Doctoral Thesis

Multiomics Characterization of the Responses to Diurnal and Seasonal Cycles in the Marine Picoeukaryote *Ostreococcus tauri*

Dissertation presented by Ana Belén Romero Losada to obtain the PhD Degree by Universidad de Sevilla.

Supervisors:

Prof. Francisco José Romero Campero

Prof. Mercedes García González

This work was supported by the research projects MINOTAUR (BIO2017-84066-R) and BLOOM (RTC-2017-6080-5) from the Spanish Ministry of Science and Innovation.

Dedicatoria . . .

word cloud

Contents

Index

Abstract

Abstract text

Abstract text

Introduction

Chronobiology

Simply by looking around, it is easy to identify many cyclic processes. We are so much used to them that we usually don't perceive how they influence our lives and bodies.

There are four environmental cycles that we all know: tides (which repeat every 12 and half hours), lunar cycles (lasting 28.5 days), days (every 24 hours) and years (every 365.25 days).

These cycles are also called circatidal, circalunar, circadian, or circannual (Numata et al., 2015) . These four rhythmic processes that arise from physical forces are quite important because they are extremely predictable. It is possible, for example, to know the exact time of a high tide years in advance. We constantly see how full moons are always followed by new moons, and everybody goes to bed sure that the sun will rise the next morning. The exact date for the next equinox is known and written in all calendars all over the world, so the change of season is never a surprise.

These four cycles affect earth with an overwhelming precision, it would be a foolish idea to think that earth living organisms don't react to such rhythmic changes. That's the main theme of Chronobiology: it is a young science that studies how these rhythmic environmental changes affect organisms(Edmunds, 1983; Kuhlman et al., 2018).

Living beings perceive environmental cyclic changes and they are able to react in advance generating endogenous biological rhythms thanks to an internal machinery that acts as a clock. Chronobiologists have found biological rhythms in a wide range of scales. It goes from a molecular level (transcription, translation, protein degradation, metabolites synthesis, etc), to a cellular and tissue level (cell division, synaptic connections, apoptosis, etc.) or even a complete organism or populations(Edmunds, 1983; Merrow et al., 2005; Sharma et al., 2022).

A surprising characteristic of theses rhythms is that they are self-sustained (Pittendrigh, 1960; Roenneberg & Merrow, 2005). For example, when the response of an organism to rhythmic environmental changes (light/dark cycles, food availability, temperature changes, etc.) is studied, it's common to find several biological functions showing rhythmic behav-

iors as well. However, some of them would will not generate rhythms in constant conditions because the change in activity is a direct acute response to changes in the environment, and therefore if the environment is constant, the activity will be constant. Rhythmicity will be maintained under constant constant environmental inputs only when the studied biological function is regulated by that machinery called clock and thus self-sustained (Fig. 2).

For that reason, every chronobiology experiment is designed as follows (Fig. 2-1): several consecutive days where the organism is exposed to the rhythmic input (called zeitgeber, which is used as synchronizer) are followed by several consecutive days where the organism is exposed to constant conditions (called free-running conditions) (Kuhlman et al., 2018). In both scenarios, data is collected every few hours, minutes or seconds depending on the complexity of the data. For example, in the case of a circadian experiment, as the ones executed in this work, the zeitgeber would be a light-dark cycle and the free-running conditions would be constant light and constant darkness. Under that experimental design, circadian processes can be detected and discerned from light or dark responding processes.

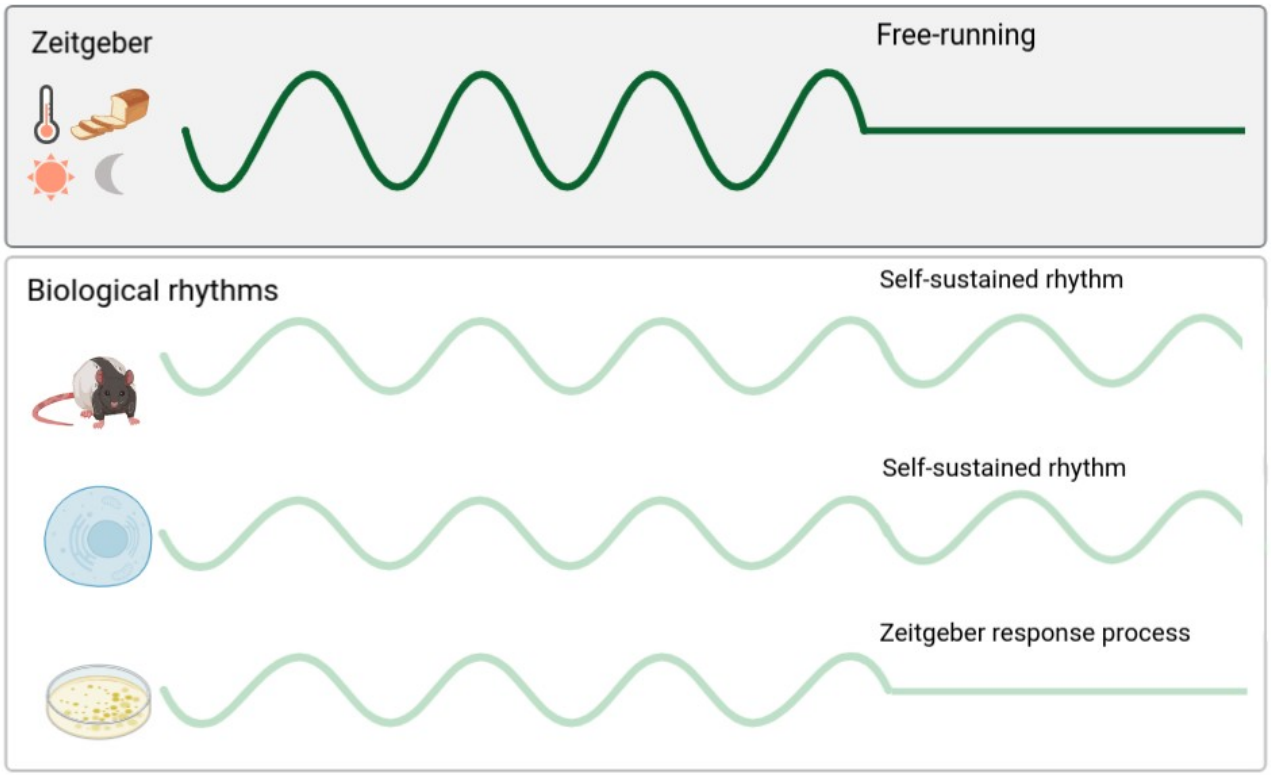

Figure 2: Under a zeitgeber (rhythmic environmental inputs as light/dark, temperature, food availability, etc) two kind of biological processes show rhythmic patterns: the ones that are self-sustained and thus regulated by an endogenous clock; the ones that are only responding to the given zeitgeber. Self-sustained processes can be discerned from zeitgeber responding processes by changing the environmental cyclic condition to a constant one (free-running condition). Under free-running conditions, only the self-sustained processes will maintain their rhythmic profiles.

Following that experimental design, chronobiologists have described self-sustained biological rhythms reacting to the four different environmental rhythms mentioned above. For example, some marine organisms show self-sustained circatidal rhythms when they are kept in laboratory tanks without its zeitgeber, in this case, the tidal changes (Rock et al., 2022). As it was observed in the marine diatom *Hantzschia amphioxys* that descends to the sand at high tides and rises to the surface at low tides (Fauré-Fremiet, 1951). In contrast, the self-sustained circalunar processes are yet more unknown (Andreatta & Tessmar-Raible, 2020). One of the most famous ones is the larvae of the insect called Ant Lion, which build small holes in the sand as traps for insects. Scientist found that the size of the traps changes showing a circalunar profile that is maintained under constant conditions (Youthed & Moran, 1969).

However, the scientific studies about circadian or circannual rhythms are much more numerous, and they have been found in a wide range of organisms (Merrow et al., 2005; Pfeuty et al., 2012; Roenneberg & Merrow, 2005). In animals, circadian rhythms are involved in activity-rest cycles (Roenneberg et al., 2022; Zee & Abbott, 2020), but hundreds of other parameters from ethology to gene expression also show circadian profiles. For example, the olfactory discrimination in mice is higher at night time, even under free-running conditions (Granados-Fuentes et al., 2006). Also fungus show circadian rhythmic phenomena, for example, *Neurospora crassa* generates asexual spores every 24 h even under constant darkness (Correa & Bell-Pedersen, 2002). Circadian rhythmic profiles in plants are found, for example, in leaf movement, growth rate, stomatal opening, as well as the expression of a wide range of genes (Merrow et al., 2005). Also all the organisms of the green lineage react in many different ways to circadian cycles (Noordally & Millar, 2015), one of the most famous ones is the 24 h-cyclic movement in the water column adjusted to their metabolism requirements (Lebert et al., 1999).

Day length or photoperiod, is a crucial signal for the circannual timing system. Seasonal changes of photoperiod have been strongly connected to reproduction. In fact, the entire animal reproductive systems are related to seasons, from gene expression profiles to anatomical structures. Hamsters kept in short days conditions have 10-fold smaller testes than the ones kept in long days conditions (Klante & Steinlechner, 1994; Nishiwaki-Ohkawa & Yoshimura, 2016). But also in plants, flowering and seed production is photoperiod-regulated (Brandoli et al., 2020; Serrano-Bueno et al., 2017).

Since photosynthetic organisms depend on light to ensure its success, they are highly synchronized with cyclic environmental changes involving light such as circadian rhythms and photoperiods. Nevertheless, the chronobiology of algae is yet barely studied compared with other organisms (Fig. 3) despite representing one of the largest polyphyletic groups in the eukaryotic domain. The already investigated genetic and molecular techniques used to identify clock components in other taxa have not been widely applied to algae yet (Noordally & Millar, 2015). This work aims to contribute to the chronobiology community by describing the circadian and photoperiodic changes in the microalgae *Ostreococcus tauri*.

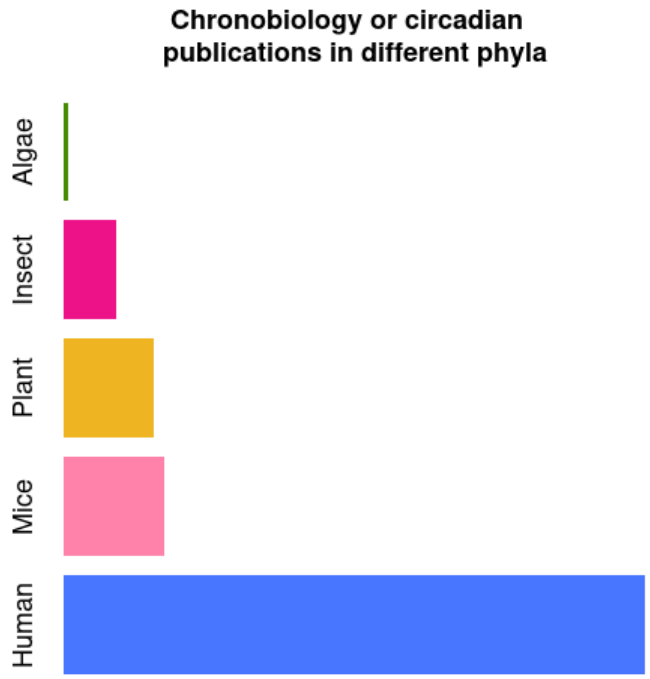

Figure 3: Number of publications found in PubMed using "chronobiology" or "circadian" keywords in their abstracts in July of 2022. For the algae group (using the generic term "algae" and "Ostreococcus tauri", "Chlamydomonas reinhardtii" and "polyedra" as main model organisms of this group), only 493 publications were found. For the insect group (including the generic term "insect" and "Drosophila melanogaster" as main model organism), 5717 publications were found. For plants (including the generic term "plant" and "Arabidopsis thaliana" as main model organism), 9734 publications were found. For mice and human, 10771 and 62305 publications were found, respectively.

Circadian research

Our society is worldwide structured in a 24 h / 7 days system. People from different countries and cultures are experiencing jet-lags, shift work, exposure to artificial light and lack of outdoor activities. Circadian clock research is gaining relevance since this phenomena has a crucial impact on human health, behavior and quality of life (Mermet et al., 2017; Roenneberg et al., 2019, 2022; Roenneberg & Merrow, 2016). Understanding the physiology, genetics and epigenetics (Ripperger & Merrow, 2011) at a laboratory level using different model organisms besides mice (as plants, fungus and microalgae ones) will also help in the comprehension of the circadian clock in humans and its variation among individuals.

The synchrony that exists between the sunrise/twilight and organisms have been so obvious for scientists that the underlying molecular mechanisms remained ignored and unexplored for centuries. The first observation indicating that daily rhythms were programmed took place in the 18th century. Jean Jacques d'Ortous De Mairan, a French astronomer, described in less than 350 words (nowadays, less than two *tweets*) how a mimosa plant inside a closet maintained its daily leaf movement (De Mairan, 1729). For some scientists, it was a clear proof that leaf movement was not controlled by light and darkness. Although De Mairan invited botanists to investigate his discovery to confirm that leaf movement was also maintained when temperature changes are avoid (which inside its closet was difficult to ensure). Nevertheless, it took 30 years to confirm what De Mairan observed, taking attention to temperature for the first time (Kuhlman et al., 2018; Roenneberg & Merrow, 2005).

The history of the circadian research have a lot of gaps between one discovery and another: the physiology of the endogenous nature of the clock in plants wasn't studied until 1832, despite De Mairan observations; similar observations in animals took another century, and 50 years more for humans (Kuhlman et al., 2018; McClung, 2006; Roenneberg & Merrow, 2005). Though circadian research and chronobiology were born in the 18th century, they took relevance and coherence in the 20th century. One of the breaking points in circadian research history was the international conference in Cold Spring Harbor, where 157 pioneers (Colin Pittendrigh, Patricia Decourse, Franz Halberg, etc) of this field met together for the fist time (Evans, 1961). From all the data shared, Pittendrigh summarized the qualities of circadian clocks in 16 generalizations that he predicted to be true in all organisms (Pittendrigh, 1960). Nowadays, 62 years later, these generalizations are still being useful though circadian research has suffered a huge development. The new technology approaches have enabled a more controlled experimental design and analysis of the data. The circadian community has already find circadian rhythms in almost all kind of organisms, even non-photosynthetic prokaryotes (Eelderink-Chen et al., 2021). Also the physiology and genetics behind the circadian clock have been broadly studied in a wide range of phyla, since the first gene of the clock was described in 1971 by Seymour Benzer and Ronald Konopka using mutant screening in *Drosophila melanogaster* (Konopka & Benzer, 1971; Takahashi, 2021). (Fig. 4)

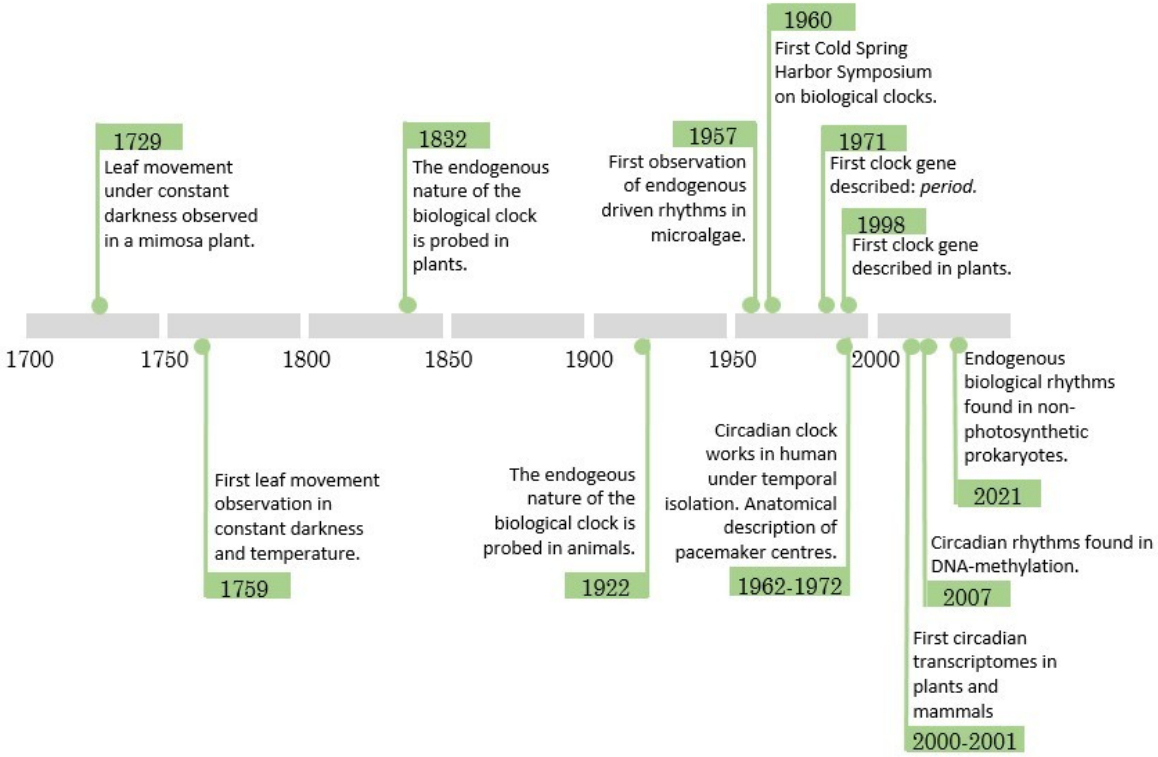

Figure 4: Timeline of circadian research. The main circadian discoveries have been listed in chronological order. (De Mairan, 1729; Eelderink-Chen et al., 2021; Evans, 1961; Konopka & Benzer, 1971; Kuhlman et al., 2018; McClung, 2006; Ripperger & Merrow, 2011; Roenneberg & Merrow, 2005; Takahashi, 2021)

As it can be observed in the timeline, circadian rhythms discoveries are very diverse. Chronobiologists usually have strong roots in other fields such as anatomy, physiology, molecular biology, genetics, ecology or even mathematics. The knowledge obtained from each field have been shared in order to obtain a better picture of circadian rhythms (Klante & Steinlechner, 1994; Merrow et al., 2005; Nishiwaki-Ohkawa & Yoshimura, 2016). Mathematics, however, have influenced the complete research field since the study of the biological rhythms as waves have been crucial. A wave's shape repeat itself over and over, maintaining several characteristics that define the wave as it is. Those characteristics are called wave parameters and are used to quantitative compare different waves. Some of them are: period (the time between two maximum points) or wavelength (distance between two maximum points), frequency (number of wave repetitions per seconds), amplitude (how much high the wave reach), phase (time point where the wave reaches its maximum high), mesor (mean level from which the wave fluctuates around), etc.

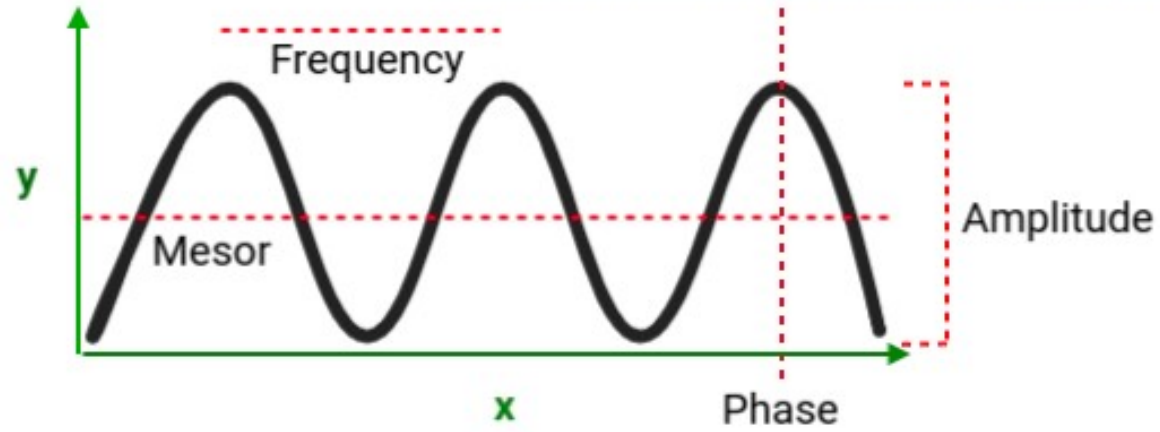

Figure 5: Graphical representation of classic waves parameters: period (the time between two maximum points or wavelength (distance between two maximum points), frequency (number of wave repetitions per seconds), amplitude (how much high the wave reach), phase (time point where the wave reaches its maximum high) and mesor (mean level from which the wave fluctuates around).

The most used ones for chronobiologists are amplitude, phase, period and mesor. The parameterization of waves using these four parameters enables to mathematically compare different groups of rhythmic data (McClung, 2006; Parsons et al., 2020). A deep study of circadian waves in *Ostreococcus tauri* was achieved during this work, using theses parameters to find and statistically validates differences between rhythms found in three biological levels (mRNAs, proteins and physiology).

Ostreococcus tauri

The green linage (Viridiplantae), comprehends two of the most important groups of oxygen photosynthetic eukaryotes: green algae and their descendants, terrestrial plants. Nowadays, the development of high-throughput sequencing has clarified the evolution history of these lineage (Bachy et al., 2022; Becker & Marin, 2009; Benites et al., 2021; Leliaert et al., 2012; Merchant et al., 2007). An early divergence of two discrete clades from an ancestral green microalgae is hypothesized. These two main clades are: the Streptophyta, including land plants and charophyte (green algae that are their closest ancestors/older sisters); and the Chlorophyta, comprising the core chlorophytes and their closest ancestors/older sisters the prasinophytes (Bachy et al., 2022; Leliaert et al., 2012; Tragin & Vaulot, 2019). (Fig. 6)

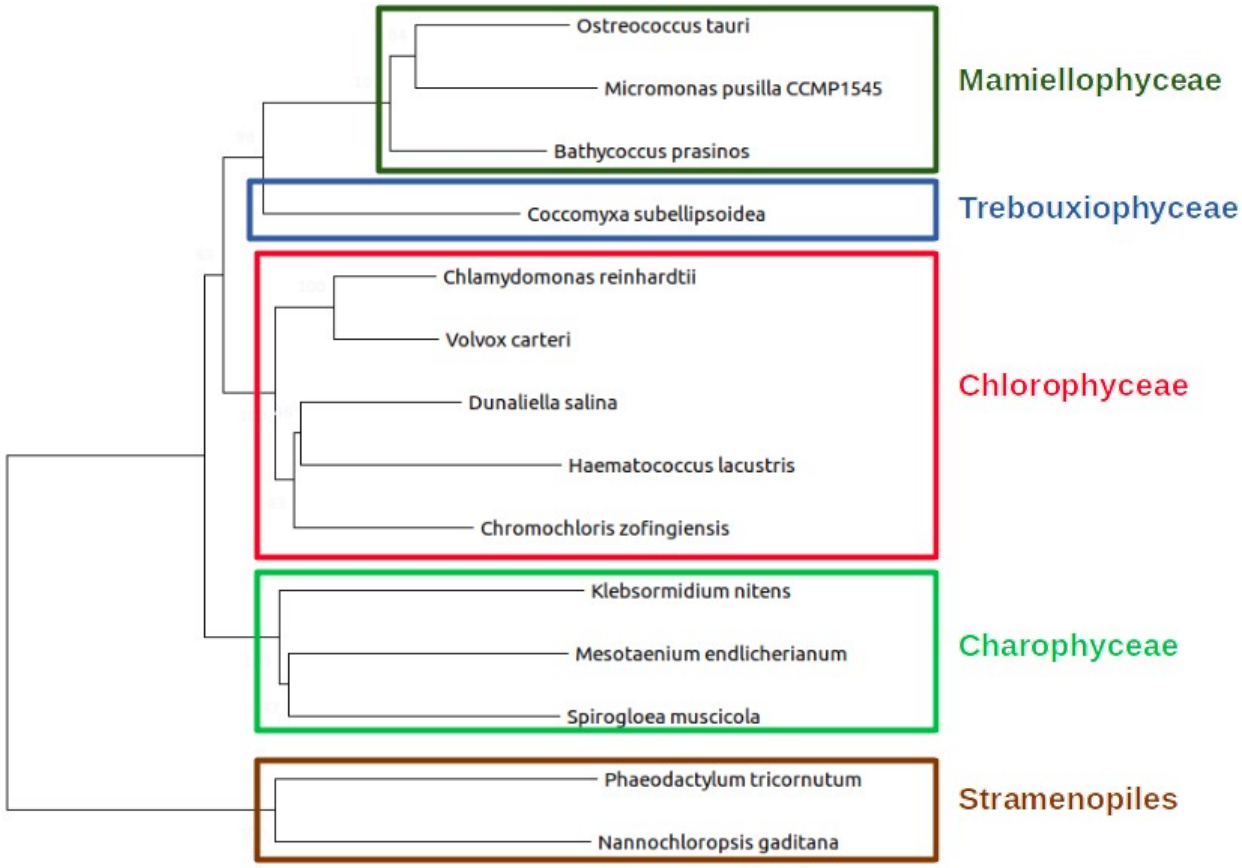

Figure 6: (mejorar arbol y escribir pie de foto)

Previous hypotheses posited the domination of the marine waters by the so-called red lineage conformed by diatoms and dinoflagellates. While the green lineage was thought to have less importance in marine water than in terrestrial environments (Worden et al., 2004). These two lineages established their differences a long time ago when the endosymbiotic events took place, giving rise to the green lineage (one single endosymbiotic event) and the red lineage (two or even more endosymbiotic events) (Leliaert et al., 2012).

However, in the last decade, metabarcoding studies have completed the previous hypotheses based in microscopic and traditional molecular techniques. Theses studies have confirmed the importance of the green lineage in marine waters. Also, the cosmopolitan distribution of prasinophytes has been described, especially the older sisters from the order Mamiellales (Collado-Fabbri et al., 2011; Demir-Hilton et al., 2011; Leconte et al., 2020; Tragin & Vaulot, 2019; Worden et al., 2004).

In that taxonomic context is placed *Ostreococcus tauri*, which is considered a green mamiellale microalgae. Its contribution to the marine phytoplankton is crucial in a wide range of oceans and seas all over the world (Benites et al., 2021; Collado-Fabbri et al., 2011; Demir-Hilton et al., 2011). As an example, different Ostreococcus strains have caused blooms in the Atlantic and Pacific Oceans (O'Kelly et al., 2003; Worden et al., 2004). Furthermore, a certain Ostreococcus strain was found to be the most prevalent Mamiellales in the Mediterranean Sea (Tragin & Vaulot, 2019). In spite of its ecologically important role (Chapman, 2013; Worden et al., 2004) and presence in natural environments, *Ostreococcus tauri* is the world's smallest free-living eukaryote known to date (around 1 µm). Due to its small size, *Ostreococcus tauri* have been "invisible" to field researchers for a long time. It was first described in a bloom that took place on the french Thau Lagoon. By that time, it was described as "undetectable" so the cells were discovered by flow cytometry. It contains a nucleus, only one chloroplast and one mitochondrion, a starch granule and a very reduced or almost non-existent cytoplasmic compartment (Moreau, H, Grimsley, N. Derelle, E, Ferraz, C, Escande, ML, Eychenié, S , Cooke, R, Piganeau, G, Desdevises, Y, Bellec, 1995).

