Haplotype-aware multiomics unveils the regulatory basis of haplodiplontic life-cycle differentiation in a cosmopolitan marine alga

Tzu-Tong Kao[1,2], Ming-Wei Lai[1], Tzu-Haw Wang[1], Chia-Ling Yang[1], Miguel J. Frada[3,4], and Chuan Ku[1]

[1] Institute of Plant and Microbial Biology, Academia Sinica, Taipei, Taiwan.
[2] Department of Biological Sciences, National Sun Yat-sen University, Kaohsiung, Taiwan.
[3] The Alexander Silberman Institute of Life Sciences, The Hebrew University of Jerusalem, Jerusalem, Israel.
[4] The Interuniversity Institute for Marine Sciences in Eilat, Eilat, Israel.

Author for correspondence:
Chuan Ku
Tel: +886-2-27871059
Email: chuanku@gate.sinica.edu.tw

Abstract

Gephyrocapsa huxleyi (formerly *Emiliania huxleyi*), a key coccolithophore alga influencing the global carbon cycle through photosynthesis and calcification, undergoes a haplodiplontic sexual life cycle with a calcifying non-flagellate diploid and a non-calcifying biflagellate haploid stage. To reveal the molecular basis of their morpho-physiological distinctions, we generated chromosome-level genome assemblies and compared the transcriptomes, proteomes, and methylomes for a pair of isogenic haploid and diploid model strains and conducted haplotype-aware analyses of their multiomic features. In addition to calcification and flagella, transcriptomes and proteomes of haploid and diploid cells modulate their differentiation in photosynthesis, sulfatases, DMSP degradation, DNA replication, and endomembrane system and transport. Haploid-diploid differential gene expression can be partially attributable to allelic imbalance (allele-specific expression) in diploid cells. Gene transcript abundance is positively associated with both CG and CHG gene-body DNA methylation, which can be inheritable, allele-specific, and differentiated between life-cycle phases. This multiomic study unravels the regulatory basis of unicellular algal life-cycle differentiation and provides valuable resources for investigating the ecologically important coccolithophore algae.

Key words: allelic imbalance, coccolithophore, *Emiliania, Gephyrocapsa huxleyi*, haplotype phasing, haplodiplontic life cycle, methylome, ploidy level

Introduction

Sexual reproduction and transitions between life-cycle phases are hallmarks of eukaryotic evolution and a major mechanism for generating genomic variations in eukaryotes (Goodenough & Heitman, 2014). Sexual reproduction requires ploidy shifts between diploid and haploid genomes through meiosis and gamete fusion, of which the genetic toolkits date back to the last eukaryotic common ancestor (Sandra K. Floyd & John L. Bowman, 2007; Goodenough & Heitman, 2014). While animals typically have diplontic life cycles with haploid genomes restricted to gametes, land plants, various algae, and other eukaryotes are characterized by haplodiplontic life cycles, where both diploid and haploid cells can divide by mitosis, express distinct attributes, and form independent populations in microbial species (Umen & Coelho,

50 2019). The differentiation in morphological and physiological attributes between life-cycle
51 phases in haplodiplontic life cycles expands the functional scope of these eukaryotes, where
52 diploid and haploid organisms can explore different niches and display distinct life strategies
53 (Frada *et al.*, 2019; de Vries *et al.*, 2021).
54 Coccolithophores are haplodiplontic, calcifying microalgae that are highly abundant and
55 ecologically significant across global oceans (De Vargas *et al.*, 2007; Taylor *et al.*, 2017; Frada
56 *et al.*, 2019; de Vries *et al.*, 2021). They contribute around 10% to global primary production as a
57 major drive of the biological carbon pump (Rousseaux & Gregg, 2013). They also synthesize
58 dimethylsulfoniopropionate (DMSP), a precursor of volatile dimethylsulfide, which may
59 influence climate through cloud formation (Charlson *et al.*, 1987; Alcolombri *et al.*, 2015;
60 Barak-Gavish *et al.*, 2018). Most notably, coccolithophores fix carbon into $CaCO_3$ through
61 calcification, forming calcite plates (coccoliths) on the cell surface. Calcification not only
62 releases CO_2 inside the cell but contributes to carbon removal from the surface ocean with heavy
63 coccoliths serving as ballast enhancing carbon export to the deep ocean (Taylor *et al.*, 2017). The
64 life cycle of coccolithophores alternates between a diploid and haploid phase with distinct
65 morphological and ecological characteristics. *Gephyrocapsa huxleyi* (*sensu* Bendif *et al*., 2019);
66 more commonly known as *Emiliania huxleyi*) is the most abundant and widespread species in the
67 modern ocean and a model in coccolithophore research (Wheeler *et al.*, 2023). Interestingly, for
68 this species, coccolith formation is limited to the diploid phase. Diploid cells regularly produce
69 heterococcoliths, while by contrast the haploid phase is not calcified and exclusively generates
70 unmineralized organic scales that coat the cell surface and that are absent in the diploid cell (Fig.
71 1). In addition, diploid cells are non-motile and haploid cells biflagellate, diploid cells appear
72 less sensitive to high light conditions (Houdan *et al.*, 2005), and strikingly, haploid cells are
73 resistant to specific lytic viruses that infect and kill diploid cells (Frada *et al.*, 2008).
74 Life cycle phase transition of *G. huxleyi* is poorly understood. A pair of diploid and
75 haploid strains, RCC1216 and RCC1217, were identified and isolated from a diploid lab culture
76 in 1999 (von Dassow *et al.*, 2009) following a spontaneous meiosis event. A major limitation in
77 past research has been the lack of a reference genome for the diploid-haploid pair. Previous
78 genome sequences of *Gephyrocapsa* are derived from a diploid strain CCMP1516 (Read *et al.*,
79 2013) or the closely diploid AWI1516 (Skeffington *et al.*, 2023). Despite the very recent origin
80 of *G. huxleyi* ~0.3 million years ago (Bendif *et al.*, 2019), its genomic structure displays

81 substantial variability, with haploid genome size estimates ranging from 99 to 168 Mb and only
82 ~70% of genes shared across different strains (Read *et al.*, 2013). Recent phylogenomic analyses
83 further that RCC1216/RCC1217 and CCMP1516/AWI1516 belong to two distinct clades (Kao *et*
84 *al.*, 2022) or even species (*G. huxleyi sensu stricto* and *G. pseudohuxleyi*, respectively) (Bendif
85 *et al.*, 2023) with distinct morphology, biogeographic range, and plastid genomes (Kao *et al.*,
86 2022). To better understand coccolithophore diploid-haploid differentiation at the molecular
87 level, here we study RCC1216 and RCC1217 as a model system using a multiomics approach.
88 Specifically, we aim to provide a comprehensive view of diploid-haploid differentiation at the
89 molecular level and novel insight into its potential regulatory mechanisms.
90

91 **Materials and Methods**

92 **Experimental Design**

93 To investigate the life cycle phase differentiation of *G. huxleyi*, the model species of
94 coccolithophores, we used a pair of diploid-haploid single-cell derived strains, RCC1216-G9
95 (G9) and RCC1217-F3 (F3). The strains RCC1216 and RCC1217 were originally obtained from
96 the Roscoff Culture Collection (RCC) in 2012. To establish a robust foundation for downstream
97 bioinformatic analyses, we constructed high-quality, haplotype-aware *de novo* genome
98 assemblies for both strains. This involved combining data from long-read (PacBio and Oxford
99 Nanopore) and linked-read (10x genomics plus Illumina) sequencing with chromosome optical
100 mapping (Bionano) to identify existing genomic variations, including copy number variations,
101 structural variations, insertion-deletions (Indels), and single-nucleotide polymorphisms (SNPs).
102 For accurate transcriptome profiling, genome annotation, and gene expression analysis, we
103 integrated data from Illumina RNA-Seq and full-length isoform sequencing (PacBio Iso-Seq).
104 We also quantified and compared global protein expression using the DIA method. Lastly, we
105 profiled genome DNA methylation, a known epigenetic regulatory mechanism for life cycle
106 phase differentiation in some eukaryotes, by identifying 5-methylcytosine at symmetric (CG and
107 CHG) and non-symmetric (CHH) sites using Illumina bisulfite sequencing and direct calling
108 from kinetic data of the PacBio HiFi long reads. By integrating the omics data from these four
109 components, our study provides a comprehensive view of diploid-haploid differentiation at the
110 molecular level.
111

112 **Cell culture preparation**

113 Single-cell derived cultures were generated using a Beckman Coulter MoFlo XDP cell sorter and
114 cultured side-by-side under controlled conditions of ~150 µmol light (16/8 hours light/dark),
115 18˚C, and 60% relative humidity in a K/2 medium supplemented with an antibiotic cocktail of
116 cefotaxime (50 ppm), carbenicillin (50 ppm), kanamycin (20 ppm), and amoxicillin (20 ppm). To
117 maintain active growth, 1 ml of the cell culture was transferred into 40 ml of fresh medium every
118 week. For DNA, RNA, and proteome extraction, the cells were harvested on the 6th morning (at
119 the late log growing phase) after the last refresh.

120

121 **Genome sequencing, assembly, and analyses**

122 *DNA isolation and genome sequencing*—Approximately 1×10^8 cells at the late log growing
123 phase were harvested and embedded in a low-melting-point agarose plug for high-molecular-
124 weight (HMW) DNA extraction. The cells were embedded in 0.75% Low-Melting-Point agarose
125 (Lonza #50101, SeaPlaque^{TM}) in SB buffer (25 mM Tris-HCl, 1 M D-sorbitol, and 25 mM
126 EDTA) and digested with pH 9.5 DB buffer (250 mM EDTA, 1% *N*-lauroylsarcosine) containing
127 1 mg/ml proteinase K at 52˚C for one day. The digested gels were subsequently washed with DB
128 buffer (52˚C, 30 mins, twice), digested with RNAase A (25µ1, 20 mg/mL) in DB buffer (1 hr,
129 37˚C), stabilized with 1x Bio-Rad wash buffer, and DNA was retrieved by drop dialysis,
130 following the Bionano Prep Cell Culture DNA Isolation protocol (#30026, Revision F). For
131 PacBio and Nanopore sequencing, an additional step of salt:chloroform gDNA wash
132 (https://www.pacb.com/wp-content/uploads/2015/09/Shared-Protocol-Guidelines-for-Using-a-
133 Salt-Chloroform-Wash-to-Clean-Up-gDNA.pdf) was performed to remove polysaccharides.
134 Genomic DNA was sequenced using PacBio (Sequel & Sequel IIe), Nanopore (MinION), 10x
135 linked-read plus Illumina (NovaSeq 6000), or only Illumina (HiSeq 2500 and MiSeq) (Table S1).
136 Additionally, chromosomes with labeled Nt.BspQI restriction sites (GCTCTTC) were optically
137 mapped using Bionano (Saphyr).

138 *Genome assembly*—A hybrid approach of *de novo* genome assembly was used,
139 integrating data from short-read, long-read, linked-read, and optical mapping. For the F3
140 genome, draft contigs were assembled with PacBio circular consensus sequences (CCS) using
141 NextDenovo v2.5.0 (Hu *et al.*, 2023). These contigs were then polished with PacBio HiFi and
142 Illumina reads using NextPolish 1.4.0 (Hu *et al.*, 2020), and non -eukaryotic contigs (Table S2)

143 were removed by blasting against NCBI non-redundant nucleotide collection (nt; downloaded on
144 Feb/19/2022). Eukaryotic contigs were scaffolded with 10x linked reads using ARCS (Yeo *et al.*,
145 2018) and with nanopore reads using ARKS-long (Coombe *et al.*, 2018). The genome was
146 further hybrid-scaffolded with Bionano *de novo* genome maps using *hybridScaffold.pl* in
147 Bionano Solve v3.6.1 and gap-filled with corrected long reads using TGS-GapCloser v1.1.1 (Xu
148 *et al.*, 2020). To validate the assembly correctness and copy number, the Illumina reads, PacBio
149 reads, Nanopore reads/NextDenovo draft contigs, and Bionano molecules were mapped back to
150 the superscaffolds. The bridged gaps were manually inspected, and only one (SS06 from
151 coordinate 4,826k to 5,035k) lacked support from any long reads. The haplotype-aware phasing
152 blocks were retrieved by calling haplotype-based variants using Freebayes v0.9.21 (Garrison &
153 Marth, 2012) with short reads and phasing the variants with long reads using WhatsHap v0.19
154 (Patterson *et al.*, 2015) In cases of partially duplicated superscaffolds (e.g. SS24 and SS26 of
155 F3), haplotype-tagged reads were extracted to reconstruct *de novo* assemblies using NextDenovo
156 and NextPolish.
157 *Genomic comparisons*—Genomic structural variations between the two genomes were
158 analyzed by mapping the Bionano molecules of one strain to the genome of the other strain using
159 *RefAligner* in Bionano Solve, and by aligning genome assemblies to each other using Minimap2
160 v2.15 (Li, 2018). Single nucleotide polymorphisms (SNPs) and short insertion-deletions (indels)
161 were identified with GATK v4.3.0.0 (McKenna *et al.*, 2010) that the variable sites were called
162 using *HaplotypeCaller* and underwent hard filtering based on the recommended GATK
163 *VariantFiltration* settings (SNPs: QD < 2.0, QUAL < 30.0, SOR > 3.0, FS > 60.0, MQ < 40.0,
164 MQRankSum < -12.5, and ReadPosRankSum < -8.0; indels: QD < 2.0, QUAL < 30.0, FS >
165 200.0, and ReadPosRankSum < -20.0). The draft of variable sites was then used in base quality
166 score recalibration using *BaseRecalibrator* and *ApplyBQSR* for subsequent variation calling.
167 Convolutional neural net filtering was applied, and variants were annotated using GATK
168 *HaplotypeCaller*, *CNNScoreVariants*, and *FilterVariantTranches*. The G9 variable sites were
169 phased with long reads using WhatsHap. The synonymous and non-synonymous variable sites of
170 each coding gene within and between G9 and F3 were analyzed using SNPGenie v2019.10.31
171 (Nelson *et al.*, 2015).
172
173 **RNA sequencing, genome annotation, and functional annotation**

