Institut für Biochemie und Biologie
Arbeitsgruppe Prof. Dr. Bernd Müller-Röber

Nuclear proteomics and transcription factor profiling

in *Chlamydomonas reinhardtii*

Dissertation
zur Erlangung des akademischen Grades
"doctor rerum naturalium"
(Dr. rer. nat.)
in der Wissenschaftsdisziplin "Molekularbiologie"

eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät
der Universität Potsdam

von

Flavia Vischi Winck

aus Campinas, Brasilien

Potsdam-Golm Summer 2011

Published online at the
Institutional Repository of the University of Potsdam:
URL http://opus.kobv.de/ubp/volltexte/2011/5390/
URN urn:nbn:de:kobv:517-opus-53909
http://nbn-resolving.de/urn:nbn:de:kobv:517-opus-53909

“Face the difficulties and challenges with open heart”

“Enfrente as difficuldades e os desafios de coração aberto”

Prof. Dr. Sérgio Marangoni

Erklärung

I hereby declare that this doctoral thesis is the result of my work carried out between winter semester 2007 and summer semester 2011 in the group of Prof. Dr. Bernd Müller-Röber at the University of Potsdam in Potsdam-Golm, Germany. This document has not been submitted for any other degree at any other university. Information derived from the published and unpublished work of others has been acknowledged in the text and the references are given. Credits over published material and adaptation of published material are indicated in the text.

Potsdam, May 11th, 2011.

Flavia Vischi Winck

Acknowledgements

This work would not be possible without the support, help and encouragements received from many people. At first, I would like to thank Prof. Dr. Bernd Müller-Röber for receiving me in his group, for sharing ideas, for giving me support, and giving me enough freedom for my scientific and intellectual development. I thank Camila Caldana for helping me to really come to Germany and during my stay here. I thank Heike Küchmeister for her kindness and help in organizing my arrival. I want to especially thank my husband, Diego, for all the moments he helped me and supported me in my work, especially with bioinformatics, for all scientific discussions we had along these years, for his partnership, his friendship and his love. Big thanks to Samuel Arvidsson for his patience, his great help with data analysis and his friendship. Thanks to all friends and colleagues which were or are part of the AG MüRö, in particular: Lilian for the friendship cultivated, to Raul and Fernando for being kind and good listeners, to Mirek, Radek, Magda, Kamil and Ola for always making us to smile; to Luiz and to Mathias for keeping me feeling Brazilian; to Ines for her kindness; to Amin and Nooshin for being so gentle; to Tripti for encouragements; to Martin for keeping things going in my absence; to Dagmar, Antje, Eike and Karina for taking care of our labs and being there always when we need. My thanks to the colleagues at the Max Planck Institute of Molecular Plant Physiology, in particular to Frederik Sommer, Jens Rupprecht, Eugenia Maximova and Armin Schlereth for helping me and advising me in the experiments involving mass spectrometry, cell cultivation, microscopy and qRT-PCR. Thanks to Aneta Koseska and Sabrina Donner for collaborating in this work for the reconstruction of GRN. I thank German citizens which through the Federal Ministry of Education and Research (BMBF) support scientific research and receive foreigners on their land. I thank the organizational team of *Go*FORSYS project for giving me the opportunity to be part of it. I thank the University of Potsdam for opening its doors for students from abroad and for trying to be better every year. I thank all my family, especially my mother Cecilia and my sister Paula for their love and support. Obrigado mãe, obrigado pai, por de uma maneira ou outra incentivarme sempre a seguir meus objetivos. Família, obrigado por vocês existirem na minha vida.

Allgemeinverständliche Zusammenfassung

Pflanzen nutzen das Sonnenlicht um Substanzen, sogenannte Kohlenhydrate, zu synthetisieren. Diese können anschließend als Energielieferant für das eigene Wachstum genutzt werden. Der aufbauende Prozess wird als Photosynthese bezeichnet. Ein wichtiges Anliegen ist deshalb zu verstehen, wie Pflanzen äußere Einflüsse wahrnehmen und die Photosynthese dementsprechend regulieren. Ihre Zellen tragen diese Informationen in den Genen. Die Pflanzen nutzen aber in der Regel nicht alle ihre Gene gleichzeitig, die sie zur Anpassung an Umwelteinflüsse besitzen. Zu meist wird nur eine Teilfraktion der gesamten Information benötigt. Wir wollten der Frage nachgehen, welche Gene die Zellen für welche Situation regulieren. Im Zellkern gibt es Proteine, sogenannte Transkriptionsfaktoren, die spezifische Gene finden können und deren Transkription modulieren. Wenn ein Gen gebraucht wird, wird seine Information in andere Moleküle übersetzt (transkribiert), sogenannte Transkripte. Die Information dieser Transkripte wird benutzt um Proteine, Makromoleküle aus Aminsäuren, zu synthetisieren. Aus der Transkription eines Gens kann eine große Zahl des Transkripts entstehen. Es ist wahrscheinlich, dass ein Gen, dass gerade gebraucht wird, mehr Transkriptmoleküle hat als andere Gene. Da die Transkriptionsfaktoren mit der Transkription der Gene interferieren können, entwickelten wir in der vorliegenden Arbeit Strategien zur Identifikation dieser im Zellkern zu findenden Proteine mittels eines „Proteomics“-Ansatzes. Wir entwickelten weiterhin eine Strategie zur Identififikation von Transkripten Transkriptionsfaktor-codierender Gene in der Zelle und in welche Menge sie vorkommen. Dieser Ansatz wird als „Transcript-Profiling“ bezeichnet. Wir fanden Zellkern-lokalisierte Proteine, die als Signalmoleküle funktionieren könnten und Transkripte, die bei unterschiedlichen Umweltbedingungen in der Zelle vorhanden waren. Wir benutzten, die oben genannten Ansätze um die einzellige Grünalge Chlamydomonas zu untersuchen. Die Informationen, die wir erhielten, halfen zu verstehen welche Transkriptionsfaktoren notwendig sind, damit Chlamydomonas bei unterschiedlichen Umweltbedingungen, wie z.B. unterschiedliche Lichtintensitäten und unterschiedlicher Konzentration von Kohlenstoffdioxid, überlebt.

Summary

The transcriptional regulation of the cellular mechanisms involves many different components and different levels of control which together contribute to fine tune the response of cells to different environmental stimuli. In some responses, diverse signaling pathways can be controlled simultaneously. One of the most important cellular processes that seem to possess multiple levels of regulation is photosynthesis. A model organism for studying photosynthesis-related processes is the unicellular green algae *Chlamydomonas reinhardtii,* due to advantages related to culturing, genetic manipulation and availability of genome sequence. In the present study, we were interested in understanding the regulatory mechanisms underlying photosynthesis-related processes. To achieve this goal different molecular approaches were followed. In order to indentify protein transcriptional regulators we optimized a method for isolation of nuclei and performed nuclear proteome analysis using shotgun proteomics. This analysis permitted us to improve the genome annotation previously published and to discover conserved and enriched protein motifs among the nuclear proteins. In another approach, a quantitative RT-PCR platform was established for the analysis of gene expression of predicted transcription factor (TF) and other transcriptional regulator (TR) coding genes by transcript profiling. The gene expression profiles for more than one hundred genes were monitored in time series experiments under conditions of changes in light intensity (200 µE m^{-2} s^{-1} to 700 µE m^{-2} s^{-1}), and changes in concentration of carbon dioxide (5% CO_2 to 0.04% CO_2). The results indicate that many TF and TR genes are regulated in both environmental conditions and groups of co-regulated genes were found. Our findings also suggest that some genes can be common intermediates of light and carbon responsive regulatory pathways. These approaches together gave us new insights about the regulation of photosynthesis and revealed new candidate regulatory genes, helping to decipher the gene regulatory networks in Chlamydomonas. Further experimental studies are necessary to clarify the function of the candidate regulatory genes and to elucidate how cells coordinately regulate the assimilation of carbon and light responses.

Contents

List of abbreviations

RNA	Ribonucleic acid
1D	One dimension
2-DE	Two-dimensional electrophoresis
C18	Chromatographic columns containing polymer chains of 18 carbons
CBI	Codon bias index
CCM	Carbon concentrating mechanism
CDSs	Coding DNA sequences
COA	Correspondence analysis
DAPI	4',6-diamidino-2-phenylindole
DB	Database
DNA	Deoxyribonucleic acid
DTT	Ditiothreitol
ENC	Effective number of codons
FAIRE	Formaldehyde-assisted isolation of regulatory elements
FC	Fold change
FDR	False discovery rate
GC	Gas chromatography
GO	Gene ontology
HC	High-CO_2
H5P	Hepes-5M-phosphate
ID	Identifier
IEF	Isoelectrofocusing
LC	Low-CO_2
LC/MS/MS	Liquid chromatography tandem mass spectrometry
LHC	Light harvesting complex
LREs	Light responsive *cis*-elements

LTQ XL	Linear ion trap quadrupole
MALDI-ToF	Matrix-assisted laser desorption/ionization -time of flight
mRNA	Messenger ribonucleic acid
MS	Mass spectrometry
NIB	Nuclei isolation buffer
NLS	Nuclear localization signals
NP-40	Nonidet P-40
PCR	Polymerase chain reaction
PIC	Pre-initiation complex
PSM	Peptide spectrum matches
PVDF	Polyvinylidene fluoride
qRT-PCR	Quantitative reverse-transcription PCR
qPCR	Quantitative PCR
RbcL	Ribulose-1,5-bisphosphate carboxylase oxygenase large subunit
RbcS	Ribulose-1,5-bisphosphate carboxylase oxygenase small subunit
RP-HPLC	Reverse phase-high performance liquid chromatography
SDS-PAGE	Sodium dodecyl sulfate polyacrylamide gel electrophoresis
siRNA	Silencing ribonucleic acid
TAP	Tris-acetate-phosphate
TEMED	Tetramethylethylenediamine
TFs	Transcription factors
TRFs	Transcription related factors
tRNAs	Transfer RNAs
TRs	Transcription regulators
TSS	Transcription starting site
TTS	Transcription termination site

List of tables

List of appendix tables

List of figures

List of appendix figures

Chapter 1 General introduction[1]

[1] The Section 1.4 presented in the following Chapter contains parts of the following manuscript:

Daniel Martins de Souza, Bruno Menezes de Oliveira, Erich Castro-Dias, **Flavia Vischi Winck**, Ricardo Shiniti Horiuchi, Paulo Aparecido Baldasso, Hugo Takeda Caetano, Natalia Katuchi Pires, Sérgio Marangoni, José Camillo Novello (2008) The untiring search for the most complete proteome representation: reviewing the methods. *Briefings in Functional Genomics and Proteomics*, 7(4): 312-321.

1.1 Gene expression in eukaryotes

Living cells are constantly subjected to environmental changes and developmental processes which function as external and internal stimuli. For example, one could think of many factors which would stimulate a cellular response, such as hormones, light, temperature, osmotic pressure, nutrient deprivation, pathogens, among others. Cells, encode in their genomes all the information which allows them to respond to a changing environment. The translation of the stimuli received into a cellular response, on a new cellular state, involves many different processes which function in a coordinated and integrated manner. One of the processes that influence the response of cells to the environmental changes is known as gene expression regulation. Since the 80´s the gene expression issue has been receiving much attention from molecular biologists mainly after discovering of evidences indicating that differential gene activation would explain how the gene expression is controlled (Brown, 1981). The process of gene transcription in eukaryotes is performed by different RNA polymerases, which transcribe different classes of genes. RNA polymerase I is nucleolar in origin and synthesizes ribosomal RNAs. RNA polymerase II is responsible for the transcription of protein coding genes and miRNA genes, whereas RNA polymerase III synthesizes tRNAs and other small structural RNA molecules, such as the 5S ribosomal RNA subunit. Additionally, flowering plants contain further two RNA polymerases, IV and V, which are involved in the synthesis of small interfering RNAs (siRNAs) (Onodera et al., 2005; Zhang et al., 2007b). In the following we focus on the mechanism of gene transcription by RNA polymerase II. It has been demonstrated that the transcriptional machinery is composed of a multi-complex of proteins which interact in order to activate gene transcription. This complex contains a group of proteins required for gene transcription, but there is evidence that the general components of the transcriptional machinery may vary depending on the physiological conditions. Among these general components, RNA polymerase II is responsible for synthesizing the mRNA molecules from protein-coding and miRNA genes, but its function is only accomplished when a pre-initiation complex (PIC) is assembled. First, the general transcription factor TFIID, a complex of TBP (TATA-binding protein) and TAFs (TBP-associated factors), binds to the core

promoter of genes permitting the complete assembling of the PIC with the participation of the general transcription factors TFIIA, TFIIB, TFIIE, TFIIF and TFIIH (Sauer et al., 1995; Verrijzer, 2001; Thomas and Chiang, 2006). The components of the PIC have been shown to pass through an extensive rearrangement of the protein-DNA interactions until complete engagement of RNA Polymerase II to the core promoter (Yakovchuk et al., 2010). Interestingly, it has been shown that protein sequences showing similarity to TBP can also participate in controlling the expression of genes which do not contain a TATA-box on their promoter region. The interaction between these transcription related factors (TRFs), transcription factors (TFs) and nucleosome remodeling factors seem to be able to control gene expression, suggesting that the recognition of different core promoters could involve different proteins complexes (Goodrich and Tjian, 2010). Complementary, the action of co-activators has been shown to be important for activating the gene transcription by direct binding to the PIC or by facilitating chromatin rearrangement and the access of the PIC to the regulatory DNA sequences (Thomas and Chiang, 2006). The role of co-activators seems to be preceded by the action of transcriptional activators, which are able to bind to specific regions of the chromatin and recruiting chromatin co-activators leading to modification of the nucleosome histones. The transcriptional activators can recruit the Mediator complex which has been shown to be an important component of gene expression regulation by targeting polymerase II and recruiting it to the PIC and favoring the engagement of PIC to the DNA (Malik and Roeder, 2010). There are indications that the presence of Mediator is necessary for the expression of all genes, but still there are controversies (Thomas and Chiang, 2006). The different modules of the Mediator complex have also been shown to be involved in the mediation of different signals to the PIC which would in fact modulate the activation of the PIC. Different modules were involved in responding or mediating signals from specific cellular pathways. The interaction of Mediator with signaling molecules or co-activators can induce to different structural configurations which could be translated into different outputs, inducing or repressing transcriptional output (Thomas and Chiang, 2006; Malik and Roeder, 2010).

1.2 Regulation of the gene expression

An important aspect of how cells respond to environmental stimuli resides in the control of gene expression through transcriptional regulation. The individual and normally differentiated cells usually do not use their whole repertoire of genes for adaptation to new environmental situations. How cells determine which genes are important in a given condition and how they are regulated constitute fascinating questions for which answers are intimately related with the basis of cell survival and development (O'Malley et al., 1977). The pattern of gene expression, in other words the assortment of which genes that are switched "on" or "off", determines the features of a cell in a certain condition. These patterns are mostly controlled through transcriptional regulation, influencing the number of genes which are transcribed and the abundance of their transcripts at any given moment (Verrijzer, 2001). In eukaryotes, the control of the cellular responses through transcriptional regulation involves the participation of different molecules such as microRNAs, transcription factors (TFs) and other transcription regulators (TRs). Transcription factors are proteins that can bind to the DNA molecules and contain in their basic structure one or more sequence-specific DNA-binding and trans-activation domain(s) which can interact with other co-activator proteins or directly with the PIC. TFs recognize specific DNA regulatory elements that are generally located in the promoter region of the genes they control, resulting in the induction or repression of gene transcription. Some TFs only execute their regulatory function after dimerization (Latchman, 1997). The nuclear DNA of eukaryotes contains many different DNA regulatory elements which can be recognized by TFs, such as core promoter elements, upstream promoter region elements and distal enhancer or silencer elements. Upstream promoter elements are usually located at a position shortly upstream of the core promoter (100-200bp) and can be recognized by TFs which in turn may interact with co-activators to facilitate the formation and recruitment of PIC components, increasing transcriptional rates (Brand et al., 1985; Nolis et al., 2009). However, the interaction of some TFs with the PIC may also lead to transcriptional repression, inhibiting basal transcription when activators are absent (Thomas and Chiang, 2006). Distal regulatory elements have also been identified and shown to be important in these processes. They can be located

thousands of base pairs away from the core promoter, and by DNA looping and interaction with TFs can fold back to the promoter region of the target genes and thereby modulate their transcriptional output (Nolis et al., 2009). Transcription regulator proteins are also involved in the regulation of transcription but, in contrast to transcription factors lack a sequence-specific DNA-binding domain. They contribute to transcriptional regulation, e.g. through chromatin remodeling, or as co-activators or co-repressors (Brivanlou and Darnell, 2002). Models of how the expression of a gene is activated or repressed date back to 1960´s and had already proposed the existence of many of the elements we today know as part of the signaling and regulatory networks (Britten and Davidson, 1969). More recent models include a vast number of possible regulatory mechanisms and confirmed that transcription factors and regulators can have positive or negative effect on the transcription initiation (Jacob and Monod, 1961; Britten and Davidson, 1969; Brivanlou and Darnell, 2002). Recently, with the possibility of genome-wide analysis of DNA *cis*-elements (DNA regions to which TFs can bind in a sequence-specific manner) and the identification of TF target genes, the analysis of gene regulatory networks has expanded. It has been revealed that most likely the previous master regulator TFs can also be regulated by other transcriptional regulators, and subsets of transcriptional cascades can be regulated by a combination of transcriptional regulators, which would result in a more flexible cellular response but also in a more complex regulatory network (Kaufmann et al., 2010). The many processes involved in the regulation of gene expression, including chromatin remodeling, transcription, mRNA processing, nuclear export of RNAs, mRNA translation in the cytoplasm and RNA degradation seem to actually function in a tightly-coupled way (Figure 1-1).

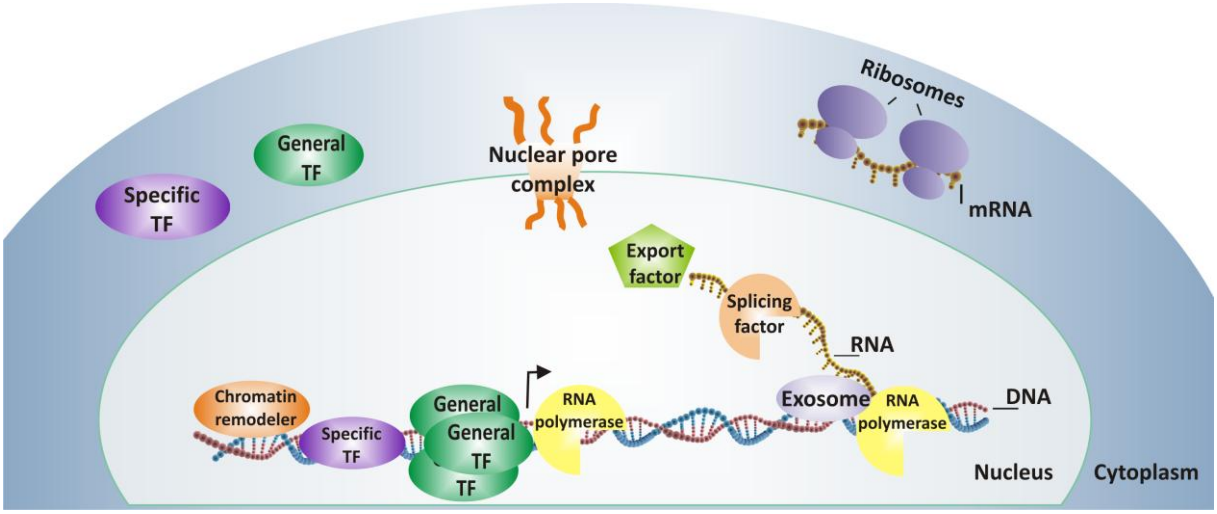


Figure 1-1. Schematic representation of the main components involved in the different levels of the regulation of eukaryotic gene expression.

(Adapted by permission from Macmillan Publishers Ltd: Nature Reviews Genetics; (Komili and Silver, 2008), copyright 2011).

Genome-wide studies, together with experimental data demonstrating molecular interactions between regulatory proteins have demonstrated that the elements or factors involved in gene expression regulation can exhibit a coordinated activity. For instance, the coordinated activity of TFs indicates that not only a combinatorial, coordinated binding of TFs to a particular gene promoter occurs but rather that a group of TFs can bind coordinately to the promoter or regulatory regions of many genes simultaneously. This would generate a coupled input which would lead to a more sensitive and organized regulatory response to different stimuli (Komili and Silver, 2008). The key role of the transcriptional regulation in driving cellular responses and determining phenotypes has been studied for many years (Doebley and Lukens, 1998) and the complexity of the signaling pathways seem to have increased during the evolution of animals and plants and it has been suggested that the number of TFs and their proportion in the genomes is correlated to the increase in complexity of the cellular responses (Levine and Tjian, 2003; Lang et al., 2010). However, there are some indications that the complexity of genetic pathways can not be explained only by natural selection, but also through non-adaptive processes (Lynch, 2007).

1.3 Chlamydomonas as a model for understanding plant gene regulatory networks

The unicellular green algae *Chlamydomonas reinhardtii* (Figure 1-2) is a model organism for the study of photosynthesis (Harris, 2001). It has recently been selected as one of the model plants to be studied with a focus on understanding photosynthesis and biomass accumulation through a systems biology approach (http://www.goforsys.de/).

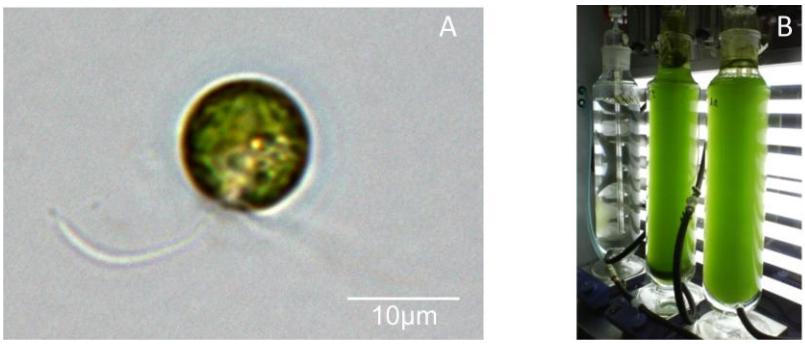


Figure 1-2. The unicellular alga *Chlamydomonas reinhardtii*.

(A) A single Chlamydomonas cell viewed by bright field microscopy. (B) Experimental setup showing Chlamydomonas cells cultured in biofermenters.

Recently, its genome was sequenced (Grossman et al., 2007; Merchant et al., 2007) allowing the annotation of TF and TR coding genes (Riaño-Pachón et al., 2007). The repertoire of TFs and TRs predicted for Chlamydomonas was found to be smaller than the one found in flowering plants, such as Arabidopsis (Riaño-Pachón et al., 2008). However, Chlamydomonas contains all the components necessary for regulating photosynthetic activity and this whole repertoire of regulators can now be analyzed, at least at the transcriptional level. The TF and TR genes that show variations on expression levels in response to environmental or developmental changes may have a regulatory role in a given cellular response. Several gene expression profiling studies of Chlamydomonas grown in different environmental conditions have been carried out. Examples include alteration of light intensity, concentration of CO_2, deprivation of sulfur and nitrogen sources (Im et al., 2003; Grossman et al., 2007; Gonzalez-Ballester et al., 2010). These studies analyzed gene expression responses in a global manner using microarray hybridizations and, recently, deep sequencing technologies. The results of such studies enormously contributed to the discovery of genes important to the cellular

responses in Chlamydomonas. However, only a few TFs of Chlamydomonas have been studied in deep (Table 1-1) and the information about the regulatory role of most of the TFs and TRs in different biotic and abiotic stresses is limited (Grossman et al., 2007; Merchant et al., 2007; Riaño-Pachón et al., 2008).

Table 1-1. Transcriptional regulators functionally characterized in Chlamydomonas.

Protein ID (JGI v.3.1)	TF family	Gene symbol	Biological process in which the protein is involved	Citation
79755	mTERF	MOC1	Mitochondrial light acclimation	(Schonfeld et al., 2004)
148821	SBP	CRR1	Copper and hypoxia induced pathways.	(Kropat et al., 2005)
159133	C2C2-CO-like	CrCO	Growth and synchronicity of cell division	(Serrano et al., 2009)
174777	RWP-RK	MID	Mating type determination	(Ferris and Goodenough, 1997); (Goodenough et al., 2007); (Lin and Goodenough, 2007)
184359	MYB-related	LCR1	Carbon Concentrating Mechanism (CCM)	(Yoshioka et al., 2004)
185576	G2-like	PSR1	Phosphate scavenging during phosphorus depletion	(Wykoff et al., 1999)
186972	C2H2	CCM1-B/CIA5	Carbon Concentrating Mechanism (CCM)	(Fukuzawa et al., 2001); (Xiang et al., 2001)
190458	Sigma70-like	CrRpoD	Transcription of plastid encoded genes	(Bohne et al., 2006)
195786	E2F-DP	E2F1	Cell cycle	(Fang et al., 2006)
195787	E2F-DP	DP1	Cell cycle	(Fang et al., 2006)
195807	RWP-RK	NIT2	Nitrate assimilation	(Camargo et al., 2007); (Schnell and Lefebvre, 1993)
117914	HSF	HSF1	Thermotolerance	(Schulz-Raffelt et al., 2007)

Due to the limited number of TFs and TRs characterized, uncovering gene regulatory networks in Chlamydomonas is a complex task. One way to accelerate the discovery of gene regulatory networks involves targeted gene expression profiling and the analysis of time-series expression data. Furthermore, this type of analysis is useful for further inferring the dynamic of the cellular responses (Gitter et al., 2010). The technologies used for global analysis of time-series gene expression, such as microarrays, have limitations, which reduce their sensitivity to identify variations in the gene expression of low abundant transcripts,

such as TF and TR genes (Sam et al., 2011). With the prediction of TF and TR genes in Chlamydomonas, more sensitive, target techniques, such as quantitative reverse-transcription PCR (qRT-PCR), can now be employed for gene expression analysis. The high sensitivity of this technique, permit the detection of variations in the expression of genes which are present in very low copy numbers in a cell. In principle, the detection of less than five copies (even one copy) of a target sequence is possible (Valasek and Repa, 2005). Furthermore, in order to identify the components and the topology of Chlamydomonas gene regulatory networks, time series gene expression analysis were carried out. We focused in detecting expression level variations of TF and TR coding genes using qRT-PCR, as well as in uncovering regulatory elements for these regulatory genes in a genome-wide scale. This approach may allow getting new and deeper insights into the regulatory mechanisms occurring in Chlamydomonas and it would support future studies on the evolution of these gene regulatory networks in plant species.

1.4 Challenges in identifying protein transcriptional regulators[1]

The experimental identification of the whole repertoire of TFs and TRs through proteomics still represents a challenge due to the fact that many of these proteins are usually in very low copy number per cells. Different strategies have been developed to increase the detection limit of such low abundant proteins. Many of them use proteomic technologies for the exhaustive identification of the whole set of proteins extracted from a cellular compartment, cell or tissue (Gygi et al., 2002; Jorrin et al., 2007; Sandhu et al., 2008). An alternative approach relies in analyzing the 'subproteomes' instead of the whole proteome. This strategy consists, primarily, of the separation of organelles or macromolecular structures of the cell, using methods such as differential purification, immunoprecipitation

[1] Modified from the following publication:

Daniel Martins de Souza, Bruno Menezes de Oliveira, Erich Castro-Dias, **Flavia Vischi Winck**, Ricardo Shiniti Horiuchi, Paulo Aparecido Baldasso, Hugo Takeda Caetano, Natalia Katuchi Pires, Sérgio Marangoni, José Camillo Novello (2008) The untiring search for the most complete proteome representation: reviewing the methods. *Briefings in Functional Genomics and Proteomics*, 7(4):312-321. ©Oxford University Press

or sucrose density gradients, with a subsequent analysis of each organellar fraction using two-dimensional electrophoresis and mass spectrometry or another proteomic techniques (Martins de Souza et al., 2008). Jung and colleagues have reviewed the application of subcellular proteomes in many different organelles and macromolecular structures including mitochondria, lysosomes, peroxisomes, endoplasmatic reticulum, Golgi apparatus and nuclei, indicating the advantages of this approach. The analysis of subcellular proteomes also contributes to the improvement of protein and gene annotation (Jung et al., 2000; Jorrin-Novo et al., 2009). In some cases, the nucleus has been sub-divided into multi-protein complexes such as the nuclear matrix, spliceosome, spindle pole and the nuclear pore complex. Moreover, the subcellular proteome analysis helped to identify protein complexes and delineate a 3D distribution of subsets of proteins in the cell, providing new clues about protein cellular trafficking and organelle specialized functions (Neubauer et al., 1998; Gerner and Sauermann, 1999). The subcellular fractionation and purification of organelles reduces the complexity of the proteome, permitting the identification and the localization of new proteins and may provide indications of possible roles for these proteins. The main drawback of this procedure is the difficulty in obtaining specific subcellular compartments free of contamination from other compartments (Jorrin-Novo et al., 2009). Different methods for the isolation of organelles have been proposed and there is no general procedure that can be applied broadly to all cell types.

Proteomic analysis of plant species had contributed to increasing the knowledge regarding the identification of proteins and their variants, but efforts have to be done for exploiting the full potential of the proteomic approach in the plant field. So far only a few plant species have been reported to be investigated by proteomic approaches (Park, 2004; Jorrin et al., 2007). In Chlamydomonas, the subproteomic studies have contributed in revealing the protein profiles of subcellular components, such as flagellum, photosynthetic apparatus, chloroplast ribosome and mitochondrial oxidative phosphorylation machinery (Stauber and Hippler, 2004; Rolland et al., 2009). The analysis of the subcellular proteome of Chlamydomonas has also identified proteins with important functions on the circadian clock (Wagner et al., 2005). In another plant species, such as *Arabidopsis thaliana*, the subproteome of many different organelles have been described, including mitochondria,

chloroplast, nuclear matrix, among others, and demonstrated to be important also for revealing the similarities of protein composition between organelles from different plant species and with other non-photosynthetic organisms (Jorrin-Novo et al., 2009). The nuclear proteome analysis of *Chlamydomonas reinhardtii* was performed in the present study and contributed to the identification of TFs and TRs, giving us an overview of the protein functional categories present in the nuclear compartment. Our approach revealed the identity of, not only, previously predicted nuclear proteins but also indicated the existence of novel proteins with unknown function.

1.5 Systems biology

With the massive information accumulated from genomic, transcriptomic and proteomic studies it is necessary to use a different approach in order to better understand the relation between the different cellular components and mechanisms which occur simultaneously in a cell. A holistic and broad interpretation of the mechanisms by which molecules interact to produce the functional properties of the cells are the essence of what is called systems biology (Alberguina and Westerhoff, 2005). By combining the capacity of generating datasets from biological systems, simulation based on principles of physics and chemistry and mathematical modeling, it has been possible to create hypotheses which try to explain the biological events in a coordinated and integrated way (Rho et al., 2008). This has been done through the development of models, molecular networks and recognizing the topology of the networks of molecular interactions detected in living cells (Zhang et al., 2007a; Zhang et al., 2008a). The improvement of such integrative analysis would most likely lead to the improvement of the capacity of creating new valid hypotheses and to perform experimental simulations to model biological systems. It is important also to emphasize that none of this would be possible without the development of the field of bioinformatics which represents an important and solid base for the future development of accessible and user friendly sources of information that could be interpreted and explored by all scientists (Zhang et al., 2008a). For establishing the connections between the components of the systems it is important to perform comparative analyses. It is then essential to identify and measure the

elements which constitute the systems (genes, metabolites, proteins, transcripts, etc.) for understanding the fundamental mechanisms underlying the regulation, interaction, metabolism and signal transduction in a cell or organism. A systems biology approach increases the possibilities of interpreting emerging characteristics and to understand the functional properties in terms of how they persist through the nonlinear interactions between the molecules of the biological systems (Alberguina and Westerhoff, 2005). A recent initiative of the German government supported the analysis of plant species in a systems biology approach. Within this project, i.e., *Go*FORSYS (http://www.goforsys.de/), the model organism Chlamydomonas is studied at different levels with the participation of different research groups in a "Core Experiment", in which the groups participate in a common experiment and work with the same sample, with the focus of collecting information of the different mechanisms involved in photosynthesis and biomass production. Within *Go*FORSYS I contributed by investigating expression levels of TF and TR encoding genes under changes of light conditions and carbon dioxide availability towards deciphering important regulators and generating information which will be integrated into further models (Grimbs et al., 2011).

1.6 Aims and structure of the present work

The main objectives of the present work were to identify the components and the structure of gene regulatory networks of *Chlamydomonas reinhardtii*. Different approaches were used to better understand gene expression regulation. I focused my work on the development of strategies for the identification of TF and TR proteins and analysis of TF and TR gene expression in response to different environmental conditions. During the work several questions arose, such as: What protein classes can be found in the nucleus of Chlamydomonas? Is it possible to experimentally identify the transcriptional regulator proteins of *Chlamydomonas reinhardtii*? Are these nuclear proteins sharing specific sequence similarities? Other important questions also appeared during the development of this work. Considering that the two main external environmental elements important for photosynthesis and biomass production are light and CO_2 availability we addressed questions related to photosynthesis-related processes: Which TFs and TRs are light responsive? Which TFs and TRs are CO_2 responsive? What are the possible functions they exert in these cellular responses? My thesis contains six Chapters and is structured as follow: Chapter 1 constitutes a general introduction containing a brief review of the most important concepts related to my work. Nevertheless, each of the following chapters 2, 3, 4 and 5, have its own introduction focused on more specific concepts, description of material and methods, results and discussion and conclusions. Supplemental information related to each chapter is included after bibliographic references as appendices. Outlook and curriculum vitae are included after appendices. Chapter 2, published on Journal of Phycology, describes the results of the optimization of a method for the isolation of nuclei from Chlamydomonas, suitable for proteomic and functional studies. Chapter 3, under revision for publication on the Proteomics journal, contains the results of the nuclear proteome analysis of Chlamydomonas, which revealed the identity of more than 700 proteins, including TF and TR that share sequence similarities. The proteomic dataset is available through the public repository PRIDE (http://www.ebi.ac.uk/pride/) and through a website I build (http://plntfdb.bio.uni-potsdam.de/ChlamyTRI). In Chapter 4, I briefly presented the results

and efforts which contributed to the Core Experiment 1 of the *Go*FORSYS project. On this experiment, several groups were working together to obtain a time series data set of biological Chlamydomonas responses to changes in light conditions. The groups involved analyzed the same sample at different levels, such as: metabolites, polysome distribution, enzyme activities, organelle transcripts, nuclear transcripts, proteome, starch accumulation, among many others. My particular contribution to this work was focused on the identification of the gene expression profile of TFs and TRs. The results of this work revealed new candidate genes which can have a role in transcriptional control on different light conditions. The data produced in this work will be integrated into a global analysis of the Chlamydomonas responses to environmental changes. Chapter 5 reports the results of the analysis of TF and TR gene expression, the genome-wide sequencing of nucleosome-depleted chromatin regions and the reconstruction of gene regulatory network for Chlamydomonas cells under carbon deprivation conditions. A manuscript summarizing these results is in preparation. The final Chapter 6 contains a general discussion of my work highlighting the main achievements and summarizing interesting new findings.

Chapter 2 Isolation of nuclei from Chlamydomonas[2]

[2] The text shown in this chapter was published in the Journal of Phycology®. Please see citation below:

Flavia Vischi Winck, Miroslaw Kwasniewski, Stefanie Wienkoop, Bernd Mueller-Roeber (2011) An optimized method for the isolation of nuclei from *Chlamydomonas reinhardtii* (CHLOROPHYCEAE), *J. Phycol.,* 47, 333–340.

2.1 Abstract

The cell nucleus harbours a large number of proteins involved in transcription, RNA processing, chromatin remodelling, nuclear signalling and ribosome assembly. The nuclear genome of the model alga *Chlamydomonas reinhardtii* has recently been sequenced and many genes encoding nuclear proteins including transcription factors and transcription regulators have been identified through computational discovery tools. However, a comprehensive annotation of nuclear proteins will require support from biochemical and proteomics data. Cellular preparations with enriched nuclei are important to assist in such analyses. Here, we describe a simple protocol for the isolation of nuclei from Chlamydomonas, based on a commercially available kit. The modifications done in the original protocol includes mainly alterations of the differential centrifugation parameters and detergent-based cell lysis. The nuclei enriched fractions obtained with the optimized protocol show low contamination with mitochondrial and plastid proteins. The protocol can be concluded within only three hours and the proteins extracted can be used for gel-based and non-gel-based proteomic approaches.

2.2 Introduction

The unicellular green alga *Chlamydomonas reinhardtii* (Chlamydomonas in the following) has been employed as a model system for the biochemical and molecular analysis of photosynthesis-related processes, flagellar function, phototaxis, lipid biosynthesis, nutritional responses and hydrogen metabolism since many years (Gorman and Levine, 1966; Rochaix, 2002; Stauber et al., 2003). It has recently been chosen by the German *Go*FORSYS consortium (www.goforsys.de) as one of its key model organisms for studying the physiological and biochemical links between photosynthesis and biomass accumulation. As numerous molecular and genetic tools are available for Chlamydomonas, gene function studies are straightforward (Martin and Goodenough, 1975; Harris, 2001). The sequence of the nuclear genome of Chlamydomonas has recently been reported (Merchant et al., 2007), facilitating the analysis of cellular processes in this model alga (Grossman et al., 2007).

In eukaryotic cells, the nucleus plays an important role in gene expression control (Fink et al., 2008). Regulation of gene expression is a multi-facetted process that besides transcriptional regulation at the promoter level involves many other layers of control including RNA splicing and processing, RNA stability control, ribosome assembly, translation and others. We have recently analyzed the Chlamydomonas genome and identified ~40 protein families representing DNA-binding transcription factors (TFs) and other transcriptional regulators (TRs) (Riaño-Pachón et al., 2007). However, a broad coverage of nuclear proteins is most likely not feasible by *in silico* analyses alone. Wet-lab experiments including e.g. proteomic approaches can provide additional information (Heazlewood et al., 2004). For example, studies performed with nuclei isolated from *Arabidopsis thaliana* allowed identifying TFs responding to cold stress (Bae et al., 2003). We envisage that proteomics will also contribute to a better definition of the nuclear proteome in Chlamydomonas. The prior isolation of nuclei in such experiments reduces the complexity of the biological samples facilitating proteomics-based discovery even of low-abundant nuclear proteins (Jung et al., 2000; Martins de Souza et al., 2008).

The interest in analyzing plant nuclear proteins is not recent and several methods for the isolation of nuclei are available (Murray, 1965; Loureiro et al., 2007). Nevertheless, protocols for the isolation of nuclei from plants often have limitations (Steck et al., 1970; Pertoft, 2000). Some of the previous protocols used protoplasts as a source of nuclei. Some others included detergents like Nonidet P40 (NP-40) in the isolation procedure, or introduced deflagellation steps as in the case of Chlamydomonas (Keller et al., 1984; Saxena et al., 1985 ; Hagstrom et al., 1997). NP-40 has been shown to interfere with DNA-protein interaction assays (Nakshatri and Bhat-Nakshatri, 1997). The deflagellation of Chlamydomonas cells is achieved by a pH shock where cells are kept at low pH (4.5) before the pH is increased again (to pH 7.0). Such a pH shift likely affects the cellular proteome profile by e.g. activating calcium uptake (Quarmby, 1996), altering the metabolism of sulfur-deprived Chlamydomonas cells (Kosourov et al., 2003), inducing dramatic changes in nuclear architecture and altering the levels of some RNA transcripts (Colon-Ramos et al., 2003).

For proteomic studies a reproducible method for sample preparation which does not largely affect the protein profile of the cell is needed. Here we present a modified protocol for the

isolation of nuclei from Chlamydomonas. The protocol can be broadly used for nuclear proteomic analyses and can be finished within three hours, with a high level of reproducibility. A first analysis of the proteins extracted from the nuclei isolates by nano-LC/MS/MS indicated the presence of a range of high- and low-abundant nuclear proteins including TFs.

2.3 Material and Methods

2.3.1 Cell culture and growth condition

Chlamydomonas reinhardtii strain CC-503 cw92 mt+ obtained from the Chlamydomonas Genetics Center (Duke University, North Carolina) was cultured in 2-L-Erlenmeyer flasks at 30°C in high-salt medium (HSM) (Sueoka, 1960) at continuous light of 75 µmol·m^{-2}·s^{-1} and continuous shaking at 100 rpm. Cells were grown to the early log phase with a maximum OD_{750nm} = 0.5 (~6 x 10^6 cells·mL^{-1}) and harvested (250 mL) by centrifugation at 3.000 *g* for 2 min at 4°C in either a fixed-angle or swinging bucket rotor.

2.3.2 Isolation of nuclei

The protocol provided in the CelLytic PN kit (Sigma-Aldrich, Steinheim, Germany) was originaly designed for the isolation of nuclei from plant leaves; we optimized it for the isolation of nuclei from Chlamydomonas cells. The original protocol was modified as outlined in the following. Unless otherwise stated, chemicals were provided by the kit. Dithiothreitol (DTT) and protease inhibitor cocktail (catalogue no. P9599) were obtained from Sigma-Aldrich. The cell pellet obtained by centrifugation (see above) was gently resuspended with a Pasteur pipette in 5 mL of 1 x NIB solution (Luthe and Quatrano, 1980; Dignam et al., 1983; Conley et al., 1994; Dyer and Herzog, 1995; Guidez et al., 1998), transferred to a 50-mL plastic conic tube and centrifuged for 2 min at 4°C at 3.000 *g* using an Eppendorf 5810R centrifuge equipped with a swinging-bucket rotor. If not indicated otherwise all subsequent centrifugations were also done using a swinging-bucket rotor. The cell pellet was resuspended in 600 µL of 1 x NIB solution and dropped into a mortar previously cooled with liquid nitrogen. Cells were macerated with a pestle until a fine soft green powder was

obtained. The powder was transferred to a 50-mL plastic conic tube and 10 mL of 1 x NIBA solution (NIB buffer containing protease inhibitors) were added. The homogenized sample was kept on ice for 10 min, filtered through one layer of Miracloth membrane (pore size 22-25µm; Calbiochem, Bad Soden, Germany) into a new 50-mL plastic conic tube, and centrifuged at 1.260 *g* for 10 min at 4°C. The pellet was resuspended in 10 mL of 1 x NIBA solution containing 1 % Triton X-100 and homogenized with a Pasteur pipette to permit lysis of the cytoplasmic membranes. The sample was kept on ice for 10 min and centrifuged at 1.000 *g* for 30 min at 4°C. The lysis was repeated once and the pellet obtained was washed by resuspension in 1 mL of 1 x NIB solution and transferred to a 2-mL plastic tube. The samples were centrifuged at 600 *g* for 10 min at 4°C in a fixed-angle rotor and the washing procedure was repeated. The final nuclei pellet was used for protein extraction or preserved at -80°C in 100 µL storage buffer provided by the CelLytic PN kit.

2.3.3 Protein extraction

Total protein extract was prepared as follows: 1 mL of culture of Chlamydomonas cells grown in HSM medium (Sueoka, 1960) was harvested at the log phase at an $OD_{750\ nm}$ = 0.5 by centrifugation at 3.000 *g* for 2 min at 4°C. The pellet was mixed with 2 mL of extraction buffer (100 mM HEPES-NaOH, pH 7.5, 4 mM EDTA, 5 mM DTT, 0.4 mM PMSF, 10 % glycerol, 0.05 % Triton X-100, 1:100 protease inhibitor cocktail for plants (Sigma-Aldrich), and subjected to cell lysis using a 10-mL Potter Elvehjem tissue grinder. The cell lysate was transferred to a new tube and centrifuged at 12.000 *g* for 10 min at 4°C. Proteins were extracted from the nuclear isolates using the solutions provided in the CelLytic PN kit according to the manufacturer´s instructions and kept at -80°C. Protein concentration was determined using the *DC* Protein Assay Kit (BioRad, München, Germany).

2.3.4 Western blot analyses

Purity of the fractions containing nuclei was investigated by Western blot analysis. Proteins (2 µg) from each sample (total cell extract and nuclear fraction) were loaded and separated on 15 % SDS-polyacrylamide gels and transferred to Immobilon-P PVDF membrane (Millipore, Schwalbach, Germany) by electro-blotting using a mini trans-blot electrophoretic

transfer cell (BioRad). Potential contaminations by plastidic and mitochondrial were tested using polyclonal antibodies raised against chloroplastic subunit H of photosystem I (Psa-H) (Agrisera, Vännäs, Sweden) and mitochondrial outer membrane porin (Heins et al., 1994) antibody kindly provided by Dr. Hans-Peter Brown, University of Hannover, Germany), respectively. The primary antibodies were used at dilutions 1:5000 and 1:1000, respectively. The enrichment for nuclear proteins was analyzed using a polyclonal antibody raised against histone H3 protein (Cell Signaling, Frankfurt am Main, Germany) at a dilution of 1:1000. The presence of ribulose-1,5-bisphosphate carboxylase/oxygenase was analyzed using an antibody raised against its large subunit (RbcL) at a dilution of 1:10000 (Spijkerman, 2008). The antibody was kindly provided by Dr. Anke Koch (University of Potsdam, Germany). Goat anti-rabbit IRDye 800-conjugated secondary antibody (1:5000 dilution; LICOR, Bad Homburg, Germany) was used in all Western blots. The blot images were analyzed with the Odyssey infrared scanner at 800 nm (LICOR).

2.3.5 Two dimensional gel electrophoresis

Frozen nuclei from a 1-L Chlamydomonas culture were thawed and centrifuged for 30 min at 10.000 *g*. The pellet was resuspended in 6 vol. (160 µL) of sample preparation buffer (9 M urea, 2 % ampholytes and 70 mM DTT). After incubation for 30 min at room temperature and centrifugation for 45 min at 15.000 *g* the supernatant was removed and precipitated with trichloracetic acid (TCA). Two-dimensional gel electrophoresis (2DE) was performed by Proteome Factory (Berlin, Germany). Thirty-six µg of protein (the whole precipitate) was applied to vertical rod gels (9 M urea, 4 % acrylamide, 0.3 % PDA, 5 % glycerol, 0.06 % TEMED and 2 % carrier ampholytes [pH 2-11], 0.02 % APS) for isoelectric focusing (IEF) at 8.820 Vh in the first dimension. After IEF, the gels were incubated for 10 min in equilibration buffer containing 125 mM trisphosphate (pH 6.8), 40 % glycerol, 65 mM DTT, and 3 % SDS, and subsequently frozen at -80°C. SDS-PAGE gels (20cm x 30cm x 0.1cm) containing 375 mM Tris-HCl buffer (pH 8.8), 12 % acrylamide, 0.2 % bisacrylamide, 0.1 % SDS and 0.03 % TEMED were prepared for the second dimension. After thawing, the equilibrated IEF gels were immediately applied to the SDS-PAGE gels. Electrophoresis was performed using 140 V for 5h 15 min until the front reached the end of the gel. After 2DE separation the gels were

stained with FireSilver (Proteome Factory, Berlin, Germany). The 2DE gel was digitized at a resolution of 150 dpi using a PowerLook 2100XL scanner (Umax, Willich, Germany) with transparency adapter. Two-dimensional image analysis was performed using ProteomWeaver software (Definiens, Munich, Germany).

2.3.6 Identification of nuclear proteins

To check for the enrichment of nuclear proteins in nuclear extracts proteins were separated by SDS polyacrylamide gel electrophoresis. Proteins isolated from the gel stained with Coomassie Brilliant Blue G (Sigma-Aldrich) were identified using a Reflex II MALDI-ToF mass spectrometer (Bruker, Bremen, Germany). The peptides were extracted from the protein bands using the protocol described by Shevchenko et al. (1996). α-Cyano-4-hydroxy cinnamic acid in a solution containing 7 vol. acetonitrile and 3 vol. 0.1 % TFA was used as matrix. Mass spectrometry was performed in reflector mode using the monoisotopic peaks produced by the autolysis of trypsin (m/z 2163.06) and matrix ion (m/z 568.14) as calibration peaks. Protein identity was determined using the list of monoisotopic peaks obtained from the analyses of the proteolytic peptides. The search was done using the Mascot tool (http://www.matrixscience.com/) against the MSDB 20060831 database. Search parameters included a maximum of two missed enzymatic cleavages and a peptide error tolerance of 10 ppm. Protein hits presenting the best score considering the significance threshold (p-value <0.05) were considered the best match.