METER FOTO DE TAURI

Meanwhile, genome sequencing approaches to understand marine phytoplankton were focus on the prokaryote component of it (Berube et al., 2018; Palenik et al., 2003). It wasn't until 2006 when the genome of *Ostreococcus tauri* was sequenced by the first time (Derelle et al., 2006). From that point, an increasing number of picoeukaryotes genomes were sequenced, including a second version of the genome of *Ostreococcus tauri* (Blanc-Mathieu et al., 2014). These contributions allowed the identification of unique aspects of its genome, showing that its simplicity goes among its reduced cell structure.

A genome size of 12.56 Mb distributed in 20 chromosomes. Nothing remarkable at first sight since Saccharomyces has similar numbers. The fascinating fact comes when the number of genes of theses two organisms are compared. With a similar genome size, *Saccharomyces cerevisiae* has around 6275 protein coding genes, while Ostreococcus tauri has 8166 (Blanc-Mathieu et al., 2014; Derelle et al., 2006; Engel et al., 2014). This makes *Ostreococcus tauri* the most gene dense free-living eukaryote known to date.

In addition to its short intergenic regions, *Ostreococcus tauri* has reduced the size of gene families keeping only one copy of each gene or even merged different genes in one. These features contribute to its intense degree of genome compaction. With the only exception of a long internal duplication on chromosome 19, hypothesized to be of recent origin due to its lack of divergence. But there is more perplexing data from the chromosome 19 and also from chromosome 2. They contain 77% of the transposable elements of the genome, they have lower G+C content and even a different codon usage in low G+C content loci of chromosome 2 (Blanc-Mathieu et al., 2014; Derelle et al., 2006). The first hypotheses about these chromosomes were that they had a different origin that the rest of the genome. Currently, only the chromosome 19 is considered an alien chromosome since the most of its protein coding genes aren't related to the green lineage. However, chromosome 2 protein coding genes are essential housekeeping genes not duplicated and related to the green lineage, so from that point it was considered a sex-related or mating-type chromosome (Benites et al., 2021; Blanc-Mathieu et al., 2014) .

Although sex is now accepted as a ubiquitous and ancestral feature of eukaryotes (Sekimoto, 2017; Swanson et al., 2011), direct observation of sex is still lacking in most unicellular eukaryotic lineages. These type of genomic regions so-called mating-type has been characterized in other Chlorophyta (Sekimoto, 2017) and, recently, in *Ostreococcus tauri,* which appears to encode two highly divergent haplotypes. These Mamiellales mating-types regions candidates are likely to be the oldest mating-type loci described to date (Benites et al., 2021; Leconte et al., 2020).

All in all, Ostreococcus tauri is proposed as a novel model organism due to its structural and genomic features. In addition, inside the green lineage there is a lot of diversity and actually it is difficult to find a model organism that can represent the hole lineage (Cock & Coelho, 2011). However, its taxonomy classification makes Ostreococcus tauri a potential green lineage ancestor (Derelle et al., 2006; Leliaert et al., 2012) and the knowledge gained using it can be easily extrapolated to a wide range of photosynthetic organisms.

Also, studies in Systems Molecular Biology often deal with the problem that complex organisms maximize the issue to study since massive data is generated in order to study complete biological systems (De Keersmaecker et al., 2006; Jamers et al., 2009; Joyce & Palsson, 2006; Weckwerth, 2011).

***Table 1**: Genomic features of different green lineage model organisms. Their genome size and protein coding genes (Blaby et al., 2014; Blanc-Mathieu et al., 2014; Craig et al., 2021; Derelle et al., 2006; Hori et al., 2014; Lamesch et al., 2012; Swarbreck et al., 2008; Yang et al., 2018) are compared with their number of predicted transcription factors. (Rayko et al., 2010; Zheng et al., 2016)*

	Genome size (Mb)	Number of protein coding genes.	Number of transcription factors.
Arabidopsis thaliana	135	27474	1779
Klebsormidium nitens	0	17055	286
Chalmydononas reinhardtii	110	14000	279
Phaeodactylum tricornutum	27.4	10567	212
Ostreococcus tauri	12.56	6275	102

Nevertheless, when Ostreococcus tauri is compared with another microalgae that are already sequenced and had been used as model organisms, it turns clear the simplification that it brings to Systems Molecular Biology studies (de los Reyes et al., 2017; Derelle et al., 2006; Krumholz et al., 2012; Le Bihan et al., 2011; Lelandais et al., 2016). *Ostreococcus tauri* allowed us to generate, analyze and interpret massive data from the complete system with less computational and experimental costs, so an holistic understanding of the chronobiology in this mammiellale was finally achieved.

Systems Biology.

For a scientist like me, it has been a passionate quarter of a century to live. First-generation DNA sequencing methods were developed only about 45 years ago. In addition, one of the most biggest international projects, the Human Genome Project, started about 20 years ago. (Ideker et al., 2001; Veenstra, 2021) In this project an extremely ambitious idea was pursued: to sequence the hole human genome. However, its impact in science transcends beyond that goal. Different groups cooperated all over the world, not only to sequence the complete human genome, but to develop new sequencing methods in order to

make the process easier and cheaper (Abascal et al., 2020). During the Human Genome Project, 13 years were needed to sequence a single complete human genome and the overall costs were 2.7 billion dollars.

Currently, with the emergence of next-generation sequencing methods, around 100000 human genomes have been sequenced in the last 6 years. The Illumina HiSeq System generates around 500 gigabase sequences per sun, dropping the cost of sequencing a complete human genome to $1500 in only 10 years of difference (Prendergast et al., 2020; Veenstra, 2021). Ultimately, it changed how science was approached.

All this advances led to a mass development of methods to sequence and identify the complete transcriptome (total mRNAs transcripts). Sequencing of complete transcriptomes offered a massive amount of information never seen before. However, genes and their products (proteins) are highly related since they interact and regulate each other forming positive and negative feedback loops, so still a lot of information was missing. Researchers began to understand organisms and process as systems with important modules (mRNA, proteins, metabolites, etc) that interact with each other forming parts of large networks.(Joyce & Palsson, 2006; Veenstra, 2021; Weckwerth, 2011).

While this holistic view of biological systems were gaining strength, traditional reductionist methods (focusing in only one gene, protein or metabolite) were the most popular and asequible ones. Scientific research was limited by the time and effort needed to complete integrate all functions that occur simultaneously within a biological system from a set of individual results, which has been almost impossible to achieve (Karahalil, 2016; Mazzocchi, 2012; Veenstra, 2021). There is a ancient Indian fable that illustrate how an holistic view and ontological reasoning contributes to knowledge, it is called “Blind men and an elephant”. A poem of John Godfrey Saxe is one of the most famous written versions of it:

“ It was six men of Indostan
To learning much inclined,
Who went to see the Elephant
(Though all of them were blind),
That each by observation
Might satisfy his mind.

The First approached the Elephant,
And happening to fall
Against his broad and sturdy side,
At once began to bawl:
"God bless me! but the Elephant
Is very like a wall!"

The Second, feeling of the tusk,
Cried, "Ho! what have we here
So very round and smooth and sharp?
To me 'tis mighty clear
This wonder of an Elephant
Is very like a spear!"

The Third approached the animal,
And happening to take
The squirming trunk within his hands,
Thus boldly up and spake:
"I see," quoth he, "the Elephant
Is very like a snake!"

The Fourth reached out an eager hand,
And felt about the knee.
"What most this wondrous beast is like
Is mighty plain," quoth he;
"Tis clear enough the Elephant
Is very like a tree!"

The Fifth, who chanced to touch the ear,
Said: "E'en the blindest man
Can tell what this resembles most;
Deny the fact who can
This marvel of an Elephant
Is very like a fan!"

The Sixth no sooner had begun
About the beast to grope,
Than, seizing on the swinging tail
That fell within his scope,
"I see," quoth he, "the Elephant
Is very like a rope!"

And so these men of Indostan
Disputed loud and long,
Each in his own opinion
Exceeding stiff and strong,
Though each was partly in the right,
And all were in the wrong! (...)”

After touching different parts of the animal, one of them concluded that the elephant was like a snake (he was touching only the elephant’s trunk), another one concluded that it was like a fan (since he was only touching the animal’s ear), and so on. Each of them were sure about their findings, but reaching an agreement was impossible since they didn’t treated their collected data as parts of a complex system, instead of independent truths.

During a long time in scientific research, systems biology studies that would bring that holistic view of living systems have been impossible to achieve due to a lack of technologies available. The development of systems biology is related to technology advances (Fig. 8) such as computational science, artificial intelligence and, the already mentioned,

next-generation sequencing methods (Ideker et al., 2001; Karahalil, 2016; Veenstra, 2021; Weckwerth, 2011).

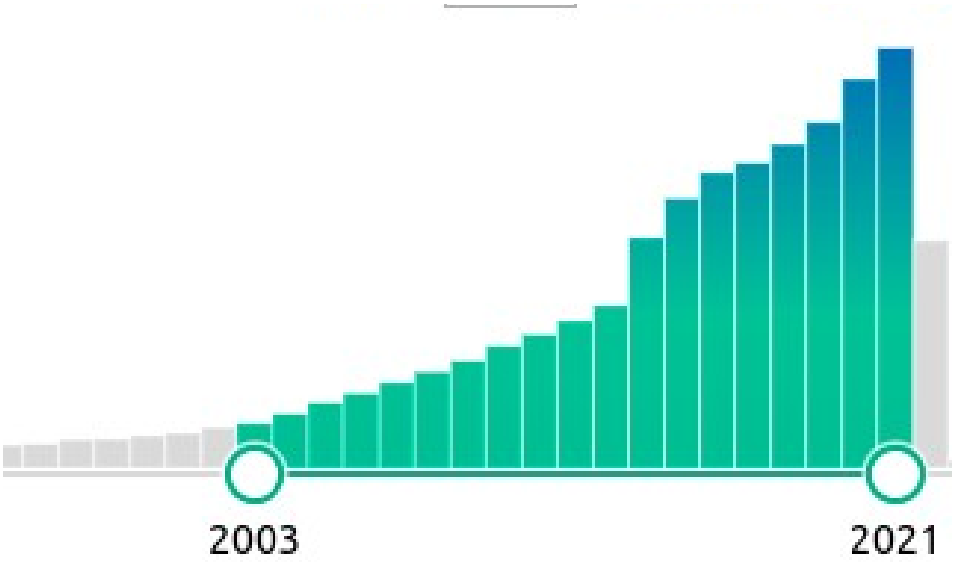

Figure 8: (es provisional, haría una mas bonita y mejor) Exponential increase in the number of publication using the term "systems biology" in PubMed since the year when the Genome Human Project was completed.

Nowadays, the so-called omics methods allow to mass measure all those important modules or biological levels that form part of the studied systems: transcriptomics, proteomics, metabolomics, etc. Systems biology aims to improve our understanding of living systems through the integration of our knowledge on how different biological components work simultaneously. The typical methodology of a systems biology study starts by obtaining omic data and its subsequent integration. Then, the results are computationally and statistically analyzed and the phenomenon observed is experimentally probed. Consequently, predictive models can be developed in order to simulate how a biological system would behave when perturbed (Jamers et al., 2009; Veenstra, 2021; Weckwerth, 2011; Zurbriggen et al., 2012).

During the progression of this work, the generation and analysis of massive data from two different omics techniques (transcriptomics and proteomics) is achieved. The main purpose is studying *Ostreococcus tauri* as a biological system by integrating the omics results with traditional physiological measurements as a biological validation of the effects observed in the computational analysis.

Specifically, the transcriptomic method used was RNA-seq, which is the main contemporary method used for this omic. The development of next-generation sequencing methods has contributed to RNA-Seq analyses enabling working with a wide variety of different classes of RNAs, not requiring transcript-specific probes (unlike microarrays, the almost obsolete previous method used in transcriptomic analyses), and not only to identify but also to quantify abundance of transcripts (Ditz et al., 2021; Veenstra, 2021; Wang et al., 2009).

Meanwhile, the field of proteomics have been directly connected with the development of mass spectrometry (MS) technology. During the first part of the century, there were two lines of development working separately for what would be known as proteomics today: identification (2-D electrophoresis gel) and quantification (using isotope tags) of proteins. Nowadays, technology has enable the development of proteins identification and quantification methods so the number of proteins that can be quantified/identified today is around several thousands (Shen et al., 2022; Veenstra, 2021). In this study, a large scale proteomic analysis is achieved by SWATH using liquid chromatography MS / MS. SWATH proteomics enables protein identification and characterization, as well as label free relative quantification (Chen et al., 2021; Ludwig et al., 2018).

Currently, the era of systems biology is increasing the amount of data generated per study and consequently, the computational and mathematical knowledge required to analyze and integrate these results increases too. Unfortunately, most laboratories are not historically designed to incorporate this requirements yet due to a lack of qualified researchers gathering solid knowledge from the different disciplines needed: computational sciences, mathematics/statistics and molecular biology. The lasts new generations of young researchers are working to develop software applications, efficient data analysis algorithms and user-friendly app-tools to enable the progress of systems biology studies making it more accessible for the hole scientific community (Coleto-Alcudia & Vega-Rodríguez, 2020; Romero-Campero et al., 2016; Romero-Losada et al., 2022).

However, systems biology studies in microalgae were recently started and there is a lack of tools for microalgae to analyze and interpret omics data. Consequently, during the progression of my doctoral thesis I aim to contribute to the progression of systems biology studies in the microalgae research community developing the web-app ALGAEFUN with

MARACAS. In that way, any researcher can analyze and functional annotate RNA-seq and ChIP-seq data without previous knowledge in computational or mathematical analysis (Romero-Losada et al., 2022).

In summary, my doctoral thesis aims to contribute to the microalgae research community with two major items: the development of free and open source tools that facilitates systems biology studies in microalgae; and the understanding of the diurnal and seasonal rhythmic changes in *Ostreococcus tauri* as a complete system knowing all the genes expressed, all the proteins present and how the altered functions are being executed.

Materials and Methods

Organism and culture growth conditions.

Organism and growth medium.

The sequenced strain of *Ostreococcus tauri,* RCC4221, was used in this study. Sterilized artificial see water (ASW) (Kester et al., 1967) supplemented nitrogen source, potassium source and vitamins were used as growing medium. A complete list of media components and concentrations needed to grow *Ostreococcus tauri* are described in Table 2.

Table 2: List of media components used in this study as growing medium for *Ostreococcus tauri* cultures.

	Concentration in solution	Concentration in medium
Solución I	400 g/L $NaNO_3$	222.22 mg/L $NaNO_3$

Solución II	2.8 g/L Na_2HPO_4	1.56 mg/L Na_2HPO_4
	10 g/L K_2HPO_4	5.56 mg/L K_2HPO_4
Solución III	5.36 g/L NH_4Cl	2.98 mg/L NH_4Cl
	10.4 g/L Fe-EDTA	5.78 mg/L Fe-EDTA
	74.4 g/L Na_2-EDTA	41.33 mg/L Na_2-EDTA
	$4.6 \cdot 10^{-2}$ g/L $ZnSO_4$	$2.56 \cdot 10^{-2}$ mg/L $ZnSO_4$
	$2.8 \cdot 10^{-2}$ g/L $CoSO_4$	$1.56 \cdot 10^{-2}$ mg/L $CoSO_4$
	$1.6 \cdot 10^{-2}$ g/L $Na_2MoO_4 \cdot 2H_2O$	$8.89 \cdot 10^{-2}$ mg/L $Na_2MoO_4 \cdot 2H_2O$
	$5.0 \cdot 10^{-3}$ g/L $CuSO_4$	$2.78 \cdot 10^{-3}$ mg/L $CuSO_4$
	$3.4 \cdot 10^{-2}$ g/L H_2SeO_3	$1.89 \cdot 10^{-2}$ mg/L H_2SeO_3
	$3.6 \cdot 10^{-2}$ g/L $MnCl_2 \cdot 4H_2O$	$2.0 \cdot 10^{-2}$ mg/L $MnCl_2 \cdot 4H_2O$
Solución IV	0,2 g/L Thiamin-HCl	0.22 mg/L Thiamin-HCl
	$1.50 \cdot 10^{-3}$ g/L Biotin	$1.67 \cdot 10^{-3}$ mg/L Biotin
	$1.50 \cdot 10^{-3}$ g/L Cyanocobalamin	$1.67 \cdot 10^{-3}$ mg/L Vitamin B12

Continuous culture conditions in photochemostats.

Photochemostats or photobiorreactors consisting in bubble column with a volume of 1.8 L of capacity (7 cm diameter, 47 cm height) were used for continuous regime cultivation in the laboratory. A detailed description of their design is represented in Figure 9. Each photochemostat is inoculated with the same cell concentration and are kept at batch regime during several days to ensure its adaptation before starting a continuous regime.

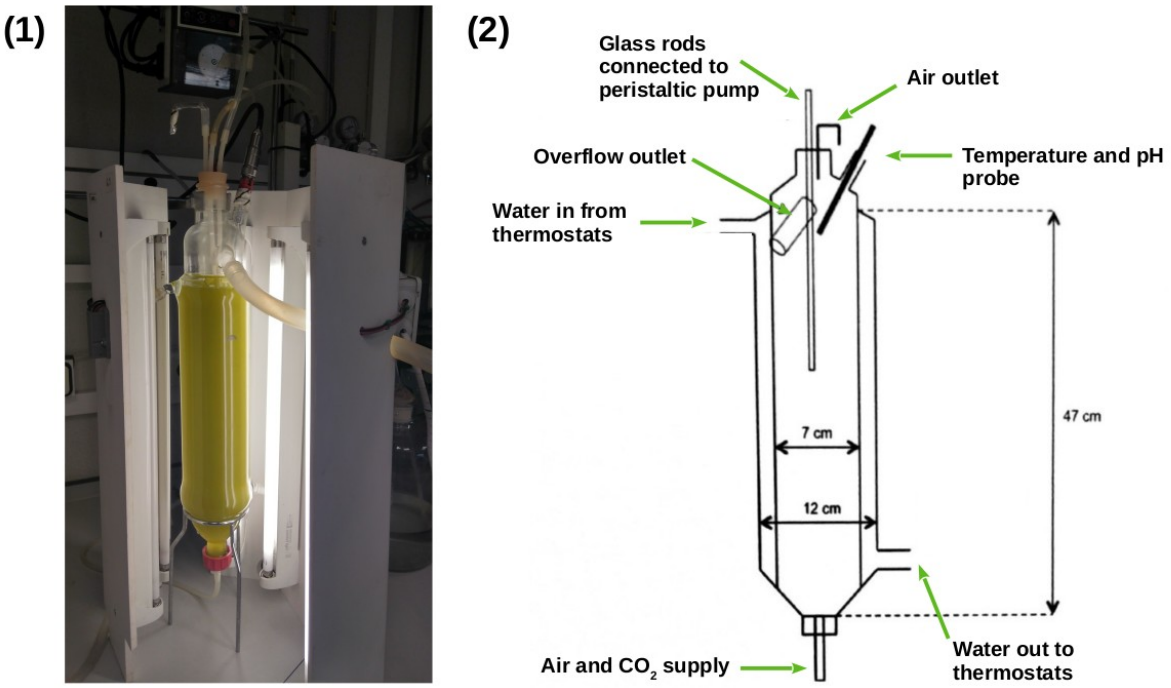

Figure 9: ***Growth system used for continous cultures****. (1) Picture of one of the photochemostats used. (2) Schematic detailed design of photochemostats.*

Photochemostats were provided with a double wall forming a water jacket circulating from a bath that kept the culture temperature constant at 20ºC. Air and CO_2 supply through the base of the chemostat ensured an optimal neumatic agitation. The air flow rate was regulated by a rotameter, adding 1 L air L culture 'min¹ (no entiendo este dato).

Culture conditions were constantly measured and computationally controlled by a LabJack: pH, dilution rate and illumination regime. For pH control, the reactor was equipped with a combined pH electrode (Crison, model 3100/225), connected to a pH controller (Crison, model pH 217/220-R1). The controller opened a electrovalve when the pH exceeded the threshold value stablished, allowing pure CO_2 to pass through along with the air stream. Once the pH value was recovered, the electrovalve was closed, leaving the culture medium in neumatic agitation exclusively by the described air supply.

To operate in continuous mode, the upper part was closed with a silicone cover fitted with glass rods of 4 mm internal diameter that reached the middle part of the reactor, allowing the supply of fresh medium. The dilution rate were maintained constant at 0.3 d^{-1} (overflow is descarded to avoid volume increase) by a peristaltic pump (P-3, Pharmacia) in order to

keep photobiorreactors in steady state. Cultures at steady state with 45 µg/mL chlorophyll content were used in our experiments. When a photobiorreactor reaches the steady state it means that the culture is growing at exponential phase but a constant concentration of cells is reached. This allows to take samples during long periods of time without changing the culture conditions.

In addition, instead of a sudden transition from dark to light and from light to dark, our Lab-Jack controlled system gradually increased light until an irradiance of 1500 µE m^{-2} s^{-1} is reached simulating the natural photoperiod (Fig. 10). Each photochemostat is illuminated using six Phillips PL-32 W/840/4p white-light fluorescent lamps. Furthermore, they are surrounded by a wooden box and a completely opaque fabric to avoid any external light input.

Experimental design.

The study focused on the two extreme photoperiods: summer long day conditions (LD, 16h light : 8h dark) and winter short day conditions (SD, 8h light : 16h dark). The experimental design (Fig. 10) consisted of three days under long day or short day conditions followed by three days under free running conditions (constant light or constant dark). Cells were harvested at specific times in the daily cycle, expressed in zeitgeber time (ZT), where ZT0 corresponds to dawn, ZT4 to 4h after dawn and so on. In this study, samples were taken every 4h (from ZT0 to ZT20) during the three days of alternating light/dark cycles, so there is a total of 6 samples for each day of sampling. No samples were collected during the first day of free running condition to allow culture acclimation. Then samples were collected once again every four hours starting at subjective dawn during two days.

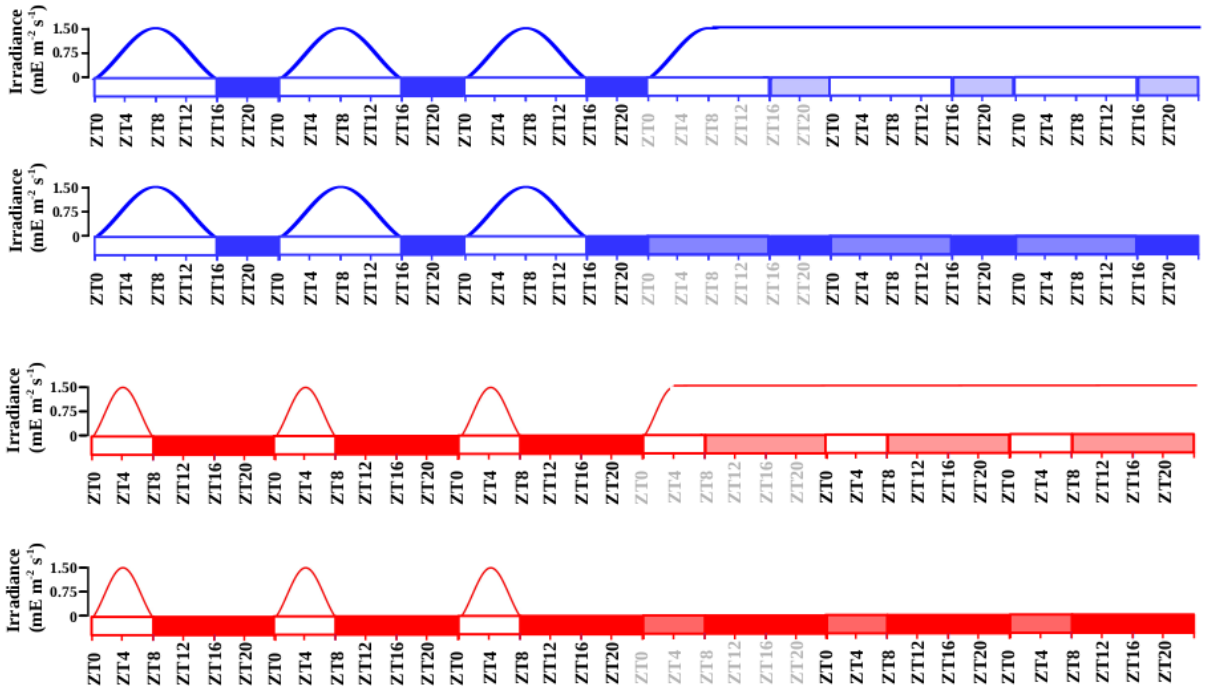

Figure 10: ***Schematic description of the experimental design.*** *Gradually increased and decreased irradiance is represented during each light phase. Long day conditions are represented in blue and short day conditions in red. Photoperiods (light periods) correspond to white rectangles and skotoperiods (dark periods) to blue/red filled rectangles. Light blue and red filled rectangles are used to represent subjective photoperiods and skotoperiods under free running conditions.*

Transcriptomic analysis

Sample Collection

From each chemostat, a volume of 50mL of cell suspension were harvested (4 min centrifugation at 5000 x g and 4ºC) for each time point. Pellets were washed using Phosphate-buffered saline solution (PBS) and flash frozen with liquid Nitrogen before stored at −80ºC.

Cell disruption

Frozen pellets were resuspended in 400 µL of disruption buffer (García-Domínguez & Florencio, 1997) and directly added to a 1,5 mL Eppendorf tube (RNAse free and phenol-proof) containing 400 µL of phenol:chloroform 1:1 and 100 µL of acid washed glass beads

(0.25–0.3 mm diameter; Braun, Melsungen, Germany). Mechanical disruption was performed by 30 min of repeated cycles of 60 s of vortexing and 60 s of incubating on ice.

RNA extraction

The tubes containing cellular extracts, disruption buffer, phenol:chloroform and acid washed glass beads were centrifuged (4 ºC) for 15 min at 13000 x g. There different phases appear after the centrifugation: an upper aqueous phase containing RNA, a white interphase containing DNA and a lower organic phase containing proteins, lipids and glass beads. The upper aqueous phase is collected and added to a new tube containing 400 µL of phenol:chloroform 1:1. This process was repeated three times more with centrifugations of 5 min, followed by a step using only chloroform to avoid phenol contamination of the samples. Finally the supernatan is incubated overnight at -20 ºC in a solution of 80 µL 10 M LiCl and 550 µL of 100% EtOH for RNA precipitation.

After incubation, samples were centrifuged 10 min at 13000 x g and 4ºC. Pellets were dryed to avoid EtOH contamination.

RNA purification

RNA purification was performed using the Isolate II RNA Plant Kit (Bioline). Washing, DNase treatment and elution were carried out following the manufacturer instructions. The final RNA concentration and integrity were measured using a bioanalyzer 2100 (Agilent RNA 6000 Nano Kit).

RNA sequencing and processing

Library was prepared in accordance with the manufacturer's instructions and the sequencing was carried out on the Illumina NextSeq500 sequencer. Approximately, 10 million 75nt long single end reads were generated for each sample. The *Ostreococcus tauri* genome sequence and annotation v3.0 were used as reference genome (Blanc-Mathieu et al., 2014). Futher computational analysis were carried out using MARACAS from ALGAEFUN with MARACAS (Romero-Losada et al., 2022).

Proteomic analysis

Sample collection

From each chemostat, a volume of 50mL of cell suspension were harvested (4 min centrifugation at 5000 x g and 4ºC) for each time point. Pellets were washed using Phosphate-buffered saline solution (PBS) and flash frozen with liquid Nitrogen before stored at −80ºC.

