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

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

Affiliations

[1] Institute of Plant and Microbial Biology, Academia Sinica, Taipei, Taiwan.
[2] The Alexander Silberman Institute of Life Sciences, The Hebrew University of Jerusalem, Israel.
[3] The Interuniversity Institute for Marine Sciences in Eilat, Eilat, Israel.
[†] Present address: Department of Biological Sciences, National Sun Yat-sen University, Kaohsiung, Taiwan.
* Corresponding author. Email: chuanku@gate.sinica.edu.tw.

Abstract

Gephyrocapsa 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 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.

Introduction

Sexual reproduction and transitions between life-cycle phases are hallmarks of eukaryotic evolution and a major mechanism for generating genomic variations in eukaryotes (*1*). 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 (*1*, *2*). 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 (*3*). The differentiation in morphological and physiological attributes between life-cycle phases in haplodiplontic life cycles expands the functional scope of these eukaryotes, where diploid and haploid organisms can explore different niches and display distinct life strategies (*4*, *5*).

Coccolithophores are haplodiplontic, calcifying microalgae that are highly abundant and ecologically significant across global oceans (*4–7*). They contribute around 10% to global primary production as a major drive of the biological carbon pump (*8*). They also synthesize dimethylsulfoniopropionate (DMSP), a precursor of volatile dimethylsulfide, which may influence climate through cloud formation (*9–11*). Most notably, coccolithophores fix carbon into $CaCO_3$ through calcification, forming calcite plates (coccoliths) on the cell surface. Calcification not only releases CO_2 inside the cell but contributes to carbon removal from the surface ocean with heavy coccoliths serving as ballast enhancing carbon export to the deep ocean (*6*). The life cycle of coccolithophores alternates between a diploid and haploid phase with distinct morphological and ecological characteristics. *Gephyrocapsa huxleyi* (*sensu* Bendif et al. 2019) (*12*); more commonly known as *Emiliania huxleyi*) is the most abundant and widespread species in the modern ocean and a model in coccolithophore research (*13*). Interestingly, for this species, coccolith formation is limited to the diploid phase. Diploid cells regularly produce heterococcoliths, while by contrast the haploid phase is not calcified and exclusively generates unmineralized organic scales that coat the cell surface and that are absent in the diploid cell (Fig. 1). In addition, diploid cells are non-motile and haploid cells biflagellate, diploid cells appear less sensitive to high light conditions (*14*), and strikingly, haploid cells are resistant to specific lytic viruses that infect and kill diploid cells (*15*).

Life cycle phase transition of *G. huxleyi* is poorly understood. A pair of diploid and haploid strains, RCC1216 and RCC1217, were identified and isolated from a diploid lab culture in 1999 (*16*) following a spontaneous meiosis event. A major limitation in past research has been the lack of a reference genome for the diploid-haploid pair. Previous genome sequences of *Gephyrocapsa* are derived from a diploid strain CCMP1516 (*17*) or the closely diploid AWI1516 (*18*). Despite the very recent origin of *G. huxleyi* ~0.3 million years ago (*12*), its genomic structure displays substantial variability, with haploid genome size estimates ranging from 99 to 168 Mb and only ~70% of genes shared across different strains (*17*). Recent phylogenomic analyses further suggest that RCC1216/RCC1217 and CCMP1516/AWI1516 belong to two distinct clades (*19*) or even species (*G. huxleyi* sensu stricto and *G. pseudohuxleyi*, respectively) (*20*) with distinct morphology, biogeographic range, and plastid genomes (*19*). To better understand coccolithophore diploid-haploid differentiation at the molecular level, here we study RCC1216 and RCC1217 as a model system using a multiomics approach. Specifically, we aim to provide a comprehensive view of diploid-haploid differentiation at the molecular level and novel insight into its potential regulatory mechanisms.

Results

Genome assemblies and annotations

The *G. huxleyi* strains RCC1216 and RCC1217 from the Roscoff Culture Collection (RCC) were originally obtained in 2012. Single cells from both strains were isolated to establish single-cell-derived cultures in 2019, namely RCC1217-F3 (F3) and RCC1216-G9 (G9). Their genomes were *de novo* assembled by integrating short reads (Illumina Inc., San Diego, CA, USA), long reads (Pacific Biosciences, Menlo Park, CA, USA; Oxford Nanopore Technologies, Oxford, UK), and optical maps of BspQI restriction sites (GCTCTTC) (BioNano Genomics, San Diego, CA, USA). These assemblies exhibit substantially improved genome contiguity compared with other assemblies of *Gephyrocapsa* (Fig. 2A). The complete genome assemblies, sequencing statistics, and read archive accessions can be accessed under NCBI BioProject PRJNA774110 (Table S1).

The genome assembly of F3 is 151 Mb, comprised of 26+1 superscaffolds (SS), where SS26 is a partial duplicate of SS24 and not present in the diploid G9. The *G. huxleyi* genomes display a high GC content (~66%) and contain a significant proportion of repetitive sequences (~36% after excluding the sequences overlapped with genes; Fig. 2B and Table 1). The telomeric repeat sequence TTAGGG, commonly found in animals and other eukaryotes (*21*) was detected in some superscaffolds but not all. The coding genes annotated utilizing both PacBio Iso-Seq full-length transcripts and Illumina RNA-seq reads amount to 35,323 in F3, with 4,005 (11%) of them absent in the CCMP1516 assembly (RefSeq GCF_000372725.1) (*17*) or its still calcifying variant AWI1516 (*18*) (Fig. 2C). A total of 827 non-coding RNA genes were identified, including 349 rRNA genes that are mostly located on SS02, SS03, SS05, and SS06 (Fig. 2D).

Genome variations and phasing

F3 and G9 genomes exhibit notable structural divergence (Fig. S1), including two chromosome-level recombination events between superscaffolds (SS01-SS05 and SS09-SS17; Fig. 2E), partial duplication of SS24 in F3, partial loss of SS01 in G9, and acquisition of partial copies of SS10, SS15, and SS18 in G9 (Fig. 2F). Noteworthy is a biased distribution of large indels (>350 bp), with a higher frequency of insertions in F3 (or deletions in G9) compared to F3 deletions (or G9 insertions), at a ratio of 444:97.

Approximately 0.33% (~491 k) of sites within the G9 genome are heterozygotic, with the majority (99.87%) of these sites having one allele corresponding to the F3 haplotype. Specific regions in the G9 genome (e.g., SS06, SS20, and partial SS01 and SS05) display low heterozygosity (Fig. 2G; Fig. S2). Comparison with a single-cell-derived strain (B4) from another RCC1216 culture (kindly provided by the lab of Assaf Vardi) indicates that regions of low heterozygosity in G9 and B4 genomes mostly do not overlap, suggesting they are due to recent loss of heterozygosity after the separation of RCC1216 and RCC1217. A total of 13,405 sites (0.0089%) have diverged between F3 and G9, primarily concentrated within the regions of low heterozygosity, presumably due to the loss of the G9 allele corresponding to the F3 haplotype. We phased the variable sites with long reads, resulting in 213 phasing blocks covering 94.90% of the haploid genome (Fig. 2H; Dataset S1). The number of recombinations is positively correlated with the length of phasing blocks, with the average recombination distance estimated to be ~56 kb (Fig. 2i).

By investigating the variation in gene copy numbers, as defined by gDNA ratios, of the two genomes, we found that only 75% of the genes possess two copies in the diploid genome relative to the haploid (Fig. 3A). Among the 25% that deviate from an expected 2:1 ratio, 12% have 3:1 ratio and 11% have a 1:1 ratio, which might be due to factors such as chromosome anomalies (e.g., aneuploidy) or gene duplications and losses. Notably, 2% of the genes have an even more pronounced deviation (with a ratio of < 0.5 or > 3.5 to 1) and were subsequently excluded from further analyses.

Transcriptomic landscapes and gene functions

RNA differential expression comparisons between G9 and F3 were conducted, utilizing four replicates of Illumina RNA-seq samples (Table S1), collected on the sixth morning after transferring to a fresh 100 ml K/2 medium (starting with 100 cells/µl). Gene copy numbers correlate with RNA transcript abundance (Figs. 3B, 3C), but there is a high variance of RNA differential expression among genes with a specific haploid-diploid gDNA ratio (Fig. 3B), suggesting the impact of copy numbers on gene expression is minor. Out of the 25,224 genes with detectable transcripts, 1,412 genes (5.60%) have transcripts only in the diploid (mean scaled transcripts per million [scaledTPM] < 5 in F3) and 2,704 genes (10.72%) only in the haploid (mean scaledTPM < 5 in G9) (Fig. 3D). Given the phylogenetic distance of haptophytes from well-studied eukaryote models (animals, fungi, and plants), our understanding of haptophyte gene functions remains very limited. Among the transcribed genes, only 53.26% (13,434 genes) are associated with known functions, and the proportions for only diploid-expressed (33.92%; 479 genes) and only haploid-expressed (39.09%; 1,057 genes) genes are even lower (Dataset S2).

The functions enriched among genes only expressed in the diploid were divided into five clusters (Fig. 3E; Dataset S3). The predominant cluster (red) primarily consists of membrane proteins, including genes implicated in coccolithogenesis, such as HCO_3^- transporters in the solute carrier 4 family, cation and proton channels, antiporters, and vesicle fusion SNARE proteins. A closely related, highly compact cluster (blue) comprises Myb-like transcription factors and Ca^{2+} channels. The third cluster (purple) mainly encompasses genes involved in carbohydrate metabolism, including the glycoside hydrolase 3 family, beta-glucan synthase, and galactosyltransferase. Two smaller clusters, each containing around 20 genes, include guanylate cyclase receptors and adenylate guanylate cyclase (yellow) and sulfatases (magenta).

The predominant functional cluster (Fig. 3F; red) of the haploid-specifically expressed genes primarily consists of genes related to flagella assembly, intra-flagellar transportation, microtubule-based movement, nuclear division, and certain cell cycle regulatory proteins. Other clusters include membrane receptors (blue; ligand-gated channels, protein kinases, blue light photoreceptors, etc.), carbohydrate metabolism (purple; glycoside hydrolase 5 family, glycosyltransferase, cellulases, etc.), sulfatases plus aminoglycan-related glycosyltransferases (yellow), and sulfotransferases (magenta).

Using Pigeon in PacBio Transcript Toolkit, we identified isoforms in 6,092 genes, with 2,145 genes (35.21%) showing isoform differential expression patterns: 1,232 (20.22%) had up-regulated isoforms in the diploid, and 1,229 (20.17%) had up-regulated isoforms in the haploid. Genes with diploid up-regulated isoforms were enriched in plastid thylakoid membrane components, such as the light-harvesting complex, cytochrome b6f complex, and photosystem II genes (Dataset S3). Those genes with haploid up-regulated isoforms

were enriched in mitotic cell cycle-related genes, such as cyclins, mitogen-activated protein kinases, and kinetochore complex components.