2.3.7 Protein pre-fractionation and nano-LC/MS/MS

Proteins extracted from two independent nuclear isolates were pre-fractionated according to the following procedure. The isolated nuclei were resuspended in urea-SDS buffer (50 mM Tris, pH 7.8, 8 M urea, 2 % SDS) and sonicated for 30 min in a bath sonicator (model Sonopuls; Bandelin, Berlin, Germany). The samples were centrifuged at 10.000 *g* for 10 min and the supernatant was transferred to a new tube. One mL Trizol was added and the proteins were extracted as described by Lee and Lo (2008). The protein pellet was resuspended in 20 µL of 1 x SDS-PAGE buffer and separated in a 12 % SDS-polyacrylamide gel (Laemmli, 1970). The gel lane was cut in 10 pieces and each fraction was subjected to

protein in-gel digestion and desalting as previously described (Shevchenko et al., 1996; Rappsilber et al., 2003). The protein digest was loaded onto a pre-column using a 1D nano flow LC system (Agilent, Böblingen, Germany). A monolithic column (Merck, Darmstadt, Germany) of 15 cm length and 0.1 mm internal diameter was coupled to an Orbitrap LTQ XL mass spectrometer (Thermo Electron, Dreieich, Germany) operated in a data-dependent mode. Peptides where eluted during a 60-min gradient from 5 % to 100 % MetOH / 0.1 % formic acid (FA) with a controlled flow rate of 0.5 µL per min. After mass spectrometric analyses, mgf files were created from the raw data and searched against an in-house developed protein database described by May et al. (2008). The protein identification was performed using Mascot software (http://www.matrixscience.com) applying the following criteria: maximum of two missed cleavages, maximum peptide error tolerance of 10 ppm, maximum fragment ion mass error tolerance of 0.8 Da. Variable modifications due to carbamidomethylation and oxidation were accepted. Peptides with an ion score greater than the calculated threshold ($p < 0.05$) were assigned as significant protein hits.

2.3.8 Microscopy

Bright-field and fluorescence microscopy were used to check intact cells and the integrity of the nuclei isolated. The nuclei isolates were resuspended in storage buffer provide by the CelLytic PN kit (Sigma-Aldrich) and stained for 10 min with Vectashield Fluorescence Mounting Medium containing 4,6-diamino-2-phenylindole (DAPI) (Vector Laboratories, Burlingame, USA). The intact cells were stained directly with DAPI for 10 min before microscopy. The slides were analyzed with an Olympus BX-51 fluorescence microscope and Olympus BX 61 epi-fluorescence microscope motorised (Olympus, Hamburg, Germany).

2.4 Results and Discussion

Initially, we used a commercially available kit, CelLytic PN from Sigma-Aldrich, for the isolation of nuclei from Chlamydomonas, and followed the manufacturer´s protocol that was originally established for the isolation of nuclei from plant leaves. The Chlamydomonas cells were disrupted by maceration in liquid nitrogen and suspended in the buffer volumes

indicated by the user´s guide. The centrifugation steps were followed without modifications in standard fixed centrifuge rotors. The CelLytic PN protocol includes a final centrifugation in either a sucrose and or Percoll gradient to increase the purity of the nuclei. However, this original procedure turned out to be unsuitable for the isolation of nuclei from Chlamydomonas cells. After the final centrifugation step the supposed nuclear fraction at the bottom of the centrifugation tube remained as a double-layered pellet that consisted of a soft, white fraction containing the nuclei and an upper green phase containing intact cells and chloroplasts that were not sufficiently removed from the nuclear preparation (data not shown). We thus optimized the available protocol to allow efficient isolation of nuclei from Chlamydomonas cells for proteome studies. Our protocol of isolation of nuclei can yield on average at least 3.5 $x10^6$ nuclei per extraction, starting from 250 ml culture volume. This finding was revealed by counting the nuclei with Thoma counting chamber. The entire procedure can be completed within three hours, is highly reproducible (Figure appendix A-1) and is suitable for gel-based or non-gel-based proteomic studies as shown by the possibility of performing high-quality 2DE gels (Figure appendix A-2) and nano-LC/MS/MS protein identification (Table 2-1). The procedure can also be performed for Chlamydomonas cw15 mutant and cc1690 strains (data not shown). A schematic outline of the protocol is given in Figure appendix A-3.

2.4.1 Cell disruption

We tested several methods for the disruption of Chlamydomonas cells including homogenization with a Potter Elvehjem tissue grinder, cell disruption by bead milling, cell suspension bio-nebulization with a BioNeb system (Glas-Col, Terre Haute, USA), and cell maceration using liquid nitrogen. In our hands, maceration in liquid nitrogen appeared to be the most simplest and efficient procedure for nuclei isolation, resulting in much less intact cells than with the other methods tested (data not shown). Keeping the fraction of intact cells low is important as they turned out to precipitate together with the released nuclei forming the upper green phase at the final centrifugation step.

Table 2-1. Examples of proteins identified by LC mass spectrometry in nuclei isolates.
Besides core histones also other nuclear-localized molecules including transcription factors and nucleolar proteins were found in the nuclear isolates.

Protein ID[#]	Functional annotation[&]	Gene ID[#]	Protein sequence covered by MS analysis (%)
161836	Histone H2B	C_270156	85.2
69	Histone H2A	fgenesh1_pm.C_scaffold_3000012	36.8
102205	Histone H2B	e_gwH.18.188.1	67.1
105	Histone H3	fgenesh1_pm.C_scaffold_3000048	50.4
97501	Histone H4	e_gwH.5.178.1	32.4
184356	HMG*	Estext fgenesh2 kg.c 280024	28
149734	ZnFinger C2H2*	Chlre2_kg.scaffold_27000003	9.7
176419	Nop10	fgenesh2_pg.C_scaffold_36000079	92.2
146324	Nop52	Chlre2_kg.scaffold_1000148	10.9

#Protein and gene sequences were obtained from JGI (http://genome.jgi-psf.org/Chlre4/Chlre4.home.html) and NCBI (http://www.ncbi.nlm.nih.gov/) databases.
&Functional annotation data are available at the JGI and NCBI databases.
*Catalogued in the Plant Transcription Factor Database (http://plntfdb.bio.uni-potsdam.de/v3.0/).

2.4.2 Detergents

In the protocol of Keller et al. (1984) Nonidet P40 was used as detergent for cell lysis; we replaced it by Triton X-100. As indicated previously by Smith and Wells (1984), 1 % Triton X-100 has the capacity to solubilize 20 - 30% of the nuclear envelope of eukaryotic cells. Triton X-100 almost completely removes the outer nuclear membrane whereas the inner nuclear membrane remains largely intact (Cox and Emili, 2006). Thus, using Triton X-100 as a detergent allows preserving internal components of the nucleus during the isolation procedure. The structural integrity of the nuclei isolated was verified by microscopy and is shown in Figure 2-1. Triton X-100 is a non-ionic detergent that can easily be removed from the samples during the clean-up procedure carried out prior to mass spectrometric analysis.

Therefore the analysis of proteins obtained from the nuclei by gel-free or gel-based proteomic techniques is straightforward (see below).

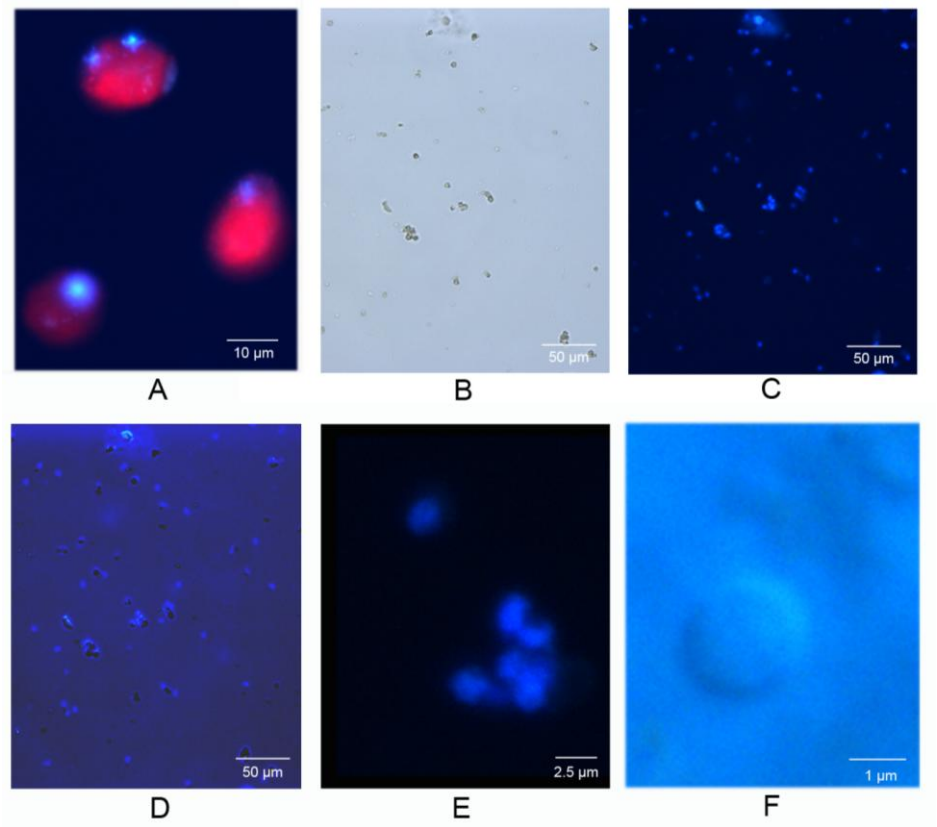


Figure 2-1. Analysis of the structural integrity of the nuclei isolated from Chlamydomonas.

(A) Fluorescence microscopy of intact Chlamydomonas cells stained with DAPI. (B) Bright field microscopy of nuclei isolated from Chlamydomonas. The same sample stained with DAPI analyzed by fluorescence microscopy is shown in (C), and the merged image is shown in (D). DAPI stained nuclei are shown 400-fold magnified in (E). Stereomicroscopy of the nucleus isolated is shown in (F).

2.4.3 Immunochemical assays

To determine the presence of non-nuclear proteins in the nuclear isolates, Western blot analyses were performed (Figure 2-2). The results indicated very low contamination of the nuclear fractions with chloroplast thylakoid membrane protein PsaH and mitochondrial outer membrane porin (Heins et al., 1994; Amunts et al., 2005). A strong enrichment for nuclear protein histone H3 was evident in isolated nuclei. Western blot analysis revealed that RbcL protein was not completely removed from the nuclear fraction during the isolation

of nuclei. However, the relative abundance of RbcL was proportionally reduced in comparison to the abundance of the nuclear proteins, demonstrating a substantial nuclear enrichment (not shown).

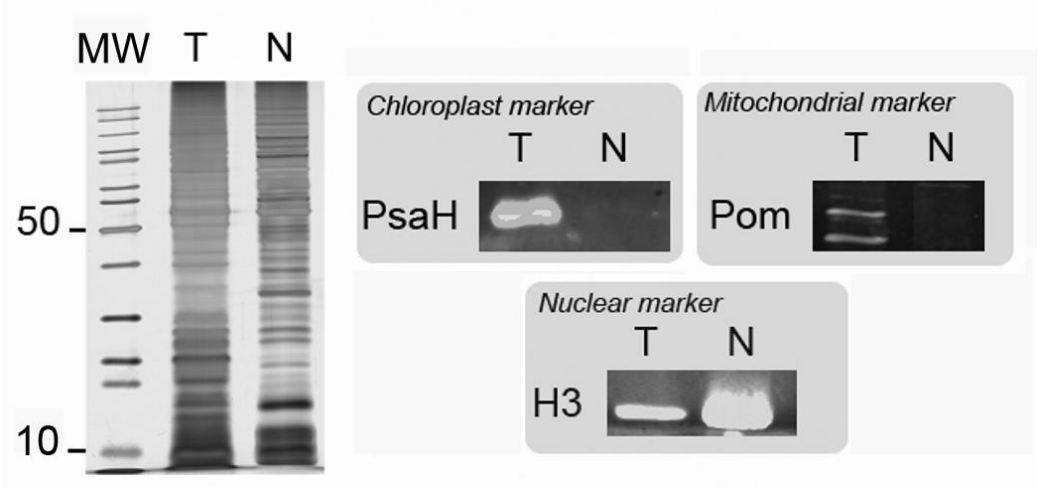


Figure 2-2. Immunochemical detection assays.

Total protein extract (T) and nuclear protein extracts (N) were separated in a 15% SDS-polyacrylamide gel as shown on the silver-stained gel on the left. For Western blot analyses, 2 µg of protein was loaded in each lane. After the transfer to a PVDF membrane and incubation with the primary and secondary antibodies, the markers were visualized by fluorescence imaging. The following marker proteins were analyzed: PsaH for chloroplasts; outer membrane porin (Pom) for mitochondria, and histone H3 (H3) for nuclei. MW, molecular weight marker (numbers indicate kDa).

2.4.4 Differential centrifugation

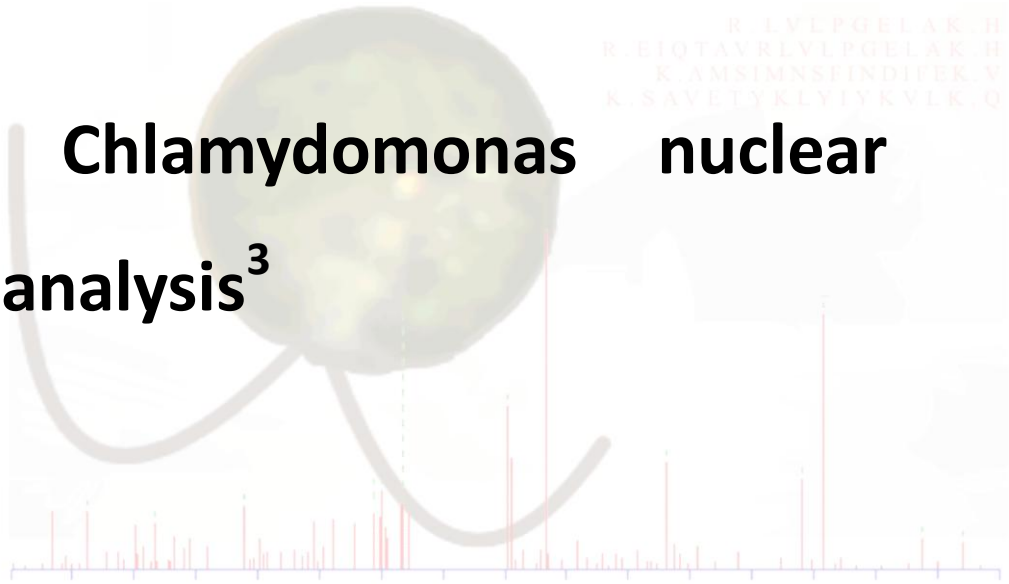
Many protocols established for the isolation of organelles include differential centrifugation steps. While optimizing the isolation protocol for Chlamydomonas nuclei we also modified some of the differential centrifugation parameters (cf. Materials and Methods). We found that low-speed centrifugation and extended centrifugation times were necessary for a successful recovery of the nuclear fraction at the bottom of the centrifugation tube. Another change we included in the protocol was the use of a swinging-bucket rotor carrying the conical tubes.

2.4.5 Identification of nuclear proteins

Proteins extracted from the nuclei were analyzed by MALDI-ToF and nano-LC/MS/MS. By performing pre-fractionation of proteins using 1D SDS-PAGE and nano-LC/MS/MS we were able to identify more than 100 proteins in a preliminary study, from a starting amount of 30 µg of nuclear protein extract. A detailed analysis of the nuclear proteome of Chlamydomonas is currently being conducted. Table 2-1 gives a list of some of the proteins identified. The proteins found are known to be present in different levels in the cell nucleus and indicate enrichment not only of highly abundant proteins such as histones but also of low-abundant proteins like transcription factors and other transcription regulators. Thus, the easy-to-implement nuclei isolation protocol reported here will be of great relevance for the identification of further constituents of the Chlamydomonas nuclear proteome in the future.

2.5 Conclusions

We present a straightforward and easy-to-implement protocol for the isolation of nuclei from Chlamydomonas cells. The protocol was optimized based on a commercial CelLytic PN kit from Sigma-Aldrich. Critical modifications of the protocol were introduced, including changes in the type of the centrifuge rotor as well as centrifugation speed and times. Efficient cell membrane lysis was obtained using Triton X-100 as detergent. The whole procedure can be finished within three hours, is highly reproducible and broadly applicable to proteomic studies. The optimized protocol was developed in order to avoid dramatic changes of the proteome profile. Noteworthy, it facilitates nuclear proteome analyses by gel-free or gel-based methods, which will aid to the characterization of nuclear proteins and gene regulatory networks in Chlamydomonas.



Chapter 3 Chlamydomonas nuclear proteome analysis[3]

[3] The following text represents a slightly extended version of a manuscript entitled "The nuclear proteome of the green alga *Chlamydomonas reinhardtii"* in revision by the journal *Proteomics* for peer review and publication. *Authors:* **Flavia Vischi Winck**, Diego Mauricio Riaño-Pachón, Frederik Sommer, Jens Rupprecht and Bernd Mueller-Roeber.

3.1 Abstract

Nuclear proteins play a central role in regulating gene expression. Their identification is important for understanding how the nuclear repertoire changes over time under different conditions and how that translates into different transcriptional states. Nuclear proteins are often underrepresented in proteomic studies due to their frequently low abundance. So far, just a few studies describing the nuclear proteome of plant species have been published. Recently, the nuclear genome sequence of the unicellular green alga *Chlamydomonas reinhardtii* has been obtained and annotated, allowing the development of further detailed studies for this organism. However, a detailed description of its nuclear proteome has not been reported so far. Here we present an analysis of the nuclear proteome of the sequenced Chlamydomonas strain cc503. Using LC-MS/MS we identified 1.108 proteins from nuclei isolates with a maximum 1% peptide spectrum false discovery rate. Besides well-known proteins (e.g. histones), transcription factors and other transcriptional regulators (e.g. tubby and HMG) were identified. We performed the prediction of the subcellular localization of these proteins using computational tools and we calculated their codon usage. Our data revealed the disagreement between different prediction approaches and suggests the needs for improvement of the computational prediction programs. The presence of protein motifs in nuclear proteins was investigated by computational tools and specific over-represented protein motifs were identified. This study provides new insights into the complexity of the nuclear environment, and reveals novel putative protein targets for further studies of nuclear mechanisms.

3.2 Introduction

The analysis of sub-proteomes in plant species has revealed the identity of many proteins from different organelles and has assisted in the recognition of their subcellular location, disclosing new molecular targets for functional studies (Rolland et al., 2009). The proteome analysis of different cellular organelles, including mitochondria, chloroplasts and nuclei, has already been carried out for *Arabidopsis thaliana*, *Oryza sativa* and *Medicago truncatula*

aided by to the publication of their genomic sequences. Similarly, with the availability of the genomic sequence of the green alga *Chlamydomonas reinhardtii* it has been possible to study its proteins in a high-throughput manner. The sub-proteomes of some of its cellular entities, including chloroplast ribosomes, flagellae, and eyespots, have previously been published, but a detailed description of the nuclear proteome was not available so far (Rolland et al., 2009; Wagner et al., 2009; Terashima et al., 2010). The functional specialization of proteins is to a great extent related to their intracellular organization (Gitai, 2005). It is therefore important to study the cellular sub-proteomes to better understand which proteins and mechanisms are involved in organelle communication and how these processes are regulated. Furthermore, analyzing the nuclear proteome is of great interest because of its close connection with the processes of gene expression regulation. Many DNA-binding proteins directly or indirectly involved in the regulation of gene expression have been shown to be at least transiently located in the nucleus (Casal and Yanovsky, 2005). *C. reinhardtii* is a unicellular green alga that has been used as model for extensive analysis of the photosynthetic machinery; recently it has also been recognized as a good experimental model in systems biology-oriented approaches (Merchant et al., 2007; May et al., 2008; Christian et al., 2009; Renberg et al., 2010; Wienkoop et al., 2010). So far little is known about the proteins involved in the regulation of nuclear gene expression in *C. reinhardtii*. The analysis of the nuclear proteome performed in the present study has revealed the identity of many proteins related to transcriptional control (e.g. transcription factors and other regulators), nucleosome biogenesis (nucleolar proteins and ribosomal proteins) and proteins with previously unknown functional annotation. Many of these proteins involved in regulatory processes have a low expression level and are supposed to be present in the nuclear compartment in very low numbers (Gygi et al., 1999). This can be partially explained by the low codon bias hypothesis (Gygi et al., 2000). In many highly expressed protein coding genes of eukaryotes and prokaryotes, not all possible codons are found, which indicates the existence of a bias for preferential use of some particular codons. In many gene sequences it was observed that they do not have all the 64 codons and only some of them are present. This can be explained by the use of alternative codons which are synonymous, which means a different codon triplet coding for the same amino acid can be

present in a gene sequence. The frequency of occurrence of synonymous codons determines what is called codon usage bias (Wright, 1990). Our findings revealed proteins of putative regulatory function (e.g., nucleolar phosphoprotein, protein kinase), with a relatively low codon bias, conversely, a higher codon bias was found among some well known highly expressed proteins (e.g., histones, tubulin). Another aspect of interest is protein localization; however, for most of the regulatory proteins from Chlamydomonas there is no experimental evidence for their subcellular localization. Many efforts have been done for the improvement of computational tools able to characterize *in silico* the protein's subcellular location (Casadio et al., 2008). Nevertheless the publicly available methods differ in their underlying biological motivation, computational method used, localization coverage and reliability (Donnes and Hoglund, 2004). Many methods use the available protein sequence information from related species for their predictions and use different algorithms for comparing those protein sequences and inferring their subcellular location (Casadio et al., 2008). Identifying the protein subcellular location is important for a better understanding of the role of proteins in organisms. However, experimentally this is a very time consuming and costly task. With the increasing number of available genomic sequences it is important to improve the predictors because many of the inferred functions of the proteins will most likely come from the use of such tools. The algorithms for predictions now include many other features combining evolutionary information and text mining (Fyshe et al., 2008). Besides the difficulty of identifying the correct subcellular location of proteins through prediction methods there are other factors limiting protein characterization in a high throughput manner. Usually, the functional characterization of proteins involves the cloning and overexpression of a certain gene and the analysis of the expressed protein. The sub-proteome analysis is an important alternative for the identification of a large number of proteins from a cellular compartment and it helps to improve protein functional annotation.

3.3 Material and methods

3.3.1 Isolation of nuclei

Nuclear-enriched fractions were obtained as previously described (Winck et al., 2011) from five independently grown cultures of *Chlamydomonas reinhardtii* (strain cc503) cultured in 2-L flasks in photoautotrophic and temperature controlled conditions. Cells were cultured in H5P medium (modified TAP medium without acetate and with 5mM HEPES as buffering agent) under continuous light (100µE m^{-2} s^{-1}) and constant supply of air with 2% CO_2 (gas mixture was provided by B-DCU Biostat fermenter (Sartorius, Goettingen, Germany) and controlled via GC system (Agilent Technologies 3000 micro-GC, Waldbronn, Germany). All steps of the nuclei isolation procedure were performed at 4°C. The cells were harvested from a 1-L culture by centrifugation for 2min at 3000 *g*, suspended in 600 µL of NIB solution and grinded in a cooled mortar. The powdered material was suspended in 10 mL of NIB solution, filtered through a Miracloth membrane (22-25µm; Calbiochem, Bad Soden, Germany) and centrifuged for 10min at 1260 *g*. The cell pellet was homogenized with 1% Triton X-100 in NIBA solution and centrifuged for 30min at 1000 *g*. This procedure was repeated once and the cell pellet was washed twice by homogenization in 1mL of NIB solution and centrifuged for 10min at 600 *g*. Five nuclei-enriched fractions were used for protein extraction as described by Lee and Lo (Lee and Lo, 2008) followed by a pre-fractionation by SDS-PAGE. The gel lane was cut into 10 pieces and each fraction was subjected to protein in-gel digestion with Trypsin and desalted with STAGE tips with C18 disks (3M, Neuss, Germany) (Shevchenko et al., 1996; Rappsilber et al., 2003).

3.3.2 Mass spectrometry

The peptides were separated and analyzed by LC/MS-MS using the Proxeon nEASY HPLC system coupled to an LTQ XL Orbitrap hybrid mass spectrometer (Thermo Scientific, Dreieich, Germany). RP-HPLC separation of peptides was performed using 75µm internal diameter Integrafrit columns (New Objective, Woburn, MA, USA) packed with 15 cm of Reprosil-Pur C18-AQ 3 µm particles (Dr. Maisch, Ammerbruch-Entringen, Germany). One-hour gradients at 350 nl/min of A (2% 2-propanol, 0.5% acetic acid) and B (80% acetonitrile,

0.5% acetic acid) ramped from 0% B to 40% B in 45 min and from 40% B to 80% B in 15 min. In between, analysis wash runs were performed to ensure no carry-over from sample to sample. The LTQ XL Orbitrap was operated in data-dependent mode taking one full MS scan (300 m/z – 1800 m/z) in the Orbitrap analyzer at a resolution of 30000 and an AGC target of 1000000, followed by 5 consecutive MS2 scans in the LTQ ion trap on the most intense ions. Parent ions with unassigned and 1+ charge state were rejected; fragmented parent ions were excluded from MS2 selection for 30 sec.

3.3.3 Protein identification

After mass spectrometry peak lists were generated from the raw data using the DTAsupercharge plugin from MSQuant software V 1.5 (http://msquant.sourceforge.net/). Default parameters were used and Mascot generic files (mgf) were created. Protein identification was performed by peptide searching against an in-house developed protein database that contains annotated protein sequences from the following sources: Chlamydomonas nuclear genome assembly versions 3.1 and 4.0 and AUGUSTUS v5 gene predictor obtained from the Department of Energy/Joint Genome Institute (DOE/JGI; http://genome.jgi-psf.org/); chloroplast genome (GenBank accession number: NC_005353) assembled by the Boyce Thompson Institute; mitochondrial genome (NC_001638) assembled by Dalhousie University. A target database was established as a non-redundant protein dataset, i.e. when two or more 100% identical protein sequences were found in the original files, only one was kept with a new protein identifier and annotated proteins that had only a single amino acid were excluded. Protein sequences from this database were concatenated to a second dataset containing the protein sequences of the target database randomly shuffled (decoy database) which was achieved using the decoy.pl script (Matrix Science, London, UK). The combined database containing both types of protein sequences (target-decoy database; 113.676 entries) is freely available from the datasets session of our website (http://plntfdb.bio.uni-potsdam.de/ChlamyTRI). Protein identification was performed using Mascot search engine (v. 2.2.03; Matrix Science) applied to target-decoy database. The search parameters were: maximum of two missed cleavages, peptide error tolerance of 10 ppm, maximum fragment ion mass error tolerance of 0.8 Da. Fixed

modifications on cysteine carbamidomethylation and variable modifications on methionine oxidation were accepted. Proteins containing peptides with an ion score greater than the calculated threshold (p<0.05) were regarded as significant hits. Those protein hits were subjected to a False Discovery Rate (FDR) analysis using the program Mayu (Reiter et al., 2009). A maximum Peptide Spectrum Matches (PSM) FDR of 1% was selected as a filtering parameter and the protein hits included in this group were selected for further analysis.

3.3.4 Prediction of protein subcellular location

Prediction of subcellular location was performed with the computational tools BaCelLo (Pierleoni et al., 2006), NucPred (Brameier et al., 2007), Predict NLS (Cokol et al., 2000) and WolfPSORT (Horton et al., 2007). The predictions were obtained only for the proteins sequences annotated from the nuclear genome sequence v3.1, mitochondria and chloroplast genomes of *C. reinhardtii*.

3.3.5 Codon usage analysis

ENC (effective number of codons), correspondence analysis (COA) and codon bias index (CBI) were calculated using the program CodonW 1.3 (John Peden, http://sourceforge.net/projects/codonw/). The analysis was performed with the CDSs obtained from the Chlamydomonas genome annotation v.3.1 (DOE/JGI), corresponding to the total proteome (9625 CDSs) and to the nuclear proteome dataset (684 CDSs) of *C. reinhardtii.* The COA analysis was used to calculate and to generate the list of putative optimal codons.

3.4 Results and discussion

3.4.1 Protein identification

By applying the protein identification and filtering parameters described in Material and methods, 1.108 protein hits were identified; protein hits and peptide coverage data are available through our website (datasets menu; Supplemental files 1 and 2; http://plntfdb.bio.uni-potsdam.de/ChlamyTRI/) which includes information about the

correspondent scores, peptides identified and protein coverage. The processed raw data (mgf files), and the list of proteins identified using Mascot are available in the PRIDE repository (Acc 13764) (http://www.ebi.ac.uk/pride/) (Barsnes et al., 2009). Protein sequences contained in the target protein database described above were subjected to the Mercator application (http://mapman.gabipd.org/web/guest/app/mercator) for automated functional protein annotation. Domain identifications and the possible classification of the proteins into families were performed by searching against the Pfam database (V. 23, http://pfam.sanger.ac.uk/). Pfam domains were assigned to approximately 52% of the proteins from the target protein database (data not shown).

3.4.2 Motif analysis

To test whether the nuclear proteins identified had common sequence features, overrepresented protein motifs were identified using the MEME suite (v.4.3.0, http://meme.nbcr.net/) (Bailey et al., 2006). From the 1.108 proteins analyzed 386 showed at least one conserved motif. In our search the twenty motifs found were subjected to alignment against the target protein database using the tool MAST (http://meme.sdsc.edu/meme4_3_0/cgi-bin/mast.cgi) (Bailey and Gribskov, 1998). This analysis allowed the identification of additional proteins sharing the same protein motif and to estimate the enrichment of the protein motifs in our nuclear proteome dataset. The results from the MAST analysis indicate that some of the motifs aligned with many other protein sequences from the total target protein database. Interestingly, other motifs were found to align with only a few other proteins suggesting that those motifs are specific of a small group of proteins. These proteins are listed in Supplemental file 2 (datasets menu; http://plntfdb.bio.uni-potsdam.de/ChlamyTRI/), which includes information about the motifs and their corresponding p-values, together with a predicted functional annotation from MapMan (Thimm et al., 2004) and InterPro (Zdobnov and Apweiler, 2001). The estimation of the enrichment of those motifs in the nuclear proteome dataset was done by comparing their frequencies in the nuclear dataset and in the target protein database (Table 3-1).

Table 3-1. Estimation of enriched protein motifs in the Chlamydomonas nuclear proteome dataset.

Highly significant protein motifs (p<0.05) found using MEME were subjected to a motif alignment search using the MAST tool revealing the frequency and existence of similar motifs in the whole cellular proteome (target protein database). Some motifs aligned to only few protein sequences indicating that they are more specific to a small group of proteins and are enriched in the nuclear proteome dataset. The number of proteins of the complete proteome which aligned to each significant motif is shown.

Motif	Motif regular expression	Motif frequency in the nuclear proteome [a]	Motif frequency in the target protein DB (E-value <10^{-5})[b]	Enrichment[c]
1	F[RK][AS]GKAPVL[VIL]ATDVAARGLD[IF]P[DG][VI]DxV[IV][NQ][YF]DFPx[DT]x[EA]DY[VI]HR[IV]GRT[GA]RAG	3	156	0,98658359
2	SIMNSFINDIFEKVATEASKLSRYNKKPTVTSREIQTAVRLVLPGELAKH	5	50	5,13023466
3	[AS]GRD[VL][ILV][GA]IAxTGSGKT[LA]AFLLPM[LI]ER[LI]LAQx	9	130	3,55170092
4	KS[AK]VETYKLYIYKVLKQVHPDTGISSKAM	6	49	6,28191999
5	RGxLxLS[QR][ILV]RYLVLDEADR[ML]LD[ML]GFEPQ[IL]	8	95	4,32019761
6	PxA[LV][VI]L[AS]PTRELAxQIxE[VE]A[ER]	6	105	2,93156266
7	D[KT][FS]AELK[VLT]KE[IV]KNGRLAM[FVL][AS][FM][LF]GFF[VA]QA[IA][VA]TGKGP[IL][AQ]NL[DA][AD]H	8	95	4,32019761
8	[IV][VM]RAL[PR]P[EQ]RQTL[LM]FSAT[MW][PT]KE	5	77	3,33132121
9	VD[TN]MI[VI]QAI[GA]LLD[DT]LDK[ED][LV]NT[YF][AV]MRVREWY[GS]WHFPE[ML][TV]KIV[TN]DN[IY]QYA[KR][CL]	3	8	19,23838
10	TGE[LF][AP]GDYG[FW]D[PT][LA]GL[GS][AK]DPET[LF]K[RW]YR[EQ][AL]ELIH[AG]RWAML[GA][AV][AL]G[IC][LI]	5	97	2,64445085
11	QQQQQQQQQP	26	1582	0,84314855
12	[ED][ED][ED][ED][ED][ED][ED][ED]EE[ED]	98	226	22,2461503
13	MAPK[KR]DEKPAT[QA]EAGAEAPAKAEAKPKAEKAAKK	5	40	6,41279332

	AKKEPSKKAAKEPKGD			
14	[EQ]LL[TN][QE][ML]DGF[DE][GA]N[GT][GN][VI][KI]V[IL][MA]ATNR[PA]D[IV]LDPALLRPGR[LF]DR[KQ][IV]	2	82	1,25127675
15	[HN]P[ED][LK][FY]TALG[IA]K[PI]PKG[VC]LL[YV]GPPGTGKTL[LI]A[RK]A[VI]A[GH][EH][TA][DG]A[PT]F	5	111	2,31091651
16	RAL[RA]RG[VP][HED]V[VL]V[AG]TPGRLID[LH]L	5	82	3,12819186
17	[IFL][SRK][VC][ASV][GA][SP]E[FL]V[QE][KMV][FY][VI]G[EV][GS]A[RS][MR]VR[DE]LFE[KLM]A[KR][ES][KHN]AP[CS][IV][IV]F[IFM]DE[IV]D[AS][IV][GA][RGT][AQ]R[FG][DAS]	4	100	2,05209386
18	PPPAP[PA]PAPPPAPPPPxPPPPPPPPPPPP	51	2206	1,186047
19	[SR][LVI][FY]V[GR][NG]L[PS]xD[VA][TRS]ED[DQ][LIV]RELF[SE]K[FY]GR[VI][KR]S[VI]R[VL]ARD[PA]	20	146	7,02771871
20	GGGGGGGG	112	0	ND

a) Motifs found using MEME (http://meme.sdsc.edu/meme4_4_0/cgi-bin/meme.cgi)

b) Alignment matches found using MAST (http://meme.sdsc.edu/meme4_4_0/cgi-bin/mast.cgi)

c) Enrichment of motifs in the nuclear proteome dataset was calculated using the following equation: (Nm/Nnp) / (Nam/Ntp), where: Nm is the number of proteins containing a motif X found in the nuclear proteome dataset; Nnp is the total number of proteins contained in the nuclear proteome dataset; Nam is the number of proteins from the total proteome dataset (target DB) which significantly aligned to the motif X; Ntp is total number of proteins contained in the total proteome dataset (target DB). Values closer to 1 indicate non-enrichment and higher values indicate an enrichment of the motif X in the nuclear proteome dataset.

3.4.3 Protein functional categories of the nuclear proteome

Some of the motifs identified are domains of well known proteins, including histones and helicases. However, other poorly characterized motifs were found. These proteins may share similar cellular locations or functions, but further experiments are needed for clarification. To identify significantly over-represented gene ontology (GO) terms in the dataset of identified proteins, we first mapped GO terms to proteins by exploiting the protein domains identified in the Pfam database, using the pfam2go mapping file (http://www.geneontology.org/external2go/*)*. We then tested for overrepresented GO terms using BiNGO (Maere et al., 2005) (with Bonferroni family wise error rate, FWER, correction and significance level threshold of 0.05) and graphically visualized the results using Cytoscape (Cline et al., 2007) (http://www.cytoscape.org/). The overrepresented GO terms for biological process, molecular function and cellular component ontologies for the nuclear proteome dataset were, respectively: gene expression (GO:0010467, corrected *p*-value 7.1e-45), structural constituent of ribosome (GO:0003735, corrected *p*-value 1.9932e-42), and ribonucleoprotein complex (GO:0005732, corrected *p*-value 2.7515e-43). Furthermore, a total of 241 out of the 386 nuclear proteins which contained significant protein motifs were also analyzed using BiNGO and Cytoscape. The most significantly overrepresented terms for biological process, molecular function and cellular component were, respectively: ribosome biogenesis (GO:0042254, *p*-value 1.3e-10), nucleic acid binding (GO:0003676, *p*-value 4.8e-44), and nucleolus (GO:0005730, *p*-value 3.7e-08). The graphical visualization of the networks of overrepresented GO terms for both protein data sets, i.e. the complete nuclear proteome data set and the sub set containing only proteins with conserved motifs, are shown in Figure 3-1 and Figure 3-2, respectively.

Figure 3-1. Network of overrepresented GO terms in the nuclear proteome dataset.

Analysis of overrepresented GO terms for the proteins identified was done using BiNGO. Terms were extracted based on the category ´molecular function´ with a significance level of 0.05; *p*-values were calculated and corrected by Bonferroni´s method. The size of the nodes (circles) is proportional to the number of proteins found within the term, and the color scale indicates the corrected *p*-value corresponding to each term.

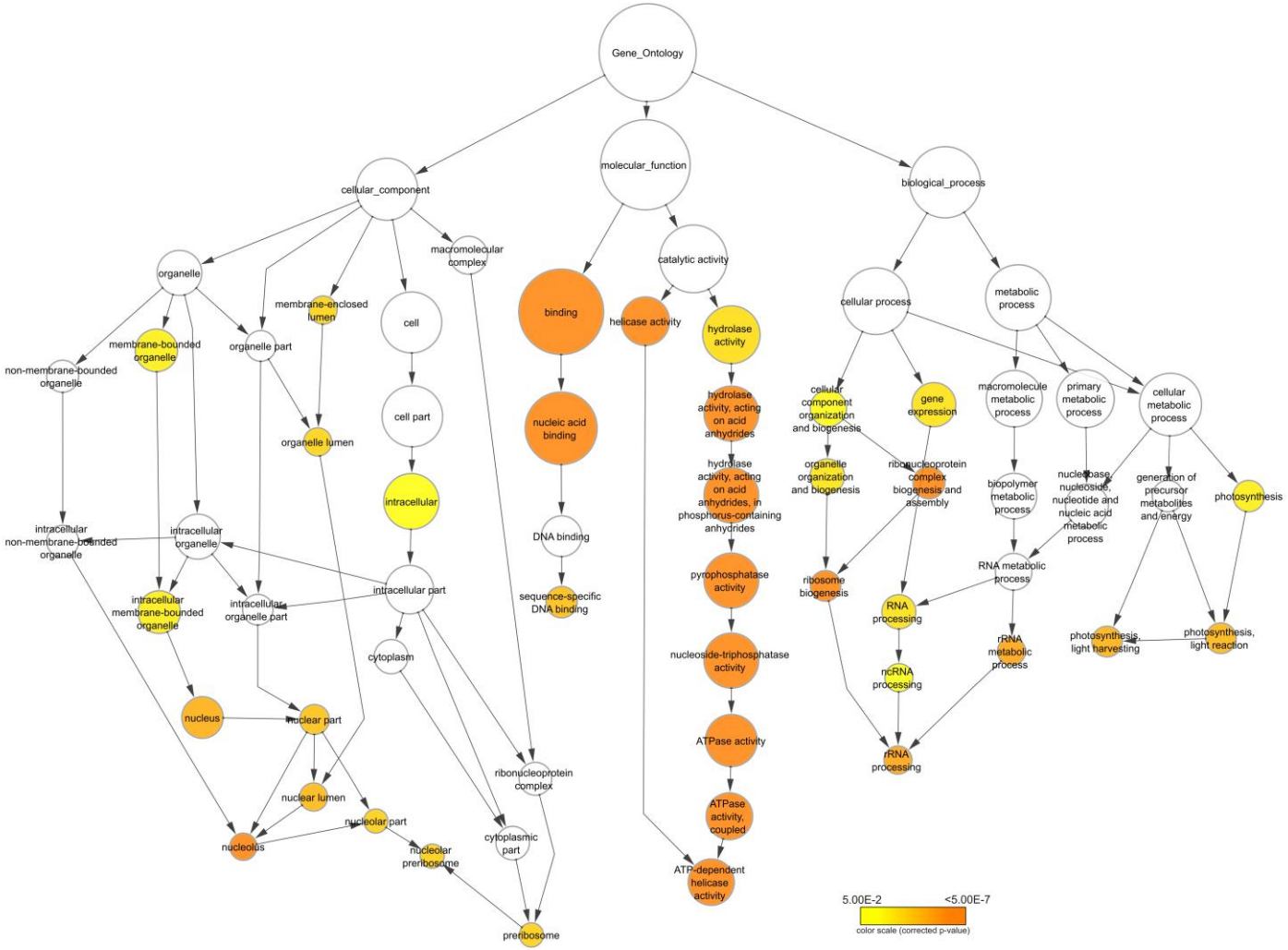


Figure 3-2. Network of overrepresented GO terms in the sub-dataset of proteins with conserved motifs.
Based on all GO categories (´cellular component´, ´molecular function´ and ´biological process´) the overrepresented GO terms for the proteins with identified motifs were calculated using BiNGO with a significance level of 0.05 and *p*-values corrected by Bonferroni´s method. The size of circles is proportional to the number of proteins found. The color scale indicates the corrected *p*-value.

Our data revealed groups of proteins that shared the same motif but did not have a defined GO term. To elucidate their possible nuclear functions we performed further analyses. The percent distribution of the functional categories of the proteins identified was calculated using GO Slimmer (http://amigo.geneontology.org/cgi-bin/amigo/slimmer) and the pre-existing generic GO slim set of ontology terms (http://www.geneontology.org/GO.slims.shtml) of the category ´molecular function´ (Figure 3-3).

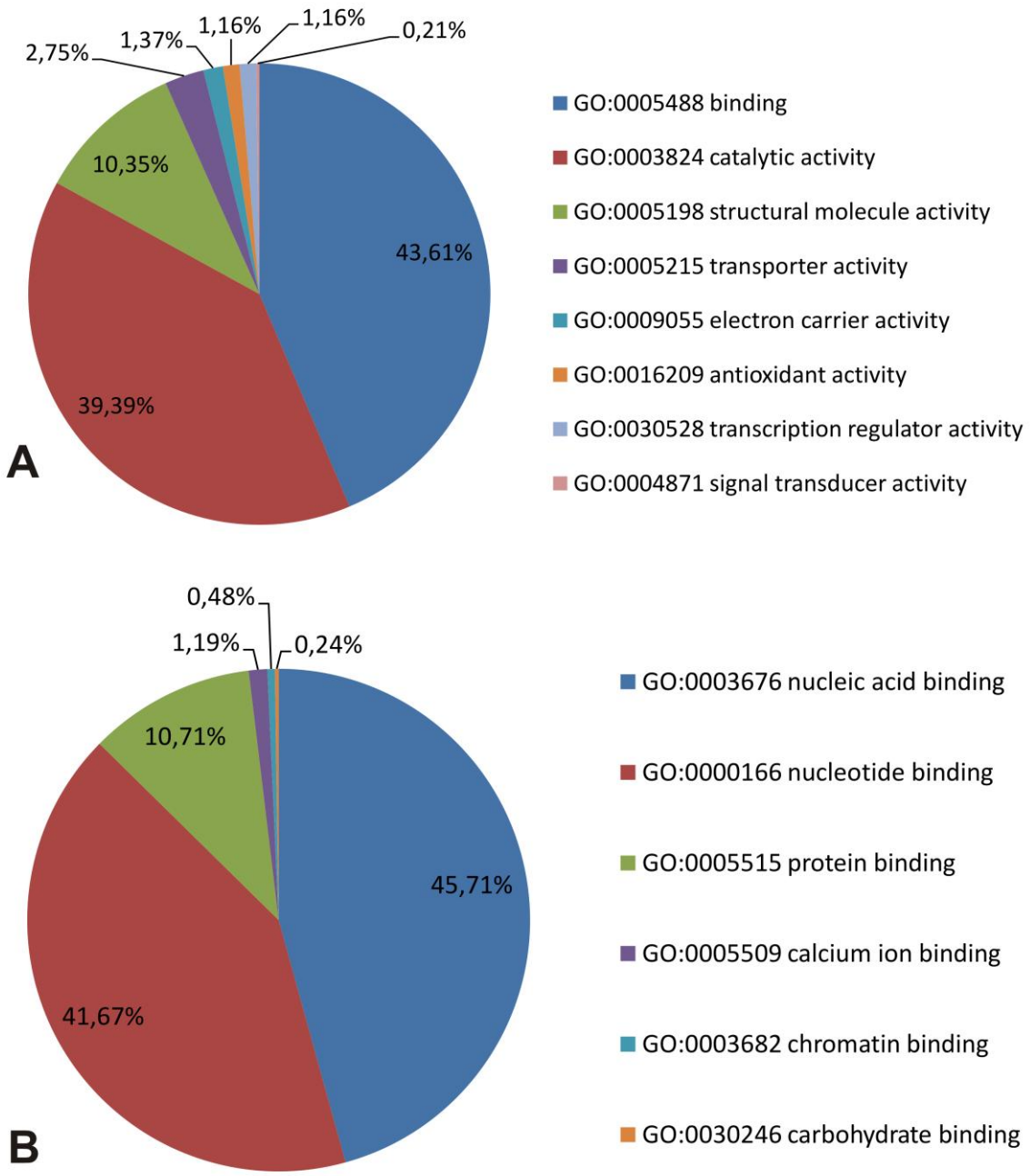


Figure 3-3. Percent distribution of the functional categories of the proteins identified.

Proteins were mapped to GO Slim terms of the molecular function category using the pre-existing generic GO set from GO Slimmer. A) Distribution of the functional categories based on the first child terms of the category ´molecular function´ from GO generic. B) Distribution of the functional categories based on the first child terms of the GO term ´binding´.

Of note, photosynthesis-related proteins also appeared in the nuclear dataset, however we can not exclude that those proteins are contaminants of the nuclear fractions. We were also

able to identify putative low-abundant proteins such as the transcription factors and transcription regulators presented in Table appendix B-1.

3.4.4 Prediction of subcellular location

The prediction of subcellular location of proteins revealed differences among the predictors. It is important to compare different tools in order to have more confidence of the subcellular predictions and in the present analysis the four tools appeared to disagree in the number of proteins predicted to be nuclear (Table 3-2). When comparing the percentage of the proteins from the whole Chlamydomonas proteome (genome annotation v.3.1, mitochondria and chloroplast genomes) the disagreement of the results revealed the need for complementary information for the judgment of the reliability of the predictions. Only 33 proteins from the total proteome of Chlamydomonas were predicted to have nuclear subcellular location when analyzed with all four tools (data not shown). Only two of them are present in the nuclear proteome dataset. The predictors were selected based in their possibility to predict the subcellular location of plant proteins and the parameters and training datasets were based on plant genome annotations. It has been shown previously that the best performing program for cellular localization of nuclear plant proteins was BaCelLo (Casadio et al., 2008). Even with the highest overall accuracy for plant proteins it seems that this tool still lacks the existence of a statistical parameter for the explanation of the significance of the predictions.

Table 3-2. Percentage of proteins predicted to be located in the nuclear compartment.

	BaCelLo	NucPred	PredictNLS	WolfPsort
Total Proteome	47,95%	1,02%	6,04%	14,90%
Nuclear proteome	31,60%	1,44%	7,20%	10,28%

3.4.5 Codon bias indexes in Chlamydomonas reinhardtii

There were previous indications that the protein expression levels can be correlated to the presence of codon bias on the corresponding genes (Gygi et al., 1999). In Yeast for example,

where more than half of the genes present a codon bias < 0.2, only a few proteins which can usually be detected in gel-based proteomics, possess codon bias lower than 0.1. In our present study of the nuclear proteome we were also interested in studying the codon usage patterns in Chlamydomonas and investigate if the codon bias indices could be of usefull for estimating protein abundance. Our analysis of the codon indices revealed that *C. reinhardtii* present a biased codon usage distribution which has also been observed previously by Naya and colleagues (Naya et al., 2001). The resulting putative optimal codons (Table 3-3) generated with the correspondence analysis (COA) were used for calculating the codon bias index (CBI).

The correspondence analysis took the overall number of codons of the input 9625 coding DNA sequences for determining the high and low biased codons. The results of the effective number of codons (ENC) and CBI analysis indicated that Chlamydomonas make use of a repertoire of preferential codons (Figure 3-4). It has been shown previously that some highly expressed genes in *C. reinhardtii* have also a lower number of codons on their composition (higher codon bias). In *C. reinhardtii* we could assume that for many genes translationally optimal codons exist, but further studies of gene transcription and translation rates and on abundances of tRNAs in Chlamydomonas would auxiliate to understand if the efficiency of transcription and codon bias are related to tRNA abundances or if it is a independent process. If the codon bias in Chlamydomonas presents a positive effect in transcription, we should expect that the high frequency of a preferred codon should be correlated to the abundance of its isoacceptor tRNA. In Chlamydomonas, the GC3 composition (G+C content at the third codon position of codons) has been shown to be the highest among all plants studied so far, reaching to approximately 77% (Wang and Roossinck, 2006). The highest similarity of synonymous codons GC3 composition was observed between Chlamydomonas and monocots, where a G or C is mostly found in the third position of the most abundant synonymous codons.

