BMC Genomics

RESEARCH

Open Access

Comparative genomics of microalgae: molecular mechanisms underlying efficient CO_2 assimilation

Jessica Di Martino[1], Franco Liberati[1], Meryam Carrus[1], Manuel Arcieri[2], Lorenzo Arcioni[2], Paolo Bottoni[2], Claudia Cafaro[3], Loretta De Giorgi[3], Antonio Fardelli[3], Paolo Franchini[1], Daniele Canestrelli[1] and Tiziana Castrignanò[1*]

Abstract

Rising atmospheric CO_2 levels necessitate the development of effective carbon capture strategies. Some species of microalgae, with their high growth rates and exceptional CO_2 fixation efficiency—up to 10–50 times greater than terrestrial plants—represent a promising biological resource. Microalgae, however, constitute a heterogeneous group distributed across different taxonomic lineages, which complicates the identification of conserved molecular traits. In this study, we conducted a comparative genomics analysis to elucidate the molecular mechanisms underlying efficient CO_2 assimilation in 29 selected species of microalgae. We curated a dataset of high-quality, annotated microalgal genomes from the NCBI Genome database, retaining only those with BUSCO gene completeness above 80%. Orthogroups were identified using OrthoFinder and annotated. A comprehensive phylogenetic tree was constructed based on single-copy orthologs. Additionally, CAFÉ was employed to assess gene family expansion and contraction in key functional categories, including light-harvesting, carbon fixation, and stress response. Our findings reveal a conserved core of functional genes across microalgae, alongside significant interspecific variability in gene copy numbers and orthogroup composition. These insights provide a comparative genomic framework to support future investigations integrating physiological and molecular data aimed at understanding CO_2 assimilation mechanisms in microalgae.

Keywords Comparative genomics, Microalgae CO_2 assimilation, Photosynthetic efficiency

Background

Atmospheric CO_2 concentrations have risen rapidly, increasing from 280 ppm in 1850 [1] to over 417 ppm in 2021, with more than half of this increase occurring in the last 30 years [2]. Currently, CO_2 levels exceed those recorded over at least the past 800,000 years [3]. Human activities now release approximately 35 gigatonnes of CO_2 per year—a dramatic increase compared to nearly 10 megatonnes two centuries ago [4]. This sharp rise in atmospheric CO_2 levels has intensified the need for efficient strategies to capture and assimilate carbon dioxide from the environment [5].

*Correspondence:
Tiziana Castrignanò
tiziana.castrignano@unitus.it
[1]Department of Ecological and Biological Sciences, University of Tuscia, Viale dell'Università, snc, Viterbo 01100, Italy
[2]Department of Computer Science, Sapienza University of Rome, Viale Regina Elena 295, Roma 00161, Italy
[3]Division of Rome, CNR-Institute of Atmospheric Pollution Research, Via Cristoforo Colombo 44, Rome IT-00147, Italy

Based on future projections of CO_2 concentrations under Representative Concentration Pathways (RCPs) [6, 7], innovative biological systems for carbon assimilation are becoming increasingly relevant. Such systems can provide effective CO_2 management without contributing to additional environmental burdens [6, 8].

Various strategies have been explored to remove atmospheric CO_2, which can be broadly classified into three main groups: (1) chemical methods, such as washing with alkaline solutions and coating activated carbon with amines; (2) direct approaches, for example, CO_2 injection into the subsurface or the oceans; and (3) biological CO_2 removal, which involves the conversion of CO_2 into organic compounds [9]. Among these, the use of microalgae for CO_2 fixation has attracted considerable attention since the 1990s, offering the opportunity to transform absorbed CO_2 into biodiesel and bioenergy. This biological approach represents an emerging system for efficient atmospheric CO_2 capture. Microalgae and cyanobacteria can grow much faster than terrestrial plants, with CO_2 fixation efficiencies estimated to be 10–50 times higher [10]. Moreover, photosynthetic microorganisms, particularly when cultivated in wastewater, can grow rapidly, produce valuable compounds, and be easily harvested [11]. However, the rate of CO_2 fixation by microalgae is influenced by various factors, including CO_2 concentration, light intensity, growth substrates, and cultivation techniques. Over the years, numerous methods have been developed to assess the potential of microalgae in CO_2 fixation, both by measuring biomass or growth rate and by directly quantifying CO_2 consumption [12–14].

This study fits into this context by focusing on the comparative genomics of microalgae to investigate the evolution of molecular mechanisms underlying their efficient CO_2 assimilation. By utilizing genome microalgae datasets and advanced bioinformatic tools (BUSCO, Orthofinder, IQTree, CAFE5), we assessed completeness in terms of expected gene content and filtered those genomes of high quality (gene completeness above 80%), identified and annotated the orthogroups among the filtered 29 species, and analyzed variations in the gene repertoire among them. Our results reveal both a conserved core of essential functions and species-specific evolutionary strategies that contribute to their CO_2 capture efficiency. These findings offer a comparative genomic framework that can support future studies integrating physiological data to better understand the molecular basis of CO_2 assimilation in microalgae.

Results and discussion

The comparative genomics analysis presented in this study can be divided into six stages: (1) selection of high-quality microalgae genomes to serve as input for the comparative genomics analysis; (2) investigation of orthologous genes among the genomes of the 29 microalgae species belonging to the three clades - Chlorophyta, Stramenopiles and Alveolata (selected in stage 1); (3) identification of orthogroups; (4) construction of the phylogenetic tree; (5) annotation of gene orthologs; (6) analysis of gene expansions-contractions in microalgae species, with a specific focus on genes associated with CO_2 capture.

Collection and selection of microalgal genomes for comparative analysis

The majority of species of microalgae relevant for CO_2 sequestration belong to four clades: Chlorophyta, Stramenopiles, Euglenozoa, and Alveolata [15–18]. Although "microalgae" is used as a descriptive term rather than representing a single monophyletic taxonomic group, classifying our microalgae at the clade level enables uniform comparisons of species from highly diverse evolutionary lineages. This approach is supported by recent taxonomic reviews [19, 20]. The bioinformatics pipeline illustrated in Fig. 1 outlines the process used to select a set of high-quality microalgal genomes, utilised for further comparative genomics.

The workflow begins with a phase of scientific curation, during which an extensive literature review was conducted to identify microalgae species that have been experimentally demonstrated to exhibit high CO_2 capture efficiency (see Table 1 - Datafile 0) [21–24]. Subsequently, the NCBI Genome database (https://www.ncbi.nlm.nih.gov/home/genomes/) was queried to retrieve all the genomes of species belonging to the four clades of interest (Alveolata, Chlorophyta, Stramenopiles, and Euglenozoa - TXT file named Datafile 1). From the list of genomes only those associated with microalgal species were selected (generating a TSV file, one sheet for each clade - Datafile 2). An additional selection step identified genomes annotated in GFF format (all records are reported in Excel spreadsheet - Datafile 3).

Assessment of genome completeness using BUSCO

Comparative genomics analyses are often sensitive to the presence of incomplete data, making the selection of high-quality datasets from representative species a critical step. This task becomes increasingly challenging as the volume of available genomic data grows, particularly given the substantial qualitative variability among them. Quantifying gene completeness allows us to identify the most reliable assemblies.

To this end, we employed the Benchmarking Universal Single-Copy Ortholog (BUSCO; http://busco.ezlab.org) [25] method to obtain intuitive quantitative measures of genome completeness by evaluating the expected gene content. The BUSCO analysis was applied to all genome accessions listed in Datafile 3, which comprised data for