Proteomic profiling

Global proteins were identified, quantified, and compared using three replicates of data-independent acquisition (DIA) data for G9 and F3 (Table S1). Due to limitations in protein detection, the number of genes with detectable proteins (7,116) is lower than that with RNA transcripts detected (25,224). There is a moderate correlation between protein and RNA expression (R = 0.52~0.56; Figs. 3A-3C). Among genes with protein products detected, 1,082 are up-regulated in the diploid, with the majority (805 genes) being specifically detected in the diploid and enriched with transportation proteins (solute carriers, major facilitators, ion transporters, and antiporters) (Fig. 4E; Dataset S3). The 599 genes with protein expression up-regulated or only detected in the haploid are grouped into three enriched functional clusters: mini-chromosome maintenance (MCM) complex proteins, protein degradation enzymes (peptidases and proteasomes), and heat-shock proteins plus mitochondrial matrix proteins (dehydrogenases and reductases) (Fig. 4F). It is notable that flagella- and cell motion-related genes that were highly upregulated in the RNA differential expression profile were not enriched in the haploid proteome. Unlike RNA, which is typically localized within the main cell body, flagella are easily detached from the cell body during microscopic observations. Therefore, the absence of these proteins may be due to the loss of flagella during the pelleting step of sample preparation.

DNA methylation profiles and gene expression

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

DNA methylation of transposable elements (TEs) has been linked to heterochromatin formation and can influence life cycle phase differentiation (*22*). However, we did not observe a higher methylation level of TEs (Fig. S3). High CG methylation was observed in some rRNA genes, but not in other repetitive sequences (Fig. S3). Of the 43 genes annotated with DNA methyltransferase activity (GO:009008), two are up-regulated in the diploid and seven in the haploid. Notably, all three N-6 adenine methyltransferases (EHUX_g4335, EHUX_g18645, EHUX_g32175) are up-regulated in the haploid. However, only a few reads showed N-6 adenine methylation signals in our PacBio HiFi reads without any specific pattern.

To analyze the correlation between DNA methylation and RNA expression, we divided genes into six groups, including unexpressed genes ("no") and five bins of expressed genes ranked by scaled transcripts per million (scaledTPM) (*23*) (low to high: groups 1 to 5). Groups with higher RNA expression are associated with lower mean CG site methylation levels upstream of the start codon and higher levels downstream (Fig. 5C). For CHG sites, higher RNA expression is linked to lower methylation levels around the start codon and higher levels in the gene body (Fig. 5D). In both life cycle phases, E3 ubiquitin-protein ligase is among the top 10% of genes with the highest mean gene body CHG methylation levels (Dataset S3). A comparison between the two life cycle phases reveals positive association between RNA expression and gene body CG or CHG methylation for specific functions (Dataset S3). For instance, genes related to histone demethylation, mitosis, endocytosis, and flagella assembly often exhibit higher RNA expression and higher methylation levels in the gene body of the haploid (Fig. 5E). Genes of certain functions with significant differences in promoter methylation levels show differential RNA expression in the opposite direction. For instance, mitochondrion-targeting and ubiquitin-dependent proteins tend to show higher RNA expression and lower methylation levels in the haploid, whereas ethylene metabolic and SNARE proteins tend to exhibit higher RNA expression and lower methylation levels in the diploid (Fig. 5F, 5G).

Genome-level allelic imbalance of RNA expression in diploid

Allelic imbalance is the phenomenon where the two alleles of a gene in a diploid cell have different RNA expression levels (*24*). Among the 10,762 heterozygous genes, 1,809 genes (16.81%) show significant signals of imbalanced RNA expression (Fig. 6A). Through comparison with the genome sequence of the haploid strain F3, we designated one G9 allele as the F3-like allele and the other as the alternative allele. Among the genes with allelic imbalance, a slightly higher proportion of F3-like alleles are up-regulated compared with the alternative alleles (Fig. 6B). Compared with genes with F3-like alleles up-regulated (23%), a much high proportion (52%) of genes with up-regulated alternative alleles also higher expression in the diploid than the haploid (Fig. 6C). This observation suggests that a fraction (approximately 400 genes) of the differentially expressed genes identified in this study may be attributed to allelic imbalanced RNA expression.

The majority of the heterozygous genes have inter-allelic SNPs and/or indels (>80%) within 2,000-bp upstream of the start codon, but the percentage with allelic imbalance is not higher than without imbalance (Fig. 6D). However, we see a higher proportion of allelic imbalanced genes have indels in their coding sequences, particularly when these indels result in frameshifts and premature stop codons (Fig. 6D). We further examined the links between allelic imbalance and allele-specific methylation. A higher percentage of genes with allele-specific methylation patterns (in either the promoter region or gene body) was observed in genes with allelic imbalance than without in both CG and CHG contexts (Fig. 6E). Leveraging PacBio long-read DNA and RNA sequencing, we were able to better detect, phase, and visualize allelic imbalance and allele-specific methylation in a haplotype-resolved manner. For example, EHUX_g5658, a sulfatase gene with transcripts only detected in G9, displays allele-specific methylation in both the promoter region and gene body. Its F3-like allele exhibits the same methylation pattern as in the F3, and similar to the F3 gene, this allele is not transcribed (Fig. 7).

Histones and histone modifications

Epigenetic modifications of histones have been proposed as key players in life cycle phase transitions (*22*). Multiple copies of each core histone protein, including H2A, H2B, H3, and H4, were identified, but the H1 linker protein, which is found in animals, fungi, plants, and kinetoplastids (Discoba) (*25*), is not present in the *G. huxleyi* or *G. pseudohuxleyi* genomes. Both haploid and diploid phases exhibited up-regulation of specific histone paralogs, as well as histone lysine methyltransferases and deacetylases (Supplemental Data S6: S6-3). Genes associated with H3K9 methylation are known to influence life cycle phase differentiation in plants, green algae, and diatoms (*22*). In our study, the transcripts of K9 histone-lysine N-methyltransferase SUV39H2 and the protein of enhancer of zeste E(z) were exclusively expressed in the diploid phase. Two other K9 histone-lysine methyltransferases (EHUX_g17408 and EHUX_g28648) and some histone arginine N-methyltransferases (PRMTs) are also up-regulated in the diploid.

Diploid-haploid physiological differentiation

We summarized the multiomic profiles of the diploid and haploid phase of *G. huxleyi* by focusing on genes related to various physiological functions (Fig. 8; Supplementary Text and Dataset S4) (*6*, *18*, *26–28*). The diploid and haploid show clear differences in certain carbonic anhydrases, transporters, and SNAP receptors (SNAREs) that are involved in coccolithogenesis. We also demonstrate that they utilize different sets of genes involved in polysaccharide and sulfate metabolism. The considerably higher concentration of DMSP in the diploid (*29*) is attributable to more genes in DMSP degradation pathways being upregulated in the haploid. In line with a more mixotrophic nutrition, genes related to "endocytic vesicle" (GO:0030139) are up-regulated and have lower CHG methylation in the haploid. On the contrary, many photosynthesis-related genes, including those for chlorophyll synthesis and light harvesting, are up-regulated in the diploid (Fig. 8; Supplementary Text).

Discussion

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

Consistent with known high variation in genome size, structure, and gene content within *Gephyrocarpsa* and haptophytes (*17*, *30*), more than 3,000 (13%) orthologous gene group identified in *G. huxleyi* are absent in *G. pseudohuxleyi* (Fig. 2c). The genomes of RCC1216 and RCC1217, a haploid-diploid pair separated by approximately twenty years, unveiled notable copy number differences at the superscaffold level and hundreds of structural variations, including two instances of chromosome-level recombination (Fig. 2E; Fig. S1). These findings highlight the prevalence of structural variation between closely related coccolithophore strains (Supplementary Text). Instead of meiosis and sexual reproduction, the high frequency of variations is more likely due to homologous recombinations during DNA repair (mitotic recombinations), as has been shown in diatoms (33). Still, the Genomic copy numbers and structural variations (Fig 2E, 2F;

Additional S3-1, S3-2) can be potential post-zygotic barriers to nuclear gene flow, offering a plausible explanation for the observed restricted nuclear gene flow among morphotypes within *Gephyrocapsa* (*19*, *31*). In line with DNA content estimation based on flow cytometry (*32*, *33*), we show a higher frequency of insertions in F3 (or deletions in G9) than F3 deletions (or G9 insertions) (Additional S3-1, S3-2), which leads to a diploid genome size smaller than twice the haploid genome size (Table 1).

The diploid-haploid differentiation in *G. huxleyi* involves genes encoding diverse cellular and biochemical functions (Fig. 8; Supplementary Text). The enrichment of transmembrane transporters among up-regulated diploid genes underscores the crucial role of ion transporters in calcification, including bicarbonate transporters (SLC4, ABCC8), calcium ion transporters and channels (CAX3, SLC24, TRPM), and H^+/Na^+ exchangers and a vacuolar-type ATPase for removal of the byproduct proton to maintain pH homeostasis. For the exocytosis of coccoliths, SNAP receptors (SNAREs) that mediate vesicle fusion with target membranes show diploid-specific expression (Fig. 8; Dataset S4) (*16*). We identified 15 carbonic anhydrases (CAs), which are pivotal for the conversion between carbon dioxide and bicarbonate and potentially regulate carbon acquisition and allocation between photosynthesis and calcification. Among the four CA families, αCAs and δCAs exhibit up-regulation in the diploid phase and are likely to be directly involved in calcification.

Regulation of $CaCO_3$ crystal growth on an organic baseplate within the coccolith vesicle involve both proteins and polysaccharides. As the major organic component of coccoliths, coccolith-associated polysaccharides (CAPs) are acidic, can contain sulfate groups, and mediate the calcification process by directing calcium ions and controlling crystal morphology (*34–36*). Transcriptomic differences between the diploid and haploid strains are marked by an enrichment of enzymes for carbohydrate metabolism (including glycosyltransferases), sulfatases, and sulfotransferases (Fig. 3; Dataset S3), suggesting that diploid and haploid cells have a major difference in their carbohydrate components that might be related to coccolithogenesis. The diploid-haploid differential expression can be regulated through allele-imbalanced transcription (Fig. 6), which is in turn associated with allele-specific methylation patterns in the diploid (Fig. 7). To further elucidate the mechanisms for epigenomic reprograming in *G. huxleyi*, it remains a major challenge to systematically manipulate diploid-haploid transitions in the lab.

In addition to coccolithogenesis, the transcriptomic and proteomic profiles reflect other physiological differences between the haploid and diploid phases (Fig. 8; Supplementary Text; Dataset S4). The presence of flagella and a likely more mixotrophic nutrition in the haploid are in contrast with more abundant chlorophyll a (*37*, *38*) and the up-regulation of photosynthesis-related genes in the diploid. The low abundance of DMSP in the haploid (*29*) is possibly due to the up-regulation of DMSP catabolic enzymes in both demethylation and cleavage pathways, including two key DMSP lyases, DddD and Alma. Given that DMSP and derived compounds are known to mediate microbial interactions (*9*, *10*), differences in DMSP metabolism imply that the two life-cycle phases might differ in their microalga-microbe interactions, as also suggested by the different bacterial composition associated with G9 and F3 cultures (Table S2). Consistent with the higher growth rate (*39*), we found that MCM, crucial for DNA replication initiation, is enriched among the haploid up-regulated proteins (Fig. 4F; Dataset S4), suggesting increased mitotic activity in haploid cells. Together these physiological differences point to potential ecological differentiation between the diploid and haploid, which merits further investigation using field-collected data.

