* and *E. pelagica* genomes share many unique gene families

(A) Genome size and total gene number for published diatom genomes compared with *Epithemia* species (dark blue). (See also Figure S1, Table S1.)

(B) Comparison of repeat content in diatom genomes showing size of the whole genome (grey dots) or the genome excluding repeat elements (orange dots). X-axis is the same as 1C.

(C) Breakdown of repeat types in diatom genomes showing amount in Mbp of the genome occupied by repeat elements of specific class, indicated by color. (See also Table S2.)

(D) Cumulative distribution of amino acid identity between pairwise orthologs from reference species. Estimated divergence time of species pair is indicated (right bar graph). (See also Figure S2.)

(E) UpSet plot depicting the number of uniquely shared orthogroups between all diatom species (first column) or subsets of 2-4 species. Orthogroups shared by *E. pelagica* and *E. clementina* are highlighted in brown. Columns are ranked by the number of uniquely shared orthogroups. (See also Figure S2.)

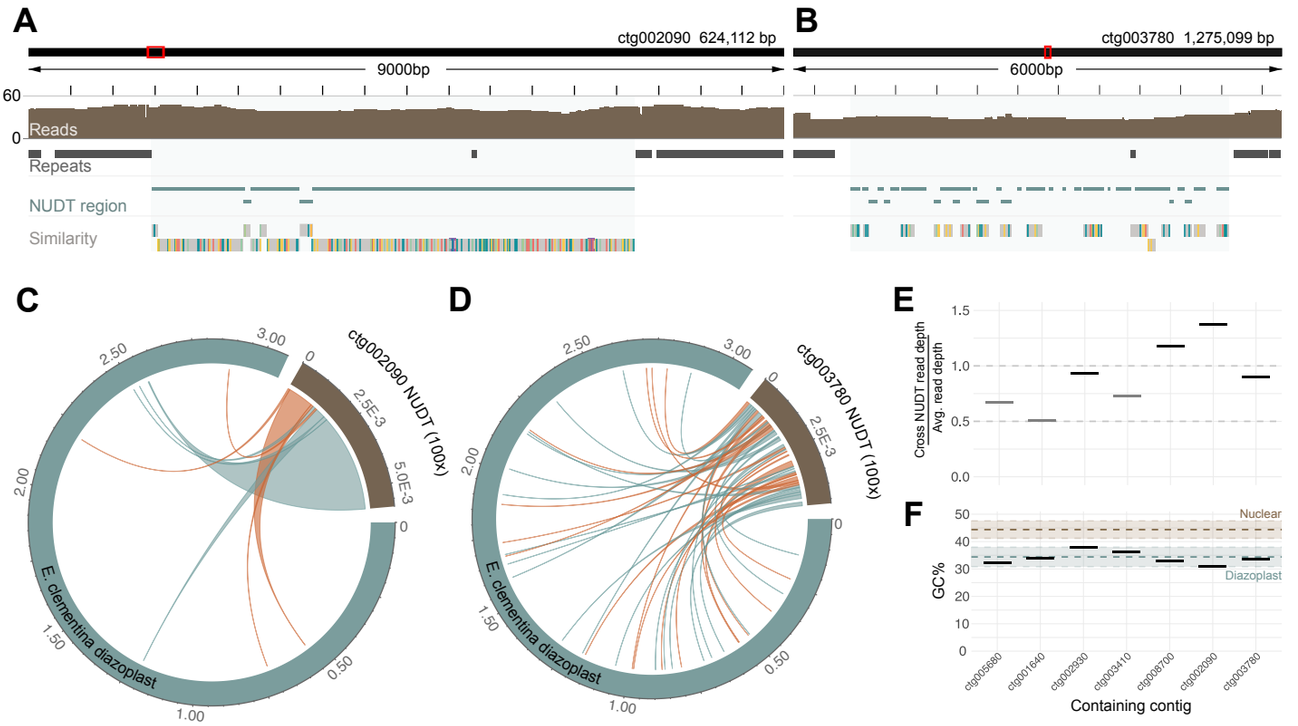

Figure 2. Detection of nuclear integrations of diazoplast DNA (NUDTs)

(A) A representative, NUDT containing *E. clementina* nuclear genome locus on contig ctg002090. Tracks shown from top to bottom: nuclear sub-region being viewed (red box) within the contig (black rectangle); length of the sub-region, with ticks every 500bp; nanopore sequencing read pileup, showing long read support across the NUDT; location of repeat masked regions (dark grey bars); locations of homology to *E. clementina* diazoplast identified by BLAST, demarcating the NUDT (blue shade); regions of homology to the E. clementina diazoplast identified by minimap2 alignment, colors represent SNVs between the diazoplast and nuclear sequence. (See also, Data S1F.)

(B) Same as A, for the NUDT on contig ctg003780. (See also, Data S1G.)

(C) Circlize plot depicting the fragmentation and rearrangement of NUDTs. The diazoplast genome (blue) and the NUDT on contig ctg002090 (brown) with chords connecting source diazoplast regions to their corresponding nuclear region, inversions in red. The length of the NUDT is depicted at 100x true relative length for ease of visualization. (See also, Figure S3A-S3E.)

(D) Same as C, for the NUDT on contig ctg003780.

(E) Ratio of long read depth of NUDT compared to average read depth for the containing contig. Heterozygous insertions (light grey bars) show approximately 0.5x depth; homozygous insertions (black bars) show approximately 1.0x depth.

(F) GC content of NUDTs, compared to mean GC content for 5kb sliding windows of the diazoplast genome (blue dashed line) and the nuclear genome (brown dashed line). Shaded regions represent mean ± 1 SD.

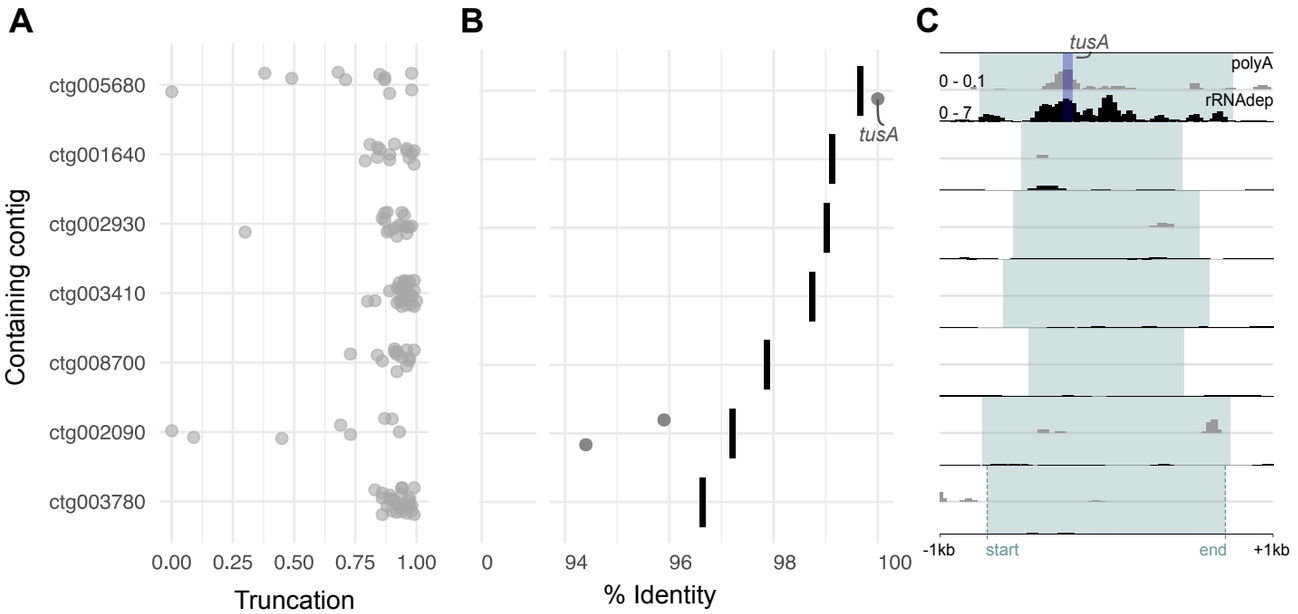

Figure 3. Most NUDTs are decaying and non-functional

(A) Truncation of diazoplast genes contained within each NUDT relative to the full-length diazoplast gene.

(B) Nucleotide identity of diazoplast genes that are <30% truncated (points) contained within each NUDT compared to identity of the full containing NUDT sequence (bars). (See also, Figures S3F and S3G.)

(C) Normalized expression across each NUDT (blue highlight) +/- 1kb of the genomic region surrounding the NUDT. For each NUDT, a pair of tracks shows RNA-seq reads after polyA enrichment of whole RNA plotted within background signal range, from 0 - 0.1 BPM (top, grey) and RNA-seq reads after rRNA depletion of whole RNA, plotted from 0 - 7 BPM (bottom, black). The region corresponding to the *tusA* gene in ctg005680 is highlighted in dark blue (See also Data S1A-S1G.)