Cell disruption

Directly onto frozen pellets, 1 mL of Trizol, 100µL of acid washed glass beads (0.25–0.3 mm diameter; Braun, Melsungen, Germany) and 40µL of Protein Inhibitor Cocktail PIC (25x) were applied. Followed by 3 disruption cycles (60s agitation-60s incubation on ice) using a Mini-Beadbeater (BioSpec Products).

Proteins extraction

Proteins were extracted using TRI Reagent (Sigma-Aldrich), according to the manufacturer's instructions. The resulting proteins pellets were washed with 2mL of 0.3 M guanidine solution in 95% EtOH. They were resuspended by 10min of sonication (falta modelo) cycles (30 s sonication - 30 s of incubating at 4 ºC) and then centrifugated at 4 ºC during 5 min at 8000 x g. This washing process was repeated twice, followed by two additional washing using 90% EtOH instead of the guanidine solution. The final pellets were resuspended in ammonium bicarbonate 50 mM/0.2% Rapidgest (Waters) and total proteins were quantified using Qubit system.

Proteins digestion.

Of each samples, 50 µg of proteins were incubated with DTT (final concentration 4.5 mM) for 30 min at 60 ºC. Then, iodoacetamide to a final concentration of 10 mM were added to continue the incubation for 30 min more, under total darkness at room temperature. The treatment with trypsin was done overnight at 37 ºC in a 1:40 trypsin:protein. After that, formic acid (concentración?) was added and incubated at 37ºC for 1h. Finally, 2% acetonitrile (v/v) were added to reach a concentration of the digested sample around 0.5 µg of protein/µl of solution.

SWATH acquisition

Equipment and data acquisition method.

The analysis were performed on a time-of-flight TOF triple quadrupole hybrid mass spectrometer MS (5600 plus, Sciex) equipped with a nano electrospray source coupled to an nanoHPLC Eksigent model 425. The Sciex software Analyst TF 1.7 was used for equipment control and data acquisition. Peptides were first loaded onto a trap column (Acclaim PepMap 100 C18, 5 µm, 100 Å, 100 µm id × 20 mm, Thermo Fisher Scientific) under isocratical order in 0.1 % formic acid/2% acetonitrile (v/v) at a flow rate of 3 µL/min for 10 min. Subsequently, they were eluted on a reversed-phase analytical column, with the built-in emitter (New Objective PicoFrit column, 75 µm id × 250 mm, packed with Reprosil-PUR 3 µm). In the case of the samples corresponding to the short day conditions, the analytical column was Acclaim PepMap 100 C18, 3 µm, 100 Å, 75 µm id × 250 mm, Thermo Fisher Scientific, coupled to a PicoTip emitter (F360-20-10-N-20_C12 from New Objective). Peptides were eluted with a linear gradient of 5-35 % (v/v) of solvent B in 120 min at a flow rate of 300 nL/min. Formic acid 0.1 % (v/v) was used as solvent A and acetonitrile with formic acid 0.1 % (v/v) were used as solvent B. The source voltage was selected at 2600 V and the temperature was maintained at 100 ºC. Gas 1 was selected at 20 PSI, gas 2 at zero, and curtain gas at 25 PSI.

For proteins identification, Data Dependent Acquisition DDA method were used. It consisted of a TOF-MS with a scan window of 400-1250 m/z (accumulation time of 250 ms) followed by 50 MS/MS with a scan window of 230-1500 m/z (accumulation time of 65 ms) and with a cycle time of 2574 s.

Library construction

The spectral library was constructed by making 1 run with a mixture of the biological replicates corresponding to each time point (ZTO, ZT4, ZT8, ZT12, ZT16, ZT20) with the DDA method described. ProteinPilot v5.0.1 software (Sciex) was used to identify the proteins in the library. A pooled search of all runs was performed. The parameters of the Paragon method were: trypsin as enzyme and iodoacetamide as cysteine alkylating agent.

The Ostreococcus tauri annotated proteome file from ORCAE (Sterck et al., 2012) linked to a Sciex Contaminants database were used in library construction. A false positive analysis (FDR) was performed and those with FDR < or equal to 1 were considered.

SWATH runs

For each sample, the equivalent of 1 µg of digested protein was injected into each run. Before that, a standard (MS synthetic peptide calibration kit from Sciex) was injected to self-calibrate the equipment, control the sensitivity and chromatographic conditions. The described DDA method was used for SWATH runs with 60 ms of accumulation time and 3.7 s of cycle time. (preguntar a rocio si cambia el metodo o no)

Data processing

The library generated by DDA (1 % FDR) was used in the analysis performed using the Sciex software PeaKView 2.2 with the microapp SWATH 2.0, together with the data obtained from the SWATH runs. Using this program, the chromatographic traces of the ions were extracted and dumped into the Marker view 1.2.1.1 program where the list of identified proteins with their corresponding areas was generated. The parameters for extraction of ions and obtaining the areas were: 10 peptides per protein, 7 transitions of each peptide, threshold of confidence of the peptides set at 90 and FDR 1%. The parameters for extraction of the ions and obtaining the areas were: 10 peptides per protein, 7 transitions of each peptide, threshold of confidence of the peptides set at 90 and FDR 1%.

The software NormalyzerDE 1.6.0 (Willforss et al., 2019) was used to test several normalization methods in order to probe which one achieved the minimum replicate variation relative to Log2. Quantile normalization was selected as normalization method based on this comparison. Data were imputed with mean (mean imputation method), which means that the missing value on a certain variable is replaced by the mean of the available cases.

Cell cycle analysis

Sample collection and cell fixation method

A volume of 1.5mL of cell suspension were harvested for each time point. These samples were diluted 1:10 in PBS to be sure the cell concentration is suitable to the assay. Two mL of these dilutions were centrifugated (condiciones??preguntar a Meriyou) and cells in the pellets were fixed with 10 mL of 100% EtOH before stored at -20ºC for at least 24h.

Cell staining method

After fixation, cell suspensions were centrifuged for 5 min at 3500 x g (room temperature) and resuspended in 1 mL of PBS. They were washed once with PBS and sonicated for 3 minutes in a Ultrasonic Cleaner (JSP, US21, ultrasonic power 50W), in order to eliminate cell clumps and aggregates before staining.

In the staining process, 2µL of the Vibrant DyeCycle Green (V35004, ThermoFisher) were added to each sample and incubated 30 min (37°C) for DNA labeling. Final stain concentration was 10 µM. Treatment with RNase was not necessary as Vibrant DyeCycle Green is a DNA-selective stain. After incubation, cells were washed (con qué??? preguntar a meriyou) and transferred to flow cytometry tubes for cell cycle analysis.

Data acquisition and processing

Flow Cytometry acquisition were performed with a BD FACS Canto II (BD Biosciences) where stained DNA were excited by a 488nm laser and emission was collected in a 530/30 nm PMT. Flow rate was low and linear amplification were established for the acquisition.

Data were analyzed using FlowJo v.10.6.1 (Becton Dickinson & Company BD). Analysis was performed using one of the univariate cell cycle platform that FlowJo provides, specifically the Watson pragmatic algorithm (Watson et al., 1987) to adjust the data to the model.

Analysis of photosynthetic activity

Sample collection

Fresh culture was harvested at the different specific times of the day. The samples were diluted 1:1 with growing medium and incubated at 20 ºC in total darkness during 10 min.

Data acquisition

In order to analyze photosynthetic parameters, Pulse-Amplitude-Modulation PAM fluorometry measurements were performed using a Waltz DUAL-PAM-100. After the darkness incubation, the non-actinic modulated light (450nm, 2.8 µE µE m^{-2} s^{-1}) was turned on in order to measure F_o (fluorescence basal level). Then, to determine F_M (the maximum fluorescence level), a saturating red light pulse of 655nm and 5000 µE m^{-2} s^{-1} is applied to the

sample during 400ms. The F_V/F_M, that corresponds to the maximum potential quantum efficiency of Photosystem II when all reaction centers were open, was calculated as:

$$F_V / F_M = (F_M - F_O) / F_M$$

Analytical determinations

Sample collection

At each time point, 50 mL of fresh culture was harvested and centrifuged at 7000 x g during 10 min. Then pellets were washed with 1% ammonium formiate (p/v) to eliminate salts from the growing medium and lyophilized.

Starch Content

Cell disruption

Approximately 2-3 mg of lyophilized biomass of each sample was added to hermetic tubes containing 1 mL of glass beads (0.25–0.3 mm diameter; Brau n, Melsungen, Germany) and 2 mL of chloroform:methanol (2:1). Three disrupting cycles (60s agitation-60s incubation on ice) were applied to the samples using Mini-Beadbeater (BioSpec Products). Then cellular extracts were separated from the beads and saved in new tubes. Cellular extracts were centrifuged for 4 min at 13000 x g and the supernatant was discarded. The addition of chloroform:methanol (2:1) and centrifugation steps were repeated until the pellets were white in order to ensure the elimination of pigments and lipids that could disturb the determination process. Finally, pigment free pellets were dryed.

Starch solubilization and digestion

Proposed protocol for plants in (Rufty & Huber, 1983) was adapted to *Ostreococcus tauri.*

Starch granules in dry pellets were alkaline solubilized with 1mL of 0.2 M KOH and heated at 100 ºC. After 30 min, samples were gradually cooled and pH was adjusted to 5.0 by adding 300 µL of 1 M acetic acid.

Starch digestion involves the breakdown by 7.4 U of α-amylase to small linear and branched oligosaccharides during a 30 min incubation at 37ºC; and the release of glucose

residues by 5 U of amyloglucosidase during 1-2 h incubation at 55 ºC. Both diluted in 0.1 M of sodium acetate pH 4.5 were added to 200 µL of each sample. Finally, in order to stop enzymatic reactions, the samples were incubated at 100 ºC for 2 min and centrifuged at 13000 x g for 10 min to discard any pellet.

Spectrophotometric quantification

The quantification of released glucose residues from starch was achieved following (Rufty & Huber, 1983) protocol. They propose the combination of two different enzymatic activities: hexoquinase that phosophorylates glucose residues; and glucose-6-phosphate dehydrogenase (G6PDH) that reduce NAD+ using the phosphorylated glucose generated by the hexoquinase. NADH absorbance can be measured at 340nm and its concentration is related to glucose residues concentration 1:1 (Fig. 11).

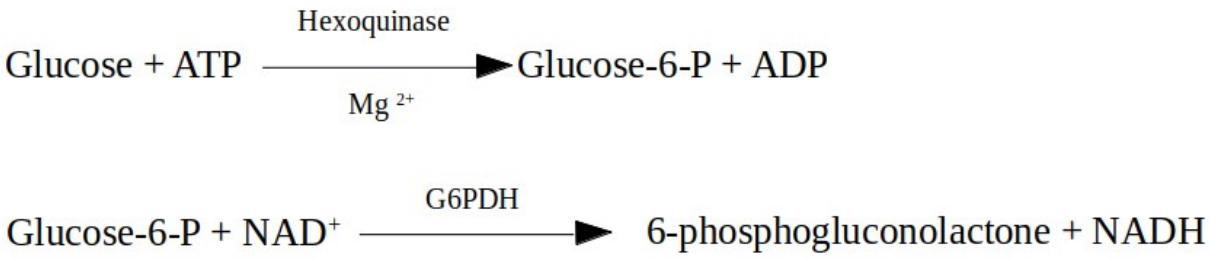

Figure 11: Enzymatic activities that link NADH production to glucose residues released from starch.

To achieve that measurement, quartz spectrophotometer cuvettes were filled with :

- 100 µL of the sample
- 500 µL of hexoquinase buffer (100mM HEPES pH 7.7, 10 mM $MgCl_2$, 0.04% BSA p/v, 1mM DTT)
- 100 µL of ATP mix (containing 10mM ATP diluted in Hepes 100 mM ph 7.7)
- 100 µL of NAD+ mix (containing 4 mM NAD+ diluted in Hepes 100 mM ph 7.7)
- 2 µL of 2.5 U μL^{-1} glucose-6-phosphate dehydrogenase
- 200 µL of destile water.

The absorbance of that mixture was measured at 340 nm twice. The first measure is followed by the addition of 5 µL of 1 U μL^{-1} hexoquinase enzyme and the incubation at 27 ºC for 20 min to ensure the efficiency of the reaction. The second measure was used to calculate the amount of NADH produced during the reaction. Using calibration curve based on processed samples that contained known concentrations of starch, NADH absorbance is related to the amount of glucose residues and consequently to the initial amount of starch in each the sample.

Carotenoid Content

Cell disruption

Four milligrams of lyophilized biomass was added to an hermetic tube containing 1 mL of glass beads (0.25–0.3 mm diameter; Braun, Melsungen, Germany) and 1 mL of pure acetone. Three disrupting cycles (60s agitation-60s incubation on ice) were applied to the samples using Mini-Beadbeater (BioSpec Products).

Carotenoids extraction

Carotenoids extraction was achieved following (Del Campo et al., 2004) proposed method. Darkness was maintain during the entire process to avoid pigments degradation.

Cellular extracts were collected and saved in new tubes after a 4 min centrifugation of the samples at 13000 x g. This process was repeated six times or until the supernatant turns colorless. Supernatans of each sample were joined in the same tube (one per each sample) and acetone were evaporated using a stream of nitrogen gas. Finally, 350 µL of acetone were added to each tube, and if was distributed in HPLC tubes following manufacturer's instructions.

Carotenoids determination and quantification

A Hitachi HPLC (Elite LaChrom), equipped with a photodiode-array detector (Hitachi L-2455) was used. Separation was performed on a Waters NovaPak C-18 (3.9×150 mm, 4 µm particle size, 60 Å pore size) column. Following CITA proposed method, the eluents used to create a gradient through the mobile phase were: eluent A (ammonium acetate 0.1 M and H_2O-methanol 15:85 v/v) and eluent B (methanol-acetonitrile-acetone 44:43:13 v/v).

Temperature was maintained constant (20 ºC) during the whole process and eluents flow at 800 µL min^{-1}.

Different carotenoids were identified following retention times and absorption profiles of previous known carotenoids analyzed. Quantification was calculated as a percentage of the total peak area.

Rhythmic patterns analysis

Rhythmic patterns detection

The R package RAIN (Thaben & Westermark, 2014) from Bioconductor was used to statistically identify rhythmic patterns in the different data collected.

Three complete diurnal cycles from both photoperiods were used to detect rhythmic patterns in the different data of this study: expression levels of genes, abundance of proteins, maximum potential quantum efficiency of Photosystem II, amount of cells in the different cell cycle phases and amount of different carotenoids and starch.

Rhythmic patterns with a single maximum point over a complete diurnal cycle were detected by setting the period parameter from RAIN to 24 hours. A similar process was used to detect more complex rhythmic patterns (the ones with two or even three maximum points per day), but changing the period parameter from RAIN to 12 and 8 hours respectively. A 0.05 p-value threshold was used in all scenarios.

In addition, the last two diurnal cycles and two consecutive days of continuous light or darkness were considered for the RAIN analysis described above. In that way, RAIN can statistically test if a similar pattern is maintained after changing the cycling light regime to a continuous light or darkness input and it prevents a bias towards any of the two conditions.

Rhythmic patterns comparison

Once that rhythmic patterns in our data were detected by RAIN, waves were characterized to enable comparison between them. This was carried out using the R package Circacompare that performs a fitting to the data as co-sinusoidal curve with a particular parametrization that is used to test for statistical significant differences between the two circadian pat-

terns under comparison. (Parsons et al., 2020). Specifically, a p-value threshold of 0.05 was used.

The significance of the global differences in the different rhythmic parameters was performed using the Mann-Whitney-Wilcoxon non parametric test implemented in the R function *wilcox.test*.

Hypothesis and Objetives

(texto de hipotesis)

Results

Chapter 1. ALGAEFUN with MARACAS: user-friendly tool for analysing and integrating omic data generated from microalgae.

In order to contribute to 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 (Blanc et al., 2012; Bowler et al., 2008; Cheng et al., 2019; Corteggiani Carpinelli et al., 2014; Hori et al., 2014; Merchant et al., 2007; Moreau et al., 2012; Morimoto et al., 2020; Ottesen et al., 2013; Palenik et al., 2007; Polle et al., 2017; Roth et al., 2017; Worden et al., 2009). This has promoted the emergence of molecular systems biology studies and the use of different omics like transcriptomics based on RNA-seq data (Hoys et al., 2021; Serrano-Pérez et al., 2022) and cistromics based on ChIP-seq data (Ngan et al., 2015; Zhao et al., 2021) in microalgae. Nonetheless, the progress of this type of studies on microalgae are limited by the lack of freely available and easy-to-use online tools to analyze, extract relevant information and integrate omics data. 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 some 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 (cita? No la encuentro) and other Bioconductor packages.

Our web portal 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 or ChIP-seq raw data from microalgae studies and produces set of differentially

expressed genes or lists of genomic loci respectively. 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 Gene Ontology (GO) (Ashburner et al., 2000; Carbon et al., 2019) and metabolic pathways enrichment analysis (Kanehisa et al., 2016; Moriya et al., 2007; Ogata et al., 1999). 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. 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

Integration of different microalgae databases.

ALGAEFUN with MARACAS supports a wide range of 14 different microalgae species that cover an ample spectrum of their phylogeny (Fig.12): *Chlamydomonas reinhardtii* (Blaby et al., 2014; Merchant et al., 2007), V*olvox carteri* (Prochnik et al., 2010), *Chromochloris zofingiensis* (Roth et al., 2017), *Dunaliella salina* (Polle et al., 2017), *Haematococcus lacustris* (Morimoto et al., 2020) (Chlorophyceae), *Coccomyxa subellipsoidea* (Blanc et al., 2012) (Trebouxiophyceae), *Ostreococcus tauri* (Blanc-Mathieu et al., 2014; Palenik et al., 2007), *Bathycoccus prasinos* (Moreau et al., 2012), *Micromonas pusilla* CCMP1545 (Worden et al., 2009) (Mamiellophyceae), *Phaeodactylum tricornutum* (Bowler et al., 2008; Yang et al., 2018), *Nannochloropsis gaditana* (Corteggiani Carpinelli et al., 2014; Radakovits et al., 2012) (Stramenopiles), *Klebsormidium nitens* (Hori et al., 2014), *Mesotaenium endlicherianum* (Cheng et al., 2019) and *Spirogloea muscicola* (Cheng et al., 2019) (Charophyceae). Supported species include microalgae used in basic scientific research, as well as those used in biotechnology industry like *H. lacustris* (Hoys et al., 2021), *N. gaditana (Ajjawi et al., 2017)* or *P. tricornutum (Cui et al., 2019; Pereira et al., 2021)* .

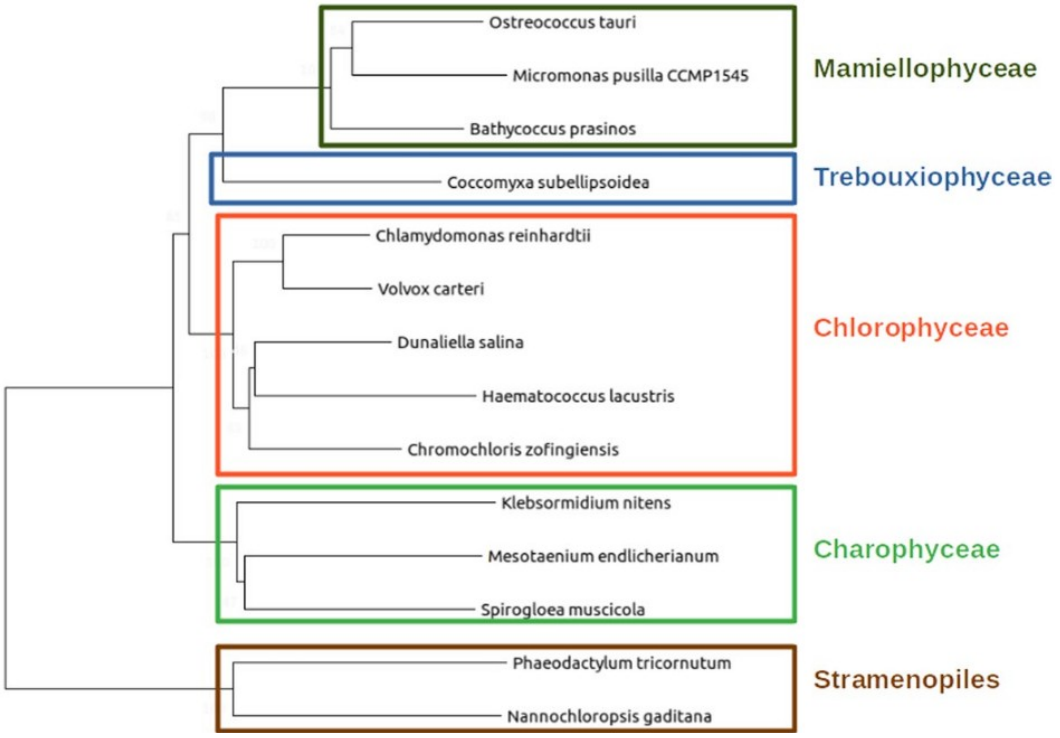

Figure 12: Phylogenetic relationship between the different microalgae species supported in ALGAEFUN with MARACAS. The Mamiellophyceae are represented using an olive green rectangle, the Trebouxiophyceae using a blue one, the Chlorophyceae using a red one, the Charophyceae using a green one and the Stramenopiles using a brown one.

One of the limiting factors mentioned before is the lack of frameworks integrating the available microalgae genome sequences and annotations. One of the goals of our tool is to overcome this issue and generate easily accessible resources, genome sequences, functional annotation and genomic feature annotation files (Gene transfer file *GTF*) for the already sequenced microalgae species. This data were systematically collected from different freely available data bases depending on the microalgae. Table 3 enumerates the different data bases used for each microalgae. Specifically, for N. gaditana and P. tricornutum, we accessed Ensembl protist (Howe et al., 2021) , a web based unicellular species genome browser storing gene annotation; for B. prasinos we used Orcae (Sterck et al., 2012) an online genome annotation resource built on the wiki philosophy; for C. reinhardtii, V. carteri, C. zofingiensis, D. salina and C. subellipsoidea the JointGenome Institute (JGI) / Phytozome (Goodstein et al., 2012), a web portal integrating omics for photosynthetic organisms was queried; for M. endlicherianum and S. muscicola a figshare associated to publication was accesed; M. pusilla, O. tauri, B. prasinos and K. nitens genome sequence and annotation was downloaded from the JGI / PhycoCosm (Grigoriev et al., 2021) , a

comparative algal genomic resource; and the genome of H. lacustris was available at NCBI genome database (cita web?).

Table 3: *Resources used to collect genome sequences, functional and gene feature annotations for each supported microalgae.*

Ensembl protists	PhycoCosm	Phytozome	Genomes NCBI datasets	Figshare associated to publication
N. gaditana	*B. prasinos*	*C. reinhardtii*	*H. lacustris*	*M. endlicherianum*
P. tricornutum	*K. nitens*	*V. carteri*		*S. muscicola*
	M. pusilla	*C. zofingiensis*		
	O. tauri	*D. salinas*		
		C. subellipsoidea		

Genome sequence and gene feature annotation files were downloaded in fasta format and GTF respectively. When necessary, different chromosome and/or scaffold files were colligated programmatically to produce a single genome file. The GTF format in the gene feature annotation files consists of a data frame with nine columns. Each line corresponds to a specific gene feature. The first eight columns must contain information related to the type of feature (3' UTR, 5' UTR, gene, CDS or mRNA), chromosome start and end positions, the strand were it is positioned, and some other attributes in a well-defined format. The ninth column is not restricted to a specific format and can contained any type of information. Nonetheless, the mappers used in MARACAS assume that this last column follows the format taken by GTF files in the data base Ensembl. In order to be able to use GTF files from other databases besides Ensembl, we developed an R script to translate any GTF file into the format followed by Ensembl and required by HISAT2 (Kim et al., 2015).

Systematic functional annotation files consisting of Gene Ontology (GO) and KEGG (Kyoto Encyclopedia of Genes and Genome) Orthology (KO) terms were also downloaded for each microalgae from the previously mentioned databases. Gene Ontology terms seeks the development of a human-readable and machine-readable hierarchical vocabulary to relate genes with their molecular functions, biological processes in which they are involved

and the cellular components where they perform their function (Ashburner et al., 2000; Carbon et al., 2019) . Complementary, KEGG Orthology terms associate genes to metabolic pathways and modules based on their orthologous relationships in sequenced genomes (Kanehisa et al., 2016; Moriya et al., 2007; Ogata et al., 1999) . However, for microalgae species lacking these annotation systems HMMER (biological sequence analysis using profile hidden Markov models) (Potter et al., 2018) was used to identify protein domains according to the PFAM (Protein Family) nomenclature (Mistry et al., 2021). PFAM terms were subsequently converted into GO terms using pfam2go (necesaria la cita? No encuentro na). KO terms were associated to genes applying KAAS (KEGG Automatic Annotation Server) (Moriya et al., 2007). Whenever possible other systematic functional annotation format were also included such as Protein Analysis Through Evolutionary Relationships (PANTHER) terms (Mi et al., 2021), used to classify based of evolutionary families and subfamilies gene products into classes capturing molecular function, biological process and metabolic pathways; Enzyme Commission numbers (EC numbers) that consists of a numerical classification identifier for enzymes related to the biochemical reactions they perform; and Eukaryotic Orthologous Groups (KOG) terms, used to identify ortholog and paralog groups of proteins (Galperin et al., 2021).

Development of functional annotation and genomic packages.

In order to use all these annotation systems in ALGAEFUN two different types of R annotation packages were developed and were made freely available from our Github repository (cito web?). On the one hand, the systematic sources of functional annotation discussed previously (GO terms and KO terms for every microalgae species and PANTHER IDs, EC numbers and KOG terms whenever available) were gather together using the function makeOrgPackage from the Bioconductor R package AnnotationForge (Carlson & Pagès, 2019) which has generated annotation packages for each microalgae. These packages are instrumental when performing functional enrichment analysis over gene sets obtained from RNA-seq data analysis.

On the other hand, gene featuring annotation of each microalgae from the previously downloaded and processed GTF files is stored applying the function makeTxDbFromGFF from the Bioconductor R package GenomicFeatures (Lawrence et al., 2013). They are central to carry out analysis over genomic loci obtained in ChIP-seq analysis.

ALGAEFUN functionalities heavily depend on these packages. We wanted to make them freely available on our Github repository, in order to enable the research community in microalgae to access them and perform omics analysis independently from the tools available from ALGAEFUN with MARACAS.

MARACAS implementation: high-throughput sequencing data processing.

The computational core of this tool consists of a parallel fully automatic computational pipeline or workflow synchronized through blackboards. This workflow is managed by the job scheduling system SLURM (Simple Linux Utility for Resource Management) and bash scripting. The inputs to our pipeline comprise the pre-computed index of the corresponding microalgae reference genome, the previously processed gene feature annotation file in GTF format (both already included in MARACAS) and the raw sequencing data in fastq format from RNA-seq and ChIP-seq microalgae studies provided by the user. In turn, this workflow produces as outputs two lists of differentially expressed genes or DEGs (activated and repressed genes) when the fastq files correspond to a RNA-seq study; or a list of genomic loci or regions significantly occupied by the transcription factor or histone modification of interest in the case of a ChIP-seq study. MARACAS requires the user to input specifications such as the microalgae of interest, the names for control and experimental conditions, number of replicates and location of raw high-throughput sequencing files. In case of analyzing already published omic data, its accession number can be loaded instead. Additionally, the user can set the statistical parameters to perform the corresponding analysis. Specifically, the fold-change and significance level cutoff thresholds for the identification of differentially expressed genes can be selected. Also, user can choose to use as read mapper software either the short read mapper HISAT2 (Hierarchical Indexing for Spliced Alignment of Transcripts 2) or the pseudoalignment method implemented in kallisto. Whereas HISAT2 is an exact method requiring several hours for processing a typical sample (Kim et al., 2015; Pertea et al., 2016), kallisto produces near-optimal gene expression quantification in only a few minutes (Bray et al., 2016). Meanwhile, in the MARACAS ChIP-seq data analysis pipeline is used the ultra-fast and memory-efficient short read mapper bowtie2 (Langmead & Salzberg, 2012). All this information is collected into a parameters file (Table 4 and Table 5), which is the main input received by our pipeline.

