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

Francisco J. Romero-Campero¹, Ignacio Perez-Hurtado¹, Eva Lucas-Reina², Jose M. Romero² and Federico Valverde²*  

Abstract  
Background: *Chlamydomonas reinhardtii* is the model organism that serves as a reference for studies in algal genomics and physiology. It is of special interest in the study of the evolution of regulatory pathways from algae to higher plants. Additionally, it has recently gained attention as a potential source for bio-fuel and bio-hydrogen production. The genome of *Chlamydomonas* is available, facilitating the analysis of its transcriptome by RNA-seq data. This has produced a massive amount of data that remains fragmented making necessary the application of integrative approaches based on molecular systems biology.  
Results: We constructed a gene co-expression network based on RNA-seq data and developed a web-based tool, ChlamyNET, for the exploration of the *Chlamydomonas* transcriptome. ChlamyNET exhibits a scale-free and small world topology. Applying clustering techniques, we identified nine gene clusters that capture the structure of the transcriptome under the analyzed conditions. One of the most central clusters was shown to be involved in carbon/nitrogen metabolism and signalling, whereas one of the most peripheral clusters was involved in DNA replication and cell cycle regulation. The transcription factors and regulators in the *Chlamydomonas* genome have been identified in ChlamyNET. The biological processes potentially regulated by them as well as their putative transcription factor binding sites were determined. The putative light regulated transcription factors and regulators in the *Chlamydomonas* genome were analyzed in order to provide a case study on the use of ChlamyNET. Finally, we used an independent data set to cross-validate the predictive power of ChlamyNET.  
(Continued on next page)
Background

The unicellular green alga *Chlamydomonas reinhardtii* (*Chlamydomonas*) is an important model organism for genomic and physiological studies in photosynthetic organisms. Due to its evolutionary position, it diverged from land-plants over a billion years ago, *Chlamydomonas* is considered a living representative of the photosynthetic organisms that gave rise to the green lineage [1]. Specifically, it has been used as a model organism to study the establishment, conservation and divergence of key biological processes in photosynthetic organisms such as the photoperiod response [2–4]. Recently, *Chlamydomonas* has attracted substantial interest for biotechnological applications in the context of bio-fuel and bio-hydrogen production [5–7]. The main advantage of using *Chlamydomonas* over higher plants is that it does not compete for agricultural land use. Additionally, *Chlamydomonas* possesses powerful genetic tools, metabolic versatility and a haploid genome. However, an important disadvantage is the lack of sufficient functional and regulatory characterization of the molecular mechanisms underpinning these processes with biotechnological interest [8].

In order to overcome this limitation its genome was sequenced and it is currently in an advanced curated state [1, 9]. The availability of its genome has facilitated the use of Next Generation Sequencing techniques, specially RNA-seq, in order to study its complete transcriptome. This has produced a massive amount of data from a variety of genotypes grown under relevant physiological conditions [10–16]. However, these studies remain fragmented without producing global insights into the organization and regulation of the *Chlamydomonas* transcriptome. The first steps towards the use of molecular systems biology methodologies to characterize the *Chlamydomonas* transcriptome has been taken [17–19]. Nevertheless, one of the most widely used tools for the integration and study of massive amounts of transcriptomic data, gene co-expression networks, have not yet been developed for *Chlamydomonas*, while gene co-expression networks have been used successfully in many other photosynthetic organisms [20–22].

Gene co-expression networks integrate fragmented transcriptomic data obtained in different studies in order to characterize patterns of coordinated gene expression at the whole transcriptome level. In gene co-expression networks nodes represent genes, being nodes connected by an edge if the corresponding genes are significantly co-expressed across appropriately chosen genotypes and physiological conditions [23]. Fundamental network concepts such as node degree, neighbourhood and clustering coefficient have important applications to unravel the organization and functioning of the represented transcriptome [24, 25]. The degree of a node, that is, the number of nodes connected to it, represents the number of genes co-expressed with the corresponding gene. Therefore, genes represented by nodes with high degrees are expected to be relevant in the transcriptome since their expression is coordinated with many others. The neighborhood of a node consists of genes co-expressed with the corresponding gene. This set of genes can be used as target genes candidates when the given gene is a transcription factor or potential regulator. The transcription factor binding sites that are responsible for the coordinated expression of genes can be identified by analyzing the significance of specific motifs in the promoters of co-expressed genes [26]. Additionally, Gene Ontology (GO) term enrichment over gene neighbourhoods can be applied to determine the potential biological processes that are carried out by the orchestrated expression of any given genes. In most gene co-expression networks the probability that a node is connected with k other nodes, \( P(k) \), follows a negative exponential distribution, \( P(k) \sim k^{-1} \). This is the defining property of scale-free networks [27]. In scale-free networks most nodes are connected with few nodes, whereas there exists a small number of highly connected nodes called hubs that dominate the network dynamics.
[28]. Genes that correspond to hub nodes play a key role in the correct functioning of biological processes and, therefore, their mutation can lead to severely affected phenotypes and even lethality [29]. The clustering coefficient of a node measures the tendency of nodes to group together around the given node, and when applied to gene co-expression networks, this concept indicates the tendency of genes to form highly co-expressed gene clusters. Scale-free networks with an average high clustering coefficient are called small world networks [28]. In this class of networks the existence of a clustering structure around hub nodes produces short paths that connect any pair of nodes. It has been often observed that biological co-expression networks are scale-free and small world networks [20, 25].

In this study we have developed ChlamyNET, a gene co-expression network and an associated web-based software tool that integrates the massive amount of RNA-seq data available for the *Chlamydomonas* transcriptome, see Additional file 1: Table S1. We have used this tool to study the organization and regulation of the algal transcriptome. ChlamyNET aims at becoming an enabling technology for researchers in the *Chlamydomonas* transcriptome, and in a wider perspective of algal transcriptomics, being the first tool of this kind existing at this date. Researchers can explore the neighbourhood of their genes of interest in ChlamyNET looking for potential targets or regulators. Additionally, our web tool can be used to determine GO terms related to biological processes, functions and components that are significantly present in the annotation of the neighbouring genes. Finally, we have used an independent experimental data set to cross-validate the predictive power of ChlamyNET.

**Results and discussion**

**Network construction and topology**

The high resolution provided by RNA-seq data and the diverse physiological conditions and genotypes analyzed allowed us to capture the co-expression relationships between genes in the *Chlamydomonas* transcriptome. In order to reduce the noise in our analysis, we only considered genes that showed significant changes in at least one comparison between a condition and its corresponding control. Data processing and selection of differentially expressed genes were performed as described in the methods section. Out of the 16624 predicted genes in the *Chlamydomonas* genome 13699 were differentially expressed in at least one of the conditions analyzed in this study. This represents 82.40 % of the algal genome, which shows that the analyzed conditions and phenotypes are diverse enough to capture the behaviour of most of the *Chlamydomonas* transcriptome.

As described in the methods section, we used a range of absolute correlation thresholds to determine the co-expression level between the selected genes [30]. A compromise between the generation of a scale-free network and a high density network was established. We observed that for increasing correlation thresholds, the density of the network decreased, whereas the fit to the scale-free property increased until the cut-off value was too restrictive and the network started to deteriorate (Fig. 1a). Indeed, the scale-free model fit exhibits a maximum at a correlation value of 0.90 with an $R^2$ equal to 0.86. According to this, we chose an absolute *Pearson correlation* threshold of 0.90 to consider that two genes are significantly co-expressed. The gene co-expression network generated for this threshold was called ChlamyNET.

ChlamyNET consists of 9171 genes or nodes exhibiting an overall of 139019 co-expression relationships or edges. ChlamyNET is composed of a major connected component consisting of 8443 genes (92.1 % of the entire network) and a multitude of small components consisting of a few genes. This global connectivity property of ChlamyNET is similar to previously constructed and analyzed networks from other organisms such as *Saccharomyces cerevisiae* [31] and *Arabidopsis thaliana* [32].

The scale-free property of ChlamyNET was corroborated by computing its degree distribution and checking that it follows a negative exponential distribution. Specifically, linear regression over the logarithmic transform of the degree distribution was used (Fig. 1b). Another topological property that we analyzed in ChlamyNET was the *clustering coefficient*, a measurement of the density of edges or co-expression relationships around genes. The distribution of the clustering coefficient in ChlamyNET was computed (Fig. 1c). The average clustering coefficient of ChlamyNET is 0.66 which is significantly high when compared to random scale-free networks, see the methods section. This shows that ChlamyNET constitutes a non-random scale-free network with a high clustering coefficient. This type of networks are called small-world networks since the minimal path length between genes is short when compared to random scale-free networks [33]. These properties are common in gene co-expression networks [20, 32]. In the case of ChlamyNET the average minimal path length between genes or the *network diameter* is 7.5 (Fig. 1d). Therefore, on average any gene on ChlamyNET can be reached from another one through approximately seven gene co-expression relationships.

The topological properties of ChlamyNET (Fig. 2a), namely scale-free and small-world properties, imply that most genes in the *Chlamydomonas* transcriptome are co-expressed with only a few other genes. However,
there exists a low number of genes that are co-expressed with a large number of other genes. These genes are called hubs and play a key role in the structure and functioning of gene co-expression networks [28]. We determined the first 1000 hubs in terms of their degree and highlighted them in ChlamyNET, observing that they are located in specific regions of the network (Fig. 2b). Nevertheless, the definition of hubs based solely on the number of genes it is co-expressed with has been found to be incomplete and the concept of authoritative hub has been introduced [34]. Following this line, in our context, an authoritative hub gene relevant to a biological process is not considered solely on the base that they have a large number of co-expressed genes. Additionally, since its co-expressed genes are involved in the same biological process they should in turn be co-expressed among themselves, establishing links in the network between them. These authoritative genes could then be responsible for bringing together genes potentially involved in a common biological process. We identified the first 1000 authoritative hubs using the HITS algorithm [34] and represented them in ChlamyNET (Fig. 2c). We observed that the most relevant authoritative hubs are located in the center of the network. Additionally, we observe that the location of regions with high clustering coefficient is not random. These regions substantially overlap with areas where hub genes are located (Fig. 2d). In order to determine whether or not these hub genes are involved in similar biological processes we performed a gene ontology (GO) term enrichment analysis (Table 1) based on orthology relationships with Arabidopsis thaliana and on the annotation of protein families and domains available in the Pfam database [35] as described in the methods section. This analysis revealed that the central part of ChlamyNET constitutes the core of the network where the most authoritative hub genes are located. These hubs interconnect key biological processes such as protein phosphorylation and response to light stimulus with carbon/nitrogen metabolism and transmembrane transport (Table 1). Protein kinases potentially involved in developmental processes such as CrMEKK (g5375), similar to the Arabidopsis MEKK gene At5g57610, and transcription factors possibly associated with circadian rhythms and photoperiodic responses like CrBbox1 (Cre03.g182700), similar to the Arabidopsis COL1 gene
At5g15850, are highly authoritative hub genes in the regulation of the Chlamydomonas transcriptome with more than 250 neighbours. Relevant enzymes in the carbon/nitrogen metabolism and transmembrane transport are also authoritative hub genes in the core of ChlamyNET indicating that their expression is highly regulated and coordinated with other biological processes. For example, the nitrate transporter NRT2.3 (Cre09.g396000), nitrate reductase NIT1 (Cre09.g410950) and starch phosphorylase CrPHS1 (Cre07.g336950) are also co-expressed with more than 250 genes.

Network clustering analysis and functional annotation

The specific location of hub genes in regions exhibiting a high clustering coefficient may reflect an underlying structure in ChlamyNET relevant to physiological functions related to the Chlamydomonas transcriptome. In ChlamyNET we can observe distinct areas composed of genes with high degree and clustering coefficient that are in turn loosely connected through other genes with low degrees and clustering coefficients (compare Figs. 2b and d). This indicates the existence of relatively isolated groups of genes whose expression are highly coordinated and, hence, are potentially involved in the same biological processes. In order to corroborate the existence of this underlying structure we applied clustering techniques over ChlamyNET using, as described in the methods section, the Pearson correlation coefficient between gene expression profiles as gene similarity measure. We compared the performance of the two most widely used clustering algorithms hierarchical clustering (HCLUST) and partition around medoids (PAM) for different number of clusters ranging from 4 to 20 clusters using the silhouette, a criterion that combines the minimization of inter-cluster similarity with the maximization of the intra-cluster similarity [36]. Our analysis concluded that the underlying structure of ChlamyNET is best described using nine clusters identified with the PAM algorithm (Fig. 3a) since this combination of clustering algorithm and number of clusters produced
the highest silhouette value of 0.28 (Fig. 3b). This identified different clusters which we highlighted with different colour codes. Each gene cluster exhibits distinct expression profiles, see Additional file 2: Figure S1.