Table 3-3. Putative optimal codons of *Chlamydomonas reinhardtii*

Codon	Amino acid	Codon	Amino acid
UUC	(Phe)	ACC	(Thr)
UCC	(Ser)	AAC	(Asn)
UAC	(Tyr)	AAG	(Lys)
UGC	(Cys)	GUC	(Val)
UCG	(Ser)	GCC	(Ala)
CCC	(Pro)	GAC	(Asp)
CAC	(His)	GGC	(Gly)
CGC	(Arg)	GUG	(Val)
CUG	(Leu)	GAG	(Glu)
CAG	(Gln)	GAG	(Glu)
AUC	(Ile)		

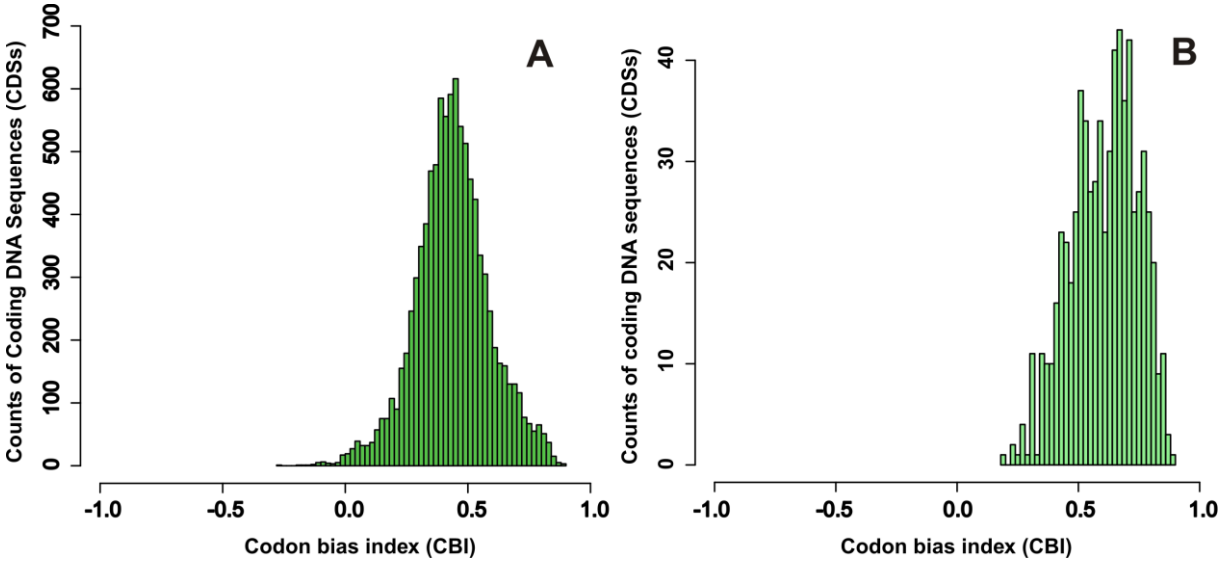


Figure 3-4. Codon bias distribution in *Chlamydomonas reinhardtii*.

(A) Most of the coding DNA sequences analyzed from the total proteome showed a relatively high codon bias (CBI > 0.1), with an average CBI of 0.4. A CBI value closest to 1 indicates that a DNA sequence contains a lower diversity of codon triplets and contains a higher usage of synonymous codons, known as "preferred" codons. The CBI values were calculated from 9625 CDSs from Chlamydomonas genome annotation V3 (DOE/JGI) using the program CodonW (http://sourceforge.net/projects/codonw/). (B) The coding DNA sequences analyzed for the nuclear proteome (684 CDSs) show a tendency of having higher codon bias compared to the total proteome, with an average CBI of 0.6.

The comparison of the ENC of the total proteome with the nuclear proteome dataset revealed that the distribution of the effective number of codons is slightly different between the two datasets. The CDSs corresponding to the nuclear proteins seem to have a higher codon usage bias in comparison to the total proteome dataset; the DNA sequences of the nuclear proteins seem to contain reduced effective number of codons (ENC) (Figure 3-5).

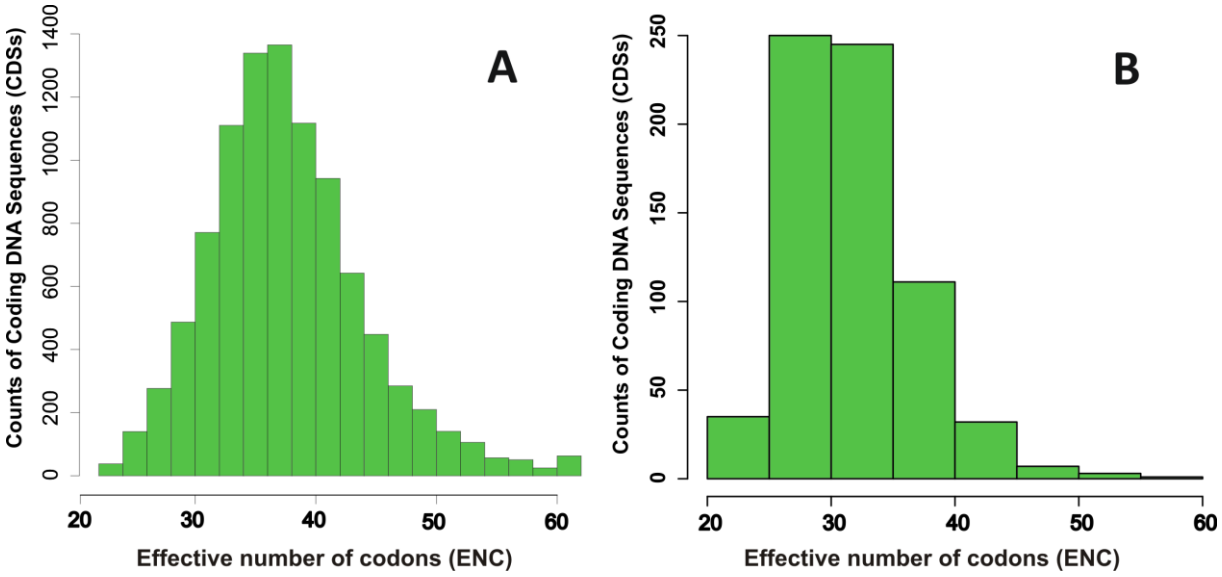


Figure 3-5. Distribution of the effective number of codons in the total annotated proteome of *Chlamydomonas reinhardtii*.

The CDSs annotated from the genome of *C. reinhardtii* v.3.1, containing most of the sequences included in our total proteome dataset (9625 CDSs) and most of the sequences which corresponded to the nuclear proteome dataset (685 CDSs), were used for the determination of the codon indexes using the program CodonW. The number of different codons found in each CDS represents the Effective Number of Codons (ENC) for this sequence. (A) Distribution of the number of different codon triplets found in the total annotated proteome. (B) ENC of the proteins identified in the nuclear proteome. The analysis indicates that sequences corresponding to the nuclear proteome dataset have a tendency to possess a lower ENC in comparison to the total proteome dataset.

These findings are showing that the codon usage in Chlamydomonas seems to present a particular and different codon bias distribution, and most of the coding sequences analyzed here showed a rather high codon bias. This could be attributed to the high GC content of the

genome, leading to higher GC3, but further studies are needed to understand if there is a general and direct correlation of codon bias and translational optimization. It is then more difficult to evaluate the validity of using the codon bias indices in Chlamydomonas for estimating protein concentrations. Further investigation on the correlation of codon usage and translational rates in Chlamydomonas would be of help to elucidate how the codon bias can influences the gene expression.

3.5 Conclusions

In the present work we analyzed the nuclear proteome of *Chlamydomonas reinhardtii* and the automated annotation combined with the analysis of GO overrepresented terms of the nuclear proteome dataset indicates that most of the proteins identified are related to nuclear functions. The identification of protein sequence motifs conserved within the group of nuclear proteins indicated the existence of enriched motifs whose function could be related to gene expression regulation, but further studies have to be done in order to confirm their function. The computational tools for prediction of protein subcellular location could not detect nuclear proteins from plant species in a consistent way, showing disagreement between predictions. This indicates that predictors should be improved, maybe by including proteins from other organisms as background models for the development of the prediction algorithms or by including conserved protein motifs in the searching strategies. We analyzed the codon bias distribution for most of the genes encoding the proteins predicted in the total proteome of Chlamydomonas and for nuclear proteome dataset obtained in the present study, looking for possible ways of estimating protein abundance. The codon bias analysis revealed that most of the protein coding sequences from *C reinhardtii* present a relatively high codon bias (> 0.1). These results make it difficult to discern which codon bias should be considered low or high for the particular case of Chlamydomonas. The codon bias indices do not seem to be good parameters for the inference of protein abundance or transcript abundance in Chlamydomonas. Our data will be helpful for further functional analyses of the Chlamydomonas proteins and their homologues from higher plants.

Chapter 4 Regulation of TF and TR gene expression by light[4]

[4] In the present Chapter the results of our contribution to the Core Experiment 1 of *Go*FORSYS project (http://www.goforsys.de) are presented. In the Core Experiment 1 different research groups were working in different platforms in order to obtain data about the cellular responses of Chlamydomonas under different light conditions. More than ten groups have worked with the same sample and our data will be integrated and made available through a Metadata repository.

4.1 Abstract

Light-regulated cellular mechanisms are closely related to photosynthesis and plant biomass accumulation. Much is known about the processes that take place within cells transforming the light energy and CO_2 into carbohydrates. Several photosynthesis-related regulatory processes occur at the level of transcription, in particular involving transcription factor (TF) and transcription regulator (TR) coding genes. However, the function of many of them remains unknown. In the present work, we analyzed the expression pattern of 245 TF and TR coding genes predicted from the annotated genome of the green algae *Chlamydomonas reinhardtii.* In a time series experiment changing the light conditions, from low light (200µE m^{-2} s^{-1}) to high light (700µE m^{-2} s^{-1}), the gene expression levels of TFs and TRs were measured through a multi-parallel quantitative reverse-transcription PCR assay. The results revealed groups of co-expressed TF and TR coding genes in response to the light shift; their possible role in light-mediated cellular responses is discussed.

4.2 Introduction

Light responsive genes are induced by light signals that reach photoreceptors which in turn activate signaling cascades. In Arabidopsis plants, it has been shown that phytochromes, representing one of several types of photoreceptors, can function as signaling molecules and, in addition to activating Ca^{2+} and cGMP dependent signaling cascades, can directly regulate the activity of transcription factors inside the nucleus. Cryptochromes also constitute a group of photoreceptors that are found in animals and plants, and are involved in most of the known blue light-induced responses (Cashmore et al., 1999). They are flavoproteins that contain two co-factors: flavin adenine dinucleotide (FAD) and methenyltetrahydrofolate (MTHF). Cryptochromes are involved in many different cellular processes, such as the control of the circadian clock, flowering and de-etiolation (Banerjee and Batschauer, 2005). In Arabidopsis, cryptochromes 1 and 2 are localized in the nucleus under light conditions, and appear to de-repress gene expression by interacting with other nucleus-localized proteins, such as constitutive photomorphogenic 1 (COP1) (Liscum et al.,

2003). Nevertheless, Arabidopsis quadruple mutants of phytochrome A and B and cryptochrome 1 and 2 under light conditions still show up-regulation of genes involved in photosynthesis and electron transfer, indicating the existence of alternative pathways which could regulate the expression of photosynthetic genes (Perelman et al., 2003).

Phototropin photoreceptors are also involved in the regulation of light-responses in land plants, and are known to function as receptor kinases capable of regulating downstream signaling cascades. Some transcriptional regulators such as transcription factors (TFs) containing zinc-finger domains function as intermediates of the signaling cascades activated by phototropins (Gyula et al., 2003). Light-induced responses in plants, such as chloroplast relocalization, stomatal opening, and phototropism are regulated by phototropins. In Arabidopsis, the mutation of phototropin 1 leads to the accumulation of light-harvesting complex II chlorophyll a/b binding protein (LHCB) and ribulose-1,5-bisphosphate carboxylase oxygenase large subunit (RbcL) when low-light-acclimated plants are transferred to high-light conditions, whereas wild type plants reduce the levels of both transcripts (Liscum et al., 2003).

Although the knowledge acquired with the studies of flowering plants is important to understand light responses, the study of algae and microalgae has revealed many particularities of these systems. In green algae genomes, including that of *Chlamydomonas reinhardtii*, no genes homologous to Arabidopsis phytochrome genes or genes for phytochrome interacting factors could be found (Riaño-Pachón et al., 2008). Thus, it is possible that, in Chlamydomonas, light responses are primarily induced by cryptochrome and phototropin photoreceptors. In Chlamydomonas, one copy of a cryptochrome gene, *CHLAMYDOMONAS PHOTOLYASE HOMOLOGUE 1* (*CPH1*), was found in its genome (Immeln et al., 2007). Furthermore, CPH1 protein was shown to undergo light-induced proteolysis *in vivo* and its blue-light-dependent autophosphorylation was verified *in vitro*. It has been suggested that the phosphorylation of CPH1 can leads to its dimerization and functional activation (Immeln et al., 2007). Phototropins in Chlamydomonas seem to be involved in sexual development, late gametogenesis, mating competence, and zygote germination (Huang and Beck, 2003). It has been shown that Chlamydomonas phototropin (PHOT) is important for the blue light induced up-regulation of light-harvesting polypeptides (e.g.

LHCBM6), glutamate-1-semialdehyde aminotransferase (GSAT) and phytoene desaturase (PDS) (Im et al., 2006).

Thus, plant photoreceptors seem to modulate the expression of light-responsive genes in a direct or indirect manner. Plants can, therefore, adapt to alterations in light conditions by changing the expression levels of genes related to photosynthesis, chloroplast development and cellular metabolism. In this biological scenario, the transcriptional regulation of gene expression plays an important role in the modulation of these cellular responses. Transcription factors (TFs) and transcription regulators (TRs) from different classes, such as bZip, MYB, bHLH, WRKY, appear among the genes early responding to alterations of light conditions, i.e., within one hour after application of the light stimulus (Casal and Yanovsky, 2005). Several TFs and TRs involved in light responses in flowering plants have been functionally characterized and their role in the regulation of gene expression investigated (Casal and Yanovsky, 2005).

The bZIP transcription factor, hypocotyl 5 (HY5), has been proposed as a key light responsive TF in all plants (Correa et al., 2008). In Arabidopsis, HY5 has been shown to bind to promoter regions containing the G-box (CACGTG) element, including the promoter of *RUBISCO CARBOXYLASE-OXYGENASE SMALL SUBUNIT* (*RbcS*) gene. In addition, HY5 interacts with the COP1 protein, a ubiquitin ligase, which ubiquitylates HY5 thus triggering its degradation. In Chlamydomonas, one Arabidopsis HY5 orthologous gene was found (Riaño-Pachón et al., 2008). However, the amino acid sequence of algal HY5 proteins differs from that of HY5 proteins found in angiosperms. Chlamydomonas HY5 does not contain the interaction site for COP1, furthermore, a homologue of *COP1* could not be detected. The absence of COP1 in Chlamydomonas indicates the presence of alternative light signaling pathways. Another TF involved in light-responses is HY5 homolog (HYH), which controls a very similar set of genes as HY5, as it binds to G-box elements located on promoter regions (Chattopadhyay et al., 1998; Casal and Yanovsky, 2005). Members of the bHLH TF family, for instance, phytochrome interacting factor (Pif3) and long hypocotyls in far red 1 (Hfr1), can also bind to G-box elements and are also involved in the light responses through phytochrome-mediated pathways in Arabidopsis (Casal and Yanovsky, 2005). Light-responsive members of the MYB TF family, bind to the DNA sequence motif AATCT which is present in the promoter of the

gene encoding light harvesting chlorophyll a/b binding protein (Lhcb1*3) (Wang et al., 1997). Several methods and approaches were used for searching conserved light-responsive *cis*-elements (LREs) on the promoter region of light responsive genes and many of them were found, but so far there is no evidence of a general, single motif present in the promoter regions of all light-regulated genes (Jiao et al., 2007). Most likely, the high flexibility and complexity of the light-regulated networks can be attributed to the combination of several different motifs found in the promoter of the light-responsive genes, which would integrate multiple inputs into a coherent response.

Chromatin remodeling factors also play an important role in the transcriptional regulation of light responses. Arabidopsis plants mutated for the *DE-ETIOLATED-1* (*DET1*) gene show an over-expression of almost half of the genes known to be induced by light. The DET1 gene seems to function as a spatio-temporal regulator of light induced genes (Maxwell et al., 2003). It is important to highlight that light responses are usually integrated with other signaling networks, influencing the activity of many of the biochemical pathways of the different cellular organelles.

Chlamydomonas cells also seem to have different mechanisms to respond to light stresses, balancing the efficiency in light capture and the photoprotective capacity. The regulation of light harvesting mechanisms in excess light involves different proteins that, are not found in vascular plants. For example, the ancient protein LHCSR which is member of the LHC superfamily was found to be important in responses to high light and for protection against photodamage (Peers et al., 2009). LHCSR transcript levels increase under high light conditions. Moreover, its expression is not coordinated with the expression of other LHCs, and its transcripts have been shown to accumulate under conditions which contribute to photo-oxidative stress, such as deprivation of carbon dioxide, high light and nutrient deprivation (iron and sulfur) (Peers et al., 2009). In Chlamydomonas, the photon conversion efficiency seems to be linked to biomass production. Mutants with reduced levels of LHC proteins showed increased biomass production. It was suggested that biomass production is reduced when cells activate their mechanism of photoprotection, particularly in high light conditions (~1000μE m^{-2} s^{-1}). The translational repression of LHC subunits by the constitutive

expression of *Nab1*, which codes for a RNA binding protein, interfered with the photoprotection mechanism and increased biomass production (Beckmann et al., 2009).

The genome-wide investigation of the Chlamydomonas gene expression patterns in different light and CO_2 conditions revealed the identity of many candidate genes involved in photosynthesis (Im et al., 2003). However, TFs and TRs represent only a very small fraction of the genes identified in such analysis, making it difficult to understand globally the regulation of expression of photosynthesis-related genes. Previous efforts have been done for determining which Chlamydomonas genes are involved in photosynthesis, and orthologous genes in Arabidopsis were found, improving the assignment of the genes into functional categories. However, the categories "signaling" and "nucleic acid transaction" obtained from genome-wide approaches have been shown to contain many genes with unknown function (Grossman et al., 2010). The discovery of many TF and TR genes in the genome of Chlamydomonas and their assignment to gene families supports further functional studies and the analysis of gene regulatory networks (Riaño-Pachón et al., 2007). However, the functions of most of them remain unknown. In order to investigate the role of regulatory genes in Chlamydomonas, we measured expression changes of the previously predicted TF and TR coding genes in a time series experiment where cultures were shifted from low light (200µE m^{-2} s^{-1}) to high light (700µE m^{-2} s^{-1}) conditions. This approach helped us to uncover the identity of many light-responsive transcriptional regulators and their possible role in light-inducible responses is discussed.

4.3 Material and methods

4.3.1 Cell culture and growth conditions

Chlamydomonas reinhardtii strain cc1690 was obtained from the Chlamydomonas Genetics Center (Duke University, North Carolina, USA) and cultured in a bioreactor system Biostat® B-DCU (Sartorius Stedim Biotech, Melsungen, Germany) equipped with pH, temperature, dO_2, pressure controllers and a turbidometer. Growth conditions were automatically adjusted keeping the concentration of CO_2 close to 5%. Cell density was kept by semi-continuous dilutions at an approximate concentration of $3x10^6$ cells·mL^{-1}. Cells were cultured

in photoautotrophic conditions in a modified TAP medium (Gorman and Levine, 1965) where the phosphates were replaced by 5mM HEPES and the acetate was omitted. The average temperature used for cooling the five liter culture vessels was 21°C. Light shift experiments with cultures transferred from 200µE m^{-2} s^{-1} (low light) to 700µE m^{-2} s^{-1} (high light) were conducted to four experimental replicates. The control bioreactor was kept at constant light at 200µE m^{-2} s^{-1}. Cells were sampled on ten time points, -60, 0, 5, 10, 20, 40, 60, 120, 240 and 480min, described here as minutes from "time zero" where the shift on light intensity was applied. Control and treatment samples were collected for the same time points and samples for transcript analysis were harvested in 2mL aliquots by centrifugation for two minutes at 3.000 *g* and the cell pellet was immediately frozen in liquid nitrogen and kept at -80°C until further use.

4.3.2 Primer design for qRT-PCR and validation

The design of primer pairs was based on the *C. reinhardtii* gene models of the genome annotation v.3.1 released by the Joint Genome Institute (DOE/JGI; http://genome.jgi-psf.org/Chlre3/Chlre3.home.html). Primer pairs were designed for the predicted transcription factors and transcription regulators genes described in the Plant Transcription Factor Database (Riaño-Pachón et al., 2007). Criteria for primer design are as follows (Udvardi et al., 2008): T_m = 60 ± 1°C, length 18 to 25 bases, preferentially on exon-exon junctions. Primers were designed to have a GC content between 45 and 55% when possible (due to the high GC content of the Chlamydomonas), generating a single PCR product sizing between 60 and 150 bp. Primers were synthesized by Eurofins MWG Operon (Ebersberg, Germany). A complementary *in silico* validation of the specificity of the primers was carried out using the computational tool QuantPrime (Arvidsson et al., 2008). Experimental validation of primer pairs was performed by checking the presence and approximate size of PCR products obtained from the amplification of cDNA synthesized from RNA extracted from *C. reinhardtii* cells grown under low light intensity (200µE m^{-2} s^{-1}) photoautotrophic conditions as described in section 4.3.1. PCR products were separated in 3% agarose gels and amplicons which did not have the correct size (i.e., deviating by more than 20bp from

the expected size, based on data extracted from the genome annotation v.3.1) were excluded from further analysis.

4.3.3 RNA extraction, cDNA synthesis and quantitative RT-PCR

Total RNA extraction was performed using the RNeasy Plant Mini Kit (Qiagen, Hilden, Germany). Briefly, the frozen cell pellet was homogenized in 450µL of lysis buffer RLT and frozen again in liquid nitrogen for three minutes. Cells were completely lysed by incubating the frozen cell pellet at 56°C for 3min. Further steps of the RNA extraction were performed as described by the manufacturer. In order to remove genomic DNA from the RNA extracts, on-column digestion of DNA was performed using DNAse I (Qiagen, Hilden, Germany) and an additional DNAse treatment was performed, on the RNA eluate, using TURBO DNAse (Ambion, Darmstadt, Germany) as indicated by the manufacturer. The integrity of the RNA was checked by electrophoresis on 2% denaturing agarose gels; RNA quality was assessed by determining the 260nm/280nm absorbance ratio using a Nanodrop (Thermo Scientific, Schwerte, Germany). Furthermore, the absence of genomic DNA in the RNA extracts was assessed by performing a q-PCR on a 1µL aliquot of each sample of total RNA using primer pair annealing to an intergenic region of chromosome 16 (forward primer 5'-TGTCTTGTGAATCCTGCCCTC-3' and reverse primer 5'-AAAGAGCTCACAAGTACACACCGA-3'). Only if the q-PCR analysis of the RNA sample resulted in a Ct value higher than 36, the RNA was used for cDNA synthesis. Three micrograms of total RNA were used for cDNA synthesis reaction employing the SuperScript III First Strand System (Invitrogen, Darmstadt, Germany) according to the manufacturer's instructions, using oligo-(dT_{20}) as primer for the synthesis of the first complementary DNA strand. The efficiency of the cDNA synthesis reaction was estimated by qRT-PCR using primer pairs for the UBIQUITIN PROTEIN LIGASE coding gene (protein ID 190824 (DOE/JGI); forward primer 5'-TTACCTGCCTTCCGATTGCGTAGC-3' and reverse primer 5'-TTACTATGCCTGAGCACGCAGCAC-3'). The cDNA samples were diluted ten times prior to the final PCR reaction which was conducted with the SYBR Green mix (Applied Biosystems, Darmstadt, Germany) within a final reaction volume of 5µL containing 5µM primers. Dilution of primers and pipetting into the PCR 384 well plates were performed with a robot to increase accuracy and throughput. The ABI PRISM 7900HT sequence detection

system (Applied Biosystems, Darmstadt, Germany) was used for the qRT-PCR reactions which were carried out as previously described (Caldana et al., 2007).

4.3.4 qRT- PCR data analysis

Raw qRT-PCR data was pre-processed with the program SDS v.2.3 (Applied Biosystems, Darmstadt, Germany) and the amplification curves were analyzed considering a threshold of 0.2 on the variation of sample fluorescence to passive dye fluorescence in order to obtain the Ct values (Cycle threshold). The baseline was taken from the 3^{rd} up to the 15^{th} cycle of amplification. Quality controls were done in order to evaluate the raw data by means of efficiency of the amplification reactions and the profile of the melting curves. Genes in which the amplification efficiency was lower than 95% for more than 25% of the reactions were excluded from further analysis. Reactions presenting multiple melting curves or melting temperatures that disagreed with the expected melting temperature were also excluded from further analysis. PCR products that experimentally did not present the same expected size and the ones that presented multiple amplicons were also excluded from further analysis. Only high quality measurements were selected for the relative gene expression analysis. All raw Ct values were normalized by the quantile method using the R package qpcrNorm that can be found through the Bioconductor project (Mar et al., 2009). Biological median of the Ct values from the four experimental replicates was calculated and the transcript levels were determined according to the delta-delta Ct method (where ΔΔCt = ΔCt treatment – ΔCt control) (Udvardi et al., 2008). The relative gene expression was calculated as $2^{-\Delta\Delta Ct}$ and presented as scaled log2 values.

4.3.5 Clustering analysis

The transcript profiles for the expressed genes were further analyzed by the quality threshold nonhierarchical clustering method (QT) (Heyer et al., 1999) using Pearson´s correlation as distance measure. The clustering process considered a minimum size cluster of five genes and a diameter of 0.5. The QT method used for the clustering is available as an application of the program Multi experiment Viewer (MeV) v.4.6.0 (Saeed et al., 2003).

4.3.6 Identification of conserved motifs

Promoter DNA sequences of 14598 genes annotated on the Chlamydomonas genome v.3.1 were obtained from the data available at the DOE Joint Genome Institute (http://genome.jgi-psf.org/chlamy/chlamy.home.html) using Perl scripts written in house. The putative promoter regions of genes were analyzed using the MEME suite (v.4.3.0, http://meme.nbcr.net/). Search parameters were: a motif could appear at any number of repetitions, the minimum and maximum width of a motif was set 4 and 15 residues, respectively, and a maximum of 10 motifs were requested on each run. Other parameters were left with their default values. Evaluation of specificity of the motifs found was carried out with MAST tool (Bailey and Gribskov, 1998).

4.4 Results and discussion

In the present work we established a qRT-PCR platform for the gene expression analysis of TF and TR genes of *C. reinhardtii*, which has opened the venue for understanding the role of many transcriptional regulators in response to different environmental conditions. By including primers for the analysis of the 245 predicted TFs and TRs, our transcript profiling revealed that expression pattern of 127 TF and TR coding genes varies with changes in light intensity. The quality control tests performed on the raw dataset for exclusion of inconsistent data were of great value for removing potential technical problematic primers and keeping only high quality expression values for further analysis. Gel images of the amplification products are available upon request. For this particular time series experiment we did not find a stable gene that could be used as a reference gene for data normalization (data not shown). Thus, data normalization by the QT method was chosen which allowed minimizing bias possibly introduced by vaguely defined reference genes for data normalization (Guenin et al., 2009).

4.4.1 Gene expression profiling

Gene expression profiles of co-expressed genes were assigned to eleven clusters by the QT method. The clusters identified show the expression profiles of the 127 TF and TR genes

modulated in the time series experiment of light shift (Figure 4-1). Our data suggested the existence of co-regulated TF and TR genes in response to the transition from low light (200µE m^{-2} s^{-1}) to high light conditions (700µE m^{-2} s^{-1}).

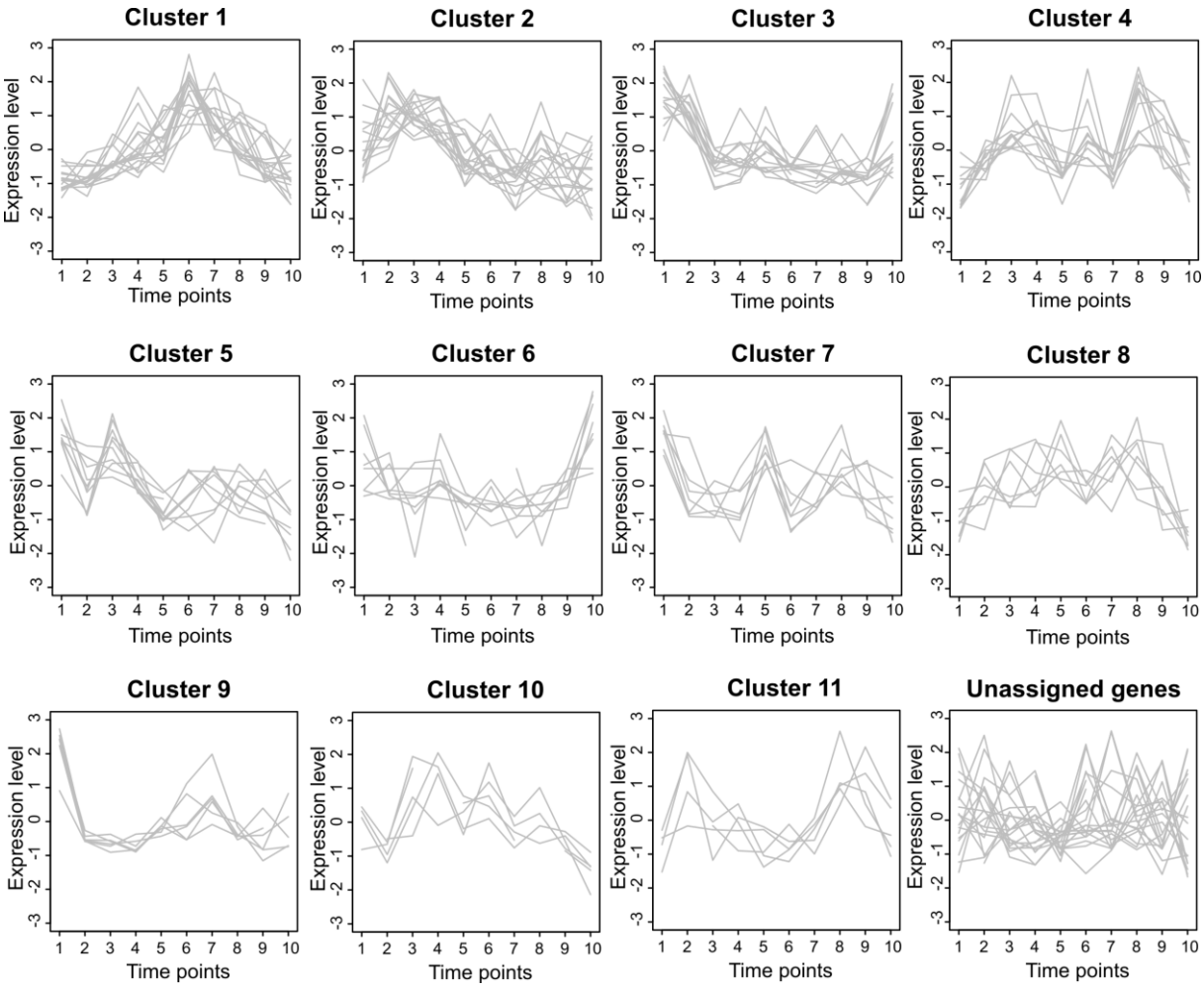


Figure 4-1. Cluster analysis of gene expression profiles for TF and TR coding genes in response to a light shift from 200µE m^{-2} s^{-1} to 700µE m^{-2} s^{-1}.

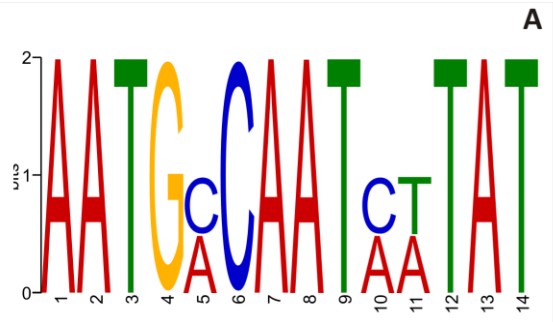
Each panel shows the expression levels (scaled, Log2) of TF and TR coding genes that, according to the QT clustering, share similar expression patterns. The gray lines represent the expression level of each individual gene at the different time points. The time points 1 to 10 correspond to the following sampling time points, respectively: -60, 0, 5, 10, 20, 40, 60, 120, 240 and 480 min relative to the time of the light shift, i.e., time point 2.

A list containing the primer sequences, information about the gene families, and the respective clusters for the genes analyzed is available in the Table appendix C-1. The expression levels of the genes analyzed by QT method are listed in Table appendix C-2.

4.4.2 Early response genes

Our clustering analysis revealed genes that had similar expression profiles and seem to be co-regulated within the first minutes (15min) after the light shift. These genes are interesting candidates, considering their possible role in the regulation of the first steps of cellular acclimation and signaling crosstalk. One of these groups of early response genes is cluster number 10. This cluster contains five genes (protein IDs: 143060, 159133, 194335, 187360 and 184359) for which we found conserved motifs located on their promoter regions using the MEME suite. The motifs found were compared to the sequences of the 14598 promoter regions of Chlamydomonas genes using motif alignment search of the MAST tool. Some of these motifs appear to be specific, and were found in the promoter region of only few genes (p-value < 0.0001) (Table appendix C-3). Among the motifs found on the promoter region of genes from Cluster 10, one in particular (Motif 7) draw our attention. Motif 7 ("AATG[AC]CAAT[AC][AT]TAT"; E-value = 5.3e+003) was found in the promoter region of two genes, protein ID 184359 and protein ID 187360, members of the MYB-related and C2C2-GATA TF families, respectively. The MAST results showed that Motif 7 is present in the promoter regions of only four genes in total at a genome-wide scale (protein IDs: 187360, 187361, 184359 and 196464; see Figure 4-2).

The TF member of the MYB-related family encodes the LCR1 protein, which is known to be induced during activation of the Carbon Concentrating Mechanism (CCM) in *C. reinhardtii.* (Yoshioka et al., 2004). The protein LCR1 (for Low-CO_2 stress responsive), appears to participate in the regulation of the expression of the carbonic anhydrase coding gene (*Cah1*). However, it is not essential for the induction of *Cah1*, and it has a role in maintaining the high expression level of *Cah1* under conditions of carbon deprivation.



B

ProteinID	Motif p-value	Site
184359	4.00e-09	AATGCCAATAATAT
187360	4.00e-09	AATGACAATCTTAT

C

ProteinID	Motif p-value	Site
187360	4.9e-08	AATGACAATCTTAT
187361	4.9e-08	ATAAGATTGTCATT
184359	4.9e-08	AATGCCAATAATAT
196464	1.4e-07	AATGTCAATAATAT

Figure 4-2. Common motif found within the promoters of early response genes from Cluster 10.

(A) MEME logo of Motif 7, present in the promoter region of MYB-related and C2C2-GATA TF genes (protein ID 184359 and protein ID 187360, respectively). (B) The conserved sequences found in the promoter region of these genes, identified using the MEME tool, and their respective p-values. (C) Motif 7 was analyzed with the MAST tool revealing two additional genes containing this motif in their promoter regions. The corresponding p-values and the motif sites found within the promoters are given. Nucleotides conserved in all sequences are highlighted in red.

The C2C2-GATA TF (protein ID 187360) identified in our transcript profiling as having the same conserved motif as *Lcr1* in its promoter has not been fully characterized in Chlamydomonas. Its orthologous gene in *Arabidopsis thaliana* (AT4G17570.1), has been shown to be up-regulated by cold and UV-B stresses (Kilian et al., 2007). Many TFs belonging to the GATA family have been reported to regulate the expression of photosynthetic genes, such as *GLYCERALDEHYDE-3-PHOSPHATE DEHYDROGENASE* (*GAP* A) and RIBULOSE-1,5-*BISPHOSPHATE CARBOXYLASE OXYGENASE SMALL SUBUNIT* (*RBCS)* (Borello et al., 1993; Jeong and Shih, 2003). Our results suggest that these two early response genes (protein ID 184359 and protein ID 187360) may be necessary for the first steps of the cellular acclimation of *C. reinhardtii* to high light conditions. However, it is not known if the conserved motif found in the promoter region of both genes contains in fact a *cis*-regulatory element. The other two genes (protein ID 187361 and 196464) which also contain Motif 7 in their promoters have not been functionally characterized so far. Nevertheless, their predicted functional annotation indicates that these genes may encode proteins with peptidase activity and phosphate/sodium transporters characteristics, respectively.

Further experimental analyses of the promoters of these genes will help to elucidate the importance of Motif 7 for light-induced responses.

4.4.3 Stress responsive genes

Among the genes responding to the light shift, two TF genes members of the heat shock factor family (HSF) were identified. These two genes, *Hsf1* and *Hfs2* (protein IDs 117914 and 122706, respectively) showed to be light responsive. Their expression levels increased after the shift from low to high light, returning to similar initial levels approximately 120min after the light shift (Figure 4-1, Cluster 1 and 8, respectively). Previous studies have shown that HSF proteins contain a DNA- binding domain at the N-terminus, a conserved hydrophobic repeat and a C-terminal transactivation domain. Heat shock factor 1 has a nuclear localization signal (conserved motif RRKRR) within its C-terminal region and two putative leucine-rich regions which may be involved in nuclear export. As far as our knowledge goes, only *Hsf1* has been functionally characterized in Chlamydomonas. It has previously been shown that the *Hsf1* transcript level increases in less than 45 min when Chlamydomonas cells are stressed by heat (40°C) (Schulz-Raffelt et al., 2007). Moreover, HSF1 interacts constitutively with HSP70A and can interact with HSP90, functioning as a master regulator of stress responses (Schmollinger et al., 2010). As it is also responsive to the light shift applied in the present study, it is possible that the HSF proteins are also an important element of the pathways of light-inducible stress responses in Chlamydomonas. Further experiments confirming HSF target genes would help to clarify the structure of the signaling pathway for heat and light stresses in Chlamydomonas and determining if HSF function is indirectly or directly related to light signals and photosynthesis-related responses.

4.4.4 Candidate regulatory genes of photosynthesis-related genes

Previous findings have demonstrated in Arabidopsis plants that the TF ELONGATED HYPOCOTYL 5 (HY5) protein, a bZIP transcription factor, interacts with G-box and GATA box elements in the promoter regions of light-responsive genes, and HY5 could be a main regulator of light-induced responses (Correa et al., 2008). HY5 protein is located in the nuclear compartment of the cells independent of the presence of light (Chattopadhyay et al.,

1998). In flowering plants, phytochromes can induce downstream signaling cascades negatively regulating COP1, which lead to accumulation of HY5 in the nucleus (Chattopadhyay et al., 1998; Hardtke et al., 2000). Phytochromes are not present in Chlamydomonas, and it is therefore likely that another pathway or components other than HY5 are involved in the regulation of main light responses.

In the following some new candidate genes that are differentialy expressed upon the light shift are discussed. Cluster 1 includes a member of the bZIP TF family (protein ID 147286) whose expression level increased after the light shift reaching its highest expression level (FC >3) around 40min after the light shift. This TF has not been characterized functionally, and the most similar sequences found in the non-redundant database of NCBI (http://www.ncbi.nlm.nih.gov/), using the BLAST tool, corresponds to hypothetical proteins from *Volvox carteri f. nagariensis* (E-value 2e-31) and *Chlorella variabilis* (E-value 6e-04). There is no evidence for ortologues in land plants, and only one orthologue was previously identified in *Ostreococcus tauri* (Ot15g01450) (Correa, et al., 2008).

Our study revealed new candidate genes which could be involved in the regulation of the light-induced responses. The expression levels of four members of the C2C2-GATA TF family were found to be modulated in response to the light shift. Two of them, protein IDs 179289 and 171659, were assigned to the gene Clusters 1 and 2, respectively, and showed changes of their expression levels within the first 60 min after the light shift. One of these GATA factors, protein ID 171659, was down-regulated after the light shift and its expression level was continuously low during high light condition. In contrast, another GATA factor (protein ID 179289) has a transient high expression during the first hour of high light, returning to similar initial expression levels after 120 min. GATA TFs are present in many different organisms. They are found for instance, in fungi, metazoans and plants. It has been shown that these factors can recognize the consensus DNA sequence WGATAR (W = T or A; R = G or A) in promoter regions (Reyes et al., 2004). The DNA binding domain of GATA factors is constituted by one or two zinc finger domains of the form $CX_2CX_{17–20}CX_2C$ followed by a highly basic region (Lowry and Atchley, 2000). Some members of this family are known to be related to light-responses. Jiao and colleagues, have shown that fast changes in the abundance of the transcripts of GATA factor genes of *Arabidopsis thaliana* seedlings occurs

in less than one hour after the shift from dark to blue-light condition (Jiao et al., 2003). Furthermore, cryptochromes have shown to be important during early flower development in Arabidopsis plants by regulating the transcription of genes containing GATA boxes in their promoters (Jiao and Meyerowitz, 2010), indicating that changes on the expression of C2C2-GATA factors may reflect the induction of blue light regulated signaling pathway(s). Due to the apparent lack of phytochromes in Chlamydomonas, one hypothesis that may explain how cells respond to light is that the cryptochromes and phototropins may have a major role in the induction of the light responses and that the light signals could be mostly induced by blue light (Jiao and Meyerowitz, 2010).

The interaction between the Arabidopsis GATA-1 factor (gene ID: At3g24050) and regulatory elements (named XXIII) located in the promoter of the *GLYCERALDEHYDE-3-PHOSPHATE DEHYDROGENASE* (*GAP* A) gene has been previously verified in vitro (Jeong and Shih, 2003). The expression levels of the genes encoding subunits A and B of GAP, *GAPA* and *GAPB* in Arabidopsis, have been shown to be coordinately modulated by light. It has been suggested that the GATA box is a general essential promoter fragment present in light-regulated photosynthesis genes (Hihara et al., 2001). Furthermore, it was previously shown that high light induces not only photosynthesis-related genes but also induces changes on the expression level of *HEAT SHOCK PROTEIN* (*HSP*) genes in cyanobacteria (Synechocystis) and also in Chlamydomonas. These observations indicate crosstalk between signaling pathways induced by light and heat (Hihara et al., 2001; Yamano et al., 2008). The study of knockout lines for these GATA factors would reveal if their target genes are indeed photosynthesis-related genes.

4.5 Conclusions

In the present study we established a platform for the analysis of the gene expression of 245 predicted TF and TR coding genes from *C. reinhardtii* through quantitative RT-PCR. A time series experiment monitoring the alterations on the gene expression patterns in response to changes of light conditions (from 200µE m^{-2} s^{-1} to 700µE m^{-2} s^{-1}) revealed groups of co-regulated genes. The expression profiles of 127 genes could be analyzed by clustering and

most of the genes could be assigned to groups of genes sharing similar expression patterns. Interestingly, among the early responsive genes we found genes involved in abiotic stress responses, the carbon concentrating mechanism and genes not functionally characterized. New candidate genes members of the bZIP and C2C2-GATA TF families were found to be regulated in response to changes in light intensity. Our results indicate that genes involved in different signaling pathways, such as stress and light-induced signaling pathways, are transcriptionally regulated during the process of cell acclimation to high light conditions. These genes may be important components for the coordination and integration of complex gene regulatory networks in Chlamydomonas. Further functional characterization of these candidate regulatory genes is necessary to elucidate their regulatory role on photosynthesis-related events.

Chapter 5 Regulation of TF and TR gene expression under carbon deprivation

5.1 Abstract

The photosynthetic unicellular algae *Chlamydomonas reinhardtii* is a model organism for the study of photosynthesis and mechanisms related to carbon metabolism. In the past years, the study of the Carbon Concentrating Mechanism (CCM) has revealed the identity of important regulatory genes. However, as far as our knowledge goes, there is no study reporting an analysis focused on recognizing the expression patterns of transcription factors and transcription regulators involved in the CCM. Transcription factors (TFs) and transcription regulators (TRs) are important players in the regulation of transcription and recently we have predicted more than 240 TF and TR to be present in the *Chlamydomonas reinhardtii* genome. For better understanding the regulation of the CCM in Chlamydomonas and uncover regulatory elements, we performed a quantitative reverse transcription polymerase chain reaction based transcript profiling of the TF and TR genes of Chlamydomonas. The gene expression patterns of 130 genes were obtained for cells grown under photoautotrophic conditions and subjected to a shift from high-CO_2 (HC) to low-CO_2 (LC) conditions. Additionally, we performed a genome-wide identification of nucleosome-depleted regions of cells under LC using a sequencing approach, revealing potential regulatory elements involved in CCM responses. The gene regulatory network for CCM was reconstructed using a mathematical modeling approach based on the expression profiling dataset and revealed new candidate genes participating in the control of the CCM. The combination of the three approaches suggested that a TF (protein ID 188627), a member of the MYB family and responsive to LC, could function as regulator of CCM-related genes. Our results suggest an additional important role of LCR1 in down-regulating the gene expression of a particular set of genes responsive to LC conditions.

5.2 Introduction

The photosynthetic algae *Chlamydomonas reinhardtii* is able to concentrate inorganic carbon (Ci) intracellular via the Carbon Concentrating Mechanism (CCM). Through this mechanism, the concentration of CO_2 increases at the site of the ribulose-1, 5 bisphophate

carboxylase oxygenase (Rubisco) enzyme which stimulates the overall rate of photosynthesis. It has been proposed that the CCM and Rubisco have co-evolved (Badger et al., 1980; Kaplan and Reinhold, 1999). The CCM is thus important for cells to survive in niches where depletion of Ci can occur rapidly, such as aquatic environments where the availability of Ci and O_2 can vary, as well as pH (Pollock et al., 2004; Spijkerman, 2008; Yamano et al., 2008). Particularly, the CCM is induced when the extracellular concentration of Ci goes below 0.5% (Fukuzawa et al., 2001). It has been shown that some proteins most likely act as HCO_3^- transporters and CO_2 channels in the plasma membrane and in the chloroplast outer membrane. Some genes, including *Ccp1*, *Ccp2*, *Lci1*, *Hla3* and *Rh1,* have been proposed to express membrane proteins that might participate in the transport of Ci into the chloroplast (Yamano and Fukuzawa, 2009). These proteins would mediate the increase of Ci concentration, and thus increasing the CO_2 availability close to the Rubisco´s active site (Moroney and Ynalvez, 2007). Furthermore, studies with the cyanobacterium *Synechocystis* PCC6803 and *Chlamydomonas acidophila* have shown that cells grown in light and CO_2-free or low CO_2 conditions are able to take up Ci faster than cells grown in darkness and in low CO_2 or CO_2-free or conditions, suggesting the existence of a cooperative effect between light and carbon signaling in the modulation of the expression of CCM-related genes (McGinn et al., 2003; Spijkerman, 2011). Similar findings were also reported for *C. reinhardtii*, revealing that the decrease of the Ci concentration is not sufficient to activate the CCM and that enhanced expression of CCM-related genes is affected by different light intensities (Yamano et al., 2008). However, since the induction of the CCM in Chlamydomonas seems to be also dependent on the light conditions, the question arised whether the cells can directly sense external concentrations of Ci or whether the induction of CCM is triggered by indirect sensing the external Ci concentration (Yamano and Fukuzawa, 2009). Mutant lines in Chlamydomonas have permitted to partially identify some of the CCM regulatory mechanisms. A mutant deficient for Ci uptake was isolated previously and a gene named *Ccm1* was shown to complement the mutant phenotype and to allow induction of the CCM. CCM1 is a hydrophilic protein that contains a C2H2-type zinc-finger motif, which may confer protein-DNA binding activity. Interestingly, the cellular level of CCM1 does not change during the acclimation to LC, therefore, a post-translational control of CCM activity

was suggested (Fukuzawa et al., 2001). Another regulator of CCM is LCR1, a TF, member of the MYB family. Mutant lines for LCR1 showed transcriptional down-regulation of CCM-related genes, such as *CARBONIC ANHYDRASE* (*Cah1*) (Dionisio-Sese et al., 1990) and *Lci1* (Low-CO_2 inducible) (Miura et al., 2004). It has been suggested that LCR1 translocates to the nucleus and binds to promoter elements of *Cah1*, regulating its transcription in response to low Ci concentrations (Yoshioka et al., 2004). The enzyme carbonic anhydrase has been previously reported to equilibrate the cellular concentration of CO_2 and HCO_3^- and it is an important component of the CCM (Badger et al., 1980). Nonetheless, there is no complete understanding of the whole regulatory network for the CCM.