174 *RNA sequencing*—Total RNA from G9 and F3 was extracted using the RNeasy Plant Mini Kit
175 (Qiagen, Hilden, Germany). Each strain had four short-read libraries prepared using the TruSeq
176 Stranded mRNA Library Prep Kit (Illumina, San Diego, CA, USA), following the manufacturer's
177 recommendations, and sequenced on Illumina NovaSeq 6000 in 150 bp paired-end mode.
178 Additionally, one PacBio Iso-seq library was prepared using the SMRTbell® prep kit (PacBio,
179 Menlo Park, CA, United States) and sequenced on the Sequel IIe.
180 *Genome annotation*—Repetitive sequences in the genome were modeled using
181 RepeatModeler2 v2.0.2a (Flynn *et al.*, 2020) and annotated with Dfam 3.5 and RepBase
182 RepeatMasker Edition (20181026) using RepeatMasker v4.1.2 (Smit *et al.*, 2015). Telomeric
183 repeats were explored and annotated with the Telomere Identification toolKit (tidk) v0.2.0 (Tree
184 of Life Kit, 2022) Both PacBio Iso-seq and Illumina RNA-seq reads were used for annotating the
185 genome, which was soft-masked with repeats. PacBio Iso-seq CCS reads were mapped to the
186 genome and then refined, clustered, and collapsed using PacificBiosciences Isoseq3 v3.8.2, with
187 protein-coding regions predicted by GeneMarkS-T v.5.1 (Tang *et al.*, 2015). The Illumina RNA-
188 seq reads were aligned to the genomes and gene models were trained and predicted using
189 GeneMark-ET and Augustus v3.4.0 in BRAKER 2.1.6 (Hoff *et al.*, 2016). The two annotations
190 were combined and renamed using TSEBRA v1.0.3 (Gabriel *et al.*, 2021) for further isoform
191 analyses, transcripts were filtered based on the Iso-Seq data using PacificBiosciences Pigeon.
192 The shared genes and ortholog groups among genomes were analyzed using the amino acid
193 datasets of haptophytes provided in the associated files of Skeffington *et al*. (2023) (DOI:
194 10.6084/m9.figshare.20463900) with OrthoFinder v2.5.2 (Emms & Kelly, 2019). Non-coding
195 RNAs were predicted and annotated with the Rfam database v14.8 using structRNAfinder v1.0
196 (Arias-Carrasco *et al.*, 2018).
197 *Functional annotation*—The functions of the predicted protein-coding transcripts were
198 annotated using the EnTAP v0.10.8 pipeline (Hart *et al.*, 2020) by searching against NCBI non-
199 redundant protein collection (nr), UniProt Swiss-Prot, and UniProt TrEMBL databases (all
200 downloaded on May/18/2022) using DIAMOND v0.9.9, and Gene Ontologies (GO), KEGG
201 pathway orthologous groups (KO), and protein domains (SMART/Pfam) were explored and
202 annotated based on the eggNOG database (downloaded on May/18/2022).
203
204 **Global protein identification and methylomic sequencing**

205 *Global protein identification*—Total proteins were extracted and digested from three replicates
206 of frozen cell pellets containing approximately 1×10^7 cells of each strain, prepared according to
207 Yun et al. (2020). The protein pellets were resuspended, filtered, and digested using the S-Trap-
208 IMAC protocol (Chen *et al.*, 2023, 2024). Mass spectrometry (MS) analyses were conducted on
209 a Thermo Orbitrap Fusion Lumos Tribrid Mass Spectrometer using DIA mode.
210 *Methylomic sequencing*—To prepare for bisulfite sequencing, cells were lysed using a
211 modified lysis buffer (Tris 0.1M, EDTA 0.05M, NaCl 0.1M, SDS 1%, N-laurylsarcosine 2%,
212 and proteinase K 1mg/ml in pH 8.0; modified from Bendif *et al*. (2019), and then genomic DNA
213 was isolated following the PacBio Salt:Chloroform gDNA Wash protocol. Two samples of each
214 strain were bisulfite converted using the EZ DNA Methylation Gold Kit (ZymoGenetics, Seattle,
215 WA, USA). The two libraries were prepared using the TruSeq DNA Methylation Kit (Illumina)
216 and subsequently sequenced using HiSeq in 150 bp and 250 bp paired-end modes, respectively.
217
218 **Statistical Analysis**
219 *Transcriptomic Analyses*—The expression of transcripts was quantified with Illumina RNA-seq
220 reads using Salmon v1.9.0 (Patro *et al.*, 2017), correcting for GC bias (--gcBias) and performing
221 30 Gibbs sampling runs (--numGibbsSamples 30). Differential expression of genes and
222 transcripts, as well as differential transcript usage, were analyzed with inferential replicate counts
223 (i.e., the Gibbs from Salmon), normalized by the scaledTPM method using the Swish method
224 carried out with the R package *fishpond* v2.0.1. The global allelic imbalance was tested using the
225 SEESAW method (Wu *et al.*, 2023). Differentially expressed genes and allelic imbalance genes
226 were identified with q-values ≤ 0.01, while G9- or F3-specific genes were those with q-values $\leq$
227 0.01 and a mean scaledTPM < 5 in the strain with fewer mapped reads.
228 *Proteomic Analyses*—The quantification and normalization of MS data were performed
229 using the MaxLFQ method as implanted in DIA-NN v1.8.1 (Demichev *et al.*, 2020). Differential
230 protein expression was analyzed based on MaxLFQ value and the number of precursors (used for
231 variance estimation) using the R package *DEqMS* v1.12.1 (Zhu *et al.*, 2020). Differentially
232 expressed genes were identified with q-values ≤ 0.05, while G9- or F3-specific genes were
233 characterized by q-values ≤ 0.05 and absence of peptide detection in two or three out of three
234 replicated samples in the strain with fewer mapped reads.

235 *Methylomic Analyses*—The bisulfite sequencing reads and PacBio HiFi reads that
236 retained kinetics were mapped to the genomes and used for methylation calling with BSBolt
237 v1.4.5 (Farrell *et al.*, 2021) and PacificBiosciences Primrose v1.3.0, respectively. Genome-wide
238 DNA methylation profiling was conducted using R package BSseq v.1.30.0 (Hansen *et al.*,
239 2012), while their association with RNA expression was analyzed using MethGet (Teng *et al.*,
240 2020). Differentially methylated CG, CHG, and CHH sites, genes, and regions were analyzed
241 using the R package *DSS* v.2.42.0 (Wu *et al.*, 2013). Differentially methylated sites were
242 detected with an FDR threshold of ≤ 0.05, while differential methylation of genes and promoter
243 regions (2000 bp upstream of genes) was determined either by mean methylation level (FDR ≤
244 0.05; Datasets S2) or by overlapping with differential methylation regions surpassing areaStats
245 thresholds. In comparisons with RNA expression, an additional set of regions, most correlated
246 with RNA expression (coordinates -350~-50 and 50-450 relative to start codon in CG context;
247 and -200~300 and 500-1500 in CHG context), were also assessed for differential methylation.
248 *Functional enrichment analyses*—Gene and transcript sets of interest, including copy
249 number variation, differential gene/protein expression, transcript usage, allelic imbalance, and
250 methylation levels, were analyzed for enriched GO, KO, SMART, and Pfam terms using the
251 function "*enricher*" in R package *clusterProfiler* v4.2.2 (Wu *et al.*, 2021). Additionally, enriched
252 GO terms were analyzed using R package *topGO* v2.46.0 (Alexa & Rahnenfuhrer, 2021) with
253 algorisms "classic" or "weight01". Enriched GO terms identified by *clusterProfiler* ($p < 0.05$)
254 were clustered using the MCL-based overlap coefficient in GOMCL v0.0.1 (Wang *et al.*, 2020).
255

256 **Results**

257 **Genome assemblies and annotations**

258 Our *de novo* genome assemblies exhibit substantially improved genome contiguity
259 compared with other assemblies of *Gephyrocapsa* (Figure 2A). The genome assembly of the
260 haploid strain F3 is 151 Mb, comprised of 26+1 superscaffolds (SS), where SS26 is a partial
261 duplicate of SS24 and not present in the diploid G9. The *G. huxleyi* genomes display a high GC
262 content (~66%) and contain a significant proportion of repetitive sequences (~36% after
263 excluding the sequences overlapped with genes; Fig. 2b and Table 1). The telomeric repeat
264 sequence TTAGGG, commonly found in animals and other eukaryotes (Fulnečková *et al.*, 2013)
265 was detected in some superscaffolds but not all. The coding genes annotated utilizing both

266 PacBio Iso-Seq full-length transcripts and Illumina RNA-seq reads amount to 35,323 in F3, with
267 4,005 (11%) of them absent in the CCMP1516 assembly (RefSeq GCF_000372725.1) (Read *et*
268 *al.*, 2013) or its still calcifying variant AWI1516 (Skeffington *et al.*, 2023) (Fig. 2c). A total of
269 827 non-coding RNA genes were identified, including 349 rRNA genes that are mostly located
270 on SS02, SS03, SS05, and SS06 (Fig. 2d).
271
272 **Genome variations and phasing**
273 F3 and G9 genomes exhibit notable structural divergence (Fig. S1), including two chromosome-
274 level recombination events between superscaffolds (SS01-SS05 and SS09-SS17; Fig. 2e), partial
275 duplication of SS24 in F3, partial loss of SS01 in G9, and acquisition of partial copies of SS10,
276 SS15, and SS18 in G9 (Fig. 2f). Noteworthy is a biased distribution of large indels (>350 bp),
277 with a higher frequency of insertions in F3 (or deletions in G9) compared to F3 deletions (or G9
278 insertions), at a ratio of 444:97.
279 Approximately 0.33% (~491 k) of sites within the G9 genome are heterozygotic, with the
280 majority (99.87%) of these sites having one allele corresponding to the F3 haplotype. Specific
281 regions in the G9 genome (e.g., SS06, SS20, and partial SS01 and SS05) display low
282 heterozygosity (Figs. 2g, S2). Comparison with a single-cell-derived strain (B4) from another
283 RCC1216 culture (kindly provided by the lab of Assaf Vardi) indicates that regions of low
284 heterozygosity in G9 and B4 genomes mostly do not overlap, suggesting they are due to recent
285 loss of heterozygosity after the separation of RCC1216 and RCC1217. A total of 13,405 sites
286 (0.0089%) have diverged between F3 and G9, primarily concentrated within the regions of low
287 heterozygosity, presumably due to the loss of the G9 allele corresponding to the F3 haplotype.
288 We phased the variable sites with long reads, resulting in 213 phasing blocks covering 94.90% of
289 the haploid genome (Fig. 2h; Dataset S1). The number of recombinations is positively correlated
290 with the length of phasing blocks, with the average recombination distance estimated to be ~56
291 kb (Fig. 2i).
292 By investigating the variation in gene copy numbers, as defined by gDNA ratios, of the
293 two genomes, we found that only 75% of the genes possess two copies in the diploid genome
294 relative to the haploid (Fig. 3a). Among the 25% that deviate from an expected 2:1 ratio, 12%
295 have 3:1 ratio and 11% have a 1:1 ratio, which might be due to factors such as chromosome
296 anomalies (e.g., aneuploidy) or gene duplications and losses. Notably, 2% of the genes have an