Since the genes in each cluster are co-expressed throughout the diverse physiological conditions integrated in this study they are likely involved in the same biological processes. In order to study the biological processes in which each gene cluster is involved we performed a Gene Ontology (GO) term enrichment over these clusters [37]. In order to overcome the limitations imposed by the sparse annotation of the *Chlamydomonas* transcriptome we combined GO terms obtained using orthology information with *Arabidopsis*, together with GO terms associated to protein families defined using conserved protein domains available from the Pfam database [35]. Since the used annotation is not based on experimental studies, the results obtained here should be taken as predictions that would need further experimental validation. This is precisely one of the main goals of GO term enrichment, namely the computational prediction of gene function to be subsequently corroborated using wet laboratory experimental work [38]. The results of our clustering and functional analysis are depicted in Fig. 4 and summarized in Table 2. The nine different gene clusters are identified with different colours and numbers following the code in Fig. 3b. In order to place our clustering analysis into a physiological, biochemical and metabolic context we used the tools and databases available from the *Plant Metabolic Network* [39]. Specifically, we used the extensive metabolic pathway information provided by *ChlamyCyc* [40] to identify the metabolic pathways contained in each cluster. In the following subsections, we present in detail four gene clusters and their functional annotation.

**Cluster 2, brown - DNA replication, chromosome organization and regulation of cell cycle**

The most cohesive gene cluster is also the smallest one. The brown cluster is located in the periphery of ChlamyNET. It presents the highest silhouette value (0.44) in the network and contains 535 genes (Fig. 3b). Our GO term enrichment analysis reveals that this cluster is involved in cell cycle processes. Specifically, it is enriched in genes required for DNA replication (GO:0006260) such as DNA polymerase POLD1 (Cre01.g015250), replication factor RFA1 (Cre16.g651000) and origin recognition complex ORC2 (Cre03.g199400); genes associated with chromosome organization (GO:0051276) like structural maintenance of chromosomes SMC4 (Cre12.g493400) and SMC2 (Cre02.g086650) and genes involved in the regulation of cell cycle process (GO:0010564) such as the cyclin A/B CYCA1 (Cre03.g207900) and the A-type cyclin CYCA1 (Cre03.g207900).

The metabolic pathways located in this cluster produce DNA and RNA precursors such as the
pyrimidine deoxyribonucleotides de novo biosynthesis pathway. For example, the formation of the DNA-specific end product dTTP starts with the hydrolysis of dUTP to produce dUMP by the dUTP pyrophosphatase DUT (Cre16.g667850), followed by the reductive methylation of dUMP catalyzed by thymidylate synthase THY (Cre17.g715900) which yields dTMP. Finally, the thymidylate kinase TMPK (Cre03.g190800) catalyzes the first phosphorylation of dTMP leading to dTTP. These three enzymes are members of this cluster (Additional file 3: Figure S2 and Table 2).

Cluster 9, blue - protein phosphorylation, carbohydrate metabolic process and transmembrane transport

The blue cluster located in the center of ChlamyNET is enriched with hub genes according to a $p$-value $< 2.210^{-16}$ obtained using Fisher’s exact test. It is the second most cohesive cluster with a silhouette value of 0.40 and 1058 genes (Fig. 3b). The most
| Cluster          | Functional Annotation          | Representative Genes                                                                 | Metabolic/Signalling Pathways                                                                 |
|------------------|-------------------------------|---------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------|
| Cluster 2 (Brown) 535 genes Silhouette 0.44 | DNA replication (GO:0006260) | Cre01.g015250 - POLD1  
Cre16.g651000 - RFA1  
Cre02.g086650 - SMC2  
Cre12.g4 934 00 - SMC4  
Cre03.g340100 - CYCA1  
Cre10.g466200 - CYCA1 | Pyrimidine deoxyribonucleotides de novo biosynthesis pathway  
Cre16.g657850 - DUT  
Cre17.g715900 - THY  
Cre03.g190800 - TMPK |
|                  | Chromosome organization (GO:00051276) |                                                                                  |                                                                                               |
|                  | Regulation of Cell Cycle (GO:00051276) |                                                                                  |                                                                                               |
| Cluster 9 (Blue) 1058 genes Silhouette 0.40 | protein phosphorylation (GO:0006468) | Cre17.g742400 - PTK17  
Cre12.g537400 - CrAUR3  
Cre08.g384750 - AMY  
Cre10.g444700 - SBE3  
Cre09.g396000 - NRT2.3  
Cre13.g564650 - MRS5 | Starch Biosynthetic Pathway  
Cre04.g215150 - SSS Sucrose  
Biosynthetic Pathway  
Cre06.g283400 - SPP Nitrogen Assimilation  
Pathway Cre09.g410750 - NII1 |
|                  | carbohydrate metabolic process (GO:0005975) |                                                                                  |                                                                                               |
|                  | transmembrane transport (GO:00059085) |                                                                                  |                                                                                               |
| Cluster 1 (Orange) 824 genes Silhouette 0.38 | vesicle-mediated transport (GO:00016192) | Cre17.g728150 - Yky6  
Cre16.g676650 - AP1G1  
Cre12.g532600 - CGL44  
Cre07.g315350 - RABGAP  
Cre09.g391500 - APG9 | TAG Biosynthetic Pathway  
Cre02.g106400 - PDAT Phospholipid  
Biosynthetic Pathway  
Cre01.g035500 - PI3K Coenzyme A Biosynthetic Pathway  
Cre01.g048050 - COAB |
|                  | GTPase activity (GO:0043087) |                                                                                  |                                                                                               |
|                  | Autophagy (GO:0000914) |                                                                                  |                                                                                               |
| Cluster 3 (Red) 1723 genes Silhouette 0.28 | protein phosphorylation (GO:0006468) | Cre02.g145500 - PTK24  
Cre12.g498650 - ALK3  
Cre12.g532550 - RPL13a  
Cre09.g400650 - RPS6  
Cre03.g207250 - GLN4 | TAG Biosynthetic Pathway  
g9572 - DGAT1 Hydrogen production  
Cre09.g396600 - HYDA2 MAP kinase cascade  
Cre10.g461150 - CrMAPKKK |
|                  | ribosome biogenesis (GO:0042254) |                                                                                  |                                                                                               |
|                  | macromolecule biosynthesis (GO:00009059) |                                                                                  |                                                                                               |
| Cluster 4 (Purple) 1174 genes Silhouette 0.26 | translation (GO:0006412) | Cre03.g199900 - EIF4E  
Cre02.g117900 - RH  
Cre16.g653050 - SpoU  
Cre10.g421600 - ThrRS g4 679 - RNase P | tRNA Charging Pathway  
g2051 - TrpS Amino Acid Biosynthesis  
Cre03.g161400 - WSN2 Pentose Phosphate Non-oxidative  
Cre12.g511900 - RPE1 TAG Biosynthetic Pathway  
Cre03.g250500 - DGAT2 |
|                  | RNA processing (GO:0006396) |                                                                                  |                                                                                               |
|                  | lipid metabolism (GO:0006269) | Cre09.g397250 - FADS  
Cre06.g295250 - PAP  
Cre03.g207250 - GLN4 |                                                                                               |
| Cluster 7 (Green) 909 genes Silhouette 0.25 | protein complex assembly (GO:0006461) | g9912 - CSNS Cre16.g663500 - CRPN10  
Cre06.g280850 - PSMB4  
Cre12.g501200 - SKP1 | Aerobic Respiration Pathway  
Cre15.g638500 - CYC1 COP9 Signalling  
g11578 - CSN6 |
|                  | response to misfolded protein (GO:00051788) |                                                                                  |                                                                                               |
| Cluster 6 (Yellow) 1351 genes Silhouette 0.24 | chromatin organization (GO:0006325) | g11636 - HDA Cre13.g590750 - HTB37  
Cre06.g280850 - PSMB4  
Cre12.g501200 - SKP1 | Chromatin Remodelling Cre13.g591200 - HTB38  
Cre13.g562400 - ABI3 |
|                  | posttranscriptional regulation (GO:00010508) |                                                                                  |                                                                                               |
| Cluster 5 (Dark Green) 567 genes Silhouette 0.21 | response to heat (GO:0009408) | Cre14.g617400 - HSP22F  
Cre08.g372100 - HSP70A  
g9841 - FKBP Cre01.g047700 - CYP40 | Stress Response  
Cre02.g098800 - ERP29  
g9861 - TOR |
|                  | protein folding (GO:0006457) |                                                                                  |                                                                                               |
significantly over-represented GO terms in this cluster are protein phosphorylation (GO:0006468) with genes such as the protein tyrosine kinases PTK17 (Cre17.g742400) and ataurora CrALR3 (Cre12.g537400), carbohydrate metabolic process (GO:0005975) including genes like the alpha amylase AMY (Cre08.g384750), and transmembrane transport (GO:0055085) including genes coding for magnesium and cobalt transport protein MRS5 (Cre13.g564650). An analysis of the metabolic context of this cluster reveals that core pathways in carbon and nitrogen metabolism are contained in it. Starch is the major reservoir of energy and carbon in photosynthetic organisms. The starch biosynthetic pathway constituted by the enzymes glucose-6-phosphate isomerase PGI (Cre03.g175400), phosphoglucomutase PGM (g2899), ADP glucose pyrophosphorylase APL (Cre16.g683450), starch synthase SSS (Cre04.g215150) and 1,4-α-glucan branching enzyme SBE3 (Cre10.g44700) is entirely contained in this cluster. In Chlamydomonas, starch is degraded to hexoses during the dark period. The derived hexoses are then used in the sucrose synthesis pathway. Key enzymes in this pathway such as glyceraldehyde 3-phosphate dehydrogenase GAP1 (Cre12.g485150) and sucrose phosphate phosphatase SPP (Cre06.g283400) are members of this cluster. The oxidative branch of the pentose phosphate pathway produces NADPH in the reactions catalyzed by glucose-6-phosphate dehydrogenase GLD2 (Cre08.g378150) and 6-phosphogluconate dehydrogenase GNID1 (Cre12.g526800), enzymes coded by genes included in this cluster. NADPH is an important source of the reducing power required by many enzymes in central metabolic pathways. The anaplerotic pathway that fixes CO₂ into oxaloacetate through the enzymes carbonic anhydrase CAH8 (Cre09.g405750) and phosphoenolpyruvate carboxylases PPC (g16646 and g11831) is also part of this cluster (Additional file 4: Figure S3). This pathway replenishes depleted Tricaboxylic Acid (TCA) cycle compounds that have been used for nitrogen assimilation or other tasks [41]. Inorganic and organic nitrogen assimilation pathways are included in this cluster (Additional file 4: Figure S3), including the nitrate transporter NRT2.3 (Cre09.g396000), nitrite transporter NARI1.4 (Cre07.g355600), nitrate reductase NIT1 (Cre09.g410950) and nitrite reductase NIT1 (Cre09.g410750) yielding ammonia as a final product. In fact, these reductases need a molybdenum cofactor and the biosynthetic pathway for molybdenum cofactor constituted by the enzymes molybdopterin synthase adenylyltransferase CNX (g10007), cyclic pyranopterin monophosphate synthase CXN2 (Cre13.g602900), molybdopterin synthase sulfurylase MoaE (Cre07.g322250) and molybdenotransferase MoeA (Cre10.g451400) is entirely included in this cluster (Additional file 4: Figure S3). Therefore, not only the enzymes, but also the pathways leading to the synthesis of the cofactors needed for nitrate assimilation are tightly co-expressed in ChlamyNET.

**Cluster 1, orange - intracellular transport, regulation of GTPase activity, autophagy and proteolysis**

The orange cluster consists of 824 genes and is located in the periphery of ChlamyNET (Fig. 4). This cluster presents a high silhouette value of 0.38 (Fig. 3b). The GO term enrichment analysis indicates that genes within this cluster are significantly involved in processes related to intracellular transport to the endoplasmic reticulum and Golgi apparatus such as vesicle-mediated transport (GO:0016192). For instance, we can find genes coding for the endosomal R-SNARE protein Yky6 (Cre17.g728150) and gammal-Adaptin AP1GI (Cre16.g676650). Genes in this cluster are also significantly related to the regulation of GTPase activity (GO:0043087) such as those coding for the rab GTPase activator protein CGL44 (Cre12.g532600) and Rab/TBC domain protein (Cre07.g315350). Autophagy (GO:0006914) and proteolysis (GO:0006508) are significant GO terms in this cluster with genes coding for the Autophagy related gene 9 ATG9 (Cre09.g391500) and ubiquitin-conjugating enzyme E2 LBC9 (Cre16.g93700). Therefore, the formation of this gene cluster suggests a connection between Rab GTPase activity and autophagy. Moreover, the positive regulation of Rab GTPase activity over autophagy has been shown in Arabidopsis [42].

