SOFTWARE

Open Access

# ALGAEFUN with MARACAS, microALGAE FUNctional enrichment tool for MicroAlgae RnA-seq and Chip-seq AnalysiS

Ana B. Romero-Losada[1,2], Christina Arvanitidou[1,2], Pedro de los Reyes[1], Mercedes García-González[1] and Francisco J. Romero-Campero[1,2*]

*Correspondence: fran@us.es
[1] Institute for Plant Biochemistry and Photosynthesis, Universidad de Sevilla – Consejo Superior de Investigaciones Científicas, Centro de Investigaciones Científicas Isla de La Cartuja, Avenida Américo Vespucio 49, 41092 Seville, Spain
Full list of author information is available at the end of the article

## Abstract

**Background:** Microalgae are emerging as promising sustainable sources for biofuels, biostimulants in agriculture, soil bioremediation, feed and human nutrients. Nonetheless, the molecular mechanisms underpinning microalgae physiology and the biosynthesis of compounds of biotechnological interest are largely uncharacterized. This hinders the development of microalgae full potential as cell-factories. The recent application of omics technologies into microalgae research aims at unraveling these systems. Nevertheless, the lack of specific tools for analysing omics raw data generated from microalgae to provide biological meaningful information are hampering the impact of these technologies. The purpose of ALGAEFUN with MARACAS consists in providing researchers in microalgae with an enabling tool that will allow them to exploit transcriptomic and cistromic high-throughput sequencing data.

**Results:** ALGAEFUN with MARACAS consists of two different tools. First, MARACAS (MicroAlgae RnA-seq and Chip-seq AnalysiS) implements a fully automatic computational pipeline receiving as input RNA-seq (RNA sequencing) or ChIP-seq (chromatin immunoprecipitation sequencing) raw data from microalgae studies. MARACAS generates sets of differentially expressed genes or lists of genomic loci for RNA-seq and ChIP-seq analysis respectively. Second, ALGAEFUN (microALGAE FUNctional enrichment tool) is a web-based application where gene sets generated from RNA-seq analysis as well as lists of genomic loci from ChIP-seq analysis can be used as input. On the one hand, it can be used to perform Gene Ontology and biological pathways enrichment analysis over gene sets. On the other hand, using the results of ChIP-seq data analysis, it identifies a set of potential target genes and analyses the distribution of the loci over gene features. Graphical representation of the results as well as tables with gene annotations are generated and can be downloaded for further analysis.

**Conclusions:** ALGAEFUN with MARACAS provides an integrated environment for the microalgae research community that facilitates the process of obtaining relevant biological information from raw RNA-seq and ChIP-seq data. These applications are designed to assist researchers in the interpretation of gene lists and genomic loci based on functional enrichment analysis. ALGAEFUN with MARACAS is publicly available on https://greennetwork.us.es/AlgaeFUN/.

**Keywords:** Microalgae, RNA-seq, ChIP-seq, Functional enrichment, GO enrichment, KEGG pathway enrichment

## Background

Microalgae are a very diverse non-monophyletic group of photosynthetic microorganisms of special interest due to their physiological plasticity and wide range of biotechnological applications. Microalgae can be found in a wide variety of different habitats, from freshwater to oceans growing under a broad range of temperature, salinity, pH and light intensity values. More than 5000 species have been identified in the oceans accounting for the production of 50% of the oxygen necessary to sustain life on Earth. Microalgae also play a central ecological role as primary producers of biomass establishing the base of aquatic food chains [1]. In the last decades, microalgae have also been of great interest for the scientific community due to the large and yet increasing number of biotechnological applications they present. Specifically, they have been described as a high yield source of carbon compounds and good candidates for the mitigation of $CO_2$ emissions. In microalgae, fixation of $CO_2$ is coupled with growth and biosynthesis of compounds of biotechnological interests such as polysaccharides, lipids, vitamins and antioxidants. Their industrial cultivation for the production of biofuels as well as bioproducts used as biostimulants in agriculture, health supplements, pharmaceuticals and cosmetics is also thoroughly explored nowadays [2, 3]. Finally, microalgae are successfully applied in wastewater treatment coupled with fixation of atmospheric $CO_2$ [4].

Due to these promising features, physiological characterization of multiple microalgae species under different cultivation regimes have been thoroughly carried out [5, 6]. Nevertheless, the molecular mechanisms underpinning the physiology of microalgae are yet poorly understood. In order to facilitate the progress in the characterization of the molecular systems regulating microalgae physiology, high throughput sequencing technologies have been recently applied to obtain the genome of a wide range of microalgae [7–19]. This has promoted the use of different omics, particularly transcriptomics based on RNA-seq data [20–22] and cistromics based on ChIP-seq data [23, 24], to initiate molecular systems biology studies in microalgae. Nonetheless, the impact of this type of studies on microalgae are hampered by the lack of freely available and easy to use online tools to analyze, extract relevant information and integrate omics data. Typically, RNA-seq and ChIP-seq analysis received as input high-throughput sequencing raw data and produce as output, respectively, sets of differentially expressed genes and genomic loci significantly bound by the protein of interest. Processing of the massive amount of high-throughput sequencing data and analysis of the resulting sets of genes and genomic loci obtained from molecular systems biology studies requires computational power, time, effort and expertise that research groups on microalgae may lack. In addition, researchers must explore different data bases separately, which makes the integration of the results and the generation of biological meaningful information more difficult. Therefore, it is imperative the development of frameworks integrating microalgae genome sequences and annotations with tools for high-throughput sequencing data analysis and functional enrichment of gene and genomic loci sets. In order to cover these microalgae research community needs and promote studies in molecular systems biology we have developed the web portal ALGAEFUN with MARACAS using the R package Shiny [25] and other Bioconductor packages. Our web portal consists of two different tools. MARACAS (MicroAlgae RnA-seq and Chip-seq AnalysiS) implements an automatic computational workflow that receives as input RNA-seq or ChIP-seq raw sequencing data from microalgae studies and produces, respectively, sets of differentially expressed genes or a list of genomic loci. These results can be further analyzed using our second tool ALGAEFUN (microAlgae FUNctional enrichment tool). On the one hand, when receiving the results from an RNA-seq analysis, sets of genes are functionally annotated by performing GO (Gene Ontology) [26] and KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways [27] enrichment analysis. On the other hand, when genomic loci from a ChIP-seq analysis are inputted, a set of potential target genes is generated together with the analysis of the distribution of the loci over gene features as well as metagene plots representing the average mapping signal. This set of potential target genes can be further studied using the features for functional enrichment analysis in ALGAEFUN as described above. ALGAEFUN with MARACAS supports a wide range of 14 different microalgae species including *Chlamydomonas reinhardtii*, *Ostreococcus tauri*, *Nannochloropsis gaditana* and *Phaeodactylum tricornutum* among others. Additionally, our tools can be easily extended to include new microalgae as their genome sequence and annotation are made available. The code for ALGAEFUN with MARACAS is publicly available at their respective GitHub repositories from the following links: https://github.com/fran-romero-campero/ALGAEFUN and https://github.com/fran-romero-campero/MARACAS.