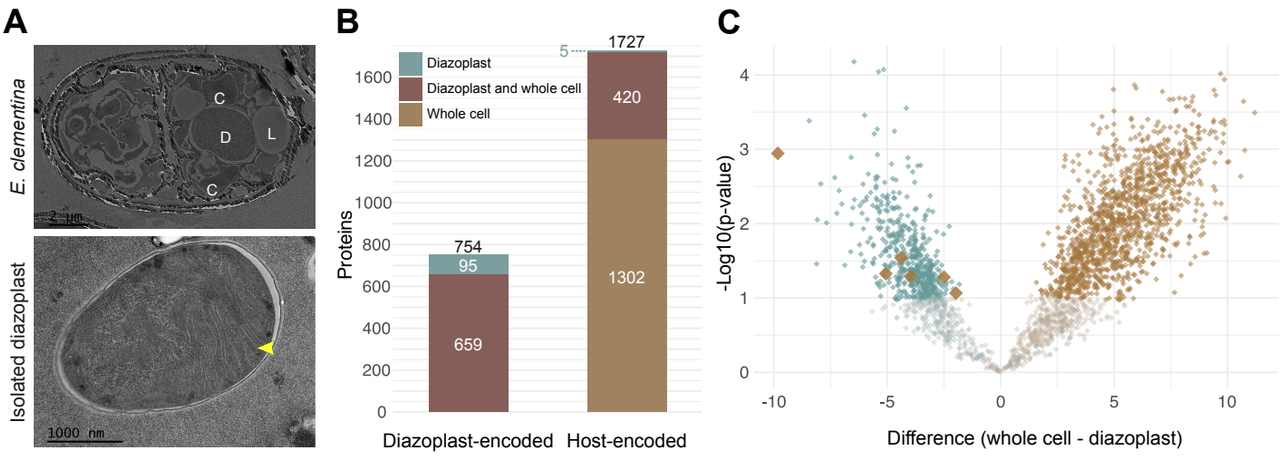

Figure 4. Few host-encoded proteins are detected in the diazoplast proteome
(A) Electron micrographs of (top) *E. clementina* cells with diazoplast (D), chloroplast lobes (C), and lipid bodies (L) indicated and (bottom) diazoplasts following purification with thylakoids (yellow arrow) indicated.
(B) Number of diazoplast-encoded (left) and host-encoded (right) proteins identified by LC-MS/MS. Total number of proteins identified from each respective proteome is shown above each stacked bar. Colored bars and numbers indicate proteins identified in purified diazoplasts only, whole cells only, or both.
(C) Volcano plot showing the enrichment of diazoplast-encoded (blue) and host-encoded (brown) proteins in whole cells or purified diazoplasts, represented by the difference between log2-transformed iBAQ values. Proteins enriched in the diazoplast are on the left side of the graph while those enriched in the host are on the right; the darker shade of each color represents significantly enriched hits. Host-encoded proteins significantly enriched in the diazoplast are shown with larger brown markers.

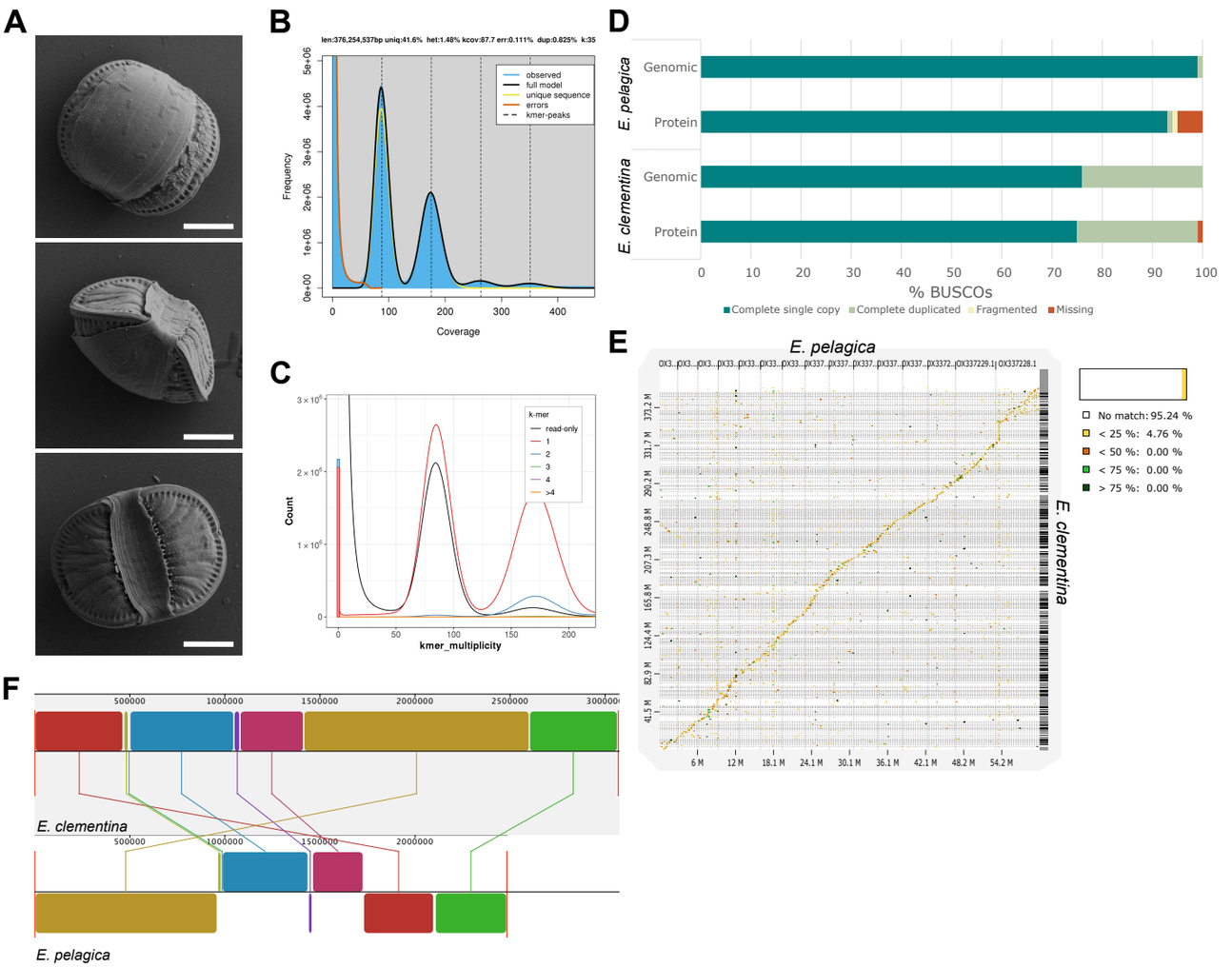

Figure S1. *E. clementina* genome assembly statistics and features

(A) Scanning electron micrographs of *E. clementina*, scale bar 5µm. Top: View looking down on the dorsal girdle band. Middle: View down the apical axis. Bottom: View of the ventral face, lined by prominent fenestral bars regularly spaced between the radial striae. The raphe lies along the strongly curved keel on the ventral margin and pinches slightly towards the dorsal margin.

(B) GenomeScope spectrum of 35-mer multiplicity collected from the Illumina sequencing reads. Peak at 1x coverage (~90) and 2x coverage (~180), consistent with a diploid genome.

(C) Merqury spectrum of k-mer multiplicity collected from the Illumina sequencing reads, stacked lines colored by number of times k-mer is seen in the genome assembly. Few k-mers within the heterozygous and homozygous peaks are read-only (black), suggesting that the assembly is not missing significant sequence present in the reads.

(D) Stramenopile-specific Benchmarking Universal Single-Copy Orthologs (BUSCOs) for *E. pelagica* and *E. clementina* genomes and proteomes. Both genomes contain all stramenopile BUSCOs, however the *E. pelagica* annotation is less complete. The genome and proteome of *E. clementina* show some duplication. This duplication of stramenopile BUSCOs is similar or less than that found in previously published genomes of diatoms, for example, *Nitzschia putrida* (GCA_016586335, 39% duplicate BUSCOs), *Fragilariopsis cylindrus* (GCA_900095095.1, 40% duplicate BUSCOs), and *Fragilaria radians* (GCA_900642245.1, 19% duplicate BUSCOs).

(E) Whole genome alignment of the *E. clementina* and *E. pelagica* genome assemblies. White indicates no sequence homology, yellow indicates alignments at <25% nucleotide identity. There is only 4.76% sequence homology between the two genomes at the nucleotide level, all at <25% identity.

(F) Genomic synteny between the whole genome alignments of the *E. clementina* and *E. pelagica* diazoplasts, showing 7 syntenic blocks.

