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)

Acknowledgments

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).

Conflict of interest

The authors declare that they have no conflicts of interest.

Abstract

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. In this study, we constructed near telomere-to-telomere (T2T) assemblies for two Trebouxiophyceae species, *Chlorella sorokiniana* NS4-2 and *C. 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 centromere sequences using CENH3 ChIP-seq and found that, similar to humans and higher plants, the centromere CENH3 signals of green algae display a pattern of hypomethylation. Interestingly, the centromeres of both species largely comprised transposable elements (TEs), although their composition showed significant differences. In *Chlorella* species, centromeres primarily consist of multiple elements, such as LINE/L1, LTR/Copia, LINE/RTEX, and Unknown families. However, in early-evolved species such as *Chamydomonas reinhardtii*, *Coccomyxa subellipsoidea*, and *Chromochloris zofingiensis*, centromere composition has evolved towards optimization with a single-element composition. Through phylogenomic and centromeric compositions analysis, We found that the centromeres of the most recent common ancestor of green algae are composed of two transposable element families, LINE/L1 and LTR/Copia. Moreover, we observed significant differences in the composition and structure of centromeres among chromosomes with strong collinearity within the *Chlorella* genus, suggesting that centromeres in green algae evolve more rapidly than non-centromeric regions. 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.

Introduction

Centromeres play a critical role in ensuring the accurate segregation of chromosomes during both mitosis and meiosis (Talbert and Henikoff, 2022). They are specialized, constricted regions on chromosomes responsible for 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 (McAinsh and Marston, 2022). The 'centromere paradox' refers to the apparent contradiction between their 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 (Henikoff et al. 2001; Melters et al. 2013). This variability highlights the rapid evolution of centromeres.

The high prevalence of satellite repeats within centromere sequences also presents a considerable challenge in genome assembly for this region, frequently leading to the substantial loss of sequence information (Miga, 2020). 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 (Cheng et al. 2021; Rautiainen et al. 2023), significant progress has been made towards the complete assembly of centromeres in many model species, including but not limited to human (Nurk et al. 2022), *Arabidopsis thaliana* (Naish et al. 2021; Wang et al. 2022), rice (Song et al. 2021), and maize (Chen et al. 2023). Notably, both human and higher plant centromeres primarily comprise numerous satellite sequences, organized into higher-order repeat arrays (Miga, 2020; Gao et al. 2023). Interestingly, the centromeres of the oldest domesticated wheat (*Triticum monococcum*) were found to predominantly feature transposable elements (TEs) (Ahmed et al. 2023). In fact, a few years ago, scholars proposed that the plant centromeric satellites might result from the sequences of TEs specifically targeting the centromere (Presting 2018; Huang et al.

2021). Through high-quality genome assembly and putative centromere sequence analysis of the moss *Physcomitrium patens* and the model green alga *Chlamydomonas reinhardtii*, it was discovered that their centromeres are indeed composed of TEs (Craig et al. 2021; Bi et al. 2024). Whether centromeres in other green algae, more distantly related to *C. 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 the T2T level, characterising and experimentally validating their centromeres. We found that the centromeres of the green algae species selected in this study are all composed of transposable elements (TEs), albeit with significant differences in composition. *Chlamydomonas reinhardtii* and *Coccomyxa subellipsoidea* predominantly feature elements from the LINE/L1 family, while *Chromochloris zofingiensis* is primarily composed of elements from the LTR/Copia family. In contrast, species within the *Chlorella* genus exhibit a more diverse centromere composition, with major constituents including families such as LTR/Copia, LINE/L1, and LINE/RTEX. This suggests that these species are in the process of optimizing one element to become the predominant constituent during evolution. Overall, our results give insight into the centromere evolution of green algae, which may be generalised to early plants, as well as providing high-quality genomic resources to facilitate future comparative analyses.

Results

Telomere-to-telomere assemblies of two *Chlorella* species

To investigate centromere evolution within the context of green algae, we assembled the genomes of two *Chlorella* species, *C. sorokiniana* NS4-2 and *C. pyrenoidosa* DBH (basionym: *Auxenochlorella pyrenoidosa*), both of which belong to the Trebouxiophyceae class, relatively distant from the Chlorophyceae class of *Chlamydomonas reinhardtii* in terms of phylogenetic distance (divergence time: 827 MYA, http://timetree.org/) (**Figure S1**) (Blanc et al. 2010). Their morphology is similar, but they vary in their chromosome numbers, with NS4-2 having 12

115 chromosomes and DBH having 13 (**Figure 1A, Figure 1B, Figure S2A and Figure**
116 **S2B**). The genomes of NS4-2 and DBH were sequenced using HiFi reads with high
117 coverage, namely 438× and 308×, respectively (**Table S1**). We assembled the HiFi
118 reads (**Figure S3**) to produce a set of 16 contigs. These contigs had a minimum length
119 of 141,360 bp and were selected based on the criterion of length exceeding 100 Kb.
120 Following this, we generated a final set of 13 contigs by excluding the genome of the
121 putative alga-benefitting commensal bacterium *Arthrobacter* sp. (Windler et al. 2022)
122 as well as the fragments of rDNA and plastid DNA (21 Kb~107 Kb). Among these
123 contigs, 11 contained the plant-type telomeric repeat unit (CCCTAAA/TTTAGGG) at
124 both ends, signifying their completeness. Additionally, two contigs were found to
125 have rDNA units at one end, as observed through collinearity analysis in comparison
126 with the closely related species *Chlorella vulgaris* which was scaffold by optical
127 mapping (Cecchin et al. 2019) (**Figure S4A**). To consolidate the rDNA ends
128 originating from the same chromosome, we merged the two contigs and inserted 100
129 Ns as spacers. As a result of this process, the NS4-2 genome now represents a T2T
130 assembly, comprised of 12 chromosomes. This conclusion is further supported by
131 karyotyping results (**Figure 1B**) and Hi-C contact maps (**Figure 1C**). In the case of
132 DBH, we initially assembled the HiFi reads using hifiasm (Cheng et al. 2021) (**Figure**
133 **S3B**), which generated 38 contigs, each exceeding 100 Kb in length. However, upon
134 closer examination, we discovered that only three of these contigs contained the
135 telomeric repeat unit at both ends. This finding suggests that achieving a T2T-level
136 assembly for DBH is more challenging compared to NS4-2 (**Figure S3A**). To
137 overcome this challenge and obtain a T2T-level assembly for DBH, we generated
138 ONT ultra-long reads (with a read length N50 of 59 Kb) at a high coverage of 368×
139 (**Table S1**). We used *NextDenovo* (Hu et al. 2023) to assemble the ONT long reads,
140 which initially yielded a set of 14 contigs. Out of the total contigs obtained, 10 of
141 them exhibited the presence of telomeric repeat units at both ends, indicating the
142 complete resolution of the telomere regions. However, two contigs only had rDNA
143 units at one end, suggesting unresolved telomere regions at the other. Additionally,
144 the remaining two contigs contained the telomeric repeat unit on one end, indicating

145 that the telomere regions of the other end were not fully resolved. To address this
146 issue and achieve a more comprehensive assembly, we utilized the HiFi contigs to
147 replace the non-resolved ONT contigs (**Figure S3B**). By incorporating these updated
148 HiFi contigs, we successfully achieved a gap-free assembly of 14 contigs. This
149 updated assembly provides a more complete representation of the genome, with
150 improved resolution and continuity. In a similar vein, for DBH, the collinearity
151 analysis also revealed two contigs with rDNA units potentially originating from the
152 same chromosome (**Figure S4B**). Following a similar approach to NS4-2, the rDNA
153 ends were merged, and 100 Ns were introduced as spacers. This led to the generation
154 of a T2T assembly with 13 chromosomes, again supported by karyotyping results
155 (**Figure S2B**) and Hi-C contact maps (**Figure S2C**). The evaluation of genome
156 completeness using Benchmarking Universal Single-Copy Orthologs (BUSCO)
157 revealed that both genomes exhibited high genome completeness, with NS4-2
158 reaching 98.3% and DBH achieving 99.0%. The assembly of these two genomes
159 represents a significant improvement compared to the existing genomes from the
160 same species (**Table 1**). NS4-2 and DBH were predicted to have 12,691 and 11,327
161 protein coding genes, respectively. NS-2 and DBH exhibit a significant collinearity
162 relationship, accompanied by chromosomal rearrangements (**Figure 1D**). All HiFi
163 reads greater than 15 Kb were selected and used to perform mitochondrial and
164 chloroplast genome assembly in Unicycler v. 0.5.0 (software optimized for circular
165 genome assemblies) with default parameters (Wick et al. 2017). The chloroplast
166 genome sizes of NS4-2 and DBH are 109,738 bp and 107,436 bp, respectively; while
167 the mitochondrial genome sizes are 53,629 bp and 69,538 bp, respectively (GenBank
168 accession: OP311589-OP311592).

169 **Putative centromere identification and CENH3 ChIP-seq validation**

170 Recent advancements in sequencing technologies have shed light on the crucial role
171 of TEs in influencing the characteristics of centromeric sequences among green algal
172 species (Craig et al. 2021; Payne et al. 2023). Notably, studies on species such as
173 *Chlamydomonas reinhardtii* CC-4532 (v. 6.1) (Craig et al. 2023) and *Coccomyxa*
174 *subellipsoidea* C-169 (v. 3.0) (Blanc et al. 2012) have revealed the prevalence of

175 LINE/L1 elements within their genomes, as well as LTR/Copia elements in
176 *Chromochloris zofingiensis* (Roth et al. 2017), further emphasizing the impact of TEs
177 in shaping these organism's centromeric regions. Upon examining the centromere
178 regions of these species, a striking abundance of TEs becomes evident. These TEs
179 form a unique and discernible structure when performing dot plots for self-alignment
180 analysis (**Figure S5**). Notably, within this structure, specific clusters stand out, such
181 as the *Zepp*-like cluster in *C. reinhardtii* and the *Zepp* cluster in *C. subellipsoidea*
182 (**Figure S5**). The presence of these distinct TE clusters further underscores the
183 significance of TEs in shaping the centromeric regions of these species. Leveraging
184 this distinctive characteristic, we conducted dot plot analyses on a per-chromosome
185 basis to pinpoint the potential centromeric regions within the genomes of NS4-2
186 (**Figure S6**) and DBH (**Figure S7**). In the case of NS4-2, we successfully validated
187 putative centromeres for all 12 chromosomes using VerityMap (Mikheenko et al.
188 2020) (**Figure S8**), with lengths ranging from 28,465 bp to 104,090 bp (average
189 length: 60 Kb) (**Table S2**). Subsequently, we employed centromeric Histone H3
190 chromatin immunoprecipitation followed by sequencing technology (CENH3 ChIP-
191 seq) to validate the predicted potential centromere regions and found them to be
192 consistent with experimental verification (**Figure 2 and Figure S9**). We then
193 proceeded to use the dot plot method (**Figure S10**) for potential centromere prediction
194 in DBH species. For DBH, we validated putative centromeres for all 13 chromosomes
195 using VerityMap (**Figure S11**), with lengths ranging from 19,117 bp to 65,960 bp
196 (average length: 39 Kb) (**Table S3**). Insights from circular plots indicate that the two
197 species exhibit a strong degree of collinearity and the putative centromeres of NS4-2
198 and DBH are composed of TE elements (**Figure 1D**). This was especially apparent for
199 NS4-2, in which the centromere regions were significantly enriched with LTR/Copia
200 elements (**Figure 2 and Figure 1D**).