358 Epigenetic modifications of DNA and histones have been proposed as key players in life
359 cycle phase transitions (*22*), with varying patterns across the tree of eukaryotes. Animals
360 typically exhibit methylation in the CG context, while land plants and the green alga
361 *Chlamydomonas* manifest methylation in CG, CHG, and CHH contexts (*40*, *41*). In
362 contrast to methylation in red algae (bias towards non-CG) (*42*), diatoms (predominantly
363 CG) (*43*, *44*), and brown algae (absent or minimal and confined to CHH) (*45*, *46*). *G.*
364 *huxleyi* methylation is found in CG (~18%) and CHG (~9%) contexts, with relatively rare
365 CHH methylation (~1%) (Fig. 5A, 5B; Fig. S3). We show that methylation patterns can be
366 heritable, as evidenced by the similarity between the F3-like haplotype in G9 and the F3
367 genome in certain regions after at least 20 years of separation (Fig. 7). Unlike
368 Chloroplastida, where gene body CG methylation up-regulates gene expression and
369 methylation in CHG contexts is typically associated with low expression (*47*), genes with
370 higher expression levels in *G. huxleyi* display increased methylation in both CG and CHG
371 contexts (Fig. 5C, 5D). Histones H2A, H2B, H3, and H4 were identified within the
372 genome, and each core histone class has specific paralogs up-regulated in either haploid or
373 diploid cells (Dataset S4). In green algae and bryophytes, the regulation of life cycle phase
374 transition has been suggested to involve histone modifications, particularly the tri-
375 methylation of lysine 27 on histone H3 protein (H3K27me3) that silences gene expression
376 (*22*). The up-regulation of H3K9 methyltransferase SUV39H2, enhancer of zeste E(z), and
377 arginine methyltransferases in the diploid of *G. huxleyi* implies a potential role of histone
378 methylation in diploid-haploid differentiation, but further studies are needed to profile
379 histone modifications and test this hypothesis.

380 This study presents a comprehensive view of the life-cycle differentiation between diploid
381 and haploid phases of *Gephyrocapsa* — globally distributed and highly abundant
382 calcifying algae with pivotal roles in marine carbon cycling. In particular, our results show
383 that (1) their genomes are dynamic and prone to variations, including chromosome-level
384 structural variations, biased indel distribution between the diploid and haploid, and loss of
385 heterozygosity in the diploid, as revealed by their highly contiguous *de novo* genome
386 assemblies; (2) Transcriptomic and proteomic differences of diploid and haploid cells
387 modulate their differentiation in coccolithogenesis, flagellar assembly and movement,
388 photosynthesis, carbohydrate metabolism, DMSP degradation, DNA replication, and
389 endomembrane system and transport; (3) Gene transcript abundance is correlated with
390 patterns of DNA methylation, which occurs in CG and CHG contexts and can be
391 inheritable, allele-specific, and differentiated between life-cycle phases. Haplodiplontic
392 life cycles are widespread across diverse lineages of eukaryotes, where coccolithophores
393 exemplify both algae and microbial eukaryotes that exist in two morphologically and
394 physiologically distinct ploidy phases. By integrating genomic, transcriptomic, proteomic,
395 and methylomic analyses, this study not only illuminates the molecular basis of diploid-
396 haploid differentiation and maintenance in coccolithophores, but lays the groundwork for
397 future research on their life cycling in marine ecosystems and its implications for the
398 distribution, ecological roles, and microbial interactions of coccolithophores.
399

400 **Materials and Methods**

401 **Experimental Design**

402 To investigate the life cycle phase differentiation of *G. huxleyi*, the model species of
403 coccolithophores, we used a pair of diploid-haploid single-cell derived strains, RCC1216-
404 G9 (G9) and RCC1217-F3 (F3). To establish a robust foundation for downstream
405 bioinformatic analyses, we constructed high-quality, haplotype-aware *de novo* genome
406 assemblies for both strains. This involved combining data from long-read (PacBio and

407 Oxford Nanopore) and linked-read (10x genomics plus Illumina) sequencing with
408 chromosome optical mapping (Bionano) to identify existing genomic variations, including
409 copy number variations, structural variations, insertion-deletions (Indels), and single-
410 nucleotide polymorphisms (SNPs). For accurate transcriptome profiling, genome
411 annotation, and gene expression analysis, we integrated data from Illumina RNA-Seq and
412 full-length isoform sequencing (PacBio Iso-Seq). We also quantified and compared global
413 protein expression using the DIA method. Lastly, we profiled genome DNA methylation,
414 a known epigenetic regulatory mechanism for life cycle phase differentiation in some
415 eukaryotes, by identifying 5-methylcytosine at symmetric (CG and CHG) and non-
416 symmetric (CHH) sites using Illumina bisulfite sequencing and direct calling from kinetic
417 data of the PacBio HiFi long reads. By integrating the omics data from these four
418 components, our study provides a comprehensive view of diploid-haploid differentiation
419 at the molecular level.

420

421 **Cell culture preparation**

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

430

431 **Genome sequencing, assembly, and analyses**

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

449 *Genome assembly*—A hybrid approach of *de novo* genome assembly was used, integrating
450 data from short-read, long-read, linked-read, and optical mapping. For the F3 genome,
451 draft contigs were assembled with PacBio circular consensus sequences (CCS) using
452 NextDenovo v2.5.0 (*48*). These contigs were then polished with PacBio HiFi and Illumina
453 reads using NextPolish 1.4.0 (*49*), and non-eukaryotic contigs (Table S2) were removed

454 by blasting against NCBI non-redundant nucleotide collection (nt; downloaded on
455 Feb/19/2022). Eukaryotic contigs were scaffolded with 10x linked reads using ARCS (*50*)
456 and with nanopore reads using ARKS-long (*51*). The genome was further hybrid-
457 scaffolded with Bionano *de novo* genome maps using *hybridScaffold.pl* in Bionano Solve
458 v3.6.1 and gap-filled with corrected long reads using TGS-GapCloser v1.1.1 (*52*). To
459 validate the assembly correctness and copy number, the Illumina reads, PacBio reads,
460 Nanopore reads/NextDenovo draft contigs, and Bionano molecules were mapped back to
461 the superscaffolds. The bridged gaps were manually inspected, and only one (SS06 from
462 coordinate 4,826k to 5,035k) lacked support from any long reads. The haplotype-aware
463 phasing blocks were retrieved by calling haplotype-based variants using Freebayes
464 v0.9.21 (*53*) with short reads and phasing the variants with long reads using WhatsHap
465 v0.19 (*54*) In cases of partially duplicated superscaffolds (e.g. SS24 and SS26 of F3),
466 haplotype-tagged reads were extracted to reconstruct *de novo* assemblies using
467 NextDenovo and NextPolish.

468 *Genomic comparisons*—Genomic structural variations between the two genomes were
469 analyzed by mapping the Bionano molecules of one strain to the genome of the other
470 strain using *RefAligner* in Bionano Solve, and by aligning genome assemblies to each
471 other using Minimap2 v2.15 (*55*). Single nucleotide polymorphisms (SNPs) and short
472 insertion-deletions (indels) were identified with GATK v4.3.0.0 (*56*) that the variable sites
473 were called using *HaplotypeCaller* and underwent hard filtering based on the
474 recommended GATK *VariantFiltration* settings (SNPs: QD < 2.0, QUAL < 30.0, SOR >
475 3.0, FS > 60.0, MQ < 40.0, MQRankSum < -12.5, and ReadPosRankSum < -8.0; indels:
476 QD < 2.0, QUAL < 30.0, FS > 200.0, and ReadPosRankSum < -20.0). The draft of
477 variable sites was then used in base quality score recalibration using *BaseRecalibrator* and
478 *ApplyBQSR* for subsequent variation calling. Convolutional neural net filtering was
479 applied, and variants were annotated using GATK *HaplotypeCaller*, *CNNScoreVariants*,
480 and *FilterVariantTranches*. The G9 variable sites were phased with long reads using
481 WhatsHap. The synonymous and non-synonymous variable sites of each coding gene
482 within and between G9 and F3 were analyzed using SNPGenie v2019.10.31 (*57*).

483

484 **RNA sequencing, genome annotation, and functional annotation**

485 *RNA sequencing*—Total RNA from G9 and F3 was extracted using the RNeasy Plant Mini
486 Kit (Qiagen, Hilden, Germany). Each strain had four short-read libraries prepared using
487 the TruSeq Stranded mRNA Library Prep Kit (Illumina, San Diego, CA, USA), following
488 the manufacturer's recommendations, and sequenced on Illumina NovaSeq 6000 in 150 bp
489 paired-end mode. Additionally, one PacBio Iso-seq library was prepared using the
490 SMRTbell® prep kit (PacBio, Menlo Park, CA, United States) and sequenced on the
491 Sequel IIe.

492 *Genome annotation*—Repetitive sequences in the genome were modeled using
493 RepeatModeler2 v2.0.2a (*58*) and annotated with Dfam 3.5 and RepBase RepeatMasker
494 Edition (20181026) using RepeatMasker v4.1.2 (*59*). Telomeric repeats were explored and
495 annotated with the Telomere Identification toolKit (tidk) v0.2.0 (*60*) Both PacBio Iso-seq
496 and Illumina RNA-seq reads were used for annotating the genome, which was soft-
497 masked with repeats. PacBio Iso-seq CCS reads were mapped to the genome and then
498 refined, clustered, and collapsed using PacificBiosciences Isoseq3 v3.8.2, with protein-
499 coding regions predicted by GeneMarkS-T v.5.1 (*61*). The Illumina RNA-seq reads were
500 aligned to the genomes and gene models were trained and predicted using GeneMark-ET
501 and Augustus v3.4.0 in BRAKER 2.1.6 (*62*). The two annotations were combined and

502 renamed using TSEBRA v1.0.3 (*63*) for further isoform analyses, transcripts were filtered
503 based on the Iso-Seq data using PacificBiosciences Pigeon. The shared genes and ortholog
504 groups among genomes were analyzed using the amino acid datasets of haptophytes
505 provided in the associated files of Skeffington et al. (2023) (*18*) (DOI:
506 10.6084/m9.figshare.20463900) with OrthoFinder v2.5.2 (*64*). Non-coding RNAs were
507 predicted and annotated with the Rfam database v14.8 using structRNAfinder v1.0 (*65*).

508 *Functional annotation*—The functions of the predicted protein-coding transcripts were
509 annotated using the EnTAP v0.10.8 pipeline (*66*) by searching against NCBI non-
510 redundant protein collection (nr), UniProt Swiss-Prot, and UniProt TrEMBL databases (all
511 downloaded on May/18/2022) using DIAMOND v0.9.9, and Gene Ontologies (GO),
512 KEGG pathway orthologous groups (KO), and protein domains (SMART/Pfam) were
513 explored and annotated based on the eggNOG database (downloaded on May/18/2022).

514

515 **Global protein identification and methylomic sequencing**

516 *Global protein identification*—Total proteins were extracted and digested from three
517 replicates of frozen cell pellets containing approximately 1×10^7 cells of each strain. Mass
518 spectrometry (MS) analyses were conducted on a Thermo Orbitrap Fusion Lumos Tribrid
519 Mass Spectrometer using DIA mode.

520 *Methylomic sequencing*—To prepare for bisulfite sequencing, cells were lysed using a
521 modified lysis buffer (Tris 0.1M, EDTA 0.05M, NaCl 0.1M, SDS 1%, N-laurylsarcosine
522 2%, and proteinase K 1mg/ml in pH 8.0; modified from Bendif et al., 2019 (*12*), and then
523 genomic DNA was isolated following the PacBio Salt:Chloroform gDNA Wash protocol.
524 Two samples of each strain were bisulfite converted using the EZ DNA Methylation Gold
525 Kit (ZymoGenetics, Seattle, WA, USA). The two libraries were prepared using the TruSeq
526 DNA Methylation Kit (Illumina) and subsequently sequenced using HiSeq in 150 bp and
527 250 bp paired-end modes, respectively.