297 even more pronounced deviation (with a ratio of < 0.5 or > 3.5 to 1) and were subsequently
298 excluded from further analyses.
299
300 **Transcriptomic landscapes and gene functions**
301 RNA differential expression comparisons between G9 and F3 were conducted, utilizing four
302 replicates of Illumina RNA-seq samples (Table S1), collected on the sixth morning after
303 transferring to a fresh 100 ml K/2 medium (starting with 100 cells/µl). Gene copy numbers
304 correlate with RNA transcript abundance (Figs. 3b, 3c), but there is a high variance of RNA
305 differential expression among genes with a specific haploid-diploid gDNA ratio (Fig. 3b),
306 suggesting the impact of copy numbers on gene expression is minor. Out of the 25,224 genes
307 with detectable transcripts, 1,412 genes (5.60%) have transcripts only in the diploid (mean scaled
308 transcripts per million [scaledTPM] < 5 in F3) and 2,704 genes (10.72%) only in the haploid
309 (mean scaledTPM < 5 in G9) (Fig. 3d). Given the phylogenetic distance of haptophytes from
310 well-studied eukaryote models (animals, fungi, and plants), our understanding of haptophyte
311 gene functions remains very limited. Among the transcribed genes, only 53.26% (13,434 genes)
312 are associated with known functions, and the proportions for only diploid-expressed (33.92%;
313 479 genes) and only haploid-expressed (39.09%; 1,057 genes) genes are even lower (Dataset
314 S2).
315 The functions enriched among genes only expressed in the diploid were divided into five
316 clusters (Fig. 3e). The predominant cluster (red) primarily consists of membrane proteins,
317 including genes implicated in coccolithogenesis, such as HCO_3^- transporters in the solute carrier
318 4 family, cation and proton channels, antiporters, and vesicle fusion SNARE proteins. A closely
319 related, highly compact cluster (blue) comprises Myb-like transcription factors and Ca^{2+}
320 channels. The third cluster (purple) mainly encompasses genes involved in carbohydrate
321 metabolism, including the glycoside hydrolase 3 family, beta-glucan synthase, and
322 galactosyltransferase. Two smaller clusters, each containing around 20 genes, include guanylate
323 cyclase receptors and adenylate guanylate cyclase (yellow) and sulfatases (magenta).
324 The predominant functional cluster (Fig. 3f; red) of the haploid-specifically expressed
325 genes primarily consists of genes related to flagella assembly, intra-flagellar transportation,
326 microtubule-based movement, nuclear division, and certain cell cycle regulatory proteins. Other
327 clusters include membrane receptors (blue; ligand-gated channels, protein kinases, blue light

328 photoreceptors, etc.), carbohydrate metabolism (purple; glycoside hydrolase 5 family,
329 glycosyltransferase, cellulases, etc.), sulfatases plus aminoglycan-related glycosyltransferases
330 (yellow), and sulfotransferases (magenta).
331 Using Pigeon in PacBio Transcript Toolkit, we identified isoforms in 6,092 genes, with
332 2,145 genes (35.21%) showing isoform differential expression patterns: 1,232 (20.22%) had up-
333 regulated isoforms in the diploid, and 1,229 (20.17%) had up-regulated isoforms in the haploid.
334 Genes with diploid up-regulated isoforms were enriched in plastid thylakoid membrane
335 components, such as the light-harvesting complex, cytochrome b6f complex, and photosystem II
336 genes (Dataset S3). Those genes with haploid up-regulated isoforms were enriched in mitotic cell
337 cycle-related genes, such as cyclins, mitogen-activated protein kinases, and kinetochore complex
338 components.
339
340 **Proteomic profiling**
341 Global proteins were identified, quantified, and compared using three replicates of data-
342 independent acquisition (DIA) data for G9 and F3 (Table S1). Due to limitations in protein
343 detection, the number of genes with detectable proteins (7,116) is lower than that with RNA
344 transcripts detected (25,224). There is a moderate correlation between protein and RNA
345 expression (R = 0.52~0.56; Figs. 3a-3c). Among genes with protein products detected, 1,082 are
346 up-regulated in the diploid, with the majority (805 genes) being specifically detected in the
347 diploid and enriched with transportation proteins (solute carriers, major facilitators, ion
348 transporters, and antiporters) (Fig. 4e; Dataset S3). The 599 genes with protein expression up-
349 regulated or only detected in the haploid are grouped into three enriched functional clusters:
350 mini-chromosome maintenance (MCM) complex proteins, protein degradation enzymes
351 (peptidases and proteasomes), and heat-shock proteins plus mitochondrial matrix proteins
352 (dehydrogenases and reductases) (Fig. 4f). It is notable that flagella- and cell motion-related
353 genes that were highly upregulated in the RNA differential expression profile were not enriched
354 in the haploid proteome. Unlike RNA, which is typically localized within the main cell body,
355 flagella are easily detached from the cell body during microscopic observations. Therefore, the
356 absence of these proteins may be due to the loss of flagella during the pelleting step of sample
357 preparation.
358

359 **DNA methylation profiles and gene expression**

360 Whole-genome 5-methylcytosine (5mC) patterns were profiled using two replicates of bisulfite
361 sequencing for the clonal populations of F3 and G9 (Table S1). DNA methylation is prevalent in
362 symmetric contexts, with CG and CHG sites accounting for 95% of the methylated sites (Fig.
363 5a). In the haploid, almost all sites are either lowly methylated (methylated reads < 20%,
364 accounting for 80% of CG and 90% of CHG) or highly methylated (methylated reads > 80%,
365 accounting for 18% of CG and 9% of CHG). The diploid has lowly methylated (76% of CG and
366 89% of CHG) and highly methylated (12% of CG and 5% of CHG) sites, along with sites
367 displaying intermediate methylation levels (methylated reads between 20% to 80%; 11% of CG
368 and 5% of CHG) (Fig. 5b). Cross-referencing with the genomic variation data indicates only a
369 negligible number of intermediate-methylated sites resulting from transitions in one of the two
370 alleles (2,090 T-C and 2,000 A-G transition sites corresponding to 0.2% of all intermediate-
371 methylated sites). Instead, they are mainly caused by differential methylation on sister
372 chromatids.

373 DNA methylation of transposable elements (TEs) has been linked to heterochromatin
374 formation and can influence life cycle phase differentiation (Vigneau & Borg, 2021). However,
375 we did not observe a higher methylation level of TEs (Fig. S3). High CG methylation was
376 observed in some rRNA genes, but not in other repetitive sequences (Fig. S3). Of the 43 genes
377 annotated with DNA methyltransferase activity (GO:009008), two are up-regulated in the diploid
378 and seven in the haploid. Notably, all three N-6 adenine methyltransferases (EHUX_g4335,
379 EHUX_g18645, EHUX_g32175) are up-regulated in the haploid. However, only a few reads
380 showed N-6 adenine methylation signals in our PacBio HiFi reads without any specific pattern.

381 To analyze the correlation between DNA methylation and RNA expression, we divided
382 genes into six groups, including unexpressed genes ("no") and five bins of expressed genes
383 ranked by scaled transcripts per million (scaledTPM) (Soneson *et al.*, 2015) (low to high: groups
384 1 to 5). Groups with higher RNA expression are associated with lower mean CG site methylation
385 levels upstream of the start codon and higher levels downstream (Fig. 5c). For CHG sites, higher
386 RNA expression is linked to lower methylation levels around the start codon and higher levels in
387 the gene body (Fig. 5d). In both life cycle phases, E3 ubiquitin-protein ligase is among the top
388 10% of genes with the highest mean gene body CHG methylation levels (Dataset S3). A
389 comparison between the two life cycle phases reveals positive association between RNA

390 expression and gene body CG or CHG methylation for specific functions (Dataset S3). For
391 instance, genes related to histone demethylation, mitosis, endocytosis, and flagella assembly
392 often exhibit higher RNA expression and higher methylation levels in the gene body of the
393 haploid (Fig. 5e). Genes of certain functions with significant differences in promoter methylation
394 levels show differential RNA expression in the opposite direction. For instance, mitochondrion-
395 targeting and ubiquitin-dependent proteins tend to show higher RNA expression and lower
396 methylation levels in the haploid, whereas ethylene metabolic and SNARE proteins tend to
397 exhibit higher RNA expression and lower methylation levels in the diploid (Figs. 5f, 5g).
398
399 **Genome-level allelic imbalance of RNA expression in diploid**
400 Allelic imbalance is the phenomenon where the two alleles of a gene in a diploid cell have
401 different RNA expression levels (Wu *et al.*, 2023). Among the 10,762 heterozygous genes, 1,809
402 genes (16.81%) show significant signals of imbalanced RNA expression (Fig. 6a). Through
403 comparison with the genome sequence of the haploid strain F3, we designated one G9 allele as
404 the F3-like allele and the other as the alternative allele. Among the genes with allelic imbalance,
405 a slightly higher proportion of F3-like alleles are up-regulated compared with the alternative
406 alleles (Fig. 6b). Compared with genes with F3-like alleles up-regulated (23%), a much high
407 proportion (52%) of genes with up-regulated alternative alleles also higher expression in the
408 diploid than the haploid (Fig. 6c). This observation suggests that a fraction (approximately 400
409 genes) of the differentially expressed genes identified in this study may be attributed to allelic
410 imbalanced RNA expression.
411 The majority of the heterozygous genes have inter-allelic SNPs and/or indels (>80%)
412 within 2,000-bp upstream of the start codon, but the percentage with allelic imbalance is not
413 higher than without imbalance (Fig. 6d). However, we see a higher proportion of allelic
414 imbalanced genes have indels in their coding sequences, particularly when these indels result in
415 frameshifts and premature stop codons (Fig. 6d). We further examined the links between allelic
416 imbalance and allele-specific methylation. A higher percentage of genes with allele-specific
417 methylation patterns (in either the promoter region or gene body) was observed in genes with
418 allelic imbalance than without in both CG and CHG contexts (Fig. 6e). Leveraging PacBio long-
419 read DNA and RNA sequencing, we were able to better detect, phase, and visualize allelic
420 imbalance and allele-specific methylation in a haplotype-resolved manner. For example,

421 EHUX_g5658, a sulfatase gene with transcripts only detected in G9, displays allele-specific
422 methylation in both the promoter region and gene body. Its F3-like allele exhibits the same
423 methylation pattern as in the F3, and similar to the F3 gene, this allele is not transcribed (Fig. 7).
424
425 **Histones and histone modifications**
426 Epigenetic modifications of histones have been proposed as key players in life cycle phase
427 transitions (Vigneau & Borg, 2021). Multiple copies of each core histone protein, including
428 H2A, H2B, H3, and H4, were identified, but the H1 linker protein, which is found in animals,
429 fungi, plants, and kinetoplastids (Discoba) (Talbert *et al.*, 2019), is not present in the *G. huxleyi*
430 or *G. pseudohuxleyi* genomes. Both haploid and diploid phases exhibited up-regulation of
431 specific histone paralogs, as well as histone lysine methyltransferases and deacetylases (Dataset
432 S4). Genes associated with H3K9 methylation are known to influence life cycle phase
433 differentiation in plants, green algae, and diatoms (Vigneau & Borg, 2021). In our study, the
434 transcripts of K9 histone-lysine N-methyltransferase SUV39H2 and the protein of enhancer of
435 zeste E(z) were exclusively expressed in the diploid phase. Two other K9 histone-lysine
436 methyltransferases (EHUX_g17408 and EHUX_g28648) and some histone arginine N-
437 methyltransferases (PRMTs) are also up-regulated in the diploid.
438
439 **Diploid-haploid physiological differentiation**
440 We summarized the multiomic profiles of the diploid and haploid phase of *G. huxleyi* by
441 focusing on genes related to various physiological functions (Fig. 8; Notes S1, S2, and Dataset
442 S4) (Mackinder *et al.*, 2011; Benner *et al.*, 2013; Taylor *et al.*, 2017; Brownlee *et al.*, 2021;
443 Skeffington *et al.*, 2023). The diploid and haploid show clear differences in certain carbonic
444 anhydrases, transporters, and SNAP receptors (SNAREs) that are involved in coccolithogenesis.
445 We also demonstrate that they utilize different sets of genes involved in polysaccharide and
446 sulfate metabolism. The considerably higher concentration of DMSP in the diploid (Heidenreich
447 *et al.*, 2019) is attributable to more genes in DMSP degradation pathways being upregulated in
448 the haploid. In line with a more mixotrophic nutrition, genes related to "endocytic vesicle"
449 (GO:0030139) are up-regulated and have lower CHG methylation in the haploid. On the
450 contrary, many photosynthesis-related genes, including those for chlorophyll synthesis and light
451 harvesting, are up-regulated in the diploid.

452

453 **Discussion**

454 Our multiomic sequencing and analyses lay a robust foundation for advancing coccolithophore
455 research. Highly contiguous *de nov*o genome assemblies serve a critical role in dissecting the
456 diploid-haploid variation in the life cycle of *G. huxleyi*. Profiling and comparing the
457 transcriptomes, proteomes, and methylomes further provide insights into potential gene
458 expression regulatory mechanisms behind morphological, physiological, and ecological
459 differentiation of diploids and haploids. This study underscores the importance of genome
460 phasing in comparative genomics and multiomics, and carries profound implications for future
461 investigations into haploid-diploid differentiation and calcification mechanisms in
462 coccolithophores.

463 Consistent with known high variation in genome size, structure, and gene content within
464 *Gephyrocarpsa* and haptophytes (Read *et al.*, 2013; Yang *et al.*, 2023), more than 3,000 (13%)
465 orthologous gene group identified in *G. huxleyi* are absent in *G. pseudohuxleyi* (Fig. 2c). The
466 genomes of RCC1216 and RCC1217, a haploid-diploid pair separated by approximately twenty
467 years, unveiled notable copy number differences at the superscaffold level and hundreds of
468 structural variations, including two instances of chromosome-level recombination (Fig. 2e; Fig.
469 S1). These findings highlight the prevalence of structural variation between closely related
470 coccolithophore strains (Note S3). Instead of meiosis and sexual reproduction, the high
471 frequency of variations is more likely due to homologous recombinations during DNA repair
472 (mitotic recombinations), as has been shown in diatoms (33). Still, the Genomic copy numbers
473 and structural variations (Fig 2e, 2f; Dateset S1) can be potential post-zygotic barriers to nuclear
474 gene flow, offering a plausible explanation for the observed restricted nuclear gene flow among
475 morphotypes within *Gephyrocapsa* (Filatov *et al.*, 2021; Kao *et al.*, 2022). In line with DNA
476 content estimation based on flow cytometry (Houdan *et al.*, 2004; Frada *et al.*, 2017), we show a
477 higher frequency of insertions in F3 (or deletions in G9) than F3 deletions (or G9 insertions)
478 (Dateset S1), which leads to a diploid genome size smaller than twice the haploid genome size
479 (Table 1).