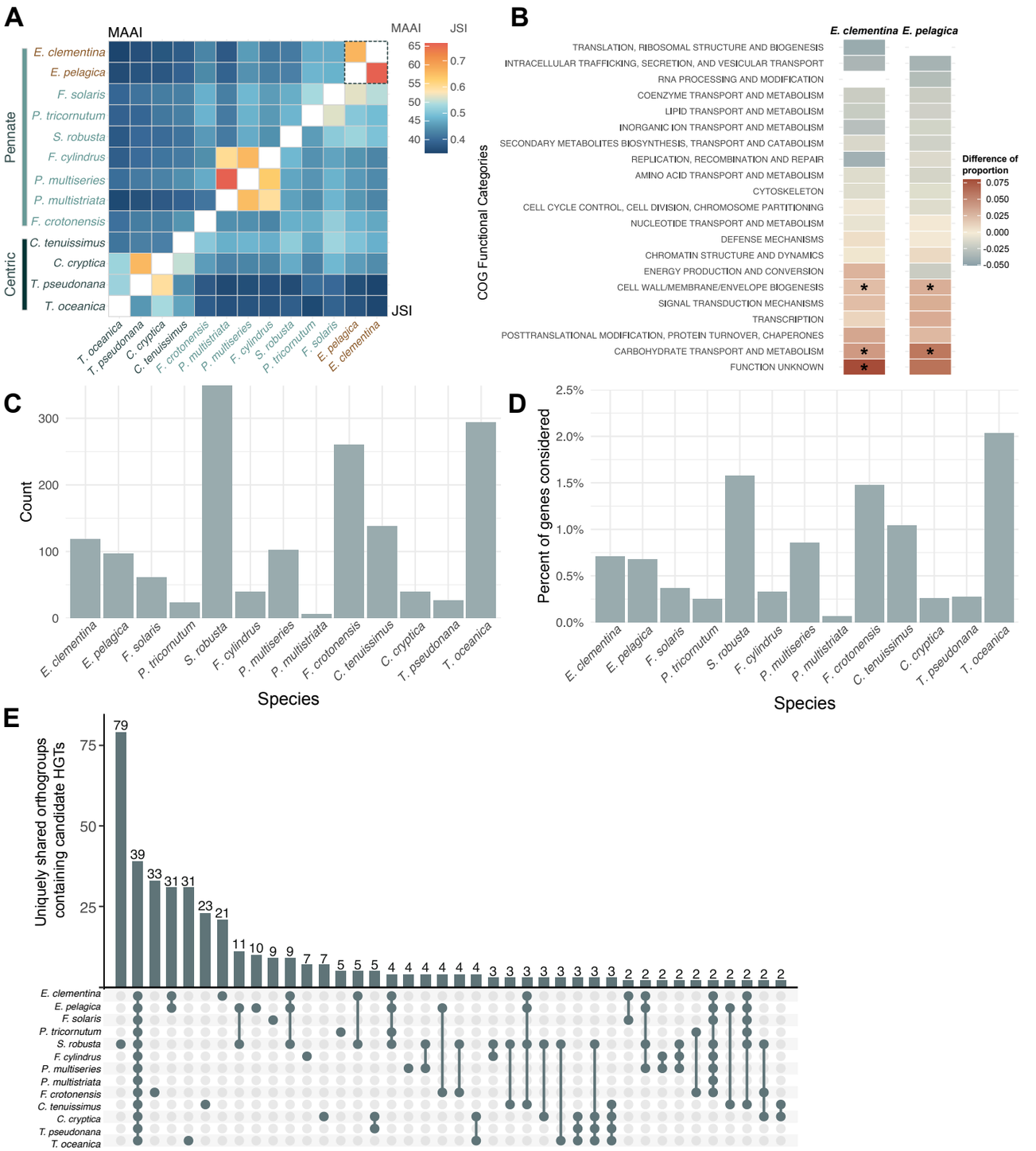

Figure S2. Gene family enrichment and HGT statistics

(A) Asymmetrical heatmap of ortholog comparisons between diatom species pairs, showing mean amino acid identity (MAAI) of pairwise orthologs (top) and Jaccard similarity index (JSI) of orthogroups (bottom).

(B) Enrichment analysis of clusters of orthologous group (COG) categories in *E. clementina* and *E. pelagica*. Color indicates difference in proportion of each COG category within genes in *Epithemia* unique gene families and the whole functionally annotated gene set. An asterisk indicates categories that were found to be statistically significantly enriched.

(C) Total count of HGT candidate genes across diatom species, considering only high-confidence candidates.

(D) Proportion of the genes considered for HGT analysis that were identified as high-confidence HGT candidates. Genes were considered for analysis if they had >10 identifiable homologues.

(E) UpSet plot depicting the number of uniquely shared, HGT candidate containing orthogroups between diatom species groupings (including singlets). Columns are ranked by the number of uniquely shared orthogroups.

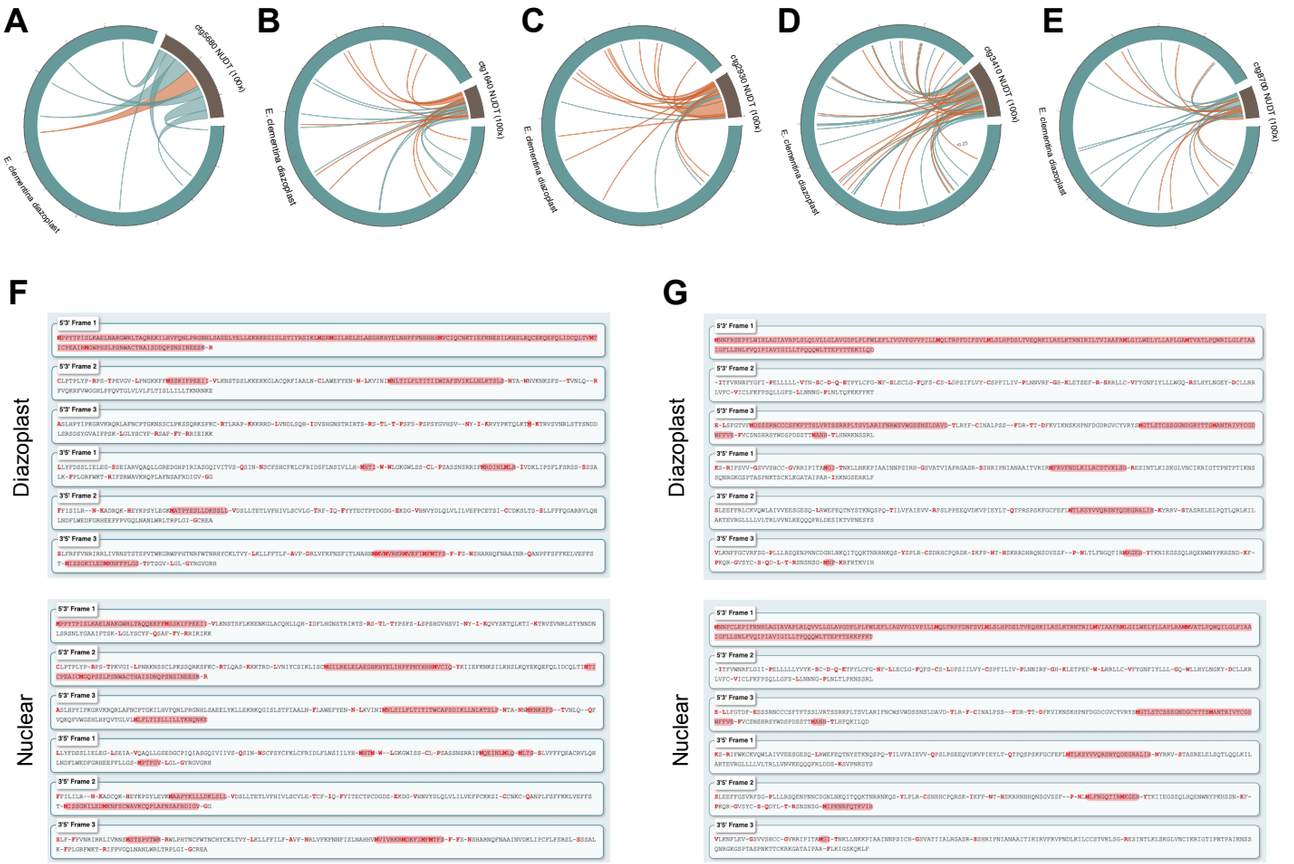

Figure S3. NUDT fragmentation and gene containing regions

(A-E) Circlize plots depicting the fragmentation and rearrangement of the NUDTs. The diazoplast genome (blue) and the NUDT on labeled contig (brown) with chords connecting source diazoplast regions to their corresponding nuclear region, inversions in red. The length of the NUDT is depicted at 100x true relative length for ease of visualization.

(F) Translation in all potential frames of the gene contained within the NUDT on contig ctg002090 (transcriptional repressor, gene ID: P3f56_RS08570). The copy within the NUDT (bottom) is untruncated (100% of the full-length gene) by nucleotide sequence and is 96% identical to the corresponding diazoplast gene (top). Compared to the diazoplast-encoded gene, the gene contained in the NUDT has a mutation that results in a premature stop codon at amino acid 39 (out of 177). Red highlight indicates a potential translation. 5’3’ Frame 1 is the native diazoplast frame.

(G) Translation in all potential frames of the gene contained within the NUDT on contig ctg002090 (low-complexity tail membrane protein, gene ID: P3F56_RS01750). The gene is 9% truncated at the 3’ terminus (91% of the full-length gene). Compared to the diazoplast-encoded gene, the gene contained in the NUDT has several non-synonymous mutations and is missing 16 amino acids at the C-terminus. Red highlight indicates a potential translation. 5’3’ Frame 1 is the native diazoplast frame.