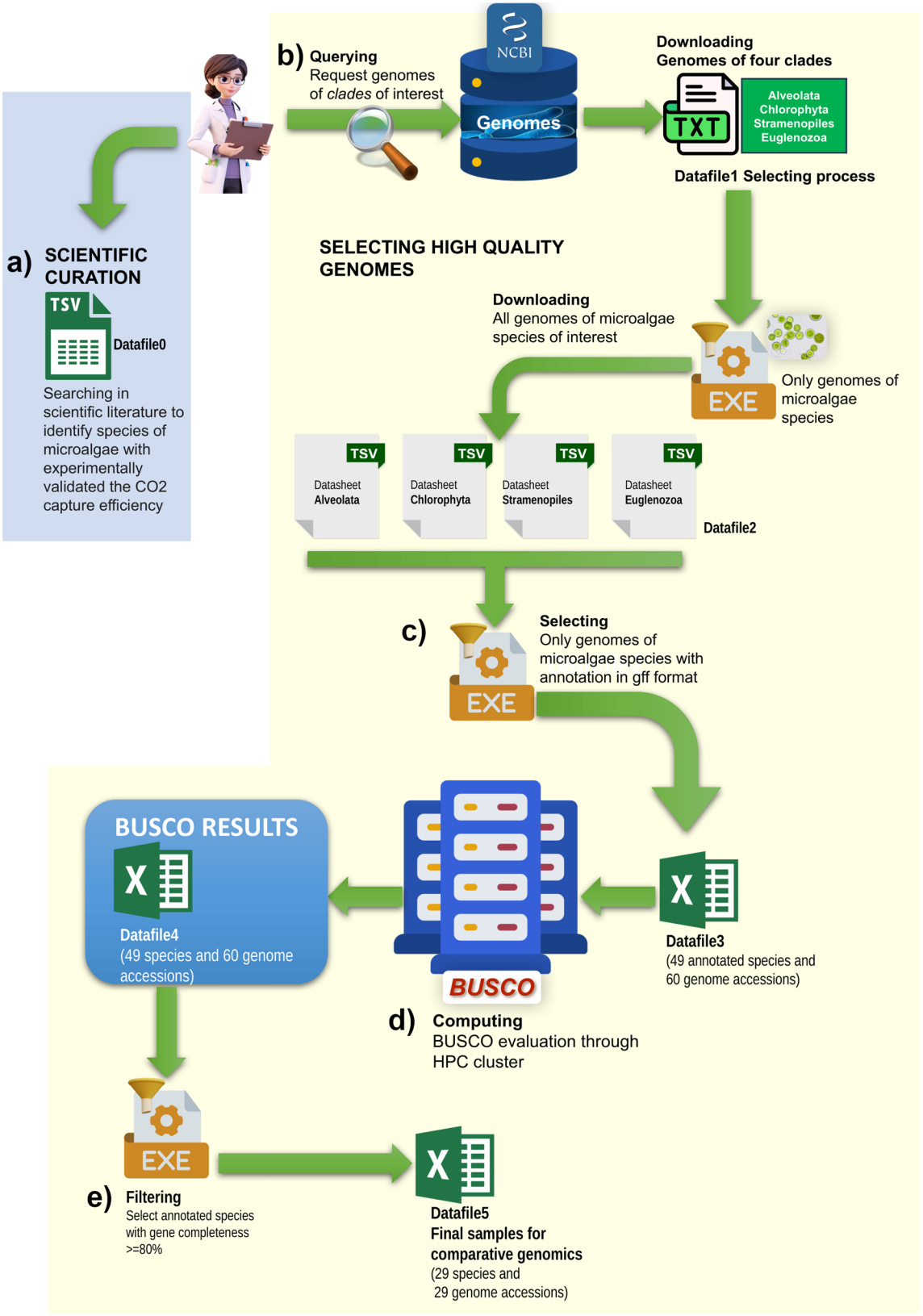

Fig. 1 (See legend on next page.)

(See figure on previous page.)

Fig. 1 Workflow for the selection of high-quality microalgae genomes for further phase of comparative genomics. It consists of the following steps: **a** Identification of species with validated CO_2 capture efficiency (Datafile 0) from scientific literature; **b** Querying the NCBI Genome database to retrieve genomes of species from four selected clades (Chlorophyta, Stramenopiles, Euglenozoa, and Alveolata); **c** Filtering genomes based on the availability of an annotation file in GFF format; **d** Assessing gene completeness using BUSCO on an HPC cluster; **e** Selecting genomes with completeness ≥ 80% based on BUSCO results. The final dataset comprises 29 species and genome accessions suitable for comparative genomic analyses

49 annotated species and 60 genome accessions. Subsequently, for each species, we selected the best genome assembly, defined as the one with the highest level of gene completeness. The results of this phase are documented in Datafile 4.

By setting a gene completeness threshold of 80% or higher, only 29 species belonging to the three clades Chlorophyta, Stramenopiles, and Alveolata met this criterion. As illustrated in Fig. 2A, which displays the clades and species whose genomes exceed the 80% threshold in terms of complete gene content. Figure 2B shows that only 15 species achieved a completeness level above 95%.

This rigorous assessment of gene completeness has ensured the use of high-quality datasets.

In particular, as shown in Fig. 2B, within the clade Alveolata only a single species exhibits a genome with a completeness level significantly exceeding 80% and even surpassing 95%, whereas numerous species belonging to the clade Chlorophyta and Stramenopiles have exceeded this threshold of excellence.

Figure 3 displays graphical representations of the detailed BUSCO results, illustrating the completeness (C), fragmentation (F), and missing ortholog (M) scores for the 29 selected species in a color-coded format.

Furthermore, for each of the best genome assemblies, XLSX files containing additional species-specific details were uploaded to Figshare, with distinctions made between complete single-copy (S) and complete duplicated (D) genes (see Table 1, Datafile 4).

The overall assembly statistics and BUSCO assessment results for the analyzed microalgal genomes are summarized in Table 2. This table provides detailed information for each species, including clade, organism name, assembly accession, assembly level, genome size, N50, and BUSCO scores (complete, single-copy complete, duplicated complete, fragmented, and missing) based on the relevant BUSCO database (Alveolata_odb10, Chlorophyta_odb10, and Stramenopiles_odb10). For example, within Alveolata, Vitrella brassicaformis (accession GCA_001179505.1) shows scaffold-level assembly with a genome size of 72.7 Mb, an N50 of 147.6 kb, and a BUSCO completeness of 96% (with 94.7% single-copy genes). Similarly, for Chlorophyta, Micromonas commoda (accession GCA_000090985.2) is represented by a complete genome assembly of 21 Mb, an N50 of 1.4 Mb and a BUSCO completeness of 99%, while Chloropicon primus shows a BUSCO completeness of 90% despite a lower percentage of complete genes. The table also reports assembly metrics for Stramenopiles species, such as Nitzschia inconspicua, which stands out with a nearly complete BUSCO score (100% completeness) and a contig-level assembly of 99.7 Mb with an N50 of 3.6 Mb.

These data not only corroborate the overall quality of the genome assemblies used in our comparative analysis, but also highlight the variability in assembly quality and completeness among different taxonomic groups.

Identification, annotation, and phylogenomic analysis of orthogroups

Orthogroups were identified using OrthoFinder [26], a tool capable of grouping protein sequences into orthogroups and detecting orthologous relationships by constructing gene trees for each group, ultimately yielding a rooted species tree.

The analysis was performed on the 29 species selected during the initial phase using the proteomes derived from the selected genomes, with a particular focus on single-copy orthologs for downstream analyses. This approach minimizes biases arising from non-coding regions or highly variable genomic sequences. The resulting orthogroups were then used to infer functional annotations and serve as a foundation for evolutionary studies.

The analysis generated numerous outputs, including a species tree (Fig. 4) and detailed orthology visualizations (Figs. 5 and 6). In Figs. 5a and 6a, each subtree represents a clade-specific topology derived from single-copy orthologs. Although the trees are displayed as composite layouts for clarity, the underlying topologies are supported by bootstrap values. To illustrate the evolutionary relationships among the analyzed species, a phylogenomic tree was constructed using only single-copy orthologs. The tree was rooted with the two species, Talaromyces marneffei and Dictyostelium discoideum AX4, as outgroups, thereby providing a clear delineation of phylogenetic relationships. The resulting phylogram (Datafile 6) (Fig. 4) reveals a distinct separation among clades, with the Alveolata group being phylogenetically closer to the Stramenopiles than to the Chlorophyta. Bootstrap values of 100 at the major nodes confirm the high statistical support, attesting to the robustness of the inferred evolutionary relationships. This tree represents the final phylogenomic result, providing a robust framework to explore macroevolutionary trends and to highlight divergence and speciation events within the analyzed clade.

Subsequently, to obtain specific gene ontology results, we divided the Orthofinder analysis into two groups

Table 1 Overview of all the data files produced in this article and their access on figshare

Label	Name of data	File types	Data repository (URL)
Datafile 0	Species selected according to the role of CO2 absorption present in scientific literature	Sheet file (.tsv)	https://doi.org/10.6084/m9.figshare.28284869
Datafile 1	Species of interest included in: *Alveolata*, *Chlorophyta*, *Euglenozoa*, *Stramenopiles*	Sheet file (.xlsx)	https://doi.org/10.6084/m9.figshare.27930891
Datafile 2	All microalgae species with genome accession and annotated/ not annotated	Sheet file (.xlsx)	https://doi.org/10.6084/m9.figshare.27930894
Datafile 3	BUSCO output (141 genomes)	Sheet file (.xlsx)	https://doi.org/10.6084/m9.figshare.27930906
Datafile 4	Summary BUSCO results (85 species and 85 genome accessions)	Sheet file (.xlsx)	https://doi.org/10.6084/m9.figshare.27930912
Datafile 5	Assembly accessions	Sheet file (.xlsx)	https://doi.org/10.6084/m9.figshare.25914637
Datafile 6	ORTHOFINDER results and phylogenetic tree	Folder	https://doi.org/10.6084/m9.figshare.28417424
Datafile 7	Input file for CAFE	.tsv and .tree	https://doi.org/10.6084/m9.figshare.28436075
Datafile 8	CAFE result (expansion and contraction)	Folder	https://doi.org/10.6084/m9.figshare.28430039
Datafile 9	Functional distribution of expanded orthogroups identified in Chlorophyta and stramenopiles-alevolata	Sheet file (.xlsx)	https://doi.org/10.6084/m9.figshare.28430045
Datafile 10	All Figures	Picture (.png)	https://doi.org/10.6084/m9.figshare.28436054

of species: one comprising the 16 Chlorophyta species (CHLO group corresponding to the green species in Fig. 4) and another including the 12 Stramenopiles and 1 Alveolata species (STR-A group corresponding to 12 species of Stramenopiles - those shown in light blue in Fig. 4 - plus one species of Alveolata shown in violet in Fig. 4). because it is evolutionarily closer to the Stramenopiles than to the Chlorophyta. This division was implemented to account for potential differences in genomic architecture and evolutionary dynamics among these taxonomic groups [27, 28] and to facilitate a more focused interpretation of the functional annotations related to CO_2 capture efficiency. Consequently, Figs. 5 and 6 present separate phylogenetic and orthologous visualizations of Orthofinder results for the CHLO group and for the STR-A group respectively.

