Near telomere-to-telomere genome assemblies of two *Chlorella* species unveil the composition and evolution of centromeres in green algae

Bo Wang[1,2†], Yanyan Jia[1,2†], Ningxin Dang[1,3†], Jie Yu[4], Stephen J. Bush[1], Shenghan Gao[1], Wenxi He[5], Sirui Wang[5], Hongtao Guo[6], Xiaofei Yang[6], Weimin Ma[4*], Kai Ye[1,2,3,5,7*]

[1]School of Automation Science and Engineering, Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, China.

[2]MOE Key Lab for Intelligent Networks & Networks Security, Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, China.

[3]Genome Institute, the First Affiliated Hospital of Xi'an Jiaotong University, Xi'an, China.

[4]College of Life Sciences, Shanghai Normal University, Shanghai, China.

[5]School of Life Science and Technology, Xi'an Jiaotong University, Xi'an, China.

[6]School of Computer Science and Technology, Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, China.

[7]Faculty of Science, Leiden University, Leiden, The Netherlands.

[†] These authors contributed equally.

* Correspondence author: kaiye@xjtu.edu.cn (K. Ye), wma@shnu.edu.cn (W. Ma)

Abstract

Background: Centromeres play a crucial and conserved role in cell division, although their composition and evolutionary history in green algae, the evolutionary ancestors of land plants, remains largely unknown.

Results: We constructed near telomere-to-telomere (T2T) assemblies for two Trebouxiophyceae species, *Chlorella sorokiniana* NS4-2 and *Chlorella pyrenoidosa* DBH, with chromosome numbers of 12 and 13, and genome sizes of 58.11 Mb and 53.41 Mb, respectively. We identified and validated their centromere sequences using CENH3 ChIP-seq and found that, similar to humans and higher plants, the centromeric CENH3 signals of green algae display a pattern of hypomethylation. Interestingly, the centromeres of both species largely comprised transposable elements, although they differed significantly in their composition. Species within the *Chlorella* genus display a more diverse centromere composition, with major constituents including members of the LTR/Copia, LINE/L1, and LINE/RTEX families. This is in contrast to green algae including *Chlamydomonas reinhardtii*, *Coccomyxa subellipsoidea*, and *Chromochloris zofingiensis*, in which centromere composition instead has a pronounced single-element composition. Moreover, we observed significant differences in the composition and structure of centromeres among chromosomes with strong collinearity within the *Chlorella* genus, suggesting that centromeric sequence evolves more rapidly than sequence in non-centromeric regions.

Conclusions: This study not only provides high-quality genome data for comparative genomics of green algae but gives insight into the composition and evolutionary history of centromeres in early plants, laying an important foundation for further research on their evolution.

Keywords

Genome hybrid assembly; Telomere-to-telomere; Centromere assembly; Transposable elements; Centromere evolution.

Background

Centromeres play a critical role in ensuring the accurate segregation of chromosomes

during both mitosis and meiosis [1], being specialized chromosomal regions associated with chromatid cohesion and correct separation during cell division. Centromeres serve as the assembly and attachment points for the kinetochore, a large protein complex that facilitates the binding of spindle microtubules to chromatids, enabling their movement towards the poles [2]. The 'centromere paradox' refers to the apparent contradiction between this essential role in chromosome segregation, which is conserved across all eukaryotes, and the significant variability in their organization and DNA composition, even within closely related species [3, 4]. This variability highlights the ('paradoxically') rapid evolution of centromeres.

The high prevalence of satellite repeats within centromere sequences also presents a considerable genome assembly challenge, frequently leading to the substantial loss of sequence information in this region [5]. This obstacle poses a significant impediment to the comprehensive investigation of centromere functionality. However, with advances in sequencing technology, in particular the PacBio high-fidelity (HiFi) and Oxford Nanopore Technologies (ONT) ultra-long methods, in conjunction with advances in assembly algorithms [6, 7], significant progress has been made towards the complete assembly of centromeres in many model species, including but not limited to human [8], *Arabidopsis thaliana* [9, 10] rice [11], and maize [12]. Notably, both human and higher plant centromeres primarily comprise numerous satellite sequences, organized into higher-order repeat arrays [5, 13]. It has previously been proposed that plant centromeric satellite sequences may have arisen from transposable element (TE) insertions specifically targeting the centromere [14, 15]. Consistent with this, the centromeres of the oldest known domesticated wheat, *Triticum monococcum*, have recently been found to predominantly feature transposable elements [16], as have those of the moss *Physcomitrium patens* and the model green alga *Chlamydomonas reinhardtii* [17, 18]. Whether centromeres in other green algae, more distantly related to *Chlamydomonas reinhardtii*, are also composed of TEs, and if so, whether the constituent elements are similar, as well as the evolutionary transitions between them, remains unclear.

In this study, we sequenced and assembled two *Chlorella* species to near-T2T level,

79 characterising and experimentally validating their centromeres, and comparing them
80 to those of *Chlamydomonas reinhardtii, Coccomyxa subellipsoidea*, and
81 *Chromochloris zofingiensis*. We found that the centromeres of the green algae species
82 selected in this study extensively comprise TEs, albeit with significant differences in
83 composition. *Chlamydomonas reinhardtii* and *Coccomyxa subellipsoidea*
84 predominantly feature elements from the LINE/L1 family, while *Chromochloris*
85 *zofingiensis* primarily comprises elements from the LTR/Copia family. In contrast,
86 species within the *Chlorella* genus exhibit a more diverse centromere composition,
87 the major constituents of which were members of the LTR/Copia, LINE/L1, and
88 LINE/RTEX families. Overall, our results give insight into the centromere evolution
89 of green algae, which may be generalised to early plants, as well as providing high-
90 quality genomic resources to facilitate future comparative analyses.

91 **Results**

92 **Near telomere-to-telomere assemblies of two *Chlorella* species**

93 To investigate centromere evolution within the context of green algae, we assembled
94 the genomes of two *Chlorella* species, *Chlorella sorokiniana* NS4-2 and *Chlorella*
95 *pyrenoidosa* DBH (basionym: *Auxenochlorella pyrenoidosa*), both of which belong to
96 the class Trebouxiophyceae (phylogenetically distant from the model green alga
97 *Chlamydomonas reinhardtii*, with a divergence time of 827 MYA; http://timetree.org/)
98 (**Supplementary Fig. S1**). The two species are morphologically similar, although
99 vary in their chromosome numbers, with *Chlorella sorokiniana* NS4-2 having 12
100 chromosomes and *Chlorella pyrenoidosa* DBH having 13 (**Fig. 1A, Fig. 1B,**
101 **Supplementary Fig. S2A and Supplementary Fig. S2B**). The genomes of *Chlorella*
102 *sorokiniana* NS4-2 and *Chlorella pyrenoidosa* DBH were sequenced using HiFi reads
103 at high coverage depths of 438 and 308-fold respectively (**Supplementary Table S1**).
104 For *Chlorella sorokiniana*, we assembled the HiFi reads (**Supplementary Fig. S3A**)
105 to produce a set of 16 contigs. These contigs had a minimum length of 141,360 bp and
106 were selected based on the criterion of length exceeding 100 Kb. Following this, we
107 generated a final set of 13 contigs by excluding the genome of the putative alga-

108 benefitting commensal bacterium *Arthrobacter* sp. [19] as well as fragments of rDNA
109 and plastid DNA (21 Kb~107 Kb). Among these contigs, 11 contained the plant-type
110 telomeric repeat unit (CCCTAAA/TTTAGGG) at both ends, signifying their
111 completeness. Additionally, two contigs were found to have rDNA units at one end,
112 as observed through collinearity analysis in comparison with the closely related
113 species *Chlorella vulgaris* which was scaffolded by optical mapping [20]
114 (**Supplementary Fig. S4A**). To consolidate the rDNA ends originating from the same
115 chromosome, we merged the two contigs and inserted 100 Ns as spacers. As a result
116 of this process, we produced a near-completely T2T 12-chromosome assembly of
117 *Chlorella sorokiniana* NS4-2, the only gaps in which were in the rDNA region of
118 chromosome 1. This conclusion was further supported by karyotyping results (**Fig. 1B**)
119 and Hi-C contact maps (**Fig. 1C**).
120 For *Chlorella pyrenoidosa* DBH, we initially assembled the HiFi reads using hifiasm
121 [6] (**Supplementary Fig. S3B**), which generated 38 contigs, each exceeding 100 Kb
122 in length. However, upon closer examination, we discovered that only three of these
123 contigs contained the telomeric repeat unit at both ends. This finding suggests that
124 achieving a T2T-level assembly for *Chlorella pyrenoidosa* DBH is more challenging
125 compared to *Chlorella sorokiniana* NS4-2 (**Supplementary Fig. S3A**). To overcome
126 this challenge, we generated ONT ultra-long reads (with a read length N50 of 59 Kb)
127 at a high coverage depth of 368-fold (**Supplementary Table S1**), assembling them
128 using NextDenovo [21] to generate a set of 14 contigs. Of these 14 contigs, 10
129 exhibited the presence of telomeric repeat units at both ends, indicating the complete
130 resolution of the telomere regions. However, two contigs only had rDNA units at one
131 end and two only had the telomeric repeat unit at one end, in both cases suggesting an
132 unresolved telomeric region at the other. To address this issue and achieve a more
133 comprehensive assembly, we utilized the HiFi contigs to replace the non-resolved
134 ONT contigs (**Supplementary Fig. S3B**). By incorporating these updated HiFi
135 contigs, we successfully achieved a gap-free assembly of 14 contigs. This updated
136 assembly provides a more complete representation of the genome, with improved
137 resolution and continuity. In a similar vein, for *Chlorella pyrenoidosa* DBH,

