Allodiploid hybridization, loss-of-heterozygosity and aneuploidy in the green alga *Auxenochlorella*, an emerging model for discovery research and bioengineering

Rory J. Craig[1,2], Marco A. Dueñas[3], Dimitrios J. Camacho[4], Sean D. Gallaher[1,4], Crysten E. Blaby-Haas[5,6], Jeffrey L. Moseley[1], Sabeeha S. Merchant[1,3,4,7]

[1] California Institute for Quantitative Biosciences (QB3), University of California, Berkeley, CA 94720, USA
[2] Department of Algal Development and Evolution, Max Planck Institute for Biology Tübingen, Tübingen, 72076, Germany
[3] Department of Plant and Microbial Biology, University of California, Berkeley, CA 94720, USA
[4] Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720, USA
[5] Molecular Foundry, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
[6] US Department of Energy Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
[7] Environmental Genomics and Systems Biology, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

Correspondence: Rory J. Craig, rory.craig@unimelb.edu.au; Sabeeha S. Merchant, sabeeha@berkeley.edu

ABSTRACT

Auxenochlorella are diploid oleaginous green algae that are highly amenable to genetic transformation by homologous recombination, providing great potential for discovery research and bioengineering. Vegetatively diploid organisms experience specific evolutionary phenomena, including allodiploid hybridization, mitotic recombination, loss-of-heterozygosity and aneuploidy; however, studies of these forces have mostly focused on fungi. Here, we present a telomere-to-telomere phased diploid genome assembly of *Auxenochlorella* UTEX 250-A, revealing a streamlined nuclear genome architecture with a haploid length of 22 Mb. We demonstrate that UTEX 250-A is an allodiploid hybrid via sequencing of *Auxenochlorella protothecoides* and *Auxenochlorella symbiontica* strains, two close relatives differentiated by extensive chromosomal rearrangements. The UTEX 250-A haplotypes are a mosaic of each parental species following mitotic recombination, and two chromosomes are trisomic, which we confirm via allele-specific transformation of a trisomic chromosome. Loss-of-heterozygosity events are likely pervasive across *Auxenochlorella* due to mitotic recombination and transient aneuploidy, and can evolve rapidly in laboratory culture. We present high-quality structural annotations, yielding ~7,500 genes per haplotype. *Auxenochlorella* have experienced extensive gene family loss and reduction, including core photosynthesis genes. Nevertheless, a minimal set of genes required for sexual reproduction is potentially present. We demonstrate periodic adenine and cytosine methylation around promoters and across gene bodies. Finally, ~10% of genes, especially those involved in DNA repair and sex, overlap antisense long noncoding RNAs, potentially representing a regulatory mechanism. These results demonstrate the generality of several evolutionary forces associated with vegetative diploidy, and provide a genomic foundation for the use of *Auxenochlorella* as a reference organism.

INTRODUCTION

Microbial organisms that under undergo periods of asexual reproduction as diploids are associated with several molecular and evolutionary phenomena. These processes are best studied in the budding yeast *Saccharomyces cerevisiae* and its close relatives, which undergo a facultatively sexual life cycle featuring long periods of asexual division as a diploid (Tsai, et al. 2008; Fischer, et al. 2021). Although inter-tetrad outcrossing is relatively rare in yeast, one interesting facet of this life cycle is occasional alloploid hybridization. Mating of haploid gametes from divergent lineages produces an allodiploid hybrid that inherits one set of chromosomes from each parent, whereas allotetraploid hybrids can arise via conjugation of divergent diploid cells or whole-genome duplication of an allodiploid individual (Sipiczki 2008; Gabaldón 2020). Although allodiploid hybrids may be incapable of meiosis due to incompatibilities between the parental chromosomes, they can persist via asexual division (Gabaldón 2020). Indeed, hybrids may exhibit increased fitness or unique phenotypes relative to their parents (i.e., heterosis), and hybrid *Saccharomyces* strains have successfully adapted to anthropomorphic environments such as alcoholic ferments (Sipiczki 2008; Peris, et al. 2018; Gallone, et al. 2019; Langdon, et al. 2019) and olive brine (Pontes, et al. 2019). Hybridization has also been linked to the evolution of pathogenicity in taxa including *Candida* yeasts (Pryszcz, et al. 2015; Schroder, et al. 2016) and *Aspergillus* filamentous fungi (Steenwyk, et al. 2020). Importantly, allodiploid hybridization is not necessarily an evolutionary dead end, since the correct pairing of homologous chromosomes and sexual compatibility can be restored via whole-genome duplication (Charron, et al. 2019). Exemplifying this process, an ancient hybridization event is thought to have preceded the whole-genome duplication in the lineage leading to *Saccharomyces* (Marcet-Houben and Gabaldón 2015). Allopolyploidy (hybridization and genome duplication) has also long been recognized as a major force in plant evolution and speciation (Soltis and Soltis 2009).

Although recombination is primarily associated with meiotic cell division, various mitotic recombination processes can occur during asexual growth. In *S. cerevisiae*, mitotic recombination occurs randomly due to double-strand breaks at a rate that is orders of magnitude lower than meiotic recombination (Mancera, et al. 2008; Sui, et al. 2020). Nonetheless, over lengthy periods of asexuality, mitotic recombination can drastically homogenize genomes via

96 loss-of-heterozygosity (LOH) events, where an allele of one chromosome is replaced by the
97 allele carried by its homologous chromosome. LOH can occur at short regions along
98 chromosomes of generally less than 10 kb via gene conversion (i.e., interstitial LOH), or at larger
99 regions that typically extend from an internal location to the telomere (terminal LOH), which are
100 the product of mitotic crossovers or the break-induced replication DNA repair pathway (Sui, et
101 al. 2020; Dutta, et al. 2021). Furthermore, in allodiploid hybrids, reciprocal mitotic crossovers
102 result in the shuffling of haplotypes between the parental sub-genomes, producing chromosomes
103 that are a mosaic of the ancestral parents (Sipiczki 2008). Mitotic recombination has also been
104 characterized in diploid oomycetes (Dale, et al. 2019) and diatoms (Bulankova, et al. 2021),
105 suggesting that it is an important evolutionary process across the tree of life.
106
107 Another phenomenon associated with asexual reproduction is the emergence of aneuploidy via
108 the loss or gain of copies of specific chromosomes. Unbalanced chromosome numbers have been
109 observed in many eukaryotes and are prevalent in several fungi (Vande Zande, et al. 2023). As
110 with hybridization, aneuploidy and its subsequent effects (e.g., gene expression changes) are
111 associated with adaptation to new environments and pathogenicity, although negative fitness
112 effects are common (Santaguida and Amon 2015). For example, whole or partial aneuploidies
113 are common among clinical isolates of *Candida albicans* (Hirakawa, et al. 2015), and an
114 aneuploidy affecting a specific chromosome arm causes antifungal drug resistance (Selmecki, et
115 al. 2006). Under laboratory evolution of *C. albicans*, chromosomal duplications are also
116 observed as transient adaptations to abrupt heat stress (Yona, et al. 2012). Transient aneuploidy
117 can result in LOH at the scale of entire chromosomes via duplication of one chromosome copy
118 and loss of the other (Yuen, et al. 2007; Andersen, et al. 2008).
119
120 These drivers of evolution have received scarce attention in chlorophyte green algae, a highly
121 diverse group that diverged from the lineage leading to land plants approximately 1 billion years
122 ago (Del Cortona, et al. 2020). Although facultatively sexual, the model green alga
123 *Chlamydomonas reinhardtii* (class Chlorophyceae) grows asexually as a haploid, with the
124 diploid stage limited to a dormant zygospore that obligately divides by meiosis (Harris 2001).
125 Similarly, asexual reproduction in the genus *Chlorella* (Trebouxiophyceae) is haploid, with a
126 possible cryptic sexual cycle that is uncharacterized (Blanc, et al. 2010; Fučiková, et al. 2015).

127 Ancient mating type loci have also been identified in haploid *Ostreococcus* species
128 (Mamelliophyceae), demonstrating the presence of haplontic life cycles at the deepest
129 evolutionary branch of the Chlorophyta (Benites, et al. 2021). Despite this, a number of
130 vegetatively diploid species have recently been reported (Figure 1A). Diploid genome assemblies
131 have been assembled for trebouxiophyte algae from the related genera *Nannochloris* (Sanders, et
132 al. 2022) and *Picochlorum* (Foflonker, et al. 2018; Becker, et al. 2020; Barten, et al. 2022; da
133 Roza, et al. 2024), and for the lichen-forming *Trebouxia lynniae* (Gazquez, et al. 2024).
134 *Chloropicon primus* (Chloropicophyceae) is trisomic (three copies) for a single chromosome and
135 is otherwise diploid (Lemieux, et al. 2019). In the Chlorophyceae, diploids from the genera
136 *Haematococcus* (Volvocales) and *Tetradesmus* (Sphaeropleales) have been reported (Calhoun, et
137 al. 2021; Marcolungo, et al. 2024). Most notably, Biondi, et al. (2024) described a diploid
138 *Tetradesmus obliquus* strain with extensive heterozygosity but one entirely homozygous
139 chromosome, patterns potentially consistent with hybridization followed by chromosome-scale
140 LOH. These results suggest that the presence of diplontic life cycles may be underestimated in
141 Chlorophyta, although the prevalence of hybridization, LOH and aneuploidy is presently
142 unknown.
143
144 *Auxenochlorella* is a genus of trebouxiophyte algae that is highly amenable to discovery research
145 and bioengineering (Goold, et al. 2024). Together with its close relatives from the genus
146 *Prototheca*, the type species *Auxenochlorella protothecoides* is diploid and targeted
147 transformation of the nuclear genome by homologous recombination is efficient (Franklin, et al.
148 2019; Moseley, et al. 2022). While *Prototheca* are obligate heterotrophs that have lost the ability
149 to photosynthesize, *Auxenochlorella* can grow robustly as a phototroph, mixotroph or
150 heterotroph, undergoing a metabolic switch that degrades the photosynthetic apparatus when
151 provided with an organic carbon source (Matsuka, et al. 1966). Resultingly, it is facile to study
152 photosynthetic gene function through reverse genetics, making *Auxenochlorella* an attractive
153 model in plant biology. Both *Auxenochlorella* and *Prototheca* are oleaginous, and
154 *Auxenochlorella* oil and biomass have generally recognized as safe (GRAS) status, enabling
155 engineering of specific lipids for bioproducts, food and renewable fuels (Franklin, et al. 2019;
156 Brooks, et al. 2024). Collectively, these features provide significant advantages relative to
157 existing models such as *C. reinhardtii*, where incorporation of transgenes mostly occurs via non-

158 homologous end-joining, making locus-specific transformation challenging (Zorin, et al. 2005).
159 Demonstrating these capabilities, we recently introduced synthetic gene constructs in to
160 *Auxenochlorella* to characterize the mechanism underlying the translation of bicistronic genes
161 (Dueñas, et al. 2024), which are widespread in green algae (Gallaher, et al. 2021).
162
163 Here, we present a fully phased, telomere-to-telomere diploid genome assembly for the
164 *Auxenochlorella* strain UTEX 250-A. We find an unusual genome architecture where six of the
165 twelve chromosome pairs are highly rearranged and non-homologous. We demonstrate that
166 UTEX 250-A is a putative allodiploid hybrid of *Auxenochlorella protothecoides* and
167 *Auxenochlorella symbiontica*, which each differ in their respective karyotypes. Despite the high
168 heterozygosity introduced by hybridization, several local genomic regions and three entire
169 chromosomes are homozygous following extensive LOH. Two chromosomes are partially
170 trisomic, one of which we experimentally verify by allele-specific transformation. We also
171 introduce a highly curated annotation of gene models, and report a streamlined nuclear genome
172 architecture and gene content. These resources demonstrate that a green algal genome can be
173 shaped by forces experienced by diplontic taxa, and provides a platform for the future use of
174 *Auxenochlorella* as a model in plant biology and bioengineering.
175

176 RESULTS & DISCUSSION

177 The genus *Auxenochlorella* and existing genomic data

178 First isolated in the 1890s as "*Chlorella*" *protothecoides* (Krüger 1894b, a), *Auxenochlorella* was
179 later established as an independent monotypic genus based on physiological, morphological and
180 genetic data (Kalina and Punčochárová 1987; Huss, et al. 1999). Indeed, *Auxenochlorella* and
181 true *Chlorella* species likely diverged more than 550 million years ago (Del Cortona, et al. 2020)
182 and clearly form distinct lineages within the taxonomically diverse order Chlorellales (Figure
183 1A). The closest relatives of *Auxenochlorella* have all lost photosynthesis; *Prototheca* is a
184 polyphyletic assemblage of heterotrophic free-living and opportunistically parasitic algae, and
185 *Helicosporidium* are obligate parasites of arthropods (Jagielski, et al. 2019). Collectively, these
186 algae form the "AHP lineage" (Ueno, et al. 2005), and phylogenetic analyses support multiple
187 independent transitions to heterotrophy, with *Auxenochlorella* enduring as the sole
188 photosynthetic lineage (Figueroa-Martinez, et al. 2015; Suzuki, et al. 2018; Guo, et al. 2022).

189 Several isolates classified as *A. protothecoides* are maintained in culture centers, including the
190 original strain isolated by W. Krüger from the sap of a poplar tree (UTEX 25 at the University of
191 Texas Culture Collection of Algae). *Auxenochlorella* have also been isolated from fresh and salt
192 water, soil, and as symbionts of freshwater hydrozoans and sponges (Figure 1B). Based on
193 ribosomal DNA (rDNA) analysis, Darienko and Pröschold (2015) described a second species, *A.*
194 *symbiontica*, with the type strain CCAP 211/61 isolated from *Hydra viridis*. We produced an 18S
195 rDNA phylogeny including some additional isolates (Table S1), which supports an association
196 between *A. protothecoides* and both tree sap and freshwater habitats, whereas most *A.*
197 *symbiontica* strains were isolated from marine environments (Figure 1B). However, the marine
198 strain UTEX 2341 groups with *A. protothecoides* and freshwater symbionts have been isolated
199 for both species, suggesting that both *A. protothecoides* and *A. symbiontica* can live in multiple
200 and overlapping environments.
201
202 Gao, et al. (2014) produced the first draft assembly of *A. protothecoides* "0710" (=UTEX 25, see
203 below), while additional draft assemblies of UTEX 25 and UTEX 2341 are available at NCBI
204 (Table S2). Several draft assemblies of *Prototheca* and *Helicosporidium* have been produced
205 (Pombert, et al. 2014; Severgnini, et al. 2018; Suzuki, et al. 2018; Bakuła, et al. 2021; Guo, et al.
206 2022). Although these assemblies are all collapsed to haploids, Guo, et al. (2022) reported
207 heterozygosity in two *Prototheca wickerhamii* genomes. Every available AHP lineage genome is
208 highly streamlined relative to ~40-60 Mb genomes of *Chlorella* species; for example, the *A.*
209 *protothecoides* 0710 assembly has a haploid length of 22.9 Mb, with ~7,000 predicted genes and
210 low repeat content (Gao, et al. 2014) (Table 1). It is interesting to note that *Picochlorum* and
211 *Nannochloris* species also possess diploid streamlined genomes (Table 1), and phylogenetic
212 placement of these species relative to *Chlorella* (including *Micractinium*) and the AHP lineage is
213 uncertain (Figure 1A). Whether the transition to a diplontic lifecycle occurred independently in
214 the AHP lineage and *Picochlorum*/*Nannochloris*, or whether these lineages may in fact form
215 sister groups, is an outstanding question.

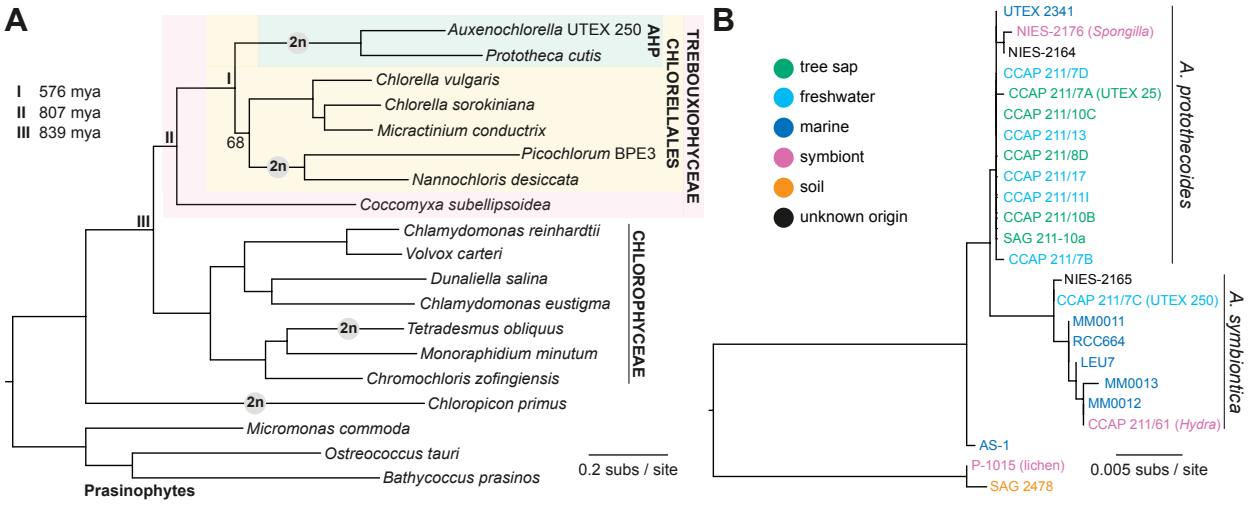


Figure 1. Phylogenetic relationships of *Auxenochlorella*. A) Maximum likelihood phylogeny based on concatenated protein alignment of 669 single-copy orthologs, using the LG+F+R5 substitution model. All nodes received ultrafast bootstrap values >95, except the node connecting *Picochlorum* and *Nannochloris* to *Chlorella* and *Micractinium* (68 support value). Putative transitions to at least partial vegetative diploidy are shown by the "2n" symbol. Divergence estimates were taken from Del Cortona, et al. (2020). B) Maximum likelihood phylogeny based on 18S ribosomal DNA alignment, using the TNe+I substitution model. See Table S1 for strain metadata.

Genome		Size (Mb)	Genes	TE genes	lncRNA genes	GC (%)	Repeats (%)	Reference
Auxenochlorella UTEX250-A	Haplotype A	22.0	7,515	86	544	63.9	6.22	
	Haplotype B	22.0	7,519	121	572	63.9	6.21	
Auxenochlorella protothecoides 0710		22.9	6,988	26	NA	63.6	5.58	Gao, et al. 2014
Prototheca cutis JCM15793		20.0	5,600	5	NA	61.3	5.84	Suzuki, et al. (2018) Kwon, et al. (2023)
Chlorella vulgaris 211/11P		40.2	10,341	357	NA	61.8	13.1	Cecchin, et al. (2019)
Chlorella sorokiniana UTEX 1602		59.4	9,489	37	NA	64.1	12.9	Arriola, et al. (2018)
Micractinium conductrix SAG 241.80		60.8	9,027	190	NA	67.3	21.5	Arriola, et al. (2018)
Nannochloris desiccata UTEX 2526		21.6	9,123	187	NA	45.0	8.12	Sanders, et al. (2022)
Picochlorum sp. BPE23	Haplotype A	14.8	7,216	11	NA	46.3	9.72	Barten, et al. (2022)
Coccomyxa subellipsoidea v3		48.9	10,836	57	NA	52.9	6.04	Blanc, et al. (2012)
Chlamydomonas reinhardtii v6.1		114.0	16,801	810	NA	64.1	26.4	Craig, et al. (2023)

Table 1. Summary statistics of representative trebouxiophyte genomes with *C. reinhardtii* as an outgroup. See methods for determination of transposable element (TE) genes included in annotations; note that these numbers are not representative of the true number of TE genes in each genome due to differences in annotation methodology.