The analysis conducted on the sixteen selected species belonging to clade Chlorophyta, and their phylogenetic tree was used to guide the identification and classification of potential orthologous genes (Fig. 5). Figure 5a displays the phylogenetic relationships among Chlorophyta species based on single-copy orthologs. The tree layout reflects the output of OrthoFinder and is optimized for visualization, which may result in apparent topological differences between clades. However, all relationships are derived from a single consensus tree and are supported by high bootstrap values. In the OrthoFinder analysis, 90.2% of all genes were assigned to 18,754 orthogroups (Datafile 6. Orthofinder tables Chlorophyta) (Fig. 5c). 50% of the genes assigned to orthogroups are found in groups comprising at least 16 genes (G50 for assigned genes = 16), whereas over 10% of the orthogroups contain all the analyzed species (n = 1,872; Fig. 5d), of which 18.5% are single-copy orthogroups (n = 347; Datafile 6 Orthofinder tables Chlorophyta; Fig. 5f).

These data demonstrate a high degree of conservation of functional genes within Chlorophyta species, suggesting the existence of a shared orthologous core that likely includes the essential functions for photosynthesis and other basic cellular activities [21, 29]. However, when examining interspecific variability (Fig. 5c), it becomes evident that the percentage of genes assigned to the different orthogroups varies considerably among species. For example, some species exhibit high values of variability (up to 96.6% in Micractinium_conductrix_1), whereas others display lower percentages (such as 79.5% in Monoraphidium_minutum_1). This variability may be influenced by differences in genome annotation completeness; however, we cannot exclude the contribution of evolutionary processes that shaped species-specific genomic repertoires.

Additional insights emerge from the analysis of the number of orthogroups containing each species (Fig. 5b): the values range from 5,675 orthogroups in Chloropicon primus (accounting for 30.3% of the total orthogroups) to 9,482 in Volvox reticuliferus (50.6%), indicating that some species share a broader core of genomic functions than others. Similarly, the analysis of species-specific orthogroups (Fig. 5d) reveals considerable diversity, with the number of orthogroups specific to each species varying from only 43 in Bathycoccus prasinos_4 (2.5% of the species' genes) to 587 in Tetradesmus obliquus_5 (13.8%). These data indicate the presence of species-specific orthogroups, which could reflect novel functions or adaptations; however, it remains necessary to disentangle these patterns from potential differences in genome annotation quality (Datafile 6. Orthofinder tables Chlorophyta).

Finally, further details on the orthogroup repertoire are illustrated in Fig. 5e and f, and 5g, which show, respectively, the distribution of genes with orthologs in all species, the comparison of orthologous multiplicities relative to Micractinium commoda, and the analysis of species-specific duplications. Micractinium commoda was chosen as the reference species for the comparison

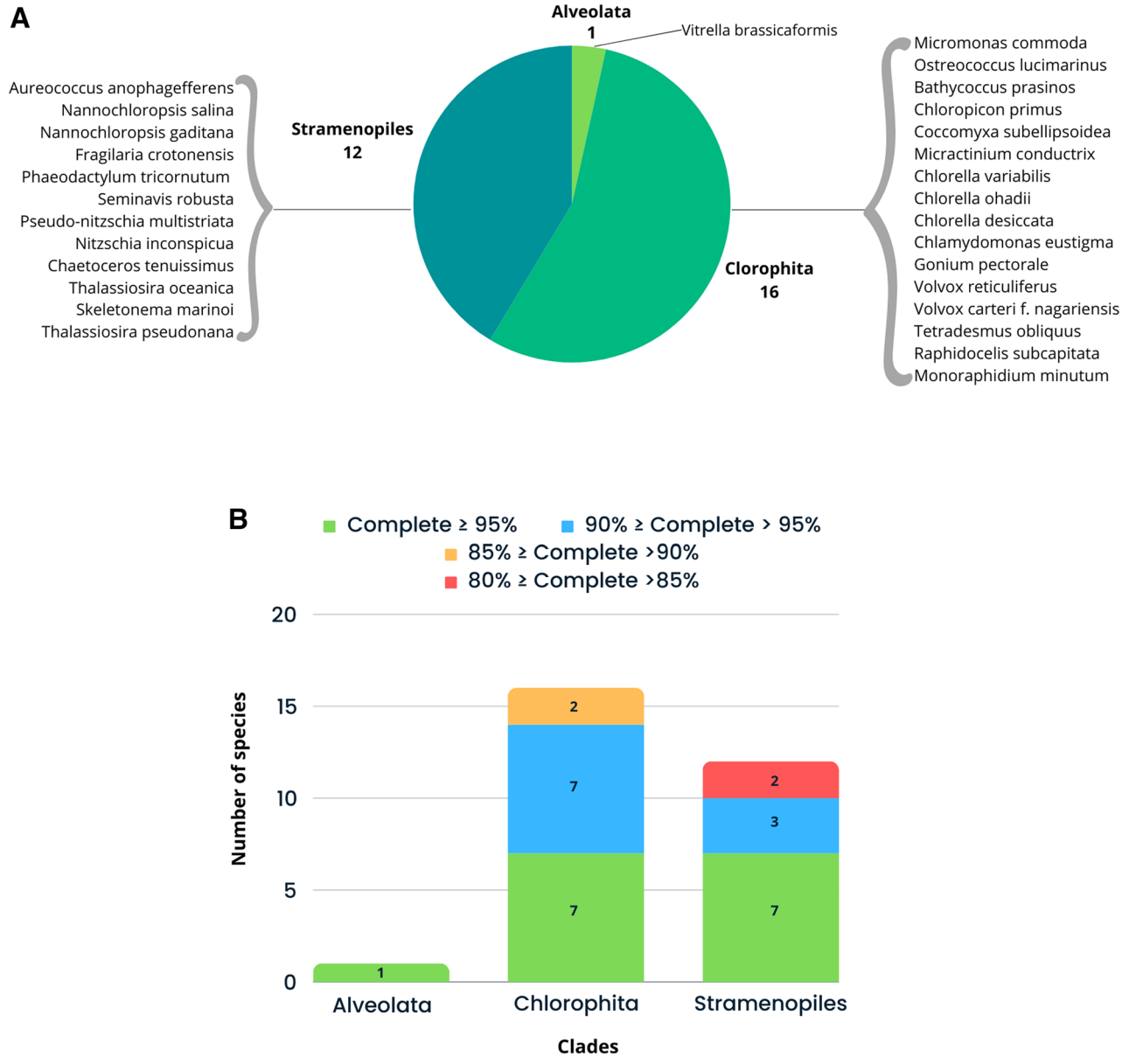

Fig. 2 **A** Species of microalgae with genomes whose level of gene completeness is at least greater than or equal to 80%. **B** Number of species with genomes with gene completeness level classified as follows: greater than or equal to 95% (green bar), between 90% and 95% (blue bar), between 85% and 90% (orange bar), and between 80% and 85% (red bar). In the legend is reported the degree of gene completeness of genome assembly for each microalga species investigated with BUSCO

of orthologous multiplicities in Fig. 5f because, among the Chlorophyta, it is the only species with a "complete genome" assembly and a gene completeness score exceeding 95%, making it the most reliable representative for this type of analysis. These complementary representations provide a comprehensive view of the orthologous gene repertoire and support the interpretation of the data in terms of conservation and diversification of genomic functions among the analyzed species.