The metabolic analysis of this cluster suggests that it is involved in triacylglycerol (TAG) biosynthesis, the major lipid reserve in plants. Many unicellular microalgae accumulate large amounts of TAG under unfavorable conditions, such as the ones leading to autophagy [43]. TAG is produced from dicylglycerol (DAG) and different acyl donors. On the one hand, DAG can be synthesized from diacylglycerol (DAG) and different acyl donors. On the one hand, DAG can be synthesized from diacylglycerol (DAG) and different acyl donors. On the other hand, phospholipids (major constituents of cellular membranes) are one of the possible donors for DAG to produce TAG. In this case, the enzyme phospholipid:DAG acyltransferase...
**PDAT** (Cre02.g106400) present in this cluster catalyzes this reaction (Additional file 5: Figure S4). The 3-phosphoinositide biosynthesis pathway is also included in this cluster. Phosphoinositides are involved in phospholipid biosynthesis as well as membrane trafficking, biological processes over-represented in this cluster. The key enzymes in this pathway are phosphatidylinositol-3-kinase PI3K (Cre01.g035500), phosphatidylinositol 4-kinase PIK1 (Cre05.g245550), phosphatidylinositol-4-phosphate 5-kinase PIP5K3 (g9964) and inositol 5-phosphatase SACI (Cre12.g537500) which are also located in this cluster (Additional file 5: Figure S4). Other important lipid metabolic reactions are the activation and deactivation of lipids achieved by the ligation or removal of acyl-CoA. These reactions are catalyzed by the enzymes long-chain-fatty-acid-CoA ligase LACS (Cre03.g182050) and acyl-CoA thioesterase ACOT (Cre01.g037350) respectively, both members of this cluster. In these reactions the common acyl carrier Coenzyme A is required, and so, key enzymes in its biosynthesis such as ketopantoate hydroxymethyltransferase PAN2 (Cre12.g508550), phosphopantothenate-cysteine ligase COAB (Cre01.g048050) and phosphopantothenoylcysteine decarboxylase COAC (Cre10.g423450) are also co-expressed in this cluster (Additional file 5: Figure S4).

**Cluster 3, red - protein phosphorylation, translation, ribosome biogenesis and macromolecule biosynthetic process**

The red cluster expands from the periphery of ChlamyNET to its core (Fig. 4). Somehow this cluster serves as an interface between the blue cluster (hub genes involved in protein phosphorylation, carbohydrate metabolic process and transmembrane transport) and the brown cluster (cell cycle processes). This cluster is the largest one including 1723 genes and presenting a moderate silhouette value of 0.28 (Fig. 3b). According to the GO term enrichment analysis, genes in this cluster are significantly involved in diverse biological processes. The three most significant processes are protein phosphorylation (GO:0006468) including genes such as the mitogen activated protein kinase PTK24 (Cre02.g145500) and aurora-like kinase ALK3 (Cre12.g498650); translation (GO:0006412) and ribosome biogenesis (GO:0042254) with genes coding for ribosomal proteins L13 RPL13a (Cre12.g532550) and S6e RPS6 (Cre09.g400650). The next significant biological process is macromolecule biosynthetic process (GO:0009059) with genes such as the glutamine synthetase GLN4 (Cre03.g207250).

The analysis of the metabolic pathways included in this cluster identified the synthesis of triacylglycerol using exclusively as acyl donors galactolipids produced by glycolipid degradation. The diacylglycerol O-acyltransferase DGAT1 (g9572) and monogalactosyl/diacylglycerol synthase FAD6 (Cre13.g590500) are thus included in this cluster (Additional file 3: Figure S2). Although no other metabolic pathway is fully represented in cluster 3, isolated key enzymes for carbon xation, hydrogen production and oxidation such as rubisco RBCS2 (Cre02.g120150) and iron hydrogenase HYDA2 (Cre09.g396600) are co-expressed within this cluster. In fact, our study suggests that this cluster is involved in signalling and transcription control rather than in metabolism. Several serine/threonine protein kinases are included in this cluster. The genes CrAUR1 (Cre16.g669800) and ALK3 (Cre12.g498650) exhibit a high sequence similarity with the and Aurora kinases in Arabidopsis AUR1 (At4g32830) and AUR3 (At2g43590) respectively. It has been described that the diversifcation of plants and aurora kinases predates the origin of land plants [44]. Here we show that this diversification may be already present in Chlamydomonas. These kinases have been shown to play a key role in cell cycle related signal transduction pathways in Arabidopsis. Several other genes similar to cyclin-dependent protein kinases are located in this cluster such as CDK11 (Cre12.g494500) and CrMAPKKK (Cre10.g461150). Cyclin-dependent protein kinases play crucial roles in the progression of the cell cycle in eukaryotes. CDK11 (Cre12.g494500) exhibits a high sequence similarity with the Arabidopsis gene CAK4 (At1g66750), which is known to be involved in the activation of cell proliferation [45]. While CrMAPKKK (Cre10.g461150) is highly similar to the Arabidopsis gene MEKK1 (At4g98500). Additionally, other genes in this cluster such as g16721, present a high similarity with the Arabidopsis Mitogen Activated Protein (MAP) kinase MAPKKK6 (At3g07980). The co-expression of these genes suggests that MAP kinase cascades are regulated not only at the posttranslational level but also at the transcriptional level in Chlamydomonas.

As it will be described in detail in the next section, this cluster is also significantly enriched in transcription factors. Several GATA transcription factors such as g7394, Cre05.g242600 and Cre08.g378800; bZIP transcription factors like Cre10.g438850 and Cre12.g498900 and the single DOF and CO-like transcription factors in Chlamydomonas CrDOF (Cre12.g521150) [46] and CrCO (g6302) [2] are members of this cluster.

A detailed description of the rest of clusters and their functional annotation is available for further exploration at the web page http://viridiplantae.ibbf.csic.es/ChlamyNet/. These results aim at providing researchers in the functional annotation of the Chlamydomonas transcriptome with a solid ground to design specific and targeted experimental studies to validate or refute the predictions produced in this clustering analysis.

**Transcription factors and transcriptional regulators in ChlamyNET**

In the previous section we performed a functional annotation of the different gene clusters identified according
to GO term enrichment and metabolic pathways analysis. In this section, we further investigate the regulatory aspects of the *Chlamydomonas* transcriptome using ChlamyNET.

One of the most important processes involved in cellular response to internal and external stimuli is transcription or gene expression. This is a highly regulated process carried out by transcription factors and transcriptional regulators. Transcription factors (TFs) bind to specific *cis*-elements in the promoters of genes to activate or repress their transcription directly. On the other hand, transcriptional regulators (TRs) are involved in gene expression control but do not bind directly to gene promoters. Transcriptional regulators modulate gene expression by interacting with transcription factors, remodeling chromatin or other indirect mechanisms [47]. A genome-wide identification and classification of transcription factors and transcriptional regulators in *Chlamydomonas* has been previously performed [48]. This classification is available on the web portals and databases PlantTFDB (http://planttfdb.cbi.pku.edu.cn/index.php?sp=Cre) [49] and PlntTFDB (http://plntfdb.bio.uni-potsdam.de/v3.0/index.php?sp id = CRE) [50]. Using this classification, we identified 118 TFs and 109 TRs in ChlamyNET, which constitute 2.48% of the total number of genes in the network (Fig. 5). These TFs and TRs are classified, according to their protein domains, into 28 and 17 gene families respectively. The TFs seem not to be randomly distributed over the clusters of ChlamyNET (Fig. 5a), whereas the TRs distribution seems to be more uniformly distributed over the net (Fig. 5b). In order to assess the statistical significance of the distribution of TFs and TRs over the clusters in ChlamyNET, we performed an enrichment analysis based on Fisher’s exact test. Indeed, no cluster was significantly enriched in TRs whereas the blue and red clusters were significantly enriched in TFs with p-values of $2.62 \times 10^{-3}$ and $2.37 \times 10^{-3}$ respectively. This complements the evidence produced by our analysis of the location of authoritative hub genes in ChlamyNET and about the key role played by the blue and red clusters in the regulation of the *Chlamydomonas* transcriptome under the conditions analyzed in this study.

In our analysis, instead of classifying genes according to their sequence as previously described [48], we have studied their co-expression patterns in order to determine groups of TFs and TRs that could exert their function over target genes in a coordinated way. Similar to the previous section we applied the two most widely used clustering algorithms, hierarchical clustering and partition around medoids, taking as distance among genes the correlation between their expression profiles. According to the silhouette criterion, the best description of the co-expression patterns among TFs and TRs is obtained with thirteen different groups identified with the partition around medoids algorithm (Fig. 6).

In order to determine the biological processes regulated by each group we applied GO term enrichment analysis over the genes directly linked to the corresponding TFs and TRs. Additionally, we performed a transcriptional factor binding site (TFBS) enrichment analysis over the promoters of these genes as described in the Methods section. In Tables 3 and 4 we show a summary of the results about the GO terms and TFBS significantly enriched in the genes directly linked to the TFs and TRs in each group.
Herein we present in detail the results of our analysis over three groups of TFs and TRs of special interest. We discuss the conservation of their function and binding sites when compared to their putative orthologs in higher plants. The results for the remaining groups of TFs and TRs identified in our analysis are available at the web page http://viridiplantae.ibvf.csic.es/ChlamyNet/.

Core metabolic regulation, group a

This constitutes a numerous group of TFs and TRs including 38 members. They are identified in Fig. 6 using green triangles. The TFs and TRs in this group are included in the cluster 9 (blue) at the center of the network where most authoritative hub genes and carbon/nitrogen core metabolic pathways are located. These TFs and TRs seem to be of key importance in the regulation of the Chlamydomonas transcriptome under the conditions of our study since they are co-expressed on average with 87.97 other genes. Some highly authoritative hub genes in ChlamyNET are members of this group such as the B-box TF CrBbox1 (Cre03.g182700), the bHLH TFs g4643 and Cre01.g011150, the SBP TF Cre16.g673250, the RWP-RK TF NIT2 (Cre03.g177700) and the MYB TF Cre03.g197100. These TFs present a normalized authoritative hub score higher than 0.8. GO term enrichment analysis over the genes directly linked to the TFs and TRs in this group suggests that they are mainly involved in core metabolism regulation and light response. Several GO terms related with metabolic processes are significantly enriched such as carbohydrate metabolic process, fatty acid biosynthetic process and nitrogen compound metabolic process. Representative genes in this group are the alpha-amylase AMA2 (Cre08.g362450), the long-chain acyl-CoA synthetase LACS2 (Cre13.g566650) and the nitrite reductase NII1 (Cre09.g410750), respectively.

Four bHLH transcription factors, Cre01.g011150, Cre14.g620850, g4643 and g4645, out of the 12 recognized members of this family in Chlamydomonas, are members of this group. Only bHLH Cre14.g620850 has similarity with genes present in higher plants. Specifically, its putative Arabidopsis ortholog is PAR1 (At1g69010) that has been shown to be involved in light response [51]. The rest show similarity with other bHLH genes present only in chlorophyceae. A bHLH binding site was found to be significantly present over the genes co-expressed with the TFs and TRs of this group (Table 3). This suggests that the binding site of bHLH TFs is conserved across the green lineage. Several genes involved in carbohydrate and nitrogen metabolism contain this binding site in their promoters, for instance the glucose-6-phosphate dehydrogenase GLD2 (Cre08.g378150) and the ammonium transporter AMT4 (Cre13.g569850).

Three bZIP TFs out of the 19 identified in the Chlamydomonas genome, Cre10.g454850, Cre12.g510200 and Cre06.g310500, are members of this group. Genes CrHY5 (Cre12.g510200) and CrHYH (Cre06.g310500) present a high similarity with the Arabidopsis genes HY5 (At5g11260) and HYH (At3g17609) respectively. These
TFs are known to bind to G-box sequences to regulate light response and metabolism in Arabidopsis [52, 53]. GO term and TFBS enrichment analysis suggest that this mechanism is already present in Chlamydomonas, since a sequence highly similar to the G-box has been found to be significantly present in the genes co-expressed with these two Chlamydomonas genes (Table 3).