480 The diploid-haploid differentiation in *G. huxleyi* involves genes encoding diverse cellular
481 and biochemical functions (Fig. 8; Notes S1, S2, Dataset S4). The enrichment of transmembrane
482 transporters among up-regulated diploid genes underscores the crucial role of ion transporters in

483 calcification, including bicarbonate transporters (SLC4, ABCC8), calcium ion transporters and
484 channels (CAX3, SLC24, TRPM), and H^+/Na^+ exchangers and a vacuolar-type ATPase for
485 removal of the byproduct proton to maintain pH homeostasis. For the exocytosis of coccoliths,
486 SNAP receptors (SNAREs) that mediate vesicle fusion with target membranes show diploid-
487 specific expression (Fig. 8; Dataset S4) (von Dassow *et al.*, 2009). We identified 15 carbonic
488 anhydrases (CAs), which are pivotal for the conversion between carbon dioxide and bicarbonate
489 and potentially regulate carbon acquisition and allocation between photosynthesis and
490 calcification. Among the four CA families, αCAs and δCAs exhibit up-regulation in the diploid
491 phase and are likely to be directly involved in calcification.

492 Regulation of $CaCO_3$ crystal growth on an organic baseplate within the coccolith vesicle
493 involve both proteins and polysaccharides. As the major organic component of coccoliths,
494 coccolith-associated polysaccharides (CAPs) are acidic, can contain sulfate groups, and mediate
495 the calcification process by directing calcium ions and controlling crystal morphology (Marsh *et*
496 *al.*, 2002; Gal *et al.*, 2016; Marzec *et al.*, 2019). Transcriptomic differences between the diploid
497 and haploid strains are marked by an enrichment of enzymes for carbohydrate metabolism
498 (including glycosyltransferases), sulfatases, and sulfotransferases (Fig. 3; Dataset S3), suggesting
499 that diploid and haploid cells have a major difference in their carbohydrate components that
500 might be related to coccolithogenesis. The diploid-haploid differential expression can be
501 regulated through allele-imbalanced transcription (Fig. 6), which is in turn associated with allele-
502 specific methylation patterns in the diploid (Fig. 7). To further elucidate the mechanisms for
503 epigenomic reprograming in *G. huxleyi*, it remains a major challenge to systematically
504 manipulate diploid-haploid transitions in the lab.

505 In addition to coccolithogenesis, the transcriptomic and proteomic profiles reflect other
506 physiological differences between the haploid and diploid phases (Fig. 8; Note S2; Dataset S4).
507 The presence of flagella and a likely more mixotrophic nutrition in the haploid are in contrast
508 with more abundant chlorophyll a under high irradiance (Hariskos et al., 2015; Ruan et al., 2023;
509 300 and 100 μmol photons m^{-2} s^{-1}, respectively) and the up-regulation of photosynthesis-related
510 genes in the diploid. The low abundance of DMSP in the haploid (Heidenreich *et al.*, 2019) is
511 possibly due to the up-regulation of DMSP catabolic enzymes in both demethylation and
512 cleavage pathways, including two key DMSP lyases, DddD and Alma. Given that DMSP and
513 derived compounds are known to mediate microbial interactions (Alcolombri *et al.*, 2015; Barak-

514 Gavish *et al.*, 2018), differences in DMSP metabolism imply that the two life-cycle phases might
515 differ in their microalga-microbe interactions, as also suggested by the different bacterial
516 composition associated with G9 and F3 cultures (Table S2). Consistent with the previous
517 observation that haploids show a higher growth rate under higher irradiance (Hariskos *et al.*,
518 2015), we found that MCM, crucial for DNA replication initiation, is enriched among the
519 haploid up-regulated proteins (Fig. 4f; Dataset S4), suggesting increased mitotic activity in F3.
520 Together these physiological differences point to potential ecological differentiation between the
521 diploid and haploid, which merits further investigation using field-collected data.
522 Epigenetic modifications of DNA and histones have been proposed as key players in life
523 cycle phase transitions (Vigneau & Borg, 2021), with varying patterns across the tree of
524 eukaryotes. Animals typically exhibit methylation in the CG context, while land plants and the
525 green alga *Chlamydomonas* manifest methylation in CG, CHG, and CHH contexts (Kawashima
526 & Berger, 2014; Lopez *et al.*, 2015). In contrast to methylation in red algae (bias towards non-
527 CG) (Lee *et al.*, 2018), diatoms (predominantly CG) (Veluchamy *et al.*, 2013; Traller *et al.*,
528 2016), and brown algae (absent or minimal and confined to CHH) (Cock *et al.*, 2010; Fan *et al.*,
529 2020). *G. huxleyi* methylation is found in CG (~18%) and CHG (~9%) contexts, with relatively
530 rare CHH methylation (~1%) (Figs. 5a, 5b, S3). We show that methylation patterns can be
531 heritable, as evidenced by the similarity between the F3-like haplotype in G9 and the F3 genome
532 in certain regions after at least 20 years of separation (Fig. 7). Unlike Chloroplastida, where gene
533 body CG methylation up-regulates gene expression and methylation in CHG contexts is typically
534 associated with low expression (Bewick & Schmitz, 2017), genes with higher expression levels
535 in *G. huxleyi* display increased methylation in both CG and CHG contexts (Fig. 5c, 5d). Histones
536 H2A, H2B, H3, and H4 were identified within the genome, and each core histone class has
537 specific paralogs up-regulated in either haploid or diploid cells (Dataset S4). In green algae and
538 bryophytes, the regulation of life cycle phase transition has been suggested to involve histone
539 modifications, particularly the tri-methylation of lysine 27 on histone H3 protein (H3K27me3)
540 that silences gene expression (Vigneau & Borg, 2021). The up-regulation of H3K9
541 methyltransferase SUV39H2, enhancer of zeste E(z), and arginine methyltransferases in the
542 diploid of *G. huxleyi* implies a potential role of histone methylation in diploid-haploid
543 differentiation, but further studies are needed to profile histone modifications and test this
544 hypothesis.

545 This study presents a comprehensive view of the life-cycle differentiation between
546 diploid and haploid phases of *Gephyrocapsa* — globally distributed and highly abundant
547 calcifying algae with pivotal roles in marine carbon cycling. In particular, our results show that
548 (1) their genomes are dynamic and prone to variations, including chromosome-level structural
549 variations, biased indel distribution between the diploid and haploid, and loss of heterozygosity
550 in the diploid, as revealed by their highly contiguous *de novo* genome assemblies; (2)
551 Transcriptomic and proteomic differences of diploid and haploid cells modulate their
552 differentiation in coccolithogenesis, flagellar assembly and movement, photosynthesis,
553 carbohydrate metabolism, DMSP degradation, DNA replication, and endomembrane system and
554 transport; (3) Gene transcript abundance is correlated with patterns of DNA methylation, which
555 occurs in CG and CHG contexts and can be inheritable, allele-specific, and differentiated
556 between life-cycle phases. Haplodiplontic life cycles are widespread across diverse lineages of
557 eukaryotes, where coccolithophores exemplify both algae and microbial eukaryotes that exist in
558 two morphologically and physiologically distinct ploidy phases. By integrating genomic,
559 transcriptomic, proteomic, and methylomic analyses, this study not only illuminates the
560 molecular basis of diploid-haploid differentiation and maintenance in coccolithophores, but lays
561 the groundwork for future research on their life cycling in marine ecosystems and its
562 implications for the distribution, ecological roles, and microbial interactions of coccolithophores.
563

564 **Acknowledgments**

565 We thank Pui-Yan Kwok, Hsiao-Jung Kao, and Hsiao-Huei Chen for their generous help with
566 the Bionano optical mapping, Irina V. Agarkova and James L. Van Etten for advice on HMW
567 DNA extraction, Chih-Horng Kuo for sharing their computing equipment, and Assaf Vardi for
568 sharing the cultures of RCC1217 and RCC1216. We are also grateful for the sequencing services
569 and technical support provided by BRCAS NGS High Throughput Genomic Core, IPMB
570 Genomic Technology Core, National Center for Genomic Medicines, and Biotools. We
571 acknowledge Yu-Fang Tseng of IPMB Cell Biology Core for flow cytometry analysis and
572 sorting services, Wann-Neng Jane and the staff of the EM division for EM sample preparation,
573 Chuan-Chih Hsu and IPMB Proteomics Core for their assistance with global proteomic
574 quantification, Academia Sinica Common Mass Spectrometry Facilities for proteomics and
575 protein modification analysis facilities, and the National Center for High-performance

576 Computing and IPMB Bioinformatics Core Lab for computer time and facilities. We appreciate
577 the consultation on genome-wide methylation analyses by the lab of Pao-Yang Chen, advice on
578 genome assembling strategies by the lab of Chung-Yen Lin, and the suggestions of Michael Borg
579 on the manuscript,. Lastly, we thank the help of other lab members with this project, especially
580 Bo-Wei Wang and Elias Miguel Barreira. This work was supported by the intramural funding of
581 the Institute of Plant and Microbial Biology (CK), Academia Sinica Career Development Award
582 (AS-CDA-110-L01 to CK), Academia Sinica Postdoctoral Scholar Program (AS-PD-11001-
583 L007 and AS-PD-11201-L01 to T.K.), and National Science and Technology Council, Taiwan
584 (108-2311-B-001-040-MY3 and 111-2611-M001-008-MY3 to C.K.). The funding bodies had no
585 role in the design of the study, in the collection, analysis, and interpretation of data, or in writing
586 the manuscript.
587

588 **Competing interests**

589 All authors declare they have no competing interests.
590

591 **Author contributions**

592 This project was initiated and conceptualized by CK and MF. The primary research activities,
593 including DNA extraction, sequencing, genome assembly, data analysis, and synthesis were
594 conducted by TK. ML was responsible for RNA purification for Illumina RNA sequencing and
595 PacBio Iso-Seq, as well as preparing cell cultures for TEM observation and global protein
596 identification. TW contributed to maintaining cell cultures, participated in genome assembly and
597 annotation, and assisted with data curation. CL played a key role in the establishment of the
598 HMW DNA extraction protocol. The initial draft of the manuscript was prepared by CK and TK.
599 All authors contributed their insights, reviewed the manuscript, and provided valuable input,
600 ultimately approving the final version. Throughout the project, CK provided supervision and
601 guidance.
602

603 **Data availability**

604 All data are available in the main text or the supplementary materials. The genome assemblies
605 have been deposited in the NCBI under BioProject PRJNA774110 and the accession numbers
606 are listed in Table S1.

607

608 **References**

609 **Alcolombri U, Ben-Dor S, Feldmesser E, Levin Y, Tawfik DS, Vardi A**. **2015**. Identification
610 of the algal dimethyl sulfide–releasing enzyme: a missing link in the marine sulfur cycle.
611 *Science* **348**: 1466–1469.

612 **Alexa A, Rahnenfuhrer J**. **2021**. topGO: Enrichment Analysis for Gene Ontology.

613 **Arias-Carrasco R, Vásquez-Morán Y, Nakaya HI, Maracaja-Coutinho V**. **2018**.
614 StructRNAfinder: an automated pipeline and web server for RNA families prediction. *BMC*
615 *Bioinformatics* **19**: 55.

616 **Barak-Gavish N, Frada MJ, Ku C, Lee PA, DiTullio GR, Malitsky S, Aharoni A, Green SJ,**
617 **Rotkopf R, Kartvelishvily E, *et al.* 2018**. Bacterial virulence against an oceanic bloom-
618 forming phytoplankter is mediated by algal DMSP. *Science Advances* **4**: eaau5716.

619 **Bendif EM, Nevado B, Wong ELY, Hagino K, Probert I, Young JR, Rickaby REM, Filatov**
620 **DA**. **2019**. Repeated species radiations in the recent evolution of the key marine
621 phytoplankton lineage *Gephyrocapsa*. *Nature Communications* **10**: 4234.

622 **Bendif EM, Probert I, Archontikis OA, Young JR, Beaufort L, Rickaby RE, Filatov D**.
623 **2023**. Rapid diversification underlying the global dominance of a cosmopolitan
624 phytoplankton. *The ISME Journal* **17**: 630–640.

625 **Benner I, Diner RE, Lefebvre SC, Li D, Komada T, Carpenter EJ, Stillman JH**. **2013**.
626 *Emiliania huxleyi* increases calcification but not expression of calcification-related genes in
627 long-term exposure to elevated temperature and pCO2. *Philosophical Transactions of the*
628 *Royal Society B: Biological Sciences* **368**: 20130049.

629 **Bewick AJ, Schmitz RJ**. **2017**. Gene body DNA methylation in plants. *Current Opinion in*
630 *Plant Biology* **36**: 103–110.