In order to obtain a deeper understanding of the regulatory mechanism in Chlamydomonas and with the purpose of obtaining a broad view of the possible regulators of the CCM we performed transcript profiling of around 250 TF and TR coding genes. These TF and TR genes were previously predicted by our group (Riaño-Pachón et al., 2007; Riaño-Pachón et al., 2008) based on the *C. reinhardtii* genome annotation v.3.1 (Merchant et al., 2007). The expression patterns of these genes were obtained for cells grown photoautotrophycally in constant illumination and subjected to a shift from High-CO_2 (HC) to Low-CO_2 (LC) conditions. Based on the results obtained novel candidate genes involved in CCM were the proposed and a gene regulatory network underlying the CCM was reconstructed. I addition, a genome-wide identification of nucleosome-depleted regions was performed, resulting in the discovery of putative regulatory elements of the CCM-related genes.

5.3 Materials and Methods

5.3.1 Cell growth conditions and strain

C. reinhardtii (strain cc503 cw92 mt+) was cultured in 2L flasks in photoautotrophic, and temperature controlled conditions. Cells were cultured in $H_{20}P$ medium (modified TAP medium without acetate supplemented with 20mM HEPES as buffering agent) under continuous light (~200µE m^{-2} s^{-1}) and constant supply of air with 5% CO_2 (gas mixture was provided by B-DCU Biostat fermenter, Sartorius, Goettingen, Germany) and controlled via a GC system (Agilent Technologies 3000 micro-GC, Waldbronn, Germany). After the cell

culture reached at an optical density of 0.5 at 750nm (approximately $3x10^6$ cells·mL^{-1}) 900mL were sampled and the concentration of CO_2 was reduced to 0.04% in air. Samples of 20mL of cell culture were then collected after 60min and 120min, and 1L of cell culture was sampled after 180min. Samples from all time points were harvested for transcript analysis by centrifugation at 3000 *g* for 2min and the cell pellet was frozen in liquid nitrogen. The samples were kept at -80°C until further use. Cells (900mL) from the last time point (180min) were directly subjected to the crosslinking step of the FAIRE method, described later on in the section 5.3.5.

5.3.2 Primer design for qRT-PCR and validation

The design of primer pairs was based on the *C. reinhardtii* gene models of the genome annotation v.3.1 released by the Joint Genome Institute (DOE/JGI; http://genome.jgi-psf.org/Chlre3/Chlre3.home.html). Primer pairs were designed for the predicted transcription factors and transcription regulators genes described in the Plant Transcription Factor Database (Riaño-Pachón et al., 2007). Primers for carbonic anhydrase coding gene (GI:159468241, GeneBank at NCBI) were designed based on the coding DNA sequence (genomic coordinates: chromosome_4: 1849052..1853308) obtained from the *C. reinhardtii* genome annotation v.4 (http://genome.jgi-psf.org/Chlre4/Chlre4.home.html). Criteria for primer design are as follows (Udvardi, et al., 2008): T_m = 60 ± 1°C, length 18 to 25 bases, preferentially on exon-exon junctions. Primers were designed to have a GC content between 45 and 55% when possible (due to the high GC content of the Chlamydomonas), generating a single PCR product sizing between 60 and 150 bp. Primers were synthesized by Eurofins MWG Operon (Ebersberg, Germany). A complementary *in silico* validation of the specificity of the primers was carried out using the computational tool QuantPrime (Arvidsson et al., 2008). Experimental validation of primer pairs was performed by checking the presence and approximate size of PCR products obtained from the amplification of cDNA synthesized from RNA extracted from *C. reinhardtii* cells grown under photoautotrophic conditions as described on section 5.3.1, except that the concentration of CO_2 was kept constant at 2%. PCR products were separated in 3% agarose gels and amplicons which did not have the correct size (i.e., deviating by more than 20bp from the expected size, based on data

extracted from the genome annotation v.3.1) were excluded from further analysis. A list with the sequences and respective identifiers of the primer pairs used in the present study is shown in Table appendix D-1.

5.3.3 RNA extraction, cDNA synthesis and quantitative RT-PCR

A quantitative reverse transcription PCR (qRT-PCR) platform was established for the analysis of the transcript levels of the 245 predicted TF and TR coding genes, *Actin* and *Cah1*. Total RNA extraction was performed using the RNeasy Plant Mini Kit (Qiagen, Hilden, Germany). Briefly, the frozen cell pellet was homogenized in 450µL of lysis buffer RLT and frozen again in liquid nitrogen for three minutes. Cells were completely lysed by incubating the frozen cell pellet at 56°C for 3min. Further steps of the RNA extraction were performed as described by the manufacturer. In order to remove genomic DNA from the RNA extracts, on-column digestion of DNA was performed using DNAse I (Qiagen, Hilden, Germany) and an additional DNAse treatment was performed, on the RNA eluate, using TURBO DNAse (Ambion, Darmstadt, Germany) as indicated by the manufacturer. The integrity of the RNA was checked by electrophoresis on 2% denaturing agarose gels; RNA quality was assessed by determining the 260/280 absorbance ratio using a Nanodrop (Thermo Scientific, Schwerte, Germany). Furthermore, the absence of genomic DNA in the RNA extracts was assessed by performing a q-PCR on a 1µL aliquot of each sample of total RNA using primer pair annealing to an intergenic region of chromosome 16 (forward primer 5'-TGTCTTGTGAATCCTGCCCTC-3' and reverse primer 5'-AAAGAGCTCACAAGTACACACCGA-3'). Only if the q-PCR analysis of the RNA sample resulted in a Ct value higher than 36, the RNA was used for cDNA synthesis. Three micrograms of total RNA was used for cDNA synthesis reaction employing the SuperScript III First Strand System (Invitrogen, Darmstadt, Germany) according to the manufacturer's instructions, using oligo-(dT_{20}) as primer for the synthesis of the first complementary DNA strand. The efficiency of the cDNA synthesis reaction was estimated by qRT-PCR using primer pairs for the UBIQUITIN PROTEIN LIGASE coding gene (protein ID 190824 (DOE/JGI); forward primer 5'-TTACCTGCCTTCCGATTGCGTAGC-3' and reverse primer 5'-TTACTATGCCTGAGCACGCAGCAC-3'). The cDNA samples were diluted ten times prior to the final PCR reaction which was conducted with the SYBR Green mix (Applied Biosystems,

Darmstadt, Germany) within a final reaction volume of 5µL containing 5µM primers. Dilution of primers and pipetting into the PCR 384 well plates were performed with a robot to increase accuracy and throughput. The ABI PRISM 7900HT sequence detection system (Applied Biosystems, Darmstadt, Germany) was used for the qRT-PCR reactions which were carried out as previously described (Caldana et al., 2007).

5.3.4 Quantitative RT-PCR data analysis

Raw qRT-PCR data were pre-processed with the program SDS v.2.3 (Applied Biosystems, Darmstadt, Germany) and the amplification curves were analyzed considering a threshold of 0.2 for the variation of the fluorescence of the sample to the fluorescence of the passive dye to obtain the C_t values (Cycle threshold). The baseline measurement was taken from the 3rd to the 15th cycle. Quality controls were done in order to evaluate the raw data by means of efficiency of the amplification reactions using log-linear regression as described previously (Ramakers et al., 2003) and the patterns of the melting curves. Genes for which the amplification efficiency was lower than 95% in more than 25% of the reactions were excluded from further analysis. Reactions presenting multiple melting peaks or melting temperatures which disagreed with the expected temperature were also excluded from further analysis. Only high quality measurements were selected for the relative gene expression analysis. All raw Ct values of TF and TR genes were normalized by the Quantile method using the R package qpcrNorm (Mar et al., 2009) available through the Bioconductor project (www.bioconductor.org/*).* Biological medians of the normalized Ct values (ΔCt) from the five biological replicates were calculated. The relative gene expression levels were calculated as delta-delta Ct values ($\Delta\Delta Ct = \Delta Ct_{LC} - \Delta Ct_{HC}$). Fold change (FC) was calculated as $2^{-\Delta\Delta Ct}$ (Udvardi et al., 2008). Expression values were normalized by genes and mean centered before the clustering analysis, performed using the MeV suite v4.6. (TM4 Microarray Software Suite) (Saeed et al., 2006). Clustering of the gene expression profiles was performed using the Quality Threshold (QT) algorithm (minimum cluster size of 4 genes, diameter 0.5, and Pearson correlation as distance measure). Expression levels for *Cah1* were determined from three experimental replicates and expression levels calculated by the ΔΔCt method as previously described (Caldana et al., 2007) using actin coding gene

(GI:159482013; forward primer: 5´-GCGGCTAACGACGGAGGAT-3´; reverse primer: 5´-CCATGACCCGCTCCTCATATC-3´) as a reference gene.

5.3.5 *Formaldehyde-assisted isolation of regulatory elements* (FAIRE)

For the identification of nucleosome-depleted chromatin regions likely containing regulatory elements we used the FAIRE method (Figure 5-1) (Nagy et al., 2003; Giresi and Lieb, 2009). To this end, 900mL of *C. reinhardtii* cell culture was sampled three hours after the shift from HC to LC and crosslinked by stirring at 150rpm for five minutes at room temperature with 100mL of 10x crosslinking buffer (10% formaldehyde, 100mM KCl, 100mM NaCl, 100mM, 100mM HEPES, pH 8.0) reaching the final concentration of 1% formaldehyde. The crosslinking reaction was quenched by adding 1M glycine to a final concentration of 125mM and stirring at 150rpm for five minutes at room temperature. Cells were harvested by centrifugation at 22°C for 3 min at 3000 *g* in 1L bottles using a Beckman coulter Avanti J series centrifuge with fixed-angle rotor model JLA-16.250 (Beckman Coulter GmbH, Krefeld, Germany). The cell pellet was suspended in 10mL of resuspension buffer (0.5% Triton X-100, 10mM HEPES, 0.5mM EDTA, 10mM KCl, 10mM NaCl), transferred to 50mL conic tubes and centrifuged at 4°C for 3min at 3000 *g*. The pellet was resuspended with 10mL of Lysis buffer (1% Triton X-100, 10mM KCl, 1.5mM $MgCl_2.6H_2O$, 0.2M sucrose, 10mM HEPES, pH 8.0) supplemented with 1:100 plant protease inhibitor cocktail (Sigma, #P9599) and fresh 1mM DTT. The samples were incubated for 10 min on ice and centrifuged at 4°C for 20 min at 1000 *g*. The lysis step was repeated once and the final pellet was suspended in 2mL of Nuclear lysis buffer (140mM NaCl, 1mM EDTA, 1% Triton X-100, 0.1% sodium deoxycholate, 1% sodium dodecylsulfate, 50mM HEPES, pH 8.0) supplemented with 1:100 plant protease inhibitor cocktail (Sigma, #P9599) and fresh 1mM DTT. An aliquot of 500µL was separated to further check chromatin crosslinking efficiency. Samples were sonicated on ice (5 cycles at 40% power, 9 x 10% cycle, 10sec/cycle) using a Sonopuls HD 2070 sonicator (Bandelin Electronic, Berlin, Germany).

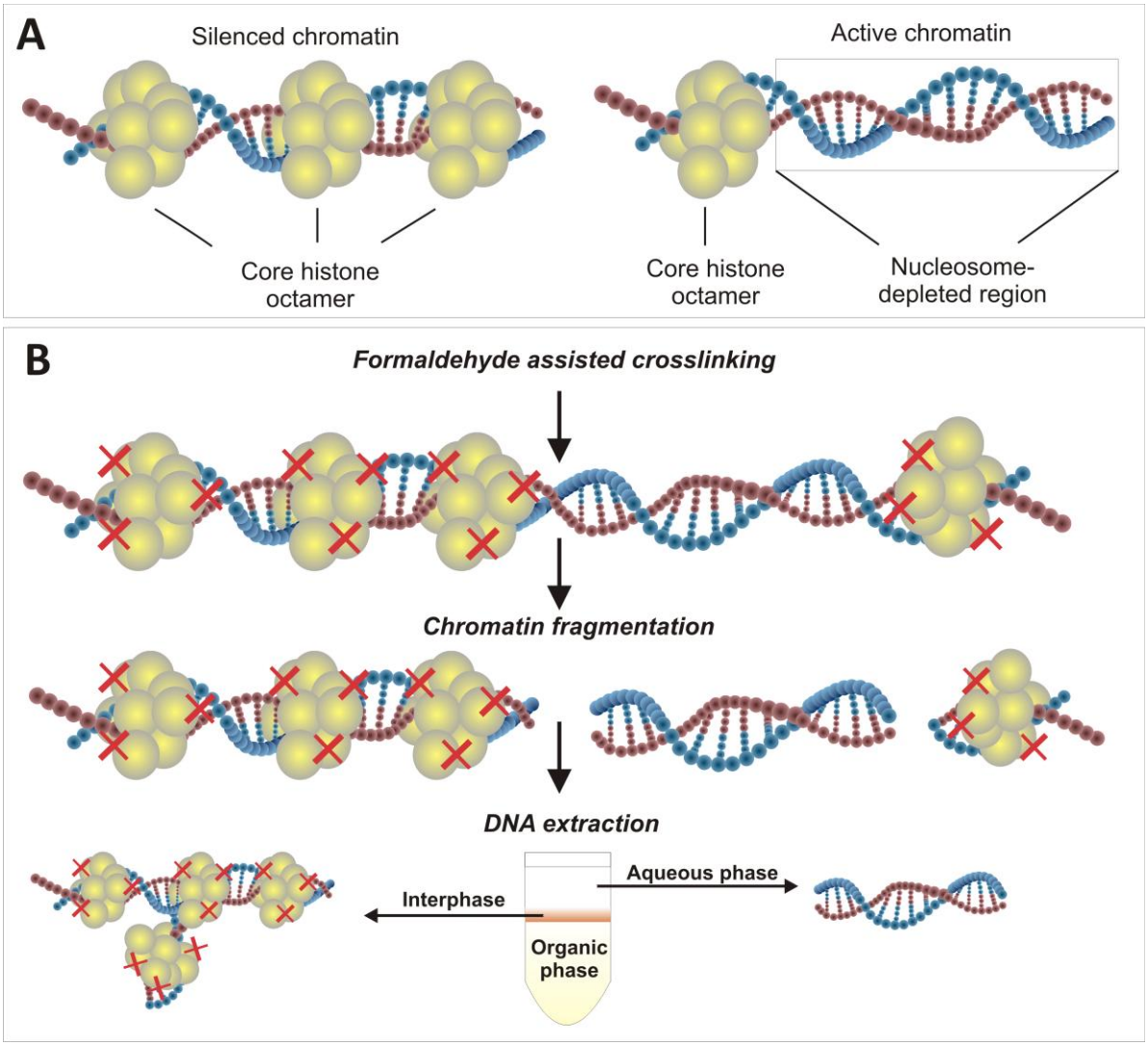


Figure 5-1. Illustration of active and silenced chromatin states and the formaldehyde-assisted isolation of regulatory elements (FAIRE).

Chromatin states are important elements for the regulation of gene expression. In the active state, chromatin shows a temporary displacement of nucleosomes, favoring the interaction of DNA-binding proteins with the genomic regulatory elements. Panel (A) shows the active and silenced chromatin, highlighting the difference between the two states. In panel (B), the FAIRE method is summarized: first, the total chromatin is crosslinked (indicated by red crosses) by formaldehyde. The crosslinked chromatin is then fragmented by sonication and phenol:chloroform DNA extraction is performed. The fragments of DNA that are isolated into the aqueous phase are the regions depleted for nucleosomes. Fragments containing high levels of crosslinked proteins are collected in the interphase.

After sonication samples were centrifuged at 4°C at 4000 *g* for 5min. The supernatant was kept at -80°C until further use. An aliquot of 500µL from the extracted sonicated chromatin

was taken for gel analysis and to serve as a control sample. Control samples were decrosslinked by overnight incubation at 65°C. Both, control and sonicated chromatin were subjected to a standard nucleic acid extraction procedure using Roti®-Phenol:Chloroform:Isoamylalcohol mixture (Roth, Karlsruhe, Germany) as previously described (Giresi and Lieb, 2009) and the samples containing the FAIRE extracted DNA fragments were kept at -80°C until further analysis.

5.3.6 Quantitative PCR of nucleosome-depleted DNA fragments isolated by FAIRE

Fragments of chromatin isolated using the FAIRE method were analyzed by quantitative PCR in order to check if the sample was enriched with defined DNA segments. We selected the *Cah1* genomic locus (chromosome_4:1849052-1853308, protein ID 24120; name: estExt_fgenesh1_pm.C_150006, *C. reinhardtii* genome annotation v.4) as a marker. The locus selected includes coding and non-coding regions (including 5´ upstream sequence). In Chlamydomonas, it has been described that the transcription of *Cah1* increases under carbon deprivation conditions. Previous studies revealed the position of putative enhancer and silencer elements in the promoter region 5´upstream of the translation initiation code (ATG) of *Cah1* (Kucho et al., 2003). Primer pairs were designed for ten segments of the genomic locus of *Cah1* (Table 5-1). A scheme of the regions covered by the primer pairs is shown in Figure 5-2.

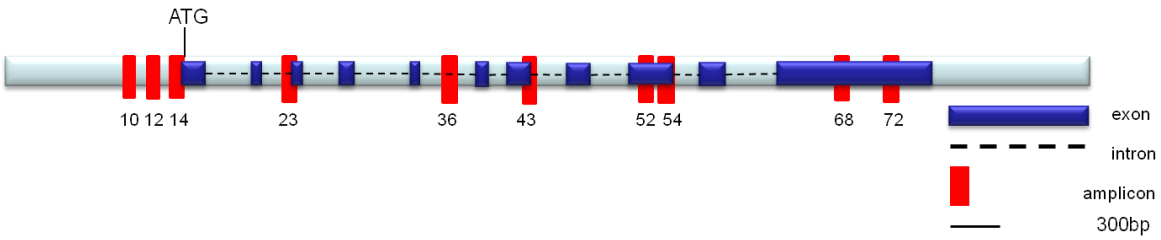


Figure 5-2. Fragments of the *Cah1* genomic locus analyzed by qPCR.

Quantitative real time PCR analysis was performed for ten DNA segments of the *Cah1* gene locus and enrichment of each fragment was measured. Red areas show the location of the DNA segments amplified and quantified. Exons are indicated by dark blue boxes and introns by dashed lines.

Table 5-1. Sequence and coordinates of the primer pairs used for enrichment analysis of FAIRE fragments through qPCR.

	*Primer pair coordinates**	*Sequences of oligonucleotides*	
Amplicon identifier**	**Start/end**	**Forward primer (5´-> 3´)**	**Reverse primer (5´-> 3´)**
10	660/729	TAGCCTTTCAAGCCGCGCCA	GCGTCAAACCTCCTTCCAACCG
12	804/867	GCCGGAACTCCAACCAAGTAGC	GCATTTCTGCATGCGCACAGT
14	937/1000	CCGCCGTACCGTTGTCACTTT	CGCAGGGTGTGGACAACTATGG
23	1574/1637	TAGGGCGATGAACGGTTCTGGT	TCCTTGCCCTCCCTGTATGTGG
36	2485/2545	TGGAGTGCGAGCACGCTTAGTA	CCAAGGCACCCAGATGCATGAC
43	2940/3008	GTTCCACTTCCACTCCACCTCG	AAAGTCAAGAGTTGCGCCCACG
52	3602/3663	CATCAAGCTGGGTGAGCTGCTG	TGAGGCTGCCCTCGTACGTTA
54	3719/3781	CATCAGCTTCGGCCAGTGGAAC	GTGGAGTTGCACTCCTTCAGGC
68	4714/4773	ATGGAGTTGGTTCCACGATGGG	ACGTCCGGCCCACTGACTTTAT
72	4996/5062	CTGTCTGGCTGGCTGGGTTGTT	TGCATTGATCGGCATGTGCGAC

*Coordinates taken from the *Cah1* genomic locus (chromosome_4:1849052-1853308)
**The position of the amplicons along the *Cah1* genomic locus is shown in Figure 5-2.

FAIRE samples were diluted to a final concentration of 50ng·µL^{-1} before the qPCR reaction and 1µL was mixed with 5µL of 2x SYBR Green mix (Applied Biosystems, Darmstadt, Germany) to a final reaction volume of 10µL containing 5µM primers. The reactions were conducted in 384 well plates and PCR amplification was conducted on an ABI PRISM 7900HT sequence detection system (Applied Biosystems, Darmstadt, Germany) as previously described (Caldana et al., 2007). To calculate the enrichment of genomic regions we used a method previously described (Giresi and Lieb, 2009). Briefly, the average of the resulting Ct values for the control and FAIRE samples were calculated. The ratio between FAIRE to control samples (FAIRE/control) was determined for each segment. For each sample the average Ct value of the adjacent region of the segment possessing the lowest resulting ratio (FAIRE/control) was taken as a reference region for calculating the ΔCt values ($Ct_{region\ of\ interest}$-$Ct_{reference\ region}$). The ΔΔCt values ($\Delta Ct_{FAIRE\ sample} - \Delta Ct_{\ control\ sample}$) were calculated and the enrichment level determined as $2^{-\Delta\Delta Ct}$. FAIRE experiments were done in triplicate and three technical replicates were done for each segment analyzed by qPCR.

5.3.7 Deep sequencing

For the genome-wide identification of nucleosome-depleted regions and putative regulatory elements of *C. reinhardtii* the DNA fragments present in the FAIRE samples were sequenced by deep sequencing technology. So far no FAIRE-seq data have been published for plant species and no protocol was available for such analysis in *C. reinhardtii*. FAIRE sample from one of the three experimental replicates analyzed previously by qPCR was selected for deep sequencing. The chromatin was extracted from cells grown for three hours under LC condition. After isolation of nucleosome-depleted genomic fragments using the FAIRE method the sample was fractionated on a 2% agarose gel and fragments between 200 to 400 bp were used for the construction of a DNA library for sequencing. DNA library preparation and DNA sequencing were performed by LGC Genomics (Berlin, Germany) according to standard protocols using Paired-end-DNA sample kit, HiSeq paired-end cluster generation kit with paired-end Flowcell and TruSeq SBS kit HT (200 cycles) according to the user guide instructions (Illumina, Eindhoven, The Netherlands). A total of 350ng of double strand DNA containing the FAIRE fragments was subjected to end repair and adenylation of the 3´ends for ligation of adaptors. The ligation products were purified from a gel and the fragments amplified by PCR (12 cycles). The final PCR product was purified with Ampure beads and the library validated and quantified using a Bioanalyzer (Agilent, Waldbronn, Germany) and Nanodrop (Thermo, Dreieich, Germany). Amplified DNA fragments were sequenced with a HiSeq2000 sequencing system (Illumina, Eindhoven, The Netherlands). Paired-end sequence reads of 50bp were mapped against the masked Chlamydomonas genome sequence v.4.0 (DOE/JGI Joint Genome Institute; http://genome.jgi-psf.org/Chlre4/Chlre4.home.html) using bowtie (Langmead et al., 2009) (URL: http://bowtie-bio.sourceforge.net/index.shtml). Mapping files on SAM format were manipulated with SAMtools (http://samtools.sourceforge.net/) and BEDtools (http://code.google.com/p/bedtools/) in order to export mapping positions into BED format. FAIRE-seq peaks were detected by using the Model-based Analysis approach, MACS (Zhang et al., 2008b). FAIRE-seq peaks were mapped to the Chlamydomonas DOE/JGI v.4.0 genome annotation in order to find the closest annotated features, thus identifying affected genes.

5.3.8 Reconstruction of gene regulatory network (GRN)

In order to reconstruct the gene regulatory network of Chlamydomonas during CCM, we used the novel IOTA (inner composition alignment) method (Donner et al., 2011). IOTA is a permutation-based asymmetric association measure which is capable to identify pairs of genes which putatively interact or are regulated, determines the directionality of this regulation, and in addition, is also capable to identify autoregulatory links. The measure was applied to the expression data of TF and TR coding genes analyzed in the present study in order to reconstruct the topology of the underlying regulatory network. IOTA determines additionally the robustness of the identified regulatory links, distinguishing also the indirect from the direct regulations, correctly identifying approximately 75% of the uni- and bidirectional regulatory links, while fixing the false positive rates at approximately 25%.

5.3.9 Motif analysis

To test whether the promoter regions of genes of interest had common sequence features, 5´ upstream regions of selected genes were retrieved from the Chlamydomonas genome annotation v.3.1 (DOE/JGI) and overrepresented sequence motifs were identified using the MEME suite (v.4.3.0, http://meme.nbcr.net/) (Bailey et al., 2006), allowing a maximum of 10 motifs to be reported and maximum motif length of 20 nucleotides. Other parameters were used as default settings. When significant motifs were found, an analysis of their specificity was conducted using MAST (http://meme.sdsc.edu/meme4_3_0/cgi-bin/mast.cgi) and a file containing 14598 promoter sequences obtained from the Chlamydomonas genome annotation v.3.1 (DOE/JGI).

5.4 Results and discussion

5.4.1 Gene expression analysis by transcript profiling of TF and TR genes

We measured changes on the TF and TR gene expression by qRT-PCR. Primer pairs were designed for the 245 TF and TR coding genes predicted previously (Riaño-Pachón et al., 2007) and validated. The results of qRT-PCR experiment showed that approximately 67% of these genes passed our criteria of quality control and were selected for further data analysis.

From these 165 genes, 130 genes coding for TFs and TRs were monitored at four different time points which correspond to 0h, 1h, 2h and 3h from the shift from High-CO_2 (HC) to Low-CO_2 (LC). Expression of thirty-five genes could not be quantified probably because their transcript abundance was below the limit of detection of the qRT-PCR or the respective genes were not expressed at all in the conditions analyzed. A graph representation of the gene expression fold change (FC) of the TF and TR genes monitored during carbon deprivation is shown in Figure 5-3.

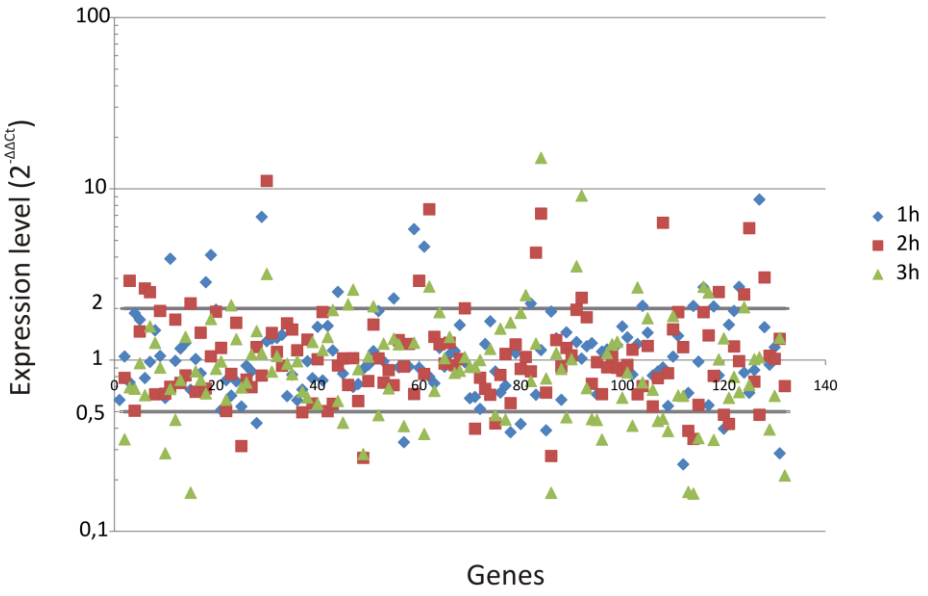


Figure 5-3. Gene expression fold changes during carbon deprivation.

Expression levels of the 130 TF and TR genes analyzed 60min, 120min and 180min after the reduction of the carbon dioxide concentration from 5% to 0.04% are shown. The darker gray lines cutting the plot area internally and horizontally represent the threshold of a 2 fold change. Dots above or below those lines represent genes exhibiting a more than 2 fold expression change.

5.4.2 Co-expressed TF and TR genes under low carbon concentration

The expression profiles of the 130 genes monitored were analyzed using the QT clustering method (using Pearson correlation as distance measure). Co-expressed genes were identified and clustered in groups showing similar expression profiles (Figure 5-4).

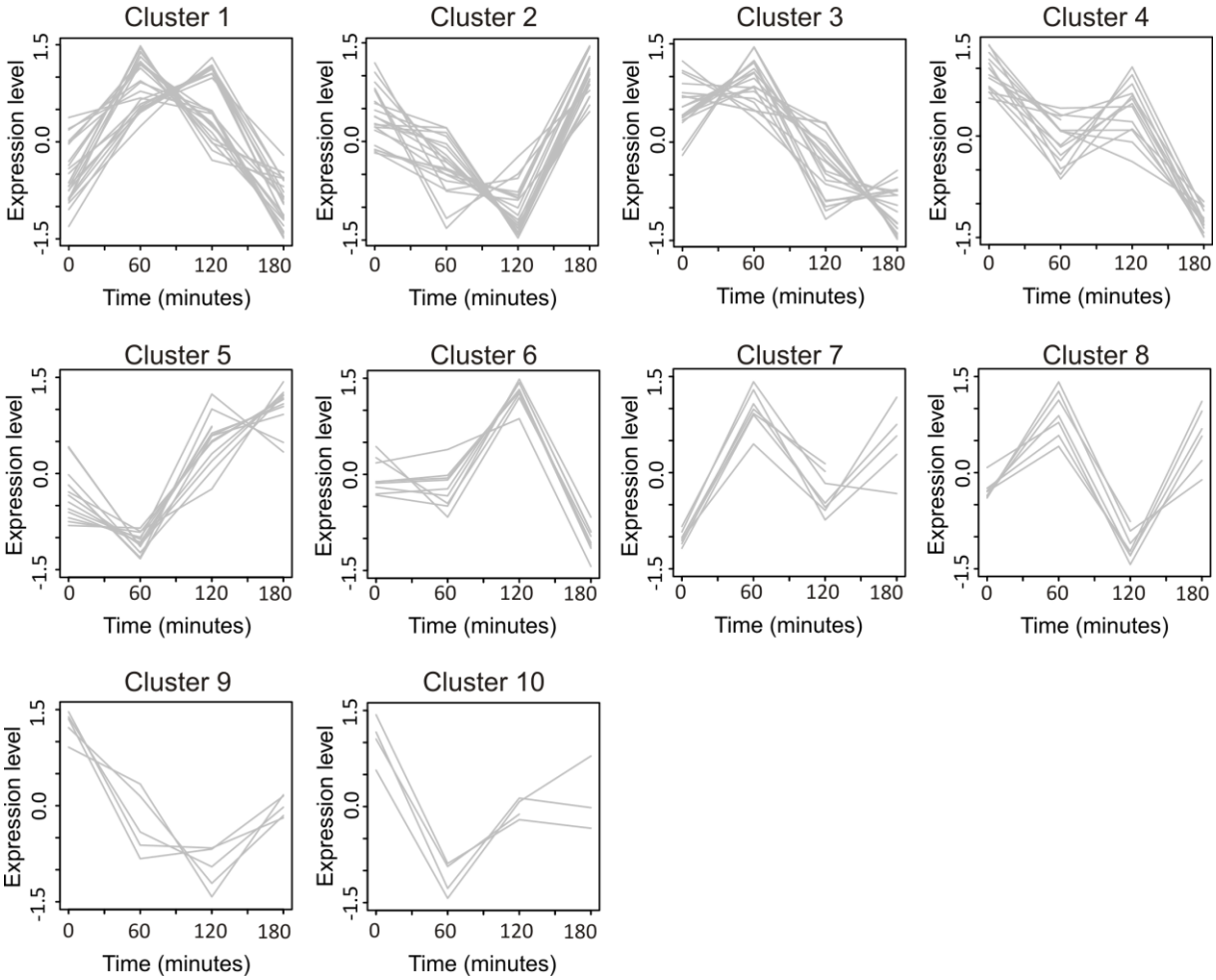


Figure 5-4. Co-expressed TF and TR genes under carbon deprivation.

The analysis of the transcript levels of TF and TR coding genes revealed co-expressed genes clustered according to similar expression profiles using the quality threshold method with Pearson correlation as distance measure. The normalized expression levels are represented in the y axis as the relative expression values (scaled, Log2) scaled by gene minima and maxima and mean centered. At time zero (0min) the carbon dioxide concentration in the growth medium was shifted from 5% to 0.04%. Samples were taken immediately before the shift (indicated as 0min) and after 60, 120 and 180 min.

Several TF and TR genes were found to be regulated by the decrease in CO_2 concentrations. One hundred and twenty one genes were grouped into ten clusters; nine genes could not be assigned to any cluster. A list with the genes and their respective cluster is presented in Table appendix D-2. Genes induced or repressed (FC > 2) are listed in Table appendix D-3.

5.4.3 Early response TF and TR genes under carbon deprivation

From the genes up-regulated after the transition from HC to LC some increased in expression already within one hour. Among these early responsive genes, we found that the following members of different TF and TR gene families: DDT (protein ID: 8155; FC > 8), CSD (protein ID: 126810; FC > 6), MYB-related (protein ID: 166618; FC > 5), HSF (protein ID: 117914; FC > 4) and SNF2 (protein ID: 111277; FC > 3), all of it were up-regulated at least 2 fold (Table appendix D-3).

The transcriptional regulator DDT was predicted to have three domains: Bromodomain, DDT and PHD and so far only one gene of this family was found in the genome of Chlamydomonas (Riaño-Pachón et al., 2007). Bromodomains have been shown to be important for chromatin remodeling through the formation of protein complexes that can recognize acetylated lysines of core histones and further causing disassembling of the nucleosome structure (Akai et al., 2010). Since more than three decades the acetylation of histones is known to directly correlate with increased rates of RNA synthesis (Marushige, 1976). However, how the histone modifications and their interplay with cofactors function to regulate the transcription of a certain gene *in vivo* is still under investigation (Jiang et al., 1997; Chen et al., 2010).

Another TF identified among the early responsive genes encodes a member of the CSD protein family and the presence of two domains, CSD and RRM, on its sequence was previously predicted (Perez-Rodriguez et al., 2009). The CSD domain is known to be related to cold shock responses and the RRM domain has been described to be related to RNA binding activities. In contrast to prokaryotic CSD proteins, eukaryotic proteins containing a CSD domain may also contain other domains and their functions are not exclusively related to cold shock responses (Graumann and Marahiel, 1998). Bacteria and vertebrate proteins containing CSD domains have been shown to be involved in many different cellular processes and are able to bind to nucleic acids, either DNA or RNA, affecting transcription and translation events (Chaikam and Karlson, 2010). A major cold shock protein in *E. coli*, CSPA, was described to be involved in mechanisms for destabilizing mRNA secondary structure favoring translation during cold acclimation. It was shown that CSPA binds to

ssRNA molecules in a cooperative way dependent of the length of the RNA molecule and concentration of CSPA and it was suggested to function as a RNA chaperone (Jiang et al., 1997). In plants, it has been shown that Arabidopsis cold shock protein CSDP1, containing the CSD domain, can complement the cold sensitivity phenotype in defective bacteria and its RNA chaperone activity was verified (Karlson and Imai, 2003; Kim et al., 2007). In our experiment, the expression of CSD was up-regulated within one hour and started to decrease after two hours under LC. Most likely, the RNA chaperone function of CSD is important during this time for favoring the translation of the new transcripts which should be produced in a fast manner due to the sudden changes on the CO_2 levels.

The MYB-related TF early induced under LC was predicted to contain one MYB DNA-binding domain. Its transcript has already been identified previously in Chlamydomonas cells under carbon deprivation conditions. However, this gene has not been functionally characterized (Asamizu et al., 2000). Members of the MYB-related TF family in Arabidospsis were showed to be involved in hydrogen peroxide responses (Desikan et al., 2001), phosphate uptake (Nilsson et al., 2007), and responses to hormones, dehydration and salt stress (Martin and Paz-Ares, 1997). In Chlamydomonas, a MYB domain-containing transcription factor, member of the G2-like TF family, has been found to be involved in the regulation of phosphate uptake and mutant lines for this gene, named *Psr1,* revealed the importance of PSR1 for cellular responses during P starvation (Moseley et al., 2009).

Our data also revealed an induced over-expression of *Hsf1* (protein ID: 117914), a member of the heat shock transcription factor (HSF) TF family. Transcription factors of the HSF family are widely distributed among eukaryotic species. In plants, this protein family usually contains many members. In contrast, for instance, *Saccharomyces cerevisae* (yeast) has only one member of this family. Also in yeast, this protein binds constitutively to heat shock elements located in the promoter region of target genes and it was demonstrated that in non-heat shock conditions it has a basal transcriptional activation (Miller and Mittler, 2006). In Arabidopsis plants, 21 genes were found to be part of the HSF family. Arabidopsis HSF proteins of class A, the largest group of HSF in Arabidopsis, contain an activation domain in the C-terminal part which has been suggested to be involved with transcriptional activation. The recognition of new domains, such as CTAD and NES, led to the identification of more

than 60 new candidate HSF proteins of class A in Arabidopsis (Kotak et al., 2004). In Chlamydomonas, two genes, *Hsf1* and *Hsf2,* have been annotated as members of the HSF protein family. A comparison of the protein sequences of HSF1 from Chlamydomonas with HSFs from flowering plants has revealed that HSF1 is phylogenetically closest to the Arabidopsis HSF proteins of class B1. However, there are indications that HSF1 of Chlamydomonas contains an extended insert sequence between its two oligomerization domains and harbor a putative nuclear export domain and aromatic, hydrophobic and acidic element domains reflecting higher similarity to Arabidopsis HSF proteins of class A. Previous studies revealed that Chlamydomas HSF1 was over-expressed during heat shock. Chlamydomonas RNAi lines generated for HSF1 showed reduced expression of HSF2 and HSP90 during heat shock and it has been suggested that HSF1 constitutively and specifically interacts with HSP70A and could be a key regulator in stress response in Chlamydomonas (Schulz-Raffelt et al., 2007). Interestingly, in Arabidopsis plants the steady-state transcript level of the different *Hsf* genes seems to be different in different tissues. The transcript level of *Hsf* genes has also been shown to vary under different conditions, such as moderate light and oxidative stress (Davletova et al., 2005; Miller and Mittler, 2006). In Chlamydomonas, previous studies revealed the identification of two EST variants for *Hsf1* in cells grown under LC condition but their specific functions were not determined (Asamizu et al., 2000).

Another gene we found highly expressed shortly after the LC shift is a TR member of the SNF2 family and its DNA sequence possess high similarity (e-value = 5e-174) to the *Rad54* gene in *Arabidopsis thaliana*. RAD54 is known to be involved in chromatin remodeling and has an important role in homologous recombination. Its overexpression enhances resistance to DNA damage in Arabidopsis (Klutstein et al., 2008). In yeast, RAD54 binds to dsDNA, causing DNA structural alterations, and modulates the interaction of proteins with DNA (Solinger et al., 2002).

The over-expression of *Hsf* in the first steps of CCM could also be an additional evidence of the role of HSF proteins in regulating gene transcription, likely recruiting remodeling factors which can together function in the transient displacement of histone octamers, as previously suggested (Erkine and Gross, 2003).

The functional analysis of mutants or RNAi lines for these genes may help to understand their role in the CCM and revealing which transcriptional regulatory function they exert in *C. reinhardtii*. Previous evidences of their role in other types of stress responses also indicate their possible function as general stress responsive genes.

5.4.4 Late responsive TF and TR genes under carbon deprivation

During carbon deprivation, cellular responses lead to the induction or repression of the expression of genes necessary for the control of the CCM (Miura et al., 2004). A master regulator of the expression of CCM-related genes, *Ccm1*, is constitutively expressed in LC and HC conditions (Fukuzawa et al., 2001) and its mutation modified the expression levels of many low carbon inducible genes, such as putative membrane proteins which have been suggested to be involved in the regulation of Ci uptake and carbon metabolism (Miura et al., 2004). One of the genes also regulated by CCM1, named *Lcr1* (Low-CO_2 stress response), is a TF member of MYB-related family which is involved in the regulation of downstream target genes important to the CCM, such as *Lci1*, *Lci6* and *Cah1* (Yoshioka et al., 2004). In the group of genes over-expressed by the cells grown for 180min under LC we found the predicted MYB-related TF (protein ID 184359) which codes for the LCR1 protein (GI:46020170, JGVI/Chlre4 transcript Id: OVA_estExt_fgenesh2_kg.C_230055), among the top five higher genes. Other TF and TR coding genes were also expressed at high expression levels and of our knowledge only *Lcr1* has been functionally characterized with respect to CCM (Figure 5-5).

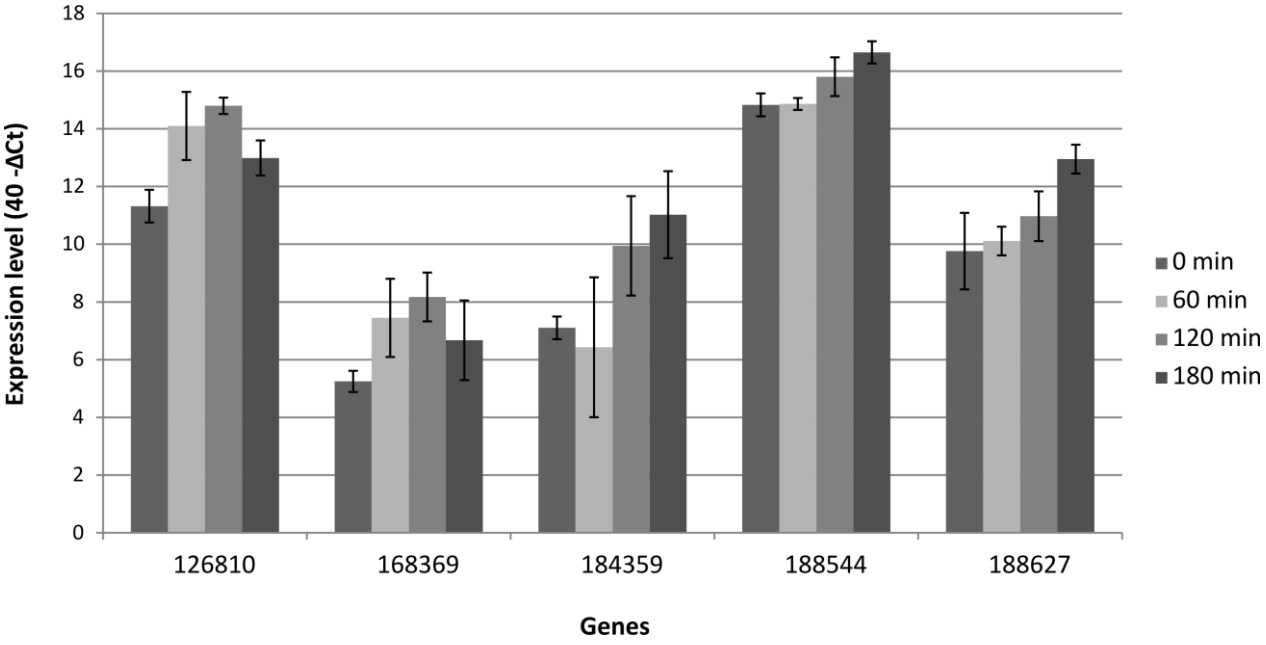


Figure 5-5. TF and TR genes showing prominent expression changes after three hours of cell growth at LC condition.

In Chlamydomonas cells grown for three hours in low CO_2concentrations (0.04%) 40 genes showed an alterated expression (FC >2) compared to cells grown in high CO_2 concentrations (5%). The graph shows the five genes for which the most prominent expression changes were observed after 180min at LC condition. Among them is the *Lcr1* (protein ID 184359), known as Low-CO_2 response regulator. The other genes represented in the plot encode members of the CSD (protein ID 126810), ABI3VP1 (protein ID 168369), SNF2 (protein ID 188544) and MYB-related (protein ID 188627) families. Error bars indicate standard deviation.

Based on our observations, it seems possible that some of these genes are involved in the initial regulation of the expression of *Cah1* since deletion mutation of *Lcr1* in Chlamydomonas did not completely abolish the expression of *Cah1* during carbon deprivation (Yoshioka et al., 2004). The function of *Lcr1* in the regulation of expression of *Cah1* has been suggested to be related to the amplification and maintenance of the levels of *Cah1* transcripts in response to low concentrations of CO_2 and no function related to the initial induction of the expression of *Cah1* was demonstrated (Yoshioka et al., 2004). A list with other TF and TR genes which showed a fold change higher than 2 after three hours from the shift from LC to HC is showed in Table appendix D-3.

5.4.5 Induction of CCM and identification of putative regulatory elements of Cah1

During the CCM, many genes have previously been shown to be induced when Chlamydomonas cells grown under LC conditions. Among them and conserved between plant species is the gene coding for the enzyme carbonic anhydrase (CAH1) (Dionisio-Sese et al., 1990). In our expression analysis, we could observe and verify the increase in the abundance of the transcripts for *Cah1*, as expected (Figure 5-6).

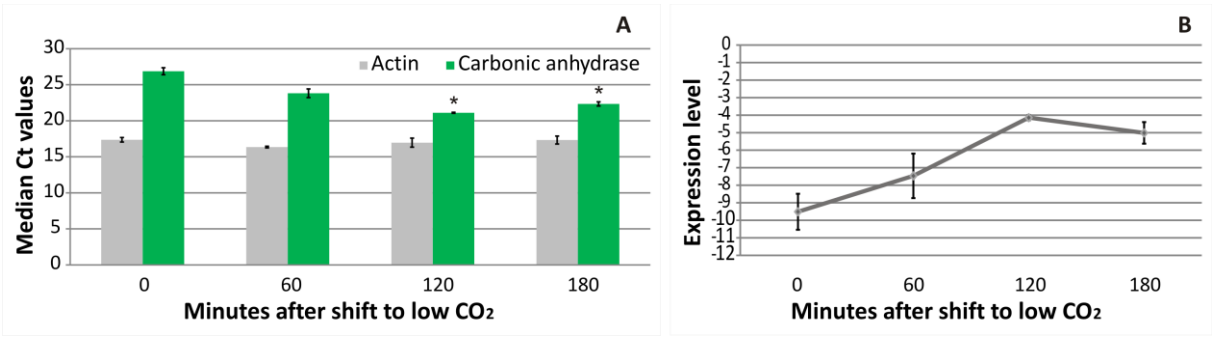


Figure 5-6. Expression levels of *Cah1*, encoding carbonic anhydrase 1.

The expression level of *Cah1* was determined by measuring its transcript abundance by qRT-PCR. The median Ct values ±SE of the gene coding for actin and carbonic anhydrase is shown in (A) demonstrating the relative stability of *ACTIN* expressionduring the shift from HC to LC. Figure A also shows the expression level of *Cah1*. *ACTIN* was used as reference gene for determining relative *Cah1* expression; the comparison of expression levels between the HC and LC conditions showed that *Cah1* is over-expressed (Log2, FC >2) at low concentration of CO_2. Time points where significant changes were detected (p-value<0.05, on dependent t-test for paired-samples) are indicated by a star (*). The relative expression levels of *Cah1* after the shift from HC to LC are shown as –ΔCt values in (B). Error bars indicate standard deviation (SD).

The over-expression of *Cah1* in LC conditions was expected as previously discussed, and this finding indicated that the CCM was induced in our experimental conditions. *Cah1* has been described to be induced by a concomitant effect of light and LC concentrations. It has been previously shown that under carbon deprivation conditions, Chlamydomonas cells grown in the dark do not show an increase in the transcript levels of *Cah1* in contrast to cells cultured in red, blue or white light (Dionisio-Sese et al., 1990). The sequence 5´upstream of *Cah1* coding region contains at least one silencer and one enhancer region (Kucho et al., 1999). In the enhancer region two putative enhancer elements, EE-1

(AGATTTTCACCGGTTGGAAGGAGGT; -293bp to -269bp from the first ATG) and EE-2 (CGACTTACGAA; -241bp to -231bp from the first ATG) were found. Previous experimental findings indicated the existence of a nuclear protein which binds to these two enhancer elements independently of the carbon dioxide concentrations and light illumination. Within the two EEs, four copies of a sequence motif (consensus sequence "GANTTNC ") were found (Kucho et al., 2003). This motif shows high sequence similarity with the plant box III motif which has been shown to be a light responsive element. In order to identify possible regulatory elements and in a further approach their genome-wide distribution in *C. reinhardtii* we analyzed the nucleosome depletion pattern of *Cah1* locus. FAIRE samples from *C. reinhardtii* grown for three hours in LC were obtained and analyzed by qPCR. I found that DNA-protein complexes are crosslinked within five minutes when cells are incubated in 1% formaldehyde (Figure 5-7). When de-crosslinked samples were subjected to the DNA extraction procedure using phenol-chloroform the recovery of DNA from the aqueous phase was possible. When DNA was crosslinked to proteins, it was not possible to extract the DNA since all complexes tend to be segregated to the interphase between the aqueous and organic phases during the DNA extraction procedure and this difference can be observed on agarose gels (Figure 5-7).