Table 4: *Explained parameters included in the parameters file used as input in MARACAS for RNA-seq analyses.*

Parameters file for RNA-seq analyses	
data_source	This parameter indicates the source of the data to be analyzed. It can take the value *FILES* when the fastq files are already located in a folder in the computer where MARACAS is installed or the value *DB* when the data has been already deposited in the GEO data base.
cluster	Parameter specifying the execution mode. *SERVER* mode executes MARACAS with a sequential analysis of the different samples. Whereas *SLURM* mode executes MARACAS in a parallel manner processing samples simultaneously in different computational nodes. In this last case SLURM needs to be installed in your computer cluster.
number_processors	Number of processors that can be used by MARACAS.
paired_end	It can take the values *FALSE* when your data is single end and *TRUE* when your data is paired end.
working_directory	It indicates the location where the analysis folder will be generated.
microalgae	Name of the microalgae of interest.
read mapper	This parameters specifies the software tool to perform read mapping. Two different options are provided *HISAT2* and *Kallisto*.
main_folder	Name of the folder that will be created at the working directory to contain the outputs from the analysis.
number_of_samples	Total number of samples to be analyzed.
control_condition_name	Name of control condition.
experimental_condition_name	Name of experimental condition.
loc_sampleN	When *paired_end: FALSE* and *data_source: FILES*, this parameter indicates the path and file name of sampleN. N will take values from 1 to *number_of_samples*.
acc_sampleN	Instead, when *data_source: DB,* this parameter specifies accession number of fastq files in GEO.
loc_sample_leftN **loc_sample_rightN**	When *paired_end: TRUE* and *data_source: FILES*, this parameters indicates the path and file name of the fastq samples containing the left and right reads, respectively.
condition_sampleN	This parameters specifies which condition name of the ones chosen in *control_condition_name* or *experimental_condition_name* correspond to each sample.
fold_change **q_value**	These parameters specify the fold-change and q-value used to determine differential expressed genes in the experimental condition when compared to the control condition.

Table 5: *Explained parameters included in the parameters file used as input in MARACAS for ChIP-seq analyses.*

Parameters file for ChIP-seq analyses	
data_source	*FILES* or *DB* as it is explained in Table 4.
cluster	*SERVER* or *SLURM* as it is explained in Table 4.
number_processors	Number of processors that can be used by MARACAS.
paired_end	*FALSE* or *TRUE* as it is explained in Table 4.
working_directory	It indicates the location where the analysis folder will be generated.
microalgae	Name of the microalgae of interest.
main_folder	Name of the folder that will be created at the working directory to contain the outputs from the analysis.
number_of_replicates	Total number of samples to be analyzed.
included_control	It can receive *yes* if your experimental design includes a control condition such as input, mock or similar. Use *no* in negative cases.
mode	Use the values *transcription_factor* or *histone_modification* to specify if your ChIP-seq data was generated for a transcription factor or a histone modification study.
transcription_factor	When *mode: transcription_factor* this parameter specifies the name of the chosen transcription factor.
histone_modification	When *mode: histone_modification* this parameter specifies the name of the chosen histone modification.
loc_chip_replicate_N **loc_control_replicate_N**	When *paired_end: FALSE* and *data_source: FILES*, this parameter indicates the path and file name of ChIP sample and, in case of *included_control: yes*, the control sample. N will take values from 1 to *number_of_replicates*.
chip_replicate_N **control_replicate_N**	Instead, when *data_source: DB,* this parameter specifies accession number of ChIP sample in GEO. In case of *included_control: yes,* control_replicate_N will specify the one for control sample in GEO.
loc_chip_replicate_left_N **loc_chip_replicate_right_N** **loc_control_replicate_left_N** **loc_control_replicate_right_N**	When *paired_end: TRUE* and *data_source: FILES,* this parameters indicates the path and file name of the ChIP samples containing the left and right reads, respectively. If *included_control: yes,* they will also indicate this information for control samples.

The first step in our workflow loads the parameters file and generates the working directory including one folder where the samples are downloaded or copied, another one where the results are generated and a third one called “logs”. In this last folder, several text files are generated containing messages written by the different processes in our pipeline that are useful to keep track of the important events. More interestingly, a different type of text file called blackboard is created and the computational processes in our workflow are given read and write permissions for it. Files functioning as blackboards are used for indirect communication between the computational processes in our parallel pipeline in order to synchronize them. In the next step, the pipeline forks into two different modes to process either RNA-seq or ChIP-seq data. The fork of our pipeline dedicated to RNA-seq analysis is represented in Fig. 13. Depending on the raw data source, the first step consists of either downloading and extracting the corresponding fastq files from a database or copying them to the specified working directory. Next, a quality control of these raw data is performed and short reads are mapped to the reference genome using HISAT2 or kallisto. When HISAT2 is used, this step generates BAM (Binary Alignment Maps) files that contain the mapped reads. The following steps take this mapping as input and perform transcripts assembly and expression quantification using STRINGTIE (Pertea et al., 2015, 2016). In this step, GTF files with the assembled transcripts and CTAB files (chemical table files) with the expression quantification values are generated for all samples. However, when kallisto is used, transcripts are mapped and quantified in the same step, so a TSV (tabular separated values) file containing the result of the quantification is generated (Bray et al., 2016). These processing steps are carried out in parallel simultaneously for each sample being synchronized using a blackboard file. These parallel processes write on the blackboard when a goal is reached (for example, BAM files are generated) in order to keep track of their progress. When all goals have been reached by all the parallel processes, the following sequential steps of the pipeline are executed. Then, differential expression analysis is carried out using the Bioconductor R packages Ballgown (Frazee et al., 2015; Pertea et al., 2016) and Limma (Ritchie et al., 2015). Specifically, DEGs are selected using the statistics based on a moderated t-student. During the final stage of this fork of our pipeline a differentially expression gene report is generated containing text files with activated and repressed genes, principal component analysis visualization, and several informative graphs such as scatter plots, volcano plots, box-plots and bar-plots.

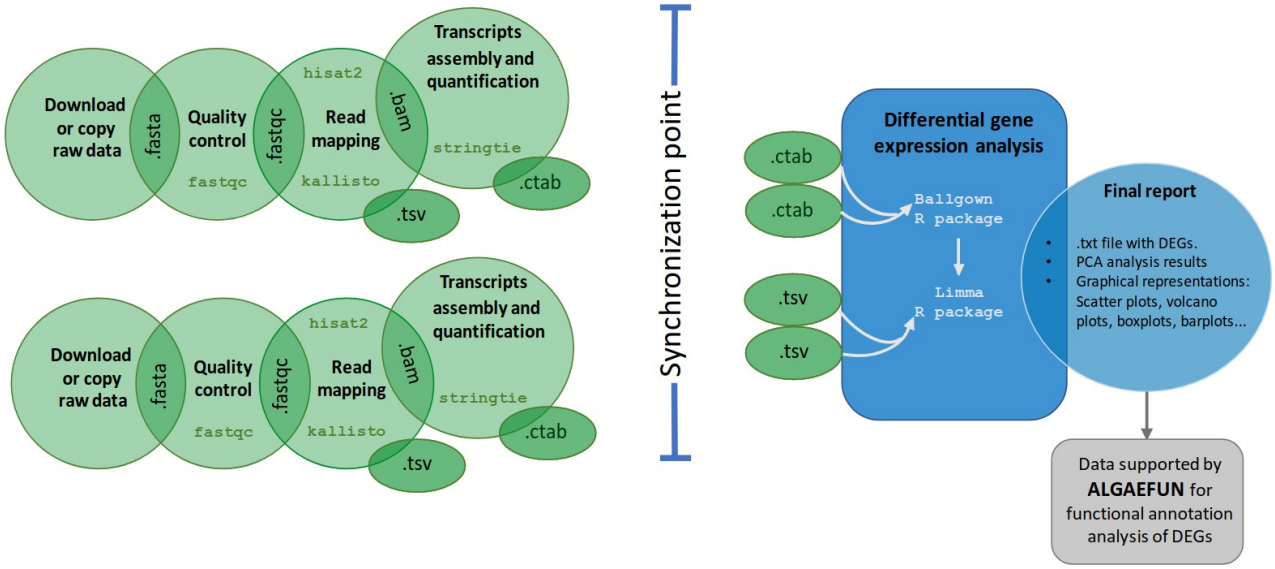

Figure 13: Workflow of the automatic pipeline for the analysis of RNA-seq data in MARACAS. The maracas-rna-seq pipeline receives as input a parameter file as described in Table 4 and Table 5. After data acquisition in fastq format, sequence quality analysis is performed using fastqc. Read mapping and gene expression quantification is then performed using HISAT2 and stringtie or kallisto depending on the user choice. 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) or TPM (Transcripts per Million) are stored in TSV 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, box-plots, scatter-plots, volcano-plots and bar-plots of individual genes 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.

The second fork of our pipeline dedicated to ChIP-seq analysis (Fig. 14) shares with the above RNA-seq fork the first steps consisting of downloading or copying the fastq raw files for the ChIP samples (data corresponding to the chromatin immunoprecipitation condition) and control samples into the generated working directory and performing a quality control. Specifically, in this part of our workflow short read mapping to the reference genome is executed using bowtie2 (Langmead & Salzberg, 2012) and the synchronization point requires that BAM files for every ChIP and control sample is generated. This synchronization point is also achieved using a blackboard. After synchronization, the last step takes as input all mappings in BAM format and performs peak calling using the software tool macs2

(Gaspar, 2018). This final step consists of the identification of genomic loci or regions significantly enriched with mapped reads indicating the genomic occupation of the transcription factor or histone modification of interest. The output files of this pipeline consists of a BED (Browser Extensible Data) file collecting the genomic loci or regions identified during the analysis and a BW (BigWIG) file containing the number of mapped reads or signal in each position of the genome.

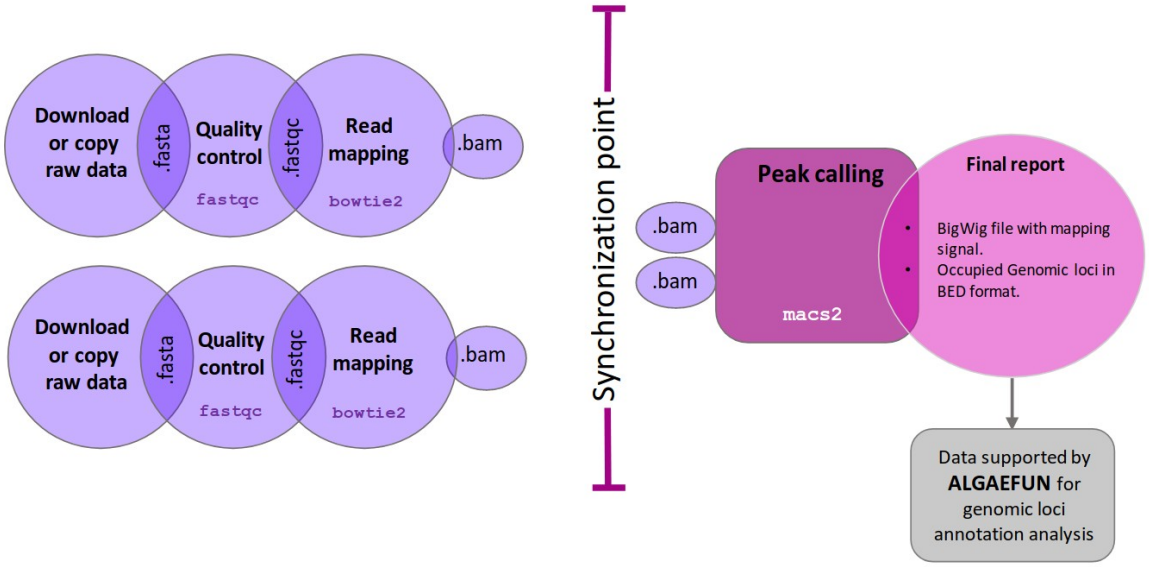

Figure 14: Workflow of the automatic pipeline for the analysis of ChIP-seq data in MARACAS. The maracas-chip-seq pipeline receives as input a parameter file as described in Table 4 and Table 5. After data acquisition in fastq format, sequence quality analysis is performed using fastqc. Read mapping to the reference genome is performed using bowtie2 and is stored in BAM format. A synchronization point ensuring the end 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.

ALGAEFUN implementation: functional annotation analysis.

The generated output from MARACAS, either sets of genes or genomic loci, can be subsequently functionally annotated using the next software tool in our portal, ALGAEFUN. Although this tool can also be used to functionally annotate genes sets or genomic loci generated independently from other tools.

The user interface of ALGAEFUN is shown in Fig. 15 and Fig. 16. Two different modes of operation depending on the kind of input data to be analyzed are implemented in ALGAEFUN.

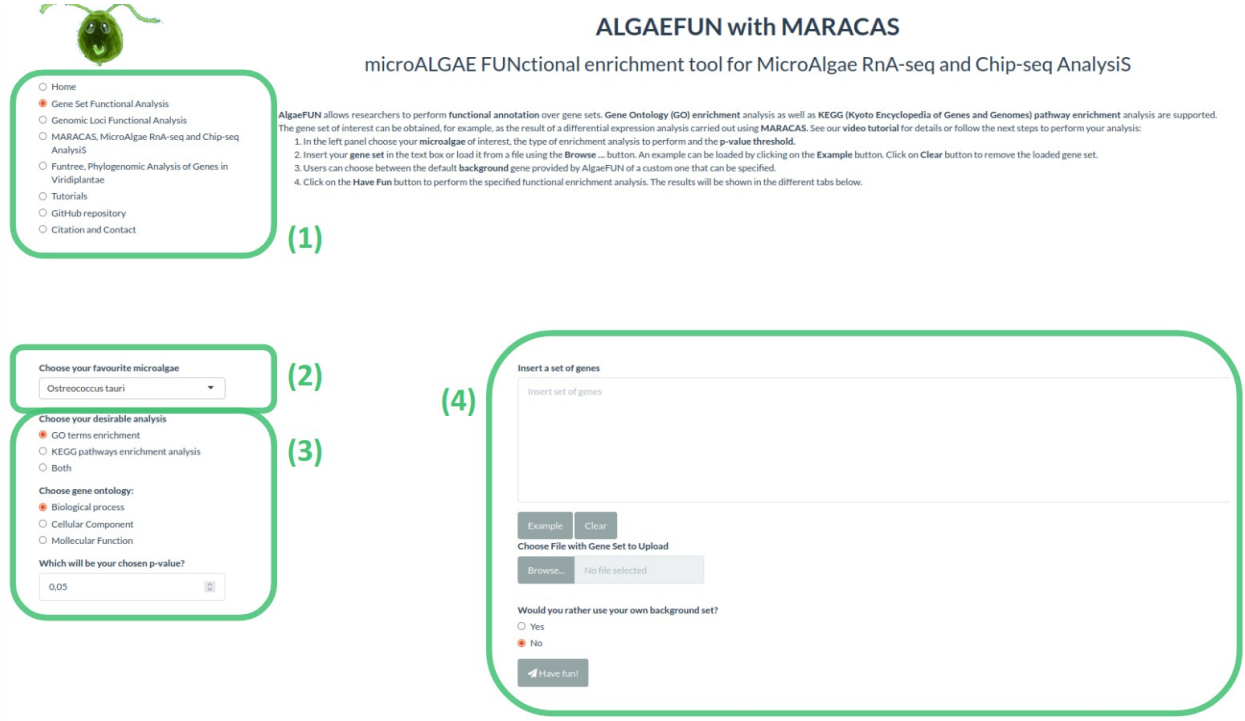

Figure 15: Microalgae Functional Annotation tool (ALGAEFUN) user interface for gene sets functional annotation. (1) Navigation bar for selecting the tool to use between the ones included in our web app; (2) Drop-down menu to select the microalgae of interest; (3) Sidebar panel to select parameters for GO and/or KEGG pathway enrichment; (4) Main panel to input set of genes to analyze and select gene background.

In one of the modes, input data consists of sets of genes obtained, for instance, from an RNA-seq study. Here after selecting this analysis mode (Fig. 15-1) the user chooses the microalgae of interest (Fig. 15-2) and whether he/she wants to carry out a GO term and/or KEGG pathway enrichment analysis at a selected significance level (Fig. 15-3). The set of genes to study can be inputted through a text box in our tool or uploading a file. Users can also choose whether to use or not his/her own background gene set or the entire microal-

gae genome (Fig. 15-4). In order to allow users to explore the functionalities of our tool and also to check the required gene id format we have included an example gene set for each microalgae. This example can be accessed and inputted in the corresponding text box by clicking on the example button (Fig. 15–4). These examples were generated during the testing of MARACAS using previously published RNA-seq data sets and analysis (los tengo q citar todos? No se ya ni cuales son algunos) and have in turn been used in the testing and validation of ALGAEFUN. The GO term and KEGG pathway enrichment analysis are carried out using the Bioconductor R package ClusterProfiler (Wu et al., 2021). This package implements statistical analysis and visualization of functional profiles for gene clusters or sets using the annotation packages developed in our tool for each microalgae integrating the systematic sources of functional annotation previously discussed. The outputs for the GO term and KEGG pathway enrichment analysis are presented in two separate tabs. The first output in the GO enrichment tab consists of a table that summarizes the results from the GO enrichment analysis carried out over the input gene set. The user can find one row for each GO term and 6 columns that represent some relevant information about the enrichment. The first column shows the GO term identifier, followed by the second column were the user can find a human readable description. Users can access more information about the GO term represented in a specific row by clicking on its identifier to be redirected to the web portal AmiGO (Carbon et al., 2009) where GO terms are described in detail. The third and fourth column represents the p-value and q-value, and the fifth one shows the enrichment value: E= (m/n) / (M/N). The value “m” is the number of genes from the inputted gene set annotated with the corresponding GO term whereas “M” is the number of genes from the background annotated with the mentioned GO term. In a similar way, “n” is the number of genes with annotation from the gene set whereas “N” is the number of genes with annotation from the gene background. Finally, the last column shows the genes from the input set of genes annotated with each GO term. The user can click them to get more information from the gene entry on the corresponding database from which annotation was retrieved for the specific microalgae under study. Furthermore, ALGAEFUN also generates several graphs that represent the GO term enrichment. Five visualization methods can help users understand the results:

- Acyclic graph: Each node stands for a GO term and the color of them indicates the level of significance (from grey, non-significant, to intense red, highly significant). An

arrow is drawn from GO term A to GO term B when A is a more general GO term than B or B is more specific than A.

- Bar-plot: Each bar represents an enriched GO term whose length corresponds to the number of genes in the gene set annotated with the given GO term. Once again, the bar color shows the level of significance (from blue, less significant, to red, more significant).
- Dot-plot: Each dot represents an enriched GO term. The x-position of the dot corresponds to the ratio between the number of genes annotated with the corresponding GO term and the total number of annotated genes in the gene set. The dot color captures the level of significance (from blue, less significant, to red, more significant).
- Enrichment Map (emap-plot): Each node represents an enriched GO term and the size of each node is proportional to the number of genes annotated with the corresponding GO term in the inputted gene set. The node colors represent the level of significance (from less significant in blue to more significant in red). These nodes are connected by edges when the corresponding GO terms are semantically related.
- Gene-concept network (cnet-plot): The beige nodes represents GO terms and the grey nodes genes. An edge is drawn from a gene to a GO term when the gene is annotated with the corresponding GO term. The size of nodes representing GO terms is proportional to the number of genes annotated with the corresponding GO term.

The outputs shown on the KEGG pathway enrichment tab consist of:

- Table summarizing the result of the KEGG pathway enrichment analysis: Each row represents a pathway significantly enriched in the inputted gene set with respect to the selected gene background. The first column represents the KEGG pathway identifier and the user can click on it to read more information about the pathway. The second column contains a human readable description. The third and fourth column present the p-value and q-value, and the fifth column displays the corresponding enrichment value E (m/n; M/N) as previously described. Finally, the last

column shows the list of genes from the inputted gene set assigned to the corresponding enriched pathway. KEGG pathways can be more informative than GO term since they are not general but specific to the corresponding organism of interest.

- KEGG pathway map: Users can choose a specific enriched pathway using a drop-down menu to generate the corresponding KEGG pathway map where genes from the inputted gene set are highlighted in red.
- Table summarizing the result of the KEGG module enrichment analysis: Each row represents a module significantly enriched in the gene set with respect to the selected gene universe. The columns in this table are organized in the same manner as the previously described. KEGG modules are distinct recurrent components of KEGG pathways in this respect they are more specific and can be more informative.

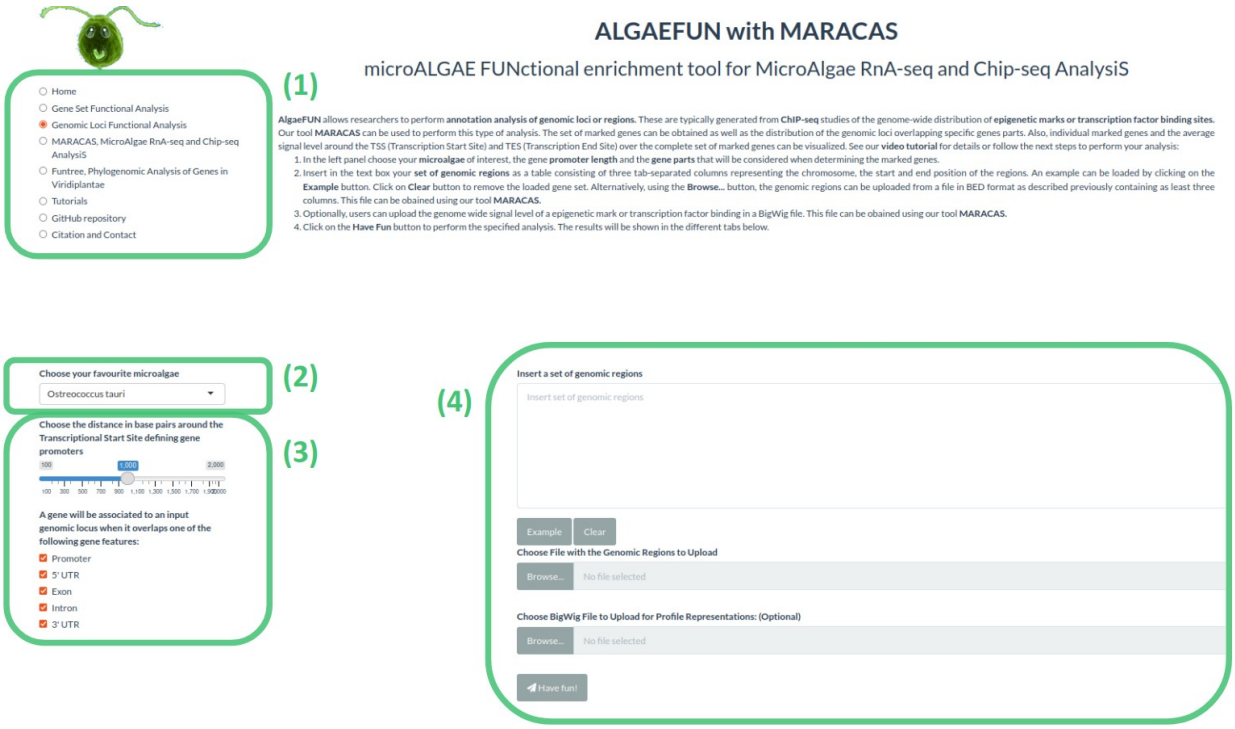

Figure 16: Microalgae Functional Annotation tool (ALGAEFUN) user interface for genomic loci annotation. (1) Navigation bar for selecting the tool to use; (2) Drop-down menu to select the microalgae of interest; (3) Sidebar panel to select parameters for the identification of the promoter regions and the gene features or parts that will be considered to assign genes to genomic loci; (4) Main panel to input set of genomic loci and signal in bigWig format obtained from a ChIP-seq analysis.

In the other mode, input data consists of genomic loci or regions obtained, for instance, from an Chip-seq study. This analysis mode is selected from the side bar panel in Fig. 16-1. Users can select their microalgae of interest using the drop-down menu from Fig. 16-2. Next, the distance around the transcriptional start site (TSS) that will be considered the promoter of each gene users must be specified. Users also need to select the gene features or parts that will be considered when assigning gene targets to genomic loci or regions (Fig. 16-3). Genomic loci or regions to analyze can be inputted through a text box in our tool or uploading a file. Additionally, a BW file containing the number of mapped reads or signal in each position of the genome can be uploaded (Fig. 16-4). Similar to the previous mode, users can explore the functionalities of our tool and check the required genomic loci or regions format using an example included for each microalgae by clicking on the example button (Fig. 16-4). These examples were generated during the testing of MARACAS using previously published ChIP-seq data sets and analysis and have in turn been used in the testing and validation of ALGAEFUN. The functional annotation of the inputted genomic loci or regions is performed using the Bioconductor R packages ChIPseeker (Yu et al., 2015) and ChIPpeakAnno (Zhu, 2013). These package implements statistical analysis and visualization of genomic loci and regions using the gene feature annotation packages generated in our tool for the microalgae previously mentioned. The outputs generated in this type of analysis consist of:

- A pie chart representing the distribution of the genomic loci or regions over the different type of gene features selected by the user such as the promoter, 3’ UTR, 5’UTR, intron or exon.
- A table enumerating the target genes associated one generated when the data comes from a RNA-seq study. It represents each gene located in the enriched genomic loci and its different annotation terms. This set of genes can be downloaded and subsequently annotated functionally using ALGAEFUN.
- A visualization of the average level of signal around TSS and transcriptional end site (TES). For each individual gene, it generates a visualization of the signal and identification of DNA motifs recognized by transcription factors and regulators in photosynthetic organisms.

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

This case of study is based on our own RNA-seq data generated to test our tools and illustrate the generation of relevant information suitable for publication in prestigious journals. It consists in a RNA-seq study carried out using *Haematococcus lacustris* which has a key role in the bio-production of astaxanthin (Hoys et al., 2021; Serrano-Pérez et al., 2022). The analysis has been performed for vegetative *Haematococcus lacustris* cells, grown both under N sufficiency and under moderate N limitation in order to unveil the transcriptomic program enhancing astaxanthin biosynthesis under N deprivation. The results obtained in this study will be used as a case of study to illustrate the type of information that ALGAEFUN with MARACAS is able to reveal from raw sequencing data.