234 **Telomere-to-telomere, phased diploid assembly of *Auxenochlorella* UTEX 250-A**
235 To serve as a foundation for discovery research and bioengineering, we sought to produce a
236 reference-quality genome assembly of *Auxenochlorella*. We selected the strain UTEX 250 based
237 on its robust performance in high cell density fermentations (Brooks, et al. 2024). Metadata on
238 UTEX 250 are sparse; the strain was isolated before 1950 in the Netherlands by A. J. Kluyver,
239 possibly from freshwater. Although described as *A. protothecoides*, UTEX 250 groups with *A.*
240 *symbiontica* based on 18S rDNA (Darienko and Pröschold 2015) (Figure 1B). Since our line of
241 the strain has been independently maintained for several years after acquisition from UTEX, we
242 performed single colony purification prior to genome sequencing, and we designate this sub-
243 clone as "UTEX 250-A".
244
245 We produced high-coverage Pacific Biosciences (PacBio) HiFi and linked-read OmniC datasets
246 for UTEX 250-A, from which we produced a genome assembly via automated and manual
247 assembly (see Methods). This effort yielded a telomere-to-telomere diploid genome assembly
248 comprising two haplotypes, each featuring 12 gapless nuclear chromosomes spanning ~22.0 Mb
249 (i.e., collectively 24 nuclear chromosomes spanning 44.0 Mb) (Table 1). Heterozygosity
250 between the two haplotypes is 2.71% (i.e., homologous chromosomal regions differ by ~27
251 single nucleotide polymorphisms per 1 kb; heterozygosity calculations exclude entirely
252 homozygous regions, see below). This extensive inter-haplotype variation made it possible to
253 fully phase the assembled chromosomes using the accurate PacBio reads, so that each haplotype
254 corresponds to a physical chromosome in UTEX 250-A. We arbitrarily label these haplotypes
255 "A" and "B". As presented below, only six of the twelve chromosome pairs exhibit one-to-one
256 homology, and the remaining six are highly rearranged between the two haplotypes.
257
258 All chromosomes terminate in the telomeric repeat (TTAGGG)n, except for one chromosome
259 arm per haplotype that terminates in a truncated rDNA array. This telomeric motif differs from
260 the $(TTTAGGG)_n$ repeat found in the trebouxiophyte genera *Chlorella* and *Coccomyxa*, which is
261 likely ancestral to green algae, suggesting an independent transition to $(TTAGGG)_n$ telomeric
262 repeats as observed in some other green algae (Fulnečková, et al. 2012). The 5S rDNA genes,
263 which generally form tandem arrays independent of the major rDNA arrays, are present as 16
264 (haplotype A) or 17 (haplotype B) standalone genes dispersed throughout the genome. Similar

265 non-tandem 5S rDNA genes have been observed in species including the yeasts
266 *Schizosaccharomyces pombe* and *Yarrowia lipolytica* (Torres-Machorro, et al. 2010).
267
268 As in *A. protothecoides* 0710, repeat content is low, with interspersed and tandem repeats
269 spanning 3.4% and 2.8% of the genome, respectively (Table 1). Among the interspersed repeats,
270 we identified potentially active families of *Metaviridae* long terminal repeat (LTR)
271 retrotransposons, and *hAT* and *EnSpm-Plavaka* cut-and-paste DNA transposons, with many
272 insertions segregating as polymorphisms between the two haplotypes (Figure S1, Table S3). The
273 most abundant repeat is an uncharacterized transposon that encodes a Fanzor protein, which are
274 RNA-guided endonucleases carried by diverse transposons in protists, fungi and giant viruses
275 (Saito, et al. 2023). Phylogenetically diverse Fanzors are sporadically found in green algal
276 genomes including *C. reinhardtii* (Bao and Jurka 2013) and *Chloropicon primus* (Yoon, et al.
277 2023), and presumably arise via lateral gene transfer.
278
279 Green algal centromeres identified thus far are transposon-enriched regions of tens to hundreds
280 of kilobases in species of *Chlorella*, *Coccomyxa* and *Chlamydomonas* (Blanc, et al. 2012; Craig,
281 et al. 2021; Wang, et al. 2024). In line with the paucity of repeats, we found no obvious repeat-
282 rich centromere candidates in UTEX 250-A. We searched for short regional centromeres,
283 nonrepetitive AT-rich regions of 1-5 kb as found in the red alga *Cyanidioschyzon merolae*
284 (Kanesaki, et al. 2015) and the diatom *Phaeodactylum tricornutum* (Diner, et al. 2017). On 9 of
285 the 12 chromosomes, we identified single AT-rich (mean 48.0% GC, relative to 63.9% GC
286 genome-wide) intergenic regions spanning 2.1 – 7.6 kb (Figure S2, Table S4). Although ChIP-
287 sequencing of the centromeric histone variant will be required to confirm centromere locations,
288 considered alongside the low repeat content and non-tandem 5S rDNA genes, these candidates
289 support a highly streamlined architecture of the UTEX 250-A genome.
290
291 We produced complete circular genome assemblies for the plastome (84.6 kb, 30.8% GC; Figure
292 S3) and mitogenome (54.0 kb, 29.0% GC; Figure S4). The UTEX 250-A plastome is broadly
293 consistent with existing assemblies from *A. protothecoides* UTEX 25 (Yan, et al. 2015; Park, et
294 al. 2022). Briefly, the plastome features 78 protein-coding genes and lacks internal repeats,
295 introns, and trans-spliced genes, resulting in a more compact architecture relative to most

296 trebouxiophyte plastomes. The mitogenome features 37 protein-coding genes, including two
297 LAGLIDADG homing endonucleases encoded by introns in *cox1*, as also observed in the *P.*
298 *wickerhamii* mitogenome (Wolff, et al. 1993).
299
300 **_Auxenochlorella_ UTEX 250 is a putative allodiploid hybrid**
301 Considering the high heterozygosity and rearranged nature of half of the UTEX 250-A diploid
302 chromosomes, we speculated that UTEX 250 may be an allodiploid hybrid between two distinct
303 parental species that possess different karyotypes. Given the close evolutionary relationship
304 between *A. protothecoides* and *A. symbiontica* (Figure 1B), we explored this hypothesis via
305 genome sequencing of representative strains of both species. We assembled gapless
306 chromosome-level assemblies for *A. protothecoides* strains UTEX 25 and UTEX 2341, and *A.*
307 *symbiontica* CCAP 211/61, from PacBio HiFi data (supplemented by Oxford Nanopore
308 Technologies, ONT, reads for CCAP 211/61; see Methods). The three assemblies were broadly
309 consistent with our UTEX 250-A assembly; each nuclear genome comprised 12 diploid
310 chromosomes, with haploid genome sizes ranging from 21.9 Mb to 23.0 Mb (Table S5).
311 However, unlike UTEX 250-A, we observed no rearrangements between the homologous
312 chromosome pairs in each genome, implying a typical diploid karyotype in these strains.
313
314 We first considered the six homologous UTEX 250-A chromosomes. Each of these six
315 chromosomes corresponds to a pair of homologous chromosomes in UTEX 25, UTEX 2341, and
316 CCAP 211/61, suggesting that they are ancestral to *Auxenochlorella*. In UTEX 250-A, three of
317 these chromosomes (1, 2 and 6) are mostly heterozygous, whereas the remaining chromosomes
318 (4, 9 and 10) are entirely homozygous (Figure 2A). We presume that the homozygous
319 chromosomes were originally heterozygous, with homozygosity arising following chromosome-
320 scale LOH events (see below). As introduced, heterozygosity in UTEX 250-A is ~2.7%,
321 although many local genomic regions are homozygous (e.g., the ~73 kb homozygous region at
322 the left arm terminus of chromosome 1, Figure 2A). In *A. protothecoides*, heterozygosity is
323 substantially lower at 0.59% (UTEX 25) and 0.62% (UTEX 2341) in heterozygous regions, with
324 both genomes also exhibiting extensive regional homozygosity. Average heterozygosity in
325 CCAP 211/61 is 1.35% in heterozygous regions, implying that *A. symbiontica* harbors greater
326 genetic diversity than *A. protothecoides*.

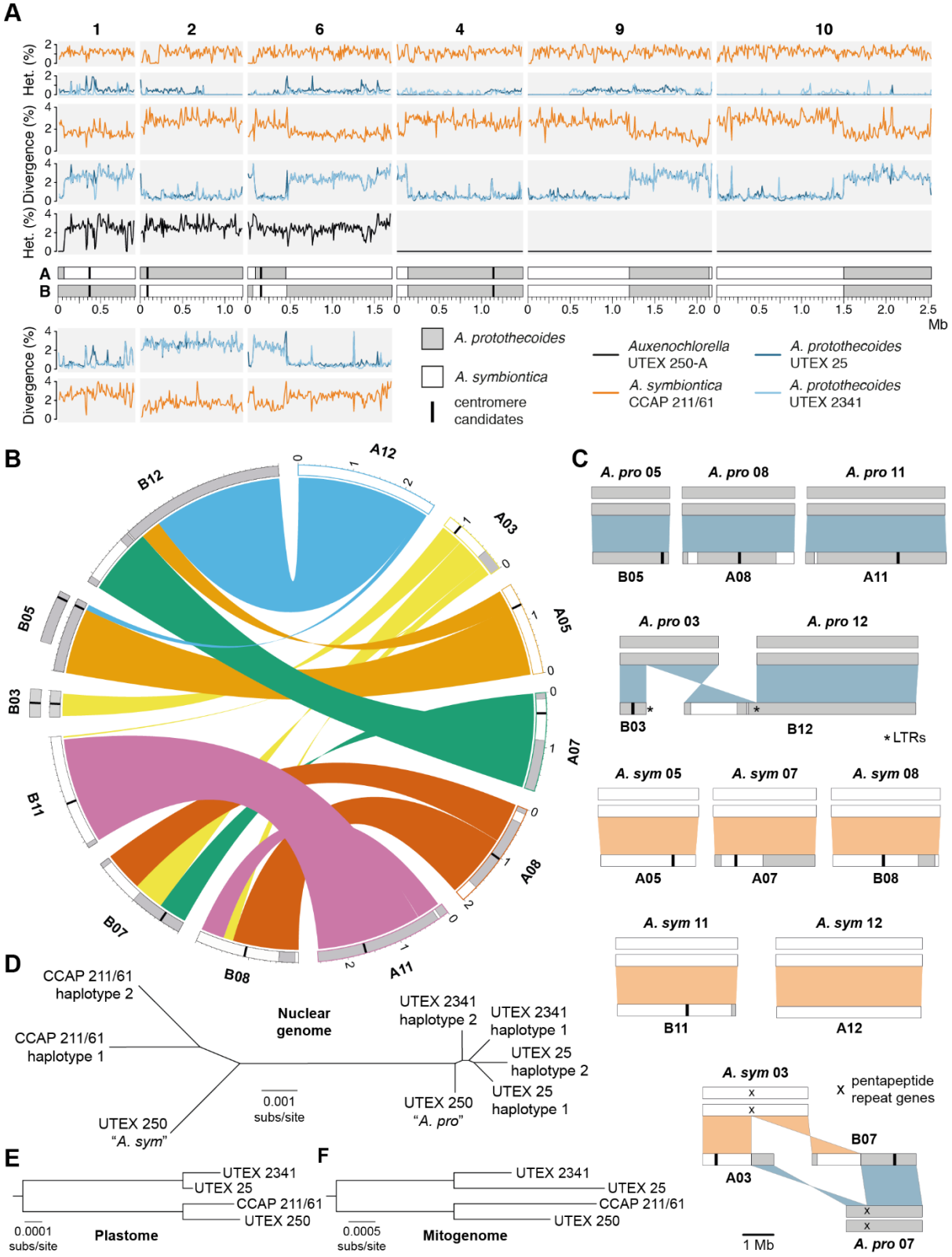
327

328 When comparing the A and B haplotypes of the UTEX 250-A chromosomes to the other
329 *Auxenochlorella* strains, we observed substantial inter-haplotype variation in genetic divergence
330 (Figure 2A). For example, on chromosome 2, haplotype A differs on average by 0.57% and
331 0.49% relative to the homologous chromosomes of UTEX 25 and UTEX 2341, respectively.
332 Meanwhile, haplotype B differs by 2.69% (UTEX 25) and 2.66% (UTEX 2341). Therefore, for
333 chromosome 2, the divergence between haplotype A and the *A. protothecoides* genomes is
334 comparable to heterozygosity between chromosome pairs in *A. protothecoides*, whereas the
335 divergence between haplotype B and *A. protothecoides* is comparable to heterozygosity in
336 UTEX 250-A. The reciprocal pattern is observed via comparison to *A. symbiontica*; divergence
337 between haplotype B and CCAP 211/61 (1.64%) is comparable to heterozygosity in CCAP
338 211/61, whereas divergence between haplotype A and CCAP 211/61 (2.84%) is comparable to
339 UTEX 250-A heterozygosity. These results are consistent with haplotype A of chromosome 2
340 originating from an *A. protothecoides*-like ancestor, and haplotype B originating from an *A.*
341 *symbiontica*-like ancestor. Extending this logic, it is possible to use genetic divergence to assign
342 a putative parent of origin to all genomic regions in UTEX 250-A (Figure 2A, B).
343

344 In contrast to chromosome 2, we observed intra-haplotype switches in assigned parentage on the
345 remaining five homologous chromosome pairs (Figure 2A). On chromosome 1, a single switch
346 corresponds to the 73 kb homozygous region at the left arm terminus, and can be attributed to a
347 LOH event in which the ancestral haplotype A copy of this region was replaced by the haplotype
348 B copy, resulting in both resembling *A. protothecoides*. Conversely, the switches on
349 chromosome 6 are reciprocal, resulting in a ~360 kb region of *A. protothecoides*-like sequence
350 on an otherwise *A. symbiontica*-like chromosome for haplotype A (and vice versa on haplotype
351 B). This mosaicism could be explained by mitotic crossover following hybridization, i.e., each
352 haplotype originally corresponded to one parent (as for chromosome 2), but have subsequently
353 undergone reciprocal exchange. The homozygous chromosomes 4, 9 and 10 also exhibit switches
354 in assigned parentage that are consistent with either LOH or crossover events occurring in a
355 heterozygous state that preceded complete LOH. Note that the A and B labels are arbitrary, and
356 the genomic mosaicism makes it impossible to correspond physical haplotypes to parents of
357 origin in UTEX 250-A.

358 We next tested whether the UTEX 250-A rearranged chromosomes correspond to the ancestral
359 karyotypes of *A. protothecoides* and *A. symbiontica*. The extent of the rearrangements among
360 these chromosomes can only be explained by multiple translocations, one fusion, and one fission
361 (Figure 2B). Remarkably, eight of the twelve chromosome copies correspond exactly to a
362 homologous chromosome pair in either *A. protothecoides* (both UTEX 25 and UTEX 2341,
363 which have an identical karyotype) or *A. symbiontica* (CCAP 211/61). Specifically, UTEX 250-
364 A chromosomes B05 (i.e., chromosome 5, haplotype B), A08 and A11 correspond to
365 homologous chromosome pairs in *A. protothecoides*, whereas A05, A07, B08, B11 and A12
366 correspond to homologous chromosome pairs in *A. symbiontica* (Figure 2C). Therefore, most of
367 the rearrangements in UTEX 250-A can be explained by direct inheritance from parental species
368 with different karyotypes. The remaining chromosomal rearrangements can be explained by
369 events that occurred following hybridization. The UTEX 250-A chromosomes B03 and B12
370 correspond to a pair of *A. protothecoides* chromosomes, suggesting fission of the ancestral *A.*
371 *protothecoides* chromosome 3 and subsequent fusion of one of the resulting chromosomal
372 fragments (which is predicted to have lacked a centromere) to the ancestral *A. protothecoides*
373 chromosome 12 (Figure 2C). Interestingly, LTR elements are present at both the *de novo*
374 terminus of B03 and the fusion site of B12, suggesting that retrotransposons may have invaded
375 (and potentially stabilized) the unprotected chromosomal ends following fission. Finally, the
376 UTEX 250-A chromosomes A03 and B07 are consistent with a reciprocal translocation between
377 the ancestral *A. protothecoides* chromosome 7 and *A. symbiontica* chromosome 3 (Figure 2C).
378 These ancestral chromosomes feature a common locus of ~6 kb that corresponds to two genes
379 that are part of the largest *Auxenochlorella*-specific gene family (see below), suggesting that the
380 translocation occurred via ectopic recombination between these nonhomologous repeats. It is
381 also noteworthy that switches in the assigned parent of origin are frequently observed on the
382 rearranged chromosomes (Figure 2B, C), implying that mitotic recombination has occurred
383 between homologous regions despite the rearrangements.
384
385 Finally, we constructed a phylogeny based on genomic regions that i) are present on UTEX 250-
386 A chromosomes that have not obviously experienced mitotic crossovers, and ii) are not affected
387 by LOH in any of the four strains. As expected, this analysis clearly divided the UTEX 250-A
388 genome into regions that clustered with either UTEX 25 and UTEX 2341 (*A. protothecoides*-

389 like) or CCAP 211/61 (*A. symbiontica*-like), with the longer branch lengths for *A. symbiontica*
390 reflecting the higher genetic diversity in the species (Figure 2D). We also constructed trees for
391 the plastome (Figure 2E) and mitogenome (Figure 2F), which demonstrated putative inheritance
392 from *A. symbiontica* in both cases.
393
394 In summary, *A. protothecoides* and *A. symbiontica* represent authentic species that are ~2.7%
395 divergent at the sequence-level and are distinguished at the chromosome-level by multiple
396 rearrangements. Assuming that the sequenced strains are representative, genetic diversity may be
397 more than two-fold higher in *A. symbiontica* than *A. protothecoides*. UTEX 250 is a putative
398 allodiploid hybrid that presumably originated via the fusion of haploid cells of *A. symbiontica*
399 and *A. protothecoides*. The ancestrally inherited haplotypes have subsequently been mixed via
400 nonreciprocal and reciprocal mitotic exchange, and four ancestrally inherited chromosomes have
401 been involved in post-hybridization rearrangements (including a reciprocal translocation
402 potentially mediated by ectopic recombination). Given the efficiency of homologous
403 recombination in *Auxenochlorella*, it is possible that the rapid emergence of chromosomal
404 differences between the closely related *A. protothecoides* and *A. symbiontica* may have been
405 driven by similar recombination events. Indeed, assuming a sexual life cycle for
406 *Auxenochlorella*, UTEX 250-A is presumably incapable of meiosis due to the non-homologous
407 nature of its chromosomes (Garagna, et al. 2014), and the karyotypic variation between the two
408 species may present a barrier to gene flow.
409
410



Legend on next page.

414 **Figure 2.** Allodiploid hybrid origin of UTEX 250. A) Heterozygosity and divergence for four
415 *Auxenochlorella* strains in 10 kb windows across the six UTEX 250-A chromosomes with one-to-one
416 homology. For divergence metrics the upper panels refer to haplotype A, lower panels to haplotype B. B)
417 Genomic rearrangements among the remaining six diploid UTEX 250-A chromosomes. C) Comparison
418 between rearranged UTEX 250-A chromosomes and diploid *A. protothecoides* and *A. symbiontica*
419 chromosomes. D) Neighbor joining phylogeny of all *Auxenochlorella* haplotypes. E) Neighbor joining
420 phylogeny of plastome sequences. F) Neighbor joining phylogeny of mitogenome sequences.
421

422 **Pervasive loss-of-heterozygosity and evolution in the laboratory**

423 LOH events are associated with several mechanisms and can affect local genomic regions
424 (interstitial LOH), large regions that extend to a chromosome terminus (terminal LOH), or entire
425 chromosomes. We observed all three LOH categories in UTEX 250-A, resulting in at least
426 35.5% of the genome being homozygous (conservatively assuming a minimum interstitial LOH
427 length of 1 kb). The three entirely homozygous chromosomes, 4, 9 and 10, collectively span
428 28.2% of the genome. Terminal LOH events affect seven chromosomal termini (e.g., the left arm
429 terminus of chromosome A01, Figure 2A), ranging from ~12 kb to ~271 kb (mean 111 kb) and
430 collectively spanning 3.5% of the genome. We observed 278 putative interstitial LOH events, 36
431 of which exceed 5 kb in length, reaching a maximum of 37.1 kb. Notably, LOH is pervasive in
432 all of the sequenced *Auxenochlorella* genomes; 23.3% of the *A. symbiontica* CCAP 211/61
433 genome is homozygous, whereas levels of homozygosity reach 67.2% (UTEX 25) and 80.3%
434 (UTEX 2341) in *A. protothecoides*.
435
436 Since all of the studied *Auxenochlorella* strains have been in culture for decades, we questioned
437 whether LOH could have accumulated in the laboratory. UTEX 25 is one of the oldest microbes
438 in culture, having been isolated in the late 19th century (Krüger 1894b, a). CCAP 211/61 was
439 isolated in 1982, and UTEX 2341 prior to 1984 (Seto, et al. 1984). Conveniently, many of these
440 strains are maintained at different culture centers, including UTEX and CCAP (Culture
441 Collection of Algae and Protozoa). Although records are uncertain, a sister lineage of UTEX 250
442 has likely been independently maintained for more than 70 years as CCAP 211/7C (Figure 3A).
443 To test whether LOH has occurred during culture, we obtained CCAP 211/7C and performed
444 high-coverage ONT sequencing.
445
446 Remarkably, LOH events affect only 19.7% of the CCAP 211/7C genome. Although
447 chromosome 10 is entirely homozygous in both UTEX 250-A and CCAP 211/7C, chromosomes