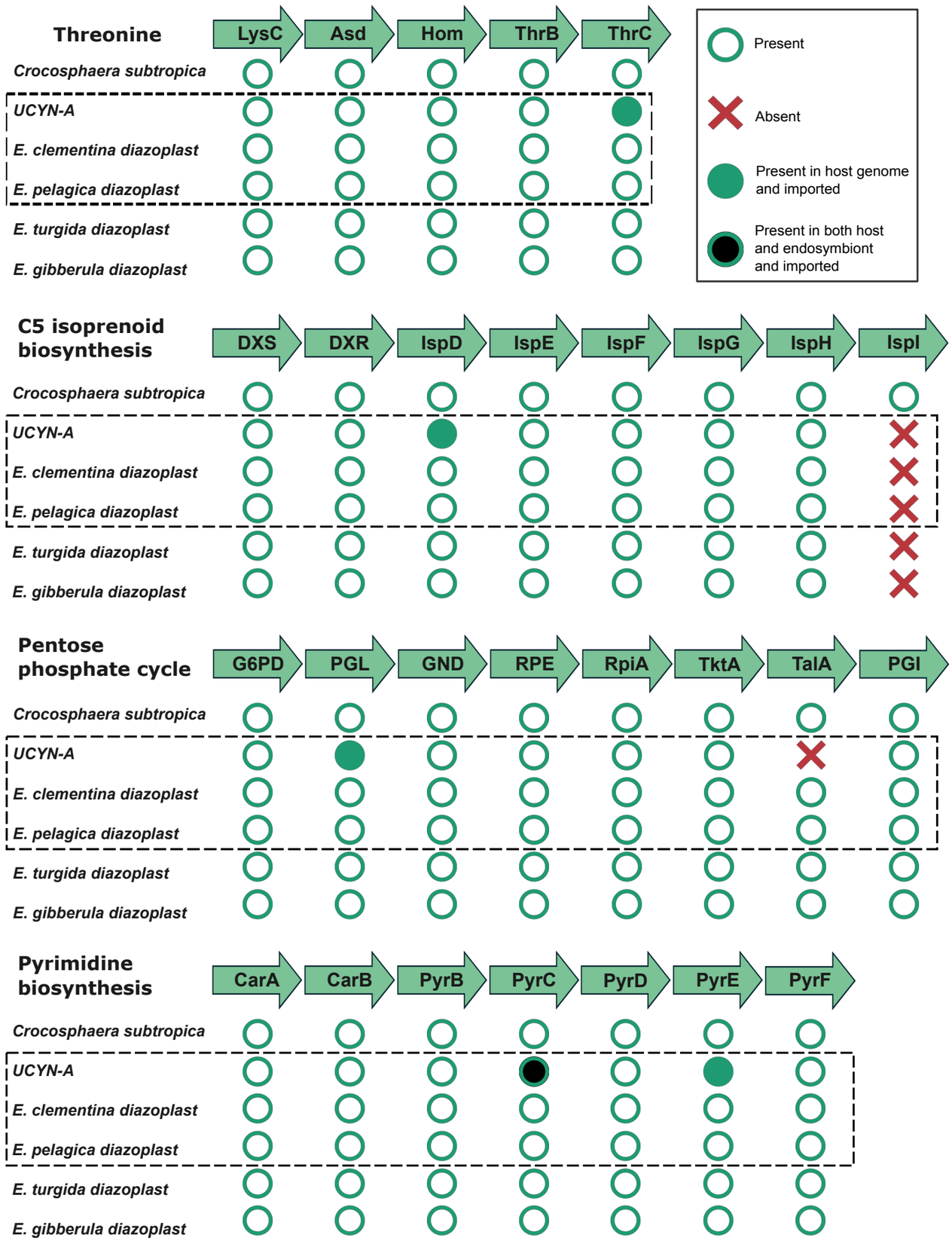

Figure S4. Comparative pathway analysis of diazoplasts and close relatives
KEGG pathway analysis of *E. clementina, E. turgida, E. gibberula* diazoplasts as well as *C. subtropica* and UCYN-A, indicating presence (green circle) or absence (red x) in the genome. Filled green circle indicates evidence for import of a host-encoded protein; filled black circle indicates presence in the endosymbiont genome and evidence for import of a host-encoded protein. A black dashed box indicates species for which data exists for both endosymbiont and host nuclear genes.

REFERENCES

1. Archibald, J.M. (2015). Endosymbiosis and Eukaryotic Cell Evolution. Curr. Biol. *25*, R911–R921. https://doi.org/10.1016/j.cub.2015.07.055.

2. Cavalier-Smith, T., and Lee, J.J. (1985). Protozoa as Hosts for Endosymbioses and the Conversion of Symbionts into Organelles,. J. Protozool. *32*, 376–379. https://doi.org/10.1111/j.1550-7408.1985.tb04031.x.

3. Cavalier-Smith, T. (1982). The origins of plastids. Biol. J. Linn. Soc. *17*, 289–306. https://doi.org/10.1111/j.1095-8312.1982.tb02023.x.

4. Martin, W., and Herrmann, R.G. (1998). Gene Transfer from Organelles to the Nucleus: How Much, What Happens, and Why? Plant Physiol. *118*, 9–17. https://doi.org/10.1104/pp.118.1.9.

5. Theissen, U., and Martin, W. (2006). The difference between organelles and endosymbionts. Curr. Biol. *16*, R1016–R1017. https://doi.org/10.1016/j.cub.2006.11.020.

6. Marin, B., M. Nowack, E.C., and Melkonian, M. (2005). A Plastid in the Making: Evidence for a Second Primary Endosymbiosis. Protist *156*, 425–432. https://doi.org/10.1016/j.protis.2005.09.001.

7. Nakayama, T., and Ishida, K. (2009). Another acquisition of a primary photosynthetic organelle is underway in *Paulinella chromatophora*. Curr. Biol. *19*, R284–R285. https://doi.org/10.1016/j.cub.2009.02.043.

8. Nowack, E.C.M., and Grossman, A.R. (2012). Trafficking of protein into the recently established photosynthetic organelles of *Paulinella chromatophora*. Proc. Natl. Acad. Sci. *109*, 5340–5345. https://doi.org/10.1073/pnas.1118800109.

9. Singer, A., Poschmann, G., Mühlich, C., Valadez-Cano, C., Hänsch, S., Hüren, V., Rensing, S.A., Stühler, K., and Nowack, E.C.M. (2017). Massive Protein Import into the Early-Evolutionary-Stage Photosynthetic Organelle of the Amoeba *Paulinella chromatophora*. Curr. Biol. *27*, 2763-2773.e5. https://doi.org/10.1016/j.cub.2017.08.010.

10. Coale, T.H., Loconte, V., Turk-Kubo, K.A., Vanslembrouck, B., Mak, W.K.E., Cheung, S., Ekman, A., Chen, J.-H., Hagino, K., Takano, Y., et al. (2024). Nitrogen-fixing organelle in a marine alga. Science *384*, 217–222. https://doi.org/10.1126/science.adk1075.

11. Morales, J., Ehret, G., Poschmann, G., Reinicke, T., Maurya, A.K., Kröninger, L., Zanini, D., Wolters, R., Kalyanaraman, D., Krakovka, M., et al. (2023). Host-symbiont interactions in *Angomonas deanei* include the evolution of a host-derived dynamin ring around the endosymbiont division site. Curr. Biol. CB *33*, 28-40.e7. https://doi.org/10.1016/j.cub.2022.11.020.

12. McCutcheon, J.P., Boyd, B.M., and Dale, C. (2019). The Life of an Insect Endosymbiont from the Cradle to the Grave. Curr. Biol. *29*, R485–R495. https://doi.org/10.1016/j.cub.2019.03.032.

13. Nowack, E.C.M. (2014). *Paulinella chromatophora* – rethinking the transition from endosymbiont to organelle. Acta Soc. Bot. Pol. *83*, 387–397. https://doi.org/10.5586/asbp.2014.049.

14. Keeling, P.J., McCutcheon, J.P., and Doolittle, W.F. (2015). Symbiosis becoming permanent: Survival of the luckiest. Proc. Natl. Acad. Sci. *112*, 10101–10103. https://doi.org/10.1073/pnas.1513346112.

15. Nowack, E.C.M., Price, D.C., Bhattacharya, D., Singer, A., Melkonian, M., and Grossman, A.R. (2016). Gene transfers from diverse bacteria compensate for reductive genome evolution in the

chromatophore of *Paulinella chromatophora*. Proc. Natl. Acad. Sci. *113*, 12214–12219. https://doi.org/10.1073/pnas.1608016113.

16. Ponce-Toledo, R.I., López-García, P., and Moreira, D. (2019). Horizontal and endosymbiotic gene transfer in early plastid evolution. New Phytol. *224*, 618–624. https://doi.org/10.1111/nph.15965.

17. Suzuki, S., Kawachi, M., Tsukakoshi, C., Nakamura, A., Hagino, K., Inouye, I., and Ishida, K. (2021). Unstable Relationship Between *Braarudosphaera bigelowii* (= *Chrysochromulina parkeae*) and Its Nitrogen-Fixing Endosymbiont. Front. Plant Sci. *12*. https://doi.org/10.3389/fpls.2021.749895.