For the Stramenopiles group, the analysis was conducted on 12 species, adding Vitrella brassicaformis (Fig. 6a) because it is evolutionarily more closer to the Stramenopiles than to the Chlorophyta (Datafile 6. Orthofinder tables Stramenopiles-Alveolata). This choice may partly account for the considerable variability observed in the data. We obtained a total of 252,503 genes, of which 225,940 (89.5%) were assigned to 26,828 orthogroups (Fig. 6c). The assignment percentages vary markedly among species: for instance, Nitzschia inconspicua exhibits an extremely high assignment rate (98.4%), whereas Vitrella brassicaformis_CCMP3155 reaches a minimum value of 78.0% (Fig. 6b). These data

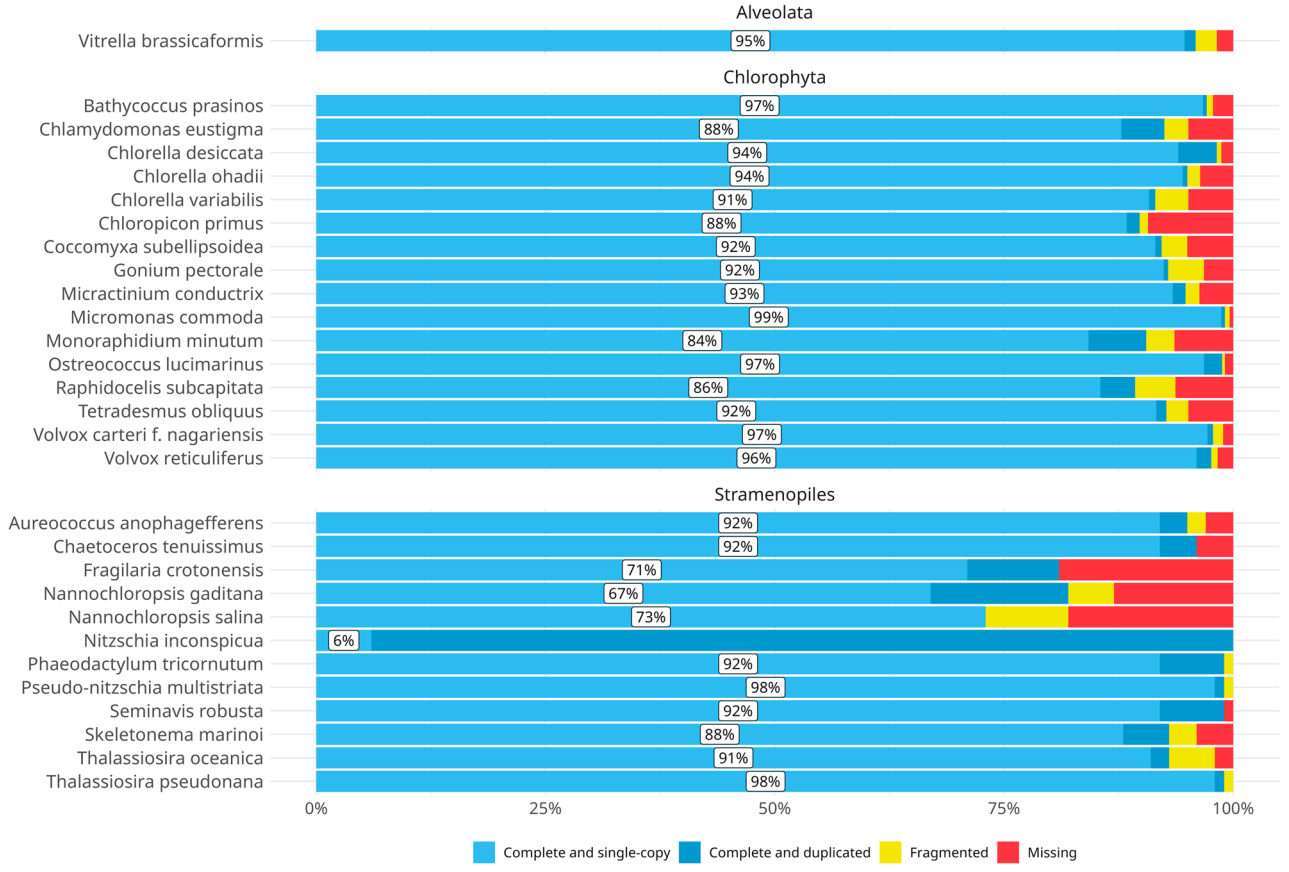

Fig. 3 Results of BUSCO completeness assessments for quality control of genomic data for the microalgae filtered. Bar graphs produced with the BUSCO plotting tool show the proportions classified as complete (C, blue), fragmented (F, yellow) and missing (M, red)

indicate variability in orthogroup coverage among the species analyzed (Fig. 6d), which may partly result from differences in genome completeness and annotation quality; in particular, the inclusion of Vitrella brassicaformis—a more distantly related species—may contribute significantly to this observed variability.

The number of orthogroups in which each species is represented also varies considerably: for example, Aureococcus anophagefferens is present in 5,021 orthogroups (representing 18.7% of the total orthogroups), whereas Nitzschia inconspicua appears in as many as 12,356 orthogroups (46.1%) (Fig. 6b). These differences, compared to Chlorophyta, may reflect variations in genome size rather than functional repertoire breadth, as species with more genes are naturally represented in a larger number of orthogroups.

Another notable aspect concerns species-specific orthogroups: the contribution of these varies significantly, ranging from 2.8% of the genes in Nannochloropsis gaditana_CCMP526 to a remarkable 41.8% in Vitrella brassicaformis_CCMP3155 (Fig. 6d). The high range of variability may be partly driven by the inclusion of Vitrella brassicaformis, a phylogenetically distant species, whose divergent genome likely inflates the number of species-specific orthogroups.

Furthermore, the analysis of the distribution of the number of genes per species in each orthogroup (Fig. 6b) shows that most groups typically contain one or two copies per species, with a tail extending toward larger group sizes, indicating gene duplication events in specific lineages. Finally, the examination of orthogroups shared among species (Fig. 6e and f) reveals that only 482 orthogroups contain genes from all analyzed species, while the number of single-copy orthogroups is extremely low (only 5). For the comparison of orthologous multiplicities shown in Fig. 6f, we selected Thalassiosira pseudonana as the reference species because its genome, assembled at the chromosome level, achieved a gene completeness of 99% according to BUSCO analysis, making it the most reliable representative for this type of comparison.

These results underscore a remarkable evolutionary complexity and a strong diversification of gene repertoires within the Stramenopiles, suggesting that, in addition to a shared functional core, there are specific strategies that contribute to the ecological and functional adaptations of each species.

Table 2 BUSCO validation (v. 5) was applied to the microalgae filtered (annotated and genomes with ≧80% completeness)

Clades	Organism	Assembly Accession	Assembly level	Genome size	N50	BUSCO complete	BUSCO Complete and single-copy (S)	BUSCO Complete and duplicated (D)	BUSCO Fragmented (F)	BUSCO Missing (M)
Alveolata	Vitrella brassicaformis	GCA_001179505.1	Scaffold	72.7 Mb	147.6 kb	96	162 (94.7%)	2 (1.2%)	4 (2.3%)	3 (1.8%)
Chlorophyta	Micromonas commoda	GCA_000090985.2	Complete Genome	21 Mb	1.4 Mb	99	1494 (98.7%)	6 (0.4%)	8 (0,5%)	6 (0,4%)
	Ostreococcus lucimarinus	GCA_000092065.1	Complete Genome	13.2 Mb	708.9 kb	99	1471 (96.8%)	30 (2.0%)	5 (0,3%)	14 (0,9%)
	Bathycoccus prasinos	GCA_002220235.1	Chromosome	15 Mb	955.7 kb	97	1472 (96.7%)	6 (0.4%)	11 (0,7%)	33 (2,2%)
	Chloropicon primus	GCA_023205875.1	Complete Genome	17.6 Mb	1.1 Mb	90	1342 (88.4%)	21 (1.4%)	14 (0,9%)	141 (9,3%)
	Coccomyxa subellipsoidea	GCA_000258705.1	Contig	48.8 Mb	2 Mb	92	1390 (91.5%)	11 (0.7%)	43 (2,8%)	76 (5%)
	Micractinium conductrix	GCA_002245815.2	Contig	60.8 Mb	1.2 Mb	95	1410 (93.4%)	21 (1.4%)	23 (1,5%)	56 (3,7%)
	Chlorella variabilis	GCA_000147415.1	Scaffold	46.2 Mb	27.9 kb	92	1379 (90.8%)	11 (0.7%)	55 (3,6%)	74 (4,9%)
	Chlorella ohadii	GCA_025026875.1	Scaffold	57.1 Mb	329 kb	95	1435 (94.5%)	8 (0.5%)	21 (1,4%)	55 (3,6%)
	Chlorella desiccata	GCA_019202925.1	Contig	20.7 Mb	1.3 Mb	98	1426 (94.0%)	64 (4.2%)	8 (0,5%)	20 (1,3%)
	Chlamydomonas eustigma	GCA_002335675.1	Scaffold	66.6 Mb	465.1 kb	93	1333 (87.8%)	71 (4.7%)	40 (2,6%)	74 (4,9%)
	Gonium pectorale	GCA_001584585.1	Scaffold	148.8 Mb	1.3 Mb	93	1403 (92.4%)	8 (0.5%)	59 (3,9%)	49 (3,2%)
	Volvox reticuliferus	GCA_019650255.1	Contig	134 Mb	1.9 Mb	98	1457 (96.0%)	24 (1.6%)	11 (0,7%)	26 (1,7%)
	Volvox carteri f. nagariensis	GCA_000143455.1	Scaffold	137.7 Mb	1.5 Mb	98	1477 (97.2%)	9 (0.6%)	17 (1,1%)	17 (1,1%)
	Tetradesmus obliquus	GCA_030272155.1	Complete Genome	104.7 Mb	6.1 Mb	92	1391 (91.6%)	17 (1.1%)	36 (2,4%)	74 (4,9%)
	Raphidocelis subcapitata	GCA_003203535.1	Scaffold	51.2 Mb	341.8 kb	89	1299 (85.5%)	58 (3.8%)	67 (4,4%)	96 (6,3%)
	Monoraphidium minutum	GCA_025201885.1	Scaffold	68.2 Mb	259.4 kb	91	1278 (84.2%)	96 (6.3%)	47 (3,1%)	97 (6,4%)
Stramenopiles	Aureococcus anophagefferens	GCA_000186865.1	Scaffold	56.7 Mb	1.4 Mb	95	92 (92.0%)	3 (3.0%)	2 (2%)	3 (3%)
	Nannochloropsis salina	GCA_004565275.1	Scaffold	23.1 Mb	24.4 kb	92	73 (73.0%)	0.0%	9 (9%)	18 (18%)
	Nannochloropsis gaditana	GCA_000240725.1	Scaffold	34 Mb	37.6 kb	82	67 (67.0%)	15 (15%)	5 (5%)	13 (13%)
	Fragilaria crotonensis	GCA_022925895.1	Contig	61.8 Mb	89.1 kb	81	71 (71.0%)	10 (10.0%)	0,00%	19 (19%)
	Phaeodactylum tricornutum	GCA_000150955.2	Chromosome	27.5 Mb	945 kb	99	92 (92.0%)	7 (7.0%)	1 (1%)	0,00%
	Seminavis robusta	GCA_903772945.1	Scaffold	125.6 Mb	50.7 kb	99	92 (92.0%)	7 (7.0%)	0,00%	1 (1%)
	Pseudo-nitzschia multistriata	GCA_900660405.1	Scaffold	56.8 Mb	141.2 kb	99	98 (98.0%)	1 (1%)	1 (1%)	0,00%
	Nitzschia inconspicua	GCA_019154785.2	Contig	99.7 Mb	3.6 Mb	100	6 (6.0%)	94 (94.0%)	0,00%	0,00%
	Chaetoceros tenuissimus	GCA_021927905.1	Scaffold	41 Mb	2.4 Mb	96	92 (92.0%)	4 (4.0%)	0,00%	4 (4%)
	Thalassiosira oceanica	GCA_000296195.2	Contig	92 Mb	3.6 kb	93	91 (91.0%)	2 (2.0%)	5 (5%)	2 (2%)
	Skeletonema marinoi	GCA_030871285.1	Contig	54.6 Mb	1.2 Mb	93	88 (88.0%)	5 (5.0%)	3 (3%)	4 (4%)
	Thalassiosira pseudonana	GCA_000149405.2	Chromosome	32.4 Mb	2 Mb	99	98 (98.0%)	1 (1.0%)	1 (1%)	0,00%