448 4 and 9 are heterozygous in CCAP 211/7C (Figure 3B). Thus, two of the chromosome-scale
449 LOH events must have occurred during culture at UTEX or in our laboratory since obtaining the
450 strain. Similarly, of the seven UTEX 250-A terminal LOH events, only five are present in CCAP
451 211/7C. For example, both terminal LOH events on chromosome A07 and the right arm terminal
452 LOH event on A08 are shared, whereas the event on the left arm terminus of A08 is unique to
453 UTEX 250-A (Figure 3B). The same pattern is observed for interstitial LOH, for example, the
454 largest interstitial event in the genome (37.1 kb at ~1.66 Mb on chromosome A08) is shared,
455 whereas the second largest (29.2 kb, ~0.69 Mb on A07) is unique to UTEX 250-A (Figure 3B).
456 Interestingly, almost all of the strain-specific LOH events are unique to UTEX 250-A,
457 suggesting that the frequency of LOH during culture at CCAP is substantially lower.
458
459 These results support the rapid accumulation of LOH events under at least some laboratory
460 conditions, as has been observed experimentally in yeast (Sui, et al. 2020; Dutta, et al. 2021).
461 Contrastingly, the UTEX 250-A chromosomal rearrangements that likely occurred following
462 hybridization are also present in CCAP 211/7C, suggesting that these events occurred prior to
463 sampling of the original strain (or at least before the UTEX and CCAP lineages were
464 established). Although it is possible that some of the LOH events in UTEX 250-A and CCAP
465 211/7C were fixed by positive selection to overcome post-hybridization incompatibles, there is
466 no obvious directionality with respect to parental species: approximately equal numbers of bases
467 within regions affected by interstitial and terminal LOH events can be assigned to *A.*
468 *protothecoides* and *A. symbiontica*. It is possible that the different frequencies of LOH events in
469 CCAP 211/7C and UTEX 250-A could be attributed to differences in culture conditions (e.g.,
470 liquid vs agar cultures that would affect clonal population sizes). We recommend
471 cryopreservation and avoiding bottlenecks in culture sizes where possible to avoid the fixation of
472 LOH events by drift.
473
474
475
476
477
478
479
480
481

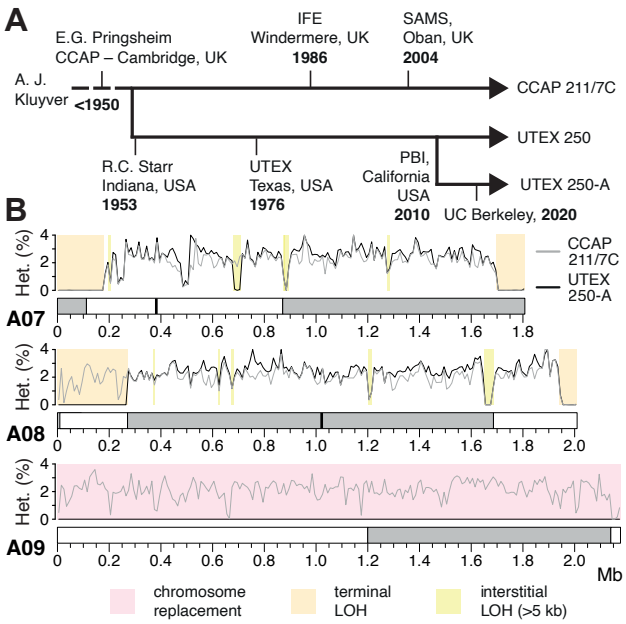


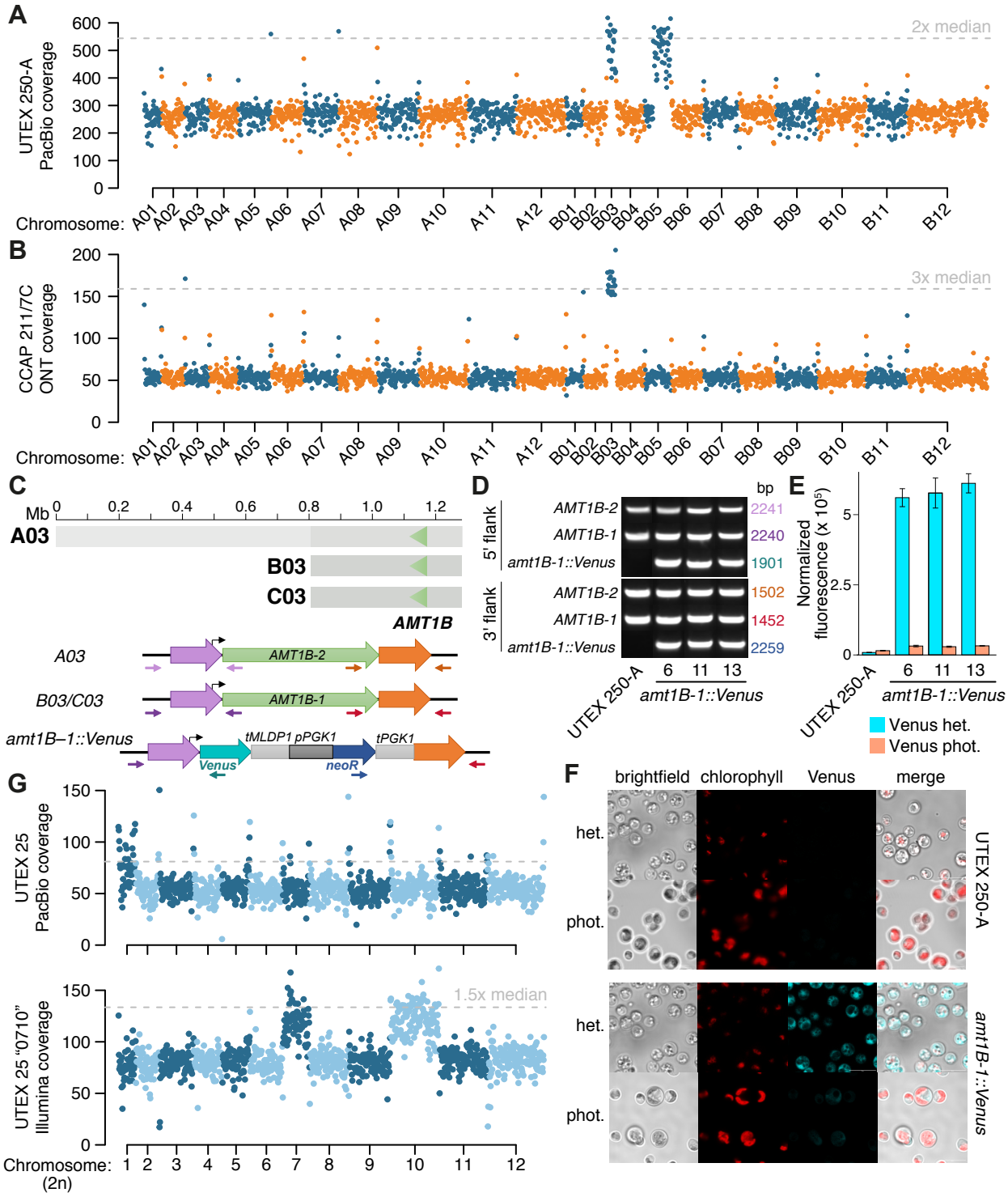
Figure 3. Loss-of-heterozygosity and evolution in the laboratory. A) Putative strain history of UTEX 250. B) Heterozygosity in 10 kb windows across three example chromosomes. Chromosomes are shaded according to parental species of origin and candidate centromeres are marked where available (see Figure 1). LOH = loss-of-heterozygosity. Different LOH categories in UTEX 250-A are shown by colored boxes.

Aneuploidy is widespread in *Auxenochlorella* and can be confirmed by allele-specific transformation

While assembling the UTEX250-A genome, we noticed that two genomic regions exhibited unexpectedly high read coverage. Specifically, when mapping the PacBio reads against the diploid assembly, the entirety of chromosome B03 and a 932 kb fragment of B05 exhibit median read coverages of 528x and 519x, respectively, approximately twice the 272x coverage of the other individual chromosomes, suggesting that these regions are trisomic when considering both A and B haplotypes (Figure 4A). The putatively duplicated part of B05 features the right arm terminus and associated telomere, and we identified reads featuring telomeric repeats that map internally to B05 precisely at the location where coverage doubles (Figure S5). Conceivably, B05 may have duplicated and then undergone fission, with the fragment retaining the centromere acquiring a de novo telomere at its left terminus, yielding a partly aneuploid chromosome. We compared read coverage between UTEX 250-A and CCAP 211/7C to explore whether the observed aneuploidies could have evolved during laboratory culture. Surprisingly, B05 exhibited normal coverage in the CCAP 211/7C ONT read dataset, whereas B03 had three times the expected coverage (163x relative to 53x genome-wide, Figure 4B), suggesting tetrasomy for this

506 chromosome (three copies of B03 and one copy of the homologous region on A03). Thus, the
507 B05 duplication appears to have occurred during culture leading to UTEX 250-A (or
508 alternatively CCAP 211/7C has reverted to disomy), whereas aneuploidy of B03 may have
509 existed prior to sampling, with subsequent copy number change during culture.
510
511 To explicitly test the predicted aneuploidy in UTEX 250-A, we took advantage of the
512 amenability of *Auxenochlorella* to genetic transformation via homologous recombination, in
513 combination with the characteristic metabolic switch of the organism. The ammonium
514 transporter gene *AMT1B*, which is highly upregulated under N-depletion in heterotrophic growth
515 in UTEX 25 (Yan, et al. 2013), is located on the right arms of chromosomes A03 and B03, with
516 a putative third copy on the duplicate of B03 (termed "C03", Figure 4C). Since A03 and
517 B03/C03 are heterozygous at this region, we designed an allele-specific construct with homology
518 arms that specifically correspond to the flanking sequences of the *AMT1B* open reading frame
519 (ORF) of the B03/C03 allele (termed *AMT1B-1*, Figure 4C). The construct replaced the *AMT1B-*
520 *1* ORF with an ORF encoding the Venus fluorescent protein, which is under the control of the
521 endogenous *AMT1B-1* promoter, and a neomycin phosphotransferase II (*neoR*) gene that served
522 as a transformation marker conferring resistance to G418 (Geneticin) (Moseley, et al. 2022;
523 Dueñas, et al. 2024). We designed allele-specific PCR primers that targeted the left and right
524 flanks of the endogenous *AMT1B* alleles, in addition to primer pairs that targeted the construct,
525 with one primer within the construct and the other in the sequence flanking the homology arms
526 (Fig. 4C). Wild-type sequences for both *AMT1B* alleles (*AMT1B-1* on B03 or C03, *AMT1B-2* on
527 A03) could be amplified from wild-type UTEX 250-A and from three representative
528 transformants (see Methods for transformation details), but the transformants also contained
529 mutant *amt1B-1::Venus* integrations, consistent with the presence of one copy of *AMT1B-2* and
530 two copies of *AMT1B-1* in the wild-type nuclear genome (Figure 4D). Transformants displayed
531 an average of 66-fold higher Venus fluorescence compared to a UTEX 250-A wild-type negative
532 control when measured in late-log phase heterotrophic cells (Figure 4E). Heterotrophic cells had
533 an average of 19-fold higher fluorescence than photoautotrophic cells, consistent with activation
534 of Venus expression by the *AMT1B-1* promoter in response to N-depletion (Figure 4E, F).
535

536 Finally, we checked for possible aneuploidies in *A. protothecoides* UTEX 25 and UTEX 2341,
537 and *A. symbiontica* CCAP 211/61, by mapping PacBio reads for each strain against one
538 representative haplotype from their respective genome assembly. For UTEX 2341 and CCAP
539 211/61, coverage estimates were consistent across all chromosomes, suggesting euploidy (Figure
540 S6). However, median read coverage for chromosome 1 (76x) of UTEX 25 was approximately
541 50% higher than median coverage across the rest of the genome (54x), implying trisomy (Figure
542 4G). We also noted that the existing *A. protothecoides* "0710" genome assembly exhibited
543 almost no variation relative to our UTEX 25 assembly, indicating that they are replicates of the
544 same strain. Surprisingly, when mapping the "0710" Illumina genomic reads to our UTEX 25
545 assembly we found no increase in coverage for chromosome 1, whereas coverage for
546 chromosomes 7 and 10 (131 and 134x, respectively) were ~50% higher than the median of 89x
547 across the rest of the genome (Figure 4G). Thus, it appears that the evolution of aneuploidy is
548 dynamic, occurring on short timescales of laboratory culture, at least in the UTEX 250 and
549 UTEX 25 genetic backgrounds. Transient aneuploidy may explain the rapid emergence of
550 whole-chromosome LOH, suggesting that it is a prominent evolutionary force in
551 *Auxenochlorella*. It is unclear whether any of the observed aneuploidies confer a selective
552 advantage; at least one additional copy of B03 has been maintained over decades of culture in
553 UTEX 250 and CCAP 211/7C, potentially indicating that it could be beneficial. There are 149
554 genes on B03, and analyses focusing on the effects of dosage on gene and protein expression will
555 be required to address this.



Legend on next page.

563 **Figure 4.** Aneuploidy in *Auxenochlorella.* A) UTEX 250-A PacBio read coverage for 20-kb windows
564 across the phased diploid genome assembly of UTEX 250-A. B) CCAP 211/7C ONT read coverage for
565 20-kb windows across the phased diploid genome assembly of UTEX 250-A. C) Schematic of
566 chromosome 3 copies (A03 and B03/C03) with green triangles marking the location of *AMT1B*. The
567 mutant locus features a gene cassette comprised of a *Venus* ORF and a 537 bp terminator sequence
568 containing the endogenous 3' UTR of *MLDP1* (*MAJOR LIPID DROPLET PROTEIN 1*), and a gene
569 cassette comprised of a *neoR* ORF flanked by the endogenous promoter/5' UTR and terminator/3' UTR
570 of *PGK1B* (*PHOSPHOGLYCERATE KINASE 1B*), which is targeted to the *AMT1B-1* allele of B03/C03.
571 *Venus* expression is under the control of the endogenous *AMT1B-1* promoter. D) PCR amplification of
572 *AMT1B-1* (B03/C03) and *AMT1B-2* (A03) 5' and 3' flanking regions using allele-specific primers in
573 wild-type UTEX 250-A and three representative *amt1B-1::Venus* transformants, with additional
574 amplification of *amt1B-1::Venus* integration in the transformants. E) Venus fluorescence detection in
575 late-log heterotrophic transformant cultures compared to wildtype UTEX 250-A negative control. F)
576 Comparison of chlorophyll autofluorescence (red) and Venus fluorescence (cyan) in wild-type UTEX
577 250-A versus a representative transformant under heterotrophic (het.) or photoautotrophic (phot.)
578 conditions using confocal fluorescence microscopy. G) UTEX 25 PacBio read coverage and strain "0710"
579 Illumina read coverage for 20-kb windows across a representative haplotype of the UTEX 25 genome
580 assembly.
581

582 **Gene family reduction and loss contribute to the streamlined *Auxenochlorella* genome**

583 To facilitate the development of *Auxenochlorella* as a model system, we aimed to produce a
584 reference quality gene annotation for the UTEX 250-A genome. We sequenced PacBio RNAseq
585 (IsoSeq) libraries from UTEX 250-A grown either autotrophically or with glucose, and multiple
586 Illumina RNAseq libraries from different growth conditions. Utilizing these datasets as evidence,
587 we annotated gene models independently for the A and B haplotypes using a combination of
588 gene predictors and extensive manual curation (see Methods). This yielded 7,515 and 7,519
589 protein-coding genes on the A and B haplotypes, respectively, with an additional 86 and 121
590 genes carried by transposable elements included in the annotation (Table 1). More than two
591 thirds of the gene models are derived from multiple full-length IsoSeq reads, enabling high
592 confidence identification of transcription start sites, terminators, and alternative splice variants.
593 Approximately 43% of genes are annotated with at least one alternative transcript, totaling
594 13,134 and 13,352 transcripts on the A and B haplotypes, respectively. Finally, 612 gene models
595 were corrected based on manual curation, including 43 bicistronic loci analyzed by Dueñas, et al.
596 (2024), which are frequently misannotated by automated approaches (Gallaher, et al. 2021).
597 Speaking to the compactness of the UTEX 250-A genome, 74% of genes encode a protein with a
598 Pfam domain (Table S6), a metric that is 25-50% in most algal genomes (Blaby-Haas and
599 Merchant 2019). Via orthology to *C. reinhardtii* proteins, one-third of genes were also assigned a

600 formal gene symbol (Table S7), which generally correspond to at least some level of functional
601 curation (Blaby, et al. 2014; Craig, et al. 2023).
602
603 Via synteny analysis, we were able to match ~98% of genes as allelic pairs between the A and B
604 haplotypes. Genes were given a unique ID featuring the haplotype and chromosome (e.g., A11,
605 B08) and a five-digit identifier that is shared between allelic pairs, e.g., despite being located on
606 rearranged chromosomes, UTEX250_A11.30955 and UTEX250_B08.30955 refer to alleles of
607 the same gene. As expected, heterozygosity between the A and B alleles at four-fold degenerate
608 sites in coding sequence is substantially higher (3.44%) than at zero-fold degenerate sites
609 (0.84%), although this nonetheless corresponds to more than 35,000 allelic variants that result in
610 amino acid differences. Only ~15% of the haplotype-specific genes encode a predicted protein
611 with a Pfam domain, and many of these genes may be evolutionarily young or misannotated
612 lncRNAs that carry a spurious ORF. Some haplotype-specific genes are the result of structural
613 variation between the A and B haplotypes, including a few copy number and presence/absence
614 variants.
615
616 Although we identified over 500 new genes in our UTEX 250-A genome relative to the existing
617 *A. protothecoides* 0710 assembly, the *Auxenochlorella* gene number remains lower than that of
618 most trebouxiophyte species, and substantially lower than that of chlorophyceaen species such as
619 *C. reinhardtii* (Table 1). We performed an orthology analysis of UTEX 250-A proteins against
620 the best quality chlorophyte annotations, arbitrarily using the A haplotype genes supplemented
621 with genes unique to the B haplotype. Figure 1A presents a phylogeny based on 669 single-copy
622 orthologs identified by this analysis, and Figure 5A shows the intersecting sets of orthogroups
623 among Chlorellales species, with *C. subellipsoidea* and *C. reinhardtii* as outgroups. Ignoring
624 species-specific orthogroups, the largest set corresponds to gene families present in all species
625 (2,974 orthogroups), many of which are presumably essential core-Chlorophyte gene families.
626 The second largest set features gene families specifically absent in the non-photosynthetic *P.*
627 *cutis* (520 orthogroups), most of which are expected to have functions related to photosynthesis
628 (Suzuki, et al. 2018). With respect to reduced gene content in *Auxenochlorella*, 297 orthogroups
629 are absent across all Chlorellales species, and 201 orthogroups are specifically absent in
630 *Auxenochlorella* UTEX 250-A and *P. cutis* but present in all other species.