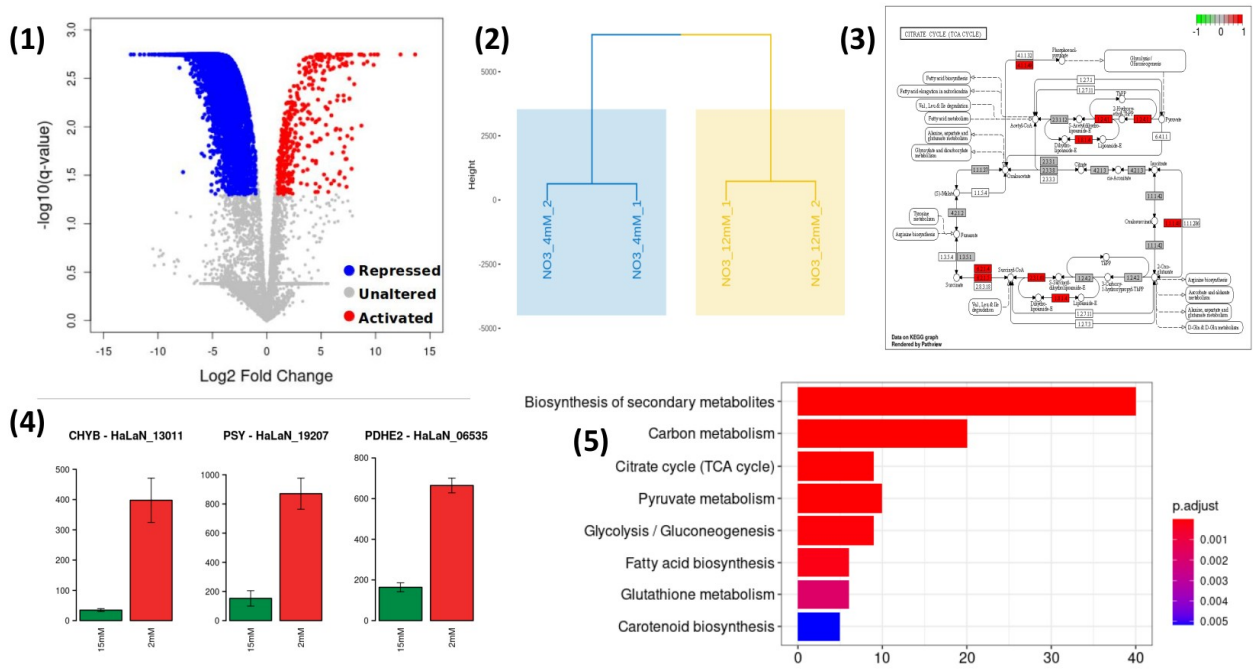

Figure 17: Results obtained from MARACAS final report (1,2,4) and from ALGAEFUN analysis (3,5). (1) Volcano-plot generated by MARACAS showing the global effect over the transcriptome as well as the activated and repressed genes detected (in red and blue color, respectively); (2) Hierarchical clustering combined with PCA generated by MARACAS; (3) KEGG representation of one of the enriched pathways generated by ALGAEFUN, genes present in the list used as input are colored in red color; (4) Bar-plots representing individual gene expression levels of key enzymes, figure generated by MARACAS. Expression level of the selected gene under N sufficiency is represented in green color. Red color is used to represent moderate N deprivation conditions; (5) Vertical bar-plot representing the enriched biological process in the set of genes used as input, figure generated by ALGAEFUN. Gradient from red to blue color is used to represent p-value.

The high-throughput sequencing raw data in fastq format were processed using MARACAS in order to obtain the set of differentially expressed genes using a criterion solely based on a fold-change value of 1.5. The report produced by MARACAS described all samples as of high quality and notified no problem during read mapping to reference genome with mapping rates greater than 86%. Scatter plots comparing gene expression between samples are also produced in the MARACAS report. High Pearson correlations greater than 98% were identified between replicates of the same condition. Accordingly, automatically hierarchical clustering combined with Principal Components Analysis PCA that leads to the identification of more stable clusters, due to the noise reduction achieved with PCA is performed by MARACAS (Fig. 17-2). It identified two clearly separated clusters constituted by the control (N sufficiency) and experimental condition samples (moderate N deprivation). Volcano plots comparing moderate N limitation with N sufficiency transcriptomes are used in the report to represent the repressed genes and activated genes identified with a fold-change threshold of two and a q-value threshold of 0.05. Moderate N limitation has shown a strong repressing effect over the transcriptome with respect to N sufficient conditions. In particular, we identified 414 activated and 5348 repressed genes (Fig. 17-1). These lists of genes can be then inputted into ALGAEFUN to determine significantly over-represented biological processes or pathways affected during experimental conditions.

As described previously, ALGAEFUN can perform two different types of functional enrichment analysis when a set of genes from a RNA-seq analysis is used as input: GO terms and KEGG pathways enrichment. The GO terms analysis can identify, for example, specific biological processes enriched in the set of genes. One of the available graphical representations of the GO enrichment results in ALGAEFUN are the bar-plot included in Figure 17-5. The bar-plot represents the biological processes enriched in the activated set of genes obtained from the *Haematococcus lacustris* RNA-seq study under moderate N limitation experiment. The experimental conditions seem to activate processes like biosynthesis of secondary metabolites, TCA cycle or pyruvate metabolism at a transcriptomic level. GO term analysis provide a general overview of the functional annotation of gene sets since they constitute universal functional terms not specific to a particular organisms. Complementary, KEGG pathway enrichment is specific to the corresponding organisms and can be more informative in some cases. It allows to identify the transcriptomic acti-

vated enzymes in enriched pathways (Fig. 17-3). The genes involved in these processes can be further studied individually and their expression level in both conditions can be compared (Fig. 17-4).

Our ALGAEFUN with MARACAS analysis described differential expression of hundreds of genes affecting key pathways converging into astaxanthin biosynthesis and storage. The affected pathways were further studied at metabolic level and a massive cell reprogramming was verified for the reddish cells (under moderate N limitation). Furthermore, a major advance has been made in discerning the underlying control mechanisms since our tool allowed us to find differentially activated enzymes in astaxanthin biosynthesis under N moderate deprivation. From this point, it was carried out the identification of the prevalent DNA sequences in the promoters of these key enzymes so common transcription factors, as bHLH, possibly regulating these key enzymes in astaxanthin biosynthesis were identified for the first time.

Case of study 2: From ChIP-seq raw sequencing data to marked genes.

As mentioned before, the development of ALGAEFUN with MARACAS has motivated our lab to generate transcriptomic as well as cistromic and epigenomic data. Our own ChIP-seq data has been already generated in our lab but it hasn't been published yet. Instead, in order to illustrate how ChIP-seq raw data can be analyzed using our tool, an already published epigenetics study in Chlamydomonas reinhardtii was re-analized.

Histone modifications play a central role in gene expression control. However, although they have been intensely studied in plants, they have been poorly characterized in microalgae. Nevertheless, in a recent publication the genome-wide distribution of the repressive mark H3K27me3 was determined in C. reinhardtii (Ngan et al., 2015). After Chip-seq raw data was re-analyzed using MARACAS, 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 provided also by MARACAS. 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 (Fig. 18-a) 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*.

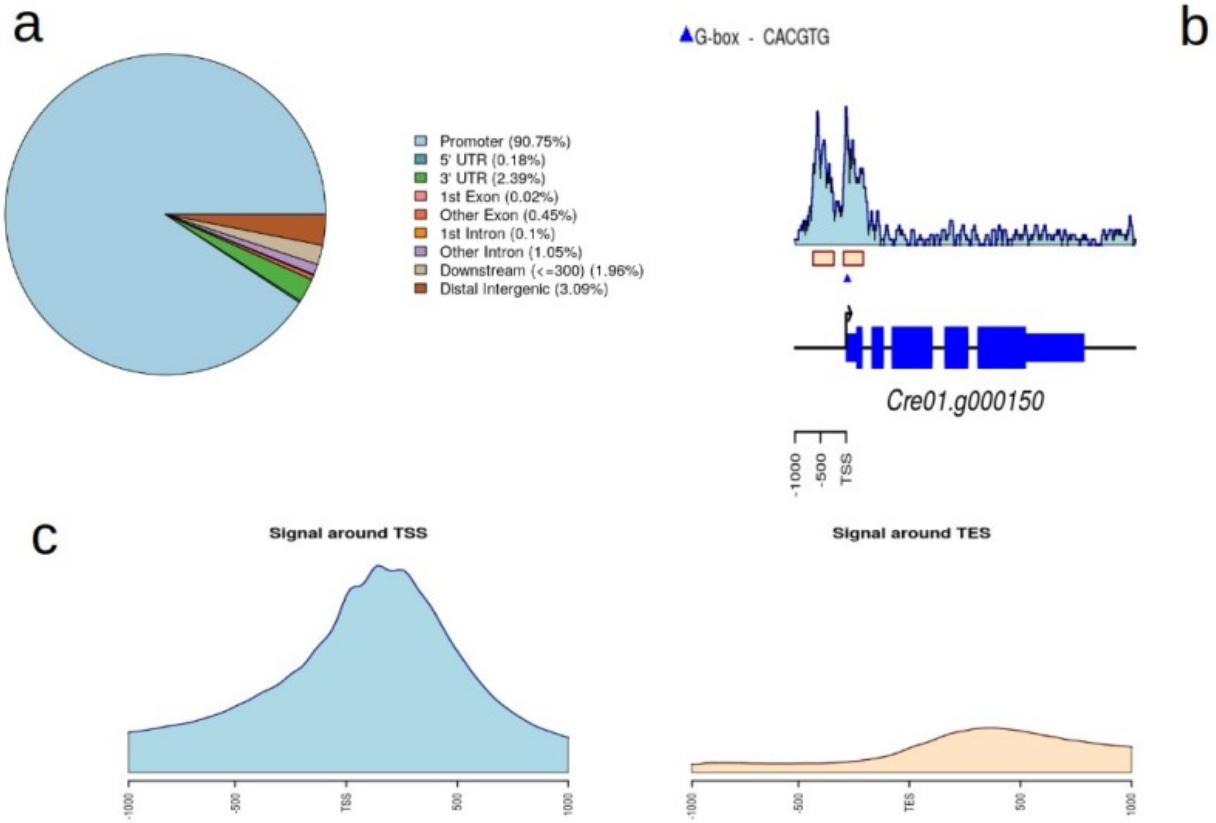

Figure 18: Summary of the outputs generated by ALGAEFUN when a genomic loci list is used as input. a) Pie chart representing the distribution of peaks or genomic loci over the different type of gene features such as the promoter, 3' UTR, 5' UTR, intron or exon; b) Visualization of the signal and identification of DNA motifs recognized by transcription factors and regulators in photosynthetic organisms for an individual gene; c) Visualization of the average level of signal around TSS and TES over the target genes.

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 Figure 18-b. 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 (Fig. 18-c). 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.

Contribution of ALGAEFUN with MARACAS to the field.

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. 8). For the model microalgae Chlamydomonas reinhardtii, researchers can find several online tools to functionally annotate set of genes, such as Algal Functional Annotation Tool (Lopez et al., 2011) and ChlamyNET (Romero-Campero et al., 2016). Only the online tool AgriGO (Tian et al., 2017) 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 χ^2 (Masseroli et al., 2004). 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 6.

Table 6: 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

ALGAEFUN with MARACAS is a constantly growing tool that will include new microalgae whenever their sequenced genomes are available. Also it has settle the bases to build numerous tools by other members of the laboratory that will for sure be a crucial toolbox for the microalgae research community doing systems biology studies. Definitely, ALGAEFUN with MARACAS has facilitated the progress of my doctoral work as well as motivated studies of molecular biology of systems in our laboratory with important contributions to the field. Since molecular systems biology has much to contribute to microalgae research, we hope that ALGAEFUN with MARACAS will reach the hands of many research groups and will be as useful to them as it has been to us.

Chapter 2: Transcriptional analysis of diurnal and seasonal cycles in *Ostreococcus tauri*

High-throughput transcriptome sequencing produced approximately 10 million short reads per sample (ANEXO 1). This allowed us to accurately estimate gene expression levels measured as FPKM (Fragments Per Kilobase of exon per Million reads mapped) in the transcriptomes corresponding to each data point of our time series. Indeed, out of the 7668 genes currently annotated in the Ostreococcus tauri genome (genome 2007 y 2014), only 3 genes were never expressed and 260 genes never exceeded an expression level of ten FPKM. This shows that practically the entire Ostreococcus tauri genome is expressed under seasonal and diurnal cycles. First, we focus in the 36 transcriptomes corresponding to the time points taken during three days under LD and SD conditions and perform a hierarchical clustering analysis. (meter definicion) The transcriptomes corresponding to the same time points during the three different days tend to cluster together (Fig. 19-1). This indicates a high circadian synchronization in our cultures. Moreover, these 36 transcriptomes assemble together into three different groups (Fig. 19-1). The first cluster corresponds to midday. The transcriptomes at time points ZT4 and ZT8 under LD and ZT4 under SD constitute this cluster. These time points correspond to the moments of maximal incident light irradiance under both LD and SD conditions. The second cluster conforms the dusk group. Here the transcriptomes at time points ZT12 and ZT16 under LD and ZT8 under SD are grouped. These time points coincide with the end of the light period in both LD and SD conditions when incident light irradiance is low. The third cluster represents night/dawn and comprises the transcriptomes at time points ZT20, ZT0 under LD and ZT12, ZT16, ZT20 and ZT0 under SD. The transcriptomes at time points in the LD and SD nights or dark periods constitute two distinct groups suggesting noticeable differences in the transcriptomic responses during the night under LD and SD conditions. It is also noteworthy the higher similarity between the dusk, night/dawn transcriptomes when compare to the midday one (Fig. 19-1).

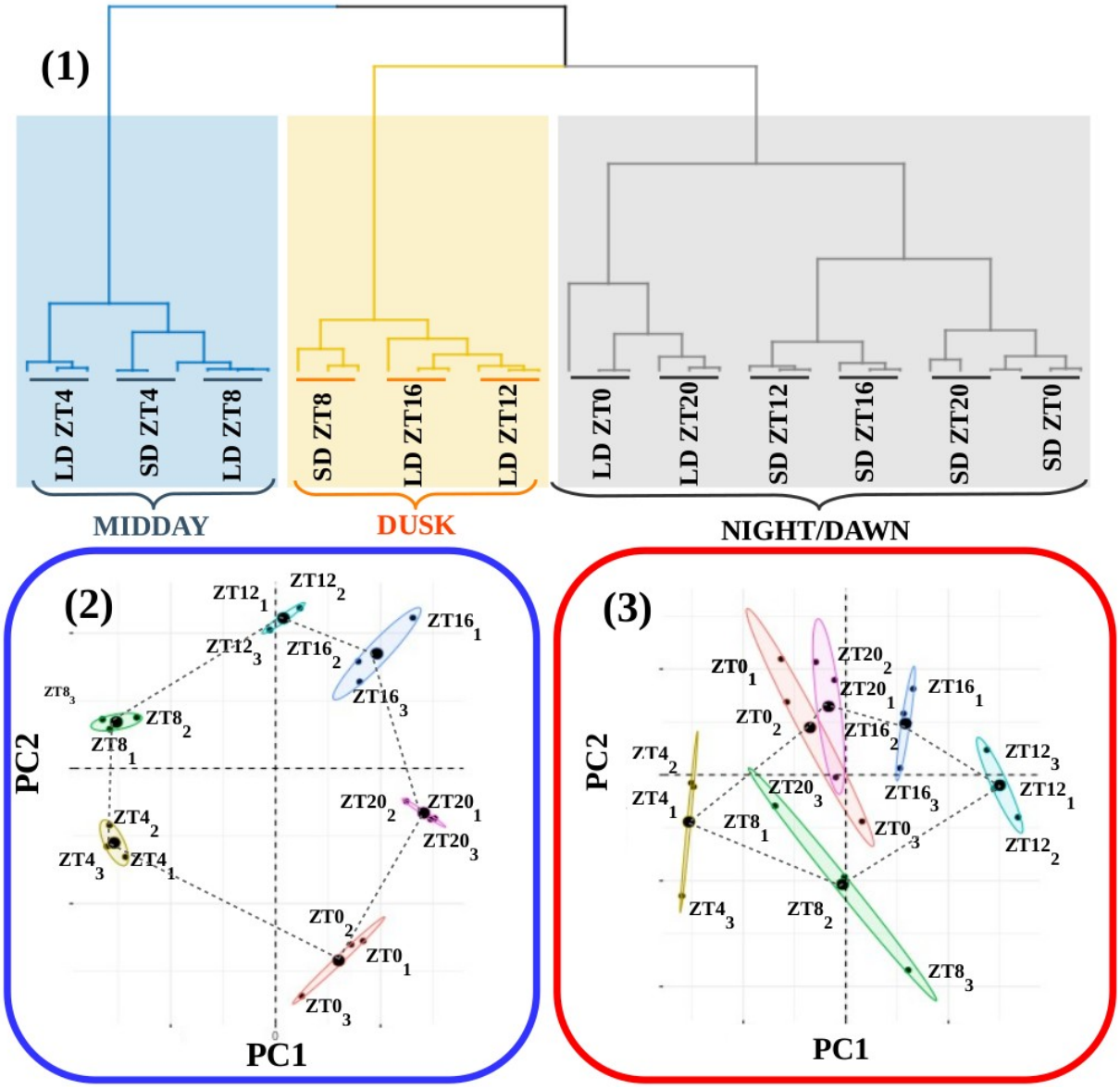

Figure 19: ***RNA reliability.*** *(A) Hierarchical clustering of the RNA-seq data corresponding to the 36 time points collected under alternating dark/light cycles simulating long and short day conditions. The three global transcriptomes corresponding to the same time points from different days cluster together showing robust rhythmicity in our cultures. Three distinct clusters are observed corresponding to midday (blue rectangle: LD ZT4, SD ZT4 and LD ZT8), dusk (yellow rectangle: SD ZT8, LD ZT16 and LD ZT12) and night/dawn (grey rectangle with two clear subclusters distinguishing between LD and SD: LD ZT0 and LD ZT20 on one hand and SD ZT12, SD ZT16, SD ZT20 and ZT0 on the other hand). (B) Principal Component Analysis of the time point global transcriptomes under long day conditions. Small dots correspond to the 2D projection of each time point global transcriptome. Big dots correspond to the average of the three replicates 2D projections for each time point. Ellipses mark the 95% confidence regions corresponding to each time point global transcriptome. (C) Principal Component Analysis of the time point global transcriptomes under short day conditions. Points and ellipses are used as described before.*

In order to obtain a deeper understanding of the underlying structure in our data we performed principal components analysis separately over the LD (Fig. 19-2) and SD (Fig. 19-3) transcriptomes. Under LD conditions, we observed that the transcriptomes corresponding to the same time point in the three different days tightly cluster together globally constituting a circular structure. Nonetheless, under SD conditions more variability is observed and the time point transcriptomes form a structure resembling an ellipse. This could indicate that whereas in LD conditions gene expression in globally cycling precisely with a similar period a more complex behavior is expected under SD conditions. Also, it is remarkable the high similarity between the transcriptomes corresponding to ZT0 and ZT20 under SD conditions that is not present under LD conditions. This suggests that the transcriptomic response at the end of a SD night is already preparing all molecular systems for the incoming light availability at dawn whereas this anticipation is not as clearly observed under LD conditions. In overall, these results support that our experimental design grants a high level of synchronization in our data allowing us to proceed to the identification and comparison of genes exhibiting rhythmic expression patterns under LD and SD conditions.

Transcriptomic characterization of diurnal rhythmic expression profiles

Most genes in Ostreococcus tauri present diurnal rhythmic expression profiles under both photoperiods

We used the bioconductor R package RAIN (Rhythmicity Analysis Incorporating Non-parametric Methods) (Thaben and Westermark, 2014), as described in Materials and methods, to identify genes exhibiting diurnal rhythmic expression patterns under the both seasonal conditions. Specifically, we used time series consisting of three days with rhythmic light / dark periods from our experiment. Independently from the photoperiod, more than 6000 genes comprising approximately 80% of the entire *Ostreococcus* genome present diurnal periodic rhythmic expression patterns. This result is in agreement with previous studies in *Ostreococcus tauri (CITA)* under a different photoperiods and other microalgae such as *Chlamydomonas reinhardtii* (Monnier et al., 2010; Zones et al. 2015). The specific rhythmic genes under each photoperiod are practically coincident (Fig. 20-A) (Supplemental Table 3—ANEXO?).

In order to further explore the remaining 20% of the genome of Ostreococcus tauri that didn't show rhythmic expression patterns under any photoperiod, we compared their highest expression values with the ones reached by rhythmic genes. We found that genes exhibiting rhythmic expression patterns under both LD and SD conditions present a maximal level of expression three times greater than genes detected as non rhythmic. This difference was significant according to a p-value of 1.45×10^{-4} computed using Mann-Whitney-Wilcoxson test (Fig. 20-B). The current methods for detecting rhythmic gene expression are known to perform optimally only for highly expressed genes (Laloum and Robison-Rechavi, 2020). Therefore, the genes identified as non rhythmic in this study could indeed be rhythmic although their low expression level have prevented our methods from detecting them.

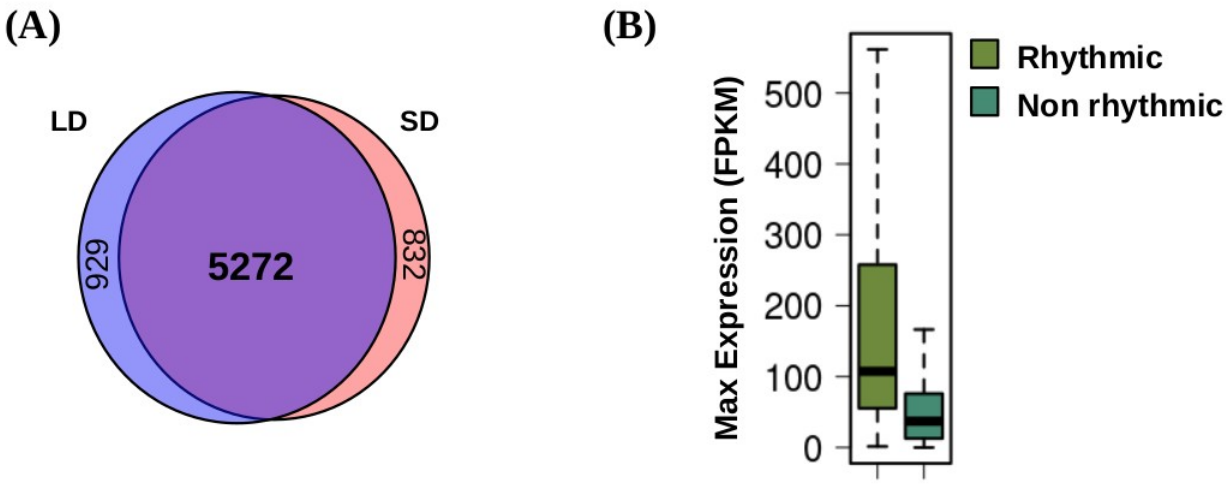

Figure 20: ***Diurnal rhythmic expression patterns detected in the transcriptome of Ostreococcus tauri under both photoperiods:*** *(A) Venn diagram comparing rhythmic genes under LD conditions (light blue circle) and SD conditions (light red circle). Genes with rhythmic expression patterns are almost identical under both photoperiodic entrainments. (B) Boxplot representing the maximum expression level of rhythmic genes (light green) and non-rhythmic genes (dark green). Differences observed are statistically probed (p-value of 1.45 × 10-4) using Mann-Whitney-Wilcoxon test. Gene expression levels are measured as FPKM (Fragments Per Kilobase of transcript per Million fragments mapped).*

Constant light and constant darkness as free-running conditions have different effects over the transcriptome of *Ostreococcus*

In order to explore the effect of the transition to continuous light on rhythmic gene expression patterns we compared three rhythmic characteristics of the expression profiles under LD or SD to the corresponding ones under LL and DD. This was carried out using the R package Circacompare. As it was mentioned in Materials and methods, Cicacompare performs a fitting to the data as co-sinusoidal curve with a particular parametrization that is used to test for statistical significant differences between the two circadian patterns under comparison.

When comparing our LD and LL data, it can be observed that most genes (97.68%) that maintain their rhythmicity under LL conditions with LD entrainment conditions presented a decrease in amplitude when transferred to LL (Fig. 21-A), being significant in more than half of them with a p-value lower than 0.05. Also a positive phase shift (increasing phase) is observed in 76.28% of the genes being significant in 14.98% of the genes exhibiting rhythmicity under constant light. However, when comparing LD with DD data the reduction in amplitud is less widespread (Fig. 21-A), being observed in 79.43% of the genes and being significant in 31.62% of the them with p-value lower than 0.05. Also, in contrast to LL, a negative phase shifts (decreasing phase) is observed in DD conditions. Specifically, 87.32% of the rhythmic genes under DD presented an anticipation in phase that was significant in 34.16% of them. Most of them present negative phase shifts between 0 and -5h of difference, whereas most of genes presenting positive phase shifts under LL conditions exhibit between 0 and 5h of difference (Fig. 21-B).

The global reduction in amplitude under LL and DD when compared to LD (Fig. 21-A) is significant with p-values 4.225738e-140 and 1.087318e-60 respectively. More precisely, the reduction in amplitude is significantly lower under LL than under DD with a p-value of 6.706823e-28.

Next we perform the analysis of the effect over the rhythmic genes under SD conditions when transferred to constant ligth. It can be observed that most genes (87.9%) that maintain their rhythmicity under LL conditions with SD entraintment presented a drastic decrease in amplitude when transferred to LL (Fig. 21-E), being significant in 34.6% of them with a p-value lower than 0.05. In agreement with LD entrainment, a positive phase shift around +5h of difference (Fig. 21-F,G) is also observed in SD entrainment samples under

LL (68.28% of the genes maintaining rhythmicity under SD and LL) being significant in 14.98% of them with a p-value lower that 0.05. However, when comparing SD with DD data, the reduction in amplitude is almost non-existent (Fig. 21-E), being observed only in 35.97% of the rhythmic genes (under SD and DD) and being significant in 23.58% of them with p-value lower than 0.05. Negative phase shifts are observed in 47.7% of the genes maintaining rhythmicity under SD and DD, being significant in 28.83% of them. However, most genes have a phase shit around 0 (Fig. 21-F), suggesting that phase shifts are not so drastic in DD after SD entrainment (Fig. 21-G).

There exists a drastic global reduction in amplitude under LL when compared to SD, which is significant with a p-value 5.951216e-65 (Fig. 21-D). In contrast to what was observed after LD entrainment, a slight but significant increase in amplitude under DD was detected when compared to SD with a p-value of 2.502664e-08 (Fig. 21-D). This suggest a highly similarity between SD and DD conditions, probably due to long periods of dark entrainment.

Free-running conditions have been widely studied in nocturnal mammals (cita paper ratones), birds(cita paper pajaros), plants (cita paper thaliana) and other organisms. The effect over the amplitude of biological rhythms has been previously observed in plants and other organisms at different biological levels and is commonly associated with a lost of synchrony (Layers of crosstalk between circadian regulation and environmental signalling in plants, Circadian Timing: From Genetics to Behavior) (Fig. 22). This suggests that LL conditions promote a larger lost of synchrony than DD conditions at the transcriptomic level in Ostreococcus under both photoperiods of entrainment.