631 **Brownlee C, Langer G, Wheeler GL**. **2021**. Coccolithophore calcification: Changing
632 paradigms in changing oceans. *Acta Biomaterialia* **120**: 4–11.

633 **Charlson RJ, Lovelock JE, Andreae MO, Warren SG**. **1987**. Oceanic phytoplankton,
634 atmospheric sulphur, cloud albedo and climate. *Nature* **326**: 655–661.

635 **Chen C-W, Lin P-Y, Lai Y-M, Lin M-H, Lin S-Y, Hsu C-C**. **2024**. TIMAHAC: streamlined
636 tandem IMAC-HILIC workflow for simultaneous and high-throughput plant
637 phosphoproteomics and N-glycoproteomics. *Molecular & Cellular Proteomics* **23**: 100762.

638 **Chen C-W, Tsai C-F, Lin M-H, Lin S-Y, Hsu C-C**. **2023**. Suspension trapping-based sample
639 preparation workflow for in-depth plant phosphoproteomics. *Analytical Chemistry* **95**:
640 12232–12239.
641 **Cock JM, Sterck L, Rouzé P, Scornet D, Allen AE, Amoutzias G, Anthouard V,**
642 **Artiguenave F, Aury J-M, Badger JH, *et al.* 2010**. The *Ectocarpus* genome and the
643 independent evolution of multicellularity in brown algae. *Nature* **465**: 617–621.
644 **Coombe L, Zhang J, Vandervalk BP, Chu J, Jackman SD, Birol I, Warren RL**. **2018**.
645 ARKS: chromosome-scale scaffolding of human genome drafts with linked read kmers.
646 *BMC Bioinformatics* **19**: 234.
647 **von Dassow P, Ogata H, Probert I, Wincker P, Da Silva C, Audic S, Claverie J-M, de**
648 **Vargas C**. **2009**. Transcriptome analysis of functional differentiation between haploid and
649 diploid cells of *Emiliania huxleyi*, a globally significant photosynthetic calcifying cell.
650 *Genome Biology* **10**: R114.
651 **Demichev V, Messner CB, Vernardis SI, Lilley KS, Ralser M**. **2020**. DIA-NN: neural
652 networks and interference correction enable deep proteome coverage in high throughput.
653 *Nature Methods* **17**: 41–44.
654 **Emms DM, Kelly S**. **2019**. OrthoFinder: phylogenetic orthology inference for comparative
655 genomics. *Genome Biology* **20**: 238.
656 **Fan X, Han W, Teng L, Jiang P, Zhang X, Xu D, Li C, Pellegrini M, Wu C, Wang Y, *et al.***
657 **2020**. Single-base methylome profiling of the giant kelp Saccharina japonica reveals
658 significant differences in DNA methylation to microalgae and plants. *New Phytologist* **225**:
659 234–249.
660 **Farrell C, Thompson M, Tosevska A, Oyetunde A, Pellegrini M**. **2021**. BiSulfite Bolt: A
661 bisulfite sequencing analysis platform. *GigaScience* **10**: giab033.
662 **Filatov DA, Bendif EM, Archontikis OA, Hagino K, Rickaby REM**. **2021**. The mode of
663 speciation during a recent radiation in open-ocean phytoplankton. *Current Biology* **31**:
664 5439-5449.e5.
665 **Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF**. **2020**.
666 RepeatModeler2 for automated genomic discovery of transposable element families.
667 *Proceedings of the National Academy of Sciences* **117**: 9451–9457.

668 **Frada MJ, Bendif EM, Keuter S, Probert I**. **2019**. The private life of coccolithophores.
669 *Perspectives in Phycology* **6**: 11–30.
670 **Frada M, Probert I, Allen MJ, Wilson WH, de Vargas C**. **2008**. The 'Cheshire Cat' escape
671 strategy of the coccolithophore *Emiliania huxleyi* in response to viral infection. *Proceedings*
672 *of the National Academy of Sciences* **105**: 15944–15949.
673 **Frada MJ, Rosenwasser S, Ben-Dor S, Shemi A, Sabanay H, Vardi A**. **2017**. Morphological
674 switch to a resistant subpopulation in response to viral infection in the bloom-forming
675 coccolithophore *Emiliania huxleyi* (JS Weitz, Ed.). *PLOS Pathogens* **13**: e1006775.
676 **Fulnečková J, Ševčíková T, Fajkus J, Lukešová A, Lukeš M, Vlček Č, Lang BF, Kim E,**
677 **Eliáš M, Sýkorová E**. **2013**. A broad phylogenetic survey unveils the diversity and
678 evolution of telomeres in eukaryotes. *Genome Biology and Evolution* **5**: 468–483.
679 **Gabriel L, Hoff KJ, Brůna T, Borodovsky M, Stanke M**. **2021**. TSEBRA: transcript selector
680 for BRAKER. *BMC Bioinformatics* **22**: 566.
681 **Gal A, Wirth R, Kopka J, Fratzl P, Faivre D, Scheffel A**. **2016**. Macromolecular recognition
682 directs calcium ions to coccolith mineralization sites. *Science* **353**: 590–593.
683 **Garrison E, Marth G**. **2012**. Haplotype-based variant detection from short-read sequencing.
684 *arXiv preprint arXiv:1207.3907*.
685 **Goodenough U, Heitman J**. **2014**. Origins of eukaryotic sexual reproduction. *Cold Spring*
686 *Harbor Perspectives in Biology* **6**: a016154–a016154.
687 **Hansen KD, Langmead B, Irizarry RA**. **2012**. BSmooth: from whole genome bisulfite
688 sequencing reads to differentially methylated regions. *Genome Biology* **13**: R83.
689 **Hariskos I, Rubner T, Posten C**. **2015**. Investigation of cell growth and chlorophyll a content
690 of the coccolithophorid alga *Emiliania huxleyi* by using simple bench-top flow cytometry.
691 *Journal of Bioprocessing & Biotechniques* **05**: 1000234.
692 **Hart AJ, Ginzburg S, Xu M (Sam), Fisher CR, Rahmatpour N, Mitton JB, Paul R,**
693 **Wegrzyn JL**. **2020**. <scp>EnTAP</scp> : Bringing faster and smarter functional annotation
694 to non-model eukaryotic transcriptomes. *Molecular Ecology Resources* **20**: 591–604.
695 **Heidenreich E, Wördenweber R, Kirschhöfer F, Nusser M, Friedrich F, Fahl K, Kruse O,**
696 **Rost B, Franzreb M, Brenner-Weiß G, *et al.* 2019**. Ocean acidification has little effect on
697 the biochemical composition of the coccolithophore *Emiliania huxleyi* (A Ianora, Ed.).
698 *PLOS ONE* **14**: e0218564.

699 **Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M**. **2016**. BRAKER1: Unsupervised
700 RNA-seq-based genome annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics*
701 **32**: 767–769.
702 **Houdan A, Billard C, Marie D, Not F, Sáez AG, Young JR, Probert I**. **2004**.
703 Holococcolithophore-heterococcolithophore (Haptophyta) life cycles: Flow cytometric
704 analysis of relative ploidy levels. *Systematics and Biodiversity* **1**: 453–465.
705 **Houdan A, Probert I, Van Lenning K, Lefebvre S**. **2005**. Comparison of photosynthetic
706 responses in diploid and haploid life-cycle phases of *Emiliania huxleyi* (Prymnesiophyceae).
707 *Marine Ecology Progress Series* **292**: 139–146.
708 **Hu J, Fan J, Sun Z, Liu S**. **2020**. NextPolish: a fast and efficient genome polishing tool for
709 long-read assembly (B Berger, Ed.). *Bioinformatics* **36**: 2253–2255.
710 **Hu J, Wang Z, Sun Z, Hu B, Ayoola AO, Liang F, Li J, Sandoval JR, Cooper DN, Ye K, *et***
711 ***al.* 2023**. An efficient error correction and accurate assembly tool for noisy long reads.
712 *bioRxiv*: 2023.03.09.531669.
713 **Kao T, Wang T, Ku C**. **2022**. Rampant nuclear–mitochondrial–plastid phylogenomic
714 discordance in globally distributed calcifying microalgae. *New Phytologist* **235**: 1394–1408.
715 **Kawashima T, Berger F**. **2014**. Epigenetic reprogramming in plant sexual reproduction. *Nature*
716 *Reviews Genetics* **15**: 613–624.
717 **Lee JM, Song HJ, Park SI, Lee YM, Jeong SY, Cho TO, Kim JH, Choi H-G, Choi CG,**
718 **Nelson WA, *et al.* 2018**. Mitochondrial and plastid genomes from coralline red algae
719 provide insights into the incongruent evolutionary histories of organelles (J Archibald, Ed.).
720 *Genome Biology and Evolution* **10**: 2961–2972.
721 **Li H**. **2018**. Minimap2: pairwise alignment for nucleotide sequences (I Birol, Ed.).
722 *Bioinformatics* **34**: 3094–3100.
723 **Lopez DA, Hamaji T, Kropat J, De Hoff P, Morselli M, Rubbi L, Fitz-Gibbon ST, Gallaher**
724 **SD, Merchant SS, Umen JG, *et al.* 2015**. Dynamic changes in the transcriptome and
725 methylome of *Chlamydomonas reinhardtii* throughout its life cycle. *Plant Physiology* **169**:
726 2730–2743.
727 **Mackinder L, Wheeler G, Schroeder D, von Dassow P, Riebesell U, Brownlee C**. **2011**.
728 Expression of biomineralization-related ion transport genes in *Emiliania huxleyi*.
729 *Environmental Microbiology* **13**: 3250–3265.

730 **Marsh ME, Ridall AL, Azadi P, Duke PJ**. **2002**. Galacturonomannan and Golgi-derived
731 membrane linked to growth and shaping of biogenic calcite. *Journal of Structural Biology*
732 **139**: 39–45.
733 **Marzec B, Walker JM, Panagopoulou M, Jhons Y, Clare D, Wheeler A, Shaver MP,**
734 **Nudelman F**. **2019**. Three-dimensional architecture and surface functionality of coccolith
735 base plates. *Journal of Structural Biology* **208**: 127–136.
736 **McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K,**
737 **Altshuler D, Gabriel S, Daly M, *et al.* 2010**. The Genome Analysis Toolkit: A MapReduce
738 framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**:
739 1297–1303.
740 **Nelson CW, Moncla LH, Hughes AL**. **2015**. SNPGenie: estimating evolutionary parameters to
741 detect natural selection using pooled next-generation sequencing data. *Bioinformatics*:
742 btv449.
743 **Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C**. **2017**. Salmon provides fast and bias-
744 aware quantification of transcript expression. *Nature Methods* **14**: 417–419.
745 **Patterson M, Marschall T, Pisanti N, Van Iersel L, Stougie L, Klau GW, Schönhuth A**.
746 **2015**. WhatsHap: weighted haplotype assembly for future-generation sequencing reads.
747 *Journal of Computational Biology* **22**: 498–509.
748 **Read BA, Kegel J, Klute MJ, Kuo A, Lefebvre SC, Maumus F, Mayer C, Miller J, Monier**
749 **A, Salamov A, *et al.* 2013**. Pan genome of the phytoplankton *Emiliania* underpins its global
750 distribution. *Nature* **499**: 209–213.
751 **Rousseaux C, Gregg W**. **2013**. Interannual variation in phytoplankton primary production at a
752 global scale. *Remote Sensing* **6**: 1–19.
753 **Ruan Z, Lu M, Lin H, Chen S, Li P, Chen W, Xu H, Qiu D**. **2023**. Different photosynthetic
754 responses of haploid and diploid *Emiliania huxleyi* (Prymnesiophyceae) to high light and
755 ultraviolet radiation. *Bioresources and Bioprocessing* **10**: 40.
756 **Sandra K. Floyd, John L. Bowman**. **2007**. The ancestral developmental tool kit of land plants.
757 **Skeffington A, Fischer A, Sviben S, Brzezinka M, Górka M, Bertinetti L, Woehle C,**
758 **Huettel B, Graf A, Scheffel A**. **2023**. A joint proteomic and genomic investigation provides
759 insights into the mechanism of calcification in coccolithophores. *Nature Communications*
760 **14**: 3749.