631 For more detailed analyses, we compared the *Auxenochlorella* UTEX 250-A annotation to the
632 high-quality gene models of *Chlorella vulgaris* (Cecchin, et al. 2019) and *C. reinhardtii* (Craig,
633 et al. 2023), excluding any genes carried by transposable elements. Considering gene family loss,
634 there are 782 orthogroups that are absent from UTEX 250-A but present in both other species,
635 featuring 1,025 *C. vulgaris* genes and 1,212 *C. reinhardtii* genes. Considering gene family size,
636 there are 1,341 orthogroups that are present in all three species and have more than one gene in
637 at least one species, featuring 2,228 UTEX 250-A genes, 2,915 *C. vulgaris* genes, and 3,272 *C.*
638 *reinhardtii* genes. Approximately 56% of these orthogroups feature a single UTEX 250-A gene,
639 relative to 35% and 29% for *C. vulgaris* and *C. reinhardtii*, respectively (Figure 5B). UTEX 250-
640 A genes also contain fewer and shorter introns (mean 5.1 introns per gene, mean length 179 bp)
641 relative to *C. vulgaris* (means 7.3 and 206 bp) and *C. reinhardtii* (means 7.8 and 272 bp) (Figure
642 5C, D). Thus, the streamlined *Auxenochlorella* UTEX 250-A genome can be explained by both
643 complete loss or lower complexity of gene families, alongside lower repeat content (Table 1) and
644 shorter gene lengths via the presence of fewer and shorter introns.
645
646 Many of the lost genes can be linked to known metabolic pathways. Two distinguishing features
647 of *Auxenochlorella* are an absolute requirement for thiamine and the inability to grow on nitrate
648 as a N source (Kalina and Punčochárová 1987). Expectedly, the nitrate reductase gene *NIA1*, and
649 *THIC1*, encoding a hydroxymethylpyrimidine phosphate synthase necessary for thiamine
650 biosynthesis, are among the genes present in *C. vulgaris* and *C. reinhardtii* but absent in UTEX
651 250-A. All other components of the thiamine biosynthesis pathway are present in UTEX 250-A,
652 enabling thiamine production via salvage of breakdown products. Similarly, as reported by Gao,
653 et al. (2014), *Auxenochlorella* cannot utilize urea as an N source and genes responsible for urea
654 assimilation (*DUR1*, *DUR2*) and transport (*DUR3*) have been specifically lost in
655 *Auxenochlorella*. Several other transporters present in at least one copy in both *C. vulgaris* and
656 *C. reinhardtii* are also absent in UTEX 250-A, including the borate transporter *BOR1* and the
657 *SLT*-family sodium/sulfate co-transporters. Other transporter gene families feature a single copy
658 in UTEX 250-A, including the *CTR*-type copper ion transporters (three genes in each of *C.*
659 *vulgaris* and *C. reinhardtii*) and the *PTB*-family sodium/phosphate symporters (four genes in *C.*
660 *vulgaris*, at least nine in *C. reinhardtii*).
661

662 The reduction in gene complexity can also be seen by BUSCO (Benchmarking Universal Single
663 Copy Orthologs) (Manni, et al. 2021) analysis against the UTEX 250-A predicted proteins. Of
664 the 1,519 chlorophyte BUSCO genes, 1,488 (98%) were identified as complete, with only nine
665 (0.6%) complete and duplicated. One protein was called as fragmented, although it was manually
666 confirmed to be a correct gene model. Of the 30 missing BUSCOs, eight could be manually
667 recovered by cross-checking against the orthogroup analysis. The remaining 22 genes were also
668 undetected when searching BUSCOs against the UTEX 250-A genome, suggesting they are
669 biologically absent and not misannotated (Table S8). Most notably, nine of the 22 missing
670 BUSCOs are part of GreenCut2, a set of almost 600 genes conserved among select
671 photosynthetic eukaryotes and absent from non-photosynthetic species (Karpowicz, et al. 2011).
672
673 Intrigued by the potential loss of core photosynthesis genes, we next queried the UTEX 250-A
674 and *C. vulgaris* annotated proteins against the entire GreenCut2 dataset, using *C. reinhardtii*
675 proteins as a reference set. GreenCut2 is divided into four categories: genes conserved among
676 green algae and land plants (CGL, i.e., the green lineage), the green lineage and the red alga *C.*
677 *merolae* (CPL), the green lineage and diatoms (CGLD), and all of these groups (CPLD). While
678 12.5% of the GreenCut2 genes were not found in *C. vulgaris*, 18.0% were absent in UTEX 250-
679 A, corresponding to the specific loss of 33 photosynthesis-related genes despite retaining
680 photosynthesis (Figure 5E; Table S9). For example, UTEX 250-A encodes a reduced repertoire
681 of light-harvesting complex proteins. Three of the *LHCA* genes that encode the antenna system
682 of photosystem I are absent, namely *LHCA4, LHCA5* and *LHCA6*, but are present in all analyzed
683 trebouxiophytes including *C. vulgaris* (Figure 5E; Table S10). This suggests that the PSI-LHCI
684 supercomplex of *Auxenochlorella* likely consists of a single LHCI tetramer (with subunits
685 LHCA1, LHCA8, LHCA7 and LHCA3) and one LHCI dimer (LHCA2 and LHCA9), as in
686 *Dunaliella* species (class Chlorophyceae) (Caspy, et al. 2020; Liu, et al. 2024). The *PSBS*
687 (GreenCut2) and *LHCSR* (not GreenCut2) genes, encoding light-harvesting complex-like
688 subunits involved in nonphotochemical quenching (Allorent, et al. 2016), are also absent in
689 *Auxenochlorella* UTEX 250-A but present in *C. vulgaris* (Cecchin, et al. 2019).
690
691 Although only examples, the absence of GreenCut2 genes in *Auxenochlorella* demonstrates that
692 they are not essential for photosynthesis despite their deep conservation in photosynthetic

693 organisms. However, it is important to note that our approach of comparison to only *C.*
694 *reinhardtii* is limited and some false negatives are expected. Small, divergent proteins with
695 organellar targeting peptides would be especially susceptible to exclusion in the orthology
696 analysis. One example is RBCX (Table S9), a protein chaperone that is essential for proper
697 assembly of the RuBisCo large subunit into the octameric $RbcL_8$ intermediate, and subsequent
698 association with RuBisCo small subunit to form the holoenzyme (Saschenbrecker, et al. 2007;
699 Liu, et al. 2010). Eukaryotic RBCX proteins are found in two distinct isoforms; RBCX-I is most
700 similar to cyanobacterial rbcX, and RBCX-II is more divergent (Saschenbrecker, et al. 2007;
701 Bracher, et al. 2015). While *Arabidopsis thaliana* has both forms, the *C. reinhardtii RBCX2A*
702 and *RBCX2B* genes both encode RBCX-II proteins (Kolesiński, et al. 2011; Bracher, et al. 2015).
703 As in the examples above, we found an ortholog of the *C. reinhardtii* RBCX genes in *C.*
704 *vulgaris,* but no ortholog in UTEX 250-A, despite its chaperone function likely being
705 indispensable. However, manual searches using the RBCX-I proteins from *A. thaliana* and
706 *Synechocystis* sp. PCC 6803 did reveal a candidate protein (UTEX250_A12.36490, Figure S7A)
707 featuring an RbcX-like domain and a 56 amino acid N-terminus plastid transit peptide predicted
708 by TargetP-2.0 (Almagro Armenteros, et al. 2019). Interestingly, *C. vulgaris* also encodes a
709 highly similar protein, suggesting that both RBCX-I and RBCX-II isoforms are present (as in *A.*
710 *thaliana,* Figure S7B), whereas UTEX 250-A has only RBCX-I and *C. reinhardtii* only RBCX-
711 II. Comparisons of monomer AlphaFold structure models of the putative RBCX-I proteins from
712 UTEX 250-A and *C. vulgaris* with the corresponding crystal structures from *Synechocystis* sp.
713 PCC 6803 (Tanaka, et al. 2007) and *A. thaliana* (Kolesiński, et al. 2013) confirmed structural
714 homology, supporting the contention that they perform the same RbcL chaperone function
715 (Figure S7C).
716
717 The putative hybrid origin of UTEX 250 suggests the presence of a cryptic life cycle stage in
718 *Auxenochlorella* species, which may include meiosis of otherwise vegetatively diploid cells and
719 the formation of gametes (or at least implies an asexual transition between diploid and haploid
720 cells). Direct evidence for sexual reproduction in the Trebouxiophyceae is scarce, although
721 Fučiková, et al. (2015) identified a core set of nine meiosis genes that are mostly present in
722 trebouxiophyte genomes. Although *Chlorella* species have never been observed to form ciliated
723 gametes, Blanc, et al. (2010) identified orthologs of *C. reinhardtii* "CiliaCut" genes, which are

724 conserved among ciliated organisms but absent from species without cilia (Merchant, et al.
725 2007). Specifically, *Chlorella variablis* has genes encoding the outer and inner dynein arm
726 proteins, but lacks most or all of the genes encoding components of the intraflagellar transport
727 particle, radial spoke and central pair complex (Blanc, et al. 2010). The cilia genes encoded by
728 *C. variablis* significantly overlap with those present in the centric diatom *Thalassiosira*
729 *pseudonana* (i.e., the "CentricCut"), which produces male gametes with motile cilia (Moore, et
730 al. 2017). Cecchin, et al. (2019) confirmed this result, finding an even more substantial intersect
731 between "MotileCut" cilia genes common to *C. vulgaris* and *T. pseudonana*, but absent in
732 *Caenorhabditis elegans* (which produces sensory cilia, but not motile cilia), suggesting that
733 *Chlorella* species may carry the necessary repertoire of genes to produce motile ciliated gametes.
734 Most recently, Gazquez, et al. (2024) demonstrated that the ciliated life stage of the lichen-
735 forming *Trebouxia lynniae* corresponds to gametes, providing one of the few characterizations of
736 a sexual lifecycle in a trebouxiophyte alga.
737
738 Only 25.3% of CiliaCut genes from *C. reinhardtii* are present in the UTEX 250-A genome,
739 relative to 38.5% in *C. vulgaris* (Figure 5F). However, 58.3% of the genes that form the intersect
740 of the MotileCut and CentricCut are present in UTEX 250-A, relative to 66.7% in *C. vulgaris*,
741 corresponding to only four genes that are uniquely absent in UTEX 250-A (Table S11).
742 Fučiková, et al. (2015) identified four of the nine core meiosis genes in the *A. protothecoides*
743 0710 genome, with *HOPP1*, *REC8*, *MER3*, *MSH4* and *MSH5* absent. We identified *HOPP1* and
744 *REC8* in UTEX 250-A (Figure 5E), suggesting that they were previously missed due to
745 misannotation. *MSH4* and *MSH5*, homologs of the bacterial mismatch repair *MutS* family, are
746 absent in some sexual species such as *Plasmodium vivax*, although *MER3*, encoding a helicase
747 involved in the formation of meiotic crossovers (Altmannova, et al. 2023), is universally present
748 in sexual species (Fučiková, et al. 2015). We also identified UTEX 250-A orthologs of two
749 critical genes for syngamy, *GEX1*, which functions in nuclear fusion (Ning, et al. 2013), and
750 *HAP2*, which functions in gamete fusion (Fedry, et al. 2017). Thus, *Auxenochlorella* species may
751 be able to reproduce sexually, possibly via the formation of ciliated gametes, although as with
752 photosynthesis the set of genes functioning in these processes has been substantially reduced.
753

754 Finally, we analyzed a small number of gene families that have specifically expanded in UTEX
755 250-A. Import of glucose and galactose in *Chlorella kessleri* is facilitated by three genes
756 encoding H+/Hexose symporters (*HUP1, HUP2* and *HUP3*) (Stadler, et al. 1995), and three
757 paralogous genes that are most similar to *C. kessleri HUP2* were identified in the *A.*
758 *protothecoides* 0710 genome (Gao, et al. 2014). We identified five *HUP* genes on both
759 haplotypes of the UTEX 250-A genome, including a tandem array of three genes on
760 chromosome A12/B12 that is disrupted by contig breaks in the 0710 assembly that resulted in
761 misannotation. For orthogroups featuring genes from UTEX 250-A, *C. vulgaris* and *C.*
762 *reinhardtii* (Figure 5B), we identified 13 cases involving UTEX 250-A specific expansion via
763 multiple gene duplications (Table S12). A tandem array of four genes on A07/B12 encode
764 predicted ribonuclease (RNAse) T2 proteins, which have diverse functions including roles in
765 stress response and pathogen defense (MacIntosh 2011), although the function of the single
766 RNAse T2 gene from *C. reinhardtii* is unknown. A tandem array of six (A01) or five (B01)
767 genes encode homologs of *C. reinhardtii* ferrireductase *FRE1*, which reduces Fe(III) to Fe(II) for
768 Fe acclimation (Allen, et al. 2007). Four of the expanded orthogroups are associated with
769 glycosylation. Three of these correspond to gene families (*HPAT*, *RRA* and xyloglucanase-113)
770 that are specifically involved in arabinose *O*-glycosylation of the Hyp-rich glycoproteins
771 (HRGPs) that form the vegetative and zygote cell walls in *C. reinhardtii* (Joo, et al. 2017). With
772 respect to proteins potentially involved in cell wall formation, it is also notable that UTEX 250-A
773 carries multiple genes encoding polyketide synthases, as reported previously for *Auxenochlorella*
774 0710 (He, et al. 2016; Heimerl, et al. 2018). These include homologs of *PKS1* from *C.*
775 *reinhardtii*, encoding a type I polyketide synthase involved in formation of the zygote cell wall
776 (Heimerl, et al. 2018), and *LAP5* from *A. thaliana*, encoding a type III polyketide synthase
777 involved in the formation of sporopollenin (Kim, et al. 2010; He, et al. 2016). Further functional
778 characterization of these expanded gene families may provide new insights into the unique
779 biology of *Auxenochlorella* species.
780
781 Of the 146 and 655 genes that are uniquely present in the AHP lineage or only in UTEX 250-A,
782 respectively, only 9.9% encode predicted proteins with an identifiable domain. Most notably, 61
783 A haplotype genes are present in five AHP/*Auxenochlorella*-specific orthogroups that are
784 associated with pentapeptide repeat domains, and a further 39 genes are present in a single

similarly associated orthogroup that otherwise features only one gene in each of the Chlorellales species (including *P. cutis*). Genes encoding pentapeptide repeat proteins are abundant in many bacterial genomes, particularly those of some cyanobacteria, although their functions remain elusive (Vetting, et al. 2006). Two highly conserved GreenCut2 proteins feature a pentapeptide repeat domain of unknown function, *CPLD59* and *CPLD44*, and are localized to the thylakoid lumen (Kieselbach, et al. 1998; Schubert, et al. 2002). Some mycobacterial pentapeptide repeat proteins adopt a fold that mimics double-stranded DNA and can compete with DNA in DNA-protein interactions (Feng, et al. 2021). Although their functions are unclear, these 100 genes are dispersed throughout the UTEX 250-A genome, and their repetitive nature can apparently mediate genomic rearrangements via ectopic recombination (Figure 2B, C).

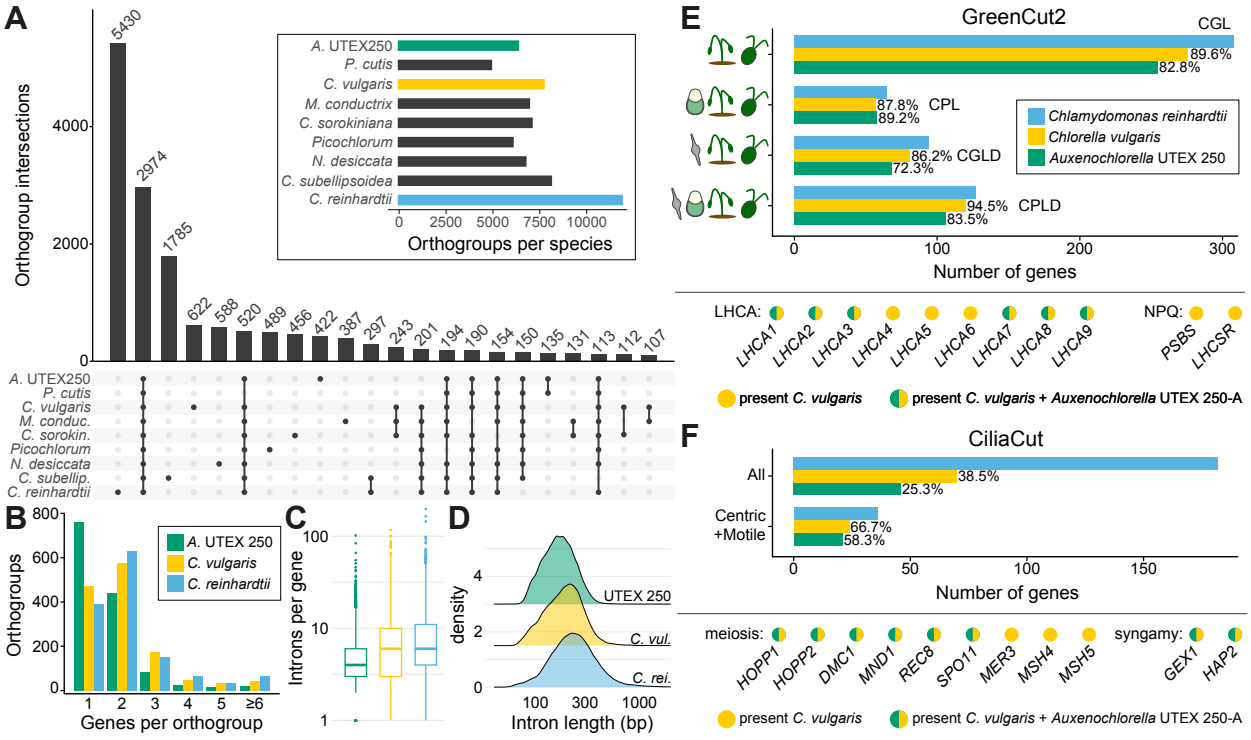


Figure 5. Gene family loss and reduction in *Auxenochlorella* UTEX 250-A. A) Upset plot showing intersection between the presence and absence of orthogroups among Trebouxiophyte algae, with *C. reinhardtii* as an outgroup. Only intersects with more than 100 orthogroups are shown. B) Number of genes per orthogroup for orthogroups that are present in each of *Auxenochlorella* UTEX 250-A, *C. vulgaris* and *C. reinhardtii*, and have more than one gene in at least one of the species. C) Introns per gene for multi-exonic genes. D) Intron length distributions. E) Presence of GreenCut2 genes in UTEX 250-A and *C. vulgaris* relative to *C. reinhardtii*. Light-harvesting complex genes of the photosystem I antenna (LHCA) and genes functioning in nonphotochemical quenching (NPQ) are highlighted; *PSBS* and *LHCSR* are present as two and three paralagous genes in *C. reinhardtii* and single orthologs in *C. vulgaris* (see Table S10). The CGL, CPL, CGLD and CPLD divisions are defined in the main text. F) Presence of CiliaCut, meiosis and syngamy genes in UTEX 250-A and *C. vulgaris* relative to *C. reinhardtii*.

808 **Periodic cytosine and adenine methylation supports widespread internal antisense**
809 **transcription**
810 The DNA modification 5-methylcytosine (5mC) is widespread across all domains of life and has
811 several essential functions (Mattei, et al. 2022), although relatively little is understood about
812 cytosine methylation in green algae. In *C. variablis*, gene bodies are extensively methylated at
813 CpG dinucleotides, whereas promoter regions are hypomethylated (Zemach, et al. 2010).
814 Although 5mC is generally restricted to centromeres and subtelomeres in *C. reinhardtii* (Lopez,
815 et al. 2015; Chaux-Jukic, et al. 2021; Craig, et al. 2023), extensive gene body methylation was
816 recently shown for the chlorophyceaen alga *T. obliquus* (Biondi, et al. 2024), suggesting that this
817 pattern may be widespread in chlorophytes. In the extremely compact genomes of prasinophytes
818 such as *Ostreococcus lucimarinus* and *Micromonas pusilla*, CpG dinucleotides in gene bodies
819 are also highly methylated, although 5mC occurs in periodic clusters that correspond to the
820 linker DNA between nucleosomes (Huff and Zilberman 2014). Notably, periodic adenine
821 methylation (6mA), mostly at ApT dinucleotides, was also discovered in *C. reinhardtii* (Fu, et al.
822 2015). Periodicity of 6mA is also due to specific methylation of nucleosome linkers, although
823 unlike for 5mC, the adenine methylation is present around transcription start sites (TSS) and not
824 in the gene body. Romero Charria, et al. (2024) demonstrated that periodic adenine methylation
825 of linker DNA downstream of TSSs is also present in both prasinophytes and *C. variablis*, as
826 well as in other distantly related taxa, suggesting that it is an ancient feature of eukaryotic
827 genomes. The function of 6mA in this context is not known; the biophysical properties of this
828 modification may play some role in promoting transcription (Bochtler and Fernandes 2021), and
829 6mA could also facilitate the coordination of promoter chromatin marks such as H3K4me3
830 (Romero Charria, et al. 2024).
831
832 We used the ONT reads from CCAP 211/7C (equivalent to UTEX 250, Figure 3A) to call both
833 5mC at CpG sites and 6mA at ApT sites across the UTEX 250-A genome. Using only the high-
834 confidence TSSs from IsoSeq-based gene models, periodicity of both modifications was evident
835 (Figure 6A). Cytosine methylation is essentially absent at sites adjacent to the majority of TSSs,
836 and universal at the hypermethylated peaks within gene bodies (on average 5mC is called at
837 ~95% of CpG sites at the highest point of the peaks). Conversely, two periodic peaks of 6mA are
838 present downstream of the TSS, although far fewer ApT sites are methylated and the signal is

839 noisier between the peaks (Figure S8). Assuming that the peaks correspond to nucleosome
840 linkers, it appears that the first two linkers downstream of the TSS are specifically associated
841 with adenine methylation, whereas cytosine methylation is present at the third linker and reaches
842 a stable maximum at subsequent linkers. These patterns closely resemble those of the
843 prasinophytes, as opposed to *C. variablis* where 6mA exhibits periodicity but 5mC is relatively
844 constant across the gene body (Romero Charria, et al. 2024). Huff and Zilberman (2014)
845 proposed that 5mC periodicity contributes to nucleosome positioning, and hypothesized that this
846 could be an adaptation to extremely compact nuclei and small cell sizes. Our results suggest this
847 unusual genomic architecture may have been present in the common ancestor of the core-
848 Chlorophyta (Trebouxiophyceae and Chlorophyceae in Figure 1A), or alternatively that it re-
849 evolved on the lineage leading to *Auxenochlorella*.
850
851 Aside from the biological importance of periodic methylation, these signatures also provide
852 evidence for promoter regions, similar to H3K4me3 ChIP-seq datasets that only exist for a
853 limited number of green algae (Strenkert, et al. 2022; Petroll, et al. 2024). Figure 6B shows a
854 representative locus featuring four genes where CpG hypomethylation clearly corresponds to the
855 regions downstream of the IsoSeq-based TSSs. Notably, one of these genes
856 (UTEX250_A08.19925) is a lncRNA transcribed from a bidirectional hypomethylated promoter
857 located within the CiliaCut gene *FAP57* (UTEX250_A08.19920), which is essential for the
858 correct assembly of inner dynein arms (Lin, et al. 2019). While manually curating genes in the
859 UTEX 250-A genome, we noticed multiple similar cases and subsequently performed a thorough
860 search for lncRNA genes that are supported by multiple IsoSeq reads (Table 1). After manual
861 verification, 165 protein-coding genes were identified with internal bidirectional promoters that
862 produced antisense lncRNAs on at least one of the two haplotypes (Table S13). For genes such
863 as *FAP57*, it is unclear whether the truncated isoforms transcribed from the internal promoter
864 could be functional, while for many others the truncated isoforms clearly have no coding
865 capacity. We also identified 615 protein-coding genes that exhibit substantial overlap with an
866 antisense lncRNA (>30% of the transcript length) transcribed from an internal, or immediately
867 adjacent, unidirectional promoter.
868