18. Sørensen, M.E.S., Zlatogursky, V.V., Onuţ-Brännström, I., Walraven, A., Foster, R.A., and Burki, F. (2023). A novel kleptoplastidic symbiosis revealed in the marine centrohelid *Meringosphaera* with evidence of genetic integration. Curr. Biol. *33*, 3571-3584.e6. https://doi.org/10.1016/j.cub.2023.07.017.

19. Hehenberger, E., Gast, R.J., and Keeling, P.J. (2019). A kleptoplastidic dinoflagellate and the tipping point between transient and fully integrated plastid endosymbiosis. Proc. Natl. Acad. Sci. *116*, 17934–17942. https://doi.org/10.1073/pnas.1910121116.

20. Larkum, A.W.D., Lockhart, P.J., and Howe, C.J. (2007). Shopping for plastids. Trends Plant Sci. *12*, 189–195. https://doi.org/10.1016/j.tplants.2007.03.011.

21. Keeling, P.J. (2013). The Number, Speed, and Impact of Plastid Endosymbioses in Eukaryotic Evolution. Annu. Rev. Plant Biol. *64*, 583–607. https://doi.org/10.1146/annurev-arplant-050312-120144.

22. Sibbald, S.J., and Archibald, J.M. (2020). Genomic Insights into Plastid Evolution. Genome Biol. Evol. *12*, 978–990. https://doi.org/10.1093/gbe/evaa096.

23. Pfitzer, E. (1871). Untersuchungen über Bau und Entwicklung der Bacillariaceen (Diatomaceen) (A. Marcus).

24. Drum, R.W., and Pankratz, S. (1965). Fine structure of an unusual cytoplasmic inclusion in the diatom genus *Rhopalodia*. Protoplasma *60*, 141–149. https://doi.org/10.1007/BF01248136.

25. DeYoe, H.R., Lowe, R.L., and Marks, J.C. (1992). Effects of Nitrogen and Phosphorus on the Endosymbiont Load of *Rhopalodia gibba* and *Epithemia turgida* (bacillariophyceae). J. Phycol. *28*, 773–777. https://doi.org/10.1111/j.0022-3646.1992.00773.x.

26. Prechtl, J., Kneip, C., Lockhart, P., Wenderoth, K., and Maier, U.-G. (2004). Intracellular Spheroid Bodies of Rhopalodia gibba Have Nitrogen-Fixing Apparatus of Cyanobacterial Origin. Mol. Biol. Evol. *21*, 1477–1481. https://doi.org/10.1093/molbev/msh086.

27. Nakayama, T., Kamikawa, R., Tanifuji, G., Kashiyama, Y., Ohkouchi, N., Archibald, J.M., and Inagaki, Y. (2014). Complete genome of a nonphotosynthetic cyanobacterium in a diatom reveals recent adaptations to an intracellular lifestyle. Proc. Natl. Acad. Sci. *111*, 11407–11412. https://doi.org/10.1073/pnas.1405222111.

28. Moulin, S.L.Y., Frail, S., Braukmann, T., Doenier, J., Steele-Ogus, M., Marks, J.C., Mills, M.M., and Yeh, E. (2024). The endosymbiont of *Epithemia clementina* is specialized for nitrogen fixation within a photosynthetic eukaryote. ISME Commun. *4*, ycae055. https://doi.org/10.1093/ismeco/ycae055.

1145 29. Ruck, E.C., Nakov, T., Alverson, A.J., and Theriot, E.C. (2016). Phylogeny, ecology, morphological
1146 evolution, and reclassification of the diatom orders Surirellales and Rhopalodiales. Mol. Phylogenet.
1147 Evol. *103*, 155–171. https://doi.org/10.1016/j.ympev.2016.07.023.

1148 30. Schvarcz, C.R., Wilson, S.T., Caffin, M., Stancheva, R., Li, Q., Turk-Kubo, K.A., White, A.E., Karl,
1149 D.M., Zehr, J.P., and Steward, G.F. (2022). Overlooked and widespread pennate diatom-diazotroph
1150 symbioses in the sea. Nat. Commun. *13*, 799. https://doi.org/10.1038/s41467-022-28065-6.

1151 31. Foster, R.A., Subramaniam, A., Mahaffey, C., Carpenter, E.J., Capone, D.G., and Zehr, J.P. (2007).
1152 Influence of the Amazon River plume on distributions of free-living and symbiotic cyanobacteria in
1153 the western tropical north Atlantic Ocean. Limnol. Oceanogr. *52*, 517–532.
1154 https://doi.org/10.4319/lo.2007.52.2.0517.

1155 32. Benson, M.E., Kociolek, P.J., Spaulding, S.A., and Smith, D.M. (2012). Pre-Neogene non-marine
1156 diatom biochronology with new data from the late Eocene Florissant Formation of Colorado, USA.
1157 Stratigraphy *9*, 121–152.

1158 33. Abresch, H., Bell, T., and Miller, S.R. (2024). Diurnal transcriptional variation is reduced in a
1159 nitrogen-fixing diatom endosymbiont. ISME J. *18*, wrae064. https://doi.org/10.1093/ismejo/wrae064.

1160 34. Keeling, P.J., and Archibald, J.M. (2008). Organelle Evolution: What's in a Name? Curr. Biol. *18*,
1161 R345–R347. https://doi.org/10.1016/j.cub.2008.02.065.

1162 35. Nakayama, T., and Inagaki, Y. (2017). Genomic divergence within non-photosynthetic
1163 cyanobacterial endosymbionts in rhopalodiacean diatoms. Sci. Rep. *7*, 13075.
1164 https://doi.org/10.1038/s41598-017-13578-8.

1165 36. Muñoz-Marín, M. del C., Shilova, I.N., Shi, T., Farnelid, H., Cabello, A.M., and Zehr, J.P. (2019).
1166 The Transcriptional Cycle Is Suited to Daytime N2 Fixation in the Unicellular Cyanobacterium
1167 "*Candidatus* Atelocyanobacterium thalassa" (UCYN-A). mBio *10*, 10.1128/mbio.02495-18.
1168 https://doi.org/10.1128/mbio.02495-18.

1169 37. Landa, M., Turk-Kubo, K.A., Cornejo-Castillo, F.M., Henke, B.A., and Zehr, J.P. (2021). Critical Role
1170 of Light in the Growth and Activity of the Marine N2-Fixing UCYN-A Symbiosis. Front. Microbiol. *12*.
1171 https://doi.org/10.3389/fmicb.2021.666739.

1172 38. Kamakura, S., Mann, D.G., Nakamura, N., and Sato, S. (2021). Inheritance of spheroid body and
1173 plastid in the raphid diatom *Epithemia* (Bacillariophyta) during sexual reproduction. Phycologia *60*,
1174 265–273. https://doi.org/10.1080/00318884.2021.1909399.

1175 39. Nieves-Morión, M., Camargo, S., Bardi, S., Ruiz, M.T., Flores, E., and Foster, R.A. (2023).
1176 Heterologous expression of genes from a cyanobacterial endosymbiont highlights substrate
1177 exchanges with its diatom host. PNAS Nexus *2*, pgad194.
1178 https://doi.org/10.1093/pnasnexus/pgad194.

1179 40. Gao, H., Kadirjan-Kalbach, D., Froehlich, J.E., and Osteryoung, K.W. (2003). ARC5, a cytosolic
1180 dynamin-like protein from plants, is part of the chloroplast division machinery. Proc. Natl. Acad. Sci.
1181 *100*, 4328–4333. https://doi.org/10.1073/pnas.0530206100.

1182 41. Zakharova, A., Tashyreva, D., Butenko, A., Morales, J., Saura, A., Svobodová, M., Poschmann, G.,
1183 Nandipati, S., Zakharova, A., Noyvert, D., et al. (2023). A neo-functionalized homolog of host
1184 transmembrane protein controls localization of bacterial endosymbionts in the trypanosomatid
1185 *Novymonas esmeraldas*. Curr. Biol. *33*, 2690-2701.e5. https://doi.org/10.1016/j.cub.2023.04.060.

1186 42. Schvarcz, C.R., Stancheva, R., Turk-Kubo, K.A., Wilson, S.T., Zehr, J.P., Edwards, K.F., Steward,
1187 G.F., Archibald, J.M., Oatley, G., Sinclair, E., et al. (2024). The genome sequences of the marine
1188 diatom *Epithemia pelagica* strain UHM3201 (Schvarcz, Stancheva & Steward, 2022) and its
1189 nitrogen-fixing, endosymbiotic cyanobacterium. Wellcome Open Res. *9*, 232.
1190 https://doi.org/10.12688/wellcomeopenres.21534.1.

1191 43. Kumar, S., Suleski, M., Craig, J.M., Kasprowicz, A.E., Sanderford, M., Li, M., Stecher, G., and
1192 Hedges, S.B. (2022). TimeTree 5: An Expanded Resource for Species Divergence Times. Mol. Biol.
1193 Evol. *39*, msac174. https://doi.org/10.1093/molbev/msac174.