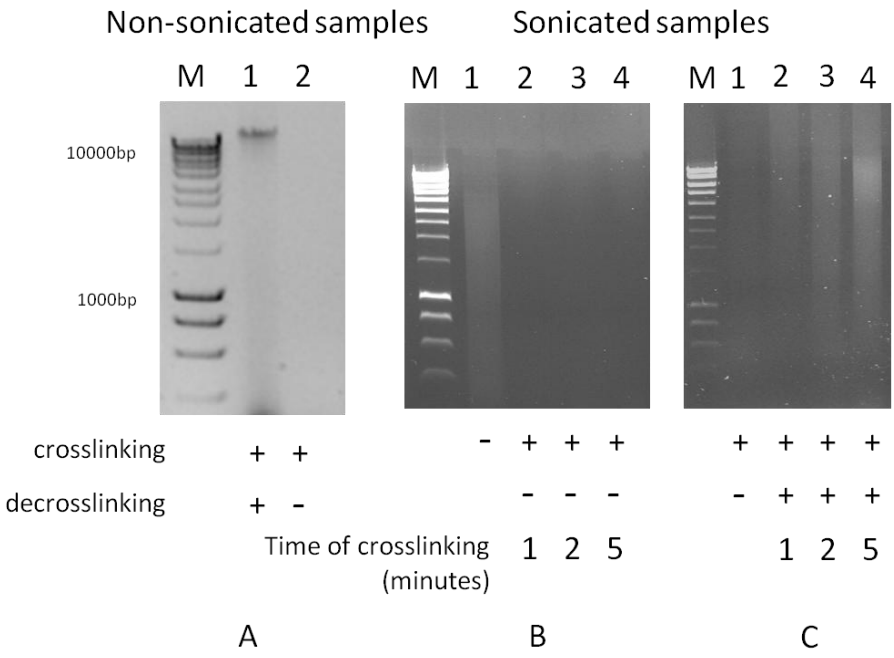


Figure 5-7. Test of crosslinking efficiency and DNA recovery from crosslinked and de-crosslinked FAIRE samples.

The efficiency of the crosslinking procedure was accessed by separation of the chromatin samples on 1% agarose gels. Samples of non-sonicated chromatin crosslinked for five minutes and de-crosslinked overnight at 65°C, were subjected to ordinary DNA extraction using phenol-chloroform (A). As can be seen in lane 1, de-crosslinked DNA was recovered from the aqueous phase by the DNA extraction procedure. Crosslinked DNA generally accumulates in the interphase between the aqueous and organic phase during the DNA extraction procedure, and it was not recovered, as expected, by the DNA extraction procedure (A, lane 2). Effects of incubation times on crosslinking of DNA were analyzed in sonicated FAIRE samples and are shown in (B) and (C). Non-crosslinked chromatin (B, lane 1) and chromatin crosslinked by incubation with 1% formaldehyde for 1 min (B, lane 2), 2min (B, lane 3) and 5min (B, lane 4) are shown. DNA crosslinking was observed within 5min of incubation with 1% formaldehyde. In (C) the de-crosslinking of the sonicated FAIRE samples permitted us to efficiently recover DNA by phenol-chloroform extraction from the samples previously crosslinked (shown in Figure B). M, molecular weight marker. Labels (+) and (-) indicate whether a certain treatment (crosslinking or de-crosslinking) was applied or not, respectively.

The FAIRE samples were subjected to qPCR analysis to check for enrichment of DNA fragments representing different regions of the genomic *Cah1* locus. Primer pairs were designed to cover the coding and non-coding genomic regions of *Cah1.* The quantitative PCR analysis revealed enrichment for DNA fragments located within the 5´upstream region of the

Cah1. The position where these enriched fragments are located matches to the region where regulatory elements have been previously found on the *Cah1* promoter during Low-CO_2 conditions (Kucho et al., 1999). We also found an enriched fragment located at the 3´UTR region of *Cah1* (Figure 5-8). Our findings suggest that it is possible to identify potential regulatory elements in the genome of *C. reinhardtii* using the FAIRE approach. Of note, a region downstream of *Cah1* was also enriched, indicating that nucleosomes are also depleted in this particular position of the chromatin and that the enriched region downstream of *Cah1* could function as a binding region recognized by DNA-binding proteins.

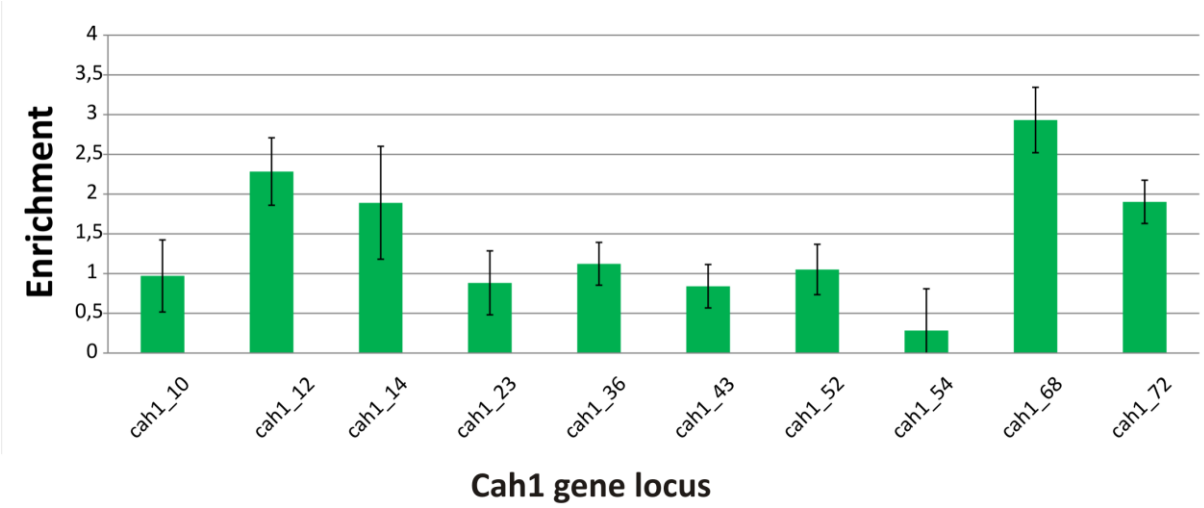


Figure 5-8. Quantitative PCR of the FAIRE fragments found in the genomic locus of *Cah1*.

DNA segments located in the promoter region (amplicons cah1_12 and cah1_14) and in a region located in the last exon of the *Cah1*, representing the 3´UTR (amplicons cah1_68 and cah1_72) of *Cah1*, were found to be enriched. Decrosslinked chromatin was used as control for calculating the enrichment of genomic fragments in the FAIRE samples. The enrichment was determined as $2^{-\Delta\Delta Ct}$ (see Material and Methods) and error bars indicate the standard error (SE) of the mean values.

5.4.6 Genome-wide mapping of nucleosome-depleted chromatin DNA segments

Our qPCR results for FAIRE samples indicated the whole set of putative regulatory elements of *C. reinhardtii* could be identified by sequencing all DNA fragments obtained with FAIRE using deep sequencing technology (FAIRE-seq). The FAIRE approach was first described by Nagy and colleagues in 2003 (Nagy et al., 2003). Initially these authors were interested in the

identification of genome-wide histone methylation patterns. It is known that histone H3 lysine 4 methylation by certain methyltranferases, e.g. SET1p from *S. cerevisae*, is associated with active chromatin, and active and repressed genes may be distinguished by di- or trimethylation of histone lysines (Boa et al., 2003; Gencheva et al., 2006). The discovery of regulatory DNA elements in formaldehyde crosslinked chromatin was a serendipitous process and it was observed when researches were trying to perform chromatin immunoprecipitation (ChIP) experiments and by a "fortunate mistake" they forgot to decrosslink the isolated chromatin used in the immunoprecipitation before subjecting it to phenol-chloroform DNA extraction. The authors observed that once they did not reverse the crosslink, there was a much greater efficiency of recovering non-coding sequences in the aqueous phase of the DNA extraction (Giresi et al., 2007; Giresi and Lieb, 2009). These observations were also confirmed by experiments using microarrays of the entire yeast genome (Hogan et al., 2006). The isolation of regulatory elements is possible due to the high number of lysines in histone proteins which makes the crosslink to formaldehyde more likely to happen within the nucleosomes than in the "open" nucleosome-depleted chromatin regions, which contains other proteins with weaker interactions to the formaldehyde and are consequently less attached to the chromatin during the DNA-protein crosslinking process (Toth and Biggin, 2000; Sutherland et al., 2008). It has been demonstrated that the regions enriched in the aqueous phase of FAIRE samples are strongly negatively correlated with the nucleosome occupancy in yeast (Hogan et al., 2006). Additional FAIRE experiments previously performed with human cell cultures showed that the enrichment with nucleosome-depleted regions was positively correlated with the presence of transcriptional start sites and long-distance regulatory elements, such as enhancers and silencers of the *Homo sapiens* genome (Giresi and Lieb, 2009). The FAIRE approach also detected nucleosome-depleted genomic regions in *Zea mays* and the active chromatin states for alleles of genes involved in the biosynthesis of anthocyanin (Louwers et al., 2009). For the study presented here, we established the FAIRE method for isolating nucleosome-depleted chromatin DNA fragments of *C. reinhardtii*, and further sequenced the DNA fragments which were isolated using the FAIRE method. The resulting sequencing reads were aligned against the masked Chlamydomonas genome sequence v.4.0 (DOE/JGI, Joint Genome Institute;

http://genome.jgi-psf.org/Chlre4/Chlre4.home.html) allowing 2 nucleotide mismatches at maximum. Computational analysis was performed using the MACS tool, revealing the regions with a significant ($p < 10^{-5}$) tag enrichment along the genome (FAIRE peak) and the regions which the accumulation of these tags, for each peak, was highest (FAIRE summit). In total, 6442 FAIRE peaks were found along the genome. This information was made available through a web genome browser (http://tartarus.uniandes.edu.co/cgi-bin/gbrowse/Chlamydomonas_v4/), where the FAIRE peaks, FAIRE summits, FAIRE-seq reads and the annotated transcripts (http://genome.jgi-psf.org/Chlre4/download/FrozenGeneCatalog_20080828_transcripts.fasta.gz) were included, thus allowing the visualization of FAIRE peaks putatively influencing particular loci (Figure 5-9).

Figure 5-9. FAIRE peaks genome browser.

On the genome browser FAIRE peaks, FAIRE summits, sequence read coverage under low CO_2 and annotated transcripts of Chlamydomonas v.4 are available. The *Cah1* locus is shown as an example here. FAIRE peaks are shown as green horizontal bars, corresponding to the nucleosome-depleted regions identified by FAIRE-seq. The *Cah1* gene structure is shown

above the FAIRE peaks (with exons indicated in yellow). This browser allows the users to track any region of the nuclear genome and the corresponding FAIRE peaks.

Our findings indicate that the genomic regions of *C. reinhardtii* depleted for nucleosomes have a tendency to be localized in intergenic regions, but close to transcription start sites (TSS) or to transcription termination sites (TTS) in most cases. We found that for the 6442 FAIRE peaks identified, 4206 FAIRE summits could be assigned to a closer gene´s coding sequence. The distance between the 5´ and 3´ extremities from the gene´s coding sequence to the FAIRE summit was calculated. Approximately 65% of the FAIRE peak summits are located up to a distance of 500bp from the extremity (either 5´start or 3´end) of a gene´s coding sequence. However, there are also peaks which are located thousands base-pairs away from the closer gene´s coding sequence. A histogram showing the distribution of the FAIRE summits and the correspondent distance to the gene´s coding sequence is plotted (Figure 5-10).

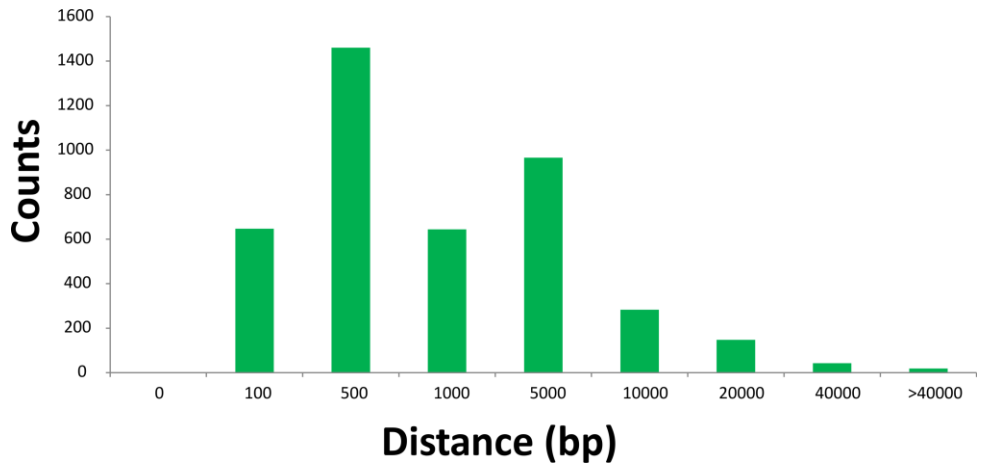


Figure 5-10. Histogram of the approximate distances of FAIRE summits from their next closest annotated gene.

Genomic coordinates were used to calculate the distances between FAIRE summits and the start and stop codons of their closest gene.

The presence or absence of FAIRE peaks close to a certain transcript was further investigated. We were interested to understand whether the presence of a FAIRE peak close to a gene´s coding sequence was associated with differential gene expression. TF and TR

genes differentially expressed (FC > 2, including up and down-regulated TF and TR genes) were included in this association analysis. The statistical analysis indicated a significant association of the two variables, (presence of FAIRE peak and fold change > 2; p-value = 0.005 using a chi-square test).

It is therefore more likely to find a FAIRE peak close to annotated transcripts of genes that are affected at the expression level by more than 2-fold, than close to genes that are only weakly affected by a treatment (here: shift from HC to LC concentrations). We did not see correlation between the presence of a FAIRE peak and exclusive up-regulation of gene expression (p-value = 0.7). The same conclusion was previously drawn for nucleosome occupancy studies at gene promoters in yeast (Hogan et al., 2006). Our experimental data suggest that the FAIRE approach reveals the identity of nucleosome-depleted regions and potential genomic regulatory elements of *C. reinhardtii*. Moreover, we identified that seventeen TF and TR genes analyzed in our transcript profiling have a correspondent FAIRE peak located closely to their coding region. The genomic coordinates of these peaks are shown in the Table appendix D-4 . It was then reasonable to assume that these seventeen genes may share similar sequence motifs in their promoter regions. To investigate this possibility, sequences of the 5´ upstream regions of the TF and TR genes for whose FAIRE peaks were identified, were analyzed using the MEME suite and the results revealed the existence of conserved motifs statistically significant (Table appendix D-5). Ten motifs identified by MEME were also analyzed using the MAST tool in order to predict their specificity and statistical significance of the conserved motifs found over a dataset containing 14598 annotated promoter sequences extracted from the Chlamydomonas genome annotation v.3.1 (DOE/JGI). The MAST analysis revealed that all seventeen promoter sequences of the TFs and TRs which contained a correspondent FAIRE peak presented p-value less than 0.0001 when compared to the whole promoter dataset, which indicates that those motifs, in the combination in which they were found on each promoter sequence, are statistically significant.

5.4.7 Chlamydomonas CCM gene regulatory network

Reconstruction of the gene regulatory network of *C. reinhardtii* grown in LC using the IOTA method (Donner et al., 2011), revealed the genes that putatively regulate the expression of other genes and the regulatory links of 122 TF and TR coding genes. This approach allowed us to visualize the topology of the CCM network (Figure 5-11), detecting the presence of already known LC responsive regulators such as LCR1 and generating new insights into the regulation of CCM. For example, in the reconstructed GRN, it was predicted that LCR1 regulates five others TF or TR coding genes (protein IDs: 93505, 120035, 152660, 160281 and 192634) ($p > 0.95$). It was also possible to identify two additional genes which could regulate a subgroup of genes predicted to be regulated by LCR1. These two genes are predicted TF gene of the MYB-related family (protein ID 188627) and a TR member of SNF_2 family (protein ID 188544).

The MYB-related TF was predicted to regulate four (protein IDs: 93505, 152660, 160281 and 192634) of the five genes putatively regulated by LCR1 with a probability higher than 0.95. These four genes are members of the following TF and TR families, respectively: SBP, Orphan, FHA and C3H. A motif search analysis using MEME was performed for the promoter region of the two MYB-related TFs which were predicted to share common target genes. One of them codes for LCR1 (protein ID 184359) and the other codes for a putative MYB TF (protein ID 188627) which has no functional annotation. The motif analysis revealed that both genes have conserved sequence motifs on their promoter; the sequences of these motifs are shown in Table appendix D-6. This finding suggests that both genes may be co-regulated by similar DNA-binding proteins if the motif found is truly a DNA cis-element sequence. Not surprisingly, not all motifs found were specific for these two sequences. When all motifs were analyzed, individually, using the MAST tool against the 14598 promoter sequences of the Chlamydomonas genes, some of them showed no statistical significance. Interestingly, one of the motifs found in the promoters of these two genes, Motif 4 (Table appendix D-6), seems to be very specific to these promoter regions. The results of MAST analysis of the Motif 4 against the whole 14598 promoter sequences of Chlamydomonas revealed that only these two genes contain this motif (p-value below

0.0001), reinforcing our suggestion of co-regulation of both MYB TFs during carbon deprivation. The TR member of SNF2 family was predicted to regulate three (protein IDs: 93505, 152660 and 192634) of the five genes predicted to be regulated by LCR1. Our transcript profiling data revealed that these putatively regulated genes are down-regulated after three hours under LC (see Table appendix D-3), indicating a possible role of LCR1, MYB-related (protein ID 188627) and SNF2 (protein ID188544) as gene expression repressors in CCM. These results gives us a new insight into the regulation of CCM-related genes, revealing new candidate regulatory genes and gives emphasis to the putative function of LCR1 as transcriptional repressor during CCM.

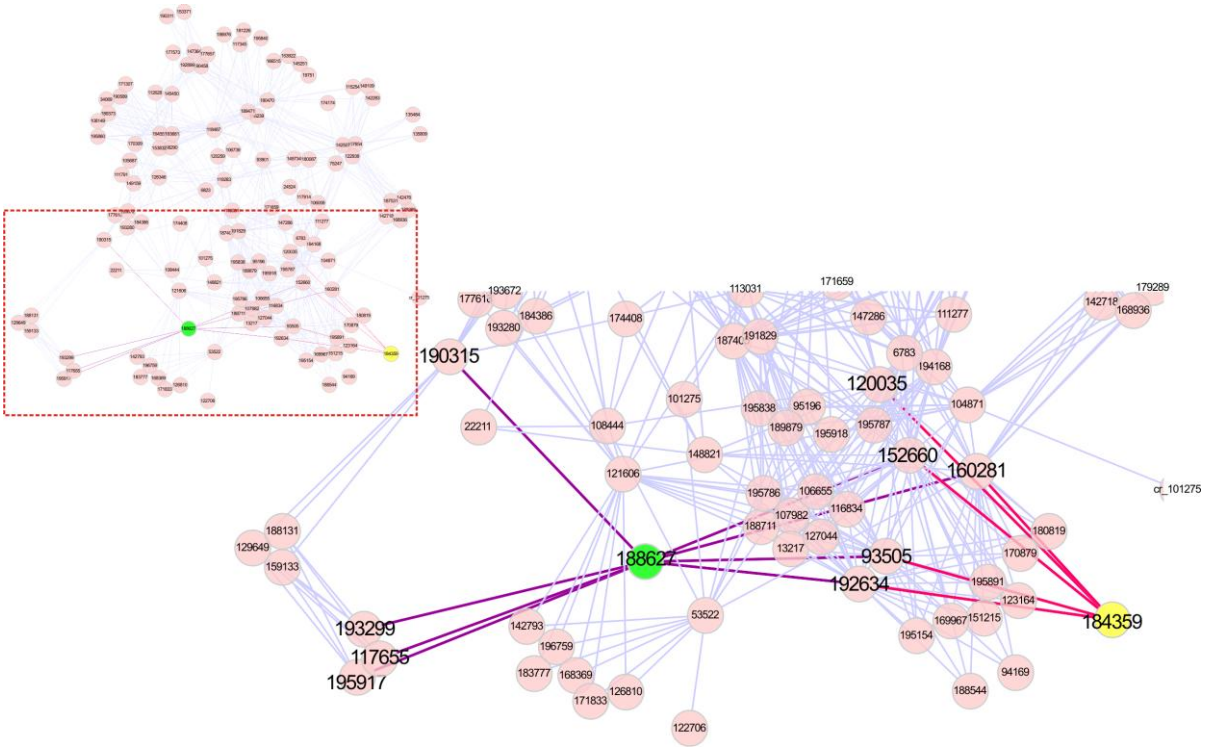


Figure 5-11. Graph representation of the reconstructed gene regulatory network of Chlamydomonas grown in LC conditions.

A gene regulatory network of Chlamydomonas was reconstructed using the IOTA method and a graphical representation of the topology of the whole network is shown in the upper left part of the figure. The area indicated by a red rectangle is zoomed out in the right part of the figure; it shows the interaction partners of the two MYB-related TFs that are highly expressed in the LC condition. One of these genes is the Low-CO_2 stress responsive regulator (LCR1, protein ID 184359; yellow node) and when compared to the other member of MYB-related family (protein ID 188627; green node) shows common interaction partners.

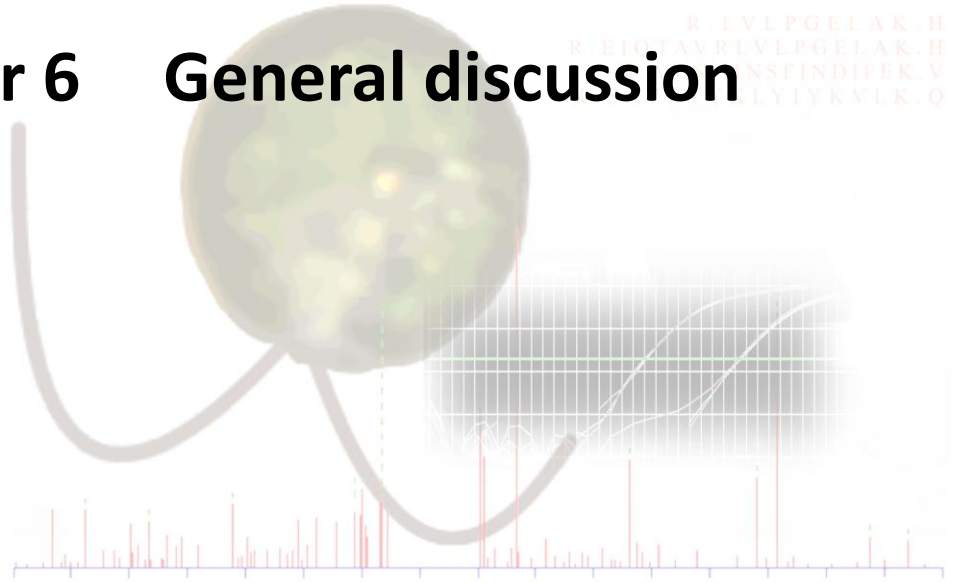
Moreover, each TF or TR from the reconstructed GRN regulates on average 4 other genes. A matrix containing all the interaction pairs obtained with IOTA method together with their probability values is available through the datasets section of our web site (http://plntfdb.bio.uni-potsdam.de/ChlamyTRI/). This information can be very useful for identifying common pathways induced or repressed under LC. It also helps in the identification of new candidate genes related to the transcriptional regulation of CCM and allow further studies focused on exploring sub-networks.

5.5 Conclusions

The cellular response of Chlamydomonas to the reduced external availability of CO_2 involves the transcriptional regulation of many genes and some of them playing a crucial role in the CCM. Until present, many different studies have been performed in order to understand how cells sense changes in the environmental conditions and how these changes lead to a new transcriptional state where a different set of genes are induced or repressed. Thus, it has been previously demonstrated that during CO_2 deprivation, Chlamydomonas triggers the expression of genes related to mechanisms related to carbon up-take, photosynthesis and photorespiration. However, there was no deep study about the changes in expression levels of the predicted transcriptional regulators or regulatory elements involved in the CCM. In our present study, we established a qRT-PCR platform for the transcript profiling of 245 TF and TR genes. Moreover, the expression profiles of 130 genes were determined in a time series experiment where the concentration of CO_2 was shifted from 5% to 0.04%. Early responsive up-regulated TF and TR genes seem to comprise a group of genes with functions related to chromatin remodeling, nucleosome displacing and alteration of RNA stability. We have found that late responsive genes include up-regulated genes which possibly execute more specific functions related to the induction of genes directly involved in the carbon metabolism, such as *Lcr1*, which was found to be present in this group. Another MYB-related TF coding gene was found to be up-regulated in the group of late responsive genes, and its promoter region contains conserved sequence motifs also found in the promoter region of *Lcr1*, suggesting that both TFs could contain similar promoter *cis*-elements and be

recognized by the same DNA-binding protein and be co-regulated. Additionally, we reconstructed theoretically the underlying regulatory network where both genes (*Lcr1* and MYB-related with protein IDs: 184359 and 188627, respectively) were predicted to regulate a very similar set of genes whose expression levels were decreased under LC. Furthermore, the predicted gene regulatory network revealed the putative interacting partners of more than one hundred genes and will serve as a basis for selecting further candidate genes and related sub-networks. We also performed a genome-wide analysis of the nucleosome occupancy of cells grown for three hours under LC using a FAIRE-seq approach. Our sequencing data revealed the existence of correlation between high gene expression fold change and presence of FAIRE peaks. It is then more likely to find a FAIRE peak close to a gene with gene expression fold change higher than 2. The TF and TR genes which presented a gene expression fold change higher than 2 and simultaneously presented a FAIRE peak, at three hours after the shift from HC to LC, also presented conserved motifs on their promoter regions. It is then suggested that the combination of expression profiling data, gene regulatory network reconstruction, and genome-wide mapping of chromatin nucleosome-depleted regions accelerates the discovery of potential transcriptional regulators, revealing new possible interactions and new target genes. Our findings revealed a number of new candidate genes related to carbon deprivation responses, opening the venue for complementary functional analysis of transcriptional regulators of CCM and regulatory elements of *C. reinhardtii.*

Chapter 6 General discussion



In the present work, different approaches were used to investigate gene regulatory networks in *Chlamydomonas reinhardtii*. The methods and platforms developed during the present work permitted to study TFs and TRs which are usually present in very low levels in the cells. The detection of the TFs and TRs either at the protein level or at the transcript levels demands high sensitivity and accurate methods. The optimization of the isolation of nuclei from Chlamydomonas permitted us to identify TFs and TRs and serve as a base for future comparative nuclear proteome analysis. An overview of the nuclear proteome of Chlamydomonas was provided and the analysis based on gene ontology terms revealed that components involved in gene expression constitute the major functional class of nuclear proteins discovered in our proteomics approach. Furthermore, a large fraction of nuclear proteins identified contained motifs classified as “binding” as a major ontology term. Several of these motifs are constituted by tandem amino acid repeats, such as polyQ or polyE. The function of tandem amino acid repeats in the protein activity is not completely understood. Previous studies in yeast and humans have suggested that proteins containing these regions may be important to transcriptional control (Simon and Hancock, 2009). In yeast, peptides containing acidic and polar amino acid repeats were significantly found within transcription factors and protein kinases (Simon and Hancock, 2009). There are evidences that modifications of the repeat length, usually the increase in the number of repeats, may lead to protein malfunctioning. In human disorders, such as Huntington's disease, the expansion of the length of the polyglutamine repeat has been shown to be related to many abnormalities, including transcriptional dysregulation (Chen, 2011). In plants, proteins of the subfamily LRR-kinases, contain leucine amino acid repeats and have shown to be involved in processes related to development and defense responses (Sun and Wang, 2011). Our proteomics data revealed that only few motifs found in the nuclear proteins identified are constituted of tandem amino acid repeats, and simultaneously enriched in the nuclear proteome of Chlamydomonas. Thus, the poly amino acid repeats of these particular proteins may be important for their functional activity in the nucleus. The function of the motifs found in the nuclear proteins remains unknown but our results suggest that these proteins are candidates for further investigation of transcriptional regulation in Chlamydomonas.

Some of the nuclear proteins with unknown functional annotation in Chlamydomonas have orthologues genes in flowering plants, such as *Arabidopsis thaliana* and Medicago. The functional characterization of these proteins would help to identify their role in nuclear processes and their functional conservation over large evolutionary time spans. The analysis of the nuclear proteome contributed to the functional annotation of Chlamydomonas proteins and confirmed that analyzing the subcellular proteome is an interesting way not only to detect the general composition of an organelle, but also to unravel some of the characteristics of their protein sequences.

The identification of the nuclear proteome was only the first step towards a deeper understanding of the complexity of the environment where transcriptional regulation occurs and revealed interesting candidate proteins specific for that compartment.

In Chlamydomonas, only few transcriptional regulators have been functionally characterized, making it difficult to get comprehensive information about gene regulatory networks. Recently, with the sequencing of the Chlamydomonas genome it became possible to explore its genomic information for the identification of candidate genes. From that point, the prediction of TFs and TRs was possible and revealed the identity of more than two hundred candidate regulatory genes (Riaño-Pachón et al., 2007).

In the present work, a set of 245 genes previously predicted to code for TFs and TRs was used as base for the analysis of time series gene expression data generated through transcript profiling using quantitative reverse-transcription PCR. This platform was used to investigate the regulation of the TF and TR coding genes under conditions of varying light intensities and carbon deprivation.

Much is known about photosynthesis and the main enzymes and proteins involved in the direct metabolic processes of using light energy for the synthesis of organic compounds. However, in Chlamydomonas, most of the transcriptional light-induced regulatory genes remain to be investigated. To our knowledge, no study focusing on the analysis of the expression patterns of TF and TR genes in Chlamydomonas had been reported until now.

By investigating the expression profiles of the TFs and TRs under different environmental conditions, such as different light intensities and different concentrations of CO_2, genes responsive to such conditions could be identified. Our transcript profiling of TF and TR genes

in short time series experiments under transition from low light (200µE m^{-2} s^{-1}) to high light (700µE m^{-2} s^{-1}) revealed the expression profile of 127 genes. Interesting results were found and showed group of light-responsive genes which are regulated within one hour after the light shift. This group contains genes which also have been shown to have a role in other cellular responses different from the light-induced ones. Genes such as *Lcr1* which is induced during induction of the carbon concentrating mechanism (CCM) and *Hsf1* which is induced during heat stress were also found to be induced by the light shift.

Interesting observations also came from the analysis of the gene expression of TFs and TRs under carbon deprivation conditions. Genes previously shown to be involved in CCM were identified as expected. However, other genes related to stress responses, such as *Hsf1*, were found to be induced by LC conditions. One new candidate gene for the regulation of the CCM-related genes is a TF of the MYB-related family (protein ID 188627) which showed an expression profile similar to that of *Lcr1*, which also codes for a MYB-related TF. As revealed previously, LCR1 is a regulator of CCM-related genes and is necessary for maintaining the expression of *Cah1* which encodes carbonic anhydrase, an enzyme highly expressed under LC conditions (Kucho et al., 1999). In the present study, a conserved motif was found in the promoter regions of both MYB-related TF genes. Our data thus indicates that both genes are co-regulated by one same protein (TF) capable to recognize this regulatory motif and to induce the expression of both genes. The reconstructed gene regulatory network for CCM showed that both MYB-related genes regulate almost the same set of target genes. However, further experiments are needed to investigate if the conserved motifs found are indeed DNA regulatory *cis*-elements and the target genes have also to be experimentally confirmed.

Furthermore, our findings suggested it seems possible that a common set of transcriptional regulators of Chlamydomonas gene regulatory networks participate in different signaling pathways as integrative regulators of different responses and function in a coordinated way to regulate photosynthesis-related genes, according to the availability of environmental light and carbon dioxide. This is not completely unexpected, and evidences from signaling crosstalk in flowering plants subjected to different environmental conditions was already found. There are evidences of the existence of common regulators involved in the signaling

pathways induced by alterations of heat and light conditions (Franklin, 2009). In Chlamydomonas, the integration of different signaling pathways is not very well understood. Nevertheless, alterations of photosynthesis and photorespiration status during the acclimation to carbon deprivation were previously observed (Vance and Spalding, 2005; Moroney and Ynalvez, 2007). We found many genes which had similar expression profiles in the experiments of carbon deprivation and light shift, and three of these genes were up-regulated (FC > 2) in both experimental conditions. These candidate genes are respectively members of the HSF (protein ID 117914), bZIP (protein ID 147286) and RWP-RK (protein ID 195917) TF families. These genes can be important candidates for the study of what I call here “light-carbon” responses, and their functional characterization would help in understanding the essential cellular mechanisms of the integrative regulatory responses.

The identification of the genomic nucleosome-depleted regions by FAIRE-seq revealed a large number of genes which could be transcriptionally regulated in Chlamydomonas and their possible regulatory elements can be identified. These data can assist in obtaining a global view of the genes that are regulated in a certain condition and this can be used as complementary information to improve the description of gene regulatory networks.

The present work indicates that regulation of cellular acclimation at the transcriptional level is multifactorial and can incorporate elements of different, but interdependent cellular responses. How the specific signaling inputs are perceived by the same set of genes but translated into regulation of specific downstream regulatory pathways needs to be elucidated in detail.

Outlook

Many new questions appeared during the development of the present work and additional work has already been initiated to address some of them. With the methods and technical platforms developed for this study more experimental (culture) conditions can be tested and comparative analyses performed at either the protein or transcript levels. It is expected that other researches will also benefit from the information reported here, e.g. when analyzing gene regulatory networks in Chlamydomonas. Functional assays using nuclear extracts, including the analysis of DNA-protein interactions have already been performed by other groups and can now be performed with the method described here for the isolation of nuclei. Studies aiming at testing the subcellular localization of functionally non-annotated proteins identified in the Chlamydomonas nuclear proteome are currently being performed. In addition, target genes of the most interesting candidate regulatory TFs can now be identified by generating knockdown cell lines and subsequent analysis of gene expression patterns using microarray hybridizations or deep sequencing technologies. Further analysis of the FAIRE-seq data will likely provide us even larger number of new candidate genes and regulatory elements which should be tested experimentally in order to confirm their regulatory role. Our website ChlamyTRI (**Chlamy**domonas **T**ranscriptional **R**egulation **I**nitiative; URL: http://plntfdb.bio.uni-potsdam.de/ChlamyTRI/) can be used for sharing information obtained in this work and for motivating future discussions on the topic of transcriptional regulation in Chlamydomonas.

References

Akai, Y., Adachi, N., Hayashi, Y., Eitoku, M., Sano, N. et al. (2010) Structure of the histone chaperone CIA/ASF1-double bromodomain complex linking histone modifications and site-specific histone eviction, *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 8153-8158.

Alberguina, L. and Westerhoff, H.V. (2005) Systems Biology: Definitions and Perspectives In Alberguina, L. and Westerhoff, H.V. (eds), *Topics in Current Genetics*. Springer-Verlag Gmbh, Heidelberg, pp. 408.

Amunts, A., Ben-Shem, A. and Nelson, N. (2005) Solving the structure of plant photosystem I--biochemistry is vital, *Photochem Photobiol Sci*, **4**, 1011-1015.

Arvidsson, S., Kwasniewski, M., Riano-Pachon, D.M. and Mueller-Roeber, B. (2008) QuantPrime--a flexible tool for reliable high-throughput primer design for quantitative PCR, *BMC bioinformatics*, **9**, 465.

Asamizu, E., Miura, K., Kucho, K., Inoue, Y., Fukuzawa, H. et al. (2000) Generation of expressed sequence tags from low-CO2 and high-CO2 adapted cells of Chlamydomonas reinhardtii, *DNA Res*, **7**, 305-307.

Badger, M.R., Kaplan, A. and Berry, J.A. (1980) Internal Inorganic Carbon Pool of Chlamydomonas reinhardtii: EVIDENCE FOR A CARBON DIOXIDE-CONCENTRATING MECHANISM, *Plant physiology*, **66**, 407-413.

Bae, M.S., Cho, E.J., Choi, E.Y. and Park, O.K. (2003) Analysis of the Arabidopsis nuclear proteome and its response to cold stress. *Plant J*. pp. 652-663.

Bailey, T.L. and Gribskov, M. (1998) Combining evidence using p-values: application to sequence homology searches, *Bioinformatics (Oxford, England)*, **14**, 48-54.

Bailey, T.L., Williams, N., Misleh, C. and Li, W.W. (2006) MEME: discovering and analyzing DNA and protein sequence motifs, *Nucleic acids research*, **34**, W369-373.

Banerjee, R. and Batschauer, A. (2005) Plant blue-light receptors, *Planta*, **220**, 498-502.

Barsnes, H., Vizcaino, J.A., Eidhammer, I. and Martens, L. (2009) PRIDE Converter: making proteomics data-sharing easy, *Nat Biotechnol*, **27**, 598-599.

Beckmann, J., Lehr, F., Finazzi, G., Hankamer, B., Posten, C. et al. (2009) Improvement of light to biomass conversion by de-regulation of light-harvesting protein translation in Chlamydomonas reinhardtii, *J Biotechnol*, **142**, 70-77.

Boa, S., Coert, C. and Patterton, H.G. (2003) Saccharomyces cerevisiae Set1p is a methyltransferase specific for lysine 4 of histone H3 and is required for efficient gene expression, *Yeast (Chichester, England)*, **20**, 827-835.

Bohne, A.V., Irihimovitch, V., Weihe, A. and Stern, D.B. (2006) Chlamydomonas reinhardtii encodes a single sigma70-like factor which likely functions in chloroplast transcription, *Curr Genet*, **49**, 333-340.

Borello, U., Ceccarelli, E. and Giuliano, G. (1993) Constitutive, light-responsive and circadian clock-responsive factors compete for the different l box elements in plant light-regulated promoters, *Plant J*, **4**, 611-619.

Brameier, M., Krings, A. and MacCallum, R.M. (2007) NucPred--predicting nuclear localization of proteins, *Bioinformatics (Oxford, England)*, **23**, 1159-1160.

Brand, A.H., Breeden, L., Abraham, J., Sternglanz, R. and Nasmyth, K. (1985) Characterization of a "silencer" in yeast: a DNA sequence with properties opposite to those of a transcriptional enhancer, *Cell*, **41**, 41-48.

Britten, R.J. and Davidson, E.H. (1969) Gene regulation for higher cells: a theory, *Science*, **165**, 349-357.

Brivanlou, A.H. and Darnell, J.E., Jr. (2002) Signal transduction and the control of gene expression, *Science*, **295**, 813-818.

Brown, D.D. (1981) Gene expression in eukaryotes, *Science*, **211**, 667-674.

Caldana, C., Scheible, W.R., Mueller-Roeber, B. and Ruzicic, S. (2007) A quantitative RT-PCR platform for high-throughput expression profiling of 2500 rice transcription factors, *Plant methods*, **3**, 7.

Camargo, A., Llamas, A., Schnell, R.A., Higuera, J.J., Gonzalez-Ballester, D. et al. (2007) Nitrate signaling by the regulatory gene NIT2 in Chlamydomonas, *The Plant cell*, **19**, 3491-3503.

Casadio, R., Martelli, P.L. and Pierleoni, A. (2008) The prediction of protein subcellular localization from sequence: a shortcut to functional genome annotation, *Briefings in functional genomics & proteomics*, **7**, 63-73.

Casal, J.J. and Yanovsky, M.J. (2005) Regulation of gene expression by light, *The International journal of developmental biology*, **49**, 501-511.

Cashmore, A.R., Jarillo, J.A., Wu, Y.J. and Liu, D. (1999) Cryptochromes: blue light receptors for plants and animals, *Science*, **284**, 760-765.

Chaikam, V. and Karlson, D.T. (2010) Comparison of structure, function and regulation of plant cold shock domain proteins to bacterial and animal cold shock domain proteins, *BMB reports*, **43**, 1-8.

Chattopadhyay, S., Ang, L.H., Puente, P., Deng, X.W. and Wei, N. (1998) Arabidopsis bZIP protein HY5 directly interacts with light-responsive promoters in mediating light control of gene expression, *The Plant cell*, **10**, 673-683.

Chen, C.M. (2011) Mitochondrial Dysfunction, Metabolic Deficits, and Increased Oxidative Stress in Huntington's Disease, *Chang Gung Med J*, **34**, 135-152.

Chen, J., Ghazawi, F.M. and Li, Q. (2010) Interplay of bromodomain and histone acetylation in the regulation of p300-dependent genes, *Epigenetics*, **5**, 509-515.

Christian, N., May, P., Kempa, S., Handorf, T. and Ebenhoh, O. (2009) An integrative approach towards completing genome-scale metabolic networks, *Molecular bioSystems*, **5**, 1889-1903.

Cline, M.S., Smoot, M., Cerami, E., Kuchinsky, A., Landys, N. et al. (2007) Integration of biological networks and gene expression data using Cytoscape, *Nature protocols*, **2**, 2366-2382.

Cokol, M., Nair, R. and Rost, B. (2000) Finding nuclear localization signals, *EMBO Rep*, **1**, 411-415.

Colon-Ramos, D.A., Salisbury, J.L., Sanders, M.A., Shenoy, S.M., Singer, R.H. et al. (2003) Asymmetric distribution of nuclear pore complexes and the cytoplasmic localization of beta2-tubulin mRNA in Chlamydomonas reinhardtii, *Dev Cell*, **4**, 941-952.

Conley, T.R., Park, S.C., Kwon, H.B., Peng, H.P. and Shih, M.C. (1994) Characterization of cis-acting elements in light regulation of the nuclear gene encoding the A subunit of chloroplast isozymes of glyceraldehyde-3-phosphate dehydrogenase from Arabidopsis thaliana, *Mol Cell Biol*, **14**, 2525-2533.

Correa, L.G., Riano-Pachon, D.M., Schrago, C.G., dos Santos, R.V., Mueller-Roeber, B. et al. (2008) The role of bZIP transcription factors in green plant evolution: adaptive features emerging from four founder genes, *PloS one*, **3**, e2944.

Cox, B. and Emili, A. (2006) Tissue subcellular fractionation and protein extraction for use in mass-spectrometry-based proteomics, *Nature protocols*, **1**, 1872-1878.

Davletova, S., Rizhsky, L., Liang, H., Shengqiang, Z., Oliver, D.J. et al. (2005) Cytosolic ascorbate peroxidase 1 is a central component of the reactive oxygen gene network of Arabidopsis, *The Plant cell*, **17**, 268-281.

Desikan, R., S, A.H.-M., Hancock, J.T. and Neill, S.J. (2001) Regulation of the Arabidopsis transcriptome by oxidative stress, *Plant physiology*, **127**, 159-172.

Dignam, J.D., Lebovitz, R.M. and Roeder, R.G. (1983) Accurate transcription initiation by RNA polymerase II in a soluble extract from isolated mammalian nuclei, *Nucleic acids research*, **11**, 1475-1489.

Dionisio-Sese, M.L., Fukuzawa, H. and Miyachi, S. (1990) Light-Induced Carbonic Anhydrase Expression in Chlamydomonas reinhardtii, *Plant physiology*, **94**, 1103-1110.

Doebley, J. and Lukens, L. (1998) Transcriptional regulators and the evolution of plant form, *The Plant cell*, **10**, 1075-1082.

Donner, S., Koseska, A., Kurths, J. and Nikoloski, Z. (2011) Inner composition alignment for inferring directed networks from short time series, *Phys. Rev. Lett.*, **under revision**.

Donnes, P. and Hoglund, A. (2004) Predicting protein subcellular localization: past, present, and future, *Genomics Proteomics Bioinformatics*, **2**, 209-215.

Dyer, R.B. and Herzog, N.K. (1995) Isolation of intact nuclei for nuclear extract preparation from a fragile B-lymphocyte cell line, *Biotechniques*, **19**, 192-195.

Erkine, A.M. and Gross, D.S. (2003) Dynamic chromatin alterations triggered by natural and synthetic activation domains, *The Journal of biological chemistry*, **278**, 7755-7764.

Fang, S.C., de los Reyes, C. and Umen, J.G. (2006) Cell size checkpoint control by the retinoblastoma tumor suppressor pathway, *PLoS genetics*, **2**, e167.

Ferris, P.J. and Goodenough, U.W. (1997) Mating type in Chlamydomonas is specified by mid, the minus-dominance gene, *Genetics*, **146**, 859-869.

Fink, J.L., Karunaratne, S., Mittal, A., Gardiner, D.M., Hamilton, N. et al. (2008) Towards defining the nuclear proteome, *Genome biology*, **9**, R15.

Franklin, K.A. (2009) Light and temperature signal crosstalk in plant development, *Current opinion in plant biology*, **12**, 63-68.

Fukuzawa, H., Miura, K., Ishizaki, K., Kucho, K.I., Saito, T. et al. (2001) Ccm1, a regulatory gene controlling the induction of a carbon-concentrating mechanism in Chlamydomonas reinhardtii by sensing CO2 availability, *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 5347-5352.

Fyshe, A., Liu, Y., Szafron, D., Greiner, R. and Lu, P. (2008) Improving subcellular localization prediction using text classification and the gene ontology, *Bioinformatics (Oxford, England)*, **24**, 2512-2517.

Gencheva, M., Boa, S., Fraser, R., Simmen, M.W., CB, A.W. et al. (2006) In Vitro and in Vivo nucleosome positioning on the ovine beta-lactoglobulin gene are related, *Journal of molecular biology*, **361**, 216-230.

Gerner, C. and Sauermann, G. (1999) Nuclear matrix proteins specific for subtypes of human hematopoietic cells, *Journal of cellular biochemistry*, **72**, 470-482.

Giresi, P.G., Kim, J., McDaniell, R.M., Iyer, V.R. and Lieb, J.D. (2007) FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin, *Genome research*, **17**, 877-885.

Giresi, P.G. and Lieb, J.D. (2009) Isolation of active regulatory elements from eukaryotic chromatin using FAIRE (Formaldehyde Assisted Isolation of Regulatory Elements), *Methods (San Diego, Calif*, **48**, 233-239.

Gitai, Z. (2005) The new bacterial cell biology: moving parts and subcellular architecture, *Cell*, **120**, 577-586.

Gitter, A., Lu, Y. and Bar-Joseph, Z. (2010) Computational methods for analyzing dynamic regulatory networks, *Methods Mol Biol*, **674**, 419-441.

Gonzalez-Ballester, D., Casero, D., Cokus, S., Pellegrini, M., Merchant, S.S. et al. (2010) RNA-seq analysis of sulfur-deprived Chlamydomonas cells reveals aspects of acclimation critical for cell survival, *The Plant cell*, **22**, 2058-2084.

Goodenough, U., Lin, H. and Lee, J.H. (2007) Sex determination in Chlamydomonas, *Semin Cell Dev Biol*, **18**, 350-361.

Goodrich, J.A. and Tjian, R. (2010) Unexpected roles for core promoter recognition factors in cell-type-specific transcription and gene regulation, *Nature reviews*, **11**, 549-558.

Gorman, D.S. and Levine, R.P. (1965) Cytochrome f and plastocyanin: their sequence in the photosynthetic electron transport chain of Chlamydomonas reinhardi, *Proceedings of the National Academy of Sciences of the United States of America*, **54**, 1665-1669.

Gorman, D.S. and Levine, R.P. (1966) Photosynthetic Electron Transport Chain of Chlamydomonas reinhardi. IV. Purification and Properties of Plastocyanin, *Plant physiology*, **41**, 1637-1642.

Graumann, P.L. and Marahiel, M.A. (1998) A superfamily of proteins that contain the cold-shock domain, *Trends in biochemical sciences*, **23**, 286-290.

Grimbs, S., Arnold, A., Koseska, A., Kurths, J., Selbig, J. et al. (2011) Spatiotemporal dynamics of the Calvin cycle: multistationarity and symmetry breaking instabilities, *Bio Systems*, **103**, 212-223.

Grossman, A.R., Croft, M., Gladyshev, V.N., Merchant, S.S., Posewitz, M.C. et al. (2007) Novel metabolism in Chlamydomonas through the lens of genomics, *Current opinion in plant biology*, **10**, 190-198.

Grossman, A.R., Karpowicz, S.J., Heinnickel, M., Dewez, D., Hamel, B. et al. (2010) Phylogenomic analysis of the Chlamydomonas genome unmasks proteins potentially involved in photosynthetic function and regulation, *Photosynthesis research*, **106**, 3-17.

Guenin, S., Mauriat, M., Pelloux, J., Van Wuytswinkel, O., Bellini, C. et al. (2009) Normalization of qRT-PCR data: the necessity of adopting a systematic, experimental conditions-specific, validation of references, *J Exp Bot*, **60**, 487-493.

Guidez, F., Li, A.C., Horvai, A., Welch, J.S. and Glass, C.K. (1998) Differential utilization of Ras signaling pathways by macrophage colony-stimulating factor (CSF) and granulocyte-macrophage CSF receptors during macrophage differentiation, *Mol Cell Biol*, **18**, 3851-3861.

Gygi, S.P., Corthals, G.L., Zhang, Y., Rochon, Y. and Aebersold, R. (2000) Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology, *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 9390-9395.