## Implementation

ALGAEFUN with MARACAS supports the analysis of the following microalgae covering an ample spectrum of their phylogeny: *Chlamydomonas reinhardtii* [7], *Volvox carteri* [8], *Chromochloris zofingiensis* [9], *Dunaliella salina* [10], *Haematococcus lacustris* [11](Chlorophyceae), *Coccomyxa subellipsoidea* [12] (Trebouxiophyceae), *Ostreococcus tauri* [13], *Bathycoccus prasinos* [14], *Micromonas pusilla CCMP1545* [15] (Mamiellophyceae), *Phaeodactylum tricornutum* [16], *Nannochloropsis gaditana* [17] (Stramenopiles), *Klebsormidium nitens* [18], *Mesotaenium endlicherianum* and *Spirogloea muscicola* [19] (Charophyceae), Fig. 1.

One of the limiting factors for the development of molecular systems biology studies in microalgae is the fragmentation of the available genomic sequences and functional annotations into different databases. To overcome this issue and generate easily accessible resources, genome sequences, functional annotation and genomic feature annotation files for the previous microalgae species were systematically collected from different freely available databases: Ensembl Protists [28], PhycoCosm [29], Phytozome [30], Genomes—NCBI Datasets [31] and a Figshare repository [32], see Table 1. The systematic use of these annotation systems in specific tools for microalgae is expected to promote the improvement of the current sparse annotations in these species. This is one of the goals of our tools, ALGAEFUN with MARACAS.

The MARACAS RNA-seq data analysis pipeline can be executed using either the short read mapper HISAT2 (Hierarchical Indexing for Spliced Alignment of Transcripts 2) [33] or the pseudoalignment method implemented in kallisto [34]. Whereas HISAT2 is an exact method requiring several hours for processing a typical sample, kallisto

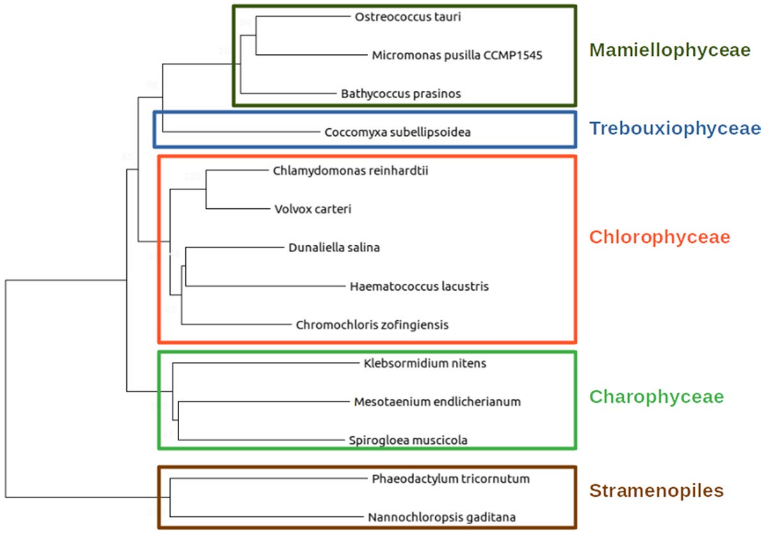


**Fig. 1** Phylogenetic relationship between the different microalgae species supported in ALGAEFUN with MARACAS. The Mamiellophyceae *Ostreococcus tauri*, *Micromonas pusilla CCMP1545* and *Bathycoccus prasinos* (olive green rectangle); the Trebouxiophyceae *Coccomyxa subellipsoidea* (blue rectangle); the Chlorophyceae *Chlamydomonas reinhardtii*, *Volvox carteri*, *Dunaliella salina, Haematococcus lacustris* and *Chromochloris zofingiensis* (red rectangle); the Charophyceae *Klebsormidium nitens*, *Mesotaenium endlicherianum* and *Spirogloea muscicola* (green rectangle); and the Stramenopiles *Phaeodactylum tricornutum*, *Nannochloropsis gaditana* (brown rectangle)

**Table 1** Resources used to collect genome sequences, functional and gene feature annotations for each supported microalga

| Ensembl protists | PhycoCosm | Phytozome | Genomes—NCBI datasets | Figshare associated to publication |
|---|---|---|---|---|
| *Nannochloropsis gaditana*<br>*(GO 41.8%; KO 47.1%)*<br>*Phaeodactylum tricornutum*<br>*(GO 58.8%; KO 40.6%)* | *Ostreococcus tauri*<br>*(GO 52.6%; KO 56.9%)*<br>*Micromonas pusilla CCMP1545*<br>*(GO 40%; KO 26.2%)*<br>*Bathycoccus prasinos*<br>*(GO 46.5%; KO 60.4%)*<br>*Klebsormidium nitens*<br>*(GO 90.3%; KO 47.9%)* | *Chlamydomonas reinhardtii*<br>*(GO 34.8%; KO 23.5%)*<br>*Volvox carteri*<br>*(GO 36.9%; KO 25.7%)*<br>Chromochloris *zofingiensis*<br>*(GO 37.9%; KO 22.9%)*<br>*Dunaliella salina*<br>*(GO 29.1%; KO 20.8%)*<br>*Coccomyxa subellipsoidea*<br>*(GO 45.1%; KO 35.6%)* | *Haematococcus lacustris*<br>*(GO 61.5%; KO 35.1%)* | *Mesotaenium endlicherianum*<br>*(GO 43.7%; KO 36.3%)*<br>*Spirogloea muscicola*<br>*(GO 43.6%; KO 33.5%)* |

The percentage of each microalga genome annotated with GO and KO terms are specified

produces near-optimal gene expression quantification in only a few minutes. The ultrafast and memory-efficient short read mapper bowtie2 [35] is used in the MARACAS ChIP-seq data analysis pipeline. Pre-computed genome indexes for each microalga are included in MARACAS to alleviate the high-cost computational tasks consisting of read mapping to reference genomes. Gene features annotation files in GTF (gene transfer format) for each microalga were also collected. Transcript assembly and estimation of gene expression is performed using stringtie [33]. Identification of differentially expressed genes and graphical representation of the results are carried out using the R

Bioconductor packages Ballgown [33] and LIMMA (Linear Models for Microarray Analysis) [36]. For ChIP-seq data analysis, the peak caller MACS2 (Model-based Analysis of ChIP-seq 2) [37] is used. MARACAS can be executed in sequential mode or in distributed/parallel mode on a computational cluster managed by the job scheduling system SLURM (Simple Linux Utility for Resource Management). In the distributed/parallel mode, synchronization among the jobs processing different samples is implemented using a blackboard system. A common blackboard file is used where parallel jobs can read and write to ensure that all samples have been processed before proceeding to the next steps. This effectively implements synchronization points during the workflow in MARACAS. Graphical representations of the workflows in MARACAS for the analysis of RNA-seq and ChIP-seq data are presented in Figs. 2 and 3 respectively.