1194 44. Kooistra, W.H.C.F., Gersonde, R., Medlin, L.K., and Mann, D.G. (2007). The Origin and Evolution of
1195 the Diatoms: Their Adaptation to a Planktonic Existence. In Evolution of Primary Producers in the
1196 Sea, P. G. Falkowski and A. H. Knoll, eds. (Academic Press), pp. 207–249.
1197 https://doi.org/10.1016/B978-012370518-1/50012-6.

1198 45. Bowler, C., Allen, A.E., Badger, J.H., Grimwood, J., Jabbari, K., Kuo, A., Maheswari, U., Martens,
1199 C., Maumus, F., Otillar, R.P., et al. (2008). The *Phaeodactylum* genome reveals the evolutionary
1200 history of diatom genomes. Nature *456*, 239–244. https://doi.org/10.1038/nature07410.

1201 46. Vancaester, E., Depuydt, T., Osuna-Cruz, C.M., and Vandepoele, K. (2020). Comprehensive and
1202 Functional Analysis of Horizontal Gene Transfer Events in Diatoms. Mol. Biol. Evol. *37*, 3243–3257.
1203 https://doi.org/10.1093/molbev/msaa182.

1204 47. Van Etten, J., and Bhattacharya, D. (2020). Horizontal Gene Transfer in Eukaryotes: Not if, but How
1205 Much? Trends Genet. *36*, 915–925. https://doi.org/10.1016/j.tig.2020.08.006.

1206 48. Lopez, J.V., Yuhki, N., Masuda, R., Modi, W., and O'Brien, S.J. (1994). Numt, a recent transfer and
1207 tandem amplification of mitochondrial DNA to the nuclear genome of the domestic cat. J. Mol. Evol.
1208 *39*, 174–190. https://doi.org/10.1007/BF00163806.

1209 49. Timmis, J.N., Ayliffe, M.A., Huang, C.Y., and Martin, W. (2004). Endosymbiotic gene transfer:
1210 organelle genomes forge eukaryotic chromosomes. Nat. Rev. Genet. *5*, 123–135.
1211 https://doi.org/10.1038/nrg1271.

1212 50. Patron, N.J., and Waller, R.F. (2007). Transit peptide diversity and divergence: A global analysis of
1213 plastid targeting signals. BioEssays *29*, 1048–1058. https://doi.org/10.1002/bies.20638.

1214 51. Oberleitner, L., Perrar, A., Macorano, L., Huesgen, P.F., and Nowack, E.C.M. (2022). A bipartite
1215 chromatophore transit peptide and N-terminal protein processing in the *Paulinella* chromatophore.
1216 Plant Physiol. *189*, 152–164. https://doi.org/10.1093/plphys/kiac012.

1217 52. Paysan-Lafosse, T., Blum, M., Chuguransky, S., Grego, T., Pinto, B.L., Salazar, G.A., Bileschi, M.L.,
1218 Bork, P., Bridge, A., Colwell, L., et al. (2023). InterPro in 2022. Nucleic Acids Res. *51*, D418–D427.
1219 https://doi.org/10.1093/nar/gkac993.

1220 53. Keeling, P.J. (2024). Horizontal gene transfer in eukaryotes: aligning theory with data. Nat. Rev.
1221 Genet. *25*, 416–430. https://doi.org/10.1038/s41576-023-00688-5.

1222 54. Muller, H.J. (1964). The relation of recombination to mutational advance. Mutat. Res. *106*, 2–9.
1223 https://doi.org/10.1016/0027-5107(64)90047-8.

1224 55. Allen, J.M., Light, J.E., Perotti, M.A., Braig, H.R., and Reed, D.L. (2009). Mutational Meltdown in
1225 Primary Endosymbionts: Selection Limits Muller's Ratchet. PLOS ONE *4*, e4969.
1226 https://doi.org/10.1371/journal.pone.0004969.

1227 56. Lhee, D., Lee, J., Ettahi, K., Cho, C.H., Ha, J.-S., Chan, Y.-F., Zelzion, U., Stephens, T.G., Price,
1228 D.C., Gabr, A., et al. (2020). Amoeba Genome Reveals Dominant Host Contribution to Plastid
1229 Endosymbiosis. Mol. Biol. Evol. *38*, 344–357. https://doi.org/10.1093/molbev/msaa206.

1230 57. Smith, D.R., Crosby, K., and Lee, R.W. (2011). Correlation between Nuclear Plastid DNA
1231 Abundance and Plastid Number Supports the Limited Transfer Window Hypothesis. Genome Biol.
1232 Evol. *3*, 365–371. https://doi.org/10.1093/gbe/evr001.

1233 58. Barbrook, A.C., Howe, C.J., and Purton, S. (2006). Why are plastid genomes retained in non-
1234 photosynthetic organisms? Trends Plant Sci. *11*, 101–108.
1235 https://doi.org/10.1016/j.tplants.2005.12.004.

1236 59. Lister, D.L., Bateman, J.M., Purton, S., and Howe, C.J. (2003). DNA transfer from chloroplast to
1237 nucleus is much rarer in *Chlamydomonas* than in tobacco. Gene *316*, 33–38.
1238 https://doi.org/10.1016/S0378-1119(03)00754-6.

1239 60. Tyra, H.M., Linka, M., Weber, A.P., and Bhattacharya, D. (2007). Host origin of plastid solute
1240 transporters in the first photosynthetic eukaryotes. Genome Biol. *8*, R212.
1241 https://doi.org/10.1186/gb-2007-8-10-r212.

1242 61. Foster, R.A., Kuypers, M.M.M., Vagner, T., Paerl, R.W., Musat, N., and Zehr, J.P. (2011). Nitrogen
1243 fixation and transfer in open ocean diatom–cyanobacterial symbioses. ISME J. *5*, 1484–1493.
1244 https://doi.org/10.1038/ismej.2011.26.

1245 62. Carpenter, E.J., and Janson, S. (2000). Intracellular cyanobacterial symbionts in the marine diatom
1246 *Climacodium frauenfeldianum* (bacillariophyceae). J. Phycol. *36*, 540–544.
1247 https://doi.org/10.1046/j.1529-8817.2000.99163.x.

1248 63. Ritchie, R.J. (2013). The ammonia transport, retention and futile cycling problem in cyanobacteria.
1249 Microb. Ecol. *65*, 180–196. https://doi.org/10.1007/s00248-012-0111-1.

1250 64. Nieves-Morión, M., and Flores, E. (2018). Multiple ABC glucoside transporters mediate sugar-
1251 stimulated growth in the heterocyst-forming cyanobacterium *Anabaena* sp. strain PCC 7120.
1252 Environ. Microbiol. Rep. *10*, 40–48. https://doi.org/10.1111/1758-2229.12603.

1253 65. Muñoz-Marín, M. del C., Luque, I., Zubkov, M.V., Hill, P.G., Diez, J., and García-Fernández, J.M.
1254 (2013). Prochlorococcus can use the Pro1404 transporter to take up glucose at nanomolar
1255 concentrations in the Atlantic Ocean. Proc. Natl. Acad. Sci. *110*, 8597–8602.
1256 https://doi.org/10.1073/pnas.1221775110.

1257 66. Leger, M.M., Petrů, M., Žárský, V., Eme, L., Vlček, Č., Harding, T., Lang, B.F., Eliáš, M., Doležal, P.,
1258 and Roger, A.J. (2015). An ancestral bacterial division system is widespread in eukaryotic
1259 mitochondria. Proc. Natl. Acad. Sci. *112*, 10239–10246. https://doi.org/10.1073/pnas.1421392112.

1260 67. Gast, R.J., Sanders, R.W., and Caron, D.A. (2009). Ecological strategies of protists and their
1261 symbiotic relationships with prokaryotic microbes. Trends Microbiol. *17*, 563–569.
1262 https://doi.org/10.1016/j.tim.2009.09.001.

1263 68. Esteves-Ferreira, A.A., Inaba, M., Obata, T., Fort, A., Fleming, G.T.A., Araújo, W.L., Fernie, A.R.,
1264 and Sulpice, R. (2017). A Novel Mechanism, Linked to Cell Density, Largely Controls Cell Division
1265 in *Synechocystis*. Plant Physiol. *174*, 2166–2182. https://doi.org/10.1104/pp.17.00729.

1266 69. Moore, K.A., Altus, S., Tay, J.W., Meehl, J.B., Johnson, E.B., Bortz, D.M., and Cameron, J.C.
1267 (2020). Mechanical regulation of photosynthesis in cyanobacteria. Nat. Microbiol. *5*, 757–767.
1268 https://doi.org/10.1038/s41564-020-0684-2.

1269 70. Liu, F., Fernie, A.R., and Zhang, Y. (2024). Can a nitrogen-fixing organelle be engineered within
1270 plants? Trends Plant Sci. https://doi.org/10.1016/j.tplants.2024.07.001.

1271 71. Elhai, J. (2023). Engineering of crop plants to facilitate bottom-up innovation: A possible role for
1272 broad host-range nitroplasts and neoplasts. Preprint at OSF, https://doi.org/10.31219/osf.io/ny2rc
1273 https://doi.org/10.31219/osf.io/ny2rc.