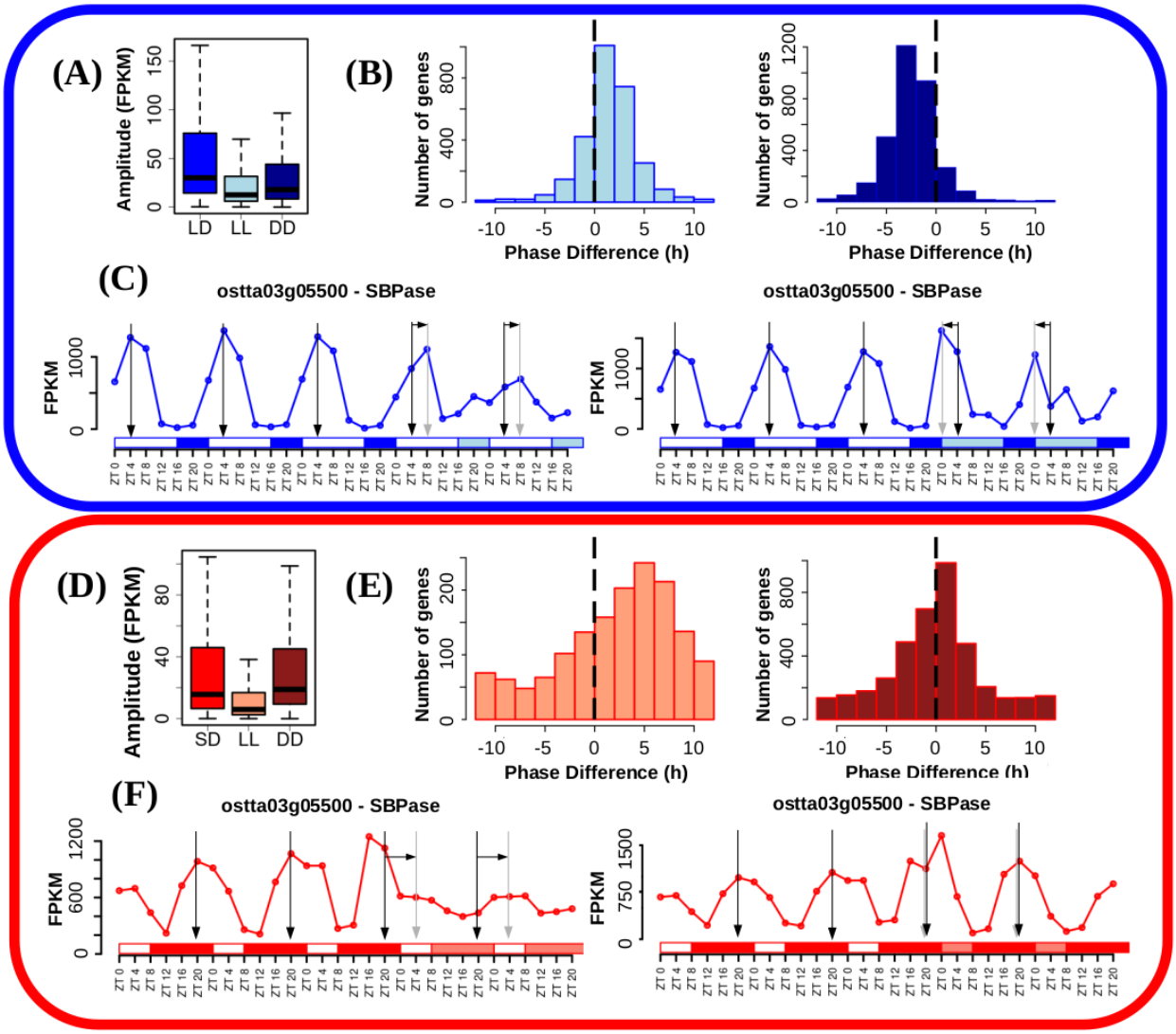

Figure 21: ***Free running conditions effects over gene expression profiles.*** *(A) Boxplot representing rhythmic genes amplitude reached under long day conditions (blue), when cultures were kept under free running conditions consisting of constant light (light blue) and when cultures were kept under free running conditions consisting of constant dark (dark blue). LL and DD amplitudes are significantly reduced with respect to LD according to p-values of 4.225738e-140 and 1.087318e-60 possibly due to a decline in culture synchrony under free running conditions. LL amplitudes are also reduced when compared to DD according to a p-value of 6.706823e-28 suggesting more severe loss of rhythmicity under LL than DD. P-values were computed using Mann-Whitney-Wilcoxon test. (B) Histograms showing the distribution of the number of genes exhibiting positive and negative phase shifts under LL (light blue, left) and DD (dark blue, right) free running conditions when compared to LD. Vertical dashed lines mark no shift. Positive forward phase shifts are observed when cultures are transferred from LD to LL whereas negative backward phase shifts are apparent when transferred to DD. (C) Gene expression profiles under LD, LL and DD of Sedoheptulose-bisphosphatase (ostta03g05500, SBPase). Vertical black arrows mark LD phases, vertical grey arrows mark LL and DD phases and horizontal black arrows represent phase shifts. SBPase illustrates how genes after LD entrainment present reduced amplitudes under LL and DD, forward phase shifts under LL and backward phase shift under DD being these changes more drastic under LL than DD. (D) Boxplot representing rhythmic genes amplitude under short day conditions (red), when cultures were kept under free running conditions consisting of constant light (light red) and when cultures*

were kept under free running conditions consisting of constant dark (dark red). LL amplitudes are significantly reduced with respect to LD according to p-values of 5.951216e-65 possibly due to a decline in culture synchrony under free running conditions. A slight but significant increase in amplitude under DD was detected when compared to SD with a p-value of 2.502664e-08. P-values were computed using Mann-Whitney-Wilcoxon test. (E) Histograms showing the distribution of the number of genes exhibiting positive and negative shifts in phase or maximum expression level time point under LL (light red, left) and DD (dark red, right) free running conditions when compared to SD. Vertical dashed lines mark no shift. Large positive forward phase shifts are observed when cultures are transferred from SD to LL whereas no substantial phase shifts are apparent when transferred to DD. (F) Gene expression profiles under LD, LL and DD of Sedoheptulose-bisphosphatase (ostta03g05500, SBPase). Vertical black arrows mark SD phases, vertical grey arrows mark LL and DD phases and horizontal black arrows represent phase shifts. SBPase illustrates how genes after SD entrainment present reduced amplitudes and forward phase shifts only under LL with no significant change under DD.

However, since plants growth is dependent on photosynthesis and thus in light, the effects of DD conditions over biological rhythms are yet poorly described. Our results bring new approaches about the effects of free running conditions in the green lineage, showing for the first time the strong desynchronizing effect of LL compared to DD conditions in a photosynthetic organism. Moreover, the effects of LL and DD observed over the transcriptome of Ostreococcus in this study agree with the ones described in mice and other nocturnal mammals. Constant light is commonly used as a circadian disruption model in those organisms (Circadian Behaviour in Neuroglobin Deficient Mice, Influence of short-term constant light on phase shift of mouse circadian locomotor activity rhythm induced by agonist and antagonist of serotonin, Desflurane anesthesia shifts the circadian rhythm phase depending on the time of day of anesthesia), which agrees with the strongly desynchronizing effect over the transcriptome of Ostreococcus. This suggest a strong dependence of photosynthetic organisms on dark periods.

Also positive and negative phase shifts are observed under LL and DD, respectively, in mice activity (Circadian Behaviour in Neuroglobin Deficient Mice) and algae-coral symbiosis content in photosynthetic pigments (Photosynthetic circadian rhythmicity patterns of Symbiodinium, [corrected] the coral endosymbiotic algae). In long term free-running conditions experiments, a slightly shorter period is described by biological rhythms under DD and a longer one under LL compared with the photoperiod of entrainment. This explains phase shifts observed in short-term experiments under free-running conditions (Circadian Timing: From Genetics to Behavior).

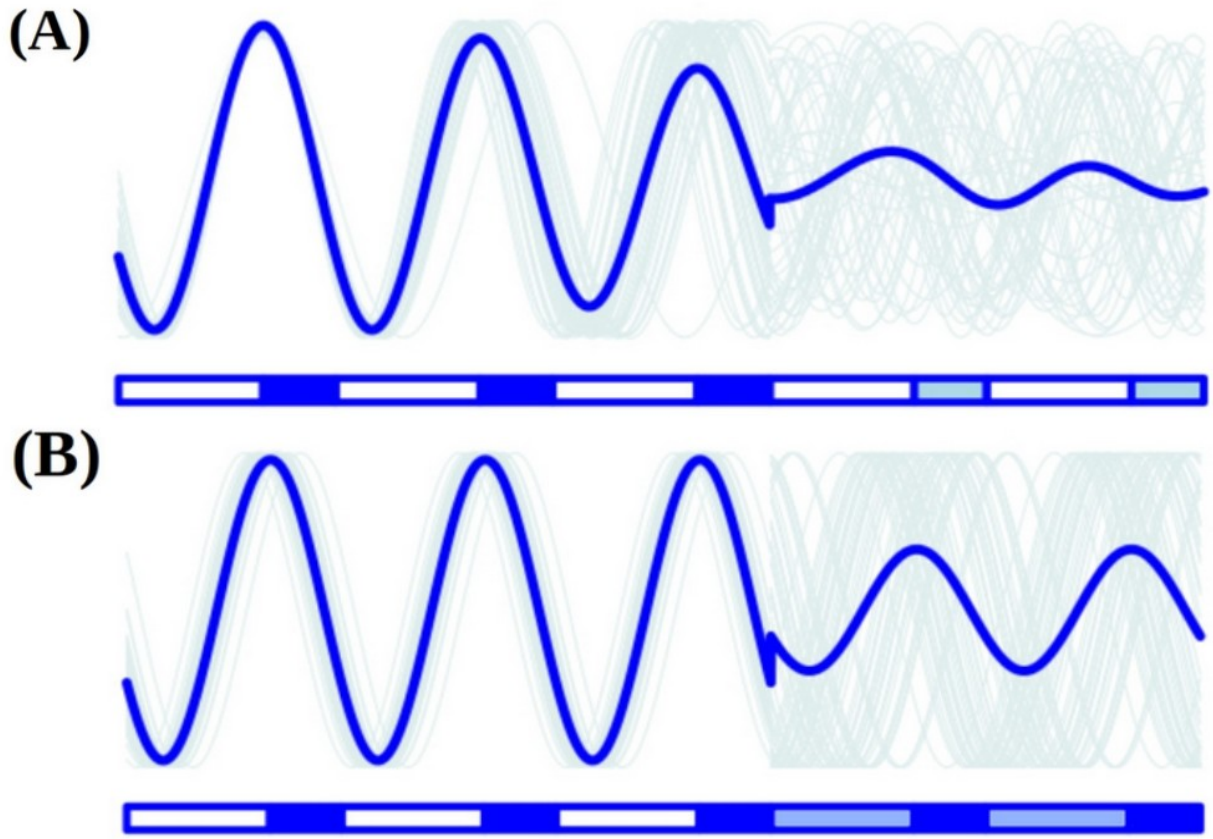

Figure 22: ***Reductions in amplitude under free running conditions can be explained by a decline in culture synchrony. (A)*** *Culture average gene expression profile under LD and constant light (LL) is represented by a thick blue line. Examples of individual cell gene expression profiles under LD and LL are represented by thin grey lines. When cultures are transferred to LL, cells get strongly desynchronized and their individual gene expression profiles become out of phase. This results in a strong reduction in the amplitude of the culture average gene expression profile although individual gene expression profiles maintain the same amplitude.* ***(B)*** *Culture average gene expression profile under LD and DD is represented by a thick blue line. Examples of individual cell gene expression profiles under LD and DD are represented by thin grey lines. When cultures are transferred to DD, cells get moderately desynchronized and their individual gene expression profiles become out of phase. This results in a slight reduction in the amplitude of the culture average gene expression profile although individual gene expression profiles maintain the same amplitude.*

Under free running conditions rhythmicity is maintained in different proportions depending on the photoperiod of entrainment

As it was described in the introduction, circadian processes are self-sustained and maintain their rhythmicity even when the given zeitgeber become constant. Following that definition, circadian genes can be detected and discerned from light or dark responding ones using our data generated under free running conditons. One of the main observations is that the maintenance of rhythmic expression profiles is dependent on the photoperiod of entrainment.

Although genes with rhythmic expression patterns under LD (6201 rhythmic genes) and SD conditions (6104 rhythmic genes) are almost coincident (Fig. 20-A), their rhythmicity is not maintained equally under free running conditions. Specifically, we found 2804 (36.57% of the entire genome) genes with a previous LD entrainment that maintain their rhythmicity under constant light (LL) conditions (Fig. 23-A). But the number of genes with a previous SD entrainment maintaining their rhythmicity under LL conditions decrease to 1526 (19.9% of the entire genome) (Fig. 23-C). However, 4006 genes with a previous SD entrainment and 3311 genes with a previous LD entrainment maintain their rhythmicity only under constant dark conditions (DD) (Fig. 23-A,C). It suggests that rhythmic expression under SD conditions is very dependent on the presence of a long period of dark whereas rhythmic expression under LD conditions due to the presence of a short period of dark is better maintained under continuous light. (cita si se puede)

Regulatory mechanisms are often composed of large networks influenced by a wide range of inputs. Circadian clocks are strongly influenced by external environmental signals but there exists a complex interplay between the clock and cell physiology as well (cita: The Circadian Clock, the Immune System, and Viral Infections: The Intricate Relationship Between Biological Time and Host-Virus Interaction /// Systems Level Understanding of Circadian Integration with Cell Physiology). Genes maintaining their rhythmic expression profiles only under LL or DD are only partially regulated by diurnal changes in the circadian clock. They are also influenced by other regulatory mechanisms as light (Fig. 23-B) or darkness (Fig.21-D).

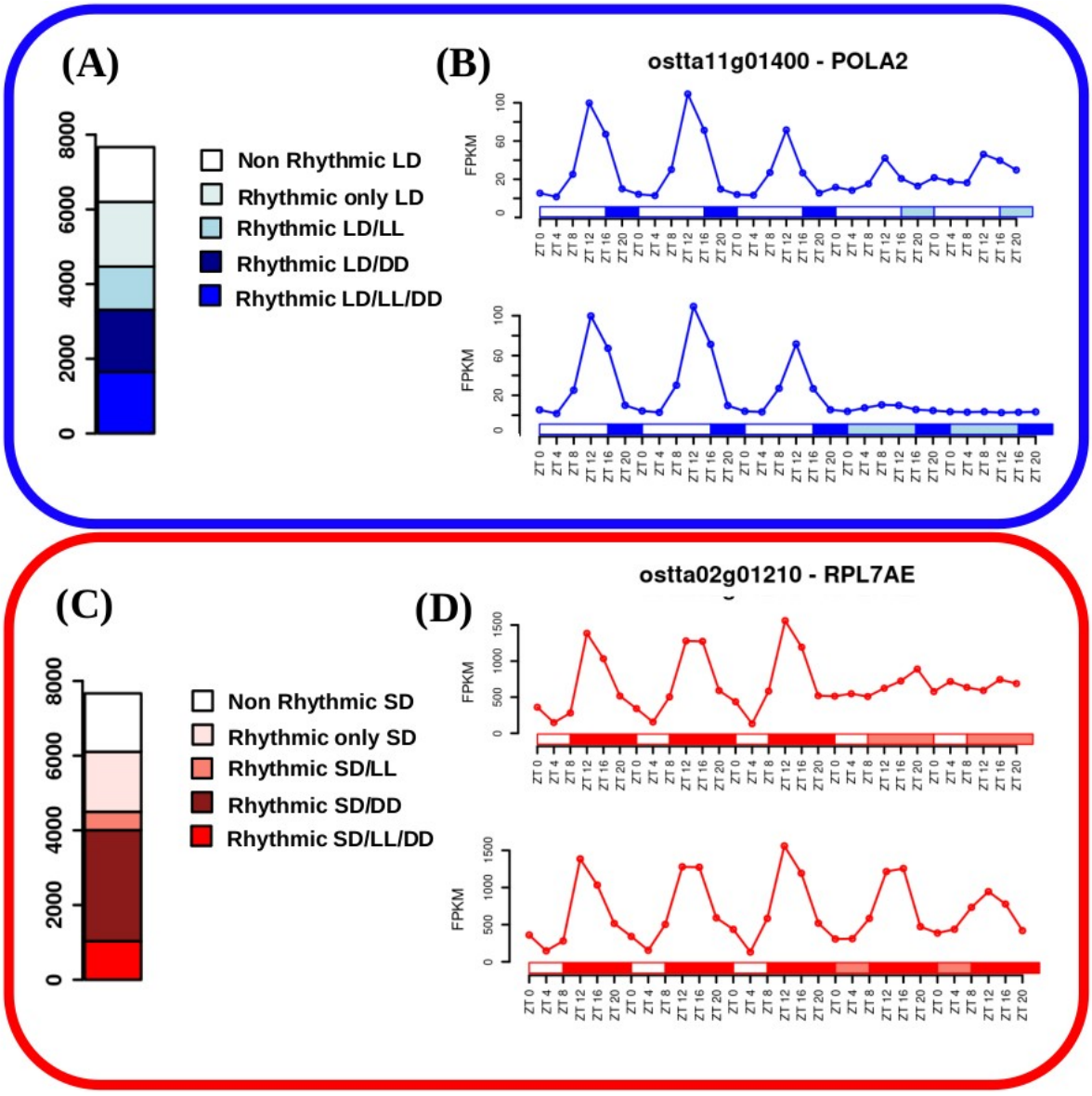

Figure 23: ***Rhythmicity maintenance under free-running conditions:*** *(C) Barplot representing with blue colors different rhythmic gene sets under LD conditions. From bottom to top: circadian genes exhibiting rhythmicity under LD, constant light (LL) and constant dark (DD); rhythmic genes under LD and DD requiring a dark input; rhythmic genes under LD and LL requiring a light input; rhythmic genes only under LD and non-rythmic genes. (D) Gene expression profiles under LD, LL and DD of DNA polymerase alpha subunit B (ostta11g01400, POLA2). POLA2 illustrates that specific genes involved in DNA replication require light maintaining rhythmicity under LL whereas being strongly repressed under DD. (E) Barplot representing with red colors different rhythmic gene sets under SD conditions. From bottom to top: circadian genes exhibiting rhythmicity under SD, LL and DD; rhythmic genes under SD and DD requiring a dark input; rhythmic genes under SD and LL requiring a light input; rhythmic genes only under SD and non-rhythmic genes. (F) Gene expression profiles under SD, LL and DD of Ribosomal protein L7Ae (ostta02g01210, RPL7AE). RPL7AE illustrates that specific genes involved in translation require dark maintaining rhythmicity under DD whereas presenting flat expression levels under LL.*

For example, one of the enriched processes in the set of genes that maintain their rhythmicity only under LL is the DNA replication (ANEXO). It agrees with previous cell cycle studies in other microalgae like *Euglena (Circadian Organization in the Algal Flagellate Euglena)* and *Chlamydomonas (Cell cycle control by timer and sizer in Chlamydomonas)* as well as *Ostreococcus (creo q los q tengo).* They suggest that cell cycle have a strong circadian clock regulation as well as G1 phase is light-dependent due to the need of light to grow in photosynthetic organisms, and it will explain why DNA replication genes need a light input to maintain rhythmicity.

Also, the main enriched processes in set of genes that maintain their rhythmicity only under DD are RNA processing and ribosome biogenesis (ANEXO). These processes, that are known have a complex regulatory mechanism influenced by the circadian clock between other regulatory machinery, are programmed at the transcriptomic level to take place during the night so translation of other proteins can be achieved during the day (cita Translation regulation in plants: an interesting past, an exciting present and a promising future). It is logical that a dark input is needed to maintain rhythmicity of those genes since their activation is dark-dependent.

In addition, comparison of circadian genes (that maintain their rhythmic expression profiles under both DD and LL) identified after LD entrainment and after SD entrainment allow us the identification of what is called bona fide circadian genes (Fig. 24-A). Bona fide circadian rhythms are the ones maintained under every light condition (both photoperiods of entrainment and both free-running conditions) and are mainly regulated by diurnal changes in the circadian clock. Our analysis identified only 350 genes (comprising 4.6% of the entire Ostreococcus genome) that didn't presented either a flat or noisy profile under any of the studied conditions. Gene expression profiles of RuBisCO activase (ostta04g02510, RCA) under the different conditions exemplifies how bona fide circadian genes maintain their rhythmicity (Fig. 24-B).

A functional enrichment analysis over this set of genes found that they were mainly involved in biological processes like photosynthesis, chlorophyll metabolic/biosynthetic process and chloroplast organization among others (Fig. 24-C). Some of those processes were known to present a circadian physiological activity in plants and microalgae like *Euglena,* but there is a lack of confirmation of this observation at the transcriptomic level in most cases (cita: RHYTHMIC PROCESSES IN PLANTS, Circadian Regulation of the Plant

Transcriptome Under Natural Conditions, Circadian Organization in the Algal Flagellate *Euglena*)(sabeis de algun paper otro paper que pueda discutir esto?). Understanding the key evolutionary position of Ostreococcus tauri in the green lineage, this results suggests that those are the processes purely transcriptionally regulated by circadian clocks since early in the evolution of the green lineage.

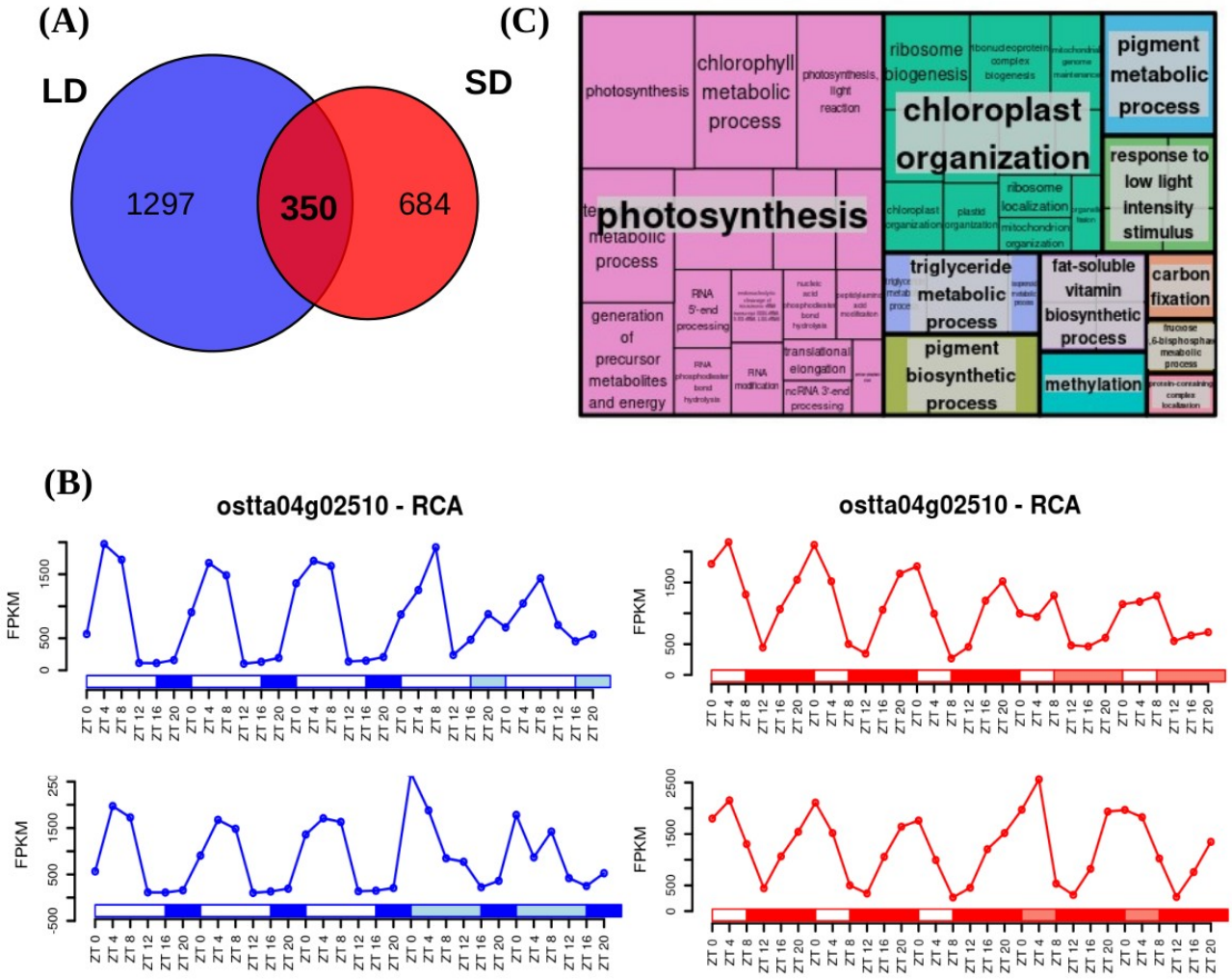

Figura 24: ***Identification of bona fide circadian genes and their functional enrichment analysis****: (A) Venn diagram comparing circadian genes identified after LD entrainment (blue circle) and after SD entrainment (red circle). Only a reduced number of genes are identified as bona fide circadian maintaining rhythmicity under the free running conditions LL and DD after both LD and SD entrainments. B) Gene expression profiles under LD, SD, LL and DD of RuBisCO Activase (ostta04g02510, RCA) exemplifying bona fide circadian genes that maintain rhythmicity under light/dark cycles as well as free running conditions. (C) Treemap summarizing the biological processes significantly enriched over the bona fide circadian genes or rhythmic genes under LD, SD, LL and DD. Rectangle sizes represent significance levels. Semantically similar biological processes are grouped together into the same colored rectangles. The most representative biological process is shown for each rectangle.*

Transcriptomic characterization of seasonal effects over gene expression profiles

Seasonal changes induce changes in amplitude and phase over gene expression profiles.

R package circacompare is used to study the global effect of the photoperiod over the amplitude and phase of rhythmic expression profiles. Comparing both photoperiods, we found that SD amplitudes are significantly reduced with respect to LD according to a p-value of $1.16e10^{-111}$ computed using Mann-Whitney-Wilcoxon (Fig. 25-A). Specifically, 2036 rhythmic genes corresponding to 46.12% exhibit a significant decrease in the amplitude of their rhythms. Whereas, only 123 corresponding to 2.33% significantly increase their amplitude in short day when compared to long day conditions. This suggests a possibly decline in culture synchrony under SD, accordingly to the previous similar synchronization observed between DD and SD conditions (Fig.21-D).

The most evident seasonal effect in the rhythmic gene expression profiles observed consists of a negative phase shift or phase anticipation. Specifically, 3424 genes comprising 64.95% of the rhythmic genes exhibited a significantly anticipated phase under short day conditions when compared to long day conditions. Phase anticipation are apparent under SD when gene phases are mostly reached around SD midnight (ZT12 to ZT16) whereas, under LD, phases are uniformly distributed from LD dusk (ZT12) to the end of the night (Fig. 25-B). Only a low number of rhythmic genes (around 200) exhibit their phase of maximum level of expression during the light period in LD.

Cyclin B (ostta01g06150, CYCB) and Delta-9 acyl-lipid desaturase 1 (ostta01g00790, ADS1) are two examples of genes exhibiting the typical phase anticipation and amplitude reduction under short day conditions when compared to long day conditions (Fig. 25-B).