138 collinearity analysis also revealed two contigs with rDNA units potentially originating
139 from the same chromosome (**Supplementary Fig. S4B**). Following a similar
140 approach to *Chlorella sorokiniana* NS4-2, the rDNA ends were merged, and 100 Ns
141 were introduced as spacers. This led to the generation of a virtually T2T assembly
142 with 13 chromosomes, again supported by karyotyping results (**Supplementary Fig.**
143 **S2B**) and Hi-C contact maps (**Supplementary Fig. S2C**).
144 Genome completeness was evaluated using Benchmarking Universal Single-Copy
145 Orthologs (BUSCO), with *Chlorella sorokiniana* NS4-2 scoring 98.3% and *Chlorella*
146 *pyrenoidosa* DBH 99.0%. The assembly of these two genomes therefore represents a
147 significant improvement compared to existing genomes for these species (**Table 1**).
148 *Chlorella sorokiniana* NS4-2 and *Chlorella pyrenoidosa* DBH were predicted to have
149 12,691 and 11,327 protein coding genes, respectively, and to exhibit a significant
150 collinearity relationship, accompanied by chromosomal rearrangements (**Fig. 1D**).
151 Finally, all HiFi reads greater than 15 Kb were selected and used to perform
152 mitochondrial and chloroplast (i.e. circular) genome assembly using Unicycler v0.5.0
153 with default parameters [22]. The chloroplast genome sizes of *Chlorella sorokiniana*
154 NS4-2 and *Chlorella pyrenoidosa* DBH are 109,738 bp and 107,436 bp, respectively;
155 while the mitochondrial genome sizes are 53,629 bp and 69,538 bp, respectively
156 (GenBank accession: OP311589-OP311592).
157
158 **Putative centromere identification and CENH3 ChIP-seq validation**
159 Recent advances in sequencing technologies have shed light on the crucial role of TEs
160 in influencing the characteristics of centromeric sequences among green algal species
161 [17, 23]. Notably, studies of *Chlamydomonas reinhardtii* CC-4532 (v. 6.1) [24] and
162 *Coccomyxa subellipsoidea* C-169 (v. 3.0) [25] have demonstrated the prevalence of
163 LINE/L1 elements within their genomes, as well as LTR/Copia elements in
164 *Chromochloris zofingiensis* [26], further emphasizing the impact of TEs in shaping
165 their organism's centromeric regions. These TEs are strikingly abundant in the
166 centromeres, forming a unique and discernible structure when performing dot plots
167 for self-alignment analysis (**Supplementary Fig. S5**). Notably, within this structure,

168 specific clusters stand out, such as the *Zepp*-like cluster in *Chlamydomonas*
169 *reinhardtii* and the *Zepp* cluster in *Coccomyxa subellipsoidea* (**Supplementary Fig.**
170 **S5**). Leveraging this distinctive characteristic, we conducted dot plot analyses on a
171 per-chromosome basis to pinpoint the potential centromeric regions within the
172 genomes of *Chlorella sorokiniana* NS4-2 (**Supplementary Fig. S6**) and *Chlorella*
173 *pyrenoidosa* DBH (**Supplementary Fig. S7**). In the case of *Chlorella sorokiniana*
174 NS4-2, we successfully validated putative centromeres for all 12 chromosomes using
175 VerityMap [27] (**Supplementary Fig. S8**), with lengths ranging from 28,465 bp to
176 104,090 bp (average length: 60 Kb) (**Supplementary Table S2**). Subsequently, we
177 employed centromeric Histone H3 chromatin immunoprecipitation followed by
178 sequencing technology (CENH3 ChIP-seq) to validate the predicted potential
179 centromere regions and found them to be consistent with experimental verification
180 (**Fig. 2 and Supplementary Fig. S9**). We then proceeded to use the dot plot method
181 (**Supplementary Fig. S10**) for potential centromere prediction in *Chlorella*
182 *pyrenoidosa* DBH species. For *Chlorella pyrenoidosa* DBH, we validated putative
183 centromeres for all 13 chromosomes using VerityMap (**Supplementary Fig. S11**),
184 with lengths ranging from 19,117 bp to 65,960 bp (average length: 39 Kb)
185 (**Supplementary Table S3**). Insights from circular plots indicate that the two species
186 exhibit a strong degree of collinearity and that the putative centromeres of *Chlorella*
187 *sorokiniana* NS4-2 and *Chlorella pyrenoidosa* DBH are composed of TE elements
188 (**Fig. 1D**). This was especially apparent for *Chlorella sorokiniana* NS4-2, in which the
189 centromere regions were significantly enriched with LTR/Copia elements (**Fig. 2 and**
190 **Fig. 1D**).
191

192 **Centromere methylation analysis in two *Chlorella* species**

193 In higher plants, such as *Arabidopsis thaliana* and *Zea mays*, satellite repeats that are
194 associated with chromatin containing CENH3 exhibit lower levels of methylation
195 when compared to the non-CENH3 centromeric regions [10, 28]. These findings
196 imply that DNA methylation is a significant factor in the epigenetic demarcation of
197 centromeric chromatin. However, it is currently unclear whether centromeres in green

198 algae exhibit similar methylation patterns. We utilized ccsmeth [29], a deep learning-
199 based circular consensus sequencing (CCS) sequencing methylation calling software,
200 to analyze the centromere methylation in the *Chlorella sorokiniana* NS4-2 genome
201 (**Fig. 3**). Interestingly, we found that the centromeric region of chromosome 1 in
202 *Chlorella sorokiniana* NS4-2 contains hypomethylated intervals (**Fig. 3A**). We more
203 closely examined the regions 50 Kb up- and downstream of the centromere and
204 observed a negative correlation between the hypomethylated regions and CENH3-
205 ChIP signals (**Fig. 3A and Supplementary Fig. S12**). We then divided the genome
206 into four categories – centromeric regions, non-centromeric regions, CENH3-ChIP
207 peak regions in centromeric regions, and non-ChIP peak regions in centromeric
208 regions – and conducted a statistical analysis of methylation frequencies. Our analysis
209 revealed that centromeric regions exhibited significantly lower methylation levels
210 compared to non-centromeric regions, which we attributed primarily to the low
211 methylation status of the CENH3-ChIP peak regions (**Fig. 3B**). For *Chlorella*
212 *pyrenoidosa* DBH, we employed HiFi and ONT sequencing to detect methylation
213 status, and similarly identified regions of low methylation within the centromeres
214 (**Supplementary Fig. S13**). We observed the same phenomenon of centromeric low
215 methylation in *Chlorella sorokiniana* NS4-2 (**Fig. 3C**). These results indicate that
216 despite the absence of tandemly repeated satellite sequences in the centromeres of the
217 two *Chlorella* genomes, their pattern of methylation nevertheless resembles that of
218 higher plants. This suggests that despite comparatively rapid centromeric evolution,
219 their sequences retained a conserved epigenetic inheritance.
220

221 **Centromere size and composition in green algae**

222 We next selected six high-quality green algal assemblies for a compositional analysis
223 of their putative centromeres: *Chlamydomonas reinhardtii* CC-4532 (v. 6.1) [24],
224 *Chromochloris zofingiensis* [26], *Coccomyxa subellipsoidea* C-169 (v. 3.0) [25],
225 *Chlorella* sp. SLA-04 [30], and the two *Chlorella* species assembled in this study (**Fig.**
226 **4A**). We used the red alga *Cyanidioschyzon merolae* (which has short low-GC
227 centromeric regions) [31, 32] as an outgroup when constructing the phylogenetic tree

228 of this lineage (**Fig. 4A**). For each of *Chlamydomonas reinhardtii*, *Chromochloris*
229 *zofingiensis*, *Coccomyxa subellipsoidea* and *Chlorella* sp. SLA-04, we identified the
230 potential centromeric loci using dot plots (**Supplementary Fig. S10 and Fig. 4C-H**)
231 and prior knowledge [17, 25, 26], with average sizes of approx. 200 Kb, 27 Kb, 23 Kb,
232 and 52 Kb, respectively (**Supplementary Table S4−S7**). We provided a detailed
233 annotation of the centromere composition of each species in **Supplementary Tables**
234 **S8 to S13**. In addition, we found a positive correlation between centromere and
235 chromosome size for the six species, with a significant Pearson correlation coefficient
236 of 0.87 (**Fig. 4B**). This fills a gap in the literature by providing data on green algae,
237 which was previously lacking [33].
238 In the centromeres of both *Chlamydomonas reinhardtii* (**Fig. 4C and Supplementary**
239 **Table S14**) and *Coccomyxa subellipsoidea* (**Fig. 4E and Supplementary Table S15**),
240 we observed a substantial number of ZeppL-1_cRei (LINE/L1) and Zepp (LINE/L1)
241 elements, respectively, consistent with previous reports [17]. Nearly all of these
242 LINE/L1 elements are distributed in the centromere region, suggesting they are
243 centromere-specific insertion (**Supplementary Table S8 and Supplementary Table**
244 **S10**). Of note, however, is that although *Chromochloris zofingiensis* is closely related
245 to *Chlamydomonas reinhardtii*, its centromeres primarily comprise members of the
246 LTR/Copia, not LINE/L1, family (**Fig. 4D and Supplementary Table S16**).
247 The TE composition of the centromeres of each species are notably diverse. In
248 *Chlorella* sp. SLA-04, the centromeres largely comprise Copia-2_Chlorella_SLA04
249 (LTR/Copia), L1-5_Chlorella_SLA04 (LINE/L1) and RTE-X-1_Chlorella_SLA04
250 (LINE/RTE-X) elements, at 37%, 22%, and 11% of the total, respectively
251 (**Supplementary Table S11 and Supplementary Table S17**). Notably, the latter two
252 elements are detected in all 13 centromeres of *Chlorella* sp. SLA-04 (**Supplementary**
253 **Fig. S14**).
254 By contrast, in *Chlorella sorokiniana* NS4-2, the centromeres largely comprise Copia-
255 1_Chlorella_NS (LTR/Copia) and Unknown-1_Chlorella_NS elements, at 50% and
256 21% of the total, respectively (**Supplementary Table S12 and Supplementary**
257 **Table S18**), with both elements identified in all 12 centromeres (**Supplementary Fig.**

258 **S15**). In *Chlorella pyrenoidosa* DBH, the largest centromeric component, at 25% of
259 the total, is Unknown-2_Chlorella_DBH (**Supplementary Table S13 and**
260 **Supplementary Table S19**), although RTE-X-2_Chlorella_DBH (LINE/RTE-X)
261 elements, which represent another 20% of the total, were the only one found in all 13
262 centromeres (**Supplementary Fig. S16**). As centromeres are 'cold spots' for
263 recombination [9, 34], it has been hypothesised that they could 'protect' TEs which
264 insert there from being purged by purifying selection [35]. This would ultimately lead
265 to the formation of TE-enriched regions specific to the centromeres, an interpretation
266 consistent with our findings (**Fig. 4F-H**).
267
268 **Centromere evolution in green algae**
269 Finally, we systematically correlated the primary constituents of each centromere
270 (that is, their TEs) across the phylogenomic tree, an approach which allows us to
271 examine the evolutionary trajectories of these sequences in different taxa (**Fig. 5A**).
272 As LINE/L1 and LTR/Copia elements were found within both the Chlorophyceae and
273 Trebouxiophyceae lineages (**Fig. 5A**), the implication is that these elements would
274 also have resided within the centromeres of the most recent common ancestor of
275 green algae.
276 The *Chlorophyceae-Chlorella* divergence is estimated to have occurred
277 approximately 827 MYA, while the *Coccomyxa-Chlorella* divergence is estimated to
278 have occurred at around 544 MYA (http://www.timetree.org/). Our analyses of
279 centromere composition suggest that for *Chlamydomonas reinhardtii*, *Chromochloris*
280 *zofingiensis*, and *Coccomyxa subellipsoidea*, the centromeres comprise a primary TE
281 component, seemingly a singular composite of many inserted elements, but that for
282 *Chlorella* species, multiple diverse TE families are maintained within the centromeres.
283 To explore this further, we performed collinearity analysis on the assemblies of three
284 *Chlorella* species (*Chlorella* sp. SLA-04, *Chlorella sorokiniana* NS4-2, and *Chlorella*
285 *pyrenoidosa* DBH), followed by illustrating the collinearity relationships and the
286 putative centromeres (**Fig. 5B**). For example, chromosome 3 of *Chlorella* sp. SLA-04,
287 chromosome 1 of *Chlorella sorokiniana* NS4-2, and chromosome 1 of *Chlorella*

288 *pyrenoidosa* DBH exhibited strong collinearity (**Supplementary Fig. S17**),
289 suggesting they share a common ancestor (**Fig. 5B**). We extracted the centromeres of
290 these three chromosomes for dot plot analysis and found significant structural (that is,
291 TE) differences among them (**Fig. 5C**). For instance, 88% of the centromere of
292 *Chlorella sorokiniana* NS4-2 chromosome 1 comprised members of the LTR/Copia
293 family, compared to 44% for *Chlorella* sp. SLA-04 chromosome 3 and 27% for
294 *Chlorella pyrenoidosa* DBH chromosome 1. The implication of these results is that
295 functional centromeres can vary greatly in their structure, and that the evolution of
296 centromeric sequences is faster than that of non-centromeric ones.