For the implementation of ALGAEFUN, functional annotation files were also downloaded for each microalga from the previously mentioned databases, Table 1. Specifically, Gene Ontology (GO) and KEGG (Kyoto Encyclopedia of Genes and Genome) Orthology (KO) terms were collected. For microalgae species lacking these annotation systems HMMER (biological sequence analysis using profile hidden Markov models) [38] was used to identify protein domains according to the PFAM (Protein Family) nomenclature. PFAM terms were subsequently converted into GO terms using pfam2go. KO terms were associated to genes applying KAAS (KEGG Automatic Annotation Server) [39]. Additionally, whenever available other systematic functional annotation

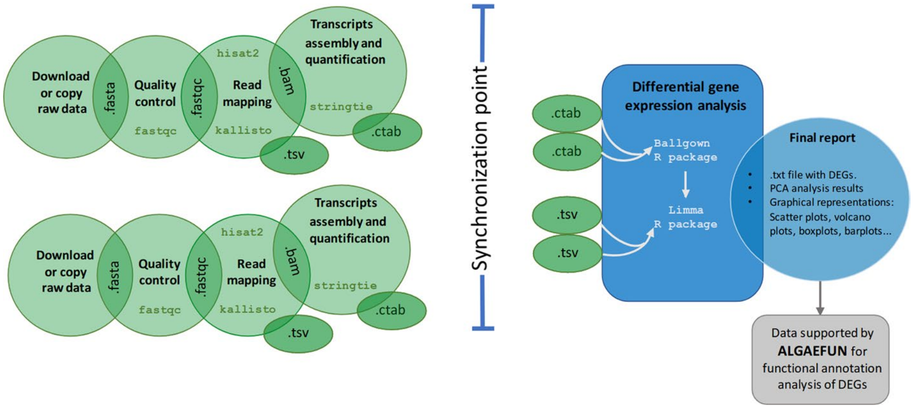


**Fig. 2** Workflow of the automatic pipeline for the analysis of RNA-seq data in MARACAS. The maracas-rna-seq pipeline receives as input a parameter file specifying the microalga of interest, the execution mode (sequential or parallel), the fastq files locations or SRA (sequence read archive) accession numbers, parameters describing the experimental design, fold_change and q_value cutoff thresholds for the identification of DEGs (differentially expressed genes). After data acquisition in fastq format, sequence quality analysis is performed using fastqc. Gene expression quantification is then performed using HISAT2 and stringtie or kallisto depending on the user choice. Read mapping to the reference genome is stored in BAM format. A synchronization point ensuring the completion of all samples processing to quantify gene expression is necessary before identifying DEGs using the R packages ballgown and limma. Gene expression estimates measured as raw count / FPKM (Fragments per Kilobase of exon and Million of mapped reads) / TPM (Transcripts per Million) are stored in tab-delimited files. Reports in html and pdf format are generated with details on sequence quality analysis, mapping process and normalization. Graphics for exploratory analysis such as principal components, boxplots, scatterplots, barplots and volcano plots are also included. These reports include links to download gene expression estimates as well as lists of activated and repressed DEGs. The outputs of this pipeline are compatible with the input formats for AlgaeFUN in order to facilitate further functional enrichment analysis and visualization

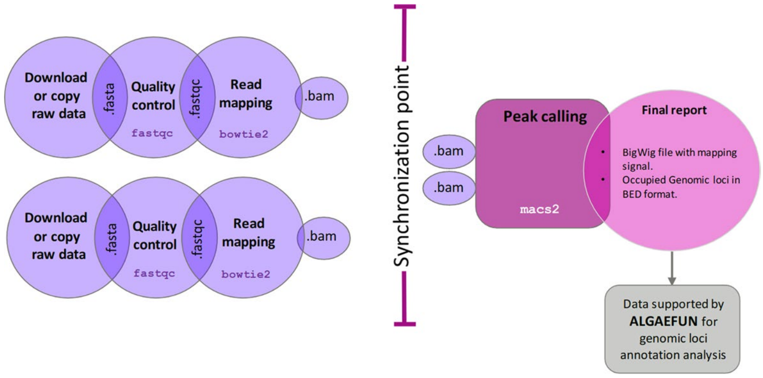

Download or copy raw data
.fasta
Quality control
fastqc
.fastqc
Read mapping
bowtie2
.bam
Download or copy raw data
.fasta
Quality control
fastqc
.fastqc
Read mapping
bowtie2
.bam
Synchronization point
.bam
.bam
Peak calling
macs2
Final report
BigWig file with mapping signal.
Occupied Genomic loci in BED format.
Data supported by ALGAEFUN for genomic loci annotation analysis

**Fig. 3** Workflow of the automatic pipeline for the analysis of ChIP-seq data in MARACAS. The maracas-chip-seq pipeline receives as input a parameter file specifying the microalga of interest, the execution mode (sequential or parallel), the fastq files locations or SRA (sequence read archive) accession numbers and the type of protein-DNA interaction under study either transcription factors or histone modifications. After data acquisition in fastq format, sequence quality analysis is performed using fastqc. Read mapping to the reference genome is performed using bowtie2 and is stored in BAM format. A synchronization point ensuring the completion of all replicates processing is necessary before carrying out peak calling with macs2. Reports in html and pdf format are generated with details on sequence quality analysis and mapping process. These reports include links to download the identified peaks in BED format and the genome wide mapping signal in bigwig format. These outputs are compatible with the input formats for AlgaeFUN in order to facilitate further annotation and visualization of the identified genomic loci significantly bound or occupied by the transcription factor or histone modification of interest

formats were also included such as Protein Analysis Through Evolutionary Relationships (PANTHER) terms [40], Enzyme Commission numbers (EC numbers) and Eukaryotic Orthologous Groups (KOG) terms [41]. In order to use all these functional annotation systems in ALGAEFUN two different types of R annotation packages were developed and were made freely available from our Github repository [42]. On the one hand, using the function makeOrgPackage from the Bioconductor R package AnnotationForge [43] we generated annotation packages for each microalga integrating the systematic sources of functional annotation discussed previously. These packages are instrumental when performing functional enrichment analysis over gene sets obtained, for instance, from a differential expression analysis based on RNA-seq data using MARACAS. On the other hand, applying the function makeTxDbFromGFF from the Bioconductor R package GenomicFeatures [44] we developed a package for each microalga storing gene features annotation available from the previously downloaded and processed GTF files. These packages are central to carry out analysis over genomic loci obtained, for example, using MARACAS and ChIP-seq data to study the genome wide occupation of specific transcription factors or histone modifications. These genomic and functional annotation packages will enable the microalgae research community to perform omics analysis independently from the tools available in ALGAEFUN with MARACAS. The interactive interface of ALGAEFUN was developed using the R package Shiny [25]. In turn, ALGAEFUN functionalities were implemented based on our annotation packages described above. The R Biocondutor packages clusterProfiler [45] and pathview [46] are used for functional enrichment analysis over gene sets. Whereas the annotation and visualization of genomic loci is performed in ALGAEFUN with the R Bioconductor packages ChIPseeker [47] and ChIPpeakAnno [48].

## Results and discussion

Our web-based applications ALGAEFUN (MicroALGAE FUNctional annotation tool) with MARACAS (MicroAlgae RnA-seq and Chip-seq AnalysiS) seek to become enabling tools that would promote molecular systems biology studies in microalgae. A wide range of microalgae species relevant both in basic research and biotechnological applications are supported. Moreover, our tools are easily extendable to include new microalgae species as their genome sequences and functional annotations become available. Analysis combining the different functionalities implemented in our tools will allow researchers to extract relevant information in the form of functional annotation enrichment analysis over gene sets and genomic loci starting from raw high-throughput sequencing data. ALGAEFUN with MARACAS has been applied recently to determine the molecular systems underpinning astaxanthin accumulation in the microalgae of industrial interest *Haematococcus lacustris* [21].