528

529 **Statistical Analysis**

530 *Transcriptomic Analyses*—The expression of transcripts was quantified with Illumina
531 RNA-seq reads using Salmon v1.9.0 (*67*), correcting for GC bias (--gcBias) and
532 performing 30 Gibbs sampling runs (--numGibbsSamples 30). Differential expression of
533 genes and transcripts, as well as differential transcript usage, were analyzed with
534 inferential replicate counts (i.e., the Gibbs from Salmon), normalized by the scaledTPM
535 method using the Swish method carried out with the R package *fishpond* v2.0.1. The
536 global allelic imbalance was tested using the SEESAW method (*24*). Differentially
537 expressed genes and allelic imbalance genes were identified with q-values ≤ 0.01, while
538 G9- or F3-specific genes were those with q-values ≤ 0.01 and a mean scaledTPM < 5 in
539 the strain with fewer mapped reads.

540 *Proteomic Analyses*—The quantification and normalization of MS data were performed
541 using the MaxLFQ method as implanted in DIA-NN v1.8.1 (*68*). Differential protein
542 expression was analyzed based on MaxLFQ value and the number of precursors (used for
543 variance estimation) using the R package *DEqMS* v1.12.1 (*69*). Differentially expressed
544 genes were identified with q-values ≤ 0.05, while G9- or F3-specific genes were
545 characterized by q-values ≤ 0.05 and absence of peptide detection in two or three out of
546 three replicated samples in the strain with fewer mapped reads.

Methylomic Analyses—The bisulfite sequencing reads and PacBio HiFi reads that retained kinetics were mapped to the genomes and used for methylation calling with BSBolt v1.4.5 (*70*) and PacificBiosciences Primrose v1.3.0, respectively. Genome-wide DNA methylation profiling was conducted using R package BSseq v.1.30.0 (*71*), while their association with RNA expression was analyzed using MethGet (*72*). Differentially methylated CG, CHG, and CHH sites, genes, and regions were analyzed using the R package *DSS* v.2.42.0 (*73*). Differentially methylated sites were detected with an FDR threshold of ≤ 0.05, while differential methylation of genes and promoter regions (2000 bp upstream of genes) was determined either by mean methylation level (FDR ≤ 0.05; Dataset S2) or by overlapping with differential methylation regions surpassing areaStats thresholds (Fig. 6E). In comparisons with RNA expression, an additional set of regions, most correlated with RNA expression (coordinates -350~-50 and 50-450 relative to start codon in CG context, see Fig. 5C; and -200~300 and 500-1500 in CHG context, see Fig. 5D), were also assessed for differential methylation.

Functional enrichment analyses—Gene and transcript sets of interest, including copy number variation, differential gene/protein expression, transcript usage, allelic imbalance, and methylation levels, were analyzed for enriched GO, KO, SMART, and Pfam terms using the function "*enricher*" in R package *clusterProfiler* v4.2.2 (*74*). Additionally, enriched GO terms were analyzed using R package *topGO* v2.46.0 (*75*)with algorisms “classic” or “weight01”. Enriched GO terms identified by *clusterProfiler* ($p < 0.05$) were clustered using the MCL-based overlap coefficient in GOMCL v0.0.1 (*76*).

References

1. U. Goodenough, J. Heitman, Origins of eukaryotic sexual reproduction. *Cold Spring Harb Perspect Biol* **6**, a016154–a016154 (2014).
2. Sandra K. Floyd, John L. Bowman, The ancestral developmental tool kit of land plants. (2007).
3. J. Umen, S. Coelho, Algal sex determination and the evolution of anisogamy. *Annu Rev Microbiol* **73**, 267–291 (2019).
4. J. de Vries, F. Monteiro, G. Wheeler, A. Poulton, J. Godrijan, F. Cerino, E. Malinverno, G. Langer, C. Brownlee, Haplo-diplontic life cycle expands coccolithophore niche. *Biogeosciences* **18**, 1161–1184 (2021).
5. M. J. Frada, E. M. Bendif, S. Keuter, I. Probert, The private life of coccolithophores. *Perspectives in Phycology* **6**, 11–30 (2019).
6. A. R. Taylor, C. Brownlee, G. Wheeler, Coccolithophore cell biology: chalking up progress. *Ann Rev Mar Sci* **9**, 283–310 (2017).
7. C. De Vargas, M.-P. Aubry, I. Probert, J. Young, “Origin and evolution of coccolithophores: from coastal hunters to oceanic farmers” in *Evolution of Primary Producers in the Sea* (Elsevier, 2007), pp. 251–285.
8. C. Rousseaux, W. Gregg, Interannual variation in phytoplankton primary production at a global scale. *Remote Sens (Basel)* **6**, 1–19 (2013).
9. U. Alcolombri, S. Ben-Dor, E. Feldmesser, Y. Levin, D. S. Tawfik, A. Vardi, Identification of the algal dimethyl sulfide–releasing enzyme: a missing link in the marine sulfur cycle. *Science (1979)* **348**, 1466–1469 (2015).
10. N. Barak-Gavish, M. J. Frada, C. Ku, P. A. Lee, G. R. DiTullio, S. Malitsky, A. Aharoni, S. J. Green, R. Rotkopf, E. Kartvelishvily, U. Sheyn, D. Schatz, A. Vardi, Bacterial virulence against an oceanic bloom-forming phytoplankter is mediated by algal DMSP. *Sci Adv* **4**, eaau5716 (2018).
11. R. J. Charlson, J. E. Lovelock, M. O. Andreae, S. G. Warren, Oceanic phytoplankton, atmospheric sulphur, cloud albedo and climate. *Nature* **326**, 655–661 (1987).
12. E. M. Bendif, B. Nevado, E. L. Y. Wong, K. Hagino, I. Probert, J. R. Young, R. E. M. Rickaby, D. A. Filatov, Repeated species radiations in the recent evolution of the key marine phytoplankton lineage *Gephyrocapsa*. *Nat Commun* **10**, 4234 (2019).
13. G. L. Wheeler, D. Sturm, G. Langer, Gephyrocapsa huxleyi (Emiliania huxleyi) as a model system for coccolithophore biology. *J Phycol* **00**, 1–7 (2023).

14. A. Houdan, I. Probert, K. Van Lenning, S. Lefebvre, Comparison of photosynthetic responses in diploid and haploid life-cycle phases of *Emiliania huxleyi* (Prymnesiophyceae). *Mar Ecol Prog Ser* **292**, 139–146 (2005).
15. M. Frada, I. Probert, M. J. Allen, W. H. Wilson, C. de Vargas, The "Cheshire Cat" escape strategy of the coccolithophore *Emiliania huxleyi* in response to viral infection. *Proceedings of the National Academy of Sciences* **105**, 15944–15949 (2008).
16. P. von Dassow, H. Ogata, I. Probert, P. Wincker, C. Da Silva, S. Audic, J.-M. Claverie, C. de Vargas, Transcriptome analysis of functional differentiation between haploid and diploid cells of *Emiliania huxleyi*, a globally significant photosynthetic calcifying cell. *Genome Biol* **10**, R114 (2009).
17. B. A. Read, J. Kegel, M. J. Klute, A. Kuo, S. C. Lefebvre, F. Maumus, C. Mayer, J. Miller, A. Monier, A. Salamov, J. Young, M. Aguilar, J.-M. Claverie, S. Frickenhaus, K. Gonzalez, E. K. Herman, Y.-C. Lin, J. Napier, H. Ogata, A. F. Sarno, J. Shmutz, D. Schroeder, C. de Vargas, F. Verret, P. von Dassow, K. Valentin, Y. Van de Peer, G. Wheeler, J. B. Dacks, C. F. Delwiche, S. T. Dyhrman, G. Glöckner, U. John, T. Richards, A. Z. Worden, X. Zhang, I. V. Grigoriev, Pan genome of the phytoplankton *Emiliania* underpins its global distribution. *Nature* **499**, 209–213 (2013).
18. A. Skeffington, A. Fischer, S. Sviben, M. Brzezinka, M. Górka, L. Bertinetti, C. Woehle, B. Huettel, A. Graf, A. Scheffel, A joint proteomic and genomic investigation provides insights into the mechanism of calcification in coccolithophores. *Nat Commun* **14**, 3749 (2023).
19. T. Kao, T. Wang, C. Ku, Rampant nuclear–mitochondrial–plastid phylogenomic discordance in globally distributed calcifying microalgae. *New Phytologist* **235**, 1394–1408 (2022).
20. E. M. Bendif, I. Probert, O. A. Archontikis, J. R. Young, L. Beaufort, R. E. Rickaby, D. Filatov, Rapid diversification underlying the global dominance of a cosmopolitan phytoplankton. *ISME J* **17**, 630–640 (2023).
21. J. Fulnečková, T. Ševčíková, J. Fajkus, A. Lukešová, M. Lukeš, Č. Vlček, B. F. Lang, E. Kim, M. Eliáš, E. Sýkorová, A broad phylogenetic survey unveils the diversity and evolution of telomeres in eukaryotes. *Genome Biol Evol* **5**, 468–483 (2013).
22. J. Vigneau, M. Borg, The epigenetic origin of life history transitions in plants and algae. *Plant Reprod* **34**, 267–285 (2021).
23. C. Soneson, M. I. Love, M. D. Robinson, Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Res* **4**, 1521 (2015).
24. E. Y. Wu, N. P. Singh, K. Choi, M. Zakeri, M. Vincent, G. A. Churchill, C. L. Ackert-Bicknell, R. Patro, M. I. Love, SEESAW: detecting isoform-level allelic imbalance accounting for inferential uncertainty. *Genome Biol* **24**, 165 (2023).
25. P. B. Talbert, M. P. Meers, S. Henikoff, Old cogs, new tricks: the evolution of gene expression in a chromatin context. *Nat Rev Genet* **20**, 283–297 (2019).
26. I. Benner, R. E. Diner, S. C. Lefebvre, D. Li, T. Komada, E. J. Carpenter, J. H. Stillman, *Emiliania huxleyi* increases calcification but not expression of calcification-related genes in long-term exposure to elevated temperature and pCO2. *Philosophical Transactions of the Royal Society B: Biological Sciences* **368**, 20130049 (2013).
27. L. Mackinder, G. Wheeler, D. Schroeder, P. von Dassow, U. Riebesell, C. Brownlee, Expression of biomineralization-related ion transport genes in *Emiliania huxleyi*. *Environ Microbiol* **13**, 3250–3265 (2011).
28. C. Brownlee, G. Langer, G. L. Wheeler, Coccolithophore calcification: Changing paradigms in changing oceans. *Acta Biomater* **120**, 4–11 (2021).
29. E. Heidenreich, R. Wördenweber, F. Kirschhöfer, M. Nusser, F. Friedrich, K. Fahl, O. Kruse, B. Rost, M. Franzreb, G. Brenner-Weiß, S. Rokitta, Ocean acidification has little effect on the biochemical composition of the coccolithophore Emiliania huxleyi. *PLoS One* **14**, e0218564 (2019).
30. P. Yang, K. Guo, Y. Yang, M. Lyu, J. Liu, X. Li, Y. Feng, Phylogeny and genetic variations of the three genome compartments in haptophytes shed light on the rapid evolution of coccolithophores. *Gene* **887**, 147716 (2023).
31. D. A. Filatov, E. M. Bendif, O. A. Archontikis, K. Hagino, R. E. M. Rickaby, The mode of speciation during a recent radiation in open-ocean phytoplankton. *Current Biology* **31**, 5439-5449.e5 (2021).
32. A. Houdan, C. Billard, D. Marie, F. Not, A. G. Sáez, J. R. Young, I. Probert, Holococcolithophore-heterococcolithophore (Haptophyta) life cycles: Flow cytometric analysis of relative ploidy levels. *Syst Biodivers* **1**, 453–465 (2004).
33. M. J. Frada, S. Rosenwasser, S. Ben-Dor, A. Shemi, H. Sabanay, A. Vardi, Morphological switch to a resistant subpopulation in response to viral infection in the bloom-forming coccolithophore Emiliania huxleyi. *PLoS Pathog* **13**, e1006775 (2017).
34. B. Marzec, J. M. Walker, M. Panagopoulou, Y. Jhons, D. Clare, A. Wheeler, M. P. Shaver, F. Nudelman, Three-dimensional architecture and surface functionality of coccolith base plates. *J Struct Biol* **208**, 127–136 (2019).
35. A. Gal, R. Wirth, J. Kopka, P. Fratzl, D. Faivre, A. Scheffel, Macromolecular recognition directs calcium ions to coccolith mineralization sites. *Science (1979)* **353**, 590–593 (2016).