●Complete≥95%; ●90%≤Complete<95%; ●85%≤Complete<90% ;●80≤Complete<85%

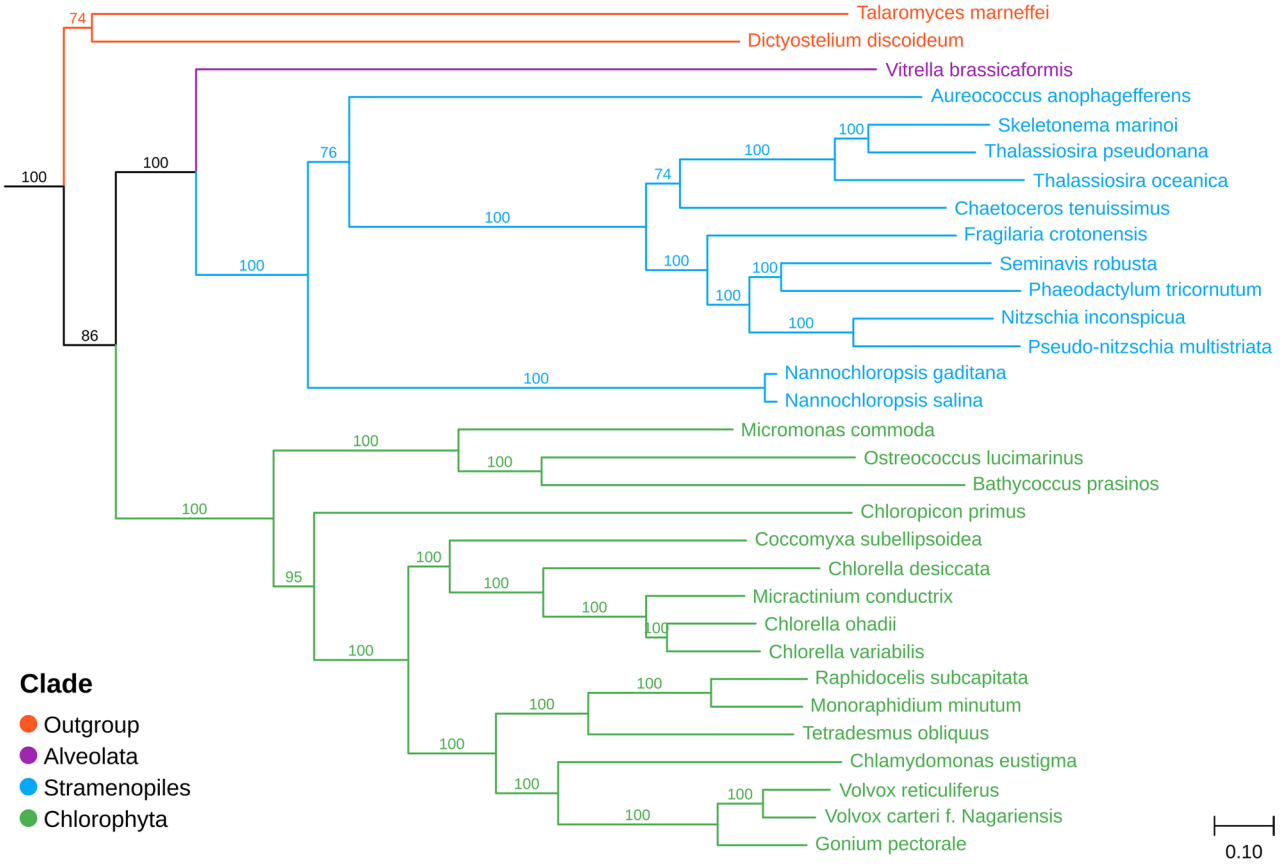

Fig. 4 A phylogenomic tree was constructed to depict the evolutionary relationships among the analyzed species. This tree was based on single-copy orthologous genes and rooted using Talaromyces marneffei, and Dictyostelium discoideum AX4 as outgroups. The root tree for species was based on single-copy orthologs, generated with Orthofinder and IQTree. Bootstrap values are also reported

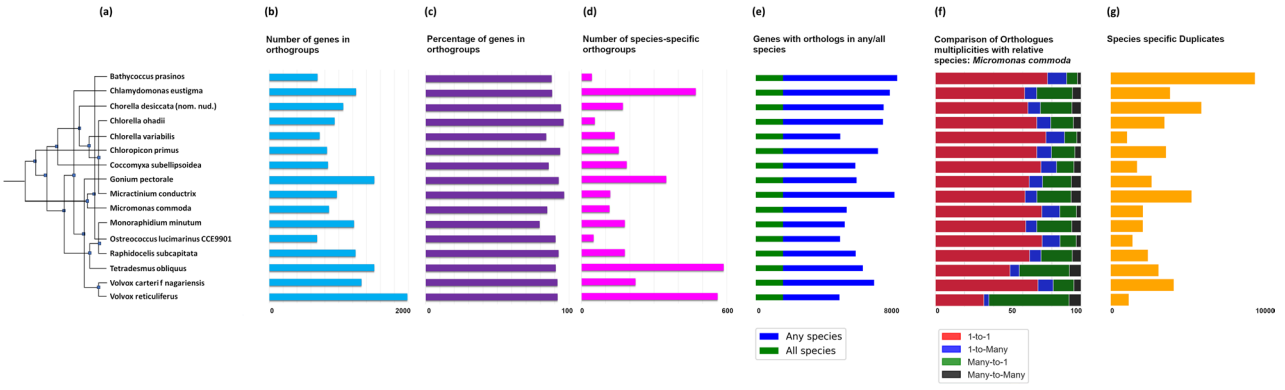

Fig. 5 Orthogroup Analysis in Chlorophyta. **a** Phylogenetic tree based on single-copy orthologs; **b** Number of genes in orthogroups per species; **c** Percentage of genes assigned to orthogroups; **d** Species-specific orthogroups; **e** Genes with orthologs across species; **f** Comparison of orthologous multiplicities relative to *Micractinium commoda* (selected for its complete genome assembly and > 95% BUSCO completeness); **g** Species-specific duplicates

Genomic expansions and contractions

In our comparative analysis, we examined the profiles of orthogroup expansions and contractions with the aim of identifying the molecular mechanisms that contribute to the optimization of CO_2 fixation. The expansion data in species belonging to the CHLO group (Fig. 7, table in Datafile 8) reveal an increase in the copy numbers of numerous key genes for photosynthetic function. In our CAFÉ analysis, we observed expansions of gene families encoding proteins involved in light-harvesting, including isoforms of chlorophyll a/b-binding proteins, early light-induced proteins, and key components of photosystems I and II such as PsaK, PsaL, and PsbR. These findings are consistent with previous studies reporting their