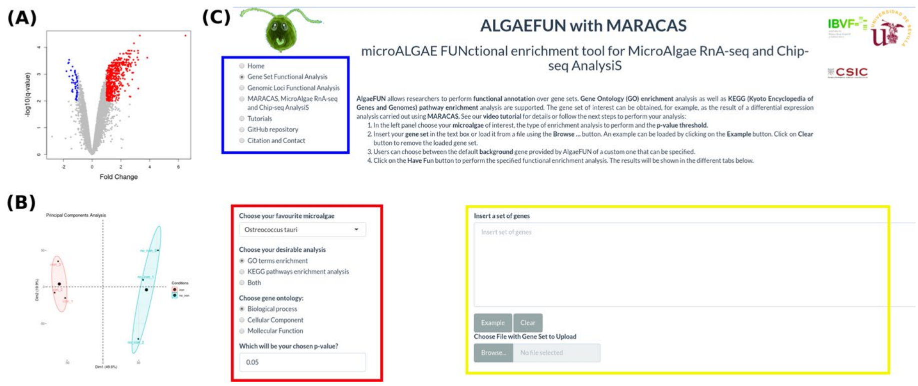
Next, we use two case studies starting from RNA-seq and ChIP-seq raw sequencing data respectively to describe the user interface and discuss the intended uses and benefits of applying ALGAEFUN with MARACAS in microalgae research.

### Case study 1: from RNA-seq raw sequencing data to biological processes and pathways

A detailed description of the steps to install and execute MARACAS is provided at the corresponding Github repository [49] and the video tutorials available on our webpage [50]. Using the automatic pipeline maracas-rna-seq, users can process raw high-throughput sequencing data in fastq format available either locally in their computers or from a data base such as Gene Expression Omnibus [51] or Sequence Read Archive [52]. This pipeline can be executed either in sequential mode or in distributed/parallel mode on a computational cluster managed by the job scheduling system SLURM. The parameter settings need to be provided in a single text file that constitutes the only input to this pipeline. Besides specifying the microalga of interest, the execution mode and fastq files locations or accession numbers, these parameters describe the experimental design, control and experimental samples. Specifically, the parameters fold_change and q_value are used to specify the fold-change and significance level cutoff thresholds for the identification of differentially expressed genes. Reports in html and pdf format are generated containing information regarding sequence quality analysis, mapping process, normalization, principal component analysis and differential gene expression. These reports include links to download gene expression estimates using raw count / FPKM (Fragments per Kilobase of exon and Million of mapped reads) / TPM (Transcripts per Million) that are potentially of interest to users and can be used for downstream analysis using other tools. Genome wide mapping signal for each sample in BigWig format is also outputed by MARACAS. These files can subsequently be loaded into AlgaeFUN for visualization and analysis. The lists of activated and repressed DEGs (Differentially Expressed Genes) can also be downloaded for further analysis using, for example, ALGAEFUN.

In order to illustrate this pipeline, we re-analysed RNA-seq data studying the response of the mamiellophyceae microalgae *Ostreococcus tauri* to iron starvation [20]. The parameter file to reproduce this analysis is provided within the MARACAS distribution bundle. The reports produced by MARACAS described all samples as of high quality and notified no problem during read mapping to the reference genome with mapping rates greater than 94%. Scatter plots comparing gene expression between samples are also produced in the MARACAS report. In this case, high Pearson correlations greater than 98% were identified between replicates of the same condition. Accordingly, the automatically performed Principal Components Analysis in MARACAS identified two clearly separated clusters constituted by the control and iron starvation samples, Fig. 4. A volcano plot is used in the report to represent the 45 repressed genes and 554 activated genes identified with a fold-change threshold of two and a q-value threshold of 0.01, Fig. 4. These lists of genes can be then inputted into ALGAEFUN to determine significantly overrepresented biological processes or pathways affected during iron starvation in *Ostreococcus*.

The graphical interface in ALGAEFUN allows users to explore the different functionalities from a navigation side panel, Fig. 4c. For this case study we selected "Gene Set Functional Analysis". Next, the microalga of interest needs to be chosen from a dropdown menu. Here, we chose *Ostreococcus tauri*. Users need to specify whether GO term and/or KEGG pathway enrichment analysis are to be performed at a selected significance level. The set of genes to analyse can be specified through a text box or by uploading the corresponding file. The use of an appropriate background gene set or gene universe is critical when performing correct functional enrichment analysis. Our tool provides, by default, the specific entire microalga genome as background gene set. These gene sets change massively depending on the microalga under study making imperative the use of their specific genome instead of common gene identifiers that do not take into account the specific genome size of each microalga. For example, the *Ostreococcus* genome codifies for fewer than 8000 genes whereas the *Chlamydomonas* genome encodes for more than 17,000 genes. Additionally, as background gene set, users can also provide a custom



ALGAEFUN with MARACAS

microALGAE FUNctional enrichment tool for MicroAlgae RnA-seq and Chip-seq AnalysiS

○ Home
● Gene Set Functional Analysis
○ Genomic Loci Functional Analysis
○ MARACAS, MicroAlgae RnA-seq and Chip-seq AnalysiS
○ Tutorials
○ GitHub repository
○ Citation and Contact

**AlgaeFUN** allows researchers to perform **functional annotation** over gene sets. **Gene Ontology (GO) enrichment** analysis as well as **KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway enrichment** analysis are supported. The gene set of interest can be obtained, for example, as the result of a differential expression analysis carried out using **MARACAS**. See our **video tutorial** for details or follow the next steps to perform your analysis:

1. In the left panel choose your **microalgae** of interest, the type of enrichment analysis to perform and the **p-value threshold.**
2. Insert your **gene set** in the text box or load it from a file using the **Browse ...** button. An example can be loaded by clicking on the **Example** button. Click on **Clear** button to remove the loaded gene set.
3. Users can choose between the default **background** gene provided by AlgaeFUN of a custom one that can be specified.
4. Click on the **Have Fun** button to perform the specified functional enrichment analysis. The results will be shown in the different tabs below.

Choose your favourite microalgae: Ostreococcus tauri

Choose your desirable analysis
● GO terms enrichment
○ KEGG pathways enrichment analysis
○ Both

Choose gene ontology:
● Biological process
○ Cellular Component
○ Mollecular Function

Which will be your chosen p-value? 0.05

Insert a set of genes: Insert set of genes

Example | Clear

Choose File with Gene Set to Upload: Browse... No file selected

**Fig. 4** RNA-seq MARACAS results and Graphical Interface in AlgaeFUN to perform Gene Set Functional Enrichment Analysis. **a** Volcano plot automatically generated by MARACAS marking activated genes in red and repressed genes in blue. **b** Principal components analysis automatically generated by MARACAS clearly separating samples into two different clusters depending on the analysed conditions. **c** AlgaeFUN Graphical Interface for functional enrichment analysis, navigation side panel for selecting the different functionalities (blue box); dropdown menu for microalgae selection and sidebar panel with parameter settings for GO and/or KEGG pathways enrichment analysis (red box); text box for gene set input and gene background selections (yellow box)

gene set in order to perform a functional enrichment analysis with a specific alternative gene universe to the complete genome of the corresponding microalga. ALGAEFUN provides gene set examples for each microalga so that users can explore our tool and check the required gene id format. These examples can be accessed and inputted in the corresponding text box by clicking on the Example button. These examples were generated during the testing of MARACAS using previously published RNA-seq data sets and have in turn been used in the testing and validation of ALGAEFUN. In this case study we decided to carry out a GO term and KEGG pathway enrichment analysis over the set of 554 activated genes under iron starvation using the default background gene set for *Ostreococcus*.