36. M. E. Marsh, A. L. Ridall, P. Azadi, P. J. Duke, Galacturonomannan and Golgi-derived membrane linked to growth and shaping of biogenic calcite. *J Struct Biol* **139**, 39–45 (2002).
37. Z. Ruan, M. Lu, H. Lin, S. Chen, P. Li, W. Chen, H. Xu, D. Qiu, Different photosynthetic responses of haploid and diploid *Emiliania huxleyi* (Prymnesiophyceae) to high light and ultraviolet radiation. *Bioresour Bioprocess* **10**, 40 (2023).
38. I. Hariskos, T. Rubner, C. Posten, Investigation of cell growth and chlorophyll a content of the coccolithophorid alga Emiliania huxleyi by using simple bench-top flow cytometry. *J Bioprocess Biotech* **05** (2015).
39. S. D. Rokitta, L. J. de Nooijer, S. Trimborn, C. de Vargas, B. Rost, U. John, Transcriptome analyses reveal differential gene expression patterns between the life-cycle stages of *Emiliania huxleyi* (haptophyta) and reflect specialization to different ecological niches. *J Phycol* **47**, 829–838 (2011).
40. T. Kawashima, F. Berger, Epigenetic reprogramming in plant sexual reproduction. *Nat Rev Genet* **15**, 613–624 (2014).
41. D. A. Lopez, T. Hamaji, J. Kropat, P. De Hoff, M. Morselli, L. Rubbi, S. T. Fitz-Gibbon, S. D. Gallaher, S. S. Merchant, J. G. Umen, M. Pellegrini, Dynamic changes in the transcriptome and methylome of *Chlamydomonas reinhardtii* throughout its life cycle. *Plant Physiol* **169**, 2730–2743 (2015).
42. J. M. Lee, H. J. Song, S. I. Park, Y. M. Lee, S. Y. Jeong, T. O. Cho, J. H. Kim, H.-G. Choi, C. G. Choi, W. A. Nelson, S. Fredericq, D. Bhattacharya, H. S. Yoon, Mitochondrial and plastid genomes from coralline red algae provide insights into the incongruent evolutionary histories of organelles. *Genome Biol Evol* **10**, 2961–2972 (2018).
43. A. Veluchamy, X. Lin, F. Maumus, M. Rivarola, J. Bhavsar, T. Creasy, K. O'Brien, N. A. Sengamalay, L. J. Tallon, A. D. Smith, E. Rayko, I. Ahmed, S. Le Crom, G. K. Farrant, J.-Y. Sgro, S. A. Olson, S. S. Bondurant, A. E. Allen, P. D. Rabinowicz, M. R. Sussman, C. Bowler, L. Tirichine, Insights into the role of DNA methylation in diatoms by genome-wide profiling in *Phaeodactylum tricornutum*. *Nat Commun* **4**, 2091 (2013).
44. J. C. Traller, S. J. Cokus, D. A. Lopez, O. Gaidarenko, S. R. Smith, J. P. McCrow, S. D. Gallaher, S. Podell, M. Thompson, O. Cook, M. Morselli, A. Jaroszewicz, E. E. Allen, A. E. Allen, S. S. Merchant, M. Pellegrini, M. Hildebrand, Genome and methylome of the oleaginous diatom *Cyclotella cryptica* reveal genetic flexibility toward a high lipid phenotype. *Biotechnol Biofuels* **9**, 258 (2016).
45. J. M. Cock, L. Sterck, P. Rouzé, D. Scornet, A. E. Allen, G. Amoutzias, V. Anthouard, F. Artiguenave, J.-M. Aury, J. H. Badger, B. Beszteri, K. Billiau, E. Bonnet, J. H. Bothwell, C. Bowler, C. Boyen, C. Brownlee, C. J. Carrano, B. Charrier, G. Y. Cho, S. M. Coelho, J. Collén, E. Corre, C. Da Silva, L. Delage, N. Delaroque, S. M. Dittami, S. Doulbeau, M. Elias, G. Farnham, C. M. M. Gachon, B. Gschloessl, S. Heesch, K. Jabbari, C. Jubin, H. Kawai, K. Kimura, B. Kloareg, F. C. Küpper, D. Lang, A. Le Bail, C. Leblanc, P. Lerouge, M. Lohr, P. J. Lopez, C. Martens, F. Maumus, G. Michel, D. Miranda-Saavedra, J. Morales, H. Moreau, T. Motomura, C. Nagasato, C. A. Napoli, D. R. Nelson, P. Nyvall-Collén, A. F. Peters, C. Pommier, P. Potin, J. Poulain, H. Quesneville, B. Read, S. A. Rensing, A. Ritter, S. Rousvoal, M. Samanta, G. Samson, D. C. Schroeder, B. Ségurens, M. Strittmatter, T. Tonon, J. W. Tregear, K. Valentin, P. von Dassow, T. Yamagishi, Y. Van de Peer, P. Wincker, The *Ectocarpus* genome and the independent evolution of multicellularity in brown algae. *Nature* **465**, 617–621 (2010).
46. X. Fan, W. Han, L. Teng, P. Jiang, X. Zhang, D. Xu, C. Li, M. Pellegrini, C. Wu, Y. Wang, M. J. S. Kaczurowski, X. Lin, L. Tirichine, T. Mock, N. Ye, Single-base methylome profiling of the giant kelp Saccharina japonica reveals significant differences in DNA methylation to microalgae and plants. *New Phytologist* **225**, 234–249 (2020).
47. A. J. Bewick, R. J. Schmitz, Gene body DNA methylation in plants. *Curr Opin Plant Biol* **36**, 103–110 (2017).
48. J. Hu, Z. Wang, Z. Sun, B. Hu, A. O. Ayoola, F. Liang, J. Li, J. R. Sandoval, D. N. Cooper, K. Ye, J. Ruan, C.-L. Xiao, D.-P. Wang, D.-D. Wu, S. Wang, An efficient error correction and accurate assembly tool for noisy long reads. *bioRxiv*, 2023.03.09.531669 (2023).
49. J. Hu, J. Fan, Z. Sun, S. Liu, NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics* **36**, 2253–2255 (2020).
50. S. Yeo, L. Coombe, R. L. Warren, J. Chu, I. Birol, ARCS: scaffolding genome drafts with linked reads. *Bioinformatics* **34**, 725–731 (2018).
51. L. Coombe, J. Zhang, B. P. Vandervalk, J. Chu, S. D. Jackman, I. Birol, R. L. Warren, ARKS: chromosome-scale scaffolding of human genome drafts with linked read kmers. *BMC Bioinformatics* **19**, 234 (2018).
52. M. Xu, L. Guo, S. Gu, O. Wang, R. Zhang, B. A. Peters, G. Fan, X. Liu, X. Xu, L. Deng, others, TGS-GapCloser: a fast and accurate gap closer for large genomes with low coverage of error-prone long reads. *Gigascience* **9**, giaa094 (2020).
53. E. Garrison, G. Marth, Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907* (2012).

54. M. Patterson, T. Marschall, N. Pisanti, L. Van Iersel, L. Stougie, G. W. Klau, A. Schönhuth, WhatsHap: weighted haplotype assembly for future-generation sequencing reads. *Journal of Computational Biology* **22**, 498–509 (2015).
55. H. Li, Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
56. A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, M. A. DePristo, The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297–1303 (2010).
57. C. W. Nelson, L. H. Moncla, A. L. Hughes, SNPGenie: estimating evolutionary parameters to detect natural selection using pooled next-generation sequencing data. *Bioinformatics*, btv449 (2015).
58. J. M. Flynn, R. Hubley, C. Goubert, J. Rosen, A. G. Clark, C. Feschotte, A. F. Smit, RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences* **117**, 9451–9457 (2020).
59. A. F. A. Smit, R. Hubley, P. Green, RepeatMasker Open-4.0. 2013--2015. [Preprint] (2015).
60. Tree of Life Kit, A Telomere Identification toolKit (tidk), *URL* (2022)p. https://github.com/tolkit/telomeric-identifier.
61. S. Tang, A. Lomsadze, M. Borodovsky, Identification of protein coding regions in RNA transcripts. *Nucleic Acids Res* **43**, e78–e78 (2015).
62. K. J. Hoff, S. Lange, A. Lomsadze, M. Borodovsky, M. Stanke, BRAKER1: Unsupervised RNA-seq-based genome annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics* **32**, 767–769 (2016).
63. L. Gabriel, K. J. Hoff, T. Brůna, M. Borodovsky, M. Stanke, TSEBRA: transcript selector for BRAKER. *BMC Bioinformatics* **22**, 566 (2021).
64. D. M. Emms, S. Kelly, OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* **20**, 238 (2019).
65. R. Arias-Carrasco, Y. Vásquez-Morán, H. I. Nakaya, V. Maracaja-Coutinho, StructRNAfinder: an automated pipeline and web server for RNA families prediction. *BMC Bioinformatics* **19**, 55 (2018).
66. A. J. Hart, S. Ginzburg, M. (Sam) Xu, C. R. Fisher, N. Rahmatpour, J. B. Mitton, R. Paul, J. L. Wegrzyn, EnTAP: Bringing faster and smarter functional annotation to non-model eukaryotic transcriptomes. *Mol Ecol Resour* **20**, 591–604 (2020).
67. R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, C. Kingsford, Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* **14**, 417–419 (2017).
68. V. Demichev, C. B. Messner, S. I. Vernardis, K. S. Lilley, M. Ralser, DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nat Methods* **17**, 41–44 (2020).
69. Y. Zhu, L. M. Orre, Y. Zhou Tran, G. Mermelekas, H. J. Johansson, A. Malyutina, S. Anders, J. Lehtiö, DEqMS: a method for accurate variance estimation in differential protein expression analysis. *Molecular & Cellular Proteomics* **19**, 1047–1057 (2020).
70. C. Farrell, M. Thompson, A. Tosevska, A. Oyetunde, M. Pellegrini, BiSulfite Bolt: A bisulfite sequencing analysis platform. *Gigascience* **10**, giab033 (2021).
71. K. D. Hansen, B. Langmead, R. A. Irizarry, BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biol* **13**, R83 (2012).
72. C.-S. Teng, B.-H. Wu, M.-R. Yen, P.-Y. Chen, MethGET: web-based bioinformatics software for correlating genome-wide DNA methylation and gene expression. *BMC Genomics* **21**, 1–10 (2020).
73. H. Wu, C. Wang, Z. Wu, A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. *Biostatistics* **14**, 232–243 (2013).
74. T. Wu, E. Hu, S. Xu, M. Chen, P. Guo, Z. Dai, T. Feng, L. Zhou, W. Tang, L. Zhan, X. Fu, S. Liu, X. Bo, G. Yu, ClusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation* **2**, 100141 (2021).
75. A. Alexa, J. Rahnenfuhrer, topGO: Enrichment Analysis for Gene Ontology. [Preprint] (2021).
76. G. Wang, D.-H. Oh, M. Dassanayake, GOMCL: a toolkit to cluster, evaluate, and extract non-redundant associations of Gene Ontology-based functions. *BMC Bioinformatics* **21**, 139 (2020).