869 To explore the potential functions of widespread antisense transcription in UTEX 250-A, we
870 analyzed the predicted functions of these genes. Manual curation of the proteins encoded by the
871 genes with internal birdirectional promoters found that potential functions associated with DNA,
872 such as repair, chromatin and replication, were most abundant, followed by functions related to
873 post-translational regulation, cilia, and the cell wall. A gene ontology (GO) enrichment analysis
874 focusing on the full set of 780 genes with evidence for antisense transcription supported these
875 observations. The most significant terms for biological processes were all related to DNA repair
876 and recombination, including the repair of double-strand breaks by homologous recombination
877 (Figure 6C, Table S14). Indeed, the list includes many fundamental DNA repair genes, including
878 *RAD51*, *RAD21*, the RECQ helicases *RECQ1* and *RECQ5*, and the DNA polymerases *POLK1*
879 and *POLQ2* (Table S13). For cellular compartments, the two most significant terms, "condensed
880 nuclear chromosome" and "cohesin complex", are both related to mitosis, meiosis and
881 recombination, whereas the next two significant terms are related to cilia. The most significant
882 molecular function term, "calcium ion binding" (Table S14), is also associated with cilia genes.
883 Indeed 14 of the CiliaCut genes (including *FAP57*) exhibit antisense transcription, alongside five
884 of the eight UTEX 250-A core meiotic and syngamy genes (Figure 5F), representing a
885 significant enrichment (χ^2= 32.05, p = 1.5x10^{-8}). The nuclear fusion gene *GEX1* represents an
886 extreme case where all of the IsoSeq reads at the locus are antisense (Figure S9).
887
888 Based on these results, it is interesting to speculate that antisense lncRNAs may play a regulatory
889 role in *Auxenochlorella*. In synchronized cultures of *C. reinhardtii*, some DNA repair genes are
890 specifically upregulated at the transition from dark to light (e.g., the ribonucleotide reductase
891 subunit *RIR2L*, which features an antisense lncRNA in UTEX 250-A), suggesting a possible role
892 in photodamage response (Strenkert, et al. 2019). Many human DNA repair genes are also
893 upregulated under genotoxic stresses such as UV radiation and ROS (Christmann and Kaina
894 2013). In *C. reinhardtii*, several of the core meiotic genes are among a small set of genes that
895 lack H3K4me3 at their TSSs during vegetative growth, supporting expression restricted to the
896 sexual cycle (Strenkert, et al. 2022). It therefore seems plausible that genes overlapped by
897 antisense lncRNAs must be specifically translated at certain points of the cell and life cycles, or
898 under specific conditions. While this gene set is enriched for putative functions in DNA repair
899 and sex, we also observe this phenomenon for many genes that are predicted to have unrelated

900 functions, suggesting that this may be a general regulatory mechanism. Antisense lncRNAs can
901 regulate expression of associated protein-coding genes by several mechanisms (Werner, et al.
902 2024). For example, in compact genomes such as *S. cerevisiae*, transcriptional interference can
903 occur via the physical collision of polymerase complexes on opposite strands (Prescott and
904 Proudfoot 2002; Hobson, et al. 2012). Experimental characterization, including differential
905 expression analyses of the protein-coding and lncRNA gene pairs at these loci, will be required
906 to test this hypothesis. However, we are aware of no similar observations in green algae,
907 suggesting widespread antisense transcription may have emerged as an unusual gene regulatory
908 mechanism during *Auxenochlorella* evolution.
909
910
911
912
913
914

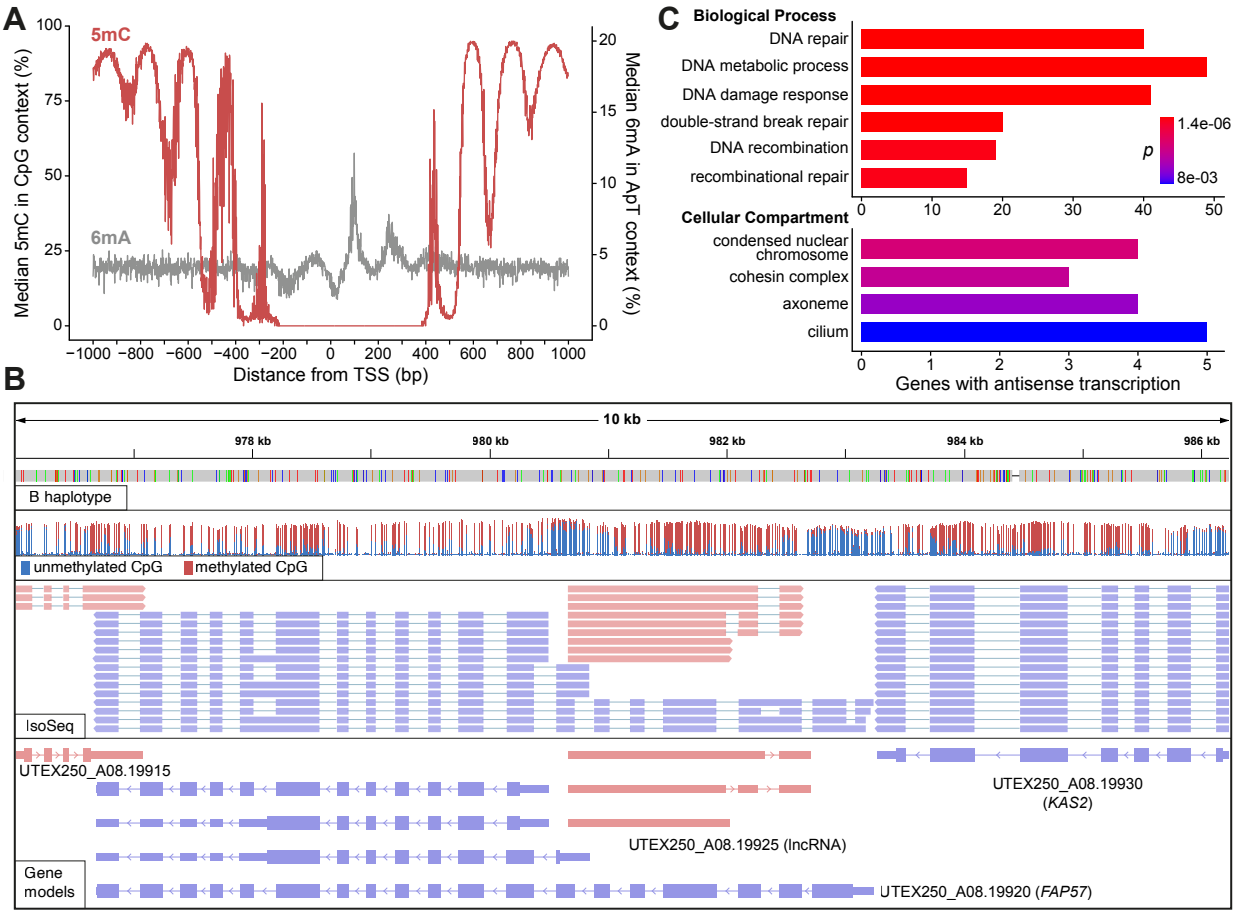


Figure 6. Promoter methylation and antisense lncRNAs in *Auxenochlorella* UTEX 250-A. A) Median per site 5mC (red) and 6mA (grey) around transcription start sites. B) IGV screenshot showing example of internal bidirectional promoter producing antisense lncRNA and truncated isoform for *FAP57* (UTEX_A08.19920). The top row shows the B haplotype of the UTEX 250-A genome mapped against the A haplotype, with colored bands corresponding to single nucleotide polymorphisms. The second row shows read coverage of ONT reads from CCAP 211/7C where only CpG sites are shown. The proportion of red to blue corresponds to the proportion of 5mC calls in the ONT reads. The third row shows IsoSeq data, where pink reads are on the forward strand, and blue reads the reverse. The final row shows the annotated gene models, where the thin lines correspond to introns, intermediate lines to UTRs, and thick lines to CDS. C) GO term enrichment analysis for 780 genes overlapped by antisense lncRNAs. See Table S14 for full results.

934 **CONCLUSIONS**

935 Species that reproduce clonally as diploids may experience specific phenomena including
936 allodiploid hybridization, mitotic recombination, loss-of-heterozygosity and aneuploidy.
937 Although these processes can act as strong evolutionary forces, for example in speciation or
938 adaptation to new environments, most of our knowledge is currently based upon yeasts and
939 filamentous fungi. In green algae, observations consistent with these phenomena have been
940 reported sporadically, including trisomy in *C. primus* (Lemieux, et al. 2019), and chromosome-
941 scale LOH and possible hybridization in *T. obliquus* (Biondi, et al. 2024). Here, we find that
942 each of these phenomena occur in the trebouxiophyte genus *Auxenochlorella*, demonstrating the
943 generality of these important molecular and evolutionary processes in vegetatively diploid
944 eukaryotes.
945
946 *Auxenochlorella* UTEX 250 is an allodiploid hybrid of two closely related species, *A.*
947 *protothecoides* and *A. symbiontica*, that are differentiated by extensive karyotypic variation and
948 possibly ecological niches. Following hybridization, the parental haplotypes have presumably
949 been shuffled by mitotic crossovers, and homogenized via LOH events mediated by
950 recombination, DNA repair, and potentially transient aneuploidy. Post hybridization, the
951 outcomes of these events can be driven by selection to resolve genomic incompatibles between
952 the divergent parental genomes in a process termed genome stabilization (Sipiczki 2008;
953 Gabaldón 2020). It is currently unclear to what extent the crossovers, LOH events and
954 aneuploidies of UTEX 250 are involved in stabilization, although at least some of the LOH
955 events that occurred prior to laboratory culture were presumably beneficial (e.g., at rDNA
956 arrays). However, our observations from authentic strains of *A. protothecoides* and *A.*
957 *symbiontica* suggest that LOH and aneuploidy are not only associated with hybridization, and are
958 generally expected to be prominent evolutionary forces in *Auxenochlorella* that can occur on the
959 timescale of laboratory culture. It is also unclear whether the sampling of the hybrid UTEX 250
960 was a rare chance event, or whether it may represent a successful hybrid lineage that is persistent
961 and potentially abundant (albeit likely sterile, assuming a sexual life cycle). Given the number of
962 *Auxenochlorella* strains in culture, and the potential to isolate new strains (Asker and Awad
963 2019; Chen, et al. 2024), it would be interesting to apply population genomics approaches to

964 investigate the prevalence of LOH, aneuploidy, and possibly other hybridization events, in the
965 genus.
966
967 *Auxenochlorella* has great potential to be developed as an algal model for fundamental discovery
968 research and bioengineering (Goold, et al. 2024). The resources produced here will serve as a
969 foundation for functional genomics and systems biology analyses, as well as the design of
970 transformation constructs for molecular manipulation of the genome. The streamlined nature of
971 *Auxenochlorella* genomes and gene content are also expected to be advantageous in this
972 direction. Gene families containing multiple members in green algae, and especially in the main
973 algal model *C. reinhardtii*, more frequently feature only a single gene in *Auxenochlorella* UTEX
974 250-A. This reduced genetic redundancy is expected to improve the likelihood of achieving
975 informative phenotypes following gene knockout or knockdown, analogous to the lower
976 redundancy of regulatory gene families in the liverwort *Marchantia polymorpha* relative to *A.*
977 *thaliana* and other angiosperms (Bowman, et al. 2017). Although hundreds of genes and gene
978 families have been entirely lost, including many genes that are expected to play roles in
979 photosynthesis, these genes could be re-introduced to the *Auxenochlorella* genome to probe their
980 functions in other species. Mutations in essential genes involved in photosynthesis are also
981 possible due to the facultative autotrophy of the species. Genomic manipulation also presents
982 opportunities to address fundamental questions related to genetic mechanisms, as demonstrated
983 by the dissection of the mechanism underlying the translation of bicistronic genes (Dueñas, et al.
984 2024). *Auxenochlorella* species are highly oleaginous and have GRAS status, and other genomic
985 features reported here could also find utility in bioengineering. The introduction of artificial
986 chromosomes is a powerful technique that enables multiple novel traits to be stacked on a single
987 locus (Birchler, et al. 2024). If the short AT-rich regions present on *Auxenochlorella*
988 chromosomes can be confirmed to function as centromeres, it may be possible to maintain
989 minichromosomes by introducing an AT-rich region, similar to the synthetic episomes that can
990 be maintained in diatoms (Diner, et al. 2017). Alternatively, the aneuploidy of UTEX 250-A and
991 UTEX 25 could be exploited. We demonstrated allele-specific knock-in to a trisomic
992 chromosome, and presumably it would be possible to sequentially manipulate this chromosome
993 with minimal impact on fitness. Collectively, the unique biology of *Auxenochlorella* among

994 green algae presents many exciting prospects for discovery in plant biology and utilization in
995 bioengineering.
996
997 **METHODS**
998 **Strains, media and growth conditions**
999 *Auxenochlorella* strains UTEX 250, UTEX 25 and UTEX 2341 were originally obtained from
1000 the University of Texas Culture Collection of Algae. A single colony of UTEX 250 was isolated
1001 and used in all experimental work (i.e., UTEX 250-A). *Auxenochlorella symbiontica* CCAP
1002 211/61 and *Auxenochlorella* CCAP 211/7C were obtained from the Culture Collection of Algae
1003 and Protozoa. Strains were typically grown in TAP (for DNA extraction) or HPv1 (for RNA)
1004 culture media using the trace element recipe of Kropat, et al. (2011). HPv1 is a modified version
1005 of TP that replaces 20 mM Tris with 20 mM HEPES. Standard liquid cultures were generally
1006 grown under constant light (100 μmol photons m^{-2} sec^{-1}) and shaken at 200 rpm. Specific growth
1007 conditions were used for preparing RNA from UTEX 250-A for the generation of IsoSeq and
1008 RNAseq datasets. These included a range of photoautotrophic, mixotrophic or heterotrophic
1009 conditions, different light conditions, CO_2 levels, and elemental adjustments to media (Table
1010 S15).
1011
1012 **Extraction and sequencing of nucleic acids**
1013 UTEX 250-A DNA was extracted at UC Davis DNA Technologies Core from a frozen cell pellet
1014 and used to prepare and sequence PacBio HiFi and OmniC linked-read libraries. Individual
1015 IsoSeq libraries were produced from RNA extracted from cultures grown in autotrophic replete
1016 and heterotrophic with 2% (w/v) glucose conditions (HPv1 media), and each library was
1017 sequenced on a Sequel II SMRT cell at UC Davis. For all remaining conditions (Table S15),
1018 Illumina RNAseq with poly(A) selection was performed on a Novoseq 6000 platform at UC
1019 Davis, generating stranded 150 bp paired-end reads.
1020
1021 UTEX 25, UTEX 2341, CCAP 211/61 and CCAP 211/7C were cultured mixotrophically in TAP
1022 medium. High molecular weight DNA was extracted using a modified version of a CTAB and
1023 Phenol:Chloroform protocol developed for *C. reinhardtii* (Chaux, et al. 2024). Cell pellets were

1024 ground in liquid nitrogen in a mortar and pestle prior to the CTAB incubation. Size selection was
1025 performed using the PacBio short read eliminator (SRE) kit following user instructions.
1026
1027 DNA from UTEX 25, UTEX 2341 and CCAP 211/61 was prepared in a multiplexed library
1028 along with five other unrelated samples and sequenced on a single PacBio Sequel II SMRT cell
1029 at QB3 Genomics, UC Berkeley. ONT sequencing for CCAP 211/61 and CCAP 211/7C was
1030 performed using the Ligation Sequencing Kit V14, two independent R10.1.4 flow cells, and a
1031 MinION Mk1B device following user instructions.
1032
1033 **Genome assembly of *Auxenochlorella* UTEX 250-A**
1034 The UTEX 250-A PacBio HiFi reads were first subsampled to remove any reads shorter than 18
1035 kb, resulting in ~0.98 Gb of total sequence. An initial phased diploid assembly was produced by
1036 passing the filtered HiFi reads and Omni-C reads to Hifiasm v0.16.1-r375 (Cheng, et al. 2022),
1037 which was run with default parameters. The initial assembly consisted of two sets of phased
1038 contigs, haplotype 1 and 2, featuring 17 and 21 nuclear contigs, respectively.
1039
1040 Several steps of manual assessment and assembly were then performed to achieve a final
1041 telomere-to-telomere phased diploid assembly. First, haplotype 1 and haplotype 2 contigs were
1042 aligned to each other using minimap2 v2.23-r1117 (Li 2021) with the parameter "-x asm20". The
1043 raw PacBio HiFi reads were also mapped against each haplotype assembly with minimap2 ("-x
1044 map-hifi"). Assembly and read-based alignments were manually inspected using IGV v2.16.2
1045 (Robinson, et al. 2011). Due to the rearranged nature of the parental haplotypes, we found that
1046 the two automated haplotype sets did not individually represent a complete copy of the haploid
1047 genome (i.e. some genomic regions were present twice in one haplotype assembly, and absent
1048 from the other). We therefore transferred contigs between the two haplotype assemblies to arrive
1049 at two sets of contigs that each represented a complete haploid genome, which were arbitrarily
1050 labelled haplotype A and B.
1051
1052 Via comparison of the haplotype A and B assemblies it was possible to identify contigs that
1053 could potentially be fused i.e. at locations where the ends of two contigs from one haplotype
1054 mapped to a single assembled region from the other haplotype. Based on manual inspection of

1055 the raw reads, four pairs of contigs were successfully fused. This process involved either
1056 removing redundant sequence from one of the contig ends, or adding additional sequence that
1057 filled the gap between the two contigs. Where gap filling was required, haplotype-specific raw
1058 reads spanning the gap were extracted, aligned with MAFFT v7.490 (Katoh and Standley 2013),
1059 and reduced to a single consensus sequence that was subsequently trimmed and inserted between
1060 the two contigs. Following a similar approach, the termini of several chromosomes were
1061 manually extended to reach the telomeric repeats. Uniquely mapped reads that exhibited soft-
1062 clipped bases extending beyond the assembled chromosome were extracted, aligned, and reduced
1063 to a consensus sequence that was trimmed and appended to the chromosome. These manual steps
1064 resulted in the A and B haplotypes each being represented by 12 gapless chromosomes that
1065 either terminated in telomeric repeats or rDNA arrays.
1066
1067 Two additional contigs represented the third copies of the trisomic chromosomes, C03 and C05.
1068 The contig corresponding to chromosome C05 was manually truncated at the location of the
1069 internal telomere (see Figure S5), since the original contig was misassembled as an exact
1070 duplicate of chromosome B05. Finally, eight short contigs were removed from the assembly; six
1071 contigs entirely featured rDNA that is already represented as truncated arrays on the assembled
1072 chromosomes, and two contigs were identified as duplicate redundant sequences that are
1073 similarly represented in the chromosomal assembly. The assembly and phasing of the entire final
1074 assembly was manually verified by examining read alignments to each haplotype.
1075
1076 The plastome and mitogenome were each identified as single linear contigs in the Hifiasm
1077 assembly. Circularization was performed manually by identifying and removing redundant
1078 sequence from one end of each of the organellar contigs.
1079
1080 **Oxford Nanopore basecalling and methylation analyses**
1081 ONT raw reads derived from CCAP 211/61 were basecalled using Dorado v0.3.2
1082 (https://github.com/nanoporetech/dorado), which was run in super accuracy basecalling mode
1083 using the command “basecaller” and the model “dna_r10.4.1_e8.2_400bps_sup@v4.2.0”. ONT
1084 reads for CCAP 211/7C were bascalled using Dorado v0.6.0 using the options
1085 “sup,5mCG_5hmCG” or “sup,6mA” to perform basecalling in super accuracy mode with the

1086 detection of 5mC and 5-hydroxymethylcytosine (5hmC) calls at CpG dinucleotides, or 6mA
1087 calls, respectively. The resulting BAM files were converted to FASTQ using samtools 1.18-1
1088 (Danecek, et al. 2021) using the "-T '*'" option to retain base modifications. The resulting
1089 FASTQ files were mapped against the UTEX 250-A assembly using minimap2 with
1090 recommended parameters for ONT 10.4.1 chemistry ("-x map-ont -k19 -w19 -U50,500 -g10k -a
1091 -y"). The 5mC status of each CpG site was then called using modkit v0.2.7
1092 (https://github.com/nanoporetech/modkit) with the command "pileup --preset traditional", which
1093 aggregates methylation calls on both strands and returns only 5mC information. The 6mA status
1094 of each ApT site was similarly called using the command "pileup --motif AT 0". Sites with less
1095 than 5 mapped reads were removed. For each site, the fraction of modified bases at each site was
1096 then extracted, and the median for each nucleotide position relative to the TSS was calculated
1097 across all genes that were predicted using IsoSeq data (see below).
1098
1099 **Genome assembly of *Auxenochlorella protothecoides* UTEX 25 and UTEX 2341, and**
1100 ***Auxenochlorella symbiontica* CCAP 211/61**
1101 The UTEX 25, UTEX 2341 and CCAP 211/61 genome assemblies were produced using Hifiasm
1102 v0.19.8-r603. For UTEX 25 and UTEX 2341, all PacBio HiFi reads were passed to Hifiasm,
1103 which was run with default parameters. For CCAP 211/61, PacBio HiFi reads were
1104 supplemented with ONT reads, which were passed to Hifiasm with the flag "--ul".
1105
1106 When run without linked-read data (e.g. Omni-C), Hifiasm produces a primary haploid
1107 assembly, as well as a diploid assembly consisting of two partially phased haplotypes. For each
1108 strain, we first mapped the two (pseudo)haplotypes of the diploid assembly to each other using
1109 minimap2 ("-x asm20"), and manually inspected the alignments to confirm that all contigs
1110 exhibited a one-to-one syntenic relationship between the two haplotypes (i.e. they are
1111 homologous without interchromosomal rearrangements as observed in UTEX 250-A). We then
1112 produced a representative haploid assembly by mapping the primary contigs to the UTEX 250-A
1113 genome assembly with minimap2 ("-x asm20") and manually inspecting alignments. For CCAP
1114 211/61, 14 primary contigs corresponded to 12 nuclear chromosomes and two circular organellar
1115 genomes. For both UTEX 25 and UTEX 2341, one pair of nuclear contigs and one pair of
1116 plastome contigs were each fused to produce the final assemblies, with all other major contigs