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

202 In higher plants, such as *Arabidopsis thaliana* and *Zea mays*, satellite repeats that are
203 associated with chromatin containing CENH3 exhibit lower levels of methylation
204 when compared to the non-CENH3 centromeric regions (Zhang et al. 2008; Wang et

al. 2022). These findings imply that DNA methylation is a significant factor in the epigenetic demarcation of centromeric chromatin. It is currently unclear whether centromeres in green algae, primarily composed of transposable elements, exhibit similar methylation patterns. We utilized ccsmeth (Ni et al. 2023), a deep learning-based circular consensus sequencing (CCS) sequencing methylation calling software, to analyze the centromere methylation in the NS4-2 genome (**Figure 3**). Interestingly, we found that the centromeric region of chromosome 1 in NS4-2 contains hypomethylated intervals (**Figure 3A**). Subsequently, we further zoomed in on the regions 50 Kb ups- and downstream of the centromere and observed a negative correlation between hypomethylated regions and ChIP signals (**Figure 3A and Figure S12**). We divided the genome into four categories: centromeric regions, non-centromeric regions, ChIP peak regions in centromeric regions, and non-ChIP peak regions in centromeric regions, and conducted a statistical analysis of methylation frequencies. Our analysis revealed that centromeric regions exhibited significantly higher methylation levels compared to non-centromeric regions, primarily attributable to the low methylation status in ChIP peak regions (**Figure 3B**). For DBH, we employed HiFi and ONT sequencing to call methylation, and similarly identified regions of low methylation within the centromeric regions (**Figure S13**). We found that the methylation resolution of HiFi data calling is higher than that of ONT in the TE-rich centromeres (**Figure S13**). We observed the same phenomenon of centromeric low methylation as in the NS4-2 species (**Figure 3C**), indicating that despite the absence of highly tandemly repeated satellite sequences in green algal centromeres, the low methylation in ChIP signal regions resembles that of higher plants. These results suggest that despite rapid evolution, centromeric sequences remain evolutionarily conserved in terms of epigenetic inheritance.

Centromere size and composition in green algae

We selected six high-quality assembled green algal genomes for the identification and compositional analysis of potential centromeres, including *Chlamydomonas reinhardtii* CC-4532 (v. 6.1) (Craig et al. 2023), *Chromochloris zofingiensis* (Rotha et al. 2017), *Coccomyxa subellipsoidea* C-169 (v. 3.0) (Blanc et al. 2012), *Chlorella* sp.

235 SLA-04 (Goemann et al. 2023), and the two *Chlorella* species assembled in this study
236 (**Figure 4A**). Additionally, we used the red alga *Cyanidioschyzon merolae* (short low-
237 GC centromeric regions) (Nozaki et al. 2007; Kanesaki et al. 2015) as an outgroup for
238 the phylogenetic tree of the lineage (**Figure 4A**). We identified the potential
239 centromere positions of the above-green species, including *C. reinhardtii*, *C.*
240 *zofingiensis*, *C. subellipsoidea* and SLA-04, using dot plot analysis and references
241 from the literature (**Table S4−S7**). We utilized RepeatMasker to annotate the
242 centromere composition of the selected species based on TE-algae library. The
243 potential centromeric regions of *C. reinhardtii*, *C. zofingiensis, C. subellipsoidea*, and
244 SLA-04 span approximately from 96 Kb to 320 Kb, 10Kb to 45 Kb, 12 Kb to 36 Kb,
245 and 14 Kb to 118 Kb in length, respectively (**Table S4−S7**), with average sizes of
246 about 200 Kb, 27 Kb, 23 Kb, and 52 Kb. The model green algae has the largest
247 genome, measuring approximately 112 Mb, and also possesses the longest average
248 centromere length. We found a positive correlation between centromere size and
249 genome size in green algae, with a Pearson correlation coefficient of 0.87 (**Figure 4B**).
250 This fills a gap in the literature by providing data on green algae, which was
251 previously lacking (Plačková et al. 2021). We provided a detailed annotation of the
252 centromere composition in the selected green algal species (**Table S8-S13**), sorted
253 based on the top five of element's proportion within centromeric regions (**Figure 4C−**
254 **H**). In the centromere sequences of model green algae (**Figure 4C and Table S14**)
255 and *C. subellipsoidea* (**Figure 4E and Table S15**), we observed a predominant
256 presence of ZeppL-1_cRei (LINE/L1) elements in *C. reinhardtii* and Zepp (LINE/L1)
257 elements in *C. subellipsoidea*, consistent with previous reports (Craig et al. 2021).
258 Nearly all of these LINE/L1 elements are distributed in the centromere region,
259 indicating they are centromere-specific elements (**Table S8 and Table S10**). It is
260 noteworthy that although the species *C. zofingiensis* demonstrates a sister relationship
261 with *C. reinhardtii* in phylogenetic analysis, the centromeric composition of *C.*
262 *zofingiensis* is primarily comprised of the LTR/Copia family (**Figure 4D and Table**
263 **S16**), in stark contrast to *C. reinhardtii* (**Figure 4C**). In species SLA-04, the
264 centromeric component predominantly consists of Copia-2_Chlorella_SLA04

265 (LTR/Copia), followed by L1-5_Chlorella_SLA04 (LINE/L1) and RTE-X-
266 1_Chlorella_SLA04 (LINE/RTE-X), accounting for 37%, 22%, and 11% respectively
267 (**Table S11 and Table S17**). In species NS4-2, the centromeric component is
268 primarily composed of Copia-1_Chlorella_NS (LTR/Copia) and Unknown-
269 1_Chlorella_NS (Unknown), accounting for 50% and 21% respectively (**Table S12**
270 **and Table S18**). In the species DBH, the predominant centromeric component is
271 Unknown-2_Chlorella_DBH (Unknown), followed by RTE-X-1_Chlorella_DBH
272 (LINE/RTE-X), RTE-X-2_Chlorella_DBH (LINE/RTE-X), and Copia-
273 1_Chlorella_DBH (LTR/Copia), accounting for 25%, 20%, 19%, and 17%
274 respectively (**Table S13 and Table S19**). The elements mentioned above in species of
275 the *Chlorella* genus are all centromere-specific components (**Figure 4F-H**).

276 **Centromere evolution in green algae**

277 We systematically correlate the primary constituents of centromeres across diverse
278 species onto the phylogenomic tree framework (**Figure 5A**). This strategic approach
279 enables us to scrutinize the evolutionary trajectories of centromeric transposon
280 sequences among distinct taxa. In our investigation, we have identified the presence
281 of LINE/L1 and LTR/Copia elements within species belonging to both Chlorophyceae
282 and Trebouxiophyceae, two major classes within the green algae lineage (**Figure 5A**).
283 This finding suggests that these elements are encompassed within the centromeric
284 sequences of the most recent common ancestor (MCRA) of green algae. An intriguing
285 observation is the exclusive presence of LINE/RTEX and Unknown family elements
286 within the centromeric composition of species belonging solely to the Chlorella genus.
287 This suggests a unique formation of the most recent common ancestor (MCRA)
288 specific to Chlorella, underscoring distinct evolutionary dynamics within this genus
289 (**Figure 5A**). Through an extensive evolutionary timeline, the Chlorophyceae-
290 Chlorella divergence is estimated to have occurred approximately 827 million years
291 ago (MYA), while the Coccomyxa-Chlorella divergence is estimated to have occurred
292 around 544 MYA (source: http://www.timetree.org/). During this evolutionary
293 journey, species such as *C. reinhardtii*, *C. zofingiensis*, and *C. subellipsoidea* have
294 independently evolved centromere-specific transposable elements (TEs) (**Figure 5A**).

295 However, within the *Chlorella* genus, multiple TE elements still persist within
296 centromeric sequences (**Figure 5A**), suggesting an ongoing process of optimization,
297 potentially towards establishing a predominant TE element as the primary constituent
298 of their centromeres. For instance, elements like NS4-2 and SLA-04 may be refining
299 LTR/Copia elements into a singular composite component, while DBH could be
300 optimizing LINE/RTEX into a singular centromeric component (**Figure 5A**). We are
301 conducting comparative genomic analysis of *Chlorella*, refining the analysis to
302 examine the centromeres between each chromosome, aiming to investigate whether
303 there is high collinearity or if they originate from the same ancestor, and to
304 understand the composition and structure of their centromeres. We conduct
305 collinearity analysis on the three species assembled at T2T-level, and then painting
306 the collinearity relationships and the putative centromeres (**Figure 5B**). The
307 chromosome 3 of SLA-04, the chromosome 1 of NS4-2, and the chromosome 1 of
308 DBH exhibit strong collinearity, indicating they originate from the same ancestor
309 (**Figure 5B**). We further extracted the centromere sequences of these three
310 chromosomes for dot plot analysis and found significant structural differences among
311 them (**Figure 5C**). The composition of TE (transposable element) families among the
312 three chromosomes also exhibits significant differences (**Figure 5C**). Taking
313 LTR/Copia as an example, in SLA-04, CEN3 comprises 44% of the selected TE
314 family within the centromere. In NS4-2, CEN1 comprises 88% of the selected TE
315 family within the centromere. Meanwhile, in DBH, CEN1 comprises 27% of the
316 selected TE family within the centromere. These results indicate that the evolution of
317 centromere sequences is significantly faster than that of non-centromeric regions.

318 ## Discussion

319 Recent advances in third-generation sequencing technologies have enabled the
320 telomere-to-telomere (T2T) assembly of a number of genomes including the model
321 plant *Arabidopsis thaliana* (Wang et al. 2022) and the model green alga
322 *Chlamydomonas reinhardtii* (Payne et al. 2023). To complement these, we employed
323 HiFi reads, ONT ultra-long reads, and stringent assembly methods to generate T2T

assemblies for *C. sorokiniana* strain NS4-2 and *C. pyrenoidosa* strain DBH. These T2T assemblies represent a significant improvement over previous genome assemblies, such as *C. sorokiniana* UTEX1230 (Hovde et al. 2018) and *C. pyrenoidosa* FACHB-9 (Fan et al. 2015), and in that respect they may serve as valuable resources for future green algae research, in particular towards understanding more complex regions of the genome, including the centromere.