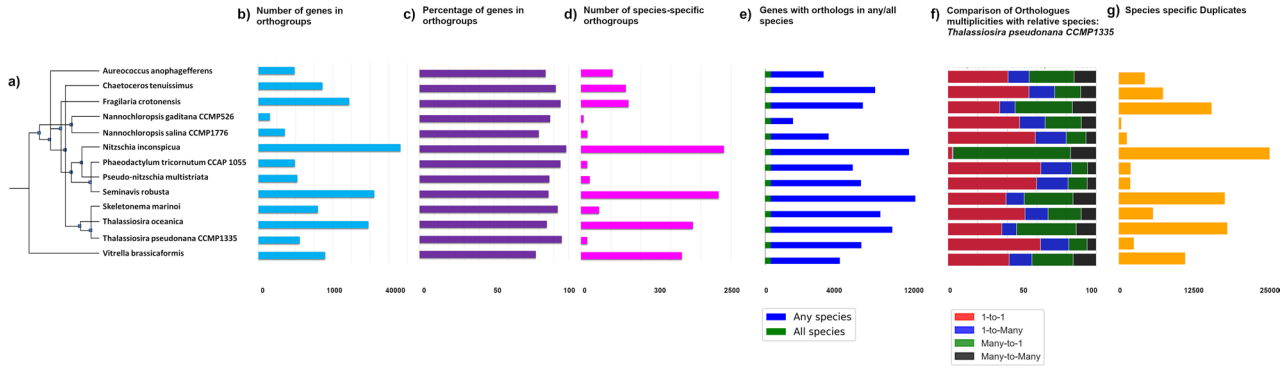

Fig. 6 Orthogroup Analysis in Stramenopiles and Alveolata. **a** Phylogenetic tree based on single-copy orthologs; **b** Number of genes in orthogroups per species; **c** Percentage of genes assigned to orthogroups; **d** Species-specific orthogroups; **e** Orthogroups shared across species; **f** Comparison of orthologous multiplicities relative to Thalassiosira pseudonana (selected for its chromosome-level assembly and 99% BUSCO completeness); **g** Species-specific duplication

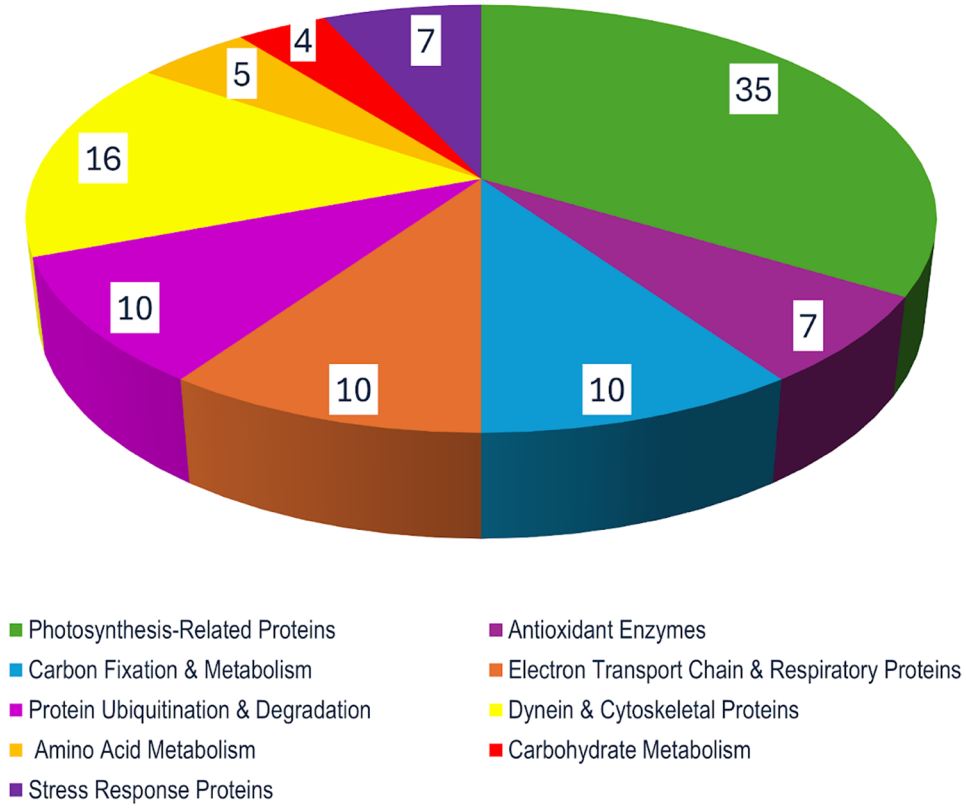

Fig. 7 Functional distribution of expanded orthogroups identified in Chlorophyta. The categories include proteins involved in photosynthesis, metabolism, stress response, protein degradation, transport, and other cellular functions. The graphical representation highlights the relative prevalence of each category

functional roles in enhancing photosynthetic performance [30, 31]. The expansion of antioxidant systems—such as gene families encoding glutathione peroxidase, manganese superoxide dismutase, and superoxide dismutase—along with increased representation of post-translational regulators (e.g., E3 ubiquitin ligases), may represent molecular strategies to mitigate oxidative stress associated with high photosynthetic rates, thereby supporting sustained CO_2 fixation efficiency under light-intensive conditions [32, 33]. A notable example is the dramatic expansion of a sulfur stress regulatory gene in Micractinium conductrix (expansion count of 1070 compared to an average of 35), which may reflect an enhanced capacity to cope with oxidative or nutrient-related stress. Given the known interactions between sulfur metabolism and photosynthetic performance, this expansion might

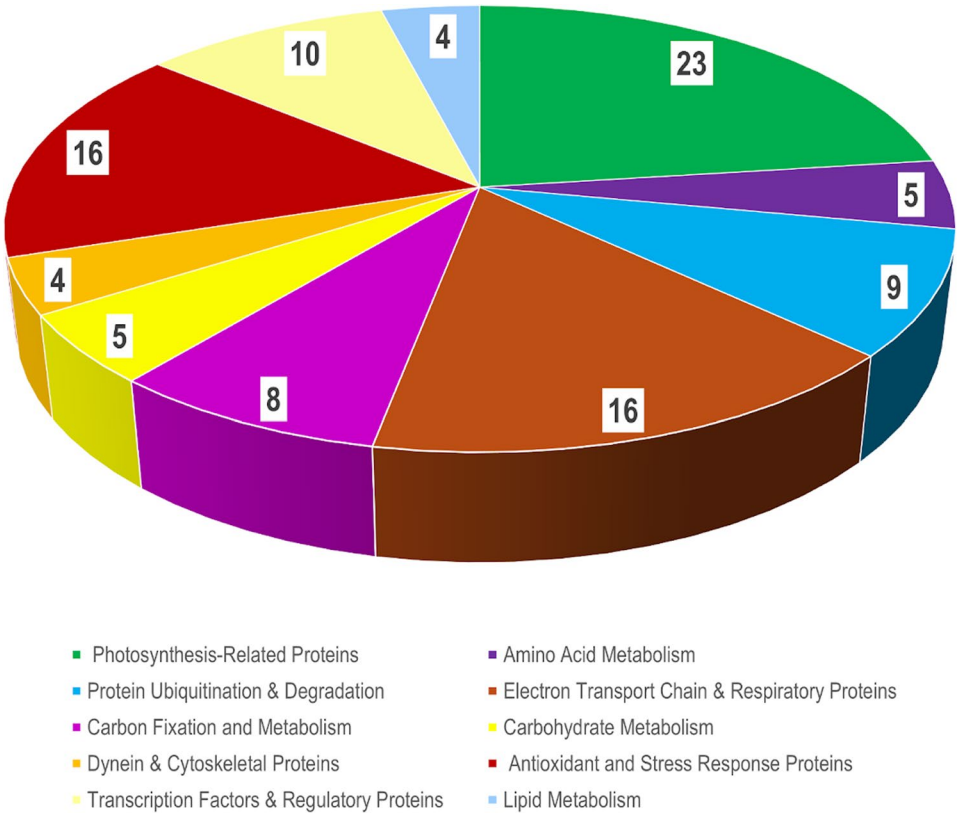

Fig. 8 Functional distribution of expanded orthogroups identified in Stramenophiles and Alveolata. The categories include proteins involved in photosynthesis, metabolism, stress response, protein degradation, transport, and other cellular functions. The graphical representation highlights the relative prevalence of each category

contribute indirectly to CO_2 assimilation efficiency under fluctuating environmental conditions [34].

In contrast, the data for species in STR-A group (Fig. 8, table in Datafile 8) show a marked expansion of orthogroups involved in processes such as light-harvesting—particularly, the fucoxanthin chlorophyll a/c binding proteins-electron transport (with numerous members of the cytochrome P450 family and components of the cytochrome b–c_1 complex), and stress response, as evidenced by the increased representation of various transcription factors and additional E3 ubiquitin-protein ligases. These results indicate an evolutionary strategy geared towards directly enhancing the mechanisms that support photosynthesis and cellular resilience [35, 36].

The details of the proteins associated with each identified category of species belonging to CHLO and to STR-A groups are reported in Datafile 8.