Table 3

| Group a - Green Triangles | Group b - Green Squares | Group c - Yellow Triangle | Group d - Blue Squares | Group e - Red Circles | Group f - Yellow Circles |
|---------------------------|-------------------------|---------------------------|-----------------------|-----------------------|-------------------------|
| g2885 (AP2-BRBSP)         | Cre13.g572450 (ARR-B)   | Cre10.g467840 (ARR-B)    | Cre10.g467840 (ARR-B) | Cre13.g572450 (ARR-B) | Cre02.g67840 (ARR-B)    |
| g1637 (ARR-B)             | Cre14.g679840 (MYB)     | Cre14.g679840 (MYB)      | Cre14.g679840 (MYB)   | Cre14.g679840 (MYB)   | Cre02.g67840 (MYB)      |
| g9099 (MYB)               | Cre01.g808050 (MYB)     | Cre01.g808050 (MYB)      | Cre01.g808050 (MYB)   | Cre01.g808050 (MYB)   | Cre01.g808050 (MYB)     |
| Cre03.g1486400 (RWP-RK)   | Cre03.g177000 (RWP-RK)  | Cre14.g679840 (MYB)      | Cre14.g679840 (MYB)   | Cre14.g679840 (MYB)   | Cre02.g67840 (MYB)      |
| Cre04.g121100 (RWP-RK)    | Cre16.g679840 (MYB)     | Cre14.g679840 (MYB)      | Cre14.g679840 (MYB)   | Cre14.g679840 (MYB)   | Cre02.g67840 (MYB)      |
| g6463 (MYB)               | Cre10.g467840 (MYB)     | Cre14.g679840 (MYB)      | Cre14.g679840 (MYB)   | Cre14.g679840 (MYB)   | Cre02.g67840 (MYB)      |
| Cre13.g176500 (MYB)       | Cre14.g679840 (MYB)     | Cre14.g679840 (MYB)      | Cre14.g679840 (MYB)   | Cre14.g679840 (MYB)   | Cre02.g67840 (MYB)      |
| Cre12.g102000 (MYB)       | Cre03.g177000 (RWP-RK)  | Cre14.g679840 (MYB)      | Cre14.g679840 (MYB)   | Cre14.g679840 (MYB)   | Cre02.g67840 (MYB)      |
| g4865 (TAF)               | Cre12.g171500 (ARRD)    | Cre14.g679840 (MYB)      | Cre14.g679840 (MYB)   | Cre14.g679840 (MYB)   | Cre02.g67840 (MYB)      |
| Cre12.g585900 (HMG)       | Cre16.g679840 (MYB)     | Cre14.g679840 (MYB)      | Cre14.g679840 (MYB)   | Cre14.g679840 (MYB)   | Cre02.g67840 (MYB)      |
| Cre13.g466800 (PI)        | Cre14.g679840 (MYB)     | Cre14.g679840 (MYB)      | Cre14.g679840 (MYB)   | Cre14.g679840 (MYB)   | Cre02.g67840 (MYB)      |
| g1628 (PI)                | Cre14.g679840 (MYB)     | Cre14.g679840 (MYB)      | Cre14.g679840 (MYB)   | Cre14.g679840 (MYB)   | Cre02.g67840 (MYB)      |
| g2763 (PHD)               | Cre14.g679840 (MYB)     | Cre14.g679840 (MYB)      | Cre14.g679840 (MYB)   | Cre14.g679840 (MYB)   | Cre02.g67840 (MYB)      |
| Cre13.g521200 (CCT)       | Cre02.g67840 (MYB)      | Cre14.g679840 (MYB)      | Cre14.g679840 (MYB)   | Cre14.g679840 (MYB)   | Cre02.g67840 (MYB)      |

Cre13.g572450 and g16739 that code for two ARR-B TFs and CrBBx1 (Cre03.g182700), that codes for a B-box TF, are present in this group. These genes exhibit high similarities with the Arabidopsis genes RR14 (At2g01760), TOCI (At5g61380) and COL1 (At5g15850), respectively. They have in common a CCT domain at the carboxyl end that directly binds to DNA [54] that...
was found to be present in the *CrCO* (g6302) gene [2]. These genes are known to be involved in light response and circadian rhythms in *Arabidopsis* [55, 56]. These functions seem to have been established already in *Chlamydomonas* constituting a link between circadian rhythms and metabolism.

Five MYB TFs are present in this group. Some of them such as *Cre14.g633789* and *Cre03.g198800* are putative orthologs of the *Arabidopsis* genes *At3g27785* and *At5g61620* that have been associated with metabolic regulation [57]. MYB TF factor binding sites have been found significantly enriched in the promoters of genes co-expressed with this group of TFs and TRs. Such as, the triacylglycerol lipase *CrTLL1* (*Cre03.g193500*) and starch phosphorylase *CrPHS1* (*Cre07.g336950*) that present sequences highly similar to MYB binding sites in their promoters (Table 3).

Finally, several genes coding for TFs from the RWP-RK family are members of this group. One of these TFs, *NIT2* (*Cre03.g177700*), has already been shown to be involved in nitrogen and carbohydrate metabolism regulation [58, 59], whereas the other remain to be studied.

### Table 4

| Transcription factors and Transcriptional Regulators | Significant GO term | Representative Genes | Significant Promoter Sequence |
|-----------------------------------------------------|---------------------|-----------------------|-------------------------------|
| Group 1 - Blue Circles:                            |                     |                       |                               |
| g638 (MYB)                                          | story modification   | *Cre03.g198800* (MYB) | Similar to NIT2 binding sites |
| g639 (MYB)                                          | DNA-dependent RNA   | *Cre03.g198800* (MYB) | Similar to NIT2 binding sites |
|                     | family              | *Cre03.g198800* (MYB) |                               |

### Table 5

| Biological processes and transcription binding sites significantly over-represented in the neighbourhood of the TFs and TRs groups in ChlamyNET | Romero-Campero et al. BMC Genomics (2016) 17:227 Page 14 of 28 |                               |                               |
|----------------------------------------------------------------------|-----------------------------------------------------------------|                               |                               |
that have been shown to be involved in nitrogen and light response [60, 61].

Not surprisingly, TFs in this group seem to constitute an intricate gene regulatory system with mutual regulations among them. For example, bHLH binding sites can be identified in the promoters of the B-box TF CreBox1 (Cre03.g182700), the bZIP TF CrHY5 (Cre12.g510200), the bHLH TF Cre01.g011150 and the MYB TFs Cre03.g198800 and Cre14.g621050. In turn, G-boxes have been found in the promoters of the bHLH genes Cre01.g011150 and Cre14.g620850 and the bZIP gene Cre10.g454850. Additionally, these TFs seem to exert their regulation in a coordinated manner over the same set of genes since both bHLH and MYB binding sites have been identified in the promoters of genes such as the nitrate transporter NRT2.3 (Cre09.g396000) and the nitrate reductase NIT1 (Cre09.g410950). Such complex interactions are also common in Arabidopsis.

**Autophagy regulation, group b**

The TFs and TRs in this group are located in the cluster 1 (orange) identified with green squares in Fig. 6. A GO term analysis of the genes directly linked to them reveals a potential regulation over processes involved in vesicle mediated transport, catabolic process, proteolysis and autophagy. In this group we can find the C3H zinc finger TF g8693 presenting a high sequence similarity with the INOSITOL-REQUIRING ENZYME-1b gene (At5g24360) from Arabidopsis. This gene is involved in the regulation of the degradation of the endoplasmic reticulum by autophagy [62]. Directly linked to this gene we can find genes involved in autophagy such as autophagy 9 ATG9 (Cre09.g391500) and proteolysis such as signal peptide peptidase-like 2 (g18126). The GATA transcription factor Cre10.g435450 is also a member of this group and its putative ortholog in Arabidopsis, BME3 (At3g54810), has been shown to be involved in response to salt stress [63]. The MYB transcription factor Cre16.g695600 is also a member of this group whose Arabidopsis putative ortholog At5g06110, is a heat shock protein involved in stress response [64]. Two genes from the chromatin remodeling family SNF2, Cre06.g287950 and Cre06.g270850, are putative orthologs of ATRX (At1g08600) and CHR8 (At2g18760), involved in DNA damage response and recombination [65]. The induction of autophagy as a response to diverse stresses has been shown in Chlamydomonas [66]. Cre03.g173165, Cre03.g174150 and g5052 are transcriptional regulators from the BTB/POZ family that are putative orthologs of ARIA (At5g19330) involved in cellular macromolecule catabolic process [67].

In fact, the OBP binding site was found to be significantly present in the promoters of the genes directly linked to the TFs and TRs in this group (Table 3). This binding site has been shown to be present in promoters of genes induced by oxidative stress in Arabidopsis [68]. This is in agreement with the reported autophagy induction by oxidative stress in Chlamydomonas [66]. Genes related to autophagy such as ATG8 (Cre16.g689650) and ATG9 (Cre09.g391500) present the OBP binding site in their promoters. Genes involved in vesicle trafficking such as Component of the Exocyst Complex SECl8 (Cre01.g003050) and Subunit f the ESCRT-I complex VPS28 (Cre16.g678100) also present the OBP binding site in their promoters (Table 3).

**Cell cycle regulation, group c**

The TFs and TRs of this group are included in the cluster 2 (brown) identified in the previous section as involved in DNA replication, chromosome organization and regulation of cell cycle. These TFs and TRs are highlighted using yellow triangles in Fig. 6. A GO term enrichment analysis over the genes directly linked to these TFs and TRs confirmed their potential regulation over these processes. In this group we can find a MYB3R TF Cre12.g522400 whose putative orthologs, based on their sequence similarity, are At5g11510 and At4g32730 in Arabidopsis and NtmybA1 and NtmybA2 in Nicotiana tabacum. These genes are involved in the G2/M transition during the cell cycle [69–71]. The single member of the YABBY family in Chlamydomonas that presents two high mobility group boxes, Cre16.g672300, belongs to this group. Its putative ortholog gene At4g11080 in Arabidopsis interacts with mitotic and meiotic chromosomes [72]. Another gene, ORCl (g11180) belonging to the PHD TF family is also a member of this group. Its putative Arabidopsis ortholog At4g14700 (Origin recognition complex) has been shown to be in the core cell cycle machinery involved in the G1/S transition [73, 74]. Several TRs potentially involved in chromatin remodeling are present in this group such as Cre03.g197700 that code for SET domain containing protein that exhibits a high sequence similarity with At1g05830 a trithorax protein in Arabidopsis [75]. The rest of TFs in this group, Cre09.g402350 and Cre12.g510200, are putative orthologs of Arabidopsis genes that have been shown to be co-transcribed with other core cell cycle regulators and TFs in Arabidopsis [76].

The E2F motif [77] was found to be the only known motif significantly enriched in the promoters of the genes directly linked to the TFs and TRs in this group (Table 3). The potential orthologs of the genes that contain in their promoters the E2F motif sequence are involved in the G1/S transition such as subunits of the origin of replication complex ORCl (g11180) and ORCA4 (Cre17.g726500), pre-initiation complex subunit CDC6 (Cre06.g292850), DNA replication initiation factor CDT1 (Cre03.g163300),
minichromosome maintenance protein MCM2 (Cre07.g338000) and DNA polymerase alpha POLA1 (Cre04.g214350) (Table 3). The presence of the E2F motif in genes regulating the S phase has been shown previously in Arabidopsis [78] and Nicotiana [77]. The gene Cre07.g323000, putative ortholog of the Arabidopsis E2F transcription factor, is not included in this group of TFs and TRs. Nevertheless, it is located in its vicinity, suggesting that it may function as an interface between regulation of cell cycle and other processes as it is the case for its Arabidopsis ortholog [74]. The two most significant de-novo motifs found in our study presents a high similarity with the octamer and hexamer motifs. The combination of these two motifs has been shown to confer S phase-specific transcriptional activation in plants [79]. Genes containing these motifs include B-type cyclin CYCB1 (Cre08.g370400) and cell division cycle protein CDC45 (Cre06.g270250). This suggests a remarkable conservation of cell cycle regulation in the plant kingdom not only limited to the TFs, TRs and their targets involved in this process but also in the cis-regulatory elements, TFBS, present in their promoters.

**Light-regulated transcription factors and transcriptional regulators in ChlamyNET, a tutorial for ChlamyNET usage**

In order to ensure the reproducibility of the results presented in this work and to facilitate further and independent studies over the Chlamydomonas transcriptome we have developed a web-based software tool also called ChlamyNET. This tool is based on WiGis, a platform for the visualization of large-scale, highly interactive graphs in a user’s web browser [80]. The software tool ChlamyNET is available from the web page http://viridiplantaie.ibvf.csic.es/ChlamyNet/. In this section we discuss a case study concerning the Chlamydomonas potentially light-regulated TFs and TRs that can be used as a tutorial for the use of ChlamyNET.

Light constitutes the most important source of energy for green algae. Therefore, light is a key environmental signal that affects profoundly Chlamydomonas growth and physiology. The study of the light-regulated TFs and TRs co-expression patterns, the biological processes controlled by them and the TFBS where they potentially exert their function can contribute to a better understanding of the response to this key environmental signal in Chlamydomonas. Previously, potential orthologs of

---

**Fig. 7** Potentially Light-regulated Transcription Factors and Transcriptional Regulators in ChlamyNET. Twenty-one TFs and TRs exhibiting a high similarity with light regulated TFs and TRs in Arabidopsis were identified in ChlamyNET. These genes are not uniformly distributed over ChlamyNET. Clusters 9 (blue), 3 (red) and 7 (green) were significantly enriched in these potentially light-regulated TFs and TRs, so that they are expected to be involved mainly in carbon/nitrogen metabolism, signalling by phosphorylation and protein degradation. The central location of several light-regulated TFs and TRs such as CHYS (Cre12510200) and CrCRY1 (Cre06.g295200) suggests that they are highly authoritative hub genes. Indeed CHYS and CrCRY1 have 133 and 57 neighbouring genes respectively.
the light-regulated TFs and TRs in Arabidopsis [81] have been identified in the Chlamydomonas genome [48]. Most of these TFs and TRs can be found in ChlamyNET. Surprisingly, these genes are not randomly distributed over ChlamyNET (Fig. 7). They are mainly located in clusters 9 (blue), 3 (red) and 7 (green) suggesting that they are specifically involved in carbon/nitrogen metabolism, signalling by phosphorylation and protein degradation. In order to identify genes in the network, the Search panel on the left of ChlamyNET can be used. Once the genes are found, they can be selected and information related to their name, putative Arabidopsis ortholog, topological indexes, protein domains (Pfam annotation) and neighbours is depicted on the Selected Gene Details panel. This information for light-regulated TFs and TRs in ChlamyNET is collected into Table 5.