In human centromeres, there are specific hypomethylated regions containing a higher concentration of the centromeric histone CENP-A. This enrichment of CENP-A suggests that these regions serves as functional kinetochore-binding sites (Logsdon et al. 2021; Altemose et al. 2022). It's worth noting that this particular characteristic is conserved in higher plants (Zhang et al. 2008; Wang et al. 2022). Interestingly, this study revealed that the same epigenetic phenomenon is also present in green algae, suggesting that this is a highly conserved epigenetic trait among eukaryotes in general. Centromeric satellite sequences have a monomer size of 100-200 base pairs, typically falling within the range of 150 to 180 bp, i.e. a length sufficient to wrap a single nucleosome (Melters et al. 2013). However, the origin of these monomers remains unclear. Analysis of ancestral crop species, such as *Gossypium raimondii* (Han et al. 2016), an ancestor species of tetraploid cotton, and *Triticum monococcum* (Ahmed et al. 2023), the first domesticated wheat species, found that their centromeres were comprised largely of retrotransposons rather than long arrays of centromeric satellites, supporting the hypothesis that centromere repeats originated from non-random retrotransposon insertions (Huang et al. 2021). We found that each of the green algae in this study had TE-rich centromeres, further supporting the hypothesis that centromeric monomers in higher plants originated from TEs in ancestral plants. Nevertheless, different taxa of green algae exhibited significant differences in their centromeric TE composition. For instance, *Chlamydomonas reinhardtii* and *Coccomyxa subellipsoidea* both consist of LINE/L1 elements, but they are composed of different subfamilies. Interestingly, the sister species of *C. reinhardtii*, *C. zofingiensis*, primarily consists of LTR/Copia sequences in its centromeric sequences (Roth et al, 2017), indicating that each species continually optimizes its specific

354 centromeric TE sequences during evolution. Despite being species of the same genus,
355 *Chlorella*, the composition and structure of centromeric sequences between
356 chromosomes derived from the common ancestor exhibit significant differences. It
357 can be said that the centromere is the fastest-evolving part of the chromosome
358 (Bensasson et al. 2008; Suzuki et al. 2020).

359 In summary, we assembled two *Chlorella* species (NS4-2 and DBH) to T2T level,
360 enabling a comprehensive analysis of their centromeric sequence composition and
361 evolution. These findings contribute to our understanding of centromere evolution in
362 green algae and, by extension, to their descendants.

363 Materials and methods

364 Culture conditions

365 *Chlorella pyrenoidosa* strain DBH (obtained from the Freshwater Algae Culture
366 Collection at the Institute of Hydrobiology) and *C. sorokiniana* strain NS4-2 (isolated
367 and purified from a water sample taken from Xueshi Lake at Shanghai Normal
368 University) were grown mixotrophically in a tris-acetate-phosphate (TAP) medium.
369 Both species were cultivated in Erlenmeyer flasks under continuous growth light
370 conditions (30-40 μmol photons m^{-2} s^{-1}) at a temperature of 25 °C, with agitation at
371 100 rpm, for maintenance purposes.

372 Karyotyping for the two algae species

373 The culture sample was harvested by centrifugation at 12,000 rpm for 2 minutes and
374 then washed sequentially with sterilized BG-11 medium and distilled water. The
375 resulting algal pellet was suspended in a 2 mM 8-hydroxyquinoline solution and kept
376 in darkness at 4 °C for 4 hours. Subsequently, the algae pellet was fixed with a
377 solution of 3 parts ethanol to 1 part acetic acid (v/v) for 24 hours to ensure fixation of
378 the algae cells in the metaphase of mitosis. After fixation, the cells were washed three
379 times with distilled water and then subjected to enzymatic digestion. The enzymatic
380 solution contained 2.5% (w/w) cellulose Onozuka R 10 (Yakult Pharmaceutical,
381 Tokyo, Japan) and 2.5% (w/w) pectinase Y23 (Yakult Pharmaceutical, Tokyo, Japan),
382 and the digestion process lasted for 18 hours at 37 °C. Following digestion, the cells

383 were homogenized at 15,000 rpm for 2 minutes and then centrifuged. The resulting
384 pellet was resuspended in 100 μL of acetic acid and vortexed for 20 seconds. The cell
385 suspension was dropped onto a sterile slide and allowed to air dry. Finally, the cells
386 were stained with an improved carbol fuchsin solution and examined under a
387 microscope to capture images of the metaphase stage.

388 **Genomic DNA extraction**

389 Samples were collected, and high molecular weight genomic DNA was prepared by
390 the CTAB method and followed by purification with QIAGEN® Genomic kit
391 (Cat#13343, QIAGEN) for regular sequencing, according to the standard operating
392 procedure provided by the manufacturer. The DNA degradation and contamination of
393 the extracted DNA was monitored on 1% agarose gels. DNA purity was then detected
394 using a NanoDrop One UV-VIS spectrophotometer (Thermo Fisher Scientific, USA),
395 of which OD 260/280 ranged from 1.8 to 2.0 and OD 260/230 from 2.0 to 2.2. DNA
396 concentration was subsequently measured using a Qubit 4.0 fluorometer (Invitrogen,
397 USA).

398 **HiFi sequencing and genome assembly**

399 SMRTbell libraries were prepared following the standard protocol of PacBio, using 15
400 Kb preparation solutions from Pacific Biosciences, USA. The library preparation
401 steps were conducted according to the method described in Wang et al. (2022).
402 Sequencing was performed on a PacBio Sequel II instrument with Sequencing Primer
403 V2 and Sequel II Binding Kit 2.0 at Xi'an Haorui Genomics Technology Co., Ltd.
404 (Xi'an, China). For the NS4-2 strain, a total of 25.87 Gb of HiFi reads were generated,
405 providing a coverage of 438×. The N50 of the reads was 15,694 bp. Regarding the
406 DBH strain, a total of 23.74 Gb of HiFi reads were generated, resulting in a coverage
407 of 308×. The N50 of the reads was 13,338 bp. The HiFi reads were assembled using
408 hifiasm v. 0.16.1-r375 (Cheng et al. 2021) with default parameters. gfatools
409 (https://github.com/lh3/gfatools) was employed to convert the sequence graphs in the
410 GFA format to FASTA format.

411 **ONT sequencing and genome assembly**

412 For the ultra-long Oxford Nanopore sequencing library, approximately 8–10 μg of
413 genomic DNA was selected (>50 kb) with the SageHLS HMW library system (Sage
414 Science, Beverly, MA), and then processed using the ligation sequencing 1D kit
415 (SQK-LSK109, Oxford Nanopore Technologies, Oxford, UK) according to the
416 manufacturer's instructions. DNA libraries (approximately 800 ng) were constructed
417 and sequenced using a PromethION (Oxford Nanopore Technologies, Oxford, UK) at
418 the Genome Center of Grandomics (Wuhan, China). A total of 22.12 Gb ONT long
419 reads with 368× coverage (read N50: 59,282 bp) were generated. To assemble these
420 into contigs, the long-read assembler NextDenovo v. 2.0 (Hu et al. 2023) was used
421 with parameters 'read_cutoff = 5k' and 'seed_cutoff = 132,382'. Contigs were then
422 polished using Nextpolish v. 1.3.1 (Hu et al. 2020) with parameters 'hifi_options = -
423 min_read_len 10k -max_read_len 45k' and 'hifi_minimap2_options = -x asm20'.

424 **Hi-C sequencing and scaffolding**

425 A Hi-C library was generated from cross-linked chromatins of algal cells using a
426 standard Hi-C protocol, and then sequenced using the NovaSeq 6000 platform. The
427 Hi-C sequencing data, as outlined in Table S1, were utilized to construct a Hi-C
428 contact map. The construction of the contact map was performed using Juicer v. 1.5
429 (Durand et al. 2016a) with parameter '-s DpnII'. The resulting contact map was
430 manually inspected and verified using Juicebox v. 1.11.08 (Durand et al. 2016b).

431 **Evaluation of the putative centromeres**

432 To assess the assembly structure of the predicted centromeric regions, we employed
433 VerityMap (Mikheenko et al. 2020). First, the PacBio HiFi reads were aligned to the
434 assembly using minimap2 v. 2.20-r1061 (Li, 2018) with '-x map-hifi' parameter. The
435 reads aligning to the centromeric regions were extracted using SAMtools v. 1.9 Li et
436 al. 2009). These extracted HiFi reads were then input into VerityMap using the
437 following parameters: '--careful --reads HiFi reads -d hifi '.

438 **Genome comparisons**

439 The synteny relationships among the chromosomes were estimated using MCScanX
440 (Wang et al. 2012) with parameter '-s 10' and visualized using Circos v. 0.69-8

441 (Krzywinski et al. 2009). Genomic alignment dot plot of chromosomes was generated
442 using Gepard v. 2.1 (Krumsiek et al. 2007). To assess genome completeness, we used
443 BUSCO v. 3.0.2 with the dataset 'chlorophyta_odb10 v. 4' (Simão et al. 2015).
444 Chromosome painting among three species of the genus Chlorella using the IAGS
445 software (Gao et al. 2022).

446 **Genome annotation**

447 For the NS4-2 and DBH genome annotations, we initially predicted the *de novo* gene
448 structure using GeneMark-ES with the parameters '--ES' (Borodovsky and Lomsadze,
449 2011). Subsequently, we employed the 'chlorella gene model' to train Augustus v. 3.3
450 (Mario and Burkhard, 2005). To identify protein-coding genes, we utilized the
451 MAKER2 pipeline (Cantarel et al. 2008), which integrated the gene models predicted
452 from GeneMark-ES and Augustus, as well as the protein sequences from closely
453 related organisms such as *Chlamydomonas reinhardtii* (GenBank accession:
454 GCF_000002595.2) (Merchant et al. 2007), *Chlorella variabilis* (GCF_000147415.1)
455 (Blanc et al. 2010), *Chlorella sorokiniana* (GCA_002245835.2) (Arriola et al. 2018),
456 and *Micractinium conductrix* (GCA_002245815.2) (Arriola et al. 2018), which are
457 known to be close relatives of NS4-2 and DBH. For the identification of repeat
458 sequences, we employed RepeatMasker v. 4.0.7 (http://www.repeatmasker.org) and
459 RepeatModeler v. 2.0.1 (Flynn et al. 2020). The TE-algae library used to annotate the
460 selected species consists of three parts: the *de novo* library of three Chlorella species
461 from RepeatModeler, a library from a previous study (Craig et al. 2021), and the
462 *Chlorophyta* Repbase library (https://www.repeatmasker.org/) (**Data S1**). The TE-
463 algae library underwent redundancy reduction using cd-hit (Li and Godzik 2006). The
464 karyoplots shown in this study were generated using KaryoploteR package in R (Gel
465 and Serra, 2017).