297 **Discussion**

298 Recent advances in long-read sequencing technologies have enabled the telomere-to-
299 telomere (T2T) assembly of a number of genomes including the model plant
300 *Arabidopsis thaliana* [10], the model green alga *Chlamydomonas reinhardtii* [23], and
301 the model moss *Physcomitrium patens* [18]. To complement these, we employed HiFi
302 reads, ONT ultra-long reads, and stringent assembly methods to generate T2T-quality
303 assemblies for *Chlorella sorokiniana* NS4-2 and *Chlorella pyrenoidosa* DBH. These
304 assemblies represent a significant improvement over previous genome assemblies,
305 such as *Chlorella sorokiniana* UTEX1230 [36] and *Chlorella pyrenoidosa* FACHB-9
306 [37], and in that respect they may serve as valuable resources for future green algae
307 research, in particular towards understanding the more complex regions of the
308 genome, including the centromere.

309 In human centromeres, there are specific hypomethylated regions containing a higher
310 concentration of the centromeric histone CENP-A, suggesting a functional role as
311 kinetochore-binding sites [38, 39]. Interestingly, this study reveals that the same
312 epigenetic phenomenon, which is conserved in higher plants [10, 28], is also present
313 in green algae, suggesting a highly conserved epigenetic trait possibly among
314 eukaryotes in general.

315 Centromeric satellite sequences typically have a monomer size of 100-200 base pairs,
316 often 150 to 180 bp, a length sufficient to wrap a single nucleosome [4]. However, in

317 plants, the origin of these monomers remains unclear. Analysis of ancestral crop
318 species, such as *Gossypium raimondii* [40], an ancestor species of tetraploid cotton,
319 and *Triticum monococcum* [16], the first domesticated wheat species, found that their
320 centromeres were comprised largely of retrotransposons rather than long arrays of
321 centromeric satellites, supporting the hypothesis that centromere repeats originated
322 from non-random retrotransposon insertions [15]. We found that each of the green
323 algae in this study had TE-rich centromeres, further supporting the hypothesis that the
324 centromeric monomers of higher plants originated from TE insertions in their
325 ancestors.
326 Nonetheless, there remain substantial variations in the composition of centromeric
327 TEs among different taxa of green algae, likely reflecting differential TE insertion,
328 expansion and purging. For example, *Chlamydomonas reinhardtii* and *Coccomyxa*
329 *subellipsoidea* both harbor LINE/L1 elements within their centromeric regions
330 although from different subfamilies [17]. Furthermore, the two *Chlorella* genomes
331 sequenced in this study, despite belonging to the same genus, differ greatly in the
332 composition and structure of their centromeres. These observations lend further
333 support to the view that the centromere is the fastest-evolving part of the chromosome
334 [41, 42].
335 In summary, we assembled two *Chlorella* species (*Chlorella sorokiniana* NS4-2 and
336 *Chlorella pyrenoidosa* DBH) to T2T level, enabling a comprehensive analysis of their
337 centromeric sequence composition and evolution. These findings contribute to our
338 understanding of centromere evolution in green algae and, by extension, to their
339 descendants.

340 ## Materials and methods

341 ### Culture conditions

342 *Chlorella pyrenoidosa* strain DBH (obtained from the Freshwater Algae Culture
343 Collection at the Institute of Hydrobiology, Chinese Academy of Sciences) and
344 *Chlorella sorokiniana* strain NS4-2 (isolated and purified from a water sample taken
345 from Xueshi Lake at Shanghai Normal University) were grown mixotrophically in a

346 tris-acetate-phosphate (TAP) medium. Both species were cultivated in Erlenmeyer
347 flasks under continuous growth light conditions (30-40 μmol photons m^{-2} s^{-1}) at a
348 temperature of 25 °C, with agitation at 100 rpm, for maintenance purposes.

349 **Karyotyping of the two *Chlorella* species**

350 Culture samples were harvested by centrifugation at 12,000 rpm for 2 min and then
351 washed sequentially with sterilized BG-11 medium and distilled water. The resulting
352 algal pellets were suspended in a 2 mM 8-hydroxyquinoline solution and kept in
353 darkness at 4 °C for 4 h. Subsequently, the algae pellets were fixed with a solution of
354 3 parts ethanol to 1 part acetic acid (v/v) for 24 h to ensure fixation of the algae cells
355 in the metaphase of mitosis. After fixation, the cells were washed three times with
356 distilled water and then subjected to enzymatic digestion. The enzymatic solution
357 contained 2.5% (w/w) cellulose Onozuka R 10 (Yakult Pharmaceutical, Tokyo, Japan)
358 and 2.5% (w/w) pectinase Y23 (Yakult Pharmaceutical, Tokyo, Japan), and the
359 digestion process lasted for 18 h at 37 °C. Following digestion, the cells were
360 homogenized at 15,000 rpm for 2 min and then centrifuged. The resulting pellets were
361 resuspended in 100 μL of acetic acid and vortexed for 20 sec. The cell suspension was
362 dropped onto a sterile slide and allowed to air dry. Finally, the cells were stained with
363 an improved carbol fuchsin solution and examined under a microscope to capture
364 images of the metaphase stage.

365

366 **Genomic DNA extraction**

367 Samples were collected, and high molecular weight genomic DNA was prepared by
368 the CTAB method followed by purification with the QIAGEN® Genomic kit
369 (Cat#13343, QIAGEN) for regular sequencing, according to the standard operating
370 procedure provided by the manufacturer. DNA degradation and contamination of the
371 extracted DNA was monitored on 1% agarose gels. DNA purity was then assessed
372 using a NanoDrop One UV-VIS spectrophotometer (Thermo Fisher Scientific, USA),
373 of which OD 260/280 ranged from 1.8 to 2.0 and OD 260/230 from 2.0 to 2.2. DNA
374 concentration was subsequently measured using a Qubit 4.0 fluorometer (Invitrogen,
375 USA).

376 **HiFi sequencing and genome assembly**

377 SMRTbell libraries were prepared following the standard protocol of PacBio, using 15
378 Kb preparation solutions from Pacific Biosciences, USA. The library preparation
379 steps were conducted according to the method described in [10]. Sequencing was
380 performed on a PacBio Sequel II instrument with Sequencing Primer V2 and Sequel
381 II Binding Kit 2.0 at Xi'an Haorui Genomics Technology Co., Ltd. (Xi'an, China).
382 For *Chlorella sorokiniana* strain NS4-2, a total of 25.87 Gb of HiFi reads were
383 generated, providing a coverage of 438×. The N50 of the reads was 15,694 bp.
384 Regarding the *Chlorella pyrenoidosa* strain DBH, a total of 23.74 Gb of HiFi reads
385 were generated, resulting in a coverage of 308×. The N50 of the reads was 13,338 bp.
386 The HiFi reads were assembled using hifiasm v. 0.16.1-r375 [6] with default
387 parameters. The gfatools (https://github.com/lh3/gfatools) was used to convert GFA-
388 format sequence graphs to FASTA format.

389 **ONT sequencing and genome assembly**

390 For the ultra-long Oxford Nanopore sequencing library, approximately 8–10 μg of
391 genomic DNA was selected (>50 kb) with the SageHLS HMW library system (Sage
392 Science, Beverly, MA), and then processed using the ligation sequencing 1D kit
393 (SQK-LSK109, Oxford Nanopore Technologies, Oxford, UK) according to the
394 manufacturer's instructions. DNA libraries (approximately 800 ng) were constructed
395 and sequenced using a PromethION (Oxford Nanopore Technologies, Oxford, UK) at
396 the Genome Center of Grandomics (Wuhan, China). A total of 22.12 Gb ONT long
397 reads with 368× coverage (read N50: 59,282 bp) were generated. To assemble these
398 into contigs, the long-read assembler NextDenovo v. 2.0 [21] was used with
399 parameters 'read_cutoff = 5k' and 'seed_cutoff = 132,382'. Contigs were then
400 polished using Nextpolish v. 1.3.1 [43] with parameters 'hifi_options = -min_read_len
401 10k -max_read_len 45k' and 'hifi_minimap2_options = -x asm20'.

402 **Hi-C sequencing and scaffolding**

403 A Hi-C library was generated from cross-linked chromatins of algal cells using a
404 standard Hi-C protocol, and then sequenced using the NovaSeq 6000 platform. The

Hi-C sequencing data, as outlined in Table S1, were utilized to construct a Hi-C contact map. The construction of the contact map was performed using Juicer v. 1.5 [44] with parameter '-s DpnII'. The resulting contact map was manually inspected and verified using Juicebox v. 1.11.08 [45].

Evaluation of the putative centromeres

To assess the assembly structure of the predicted centromeric regions, we employed VerityMap [27]. First, the PacBio HiFi reads were aligned to the assembly using minimap2 v. 2.20-r1061 [46] with the '-x map-hifi' parameter. Reads aligning to the centromeric regions were extracted using SAMtools v. 1.9 [47]. These extracted HiFi reads were then input into VerityMap using the following parameters: '--careful --reads HiFi reads -d hifi'.