The integrated analysis of expansion and contraction data suggests that the optimization of CO_2 assimilation does not rely solely on an increase in gene copy numbers, but also on a qualitative refinement of existing components. For instance, among Chlorophyta, some species exhibit clear duplications of key genes, whereas others - such as Tetradesmus obliquus and Chlorella ohadii, species known in the literature for their high CO_2 fixation capacity [37] - do not show evident orthogroup expansions. This observation suggests that these species might optimize CO_2 assimilation through mechanisms other than gene family expansion - such as transcriptional regulation, post-translational control, or functional divergence of existing proteins - highlighting that efficient photosynthesis can arise from different molecular strategies beyond genomic duplication [38].

In summary, our results highlight expansions in gene families related to light capture, carbon fixation, and oxidative stress response in both Chlorophyta and Stramenopiles. While these may contribute to photosynthetic efficiency, further studies—including comparisons with non-microalgal taxa—are needed to determine whether these expansions represent convergent adaptations for enhanced CO_2 assimilation. These strategies, integrated with a potential genomic "refinement" via targeted contractions, delineate a complex evolutionary interplay that enables high photosynthetic efficiency. The data presented here provide a solid foundation for future functional and comparative studies aimed at further elucidating the mechanisms underlying CO_2 assimilation optimization in microalgae.

Materials and methods

This study focused on the comparative genomics of microalgal species with annotated reference genomes, available from the NCBI Genome archive (https://www.ncbi.nlm.nih.gov/home/genomes).

Data collection and curation

The term "microalgae" is used descriptively rather than as a monophyletic taxonomic group. It refers to a heterogeneous ensemble of microscopic photosynthetic organisms from diverse evolutionary lineages, including green, brown, and red algae [21, 29].

A stepwise protocol was employed to select the microalgae of interest, i.e. those with a well annotated reference genome. Details about the filtering process are shown in Fig. 1 and described in the Result and Discussion subparagraph titled 'Collection and Selection of Microalgal Genomes for Comparative Analysis'.

All the analyses presented in this article are computationally demanding and were therefore performed on an HPC platform [39–41].

BUSCO analysis for genome completeness

All genome selection, filtering, and annotation processing steps were executed using command-line tools and custom scripts. The high-throughput computational workload required for BUSCO [42] evaluations was conducted on an HPC cluster to ensure efficient processing of large genome datasets. Data curation and file management were performed using standard bioinformatics tools, including Python, Bash scripting, and spreadsheet software for metadata organization.

For many of the selected organisms, multiple genomic accessions were available. To identify the highest-quality accession for our analysis, we applied the Benchmarking Universal Single-Copy Orthologs (BUSCO) method v5.4.4 [25] (final stage of Fig. 1). These evaluations allowed us to compare each assembly in terms of gene completeness, enabling the selection of the assembly with the highest score. We acknowledge, however, that the reliance on BUSCO-based completeness may introduce a bias against species with naturally reduced genomes, potentially underrepresenting organisms with minimal but efficient gene sets for CO_2 assimilation. BUSCO provides a quantitative measure of genome quality and completeness based on evolutionary predictions of gene content derived from databases of nearly universal and highly conserved protein orthologs.

BUSCO Analyses were conducted on genomes and using four ortholog gene databases: Alveolata_odb10, Chlorophyta_odb10, Stramenopiles_odb10, and Euglenozoa_odb10. The BUSCO evaluation process quantified gene completeness for each genome, producing a summary of the completeness scores.

As a result of application of workflow in Fig. 1, we obtained a dataset of 29 species, with their highest-quality genome assembly. These selection of microalgae species in terms of gene completeness and information about the quality of each genome assembly are summarized in Table 1. Accession numbers, assembly levels, genome sizes, and N50 values are also reported.

This curated and high-quality dataset was subsequently used for downstream comparative genomic analyses, ensuring a robust foundation for evolutionary and functional studies of microalgae species with high CO_2 capture efficiency.

Orthogroup identification and phylogenomic tree construction

The downstream analysis proceeded with orthogroup detection and phylogenomic tree construction to investigate the evolutionary relationships and functional conservation of genes among the selected microalgal species. Orthogroups were identified using OrthoFinder v2.5.2 [26], employing the STAG algorithm [43] for species tree inference and STRIDE [44] for rooting. OrthoFinder thus clusters protein sequences into orthogroups via gene tree construction. This analysis provided information on conserved gene families that are essential to understand shared biological pathways and molecular functions among taxa (see result files in Table 2, Datafile 6).

To construct the phylogenetic tree (Fig. 4), we used the proteomes of the 29 examined species. We chose to use proteomes instead of complete genomes because protein sequences directly represent coding regions, providing more reliable information for the identification of orthologs and reducing biases arising from the presence of non-coding regions or possible assembly errors. Following a series of key steps, the 29 microalgal proteomes were retrieved from the NCBI Genome archive, together with those of two other organisms selected as outgroups: Talaromyces marneffei (Ascomycota) and Dictyostelium discoideum AX4 (Amoebozoa). These outgroups were chosen because their phylogenetic distance from microalgae ensures reliable rooting of the tree, thus avoiding potential biases that could arise from the use of too closely related species. In particular, T. marneffei, as an ascomycete fungus, and D. discoideum, belonging to the Amoebozoa, represent external lineages that clearly delineate the study group. Subsequently, orthogroups were identified in all proteomes using OrthoFinder v2.5.5 and orthologous sequences were aligned with MAFFT [45] (included in OrthoFinder package). Individual phylogenetic trees were generated using IQ-TREE v2.1.2 [46] for each one of the 21,880 orthogroups identified by OrthoFinder. The WAG+G4 [47] substitution model (Whelan & Goldman, with a 4-category discrete Gamma model) was selected as the best substitution matrix by

ModelFinder. A global species tree is then obtained from the set of single-copy ortholog gene trees using the STAG and STRIDE procedures implemented in OrthoFinder, with branch lengths estimated by FastTree [48]. Local bootstrap values are approximated using the Shimodaira-Hasegawa test [49] on all splits of the tree, estimating the statistical confidence of tree branches by resampling site-wise likelihoods to determine if the selected topology is significantly better than alternative configurations. This phylogram is based solely on substitution rates of single-copy orthologs.

It is important to note that the overall phylogenetic tree (Fig. 4) was constructed from proteomes representing all the species belonging to three phyla (Chlorophyta, Stramenopiles, and Alveolata).

After constructing the overall phylogenetic tree, we further investigated orthologous relationships among the selected species by performing separate Orthofinder analyses for the CHLO group and the STR-A group. This phyla-oriented approach improves the detection of lineage-specific genomic variations and adaptive strategies, thus providing a more targeted interpretation of functional annotations related to CO_2 capture efficiency of the species of two groups.

For both two groups, we implemented ad hoc Python scripts in order to visualize Orthofinder results as a series of bar graphs (Figs. 5 and 6), illustrating several important aspects which we discussed in Result and Discussion sub-paragraph 'Identification, Annotation, and Phylogenomic Analysis of Orthogroups'.

Gene family expansion and contraction

To investigate the evolutionary dynamics of protein family size variations (expansion or contraction), we used the CAFÉ release 5 v1.1 software [50], which models gene birth and death events along a user-defined phylogenetic tree. CAFÉ employs a stochastic model to estimate whether a protein family has undergone expansion, contraction, or remained unchanged along each branch of the tree, based on the statistical distribution of family sizes (p-value ≤ 0.001).

The software was launched on the the two microalgae results (the two datasets derived from two Orthofinder runs, one for Chlorophyta species - green species in Fig. 4 - and one for Stramenopiles plus Alveolata - light blue and violet species in Fig. 4 respectively) to identify evolving gene families and determine the branches where these changes occurred. The analysis required two input files: (1) a count data file, generated through clustering analysis (e.g., using OrthoFinder) and formatted to include gene family descriptions, unique identifiers, and gene counts across taxa, and (2) a phylogenetic tree file, which had to be a binary, rooted, ultrametric tree in Newick format, typically inferred through molecular dating methods. The two input data files for CAFE are reported in Table 2 and Datafile 7. Results of CAFE execution (Base_change.tab) contain gene families. Positive values were classified as expanded, while the negative ones were considered contracted. This allowed for the identification of gene families relevant to specific biological processes [26].

Ad-hoc Python scripts were used to annotate orthogroups. The information about the annotation contains orthogroup ID, accession number, protein description, and species (Datafile 9). Through an accurate investigation we selected proteins of interest, based on their potential role in CO_2 fixation, carbon metabolism, and related metabolic pathways. Using the same phylogenetic framework, we further examined whether gene families associated with CO_2 capture or metabolism had undergone significant variations among the studied species.

Acknowledgements

The project was implemented under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.4, Call for tender No. 3138 of 16 December 2021, rectified by Decree n.3175 of 18 December 2021 of the Italian Ministry of University and Research funded by the European Union-Next Generation EU. Project code CN_00000033, Concession Decree No.1034 of 17 June 2022 adopted by the Italian Ministry of University and Research, CUPJ83C22000860007, Project title 'National Biodiversity'.

Authors' contributions

TC, DC, AF Conceptualization; JDM, FL, MA AND LA Material sampling; JDM, FL, MA AND LA Methodology; JDM, FL, MA, LA, MC, computational bioinformatics analysis; TC, JDM, PF, CC, LDG, AF and PB writing-original draft preparation; all authors writing-review and editing; all authors supervision. All authors read and approved the final manuscript.

Funding

Not applicable.