1274 72. Bombar, D., Heller, P., Sanchez-Baracaldo, P., Carter, B.J., and Zehr, J.P. (2014). Comparative
1275 genomics reveals surprising divergence of two closely related strains of uncultivated UCYN-A
1276 cyanobacteria. ISME J. *8*, 2530–2542. https://doi.org/10.1038/ismej.2014.167.

1277 73. McCutcheon, J.P., and Moran, N.A. (2012). Extreme genome reduction in symbiotic bacteria. Nat.
1278 Rev. Microbiol. *10*, 13–26. https://doi.org/10.1038/nrmicro2670.

1279 74. Stein-Taylor, J.R. and Phycological Society of America (1973). Handbook of Phycological Methods:
1280 Culture methods and growth measurements (University Press Cambridge [England]).

1281 75. Workman, Rachel, Timp, Winston, Fedak, Renee, Kilburn, Duncan, Hao, Stephanie, and Liu, Kelvin
1282 (2018). High Molecular Weight DNA Extraction from Recalcitrant Plant Species for Third Generation
1283 Sequencing. Protoc. Exch. https://doi.org/10.1038/protex.2018.059.

1284 76. Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor.
1285 Bioinformatics *34*, i884–i890. https://doi.org/10.1093/bioinformatics/bty560.

1286 77. Ranallo-Benavidez, T.R., Jaron, K.S., and Schatz, M.C. (2020). GenomeScope 2.0 and Smudgeplot
1287 for reference-free profiling of polyploid genomes. Nat. Commun. *11*, 1432.
1288 https://doi.org/10.1038/s41467-020-14998-3.

1289 78. Marçais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of
1290 occurrences of k-mers. Bioinformatics *27*, 764–770. https://doi.org/10.1093/bioinformatics/btr011.

1291 79. De Coster, W., D'Hert, S., Schultz, D.T., Cruts, M., and Van Broeckhoven, C. (2018). NanoPack:
1292 visualizing and processing long-read sequencing data. Bioinformatics *34*, 2666–2669.
1293 https://doi.org/10.1093/bioinformatics/bty149.

1294 80. Andrew, S. (2019). Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput
1295 Sequence Data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

1296 81. Hu, J., Wang, Z., Sun, Z., Hu, B., Ayoola, A.O., Liang, F., Li, J., Sandoval, J.R., Cooper, D.N., Ye,
1297 K., et al. (2024). NextDenovo: an efficient error correction and accurate assembly tool for noisy long
1298 reads. Genome Biol. *25*, 107. https://doi.org/10.1186/s13059-024-03252-4.

1299 82. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics *34*, 3094–
1300 3100. https://doi.org/10.1093/bioinformatics/bty191.

1301 83. Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.
1302 Preprint at arXiv.

1303 84. Vaser, R., Sović, I., Nagarajan, N., and Šikić, M. (2017). Fast and accurate de novo genome
1304 assembly from long uncorrected reads. Genome Res. *27*, 737–746.
1305 https://doi.org/10.1101/gr.214270.116.

1306 85. Zimin, A.V., Puiu, D., Luo, M.-C., Zhu, T., Koren, S., Marçais, G., Yorke, J.A., Dvořák, J., and
1307 Salzberg, S.L. (2017). Hybrid assembly of the large and highly repetitive genome of *Aegilops*

tauschii, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. Genome Res. *27*, 787–792. https://doi.org/10.1101/gr.213405.116.

86. Zimin, A.V., and Salzberg, S.L. (2020). The genome polishing tool POLCA makes fast and accurate corrections in genome assemblies. PLOS Comput. Biol. *16*, e1007981. https://doi.org/10.1371/journal.pcbi.1007981.

87. Laetsch, D.R., and Blaxter, M.L. (2017). BlobTools: Interrogation of genome assemblies. Preprint at F1000Research, https://doi.org/10.12688/f1000research.12232.1 https://doi.org/10.12688/f1000research.12232.1.

88. Mikheenko, A., Prjibelski, A., Saveliev, V., Antipov, D., and Gurevich, A. (2018). Versatile genome assembly evaluation with QUAST-LG. Bioinformatics *34*, i142–i150. https://doi.org/10.1093/bioinformatics/bty266.

89. Rhie, A., Walenz, B.P., Koren, S., and Phillippy, A.M. (2020). Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol. *21*, 245. https://doi.org/10.1186/s13059-020-02134-9.

90. Manni, M., Berkeley, M.R., Seppey, M., Simão, F.A., and Zdobnov, E.M. (2021). BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. Mol. Biol. Evol. *38*, 4647–4654. https://doi.org/10.1093/molbev/msab199.

91. Flynn, J.M., Hubley, R., Goubert, C., Rosen, J., Clark, A.G., Feschotte, C., and Smit, A.F. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. Proc. Natl. Acad. Sci. *117*, 9451–9457. https://doi.org/10.1073/pnas.1921046117.

92. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinformatics *10*, 421. https://doi.org/10.1186/1471-2105-10-421.

93. Sayers, E.W., Bolton, E.E., Brister, J.R., Canese, K., Chan, J., Comeau, D.C., Connor, R., Funk, K., Kelly, C., Kim, S., et al. (2022). Database resources of the national center for biotechnology information. Nucleic Acids Res. *50*, D20–D26. https://doi.org/10.1093/nar/gkab1112.

94. Mitra, R., Li, X., Kapusta, A., Mayhew, D., Mitra, R.D., Feschotte, C., and Craig, N.L. (2013). Functional characterization of piggyBat from the bat *Myotis lucifugus* unveils an active mammalian DNA transposon. Proc. Natl. Acad. Sci. *110*, 234–239. https://doi.org/10.1073/pnas.1217548110.

95. Kriventseva, E.V., Kuznetsov, D., Tegenfeldt, F., Manni, M., Dias, R., Simão, F.A., and Zdobnov, E.M. (2019). OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. Nucleic Acids Res. *47*, D807–D811. https://doi.org/10.1093/nar/gky1053.

96. Stanke, M., Diekhans, M., Baertsch, R., and Haussler, D. (2008). Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics *24*, 637–644. https://doi.org/10.1093/bioinformatics/btn013.

97. Stanke, M., Schöffmann, O., Morgenstern, B., and Waack, S. (2006). Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinformatics *7*, 62. https://doi.org/10.1186/1471-2105-7-62.

98. Gabriel, L., Hoff, K.J., Brůna, T., Borodovsky, M., and Stanke, M. (2021). TSEBRA: transcript selector for BRAKER. BMC Bioinformatics *22*, 566. https://doi.org/10.1186/s12859-021-04482-0.

1350 99. Brůna, T., Hoff, K.J., Lomsadze, A., Stanke, M., and Borodovsky, M. (2021). BRAKER2: automatic
1351 eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein
1352 database. NAR Genomics Bioinforma. *3*, lqaa108. https://doi.org/10.1093/nargab/lqaa108.

1353 100. Brůna, T., Lomsadze, A., and Borodovsky, M. (2020). GeneMark-EP+: eukaryotic gene prediction
1354 with self-training in the space of genes and proteins. NAR Genomics Bioinforma. *2*, lqaa026.
1355 https://doi.org/10.1093/nargab/lqaa026.

1356 101. Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y.O., and Borodovsky, M. (2005). Gene
1357 identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Res. *33*, 6494–
1358 6506. https://doi.org/10.1093/nar/gki937.

1359 102. Iwata, H., and Gotoh, O. (2012). Benchmarking spliced alignment programs including Spaln2, an
1360 extended version of Spaln that incorporates additional species-specific features. Nucleic Acids Res.
1361 *40*, e161. https://doi.org/10.1093/nar/gks708.

1362 103. Gotoh, O., Morita, M., and Nelson, D.R. (2014). Assessment and refinement of eukaryotic gene
1363 structure prediction with gene-structure-aware multiple protein sequence alignment. BMC
1364 Bioinformatics *15*, 189. https://doi.org/10.1186/1471-2105-15-189.

1365 104. Bruna, T., Lomsadze, A., and Borodovsky, M. (2024). A new gene finding tool GeneMark-ETP
1366 significantly improves the accuracy of automatic annotation of large eukaryotic genomes. Preprint at
1367 bioRxiv, https://doi.org/10.1101/2023.01.13.524024 https://doi.org/10.1101/2023.01.13.524024.

1368 105. Kim, D., Paggi, J.M., Park, C., Bennett, C., and Salzberg, S.L. (2019). Graph-based genome
1369 alignment and genotyping with HISAT2 and HISAT-genotype. Nat. Biotechnol. *37*, 907–915.
1370 https://doi.org/10.1038/s41587-019-0201-4.

1371 106. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane,
1372 T., McCarthy, S.A., Davies, R.M., et al. (2021). Twelve years of SAMtools and BCFtools.
1373 GigaScience *10*, giab008. https://doi.org/10.1093/gigascience/giab008.

1374 107. Pertea, G., and Pertea, M. (2020). GFF Utilities: GffRead and GffCompare. Preprint at
1375 F1000Research, https://doi.org/10.12688/f1000research.23297.1
1376 https://doi.org/10.12688/f1000research.23297.1.