Gygi, S.P., Rist, B., Griffin, T.J., Eng, J. and Aebersold, R. (2002) Proteome analysis of low-abundance proteins using multidimensional chromatography and isotope-coded affinity tags, *Journal of proteome research*, **1**, 47-54.

Gygi, S.P., Rochon, Y., Franza, B.R. and Aebersold, R. (1999) Correlation between protein and mRNA abundance in yeast, *Mol Cell Biol*, **19**, 1720-1730.

Gyula, P., Schafer, E. and Nagy, F. (2003) Light perception and signalling in higher plants, *Current opinion in plant biology*, **6**, 446-452.

Hagstrom, J.E., Ludtke, J.J., Bassik, M.C., Sebestyen, M.G., Adam, S.A. et al. (1997) Nuclear import of DNA in digitonin-permeabilized cells, *J Cell Sci*, **110 (Pt 18)**, 2323-2331.

Hardtke, C.S., Gohda, K., Osterlund, M.T., Oyama, T., Okada, K. et al. (2000) HY5 stability and activity in arabidopsis is regulated by phosphorylation in its COP1 binding domain, *EMBO J*, **19**, 4997-5006.

Harris, E.H. (2001) Chlamydomonas as a Model Organism, *Annual review of plant physiology and plant molecular biology*, **52**, 363-406.

Heazlewood, J.L., Tonti-Filippini, J.S., Gout, A.M., Day, D.A., Whelan, J. et al. (2004) Experimental analysis of the Arabidopsis mitochondrial proteome highlights signaling and regulatory components, provides assessment of targeting prediction programs, and indicates plant-specific mitochondrial proteins, *The Plant cell*, **16**, 241-256.

Heins, L., Mentzel, H., Schmid, A., Benz, R. and Schmitz, U.K. (1994) Biochemical, molecular, and functional characterization of porin isoforms from potato mitochondria, *The Journal of biological chemistry*, **269**, 26402-26410.

Heyer, L.J., Kruglyak, S. and Yooseph, S. (1999) Exploring expression data: identification and analysis of coexpressed genes, *Genome research*, **9**, 1106-1115.

Hihara, Y., Kamei, A., Kanehisa, M., Kaplan, A. and Ikeuchi, M. (2001) DNA microarray analysis of cyanobacterial gene expression during acclimation to high light, *The Plant cell*, **13**, 793-806.

Hogan, G.J., Lee, C.K. and Lieb, J.D. (2006) Cell cycle-specified fluctuation of nucleosome occupancy at gene promoters, *PLoS genetics*, **2**, e158.

Horton, P., Park, K.J., Obayashi, T., Fujita, N., Harada, H. et al. (2007) WoLF PSORT: protein localization predictor, *Nucleic acids research*, **35**, W585-587.

Huang, K. and Beck, C.F. (2003) Phototropin is the blue-light receptor that controls multiple steps in the sexual life cycle of the green alga Chlamydomonas reinhardtii, *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 6269-6274.

Im, C.S., Eberhard, S., Huang, K., Beck, C.F. and Grossman, A.R. (2006) Phototropin involvement in the expression of genes encoding chlorophyll and carotenoid biosynthesis enzymes and LHC apoproteins in Chlamydomonas reinhardtii, *Plant J*, **48**, 1-16.

Im, C.S., Zhang, Z., Shrager, J., Chang, C.W. and Grossman, A.R. (2003) Analysis of light and CO(2) regulation in Chlamydomonas reinhardtii using genome-wide approaches, *Photosynthesis research*, **75**, 111-125.

Immeln, D., Schlesinger, R., Heberle, J. and Kottke, T. (2007) Blue light induces radical formation and autophosphorylation in the light-sensitive domain of Chlamydomonas cryptochrome, *The Journal of biological chemistry*, **282**, 21720-21728.

Jacob, F. and Monod, J. (1961) On the Regulation of Gene Activity, *Cold Spring Harbor Symp Quant Biol*, **26**, 193-211.

Jeong, M.J. and Shih, M.C. (2003) Interaction of a GATA factor with cis-acting elements involved in light regulation of nuclear genes encoding chloroplast glyceraldehyde-3-phosphate dehydrogenase in Arabidopsis, *Biochem Biophys Res Commun*, **300**, 555-562.

Jiang, W., Hou, Y. and Inouye, M. (1997) CspA, the major cold-shock protein of Escherichia coli, is an RNA chaperone, *The Journal of biological chemistry*, **272**, 196-202.

Jiao, Y., Lau, O.S. and Deng, X.W. (2007) Light-regulated transcriptional networks in higher plants, *Nature reviews*, **8**, 217-230.

Jiao, Y. and Meyerowitz, E.M. (2010) Cell-type specific analysis of translating RNAs in developing flowers reveals new levels of control, *Molecular systems biology*, **6**, 419.

Jiao, Y., Yang, H., Ma, L., Sun, N., Yu, H. et al. (2003) A genome-wide analysis of blue-light regulation of Arabidopsis transcription factor gene expression during seedling development, *Plant physiology*, **133**, 1480-1493.

Jorrin-Novo, J.V., Maldonado, A.M., Echevarria-Zomeno, S., Valledor, L., Castillejo, M.A. et al. (2009) Plant proteomics update (2007-2008): Second-generation proteomic techniques, an appropriate experimental design, and data analysis to fulfill MIAPE standards, increase plant proteome coverage and expand biological knowledge, *Journal of proteomics*, **72**, 285-314.

Jorrin, J.V., Maldonado, A.M. and Castillejo, M.A. (2007) Plant proteome analysis: a 2006 update, *Proteomics*, **7**, 2947-2962.

Jung, E., Heller, M., Sanchez, J.C. and Hochstrasser, D.F. (2000) Proteomics meets cell biology: the establishment of subcellular proteomes, *Electrophoresis*, **21**, 3369-3377.

Kaplan, A. and Reinhold, L. (1999) Co2 Concentrating Mechanisms in Photosynthetic Microorganisms, *Annual review of plant physiology and plant molecular biology*, **50**, 539-570.

Karlson, D. and Imai, R. (2003) Conservation of the cold shock domain protein family in plants, *Plant physiology*, **131**, 12-15.

Kaufmann, K., Pajoro, A. and Angenent, G.C. (2010) Regulation of transcription in plants: mechanisms controlling developmental switches, *Nature reviews*, **11**, 830-842.

Keller, L.R., Schloss, J.A., Silflow, C.D. and Rosenbaum, J.L. (1984) Transcription of alpha- and beta-tubulin genes in vitro in isolated Chlamydomonas reinhardi nuclei, *J Cell Biol*, **98**, 1138-1143.

Kilian, J., Whitehead, D., Horak, J., Wanke, D., Weinl, S. et al. (2007) The AtGenExpress global stress expression data set: protocols, evaluation and model data analysis of UV-B light, drought and cold stress responses, *Plant J*, **50**, 347-363.

Kim, J.S., Park, S.J., Kwak, K.J., Kim, Y.O., Kim, J.Y. et al. (2007) Cold shock domain proteins and glycine-rich RNA-binding proteins from Arabidopsis thaliana can promote the cold adaptation process in Escherichia coli, *Nucleic acids research*, **35**, 506-516.

Klutstein, M., Shaked, H., Sherman, A., Avivi-Ragolsky, N., Shema, E. et al. (2008) Functional conservation of the yeast and Arabidopsis RAD54-like genes, *Genetics*, **178**, 2389-2397.

Komili, S. and Silver, P.A. (2008) Coupling and coordination in gene expression processes: a systems biology view, *Nature reviews*, **9**, 38-48.

Kosourov, S., Seibert, M. and Ghirardi, M.L. (2003) Effects of extracellular pH on the metabolic pathways in sulfur-deprived, H2-producing Chlamydomonas reinhardtii cultures, *Plant Cell Physiol*, **44**, 146-155.

Kotak, S., Port, M., Ganguli, A., Bicker, F. and von Koskull-Doring, P. (2004) Characterization of C-terminal domains of Arabidopsis heat stress transcription factors (Hsfs) and identification of a new signature combination of plant class A Hsfs with AHA and NES motifs essential for activator function and intracellular localization, *Plant J*, **39**, 98-112.

Kropat, J., Tottey, S., Birkenbihl, R.P., Depege, N., Huijser, P. et al. (2005) A regulator of nutritional copper signaling in Chlamydomonas is an SBP domain protein that recognizes the GTAC core of copper response element, *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 18730-18735.

Kucho, K., Ohyama, K. and Fukuzawa, H. (1999) CO(2)-responsive transcriptional regulation of CAH1 encoding carbonic anhydrase is mediated by enhancer and silencer regions in Chlamydomonas reinhardtii, *Plant physiology*, **121**, 1329-1338.

Kucho, K., Yoshioka, S., Taniguchi, F., Ohyama, K. and Fukuzawa, H. (2003) Cis-acting elements and DNA-binding proteins involved in CO2-responsive transcriptional activation of Cah1 encoding a periplasmic carbonic anhydrase in Chlamydomonas reinhardtii, *Plant physiology*, **133**, 783-793.

Laemmli, U.K. (1970) Cleavage of structural proteins during the assembly of the head of bacteriophage T4, *Nature*, **227**, 680-685.

Lang, D., Weiche, B., Timmerhaus, G., Richardt, S., Riano-Pachon, D.M. et al. (2010) Genome-wide phylogenetic comparative analysis of plant transcriptional regulation: a timeline of loss, gain, expansion, and correlation with complexity, *Genome Biol Evol*, **2**, 488-503.

Langmead, B., Trapnell, C., Pop, M. and Salzberg, S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome, *Genome biology*, **10**, R25.

Latchman, D.S. (1997) Transcription factors: an overview, *The international journal of biochemistry & cell biology*, **29**, 1305-1312.

Lee, F.W. and Lo, S.C. (2008) The use of Trizol reagent (phenol/guanidine isothiocyanate) for producing high quality two-dimensional gel electrophoretograms (2-DE) of dinoflagellates, *Journal of microbiological methods*, **73**, 26-32.

Levine, M. and Tjian, R. (2003) Transcription regulation and animal diversity, *Nature*, **424**, 147-151.

Lin, H. and Goodenough, U.W. (2007) Gametogenesis in the Chlamydomonas reinhardtii minus mating type is controlled by two genes, MID and MTD1, *Genetics*, **176**, 913-925.

Liscum, E., Hodgson, D.W. and Campbell, T.J. (2003) Blue light signaling through the cryptochromes and phototropins. So that's what the blues is all about, *Plant physiology*, **133**, 1429-1436.

Loureiro, J., Rodriguez, E., Dolezel, J. and Santos, C. (2007) Two new nuclear isolation buffers for plant DNA flow cytometry: a test with 37 species, *Annals of botany*, **100**, 875-888.

Louwers, M., Bader, R., Haring, M., van Driel, R., de Laat, W. et al. (2009) Tissue- and expression level-specific chromatin looping at maize b1 epialleles, *The Plant cell*, **21**, 832-842.

Lowry, J.A. and Atchley, W.R. (2000) Molecular evolution of the GATA family of transcription factors: conservation within the DNA-binding domain, *J Mol Evol*, **50**, 103-115.

Luthe, D.S. and Quatrano, R.S. (1980) Transcription in Isolated Wheat Nuclei: I. ISOLATION OF NUCLEI AND ELIMINATION OF ENDOGENOUS RIBONUCLEASE ACTIVITY, *Plant physiology*, **65**, 305-308.

Lynch, M. (2007) The evolution of genetic networks by non-adaptive processes, *Nature reviews*, **8**, 803-813.

Maere, S., Heymans, K. and Kuiper, M. (2005) BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks, *Bioinformatics (Oxford, England)*, **21**, 3448-3449.

Malik, S. and Roeder, R.G. (2010) The metazoan Mediator co-activator complex as an integrative hub for transcriptional regulation, *Nature reviews*, **11**, 761-772.

Mar, J.C., Kimura, Y., Schroder, K., Irvine, K.M., Hayashizaki, Y. et al. (2009) Data-driven normalization strategies for high-throughput quantitative RT-PCR, *BMC bioinformatics*, **10**, 110.

Martin, C. and Paz-Ares, J. (1997) MYB transcription factors in plants, *Trends Genet*, **13**, 67-73.

Martin, N.C. and Goodenough, U.W. (1975) Gametic differentiation in Chlamydomonas reinhardtii. I. Production of gametes and their fine structure, *J Cell Biol*, **67**, 587-605.

Martins de Souza, D., Oliveira, B.M., Castro-Dias, E., Winck, F.V., Horiuchi, R.S. et al. (2008) The untiring search for the most complete proteome representation: reviewing the methods, *Briefings in functional genomics & proteomics*, **7**, 312-321.

Marushige, K. (1976) Activation of chromatin by acetylation of histone side chains, *Proceedings of the National Academy of Sciences of the United States of America*, **73**, 3937-3941.

Maxwell, B.B., Andersson, C.R., Poole, D.S., Kay, S.A. and Chory, J. (2003) HY5, Circadian Clock-Associated 1, and a cis-element, DET1 dark response element, mediate DET1 regulation of chlorophyll a/b-binding protein 2 expression, *Plant physiology*, **133**, 1565-1577.

May, P., Wienkoop, S., Kempa, S., Usadel, B., Christian, N. et al. (2008) Metabolomics- and proteomics-assisted genome annotation and analysis of the draft metabolic network of Chlamydomonas reinhardtii, *Genetics*, **179**, 157-166.

McGinn, P.J., Price, G.D., Maleszka, R. and Badger, M.R. (2003) Inorganic carbon limitation and light control the expression of transcripts related to the CO2-concentrating mechanism in the cyanobacterium Synechocystis sp. strain PCC6803, *Plant physiology*, **132**, 218-229.

Merchant, S.S., Prochnik, S.E., Vallon, O., Harris, E.H., Karpowicz, S.J. et al. (2007) The Chlamydomonas genome reveals the evolution of key animal and plant functions, *Science*, **318**, 245-250.

Miller, G. and Mittler, R. (2006) Could heat shock transcription factors function as hydrogen peroxide sensors in plants?, *Annals of botany*, **98**, 279-288.

Miura, K., Yamano, T., Yoshioka, S., Kohinata, T., Inoue, Y. et al. (2004) Expression profiling-based identification of CO2-responsive genes regulated by CCM1 controlling a carbon-concentrating mechanism in Chlamydomonas reinhardtii, *Plant physiology*, **135**, 1595-1607.

Moroney, J.V. and Ynalvez, R.A. (2007) Proposed carbon dioxide concentrating mechanism in Chlamydomonas reinhardtii, *Eukaryotic cell*, **6**, 1251-1259.

Moseley, J.L., Gonzalez-Ballester, D., Pootakham, W., Bailey, S. and Grossman, A.R. (2009) Genetic interactions between regulators of Chlamydomonas phosphorus and sulfur deprivation responses, *Genetics*, **181**, 889-905.

Murray, K. (1965) The Basic Proteins of Cell Nuclei, *Annu Rev Biochem*, **34**, 209-246.

Nagy, P.L., Cleary, M.L., Brown, P.O. and Lieb, J.D. (2003) Genomewide demarcation of RNA polymerase II transcription units revealed by physical fractionation of chromatin, *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 6364-6369.

Nakshatri, H. and Bhat-Nakshatri, P. (1997) Differential effect of Nonidet P40 on DNA binding of transcription factors, *Anal Biochem*, **249**, 103-104.

Naya, H., Romero, H., Carels, N., Zavala, A. and Musto, H. (2001) Translational selection shapes codon usage in the GC-rich genome of Chlamydomonas reinhardtii, *FEBS Lett*, **501**, 127-130.

Neubauer, G., King, A., Rappsilber, J., Calvio, C., Watson, M. et al. (1998) Mass spectrometry and EST-database searching allows characterization of the multi-protein spliceosome complex, *Nature genetics*, **20**, 46-50.

Nilsson, L., Muller, R. and Nielsen, T.H. (2007) Increased expression of the MYB-related transcription factor, PHR1, leads to enhanced phosphate uptake in Arabidopsis thaliana, *Plant, cell & environment*, **30**, 1499-1512.

Nolis, I.K., McKay, D.J., Mantouvalou, E., Lomvardas, S., Merika, M. et al. (2009) Transcription factors mediate long-range enhancer-promoter interactions, *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 20222-20227.

O'Malley, B.W., Towle, H.C. and Schwartz, R.J. (1977) Regulation of gene expression in eucaryotes, *Annual review of genetics*, **11**, 239-275.

Onodera, Y., Haag, J.R., Ream, T., Nunes, P.C., Pontes, O. et al. (2005) Plant nuclear RNA polymerase IV mediates siRNA and DNA methylation-dependent heterochromatin formation, *Cell*, **120**, 613-622.

Park, O.K. (2004) Proteomic studies in plants, *Journal of biochemistry and molecular biology*, **37**, 133-138.

Peers, G., Truong, T.B., Ostendorf, E., Busch, A., Elrad, D. et al. (2009) An ancient light-harvesting protein is critical for the regulation of algal photosynthesis, *Nature*, **462**, 518-521.

Perelman, S., Mazzella, M.A., Muschietti, J., Zhu, T. and Casal, J.J. (2003) Finding unexpected patterns in microarray data, *Plant physiology*, **133**, 1717-1725.

Perez-Rodriguez, P., Riaño-Pachón, D.M., Correa, L.G., Rensing, S.A., Kersten, B. et al. (2009) PlnTFDB: updated content and new features of the plant transcription factor database, *Nucleic acids research*, **38**, D822-827.

Pertoft, H. (2000) Fractionation of cells and subcellular particles with Percoll, *J Biochem Biophys Methods*, **44**, 1-30.

Pierleoni, A., Martelli, P.L., Fariselli, P. and Casadio, R. (2006) BaCelLo: a balanced subcellular localization predictor, *Bioinformatics (Oxford, England)*, **22**, e408-416.

Pollock, S.V., Prout, D.L., Godfrey, A.C., Lemaire, S.D. and Moroney, J.V. (2004) The Chlamydomonas reinhardtii proteins Ccp1 and Ccp2 are required for long-term growth, but are not necessary for efficient photosynthesis, in a low-CO2 environment, *Plant molecular biology*, **56**, 125-132.

Quarmby, L.M. (1996) Ca2+ influx activated by low pH in Chlamydomonas, *J Gen Physiol*, **108**, 351-361.

Ramakers, C., Ruijter, J.M., Deprez, R.H. and Moorman, A.F. (2003) Assumption-free analysis of quantitative real-time polymerase chain reaction (PCR) data, *Neurosci Lett*, **339**, 62-66.

Rappsilber, J., Ishihama, Y. and Mann, M. (2003) Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics, *Analytical chemistry*, **75**, 663-670.

Reiter, L., Claassen, M., Schrimpf, S.P., Jovanovic, M., Schmidt, A. et al. (2009) Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry, *Mol Cell Proteomics*, **8**, 2405-2417.

Renberg, L., Johansson, A.I., Shutova, T., Stenlund, H., Aksmann, A. et al. (2010) A metabolomic approach to study major metabolite changes during acclimation to limiting CO2 in Chlamydomonas reinhardtii, *Plant physiology*, **154**, 187-196.

Reyes, J.C., Muro-Pastor, M.I. and Florencio, F.J. (2004) The GATA family of transcription factors in Arabidopsis and rice, *Plant physiology*, **134**, 1718-1732.

Rho, S., You, S., Kim, Y. and Hwang, D. (2008) From proteomics toward systems biology: integration of different types of proteomics data into network models, *BMB reports*, **41**, 184-193.

Riaño-Pachón, D.M., Correa, L.G., Trejos-Espinosa, R. and Mueller-Roeber, B. (2008) Green transcription factors: a chlamydomonas overview, *Genetics*, **179**, 31-39.

Riaño-Pachón, D.M., Ruzicic, S., Dreyer, I. and Mueller-Roeber, B. (2007) PlnTFDB: an integrative plant transcription factor database, *BMC bioinformatics*, **8**, 42.

Rochaix, J.D. (2002) Chlamydomonas, a model system for studying the assembly and dynamics of photosynthetic complexes, *FEBS Lett*, **529**, 34-38.

Rolland, N., Atteia, A., Decottignies, P., Garin, J., Hippler, M. et al. (2009) Chlamydomonas proteomics, *Current opinion in microbiology*, **12**, 285-291.

Saeed, A.I., Bhagabati, N.K., Braisted, J.C., Liang, W., Sharov, V. et al. (2006) TM4 microarray software suite, *Methods in enzymology*, **411**, 134-193.

Saeed, A.I., Sharov, V., White, J., Li, J., Liang, W. et al. (2003) TM4: a free, open-source system for microarray data management and analysis, *Biotechniques*, **34**, 374-378.

Sam, L.T., Lipson, D., Raz, T., Cao, X., Thompson, J. et al. (2011) A comparison of single molecule and amplification based sequencing of cancer transcriptomes, *PloS one*, **6**, e17305.

Sandhu, C., Hewel, J.A., Badis, G., Talukder, S., Liu, J. et al. (2008) Evaluation of data-dependent versus targeted shotgun proteomic approaches for monitoring transcription factor expression in breast cancer, *Journal of proteome research*, **7**, 1529-1541.

Sauer, F., Hansen, S.K. and Tjian, R. (1995) Multiple TAFIIs directing synergistic activation of transcription, *Science*, **270**, 1783-1788.

Saxena, P.K., Fowke, L.C. and King, J. (1985) An efficient procedure for isolation of nuclei from plant protoplasts, *Protoplasma*, **128**, 184-189.

Schmollinger, S., Strenkert, D. and Schroda, M. (2010) An inducible artificial microRNA system for Chlamydomonas reinhardtii confirms a key role for heat shock factor 1 in regulating thermotolerance, *Curr Genet*, **56**, 383-389.

Schnell, R.A. and Lefebvre, P.A. (1993) Isolation of the Chlamydomonas regulatory gene NIT2 by transposon tagging, *Genetics*, **134**, 737-747.

Schonfeld, C., Wobbe, L., Borgstadt, R., Kienast, A., Nixon, P.J. et al. (2004) The nucleus-encoded protein MOC1 is essential for mitochondrial light acclimation in Chlamydomonas reinhardtii, *The Journal of biological chemistry*, **279**, 50366-50374.

Schulz-Raffelt, M., Lodha, M. and Schroda, M. (2007) Heat shock factor 1 is a key regulator of the stress response in Chlamydomonas, *Plant J*, **52**, 286-295.

Serrano, G., Herrera-Palau, R., Romero, J.M., Serrano, A., Coupland, G. et al. (2009) Chlamydomonas CONSTANS and the evolution of plant photoperiodic signaling, *Curr Biol*, **19**, 359-368.

Shevchenko, A., Wilm, M., Vorm, O. and Mann, M. (1996) Mass spectrometric sequencing of proteins silver-stained polyacrylamide gels, *Analytical chemistry*, **68**, 850-858.

Simon, M. and Hancock, J.M. (2009) Tandem and cryptic amino acid repeats accumulate in disordered regions of proteins, *Genome biology*, **10**, R59.

Solinger, J.A., Kiianitsa, K. and Heyer, W.D. (2002) Rad54, a Swi2/Snf2-like recombinational repair protein, disassembles Rad51:dsDNA filaments, *Molecular cell*, **10**, 1175-1188.

Spijkerman, E. (2008) What physiological acclimation supports increased growth at high CO2 conditions?, *Physiol Plant*, **133**, 41-48.

Spijkerman, E. (2011) The expression of a carbon concentrating mechanism in Chlamydomonas acidophila under variable phosphorus, iron, and CO(2) concentrations, *Photosynthesis research*.

Stauber, E.J., Fink, A., Markert, C., Kruse, O., Johanningmeier, U. et al. (2003) Proteomics of Chlamydomonas reinhardtii light-harvesting proteins, *Eukaryotic cell*, **2**, 978-994.

Stauber, E.J. and Hippler, M. (2004) Chlamydomonas reinhardtii proteomics, *Plant Physiol Biochem*, **42**, 989-1001.

Steck, T.L., Straus, J.H. and Wallach, D.F. (1970) A model for the behavior of vesicles in density gradients: implications for fractionation, *Biochim Biophys Acta*, **203**, 385-393.

Sueoka, N. (1960) Mitotic Replication of Deoxyribonucleic Acid in Chlamydomonas Reinhardi, *Proceedings of the National Academy of Sciences of the United States of America*, **46**, 83-91.

Sun, X. and Wang, G.L. (2011) Genome-wide identification, characterization and phylogenetic analysis of the rice LRR-kinases, *PloS one*, **6**, e16079.

Sutherland, B.W., Toews, J. and Kast, J. (2008) Utility of formaldehyde cross-linking and mass spectrometry in the study of protein-protein interactions, *J Mass Spectrom*, **43**, 699-715.

Terashima, M., Specht, M., Naumann, B. and Hippler, M. (2010) Characterizing the anaerobic response of Chlamydomonas reinhardtii by quantitative proteomics, *Mol Cell Proteomics*, **9**, 1514-1532.

Thimm, O., Blasing, O., Gibon, Y., Nagel, A., Meyer, S. et al. (2004) MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes, *Plant J*, **37**, 914-939.

Thomas, M.C. and Chiang, C.M. (2006) The general transcription machinery and general cofactors, *Critical reviews in biochemistry and molecular biology*, **41**, 105-178.

Toth, J. and Biggin, M.D. (2000) The specificity of protein-DNA crosslinking by formaldehyde: in vitro and in drosophila embryos, *Nucleic acids research*, **28**, e4.

Udvardi, M.K., Czechowski, T. and Scheible, W.R. (2008) Eleven golden rules of quantitative RT-PCR, *The Plant cell*, **20**, 1736-1737.

Valasek, M.A. and Repa, J.J. (2005) The power of real-time PCR, *Adv Physiol Educ*, **29**, 151-159.

Vance, P. and Spalding, M.H. (2005) Growth, photosynthesis, and gene expression in Chlamydomonas over a range of CO2 concentrations and CO2/O2 ratios: CO2 regulates multiple acclimation states1, *Can. J. Bot.* , **83**, 796–809.

Verrijzer, C.P. (2001) Transcription factor IID--not so basal after all, *Science*, **293**, 2010-2011.

Wagner, V., Boesger, J. and Mittag, M. (2009) Sub-proteome analysis in the green flagellate alga Chlamydomonas reinhardtii, *Journal of basic microbiology*, **49**, 32-41.

Wagner, V., Gessner, G. and Mittag, M. (2005) Functional proteomics: a promising approach to find novel components of the circadian system, *Chronobiology international*, **22**, 403-415.

Wang, L. and Roossinck, M.J. (2006) Comparative analysis of expressed sequences reveals a conserved pattern of optimal codon usage in plants, *Plant molecular biology*, **61**, 699-710.

Wang, Z.Y., Kenigsbuch, D., Sun, L., Harel, E., Ong, M.S. et al. (1997) A Myb-related transcription factor is involved in the phytochrome regulation of an Arabidopsis Lhcb gene, *The Plant cell*, **9**, 491-507.

Wienkoop, S., Weiss, J., May, P., Kempa, S., Irgang, S. et al. (2010) Targeted proteomics for Chlamydomonas reinhardtii combined with rapid subcellular protein fractionation, metabolomics and metabolic flux analyses, *Molecular bioSystems*, **6**, 1018-1031.

Winck, F.V., Kwasniewski, M., Wienkoop, S. and Mueller-Roeber, B. (2011) An optimized method for the isolation of nuclei from Chlamydomonas reinhardtii (CHLOROPHYCEAE), *J Phycol* **47** 333–340.

Wright, F. (1990) The 'effective number of codons' used in a gene, *Gene*, **87**, 23-29.

Wykoff, D.D., Grossman, A.R., Weeks, D.P., Usuda, H. and Shimogawara, K. (1999) Psr1, a nuclear localized protein that regulates phosphorus metabolism in Chlamydomonas, *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 15336-15341.

Xiang, Y., Zhang, J. and Weeks, D.P. (2001) The Cia5 gene controls formation of the carbon concentrating mechanism in Chlamydomonas reinhardtii, *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 5341-5346.

Yakovchuk, P., Gilman, B., Goodrich, J.A. and Kugel, J.F. (2010) RNA polymerase II and TAFs undergo a slow isomerization after the polymerase is recruited to promoter-bound TFIID, *Journal of molecular biology*, **397**, 57-68.

Yamano, T. and Fukuzawa, H. (2009) Carbon-concentrating mechanism in a green alga, Chlamydomonas reinhardtii, revealed by transcriptome analyses, *Journal of basic microbiology*, **49**, 42-51.

Yamano, T., Miura, K. and Fukuzawa, H. (2008) Expression analysis of genes associated with the induction of the carbon-concentrating mechanism in Chlamydomonas reinhardtii, *Plant physiology*, **147**, 340-354.

Yoshioka, S., Taniguchi, F., Miura, K., Inoue, T., Yamano, T. et al. (2004) The novel Myb transcription factor LCR1 regulates the CO2-responsive gene Cah1, encoding a periplasmic carbonic anhydrase in Chlamydomonas reinhardtii, *The Plant cell*, **16**, 1466-1477.

Zdobnov, E.M. and Apweiler, R. (2001) InterProScan--an integration platform for the signature-recognition methods in InterPro, *Bioinformatics (Oxford, England)*, **17**, 847-848.

Zhang, M., Ouyang, Q., Stephenson, A., Kane, M.D., Salt, D.E. et al. (2008a) Interactive analysis of systems biology molecular expression data, *BMC systems biology*, **2**, 23.

Zhang, S., Jin, G., Zhang, X.S. and Chen, L. (2007a) Discovering functions and revealing mechanisms at molecular level from biological networks, *Proteomics*, **7**, 2856-2869.

Zhang, X., Henderson, I.R., Lu, C., Green, P.J. and Jacobsen, S.E. (2007b) Role of RNA polymerase IV in plant small RNA metabolism, *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 4536-4541.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S. et al. (2008b) Model-based analysis of ChIP-Seq (MACS), *Genome biology*, **9**, R137.

Appendix A. Isolation of nuclei from Chlamydomonas

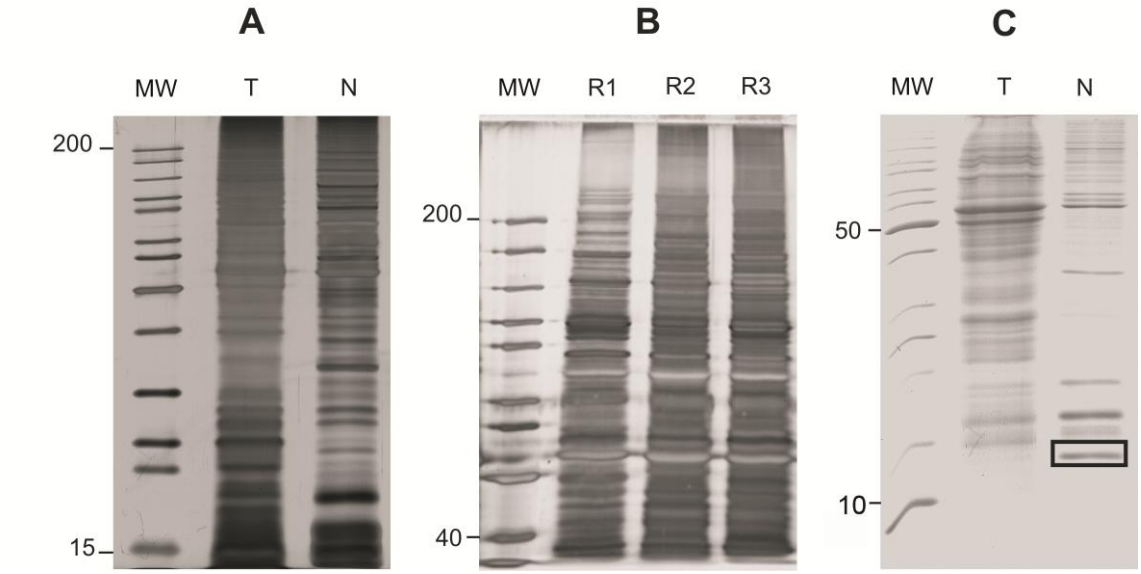


Figure appendix A-1. Enrichment in nuclear proteins and reproducibility of the protein profiles from isolated nuclei.

(A) SDS-PAGE analysis of proteins obtained from whole-cell extracts (lane T) versus those obtained from nuclei (lane N). As expected, the nuclear protein pattern markedly differs from that of whole-cell extracts. (B) Reproducibility of the protein profiles obtained from Chlamydomonas nuclei. Proteins obtained from three independent replicates of nuclear isolates (R1, R2 and R3) were separated by SDS-PAGE and stained with silver nitrate. (C) Proteins from whole-cell and nuclear extracts stained with silver nitrate. The square indicates the gel band that was subjected to protein identification by MALDI-ToF mass spectrometry, resulting in the identification of the protein histone H4. MW, molecular weight in kDa.

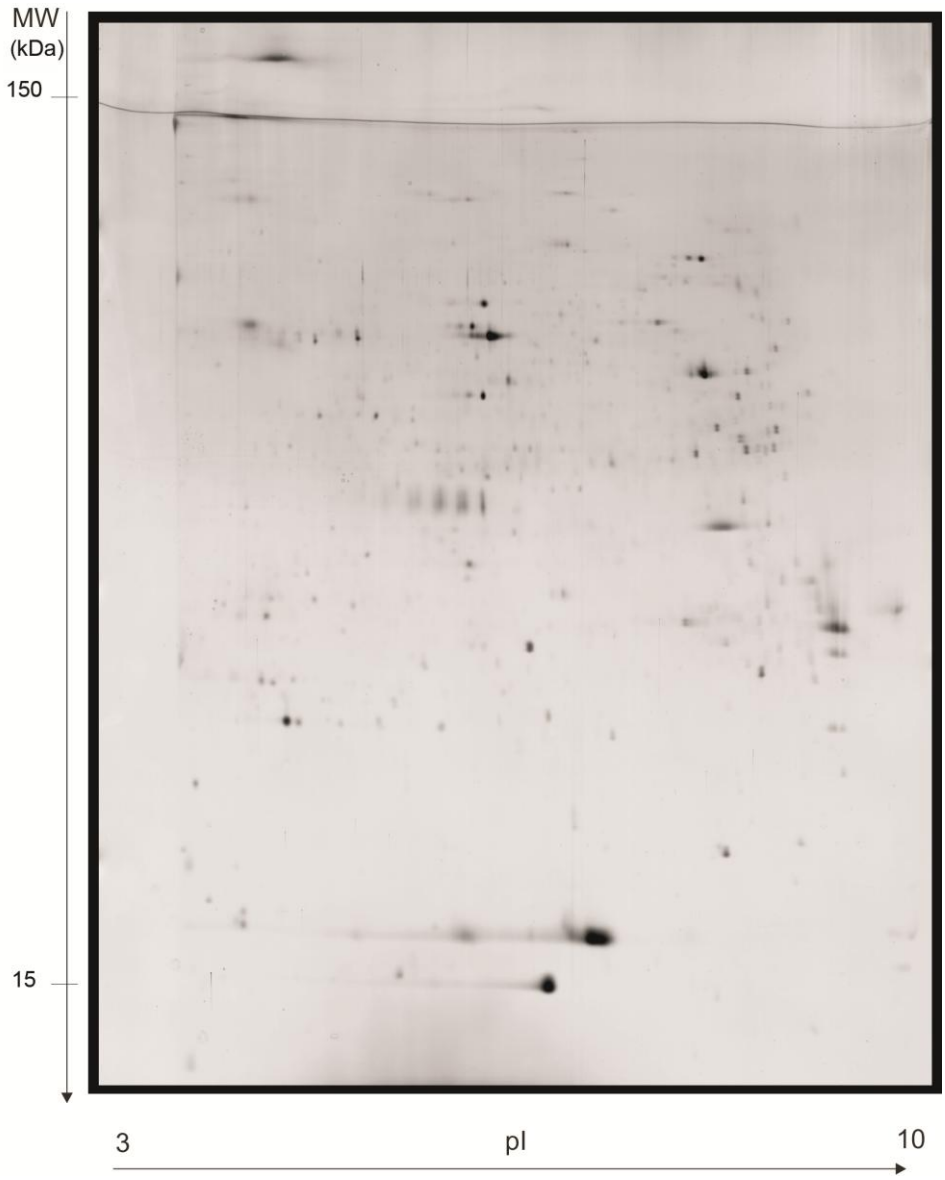


Figure appendix A-2. Two-dimensional electrophoresis of the nuclear proteins of Chlamydomonas.

High-resolution separation of proteins was straightforward using nuclear extracts isolated following the optimized procedure described in this report. A starting amount of ~100 µg total nuclear protein yielded approximately 450 protein spots.

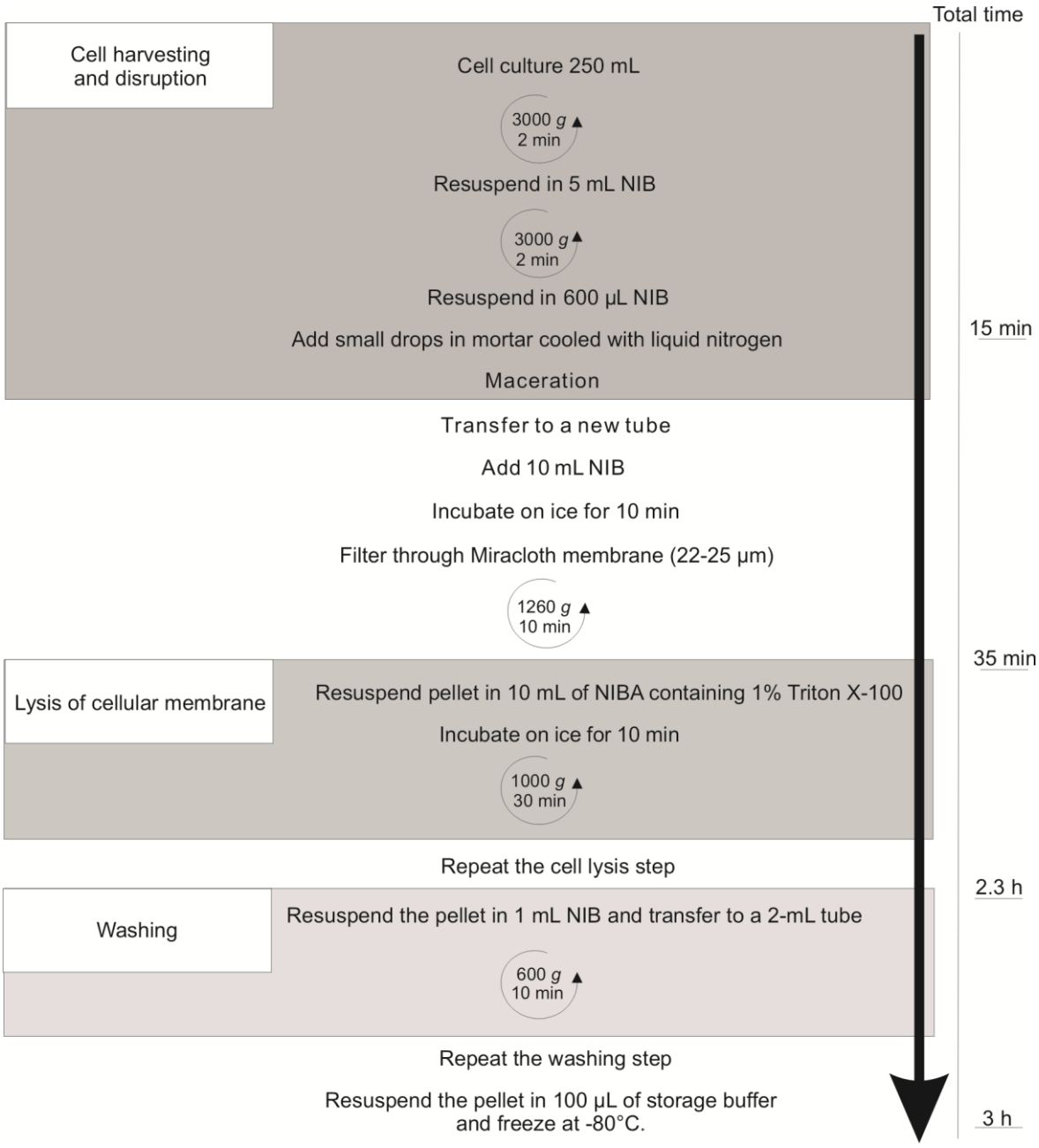


Figure appendix A-3. Scheme of the protocol used for the isolation of nuclei from Chlamydomonas.

Appendix B. Nuclear proteome of Chlamydomonas

Table appendix B-1. Transcription factors and transcription regulators identified in the nuclear proteome

Protein ID	TF/TR family (PlntTFDB)	Pfam family(s)	IPR family(s)	Putative functional annotation[a]
jgi|Chlre3|107210	MYB-related	DnaJ, Myb_DNA-binding	DnaJ, Myb_DNA domain, Zuotin Related FACTOR4.	Protein sequence (jgi|Chlre3|107210) is similar (E-value 1e-114) to the GlsA protein (XP_002953749.1) from *Volvox carteri* which interacts with Hsp70 and is required for asymmetric cell division [1].
jgi|Chlre3|129649	MYB	Myb_DNA-binding, DUF3351	Myb_DNA-binding, Homeodomain_like, SANT.	In plants, MYB transcription factors have been described to be involved in controlling secondary metabolism, regulating cellular morphogenesis and serving in signal transduction pathways of growth regulators [2].
jgi|Chlre3|168936	MYB-related	Myb_DNA-binding	Myb_DNA-binding, SWI/SNF COMPLEX-RELATED,Homeodomain-like, JAB1/MPN domain.	
jgi|Chlre3|135484	SNF2	SNF2_N	SNF2_N, DEXDc, Helicase_C, Homeodomain_like, BchI-ChlI: magnesium chelatase ATPase.	The SNF2 transcription regulator family includes proteins with roles in cellular processes such as transcriptional regulation, recombination and DNA repair [3].
jgi|Chlre3|6783	SNF2	SNF2_N	SNF2_N, Helicase_C, Vps54.	

Protein ID	TF/TR family (PlntTFDB)	Pfam family(s)	IPR family(s)	Putative functional annotation[a]
jgi|Chlre3|142283	HMG	HMG_box	HMG_box, SWI/SNF-RELATED CHROMATIN BINDING PROTEIN.	Transcription regulators of the HMG family contain the HMG-box which mediates the non-sequence-specific interaction of HMGB proteins with linear double-strand DNA [4]. In plants it has been shown that HMG proteins localize to the nucleus and accumulate in the nucleolus. The spatio-temporal expression of *HMGA* genes suggests a role for HMGA proteins in growth and differentiation [5].
jgi|Chlre3|145251	HMG	HMG_box	HMG-box, SWI/SNF-RELATED CHROMATIN BINDING PROTEIN.	
jgi|Chlre3|184386	HMG	HMG_box	HMG-box, SWI/SNF-RELATED CHROMATIN BINDING PROTEIN	
jgi|Chlre4|290277|au.g8318_t1	C3H	RRM_1	RNA-binding domain, RRM, CCCH zinc finger	Members of this family have been shown in Arabidopsis to be expressed throughout the embryo from globular to late cotyledon stage [6].
jgi|Chlre3|142476	C3H	zf-CCCH	zf-CCCH, U2AUXFACTOR.	
jgi|Chlre3|144201	IWS1	Med26	IWS1_C	Transcription regulator domain found in the transcription elongation factors TFIIS [7]. In Arabidopsis it has been shown that IWS1 interacts with transcription factor BES1 and is required for brassinosteroid-induced gene expression [8].
jgi|Chlre3|149734	C2H2	NA	zf-C2H2	C_2H_2 zinc fingers (ZF) display a wide range of functions, from DNA or RNA binding to the involvement in protein-protein interactions. Therefore ZF proteins not only act in transcriptional regulation, either directly or through site-specific modification and/or regulation of chromatin, but also participate in RNA metabolism and in other cellular functions that probably require specific protein contacts of the ZF domain.There are many types of zinc finger proteins, classified according to the number and order of the Cys and His residues that bind the Zinc ion [9].
jgi|Chlre3|17453	TUB	Tub	TUBBY-LIKE PROTEIN 2, TUBBY-RELATED	In mammals, TUBBY-like proteins play an important role in maintenance and function of neuronal cells during post differentiation and development. Tubby-like TULP2 from Chlamydomonas was proposed to be involved in signaling pathways from G protein-coupled receptors (GPCRs). Tubby-like transcripts have been found to be upregulated during flagella regeneration but the protein could not be found in flagella or cilia [10].

Protein ID	TF/TR family (PlntTFDB)	Pfam family(s)	IPR family(s)	Putative functional annotation[a]
jgi|Chlre3|194816	SWI/SNF-BAF60b	DEK_C	BRG-1 ASSOCIATED FACTOR 60 (BAF60), DEK_C, SWIB, SWIB/MDM2 domain	Members of the SNF family are known to function in the form of complexes (SWI/SNF) in chromatin remodelling and nucleosome mobilization. BAF60 protein domains have been shown to interact with multiple transcription factors and are thought to bridge interactions between these transcription factors and SWI/SNF complexes, thereby allowing the recruitment of SWI/SNF to target genes [11].
jgi|Chlre3|195341	TRAF	BTB	BTB, PTHR23230:SF181, BTB/POZ_fold, Kelch_related	Proteins containing the BTB domain, a protein-protein interaction motif, have been shown to participate in transcription repression, cytoskeleton regulation, tetramerization and gating of ion channels and protein degradation [12].
jgi|Chlre4|283773|au.g2242_t1	bZIP	NA	bZIP_2	In plants, the bZIP transcription factors have been described to regulate several processes including pathogen defence, light and stress signalling, seed maturation, flower development, cell elongation, nitrogen/carbon balance control, pathogen defence, energy metabolism, unfolded protein response, hormone and sugar signalling and seed storage protein gene regulation [13].
jgi|Chlre4|288402|au.g6571_t1	PHD	NA	PHD, FYVE/PHD zinc finger, coiled-coil	PHD proteins seem to be found universally in the nucleus, and their functions tend to lie in the control of chromatin or transcription [14]. There is evidence showing that the association between Polycomb Repressive Complex 2 (PRC2)-like complexes with Plant Homeo Domain (PHD)-finger proteins in animals and plants is needed for efficient deposition of histone H3 trimethylated at lysine 27 (H3K27me3) and transcriptional repression [15].
jgi|Chlre4|343150|estExt_fgenesh1_pg.C_chromosome_20875	SBP	BTB, SBP	BTB, EF-hand, KELCH-RELATED PROTEIN, POZ domain, SBP, SBT domain	SBP-box genes are present in every major plant taxon and include genes with and without a *miR156* target site. This observation suggests that *miR156* and its SBP-box targets play an important role in vegetative phase change throughout the plant kingdom [16].

Protein ID	TF/TR family (PlntTFDB)	Pfam family(s)	IPR family(s)	Putative functional annotation[a]
jgi|Chlre3|126810	CSD	RRM_1, CSD	CSD, RRM_1	Using Arabidopsis as a model system four unique *CSD* genes were found to be differentially regulated in response to low temperature [17]. In bacteria (e.g. *Escherichia coli*), cold shock proteins (CSPs) function as transcription antiterminators or translational enhancers at low temperature by destabilizing RNA secondary structure. In wheat, nucleic acid-binding protein WCSP1 was found to contain a cold shock domain (CSD) that has high similarity CSP from *E. coli* [18, 19].

PlntTFDB: PlantTFDatabase (http://plntfdb.bio.uni-potsdam.de/v3.0/); NA : Not assigned; ID: identifier; TF: Transcription factor; TR: Transcription regulator; IPR: InterPro (http://www.ebi.ac.uk/Tools/InterProScan/); Pfam v23: Protein families (http://pfam.sanger.ac.uk/).

a: For references mentioned in Table appendix B-1, please see below.

[1] Miller, S. M., Kirk, D. L., glsA, a Volvox gene required for asymmetric division and germ cell specification, encodes a chaperone-like protein. Development 1999, 126, 649-658.

[2] Martin, C., Paz-Ares, J., MYB transcription factors in plants. Trends Genet 1997, 13, 67-73.

[3] Eisen, J. A., Sweder, K. S., Hanawalt, P. C., Evolution of the SNF2 family of proteins: subfamilies with distinct sequences and functions. Nucleic Acids Res 1995, 23, 2715-2723.

[4] Grasser, K. D., Launholt, D., Grasser, M., High mobility group proteins of the plant HMGB family: dynamic chromatin modulators. Biochim Biophys Acta 2007, 1769, 346-357.

[5] Zhao, J., Paul, L. K., Grafi, G., The maize HMGA protein is localized to the nucleolus and can be acetylated in vitro at its globular domain, and phosphorylation by CDK reduces its binding activity to AT-rich DNA. Biochim Biophys Acta 2009, 1789, 751-757.

[6] Li, Z., Thomas, T. L., PEI1, an embryo-specific zinc finger protein gene required for heart-stage embryo formation in Arabidopsis. Plant Cell 1998, 10, 383-398.

[7] Booth, V., Koth, C.M., Edwards, A.M., Arrowsmith, C.H., Structure of a conserved domain common to the transcription factors TFIIS, elongin A, and CRSP70. J Biol Chem. 2000, 275, 31266-31268.