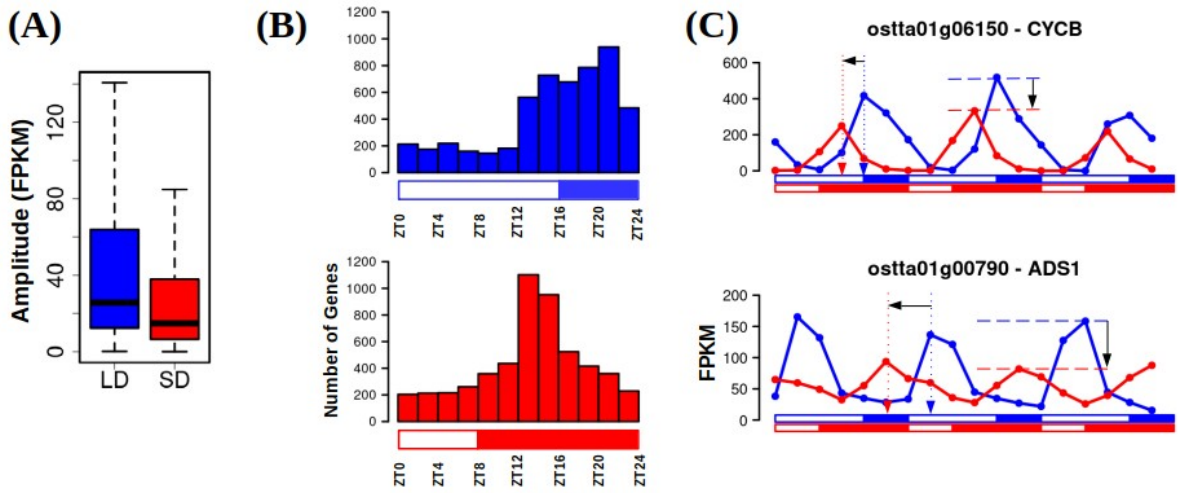

Figure 25: ***Photoperiod changes cause amplitude reductions and phase shifts over gene expression profiles.*** *(A) Boxplot representing rhythmic genes amplitude or maximum expression level reached under LD and under SD. SD amplitudes are significantly reduced with respect to LD according to a p-value of* 1.16×10^{-111} *computed using Mann-Whitney-Wilcoxon test possibly due to a decline in culture synchrony under SD. (B) Histograms showing the distribution of the number of genes with phase or maximum expression level at specific time points during the day under LD conditions (blue, top) and SD conditions (red, bottom). Backward phase shifts are apparent under SD when gene phases are mostly reached around SD midnight ZT12 to ZT16 whereas, under LD, phases are uniformly distributed from dusk (ZT12) to the end of the night. (C) Gene expression profiles under LD (blue line) and SD (red line) of Cyclin B (ostta01g06150, CYCB, top) and Delta-9 acyl-lipid desaturase 1 (ostta01g00790, ADS1, bottom). Blue and red vertical dotted arrows mark LD and SD phases. Horizontal black arrow represents backward phase shifts under SD when compared to LD. Blue and red horizontal dashed lines mark LD and SD amplitudes. Vertical black arrows represent the reductions in amplitude under SD with respect to LD.*

Seasonal changes induce complex rhythmic expression profiles

When comparing SD and LD conditions, another significant phenomenon affects the transcriptome of *Ostreococcus.* Under LD conditions almost every rhythmic gene (5825 genes covering 75.97% of the entire genome) reaches its maximum level of expression once a day, presenting one single peak each 24h in their expression profile (Fig. 26-D, E). Under SD conditions the number of genes presenting one single peak of expression per day decreased to 4249 (55.41% of the entire genome). Also, an increasing number of genes, namely 1855 genes, presented a more complex rhythmic expression profile with two peaks of expression per day (Fig. 26-D, E). That means, there is an increasing number of genes that reach their maximum level expression twice a day (every 12h) as the photoperiod get shorter.

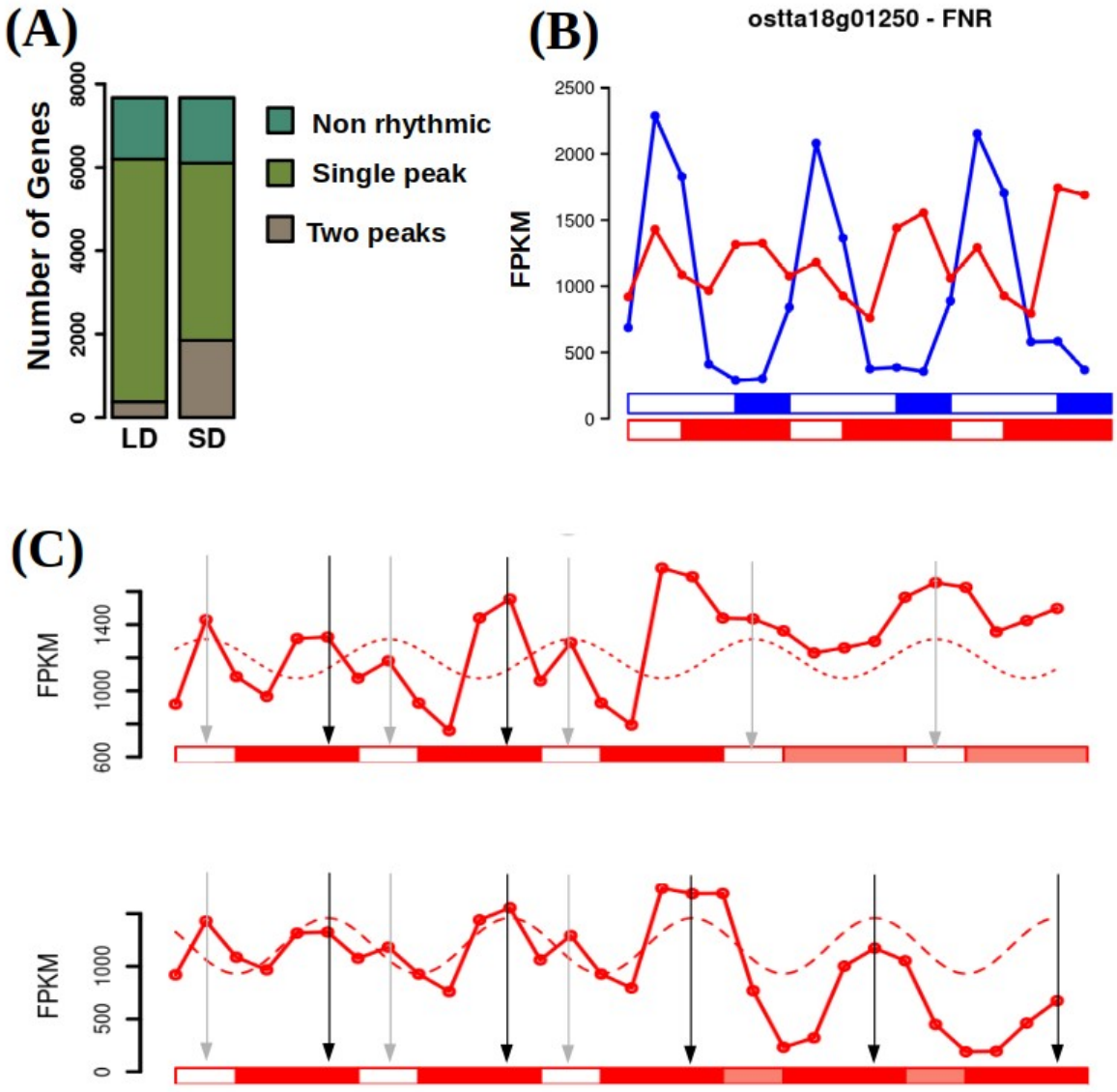

Figure 26: ***Emergence of complex expression profiles under SD.*** *(A) Barplot representing in different green colors the number of non rhythmic, single peak rhythmic and two peaks rhythmic genes under LD and SD conditions. An increase in the number of two peaks rhythmic genes is observed as the photoperiod shortens. (E) Gene expression profiles under LD (blue line) and SD (red line) of Ferredoxin-NADP+ reductase (ostta18g01250, FNR). This gene illustrates how specific single peak rhythmic expression patterns under LD conditions become two peaks rhythmic expression profiles under SD conditions. (F) Gene expression profiles under SD and free running conditions consisting of constant light (LL) or constant dark (DD) of Ferredoxin-NADP+ reductase (ostta18g01250, FNR). This gene exemplifies how two peaks expression patterns under SD conditions could emerge as the combination of two distinct rhythmic profiles. One depending on the photoperiod (dotted line) with phase marked with a grey vertical arrow maintaining its rhythmicity only under LL (top). Another expression profile is apparent depending on the skotoperiod (dashed line) with phase marked with a black vertical arrow maintaining its rhythmicity only under DD (bottom).*

This phenomenon can be found also in an already published microarray set of data generated from cultures of Ostreococcus tauri under 12 h of light- 12 h of dark cycle. (CITA) Most of the rhythmic genes, presented a simple rhythmic expression pattern with a 24 hours period and a single expression peak. However, 1171 genes presented a complex rhythmic expression pattern presenting two expression peaks with an apparent period of 12 hours. This photoperiod can be considered an intermediate step between the two extreme photoperiods studied (LD and SD) and, indeed, the data show an intermediate number of genes with two peaks of expression per day. Also in agreement with our results, this effect can be found in data generated from other organisms like Chlamydomonas under neutral day (CITA).

The two peaks presented by complex expression profiles under SD behave differently under free-running conditions. One of the peaks is maintained only under constant light and the other one only under constant darkness (Fig. 26-C). This suggests how two peaks expression patterns under SD conditions could emerge as the combination of two distinct rhythmic profiles: one depending on the photoperiod (maintaining its rhythmicity only under LL) and another apparently depending on the skotoperiod (maintaining its rhythmicity only under DD).

Biological rhythms with 12 h periods (two peaks per day) have been described in different organisms, from marine animals to mammals including humans. It is hypothesized that 12 h rhythm of gene expression and metabolism in terrestrial organisms is reminiscent of the ~ 12-h circatidal rhythms of coastal and estuarine organisms ((el de 12h)). The maintenance of 12-h rhythms after evolving to live on land is hypothesized to provide an advantage in the adaptation to metabolic stress that peak at transition periods during the diurnal cycle. (A Cell-Autonomous Mammalian 12-hour Clock Coordinates Metabolic and Stress Rhythms , Revealing the hidden reality of the mammalian 12-h ultradian rhythms, 12-h clock regulation of genetic information flow by XBP1s). However, the 12 h cycles emerging only under shorter photoperiods have gone unnoticed (although they can be found reanalyzing already published transcriptomic data(cita chlamy y tauri) and thus there isn't an hypothesis about its biological role yet.

Bibliography

Abascal, F., Acosta, R., Addleman, N. J., Adrian, J., Afzal, V., Aken, B., Akiyama, J. A., Jammal, O. Al, Amrhein, H., Anderson, S. M., Andrews, G. R., Antoshechkin, I., Ardlie, K. G., Armstrong, J., Astley, M., Banerjee, B., Barkal, A. A., Barnes, I. H. A., Barozzi, I., … Myers, R. M. (2020). Perspectives on ENCODE. In *Nature* (Vol. 583, Issue 7818). https://doi.org/10.1038/s41586-020-2449-8

Ajjawi, I., Verruto, J., Aqui, M., Soriaga, L. B., Coppersmith, J., Kwok, K., Peach, L., Orchard, E., Kalb, R., Xu, W., Carlson, T. J., Francis, K., Konigsfeld, K., Bartalis, J., Schultz, A., Lambert, W., Schwartz, A. S., Brown, R., & Moellering, E. R. (2017). Lipid production in Nannochloropsis gaditana is doubled by decreasing expression of a single transcriptional regulator. *Nature Biotechnology*, *35*, 647–652. https://doi.org/10.1038/nbt.3865

Andreatta, G., & Tessmar-Raible, K. (2020). The Still Dark Side of the Moon: Molecular Mechanisms of Lunar-Controlled Rhythms and Clocks. In *Journal of Molecular Biology* (Vol. 432, Issue 12). https://doi.org/10.1016/j.jmb.2020.03.009

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene ontology: Tool for the unification of biology. *Nature Genetics*, *25*, 25–29. https://doi.org/10.1038/75556

Bachy, C., Wittmers, F., Muschiol, J., Hamilton, M., Henrissat, B., & Worden, A. Z. (2022). The Land-Sea Connection: Insights Into the Plant Lineage from a Green Algal Perspective. In *Annual Review of Plant Biology* (Vol. 73). https://doi.org/10.1146/annurev-arplant-071921-100530

Becker, B., & Marin, B. (2009). Streptophyte algae and the origin of embryophytes. *Annals of Botany*, *103*(7), 999–1004. https://doi.org/10.1093/aob/mcp044

Benites, L. F., Bucchini, F., Sanchez-Brosseau, S., Grimsley, N., Vandepoele, K., & Piganeau, G. (2021). Evolutionary Genomics of Sex-Related Chromosomes at the Base of the Green Lineage. *Genome Biology and Evolution*, *13*(10). https://doi.org/10.1093/gbe/evab216

Berube, P. M., Biller, S. J., Hackl, T., Hogle, S. L., Satinsky, B. M., Becker, J. W., Braakman, R., Collins, S. B., Kelly, L., Berta-Thompson, J., Coe, A., Bergauer, K., Bouman, H. A., Browning, T. J., De Corte, D., Hassler, C., Hulata, Y., Jacquot, J. E., Maas, E. W., … Chisholm, S. W. (2018). Data descriptor: Single cell genomes of Prochlorococcus, Synechococcus, and sympatric microbes from diverse marine environments. *Scientific Data*, *5*. https://doi.org/10.1038/sdata.2018.154

Blaby, I. K., Blaby-Haas, C. E., Tourasse, N., Hom, E. F. Y., Lopez, D., Aksoy, M., Grossman, A., Umen, J., Dutcher, S., Porter, M., King, S., Witman, G. B., Stanke, M., Harris, E. H., Goodstein, D., Grimwood, J., Schmutz, J., Vallon, O., Merchant, S. S., & Prochnik, S. (2014). The Chlamydomonas genome project: A decade on. In *Trends in Plant Science* (Vol. 19, Issue 10). https://doi.org/10.1016/j.tplants.2014.05.008

Blanc-Mathieu, R., Verhelst, B., Derelle, E., Rombauts, S., Bouget, F.-Y., Carré, I., Château, A., Eyre-Walker, A., Grimsley, N., Moreau, H., Piégu, B., Rivals, E., Schackwitz, W., Van de Peer, Y., & Piganeau, G. (2014). An improved genome of the model marine alga Ostreococcus tauri unfolds by assessing Illumina de novo assemblies. *BMC Genomics*, *15*(1), 1103. https://doi.org/10.1186/1471-2164-15-1103

Blanc, G., Agarkova, I., Grimwood, J., Kuo, A., Brueggeman, A., Dunigan, D. D., Gurnon, J., Ladunga, I., Lindquist, E., Lucas, S., Pangilinan, J., Pröschold, T., Salamov, A., Schmutz, J., Weeks, D., Yamada, T., Lomsadze, A., Borodovsky, M., Claverie, J. M., … Van Etten, J. L. (2012). The genome of the polar eukaryotic microalga Coccomyxa subellipsoidea reveals traits of cold adaptation. *Genome Biology*, *13*(5). https://doi.org/10.1186/gb-2012-13-5-r39

Bowler, C., Allen, A. E., Badger, J. H., Grimwood, J., Jabbari, K., Kuo, A., Maheswari, U., Martens, C., Maumus, F., Otillar, R. P., Rayko, E., Salamov, A., Vandepoele, K., Beszteri, B., Gruber, A., Heijde, M., Katinka, M., Mock, T., Valentin, K., … Grigoriev, I. V. (2008). The Phaeodactylum genome reveals the evolutionary history of diatom genomes. *Nature*, *456*(7219). https://doi.org/10.1038/nature07410

Brandoli, C., Petri, C., Egea-Cortines, M., & Weiss, J. (2020). Gigantea: Uncovering new functions in flower development. In *Genes* (Vol. 11, Issue 10, pp. 1–15). MDPI AG. https://doi.org/10.3390/genes11101142

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, *34*(5). https://doi.org/10.1038/nbt.3519

Carbon, S., Douglass, E., Dunn, N., Good, B., Harris, N. L., Lewis, S. E., Mungall, C. J., Basu, S., Chisholm, R. L., Dodson, R. J., Hartline, E., Fey, P., Thomas, P. D., Albou, L. P., Ebert, D., Kesling, M. J., Mi, H., Muruganujan, A., Huang, X., … Westerfield, M. (2019). The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research*, *47*(D1), D330–D338. https://doi.org/10.1093/nar/gky1055

Carbon, S., Ireland, A., Mungall, C. J., Shu, S., Marshall, B., Lewis, S., Lomax, J., Mungall, C., Hitz, B., Balakrishnan, R., Dolan, M., Wood, V., Hong, E., & Gaudet, P. (2009). AmiGO: Online access to ontology and annotation data. *Bioinformatics*, *25*(2), 288–289. https://doi.org/10.1093/bioinformatics/btn615

Carlson, M., & Pagès, H. (2019). *AnnotationForge: Tools for building SQLite-based annotation data packages* (1.26.0.).

Chapman, R. L. (2013). Algae: The world's most important "plants"-an introduction. *Mitigation and Adaptation Strategies for Global Change*, *18*(1). https://doi.org/10.1007/s11027-010-9255-9

Chen, M. X., Zhang, Y., Fernie, A. R., Liu, Y. G., & Zhu, F. Y. (2021). SWATH-MS-Based Proteomics: Strategies and Applications in Plants. In *Trends in Biotechnology* (Vol. 39, Issue 5, pp. 433–437). Elsevier Ltd. https://doi.org/10.1016/j.tibtech.2020.09.002

Cheng, S., Xian, W., Fu, Y., Marin, B., Keller, J., Wu, T., Sun, W., Li, X., Xu, Y., Zhang, Y., Wittek, S., Reder, T., Günther, G., Gontcharov, A., Wang, S., Li, L., Liu, X., Wang, J., Yang, H., … Melkonian, M. (2019). Genomes of Subaerial Zygnematophyceae Provide Insights into Land Plant Evolution. *Cell*, *179*(5). https://doi.org/10.1016/j.cell.2019.10.019

Cock, J. M., & Coelho, S. M. (2011). Algal models in plant biology. In *Journal of Experimental Botany* (Vol. 62, Issue 8). https://doi.org/10.1093/jxb/err117

Coleto-Alcudia, V., & Vega-Rodríguez, M. A. (2020). Artificial Bee Colony algorithm based on Dominance (ABCD) for a hybrid gene selection method. *Knowledge-Based Systems*, *205*. https://doi.org/10.1016/j.knosys.2020.106323

Collado-Fabbri, S., Vaulot, D., & Ulloa, O. (2011). Structure and seasonal dynamics of the eukaryotic picophytoplankton community in a wind-driven coastal upwelling ecosystem. *Limnology and Oceanography*, *56*(6). https://doi.org/10.4319/lo.2011.56.6.2334

Correa, A., & Bell-Pedersen, D. (2002). Distinct signaling pathways from the circadian clock participate in regulation of rhythmic conidiospore development in Neurospora crassa. *Eukaryotic Cell*, *1*(2). https://doi.org/10.1128/EC.1.2.273-280.2002

Corteggiani Carpinelli, E., Telatin, A., Vitulo, N., Forcato, C., D'Angelo, M., Schiavon, R., Vezzi, A., Giacometti, G. M., Morosinotto, T., & Valle, G. (2014). Chromosome scale genome assembly and transcriptome profiling of nannochloropsis gaditana in nitrogen depletion. *Molecular Plant*, *7*(2). https://doi.org/10.1093/mp/sst120

Craig, R. J., Hasan, A. R., Ness, R. W., & Keightley, P. D. (2021). Comparative genomics of Chlamydomonas. *Plant Cell*, *33*(4). https://doi.org/10.1093/plcell/koab026

Cui, Y., Thomas-Hall, S. R., & Schenk, P. M. (2019). Phaeodactylum tricornutum microalgae as a rich source of omega-3 oil: Progress in lipid induction techniques towards industry adoption. In *Food Chemistry* (Vol. 297). https://doi.org/10.1016/j.foodchem.2019.06.004

De Keersmaecker, S. C. J., Thijs, I. M. V., Vanderleyden, J., & Marchal, K. (2006). Integration of omics data: How well does it work for bacteria? In *Molecular Microbiology* (Vol. 62, Issue 5). https://doi.org/10.1111/j.1365-2958.2006.05453.x

de los Reyes, P., Romero-Campero, F. J., Ruiz, M. T., Romero, J. M., & Valverde, F. (2017). Evolution of Daily Gene Co-expression Patterns from Algae to Plants. *Frontiers in Plant Science*. https://doi.org/10.3389/fpls.2017.01217

De Mairan, J. J. . (1729). Observation Botanique. *Histoire de l'Academie Royale Des Sciences*.

Del Campo, J. A., Rodríguez, H., Moreno, J., Vargas, M. Á., Rivas, J., & Guerrero, M. G. (2004). Accumulation of astaxanthin and lutein in Chlorella zofingiensis (Chlorophyta). *Applied Microbiology and Biotechnology*, *64*(6). https://doi.org/10.1007/s00253-003-1510-5

Demir-Hilton, E., Sudek, S., Cuvelier, M. L., Gentemann, C. L., Zehr, J. P., & Worden, A. Z. (2011). Global distribution patterns of distinct clades of the photosynthetic picoeukaryote Ostreococcus. *ISME Journal*, *5*(7). https://doi.org/10.1038/ismej.2010.209

Derelle, E., Ferraz, C., Rombauts, S., Rouze, P., Worden, A. Z., Robbens, S., Partensky, F., Degroeve, S., Echeynie, S., Cooke, R., Saeys, Y., Wuyts, J., Jabbari, K., Bowler, C., Panaud, O., Piegu, B., Ball, S. G., Ral, J.-P., Bouget, F.-Y., … Moreau, H. (2006). Genome analysis of the smallest free-living eukaryote Ostreococcus tauri unveils many unique features. *Proceedings of the National Academy of Sciences*, *103*(31), 11647–11652. https://doi.org/10.1073/pnas.0604795103

Ditz, B., Boekhoudt, J. G., Aliee, H., Theis, F. J., Nawijn, M., Brandsma, C.-A., Hiemstra, P. S., Timens, W., Tew, G. W., Grimbaldeston, M. A., Neighbors, M., Guryev, V., van den Berge, M., & Faiz, A. (2021). Comparison of genome-wide gene expression profiling by RNA Sequencing versus microarray in bronchial biopsies of COPD patients before and after inhaled corticosteroid treatment: does it provide new insights? *ERJ Open Research*, *7*(2), 00104–02021. https://doi.org/10.1183/23120541.00104-2021

Edmunds, L. N. (1983). Chronobiology at the cellular and molecular levels: Models and mechanisms for circadian timekeeping. *American Journal of Anatomy*, *168*(4). https://doi.org/10.1002/aja.1001680404

Eelderink-Chen, Z., Bosman, J., Sartor, F., Dodd, A. N., Kovács, Á. T., & Merrow, M. (2021). A circadian clock in a nonphotosynthetic prokaryote. *Science Advances*, *7*(2). https://doi.org/10.1126/sciadv.abe2086

Engel, S. R., Dietrich, F. S., Fisk, D. G., Binkley, G., Balakrishnan, R., Costanzo, M. C., Dwight, S. S., Hitz, B. C., Karra, K., Nash, R. S., Weng, S., Wong, E. D., Lloyd, P., Skrzypek, M. S., Miyasato, S. R., Simison, M., & Cherry, J. M. (2014). The Reference Genome Sequence of Saccharomyces cerevisiae: Then and Now. *G3: Genes, Genomes, Genetics*, *4*(3). https://doi.org/10.1534/g3.113.008995

Evans, D. R. (1961). Biological Clocks. Volume XXV. Cold Spring Harbor Symposia on Quantitative Biology. . *The Quarterly Review of Biology*, *36*(3). https://doi.org/10.1086/403447

Fauré-Fremiet, E. (1951). The tidal rhythm of the diatom Hantzschia amphioxys. *Biological Bulletin*, *100*(3), 173–177.

Frazee, A. C., Pertea, G., Jaffe, A. E., Langmead, B., Salzberg, S. L., & Leek, J. T. (2015). Ballgown bridges the gap between transcriptome assembly and expression analysis. *Nature Biotechnology*, *33*, 243. https://doi.org/10.1038/nbt.3172

Galperin, M. Y., Wolf, Y. I., Makarova, K. S., Alvarez, R. V., Landsman, D., & Koonin, E. V. (2021). COG database update: Focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Research*, *49*(D1). https://doi.org/10.1093/nar/gkaa1018

García-Domínguez, M., & Florencio, F. J. (1997). Nitrogen availability and electron transport control the expression of glnB gene (encoding PII protein) in the cyanobacterium Synechocystis sp. PCC 6803. *Plant Molecular Biology*, *35*(6). https://doi.org/10.1023/A:1005846626187

Gaspar, J. M. (2018). Improved peak-calling with MACS2. *BioRxiv*.