1117 corresponding to near-complete chromosomes. Contig fusion and circularization of the
1118 organellar contigs was performed using the manual assembly approaches described above. As
1119 with the UTEX 250-A assembly, additional short contigs featuring rDNA or redundant sequence
1120 were removed.
1121
1122 **Quantification of heterozygosity and divergence among strains**
1123 To quantify heterozygosity (i.e. individual-level genetic diversity) between the A and B
1124 haplotypes of the UTEX 250-A assembly, the haplotype A and B chromosomes were aligned
1125 against each other using minimap2 ("-x asm20"). Variants segregating between the A and B
1126 haplotypes were then identified using the minimap2 script paftools.js and the parameters "call -l
1127 2000 -L 10000" (i.e., a minimum alignment length to compute coverage of 2 kb, and a minimum
1128 alignment length to call variants of 10 kb). Chromosomes were divided in to 10 kb windows, and
1129 heterozygosity was calculated based on the number of single nucleotide variants (mapping
1130 quality of 60) relative to the number of alignable sites per window. Windows with <2000 aligned
1131 sites were excluded. Zero-fold and four-fold degenerate sites were extracted from the coding
1132 sequence of the primary transcript of each A haplotype gene model using the degenotate tool
1133 (https://github.com/harvardinformatics/degenotate). All overall heterozygosity values were
1134 reported after removing homozygous genomic tracts that are putatively the product of LOH (see
1135 below).
1136
1137 Since the UTEX 25, UTEX 2341, and CCAP 211/61 genomes were assembled to the haploid-
1138 level, a different approach was used to quantify heterozygosity. Raw PacBio HiFi reads for each
1139 strain were mapped against the A haplotype of the UTEX 250-A assembly using minimap2 ("-x
1140 map-hifi"). DeepVariant v1.6.0 (Poplin, et al. 2018) was then used to call variants directly from
1141 the PacBio HiFi read alignments for each strain independently, using the parameter "model_type
1142 PACBIO". Heterozygosity was calculated in 10 kb windows by identifying heterozygous single
1143 nucleotide variants segregating among the reads of the strain in question. Callable sites for each
1144 strain were determined by mapping the haploid genome assembly of the relevant strain against
1145 the A haplotype of UTEX 250 with minimap2 ("-x asm20") and identifying alignment blocks
1146 with paftools.js ("call -l 2000 -L 10000"). Genetic divergence for a given strain was calculated in
1147 10 kb windows as the proportion of variant sites relative to the UTEX 250-A haplotype A.

1148 Divergence was averaged over both haplotypes for each strain (i.e., a site where only one of the
1149 two haplotypes varied relative to the UTEX 250-A reference contributed half as much to
1150 divergence as a site where both haplotypes varied). Divergence was also calculated relative to
1151 the UTEX 250 B haplotype using the same approach.
1152
1153 Neighbor joining trees for nuclear genome, plastome and mitogenome sites were produced using
1154 MEGA v11.0.13 (Tamura, et al. 2021) with the Tamura Nei substitution model and default
1155 parameters. For the organelles, either the plastome or mitogenome assembly of each strain was
1156 aligned to the UTEX 250-A plastome or mitogenome assembly with minimap2 ("-x asm20"),
1157 and variant and invariant sites were called with paftools.js ("call -l 2000 -L 10000"). To be
1158 included in the nuclear genome analysis, a site had to meet two criteria, i) to be on a UTEX 250-
1159 A chromosome that had not undergone any reciprocal crossovers, and ii) to be in a region that
1160 was not affected by a LOH event in any of the strains. These criteria were enforced to ensure that
1161 ancestral evolutionary relationships among haplotypes were recovered, and resulted in the
1162 analysis of 651,399 sites on chromosome A11. Variants were extracted from either the
1163 paftools.js analysis (for variants segregating between the UTEX 250 A and B haplotypes) or the
1164 DeepVariant analyses (for variants segregating within and between the other strains). Variants in
1165 UTEX 25, UTEX 2341 and CCAP 211/61 were phased relative to their respective genome
1166 assemblies (note that these may represent pseudohaplotypes with a small number of phase
1167 changes per chromosome).
1168
1169 **Loss-of-heterozygosity events**
1170 LOH events in the UTEX 250-A genome were called from the alignment of the A and B
1171 haplotype assemblies and paftools.js analysis (see above). LOH events were arbitrarily assigned
1172 to regions that featured no heterozygous sites over a stretch of 1 kb or more. Events were
1173 designated as terminal if they extended to a chromosome end, or interstitial if they fell within a
1174 chromosome.
1175
1176 LOH events were similarly called as invariant regions of at least 1 kb for UTEX 25, UTEX 2341
1177 and CCAP 211/61 based on heterozygous variants identified from the DeepVariant analysis.
1178 CCAP 211/7C heterozygous sites and LOH events were also identified using DeepVariant, ran

1179 with the parameter “model_type ONT_R104” and supplied with an alignment of the raw ONT
1180 reads against the UTEX 250-A haplotype A assembly (minimap2 “-x map-ont”).
1181
1182 **Aneuploidy events and allele-specific transformation**
1183 The raw reads for all available strains were first mapped against the relevant genome assembly to
1184 calculate coverage. For PacBio datasets, the UTEX 250-A reads were mapped against the phased
1185 UTEX 250-A assembly, whereas UTEX 25, UTEX 2341 and CCAP 211/61 reads were mapped
1186 against the representative haplotype from their respective haploid assemblies, all using minimap2
1187 (“-x map-hifi”). CCAP 211/7C ONT reads were mapped against the UTEX 250-A assembly
1188 using minimap2 (“-x map-ont”), and the “0710” reads were mapped against the UTEX 25
1189 assembly using bwa mem (Li 2013). The coverage at each nucleotide was calculated using the
1190 samtools “depth” command, and the average coverage per 20-kb sliding window was calculated.
1191
1192 Transformation, fluorescence measurements and microscopy were performed following Dueñas,
1193 et al. (2024). Full details are provided in Supplementary Note 1, and a detailed description of all
1194 templates, primers and plasmids used in this study can be found in Tables S16 and S17.
1195
1196 ***Auxenochlorella* UTEX 250-A gene model annotation**
1197 Gene models were annotated using a hybrid approach that utilized both the IsoSeq and RNAseq
1198 datasets, followed by extensive manual annotation. The A and B haplotypes of the UTEX 250
1199 assembly were annotated independently, except for the entirely homozygous chromosomes (04,
1200 09 and 10).
1201
1202 IsoSeq reads were pre-processed using the IsoSeq3 pipeline v3.8.1
1203 (https://github.com/PacificBiosciences/IsoSeq). Primers and poly(A) tails were first removed
1204 from circular consensus sequencing (CCS) reads. The resulting reads from the two libraries were
1205 then combined and mapped against the UTEX 250-A genome assembly using minimap2 (“-x
1206 splice:hq”). These read mappings were used to partition the reads into sets that either primarily
1207 mapped to the A haplotype (>=20 mapping quality), primarily mapped to the B haplotype, or
1208 mapped to neither with high quality (<20 mapping quality; this set was enriched for reads
1209 derived from homozygous genomic regions). Read clustering was then performed individually

1210 on each read set ("isoseq3 cluster –use-qvs"), to avoid clustering reads that derived from
1211 different alleles of the same gene. The three clustered read sets were then re-combined and
1212 aligned collectively to the UTEX 250-A assembly using pbmm2 v1.9.0 ("--preset ISOSEQ").
1213 Unique isoforms were called using the isoseq3 collapse command with the options "--do-not-
1214 collapse-extra-5exons --max-5p-diff 20 --max-3p-diff 20". These options were selected
1215 following manual assessment of the quality of the IsoSeq data, and effectively enabled
1216 alternative TSSs (within a different exon of a longer isoform, or >20 bp away if on the same
1217 exon) and termination sites (>20 bp if on the same exon) to be preserved as independent
1218 isoforms.
1219
1220 Gene models were predicted from the isoseq3 output using SQANTI3 v5.1 in QC mode with
1221 default parameters (Pardo-Palacios, et al. 2024), which utilizes GeneMarkS-T (Tang, et al. 2015)
1222 to predict ORFs from isoforms. The isoseq3 pipeline groups neighboring genes under a single
1223 gene model if any of their isoforms overlap on the same strand, which is a frequent occurrence in
1224 compact genomes such as that of *Auxenochlorella*. We thus had to "decouple" the isoforms of
1225 incorrectly merged genes based on comparison of the protein-coding coordinates of each isoform
1226 (i.e. isoforms which did not share coding sequence were separated to independent gene models).
1227 Next, we removed any isoforms that were supported by <5 reads or that were derived from <5%
1228 of the total reads associated with a given gene. This step minimized the number of false isoforms
1229 derived from 5' truncated reads. Finally, the primary isoform for each gene was selected based
1230 on a combination of read abundance and length criteria. For the majority of genes with multiple
1231 isoforms, the isoform with the longest coding sequence was also the most abundant and was
1232 deemed the primary isoform (where multiple isoforms had the same coding length but differed in
1233 their UTRs, the most abundant was selected). For genes where the longest CDS was not the most
1234 abundant, it was deemed the primary isoform if it was supported >=50% of the reads associated
1235 with the most abundant isoform. If this was not the case, the most abundant isoform was selected
1236 if its coding length was >=80% of the isoform with the longest coding length. In all other cases,
1237 the longest isoform was selected to avoid overly short isoforms being selected. since they may be
1238 overrepresented in IsoSeq data.
1239

1240 We noticed that many ORFs were truncated due to predicted initiation at an internal ATG codon.
1241 To perform ORF extension, we first built a model of the Kozak sequence surrounding annotated
1242 initiation sites from primary transcripts that had no in-frame upstream ATG codons (i.e. ORFs
1243 that could not be extended), based on the assumption that most of these annotations represented
1244 true initiation sites. The strength of the Kozak sequence surrounding in-frame upstream ATG
1245 codons for the remaining transcripts was then calculated by comparison to this Kozak model (see
1246 Cross (2015) and Dueñas, et al. (2024) for details). ORF extension was performed if an upstream
1247 in-frame start codon existed with a Kozak score greater than the 25th percentile of the scores
1248 from which the Kozak model was built, or if the upstream Kozak score was simply higher than
1249 that of the annotated start codon.
1250
1251 In parallel to the above annotation, we used high confidence gene models derived from the
1252 photoautotrophic IsoSeq dataset to train the AUGUSTUS v3.3.2 gene predictor tool (Stanke, et
1253 al. 2008), following the protocols described by Hoff and Stanke (2019). Pre-processing and gene
1254 model annotation using isoseq3 and SQANTI3 was performed as above. The training set of gene
1255 models was produced by extracting the most abundant isoform supported by a minimum of 10
1256 reads from gene models derived from IsoSeq reads that mapped uniquely to the A haplotype.
1257 Accurate AUGUSTUS training requires information on the noncoding DNA flanking the
1258 training set genes. To avoid including unannotated genes in the flanking regions, we performed a
1259 preliminary de novo gene annotation using BRAKER v2.1.2 (Hoff, et al. 2016). Each RNAseq
1260 dataset was mapped against the UTEX 250-A assembly using STAR v2.7.10a ("--twopassMode
1261 Basic --alignIntronMax 5000") (Dobin, et al. 2013), and subsequently passed to the braker.pl
1262 script using default parameters. We then extracted BRAKER gene models that had no overlap
1263 with the IsoSeq-derived training gene set using bedtools v2.30.0 (Quinlan and Hall 2010). The
1264 combined dataset of IsoSeq training genes and nonoverlapping BRAKER genes was then used to
1265 compute the flanking region length (1,366 bp), and the final training set was produced by
1266 extracting the training set genes with flanking DNA. This dataset was then split randomly into a
1267 gene set for training (N=3,444 genes) and for testing the trained parameters (N=1,000). Training
1268 was performed using both the standard hidden-Markov model and the conditional random field
1269 (CRF) approach, which can result in more accurate gene prediction but is less robust to errors in
1270 the training dataset. We selected the CRF parameters based on increased sensitivity when applied

1271 to the test dataset. A final RNAseq-based gene model prediction was then performed by passing
1272 the RNAseq BAM files (see above) to BRAKER using the trained CRF parameters ("—
1273 useexisting").
1274
1275 The IsoSeq and AUGUSTUS-based gene models were then semi-manually integrated. IsoSeq-
1276 based models were given precedence, and an AUGUSTUS model was automatically retained if it
1277 did not share any same-strand coding exons with an IsoSeq model. AUGUSTUS models that
1278 exactly matched an IsoSeq model (with the exception of the first and/or last exon, which was
1279 often incorrectly predicted by AUGUSTUS) were immediately filtered. All other AUGUSTUS
1280 models, which could have partial overlap with an IsoSeq model, were manually curated using all
1281 IsoSeq and RNAseq evidence.
1282
1283 Annotation of the plastome and mitogenome was performed using the GeSeq webserver (Tillich,
1284 et al. 2017), which was provided with existing annotations from *A. protothecoides* "0710" and
1285 *Prototheca wickerhamii* (Yan, et al. 2015). Additional genes were manually annotated by
1286 searching for long ORFs in the unannotated sequence.
1287
1288 **Repeat annotation**
1289 Preliminary annotation of interspersed repeats was performed by running RepeatModeler 2.0.5
1290 (Flynn, et al. 2020) on the UTEX 250-A genome with the parameter "–LTRStruct". Consensus
1291 sequences for specific transposon families were then manually curated following Goubert, et al.
1292 (2022) (see Table S3, Figure S1). The manually curated models were then combined with the
1293 automated models and passed to RepeatMasker v4.1.6 (https://github.com/Dfam-
1294 consortium/RepeatMasker) ("–gccalc"). Tandem Repeats Finder v4.10.0 was run on the UTEX
1295 250-A genome to identify tandem repeats using the parameters "2 7 7 80 10 50 2000 -f -d -m -
1296 ngs". Total repeat density was calculated by combining the output of RepeatMasker and TRF.
1297 Transposon genes (see Table 1, Figure S1) were identified and marked in the GFF3 annotation
1298 by intersecting the coordinates of manually curated transposons and gene model coding
1299 sequence, followed by manual inspection.
1300

1301 Repeat densities for the genome assemblies in Table 1 were similarly performed by combining
1302 the output of RepeatModeler/RepeatMasker and TRF. RepeatModeler was not run for *A.*
1303 *protothecoides* 0710, which was repeat masked using the UTEX 250-A repeat library (see
1304 above), nor *C. reinhardtii*, for which an extensive repeat library exists (Craig, et al. 2021).
1305
1306 **Gene family and phylogenetic analyses**
1307 Gene family analyses were performed using the 19 species shown in Figure 1A. Protein
1308 sequences for each species were reduced to their primary isoforms, and where possible genes on
1309 the organelle genomes were removed. Annotated transposon proteins were included if known.
1310 For UTEX 250-A, all A haplotype genes were arbitrarily selected and supplemented by genes
1311 that are putatively unique to the B haplotype (see below). For *Picochlorum* sp. BPE23, the genes
1312 from only haplotype A were arbitrarily used. Metadata for protein datasets are presented in Table
1313 S18.
1314
1315 OrthoFinder v2.5.5 was run on the 19 protein sets using the Diamond search method ("-S
1316 diamond_ultra_sens") (Emms and Kelly 2019). Orthogroups featuring a known transposon
1317 protein were filtered out (the results of which are reflected in Table 1). GreenCut2 and CiliaCut
1318 orthologs were extracted for UTEX 250-A and *C. vulgaris* by comparison to *C. reinhardtii* (i.e.,
1319 from the "orthologues" output directory of OrthoFinder). For the specific genes shown in Figure
1320 E and F (LHCA, NPQ, meiosis and syngamy), the gene trees for the underlying orthogroups
1321 were manually validated.
1322
1323 The species tree in Figure 1A was produced from a concatenated alignment of the 669 single-
1324 copy orthologs identified by OrthoFinder. For each ortholog, proteins were aligned using
1325 MAFFT v7.525 ("–maxiterate 1000 –localpair"), and the resulting alignments were trimmed
1326 using trimAl v1.4.rev22 ("–automated1") (Capella-Gutiérrez, et al. 2009). The trimmed
1327 alignments were concatenated and a maximum likelihood phylogeny was produced using IQ-
1328 TREE v2.3.0 ("-m MFP -bb 1000") (Minh, et al. 2020). For Figure 1B, rDNA sequences were
1329 accessed from NCBI (Table S1), aligned with MAFFT ("L-INS-i"), manually trimmed, and
1330 passed to IQ-TREE as above.
1331

1332 **Gene IDs and functional annotation**

1333 Allelic gene pairs were determined by running SynChro (Drillon, et al. 2014), a tool for detecting
1334 syntenic orthologs between genomes, on the A and B haplotype genomes and gene models.
1335 Genes were assigned a unique 5-digit number starting with "00005" for the first gene on
1336 chromosome A01 (i.e., UTEX250_A01.00005), with each subsequent gene along the A
1337 chromosomes receiving a number that increased in increments of five (i.e.,
1338 UTEX250_A01.00010, UTEX250_A01.00015, etc.). Genes on the B haplotype that had been
1339 identified as part of an allelic pair were assigned the gene number of the A gene, regardless of
1340 their order on the B chromosomes. Genes that were determined to be unique to haplotype B were
1341 given new gene numbers that extended beyond the last number that was used on haplotype A.
1342 Genes on the plastome were given unique numbers starting at "80000" (e.g.,
1343 UTEX250_xCp.80000), and genes on the mitogenome starting at "90000" (e.g.,
1344 UTEX250_xMt.90000). Isoforms were distinguished by adding a number to the gene ID, with
1345 the primary isoform labelled as "1", e.g., UTEX250_A01.00030.1.
1346

1347 In addition to IDs, genes were also assigned a unique name or "gene symbol" where available
1348 following the nomenclature of *C. reinhardtii* (Craig, et al. 2023). Gene symbols were transferred
1349 based on the gene orthology relationships between *C. reinhardtii* and *Auxenochlorella* UTEX
1350 250-A determined by OrthoFinder ("Orthologues" output directory). For genes with 1:1
1351 orthology the *C. reinhardtii* symbol was directly transferred to UTEX 250-A. For multicopy
1352 gene families in *C. reinhardtii* that are represented by a single gene in UTEX 250-A, either the
1353 symbol with the lowest number was used (e.g., *FOX1*/*FOX2* gene pair orthologous to *FOX1* in
1354 UTEX 250-A), or if letters are used, they were removed (e.g., *CMT1A*/*CMT1B* gene pair
1355 orthologous to *CMT1* in UTEX 250-A). For the reverse situation, letters were appended to the
1356 UTEX 250-A gene symbol where possible (e.g., *LPAAT1* orthologous to *LPAAT1A*/*LPAAT1B*
1357 gene pair in UTEX 250-A). Genes on the nuclear genome in *C. reinhardtii* but on an organelle
1358 genome in UTEX 250-A were changed accordingly (e.g., *CHLI1* orthologous to *chlI* in UTEX
1359 250-A plastome). Several gene symbols were manually curated, including a small number of
1360 genes that are absent from *C. reinhardtii* (e.g., the photosystem I subunit gene *psaM*), and all of
1361 the histone genes (for which orthology relationships are difficult to determine). All gene symbol
1362 relationships between *C. reinhardtii* and UTEX 250-A are presented in Table S7.