[8] Li, L., Ye, H., Guo, H., Yin, Y., Arabidopsis IWS1 interacts with transcription factor BES1 and is involved in plant steroid hormone brassinosteroid regulated gene expression. Proc Natl Acad Sci U S A 2010, 107, 3918-3923.

[9] Ciftci-Yilmaz, S., Mittler, R., The zinc finger network of plants. Cell Mol Life Sci 2008, 65, 1150-1160.

[10] Stolc, V., Samanta, M. P., Tongprasit, W., Marshall, W. F., Genome-wide transcriptional analysis of flagellar regeneration in Chlamydomonas reinhardtii identifies orthologs of ciliary disease genes. Proc Natl Acad Sci U S A 2005, 102, 3703-3707.

[11] Lores, P., Visvikis, O., Luna, R., Lemichez, E., Gacon, G., The SWI/SNF protein BAF60b is ubiquitinated through a signalling process involving Rac GTPase and the RING finger protein Unkempt. FEBS J 2010, 277, 1453-1464.

[12] Stogios, P. J., Downs, G. S., Jauhal, J. J., Nandra, S. K., Prive, G. G., Sequence and structural analysis of BTB domain proteins. Genome Biol 2005, 6, R82.

[13] Correa, L. G., Riaño-Pachón, D. M., Schrago, C. G., dos Santos, R. V., et al., The role of bZIP transcription factors in green plant evolution: adaptive features emerging from four founder genes. PLoS One 2008, 3, e2944.

[14] Bienz, M., The PHD finger, a nuclear protein-interaction domain. Trends Biochem Sci 2006, 31, 35-40.

[15] Hennig, L., Derkacheva, M., Diversity of Polycomb group complexes in plants: same rules, different players? Trends Genet 2009, 25, 414-423.

[16] Poethig, R. S., Small RNAs and developmental timing in plants. Curr Opin Genet Dev 2009, 19, 374-378.

[17] Karlson, D., Imai, R., Conservation of the cold shock domain protein family in plants. Plant Physiol 2003, 131, 12-15.

[18] Karlson, D., Nakaminami, K., Toyomasu, T., Imai, R., A cold-regulated nucleic acid-binding protein of winter wheat shares a domain with bacterial cold shock proteins. J Biol Chem 2002, 277, 35248-35256.

[19] Nakaminami, K., Karlson, D. T., Imai, R., Functional conservation of cold shock domains in bacteria and higher plants. Proc Natl Acad Sci U S A 2006, 103, 10122-10127.

Appendix C. Analysis of TF and TR genes regulated by light

Table appendix C-1. List of identified sets of genes which were clustered according to their similar expression pattern during shift from low light (200µE m^{-2} s^{-1}) to high light (700µE m^{-2} $s^{-}1$) conditions

Protein ID	Cluster	Gene name	TF family	Forward primer (5´-> 3´)	Reverse primer (5´-> 3´)
116834	1	e_gwW.1.432.1	E2F-DP	ACGAGCAGGAAAGGGAGCTAGATG	TTCTCAGACAACGCCTGTATCGCC
117914	1	e_gwW.22.121.1	HSF	ATGGATTTCGCAAAGTGGACCCCG	TTCCTTCTTGCCCCGCACAAAGTG
129649	1	estExt_gwp_1H.C_190094	MYB	GGAAGTCTGCCAAGCAGTGCAAAG	ATGCAGCAGCTTCTCATCCTCCTC
135484	1	estExt_gwp_1W.C_110128	SNF2	ACGCTGACATTGACGCCATCATC	TCCGTGAACTGCTGCATCTTCTGG
142793	1	Chlre2_kg.scaffold_6000283	SET	TATAGCTCCTGAACGTGGCTCTGC	AAAACCAAGGCCCACGACCATTC
146239	1	Chlre2_kg.scaffold_1000063	C3H	TGGAGGCGAAGAAGAAGGACAAGG	TGGTCGGCAGAAAAATGTTGCCC
147286	1	Chlre2_kg.scaffold_18000211	bZIP	AATGGAGCTTGGACTCGCAGAGAG	TGCCATTTAGGTAGGCTGCTAGGG
149251	1	Chlre2_kg.scaffold_28000199	CCAAT	CGAAGACATTGAATCCCAGCCACC	AAGTCCATAGTCTTGTCCGCACCC
149876	1	Chlre2_kg.scaffold_27000145	C3H	TGTACCCCAAGTCGGACTCCATTG	ATACAGGAAGTCGAGCACACTGCC
17550	1	fgenesh1_pg.C_scaffold_36000015	C2H2	TGGCATCCAGACAAGAACGCACAC	ATACCAGGCTCGCTCGTGCTTATC
179289	1	fgenesh2_pg.C_scaffold_66000003	C2C2-GATA	TAGCTACGCATGATGGCAACGGTG	TCAGCGCAACCTTGGTGGTGATAG
186976	1	estExt_fgenesh2_pg.C_50434	PLATZ	GACACAGCAAGTACGCTTACCAGC	AGTAAAATAGCCACCTCCAGCCGC

Protein ID	Cluster	Gene name	TF family	Forward primer (5´-> 3´)	Reverse primer (5´-> 3´)
195786	1	BIS_E2F	E2F-DP	ATCCGCTGGAAGATGTCACAGGAG	GCCGAGGTCATGAAGCGAATGTTG
195787	1	BIS_DP	E2F-DP	ACCAGCGAGACCAAGCTGTTTCTG	GAAATCTTCACCTCCACTGTGGCG
195891	1	FER_158168	RWP-RK	ACCGCCTGAGCAATGCTTTATCG	GCTCATTAGCATGTCCATGCAGCC
195917	1	FER_171222	RWP-RK	AGCCATAGCCATAGCTACAGCCAC	GGTGAAAAGAAAGGATGCGGCTCC
93505	1	fgenesh1_est.C_scaffold_5000041	SBP	GTTGACAAGTGCAACCAGGCGATG	TATCCGAAGCCGCTGATACTGCTC
116056	2	e_gwW.16.94.1	MYB	ATACGGCAACAAATGGGCCAGC	CAGTTCCAGCGGTTCTTGATGTGG
117345	2	e_gwW.19.26.1	SNF2	CGCTCCTGCAACACCATGAAGAAG	TGCAAAACACCACCTGCTCCAC
126810	2	estExt_gwp_1H.C_30224	CSD	AAAGGCTTCGGCTTCATCACGC	TCCGAGTTGATGTTGGTCTGGTGC
127044	2	estExt_gwp_1H.C_60003	C2C2-GATA	AGTGCCCTACACCAGCTTTAAGCC	CTTTCCGGTGAGCAAAGCTTCAGG
148821	2	Chlre2_kg.scaffold_24000220	SBP	GGTTTGGCAGCATTCAGCATTGTG	CAGCATCAACTCGCGTTGAGATCG
171659	2	fgenesh2_pg.C_scaffold_1000280	C2C2-GATA	TGTTGTTGGTTGCCGCGTTTGG	TACGAACAAATCGCGCCCACAG
181226	2	fgenesh2_pg.C_scaffold_692000001	FHA	TCTGCCAGCAAAAGTTTGGCTCC	TTGCGACCGATTCGGAATGGAAC
183777	2	estExt_fgenesh2_kg.C_10285	bHLH	CCGCCGAACTCGAATCAATGAGAG	TGATGACCTCCTCCAGAAAGCACG
190311	2	estExt_fgenesh2_pg.C_180098	FHA	CCATTGAAGGTGACAAGGTGCTGG	ACTCATCACCGAACACGATCTCCC
192899	2	estExt_fgenesh2_pg.C_330097	HMG	AGAAGAAGTTCCAGGACTTGGCGG	TAAGCAGCTCCACACCATACGCTC
193280	2	estExt_fgenesh2_pg.C_400083	FHA	AGATGACGCTGTGGAAGCTTGTCC	TTGCGACCGATTCGGAATGGAAC
193681	2	estExt_fgenesh2_pg.C_450107	PHD	AAAGCACCACTAGCTGCCATCCTG	TCTCGCCGTTCTAGCGACTCATAC
194555	2	estExt_fgenesh2_pg.C_690037	C3H	AGTGCCAGGTGAACAACGGTGTAG	ACGCCTTTGACGAAGTCGTAGCAG
194959	2	estExt_fgenesh2_pg.C_800016	HMG	TGTTCGCAAGCTGAAACCCACC	ATGGCAATCTCCTGCTCCTGGTTC
6823	2	fgenesh1_pg.C_scaffold_7000374	SET	ACAAGCAGCAACAACAGCAGCAG	ATCACATCAGGTTCCCGATGCACC
94169	2	fgenesh1_est.C_scaffold_10000094	C3H	ATCAATGCACACTAGGGAAGCCGC	AAAGACAGTCCGTCACTCCACGTC
97080	2	e_gwH.7.97.1	Jumonji	ACATCAAAACCGTGCAGCTCCC	AGTTGTAGCCGCAGTTGTAGCCAG
104871	3	e_gwH.34.98.1	FHA	AGCGCCTTCATCTACGATCCC	CTTGATGCGCTGCTTGTTCAG
108149	3	e_gwH.74.33.1	SBP	CGGTCGCAGCTTGCAAAACAG	TGGTCCTCGCACAGCTTGAAG

Protein ID	Cluster	Gene name	TF family	Forward primer (5´-> 3´)	Reverse primer (5´-> 3´)
112237	3	e_gwW.5.181.1	MYB	CAAGTGGACCGAAATCGCCG	GTTCACCCATCTCAGCCTGC
119209	3	e_gwW.27.84.1	C3H	AACTTCAAGGTGCTGCGCTG	AGAAGCCCTGCTTGTAGTCGG
130971	3	estExt_gwp_1H.C_300115	G2-like	TTGAGAGCGCGGTGATCAAGTTGG	ATCTCCTGCATGATGGTCTTGGGC
142507	3	Chlre2_kg.scaffold_3000480	MYB-related	AGGCGTACCAAGTGATAGCCGATG	CCATTCGCACTAGCAACTAGCGAC
145251	3	Chlre2_kg.scaffold_13000164	HMG	CTCACGCCCTACATCTTGTTCTTC	TTCCACAACTGCCCCATCTTGG
149109	3	Chlre2_kg.scaffold_28000057	MYB-related	ACACCGACTTACACAGAGACGTGG	TAAGTGTCCCGACGCACAATCTCG
149159	3	Chlre2_kg.scaffold_28000107	Orphans	TGACAAGGTCATTCGCTACGCCAC	ATCTTCCACGCCTCAATCTCCTCC
153832	3	Chlre2_kg.scaffold_68000004	SET	AACCATCATTGTGAGGCGGCGAAC	ATGTGCGCCACGTAGTTTAGCAGC
180470	3	fgenesh2_pg.C_scaffold_87000015	Orphans	AGGCGCAGCATCTCAAGGTTAAGC	ATCTGGCTCACGTAGTTTCCCTCG
184386	3	estExt_fgenesh2_kg.C_280024	HMG	ATCAAGGCGCAGAACCCTGACAAG	TCCGACATCCCCTTCCAAATGGAC
188544	3	estExt_fgenesh2_pg.C_100357	SNF2	AGGCGACACTGAGTTACCTGATGC	TGGAGATGATGCGGGGTACTATGC
22211	3	fgenesh1_pg.C_scaffold_94000028	C3H	AACCACCGGACAAACGAGACG	AGATCTGGCCCTGCATCGTGAC
106655	4	e_gwH.45.68.1	MYB	ACAAGTGGTCCGTTATCGCCAAGC	GGTTCTTGACTGCGTTGTCTGTCC
106739	4	e_gwH.46.36.1	SBP	AGGCGCACTTCAAAGCCGAGTATG	GGCTGGAACTTGTTACACTGCTGG
142476	4	Chlre2_kg.scaffold_3000449	C3H	AGTGCAAGAGTGACCGAAAATGGC	CGGCACGCTCCAATCTTGAAGTAG
152660	4	Chlre2_kg.scaffold_48000084	Orphans	GGGAAAAAGCTTAAGGAGCAGGGC	ATCGCGTTGCGATTTGACCACC
168736	4	fgenesh2_pg.C_scaffold_2000382	PHD	AACTGCTGGAGGAAGAGAAACCGC	AAGATTTCGTACTCGCCTGCCGTG
177618	4	fgenesh2_pg.C_scaffold_43000072	SNF2	ATTCGGGAGCCGAACAACACCTAC	ATGTGCCCGATTTTATCGCCGC
187405	4	estExt_fgenesh2_pg.C_60109	PHD	TGGACTCGATTACGCCACAGAACC	ACGTGCTGCTTTACATTGGGTGGG
190589	4	estExt_fgenesh2_pg.C_190258	C2H2	ATCTAGCGAGCGGCACAACGAATC	GAAGCACCTAAGAAACGGAAGGGC
191464	4	estExt_fgenesh2_pg.C_230178	SBP	TTTGCAGTGGACTGGCACATGAC	ATCAGATCCAGCAGGTTCAGCTCG
195154	4	estExt_fgenesh2_pg.C_940001	SET	AAGGATGTTGATCCGCACGCGAAG	AGCTCCAAAAGCTGAGCCACTGAC
195838	4	FER_fgenesh1_est.C_scaffold_27000019	RWP-RK	TTCCGTGAGGAGGTGTACGACAAC	ATTGGCCTGGCGCAGCTTCTTAAC
112628	5	e_gwW.3.655.1	MYB-related	ACATTATCACGAAGCAACGGGAGC	ATTTTCCGCCAGGCACGTCCATAG

Protein ID	Cluster	Gene name	TF family	Forward primer (5´-> 3´)	Reverse primer (5´-> 3´)
119283	5	e_gwW.27.121.1	MYB	AAGAACGACCTTCGGCCCAACATC	AGCACGTACTCCTCCTTGAGTGAC
120035	5	e_gwW.31.178.1	C2C2-Dof	AGCATGGACACCAAATTCTGCTAC	CGATGCTCATTGCTTGCCTTG
149807	5	Chlre2_kg.scaffold_27000076	PHD	TCGCCGCGCTAAAGTCTCATTG	TAGTCCATCATCACGTCACGCAGC
153371	5	Chlre2_kg.scaffold_60000060	SET	TCTGCAAGCCCAACAGCAACAG	TGCTGCAAGGCAAGTCCGTATTG
158359	5	acegs_kg.scaffold_28000085	Orphans	AGTGCGTGGACGTATACAGCAAGG	TTGTAGGTCCAGGAAGACGCAGAC
177657	5	fgenesh2_pg.C_scaffold_44000001	AP2-EREBP	ATGGGAAGCCCGCAAGCGATAAAC	GTTCTTACGATTCCAGCAGACGCC
189471	5	estExt_fgenesh2_pg.C_150166	CCAAT	TCACGTCAAATGCGGAAAGCAGC	GCCACAAGCGCATCCAAGAACATC
193672	5	estExt_fgenesh2_pg.C_450097	bZIP	TGGAGCTGCCAAAGCTACTAAGCC	TGACCCTTGTTGCCCATTCCATCG
53522	5	estExt_GenewiseH_1.C_70037	FHA	AACGGCCAGTACTTTGTCAGCAGC	TGGAAGCCGTACCACTGCATCTTG
101275	6	e_gwH.16.88.1	CCAAT	GGGCGACGATGCAAAGGAATTCAG	GCACTCTGCAAGTGCTGTAAGAGC
121606	6	e_gwW.50.58.1	SBP	TCAACAGTGCTCGCGTTTCCAC	ATCATGCGGCACGTCCTGTTTG
142283	6	Chlre2_kg.scaffold_3000256	HMG	TGAAGTCCGAGAACCCTGGCATTG	TCGTACTCCTTCTTGTCGTCAGCG
147364	6	Chlre2_kg.scaffold_19000038	CCAAT	AAGCAAAAGGTGGACAAGGCCG	ACGCCTTCTCAATGCTGTTGTGC
160281	6	acegs_kg.scaffold_89000008	FHA	CTCGTGAAGGCGAATTTCTCAGGC	ATCGCTACGCGCAACAATGGTC
177573	6	fgenesh2_pg.C_scaffold_43000027	RWP-RK	TGGCCGACCTGAAGCTGTACTATG	TTGAAGCACGACAGGCTGAAGC
188131	6	estExt_fgenesh2_pg.C_90026	SNF2	CATCTTTGACAGCGACTGGAACCC	CCGCTCGTCACAAAGCGGTAAATG
191829	6	estExt_fgenesh2_pg.C_250060	FHA	TTTTGACGCCGCAGAAATGGACGC	TTGGTTTTCTTTGTCCGCCCGAGC
107654	7	e_gwH.64.49.1	C3H	ACTGCAAATTCGCAGACACATGC	GAAGCCGCACTCGTTGTTCTTG
126346	7	estExt_gwp_1H.C_70177	Orphans	GTGAACGATCTGCGAGAGAACGTG	TGTGGAACACCTGTATGATGCGGC
13217	7	fgenesh1_pg.C_scaffold_18000047	SNF2	AAGACAGCCACCGTCATCACCTTC	TCCCAGAACTCCAGCATGGATTGC
188711	7	estExt_fgenesh2_pg.C_110254	Jumonji	ATGCCCCTGTACCTCTTCGACAAG	AAACTGCTCCGGTACGTGGTAGTC
75247	7	gwW.45.133.1	AP2-EREBP	CAAGTACGCGGCTGAGATCC	TAGGTTCCAAGCCACAAGCGG
93901	7	fgenesh1_est.C_scaffold_2000026	Orphans	AGACCTACAAGCCCGAGGACAG	TGACCAGGAACGCCTCAATGC
95196	7	fgenesh1_est.C_scaffold_28000028	bZIP	TGATGAGCAGCAAGGACGAGTACG	TCTGACCCGAATCGTCGGAGAAAG

Protein ID	Cluster	Gene name	TF family	Forward primer (5´-> 3´)	Reverse primer (5´-> 3´)
122706	8	e_gwW.71.31.1	HSF	AGCGGCAAAAGCTTCATCGTCTG	TGGCGCACAAAGGAGGAGAAGTTG
142718	8	Chlre2_kg.scaffold_6000208	FHA	ATAAAGAAAAGGGCGGCTCGGAGG	CCCTGGCAAGCTGTGAAAATCCTG
166327	8	fgenesh2_pg.C_scaffold_7000246	SBP	AGGCGCACATGAAGTCTATGTCCC	TCAACACGGTAAGAACCCCACTGC
170309	8	fgenesh2_pg.C_scaffold_13000060	Jumonji	ATCAATGTGCCCATCCGGCAGAAC	CTTTGAACTCCGCCACAGACATCG
195358	8	estExt_fgenesh2_pg.C_70a0001	C3H	AACGCCTTCCTCACCATTGACGAG	AGTTGTAGAAGATCTGCCAGCCCC
195860	8	FER_estExt_fgenesh2_pg.C_260253	RWP-RK	TGAACAGCCATTGTGGACCCAACC	TCGGACGCTTCCAACATTTGGTG
196759	8	CNA_query	SNF2	TAGCACGTCCCAGCCTGCTTTTAG	ATGCATTGTAAGCCCCGCTCCAAG
134186	9	estExt_gwp_1W.C_30096	C3H	TGCCGTTCGCTTGCTTCATCTG	GGTCGTCTTGGCGTTGTGTTTGAG
149734	9	Chlre2_kg.scaffold_27000003	C2H2	AGCAGTAGCGGCAGTGCAAATG	TATGTGCCCCAACACGCTTCCTTG
178843	9	fgenesh2_pg.C_scaffold_56000076	PHD	TCGAGTACGACATAAACGACCCCG	ATGTAGTCAACCTGCTTGCGCTGG
180087	9	fgenesh2_pg.C_scaffold_82000073	SNF2	AACCCCAAGATGAAGCTGGTGGAG	ATCAGGTTGTTCTGGATGGGCGTG
187531	9	estExt_fgenesh2_pg.C_60308	bZIP	CAGGTGATTGCAAGCCTGAAGGAC	GCACTTCAATGCCTCGTCGATCTC
188572	9	estExt_fgenesh2_pg.C_110041	bZIP	CAGCGACACCTTCGGGTTAATG	TTCGGTATCCATGCTGCTGC
143060	10	Chlre2_kg.scaffold_4000183	PHD	ACCTCCAGTCAGGCATTCCTGTTC	TTTTGTAGTGCCGCACCACCAC
159133	10	acegs_kg.scaffold_38000029	C2C2-CO-like	CGGCGACGTTCAAGGAAATCAAGG	ATGTCGAGCCAGTTGTTGTCCAGG
184359	10	estExt_fgenesh2_kg.C_230055	MYB-related	GCGCACAGTCCCTTCGTCTAATTC	ATCTCGGTCCAACTGTTACCCAGC
187360	10	estExt_fgenesh2_pg.C_60027	C2C2-GATA	TGTTCGGCGGCGTAAGTAACTTCG	TCATCCGACCATTGACATCCACCC
194335	10	estExt_fgenesh2_pg.C_630030	MYB-related	TGCTGGTGCGAAGCTTGTGTTC	ATCCAACTCAAAAGCCGCCCAGTC
108444	11	e_gwH.81.18.1	SBP	TCTGCCAGCAATGCGGAAAGTG	AGAGTGGCGATGCAGCTCTTCTTG
115254	11	e_gwW.13.162.1	SBP	AGTTCGACGGCAGCAGGCGCA	CAGCTTCTTGGCTCGCCGCATGTT
135809	11	estExt_gwp_1W.C_130065	CCAAT	TGCGCTAATGACACGGTGGAGATG	TGGACAGAAGCTGGATGAACTCGG
145462	11	Chlre2_kg.scaffold_14000095	TAZ	CCAGTGCTTCACAGACCAAAAGGG	TCGTTTTTCCGTTTCACCAGGTCG
195840	11	FER_158808	RWP-RK	ACAGCAGCAGCATCAGGCGAATAG	AATGTAGCGGTAGCGACTGCGAAG
106059	Unassigned	e_gwH.39.39.1	PLATZ	TTCCTGTCCGCAACCTACTTCCAG	GCACTCGTTCTTCTTGTGATGCCG

Protein ID	Cluster	Gene name	TF family	Forward primer (5´-> 3´)	Reverse primer (5´-> 3´)
111277	Unassigned	e_gwW.8.20.1	SNF2	TGTTCCTGGACTCAAGCAACACGG	ATGCGGCACCTGAAACTTGTCAGC
111791	Unassigned	e_gwW.7.372.1	CCAAT	ATCACGTCGGAGGCTAGTGACAAG	CGTCGCCGTTAATTGTTTTCCGC
113031	Unassigned	e_gwW.6.283.1	E2F-DP	ACCAGCGAGACCAAGCTGTTTCTG	GAAATCTTCACCTCCACTGTGGCG
118487	Unassigned	e_gwW.23.151.1	SET	TTCTCCAAGCGGATGTACGCCAAC	TCCAACACCTCGCCAATGTACTCG
122939	Unassigned	e_gwW.73.33.1	PHD	AGGGCTACGACGACACGCTTTG	AATTTCCTCCAGTGGCGGGTCCAG
145450	Unassigned	Chlre2_kg.scaffold_14000083	SNF2	AACTGGCACGCTGAGATGGACAAG	TGGAACAGGTGTTGCTGACGAAGC
151215	Unassigned	Chlre2_kg.scaffold_33000083	HMG	TATTCAGCAACGCTATGCGGGCAG	TGACCTCGCCGAAATCAATGCCAG
166515	Unassigned	fgenesh2_pg.C_scaffold_7000434	Pseudo ARR-B	ATGCCAGAGGTTACCGGGTTTGAC	ATCGGCAAAGGTCTCTCCATGCAC
166618	Unassigned	fgenesh2_pg.C_scaffold_5000087	MYB-related	AGAGGAGGAGGCAGTTTTGTGTCG	AGCGTGGAGTGCCAAATGTTCTTG
167010	Unassigned	fgenesh2_pg.C_scaffold_3000020	AP2-EREBP	ATTCAACACAGCGGAGGAGGC	TCGGGGAAGTTGCACTTAGCG
168369	Unassigned	fgenesh2_pg.C_scaffold_2000015	ABI3VP1	AACAACAGCAGCCGCATGTACC	AACACCATTACATCGCCCACCTCC
171833	Unassigned	fgenesh2_pg.C_scaffold_1000454	SBP	AGCGAAAGCAGCAGTTGAGGAAGC	ATCTGTCAGGTCAGCAGTGCATCC
174174	Unassigned	fgenesh2_pg.C_scaffold_24000249	FHA	CAACGGCACCTTTTTGAACGGC	ACGTCCTTCTCCAACAGCTCGTAG
182666	Unassigned	estExt_fgenesh2_kg.C_20019	C3H	AGTCTTTGAGTGCTGGCTGCATCC	AAAGAAGCACACGCGCCTCTTG
188627	Unassigned	estExt_fgenesh2_pg.C_110140	MYB-related	GGTTTTGACGCTGTGCAATGCAAC	ATGGCCCGCTCACTCTTGTGTAAC
189373	Unassigned	estExt_fgenesh2_pg.C_140328	PHD	ACACATAGCACTCATTGCCCTGCC	AGACTGCGAATGCGAAACGAGC
189879	Unassigned	estExt_fgenesh2_pg.C_10357	HMG	ATTCCCCACCGACGACGATATCAG	GTTCTTCTCGTTCTTTGCCGCCAG
195918	Unassigned	FER_rwp13	RWP-RK	ACGACGAAGACTTCGGCTTTGACG	TGAAAGACCTTTCGCAACGCATCC
34069	Unassigned	estExt_fgenesh1_pg.C_280053	FHA	CCGGTGCCTAATGCTGTTCAATCC	TGCCGGAATAAACCTCAAGGCTCG

Table appendix C-2. Gene expression fold change ($2^{-\Delta\Delta Ct}$) of TF and TR genes induced by shift from low light to high light conditions

Protein ID	TP1	TP2	TP3	TP4	TP5	TP6	TP7	TP8	TP9	TP10
101275	0.7896904	0.52513034	0.7305653	0.74788141	0.25806737	0.56282979	0.20781464	0.38899435	0.55232555	0.92975221
104871	2.84423229	1.63841612	0.90550061	1.0510042	0.85029786	0.90654811	0.73569944	1.57500481	0.87855867	1.17956449
106059	0.56346081	1.53179041	0.98576917	0.58120674	0.38480437	0.41571459	0.55876791	0.37359465	0.55128582	1.18863983
106655	0.47157031	0.26525935	1.28560447	0.68059347	0.19785015	0.678016	0.23599371	0.7031912	0.3949572	0.04857584
106739	0.5709591	0.9599351	1.11773291	1.05537569	0.92927736	0.85239371	0.64906885	1.50198998	1.25449654	0.44441047
107654	1.59975597	0.77024304	0.80522684	0.7364587	1.21972311	1.31689938	1.1790372	1.14068646	1.30934083	0.49188763
108149	1.13011624	1.41738025	0.92700186	0.97598688	0.66633961	0.63690781	0.57171765	0.71615687	0.6493031	1.33967241
108444	0.61078555	1.2146352	0.98794607	1.12356753	0.73345407	0.68842758	0.93709491	1.26076639	1.35355628	1.0926891
111277	1.23808202	0.77607876	0.72326954	0.99800179	0.74844442	1.97189112	0.58695968	0.7629093	1.19247061	0.24396055
111791	1.18133795	1.04506722	0.76793916	0.87394163	0.54096145	1.18445527	0.79615836	0.83242049	0.54691034	0.58581724
112237	1.27999207	1.00158789	0.702554	1.26231716	0.87377321	0.59741873	NA	NA	0.28679676	0.66581255
112628	0.44901844	0.12340842	0.47198962	0.36390925	0.0576972	0.15101125	0.34139682	0.13077296	0.32807866	0.13796799
113031	0.55567626	1.19666758	1.08085848	1.24365462	0.71514677	0.87618422	0.7784525	0.84998044	0.7132557	1.0772331
115254	0.67784982	1.18485074	0.97035702	0.80062579	0.55411906	0.65056473	0.78933333	0.98550698	0.77762317	0.72968895
116056	1.07131355	1.02394231	1.32644002	0.92633521	0.73573866	0.83689217	0.85552711	0.68288285	0.39761762	0.54388774
116834	0.53183151	0.72736364	0.98945205	0.72614529	1.49016614	2.52268516	1.57012952	1.94225941	1.47778089	0.64430365
117345	1.7748349	1.29693622	1.05023798	1.12362451	0.75372138	0.53002566	0.37587468	0.86228669	0.71664266	0.71761306
117914	1.04362148	0.47462244	0.99961202	0.8150301	2.12493591	1.87549696	2.07080845	1.83577977	0.76076438	0.32351922
118487	0.80377245	0.6949516	1.19953112	0.6644458	0.61880029	0.64408846	0.52114413	0.88737099	1.28944126	0.23654786
119209	1.87791848	1.97364197	0.8333327	0.73375744	0.95221114	NA	0.43285251	0.66340766	0.5525513	NA
119283	0.78640693	0.17372528	0.98746278	0.50752256	0.36608748	0.54257951	0.39371457	0.25417279	0.18710646	0.06721531
120035	1.91658378	0.8580237	2.07918339	0.91286826	0.35718392	1.27674404	0.63668642	0.30672803	0.17256861	NA
121606	1.64458503	1.72306611	1.05717816	1.84449826	1.4648231	1.35408492	1.43086175	1.47071596	1.54299747	1.59308891
122706	0.43030212	0.49448351	1.0258381	0.59120233	0.89333537	0.91198922	0.76719294	1.5718088	0.16922918	0.2063569

Protein ID	TP1	TP2	TP3	TP4	TP5	TP6	TP7	TP8	TP9	TP10
122939	0.94148201	1.56225188	0.72053684	0.91604113	1.10703403	1.36451865	3.15317544	1.33247846	1.49328588	0.96916039
126346	1.15088795	0.39763754	0.77872713	0.64964279	0.98083284	0.34724145	0.9097805	1.44604246	0.33794809	0.19953827
126810	NA	1.33461889	1.03337349	0.95619084	0.93532138	0.61861221	0.50206065	0.66562223	0.73745676	0.07947688
127044	0.30824415	0.49935792	1.26211933	1.18510177	0.80050787	0.48270075	0.29424137	0.81122541	0.29795879	0.15645679
129649	0.62300536	0.82455253	0.86861576	1.13892885	0.81908711	1.09153092	1.10674548	0.77868415	0.68337585	0.4948516
130971	1.18264106	0.9154992	0.47865857	0.29168443	0.29288904	0.29976502	0.25638359	0.11572898	0.2170736	0.55023684
13217	1.55309814	1.07192668	0.92918246	0.47027043	1.26339887	0.58528246	0.78269611	1.18501473	1.235041	1.09413793
134186	2.32001515	0.8376301	0.66157572	0.70033562	1.02946718	1.08590639	1.53687724	0.9388957	0.70703584	1.57337795
135484	1.34974856	1.2122895	1.15092982	2.52644292	1.36906332	4.27251401	4.37846734	0.80144457	0.49818973	0.95812075
135809	1.2782662	NA	1.2600875	0.91720175	0.89877306	1.27971417	0.8738637	1.5434698	2.33685567	1.62331901
142283	0.8211962	0.66355034	0.6868595	0.91768875	0.56762716	0.57348803	0.48790294	0.57478215	1.22726694	2.5611115
142476	0.40023469	0.93306007	1.00464978	1.25296881	0.70417213	1.46597781	0.5918822	1.13386756	1.27914089	0.89744173
142507	1.23282946	1.32134011	0.64481144	0.7077454	0.53215674	0.63802112	0.51079648	0.35550514	0.43239226	1.68679725
142718	0.72889769	0.68263762	1.12967437	1.18679498	0.94936959	0.83184529	0.97669269	1.16823206	0.76574098	0.79426984
142793	0.40780421	0.58552354	0.79300807	1.0134452	0.68891942	1.57614315	2.13747977	1.85732308	0.7487025	0.95812525
143060	1.00573271	0.70231666	1.34182727	NA	1.04031375	1.10334088	0.71816872	NA	0.6257341	0.45711294
145251	4.63297583	4.38810166	1.04075764	0.73798025	4.17937229	1.12233785	1.06419803	0.85346093	0.6025209	1.4456363
145450	1.27280584	1.23985942	1.00583701	0.78221872	1.02289362	1.32997638	0.91463065	1.37480477	0.92591834	1.74894387
145462	0.83405486	1.00011852	0.94750547	0.93340461	0.9492456	0.67455459	1.11121432	2.33224504	1.30890777	0.71125549
146239	0.59732068	1.20127696	1.07093669	1.65724598	1.55002368	2.54077411	1.70777013	1.46041194	1.02728846	0.89206116
147286	0.7794009	0.6540224	1.17028212	1.42410412	2.63405249	3.81330902	2.44667176	1.37169003	1.23508203	0.78285672
147364	2.69370827	0.8972154	0.60345106	1.19858882	0.72405209	0.48495818	1.06570908	0.49012583	0.58786188	2.76114883
148821	NA	1.52248756	0.93642346	0.71115668	0.49128201	0.18996571	0.223682	0.63685035	0.4926564	0.2454932
149109	1.2293308	1.0472516	0.88583026	0.99332164	1.01180812	0.77958349	0.72197469	0.83485827	0.80323572	1.27929462
149159	3.92087653	2.56850868	0.92081795	1.14673423	1.88969573	1.22953451	1.26686006	1.1216815	1.42740465	0.87719667

Protein ID	TP1	TP2	TP3	TP4	TP5	TP6	TP7	TP8	TP9	TP10
149251	0.66685472	0.36490262	1.4032053	2.42993499	1.61755552	1.72996362	1.71208128	1.13389452	0.75452191	0.68005942
149734	6.79397642	0.86666989	0.97809985	0.31782006	1.89855262	0.68120667	1.51936371	0.87579758	0.92309867	1.94858022
149807	1.85099823	0.11640779	4.68311947	1.00619606	0.75402307	NA	0.3094209	NA	0.80485785	NA
149876	0.56372206	0.49339694	1.1176693	0.58523129	0.92075927	1.36135126	2.48348705	1.24164347	0.46508422	NA
151215	NA	1.5178823	1.40153723	2.23015258	0.49472456	2.86137311	NA	1.4419405	3.88787587	0.22959445
152660	0.82075619	1.10313082	1.36876647	1.46935758	1.41529457	1.45206404	1.03105371	1.72090446	1.12272715	1.16890719
153371	1.55385586	0.88449461	1.14890442	1.15520833	0.90434445	0.67420293	0.42147864	0.9805759	1.00229081	0.72002444
153832	1.54417869	1.23723325	0.82628092	1.0243108	0.87368222	0.83604484	0.84028297	0.89163013	0.69795043	0.79815522
158359	0.6663687	0.59803582	0.58589946	0.31153331	0.12915853	0.2857693	0.41796435	0.27347261	0.1875849	0.04470135
159133	1.33412889	0.83023652	1.49853787	1.16947624	1.32714031	1.67134003	1.2724923	1.60743896	0.98506455	0.36484116
160281	1.27760792	2.08556381	0.53179199	1.56498302	0.71571683	1.59523279	0.47306328	0.46772278	1.45290438	3.92485071
166327	0.58593787	1.44542968	0.91490446	1.68736308	1.59002739	1.22941674	0.87126329	1.52116388	1.14223832	0.63264256
166515	1.94511501	3.19500765	1.09508836	1.101977	1.57960054	1.21416407	1.57743356	1.36956991	1.50566262	0.84410018
166618	0.57214138	0.26426213	1.29302375	0.32763717	0.11754903	0.21961581	0.11890227	1.23285272	0.33106732	NA
167010	1.71422389	0.48948387	0.96773653	0.92012942	1.13869307	0.78437476	0.79765265	0.59633586	1.37892569	0.40029761
168369	0.8495768	1.72894686	0.96144247	1.17437573	1.01329779	1.90220853	1.71945594	1.4773304	1.53245241	0.60098135
168736	0.28056591	0.58403214	0.96129791	0.62028495	0.39960199	0.94198219	0.62309225	1.86238118	0.79281109	0.3321379
170309	1.45693923	1.60664779	1.32729952	1.47213734	3.18136551	1.65771029	2.54474572	1.26313477	0.99736256	0.1318405
171659	0.57129335	1.05574396	1.05465828	0.93695076	1.00187011	0.80877351	0.59542404	0.62617061	0.30104459	0.19474857
171833	1.08750096	0.30127286	NA	1.83457627	0.38760255	1.74081245	NA	0.24329632	NA	1.11472166
174174	3.13339094	2.48085097	0.62610369	0.64136908	1.08457937	1.22595468	1.53386806	0.79099927	0.71925138	4.24833842
17550	NA	0.33909593	0.92930766	2.08586937	1.5503927	3.10789491	1.91007272	1.30419268	0.61585745	0.42221127
177573	1.89275101	0.89943074	0.87737554	0.84279411	0.78116457	0.49128232	0.60307472	0.99836766	1.05399946	1.59293718
177618	0.52734629	0.51635761	1.47683903	1.49393971	0.58865654	0.95796857	0.79501591	0.91475238	0.86319991	0.38358384
177657	1.11458809	0.98377501	0.8754521	0.80023865	0.48217916	0.67903795	0.88732096	0.8639056	0.68350492	0.14484007

Protein ID	TP1	TP2	TP3	TP4	TP5	TP6	TP7	TP8	TP9	TP10
178843	2.68271249	0.95993945	0.79928774	1.20812138	1.27422931	2.92240478	3.92430661	1.47046007	0.30439552	0.8181563
179289	0.37059798	0.34239231	0.97223277	0.7249447	1.27969971	1.34840506	1.2356583	0.78367823	0.5912445	0.06056354
180087	7.57069526	1.15958244	1.13333125	0.51386446	2.02339017	1.16178642	3.50767413	1.35980078	3.10503888	1.32699255
180470	2.06735786	1.63073175	0.85794302	0.91452427	1.5694393	1.21533979	1.12976955	NA	0.94677988	1.25049948
181226	0.62348932	1.57004046	1.1084359	0.85312333	0.4754109	0.63440784	0.19584019	0.34665726	0.60177891	0.72442133
182666	NA	0.39345209	0.85538414	0.49032849	0.72916298	0.86922933	1.22644272	1.14942985	0.52347407	NA
183777	0.92736543	0.84520552	1.02796576	0.96483254	0.69242693	0.63712378	0.58767658	1.17138674	0.58379208	0.18867203
184359	0.33020041	0.71399744	7.16376486	6.3985888	1.85849289	2.6038817	0.45266739	2.07748363	1.64985341	0.13586935
184386	1.45304631	1.23355299	0.61569066	0.47647636	0.43001548	0.21022742	0.2413842	0.20403639	0.61551225	1.57364754
186976	0.99811525	0.74796524	1.16982866	1.50841468	2.3099792	5.73423959	3.53214642	1.85297795	3.13341753	1.99902627
187360	1.00704482	0.5965528	1.09336951	1.59715861	1.20411871	1.11013092	0.87733699	0.77359368	0.84162982	0.55974592
187405	0.39580568	1.21392386	1.24537358	1.64391707	0.95966905	1.10766794	0.76392166	1.88653835	1.93540284	1.00520195
187531	4.05733038	1.4000503	1.08053367	0.76933578	1.5286362	2.47580642	1.91427466	1.63338003	0.8323471	0.9238699
188131	1	1	1	1	0.50290764	NA	1	0.50078595	1	1
188544	0.95299635	1.73394237	0.84002705	0.68552453	0.92211258	0.63101719	1.08268158	0.52408318	0.19118207	0.73102123
188572	9.72829329	1.23639889	0.92245375	1.16284171	1.58115437	2.40110285	4.73695478	1.28340939	2.25608189	NA
188627	0.5965444	0.736275	0.8180666	0.82647104	0.47356142	1.26459549	0.72711114	0.66976111	0.32785525	0.83088569
188711	2.09093134	1.26164992	1.2488882	1.920801	2.43211701	1.04666061	1.44716943	2.1764254	1.74130096	1.4405322
189373	0.66983611	0.70035196	0.88835202	0.93775354	0.87587082	0.88273966	0.87439613	1.40426282	1.01866831	1.26177172
189471	1.5044385	0.62173811	0.65092481	0.51317319	0.1850219	0.21472808	0.28938359	0.55257606	0.41075617	0.61638907
189879	0.81924677	0.77436144	0.74119904	0.6742289	0.72583324	0.75847757	1.45709256	0.89481802	0.94351971	0.88629565
190311	1.06684758	1.01164171	0.8687551	0.97396864	0.91193237	0.60720211	0.68311303	0.61238819	0.44800486	0.59586877
190589	0.56854082	0.87827376	0.83892928	0.86058019	0.57114842	0.86692031	0.79932538	1.12745348	0.91863423	0.86865213
191464	0.5844227	0.54672769	1.23114332	0.55416786	0.35561099	2.55150082	0.44100005	1.03997408	0.98450388	NA
191829	0.53348738	0.62084946	0.57524323	0.76041613	0.54104572	0.45354926	0.38633181	0.41047529	0.51101718	2.05711501

Protein ID	TP1	TP2	TP3	TP4	TP5	TP6	TP7	TP8	TP9	TP10
192899	0.7409543	0.78062816	0.81158091	0.70814167	0.72985443	0.60869194	0.64218967	0.57498154	0.55113303	0.70082122
193280	0.77716747	0.80823569	0.98984315	1.18470099	0.36146046	0.65387185	0.14622524	0.36833131	0.86317538	0.76766736
193672	1.51320409	1.10695453	1.16919281	0.98119023	0.68809231	0.55038743	0.70849863	1.13172069	0.89656862	0.69203144
193681	0.59509921	1.22565869	1.04792739	0.97158439	0.57450755	0.94378658	0.63316254	0.96777945	0.46078448	0.80241764
194335	NA	0.92753355	1.00866503	2.69542293	1.0409576	2.98337539	1.27660849	1.61448737	0.6647182	0.18740174
194555	0.55052653	0.61538471	0.8909328	0.9245844	0.47088131	0.47492229	0.25989782	0.7098914	0.6235612	0.55560089
194959	0.16113595	1.66641734	1.09813623	0.62306532	0.22194838	0.36760917	0.19002181	0.58292862	0.29772362	0.28470005
195154	0.29798755	0.61708078	1.36607448	0.88290713	0.32132411	0.77144418	0.41677345	2.02904019	0.57915438	0.09742007
195358	0.73901082	1.39298887	0.90729938	0.89878637	2.10193473	0.98792471	2.09393551	1.62418908	1.12584227	0.42143477
195786	0.31010749	0.62445964	0.85528261	1.25217649	1.54119026	2.48613747	1.27325044	2.08007735	1.79435938	0.13854064
195787	0.61664931	0.79123298	1.097307	1.29982754	1.35302415	2.85185156	1.56892065	1.94046354	1.68790692	0.79066705
195838	0.38572794	0.82003969	0.95368802	0.86832484	0.80635347	0.76388714	0.76566909	1.67395888	1.45831469	0.86551854
195840	1.85390923	3.49546994	1.19679248	2.13127128	1.92759021	1.60882976	1.63356711	2.88372888	2.69027821	1.28282127
195860	0.81170588	1.04575909	0.99022993	1.13450325	1.32070904	0.980578	1.32867683	1.5329489	1.4968071	0.62908219
195891	1.60446353	1.35414796	1.31448374	2.13329484	1.98125016	5.73620621	4.17261129	3.82400749	1.15228208	3.09137971
195917	1.23993355	0.58325848	5.15696073	44.6466671	59.3107046	628.442913	95.1037423	67.7454687	19.8615278	15.9403272
195918	1.9029511	2.82818722	1.0689547	0.86361511	1.64325087	1.8905412	1.8724701	3.38033129	1.4615636	3.33287746
196759	0.39794108	1.66875999	1.83893027	1.46891747	1.37080388	1.2661196	1.71677711	1.20584814	1.25206917	0.27348878
22211	5.71927855	3.9365218	0.99892373	1.43064285	2.60456656	2.0268739	3.71008447	1.81711776	1.69350169	2.37375264
34069	1.56998224	0.88209016	0.83262091	0.76193522	0.48921281	0.20272558	0.45633565	0.88388279	1.02540753	0.75267985
53522	1.16706769	0.75794266	1.16598637	0.85998547	0.52046014	0.89605668	0.88802968	0.66750425	0.52560856	0.19588627
6823	0.46643471	0.96936021	1.51772032	1.44349231	0.48831994	0.26491522	0.15149097	0.27582732	0.28368103	0.89299367
75247	4.38446334	1.57745066	0.9823673	0.80568361	4.34758663	1.34553738	2.64565335	1.76541251	1.59430752	1.6968672
93505	0.67641485	0.69963658	1.03322017	2.47172864	3.54865714	5.46316954	3.27357689	2.11890645	1.88758179	2.09342791
93901	2.78440826	2.67592096	0.72341991	0.49194545	2.4577488	0.42162111	0.80176281	1.41764925	0.89107393	0.376759

Protein ID	TP1	TP2	TP3	TP4	TP5	TP6	TP7	TP8	TP9	TP10
94169	0.86184226	0.91842683	1.43467818	1.5883382	0.61944461	1.36347307	0.68424173	0.4028903	0.36652347	0.32970394
95196	1.48502128	0.81891171	0.79078991	0.83089199	1.13757281	0.80403675	0.68375713	0.95956878	0.66191312	0.47333593
97080	0.57399406	0.9424158	0.84527641	0.91553251	0.61549691	0.76552784	0.47552933	0.37595308	0.40819582	0.38098729

Table appendix C-3. Motif analysis of promoter regions of light-responsive genes found in Cluster 10

Sequence	Motif	Start	P-value	Left flank	Motif site	Right flank	Motif regular expression	MAST position p-value
184359	1	82	1.74e-08	TTGAAGGTGC	ATACAGGTGCTTCA	CGGGCTTTGC	[AC]TA[CT]A[GA][GCT]T[GA][CT][AGT][TA][CT]A	4.2e-08
143060	1	1049	5.02e-08	AAATACCTTC	ATACAATTGCGTCA	ATCAGTGTGT	[AC]TA[CT]A[GA][GCT]T[GA][CT][AGT][TA][CT]A	1.9e-07
187360	1	40	8.97e-08	GTCCCCGTGC	ATATAAGTGTGTCA	AAAAGCGGAC	[AC]TA[CT]A[GA][GCT]T[GA][CT][AGT][TA][CT]A	> 1e-04
159133	1	616	2.93e-07	GATAATGCAG	CTACAGCTGCATCA	GCACACCTCG	[AC]TA[CT]A[GA][GCT]T[GA][CT][AGT][TA][CT]A	>1e-04
194335	1	1302	5.34e-07	TCTGTGATGT	ATATAGGTACAATA	CTACACTGTT	[AC]TA[CT]A[GA][GCT]T[GA][CT][AGT][TA][CT]A	>1e-04
143060	2	1279	3.76e-09	TTCGCCTTGG	CACACGTAAACCACA	ATCGGCCTCG	CACA[CA][ACG][TC]A[ACT]ACCACA	9.8e-09
159133	2	1471	3.76e-09	TCCCCAACAA	CACAAATATACCACA	CCTCTATACG	CACA[CA][ACG][TC]A[ACT]ACCACA	9.8e-09
194335	2	193	1.75e-08	CCACACCCAC	CACACCCACACCACA	CCCACACCCA	CACA[CA][ACG][TC]A[ACT]ACCACA	4.8e-06
194335	3	1222	9.20e-09	GTGCTGCACC	AGATGTGTTGCATCC	GGGAGACCGG	A[GT]A[TC][GT]T[GC]T[TA][GC]C[AGT][TG][CG][AC]	2.4e-08
187360	3	728	1.86e-08	GCCTAAGCAG	AGATGTCTAGCATGA	AGCCGTCTTA	A[GT]A[TC][GT]T[GC]T[TA][GC]C[AGT][TG][CG][AC]	4.4e-08
143060	3	1415	3.64e-08	GCAAGATTAC	AGATGTGTTCCGTCA	GAGCGTGCAA	A[GT]A[TC][GT]T[GC]T[TA][GC]C[AGT][TG][CG][AC]	9.3e-08
159133	3	1065	3.64e-08	GATGGCGTGG	ATATGTGTTGCAGCA	AGGATAAAGG	A[GT]A[TC][GT]T[GC]T[TA][GC]C[AGT][TG][CG][AC]	9.3e-08
184359	3	920	1.93e-07	TTACGGTGGG	AGACTTGTAGCTTGA	GGAGCCGAGC	A[GT]A[TC][GT]T[GC]T[TA][GC]C[AGT][TG][CG][AC]	>1e-04
184359	4	316	5.40e-09	TGCGGCGTCA	ACTGAAATTGAGA	AACAGCATGC	ACTG[AC][AC]ATTG[AC]GA	2.3e-08
143060	4	1013	2.78e-08	CCGACAACCG	ACTGCAATTGAGA	CTATACCGTA	ACTG[AC][AC]ATTG[AC]GA	7.8e-08
159133	4	1298	5.86e-08	TGCACGTACC	ACTGACATTGCGA	TAACATATAA	ACTG[AC][AC]ATTG[AC]GA	1.2e-07
159133	5	202	8.97e-08	GGGCCGCGTG	TATGTATGCAT	GTACTGGCTA	[TC]A[TC][GC][TC][AG][TC]G[CT]AT	>1e-04
194335	5	517	3.03e-07	GCGCTCACCC	TATCTATGTAT	CACCAGGCAA	[TC]A[TC][GC][TC][AG][TC]G[CT]AT	>1e-04
143060	5	929	7.90e-07	TGTTTGGCAG	CATGTATGCAT	CAAGTCGTTT	[TC]A[TC][GC][TC][AG][TC]G[CT]AT	>1e-04
187360	5	385	4.10e-06	TTATTTACGC	TACGTGTGCAT	ACCTGACTTG	[TC]A[TC][GC][TC][AG][TC]G[CT]AT	>1e-04
184359	5	288	5.03e-06	GCGGCGGCGC	TATCCACGCAT	GCTTTTTTGC	[TC]A[TC][GC][TC][AG][TC]G[CT]AT	>1e-04
184359	6	243	1.19e-05	AGCGTGCTGG	CATCCACA	TGGGCAAGGT	CA[TC]C[CA]ACA	>1e-04