466 **Phylogenetic tree construction**

467 The phylogenetic trees depicted in Figure 1 and Figure 4 were generated utilizing
468 OrthoFinder v. 2.2.7 with the default parameters (Emms and Kelly, 2019). Figure
469 S1's phylogenetic tree was constructed using MEGA v. 7 softwareand employed the
470 Kimura 2-parameter model, accompanied by 1000 bootstrap replicates (Kumar et al.

471 2016). For Figure S13, the phylogenetic tree was crafted using RAxML-NG v. 1.1
472 (Kozlov et al. 2019) with the specified parameters '--all --bs-trees 1000', and the
473 HKY+G4 model was chosen through ModelTest-NG v. 0.1.7 (Darriba et al. 2020).

474 **PacBio DNA methylation analysis**

475 Methylation for NS4-2 and DBH was inferred with ccsmeth v.0.4.1 (Ni et al. 2023),
476 using kinetics features from PacBio CCS reads. Methylation of CCS reads was
477 predicted using 'ccsmeth call_mods' with the model
478 'model_ccsmeth_5mCpG_call_mods_attbigru2s_b21.v2.ckpt'. Then, we used
479 'ccsmeth align_hifi' to align the reads to their respective genome. The methylation
480 frequencies were calculated by using 'ccsmeth call_freqb' with the aggregate model
481 'model_ccsmeth_5mCpG_aggregate_attbigru_b11.v2p.ckpt'.

482 **ONT DNA methylation analysis**

483 Nanopolish v. 0.13.2 with the parameters 'call-methylation --methylation cpg' was
484 used to measure the frequency of CpG methylation. The ONT reads were aligned to
485 whole-genome assemblies via minimap2 (Li, 2018). The script
486 'calculate_methylation_frequency.py', provided in the methplotlib package (De
487 Coster et al. 2020), was then used to generate the methylation frequency. Methylation
488 density is defined as the number of occurrences with a methylation frequency greater
489 than 0.8 within genomic sliding windows (100-base pair window advanced at 50-base
490 pair increments).

491 **CENH3 ChIP-seq**

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

499 Shanghai Jiayin Biotechnology Co., Ltd conducted ChIP assays following a
500 modified version of the standard crosslinking ChIP protocol. In summary, the culture

501 sample was harvested by centrifugation at 12,000 rpm for 2 minutes and sequentially
502 washed with sterilized BG-11 and distilled water. Subsequently, approximately 4g of
503 algae pellet was obtained and ground in liquid nitrogen. The ground sample was fixed
504 with 1% formaldehyde at room temperature for 15 minutes, followed by a 5-minute
505 treatment with 0.125 M glycine. After fixation, the sample was washed, resuspended
506 in a lysis buffer, and sonicated to generate DNA fragments. Following sonication,
507 immunoprecipitation was carried out using the custom antibody NS4-2_CENH3. The
508 immunoprecipitated complex was washed, and DNA was extracted. DNA purification
509 was performed using the Universal DNA Purification Kit (#DP214). A ChIP-seq
510 library was prepared by employing the ChIP-Seq DNA sample preparation kit
511 (NEBNext® UltraTMII DNA), adhering to the manufacturer's instructions. The
512 extracted DNA was ligated to specific adaptors and subsequently subjected to deep
513 sequencing on the Illumina Novaseq 6000 platform in 150bp paired-end mode.

514 The ChIP/input paired-end reads were mapped to the NS4-2 genome using the
515 'mem' algorithm of BWA (Li and Durbin 2009). We determined the ChIP/input
516 coverage ratio by employing the deeptools function 'bamCompare', with a threshold
517 of MAPQ ≥ 30. The output format was in BigWig. We carried out peak calling using
518 MACS2 (Zhang et al. 2008) with the parameters '-t ChIP.bam -c input.bam -f BAM --
519 outdir chipseq -n chipseq -B --nomodel --keep-dup 1 -g 58000000'.

520 **Data availability**

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

accessible at https://ngdc.cncb.ac.cn/gsa.

Author contributions

KY and WM conceived and designed the project. KY and BW designed the experiments. BW, YJ, SG, and HG analyzed the results. BW, YJ, ND, JY, SB, SG, WH, SW, and HG performed all the experiments and analyzed the data. BW, YJ and ND created the figures. BW, YJ and SB wrote the manuscript. SB, XY, WM, and KY reviewed the manuscript.

Table 1. Comparison genomic features of *Chlorella sorokiniana* NS4-2 and *C. pyrenoidosa* DBH assemblies.

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 (count)	12	NA	13	NA
No. of contigs (count)	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
Protein coding genes (count)	12,691	12,871	11,327	10,283
Repeat content (%)	8.84	8.89	8.91	5.46

[a] UTEX1230 is a strain of *C. sorokiniana* (Hovde et al. 2018).
[b] FACHB-9 is a strain of *C. pyrenoidosa* (Fan et al. 2015).

Figure legends

Figure 1. T2T genome assemblies and circular plot of *Chlorella sorokiniana* NS4-2 and *C. pyrenoidosa* DBH. (A) Micrograph of *C. sorokiniana* NS4-2, bar = 20 μm; **(B)** The karyotype diagram indicates that NS4-2 has 12 chromosomes; **(C)** Hi-C contact map of the NS4-2 T2T genome assembly (chromosome 1-12); **(D)** Circular plot shows the T2T genome features of NS4-2 and DBH. Track a shows the karyotype distribution of the assembled chromosomes; track b, GC density; track c, gene density; track d, LINE element density; track e, LTR element density; track f, putative centromeres. Densities was calculated in 50-Kb windows. The innermost of the circular plots represents the synteny relationship between NS4-2 and DBH.

557 **Figure 2. CENH3 binding defines centromeres in *Chlorella sorokiniana* NS4-2**.
558 Schematic overview of the chromosomes of NS4-2 showing the normalized CENH3
559 ChIP-seq read coverage (RPGC normalization in 1-Kb bins with 3-Kb smoothing,
560 plot generated using kpPlotBigWig function in karyoploteR), predicted centromere
561 regions, gene density, and LTR/Copia density. Densities was calculated in 5-Kb
562 windows.
563
564 **Figure 3. Centromeric methylation features of *Chlorella sorokiniana* NS4-2. (A)**
565 Methylation density distribution chart of chrosomome 1 in NS4-2. The red area
566 represents the centromere region. Zoomed-in overview of the chromosome 1 of NS4-
567 2 showing the methylation density distribution, CENH3 ChIP-seq signals (plot
568 generated using kpPlotBAMCoverage function in karyoploteR), as well as the regions
569 of centromere (green band) and ChIP-seq peaks called by MACS2 (light green); **(B)**
570 The box- and violin plots display a comparison of methylation density in NS4-2
571 among centromere (CEN), non-centromere (non-CEN), peaks within the centromere
572 (CEN_peaks), and non-peaks within the centromere (CEN_nonPeaks); **(C)** The box-
573 and violin plots display a comparison of methylation density in DBH among CEN and
574 nonCEN.
575
576 **Figure 4. Centromeric size and compositions in green algae. (A)** Phylogenomic
577 tree of six green algae species with the red algae *Cyanidioschyzon merolae* utilized as
578 an outgroup species. The figure on the right illustrates a comparison of genome size
579 and centromere length among these selected species; **(B)** Relationship between the
580 genome size of each chromosome (log-transformed) and each centromere size
581 (logtransformed). The figure was generated using basic plot functions in R; **(C-H)**
582 The distribution of different repeat subfamilies in centromeres (CEN; dark blue) and
583 across the genome (non-CEN; light gray) of *Chlamydomonas reinhardtii* (CR) **(C)**,
584 *Chromochloris zofingiensis* (CZ) **(D)**, *Coccomyxa subellipsoidea* (CS) **(E)**, *Chlorella*
585 *sorokiniana* (SLA-04) **(F)**, *Chlorella pyrenoidosa* (DBH) **(G)** and *Chlorella*
586 *sorokiniana* (NS4-2) **(H)**.

587

588 **Figure 5. Centromeric TEs evolutionary in green algae and *Chlorella* species. (A)**
589 The representation of the main centromeric TE families of the selected species on the
590 phylogenetic tree. MCRA stands for most recent common ancestor; **(B)** Synteny
591 analyses of the 13 chromosomes of SLA-04, 12 chromosomes of NS4-2 and 13
592 chromosomes of DBH. The corresponding syntenic regions are marked using the
593 same color. Approximate locations of putative centromeres are indicated by ellipses.
594 The red circles represent the collinearity of the whole chromosomes; **(C)** The dotplot
595 display of the centromeres showing the collinearity of entire chromosomes from three
596 *Chlorella* species, along with the composition of TEs. Referring to the legend of (A),
597 different colors represent different TE families.

598

599 **Supporting information**

600 **Figure S1.** Maximum likelihood phylogenetic tree based on the analysis of 18S gene
601 sequences. GenBank accession numbers in brackets.

602 **Figure S2.** T2T genome assemblies and centromere identification of *Chlorella*
603 *pyrenoidosa* strain DBH. **(A)** Micrograph of DBH,bar = 20 μm; **(B)** The karyotype
604 diagram indicates that DBH has 13 chromosomes; **(C)** Hi-C contact map of the DBH
605 T2T genome assembly (chromosome 1-13); **(D)** Schematic overview of the
606 chromosomes of DBH showing the predicted centromere regions (green band).

607 **Figure S3.** High-quality genome assembly pipelines. **(A)** *Chlorella sorokiniana* NS4-
608 2 genome assembly; **(B)** *C. pyrenoidosa* DBH genome assembly.

609 **Figure S4.** Demonstration of collinearity between chromosomes. **(A)** Between two
610 contigs of *Chlorella sorokiniana* NS4-2 and chromosome 2 of *C. vulgaris*; **(B)**
611 Between two contigs of *C. pyrenoidosa* DBH and chromosome 2 of *C. vulgaris.*

612 **Figure S5.** Self-Dot plot showing potential centromeric features and regions of two
613 previous assemblies, *Chlamydomonas reinhardtii* and *Coccomyxa subellipsoidae*.

614 **Figure S6.** Self-dot plot of the 12 chromosomes of *Chlorella sorokiniana* NS4-2, with
615 green boxes indicating potential centromeric regions.

616 **Figure S7.** Self-dot plot of the 13 chromosomes of *Chlorella pyrenoidosa* DBH, with

617 green boxes indicating potential centromeric regions.

618 **Figure S8.** Evaluation of the 12 putative centromere assemblies of NS4-2 using
619 VerityMap. The xaxis is centromere position, the y-axis is percent of deviated reads.
620 Only a low percentage (12.5%) of deviated reads were detected in CEN11, indicating
621 no potential large structural errors in the assembly.

622 **FigureS9.** ChIP/Input ratio of the 12 chromosomes of *Chlorella sorokiniana* NS4-2.
623 Schematic overview showing the ratio (plot generated using the kpPlotBigWig
624 function in karyoploteR) and the putative centromere regions (red band).