According to this information several light-regulated TFs are highly authoritative hub genes in ChlamyNET such as CrGBF1 (Cre01.043150) and CrHY5 (Cre12.510200) that are co-expressed with more than 130 genes. These genes are involved in photomorphogenesis in Arabidopsis, yet their function in Chlamydomonas is unknown. Others light-regulated TFs and TRs that constitute hub genes that are co-expressed with more than 50 genes, are CrCRY1 (Cre06.g295200), CrCO (g6302), CrLHY (Cre06.g275350) and the different subunits of the 26S proteasome CrRPN12A (Cre17.g708300), CrRPT5A (Cre10.g439150) and CrRPN7 (Cre13.g581450). CrCRY1, also known as CPH1, codes for a putative ortholog of CRY1 in Arabidopsis and it is a well known photoreceptor that responds to light stimulus [82]. On the other hand, CrCO expression is affected by photoperiod and regulates carbon metabolism and cell cycle progression [2]. Silencing and overexpression of these genes have been shown to massively disrupt Chlamydomonas cell growth and proliferation supporting their function as hubs in the network [2]. The potential role of CrLHY in circadian rythms and the proteolytic function of CrRPN12A, CrRPT5A and CrRPN7 are yet to be tested experimentally.

Table 5 Potentially Light Regulated TFs and TRs in ChlamyNET. Their potential Arabidopsis ortholog and topological indexes are indicated as well

| Chlamydomonas gene | Putative Arabidopsis Ortholog | Number of neighbours | Normalized hub score | Clusteringcoefficient |
|--------------------|-------------------------------|----------------------|----------------------|-----------------------|
| Cre06.g295200 CPH1 / CrCRY1 | At4g08920 CRYPTOCHROME 1 | 57 | 8.12 x 10^-5 | 0.39 |
| Cre01.g043150 CrGBF1 | At4g36730 G-BOX BINDING FACTOR 1 | 182 | 7.36 x 10^-7 | 0.31 |
| Cre12.g510200CrHY5 | At5g11260 ELONGATED HYPOCOTYL 5 | 133 | 0.32 | 0.47 |
| Cre06.g310500CrHYH | At3g17609 HY5-HOMOLOG | 39 | 0.12 | 0.71 |
| Cre12.g521150GDOF | At5g39660 CYCLING DOF FACTOR 2 | 27 | 3.79 x 10^-8 | 0.34 |
| Cr3102CcCO | At5g15840 CONSTANS | 58 | 2.29 x 10^-4 | 0.40 |
| Cre02.g094150CrPBR9 | At2g46790 PSEUDO-RESPONSE REGULATOR 9 | 7 | 7.72 x 10^-7 | 0.0 |
| Cre06.g275350CrLHY | At1g01060 LATE ELONGATED HYPOCOTYL | 78 | 5.86 x 10^-7 | 0.38 |
| Cr1542CrLUX | At3g46640 LUX | 38 | 2.23 x 10^-7 | 0.40 |
| Cre06.g277350CHD1 | At4g38130 HISTONE DEACETYLASE 1 | 1 | 5.12 x 10^-18 | 0.0 |
| Cr1673CrTOC1 | At5g1380 TIMING OF CAB EXPRESSION 1 | 3 | 1.49 x 10^-5 | 0.33 |
| Cre14.g617350CrUV8 | At5g63860 UVB-RESISTANCE 8 | 1 | 4.09 x 10^-14 | 0.0 |
| Cre05.g23400CrCOP11 | At3g61140 CONSTITUTIVE PHOTOMORPHOGENIC 11 | 36 | 7.81 x 10^-14 | 0.37 |
| Cre14.g60880CrCOP9 | At4g14110 CONSTITUTIVE PHOTOMORPHOGENIC 9 | 11 | 1.26 x 10^-12 | 0.27 |
| Cre17.g708300CrRPN12A | At1g64520 REGULATORY PARTICLE NON-ATPASE 12A | 57 | 7.44 x 10^-14 | 0.39 |
| Cre10.g439150CrRPT5A | At3g05330 REGULATORY PARTICLE TRIPLE-A ATPASE 5A | 51 | 7.49 x 10^-14 | 0.45 |
| Cre13.g581450CrRPN7 | At4g24820 REGULATORY PARTICLE NON-ATPASE 7 | 59 | 1.52 x 10^-13 | 0.28 |
| Cre06.g275650CrRPN3A | At1g20200 REGULATORY PARTICLE NON-ATPASE 3A | 14 | 2.80 x 10^-15 | 0.41 |
| Cre16.g663500CrRPN10 | At4g38630 REGULATORY PARTICLE NON-ATPASE 10 | 4 | 3.48 x 10^-14 | 0.0 |
| Cre04.g216600CrRPT6A | At5g19990 REGULATORY PARTICLE TRIPLE-A ATPASE 6A | 19 | 1.76 x 10^-12 | 0.25 |
| Cre07.g329700CrRPT2A | At4g29040 REGULATORY PARTICLE TRIPLE-A ATPASE 2A | 3 | 3.44 x 10^-16 | 0.33 |
potentially light-regulated genes CrHYH (Cre06.g310500), CrDOF (Cre12.g521150), CrLUX (g1542) and CrCOP11 (Cre05.g234300) whose putative Arabidopsis orthologs are involved in photomorphogenesis, photoperiod response, circadian rhythms and protein degradation respectively are co-expressed with around 30 other genes. Recently, CrDOF expression has been shown to be influenced by circadian rhythms and the photoperiod whereas it directly regulates the expression of CrCO [46]. The rest of potentially light-regulated TFs and TRs identified in ChlamyNET are co-expressed with fewer than 20 other genes and are not considered hubs in the network. Most of these genes exhibit a high clustering coefficient in ChlamyNET suggesting a high level of coordination among their co-expressed genes.

In order to study co-expression patterns among a set of selected genes we can generate heatmaps using ChlamyNET. In this type of graphs we represent the correlation among expression profiles of a selected set of genes. Heatmaps can be generated in ChlamyNET using the Analysis section located in the Search panel once a set of genes have been selected. Additionally, genes at distance one, two or three from the selected genes can be included in the heatmap. As an example, a heatmap depicting the co-expression patterns among the potentially light-regulated TFs and TRs previously selected is presented in Fig. 8. We can observe three different groups. Genes in the same group exhibit a high positive correlation (red/yellow colours) whereas genes from different groups present a very low negative correlation (blue/purple colours). In group 1 we can distinguish two subgroups. On the one hand, in subgroup 1a it can be observed that CrCRY1, CrPRR9, CrCO, CrHY5 and CrHYH are highly co-expressed. This could indicate coregulation of these genes by the same factors or regulation among themselves. On the other hand, in subgroup 1b we find that CrLUX, CrLHY, CrGBF1 and CrDOF have a very high co-expression value. These two subgroups are also co-expressed and hence form a single group. Group 2 is composed of three genes, CrLVR8, CrTOC1 and CrHD1. This group of genes presents a negative correlation with genes in subgroup 1b which makes them constitute a separate group. For instance, CrLHY and CrTOC1 are negatively co-expressed as expected if they are true orthologs of the Arabidopsis circadian genes LHY/CCA1 and TOC1. These genes form a negative feedback loop that constitutes the core of the circadian clock in Arabidopsis. Group 3 exhibits a very low negative correlation with group 1 and a moderate negative correlation with group 2. The genes in this group are different members of the 26S proteasome and signalosome subunits such as CrRPN7 and CrCOP9.

Fig. 8 Heatmap Representing the Co-expression Patterns among the Potentially Light-regulated TFs and TRs in ChlamyNET. High positive correlation between the corresponding gene profiles is represented by red/yellow colours, low negative correlation is represented by blue/purple colours. Three different groups are apparent. The first group can be divided into two subgroups. We can observe negative correlations between genes in the subgroup 1b and genes in the second group such as between CrLHY and CrTOC1 which indicates that these two genes may be true orthologs of the circadian clock Arabidopsis genes LHY/CCA1 and TOC1. Very low negative correlations are observed between genes in Group 1 and genes in Group 3. Genes coding for different 26S proteasome and signalosome subunits such as CrRPN7 and CrCOP9 can be found in Group 3. Their putative Arabidopsis orthologs have been described to degrade proteins involved in light response that exhibit a high sequence similarity with those coded by genes in Group 1.
Their putative *Arabidopsis* orthologs have been described to degrade proteins involved in light response [61]. This suggests an antagonist expression pattern between light-regulated TFs and different subunits of the 26S proteasome and signalosome that is already established in *Chlamydomonas*. This would need further research and experimental validation, but can be an important clue for a preliminary investigation.

The biological processes potentially controlled by the light-regulated TFs and TRs in ChlamyNET can be deduced by applying GO terms enrichment over their co-expressed genes. This can be performed by using the Analysis section located in the Search panel once the neighbouring genes have been selected. As described in the Methods section we can combine the significative GO term identified, based on orthology, with those determined based on conserved protein domains (Additional file 6: Table S2). According to this methodology, the potentially light-regulated TFs and TRs are co-expressed with genes involved in ion transport, for example the nitrate transporter NRT2.2 (Cre09.g410800) and Mo-molybdopterin cofactor biosynthesis such as MoeA (Cre10.g451400), which produces essential cofactors for the nitrate reductase, a key enzyme in the nitrate metabolism. Additionally, carbohydrate metabolism appears as a significative GO term, including genes involved in starch and glucose degradation such as the starch phosphorylase CrPHS2 (Cre12.g552200), the alpha-amylose AMY (Cre08.g384750) and the glucose-6-phosphate dehydrogenase GLD2 (Cre08.g378150). Finally, protein phosphorylation is another relevant significative GO term with genes potentially involved in cell cycle control such as the mitogen-activated kinase kinase kinases CrMKKK1 (Cre07.g347000) and CrMKKK2 (Cre02.g108650). Therefore, this analysis suggests a potential regulation of carbon/nitrogen metabolism and cell cycle through protein phosphorylation by these potentially light-regulated TFs and TRs that needs to be experimentally validated.

Finally, selecting the switch Promoter Sequence Enrichment in the Search panel, a transcriptional factor binding site (TFBS) enrichment analysis over the promoters of selected genes can be performed. In this case, several significative light-regulated TFBS in *Arabidopsis* such as SORLIP2, SORLIP3 and SORLREP5 [83] were identified. For example, the CrGBF1, CrHYH and CrLUX genes present the sequence SORLIP3 in their promoters. This is in agreement with their high co-expression values (Fig. 8). The presence of these TFBS in the promoters of the genes studied here suggests a high conservation of light regulated TFBS across the green lineage.

A more detailed presentation of this case study is available from the web page of ChlamyNET, http://viridi-plantae.ibv.csic.es/ChlamyNET/.

**Experimental validation**

ChlamyNET aims at becoming an enabling technology for researchers on the *Chlamydomonas* transcriptome. For example, ChlamyNET can be used to predict changes in gene expression. When a specific gene is mutated or overexpressed, ChlamyNET predicts that the expression of genes located in the neighbourhood of the mutated or overexpressed gene will be affected, whereas genes in distant regions will not substantially change their expression profile. Specifically, another application of ChlamyNET is predicting targets of specific TFs. TFs and their targets tend to be strongly co-expressed since TFs directly regulate the expression of their target genes. Therefore, the targets of a TF should be contained in its neighbourhood, possibly directly linked to it. Additionally, when the sequences recognised by a TF are well characterized, the identification of these sequences in the promoters of genes co-expressed with it provides more convincing evidence for these genes being direct targets of the corresponding TF. In this way, the analysis of gene neighbourhoods and the significance of TFBS in gene promoters can be studied using ChlamyNET constituting a powerful tool for gene expression analysis.