Acknowledgments

We thank Pui-Yan Kwok, Hsiao-Jung Kao, and Hsiao-Huei Chen for their generous help with the Bionano optical mapping, Irina V. Agarkova and James L. Van Etten for advice on HMW DNA extraction, Chih-Horng Kuo for sharing their computing equipment, and Assaf Vardi for sharing the cultures of RCC1217 and RCC1216. We are also grateful for the sequencing services and technical support provided by BRCAS NGS High Throughput Genomic Core, IPMB Genomic Technology Core, National Center for Genomic

Medicines, and Biotools. We acknowledge Yu-Fang Tseng of IPMB Cell Biology Core for flow cytometry analysis and sorting services, Wann-Neng Jane and the staff of the EM division for EM sample preparation, Chuan-Chih Hsu and IPMB Proteomics Core for their assistance with global proteomic quantification, and the National Center for High-performance Computing and IPMB Bioinformatics Core Lab for computer time and facilities. We appreciate the consultation on genome-wide methylation analyses by the lab of Pao-Yang Chen, advice on genome assembling strategies by the lab of Chung-Yen Lin, and the suggestions of Michael Borg on the manuscript,. Lastly, we thank the help of other lab members with this project, especially Bo-Wei Wang and Elias Miguel Barreira.

Funding:
This work was supported by the intramural funding of the Institute of Plant and Microbial Biology (CK), Academia Sinica Career Development Award (AS-CDA-110-L01 to CK), Academia Sinica Postdoctoral Scholar Program (AS-PD-11001-L007 and AS-PD-11201-L01 to T.K.), and National Science and Technology Council, Taiwan (108-2311-B-001-040-MY3 and 111-2611-M001-008-MY3 to C.K.). The funding bodies had no role in the design of the study, in the collection, analysis, and interpretation of data, or in writing the manuscript.

Author contributions:
Conceptualization: CK and MF. Methodology: TK, ML, and CL. Investigation: TK, ML, and TW. Formal Analysis: TK. Visualization: TK. Data curation: TK, TW, ML, and CK. Supervision: CK. Writing—original draft: TK, CK. Writing—review & editing: TK, CK, MF, TW, ML, and CL.

Figures and Tables

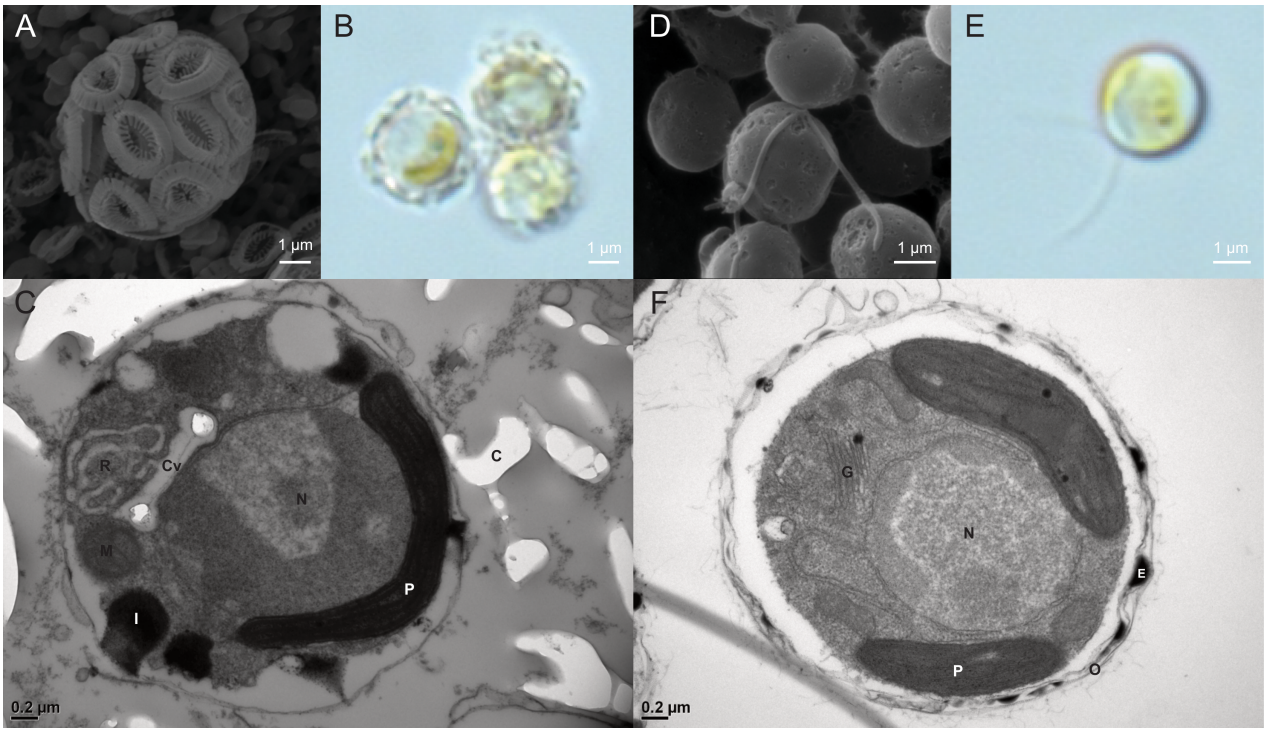

Fig. 1. Morphology of *Gephyrocapsa huxleyi* RCC1216-G9 (diploid) and RCC1217-F3 (haploid). Scanning electron microscopy (**A** and **D**), differential interference contrast (**B** and **E**), and transmission electron microscopy (**C** and **F**) images of G9 and F3. C: empty space created by a fallen 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.

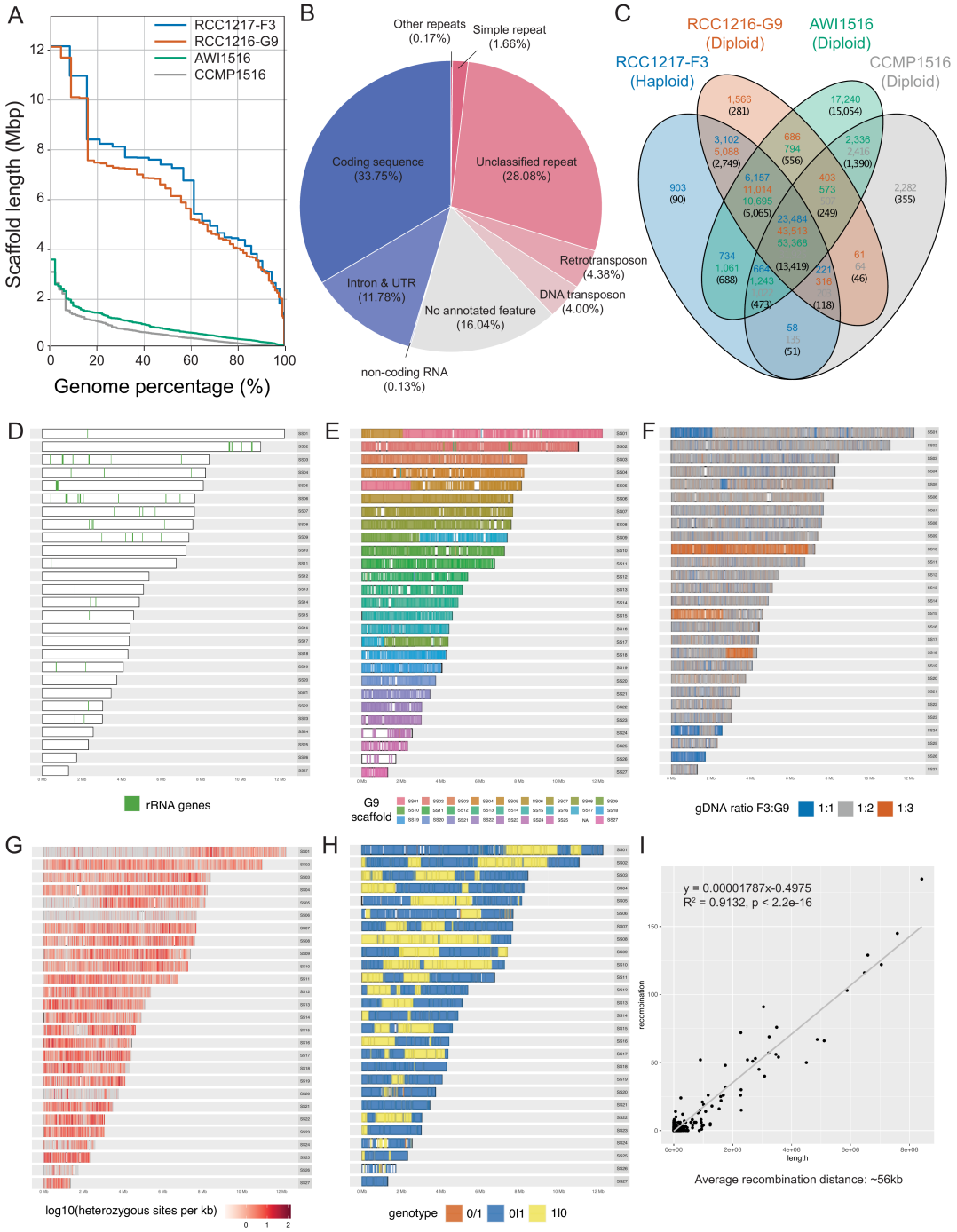

Fig. 2. Genome assemblies, genomic features, structure variations, heterozygosity, and recombination rate. (**A**) N statistics (scaffold length at different percentages of genome assembly) of RCC1217-F3 and RCC1216-G9 (*G. huxleyi*) compared with CCMP1516 and AWI1516 (*G. pseudohuxleyi*). (**B**) Relative length percentage of coding and non-coding sequences of genes (blues) and different classes of repeats (reds) in F3. (**C**) Numbers of genes and orthogroups (in parenthesis) unique to or shared among the strains. (**D**) Distribution of rDNAs in F3. (**E**) Mapping of G9 superscaffolds to the F3 genome, revealing interchromosomal recombinations and large-scale indels. White: unmappable regions. (**F**) Distribution of gene copy number variations across the genome, showing partial aneuploidy. (**G**) Frequency of G9 heterozygous sites across the genome, with lower heterozygosity in certain

genomic regions. (**H**) Genomic regions of F3 that have been phased into different G9 alleles (blue and yellow). Orange: un-phased regions in phasing blocks. White: regions outside of phasing blocks. (**I**) Correlation between the length of phasing blocks and the number of recombination sites.