The outputs of the functional enrichment analysis are presented in the graphical interface in different tabs. Downloadable tables are generated consisting of columns with GO/KEGG term identifiers, human readable descriptions, p-values, q-values, enrichment values and the list of genes associated in the input set with the corresponding GO/KEGG term. Gene names can be clicked to access their annotation from different data bases. Furthermore, ALGAEFUN also generates several graphs that represent the GO/KEGG term enrichment that can be visualized and downloaded from the different tabs such as acyclic graphs, barplots, dotplots, enrichment maps, gene-concept networks and KEGG pathway maps. Our results were in agreement with the published ones [20] identifying ribosome biogenesis and DNA metabolic process as key biological processes overrepresented in the set of activated genes.

### Case study 2: from ChIP-seq raw sequencing data to marked genes

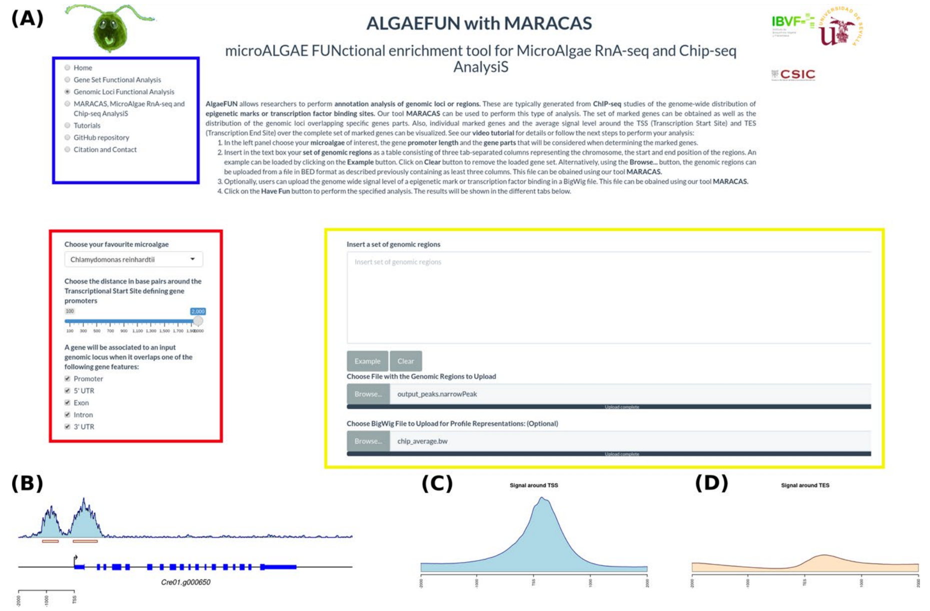
Similar to the previous case study, MARACAS contains an automatic pipeline that can be executed in sequential or parallel mode, maracas-chip-seq. This pipeline implements ChIP-seq raw data analysis providing information on the interaction between proteins such as transcription factors and histone modifications with DNA cis-regulatory elements controlling gene expression. In this respect, the analysis of ChIP-seq data in microalgae would contribute to characterize the molecular mechanisms underpinning gene expression control in microalgae beyond the sole estimation of gene expression generated from RNA-seq data analysis.

This pipeline starts from raw high-throughput sequencing data in fastq format generated in ChIP-seq experiments and produces lists of genomic loci occupied by the histone modification or bound by the transcription factor under study. For this type of analysis, a parameter file needs to be provided specifying the microalga of interest, execution mode, the fastq files locations or accession numbers, whether an input or mock control sample is included and whether the data has been generated for a histone modification or transcription factor. The results of the pipeline include reports in html and pdf format with information regarding sequence quality analysis, mapping process and identification of the genomic loci occupied by the histone modification or bound by the transcription factor under study. The coordinates of these genomic loci are generated in BED (Browser Extensible Data) format and the genome wide mapping signal is produced in a BigWig (Big Wiggle) file. These formats are supported by ALGAEFUN so that these files can be uploaded directly to proceed with the analysis.

To illustrate this pipeline, we re-analysed ChIP-seq data studying the genome wide distribution of the histone modification H3K4me3 in the chlorophyceae microalgae *Chlamydomonas reinhardtii* [24] using the parameter file provided within the MARACAS distribution bundle. The reports produced by MARACAS did not inform of any issue during the processing, leading to the generation of the genomic loci in BED format and the mapping signal in BigWig format.

The graphical interface in ALGAEFUN allows users to input these files to study the genome-wide distribution of specific transcription factors or histone modifications like the H3K4me3 in this case. First, users need to select "Genomic Loci Functional Analysis" from the navigation side panel, Fig. 5a. Next the microalga of interest has to be chosen from a dropdown menu, *Chlamydomonas reinhardtii* in this example. Users must also set the distance around the transcription start site (TSS) defining gene promoters as well as the gene features that a genomic locus needs to overlap to consider the corresponding gene a target; promoter, 5',3'- UTR (5',3'- Untranslated Regions), Exon and Intron, Fig. 5a. The set of genomic loci must be specified by either uploading a BED file or by pasting them in the corresponding text box. Optionally, the genome wide mapping signal can be uploaded in BigWig format, Fig. 5a.

For some microalgae an example consisting of a list of genomic loci can be accessed and inputted in the text box by clicking on the Example button. This would allow users to explore the type of results generated by ALGAEFUN when analysing the outcomes of a



ALGAEFUN with MARACAS

microALGAE FUNctional enrichment tool for MicroAlgae RnA-seq and Chip-seq AnalysiS

**AlgaeFUN** allows researchers to perform **annotation analysis of genomic loci or regions.** These are typically generated from **ChIP-seq** studies of the genome-wide distribution of **epigenetic marks or transcription factor binding sites.** Our tool **MARACAS** can be used to perform this type of analysis. The set of marked genes can be obtained as well as the distribution of the genomic loci overlapping specific genes parts. Also, individual marked genes and the average signal level around the TSS (Transcription Start Site) and TES (Transcription End Site) over the complete set of marked genes can be visualized. See our **video tutorial** for details or follow the next steps to perform your analysis:

1. In the left panel choose your **microalgae** of interest, the gene **promoter length** and the **gene parts** that will be considered when determining the marked genes.
2. Insert in the text box your **set of genomic regions** as a table consisting of three tab-separated columns representing the chromosome, the start and end position of the regions. An example can be loaded by clicking on the **Example** button. Click on **Clear** button to remove the loaded gene set. Alternatively, using the **Browse...** button, the genomic regions can be uploaded from a file in BED format as described previously containing as least three columns. This file can be obained using our tool **MARACAS.**
3. Optionally, users can upload the genome wide signal level of a epigenetic mark or transcription factor binding in a BigWig file. This file can be obained using our tool **MARACAS.**
4. Click on the **Have Fun** button to perform the specified analysis. The results will be shown in the different tabs below.

Choose your favourite microalgae: Chlamydomonas reinhardtii

Choose the distance in base pairs around the Transcriptional Start Site defining gene promoters: 100 – 2,000

A gene will be associated to an input genomic locus when it overlaps one of the following gene features:

- ☑ Promoter
- ☑ 5' UTR
- ☑ Exon
- ☑ Intron
- ☑ 3' UTR

Insert a set of genomic regions

Choose File with the Genomic Regions to Upload: output_peaks.narrowPeak

Choose BigWig File to Upload for Profile Representations: (Optional): chip_average.bw

**Fig. 5** AlgaeFUN Graphical Interface and Results for Genomic Loci Annotation generated from ChIP-seq Analysis. **a** Navigation side panel for selecting the different functionalities (blue box); dropdown menu for microalgae selection and sidebar panel with parameter settings for identifying genes associated with the analysed genomic loci (red box); text box and upload buttons for genomic loci and signal level (yellow box). **b** Visualization of signal distribution over a specific gene and identified peaks in MARACAS. **c** Average signal level around the TSS (Transcription Start Site) for the identified target genes generated as one of the output graphs in AlgaeFUN. **d** Average signal level around the TES (Transcription End Site) for the identified target genes generated as one of the output graphs in AlgaeFUN

ChIP-seq experiment. These examples were generated during the testing of MARACAS using previously published ChIP-seq data sets and have in turn been used in the testing and validation of ALGAEFUN.