625 **FigureS10.** Using chromosome 10 of *Chlorella sorokiniana* NS4-2 as an example to
626 demonstrate the dot plot method for identifying potential centromere positions in the
627 genome.

628 **Figure S11.** Evaluation of the 13 putative centromere assemblies of *Chlorella*
629 *pyrenoidosa* DBH using VerityMap. The xaxis is centromere position, the y-axis is
630 percent of deviated reads.

631 **Figure S12.** Centromeric methylation features of *Chlorella sorokiniana* NS4-2.
632 Schematic overview depicting each centromere along with its 50 Kb upstream and
633 downstream regions of NS4-2. The illustration shows the distribution of methylation
634 density (determined by HiFi data), CENH3 ChIP-seq signals (plot generated using
635 kpPlotBAMCoverage function in karyoploteR), and the regions of centromeres (green
636 bands).

637 **Figure S13.** Centromeric methylation features of *Chlorella pyrenoidosa* DBH.
638 Schematic overview depicting each centromere along with its 50 Kb upstream and
639 downstream regions of DBH. The illustration shows the distribution of methylation
640 density determined by HiFi data, methylation density determined by ONT data, and
641 the regions of centromere (green band).

642 **Table S1**. Details of the sequencing data for *Chlorella sorokiniana* NS4-2 and *C.*
643 *pyrenoidosa* DBH.

644 **Table S2.** *Chlorella sorokiniana* NS4-2 putative centromeric coordinates.

645 **Table S3.** *Chlorella pyrenoidosa* DBH putative centromeric coordinates.

646 **Table S4**. *Chlamydomonas reinhardtii* CC4532 putative centromeric coordinates.

Table S5. *Chromochloris zofingiensis* putative centromeric coordinates.

Table S6. *Coccomyxa subellipsoidea* C169 putative centromeric coordinates.

Table S7. *Chlorella sp.* SLA-04 putative centromeric repeat annotations.

Table S8. Quantitative analysis of the number, length, and content of various transposable element families at the centromere of *Chlamydomonas reinhardtii*.

Table S9. Quantitative analysis of the number, length, and content of various transposable element families at the centromere of *Chromochloris zofingiensis*.

Table S10. Quantitative analysis of the number, length, and content of various transposable element families at the centromere of *Coccomyxa subellipsoidea* C-169.

Table S11. Quantitative analysis of the number, length, and content of various transposable element families at the centromere of *Chlorella* sp. SLA-04.

Table S12. Quantitative analysis of the number, length, and content of various transposable element families at the centromere of *Chlorella sorokiniana* NS4-2.

Table S13. Quantitative analysis of the number, length, and content of various transposable element families at the centromere of *Chlorella sorokiniana* DBH.

Table S14. *Chlamydomonas reinhardtii* putative centromeric repeat annotations.

Table S15. *Coccomyxa subellipsoidea* putative centromeric repeat annotations.

Table S16. *Chromochloris zofingiensis* putative centromeric repeat annotations.

Table S17. *Chlorella sp.* SLA-04 putative centromeric repeat annotations.

Table S18. *Chlorella sorokiniana* NS4-2 putative centromeric repeat annotations.

Table S19. *Chlorella pyrenoidosa* DBH putative centromeric repeat annotations.

Data S1. TE library used in this study.

References

Ahmed, H.I., Heuberger, M., Schoen, A., Koo, D.H., Quiroz-Chavez, J., Adhikari, L., Raupp, J., Cauet, S., Rodde, N., Cravero, C., Callot, C., Lazo, G.R., Kathiresan, N., Sharma, P.K., Moot, I., Yadav, I.S., Singh, L., Saripalli, G., Rawat, N., Datla, R., Athiyannan, N., Ramirez-Gonzalez, R.H., Uauy, C., Wicker, T., Tiwari, V.K., Abrouk, M., Poland, J., Krattinger, S. G. (2023). Einkorn genomics sheds light on history of the oldest domesticated wheat. Nature **620**, 830-838.

Altemose, N., Logsdon, G. A., Bzikadze, A. V., Sidhwani, P., Langley, S. A.,

Caldas, G. V., Hoyt, S. J., Uralsky, L., Ryabov, F. D., Shew, C. J., Sauria, M. E. G., Borchers, M., Gershman, A., Mikheenko, A., Shepelev, V. A., Dvorkina, T., Kunyavskaya, O., Vollger, M. R., Rhie, A., McCartney, A. M., Asri, M., Lorig-Roach, R., Shafin, K., Lucas, J.K., Aganezov, S., Olson, D., de Lima, L.G., Potapova, T., Hartley, G.A., Haukness, M., Kerpedjiev, P., Gusev, F., Tigyi, K., Brooks, S., Young, A., Nurk, S., Koren, S., Salama, S.R., Paten, B., Rogaev, E.I., Streets, A., Karpen, G.H, Dernburg, A.F., Sullivan, B.A., Straight, A.F., Wheeler, T.J., Gerton, J.L., Eichler, E.E., Phillippy, A.M., Timp, W., Dennis, M.Y., O'Neill, R.J., Zook, J.M., Schatz, M.C., Pevzner, P.A., Diekhans, M., Langley, C.H., Alexandrov, I.A., and Miga, K.H. (2022). Complete genomic and epigenetic maps of human centromeres. Science **376**, eabl4178.

Arriola, M.B., Velmurugan, N., Zhang, Y., Plunkett, M.H., Hondzo, H., and Barney, B.M. (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.

Bensasson, D., Zarowiecki, M., Burt, A., and Koufopanou, V. (2008). Rapid evolution of yeast centromeres in the absence of drive. Genetics **178**, 2161-2167.

Blanc, G., Agarkova, I., Grimwood, J., Kuo, A., Brueggeman, A., Dunigan, D.D., Gurnon, J., Ladunga, I., Lindquist, E., Lucas, S., Pangilinan, J., Pröschold, T., Salamov, A., Schmutz, J., Weeks, D., Yamada, T., Lomsadze, A., Borodovsky, M., Claverie, J.M., Grigoriev, I.V., and Van Etten, J.L. (2012). The genome of the polar eukaryotic microalga *Coccomyxa subellipsoidea* reveals traits of cold adaptation. Genome Biol **13**, R39.

Blanc, G., Duncan, G., Agarkova, I., Borodovsky, M., Gurnon, J., Kuo, A., Lindquist, E., Lucas, S., Pangilinan, J., Polle, J., Salamov, A., Terry, A., Yamada, T., Dunigan, D.D., Grigoriev, I.V., Claverie, J.M., and Van Etten, J.L. (2010). The *Chlorella variabilis* NC64A genome reveals adaptation to photosymbiosis, coevolution with viruses, and cryptic sex. Plant Cell **22**, 2943-2955.

Bi, G., Zhao, S., Yao, J., Wang, H., Zhao, M., Sun, Y., Hou, X., Haas, F. B., Varshney, D., Prigge, M., Rensing, S. A., Jiao, Y., Ma, Y., Yan, J., and Dai, J. (2024). Near telomere-to-telomere genome of the model plant Physcomitrium patens. Nat Plants 10,327-343.

Borodovsky, M., and Lomsadze, A. (2011). Eukaryotic gene prediction using GeneMark.hmm-E and GeneMark-ES. Curr Protoc Bioinformatics **4**, 4.6.1-4.6.10.

Cantarel, B.L., Korf, I., Robb, S.M., Parra, G., Ross, E., Moore, B., Holt, C., Sánchez Alvarado, A., and Yandell, M. (2008). MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res **18**, 188-196.

Cecchin, M., Marcolungo, L., Rossato, M., Girolomoni, L., Cosentino, E., Cuine, S., Li-Beisson, Y., Delledonne, M., and Ballottari, M. (2019). *Chlorella vulgaris* genome assembly and annotation reveals the molecular basis for

723 metabolic acclimation to high light conditions. Plant J **100**, 1289-1305.
724 **Chen, J., Wang, Z., Tan, K., Huang, W., Shi, J., Li, T., Hu, J., Wang, K., Wang, C.,**
725 **Xin, B., Zhao, H., Song, W., Hufford, M.B., Schnable, J.C., Jin, W., and Lai,**
726 **J.** (2023). A complete telomere-to-telomere assembly of the maize genome. **Nat**
727 **Genet** 55, 1221-1231.
728 **Cheng, H., Concepcion, G.T., Feng, X., Zhang, H., and Li, H.** (2021). Haplotype-
729 resolved *de novo* assembly using phased assembly graphs with hifiasm. Nat
730 Methods **18**, 170-175.
731 **Craig, R.J., Hasan, A.R., Ness, R.W., and Keightley, P.D.** (2021). Comparative
732 genomics of *Chlamydomonas*. Plant Cell **33**, 1016-1041.
733 **Craig, R.J., Gallaher, S.D., Shu, S., Salomé, P.A., Jenkins, J.W., Blaby-Haas, C.E.,**
734 **Purvine, S.O., O'Donnell, S., Barry, K., Grimwood, J., Strenkert, D., Kropat,**
735 **J., Daum, C., Yoshinaga, Y., Goodstein, D.M., Vallon, O., Schmutz, J., and**
736 **Merchant, S.S.** (2023). The *Chlamydomonas* Genome Project, version 6:
737 Reference assemblies for mating-type plus and minus strains reveal extensive
738 structural mutation in the laboratory. Plant Cell **35**, 644-672.
739 **Darriba, D., Posada, D., Kozlov, A. M., Stamatakis, A., Morel, B., and Flouri, T.**
740 (2020). ModelTest-NG: A new and scalable tool for the selection of DNA and
741 protein evolutionary models. Mol Biol Evol **37**, 291-294.
742 **Dias, G.B., Svartman, M., Delprat, A., Ruiz, A., and Kuhn, G.C.** (2014). Tetris is a
743 foldback transposon that provided the building blocks for an emerging satellite
744 DNA of *Drosophila virilis*. Genome Biol Evol **6**, 1302-1313.
745 **Durand, N.C., Robinson, J.T., Shamim, M.S., Machol, I., Mesirov, J.P., Lander,**
746 **E.S., and Aiden, E.L.** (2016a). Juicebox provides a visualization system for Hi-
747 C contact maps with unlimited zoom. Cell Syst **3**, 99-101.
748 **Durand, N.C., Shamim, M.S., Machol, I., Rao, S.S., Huntley, M.H., Lander, E.S.,**
749 **and Aiden, E.L.** (2016b). Juicer provides a one-click system for analyzing loop-
750 resolution Hi-C experiments. Cell Syst **3**, 95-98.
751 **Emms, D. M., and Kelly, S.** (2019). OrthoFinder: phylogenetic orthology inference
752 for comparative genomics. Genome Biol **20**, 238.
753 **Fan, J., Ning, K., Zeng, X., Luo, Y., Wang, D., Hu, J., Li, J., Xu, H., Huang, J.,**
754 **Wan, M., Wang, W., Zhang, D., Shen, G., Run, C., Liao, J., Fang, L., Huang,**
755 **S., Jing, X., Su, X., Wang, A., Bai, L., Hu, Z. X, J., and Li, Y.** (2015).
756 Genomic foundation of starch-to-lipid switch in oleaginous *Chlorella* spp. Plant
757 Physiol **169**, 2444-2461.
758 **Flynn, J.M., Hubley, R., Goubert, C., Rosen, J., Clark, A.G., Feschotte, C., and**
759 **Smit, A.F.** (2020). RepeatModeler2 for automated genomic discovery of
760 transposable element families. Proc Natl Acad Sci U S A **117**, 9451-9457.
761 **Gao, S., Yang, X., Guo, H., Zhao, X., Wang, B., and Ye, K.** (2023). HiCAT: A tool
762 for automatic annotation of centromere structure. Genome Biol **24**, 58.
763 **Gao, S., Yang, X., Sun, J., Zhao, X., Wang, B., and Ye, K.** (2022). IAGS: Inferring
764 ancestor genome structure under a wide range of evolutionary scenarios. Mol
765 Biol Evol **39**, msac041.
766 **Gel, B., and Serra, E.** (2017). karyoploteR: an R/Bioconductor package to plot