Genome comparisons

The synteny relationships among the chromosomes were estimated using MCScanX [48] with parameter '-s 10' and visualized using Circos v. 0.69-8 [49]. Genomic alignment dot plot of chromosomes was generated using Gepard v. 2.1 [50]. To assess genome completeness, we used BUSCO v. 3.0.2 with the dataset 'chlorophyta_odb10 v. 4' [51]. Chromosome painting among three species of the genus *Chlorella* was performed using IAGS [52] with the GMP model.

Genome annotation

For the *Chlorella sorokiniana* NS4-2 and *Chlorella pyrenoidosa* DBH genome annotations, we initially predicted the *de novo* gene structure using GeneMark-ES with the parameters '--ES' [53]. Subsequently, we employed the 'chlorella gene model' to train Augustus v. 3.3 [54]. To identify protein-coding genes, we utilized the MAKER2 pipeline [55], which integrated the gene models predicted from GeneMark-ES and Augustus, as well as the protein sequences from four closely related organisms: *Chlamydomonas reinhardtii* (GenBank accession: GCF_000002595.2) [56], *Chlorella variabilis* (GCF_000147415.1) [57], *Chlorella sorokiniana* strain

435 1602 (GCA_002245835.2) [58], and *Micractinium conductrix* (GCA_002245815.2)
436 [58]. For the identification of repeat sequences, we employed RepeatMasker v. 4.0.7
437 (http://www.repeatmasker.org) and RepeatModeler v. 2.0.1 [59]. The TE-algae library
438 used to annotate the selected species consists of three parts: the *de novo* library of
439 three *Chlorella* species from RepeatModeler, a library from a previous study [17], and
440 the *Chlorophyta* Repbase library (https://www.repeatmasker.org/) (**Supplementary**
441 **Dataset S1**). The TE-algae library underwent redundancy reduction using cd-hit [60].
442 The karyoplots shown in this study were generated using the KaryoploteR package in
443 R [61].
444
445 **Phylogenetic tree construction**
446 The phylogenetic trees depicted in Fig. 4 and Fig. 5 were generated utilizing
447 OrthoFinder v. 2.2.7 with the default parameters [62]. Supplementary Fig. S1's
448 phylogenetic tree was constructed using MEGA v. 7 software and employed the
449 Kimura 2-parameter model, accompanied by 1000 bootstrap replicates [63].
450
451 **PacBio DNA methylation analysis**
452 Methylation for *Chlorella sorokiniana* NS4-2 and *Chlorella pyrenoidosa* DBH was
453 inferred with ccsmeth v.0.4.1 [29], using kinetic features from PacBio CCS reads.
454 Methylation of CCS reads was predicted using 'ccsmeth call_mods' with the model
455 'model_ccsmeth_5mCpG_call_mods_attbigru2s_b21.v2.ckpt'. Then, we used
456 'ccsmeth align_hifi' to align the reads to their respective genome. The methylation
457 frequencies were calculated by using 'ccsmeth call_freqb' with the aggregate model
458 'model_ccsmeth_5mCpG_aggregate_attbigru_b11.v2p.ckpt'.
459
460 **ONT DNA methylation analysis**
461 Nanopolish v. 0.13.2 with the parameters 'call-methylation --methylation cpg' was
462 used to measure the frequency of CpG methylation. The ONT reads were aligned to
463 whole-genome assemblies using minimap2 [44] with default parameters. The script
464 'calculate_methylation_frequency.py', provided in the methplotlib package [64], was

465 then used to generate the methylation frequency. Methylation density is defined as the
466 number of occurrences with a methylation frequency greater than 0.8 within genomic
467 sliding windows (100-base pair window advanced at 50-base pair increments).
468

469 **CENH3 ChIP-seq**

470 We first used BLASTP to identify the amino acid sequence of CENH3 in *Chlorella*
471 *sorokiniana* NS4-2, with the reference sequence HTR12 from *Arabidopsis thaliana*
472 (Genbank accession: NP_001322867.1), and the '-evalue' parameter set to 1e-30. We
473 selected the coding gene NS|00958 as the unique matching result for the CENH3
474 sequence of the *Chlorella sorokiniana* NS4-2 species for subsequent antibody
475 preparation. An antigen with the peptide sequence 'MARTKQTKPSPSKRGA'
476 corresponding to the N terminus of NS4-2_CENH3 was used to produce the antibody
477 (Beijing QiWei YiCheng Tech Co., Ltd.).

478 Shanghai Jiayin Biotechnology Co., Ltd conducted ChIP assays following a
479 modified version of the standard crosslinking ChIP protocol. In summary, the culture
480 sample was harvested by centrifugation at 12,000 rpm for 2 min and sequentially
481 washed with sterilized BG-11 and distilled water. Subsequently, approximately 4g of
482 algae pellet was obtained and ground in liquid nitrogen. The ground sample was fixed
483 with 1% formaldehyde at room temperature for 15 min, followed by a 5-minute
484 treatment with 0.125 M glycine. After fixation, the sample was washed, resuspended
485 in a lysis buffer, and sonicated to generate DNA fragments. Following sonication,
486 immunoprecipitation was carried out using the custom antibody NS4-2_CENH3. The
487 immunoprecipitated complex was washed, and DNA was extracted. DNA purification
488 was performed using the Universal DNA Purification Kit (#DP214). A ChIP-seq
489 library was prepared by employing the ChIP-Seq DNA sample preparation kit
490 (NEBNext® UltraTMII DNA), adhering to the manufacturer's instructions. The
491 extracted DNA was ligated to specific adaptors and subsequently subjected to deep
492 sequencing on the Illumina Novaseq 6000 platform in 150bp paired-end mode. The
493 raw reads were trimmed using the Trimmomatic program [65] with default parameters.

494 The ChIP/input paired-end clean reads were mapped to the *Chlorella sorokiniana*

495 NS4-2 genome using BWA-mem [66] with default parameters. We determined the
496 ChIP/input coverage ratio by employing the DeepTools function 'bamCompare', with
497 a threshold of MAPQ ≥ 30. The output format was in BigWig. We carried out peak
498 calling using MACS2 [67] with the parameters '-t ChIP.bam -c input.bam -f BAM --
499 outdir chipseq -n chipseq -B --nomodel --keep-dup 1 -g 58000000'.

500

501 **Table 1. Genomic features of the *Chlorella sorokiniana* NS4-2 and *Chlorella***
502 ***pyrenoidosa* DBH assemblies.**

503

Genome features	NS4-2	UTEX1230[a]	DBH	FACHB-9[b]
Nuclear genome size (Mb)	58.11	58.36	53.41	56.77
No. of chromosomes	12	NA	13	NA
No. of contigs	13	18	14	8193
GC content (%)	63.95	63.95	66.94	66.71
Contigs N50 (bp)	4,352,225	3,818,101	4,302,578	9757
Scaffolds N50 (bp)	4,714,146	4,091,730	4,383,467	1,392,758
No. of protein-coding genes	12,691	12,871	11,327	10,283
Repeat content (%)	8.84	8.89	8.91	5.46

504

505 [a] UTEX1230 is a strain of *Chlorella sorokiniana* [36].
506 [b] FACHB-9 is a strain of *Chlorella pyrenoidosa* [37].

507

508 **Figure legends**

509 **Fig. 1. T2T genome assemblies of *Chlorella sorokiniana* NS4-2 and *Chlorella***
510 ***pyrenoidosa* DBH. (A)** Micrograph of *C. sorokiniana* NS4-2, with bar length = 20 μm;
511 **(B)** Karyotype diagram indicating that *C. sorokiniana* NS4-2 has 12 chromosomes; **(C)**
512 Hi-C contact map of the *C. sorokiniana* NS4-2 T2T assembly; **(D)** Circular plot
513 showing the principal genomic features of both the *C. sorokiniana* NS4-2 and *C.*
514 *pyrenoidosa* DBH assemblies: karyotype distribution (track a); GC density (track b);
515 gene density (track c); LINE element density (track d); LTR element density (track e);
516 putative centromeric loci (track f). Densities were calculated in 50 Kb windows. The
517 innermost of the circular plots represents the synteny relationship between *C.*
518 *sorokiniana* NS4-2 and *C. pyrenoidosa* DBH.

519

520 **Fig. 2. CENH3 binding of the centromeres in *Chlorella sorokiniana* NS4-2**.

521 Schematic overview of the chromosomes of *C. sorokiniana* NS4-2 showing, from top
522 to bottom per chromosome, the normalized CENH3 ChIP-seq read coverage,
523 predicted centromere regions, gene density, and LTR/Copia element density. Densities
524 were calculated in 5 Kb windows, and normalisation conducted in 1 Kb bins with 3
525 Kb smoothing.
526
527 **Fig. 3. Centromeric methylation features of *Chlorella sorokiniana* NS4-2. and**
528 ***Chlorella pyrenoidosa* DBH. (A)** Methylation density distribution of *C. sorokiniana*
529 NS4-2 chromosome 1. The red area represents the centromere. The close-up view
530 shows, from top to bottom, the distribution of methylation density, CENH3 ChIP-seq
531 signals, and the location of centromeric regions (parallel red band). Within the
532 centromeric region, ChIP-seq peaks called by MACS2 are shown in light green and
533 non-peaks in orange; **(B)** Box and violin plots showing methylation density in *C.*
534 *sorokiniana* NS4-2 for centromeric (CEN) and non-centromeric (non-CEN) regions,
535 and for peaks (CEN_peaks) and non-peaks (CEN_nonPeaks) within the former; **(C)**
536 Box and violin plots showing methylation density in *C. pyrenoidosa* DBH for CEN
537 and non-CEN regions. The statistical analysis was conducted via the Wilcoxon test.
538
539 **Fig. 4. Centromeric size and composition in green algae. (A)** Phylogenomic tree of
540 six green algae species with the red algae *Cyanidioschyzon merolae* utilized as an
541 outgroup. The figure on the right illustrates a comparison of genome size and
542 centromere length among these selected species; **(B)** Relationship between the (log-
543 transformed) lengths of chromosome and centromere; **(C-H)** The distribution of
544 different repeat subfamilies in centromeres (CEN; dark blue) and across the genome
545 (non-CEN; light gray) for *Chlamydomonas reinhardtii* **(C)**, *Chromochloris*
546 *zofingiensis* **(D)**, *Coccomyxa subellipsoidea* **(E)**, *Chlorella* sp. SLA-04 **(F)**, *Chlorella*
547 *sorokiniana* NS4-2 **(G)** and *Chlorella pyrenoidosa* DBH **(H)**.
548
549 **Fig. 5. Evolution of centromeric transposable elements in green algae.**
550 **(A)** Predominant transposable elements (TEs) in the centromeres of green algal

species. Different colors represent different TE families. MRCA represents the most recent common ancestor; **(B)** Synteny analyses of the 13 chromosomes of *Chlorella* sp. SLA-04, 12 chromosomes of *Chlorella sorokiniana* NS4-2 and 13 chromosomes of *Chlorella pyrenoidosa* DBH. The corresponding syntenic regions are marked using the same color. Approximate locations of putative centromeres are indicated by ellipses. The red circles represent the collinearity of the whole chromosomes; **(C)** Dot plot representation of the centromeres showing the collinearity of entire chromosomes from three *Chlorella* species, along with their TE composition. As with panel (A), different colors represent different TE families.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The whole genome sequence data reported in this paper have been deposited in the Genome Warehouse in National Genomics Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation (GWH: GWHBKBA00000000 for *Chlorella sorokiniana* NS4-2; GWH: GWHBKBB00000000 for *Chlorella pyrenoidosa* DBH) and is publicly accessible at https://ngdc.cncb.ac.cn/gwh. The genome annotations of *Chlorella sorokiniana* NS4-2 and *Chlorella pyrenoidosa* DBH have been deposited in Figshare (10.6084/m9.figshare.24503881). The raw data for the PacBio HiFi reads, ONT long-reads, Hi-C Illumina short reads and ChIP-seq data have been deposited in the Genome Sequence Archive at the National Genomics Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation (GSA: CRA007945), and are publicly accessible at https://ngdc.cncb.ac.cn/gsa.