In this case study we uploaded to ALGAEFUN the 12,814 genomic loci identified by MARACAS as significantly occupied by H3K4me3 in the *Chlamydomonas* genome under standard growth conditions and the corresponding genome wide mapping signal file in BigWig format. We considered as gene promoter the region two kilobases around the TSS and selected all the gene features to determine the H3K4me3 marked genes. The outputs are presented in the graphical interface in different tabs. A downloadable table with the marked genes and their available annotation is generated. This gene list can in turn be analysed by ALGAEFUN to perform a GO term and/or pathways enrichment analysis. We identified 11,558 H3K4me3 marked genes. Graphs representing the distribution of the genomic loci overlapping different gene features and the distance distribution upstream and downstream from genes TSS are also represented. In agreement with previously published results, we found that 90.75% of the genomic loci occupied by H3K4me3 are located at gene promoters in *Chlamydomonas*. As in this case study, when a BigWig file with the genome wide mapping signal is provided, specific marked genes can be selected to visualize the signal profile over their gene bodies and promoters. A gene example presenting two H3K4me3 peaks on its promoter is depicted to illustrate this functionality in Fig. 5b. Moreover, DNA motifs recognized by specific transcription factors and regulators in photosynthetic organisms can be identified in the promoter of the selected gene. Finally, a visualization of the average level of signal around Transcriptional Start Site (TSS) and Transcriptional End Site (TES) across all marked genes is generated. For the case of H3K4me3 in *Chlamydomonas* we obtained further evidence showing that this epigenetic mark specifically and exclusively locates at the TSS of marked genes and not at the TES, Fig. 5c, d.

As described above, ALGAEFUN with MARACAS constitutes one of the first steps that has been taken for the development of tools that would enable the microalgae research community to exploit high throughput next generation sequencing data by applying systems biology techniques. The first difference between ALGAEFUN with MARACAS with respect to already existing tools consists in the wide range of supported microalgae species, Fig. 1. For the model microalgae *Chlamydomonas reinhardtii*, researchers can find several online tools to functionally annotate set of genes, such as Algal Functional Annotation Tool [53] and ChlamyNET [54]. Only the online tool AgriGO [55] offers the possibility of analysing a restrictive number of different microalgae species beyond *Chlamydomonas*. The second biggest difference between ALGAEFUN and other tools is the annotation systems they use. Most available functional enrichment tools can only perform functional annotation of gene sets based exclusively on Gene Ontology (GO) enrichment analysis. The identification of significantly enriched Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways in the inputted sets of genes is only supported by ALGAEFUN and Algal Functional Annotation Tool. A fundamental difference between ALGAEFUN and other tools consists of the statistical tests. Whereas AgriGO and ChlamyNET are based on Fisher's exact test, ALGAEFUN and Algal Functional Annotation Tool compute statistical significance according to Hypergeometric tests. It has been shown that, in general, the hypergeometric test has more statistical power than Fisher's exact and $\chi^2$ [56]. Moreover, none of these tools can be used as a complete and integrated tool to process high-throughput sequencing raw data from RNA-seq or ChIP-seq experiments, or functionally annotate genomic loci obtained from a ChIP-seq analysis. In this respect, ALGAEFUN with MARACAS improves and implements several novel functionalities of similar already existing software tools, Table 2.

As future work and perspective, ALGAEFUN with MARACAS will be further developed and extended with new tools implementing automatic pipelines to analyse, from raw data to functional analysis, omics technologies recently being applied in microalgae research such as proteomics and metabolomics.

## Conclusion

The main contributions of the tool introduced herein, ALGAEFUN with MARACAS, are enumerated below:

1. ALGAEFUN with MARACAS provides a platform specifically designed for microalgae to analyse raw high-throughput sequencing data from RNA-seq and ChIP-seq studies and functionally annotate the resulting genes sets and/or genomic loci. Our goal consists of developing freely available and easy to use tools that would enable the microalgae research community to perform molecular systems biology analysis.
2. In order to interpret the biological relevance of the gene sets and genomic loci obtained from RNA-seq and ChIP-seq data analysis ALGAEFUN with MARACAS provides simple and informative graphical representations of the GO functional and KEGG pathways enrichment results.
3. Most similar annotation tools only support functional enrichment analysis for gene sets and are restricted to the model microalgae *Chlamydomonas*. In contrast, our annotation tool ALGAEFUN also supports results obtained from ChIP-seq analysis and a wide range of different microalgae species.

**Table 2** Comparison between ALGAEFUN with MARACAS and other functional enrichment analysis tools

| | **Algal functional annotation tool** | **AgriGO** | **ChlamyNET** | **ALGAEFUN with MARACAS** |
|---|---|---|---|---|
| Gene sets as input | *YES* | *YES* | *YES* | *YES* |
| Genomic loci as input | *NO* | *NO* | *NO* | *YES* |
| GO enrichment | *YES* | *YES* | *YES* | *YES* |
| KEGG pathways enrichment | *YES* | *NO* | *NO* | *YES* |
| Several microalgae | *NO* | *YES* | *NO* | *YES* |
| Statistical test | Hypergeometric tests | Fisher's exact test | Fisher's exact test | Hypergeometric tests |

### Availability and requirements

Project name: AlgaeFUN with MARACAS

Project home page: https://greennetwork.us.es/AlgaeFUN/

Operating system(s): Platform independent

Programming language: Bash, R, shiny

Other requirements: none

License: GPL-3.0 License

Any restrictions to use by non-academics: GPL-3.0 License

### Abbreviations

AlgaeFUN: MicroAlgae functional annotation tool; BED: Browser extensible data; ChIP-seq: Chromatin immunoprecipitation sequencing; DEGs: Differentially expressed genes; EC: Enzyme Commission; FDR: False discovery rate; FPKM: Fragments per kilobase of exon and million of mapped reads; GO: Gene Ontology; HISAT2: Hierarchical indexing for spliced alignment of transcripts 2; HMMER: Biological sequence analysis using profile Hidden Markov Models; KAAS: KEGG automatic annotation server; KEGG: Kyoto Encyclopedia of Genes and Genomes; KOG: Eukaryotic orthologous groups; KO: KEGG orthology; LIMMA: Linear models for microarray analysis; MACS2: Model-based analysis of ChIP-seq 2; MARACAS: MicroAlgae RnA-seq and Chip-seq AnalysiS; RNA-seq: RNA sequencing; PANTHER: Protein analysis through evolutionary relationships; PFAM: Protein family; SLURM: Simple Linux utility for resource management; TES: Transcription end site; TPM: Transcripts per million; TSS: Transcription start site; UTR: Untranslated region.

### Acknowledgements

This work was supported by the research project MINOTAUR (BIO2017-84066-R) from the Spanish Ministry of Science and Innovation and the project 3539/0145 funded by AlgaEnergy. We would like to acknowledge Holly Kay and Gerben van Ooijen for assistance with *Ostreococcus tauri* functional annotation.