767 customizable genomes displaying arbitrary data. Bioinformatics **33**, 3088-3090.
768 **Goemann, C. L., Wilkinson, R., Henriques, W., Bui, H., Goemann, H. M.,**
769 **Carlson, R. P., Viamajala, S., Gerlach, Robin., and Wiedenheft, B.** (2023).
770 Genome sequence, phylogenetic analysis, and structure-based annotation reveal
771 metabolic potential of *Chlorella* sp. SLA-04. Algal Res **69**, 102943.
772 **Han, J., Masonbrink, R.E., Shan, W., Song, F., Zhang, J., Yu, W., Wang, K., Wu,**
773 **Y., Tang, H., Wendel, J.F., and Wang, K.** (2016). Rapid proliferation and
774 nucleolar organizer targeting centromeric retrotransposons in cotton. Plant J **88**,
775 992-1005.
776 **Henikoff, S., Ahmad, K., and Malik, H.S.** (2001). The centromere paradox: stable
777 inheritance with rapidly evolving DNA. Science **293**, 1098-1102.
778 **Hovde, B.T., Hanschen, E.R., Tyler, C.R. S., Lo, C.C., Kunde, Y., Davenport, K.,**
779 **Daligault, H., Msanne, J., Canny, S., Eyun, S.I., Riethoven, J.M., Polle, J.,**
780 **Starkenburg, S.R.** (2018). Genomic characterization reveals significant
781 divergence within *Chlorella sorokiniana* (Chlorellales, Trebouxiophyceae). Algal
782 Res **35**, 449-461.
783 **Hu, J., Fan, J., Sun, Z., and Liu, S.** (2020). NextPolish: a fast and efficient genome
784 polishing tool for long-read assembly. Bioinformatics **36**, 2253-2255.
785 **Hu, J., Wang, Z., Sun, Z., Hu, B., Ayoola, A.O., Liang, F., Li, J., Sandoval, J.R.,**
786 **Cooper, D.N., Ye, K., Ruan, J., Xiao, C., Wang, D., Wu, D., and Wang, S.**
787 (2023). An efficient error correction and accurate assembly tool for noisy long
788 reads. bioRxiv, 2023.03.09.531669.
789 **Huang, Y., Ding, W., Zhang, M., Han, J., Jing, Y., Yao, W., Hasterok, R., Wang,**
790 **Z., and Wang, K.** (2021). The formation and evolution of centromeric satellite
791 repeats in *Saccharum* species. Plant J **106**, 616-629.
792 **Kanesaki, Y., Imamura, S., Matsuzaki, M., and Tanaka, K.** (2015). Identification
793 of centromere regions in chromosomes of a unicellular red alga,
794 *Cyanidioschyzon merolae*. FEBS Lett **589**, 1219-1224.
795 **Kozlov, A.M., Darriba, D., Flouri, T., Morel, B., and Stamatakis, A.** (2019).
796 RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood
797 phylogenetic inference. Bioinformatics **35**, 4453-4455.
798 **Krumsiek, J., Arnold, R., and Rattei, T.** (2007). Gepard: a rapid and sensitive tool
799 for creating dot plots on a genome scale. Bioinformatics **23**, 1026-1028.
800 **Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D.,**
801 **Jones, S.J., and Marra, M.A**. (2009). Circos: an information aesthetic for
802 comparative genomics. Genome Res **19**, 1639-1645.
803 **Kumar, S., Stecher, G., and Tamura, K**. (2016). MEGA7: Molecular Evolutionary
804 Genetics Analysis Version 7.0 for Bigger Datasets. Mol Biol Evol **33**, 1870-1874.
805 **Li, H.** (2018). Minimap2: pairwise alignment for nucleotide sequences.
806 Bioinformatics, **34**, 3094-3100.
807 **Li, H., Durbin, R.** (2009). Fast and accurate short read alignment with Burrows-
808 Wheeler transform. Bioinformatics **25**, 1754-1760.
809 **Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G.,**
810 **Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing**

Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics **25**, 2078-2079.

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics **22**, 1658-1659.

Logsdon, G.A., Vollger, M.R., Hsieh, P., Mao, Y., Liskovykh, M.A., Koren, S., Nurk, S., Mercuri, L., Dishuck, P.C., Rhie, A., de Lima, L.G., Dvorkina, T., Porubsky, D., Harvey, W.T., Mikheenko, A., Bzikadze, A.V., Kremitzki, M., Graves-Lindsay, T.A., Jain, C., Hoekzema, K., Murali, S.C., Munson, K.M., Baker, C., Sorensen, M., Lewis, A.M., Surti, U., Gerton, J.L., Larionov, V., Ventura, M., Miga, K.H., Phillippy, A.M., Eichler, E.E. (2021). The structure, function and evolution of a complete human chromosome 8. **Nature** 593, 101-107.

Stanke, M., and Morgenstern, B. (2005). AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Res **33**, W465-W467.

Melters, D.P., Bradnam, K.R., Young, H.A., Telis, N., May, M.R., Ruby, J.G., Sebra, R., Peluso, P., Eid, J., Rank, D., Garcia, J.F., DeRisi, J.L., Smith, T., Tobias, C., Ross-Ibarra, J., Korf, I., and Chan, S.W. (2013). Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. Genome Bio **14**, 1-20.

Merchant, S.S., Prochnik, S.E., Vallon, O., Harris, E.H., Karpowicz, S.J., Witman, G. B., Terry, A., Salamov, A., Fritz-Laylin, L.K., Maréchal-Drouard, L., Marshall, W.F., Qu, L.H., Nelson, D.R., Sanderfoot, A.A., Spalding, M.H., Kapitonov, V.V., Ren, Q., Ferris, P., Lindquist, E., Shapiro, H., ··· Grossman, A.R. (2007). The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. Science **318**, 245-250.

McAinsh, A.D., and Marston, A.L. (2022). The four causes: the functional architecture of centromeres and kinetochores. Annu Rev Genet **56**, 279-314.

Michael, T.P. (2014). Plant genome size variation: bloating and purging DNA. Brief Funct Genomics **13**, 308-317.

Miga, K.H. (2020). Centromere studies in the era of 'telomere-to-telomere' genomics. Exp Cell Res **394**, 112127.

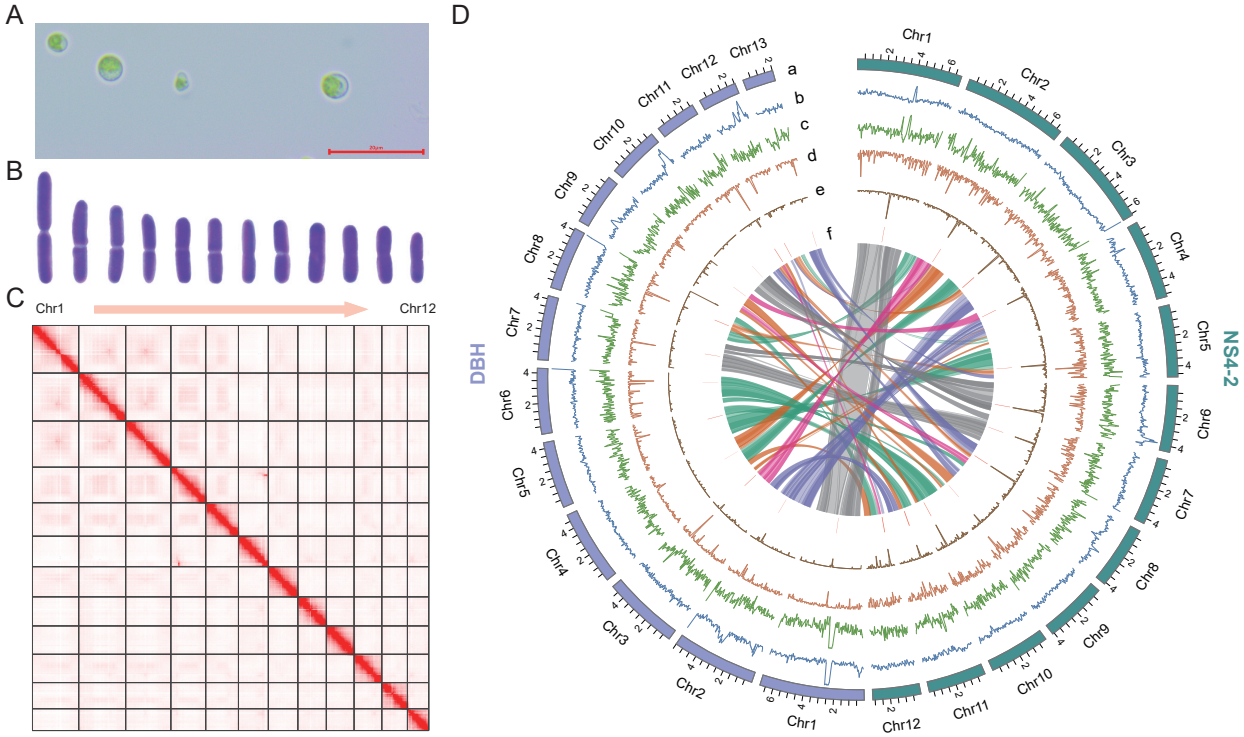
Mikheenko, A., Bzikadze, A.V., Gurevich, A., Miga, K.H., and Pevzner, P.A. (2020). TandemTools: mapping long reads and assessing/improving assembly quality in extra-long tandem repeats. Bioinformatics **36**, i75-83.