In this section, we present an experimental validation of the predictive power of ChlamyNET based on one of the most widely used methodologies, cross-validation. In this methodology, once a predictor has been constructed, an independent data set is used to evaluate its performance. As it is common in the study of *Chlamydomonas*, the data used to construct ChlamyNET was obtained from algae grown on continuous light (LL) conditions without a dark period. In spite of this, probably due to the fact that the different experiments were carried out with different light intensities, we were able to identify patterns of co-expression between TFs and TRs that potentially respond to light stimuli as discussed in the previous section. Additionally, none of the genotypes used in the construction of ChlamyNET consisted of an overexpressor of a gene involved in photoperiod response. Therefore, in order to use a totally independent data set from those used to construct ChlamyNET we generated RNA-seq data from algae grown in long day (LD) conditions (16 hours of light and 8 hours of dark) and in short day (SD) conditions (8 hours of light and 16 hours of dark). Two different genotypes were analyzed. As wild type we used the *Chlamydomonas* strain CW15 and for a strain affecting the response to photoperiod we chose algae that overexpressed the *CrDOF* gene under a nitrate inducible promoter. As discussed in the previous section the gene *CrDOF* (Cre12.g521150) is a potentially light-regulated transcription factor that has recently been shown to respond to photoperiod and circadian rhythms [46]. This last *Chlamydomonas* strain was called *CrDOF*Fin. For more details on algae material, growth conditions and RNA-seq see the Methods.
section. Using the RNA-seq data analysis protocol described in the Methods section we determined the gene expression level fold-change when comparing CrDOF to CW15 (Fig. 9). According to the predictions provided by ChlamyNET the 216 genes in the neighbourhood at distance two from the CrDOF gene, yellow dots in Fig. 9a, are expected to increase their expression. Indeed, 69.44% of these CrDOF neighbouring genes increased their expression level in the CrDOF strain when compared to the wild type strain CW15 in LD conditions. Actually, the CrDOF neighbouring genes showed an average fold-change increase of 2.7 which is significantly higher than the fold-change in the rest of ChlamyNET with a p-value of $5.63 \times 10^{-3}$. Additionally, we identified in ChlamyNET those genes that increased their expression level a fold-change of four in the genotype CrDOF when compared to the wild type CW15 in LD and SD conditions (Fig. 9b and c). The neighbourhood of the gene CrDOF at distance two was shown to be significantly enriched with a p-value of 0.029 in genes that increase their expression level a fold-change of 4 in LD conditions. This further supports the predictive power of ChlamyNET in LD conditions (Fig. 9b). Nevertheless, the highly activated genes in SD conditions appeared scattered all over ChlamyNET without concentrating in the CrDOF neighbourhood at distance two (Fig. 9c). This shows a limitation in the predictive power of ChlamyNET. It should be recalled that all data collected for the construction of ChlamyNET correspond to LL conditions. The data from LD conditions is somehow similar to these data and therefore, ChlamyNET was a good predictor for this condition. On the other hand, SD conditions, in which the dark period is longer than the light period, represent very different conditions from those

**Fig. 9** Experimental Cross-validation of the Predictive Power of ChlamyNET using RNA-seq Data from Algae Overexpressing the CrDOF gene. (a) The CrDOF gene (identified as a green diamond in ChlamyNET) has a neighbourhood at distance two consisting of 216 genes represented in yellow. These genes showed an average fold-change increase of 2.7 which is significantly higher than the fold-change in the rest of ChlamyNET with a p-value of $5.63 \times 10^{-3}$. (b) Genes increasing their expression level in LD conditions at least by a four fold-change in the CrDOF genotype when compared to the wild type CW15 are represented in red. Note that the neighbourhood of the CrDOF gene, represented in green, is enriched in these genes according to a p-value of 0.029 obtained using Fisher’s exact test. (c) Genes increasing their expression level in SD conditions at least by a four fold-change in the CrDOF genotype when compared to the wild type CW15 are represented in red. These genes tend to group around the CrDOF gene, represented in green. (d) Inhibited genes in LD conditions in the CrDOF genotype when compared to the wild type CW15 with at least a four fold-change are represented in blue. Note that cluster 2 (brown) involved in DNA replication and cell cycle processes is significantly enriched in these genes.
used for the data with which ChlamyNET was generated. Therefore, ChlamyNET performs poorly as a predictor for this condition.

We also identified in ChlamyNET those genes that showed a 4-fold decrease in expression level in CrDOFfin compared to the wild type CW15 in LD conditions (Fig. 9d). These genes were not located in the neighborhood of CrDOF, suggesting that it acts as a direct activator and, possibly, as an indirect repressor in LD conditions. Instead, cluster 2 (brown) was significantly enriched with these highly inhibited genes with a p-value of $2.2 \times 10^{-16}$. This provides evidence about CrDOF being involved in cell-cycle regulation, which indeed was experimentally validated [46].

In order to illustrate the effect of the overexpression of CrDOF over its neighboring genes we chose three genes at distance two from it and represented their expression level in the four conditions, CrDOFfin and CW15 grown in LD and SD. We chose three genes that represent the main biological processes affected by the potentially light-regulated TFs and TRs according to the results from the previous section. These genes are a fatty acid desaturase FAD6 (Cre13.g590500) involved in carbon metabolism, a glutamate dehydrogenase GDH2 (Cre05.g232150) involved in nitrogen metabolism and a serine/threonine kinase MAPKK2 (Cre16.g684450) possibly involved in cell cycle regulation (Fig. 10). As expected, these genes increased their expression levels in the CrDOFfin genotype when compared to CW15. Additionally, we chose three genes located in the cluster 4 (purple), far away from CrDOF, to show that the expression level of distant genes tend to remain unchanged. These genes are a glycinamide ribonucleotide synthetase CrGARS (g18106), a phosphoribosylglycinamide formyltransferase PGFT (Cre12.g550700), both involved in purine biosynthesis and a plastid TF PTAC3.

![Fig. 10](image_url) Expression Levels (FPKM) of several CrDOF Neighbouring Genes and Distant Genes in the CrDOFfin and CW15 Genotypes Grown in LD and SD Conditions. Three genes in the neighbourhood at distance two from CrDOF were chosen to illustrate the correct prediction provided by ChlamyNET with respect to their increase in expression level in the CrDOFfin genotype when compared to CW15. These genes are fatty acid desaturase FAD6 (Cre13.g590500) involved in carbon metabolism, glutamate dehydrogenase GDH2 (Cre05.g232150) involved in nitrogen metabolism and serine/threonine kinase MAPKK2 (Cre16.g684450) possibly involved in cell cycle regulation. Additionally, we selected three genes from the purple cluster located far away from the CrDOF gene in ChlamyNET to show that distant genes expression is not substantially affected by CrDOF overexpression. These genes are a glycinamide ribonucleotide synthetase CrGARS (g18106), a phosphoribosylglycinamide formyltransferase PGFT (Cre12.g550700), both involved in purine biosynthesis and a plastid TF PTAC3.
(Cre12.g497350) involved in regulation of plastid genes. All these experimental tests validated the predictive power of ChlamyNET.

Conclusions

This gene co-expression network and its associated web-based tool ChlamyNET constitute one of the first integrative approaches to the study of the *Chlamydomonas* transcriptome. They aim at providing researchers with an enabling technology that will allow them to study gene co-expression patterns, determine significant biological processes, molecular functions and cellular components for a set of genes of interest as well as to identify significant TFBS in the promoters of a given set of genes.

In this work, we have shown that ChlamyNET exhibits non-random topological properties, namely scale-free and small-world properties. This suggests that the *Chlamydomonas* transcriptome possesses relevant characteristics related to error tolerance, vulnerability and information propagation [28]. On the other hand, the scale-free property implies robustness against random gene mutations or error tolerance, which means that since most genes are only co-expressed with a few other genes, a random mutation is likely to affect a non-important gene altering the expression of a reduced number of other genes. Nevertheless, the existence of key authoritative hub genes produces fragility or vulnerability to targeted attacks against this type of genes. The removal or mutation of an authoritative hub gene would affect a large number of other genes co-expressed with it, massively disrupting the functioning of the *Chlamydomonas* transcriptome. This can lead to lethality or defective growth. For example, this has been shown for the authoritative hub gene CrCO (g6302) whose over-expression and silencing are detrimental for cellular growth [2]. Additionally, the small-world property facilitates a quick spreading of information throughout ChlamyNET.

The analysis of the location of hub genes and genes with high clustering coefficient shows that both of them group together in specific regions of ChlamyNET. This indicates the existence of gene clusters whose expressions are highly coordinated, possibly to perform related biological processes. Indeed, we identified nine gene clusters that present a high intra-cluster and a low inter-cluster co-expression. Among these clusters we highlight the results obtained in two of them, clusters 1 and 2.

The most central cluster (blue cluster), where most authoritative hub genes are located, is significantly enriched in genes involved in carbon/nitrogen metabolism induced by light. Several bHLH TFs are contained in this cluster and only one of them, *CrbHLH1* (Cre14.g620850), presents a potential Arabidopsis ortholog, *PARI* (At1g69010), which has been shown to be involved in light response, and indeed the E-box sequence, a bHLH binding site, was found to be significantly present in the promoters of genes in this cluster. Two TFs from the bZIP family are also located in this cluster, *CrHY5* (Cre12.g510200) and *CrHYH* (Cre06.g310500), potential orthologs of the Arabidopsis *HY5* (At5g11260) and *HYH* (At3g17609) genes, respectively. Again, the G-box sequence, a TFBS recognized by *HY5* and *HYH* in Arabidopsis, was found to be significantly present in the promoters of their co-expressed genes. Additionally, this cluster contains two ARR-B TFs (Cre13.g572450 and g16739) and the B-box TF *CrBox1* (Cre03.g182700) whose potential Arabidopsis orthologs *RR14* (At2g01760), TOC1 (At5g61380) and COL1 (At5g15850) are involved in circadian rhythms. This suggests a key role played by light in the regulation of central metabolism in *Chlamydomonas* mediated by TFs from the bHLH and bZIP families. It also provides evidence for an input from circadian rhythms exerted by genes from the ARR-B and B-box families.

One of the most peripheral clusters in ChlamyNET, cluster 2 (brown), was shown to be involved in DNA replication and transitions between the cell cycle phases. This cluster contains potential orthologs of Arabidopsis genes involved in the G1/S transition such as origin of replication complex ORC1 (g11180) and ORC4 (Cre17.g726500), pre-initiation complex subunit CDC6 (Cre06.g292850), DNA replication initiation factor CDT1 (Cre03.g163300), minichromosome maintenance protein MCM2 (Cre07.g338000) and DNA polymerase alpha POLA1 (Cre04.g214350). The E2F sequence was found to be significantly present in the promoters of these genes. Additionally, the presence of a combination of the octamer and hexamer motifs was significantly enriched in the promoters of the genes from this cluster such as the B-type cyclin *CYCB1* (Cre08.g370400). The E2F sequence and the combination of octamer and hexamer motifs have been shown to confer S phase-specific transcriptional activation in higher plants.

These results suggest that key elements in the regulation of cell cycle, light response and carbon/nitrogen metabolism are already established in *Chlamydomonas* and conserved in higher plants such as Arabidopsis. The conserved elements are not only limited to TFs, TRs and their targets but also include the cis-regulatory elements, TFBS, present in their promoters.

The web-based software tool ChlamyNET (http://viridiplantae.ibvf.csic.es/ChlamyNet/) was developed to ensure the reproducibility of the results presented in this work and to facilitate further and independent studies.
over the *Chlamydomonas* transcriptome. We used potentially light-regulated TFs and TRs to illustrate its functionalities. Our case study suggests that these genes regulate carbon/nitrogen metabolism and cell cycle. Additionally, light regulated TFBS in *Arabidopsis* such as SORLIP2, SORLIP3 and SORLREP5 were identified in their gene promoters. Several other cases studies are available from the web page of ChlamyNET, http://viridiplantae.ibv.fcsic.es/ChlamyNET/.

Finally, RNA-seq data from algae overexpressing the transcription factor CrDOF involved in photoperiod response were used as an independent data set to experimentally cross-validate the predictive power of ChlamyNET.

**Methods**

**Data acquisition and processing**

In this study we used RNA-seq data from the *Chlamydomonas* transcriptome publicly available at the *Sequence Read Archive* (SRA, http://www.ncbi.nlm.nih.gov/Traces/sra/) [84], a database resource at the *National Center for Biotechnology Information* (NCBI) that stores more than 500 TeraBases of next-generation sequencing data. We collected more than 287 GigaBytes of information produced by seven different studies [10–16] consisting of 50 samples representing eight different genotypes under diverse physiological conditions, see Additional file 1: Table S1. These data provide a general overview of the *Chlamydomonas* transcriptome in a physiologically relevant context. All these data were obtained using the same next-generation sequencing platform, Illumina Genome Analyzer [85], in order to facilitate the comparison between samples from different experiments.

The available *Chlamydomonas* sequenced genome (version 5.3) [1] was downloaded from Phytozome (http://www.phytozome.net/) [86], a web-based platform for green plant genomics, in order to be used as the reference genome in our study. Additionally, we also obtained from the same web resource the corresponding Augustus u11.6 annotation for the assembly of the sample specific transcriptomes using the software package Cufflinks. The whole transcriptome identified in all the samples was integrated and stored in a GTF (gene transfer format) file using Cuffmerge, a utility program within the Cufflinks package. We performed this refinement of the currently available annotated *Chlamydomonas* transcriptome in order to avoid incomplete or incorrect annotation that could reduce accuracy [90] in our study.

Finally, the gene expression levels in the different conditions integrated in our study were estimated using Cuffdiff, a program included in the Cufflinks package. In order to avoid biases due to transcript length and the total number of reads generated in each experiment, we used as unit of measurement *fragments per kilobase of transcript per million mapped fragments* (FPKM) [90, 94]. Additionally, recent suggestions for normalization methods [95] that reduce the bias due to the non-uniform distribution of mapped reads within transcripts were taken into account by setting the corresponding parameters in Cuffdiff. These normalizations remove the biases in the data while preserving the variation in gene expression that occurs because of biologically relevant changes in transcription, allowing the comparison of gene expression across multiple experiments. Subsequent analysis and visualization of the results were performed using the R package cummeRbund [96].