### Authors' contributions

ABRL, FJRC, CA and PdlR co-developed all the components of MARACAS, ALGAEFUN and the different annotation R packages. FJRC and MGG selected the microalgae species to be included in ALGAEFUN with MARACAS, designed the application interface and wrote the manuscript. All authors read and approved the final manuscript.

### Funding

This work was supported by the research project MINOTAUR (BIO2017-84066-R) from the Spanish Ministry of Science and Innovation and the project 3539/0145 funded by AlgaEnergy. The funding bodies did not play any role in the design of the study, collection, analysis, interpretation of data or in writing the manuscript.

### Availability of data and materials

The code for ALGAEFUN with MARACAS together with the R scripts used in data processing are publicly available at their respective GitHub repositories from the following links: https://github.com/fran-romero-campero/ALGAEFUN and https://github.com/fran-romero-campero/MARACAS. A fully reproducible computational capsule for MARACAS has been deposited at the Code Ocean platform https://doi.org/10.24433/CO.0334617.v2. This capsule is versioned and contains code, data, environment, and the associated results corresponding to a standard RNA-seq and ChIP-seq analysis performed with MARACAS. The capsule is open, exportable, reproducible, and interoperable. The ***Ostreococcus tauri*** RNA-seq data set used in the first case study is freely available on the SRA database identified with accession number SRP066656. The ***Chlamydomonas reinhardtii*** ChIP-seq data set used in the second case study is freely available on the GEO database identified with the accession number GSE59629.

## Declarations

### Ethics approval and consent to participate

Not applicable.

### Consent for publication

Not applicable.

### Competing interests

The authors declare that they have no competing interests.

### Author details

[1]Institute for Plant Biochemistry and Photosynthesis, Universidad de Sevilla – Consejo Superior de Investigaciones Científicas, Centro de Investigaciones Científicas Isla de La Cartuja, Avenida Américo Vespucio 49, 41092 Seville, Spain. [2]Department of Computer Science and Artificial Intelligence, University of Sevilla, Escuela Técnica Superior en Ingeniería Informática, Avenida Reina Mercedes s/n, 41012 Seville, Spain.

Received: 24 August 2021 Accepted: 17 March 2022
Published online: 31 March 2022

## References

1. Chapman RL. Algae: the world's most important "plants"—an introduction. Mitig Adapt Strateg Glob Change. 2013;18:5–12. https://doi.org/10.1007/s11027-010-9255-9.
2. Chen H, Li T, Wang Q. Ten years of algal biofuel and bioproducts: gains and pains. Planta. 2019;249:195–219. https://doi.org/10.1007/s00425-018-3066-8.
3. Lee SM, Ryu CM. Algae as new kids in the beneficial plant microbiome. Front Plant Sci. 2021;12:91. https://doi.org/10.3389/fpls.2021.599742.
4. Shahid A, Malik S, Zhu H, Xu J, Nawaz MZ, Nawaz S, Alam MA, Mehmood MA. Cultivating microalgae in wastewater for biomass production, pollutant removal, and atmospheric carbon mitigation; a review. Sci Total Environ. 2020;704:135303. https://doi.org/10.1016/j.scitotenv.2019.135303.
5. Al Jabri H, Taleb A, Touchard R, Saadaoui I, Goetz V, Pruvost J. Cultivating microalgae in desert conditions: evaluation of the effect of light-temperature summer conditions on the growth and metabolism of nannochloropsis QU130. Appl Sci. 2021;11:3799. https://doi.org/10.3390/app11093799.
6. Patil PP, Vass I, Kodru S, Szabó M. A multi-parametric screening platform for photosynthetic trait characterization of microalgae and cyanobacteria under inorganic carbon limitation. PLoS ONE. 2020;15:e0236188. https://doi.org/10.1371/journal.pone.0236188.
7. Merchant SS, Prochnik SE, Vallon O, Harris EH, Karpowicz SJ, Witman GB, et al. The Chlamydomonas genome reveals the evolution of key animal and plant functions. Science. 2007;318:245–50. https://doi.org/10.1126/science.1143609.
8. Prochnik SE, Umen J, Nedelcu AM, Hallmann A, Miller SM, Nishii I, et al. Genomic analysis of organismal complexity in the multicellular green alga Volvox carteri. Science. 2010;329:223–6. https://doi.org/10.1126/science.1188800.
9. Roth MS, Cokus SJ, Gallaher SD, Walter A, Lopez D, Erickson E, et al. Chromosome-level genome assembly and transcriptome of the green alga *Chromochloris zofingiensis* illuminates astaxanthin production. Proc Natl Acad Sci USA. 2017;114:E4296–305. https://doi.org/10.1073/pnas.1619928114.
10. Polle JEW, Barry K, Cushman J, Schmutz J, Tran D, Hathwaik LT, et al. Draft nuclear genome sequence of the halophilic and beta-carotene-accumulating green alga *Dunaliella salina* strain CCAP19/18. Genome Announc. 2017;5:e01105-e1117. https://doi.org/10.1128/genomeA.01105-17.
11. Morimoto D, Yoshida T, Sawayama S. Draft genome sequence of the astaxanthin-producing microalga *Haematococcus lacustris* strain NIES-144. Microbiol Resour Announc. 2020;9:e00128-e220. https://doi.org/10.1128/MRA.00128-20.
12. Blanc G, Agarkova I, Grimwood J, Kuo A, Brueggeman A, Dunigan DD, et al. The genome of the polar eukaryotic microalga *Coccomyxa subellipsoidea* reveals traits of cold adaptation. Genome Biol. 2012;13:R39. https://doi.org/10.1186/gb-2012-13-5-r39.
13. Palenik B, Grimwood J, Aerts A, Rouzé P, Salamov A, Putnam N, et al. The tiny eukaryote Ostreococcus provides genomic insights into the paradox of plankton speciation. Proc Natl Acad Sci USA. 2007;104:7705–10. https://doi.org/10.1073/pnas.0611046104.
14. Moreau H, Verhelst B, Couloux A, Derelle E, Rombauts S, Grimsley N, et al. Gene functionalities and genome structure in *Bathycoccus prasinos* reflect cellular specializations at the base of the green lineage. Genome Biol. 2012;13:R74. https://doi.org/10.1186/gb-2012-13-8-r74.
15. Worden AZ, Lee JH, Mock T, Rouzé P, Simmons MP, Aerts AL, et al. Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes Micromonas. Science. 2009;324:268–72. https://doi.org/10.1126/science.1167222.
16. Bowler C, Allen AE, Badger JH, Grimwood J, Jabbari K, Kuo A, et al. The Phaeodactylum genome reveals the evolutionary history of diatom genomes. Nature. 2008;456:239–44. https://doi.org/10.1038/nature07410.
17. Corteggiani Carpinelli E, Telatin A, Vitulo N, Forcato C, D'Angelo M, Schiavon R, et al. Chromosome scale genome assembly and transcriptome profiling of *Nannochloropsis gaditana* in nitrogen depletion. Mol Plant. 2014;7:323–35. https://doi.org/10.1093/mp/sst120.
18. Hori K, Maruyama F, Fujisawa T, Togashi T, Yamamoto N, Seo M, et al. *Klebsormidium flaccidum* genome reveals primary factors for plant terrestrial adaptation. Nat Commun. 2014;5:3978. https://doi.org/10.1038/ncomms4978.
19. Cheng S, Xian W, Fu Y, Marin B, Keller J, Wu T, et al. Genomes of subaerial Zygnematophyceae provide insights into land plant evolution. Cell. 2019;179:1057–67. https://doi.org/10.1016/j.cell.2019.10.019.
20. Lelandais G, Scheiber I, Paz-Yepes J, Lozano JC, Botebol H, Pilatova J, et al. *Ostreococcus tauri* is a new model green alga for studying iron metabolism in eukaryotic phytoplankton. BMC Genomics. 2016;17:319. https://doi.org/10.1186/s12864-016-2666-6.
21. Hoys C, Romero-Losada AB, Del Río E, Guerrero MG, Romero-Campero FJ, García-González M. Unveiling the underlying molecular basis of astaxanthin accumulation in Haematococcus through integrative metabolomic-transcriptomic analysis. Bioresour Technol. 2021;332: 125150. https://doi.org/10.1016/j.biortech.2021.125150.
22. Monte I, Kneeshaw S, Franco-Zorrilla JM, Chini A, Zamarreño AM, García-Mina JM, Solano R. An ancient COI1-independent function for reactive electrophilic oxylipins in thermotolerance. Curr Biol. 2020;30:962–71. https://doi.org/10.1016/j.cub.2020.01.023.
23. Zhao X, Rastogi A, Deton-Cabanillas AF, Mohamed OA, Cantrel C, Lombard B, et al. Genome wide natural variation of H3K27me3 selectively marks genes predicted to be important for cell differentiation in *Phaeodactylum tricornutum*. New Phytol. 2021;229:3208–20. https://doi.org/10.1111/nph.17129.
24. Ngan CY, Wong CH, Choi C, Yoshinaga Y, Louie K, Jia J, et al. Lineage-specific chromatin signatures reveal a regulator of lipid metabolism in microalgae. Nat Plants. 2015;1:15107. https://doi.org/10.1038/nplants.2015.107.