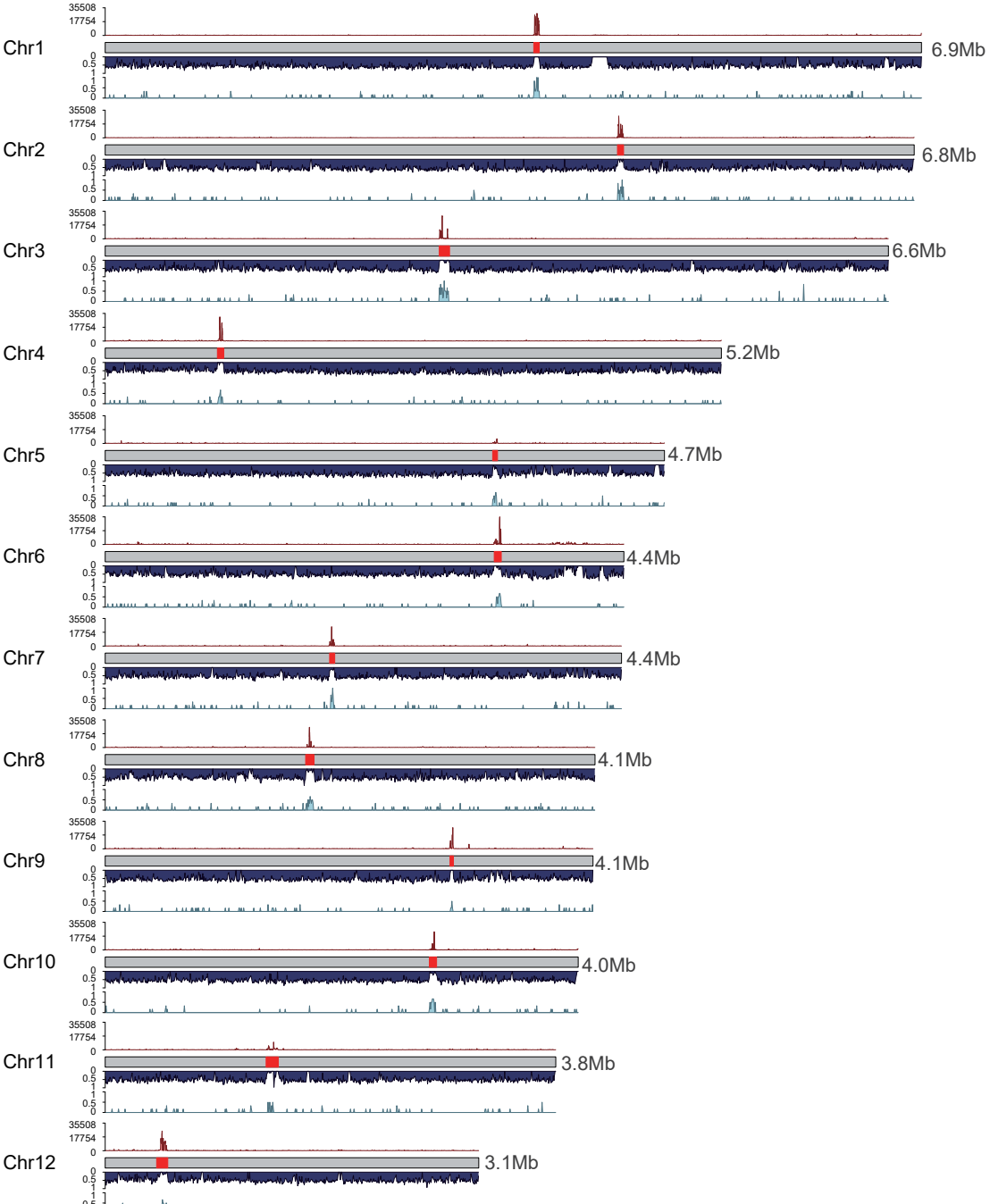
Naish, M., Alonge, M., Wlodzimierz, P., Tock, A.J., Abramson, B.W., Schmücker, A., Mandáková, T., Jamge, B., Lambing, C., Kuo, P., Yelina, N., Hartwick, N., Colt, K., Smith, L. M., Ton, J., Kakutani, T., Martienssen, R.A., Schneeberger, K., Lysak, M.A., Berger, F., ··· Henderson, I.R. (2021). The genetic and epigenetic landscape of the Arabidopsis centromeres. Science **374**, eabi7489.

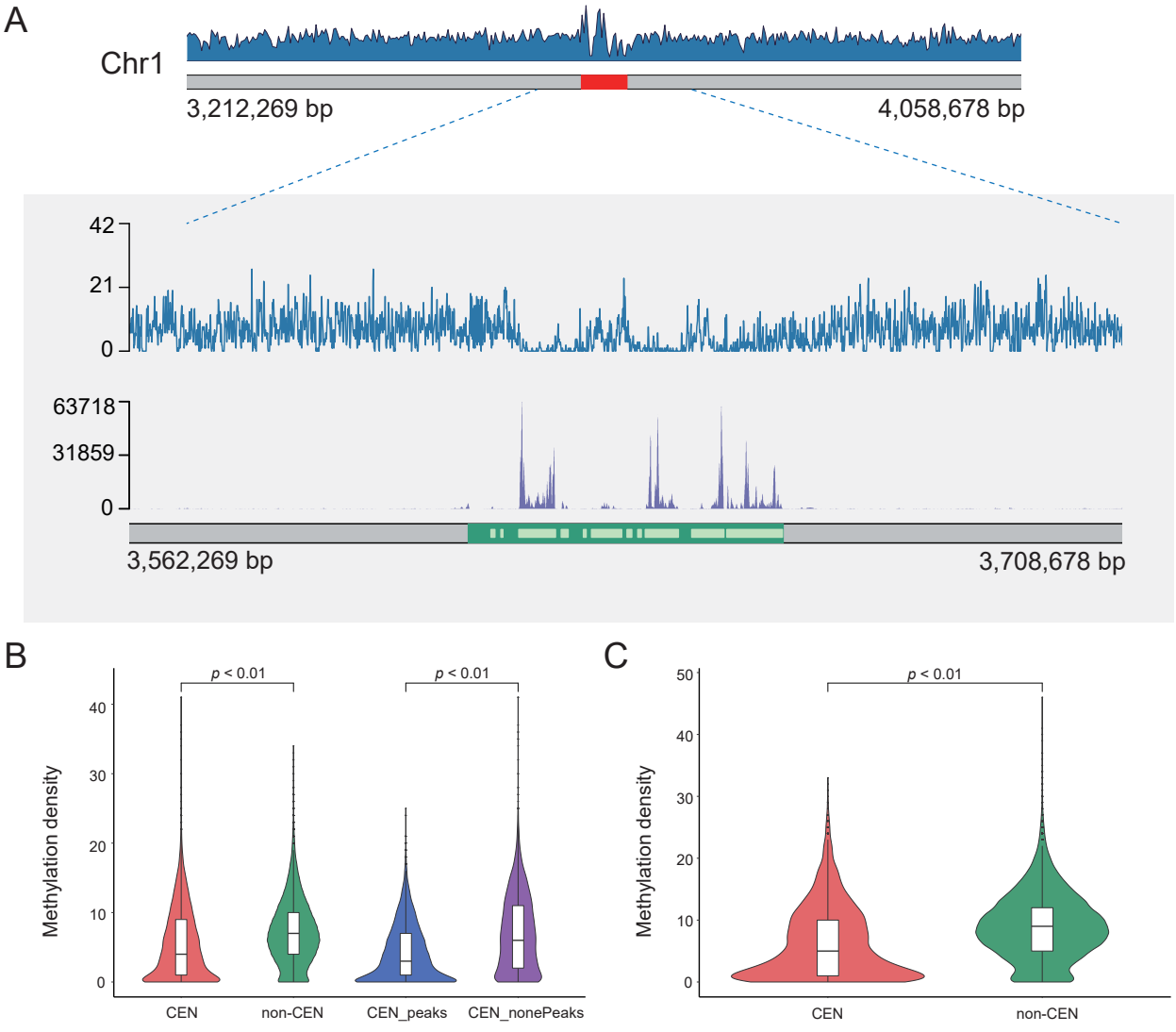
Ni, P., Nie, F., Zhong, Z., Xu, J., Huang, N., Zhang, J., Zhao, H., Zou, Y., Huang, Y., Li, J., Xiao, C.L., Luo, F., and Wang, J. (2023). DNA 5-methylcytosine detection and methylation phasing using PacBio circular consensus sequencing.