1377 108. Dainat, J. (2022). AGAT: Another Gff Analysis Toolkit to handle annotations in any GTF/GFF format.
1378 Version v1.0.0 (Zenodo). https://doi.org/10.5281/zenodo.11106497
1379 https://doi.org/10.5281/zenodo.11106497.

1380 109. Cantalapiedra, C.P., Hernández-Plaza, A., Letunic, I., Bork, P., and Huerta-Cepas, J. (2021).
1381 eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the
1382 Metagenomic Scale. Mol. Biol. Evol. *38*, 5825–5829. https://doi.org/10.1093/molbev/msab293.

1383 110. Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S.K., Cook, H., Mende,
1384 D.R., Letunic, I., Rattei, T., Jensen, L.J., et al. (2019). eggNOG 5.0: a hierarchical, functionally and
1385 phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic
1386 Acids Res. *47*, D309–D314. https://doi.org/10.1093/nar/gky1085.

1387 111. Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterProfiler: an R Package for Comparing
1388 Biological Themes Among Gene Clusters. OMICS J. Integr. Biol. *16*, 284–287.
1389 https://doi.org/10.1089/omi.2011.0118.

1390 112. Tanaka, T., Maeda, Y., Veluchamy, A., Tanaka, M., Abida, H., Maréchal, E., Bowler, C., Muto, M.,
1391 Sunaga, Y., Tanaka, M., et al. (2015). Oil Accumulation by the Oleaginous Diatom *Fistulifera solaris*

as Revealed by the Genome and Transcriptome. Plant Cell *27*, 162–176. https://doi.org/10.1105/tpc.114.135194.

113. Hongo, Y., Kimura, K., Takaki, Y., Yoshida, Y., Baba, S., Kobayashi, G., Nagasaki, K., Hano, T., and Tomaru, Y. (2021). The genome of the diatom *Chaetoceros tenuissimus* carries an ancient integrated fragment of an extant virus. Sci. Rep. *11*, 22877. https://doi.org/10.1038/s41598-021-00565-3.

114. Osuna-Cruz, C.M., Bilcke, G., Vancaester, E., De Decker, S., Bones, A.M., Winge, P., Poulsen, N., Bulankova, P., Verhelst, B., Audoor, S., et al. (2020). The *Seminavis robusta* genome provides insights into the evolutionary adaptations of benthic diatoms. Nat. Commun. *11*, 3320. https://doi.org/10.1038/s41467-020-17191-8.

115. Zepernick, B.N., Truchon, A.R., Gann, E.R., and Wilhelm, S.W. (2022). Draft Genome Sequence of the Freshwater Diatom *Fragilaria crotonensis* SAG 28.96. Microbiol. Resour. Announc. *11*, e00289-22. https://doi.org/10.1128/mra.00289-22.

116. Paajanen, P., Strauss, J., van Oosterhout, C., McMullan, M., Clark, M.D., and Mock, T. (2017). Building a locally diploid genome and transcriptome of the diatom *Fragilariopsis cylindrus*. Sci. Data *4*, 170149. https://doi.org/10.1038/sdata.2017.149.

117. Armbrust, E.V., Berges, J.A., Bowler, C., Green, B.R., Martinez, D., Putnam, N.H., Zhou, S., Allen, A.E., Apt, K.E., Bechner, M., et al. (2004). The Genome of the Diatom *Thalassiosira pseudonana*: Ecology, Evolution, and Metabolism. Science *306*, 79–86. https://doi.org/10.1126/science.1101156.

118. Roberts, W.R., Downey, K.M., Ruck, E.C., Traller, J.C., and Alverson, A.J. (2020). Improved Reference Genome for *Cyclotella cryptica* CCMP332, a Model for Cell Wall Morphogenesis, Salinity Adaptation, and Lipid Production in Diatoms (Bacillariophyta). G3 GenesGenomesGenetics *10*, 2965–2974. https://doi.org/10.1534/g3.120.401408.

119. Ferrante, M.I., Broccoli, A., and Montresor, M. (2023). The pennate diatom *Pseudo-nitzschia multistriata* as a model for diatom life cycles, from the laboratory to the sea. J. Phycol. *59*, 637–643. https://doi.org/10.1111/jpy.13342.

120. Lommer, M., Specht, M., Roy, A.-S., Kraemer, L., Andreson, R., Gutowska, M.A., Wolf, J., Bergner, S.V., Schilhabel, M.B., Klostermeier, U.C., et al. (2012). Genome and low-iron response of an oceanic diatom adapted to chronic iron limitation. Genome Biol. *13*, R66. https://doi.org/10.1186/gb-2012-13-7-r66.

121. Emms, D.M., and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. *20*, 238. https://doi.org/10.1186/s13059-019-1832-y.

122. Conway, J.R., Lex, A., and Gehlenborg, N. (2017). UpSetR: an R package for the visualization of intersecting sets and their properties. Bioinformatics *33*, 2938–2940. https://doi.org/10.1093/bioinformatics/btx364.

123. Rohwer, R.R., Hamilton, J.J., Newton, R.J., and McMahon, K.D. (2018). TaxAss: Leveraging a Custom Freshwater Database Achieves Fine-Scale Taxonomic Resolution. mSphere *3*, 10.1128/msphere.00327-18. https://doi.org/10.1128/msphere.00327-18.

124. Tsuji, J., Frith, M.C., Tomii, K., and Horton, P. (2012). Mammalian NUMT insertion is non-random. Nucleic Acids Res. *40*, 9073–9088. https://doi.org/10.1093/nar/gks424.

1432 125. Marçais, G., Delcher, A.L., Phillippy, A.M., Coston, R., Salzberg, S.L., and Zimin, A. (2018).
1433 MUMmer4: A fast and versatile genome alignment system. PLOS Comput. Biol. *14*, e1005944.
1434 https://doi.org/10.1371/journal.pcbi.1005944.

1435 126. Ramírez, F., Dündar, F., Diehl, S., Grüning, B.A., and Manke, T. (2014). deepTools: a flexible
1436 platform for exploring deep-sequencing data. Nucleic Acids Res. *42*, W187–W191.
1437 https://doi.org/10.1093/nar/gku365.

1438 127. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic
1439 features. Bioinformatics *26*, 841–842. https://doi.org/10.1093/bioinformatics/btq033.

1440 128. Rice, P.M., Bleasby, A.J., and Ison, J.C. EMBOSS User's Guide: Practical Bioinformatics with
1441 EMBOSS (Cambridge University Press).

1442 129. Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov,
1443 J.P. (2011). Integrative genomics viewer. Nat. Biotechnol. *29*, 24–26.
1444 https://doi.org/10.1038/nbt.1754.

1445 130. Gu, Z., Gu, L., Eils, R., Schlesner, M., and Brors, B. (2014). circlize implements and enhances
1446 circular visualization in R. Bioinformatics *30*, 2811–2812.
1447 https://doi.org/10.1093/bioinformatics/btu393.

1448 131. Buchfink, B., Reuter, K., and Drost, H.-G. (2021). Sensitive protein alignments at tree-of-life scale
1449 using DIAMOND. Nat. Methods *18*, 366–368. https://doi.org/10.1038/s41592-021-01101-x.

1450 132. Katoh, K., and Standley, D.M. (2013). MAFFT Multiple Sequence Alignment Software Version 7:
1451 Improvements in Performance and Usability. Mol. Biol. Evol. *30*, 772–780.
1452 https://doi.org/10.1093/molbev/mst010.

1453 133. Capella-Gutiérrez, S., Silla-Martínez, J.M., and Gabaldón, T. (2009). trimAl: a tool for automated
1454 alignment trimming in large-scale phylogenetic analyses. Bioinformatics *25*, 1972–1973.
1455 https://doi.org/10.1093/bioinformatics/btp348.

1456 134. Price, M.N., Dehal, P.S., and Arkin, A.P. (2010). FastTree 2 – Approximately Maximum-Likelihood
1457 Trees for Large Alignments. PLOS ONE *5*, e9490. https://doi.org/10.1371/journal.pone.0009490.

1458 135. Stephens, T.G., Bhattacharya, D., Ragan, M.A., and Chan, C.X. (2016). PhySortR: a fast, flexible
1459 tool for sorting phylogenetic trees in R. PeerJ *4*, e2038. https://doi.org/10.7717/peerj.2038.

1460 136. Rancurel, C., Legrand, L., and Danchin, E.G.J. (2017). Alienness: Rapid Detection of Candidate
1461 Horizontal Gene Transfers across the Tree of Life. Genes *8*, 248.
1462 https://doi.org/10.3390/genes8100248.

1463 137. Tyanova, S., Temu, T., and Cox, J. (2016). The MaxQuant computational platform for mass
1464 spectrometry-based shotgun proteomics. Nat. Protoc. *11*, 2301–2319.
1465 https://doi.org/10.1038/nprot.2016.136.

1466 138. Tyanova, S., Temu, T., Sinitcyn, P., Carlson, A., Hein, M.Y., Geiger, T., Mann, M., and Cox, J.
1467 (2016). The Perseus computational platform for comprehensive analysis of (prote)omics data. Nat.
1468 Methods *13*, 731–740. https://doi.org/10.1038/nmeth.3901.