25. Chang W, Cheng J, Allaire JJ, Sievert C, Schloerke B, Xie Y, Allen J, McPherson J, Dipert A, Borges B. shiny: web application framework for R. R package version 1.6.0. 2021 http://shiny.rstudio.com/
26. Carbon S, Douglass E, Dunn N, Good B, Harris NL, Lewis SE, et al. The gene ontology resource: 20 years and still going strong. Nucleic Acids Res. 2019;47:D330–8. https://doi.org/10.1093/nar/gky1055.
27. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 2016;44:D457–62. https://doi.org/10.1093/nar/gkv1070.
28. Ensembl Protists realease 51. EMBL-EBI. 2021. https://protists.ensembl.org Accessed Aug 2021.
29. Grigoriev IV, Hayes RD, Calhoun S, Kamel B, Wang A, Ahrendt S, Dusheyko S, Nikitin R, Mondo SJ, Salamov A, Shabalov I, Kuo A. PhycoCosm, a comparative algal genomics resource. Nucleic Acids Res. 2021;49:D1004–11. https://doi.org/10.1093/nar/gkaa898.
30. Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, Rokhsar DS. Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res. 2012;40:D1178–86. https://doi.org/10.1093/nar/gkr944.
31. Genomes – NCBI Datasets Beta Accession Number GCA_011766145.1. 2021. https://www.ncbi.nlm.nih.gov/datasets/genomes/ Accessed Aug 2021.
32. Figshare repository for subaerial Zygnematophyceae. 2019. https://figshare.com/articles/dataset/Genomes_of_subaerial_Zygnematophyceae_provide_insights_into_land_plant_evolution/9911876/1 Accessed Aug 2021.
33. Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nat Protoc. 2016;11:1650–67. https://doi.org/10.1038/nprot.2016.095.
34. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016;34:525–7. https://doi.org/10.1038/nbt.3519.
35. Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9. https://doi.org/10.1038/nmeth.1923.
36. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43: e47. https://doi.org/10.1093/nar/gkv007.
37. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson D, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9:R137. https://doi.org/10.1186/gb-2008-9-9-r137.
38. Mistry J, Finn RD, Eddy SR, Bateman A, Punta M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. Nucleic Acids Res. 2013;41:e121. https://doi.org/10.1093/nar/gkt263.
39. Moriya Y, Itoh M, Okuda S, Yoshizawa A, Kanehisa M. KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res. 2007;35:W182–5. https://doi.org/10.1093/nar/gkm321.
40. Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, et al. PANTHER: a library of protein families and subfamilies indexed by function. Genome Res. 2003;13:2129–41. https://doi.org/10.1101/gr.772403.
41. Galperin M, Wolf Y, Makarova KS, Vera-Álvarez R, Landsman D, Koonin EV. COG database update: focus on microbial diversity, model organisms, and widespread pathogens. Nucleic Acids Res. 2021;49:D274–81. https://doi.org/10.1093/nar/gkaa1018.
42. ALGAEFUN Github repository 2021. https://github.com/fran-romero-campero/ALGAEFUN Accessed Aug 2021.
43. Carlson M, Pagès H. AnnotationForge: tools for building SQLite-based annotation data packages. R package version 1.34.0. 2021 https://bioconductor.org/packages/AnnotationForge
44. Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ. Software for computing and annotating genomic ranges. PloS Comput Biol. 2013;9:e1003118. https://doi.org/10.1371/journal.pcbi.1003118.
45. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. The Innovation. 2021. https://doi.org/10.1016/j.xinn.2021.100141.
46. Luo W, Brouwer C. Pathview: an R/Bioconductor package for pathway-based data integration and visualization. Bioinformatics. 2013;15:1830–1. https://doi.org/10.1093/bioinformatics/btt285.
47. Yu G, Wang LG, He QY. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. Bioinformatics. 2015;31:2382–3. https://doi.org/10.1093/bioinformatics/btv145.
48. Zhu LJ. Integrative analysis of ChIP-chip and ChIP-seq dataset. Methods Mol Biol. 2013;1067:105–24. https://doi.org/10.1007/978-1-62703-607-8_8.
49. MARACAS Github repository 2021. https://github.com/fran-romero-campero/MARACAS. Accessed Aug 2021.
50. ALGAEFUN with MARACAS webpage 2021. https://greennetwork.us.es/AlgaeFUN/. Accessed Aug 2021.
51. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res. 2013;41:D991–5. https://doi.org/10.1093/nar/gks1193.
52. Leinonen R, Sugarawa H, Shumway M. The sequence read archive. Nucleic Acids Res. 2011;39:D19–21. https://doi.org/10.1093/nar/gkq1019.
53. Lopez D, Casero D, Cokus SJ, Merchant SS, Pellegrini M. Algal functional annotation tool: a web-based analysis suite to functionally interpret large gene lists using integrated annotation and expression data. BMC Bioinform. 2011;12:282. https://doi.org/10.1186/1471-2105-12-282.
54. Romero-Campero FJ, Perez-Hurtado I, Lucas-Reina E, Romero JM, Valverde F. ChlamyNET: A Chlamydomonas gene co-expression network reveals global properties of the transcriptome and the early setup of key co-expression patterns in the green lineage. BMC Genomics. 2016;17:227. https://doi.org/10.1186/s12864-016-2564-y.
55. Tian T, Liu Y, Yan H, You Q, Yi X, Du Z, Xu W, Su Z. AgriGO v2.0: a GO analysis toolkit for the agricultural community, 2017 update. Nucleic Acids Res. 2017;45:W122–9. https://doi.org/10.1093/nar/gkx382.
56. Masseroli M, Martucci D, Pinciroli F. GFINDer: Genome Function INtegrated Discoverer through dynamic annotation, statistical analysis, and mining. Nucleic Acids Res. 2004;32:W293–300.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.