761 **Smit AFA, Hubley R, Green P**. **2015**. RepeatMasker Open-4.0. 2013--2015.
762 **Soneson C, Love MI, Robinson MD**. **2015**. Differential analyses for RNA-seq: transcript-level
763 estimates improve gene-level inferences. *F1000Research* **4**: 1521.
764 **Talbert PB, Meers MP, Henikoff S**. **2019**. Old cogs, new tricks: the evolution of gene
765 expression in a chromatin context. *Nature Reviews Genetics* **20**: 283–297.
766 **Tang S, Lomsadze A, Borodovsky M**. **2015**. Identification of protein coding regions in RNA
767 transcripts. *Nucleic Acids Research* **43**: e78–e78.
768 **Taylor AR, Brownlee C, Wheeler G**. **2017**. Coccolithophore cell biology: chalking up
769 progress. *Annual Review of Marine Science* **9**: 283–310.
770 **Teng C-S, Wu B-H, Yen M-R, Chen P-Y**. **2020**. MethGET: web-based bioinformatics software
771 for correlating genome-wide DNA methylation and gene expression. *BMC genomics* **21**: 1–
772 10.
773 **Traller JC, Cokus SJ, Lopez DA, Gaidarenko O, Smith SR, McCrow JP, Gallaher SD,**
774 **Podell S, Thompson M, Cook O, *et al.* 2016**. Genome and methylome of the oleaginous
775 diatom *Cyclotella cryptica* reveal genetic flexibility toward a high lipid phenotype.
776 *Biotechnology for Biofuels* **9**: 258.
777 **Tree of Life Kit**. **2022**. A Telomere Identification toolKit (tidk). *URL*:
778 https://github.com/tolkit/telomeric-identifier.
779 **Umen J, Coelho S**. **2019**. Algal sex determination and the evolution of anisogamy. *Annual*
780 *Review of Microbiology* **73**: 267–291.
781 **De Vargas C, Aubry M-P, Probert I, Young J**. **2007**. Origin and evolution of
782 coccolithophores: from coastal hunters to oceanic farmers. In: Evolution of Primary
783 Producers in the Sea. Elsevier, 251–285.
784 **Veluchamy A, Lin X, Maumus F, Rivarola M, Bhavsar J, Creasy T, O'Brien K,**
785 **Sengamalay NA, Tallon LJ, Smith AD, *et al.* 2013**. Insights into the role of DNA
786 methylation in diatoms by genome-wide profiling in *Phaeodactylum tricornutum*. *Nature*
787 *Communications* **4**: 2091.
788 **Vigneau J, Borg M**. **2021**. The epigenetic origin of life history transitions in plants and algae.
789 *Plant Reproduction* **34**: 267–285.

790 **de Vries J, Monteiro F, Wheeler G, Poulton A, Godrijan J, Cerino F, Malinverno E,**
791 **Langer G, Brownlee C**. **2021**. Haplo-diplontic life cycle expands coccolithophore niche.
792 *Biogeosciences* **18**: 1161–1184.
793 **Wang G, Oh D-H, Dassanayake M**. **2020**. GOMCL: a toolkit to cluster, evaluate, and extract
794 non-redundant associations of Gene Ontology-based functions. *BMC Bioinformatics* **21**:
795 139.
796 **Wheeler GL, Sturm D, Langer G**. **2023**. *Gephyrocapsa huxleyi* (*Emiliania huxleyi*) as a model
797 system for coccolithophore biology. *Journal of Phycology* **00**: 1–7.
798 **Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, Feng T, Zhou L, Tang W, Zhan L, *et al.* 2021**.
799 ClusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation*
800 **2**: 100141.
801 **Wu EY, Singh NP, Choi K, Zakeri M, Vincent M, Churchill GA, Ackert-Bicknell CL,**
802 **Patro R, Love MI**. **2023**. SEESAW: detecting isoform-level allelic imbalance accounting
803 for inferential uncertainty. *Genome Biology* **24**: 165.
804 **Wu H, Wang C, Wu Z**. **2013**. A new shrinkage estimator for dispersion improves differential
805 expression detection in RNA-seq data. *Biostatistics* **14**: 232–243.
806 **Xu M, Guo L, Gu S, Wang O, Zhang R, Peters BA, Fan G, Liu X, Xu X, Deng L, *et al.***
807 **2020**. TGS-GapCloser: a fast and accurate gap closer for large genomes with low coverage
808 of error-prone long reads. *GigaScience* **9**: giaa094.
809 **Yang P, Guo K, Yang Y, Lyu M, Liu J, Li X, Feng Y**. **2023**. Phylogeny and genetic variations
810 of the three genome compartments in haptophytes shed light on the rapid evolution of
811 coccolithophores. *Gene* **887**: 147716.
812 **Yeo S, Coombe L, Warren RL, Chu J, Birol I**. **2018**. ARCS: scaffolding genome drafts with
813 linked reads (C Sahinalp, Ed.). *Bioinformatics* **34**: 725–731.
814 **Yun G, Park J-M, Duong V-A, Mok J-H, Jeon J, Nam O, Lee J, Jin E, Lee H**. **2020**.
815 Proteomic profiling of *Emiliania huxleyi* using a three-dimensional separation method
816 combined with tandem mass spectrometry. *Molecules* **25**: 3028.
817 **Zhu Y, Orre LM, Zhou Tran Y, Mermelekas G, Johansson HJ, Malyutina A, Anders S,**
818 **Lehtiö J**. **2020**. DEqMS: a method for accurate variance estimation in differential protein
819 expression analysis. *Molecular & Cellular Proteomics* **19**: 1047–1057.
820

Figures and Tables

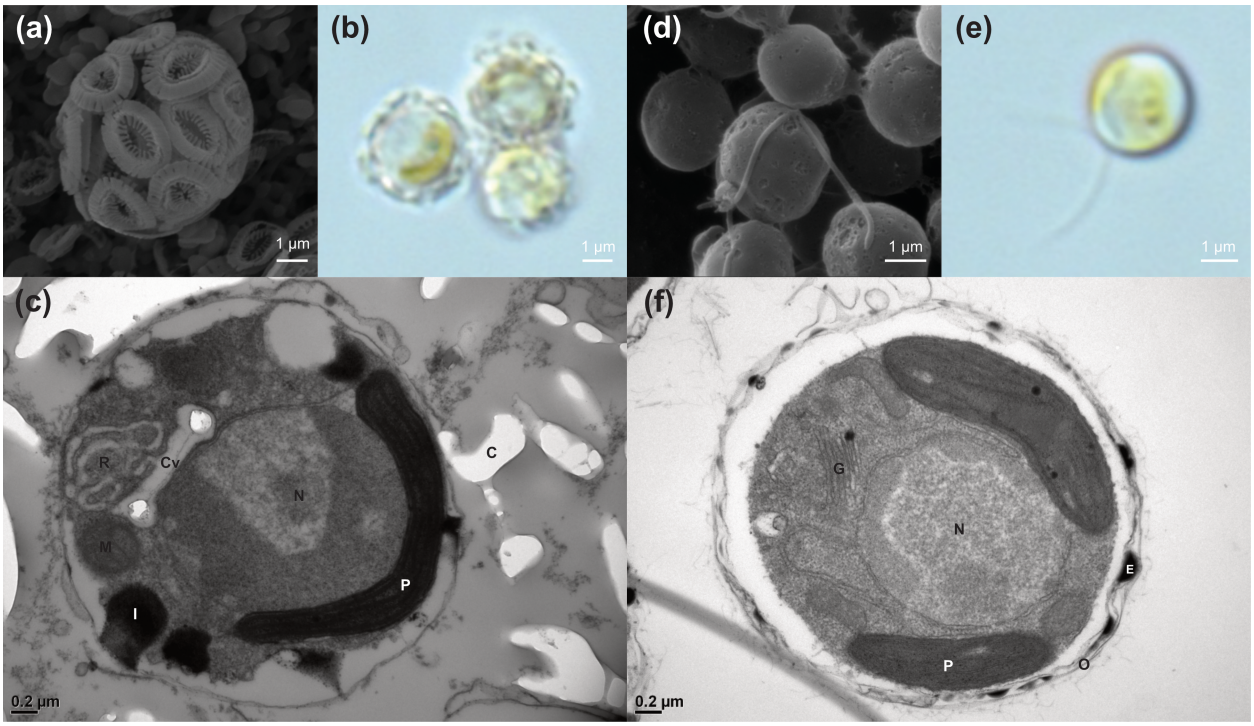
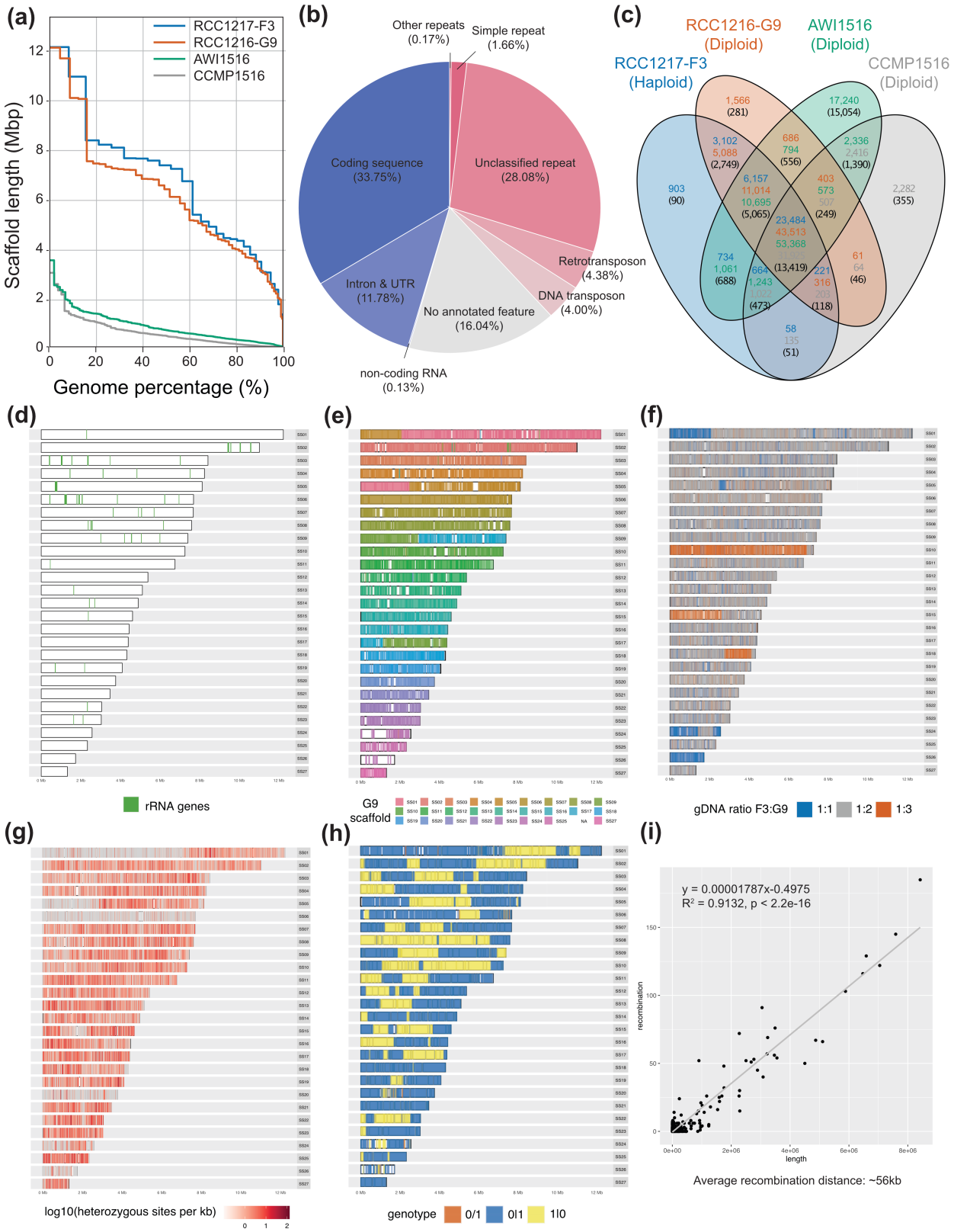


Fig. 1. Morphology of *Gephyrocapsa huxleyi* RCC1216-G9 (diploid) and RCC1217-F3 (haploid). Scanning electron microscopy (a and d), differential interference contrast (b and e), and transmission electron microscopy (c and f) images of G9 and F3. C: empty space created by a fallen coccolith (calcite scale); Cv: coccolith vesicle; E: electron-dense particle; G: Golgi body; I: ion-dense vacuole; M: mitochondrion; N: nucleus; O: organic scale; P: plastid; R: reticular body.