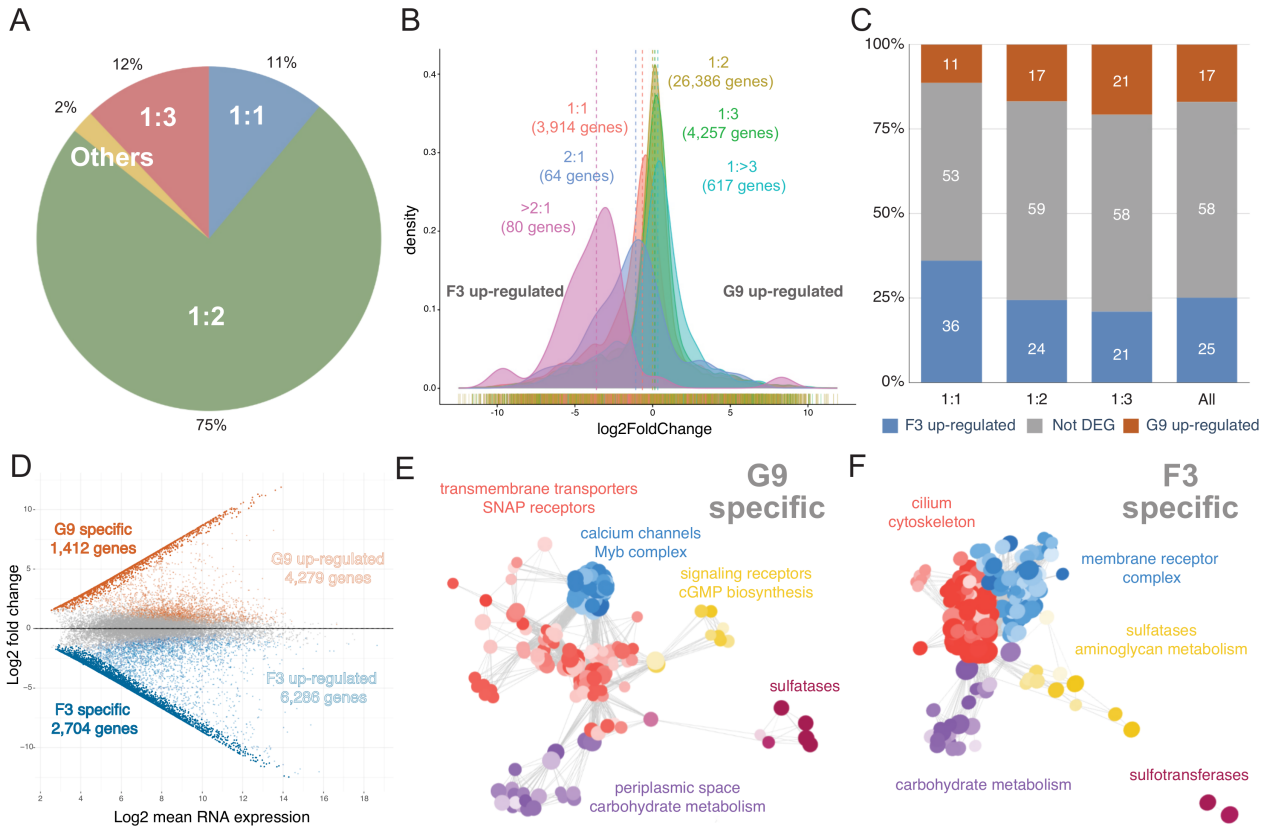

Fig. 3. Gene copy number variation and transcriptomic differentiation. (**A**) Relative abundance of genes with different G9:F3 genomic DNA (gDNA) ratios. (**B**) Distribution patterns of RNA expression ratios for gene groups with different gDNA ratios. (**C**) Percentage of up- and down-regulated genes 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.

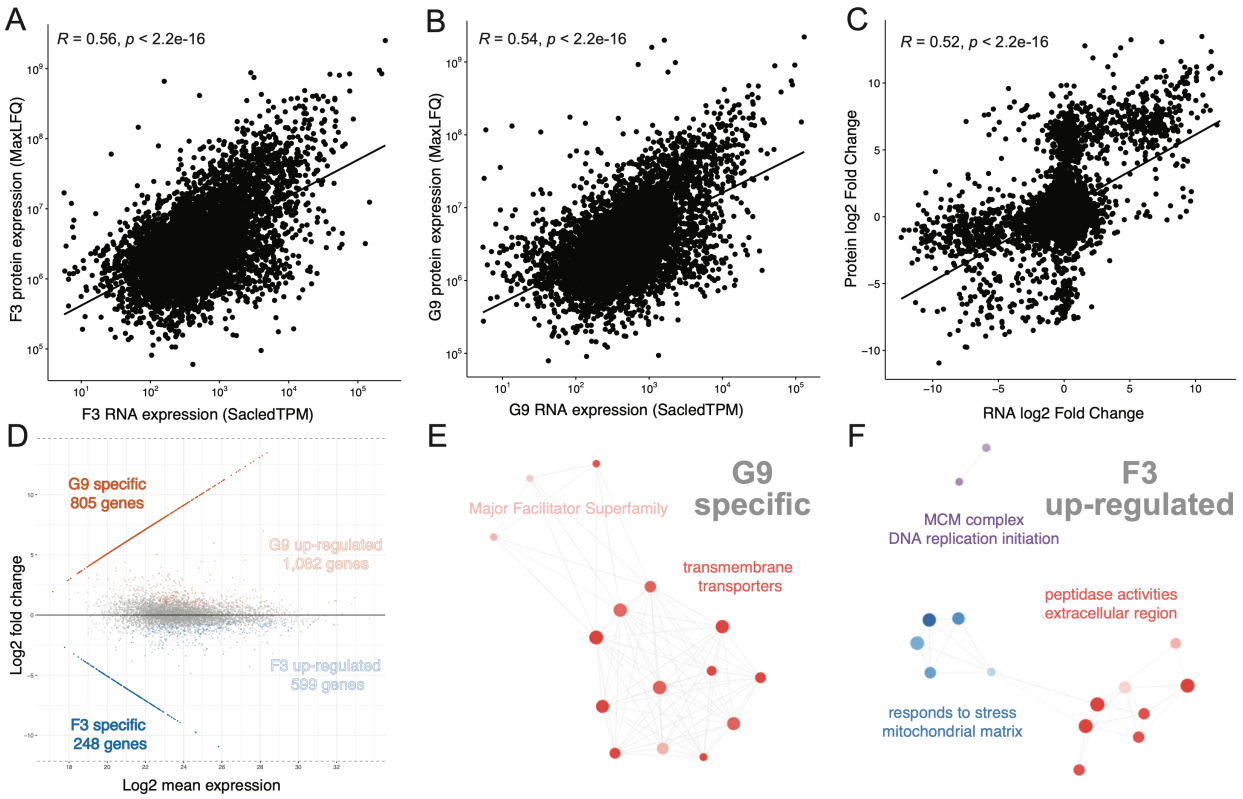

Fig. 4. Protein-RNA correlations and proteomic differentiation. (**A, B**) Correlation between RNA and protein expression in F3(**A**) and G9 (**B**). (**C**) Correlation between RNA and protein $\log_2$ fold change between F3 and G9. (**D**) MA plot of protein expression, highlighting genes specifically 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.

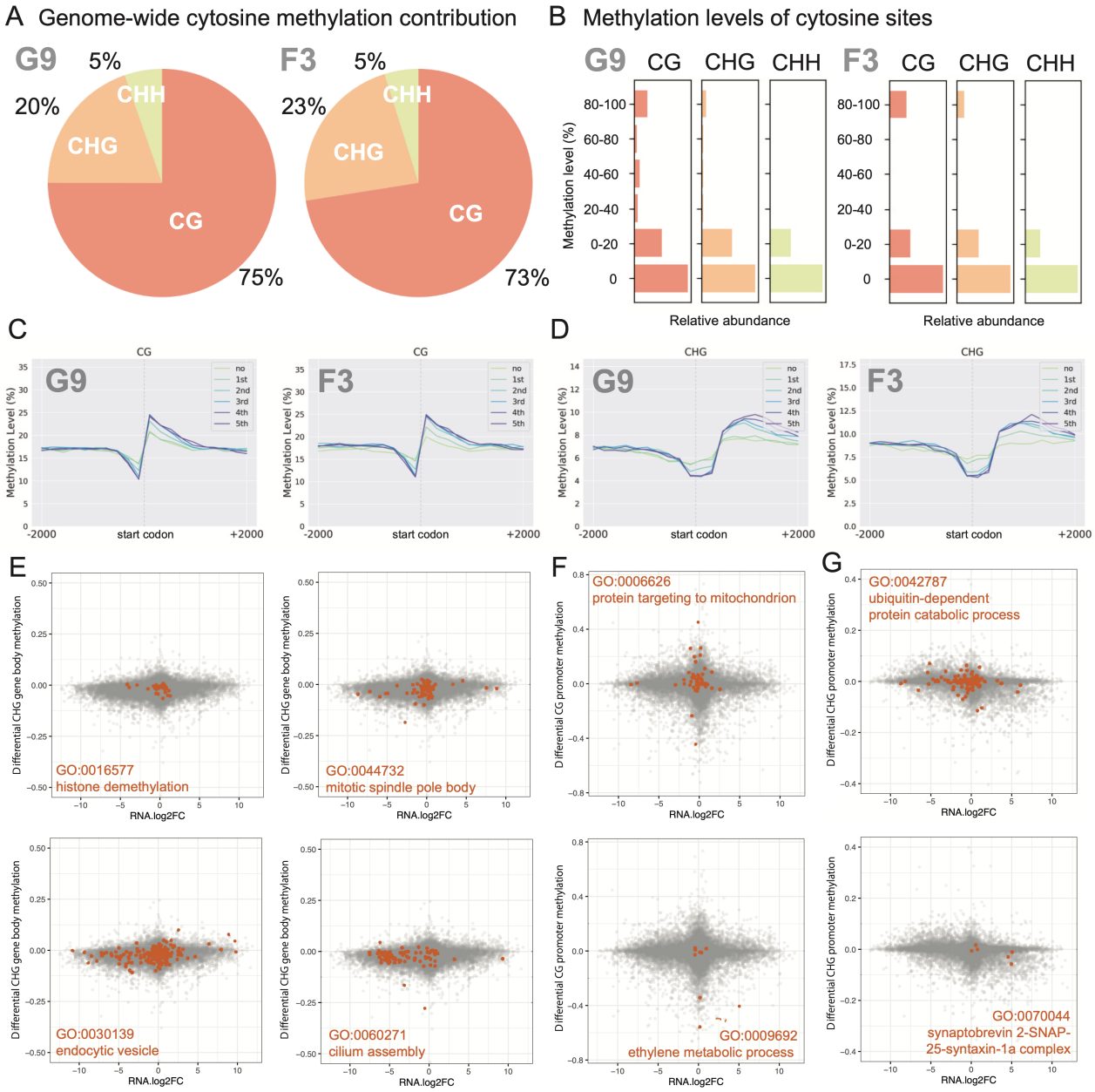

Fig. 5. Genome-wide methylcytosine (mC) patterns, levels, and correlations with RNA expression. (**A**) Percentage of mC contribution in different sequence contexts: CG, CHG, and CHH. (**B**) Relative abundance of sites with different mC levels in CG, CHG, and CHH sequence contexts. (**C**, **D**) Average 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).