Competing interests

The authors declare no competing interests.

Funding

This study was supported by the National Key R&D Program of China (2022YFC3400300) and the National Natural Science Foundation of China (32200510; 32070663; 32170257; 32125009; 62172325).

Authors' contributions

K.Y. and W.M. conceived and designed the project. K.Y. and B.W. designed the experiments. B.W., Y.J., S.G., H.G. and X.Y. analyzed the results. B.W., Y.J., N.D., J.Y., S.B, S.G., W.H., S.W., and H.G. performed all the experiments and analyzed the data. B.W., Y.J. and N.D. prepared the figures. B.W., Y.J. and S.B. wrote the manuscript text. All authors reviewed and approved the manuscript.

References

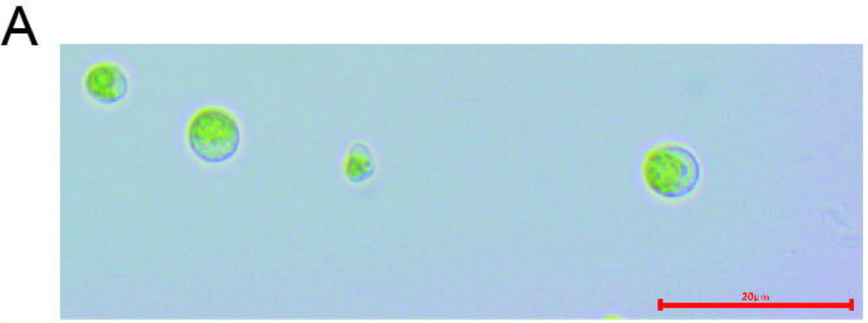
1. Talbert PB, Henikoff S. What makes a centromere? Exp Cell Res. 2020;389(2):111895.
2. McAinsh AD, Marston AL. The four causes: the functional architecture of centromeres and kinetochores. Annu Rev Genet. 2022;56:279-314.
3. Henikoff S, Ahmad K, Malik HS. The centromere paradox: stable inheritance with rapidly evolving DNA. Science. 2001;293(5532):1098-102.
4. Melters DP, Bradnam KR, Young HA, Telis N, May MR, Ruby JG, et al. Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. Genome Biol. 2013;14(1):R10.
5. Miga KH. Centromere studies in the era of 'telomere-to-telomere' genomics. Exp Cell Res. 2020;394(2):112127.
6. Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods. 2021;18(2):170-5.
7. Rautiainen M, Nurk S, Walenz BP, Logsdon GA, Porubsky D, Rhie A, et al. Telomere-to-telomere assembly of diploid chromosomes with Verkko. Nat Biotechnol. 2023;41(10):1474-82.
8. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, et al. The complete sequence of a human genome. Science. 2022;376(6588):44-53.
9. Naish M, Alonge M, Wlodzimierz P, Tock AJ, Abramson BW, Schmücker A, et al. The genetic and epigenetic landscape of the *Arabidopsis* centromeres. Science.

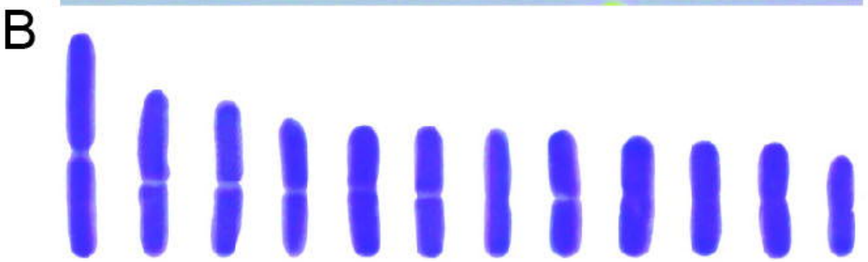
2021;374(6569):eabi7489.
10. Wang B, Yang X, Jia Y, Xu Y, Jia P, Dang N, et al. High-quality *Arabidopsis thaliana* genome assembly with Nanopore and HiFi long reads. Genomics Proteomics Bioinformatics. 2022;20(1):4-13.
11. Song JM, Xie WZ, Wang S, Guo YX, Koo DH, Kudrna D, et al. Two gap-free reference genomes and a global view of the centromere architecture in rice. Mol Plant. 2021;14(10):1757-67.
12. Chen J, Wang Z, Tan K, Huang W, Shi J, Li T, et al. A complete telomere-to-telomere assembly of the maize genome. Nat Genet. 2023;55(7):1221-31.
13. Gao S, Yang X, Guo H, Zhao X, Wang B, Ye K. HiCAT: a tool for automatic annotation of centromere structure. Genome Biol. 2023;24(1):58.
14. Presting GG. Centromeric retrotransposons and centromere function. Curr Opin Genet Dev. 2018;49:79-84.
15. Huang Y, Ding W, Zhang M, Han J, Jing Y, Yao W, et al. The formation and evolution of centromeric satellite repeats in *Saccharum* species. Plant J. 2021;106(3):616-29.
16. Ahmed HI, Heuberger M, Schoen A, Koo DH, Quiroz-Chavez J, Adhikari L, et al. Einkorn genomics sheds light on history of the oldest domesticated wheat. Nature. 2023;620(7975):830-8.
17. Craig RJ, Hasan AR, Ness RW, Keightley PD. Comparative genomics of *Chlamydomonas*. Plant Cell. 2021;33(4):1016-41.
18. Bi G, Zhao S, Yao J, Wang H, Zhao M, Sun Y, et al. Near telomere-to-telomere genome of the model plant *Physcomitrium patens*. Nat Plants. 2024;10(2):327-43.
19. Windler M, Stuart R, Deutzmann JS, Mayali X, Navid A, D'haeseleer P, et al. Bacterial exometabolites influence *Chlamydomonas* cell cycle and double algal productivity. FEMS Microbiol Ecol. 2022;98(9):fiac091.
20. Cecchin M, Marcolungo L, Rossato M, Girolomoni L, Cosentino E, Cuine S, et al. *Chlorella vulgaris* genome assembly and annotation reveals the molecular basis for metabolic acclimation to high light conditions. Plant J. 2019;100(6):1289-305.
21. Hu J, Wang Z, Sun Z, Hu B, Ayoola AO, Liang F, et al. An efficient error correction and accurate assembly tool for noisy long reads. bioRxiv. doi:10.1101/2023.03.09.531669v1.
22. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol. 2017;13(6):e1005595.
23. Payne ZL, Penny GM, Turner TN, Dutcher SK. A gap-free genome assembly of *Chlamydomonas reinhardtii* and detection of translocations induced by CRISPR-mediated mutagenesis. Plant Commun. 2023;4(2):100493.
24. Craig RJ, Gallaher SD, Shu S, Salomé PA, Jenkins JW, Blaby-Haas CE, Purvine SO, et al. The *Chlamydomonas* Genome Project, version 6: Reference assemblies for mating-type plus and minus strains reveal extensive structural mutation in the laboratory. Plant Cell. 2023;35(2):644-72.
25. Blanc G, Agarkova I, Grimwood J, Kuo A, Brueggeman A, Dunigan DD, et al. The genome of the polar eukaryotic microalga *Coccomyxa subellipsoidea* reveals