1363 Protein domains were identified by running InterProScan v5.67-99.0 ("-dp -goterms").
1364
1365 **Identification of antisense long noncoding RNA genes**
1366 LncRNA genes were annotated based on IsoSeq data using two approaches. First, IsoSeq-based
1367 gene models that had initially been annotated as protein-coding but had no orthology to proteins
1368 from other species, nor a recognized Pfam or CDD domain, were reassigned as lncRNAs if they
1369 overlapped coding sequence of another gene that did meet one of these criteria. Second, IsoSeq
1370 gene models supported by at least 5 reads for which no ORF had been annotated by SQANTI3
1371 were added. The first category of genes had been assigned gene IDs as described above, and
1372 retained these IDs as lncRNAs instead of protein-coding genes. New gene IDs starting from
1373 "50000" and increasing in increments of 5 were introduced for the second category of lncRNA
1374 genes.
1375
1376 The coordinates of protein-coding genes with either a homolog in at least one other species or a
1377 recognized domain were then intersected with the lncRNA gene coordinates, and those with
1378 antisense overlap spanning at least 30% of the protein-coding gene length were selected for
1379 manual curation. Validated gene models with antisense lncRNAs were then divided into those
1380 with internal birdirectional promoters, or those with internal or immediately adjacent promoters
1381 without evidence for birdirectional activity. Gene ontology enrichment analysis was performed
1382 using topGO (Alexa and Rahnenfuhrer 2023) using the GO term output from InterProScan and
1383 the classic Fisher Test option.
1384
1385
1386
1387
1388
1389
1390
1391
1392
1393

DATA AVAILABILITY

All sequencing data associated with *Auxenochlorella* UTEX 250-A is available from NCBI under the BioSample SAMN45466464. The UTEX 250-A genome and gene annotations are specifically available under the BioProjects PRJNA1195245 (Haplotype A) and PRJNA1195244 (Haplotype B).

FUNDING

This work was supported by the Department of Energy (DOE) Office of Science, Biological and Environmental Research program under award no. DE-SC0023027 (to SSM and JLM). MAD was supported, in part, by the Newton Graduate Fellowship in Synthetic Biology (QB3-Berkeley).

REFERENCES

Alexa A, Rahnenfuhrer J. 2023. topGO: enrichment analysis for gene ontology. Version R package version 2.54.0.

Allen MD, del Campo JA, Kropat J, Merchant SS. 2007. *FEA1*, *FEA2*, and *FRE1*, encoding two homologous secreted proteins and a candidate ferrireductase, are expressed coordinately with *FOX1* and *FTR1* in iron-deficient *Chlamydomonas reinhardtii*. Eukaryot Cell 6:1841-1852.

Allorent G, Lefebvre-Legendre L, Chappuis R, Kuntz M, Truong TB, Niyogi KK, Ulm R, Goldschmidt-Clermont M. 2016. UV-B photoreceptor-mediated protection of the photosynthetic machinery in *Chlamydomonas reinhardtii*. Proc Natl Acad Sci U S A 113:14864-14869.

Almagro Armenteros JJ, Salvatore M, Emanuelsson O, Winther O, von Heijne G, Elofsson A, Nielsen H. 2019. Detecting sequence signals in targeting peptides using deep learning. Life Sci Alliance 2.

Altmannova V, Firlej M, Muller F, Janning P, Rauleder R, Rousova D, Schaffler A, Bange T, Weir JR. 2023. Biochemical characterisation of Mer3 helicase interactions and the protection of meiotic recombination intermediates. Nucleic Acids Res 51:4363-4384.

Andersen MP, Nelson ZW, Hetrick ED, Gottschling DE. 2008. A genetic screen for increased loss of heterozygosity in *Saccharomyces cerevisiae*. Genetics 179:1179-1195.

Arriola MB, Velmurugan N, Zhang Y, Plunkett MH, Hondzo H, Barney BM. 2018. Genome sequences of *Chlorella sorokiniana* UTEX 1602 and *Micractinium conductrix* SAG 241.80: implications to maltose excretion by a green alga. Plant J 93:566-586.

Asker D, Awad TS. 2019. Isolation and characterization of a novel lutein-producing marine microalga using high throughput screening. Food Res Int 116:660-667.

1434
1435 Bakuła Z, Siedlecki P, Gromadka R, Gawor J, Gromadka A, Pomorski JJ, Panagiotopoulou H,
1436 Jagielski T. 2021. A first insight into the genome of *Prototheca wickerhamii*, a major causative
1437 agent of human protothecosis. Bmc Genomics 22:168.
1438
1439 Bao W, Jurka J. 2013. Homologues of bacterial TnpB_IS605 are widespread in diverse
1440 eukaryotic transposable elements. Mob DNA 4:12.
1441
1442 Barten R, van Workum DM, de Bakker E, Risse J, Kleisman M, Navalho S, Smit S, Wijffels RH,
1443 Nijveen H, Barbosa MJ. 2022. Genetic mechanisms underlying increased microalgal
1444 thermotolerance, maximal growth rate, and yield on light following adaptive laboratory
1445 evolution. BMC Biol 20:242.
1446
1447 Becker SA, Spreafico R, Kit JL, Brown R, Likhogrud M, Fang W, Posewitz MC, Weissman JC,
1448 Radakovits R. 2020. Phased diploid genome sequence for the fast-growing microalga
1449 *Picochlorum celeri*. Microbiol Resour Announc 9.
1450 Benites LF, Bucchini F, Sanchez-Brosseau S, Grimsley N, Vandepoele K, Piganeau G. 2021.
1451 Evolutionary genomics of sex-related chromosomes at the base of the green lineage. Genome
1452 Biol Evol 13.
1453
1454 Biondi TC, Kruse CPS, Koehler SI, Kwon T, Davis AK, Eng W, Kunde Y, Gleasner CD, Mak
1455 KTY, Polle J, et al. 2024. The telomere-to-telomere, gapless, phased diploid genome and
1456 methylome of the green alga *Scenedesmus obliquus* UTEX 3031 reveals significant
1457 heterozygosity and genetic divergence of the haplotypes. Algal Research 79.
1458
1459 Birchler JA, Kelly J, Singh J, Liu H, Zhang Z, Char SN, Sharma M, Yang H, Albert PS, Yang B.
1460 2024. Synthetic minichromosomes in plants: past, present, and promise. Plant J.
1461
1462 Blaby IK, Blaby-Haas CE, Tourasse N, Hom EF, Lopez D, Aksoy M, Grossman A, Umen J,
1463 Dutcher S, Porter M, et al. 2014. The Chlamydomonas genome project: a decade on. Trends
1464 Plant Sci 19:672-680.
1465
1466 Blaby-Haas CE, Merchant SS. 2019. Comparative and functional algal genomics. Annu Rev
1467 Plant Biol 70:605-638.
1468
1469 Blanc G, Agarkova I, Grimwood J, Kuo A, Brueggeman A, Dunigan DD, Gurnon J, Ladunga I,
1470 Lindquist E, Lucas S, et al. 2012. The genome of the polar eukaryotic microalga *Coccomyxa*
1471 *subellipsoidea* reveals traits of cold adaptation. Genome Biology 13.
1472
1473 Blanc G, Duncan G, Agarkova I, Borodovsky M, Gurnon J, Kuo A, Lindquist E, Lucas S,
1474 Pangilinan J, Polle J, et al. 2010. The Chlorella variabilis NC64A genome reveals adaptation to
1475 photosymbiosis, coevolution with viruses, and cryptic sex. Plant Cell 22:2943-2955.
1476
1477 Bochtler M, Fernandes H. 2021. DNA adenine methylation in eukaryotes: Enzymatic mark or a
1478 form of DNA damage? Bioessays 43:e2000243.
1479

1480 Bowman JL, Kohchi T, Yamato KT, Jenkins J, Shu S, Ishizaki K, Yamaoka S, Nishihama R,
1481 Nakamura Y, Berger F, et al. 2017. Insights into land plant evolution garnered from the
1482 *Marchantia polymorpha* genome. Cell 171:287-304 e215.
1483
1484 Bracher A, Hauser T, Liu C, Hartl FU, Hayer-Hartl M. 2015. Structural analysis of the Rubisco-
1485 assembly chaperone RbcX-II from *Chlamydomonas reinhardtii*. PLoS One 10:e0135448.
1486
1487 Brooks G, Franklin S, Avila J, Decker SM, Baliu E, Rakitsky W, Piechocki J, Zdanis D, Norris
1488 LM. 2024. Microalgal flour. United States US12037630B2, filed November 5, 2021, and issued
1489 July 16, 2024.
1490
1491 Bulankova P, Sekulić M, Jallet D, Nef C, van Oosterhout C, Delmont TO, Vercauteren I, Osuna-
1492 Cruz CM, Vancaester E, Mock T, et al. 2021. Mitotic recombination between homologous
1493 chromosomes drives genomic diversity in diatoms. Curr Biol 31:3221-3232 e3229.
1494
1495 Calhoun S, Bell TAS, Dahlin LR, Kunde Y, LaButti K, Louie KB, Kuftin A, Treen D, Dilworth
1496 D, Mihaltcheva S, et al. 2021. A multi-omic characterization of temperature stress in a
1497 halotolerant *Scenedesmus* strain for algal biotechnology. Commun Biol 4:333.
1498
1499 Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated
1500 alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972-1973.
1501
1502 Caspy I, Malavath T, Klaiman D, Fadeeva M, Shkolnisky Y, Nelson N. 2020. Structure and
1503 energy transfer pathways of the *Dunaliella Salina* photosystem I supercomplex. Biochim
1504 Biophys Acta Bioenerg 1861:148253.
1505
1506 Cecchin M, Marcolungo L, Rossato M, Girolomoni L, Cosentino E, Cuine S, Li-Beisson Y,
1507 Delledonne M, Ballottari M. 2019. *Chlorella vulgaris* genome assembly and annotation reveals
1508 the molecular basis for metabolic acclimation to high light conditions. Plant J 100:1289-1305.
1509
1510 Charron G, Marsit S, Henault M, Martin H, Landry CR. 2019. Spontaneous whole-genome
1511 duplication restores fertility in interspecific hybrids. Nat Commun 10:4126.
1512
1513 Chaux F, Agier N, Eberhard S, Xu Z. 2024. Extraction and selection of high-molecular-weight
1514 DNA for long-read sequencing from *Chlamydomonas reinhardtii*. PLoS One 19:e0297014.
1515
1516 Chaux-Jukic F, O'Donnell S, Craig RJ, Eberhard S, Vallon O, Xu Z. 2021. Architecture and
1517 evolution of subtelomeres in the unicellular green alga *Chlamydomonas reinhardtii*. Nucleic
1518 Acids Res 49:7571-7587.
1519
1520 Chen H, Sosa A, Chen F. 2024. Growth and cell size of microalga *Auxenochlorella*
1521 *protothecoides* AS-1 under different trophic modes. Microorganisms 12.
1522
1523 Cheng H, Jarvis ED, Fedrigo O, Koepfli KP, Urban L, Gemmell NJ, Li H. 2022. Haplotype-
1524 resolved assembly of diploid genomes without parental data. Nat Biotechnol 40:1332-1335.
1525

1526 Christmann M, Kaina B. 2013. Transcriptional regulation of human DNA repair genes following
1527 genotoxic stress: trigger mechanisms, inducible responses and genotoxic adaptation. Nucleic
1528 Acids Res 41:8403-8420.
1529
1530 Craig RJ, Gallaher SD, Shu S, Salome PA, Jenkins JW, Blaby-Haas CE, Purvine SO, O'Donnell
1531 S, Barry K, Grimwood J, et al. 2023. The Chlamydomonas Genome Project, version 6:
1532 Reference assemblies for mating-type plus and minus strains reveal extensive structural mutation
1533 in the laboratory. Plant Cell 35:644-672.
1534
1535 Craig RJ, Hasan AR, Ness RW, Keightley PD. 2021. Comparative genomics of *Chlamydomonas*.
1536 Plant Cell 33:1016-1041.
1537
1538 Cross FR. 2015. Tying down loose ends in the *Chlamydomonas* genome: functional significance
1539 of abundant upstream open reading frames. G3 (Bethesda) 6:435-446.
1540 da Roza PA, Muller H, Sullivan GJ, Walker RSK, Goold HD, Willows RD, Palenik B, Paulsen
1541 IT. 2024. Chromosome-scale assembly of the streamlined picoeukaryote *Picochlorum* sp.
1542 SENEW3 genome reveals Rabl-like chromatin structure and potential for C(4) photosynthesis.
1543 Microb Genom 10.
1544
1545 Dale AL, Feau N, Everhart SE, Dhillon B, Wong B, Sheppard J, Bilodeau GJ, Brar A, Tabima
1546 JF, Shen D, et al. 2019. Mitotic Recombination and Rapid Genome Evolution in the Invasive
1547 Forest Pathogen Phytophthora ramorum. mBio 10.
1548
1549 Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T,
1550 McCarthy SA, Davies RM, et al. 2021. Twelve years of SAMtools and BCFtools. Gigascience
1551 10:giab008.
1552
1553 Darienko T, Pröschold T. 2015. Genetic variability and taxonomic revision of the genus
1554 *Auxenochlorella* (Shihira et Krauss) Kalina et Puncocharova (Trebouxiophyceae, Chlorophyta). J
1555 Phycol 51:394-400.
1556
1557 Del Cortona A, Jackson CJ, Bucchini F, Van Bel M, D'Hondt S, Skaloud P, Delwiche CF, Knoll
1558 AH, Raven JA, Verbruggen H, et al. 2020. Neoproterozoic origin and multiple transitions to
1559 macroscopic growth in green seaweeds. Proc Natl Acad Sci U S A 117:2551-2559.
1560
1561 Diner RE, Nodding CM, Lian NC, Kang AK, McQuaid JB, Jablanovic J, Espinoza JL, Nguyen
1562 NA, Anzelmatti MA, Jansson J, et al. 2017. Diatom centromeres suggest a mechanism for
1563 nuclear DNA acquisition. Proc Natl Acad Sci U S A 114:E6015-E6024.
1564
1565 Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras
1566 TR. 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29:15-21.
1567
1568 Drillon G, Carbone A, Fischer G. 2014. SynChro: a fast and easy tool to reconstruct and
1569 visualize synteny blocks along eukaryotic chromosomes. PLoS One 9:e92621.
1570

1571 Dueñas MA, Craig RJ, Gallaher SD, Moseley JL, Merchant SS. 2024. Leaky ribosomal scanning
1572 enables tunable translation of bicistronic ORFs in green algae. bioRxiv.
1573 https://doi.org/10.1101/2024.07.24.605010
1574
1575 Dutta A, Dutreux F, Schacherer J. 2021. Loss of heterozygosity results in rapid but variable
1576 genome homogenization across yeast genetic backgrounds. Elife 10.
1577
1578 Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative
1579 genomics. Genome Biol 20:238.
1580
1581 Fedry J, Liu Y, Pehau-Arnaudet G, Pei J, Li W, Tortorici MA, Traincard F, Meola A, Bricogne
1582 G, Grishin NV, et al. 2017. The ancient gamete fusogen HAP2 is a eukaryotic class II fusion
1583 protein. Cell 168:904-915 e910.
1584 Feng L, Mundy JEA, Stevenson CEM, Mitchenall LA, Lawson DM, Mi K, Maxwell A. 2021.
1585 The pentapeptide-repeat protein, MfpA, interacts with mycobacterial DNA gyrase as a DNA T-
1586 segment mimic. Proc Natl Acad Sci U S A 118.
1587
1588 Figueroa-Martinez F, Nedelcu AM, Smith DR, Reyes-Prieto A. 2015. When the lights go out:
1589 the evolutionary fate of free-living colorless green algae. New Phytologist 206:972-982.
1590
1591 Fischer G, Liti G, Llorente B. 2021. The budding yeast life cycle: More complex than
1592 anticipated? Yeast 38:5-11.
1593
1594 Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF. 2020.
1595 RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl
1596 Acad Sci U S A 117:9451-9457.
1597
1598 Foflonker F, Mollegard D, Ong M, Yoon HS, Bhattacharya D. 2018. Genomic analysis of
1599 *Picochlorum* species reveals how microalgae may adapt to variable environments. Mol Biol Evol
1600 35:2702-2711.
1601
1602 Franklin S, Somanchi A, Wee J, Rudenko G, Moseley JL, Rakitsky W, Zhao X, Bhat R. 2019.
1603 Tailored oils produced from recombinant oleaginous microorganisms. United States
1604 US20190100780A1, filed September 7, 2018, and issued April 4, 2019.
1605
1606 Fu Y, Luo GZ, Chen K, Deng X, Yu M, Han D, Hao Z, Liu J, Lu X, Dore LC, et al. 2015. *N*6-
1607 methyldeoxyadenosine marks active transcription start sites in *Chlamydomonas*. Cell 161:879-
1608 892.
1609
1610 Fučiková K, Pažoutová M, Rindi F. 2015. Meiotic genes and sexual reproduction in the green
1611 algal class Trebouxiophyceae (Chlorophyta). J Phycol 51:419-430.
1612
1613 Fulnečková J, Hasíková T, Fajkus J, Lukešová A, Eliáš M, Sýkorová E. 2012. Dynamic
1614 evolution of telomeric sequences in the green algal order Chlamydomonadales. Genome Biology
1615 and Evolution 4:248-264.
1616

1617 Gabaldón T. 2020. Hybridization and the origin of new yeast lineages. FEMS Yeast Res 20.
1618
1619 Gallaher SD, Craig RJ, Ganesan I, Purvine SO, McCorkle SR, Grimwood J, Strenkert D, Davidi
1620 L, Roth MS, Jeffers TL, et al. 2021. Widespread polycistronic gene expression in green algae.
1621 Proc Natl Acad Sci U S A 118.
1622
1623 Gallone B, Steensels J, Mertens S, Dzialo MC, Gordon JL, Wauters R, Thesseling FA,
1624 Bellinazzo F, Saels V, Herrera-Malaver B, et al. 2019. Interspecific hybridization facilitates
1625 niche adaptation in beer yeast. Nat Ecol Evol 3:1562-1575.
1626
1627 Gao C, Wang Y, Shen Y, Yan D, He X, Dai J, Wu Q. 2014. Oil accumulation mechanisms of the
1628 oleaginous microalga *Chlorella protothecoides* revealed through its genome, transcriptomes, and
1629 proteomes. BMC Genomics 15:582.
1630 Garagna S, Page J, Fernandez-Donoso R, Zuccotti M, Searle JB. 2014. The Robertsonian
1631 phenomenon in the house mouse: mutation, meiosis and speciation. Chromosoma 123:529-544.
1632
1633 Gazquez A, Bordenave CD, Montero-Pau J, Pérez-Rodrigo M, Marco F, Martínez-Alberola F,
1634 Muggia L, Barreno E, Carrasco P. 2024. From spores to gametes: A sexual life cycle in a
1635 symbiotic Trebouxia microalga. Algal Research 84.
1636
1637 Goold HD, Moseley JL, Lauersen KJ. 2024. The synthetic future of algal genomes. Cell Genom
1638 4:100505.
1639
1640 Goubert C, Craig RJ, Bilat AF, Peona V, Vogan AA, Protasio AV. 2022. A beginner's guide to
1641 manual curation of transposable elements. Mob DNA 13:7.
1642
1643 Guo J, Jian J, Wang L, Xiong L, Lin H, Zhou Z, Sonnenschein EC, Wu W. 2022. Genome
1644 sequences of two strains of *Prototheca wickerhamii* provide insight into the protothecosis
1645 evolution. Front Cell Infect Microbiol 12:797017.
1646
1647 Harris EH. 2001. *Chlamydomonas* as a Model Organism. Annu Rev Plant Physiol Plant Mol Biol
1648 52:363-406.
1649
1650 He X, Dai J, Wu Q. 2016. Identification of sporopollenin as the outer layer of vell eall in
1651 microalga *Chlorella protothecoides*. Front Microbiol 7:1047.
1652
1653 Heimerl N, Hommel E, Westermann M, Meichsner D, Lohr M, Hertweck C, Grossman AR,
1654 Mittag M, Sasso S. 2018. A giant type I polyketide synthase participates in zygospore maturation
1655 in *Chlamydomonas reinhardtii*. Plant J 95:268-281.
1656
1657 Hirakawa MP, Martinez DA, Sakthikumar S, Anderson MZ, Berlin A, Gujja S, Zeng Q, Zisson
1658 E, Wang JM, Greenberg JM, et al. 2015. Genetic and phenotypic intra-species variation in
1659 *Candida albicans*. Genome Res 25:413-425.
1660
1661 Hobson DJ, Wei W, Steinmetz LM, Svejstrup JQ. 2012. RNA polymerase II collision interrupts
1662 convergent transcription. Mol Cell 48:365-374.

Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. 2016. BRAKER1: unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. Bioinformatics 32:767-769.

Hoff KJ, Stanke M. 2019. Predicting genes in single genomes with AUGUSTUS. Curr Protoc Bioinformatics 65:e57.

Huff JT, Zilberman D. 2014. Dnmt1-independent CG methylation contributes to nucleosome positioning in diverse eukaryotes. Cell 156:1286-1297.

Huss VAR, Frank C, Hartmann EC, Hirmer M, Kloboucek A, Seidel BM, Wenzeler P, Kessler E. 1999. Biochemical taxonomy and molecular phylogeny of the genus *Chlorella* sensu lato (Chlorophyta). Journal of Phycology 35:587-598.

Jagielski T, Bakuła Z, Gawor J, Maciszewski K, Kusber WH, Dylag M, Nowakowska J, Gromadka R, Karnkowska A. 2019. The genus *Prototheca* (Trebouxiophyceae, Chlorophyta) revisited: Implications from molecular taxonomic studies. Algal Research-Biomass Biofuels and Bioproducts 43.