**Selection of differentially expressed genes**

The selection of differentially expressed genes was performed using the standard methodology applied to the analysis of RNA-seq data described in [97]. The logarithm of the levels of expression measured in FPKM were computed and the delta method to estimate the variance of the log odds was used. Those genes that exhibited an adjusted p-value for the multiple testing lower than 0.05 were considered to be differentially expressed.

**Gene Co-expression criterion and network construction**

We used the absolute value of the *Pearson correlation coefficient* between gene expression profiles across the different conditions to determine the level of co-expression between the selected genes [30]. For each possible correlation value *cor*, we represented the co-expression relationships between genes that exceed this value in the *Chlamydomonas* transcriptome using an
undirected weighted network $G_{\text{cor}} = (V_{\text{cor}}, E_{\text{cor}})$. The nodes or vertices in $V_{\text{cor}}$ correspond to the genes. An undirected edge $(g_1, g_2)$ in $E_{\text{cor}}$ with associated weight $w > \text{cor}$ indicates the existence of a significant co-expression relationship between genes $g_1$ and $g_2$ with an absolute value of the Pearson correlation coefficient between the corresponding expression profiles equal to $w$.

In order to decide which correlation value is high enough to consider that two genes are significantly co-expressed we used a criterion that establishes a compromise between the generation of a scale-free network and a high network density. Most biological networks characterized so far are scale-free, which makes this property the most common metric for the rational selection of a gene correlation threshold. In order to facilitate the detection of clusters or modules of genes in the constructed network we also added the restriction of generating a network with a high density [25].

A range of correlation thresholds were considered. For each possible correlation cut-off value we determined how close the corresponding network was to fulfill the scale-free condition by computing the $R^2$ of the linear regression for the logarithmic transform of the node degree distribution. Additionally, for each possible cut-off value we used the average node degree as a measurement of the network density.

Network visualization
Graphical representations of the network were performed using Cytoscape [98], a software package for network visualization and data integration. Specifically, the organic layout method was applied to visualize ChlamyNET. This algorithm consists of a variant of the force-directed layout. Nodes produce repulsive forces whereas edges induce attractive forces. Nodes are then placed such that the sum of these forces are minimised. The organic layout has the effect of exposing the clustering structure of a network. In particular, this layout tends to locate tightly connected nodes with many interactions or hub nodes together in central areas of the network.

Significance of topological properties
In order to determine the statistical significance of the scale-free property of ChlamyNET we generated $10^4$ random networks with the same number of nodes and edges as ChlamyNET following the Erdős and Rényi random graph model [99]. None of these random networks exhibited a scale-free topology similar to ChlamyNET. This indicates that the scale-free topology of ChlamyNET is not random but rather it is the product of a self-organizing process. It has been suggested that scale-free networks emerge from a growth process by which newly added nodes preferentially attach to already existing nodes with a high number of neighbours [27]. In the case of ChlamyNET the scale-free feature can be a consequence of two mechanisms in the evolution of gene co-expression networks: (i) gene co-expression networks are not static, instead new genes may appear; and (ii) new genes are preferentially co-expressed with genes that already exhibit a large number of co-expressed genes.

We also studied the clustering coefficient in ChlamyNET, a measurement of the density of edges or co-expression relationships around genes. The clustering coefficient of a gene is calculated as the ratio of the actual number of co-expression relationships among all its neighbours and the maximal possible number of such co-expression relationships.

In order to determine that the global clustering coefficient of ChlamyNET is significantly high we generated $10^4$ random scale-free networks with the same number of nodes and edges as ChlamyNET following the Barabási random scale-free graph model [27]. None of these random networks exhibited a clustering coefficient higher than ChlamyNET.

Clustering techniques
In a general way, clustering techniques aim at identifying groups or clusters whose individuals exhibit high similarities, whereas individuals from different groups or clusters present low similarities. When clustering techniques are applied to co-expression networks, the similarity among genes is measured using the correlation among the corresponding gene profiles or co-expression. Therefore, the goal of clustering techniques, when applied to gene co-expression data, consists on identifying disjoint groups or clusters of genes so that the co-expression between genes in the same cluster is high (intra-cluster similarity) whereas the co-expression between genes from different clusters is low (inter-cluster similarity) [36]. In this respect, the silhouette [36] a criterion that combines the minimization of inter-cluster similarity with the maximization of the intra-cluster similarity, is one of the most popular measurements for the assessment of a clustering analysis. In our study we used this criterion to determine which clustering algorithm and number of clusters best describes the underlying structure in ChlamyNET. We compared the performance of the two most widely used clustering algorithms, hierarchical clustering and partition around medoids (PAM) for different number of clusters ranging from 4 to 20 clusters using the R package clValid [36].

Gene ontology term enrichment analysis
The Chlamydomonas transcriptome is very sparsely annotated since experimental validation of the different computationally predicted functions is still missing for most genes. In order to overcome this lack of GO term
annotation, we followed two different complementary approaches. In our first approach, we assigned to each *Chlamydomonas* gene the GO terms associated with its potential *Arabidopsis* ortholog based on sequence similarity. In our second approach, we used the annotation about protein domains and tools available from the Pfam database [35] to determine the protein family to which each *Chlamydomonas* gene belongs to. The GO terms associated to the identified protein family were then assigned to the corresponding gene. Our methodology for the identification of the GO terms over-represented in each cluster is a combination of both approaches. We identified as over-represented GO terms those found by both approaches or by only one of them with a very high statistical significance (a p-value lower than 10^{-6}). The R package topGO was used to perform GO term enrichment using Fisher's exact test. As gene background we selected the entire *Chlamydomonas* gene set as identified in the Phytozome 9.1 database.

Transcription factor binding site enrichment analysis

Transcription Factor Binding Sites (TFBS) enrichment analysis was performed using HOMER [100] and the known TFBS sequences in plants from the databases AGRIS [101], JASPAR [102] and AthaMap [103]. The findMotifs.pl script, applying the default parameters, was used to perform a known and de-novo motif over-representation analysis.

The background used for the over-representation analysis consists of all the gene promoters annotated in the current version of the *Chlamydomonas* genome. These data were downloaded using the BioMart functionality associated with Phytozome.

Alga material, growth conditions and RNA sequencing

Two independent biological replicates of *Chlamydomonas reinhardtii* wild type CW15 [104], and the transgenic line *CrdOFin* [46], were grown in flasks with the induction media Sueoka NO3- [105] under LD (16 h light/8 h dark) or SD (8 h light/16 h dark) conditions at 50 E light intensity with 22 \(^{\circ}\)C (during light period) and 18 \(^{\circ}\)C (during night period) in a model SG-1400 phototron (Radiber SA, Spain). Algal cells were grown during 4 days and then, were harvested 4 hours after the light went on, which was considered at Zeitgeber time zero (ZT0). The RNA isolation was performed by TRIZOL (Invitrogen) method following the manufacturer instructions. RNA quality was tested employing a ND-1000 Spectrophotometer (Nanodrop). Library preparation was carried out following the manufacturer's recommendations. Sequencing of RNA libraries was performed with the Illumina HiSeq 2000 sequencer, yielding approximately 40 million 50 bp long reads for each sample.

Availability of supporting data

The RNA-seq data set used to cross-validate the predictive power of ChlamyNET is available at European Nucleotide Archive identified with the accession number PRJEB6682.

The processed RNA-seq data, R scripts used in the construction and analysis of our network as well as the network itself in gml format are available from the web page http://viridiplantae.ibvf.csic.es/ChlamyNet/.

Additional files

Additional file 1: Table S1. Brief description of the data used in this analysis including genotypes, conditions and sequence information. (PDF 19 kb)

Additional file 2: Figure S1. Normalized expression of every gene in each cluster is represented in grey. For each cluster the red line represents the mean positive expression profile whereas the blue line represents the mean negative expression profile. Recall that the absolute value of the correlation is used to define co-expression. Note that the expression profiles for each cluster are distinct. (TIFF 7917 kb)

Additional file 3: Figure S2. Cluster 1 (orange) is enriched in genes involved in lipid metabolism, including the pentose phosphate pathway. The TAG biosynthetic pathway that preferentially uses as acyl donors galactolipids is also a member of this cluster.

Additional file 4: Figure S3. Cluster 2 (Brown), 3 (Red) and 7 (Green) from Figure 4. Analysis of the *Chlamydomonas* metabolic pathways revealed that cluster 2 (brown) in figure 4 is enriched in metabolic processes involved in the DNA and RNA synthesis such as the pyrimidine synthesis pathways. Nevertheless, clusters 3 (red) and 7 (green) were not so significantly enriched in metabolic pathways. Nevertheless, in cluster 3 (red), with a low significance, the synthesis pathway for triacylglycerol using galactolipids produced by galactolipid desaturation as acyl donors, was identified. In cluster 7 (green), the aerobic respiration pathway was fully included. (PNG 50 kb)

Additional file 5: Figure S4. Metabolic Pathways Contained in Clusters 8 (turquoise), the two most important metabolic pathways in photosynthetic organisms, the Calvin cycle and the TCA cycle, have a significant number of genes represented. (PNG 98 kb)

Additional file 6: Table S2. Biological Processes Potentially Controlled by the Light-regulated TFs and TRs in ChlamyNET. (PDF 8 kb)

Competing interests

The authors declare that they have no competing interests.

Author's contributions

FJR performed the bioinformatics analysis, contributed to the design and development of the web-based tool ChlamyNET and wrote the paper. ELR performed the experimental work for the validation of ChlamyNET and
References

1. Merchant S, Prochnik S, Vallon O, Harris E, Karpowicz S, Witman G, Terry A, Salmov A, Fritz-Laylin L, Marechal-Drouard L. The Chlamydomonas genome reveals the evolution of key animal and plant functions. Science. 2007;318(5848):245–50.

2. Serrano G, Herrera-Palau R, Romero JM, Serrano A, Coupland G, Valverde F. Evolutionarily conserved photoperiod mechanisms in plants: when did plant photoperiodic signaling appear? Plant Signal. Signaling. 2009;19(10):672.

3. Romero JM, Valverde F. Evolutionarily conserved photoperiod mechanisms in plants: when did plant photoperiodic signaling appear? Plant Signal. Signaling. 2009;19(10):672.

4. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol. 2005;4:Article 17.

5. Watt S, Strogatz S. Collective dynamics of ‘small-world’ networks. Nature. 2001;411(6833):41.

6. Altocumulus: A New View of the Yeast Protein Interaction Network. PLoS One. 2014;9(4):e95176.

7. Liang Y, Cai B, Chen F, Wang G, Wang M, Zhong Y, Z C. Construction and validation of a gene co-expression network in grapevine (Vitis vinifera. L). Horticulture Res. 2014;1:14040.

8. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol. 2005;4:Article 17.

9. Dong J, Horvath S. Understanding network concepts in modules. BMC Syst Biol. 2007;1:24.

10. Aoki K, Ogata Y, Shibata D. Approaches for Extracting Practical Information from Gene Coexpression Networks in Plant Biology. Plant Cell Physiol. 2007;48(3):381–90.

11. Gao Z, Zhao R, Ruan J. A genome-wide cis-regulatory element discovery method based on promoter sequences and gene co-expression networks. BMC Genomes. 2013;14:52.

12. Barabasi A, Albert R. Emergence of Scaling in Random Networks. Science. 1999;286:509–12.

13. Kroopat J, Hong-Hermersdorf A, Casero D, Ent P, Castruita M, Pellegrini M, Merchant SS, DM. A revised mineral nutrient supplement increases biomass and growth rate in Chlamydomonas reinhardtii. Plant J. 2011;66:770–80.

14. Castruita M, Casero D, Karpowicz S, Kroopat J, Vieler A, Hsieh S, Yan W, Cokus S, Luo J, Benning C, Pellegrini M, Merchant S. Systems biology approach in Chlamydomonas reveals connections between copper nutrition and multiple metabolic steps. Plant Cell. 2011;23(4):1273–92.

15. Miller R, Wu G, Deshpande R, Vieler A, Gartner K, Li X, Moellinger ER, Zauner S, Cornish AJ, Liu B, Bullard S, Sears BB, Kuo M, Hegg EL, Shachar-Hill Y, Shiu S, Benning C. Changes in Transcript Abundance in Chlamydomonas reinhardtii following Nitrogen Deprivation Predict Distension of Metabolism. Plant Physiol. 2010;154:1727–52.

16. Gonzalez-Ballester D, Casero C, Cokus S, Pellegrini M, Merchant SS, Grossman A. RNA-Seq Analysis of Sulfur-Deprived Chlamydomonas Cells Reveals Aspects of Acclimation Critical for Cell Survival. Plant Cell. 2010;22:2058–84.

17. DallMolin C, Quek L, Palfreyman R, Nielsen L. AlgaGEM – a genome-scale metabolic reconstruction of algae based on the Chlamydomonas reinhardtii genome. BMC Genomes. 2011;12:SS.