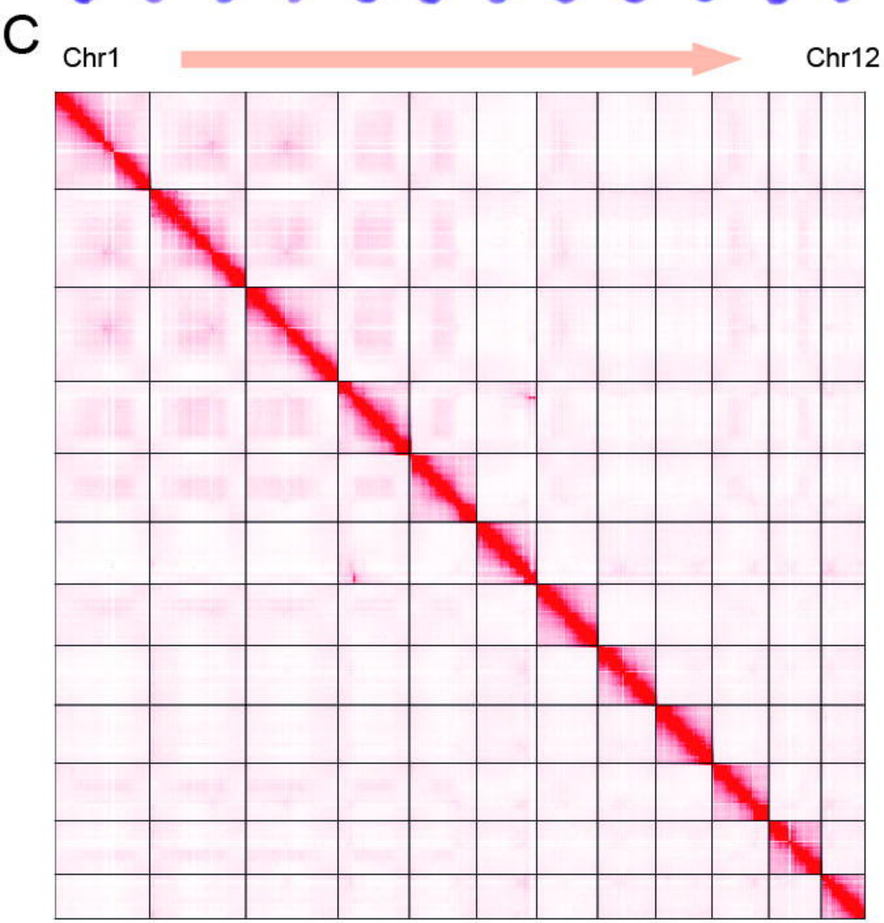
661 traits of cold adaptation. Genome Biol. 2012;13(5):R39.
662 26. Roth MS, Cokus SJ, Gallaher SD, Walter A, Lopez D, Erickson E, et al.
663 Chromosome-level genome assembly and transcriptome of the green alga
664 *Chromochloris zofingiensis* illuminates astaxanthin production. Proc Natl Acad
665 Sci U S A. 2017;114(21):E4296-E4305.
666 27. Mikheenko A, Bzikadze AV, Gurevich A, Miga KH, Pevzner PA. TandemTools:
667 mapping long reads and assessing/improving assembly quality in extra-long
668 tandem repeats. Bioinformatics. 2020;36 Suppl 1:i75-i83.
669 28. Zhang W, Lee HR, Koo DH, Jiang J. Epigenetic modification of centromeric
670 chromatin: hypomethylation of DNA sequences in the CENH3-associated
671 chromatin in *Arabidopsis thaliana* and maize. Plant Cell. 2008;20(1):25-34.
672 29. Ni P, Nie F, Zhong Z, Xu J, Huang N, Zhang J, et al. DNA 5-methylcytosine
673 detection and methylation phasing using PacBio circular consensus sequencing.
674 Nat Commun. 2023;14(1):4054.
675 30. Goemann CL, Wilkinson R, Henriques W, Bui H, Goemann HM, Carlson RP, et al.
676 Genome sequence, phylogenetic analysis, and structure-based annotation reveal
677 metabolic potential of *Chlorella* sp. SLA-04. Algal Res. 2023;69:102943.
678 31. Nozaki H, Takano H, Misumi O, Terasawa K, Matsuzaki M, Maruyama S, et al. A
679 100%-complete sequence reveals unusually simple genomic features in the hot-
680 spring red alga *Cyanidioschyzon merolae*. BMC Biol. 2007;5:28.
681 32. Kanesaki Y, Imamura S, Matsuzaki M, Tanaka K. Identification of centromere
682 regions in chromosomes of a unicellular red alga, *Cyanidioschyzon merolae*.
683 FEBS Lett. 2015;589(11):1219-24.
684 33. Plačková K, Bureš P, Zedek F. Centromere size scales with genome size across
685 Eukaryotes. Sci Rep. 2021;11(1):19811.
686 34. Talbert PB, Henikoff S. Centromeres convert but don't cross. PLoS Biol.
687 2010;8(3):e1000326.
688 35. Michael TP. Plant genome size variation: bloating and purging DNA. Brief Funct
689 Genomics. 2014;13(4):308-17.
690 36. Hovde BT, Hanschen ER, Tyler CRS, Lo CC, Kunde Y, Davenport K, et al.
691 Genomic characterization reveals significant divergence within *Chlorella*
692 *sorokiniana* (Chlorellales, Trebouxiophyceae). Algal Res. 2018;35:449-61.
693 37. Fan J, Ning K, Zeng X, Luo Y, Wang D, Hu J, et al. Genomic foundation of
694 starch-to-lipid switch in oleaginous *Chlorella* spp. Plant Physiol.
695 2015;169(4):2444-61.
696 38. Logsdon GA, Vollger MR, Hsieh P, Mao Y, Liskovykh MA, Koren S, et al. The
697 structure, function and evolution of a complete human chromosome 8. Nature.
698 2021;593(7857):101-7.
699 39. Altemose N, Logsdon GA, Bzikadze AV, Sidhwani P, Langley SA, Caldas GV, et
700 al. Complete genomic and epigenetic maps of human centromeres. Science.
701 2022;376(6588):eabl4178.
702 40. Han J, Masonbrink RE, Shan W, Song F, Zhang J, Yu W, et al. Rapid proliferation
703 and nucleolar organizer targeting centromeric retrotransposons in cotton. Plant J.
704 2016;88(6):992-1005.

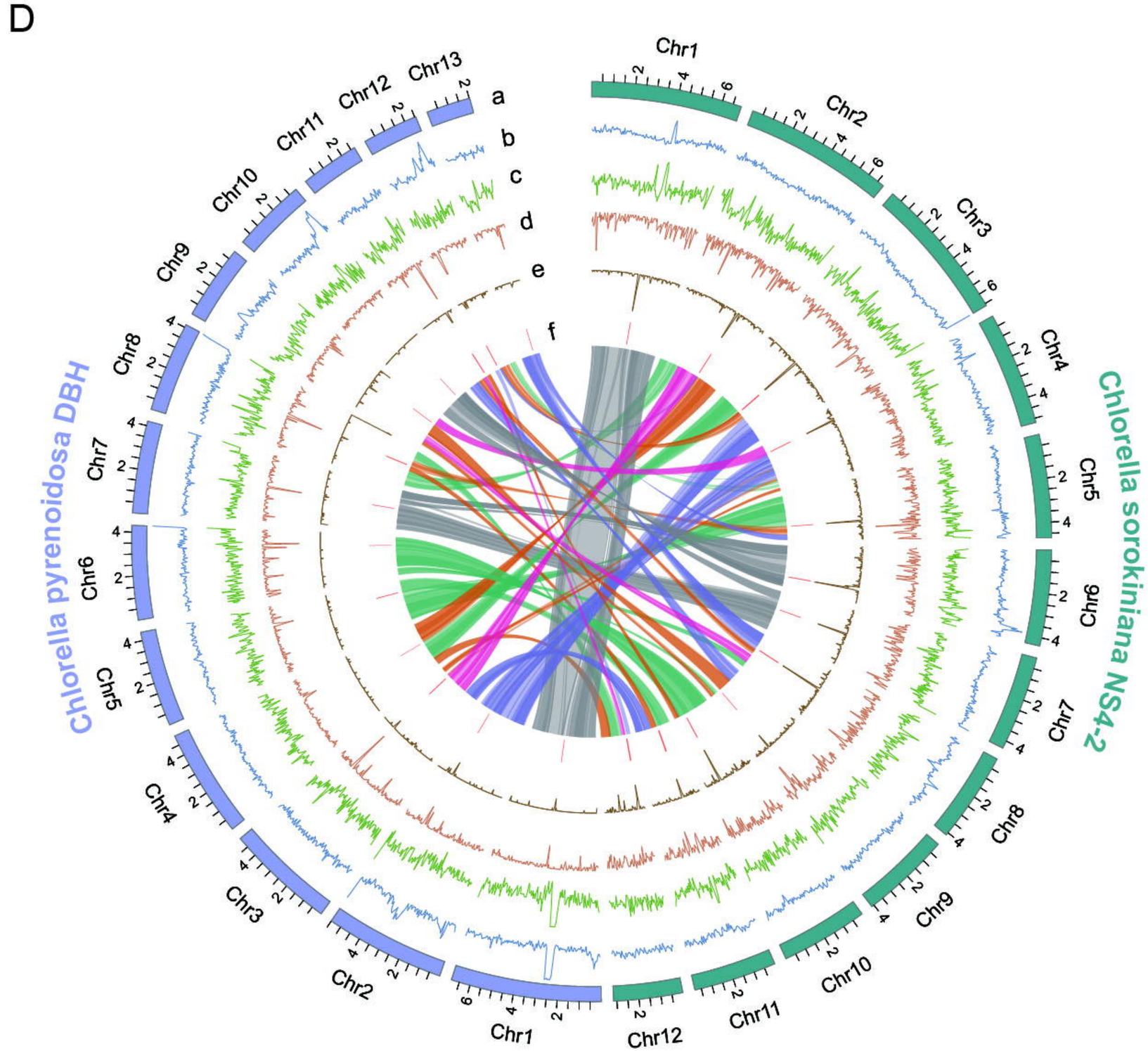
705 41. Suzuki Y, Myers EW, Morishita S. Rapid and ongoing evolution of repetitive
706 sequence structures in human centromeres. Sci Adv. 2020;6(50):eabd9230.
707 42. Bensasson D, Zarowiecki M, Burt A, Koufopanou V. Rapid evolution of yeast
708 centromeres in the absence of drive. Genetics. 2008;178(4):2161-7.
709 43. Hu J, Fan J, Sun Z, Liu S. NextPolish: a fast and efficient genome polishing tool
710 for long-read assembly. Bioinformatics. 2020 ;36(7):2253-5.
711 44. Durand NC, Shamim MS, Machol I, Rao SS, Huntley MH, Lander ES, et al.
712 Juicer provides a one-click system for analyzing loop-resolution Hi-C
713 experiments. Cell Syst. 2016;3(1):95-8.
714 45. Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, et al.
715 Juicebox provides a visualization system for Hi-C contact maps with unlimited
716 zoom. Cell Syst. 2016;3(1):99-101.
717 46. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics.
718 2018;34(18):3094-100.
719 47. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence
720 Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9.
721 48. Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, et al. MCScanX: a toolkit for
722 detection and evolutionary analysis of gene synteny and collinearity. Nucleic
723 Acids Res. 2012;40(7):e49.
724 49. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al.
725 Circos: an information aesthetic for comparative genomics. Genome Res.
726 2009;19(9):1639-45.
727 50. Krumsiek J, Arnold R, Rattei T. Gepard: a rapid and sensitive tool for creating
728 dotplots on genome scale. Bioinformatics. 2007;23(8):1026-8.
729 51. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO:
730 assessing genome assembly and annotation completeness with single-copy
731 orthologs. Bioinformatics. 2015;31(19):3210-2.
732 52. Gao S, Yang X, Sun J, Zhao X, Wang B, Ye K. IAGS: Inferring ancestor genome
733 structure under a wide range of evolutionary scenarios. Mol Biol Evol.
734 2022;39(3):msac041.
735 53. Borodovsky M, Lomsadze A. Eukaryotic gene prediction using GeneMark.hmm-E
736 and GeneMark-ES. Curr Protoc Bioinformatics. 2011;4, 4.6.1-4.6.10.
737 54. Stanke M, Morgenstern B. AUGUSTUS: a web server for gene prediction in
738 eukaryotes that allows user-defined constraints. Nucleic Acids Res. 2005;33(Web
739 Server issue):W465-7.
740 55. Cantarel BL, Korf I, Robb SM, Parra G, Ross E, Moore B et al. MAKER: an easy-
741 to-use annotation pipeline designed for emerging model organism genomes.
742 Genome Res. 2008;18(1):188-96.
743 56. Merchant SS, Prochnik SE, Vallon O, Harris EH, Karpowicz SJ, Witman GB, et al.
744 The *Chlamydomonas* genome reveals the evolution of key animal and plant
745 functions. Science. 2007;318(5848):245-50.
746 57. Blanc G, Duncan G, Agarkova I, Borodovsky M, Gurnon J, Kuo A, et al. The
747 *Chlorella variabilis* NC64A genome reveals adaptation to photosymbiosis,
748 coevolution with viruses, and cryptic sex. Plant Cell. 2010;22(9):2943-55.