(a)
RCC1217-F3
RCC1216-G9
AWI1516
CCMP1516
Scaffold length (Mbp)
Genome percentage (%)
(b)
Other repeats (0.17%)
Simple repeat (1.66%)
Coding sequence (33.75%)
Unclassified repeat (28.08%)
Retrotransposon (4.38%)
DNA transposon (4.00%)
Intron & UTR (11.78%)
No annotated feature (16.04%)
non-coding RNA (0.13%)
(c)
RCC1216-G9 (Diploid)
AWI1516 (Diploid)
RCC1217-F3 (Haploid)
CCMP1516 (Diploid)
(d)
rRNA genes
(e)
G9 scaffold
(f)
gDNA ratio F3:G9 1:1 1:2 1:3
(g)
log10(heterozygous sites per kb)
(h)
genotype 0/1 0|1 1|0
(i)
y = 0.00001787x-0.4975
R² = 0.9132, p < 2.2e-16
recombination
length
Average recombination distance: ~56kb

11

12 **Fig. 2.** Genome assemblies, genomic features, structure variations, heterozygosity, and
13 recombination rate. (a) N statistics (scaffold length at different percentages of genome assembly)
14 of RCC1217-F3 and RCC1216-G9 (*G. huxleyi*) compared with CCMP1516 and AWI1516 (*G.*
15 *pseudohuxleyi*). (b) Relative length percentage of coding and non-coding sequences of genes
16 (blues) and different classes of repeats (reds) in F3. (c) Numbers of genes and orthogroups (in
17 parenthesis) unique to or shared among the strains. (d) Distribution of rDNAs in F3. (e) Mapping
18 of G9 superscaffolds to the F3 genome, revealing interchromosomal recombinations and large-
19 scale indels. White: unmappable regions. (f) Distribution of gene copy number variations across
20 the genome, showing partial aneuploidy. (g) Frequency of G9 heterozygous sites across the
21 genome, with lower heterozygosity in certain genomic regions. (h) Genomic regions of F3 that
22 have been phased into different G9 alleles (blue and yellow). Orange: un-phased regions in
23 phasing blocks. White: regions outside of phasing blocks. (i) Correlation between the length of
24 phasing blocks and the number of recombination sites.
25

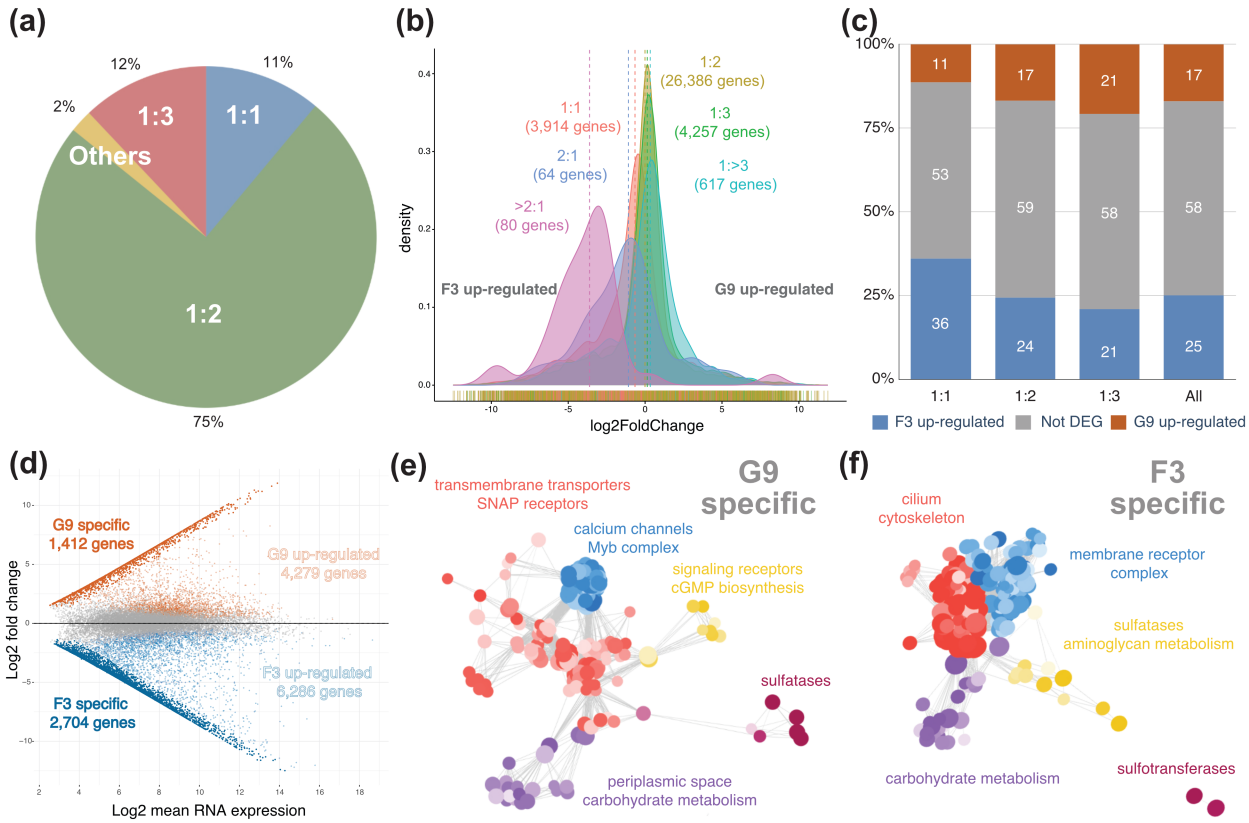


Fig. 3. Gene copy number variation and transcriptomic differentiation. (a) Relative abundance of genes with different G9:F3 genomic DNA (gDNA) ratios. (b) Distribution patterns of RNA expression ratios for gene groups with different gDNA ratios. (c) Percentage of up- and down-regulated genes across different gDNA ratios. (d) MA plot of RNA expression, highlighting genes specifically expressed or up-regulated in each life cycle phases. (e) Clusters of enriched GO terms of diploid-specific transcripts. Each dot represents an enriched term, with darker colors indicating smaller p-values. Dots with similarity > 0.5 are linked by a grey line. Clusters with ≤ 5 genes were excluded. (f) Clusters of enriched GO terms of haploid-specific transcripts.

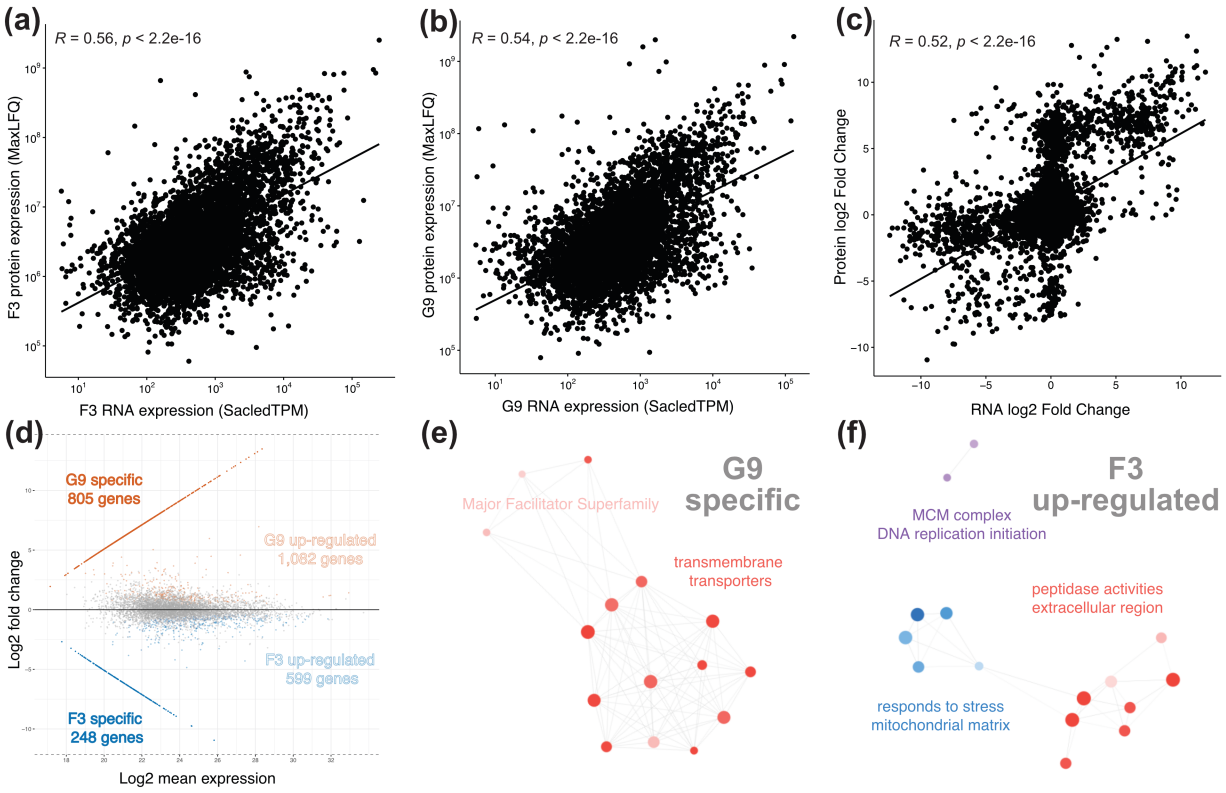


Fig. 4. Protein-RNA correlations and proteomic differentiation. (a, b) Correlation between RNA and protein expression in F3 (a) and G9 (b). (c) Correlation between RNA and protein $\log_2$ fold change between F3 and G9. (d) MA plot of protein expression, highlighting genes specifically expressed or up-regulated in each life cycle phases. (e) Clusters of enriched GO terms of diploid-specific proteins. Each dot represents an enriched term, with darker colors indicating smaller p-values. Dots with similarity > 0.5 are linked by a grey line. Clusters with ≤ 5 genes were excluded. (f) Clusters of enriched GO terms of up-regulated proteins in the haploid.

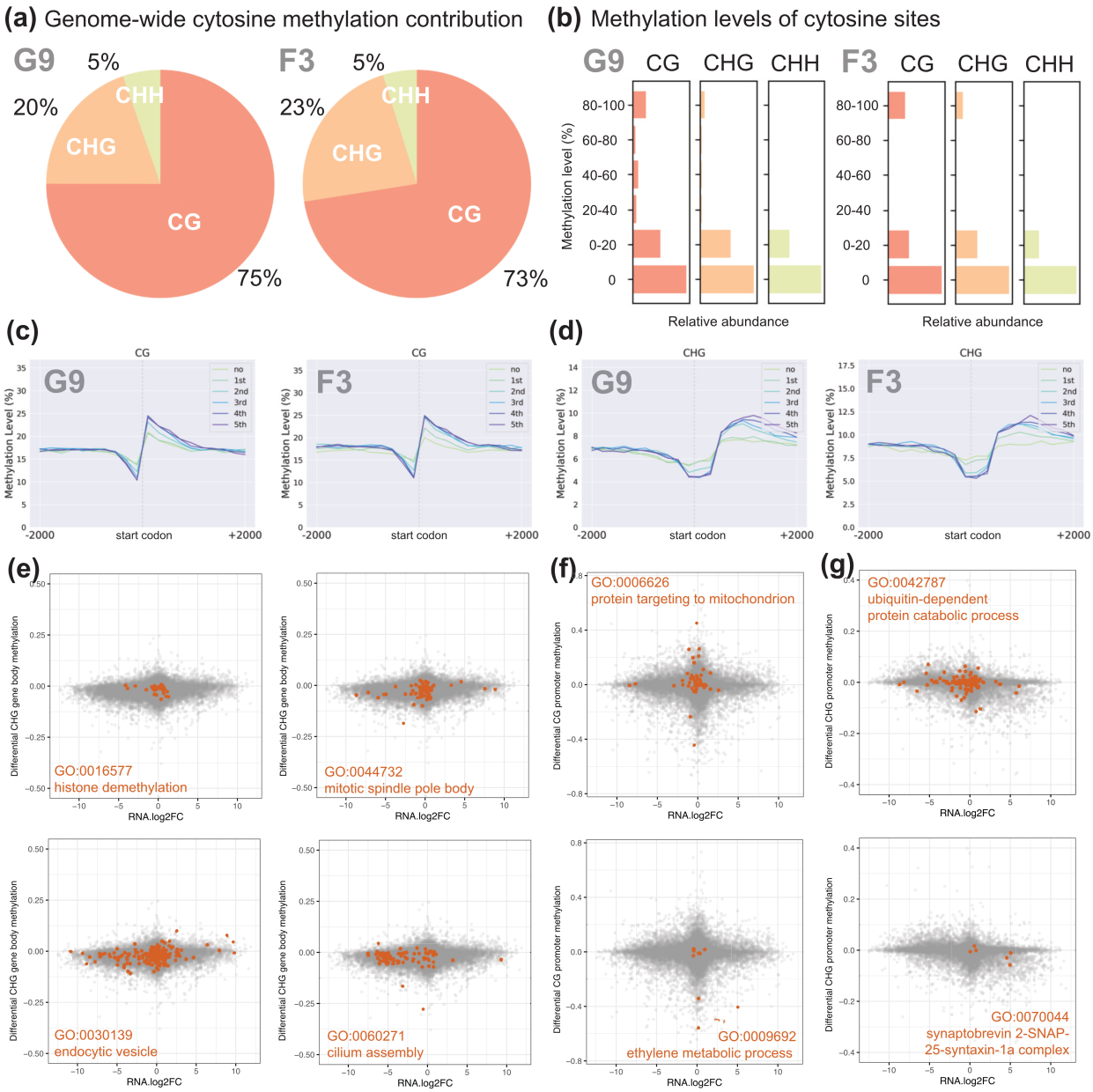


Fig. 5. Genome-wide methylcytosine (mC) patterns, levels, and correlations with RNA expression. (a) Percentage of mC contribution in different sequence contexts: CG, CHG, and CHH. (b) Relative abundance of sites with different mC levels in CG, CHG, and CHH sequence contexts. (c, d) Average mC level profiles (CG [c] or CHG [d] context) around genes in six categories of RNA expression (5th: highest level). (e) Selected enriched GO terms among genes with higher RNA and higher gene-body mC levels in one life cycle phase (CHG context). (f, g) Selected enriched GO terms among genes with higher RNA and lower promoter-region mC levels in one life cycle phase (CG [f] or CHG [g] context).