18. Lopez D, Casero D, Cokus S, Merchant S, Pellegrini M. Algal Functional Annotation Tool a web-based analysis suite to functionally interpret large gene lists using integrated annotation and expression data. BMC Bioinf. 2011;12:282.

19. Zheng H, Chiang-Hsieh Y, Chen C, Hsu B, Liu T, Chen C, Chang W. AlgaePath: comprehensive analysis of metabolic pathways using transcript abundance data from next-generation sequencing in green algae. BMC Genomes. 2014;15:196.

20. Bassel G, Lan H, Glaab E, Gibbs D, Gerjets T, Kranzogor N, Bonner A, Holdsworth M, Provart N. Genome-wide network model capturing seed germination reveals coordinated regulation of plant cellular phase transitions. Proc Natl Acad Sci. 2011;108(23):9709–14.

21. Cai B, Li C, Huang J. Systematic identification of cell-wall related genes in Arabidopsis thaliana. J Exp Bot. 2011;62(8):2453–69.

22. Batada N, Reguly T, Breitkreutz A, Boucher L, Breitkreutz BJ. Stratus Not Altocumulus: A New View of the Yeast Protein Interaction Network. PLoS One. 2014;9(4):e95176.

23. Zheng H, Chiang-Hsieh Y, Chen C, Hsu B, Liu T, Chen C, Chang W. AlgaePath: comprehensive analysis of metabolic pathways using transcript abundance data from next-generation sequencing in green algae. BMC Genomes. 2014;15:196.

24. Dong J, Horvath S. Understanding network concepts in modules. BMC Syst Biol. 2007;1:24.

25. Aoki K, Ogata Y, Shibata D. Approaches for Extracting Practical Information from Gene Coexpression Networks in Plant Biology. Plant Cell Physiol. 2007;48(3):381–90.

26. Gao Z, Zhao R, Ruan J. A genome-wide cis-regulatory element discovery method based on promoter sequences and gene co-expression networks. BMC Genomes. 2013;14:52.

27. Barabasi A, Albert R. Emergence of Scaling in Random Networks. Science. 1999;286:509–12.

28. Wang X, Chen G. Complex networks: small-world, scale-free, and beyond. IEEE Circuits Syst Mag. 2003;3:6–20.

29. Jeong H, Mason S, Barabasi A, Oltvai Z. Lethality and centrality in protein networks. Nature. 2001;411(6833):41–2.

30. Song L, Langfield P, H S. Comparison of co-expression measures: mutual information, correlation, and model based indices. BMC Bioinf. 2012;13:328.

31. Batada N, Reguly T, Breitkreutz A, Boucher L, Breitkreutz BJ. Stratus Not Altocumulus: A New View of the Yeast Protein Interaction Network. PLoS One. 2014;9(4):e95176.

32. Mao L, Van Hemert J, Suddothus DD, Dickerson J. Arabidopsis gene co-expression network and its functional modules. BMC Bioinf. 2009;10:346.

33. Watts D, Strogatz S. Collective dynamics of ‘small-world’ networks. Nature. 1999;393:440–2.

34. Kivilcim B. Authoritative Sources in a Hyperlinked Environment. J ACM. 1999;46(5):604–12.

35. Punta M, Cogoli P, Eberhardt R, Mistri J, Tate J, Bournsell C, Pang N. Protein networks. Nature. 2006;441(7090):190–9.

36. Brock G, Pihur V, Datta S. Datta S: clValid: An R Package for Cluster Validation. J Stat Softw. 2008;25:4.
37. Rhee S, Wood V, Dolinski K, Dragic S. Use and misuse of the gene ontology annotations. Nat Rev Genet. 2008;9:509–15.

38. Marcotte E, Date S. Exploiting big biology: integrating large-scale biological data for function inference. Brief Bioinform. 2002;3:63–74.

39. Chae L, Lee I, Shin J, Rhee S. Towards understanding how molecular networks evolve in plants. Curr Opin Plant Biol. 2012;15(2):177–84.

40. May P, Christian J, Kempa S, Walther D. ChlamyCyc: an integrative systems biology database and web-portal for Chlamydomonas reinhardtii. BMC Genomics. 2009;10:209.

41. Gennidiakis S, Raso S, Greenham K, Uihirg R, O’Leary B, Snedden W, Lu C, Plaxton W. Bacterial- and plant-type phosphoenolpyruvate carboxylase polypeptides interact in the hetero-oligomeric Class-2 PEPC complex of developing castor oil seeds. Plant J. 2007;52(5):839–49.

42. Kwon S, Cho H, Kim S, Park O. The Rab GTPase RabG3b Positively Regulates Autophagy and Immunity-Associated Hypersensitive Cell Death in Arabidopsis. Plant Physiol. 2013;161(4):1722–36.

43. Yoon K, Han D, Li Y, Sommerfeld M, Hu Q. Phospholipid:Diacylglycerol Acyltransferase II is a Multifunctional Enzyme Involved in Membrane Lipid Turnover and Degradation While Synthesizing Triacylglycerol in the Unicellular Green Microalga Chlamydomonas reinhardtii. Plant Cell. 2012;24(9):3708–24.

44. Demidov D, VanDammme D, Geelen D, Blatter F, Houbena A. Identi cation and Dynamics of Two Classes of Aurora-Like Kinases in Arabidopsis and Other Plants. Plant Cell. 2005;17:386–48.

45. Umeda M, Shimotohno A, Y M. Control of Cell Division and Transcription by Cyclin-dependent Kinase-activating Kinases in Plants. Plant Cell Physiol. 2005;46(9):1437–42.

46. Lucas-Reina E, Romero-Campero F, Romero J, Valverde F. An evolutionarily conserved DOF-CONSTANS module controls plant photoperiodic signalling. Plant Physiol. 2015;168(2):561–74.

47. Kaufmann K, Papiero A, Angenet G. Regulation of transcription in plants: mechanisms controlling developmental switches. Nat Rev Genet. 2010;11:830–42.

48. Riaño-Pachón D, Guedes-Corea L, Trejo-Espinosa R, MR B. Green Transcription Factors: A Chlamydomonas Overview. Genetics. 2008;179:31–9.

49. Jin J, Zhang H, Kong L, Gao G, Luo J. PlantTFDB 3.0: a portal for the functional and evolutionary study of plant transcription factors. Nucleic Acids Res. 2013;41(D1):D182–7.

50. Pérez-Rodríguez P, Riaño-Pachón DM, Correa LG, Rensing SA, Kersten B, Mueller-Roeber B. PlantTFDB: updated content and new features of the plant transcription factor database. Nucleic Acids Res. 2010;38 suppl 1:D282–7.

51. Cifuentes-Esquivel N, Bou-Torrent J, Calstyan A, Galleri M, Sessa G, Salà-Martet M, Roig-Villanova I, Ruberti I, Mart nez-García J. The bHLH proteins NtmybA1 and NtmybA2 causes downregulation of G2/M phase-expressed genes and negatively affects a eXeR cells division and expansion in tobacco. Plant Signal Behav. 2013;8:e26780.

52. Pedersen D, Coppins F, Ma A, Antosch M, Matix B, Merkler T, Beemster G, Houben A, Grasser K. The plant-specific family of DNA-binding proteins containing three HMGB-box domains interacts with mitotic and meiotic chromosome. New Phytol. 2011;192(3):577–89.

53. Van Leene J, Hollunder J, Eeckhout D, Persiau G, Van De Slijke E, Stals H, Van Stalle L, Pharazyn A, Ferreira P, Eloy N, Renne C, Meyer C, Faure J, Steinbrenner J, Isterdael G, Verkest A, Neirynck S, Buel Y, De Bodt S, Maere S, Laukens K, Phyrazyn A, Ferreira P, Eloy N, Renne C, Meyer C, Faure J, Steinbrenner J, Beyerin J, Larkin J, Van de Peere Y, Hilson P, Kuiper M, De Veylder L, Van Onckelen H, Inze D, Witters E, De Jaeger G. Targeted interactomics reveals a complex core cell cycle machinery in Arabidopsis thaliana. Mol Syst Biol. 2010;6:397.

54. Gutierrez C. The Arabidopsis Cell Division Cycle. Arabidopsis Book. 2009;7:e0120.

55. Avramova Z. Evolution and pleiotropy of TRITHORAX function in Arabidopsis. Int J Dev Biol. 2005;49(3–4):71–81.

56. Wang Y, Yang M. Silico Identi cation of Co- transcribed Core Cell Cycle Regulators and Transcription Factors in Arabidopsis. J Integr Plant Biol. 2007;49(8):1253–60.

57. Chaboute M, Clement B, Sekine M,Philippis G, Chabert-Gicot N. Cell cycle regulation of the tobacco ribonucleotide reductase small subunit gene is mediated by E2F-like elements. Plant Cell. 2000;12(10):1987–2000.

58. de Jager S, Menges M, Baur U, Murray J. Arabidopsis E2F1 binds a composite regulatory element present in the promoter of S-phase-regulated gene AtCD64 and is a member of a multigene family with differential activities. Plant Mol Biol. 2001;47(4):555–68.

59. Taoka K, Kaya H, Nakayama T, Araki T, Meshi T, Iwabuchi M. Identification of three kinds of mutually related composite elements conferring S phase- specific transcriptional activation. Plant J. 1999;18(6):511–23.
80. Gretarsson B, Bostandjieva S, Donovan J, Hilder T. WiGis: A framework for Web-based interactive graph visualizations. In: International Symposium on Graph Drawing. 2009. p. 119–43.
81. Jiao Y, Yang H, Ma L, Sun N, Yu H, Liu T, Gao Y, Gu H, Chen Z, Wada M, Gerstein M, Zhao M, Qu L, Deng X. A Genome-Wide Analysis of Blue-Light Regulation of Arabidopsis Transcription Factor Gene Expression During Seedling Development. Plant Physiol. 2003;133(4):1480–93.
82. Reisdorph N, Small G. The CPH1 Gene of Chlamydomonas reinhardtii Encodes Two Forms of Cryptochrome Whose Levels Are Controlled by Light-Induced Proteolysis. Plant Physiol. 2004;134:154–64.
83. Hudson M, Quail P. Identification of Promoter Motifs Involved in the Network of Phytochrome A-Regulated Gene Expression By Combined Analysis of Genomic Sequence and Microarray Data. Plant Physiol. 2003; 133(4):1605–16.
84. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Church DM, DiCuccio M, Edgar R, Federhen S, Helmbold W, Kenton DL, Khovayko O, Lipman DJ, Madden TL, Maglott DR, Ostell J, Poustis JU, Pruitt KD, Schuler GD, Schriml LM, Sequeira E, Sherry ST, Sirotkin K, Starchenko G, Suzek TO, Tatusov R, Tatusova TA; L W, Y. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2005;33:D39–45.
85. Bentley DR, Balasubramanian S, Sowldow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell G. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008; 456(7218):53–9.
86. Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazio J, Mitros T, Dirks W, Heilsten U, Putnam N, S RD. Phytosome: a comparative platform for green plant genomics. Nucleic Acids Res. 2012;40:D1178–86.
87. Garber M, Grabherr M, Guttmann M, Trapnell C. Computational methods for transcriptome annotation and quantification using RNA-seq. Nat Methods. 2011;8(6):649–77.
88. Trapnell C, Roberts A, Go L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn J, L P. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 2012;7(3):562–78.
89. Trapnell C, Pachter L, Salzberg S. TopHat: discovering splice junctions with RNA-seq. Bioinformatics. 2009;25:1105–11.
90. Cea T. Transcript assembly and quanti cation by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010;28:511–5.
91. Ewing B, Hillier L, Wendel MC, Green P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. 1998;8(3):175–85.
92. Ewing B, Green P. Base-calling of automated sequencer traces using phred, II. Error probabilities. Genome Res. 1998;8(3):186–94.
93. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-e cient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10:R25.
94. Mortazavi A, Williams B, McCue K, Schae er E, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-seq. Nat Methods. 2008;5:621–8.
95. Roberts A, Trapnell C, Donaghey J, Rinn J, Pachter L. Improving RNA-Seq expression estimates by correcting for fragment bias. Genome Biol. 2011;12: R22.
96. Go L, Trapnell C. cummeRbund: Analysis, exploration, manipulation, and visualization of Cufflinks high-throughput sequencing data. 2011. [R package version 1.2.0].
97. Bullard J, Purdom E, Hansen K, Dudoit S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinf. 2010;11(94):1–13.
98. Smoot M, Ono K, Ruscheinski J, Peng-Liang W, Ideker T. Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics. 2011; 27(3):431–2.
99. Erdos P, Renyi A. On the evolution of random graphs. Publ Math Inst Hung Acad Sci. 1960;5:17–61.
100. Heinz S, Benner C, Spann N, Bertolino E, Lin Y, Laslo P, Cheng J, Murre C, Singh H, Glass C. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell. 2010;38(4):576–89.
101. Davuluri R, Sun H, Palaniswamy S, Matthews N, Molina C, Kurtz M, Grotewold E. AGRIS: Arabidopsis Gene Regulatory Information Server, an information resource of Arabidopsis cis-regulatory elements and transcription factors. BMC Bioinf. 2003;4:25.