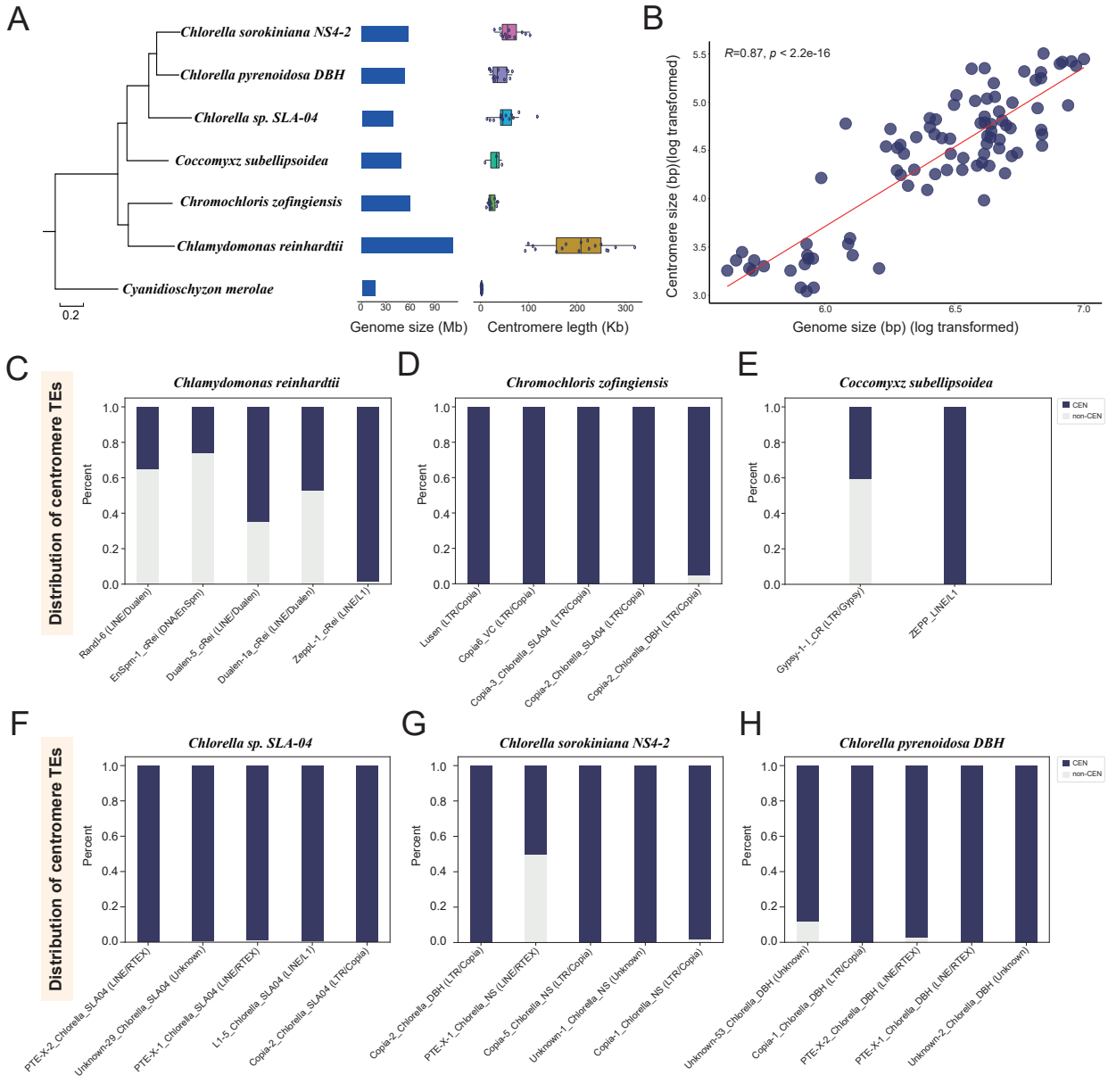
855 Nat Commun **14**, 4054.
856 **Nozaki, H., Takano, H., Misumi, O., Terasawa, K., Matsuzaki, M., Maruyama, S.,**
857 **Nishida, K., Yagisawa, F., Yoshida, Y., Fujiwara, T., Takio, S., Tamura, K.,**
858 **Chung, S.J., Nakamura, S., Kuroiwa, H., Tanaka, K., Sato, N., and Kuroiwa,**
859 **T.** (2007). A 100%-complete sequence reveals unusually simple genomic
860 features in the hot-spring red alga *Cyanidioschyzon merolae*. BMC Biol **5**, 1-8.
861 **Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A.V., Mikheenko, A., ...**
862 **& Phillippy, A.M.** (2022). The complete sequence of a human genome. Science
863 **376**, 44-53.
864 **Payne, Z.L., Penny, G.M., Turner, T.N., and Dutcher, S.K**. (2023). A gap-free
865 genome assembly of *Chlamydomonas reinhardtii* and detection of translocations
866 induced by CRISPR-mediated mutagenesis. Plant Commun **4**, 100493.
867 **Plačková, K., Bureš, P., and Zedek, F.** (2021). Centromere size scales with genome
868 size across Eukaryotes. Sci Rep **11**, 19811.
869 **Presting, G.G.** (2018). Centromeric retrotransposons and centromere function. Curr
870 Opin Genet Dev **49**, 79-84.
871 **Rautiainen, M., Nurk, S., Walenz, B.P., Logsdon, G.A., Porubsky, D., Rhie, A.,**
872 **Eichler, E.E., Phillippy, A.M., and Koren, S.** (2023). Telomere-to-telomere
873 assembly of diploid chromosomes with Verkko. Nat Biotechnol **41**, 1474-1482.
874 **Roth, M. S., Cokus, S. J., Gallaher, S. D., Walter, A., Lopez, D., Erickson, E.,**
875 **Endelman, B., Westcott, D., Larabell, C. A., Merchant, S. S., Pellegrini, M.,**
876 **and Niyogi, K. K.** (2017). Chromosome-level genome assembly and
877 transcriptome of the green alga *Chromochloris zofingiensis* illuminates
878 astaxanthin production. Proc Natl Acad Sci U S A **114**, E4296-E4305.
879 **Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov,**
880 **E.M.** (2015). BUSCO: assessing genome assembly and annotation completeness
881 with single-copy orthologs. Bioinformatics **31**, 3210-3212.
882 **Song, J.M., Xie, W. Z., Wang, S., Guo, Y.X., Koo, D.H., Kudrna, D., Gong, C.,**
883 **Huang, Y., Feng, J.W., Zhang, W., Zhou, Y., Zuccolo, A., Long, E., Lee, S.,**
884 **Talag, J., Zhou, R., Zhu, X.T., Yuan, D., Udall, J., Xie, W., ··· Chen, L.L.**
885 (2021). Two gap-free reference genomes and a global view of the centromere
886 architecture in rice. Mol Plant **14**, 1757-1767.
887 **Suzuki, Y., Myers, E. W., and Morishita, S.** (2020). Rapid and ongoing evolution of
888 repetitive sequence structures in human centromeres. Sci Adv **6**, eabd9230.
889 **Talbert, P.B., and Henikoff, S.** (2010) Centromeres convert but don't cross. PLoS
890 Biol **8**, e1000326.
891 **Talbert, P. B., and Henikoff, S.** (2020). What makes a centromere?. Exp Cell Res
892 **389**, 111895.
893 **Wang, B., Yang, X., Jia, Y., Xu, Y., Jia, P., Dang, N., Wang, S., Xu, T., Zhao, X.,**
894 **Gao, S., Dong, Q., and Ye, K.** (2022). High-quality Arabidopsis thaliana
895 genome assembly with nanopore and HiFi long reads. Genom Proteom Bioinf **20**,
896 4-13.
897 **Wang, Y., Tang, H., Debarry, J.D., Tan, X., Li, J., Wang, X., Lee, T.H., Jin, H.,**
898 **Marler, B., Guo, H., Kissinger, J.C., and Paterson, A.H.** (2012). MCScanX: a

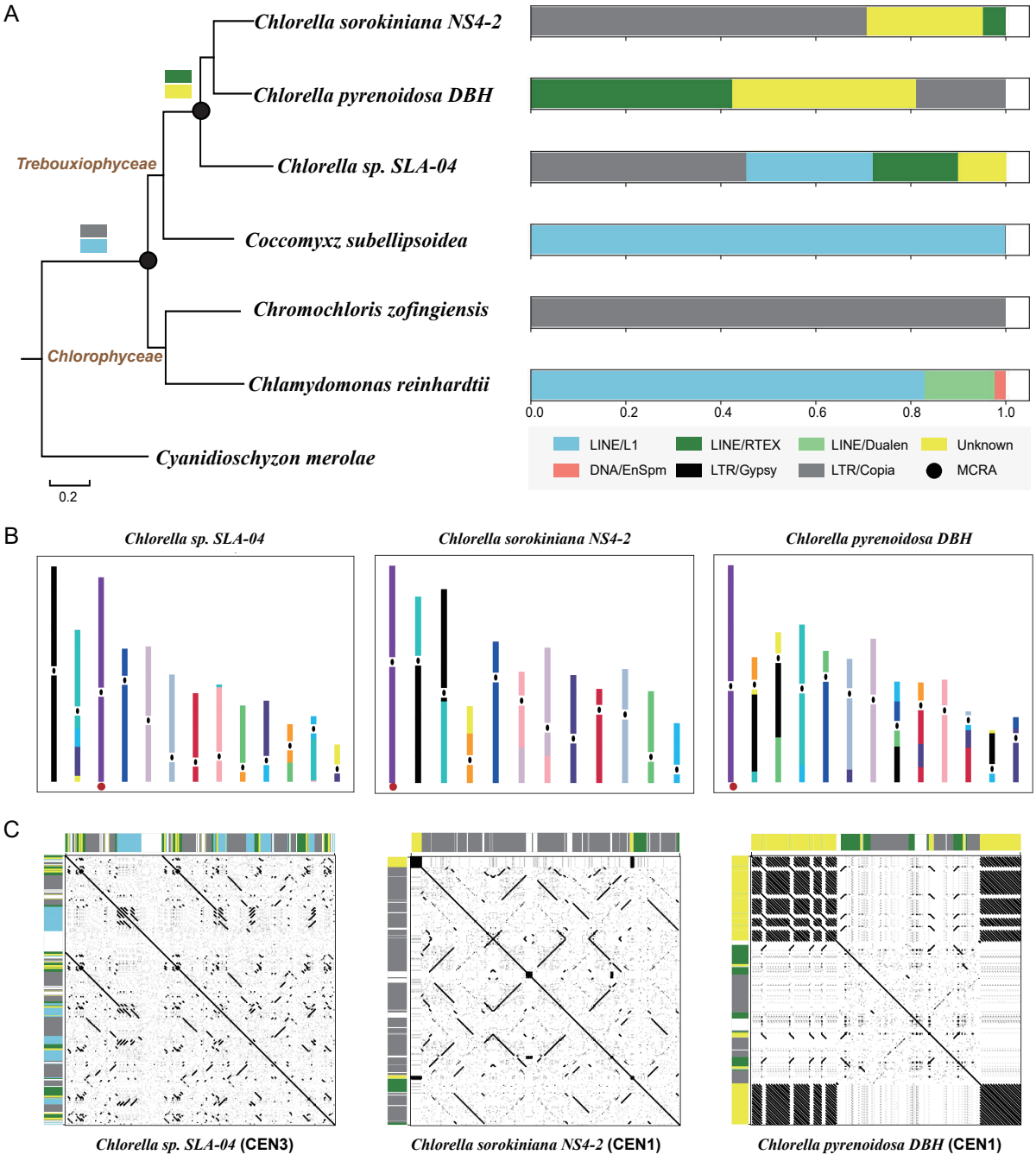
899 toolkit for detection and evolutionary analysis of gene synteny and collinearity.
900 Nucleic Acids Res **40**, e49.
901 **Wick, R.R., Judd, L.M., Gorrie, C.L., and Holt, K.E.** (2007). Unicycler: Resolving
902 bacterial genome assemblies from short and long sequencing reads. PLoS
903 Comput Biol **13**, e1005595.
904 **Windler, M., Stuart, R., Deutzmann, J.S., Mayali, X., Navid, A., D'haeseleer, P.,**
905 **Marcu, O.E., Lipton, M., Nicora, C., and Spormann, A.M.** (2022). Bacterial
906 exometabolites influence *Chlamydomonas* cell cycle and double algal
907 productivity. FEMS Microbiol Ecol **98**, fiac091.
908 **Zhang, W., Lee, H. R., Koo, D. H., and Jiang, J**. (2008). Epigenetic modification of
909 centromeric chromatin: hypomethylation of DNA sequences in the CENH3-
910 associated chromatin in Arabidopsis thaliana and maize. Plant Cell **20**, 25-34.
911 **Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E.,**
912 **Nusbaum, C., Myers, R.M., Brown, M., Li, W., and Liu, X.S.** (2008). Model-
913 based analysis of ChIP-Seq (MACS). Genome Biol **9**, R137.


A
B
C
Chr1
Chr12
D
DBH
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
0.5
1
1
0.5
0


A
Chr1
3,212,269 bp
4,058,678 bp
42
21
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_nonePeaks
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
Coccomyxz 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
Percent
Randi-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
Coccomyxz subellipsoidea
Gypsy-1-I_CR (LTR/Gypsy)
ZEPP_LINE/L1
CEN
non-CEN
F
Chlorella sp. SLA-04
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
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
Coccomyxz 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
MCRA
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)