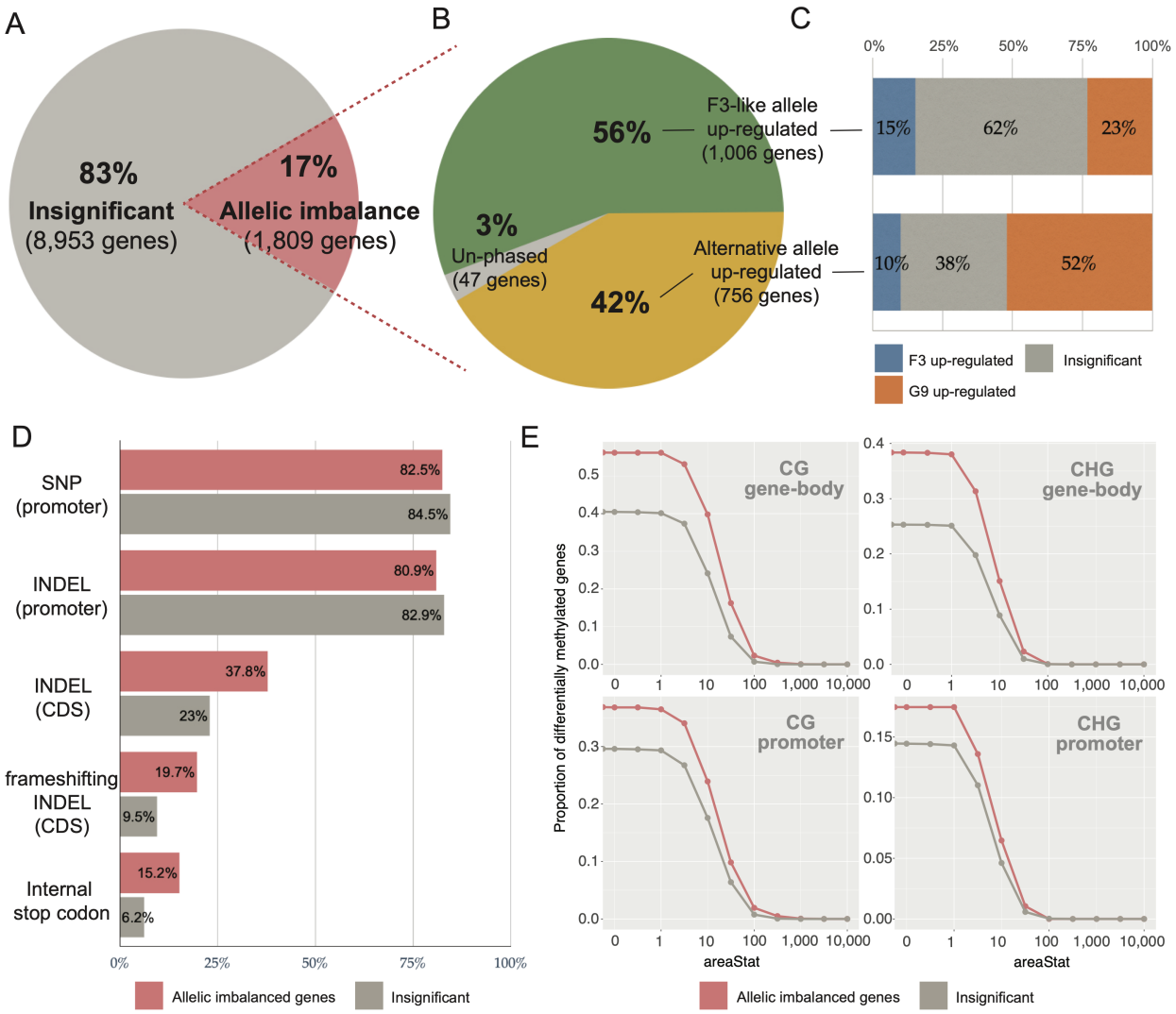

Fig. 6. Allelic imbalance in RNA expression and its relationships with genomic variation and allele-specific methylation. (**A**) Percentage of allelic imbalanced genes among the heterozygous genes with exactly two copies in the G9 genome and with detectable transcripts. (**B**) Percentage of genes with 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.

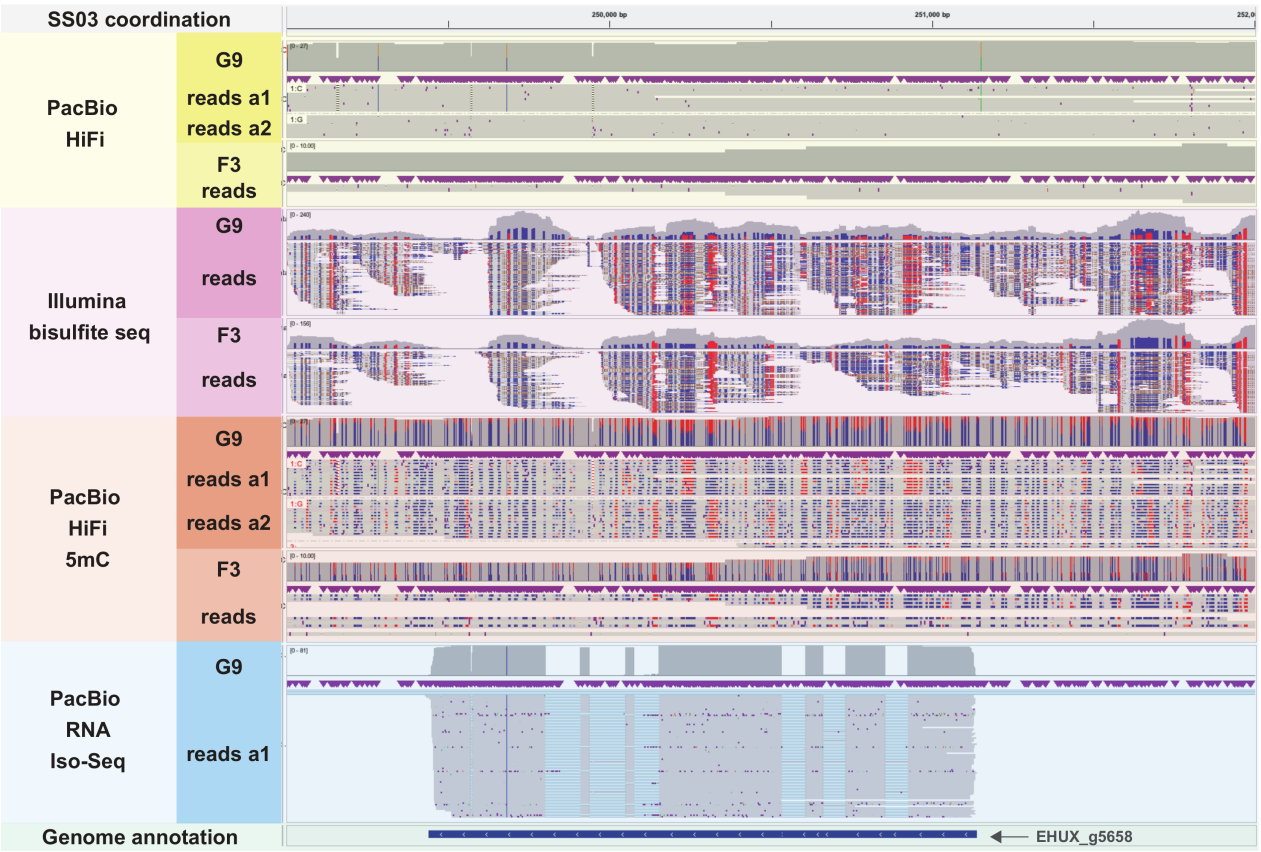

Fig. 7. Allele-specific methylation and allele-imbalanced RNA expression in a *G. huxleyi* gene. Genomic, methylomic, and transcriptomic reads were mapped onto SS03 (249,000–252,000) containing EHUX_g5658, a sulfatase gene located on the Crick strand. Track yellow (top): PacBio HiFi reads (variable sites: 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.

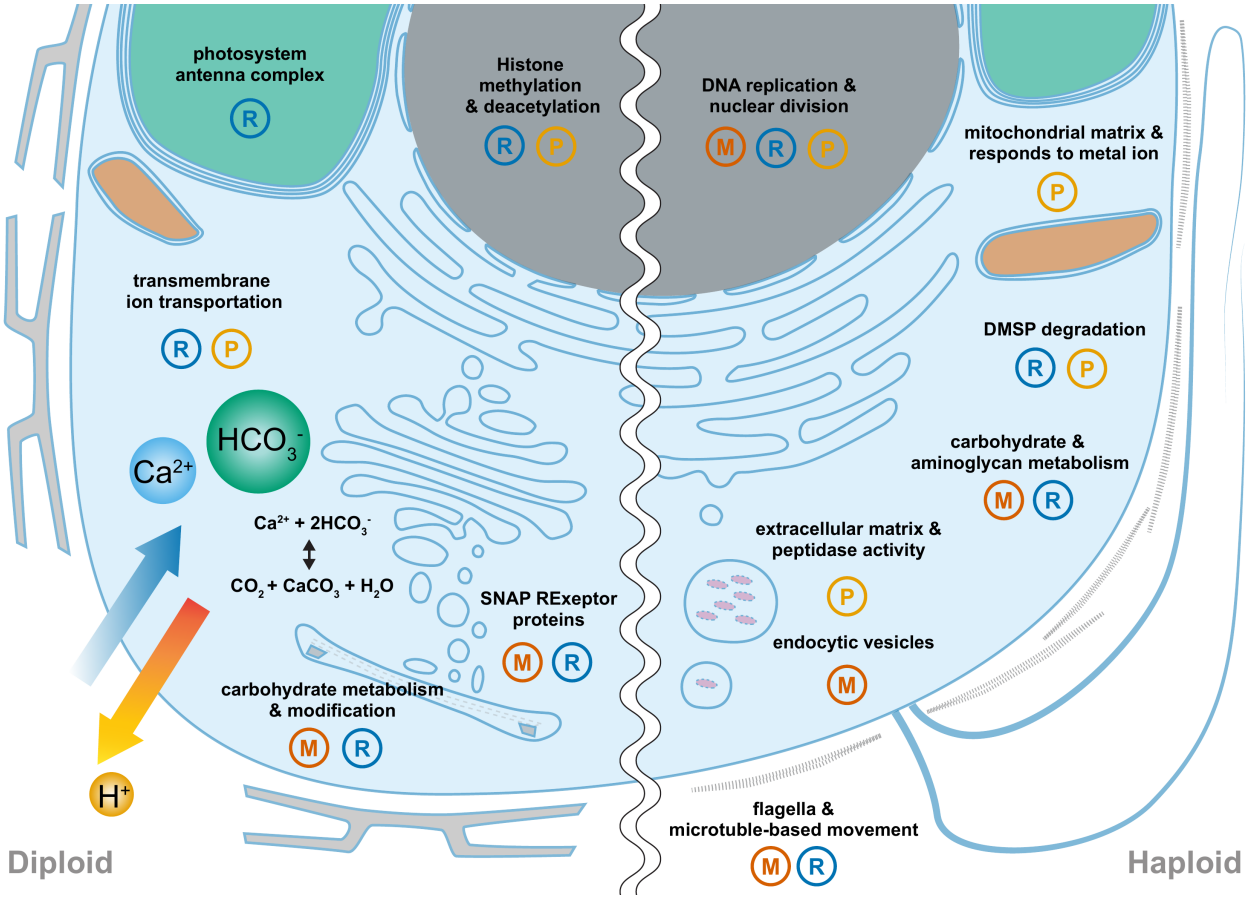

Fig. 8. Summary of molecular differentiation between the diploid (left) and haploid (right) cell. The types of data supporting each functional category are indicated. M: DNA cytosine methylation; R: mRNA expression; P: protein expression.

902 **Table 1. Statistics of the haploid and diploid genome assemblies of *Gephyrocapsa* spp.**
903

	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

904
905
906
907
908