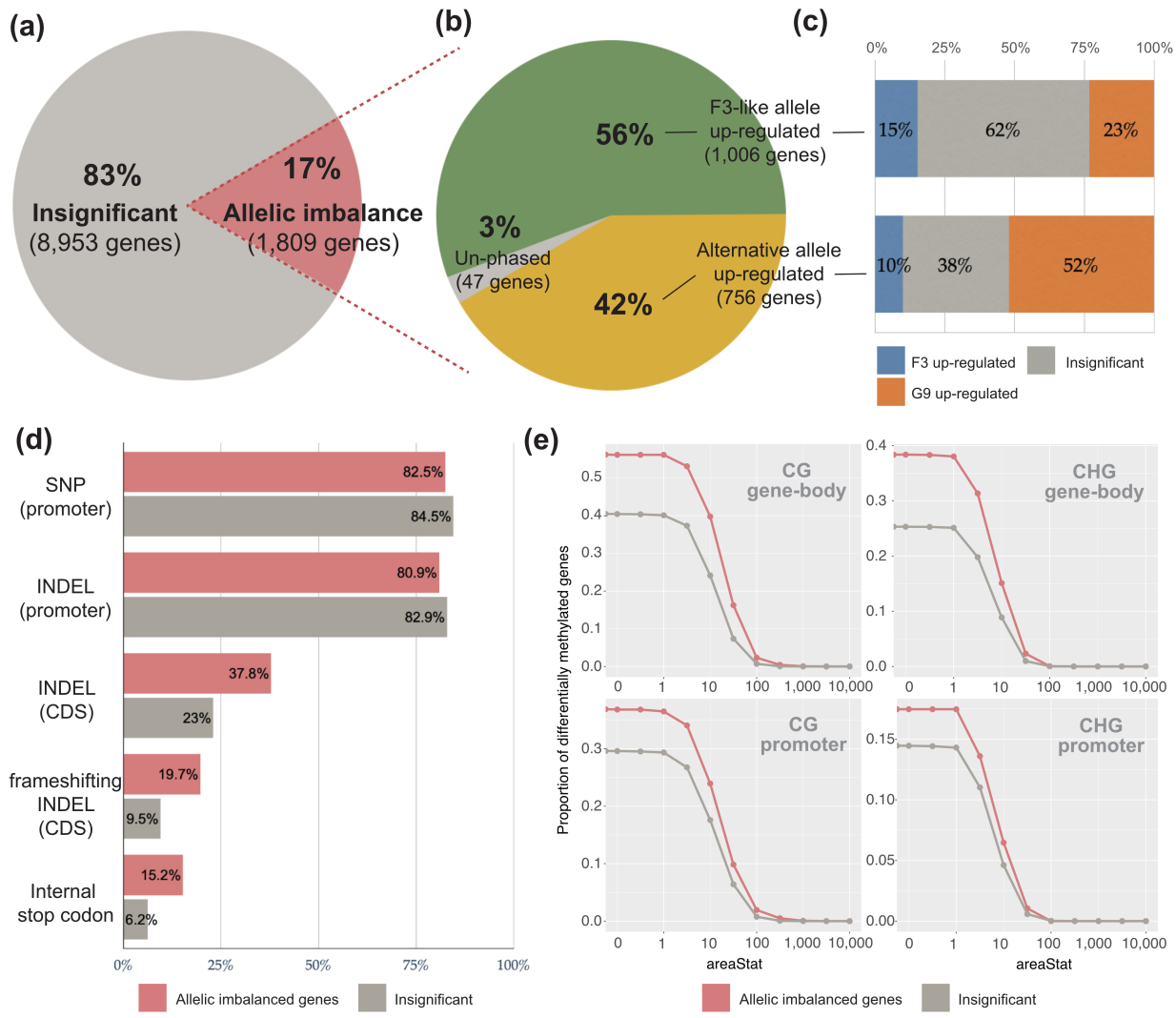


Fig. 6. Allelic imbalance in RNA expression and its relationships with genomic variation and allele-specific methylation. (a) Percentage of allelic imbalanced genes among the heterozygous genes with exactly two copies in the G9 genome and with detectable transcripts. (b) Percentage of genes with up-regulated F3-like or alternative alleles among the allelic imbalanced genes. (c) The percentage of F3 up-regulated genes and G9 up-regulated genes among the genes with up-regulated F3-like or alternative alleles. (d) Percentage of genes with and without significant allelic imbalanced expression that contain single-nucleotide-polymorphisms (SNPs), insertion-deletions (INDELs), frameshifts, or internal stop codons. (e) Proportion of genes with and without significant allelic imbalance that have allele-specific methylated regions. AreaStat: the sum of t-statistics for the sites within the region.

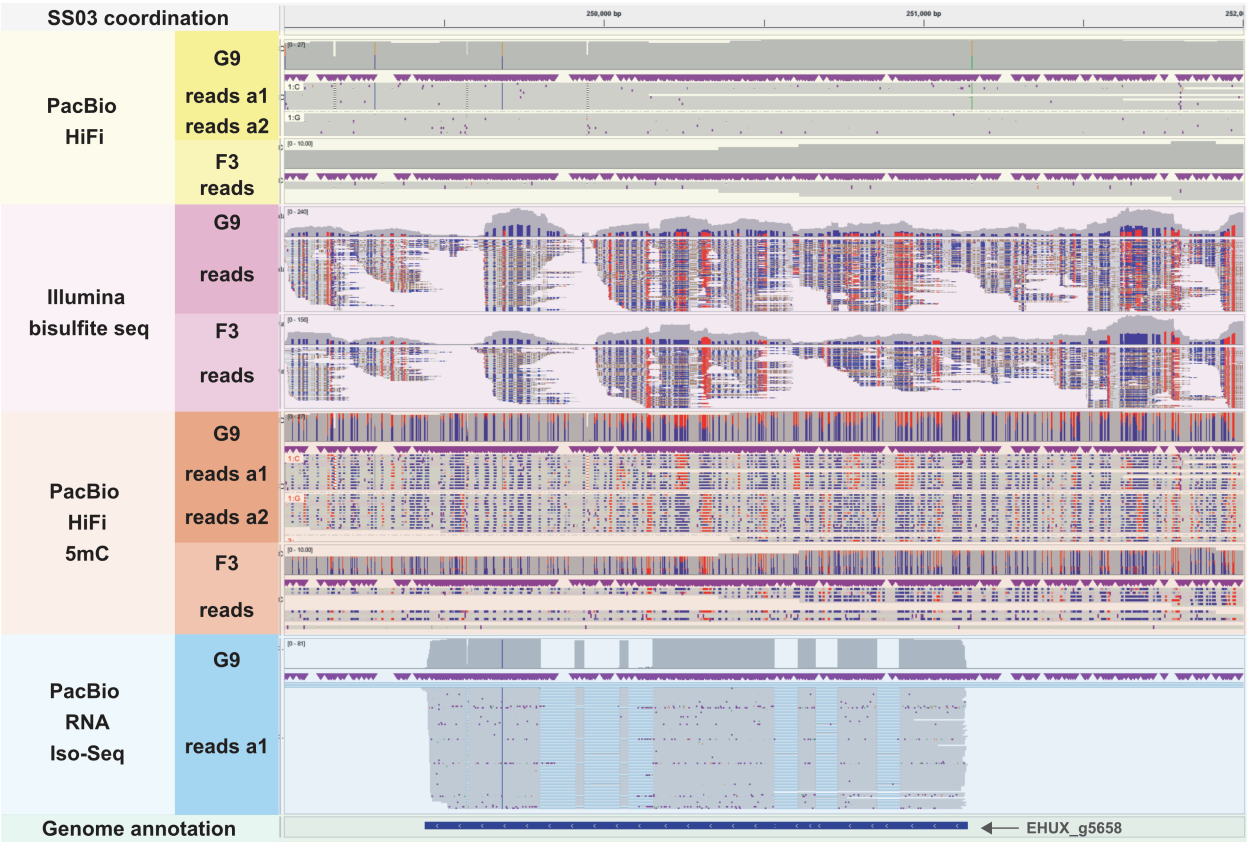


Fig. 7. Allele-specific methylation and allele-imbalanced RNA expression in a *G. huxleyi* gene. Genomic, methylomic, and transcriptomic reads were mapped onto SS03 (249,000–252,000) containing EHUX_g5658, a sulfatase gene located on the Crick strand. Track yellow (top): PacBio HiFi reads (variable sites: A [green], C [blue], G [yellow], T [red], insertions [purple vertical bars], and deletions [black dashes]). Track pink: Illumina bisulfite sequencing reads (5mC [red], C [blue], or antisense [methylation status unknown; grey]). Track orange: PacBio HiFi reads with methylation calling (5mC probability > 50% [red] or 5mC probability < 50% [blue]). Track blue: PacBio Iso-Seq reads (transcript [grey] or intron [blue]). Track green: gene annotation.

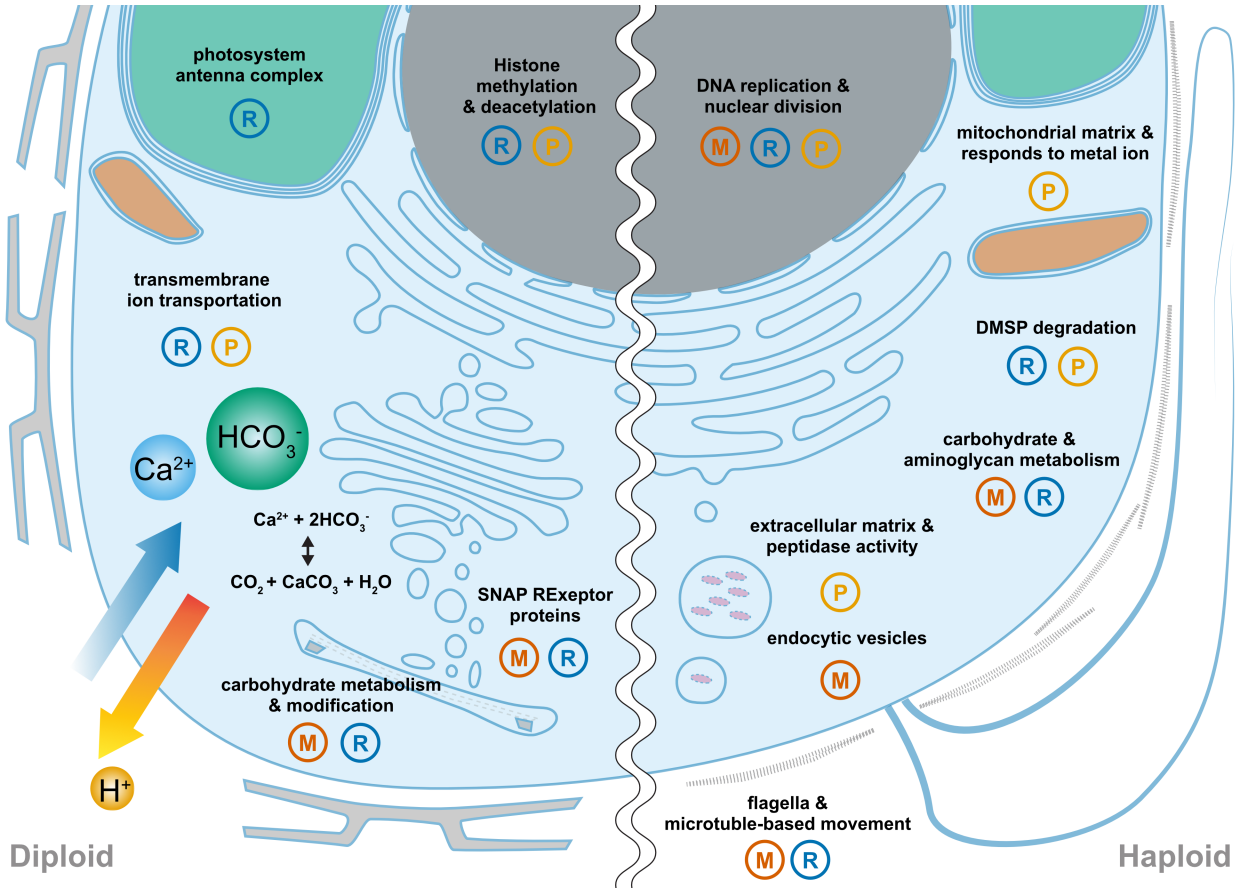


Fig. 8. Summary of molecular differentiation between the diploid (left) and haploid (right) cell. The types of data supporting each functional category are indicated. M: DNA cytosine methylation; R: mRNA expression; P: protein expression.

82 **Table 1. Statistics of the haploid and diploid genome assemblies of *Gephyrocapsa* spp.**

83

	CCMP1516	AWI1516	RCC1217-F3	RCC1216-G9
Species	*G. pseudohuxleyi*	*G. pseudohuxleyi*	*G. huxleyi*	*G. huxleyi*
Ploidy	Diploid	Diploid	Haploid	Diploid
Genome size	167.68 Mb	195.82 Mb	150.96 Mb	279.23 Mb
GC content	65.67 %	66.39%	65.99 %	66.04 %
# Scaffold	7795	600	27	54
# Contigs	16921	600	27	65
Gap	7.01 %	0.00 %	0.00 %	0.19 %
All repeats	44.4 %	35.82 %	47.22 %	47.45 %
Retroelements	1.11 %	1.83 %	4.83 %	4.81 %
DNA transposons	2.98 %	4.94 %	4.25 %	4.56 %
Unclassified repeats	30.94 %	27.27 %	33.09 %	32.18 %
# Genes	38549	87310	35323	62647
# Transcripts	38559	129544	54163	76986
Reference	Read et al., 2013	Skeffington et al., 2023	This study	This study