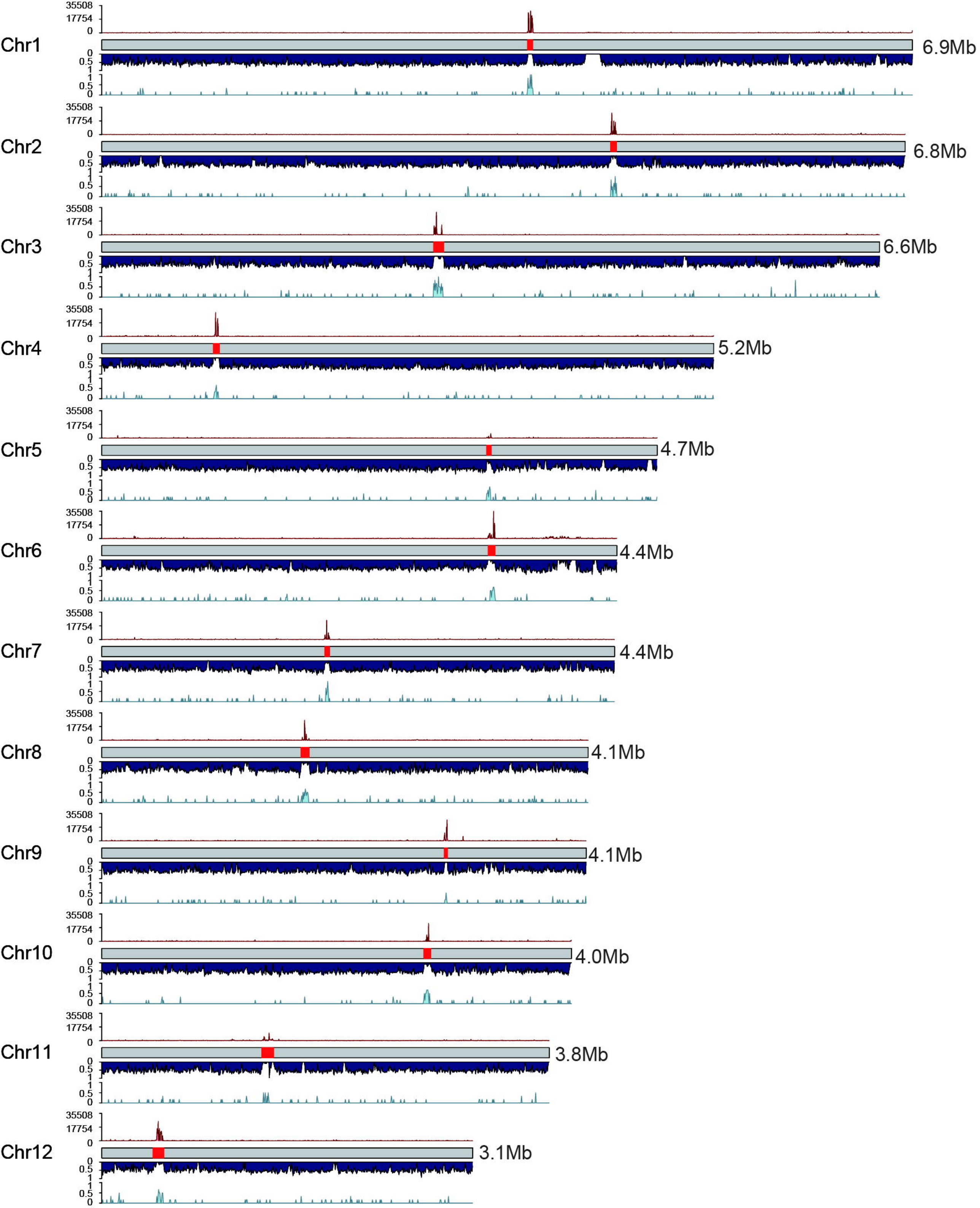
749 58. Arriola MB, Velmurugan N, Zhang Y, Plunkett MH, Hondzo H, Barney BM.
750 Genome sequences of *Chlorella sorokiniana* UTEX 1602 and *Micractinium*
751 *conductrix* SAG 241.80: implications to maltose excretion by a green alga. Plant
752 J. 2018;93(3):566-86.
753 59. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, et al.
754 RepeatModeler2 for automated genomic discovery of transposable element
755 families. Proc Natl Acad Sci U S A. 2020;117(17):9451-7.
756 60. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of
757 protein or nucleotide sequences. Bioinformatics. 2006;22(13):1658-9.
758 61. Gel B, Serra E. karyoploteR: an R/Bioconductor package to plot customizable
759 genomes displaying arbitrary data. Bioinformatics. 2017;33(19):3088-90.
760 62. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for
761 comparative genomics. Genome Biol. 2019;20(1):238.
762 63. Kumar S, Stecher G, Tamura K. MEGA7: molecular evolutionary genetics
763 analysis version 7.0 for bigger datasets. Mol Biol Evol. 2016;33(7):1870-4.
764 64. De Coster W, Stovner EB, Strazisar M. Methplotlib: analysis of modified
765 nucleotides from nanopore sequencing. Bioinformatics. 2020;36(10):3236-8.
766 65. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina
767 sequence data. Bioinformatics. 2014;30(15):2114-20.
768 66. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler
769 transform. Bioinformatics. 2009;25(14):1754-60.
770 67. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-
771 based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9(9):R137.
772