Gerotto, C., & Morosinotto, T. (2013). Evolution of photoprotection mechanisms upon land colonization: Evidence of PSBS-dependent NPQ in late Streptophyte algae. *Physiologia Plantarum*, *149*(4). https://doi.org/10.1111/ppl.12070

Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., & Rokhsar, D. S. (2012). Phytozome: A comparative platform for green plant genomics. *Nucleic Acids Research*, *40*(D1), D1178–D1186. https://doi.org/10.1093/nar/gkr944

Granados-Fuentes, D., Tseng, A., & Herzog, E. D. (2006). A circadian clock in the olfactory bulb controls olfactory responsivity. *Journal of Neuroscience*, *26*(47). https://doi.org/10.1523/JNEUROSCI.3445-06.2006

Grigoriev, I. V., Hayes, R. D., Calhoun, S., Kamel, B., Wang, A., Ahrendt, S., Dusheyko, S., Nikitin, R., Mondo, S. J., Salamov, A., Shabalov, I., & Kuo, A. (2021). PhycoCosm, a comparative algal genomics resource. *Nucleic Acids Research*, *49*(D1). https://doi.org/10.1093/nar/gkaa898

Hori, K., Maruyama, F., Fujisawa, T., Togashi, T., Yamamoto, N., Seo, M., Sato, S., Yamada, T., Mori, H., Tajima, N., Moriyama, T., Ikeuchi, M., Watanabe, M., Wada, H., Kobayashi, K., Saito, M., Masuda, T., Sasaki-Sekimoto, Y., Mashiguchi, K., … Ohta, H. (2014). Klebsormidium flaccidum genome reveals primary factors for plant terrestrial adaptation. *Nature Communications*, *5*. https://doi.org/10.1038/ncomms4978

Howe, K. L., Achuthan, P., Allen, J., Allen, J., Alvarez-Jarreta, J., Ridwan Amode, M., Armean, I. M., Azov, A. G., Bennett, R., Bhai, J., Billis, K., Boddu, S., Charkhchi, M., Cummins, C., da Rin Fioretto, L., Davidson, C., Dodiya, K., El Houdaigui, B., Fatima, R., … Flicek, P. (2021). Ensembl 2021. *Nucleic Acids Research*, *49*(D1). https://doi.org/10.1093/nar/gkaa942

Hoys, C., Romero-Losada, A. B., del Río, E., Guerrero, M. G., Romero-Campero, F. J., & García-González, M. (2021). Unveiling the underlying molecular basis of astaxanthin accumulation in Haematococcus through integrative metabolomic-transcriptomic analysis. *Bioresource Technology*, *332*. https://doi.org/10.1016/j.biortech.2021.125150

Ideker, T., Galitski, T., & Hood, L. (2001). A new approach to decoding life: Systems biology. In *Annual Review of Genomics and Human Genetics* (Vol. 2). https://doi.org/10.1146/annurev.genom.2.1.343

Jamers, A., Blust, R., & De Coen, W. (2009). Omics in algae: Paving the way for a systems biological understanding of algal stress phenomena? In *Aquatic Toxicology* (Vol. 92, Issue 3). https://doi.org/10.1016/j.aquatox.2009.02.012

Joyce, A. R., & Palsson, B. (2006). The model organism as a system: Integrating “omics” data sets. In *Nature Reviews Molecular Cell Biology* (Vol. 7, Issue 3). https://doi.org/10.1038/nrm1857

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., & Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, *44*(D1), D457–D462. https://doi.org/10.1093/nar/gkv1070

Karahalil, B. (2016). Overview of Systems Biology and Omics Technologies. *Current Medicinal Chemistry*, *23*(37), 4221–4230. https://doi.org/10.2174/0929867323666160926150617

Kester, D. R., Duedall, I. W., Connors, D. N., & Pytkowicz, R. M. (1967). PREPARATION OF ARTIFICIAL SEAWATER. *Limnology and Oceanography*, *12*(1), 176–179. https://doi.org/10.4319/LO.1967.12.1.0176

Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: A fast spliced aligner with low memory requirements. *Nature Methods*, *12*, 357–360. https://doi.org/10.1038/nmeth.3317

Klante, G., & Steinlechner, S. (1994). Light Irradiance and Wavelength as Seasonal Cues for Djungarian Hamsters. *Biological Rhythm Research*, *25*(4). https://doi.org/10.1080/09291019409360310

Konopka, R. J., & Benzer, S. (1971). Clock mutants of Drosophila melanogaster. *Proceedings of the National Academy of Sciences of the United States of America*, *68*(9). https://doi.org/10.1073/pnas.68.9.2112

Krumholz, E. W., Yang, H., Weisenhorn, P., Henry, C. S., & Libourel, I. G. L. (2012). Genome-wide metabolic network reconstruction of the picoalga Ostreococcus. *Journal of Experimental Botany*, *63*(6), 2353–2362. https://doi.org/10.1093/jxb/err407

Kuhlman, S. J., Craig, L. M., & Duffy, J. F. (2018). Introduction to chronobiology. *Cold Spring Harbor Perspectives in Biology*, *10*(9). https://doi.org/10.1101/cshperspect.a033613

Lamesch, P., Berardini, T. Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., Muller, R., Dreher, K., Alexander, D. L., Garcia-Hernandez, M., Karthikeyan, A. S., Lee, C. H., Nelson, W. D., Ploetz, L., Singh, S., Wensel, A., & Huala, E. (2012). The Arabidopsis Information Resource (TAIR): Improved gene annotation and new tools. *Nucleic Acids Research*, *40*(D1). https://doi.org/10.1093/nar/gkr1090

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4). https://doi.org/10.1038/nmeth.1923

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M. T., & Carey, V. J. (2013). Software for Computing and Annotating Genomic Ranges. *PLoS Computational Biology*, *9*(8). https://doi.org/10.1371/journal.pcbi.1003118

Le Bihan, T., Martin, S. F., Chirnside, E. S., van Ooijen, G., Barrios-LLerena, M. E., O'Neill, J. S., Shliaha, P. V., Kerr, L. E., & Millar, A. J. (2011). Shotgun proteomic analysis of the unicellular alga Ostreococcus tauri. *Journal of Proteomics*, *74*(10). https://doi.org/10.1016/j.jprot.2011.05.028

Lebert, M., Porst, M., & Häder, D. P. (1999). Circadian rhythm of gravitaxis in Euglena gracilis. *Journal of Plant Physiology*, *155*(3). https://doi.org/10.1016/S0176-1617(99)80115-1

Leconte, J., Benites, L. F., Vannier, T., Wincker, P., Piganeau, G., & Jaillon, O. (2020). Genome resolved biogeography of mamiellales. *Genes*, *11*(1). https://doi.org/10.3390/genes11010066

Lelandais, G., Scheiber, I., Paz-Yepes, J., Lozano, J.-C., Botebol, H., Pilátová, J., Žárský, V., Léger, T., Blaiseau, P.-L., Bowler, C., Bouget, F.-Y., Camadro, J.-M., Sutak, R., & Lesuisse, E. (2016). Ostreococcus tauri is a new model green alga for studying iron metabolism in eukaryotic phytoplankton. *BMC Genomics*, *17*(1), 319. https://doi.org/10.1186/s12864-016-2666-6

Leliaert, F., Smith, D. R., Moreau, H., Herron, M. D., Verbruggen, H., Delwiche, C. F., & De Clerck, O. (2012). Phylogeny and Molecular Evolution of the Green Algae. *Critical Reviews in Plant Sciences*, *31*(1). https://doi.org/10.1080/07352689.2011.615705

Li, X. P., Björkman, O., Shih, C., Grossman, A. R., Rosenquist, M., Jansson, S., & Niyogi, K. K. (2000). A pigment-binding protein essential for regulation of photosynthetic light harvesting. *Nature*, *403*(6768). https://doi.org/10.1038/35000131

Lopez, D., Casero, D., Cokus, S. J., Merchant, S. S., & Pellegrini, M. (2011). Algal Functional Annotation Tool: A web-based analysis suite to functionally interpret large gene lists using integrated annotation and expression data. *BMC Bioinformatics*, *12*, 282. https://doi.org/10.1186/1471-2105-12-282

Ludwig, C., Gillet, L., Rosenberger, G., Amon, S., Collins, B. C., & Aebersold, R. (2018). Data-independent acquisition-based SWATH - MS for quantitative proteomics: a tutorial. *Molecular Systems Biology*, *14*(8). https://doi.org/10.15252/msb.20178126

Masseroli, M., Martucci, D., & Pinciroli, F. (2004). GFINDer: Genome Function INtegrated Discoverer through dynamic annotation, statistical analysis, and mining. *Nucleic Acids Research*, *32*(WEB SERVER ISS.). https://doi.org/10.1093/nar/gkh432

Mazzocchi, F. (2012). Complexity and the reductionism-holism debate in systems biology. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, *4*(5), 413–427. https://doi.org/10.1002/wsbm.1181

McClung, C. R. (2006). Plant circadian rhythms. In *Plant Cell* (Vol. 18, Issue 4). https://doi.org/10.1105/tpc.106.040980

Merchant, S. S., Prochnik, S. E., Vallon, O., Harris, E. H., Karpowicz, S. J., Witman, G. B., Terry, A., Salamov, A., Fritz-Laylin, L. K., Maréchal-Drouard, L., Marshall, W. F., Qu, L. H., Nelson, D. R., Sanderfoot, A. A., Spalding, M. H., Kapitonov, V. V., Ren, Q., Ferris, P., Lindquist, E., … Zhou, K. (2007). The Chlamydomonas genome reveals the evolution of key animal and plant functions. *Science*, *318*(5848). https://doi.org/10.1126/science.1143609

Mermet, J., Yeung, J., & Naef, F. (2017). Systems chronobiology: Global analysis of gene regulation in a 24-hour periodic world. *Cold Spring Harbor Perspectives in Biology*, *9*(3). https://doi.org/10.1101/cshperspect.a028720

Merrow, M., Spoelstra, K., & Roenneberg, T. (2005). The circadian cycle: Daily rhythms from behaviour to genes. In *EMBO Reports* (Vol. 6, Issue 10). https://doi.org/10.1038/sj.embor.7400541

Mi, H., Ebert, D., Muruganujan, A., Mills, C., Albou, L. P., Mushayamaha, T., & Thomas, P. D. (2021). PANTHER version 16: A revised family classification, tree-based classification tool, enhancer regions and extensive API. *Nucleic Acids Research*, *49*(D1). https://doi.org/10.1093/nar/gkaa1106

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., Tosatto, S. C. E., Paladin, L., Raj, S., Richardson, L. J., Finn, R. D., & Bateman, A.

(2021). Pfam: The protein families database in 2021. *Nucleic Acids Research*, *49*(D1). https://doi.org/10.1093/nar/gkaa913

Moreau, H, Grimsley, N. Derelle, E, Ferraz, C, Escande, ML, Eychenié, S , Cooke, R, Piganeau, G, Desdevises, Y, Bellec, L. (1995). *A new marine picoeucaryote: Ostreococcus tauri gen. et sp. nov. (Chlorophyta, Prasinophyceae)*. *34*(4), 285–292.

Moreau, H., Verhelst, B., Couloux, A., Derelle, E., Rombauts, S., Grimsley, N., Van Bel, M., Poulain, J., Katinka, M., Hohmann-Marriott, M. F., Piganeau, G., Rouzé, P., Da Silva, C., Wincker, P., Van de Peer, Y., & Vandepoele, K. (2012). Gene functionalities and genome structure in Bathycoccus prasinos reflect cellular specializations at the base of the green lineage. *Genome Biology*, *13*(8). https://doi.org/10.1186/gb-2012-13-8-r74

Morimoto, D., Yoshida, T., & Sawayama, S. (2020). Draft Genome Sequence of the Astaxanthin-Producing Microalga Haematococcus lacustris Strain NIES-144. *Microbiology Resource Announcements*, *9*(23). https://doi.org/10.1128/mra.00128-20

Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C., & Kanehisa, M. (2007). KAAS: An automatic genome annotation and pathway reconstruction server. *Nucleic Acids Research*, *35*(SUPPL.2), 182–185. https://doi.org/10.1093/nar/gkm321

Ngan, C. Y., Wong, C. H., Choi, C., Yoshinaga, Y., Louie, K., Jia, J., Chen, C., Bowen, B., Cheng, H., Leonelli, L., Kuo, R., Baran, R., Garciá-Cerdán, J. G., Pratap, A., Wang, M., Lim, J., Tice, H., Daum, C., Xu, J., … Wei, C. L. (2015). Lineage-specific chromatin signatures reveal a regulator of lipid metabolism in microalgae. *Nature Plants*, *1*, 15107. https://doi.org/10.1038/nplants.2015.107

Nishiwaki-Ohkawa, T., & Yoshimura, T. (2016). Molecular basis for regulating seasonal reproduction in vertebrates. In *Journal of Endocrinology* (Vol. 229, Issue 3). https://doi.org/10.1530/JOE-16-0066

Noordally, Z. B., & Millar, A. J. (2015). Clocks in algae. *Biochemistry*, *54*(2), 171–183. https://doi.org/10.1021/bi501089x

Numata, H., Miyazaki, Y., & Ikeno, T. (2015). Common features in diverse insect clocks. *Zoological Letters*, *1*(1). https://doi.org/10.1186/s40851-014-0003-y

O'Kelly, C. J., Sieracki, M. E., Thier, E. C., & Hobson, I. C. (2003). A transient bloom of Ostreococcus (Chlorophyta, Prasinophyceae) in West Neck Bay, Long Island, New York. *Journal of Phycology*, *39*(5). https://doi.org/10.1046/j.1529-8817.2003.02201.x

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., & Kanehisa, M. (1999). KEGG: Kyoto encyclopedia of genes and genomes. In *Nucleic Acids Research*. https://doi.org/10.1093/nar/27.1.29

Ottesen, E. A., Young, C. R., Eppley, J. M., Ryan, J. P., Chavez, F. P., Scholin, C. A., & DeLong, E. F. (2013). Pattern and synchrony of gene expression among sympatric marine microbial populations. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(6). https://doi.org/10.1073/pnas.1222099110

Palenik, B., Brahamsha, B., Larimer, F. W., Land, M., Hauser, L., Chain, P., Lamerdin, J., Regala, W., Allen, E. E., McCarren, J., Paulsen, I., Dufresne, A., Partensky, F., Webb, E. A., & Waterbury, J. (2003). The genome of a motile marine Synechococcus. *Nature*, *424*(6952). https://doi.org/10.1038/nature01943

Palenik, B., Grimwood, J., Aerts, A., Rouzé, P., Salamov, A., Putnam, N., Dupont, C., Jorgensen, R., Derelle, E., Rombauts, S., Zhou, K., Otillar, R., Merchant, S. S., Podell, S., Gaasterland, T., Napoli, C., Gendler, K., Manuell, A., Tai, V., … Grigoriev, I. V. (2007). The tiny eukaryote Ostreococcus provides genomic insights into the paradox of plankton speciation. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(18). https://doi.org/10.1073/pnas.0611046104

Parsons, R., Parsons, R., Garner, N., Oster, H., & Rawashdeh, O. (2020). CircaCompare: A method to estimate and statistically support differences in mesor, amplitude and phase, between circadian rhythms. *Bioinformatics*, *36*(4). https://doi.org/10.1093/bioinformatics/btz730

Peers, G., Truong, T. B., Ostendorf, E., Busch, A., Elrad, D., Grossman, A. R., Hippler, M., & Niyogi, K. K. (2009). An ancient light-harvesting protein is critical for the regulation of algal photosynthesis. *Nature*, *462*(7272). https://doi.org/10.1038/nature08587

Pereira, H., Sá, M., Maia, I., Rodrigues, A., Teles, I., Wijffels, R. H., Navalho, J., & Barbosa, M. (2021). Fucoxanthin production from Tisochrysis lutea and Phaeodactylum tricornutum at industrial scale. *Algal Research*, *56*. https://doi.org/10.1016/j.algal.2021.102322

Pertea, M., Kim, D., Pertea, G. M., Leek, J. T., & Salzberg, S. L. (2016). Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nature Protocols*, *11*(9). https://doi.org/10.1038/nprot.2016.095

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T. C., Mendell, J. T., & Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, *33*, 290–295. https://doi.org/10.1038/nbt.3122

Pfeuty, B., Thommen, Q., Corellou, F., Djouani-Tahri, E. B., Bouget, F. Y., & Lefranc, M. (2012). Circadian clocks in changing weather and seasons: Lessons from the picoalga ostreococcus tauri. *BioEssays*, *34*(9), 781–790. https://doi.org/10.1002/bies.201200012

Pittendrigh, C. S. (1960). Circadian rhythms and the circadian organization of living systems. *Cold Spring Harbor Symposia on Quantitative Biology*, *25*. https://doi.org/10.1101/SQB.1960.025.01.015

Polle, J. E. W., Barry, K., Cushman, J., Schmutz, J., Tran, D., Hathwaik, L. T., Yim, W. C., Jenkins, J., McKie-Krisberg, Z., Prochnik, S., Lindquist, E., Dockter, R. B., Adam, C., Molina, H., Bunkenborg, J., Jin, E. S., Buchheim, M., & Magnuson, J. (2017). Draft nuclear genome sequence of the halophilic and beta-carotene-accumulating green alga Dunaliella salina strain CCAP19/18. In *Genome Announcements* (Vol. 5, Issue 43). https://doi.org/10.1128/genomeA.01105-17

Potter, S. C., Luciani, A., Eddy, S. R., Park, Y., Lopez, R., & Finn, R. D. (2018). HMMER web server: 2018 update. *Nucleic Acids Research*, *46*(W1). https://doi.org/10.1093/nar/gky448

Prendergast, S. C., Strobl, A. C., Cross, W., Pillay, N., Strauss, S. J., Ye, H., Lindsay, D., Tirabosco, R., Chalker, J., Mahamdallie, S. S., Sosinsky, A., Flanagan, A. M., & Amary, F. (2020). Sarcoma and the 100,000 Genomes Project: our experience and changes to practice. *Journal of Pathology: Clinical Research*, *6*(4). https://doi.org/10.1002/cjp2.174

Prochnik, S. E., Umen, J., Nedelcu, A. M., Hallmann, A., Miller, S. M., Nishii, I., Ferris, P., Kuo, A., Mitros, T., Fritz-Laylin, L. K., Hellsten, U., Chapman, J., Simakov, O., Rensing, S. A., Terry, A., Pangilinan, J., Kapitonov, V., Jurka, J., Salamov, A., … Rokhsar, D. S. (2010). Genomic analysis of organismal complexity in the multicellular green alga volvox carteri. *Science*, *329*(5988). https://doi.org/10.1126/science.1188800

Radakovits, R., Jinkerson, R. E., Fuerstenberg, S. I., Tae, H., Settlage, R. E., Boore, J. L., & Posewitz, M. C. (2012). Draft genome sequence and genetic transformation of the oleaginous alga Nannochloropis gaditana. *Nature Communications*, *3*. https://doi.org/10.1038/ncomms1688

Rayko, E., Maumus, F., Maheswari, U., Jabbari, K., & Bowler, C. (2010). Transcription factor families inferred from genome sequences of photosynthetic stramenopiles. *New Phytologist*, *188*(1). https://doi.org/10.1111/j.1469-8137.2010.03371.x

Ripperger, J. A., & Merrow, M. (2011). Perfect timing: Epigenetic regulation of the circadian clock. In *FEBS Letters* (Vol. 585, Issue 10). https://doi.org/10.1016/j.febslet.2011.04.047

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, *43*(7), e47. https://doi.org/10.1093/nar/gkv007

Rock, A., Wilcockson, D., & Last, K. S. (2022). Towards an Understanding of Circatidal Clocks. In *Frontiers in Physiology* (Vol. 13). https://doi.org/10.3389/fphys.2022.830107

Roenneberg, T., Foster, R. G., & Klerman, E. B. (2022). The circadian system, sleep, and the health/disease balance: a conceptual review. *Journal of Sleep Research*, *April*, 1–14. https://doi.org/10.1111/jsr.13621

Roenneberg, T., & Merrow, M. (2005). Circadian clocks - The fall and rise of physiology. In *Nature Reviews Molecular Cell Biology* (Vol. 6, Issue 12). https://doi.org/10.1038/nrm1766

Roenneberg, T., & Merrow, M. (2016). The circadian clock and human health. In *Current Biology* (Vol. 26, Issue 10). https://doi.org/10.1016/j.cub.2016.04.011

Roenneberg, T., Pilz, L. K., Zerbini, G., & Winnebeck, E. C. (2019). Chronotype and social jetlag: A (self-) critical review. In *Biology* (Vol. 8, Issue 3). https://doi.org/10.3390/biology8030054

Romero-Campero, F. J., Perez-Hurtado, I., Lucas-Reina, E., Romero, J. M., & Valverde, F. (2016). 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*, *17*, 227. https://doi.org/10.1186/s12864-016-2564-y

Romero-Losada, A. B., Arvanitidou, C., de los Reyes, P., García-González, M., & Romero-Campero, F. J. (2022). ALGAEFUN with MARACAS, microALGAE FUNctional enrichment tool for MicroAlgae RnA-seq and Chip-seq AnalysiS. *BMC Bioinformatics*, *23*(1). https://doi.org/10.1186/s12859-022-04639-5

Roth, M. S., Cokus, S. J., Gallaher, S. D., Walter, A., Lopez, D., Erickson, E., Endelman, B., Westcott, D., Larabell, C. A., Merchant, S. S., Pellegrini, M., & Niyogi, K. K. (2017). Chromosome-level genome assembly and transcriptome of the green alga Chromochloris zofingiensis illuminates astaxanthin production. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(21), E4296–E4305. https://doi.org/10.1073/pnas.1619928114

Rufty, T. W., & Huber, S. C. (1983). Changes in Starch Formation and Activities of Sucrose Phosphate Synthase and Cytoplasmic Fructose-1,6-bisphosphatase in Response to Source-Sink Alterations. *Plant Physiology*, *72*(2). https://doi.org/10.1104/pp.72.2.474

Sekimoto, H. (2017). Sexual reproduction and sex determination in green algae. *Journal of Plant Research*, *130*(3). https://doi.org/10.1007/s10265-017-0908-6

Serrano-Bueno, G., Romero-Campero, F. J., Lucas-Reina, E., Romero, J. M., & Valverde, F. (2017). Evolution of photoperiod sensing in plants and algae. *Current Opinion in Plant Biology*, *37*, 10–17. https://doi.org/10.1016/j.pbi.2017.03.007

Serrano-Pérez, E., Romero-Losada, A. B., Morales-Pineda, M., García-Gómez, M. E., Couso, I., García-González, M., & Romero-Campero, F. J. (2022). Transcriptomic and

Metabolomic Response to High Light in the Charophyte Alga Klebsormidium nitens. *Frontiers in Plant Science*, *13*. https://doi.org/10.3389/fpls.2022.855243

Sharma, A., Tripathi, V., & Kumar, V. (2022). Control and adaptability of seasonal changes in behavior and physiology of latitudinal avian migrants: Insights from laboratory studies in Palearctic-Indian migratory buntings. *Journal of Experimental Zoology Part A: Ecological and Integrative Physiology*, 35677956. https://doi.org/10.1002/jez.2631

Shen, M., Chang, Y. T., Wu, C. T., Parker, S. J., Saylor, G., Wang, Y., Yu, G., Van Eyk, J. E., Clarke, R., Herrington, D. M., & Wang, Y. (2022). Comparative assessment and novel strategy on methods for imputing proteomics data. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-022-04938-0

Sterck, L., Billiau, K., Abeel, T., Rouzé, P., & Van de Peer, Y. (2012). ORCAE: online resource for community annotation of eukaryotes. *Nature Methods*. https://doi.org/10.1038/nmeth.2242

Swanson, W. J., Aagaard, J. E., Vacquier, V. D., Monné, M., Sadat Al Hosseini, H., & Jovine, L. (2011). The molecular basis of sex: Linking yeast to human. *Molecular Biology and Evolution*, *28*(7). https://doi.org/10.1093/molbev/msr026

Swarbreck, D., Wilks, C., Lamesch, P., Berardini, T. Z., Garcia-Hernandez, M., Foerster, H., Li, D., Meyer, T., Muller, R., Ploetz, L., Radenbaugh, A., Singh, S., Swing, V., Tissier, C., Zhang, P., & Huala, E. (2008). The Arabidopsis Information Resource (TAIR): Gene structure and function annotation. *Nucleic Acids Research*, *36*(SUPPL. 1). https://doi.org/10.1093/nar/gkm965

Takahashi, J. S. (2021). The 50th anniversary of the konopka and benzer 1971 paper in PNAS: “Clock mutants of drosophila melanogaster.” In *Proceedings of the National Academy of Sciences of the United States of America* (Vol. 118, Issue 39). https://doi.org/10.1073/pnas.2110171118

Thaben, P. F., & Westermark, P. O. (2014). Detecting rhythms in time series with rain. *Journal of Biological Rhythms*, *29*(6). https://doi.org/10.1177/0748730414553029

Tian, T., Liu, Y., Yan, H., You, Q., Yi, X., Du, Z., Xu, W., & Su, Z. (2017). AgriGO v2.0: A GO analysis toolkit for the agricultural community, 2017 update. *Nucleic Acids Research*, *45*(W1), W122–W129. https://doi.org/10.1093/nar/gkx382

Tragin, M., & Vaulot, D. (2019). Novel diversity within marine Mamiellophyceae (Chlorophyta) unveiled by metabarcoding. *Scientific Reports*, *9*(1). https://doi.org/10.1038/s41598-019-41680-6

Veenstra, T. D. (2021). Omics in Systems Biology: Current Progress and Future Outlook. In *Proteomics* (Vol. 21, Issues 3–4). https://doi.org/10.1002/pmic.202000235

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: A revolutionary tool for transcriptomics. In *Nature Reviews Genetics* (Vol. 10, Issue 1). https://doi.org/10.1038/nrg2484

Watson, J. V., Chambers, S. H., & Smith, P. J. (1987). A pragmatic approach to the analysis of DNA histograms with a definable G1 peak. *Cytometry*, *8*(1), 1–8. https://doi.org/10.1002/CYTO.990080101

Weckwerth, W. (2011). Green systems biology - From single genomes, proteomes and metabolomes to ecosystems research and biotechnology. In *Journal of Proteomics* (Vol. 75, Issue 1). https://doi.org/10.1016/j.jprot.2011.07.010

Willforss, J., Chawade, A., & Levander, F. (2019). NormalyzerDE: Online Tool for Improved Normalization of Omics Expression Data and High-Sensitivity Differential Expression Analysis. *Journal of Proteome Research*, *18*(2). https://doi.org/10.1021/acs.jproteome.8b00523

Worden, A. Z., Lee, J. H., Mock, T., Rouzé, P., Simmons, M. P., Aerts, A. L., Allen, A. E., Cuvelier, M. L., Derelle, E., Everett, M. V., Foulon, E., Grimwood, J., Gundlach, H., Henrissat, B., Napoli, C., McDonald, S. M., Parker, M. S., Rombauts, S., Salamov, A., … Grigoriev, I. V. (2009). Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes micromonas. *Science*, *324*(5924). https://doi.org/10.1126/science.1167222

Worden, A. Z., Nolan, J. K., & Palenik, B. (2004). Assessing the dynamics and ecology of marine picophytoplankton: The importance of the eukaryotic component. *Limnology and Oceanography*, *49*(1). https://doi.org/10.4319/lo.2004.49.1.0168

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., Fu, X., Liu, S., Bo, X., & Yu, G. (2021). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation*, *2*(3). https://doi.org/10.1016/j.xinn.2021.100141

Yang, M., Lin, X., Liu, X., Zhang, J., & Ge, F. (2018). Genome Annotation of a Model Diatom Phaeodactylum tricornutum Using an Integrated Proteogenomic Pipeline. *Molecular Plant*, *11*(10). https://doi.org/10.1016/j.molp.2018.08.005

Youthed, G. J., & Moran, V. C. (1969). The lunar-day activity rhythm of myrmeleontid larvae. *Journal of Insect Physiology*, *15*(7). https://doi.org/10.1016/0022-1910(69)90235-2

Yu, G., Wang, L. G., & He, Q. Y. (2015). ChIP seeker: An R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics*, *31*(14), 2382–2383. https://doi.org/10.1093/bioinformatics/btv145

Zee, P. C., & Abbott, S. M. (2020). Circadian Rhythm Sleep-Wake Disorders. In *CONTINUUM Lifelong Learning in Neurology* (Vol. 26, Issue 4). https://doi.org/10.1212/CON.0000000000000884

Zhao, X., Rastogi, A., Deton Cabanillas, A. F., Ait Mohamed, O., Cantrel, C., Lombard, B., Murik, O., Genovesio, A., Bowler, C., Bouyer, D., Loew, D., Lin, X., Veluchamy, A., Vieira, F. R. J., & Tirichine, L. (2021). Genome wide natural variation of H3K27me3 selectively marks genes predicted to be important for cell differentiation in Phaeodactylum tricornutum. *New Phytologist*, *229*(6). https://doi.org/10.1111/nph.17129

Zheng, Y., Jiao, C., Sun, H., Rosli, H. G., Pombo, M. A., Zhang, P., Banf, M., Dai, X., Martin, G. B., Giovannoni, J. J., Zhao, P. X., Rhee, S. Y., & Fei, Z. (2016). iTAK: A Program for Genome-wide Prediction and Classification of Plant Transcription Factors, Transcriptional Regulators, and Protein Kinases. In *Molecular Plant* (Vol. 9, Issue 12). https://doi.org/10.1016/j.molp.2016.09.014

Zhu, L. J. (2013). Integrative analysis of ChIP-chip and ChIP-seq dataset. *Methods in Molecular Biology*, *1067*. https://doi.org/10.1007/978-1-62703-607-8_8

Zurbriggen, M. D., Moor, A., & Weber, W. (2012). Plant and bacterial systems biology as platform for plant synthetic bio(techno)logy. *Journal of Biotechnology*, *160*(1–2). https://doi.org/10.1016/j.jbiotec.2012.01.014