Data availability

All files produced in this article are stored on figshare ([https://figshare.com/projects/Comparative_genomics_of_microalgae/206506](https:/figshare.com/projects/Comparative_genomics_of_microalgae/206506)) and details can be found at each row in Table 2.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 12 July 2025 / Accepted: 27 February 2026
Published online: 01 April 2026

References

1. Etheridge DM, et al. Natural and anthropogenic changes in atmospheric CO2 over the last 1000 years from air in Antarctic ice and firn. J Geophys Research: Atmos. 1996;101D2:4115–28.
2. Fellendorf A. Trading the temperature-voluntary carbon offsetting as a climate change mitigation tool for developing countries: lessons from cook-stove projects in Nepal. Wien: Diss; 2018.

3. Lüthi D, et al. High-resolution carbon dioxide concentration records 650,000–800,000 years before present. Nature. 2008;453(7193):379–82.
4. Ritchie H. Pablo Rosado, and Max Roser. Greenhouse gas emissions. Our world data. 2020.
5. Szulejko JE, Kumar P, Deep A, Kim KH. Global warming projections to 2100 using simple CO_2 greenhouse gas modeling and comments on CO_2 climate sensitivity factor. Atmospheric Pollution Res. 2017;8(1):136–40.
6. Nazarenko L, et al. Future Climate Change under RCP Emission Scenarios with GISS ModelE2. J Adv Model Earth Syst. 2015;7(1):244–67.
7. Riahi K, et al. The Shared Socioeconomic Pathways and Their Energy, Land Use, and Greenhouse Gas Emissions Implications: An Overview. Glob Environ Change. 2017;42:153–68.
8. Walsh B, et al. Pathways for balancing CO_2 emissions and sinks. Nat Commun. 2017;8:14856.
9. Ho SH, Chen CY, Chang JS. Effect of light intensity and nitrogen starvation on CO_2 fixation of *Chlorella vulgaris* and the effects of different photobioreactor configurations. Bioresour Technol. 2010;101(13):5510–6.
10. Costa JAV, Colla LM, Duarte FA, Reis MS. *Spirulina platensis* growth in open raceway ponds using carbon dioxide from ethanol production. World J Microbiol Biotechnol. 2000;16:445–8.
11. Lopez P, et al. Enhancement of photosynthetic capacity and biomass yield in *Spirulina platensis* through carbon flux regulation. Biotechnol Adv. 2009;27(4):531–40.
12. Ho S-H, Chen C-Y, Jo-Shu C. Effect of light intensity and nitrogen starvation on CO2 fixation and lipid/carbohydrate production of an indigenous microalga Scenedesmus obliquus CNW-N. Bioresour Technol. 2012;113:244–52.
13. Skjanes K, Rebours C, Lindblad P. Potential for greenhouse gas mitigation through carbon fixation by cyanobacteria and microalgae. AMB Express. 2007;1:20.
14. Sydney, Eduardo Bittencourt, et al. "RETRACTED: Potential carbon dioxide fixation by industrially important microalgae." (2010):5892–6. https://doi.org/10.1016/j.biortech.2010.02.088
15. Paul S, Kumari S, Pandey A, Rana S, Bhatnagar A. Review on the recent structural advances in open and closed systems for carbon capture through algae. Energy Nexus. 2021;4:100032.
16. Kassim MA, Meng TK. Carbon dioxide (CO_2) biofixation by microalgae and its potential for biorefinery and biofuel production. Sci Total Environ. 2017;584–585:1121–9.
17. Kumar K, Dasgupta CN, Nayak B, Lindblad P, Das D. Development of suitable photobioreactors for CO_2 sequestration addressing global warming using green algae and cyanobacteria. Bioresour Technol. 2011;102(8):4945–53.
18. Zhang X, Lei Z, Ge S, Ji B, Zhang B, Editorial. Algae and microalgae-bacteria based technology for sustainable wastewater treatment. Front Microbiol. 2023;14:1263955.
19. Adl SM, et al. The revised classification of eukaryotes. J Eukaryot Microbiol. 2012;59(5):429–93.
20. Burki F, et al. The new tree of eukaryotes. Trends Ecol Evol. 2020;35(1):43–55.
21. Spolaore P, Joannis-Cassan C, Duran E, Isambert A. Commercial applications of microalgae. J Biosci Bioeng. 2006;101(2):87–96.
22. Borowitzka MA. High-value products from microalgae—Their development and commercialization. J Appl Phycol. 2013;25:743–56.
23. Wijffels RH, Barbosa MJ. An outlook on microalgal biofuels. Science. 2010;329(5993):796–9.
24. Grima EM, Belarbi EH, Acién Fernández FG, Robles Medina A, Chisti Y. Recovery of microalgal biomass and metabolites: process options and economics. Biotechnol Adv. 2003;20(7–8):491–515.
25. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31(19):3210–2.
26. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 2019;20:238.
27. Adl SM, Simpson AGB, Lane CE, Lukes J, Bass D, Bowser SS, Brown MW, Burki F, Dunthorn M, Hampl V, et al. The Revised Classification of Eukaryotes. J Eukaryot Microbiol. 2012;59(5):429–93.
28. Burki F, Roger AJ, Brown MW, Simpson AGB. The new tree of eukaryotes. Trends Ecol Evol. 2020;35(1):43–55.
29. Borowitzka MA. High-value products from microalgae—Their development and commercialisation. J Appl Phycol. 2013;25:743–56.
30. García JL, de Vicente M, Galán B. Unconventional high-value products from microalgae: A review. Bioresour Technol. 2021;329:124872.
31. Wells ML, Potin P, Craigie JS, Raven JA, Merchant SS, Helliwell KE, Smith AG, Camire ME, Brawley SH. Algae as nutritional and functional food sources: revisiting our understanding. J Appl Phycol. 2017;29:949–82.
32. Apel K, Hirt H. Reactive oxygen species: metabolism, oxidative stress, and signal transduction. Annu Rev Plant Biol. 2004;55:373–99.
33. Mittler R. Oxidative stress, antioxidants and stress tolerance. Trends Plant Sci. 2002;7(9):405–10.
34. Giordano M, Beardall J, Raven JA. CO_2 concentrating mechanisms in algae: mechanisms, environmental modulation, and evolution. Annu Rev Plant Biol. 2005;56:99–131.
35. Bowler C, Allen AE, Badger JH, Grimwood J, Jabbari K, Kuo A, Maheswari U, Martens C, Maumus F, Otillar RP, et al. The Phaeodactylum genome reveals the evolutionary history of diatom genomes. Nature. 2008;456:239–44.
36. Armbrust EV. The life of diatoms in the world's oceans. Nature. 2004;459:589–95.
37. Costa JAV, Colla LM, Duarte FA, Reis MS. *Spirulina platensis* growth in open raceway ponds using carbon dioxide from ethanol production. World J Microbiol Biotechnol. 2000;16(5):445–8.
38. Ho SH, Chen CY, Chang JS. Effect of light intensity and nitrogen starvation on CO_2 fixation of *Chlorella vulgaris* and the effects of different photobioreactor configurations. Bioresour Technol. 2010;101(13):5510–6. https://doi.org/10.1016/j.biortech.2010.02.004.
39. Castrignanò T, Gioiosa S, Flati T, et al. ELIXIR-IT HPC@CINECA: high performance computing resources for the bioinformatics community. BMC Bioinformatics. 2020;21(Suppl 10):352.
40. Pieroni M, Madeddu F, Di Martino J, Arcieri M, Parisi V, Bottoni P, Castrignanò T. MD-Ligand-Receptor: A High-Performance Computing Tool for Characterizing Ligand-Receptor Binding Interactions in Molecular Dynamics Trajectories. Int J Mol Sci. 2023;24(14):11671.
41. Chiara M, Gioiosa S, Chillemi G, D'Antonio M, Flati T, Picardi E, Zambelli F, Horner DS, Pesole G, Castrignanò T. CoVaCS: a consensus variant calling system. BMC Genomics. 2018;19(1):120. https://doi.org/10.1186/s12864-018-4508-1.
42. Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. BUSCO Update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Mol Biol Evol. 2021;38(10):4647–54.
43. Emms DM, Kelly S. STAG: Species Tree Inference from All Genes, *bioRxiv*. 2018.
44. David M, Emms S, Kelly STRIDE. Species Tree Root Inference from Gene Duplication Events, *Molecular Biology and Evolution*. 2017:34(12);3267–3278.
45. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30(4):772–80.
46. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32(1):268–74.
47. Simon Whelan N, Goldman A, General Empirical. Model of Protein Evolution Derived from Multiple Protein Families Using a Maximum-Likelihood Approach, *Molecular Biology and Evolution*. 2001:18(5);691–699.
48. Price MN, Dehal PS, Arkin AP. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol Biol Evol. 2009;26(7):1641–50.
49. Hidetoshi, Shimodaira. An Approximately Unbiased Test of Phylogenetic Tree Selection, *Systematic Biology*. 2002:51(3); 492–508.
50. de Bie T, Cristianini N, Demuth JP, Hahn MW. CAFE: a computational tool for the study of gene family evolution. Bioinformatics. 2006;22(10):1269–71.

Publisher's Note