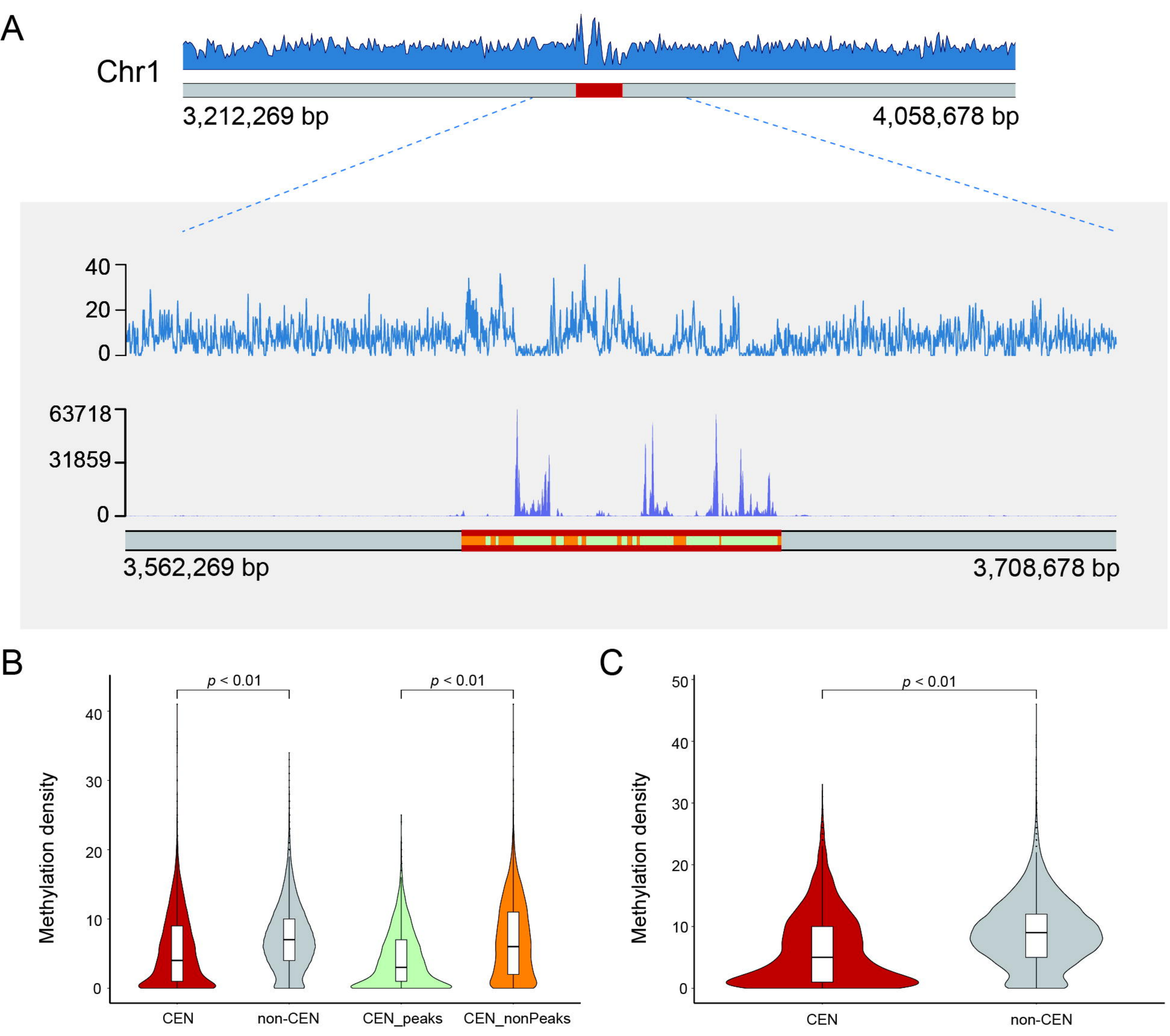

A
20μm

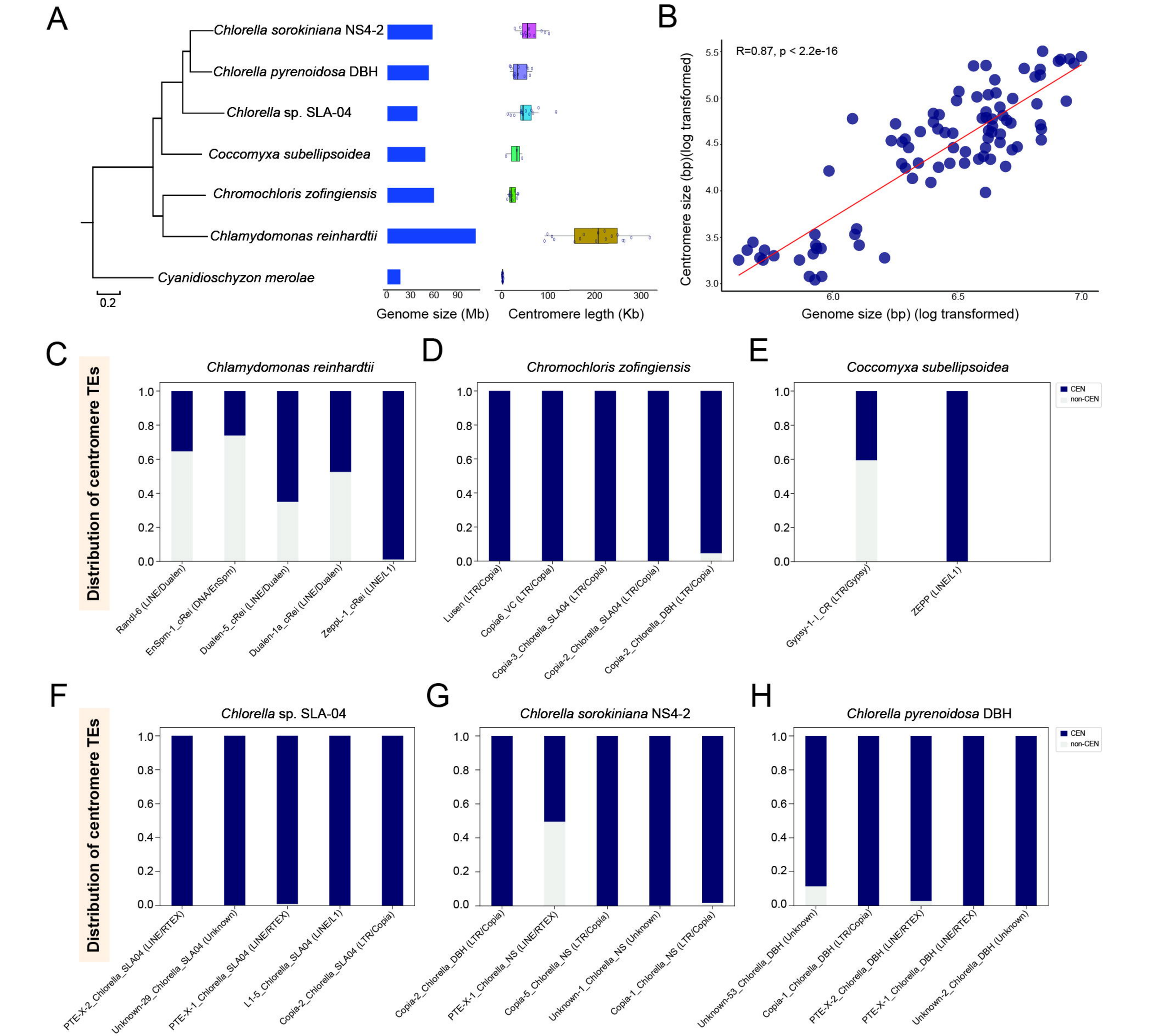

B

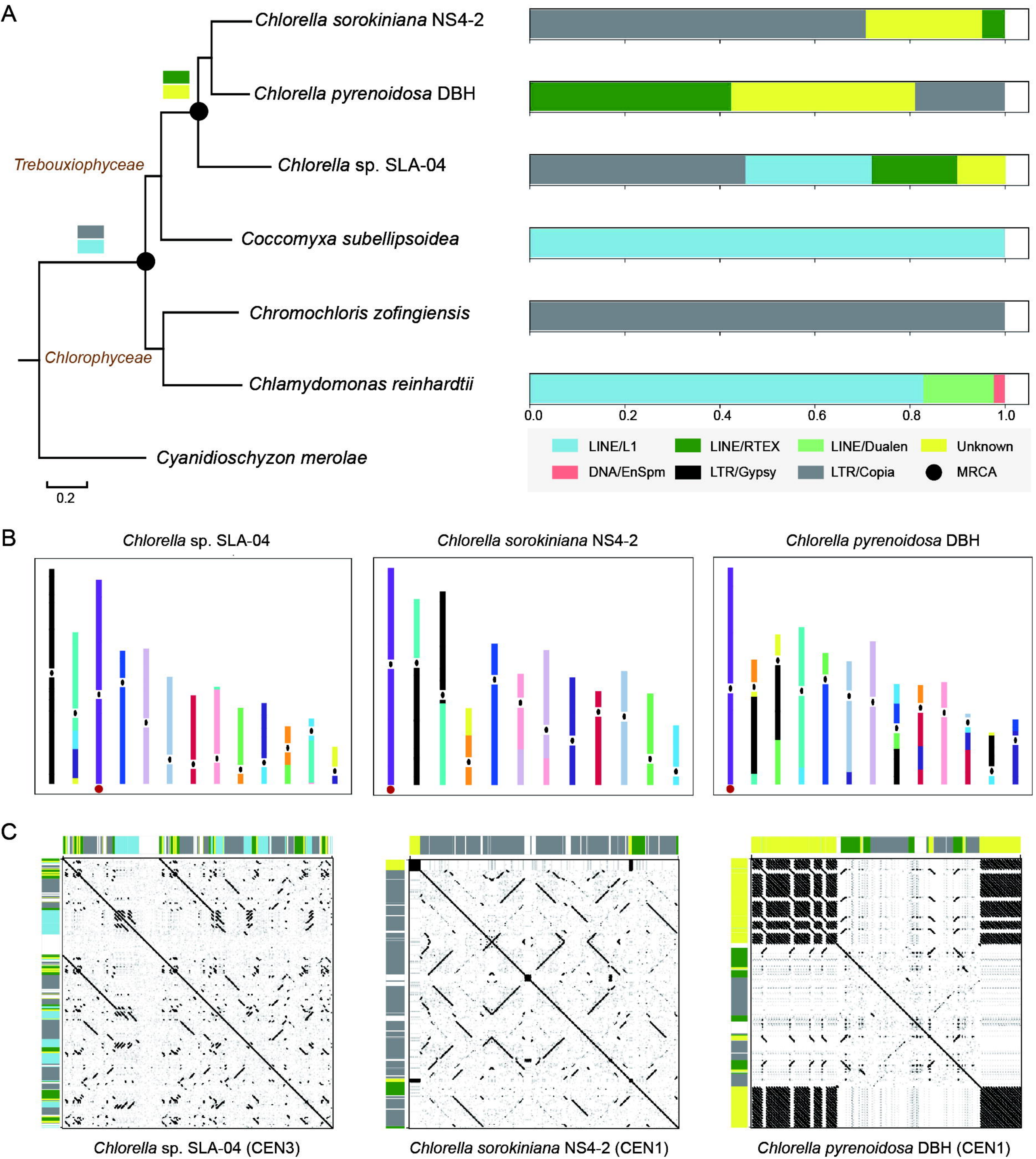

C
Chr1
Chr12


D
Chlorella pyrenoidosa DBH
Chlorella sorokiniana NS4-2
Chr1
Chr2
Chr3
Chr4
Chr5
Chr6
Chr7
Chr8
Chr9
Chr10
Chr11
Chr12
Chr13
a
b
c
d
e
f


Chr1
6.9Mb
Chr2
6.8Mb
Chr3
6.6Mb
Chr4
5.2Mb
Chr5
4.7Mb
Chr6
4.4Mb
Chr7
4.4Mb
Chr8
4.1Mb
Chr9
4.1Mb
Chr10
4.0Mb
Chr11
3.8Mb
Chr12
3.1Mb
35508
17754
0
0.5
1


A
Chr1
3,212,269 bp
4,058,678 bp
40
20
0
63718
31859
0
3,562,269 bp
3,708,678 bp
B
p < 0.01
p < 0.01
Methylation density
40
30
20
10
0
CEN
non-CEN
CEN_peaks
CEN_nonPeaks
C
p < 0.01
Methylation density
50
40
30
20
10
0
CEN
non-CEN


A
Chlorella sorokiniana NS4-2
Chlorella pyrenoidosa DBH
Chlorella sp. SLA-04
Coccomyxa subellipsoidea
Chromochloris zofingiensis
Chlamydomonas reinhardtii
Cyanidioschyzon merolae
0.2
0 30 60 90
Genome size (Mb)
0 100 200 300
Centromere legth (Kb)
B
R=0.87, p < 2.2e-16
Centromere size (bp)(log transformed)
Genome size (bp) (log transformed)
C
Chlamydomonas reinhardtii
Distribution of centromere TEs
Randl-6 (LINE/Dualen)
EnSpm-1_cRei (DNA/EnSpm)
Dualen-5_cRei (LINE/Dualen)
Dualen-1a_cRei (LINE/Dualen)
ZeppL-1_cRei (LINE/L1)
D
Chromochloris zofingiensis
Lusen (LTR/Copia)
Copia6_VC (LTR/Copia)
Copia-3_Chlorella_SLA04 (LTR/Copia)
Copia-2_Chlorella_SLA04 (LTR/Copia)
Copia-2_Chlorella_DBH (LTR/Copia)
E
Coccomyxa subellipsoidea
CEN
non-CEN
Gypsy-1-I_CR (LTR/Gypsy)
ZEPP (LINE/L1)
F
Chlorella sp. SLA-04
Distribution of centromere TEs
PTE-X-2_Chlorella_SLA04 (LINE/RTEX)
Unknown-29_Chlorella_SLA04 (Unknown)
PTE-X-1_Chlorella_SLA04 (LINE/RTEX)
L1-5_Chlorella_SLA04 (LINE/L1)
Copia-2_Chlorella_SLA04 (LTR/Copia)
G
Chlorella sorokiniana NS4-2
Copia-2_Chlorella_DBH (LTR/Copia)
PTE-X-1_Chlorella_NS (LINE/RTEX)
Copia-5_Chlorella_NS (LTR/Copia)
Unknown-1_Chlorella_NS (Unknown)
Copia-1_Chlorella_NS (LTR/Copia)
H
Chlorella pyrenoidosa DBH
CEN
non-CEN
Unknown-53_Chlorella_DBH (Unknown)
Copia-1_Chlorella_DBH (LTR/Copia)
PTE-X-2_Chlorella_DBH (LINE/RTEX)
PTE-X-1_Chlorella_DBH (LINE/RTEX)
Unknown-2_Chlorella_DBH (Unknown)


A
Chlorella sorokiniana NS4-2
Chlorella pyrenoidosa DBH
Chlorella sp. SLA-04
Coccomyxa subellipsoidea
Chromochloris zofingiensis
Chlamydomonas reinhardtii
Cyanidioschyzon merolae
Trebouxiophyceae
Chlorophyceae
0.2
0.0
0.2
0.4
0.6
0.8
1.0
LINE/L1
LINE/RTEX
LINE/Dualen
Unknown
DNA/EnSpm
LTR/Gypsy
LTR/Copia
MRCA
B
Chlorella sp. SLA-04
Chlorella sorokiniana NS4-2
Chlorella pyrenoidosa DBH
C
Chlorella sp. SLA-04 (CEN3)
Chlorella sorokiniana NS4-2 (CEN1)
Chlorella pyrenoidosa DBH (CEN1)