Joo S, Nishimura Y, Cronmiller E, Hong RH, Kariyawasam T, Wang MH, Shao NC, El Akkad SE, Suzuki T, Higashiyama T, et al. 2017. Gene regulatory networks for the haploid-to-diploid transition of *Chlamydomonas reinhardtii*. Plant Physiology 175:314-332.

Kalina T, Punčochárová M. 1987. Taxonomy of the subfamily Scotiellocystoideae Fott 1976 (Chlorellaceae, Chlorophyceae). Algol Stud 45:473-521.

Kanesaki Y, Imamura S, Matsuzaki M, Tanaka K. 2015. Identification of centromere regions in chromosomes of a unicellular red alga, *Cyanidioschyzon merolae*. FEBS Letters 589:1219-1224.

Karpowicz SJ, Prochnik SE, Grossman AR, Merchant SS. 2011. The GreenCut2 resource, a phylogenomically derived inventory of proteins specific to the plant lineage. J Biol Chem 286:21427-21439.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772-780.

Kieselbach T, Mant A, Robinson C, Schroder WP. 1998. Characterisation of an *Arabidopsis* cDNA encoding a thylakoid lumen protein related to a novel 'pentapeptide repeat' family of proteins. FEBS Lett 428:241-244.

Kim SS, Grienenberger E, Lallemand B, Colpitts CC, Kim SY, Souza Cde A, Geoffroy P, Heintz D, Krahn D, Kaiser M, et al. 2010. *LAP6/POLYKETIDE SYNTHASE A* and *LAP5/POLYKETIDE SYNTHASE B* encode hydroxyalkyl alpha-pyrone synthases required for pollen development and sporopollenin biosynthesis in *Arabidopsis thaliana*. Plant Cell 22:4045-4066.

Kolesiński P, Golik P, Grudnik P, Piechota J, Markiewicz M, Tarnawski M, Dubin G, Szczepaniak A. 2013. Insights into eukaryotic Rubisco assembly - crystal structures of RbcX chaperones from *Arabidopsis thaliana*. Biochim Biophys Acta 1830:2899-2906.

Kolesiński P, Piechota J, Szczepaniak A. 2011. Initial characteristics of RbcX proteins from *Arabidopsis thaliana*. Plant Mol Biol 77:447-459.

Kropat J, Hong-Hermesdorf A, Casero D, Ent P, Castruita M, Pellegrini M, Merchant SS, Malasarn D. 2011. A revised mineral nutrient supplement increases biomass and growth rate in *Chlamydomonas reinhardtii*. Plant J 66:770-780.

Krüger W. 1894a. Beiträge zur Kenntnis der Organismen des Saftflusses (sog. Schleimflusses) der Laubbäume. II. Über zwei aus Saftflüssen rein gezüchtete Algen. Zopf Beitr Physiol Morphol Organ 4:91-116.

Krüger W. 1894b. Kurze Charakteristik einiger niederer Organismen im Saftflusse der Laubbäume. Hedwigia 33:241-266.

Kwon T, Hanschen ER, Hovde BT. 2023. Addressing the pervasive scarcity of structural annotation in eukaryotic algae. Sci Rep 13:1687.

Langdon QK, Peris D, Baker EP, Opulente DA, Nguyen HV, Bond U, Goncalves P, Sampaio JP, Libkind D, Hittinger CT. 2019. Fermentation innovation through complex hybridization of wild and domesticated yeasts. Nat Ecol Evol 3:1576-1586.

Lemieux C, Turmel M, Otis C, Pombert JF. 2019. A streamlined and predominantly diploid genome in the tiny marine green alga *Chloropicon primus*. Nat Commun 10:4061.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Arxiv. https://doi.org/10.48550/arXiv.1303.3997

Li H. 2021. New strategies to improve minimap2 alignment accuracy. Bioinformatics 37:4572-4574.

Lin J, Le TV, Augspurger K, Tritschler D, Bower R, Fu G, Perrone C, O'Toole ET, Mills KV, Dymek E, et al. 2019. FAP57/WDR65 targets assembly of a subset of inner arm dyneins and connects to regulatory hubs in cilia. Mol Biol Cell 30:2659-2680.

Liu C, Young AL, Starling-Windhof A, Bracher A, Saschenbrecker S, Rao BV, Rao KV, Berninghausen O, Mielke T, Hartl FU, et al. 2010. Coupled chaperone action in folding and assembly of hexadecameric Rubisco. Nature 463:197-202.

Liu HW, Khera R, Grob P, Gallaher SD, Purvine SO, Nicora CD, Lipton MS, Niyogi KK, Nogales E, Iwai M, et al. 2024. Fe starvation induces a second LHCI tetramer to photosystem I in green algae. bioRxiv. https://doi.org/10.1101/2024.12.11.624522

1754 Lopez D, Hamaji T, Kropat J, De Hoff P, Morselli M, Rubbi L, Fitz-Gibbon S, Gallaher SD,
1755 Merchant SS, Umen J, et al. 2015. Dynamic changes in the transcriptome and methylome of
1756 *Chlamydomonas reinhardtii* throughout Its life cycle. Plant Physiology 169:2730-2743.
1757
1758 MacIntosh GC. 2011. RNase T2 family: enzymatic properties, functional diversity, and evolution
1759 of ancient ribonucleases. In: Nicholson AW, editor. Ribonucleases. p. 89-114.
1760
1761 Mancera E, Bourgon R, Brozzi A, Huber W, Steinmetz LM. 2008. High-resolution mapping of
1762 meiotic crossovers and non-crossovers in yeast. Nature 454:479-485.
1763 Manni M, Berkeley MR, Seppey M, Simao FA, Zdobnov EM. 2021. BUSCO Update: Novel and
1764 Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of
1765 Eukaryotic, Prokaryotic, and Viral Genomes. Mol Biol Evol 38:4647-4654.
1766
1767 Marcet-Houben M, Gabaldón T. 2015. Beyond the whole-genome duplication: phylogenetic
1768 evidence for an ancient interspecies hybridization in the baker's yeast lineage. PLoS Biol
1769 13:e1002220.
1770
1771 Marcolungo L, Bellamoli F, Cecchin M, Lopatriello G, Rossato M, Cosentino E, Rombauts S,
1772 Delledonne M, Ballottari M. 2024. *Haematococcus lacustris* genome assembly and annotation
1773 reveal diploid genetic traits and stress-induced gene expression patterns. Algal Research-
1774 Biomass Biofuels and Bioproducts 80.
1775
1776 Matsuka M, Otsuka H, Hase E. 1966. Changes in contents of carbohydrate and fatty acid in cells
1777 of *Chlorella protothecoides* during processes of de- and re-generation of chloroplasts. Plant and
1778 Cell Physiology 7:651-662.
1779
1780 Mattei AL, Bailly N, Meissner A. 2022. DNA methylation: a historical perspective. Trends
1781 Genet 38:676-707.
1782
1783 Merchant SS, Prochnik SE, Vallon O, Harris EH, Karpowicz SJ, Witman GB, Terry A, Salamov
1784 A, Fritz-Laylin LK, Marechal-Drouard L, et al. 2007. The *Chlamydomonas* genome reveals the
1785 evolution of key animal and plant functions. Science 318:245-250.
1786
1787 Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear
1788 R. 2020. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the
1789 genomic era. Mol Biol Evol 37:1530-1534.
1790
1791 Moore ER, Bullington BS, Weisberg AJ, Jiang Y, Chang J, Halsey KH. 2017. Morphological
1792 and transcriptomic evidence for ammonium induction of sexual reproduction in *Thalassiosira*
1793 *pseudonana* and other centric diatoms. PLoS One 12:e0181098.
1794
1795 Moseley JL, Lee B, Im CS, Kim J, Kim D, Bhat R. 2022. Production of lipids and terpenoids in
1796 *Auxenochlorella protothecoides*. United States US12037630B2, filed November 5, 2021, and
1797 issued July 16, 2024.
1798

1799 Ning J, Otto TD, Pfander C, Schwach F, Brochet M, Bushell E, Goulding D, Sanders M,
1800 Lefebvre PA, Pei J, et al. 2013. Comparative genomics in *Chlamydomonas* and *Plasmodium*
1801 identifies an ancient nuclear envelope protein family essential for sexual reproduction in protists,
1802 fungi, plants, and vertebrates. Genes Dev 27:1198-1215.
1803
1804 Pardo-Palacios FJ, Arzalluz-Luque A, Kondratova L, Salguero P, Mestre-Tomas J, Amorin R,
1805 Estevan-Morio E, Liu T, Nanni A, McIntyre L, et al. 2024. SQANTI3: curation of long-read
1806 transcriptomes for accurate identification of known and novel isoforms. Nat Methods. 21:793-
1807 797
1808 Park SH, Kyndt JA, Brown JK. 2022. Comparison of *Auxenochlorella protothecoides* and
1809 *Chlorella* spp. chloroplast genomes: evidence for endosymbiosis and horizontal virus-like gene
1810 transfer. Life (Basel) 12.
1811
1812 Peris D, Perez-Torrado R, Hittinger CT, Barrio E, Querol A. 2018. On the origins and industrial
1813 applications of *Saccharomyces cerevisiae* x *Saccharomyces kudriavzevii* hybrids. Yeast 35:51-
1814 69.
1815
1816 Petroll R, Papareddy RK, Krela R, Laigle A, Riviere Q, Bišová K, Mozgová I, Borg M. 2024.
1817 The expansion and diversification of epigenetic regulatory networks underpins major transitions
1818 in the evolution of land plants. bioRxiv. https://doi.org/10.1101/2024.09.22.614159
1819
1820 Pombert JF, Blouin NA, Lane C, Boucias D, Keeling PJ. 2014. A lack of parasitic reduction in
1821 the obligate parasitic green alga *Helicosporidium*. PLoS Genet 10:e1004355.
1822
1823 Pontes A, Cadez N, Goncalves P, Sampaio JP. 2019. A quasi-domesticate relic hybrid population
1824 of *Saccharomyces cerevisiae* x *S. paradoxus* adapted to olive brine. Front Genet 10:449.
1825
1826 Poplin R, Chang PC, Alexander D, Schwartz S, Colthurst T, Ku A, Newburger D, Dijamco J,
1827 Nguyen N, Afshar PT, et al. 2018. A universal SNP and small-indel variant caller using deep
1828 neural networks. Nat Biotechnol 36:983-987.
1829
1830 Prescott EM, Proudfoot NJ. 2002. Transcriptional collision between convergent genes in budding
1831 yeast. Proc Natl Acad Sci U S A 99:8796-8801.
1832
1833 Pryszcz LP, Nemeth T, Saus E, Ksiezopolska E, Hegedusova E, Nosek J, Wolfe KH, Gacser A,
1834 Gabaldon T. 2015. The genomic aftermath of hybridization in the opportunistic pathogen
1835 *Candida metapsilosis*. PLoS Genet 11:e1005626.
1836
1837 Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic
1838 features. Bioinformatics 26:841-842.
1839
1840 Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP.
1841 2011. Integrative Genomics Viewer. Nature Biotechnology 29:24-26.
1842
1843 Romero Charria P, Navarrete C, Ovchinnikov V, Sarre LA, Shabardina V, Casacuberta E, Lara-
1844 Astiaso D, Sebé-Pedrós A, de Mendoza A. 2024. Adenine DNA methylation associated to

1845 transcription is widespread across eukaryotes. bioRxiv.
1846 https://doi.org/10.1101/2024.10.28.620566
1847
1848 Saito M, Xu P, Faure G, Maguire S, Kannan S, Altae-Tran H, Vo S, Desimone A, Macrae RK,
1849 Zhang F. 2023. Fanzor is a eukaryotic programmable RNA-guided endonuclease. Nature
1850 620:660-668.
1851
1852 Sanders CK, Hanschen ER, Biondi TC, Hovde BT, Kunde YA, Eng WL, Kwon T, Dale T. 2022.
1853 Phylogenetic analyses and reclassification of the oleaginous marine species *Nannochloris* sp.
1854 "desiccata" (Trebouxiophyceae, Chlorophyta), formerly *Chlorella desiccata*, supported by a
1855 high-quality genome assembly. J Phycol 58:436-448.
1856
1857 Santaguida S, Amon A. 2015. Short- and long-term effects of chromosome mis-segregation and
1858 aneuploidy. Nat Rev Mol Cell Biol 16:473-485.
1859
1860 Saschenbrecker S, Bracher A, Rao KV, Rao BV, Hartl FU, Hayer-Hartl M. 2007. Structure and
1861 function of RbcX, an assembly chaperone for hexadecameric Rubisco. Cell 129:1189-1200.
1862
1863 Schroder MS, Martinez de San Vicente K, Prandini TH, Hammel S, Higgins DG, Bagagli E,
1864 Wolfe KH, Butler G. 2016. Multiple origins of the pathogenic yeast *Candida orthopsilosis* by
1865 separate hybridizations between two parental species. PLoS Genet 12:e1006404.
1866
1867 Schubert M, Petersson UA, Haas BJ, Funk C, Schroder WP, Kieselbach T. 2002. Proteome map
1868 of the chloroplast lumen of *Arabidopsis thaliana*. J Biol Chem 277:8354-8365.
1869
1870 Selmecki A, Forche A, Berman J. 2006. Aneuploidy and isochromosome formation in drug-
1871 resistant *Candida albicans*. Science 313:367-370.
1872
1873 Seto A, Wang HL, Hesseltine CW. 1984. Culture conditions affect eicosapentaenoic acid content
1874 of *Chlorella minutissima*. Journal of the American Oil Chemists Society 61:892-894.
1875
1876 Severgnini M, Lazzari B, Capra E, Chessa S, Luini M, Bordoni R, Castiglioni B, Ricchi M,
1877 Cremonesi P. 2018. Genome sequencing of Prototheca zopfii genotypes 1 and 2 provides
1878 evidence of a severe reduction in organellar genomes. Sci Rep 8:14637.
1879
1880 Sipiczki M. 2008. Interspecies hybridization and recombination in *Saccharomyces* wine yeasts.
1881 FEMS Yeast Res 8:996-1007.
1882
1883 Soltis PS, Soltis DE. 2009. The role of hybridization in plant speciation. Annu Rev Plant Biol
1884 60:561-588.
1885
1886 Stadler R, Wolf K, Hilgarth C, Tanner W, Sauer N. 1995. Subcellular localization of the
1887 inducible *Chlorella* HUP1 monosaccharide-H+ symporter and cloning of a co-Induced galactose-
1888 H+ symporter. Plant Physiology 107:33-41.
1889

1890 Stanke M, Diekhans M, Baertsch R, Haussler D. 2008. Using native and syntenically mapped
1891 cDNA alignments to improve *de novo* gene finding. Bioinformatics 24:637-644.
1892
1893 Steenwyk JL, Lind AL, Ries LNA, Dos Reis TF, Silva LP, Almeida F, Bastos RW, Fraga da
1894 Silva TFC, Bonato VLD, Pessoni AM, et al. 2020. Pathogenic allodiploid hybrids of *Aspergillus*
1895 fungi. Curr Biol 30:2495-2507 e2497.
1896
1897 Strenkert D, Schmollinger S, Gallaher SD, Salome PA, Purvine SO, Nicora CD, Mettler-
1898 Altmann T, Soubeyrand E, Weber APM, Lipton MS, et al. 2019. Multiomics resolution of
1899 molecular events during a day in the life of Chlamydomonas. Proc Natl Acad Sci U S A
1900 116:2374-2383.
1901
1902 Strenkert D, Yildirim A, Yan J, Yoshinaga Y, Pellegrini M, O'Malley RC, Merchant SS, Umen
1903 JG. 2022. The landscape of Chlamydomonas histone H3 lysine 4 methylation reveals both
1904 constant features and dynamic changes during the diurnal cycle. Plant J 112:352-368.
1905
1906 Sui Y, Qi L, Wu JK, Wen XP, Tang XX, Ma ZJ, Wu XC, Zhang K, Kokoska RJ, Zheng DQ, et
1907 al. 2020. Genome-wide mapping of spontaneous genetic alterations in diploid yeast cells. Proc
1908 Natl Acad Sci U S A 117:28191-28200.
1909
1910 Suzuki S, Endoh R, Manabe R, Ohkuma M, Hirakawa Y. 2018. Multiple losses of
1911 photosynthesis and convergent reductive genome evolution in the colourless green algae
1912 *Prototheca*. Scientific Reports 8.
1913
1914 Tamura K, Stecher G, Kumar S. 2021. MEGA11: Molecular Evolutionary Genetics Analysis
1915 Version 11. Mol Biol Evol 38:3022-3027.
1916
1917 Tanaka S, Sawaya MR, Kerfeld CA, Yeates TO. 2007. Structure of the RuBisCO chaperone
1918 RbcX from *Synechocystis* sp. PCC6803. Acta Crystallogr D Biol Crystallogr 63:1109-1112.
1919
1920 Tang S, Lomsadze A, Borodovsky M. 2015. Identification of protein coding regions in RNA
1921 transcripts. Nucleic Acids Res 43:e78.
1922
1923 Tillich M, Lehwark P, Pellizzer T, Ulbricht-Jones ES, Fischer A, Bock R, Greiner S. 2017.
1924 GeSeq - versatile and accurate annotation of organelle genomes. Nucleic Acids Res 45:W6-W11.
1925
1926 Torres-Machorro AL, Hernandez R, Cevallos AM, López-Villaseñor I. 2010. Ribosomal RNA
1927 genes in eukaryotic microorganisms: witnesses of phylogeny? FEMS Microbiol Rev 34:59-86.
1928
1929 Tsai IJ, Bensasson D, Burt A, Koufopanou V. 2008. Population genomics of the wild yeast
1930 *Saccharomyces paradoxus*: Quantifying the life cycle. Proc Natl Acad Sci U S A 105:4957-
1931 4962.
1932
1933 Ueno R, Hanagata N, Urano N, Suzuki M. 2005. Molecular phylogeny and phenotypic variation
1934 in the heterotrophic green algal genus *Prototheca* (Trebouxiophyceae, Chlorophyta). Journal of
1935 Phycology 41:1268-1280.

1936
1937 Vande Zande P, Zhou X, Selmecki A. 2023. The dynamic fungal genome: polyploidy,
1938 aneuploidy and copy number variation in response to stress. Annu Rev Microbiol 77:341-361.
1939
1940 Vetting MW, Hegde SS, Fajardo JE, Fiser A, Roderick SL, Takiff HE, Blanchard JS. 2006.
1941 Pentapeptide repeat proteins. Biochemistry 45:1-10.
1942
1943 Wang B, Jia Y, Dang N, Yu J, Bush SJ, Gao S, He W, Wang S, Guo H, Yang X, et al. 2024.
1944 Near telomere-to-telomere genome assemblies of two *Chlorella* species unveil the composition
1945 and evolution of centromeres in green algae. BMC Genomics 25:356.
1946
1947 Werner A, Kanhere A, Wahlestedt C, Mattick JS. 2024. Natural antisense transcripts as versatile
1948 regulators of gene expression. Nat Rev Genet 25:730-744.
1949
1950 Wolff G, Burger G, Lang BF, Kück U. 1993. Mitochondrial genes in the colourless alga
1951 *Prototheca wickerhamii* resemble plant genes in their exons but fungal genes in their introns.
1952 Nucleic Acids Res 21:719-726.
1953
1954 Yan D, Dai J, Wu Q. 2013. Characterization of an ammonium transporter in the oleaginous alga
1955 *Chlorella protothecoides*. Appl Microbiol Biotechnol 97:919-928.
1956
1957 Yan D, Wang Y, Murakami T, Shen Y, Gong JH, Jiang HF, Smith DR, Pombert JF, Dai JBA,
1958 Wu QY. 2015. *Auxenochlorella protothecoides* and *Prototheca wickerhamii* plastid genome
1959 sequences give insight into the origins of non-photosynthetic algae. Scientific Reports 5:14465.
1960
1961 Yona AH, Manor YS, Herbst RH, Romano GH, Mitchell A, Kupiec M, Pilpel Y, Dahan O. 2012.
1962 Chromosomal duplication is a transient evolutionary solution to stress. Proc Natl Acad Sci U S A
1963 109:21010-21015.
1964
1965 Yoon PH, Skopintsev P, Shi H, Chen L, Adler BA, Al-Shimary M, Craig RJ, Loi KJ, DeTurk
1966 EC, Li Z, et al. 2023. Eukaryotic RNA-guided endonucleases evolved from a unique clade of
1967 bacterial enzymes. Nucleic Acids Res 51:12414-12427.
1968
1969 Yuen KW, Warren CD, Chen O, Kwok T, Hieter P, Spencer FA. 2007. Systematic genome
1970 instability screens in yeast and their potential relevance to cancer. Proc Natl Acad Sci U S A
1971 104:3925-3930.
1972
1973 Zemach A, McDaniel IE, Silva P, Zilberman D. 2010. Genome-wide evolutionary analysis of
1974 eukaryotic DNA methylation. Science 328:916-919.
1975
1976 Zorin B, Hegemann P, Sizova I. 2005. Nuclear-gene targeting by using single-stranded DNA
1977 avoids illegitimate DNA integration in *Chlamydomonas reinhardtii*. Eukaryot Cell 4:1264-1272.
1978