Sequence	Motif	Start	P-value	Left flank	Motif site	Right flank	Motif regular expression	MAST position p-value
143060	6	421	1.19e-05	AATACCTTGA	CATCCACA	GCGACGGGGC	CA[TC]C[CA]ACA	>1e-04
187360	6	330	1.19e-05	TCCGGCTCCC	CATCCACA	AGAAGTTTCC	CA[TC]C[CA]ACA	>1e-04
194335	6	130	2.80e-05	ACAATCACCA	CACCCACA	CCCACACACC	CA[TC]C[CA]ACA	>1e-04
159133	6	473	4.84e-05	GGCCGCGACC	CACCAACA	ACCGACCTAT	CA[TC]C[CA]ACA	>1e-04
184359	7	1211	4.00e-09	TACCAAGGAG	AATGCCAATAATAT	TGAGCTCGCG	AATG[AC]CAAT[AC][AT]TAT	4.9e-08
187360	7	1368	4.00e-09	GCGTGTCTTG	AATGACAATCTTAT	GAACATTTGA	AATG[AC]CAAT[AC][AT]TAT	4.9e-08
184359	8	506	1.57e-08	CCACTTCCGC	CTAATACAATCCGCC	TCCCGCCTGG	CT[AT][ACGTA][CGT]ACAA[GT][CGT]CGC[CT]	1.6e-08
143060	8	187	4.41e-08	GGGCCGCCAG	CTTGCACAAGCCGCC	TGCCGCACCC	CT[AT][ACGTA][CGT]ACAA[GT][CGT]CGC[CT]	4.5e-08
187360	8	819	1.41e-07	GGGACGCCCT	CTTTGACAAGTCGCC	AAGTAAGGGC	CT[AT][ACGTA][CGT]ACAA[GT][CGT]CGC[CT]	1.4e-07
194335	8	701	1.86e-07	GTGGCGCTCA	CTACCACAATGCGCT	CTAAAAGGGT	CT[AT][ACGTA][CGT]ACAA[GT][CGT]CGC[CT]	8.7e-05
194335	9	1385	6.19e-09	TTGTAGTCCG	CAATAAGCCTCGCAA	ACCAGCACCG	[CA][AT][AT][TG][ACT][AC]G[CG]CT[GACTA]G[AC][AG]A	1.0e-05
187360	9	992	1.86e-08	GGTGGCTTTT	CAAGAAGCCTGGAAA	CTAAGGAGCT	[CA][AT][AT][TG][ACT][AC]G[CG]CT[GACTA]G[AC][AG]A	5.7e-08
184359	9	1534	1.35e-07	ACACCAAAAT	CTATTCGCCTTGAAA	CCAACGGACC	[CA][AT][AT][TG][ACT][AC]G[CG]CT[GACTA]G[AC][AG]A	>1e-04
143060	9	668	1.35e-07	TGCCCGGGAC	CATTCAGCCTGGCAA	CTGGTAGCGG	[CA][AT][AT][TG][ACT][AC]G[CG]CT[GACTA]G[AC][AG]A	>1e-04
159133	9	1315	4.84e-07	TTGCGATAAC	ATATAAGGCTAGAGA	TCAAGTTAAA	[CA][AT][AT][TG][ACT][AC]G[CG]CT[GACTA]G[AC][AG]A	>1e-04
184359	10	1505	4.67e-06	CAGCAGCTGT	AAATCATT	TGCACCAGCA	AAATCATT	>1e-04
143060	10	969	4.67e-06	AACAGGGCAC	AAATCATT	TTGGCGCGAG	AAATCATT	>1e-04

Appendix D. Analysis of TF and TR genes under carbon deprivation

Table appendix D-1. Primer pairs of the TF and TR genes investigated in carbon deprivation and *Cah1*

Protein ID	Gene name	Gene family	Forward primer (5´->3´)	Reverse primer (5´->3´)
101275	e_gwH.16.88.1	CCAAT	GGGCGACGATGCAAAGGAATTCAG	GCACTCTGCAAGTGCTGTAAGAGC
104871	e_gwH.34.98.1	FHA	AGCGCCTTCATCTACGATCCC	CTTGATGCGCTGCTTGTTCAG
105687	e_gwH.38.105.1	C3H	CACGTGATGCACAAGCCCAGT	CGTGGTCGCAGTCCATCAGTC
106059	e_gwH.39.39.1	PLATZ	TTCCTGTCCGCAACCTACTTCCAG	GCACTCGTTCTTCTTGTGATGCCG
106655	e_gwH.45.68.1	MYB	ACAAGTGGTCCGTTATCGCCAAGC	GGTTCTTGACTGCGTTGTCTGTCC
106739	e_gwH.46.36.1	SBP	AGGCGCACTTCAAAGCCGAGTATG	GGCTGGAACTTGTTACACTGCTGG
107654	e_gwH.64.49.1	C3H	ACTGCAAATTCGCAGACACATGC	GAAGCCGCACTCGTTGTTCTTG
107982	e_gwH.71.56.1	AP2-EREBP	AGATACGCGACCCGCTCAAGG	CGCTGTCAAATGTGCCCAGCC
108149	e_gwH.74.33.1	SBP	CGGTCGCAGCTTGCAAAACAG	TGGTCCTCGCACAGCTTGAAG
108444	e_gwH.81.18.1	SBP	TCTGCCAGCAATGCGGAAAGTG	AGAGTGGCGATGCAGCTCTTCTTG
111277	e_gwW.8.20.1	SNF2	TGTTCCTGGACTCAAGCAACACGG	ATGCGGCACCTGAAACTTGTCAGC
111791	e_gwW.7.372.1	CCAAT	ATCACGTCGGAGGCTAGTGACAAG	CGTCGCCGTTAATTGTTTTCCGC
112628	e_gwW.3.655.1	MYB-related	ACATTATCACGAAGCAACGGGAGC	ATTTTCCGCCAGGCACGTCCATAG

Protein ID	Gene name	Gene family	Forward primer (5´->3´)	Reverse primer (5´->3´)
113031	e_gwW.6.283.1	E2F-DP	ACCAGCGAGACCAAGCTGTTTCTG	GAAATCTTCACCTCCACTGTGGCG
115254	e_gwW.13.162.1	SBP	AGTTCGACGGCAGCAGGCGCA	CAGCTTCTTGGCTCGCCGCATGTT
116834	e_gwW.1.432.1	E2F-DP	ACGAGCAGGAAAGGGAGCTAGATG	TTCTCAGACAACGCCTGTATCGCC
117345	e_gwW.19.26.1	SNF2	CGCTCCTGCAACACCATGAAGAAG	TGCAAAACACCACCTGCTCCAC
117655	e_gwW.20.153.1	CCAAT	TGCCAAGGCGTGTGAGATGTTC	GGCACAATGTCGATCAGGAAGTC
117914	e_gwW.22.121.1	HSF	ATGGATTTCGCAAAGTGGACCCCG	TTCCTTCTTGCCCCGCACAAAGTG
118487	e_gwW.23.151.1	SET	TTCTCCAAGCGGATGTACGCCAAC	TCCAACACCTCGCCAATGTACTCG
119283	e_gwW.27.121.1	MYB	AAGAACGACCTTCGGCCCAACATC	AGCACGTACTCCTCCTTGAGTGAC
120035	e_gwW.31.178.1	C2C2-Dof	AGCATGGACACCAAATTCTGCTAC	CGATGCTCATTGCTTGCCTTG
120259	e_gwW.35.20.1	Jumonji	TACCTGGCATACCTCCAACACG	AGGTGCCGTCAAACACATACAGC
121606	e_gwW.50.58.1	SBP	TCAACAGTGCTCGCGTTTCCAC	ATCATGCGGCACGTCCTGTTTG
122706	e_gwW.71.31.1	HSF	AGCGGCAAAAGCTTCATCGTCTG	TGGCGCACAAAGGAGGAGAAGTTG
122939	e_gwW.73.33.1	PHD	AGGGCTACGACGACACGCTTTG	AATTTCCTCCAGTGGCGGGTCCAG
123164	e_gwW.83.55.1	MYB-related	GGATGGTTCTCGCTTACAGACTCG	TAGAGTTTCAGCGCCTCCACGAAG
126346	estExt_gwp_1H.C_70177	Orphans	GTGAACGATCTGCGAGAGAACGTG	TGTGGAACACCTGTATGATGCGGC
126810	estExt_gwp_1H.C_30224	CSD	AAAGGCTTCGGCTTCATCACGC	TCCGAGTTGATGTTGGTCTGGTGC
127044	estExt_gwp_1H.C_60003	C2C2-GATA	AGTGCCCTACACCAGCTTTAAGCC	CTTTCCGGTGAGCAAAGCTTCAGG
129649	estExt_gwp_1H.C_190094	MYB	GGAAGTCTGCCAAGCAGTGCAAAG	ATGCAGCAGCTTCTCATCCTCCTC
130971	estExt_gwp_1H.C_300115	G2-like	TTGAGAGCGCGGTGATCAAGTTGG	ATCTCCTGCATGATGGTCTTGGGC
13217	fgenesh1_pg.C_scaffold_18000047	SNF2	AAGACAGCCACCGTCATCACCTTC	TCCCAGAACTCCAGCATGGATTGC
135484	estExt_gwp_1W.C_110128	SNF2	ACGCTGACATTGACGCCATCATC	TCCGTGAACTGCTGCATCTTCTGG
135809	estExt_gwp_1W.C_130065	CCAAT	TGCGCTAATGACACGGTGGAGATG	TGGACAGAAGCTGGATGAACTCGG
142283	Chlre2_kg.scaffold_3000256	HMG	TGAAGTCCGAGAACCCTGGCATTG	TCGTACTCCTTCTTGTCGTCAGCG
142476	Chlre2_kg.scaffold_3000449	C3H	AGTGCAAGAGTGACCGAAAATGGC	CGGCACGCTCCAATCTTGAAGTAG

Protein ID	Gene name	Gene family	Forward primer (5´->3´)	Reverse primer (5´->3´)
142507	Chlre2_kg.scaffold_3000480	MYB-related	AGGCGTACCAAGTGATAGCCGATG	CCATTCGCACTAGCAACTAGCGAC
142718	Chlre2_kg.scaffold_6000208	FHA	ATAAAGAAAAGGGCGGCTCGGAGG	CCCTGGCAAGCTGTGAAAATCCTG
142793	Chlre2_kg.scaffold_6000283	SET	TATAGCTCCTGAACGTGGCTCTGC	AAAACCAAGGCCCACGACCATTC
145251	Chlre2_kg.scaffold_13000164	HMG	CTCACGCCCTACATCTTGTTCTTC	TTCCACAACTGCCCCATCTTGG
145450	Chlre2_kg.scaffold_14000083	SNF2	AACTGGCACGCTGAGATGGACAAG	TGGAACAGGTGTTGCTGACGAAGC
146239	Chlre2_kg.scaffold_1000063	C3H	TGGAGGCGAAGAAGAAGGACAAGG	TGGTCGGCAGAAAAATGTTGCCC
147286	Chlre2_kg.scaffold_18000211	bZIP	AATGGAGCTTGGACTCGCAGAGAG	TGCCATTTAGGTAGGCTGCTAGGG
147364	Chlre2_kg.scaffold_19000038	CCAAT	AAGCAAAAGGTGGACAAGGCCG	ACGCCTTCTCAATGCTGTTGTGC
148821	Chlre2_kg.scaffold_24000220	SBP	GGTTTGGCAGCATTCAGCATTGTG	CAGCATCAACTCGCGTTGAGATCG
149109	Chlre2_kg.scaffold_28000057	MYB-related	ACACCGACTTACACAGAGACGTGG	TAAGTGTCCCGACGCACAATCTCG
149159	Chlre2_kg.scaffold_28000107	Orphans	TGACAAGGTCATTCGCTACGCCAC	ATCTTCCACGCCTCAATCTCCTCC
149734	Chlre2_kg.scaffold_27000003	C2H2	AGCAGTAGCGGCAGTGCAAATG	TATGTGCCCCAACACGCTTCCTTG
151215	Chlre2_kg.scaffold_33000083	HMG	TATTCAGCAACGCTATGCGGGCAG	TGACCTCGCCGAAATCAATGCCAG
152660	Chlre2_kg.scaffold_48000084	Orphans	GGGAAAAAGCTTAAGGAGCAGGGC	ATCGCGTTGCGATTTGACCACC
153371	Chlre2_kg.scaffold_60000060	SET	TCTGCAAGCCCAACAGCAACAG	TGCTGCAAGGCAAGTCCGTATTG
153832	Chlre2_kg.scaffold_68000004	SET	AACCATCATTGTGAGGCGGCGAAC	ATGTGCGCCACGTAGTTTAGCAGC
153922	Chlre2_kg.scaffold_69000044	C3H	ATAGCAACAGCGCTGAGCAGAC	TGTTGCTGATGCAGGAGTACCG
159133	acegs_kg.scaffold_38000029	C2C2-CO-like	CGGCGACGTTCAAGGAAATCAAGG	ATGTCGAGCCAGTTGTTGTCCAGG
160281	acegs_kg.scaffold_89000008	FHA	CTCGTGAAGGCGAATTTCTCAGGC	ATCGCTACGCGCAACAATGGTC
166327	fgenesh2_pg.C_scaffold_7000246	SBP	AGGCGCACATGAAGTCTATGTCCC	TCAACACGGTAAGAACCCCACTGC
166515	fgenesh2_pg.C_scaffold_7000434	Pseudo ARR-B	ATGCCAGAGGTTACCGGGTTTGAC	ATCGGCAAAGGTCTCTCCATGCAC
166618	fgenesh2_pg.C_scaffold_5000087	MYB-related	AGAGGAGGAGGCAGTTTTGTGTCG	AGCGTGGAGTGCCAAATGTTCTTG
168290	fgenesh2_pg.C_scaffold_4000341	Orphans	ACGACGAGGATTACGGCTATGG	GGCCTCTGCTCTATTTGGTCGAAG
168369	fgenesh2_pg.C_scaffold_2000015	ABI3VP1	AACAACAGCAGCCGCATGTACC	AACACCATTACATCGCCCACCTCC

Protein ID	Gene name	Gene family	Forward primer (5´->3´)	Reverse primer (5´->3´)
168936	fgenesh2_pg.C_scaffold_9000075	MYB-related	AGCCAGAAGGACATGGACAACC	CGTGCACGTTGGTCTTGAACAG
169967	fgenesh2_pg.C_scaffold_12000026	bZIP	GCGCATTGAGCTGATGCAGATG	TGATGATGCGAGTGCGTGTG
170309	fgenesh2_pg.C_scaffold_13000060	Jumonji	ATCAATGTGCCCATCCGGCAGAAC	CTTTGAACTCCGCCACAGACATCG
170879	fgenesh2_pg.C_scaffold_14000329	AP2-EREBP	TACCGCGGAGTCACCCGACA	TCCGTGAACTCACCCAGCAGCA
171307	fgenesh2_pg.C_scaffold_16000183	bHLH	CTGCTACTATGTGCAGGCAAGTCG	ATAAAGAAGCAGCCGACCGCAGTG
171659	fgenesh2_pg.C_scaffold_1000280	C2C2-GATA	TGTTGTTGGTTGCCGCGTTTGG	TACGAACAAATCGCGCCCACAG
171833	fgenesh2_pg.C_scaffold_1000454	SBP	ATCCTTGGAGCGAAAGCAGCAG	TCTGTCAGGTCAGCAGTGCATC
173312	fgenesh2_pg.C_scaffold_20000161	FHA	TATCTTTCGCCGACGACGACTC	GAAGCTCGCACAAAGTCCTTGC
174174	fgenesh2_pg.C_scaffold_24000249	FHA	CAACGGCACCTTTTTGAACGGC	ACGTCCTTCTCCAACAGCTCGTAG
174408	fgenesh2_pg.C_scaffold_28000008	Pseudo ARR-B	ATGTATGCGGTGGACGCGAATG	ATGCGGCGGCATCAGTGATATGAC
17550	fgenesh1_pg.C_scaffold_36000015	C2H2	TGGCATCCAGACAAGAACGCACAC	ATACCAGGCTCGCTCGTGCTTATC
177573	fgenesh2_pg.C_scaffold_43000027	RWP-RK	TGGCCGACCTGAAGCTGTACTATG	TTGAAGCACGACAGGCTGAAGC
177618	fgenesh2_pg.C_scaffold_43000072	SNF2	ATTCGGGAGCCGAACAACACCTAC	ATGTGCCCGATTTTATCGCCGC
177657	fgenesh2_pg.C_scaffold_44000001	AP2-EREBP	ATGGGAAGCCCGCAAGCGATAAAC	GTTCTTACGATTCCAGCAGACGCC
179289	fgenesh2_pg.C_scaffold_66000003	C2C2-GATA	TAGCTACGCATGATGGCAACGGTG	TCAGCGCAACCTTGGTGGTGATAG
180087	fgenesh2_pg.C_scaffold_82000073	SNF2	AACCCCAAGATGAAGCTGGTGGAG	ATCAGGTTGTTCTGGATGGGCGTG
180245	fgenesh2_pg.C_scaffold_80000049	bZIP	AACCCAGAAACGCTACGGACAG	TCGCGCTTGCTCATGTCATACG
180470	fgenesh2_pg.C_scaffold_87000015	Orphans	AGGCGCAGCATCTCAAGGTTAAGC	ATCTGGCTCACGTAGTTTCCCTCG
180819	fgenesh2_pg.C_scaffold_111000009	CCAAT	CGTGTCAGACCTGGTGCTCTGG	AGCACCTCGGCCGTGAGGTA
181226	fgenesh2_pg.C_scaffold_692000001	FHA	TCTGCCAGCAAAAGTTTGGCTCC	TTGCGACCGATTCGGAATGGAAC
183777	estExt_fgenesh2_kg.C_10285	bHLH	CCGCCGAACTCGAATCAATGAGAG	TGATGACCTCCTCCAGAAAGCACG
184359	estExt_fgenesh2_kg.C_230055	MYB-related	GCGCACAGTCCCTTCGTCTAATTC	ATCTCGGTCCAACTGTTACCCAGC
184386	estExt_fgenesh2_kg.C_280024	HMG	ATCAAGGCGCAGAACCCTGACAAG	TCCGACATCCCCTTCCAAATGGAC
186976	estExt_fgenesh2_pg.C_50434	PLATZ	GACACAGCAAGTACGCTTACCAGC	AGTAAAATAGCCACCTCCAGCCGC

Protein ID	Gene name	Gene family	Forward primer (5´->3´)	Reverse primer (5´->3´)
187360	estExt_fgenesh2_pg.C_60027	C2C2-GATA	TGTTCGGCGGCGTAAGTAACTTCG	TCATCCGACCATTGACATCCACCC
187405	estExt_fgenesh2_pg.C_60109	PHD	TGGACTCGATTACGCCACAGAACC	ACGTGCTGCTTTACATTGGGTGGG
187531	estExt_fgenesh2_pg.C_60308	bZIP	CAGGTGATTGCAAGCCTGAAGGAC	GCACTTCAATGCCTCGTCGATCTC
188131	estExt_fgenesh2_pg.C_90026	SNF2	CATCTTTGACAGCGACTGGAACCC	CCGCTCGTCACAAAGCGGTAAATG
188544	estExt_fgenesh2_pg.C_100357	SNF2	AGGCGACACTGAGTTACCTGATGC	TGGAGATGATGCGGGGTACTATGC
188627	estExt_fgenesh2_pg.C_110140	MYB-related	GGTTTTGACGCTGTGCAATGCAAC	ATGGCCCGCTCACTCTTGTGTAAC
188711	estExt_fgenesh2_pg.C_110254	Jumonji	ATGCCCCTGTACCTCTTCGACAAG	AAACTGCTCCGGTACGTGGTAGTC
189373	estExt_fgenesh2_pg.C_140328	PHD	ACACATAGCACTCATTGCCCTGCC	AGACTGCGAATGCGAAACGAGC
189471	estExt_fgenesh2_pg.C_150166	CCAAT	TCACGTCAAATGCGGAAAGCAGC	GCCACAAGCGCATCCAAGAACATC
189879	estExt_fgenesh2_pg.C_10357	HMG	ATTCCCCACCGACGACGATATCAG	GTTCTTCTCGTTCTTTGCCGCCAG
190311	estExt_fgenesh2_pg.C_180098	FHA	CCATTGAAGGTGACAAGGTGCTGG	ACTCATCACCGAACACGATCTCCC
190315	estExt_fgenesh2_pg.C_180104	Orphans	AAGCCTCGTTCTGTGCATCGTC	ACCCGGGTTAGCGGAATGAATC
190458	estExt_fgenesh2_pg.C_190061	Sigma70-like	ATCCGTCAGATTGAGGCCAAGG	AGCTTGCCGCTCACGTCAATAG
190589	estExt_fgenesh2_pg.C_190258	C2H2	ATCTAGCGAGCGGCACAACGAATC	GAAGCACCTAAGAAACGGAAGGGC
191829	estExt_fgenesh2_pg.C_250060	FHA	TTTTGACGCCGCAGAAATGGACGC	TTGGTTTTCTTTGTCCGCCCGAGC
192634	estExt_fgenesh2_pg.C_310136	C3H	CTGGATTTGTCAACGGCTGTGC	ACAAAGACAGCGTGGCCTAACG
192899	estExt_fgenesh2_pg.C_330097	HMG	AGAAGAAGTTCCAGGACTTGGCGG	TAAGCAGCTCCACACCATACGCTC
193280	estExt_fgenesh2_pg.C_400083	FHA	AGATGACGCTGTGGAAGCTTGTCC	TTGCGACCGATTCGGAATGGAAC
193299	estExt_fgenesh2_pg.C_410029	Orphans	GTCGTCATCCTGTTCTCGGACATC	TGAAGCGGTCTGTGAGCTTGTC
193672	estExt_fgenesh2_pg.C_450097	bZIP	TGGAGCTGCCAAAGCTACTAAGCC	TGACCCTTGTTGCCCATTCCATCG
193681	estExt_fgenesh2_pg.C_450107	PHD	AAAGCACCACTAGCTGCCATCCTG	TCTCGCCGTTCTAGCGACTCATAC
194168	estExt_fgenesh2_pg.C_570011	Orphans	ACCTCTTCACCAACCTGTTCACGC	AAGCCCCAGGAAGATGAAGGTCAG
194555	estExt_fgenesh2_pg.C_690037	C3H	AGTGCCAGGTGAACAACGGTGTAG	ACGCCTTTGACGAAGTCGTAGCAG
195154	estExt_fgenesh2_pg.C_940001	SET	AAGGATGTTGATCCGCACGCGAAG	AGCTCCAAAAGCTGAGCCACTGAC

Protein ID	Gene name	Gene family	Forward primer (5´->3´)	Reverse primer (5´->3´)
195786	BIS_E2F	E2F-DP/BIS_E2F	ATCCGCTGGAAGATGTCACAGGAG	GCCGAGGTCATGAAGCGAATGTTG
195787	BIS_DP	E2F-DP	ACCAGCGAGACCAAGCTGTTTCTG	GAAATCTTCACCTCCACTGTGGCG
195838	FER_fgenesh1_est.C_scaffold_27000019	RWP-RK	TTCCGTGAGGAGGTGTACGACAAC	ATTGGCCTGGCGCAGCTTCTTAAC
195840	FER_158808	RWP-RK	ACAGCAGCAGCATCAGGCGAATAG	AATGTAGCGGTAGCGACTGCGAAG
195860	FER_estExt_fgenesh2_pg.C_260253	RWP-RK	TGAACAGCCATTGTGGACCCAACC	TCGGACGCTTCCAACATTTGGTG
195891	FER_158168	RWP-RK	ACCGCCTGAGCAATGCTTTATCG	GCTCATTAGCATGTCCATGCAGCC
195917	FER_171222	RWP-RK	AGCCATAGCCATAGCTACAGCCAC	GGTGAAAAGAAAGGATGCGGCTCC
195918	FER_rwp13	RWP-RK	ACGACGAAGACTTCGGCTTTGACG	TGAAAGACCTTTCGCAACGCATCC
196759	CNA_query	SNF2	TAGCACGTCCCAGCCTGCTTTTAG	ATGCATTGTAAGCCCCGCTCCAAG
19751	fgenesh1_pg.C_scaffold_52000016	AP2-EREBP	TGCCTCCACCGTACCACAACATTG	TAGCCATAGCCATACGTCGTCGTC
22211	fgenesh1_pg.C_scaffold_94000028	C3H	AACCACCGGACAAACGAGACG	AGATCTGGCCCTGCATCGTGAC
34069	estExt_fgenesh1_pg.C_280053	FHA	CCGGTGCCTAATGCTGTTCAATCC	TGCCGGAATAAACCTCAAGGCTCG
53522	estExt_GenewiseH_1.C_70037	FHA	AACGGCCAGTACTTTGTCAGCAGC	TGGAAGCCGTACCACTGCATCTTG
6783	fgenesh1_pg.C_scaffold_7000334	SNF2	AAGTCAAACGGACCGGATGAGG	TGTTGGGCGTACTGCTGTTGTTG
6823	fgenesh1_pg.C_scaffold_7000374	SET	ACAAGCAGCAACAACAGCAGCAG	ATCACATCAGGTTCCCGATGCACC
75247	gwW.45.133.1	AP2-EREBP	CAAGTACGCGGCTGAGATCC	TAGGTTCCAAGCCACAAGCGG
8155	fgenesh1_pg.C_scaffold_6000122	DDT	ACACGCACTTCAATGCAGTGGC	AAGCAGGTACTCCACAATGCGCTC
93505	fgenesh1_est.C_scaffold_5000041	SBP	GTTGACAAGTGCAACCAGGCGATG	TATCCGAAGCCGCTGATACTGCTC
93901	fgenesh1_est.C_scaffold_2000026	Orphans	AGACCTACAAGCCCGAGGACAG	TGACCAGGAACGCCTCAATGC
94169	fgenesh1_est.C_scaffold_10000094	C3H	ATCAATGCACACTAGGGAAGCCGC	AAAGACAGTCCGTCACTCCACGTC
95196	fgenesh1_est.C_scaffold_28000028	bZIP	TGATGAGCAGCAAGGACGAGTACG	TCTGACCCGAATCGTCGGAGAAAG
24120	estExt_fgenesh1_pm.C_150006	carbonic anhydrases	CTCGGTCCAAGTCATCAACAACGG	GCGTAGTTGTAAGTCCACTGCACC

Table appendix D-2. Clusters of transcription factor and transcription regulator gene expression profiles in carbon deprivation conditions

Protein ID	Gene name (DOE/JGI genome v.3.1)	Gene family	Cluster
106059	e_gwH.39.39.1	PLATZ	1
106655	e_gwH.45.68.1	MYB	1
107982	e_gwH.71.56.1	AP2-EREBP	1
111277	e_gwW.8.20.1	SNF2	1
117914	e_gwW.22.121.1	HSF	1
118487	e_gwW.23.151.1	SET	1
126810	estExt_gwp_1H.C_30224	CSD	1
127044	estExt_gwp_1H.C_60003	C2C2-GATA	1
129649	estExt_gwp_1H.C_190094	MYB	1
13217	fgenesh1_pg.C_scaffold_18000047	SNF2	1
142793	Chlre2_kg.scaffold_6000283	SET	1
147286	Chlre2_kg.scaffold_18000211	bZIP	1
152660	Chlre2_kg.scaffold_48000084	Orphans	1
171659	fgenesh2_pg.C_scaffold_1000280	C2C2-GATA	1
171833	fgenesh2_pg.C_scaffold_1000454	SBP	1
183777	estExt_fgenesh2_kg.C_10285	bHLH	1
187360	estExt_fgenesh2_pg.C_60027	C2C2-GATA	1
188131	estExt_fgenesh2_pg.C_90026	SNF2	1
192634	estExt_fgenesh2_pg.C_310136	C3H	1
195787	BIS_DP	E2F-DP	1
196759	CNA_query	SNF2	1
93505	fgenesh1_est.C_scaffold_5000041	SBP	1
107654	e_gwH.64.49.1	C3H	2
115254	e_gwW.13.162.1	SBP	2
120035	e_gwW.31.178.1	C2C2-Dof	2
122939	e_gwW.73.33.1	PHD	2
142507	Chlre2_kg.scaffold_3000480	MYB-related	2
145251	Chlre2_kg.scaffold_13000164	HMG	2
147364	Chlre2_kg.scaffold_19000038	CCAAT	2
149109	Chlre2_kg.scaffold_28000057	MYB-related	2
149734	Chlre2_kg.scaffold_27000003	C2H2	2
153922	Chlre2_kg.scaffold_69000044	C3H	2
166515	fgenesh2_pg.C_scaffold_7000434	Pseudo ARR-B	2
174174	fgenesh2_pg.C_scaffold_24000249	FHA	2
174408	fgenesh2_pg.C_scaffold_28000008	Pseudo ARR-B	2
177657	fgenesh2_pg.C_scaffold_44000001	AP2-EREBP	2
180087	fgenesh2_pg.C_scaffold_82000073	SNF2	2
190458	estExt_fgenesh2_pg.C_190061	Sigma70-like	2
192899	estExt_fgenesh2_pg.C_330097	HMG	2
19751	fgenesh1_pg.C_scaffold_52000016	AP2-EREBP	2
6823	fgenesh1_pg.C_scaffold_7000374	SET	2
75247	gwW.45.133.1	AP2-EREBP	2
108149	e_gwH.74.33.1	SBP	3

Protein ID	Gene name (DOE/JGI genome v.3.1)	Gene family	Cluster
111791	e_gwW.7.372.1	CCAAT	3
117345	e_gwW.19.26.1	SNF2	3
146239	Chlre2_kg.scaffold_1000063	C3H	3
149159	Chlre2_kg.scaffold_28000107	Orphans	3
153832	Chlre2_kg.scaffold_68000004	SET	3
168290	fgenesh2_pg.C_scaffold_4000341	Orphans	3
171307	fgenesh2_pg.C_scaffold_16000183	bHLH	3
177618	fgenesh2_pg.C_scaffold_43000072	SNF2	3
184386	estExt_fgenesh2_kg.C_280024	HMG	3
187405	estExt_fgenesh2_pg.C_60109	PHD	3
189373	estExt_fgenesh2_pg.C_140328	PHD	3
189471	estExt_fgenesh2_pg.C_150166	CCAAT	3
190589	estExt_fgenesh2_pg.C_190258	C2H2	3
191829	estExt_fgenesh2_pg.C_250060	FHA	3
193280	estExt_fgenesh2_pg.C_400083	FHA	3
193672	estExt_fgenesh2_pg.C_450097	bZIP	3
194555	estExt_fgenesh2_pg.C_690037	C3H	3
195840	FER_158808	RWP-RK	3
195860	FER_estExt_fgenesh2_pg.C_260253	RWP-RK	3
101275	e_gwH.16.88.1	CCAAT	4
142718	Chlre2_kg.scaffold_6000208	FHA	4
160281	acegs_kg.scaffold_89000008	FHA	4
173312	fgenesh2_pg.C_scaffold_20000161	FHA	4
179289	fgenesh2_pg.C_scaffold_66000003	C2C2-GATA	4
181226	fgenesh2_pg.C_scaffold_692000001	FHA	4
186976	estExt_fgenesh2_pg.C_50434	PLATZ	4
187531	estExt_fgenesh2_pg.C_60308	bZIP	4
189879	estExt_fgenesh2_pg.C_10357	HMG	4
193681	estExt_fgenesh2_pg.C_450107	PHD	4
195838	FER_fgenesh1_est.C_scaffold_27000019	RWP-RK	4
195918	FER_rwp13	RWP-RK	4
93901	fgenesh1_est.C_scaffold_2000026	Orphans	4
95196	fgenesh1_est.C_scaffold_28000028	bZIP	4
106739	e_gwH.46.36.1	SBP	5
120259	e_gwW.35.20.1	Jumonji	5
123164	e_gwW.83.55.1	MYB-related	5
148821	Chlre2_kg.scaffold_24000220	SBP	5
151215	Chlre2_kg.scaffold_33000083	HMG	5
166327	fgenesh2_pg.C_scaffold_7000246	SBP	5
169967	fgenesh2_pg.C_scaffold_12000026	bZIP	5
170879	fgenesh2_pg.C_scaffold_14000329	AP2-EREBP	5
180245	fgenesh2_pg.C_scaffold_80000049	bZIP	5
180819	fgenesh2_pg.C_scaffold_111000009	CCAAT	5
184359	estExt_fgenesh2_kg.C_230055	MYB-related	5
195154	estExt_fgenesh2_pg.C_940001	SET	5
195891	FER_158168	RWP-RK	5

Protein ID	Gene name (DOE/JGI genome v.3.1)	Gene family	Cluster
104871	e_gwH.34.98.1	FHA	6
113031	e_gwW.6.283.1	E2F-DP	6
116834	e_gwW.1.432.1	E2F-DP	6
142476	Chlre2_kg.scaffold_3000449	C3H	6
168936	fgenesh2_pg.C_scaffold_9000075	MYB-related	6
188711	estExt_fgenesh2_pg.C_110254	Jumonji	6
194168	estExt_fgenesh2_pg.C_570011	Orphans	6
195786	BIS_E2F	E2F-DP/BIS_E2F {E2F1}	6
6783	fgenesh1_pg.C_scaffold_7000334	SNF2	6
117655	e_gwW.20.153.1	CCAAT	7
159133	acegs_kg.scaffold_38000029	C2C2-CO-like	7
166618	fgenesh2_pg.C_scaffold_5000087	MYB-related	7
190315	estExt_fgenesh2_pg.C_180104	Orphans	7
193299	estExt_fgenesh2_pg.C_410029	Orphans	7
195917	FER_171222	RWP-RK	7
8155	fgenesh1_pg.C_scaffold_6000122	DDT	7
112628	e_gwW.3.655.1	MYB-related	8
130971	estExt_gwp_1H.C_300115	G2-like	8
145450	Chlre2_kg.scaffold_14000083	SNF2	8
153371	Chlre2_kg.scaffold_60000060	SET	8
170309	fgenesh2_pg.C_scaffold_13000060	Jumonji	8
177573	fgenesh2_pg.C_scaffold_43000027	RWP-RK	8
190311	estExt_fgenesh2_pg.C_180098	FHA	8
105687	e_gwH.38.105.1	C3H	9
119283	e_gwW.27.121.1	MYB	9
121606	e_gwW.50.58.1	SBP	9
142283	Chlre2_kg.scaffold_3000256	HMG	9
22211	fgenesh1_pg.C_scaffold_94000028	C3H	9
108444	e_gwH.81.18.1	SBP	10
122706	e_gwW.71.31.1	HSF	10
126346	estExt_gwp_1H.C_70177	Orphans	10
17550	fgenesh1_pg.C_scaffold_36000015	C2H2	10
135484	estExt_gwp_1W.C_110128	SNF2	Unassigned
135809	estExt_gwp_1W.C_130065	CCAAT	Unassigned
168369	fgenesh2_pg.C_scaffold_2000015	ABI3VP1	Unassigned
180470	fgenesh2_pg.C_scaffold_87000015	Orphans	Unassigned
188544	estExt_fgenesh2_pg.C_100357	SNF2	Unassigned
188627	estExt_fgenesh2_pg.C_110140	MYB-related	Unassigned
34069	estExt_fgenesh1_pg.C_280053	FHA	Unassigned
53522	estExt_GenewiseH_1.C_70037	FHA	Unassigned
94169	fgenesh1_est.C_scaffold_10000094	C3H	Unassigned

Table appendix D-3. *C. reinhardtii* transcription factors and transcription regulators regulated (FC > 2) after 180min under carbon deprivation conditions

Protein ID	Gene name (DOE/JGI genome annotation v.3.1)	TF family	Cluster	FC 60min	FC 120min	FC 180min
195840	FER_158808	RWP-RK	3	0.643151	0.346193	0.166138
186976	estExt_fgenesh2_pg.C_50434	PLATZ	4	0.389047	0.275548	0.167957
113031	e_gwW.6.283.1	E2F-DP	6	1.260075	2.144672	0.168229
195838	FER_fgenesh1_est.C_scaffold_27000019	RWP-RK	4	0.246324	0.385934	0.169866
95196	fgenesh1_est.C_scaffold_28000028	bZIP	4	0.28552	0.704606	0.211916
149159	Chlre2_kg.scaffold_28000107	Orphans	3	0.724528	0.268482	0.282718
108149	e_gwH.74.33.1	SBP	3	1.057589	0.636253	0.285618
195918	FER_rwp13	RWP-RK	4	0.544746	0.809348	0.341931
189879	estExt_fgenesh2_pg.C_10357	HMG	4	0.631136	0.632089	0.343505
101275	e_gwH.16.88.1	CCAAT	4	0.585307	0.786567	0.344606
195860	FER_estExt_fgenesh2_pg.C_260253	RWP-RK	3	2.07106	0.547271	0.350195
168290	fgenesh2_pg.C_scaffold_4000341	Orphans	3	0.902705	0.828699	0.370488
194555	estExt_fgenesh2_pg.C_690037	C3H	3	0.904651	0.837677	0.385147
93505	fgenesh1_est.C_scaffold_5000041	SBP	1	1.557178	1.060538	0.393198
160281	acegs_kg.scaffold_89000008	FHA	4	0.907589	0.916513	0.411621
192634	estExt_fgenesh2_pg.C_310136	C3H	1	1.357989	1.150915	0.413509
147286	Chlre2_kg.scaffold_18000211	bZIP	1	2.502103	1.022082	0.430682
193681	estExt_fgenesh2_pg.C_450107	PHD	4	0.813236	0.784364	0.441383
189471	estExt_fgenesh2_pg.C_150166	CCAAT	3	1.256141	0.972309	0.447188
111277	e_gwW.8.20.1	SNF2	1	3.908052	1.723852	0.447869

Protein ID	Gene name (DOE/JGI genome annotation v.3.1)	TF family	Cluster	FC 60min	FC 120min	FC 180min
179289	fgenesh2_pg.C_scaffold_66000003	C2C2-GATA	4	0.6461	1.08604	0.449293
189373	estExt_fgenesh2_pg.C_140328	PHD	3	1.204991	0.727616	0.452294
194168	estExt_fgenesh2_pg.C_570011	Orphans	6	0.858053	6.334243	0.455064
187531	estExt_fgenesh2_pg.C_60308	bZIP	4	0.5865	1.176443	0.463547
152660	Chlre2_kg.scaffold_48000084	Orphans	1	1.127354	1.023611	0.477008
177618	fgenesh2_pg.C_scaffold_43000072	SNF2	3	1.685404	0.426737	0.4771
53522	estExt_GenewiseH_1.C_70037	FHA	Unassigned	2.678312	2.41764	2.033292
151215	Chlre2_kg.scaffold_33000083	HMG	5	0.938635	1.611377	2.059172
120035	e_gwW.31.178.1	C2C2-Dof	2	0.763815	0.830049	2.088145
147364	Chlre2_kg.scaffold_19000038	CCAAT	2	0.834642	0.716048	2.112174
180819	fgenesh2_pg.C_scaffold_111000009	CCAAT	5	0.423561	1.042314	2.392005
195917	FER_171222	RWP-RK	7	2.663892	1.392326	2.480409
148821	Chlre2_kg.scaffold_24000220	SBP	5	0.712728	1.02503	2.575971
192899	estExt_fgenesh2_pg.C_330097	HMG	2	0.823451	0.630828	2.651903
195891	FER_158168	RWP-RK	5	0.979645	1.903212	2.677956
168369	fgenesh2_pg.C_scaffold_2000015	ABI3VP1	Unassigned	4.597668	7.593528	2.68601
126810	estExt_gwp_1H.C_30224	CSD	1	6.858036	11.12799	3.181677
188544	estExt_fgenesh2_pg.C_100357	SNF2	Unassigned	1.021668	1.966845	3.526228
188627	estExt_fgenesh2_pg.C_110140	MYB-related	Unassigned	1.274458	2.31237	9.118325
184359	estExt_fgenesh2_kg.C_230055	MYB-related	5	0.627714	7.168026	15.14901

Table appendix D-4. FAIRE peaks found for the TF and TR genes analyzed under carbon deprivation

			Genomic coordinates (DOE/ JGI genome annotation v.4)						
Protein ID (DOE/ JGI genome annotation v.3.1)	**Expression level (FC) 3h under Low-CO_2**	**Peak name**	**Chromosome**	**Peak start**	**Peak end**	**Peak length**	**Score (-10*LOG10(pvalue))**	**Fold enrichment**	**Peak summit**
104871	0.69401455135	MACS_peak_3848	chromosome_3	610767	611705	939	137.66	10.38	607
126810	3.1816773588	MACS_peak_4726	chromosome_6	2429964	2430753	790	58.72	7.48	363
145450	1.9580680185	MACS_peak_4095	chromosome_3	4794908	4796574	1667	254.36	8.68	909
148821	2.5759705651	MACS_peak_3590	chromosome_2	6650167	6651217	1051	88.91	6.92	181
148821	2.5759705651	MACS_peak_3591	chromosome_2	6655557	6656727	1171	70.25	5.89	701
153371	1.2459582563	MACS_peak_3149	chromosome_17	6376236	6376819	584	71.7	7.23	363
153371	1.2459582563	MACS_peak_3150	chromosome_17	6379034	6379676	643	55.09	5.34	385
17550	NA	MACS_peak_4506	chromosome_5	2018153	2018856	704	58.7	6.68	386
183777	1.2483883878	MACS_peak_2109	chromosome_14	1896227	1896655	429	59.51	6.03	149
183777	1.2483883878	MACS_peak_2110	chromosome_14	1897510	1898345	836	55.53	4.36	386
184359	15.149008985	MACS_peak_3696	chromosome_2	8323928	8324887	960	115.95	6.74	471
184386	0.78060583537	MACS_peak_2569	chromosome_16	3417744	3418551	808	59.2	5.07	498
190458	1.2788823095	MACS_peak_4132	chromosome_3	5474845	5475473	629	63.76	6.8	194
194168	0.45506384264	MACS_peak_2240	chromosome_15	178280	178927	648	65.06	7.13	455
194555	0.38514740986	MACS_peak_4905	chromosome_6	5312891	5313717	827	52.9	3.84	529
195840	0.16613764327	MACS_peak_4934	chromosome_6	5805941	5806433	493	68.42	6.68	250
195840	0.16613764327	MACS_peak_4935	chromosome_6	5815166	5816052	887	53.66	7.08	358
195860	0.35019519121	MACS_peak_4884	chromosome_6	4971627	4972178	552	123.93	9.33	369
195891	2.6779563109	MACS_peak_3813	chromosome_3	201643	202283	641	125.62	8.5	219
195891	2.6779563109	MACS_peak_3814	chromosome_3	210670	212073	1404	141.11	6.71	1009
93505	0.39319779781	MACS_peak_3473	chromosome_2	4821207	4822768	1562	147.03	8.25	537
95196	0.21191620003	MACS_peak_2595	chromosome_16	3912993	3913924	932	94.1	7.29	411

Table appendix D-5. Motifs found in the 5´upstream regions of the TF and TR genes which presented FAIRE peaks

Protein ID (JGI v.3.1)	TranscriptID (JGI v.4)	MotifID	Motif p-value	Left_flank	Site	Right_flank	Gene model name	TF family
104871	104871	motif_1	9.67E-08	CCGCGCGTGC	CGCAGCTGCTGCGGC	GGCTGGCGCT	e_gwH.34.98.1	FHA
104871	104871	motif_3	5.94E-06	TCGCGCACAG	ACCATACAAACACA	CCAAATGTAT	e_gwH.34.98.1	FHA
104871	104871	motif_4	1.14E-10	TGGGTTTGGT	TCAGTTTGGGCTGGTATCTA	CACCCCCCCC	e_gwH.34.98.1	FHA
104871	104871	motif_5	1.15E-06	GATGGATGCT	TTTTCTCTTG	ATGGCAGTAT	e_gwH.34.98.1	FHA
104871	104871	motif_6	2.54E-05	TATTCGCCGG	GTTTGGTGAT	CGCGCACAGA	e_gwH.34.98.1	FHA
104871	104871	motif_7	6.35E-10	CGGTGGTGGT	GGCGGTGGTGGTGGTGGTG	GACAGGGCGG	e_gwH.34.98.1	FHA
104871	104871	motif_8	1.04E-07	ATGTGTGGCT	CTCCTTAGCCATGAAACAT	GACTTTACAA	e_gwH.34.98.1	FHA
104871	104871	motif_9	2.62E-09	TGGGCCCGCA	GGCGACGGGGCCGCAGGCGC	CGGCCCCCCC	e_gwH.34.98.1	FHA
126810	126810	motif_1	2.01E-05	TGGCCTTTCG	AGCATTTGCAGCCGG	GCAGGGCTCT	estExt_gwp_1H.C_30224	CSD
126810	126810	motif_2	1.69E-07	CACAAATATT	TATTAATAAATT	AAAAGCAAAT	estExt_gwp_1H.C_30224	CSD
126810	126810	motif_3	5.04E-08	ATAAATTAAA	AGCAAATAAAGACA	CAGAGTCAGC	estExt_gwp_1H.C_30224	CSD
126810	126810	motif_4	1.20E-07	ACCACTTGAA	CGTCTTTGACGTGGTATAAA	TGCATCTGTC	estExt_gwp_1H.C_30224	CSD
126810	126810	motif_6	1.20E-05	CACACCAGCT	GTTTGCTCAT	GCCTTGCCGT	estExt_gwp_1H.C_30224	CSD
126810	126810	motif_7	3.25E-05	GTGTTTCTTG	GCCGTTTGCGCAGGTCCTT	GTGGCACCCA	estExt_gwp_1H.C_30224	CSD
145450	206676	motif_1	6.42E-06	CGGCACCGGC	ATCAGTTGCCGCTGC	ATCGGCATGC	Chlre2_kg.scaffold_14000083	SNF2
145450	206676	motif_2	1.69E-07	TGTATGCTTA	AATAAATTATGT	CAGGCTCTTT	Chlre2_kg.scaffold_14000083	SNF2
145450	206676	motif_3	4.49E-06	CGGCCAAGTA	ATCGAGCGAATACA	CCCGAAAGCG	Chlre2_kg.scaffold_14000083	SNF2
145450	206676	motif_4	6.68E-09	GCGCGGGACT	TCAATTTGGTGTATTATATT	ATCAAATTGT	Chlre2_kg.scaffold_14000083	SNF2
145450	206676	motif_6	1.57E-06	GGCCCCCTTA	ATTTGAGGAT	TCGATATCAT	Chlre2_kg.scaffold_14000083	SNF2
145450	206676	motif_7	1.83E-07	GGGGCCTGGT	GGCAGTGCTGGCGGTGGTG	CGGCGGGCAG	Chlre2_kg.scaffold_14000083	SNF2
145450	206676	motif_9	1.23E-06	TCGGCGCATC	GGCGACCCAGCCATCAGTGC	GGGTTCGTCG	Chlre2_kg.scaffold_14000083	SNF2
148821	195928	motif_1	9.21E-07	CTCGCCACAG	AGGCTCTGCTGGTGC	GGCTGACTTA	Chlre2_kg.scaffold_24000220	SBP
148821	195928	motif_2	4.93E-08	TGTTTCACAT	AAATAATAATGT	CCTGTTCCCA	Chlre2_kg.scaffold_24000220	SBP
148821	195928	motif_3	1.82E-07	ACCATCTTCT	AGCAAGCAAATAAA	GGCAACGCTG	Chlre2_kg.scaffold_24000220	SBP
148821	195928	motif_5	1.15E-06	CCCCCGTTTC	TTTCCTCTTG	CACTCGACCG	Chlre2_kg.scaffold_24000220	SBP
148821	195928	motif_6	1.81E-05	CAGATTGACT	ATTTGTGTAT	GCCGTCGCTG	Chlre2_kg.scaffold_24000220	SBP
148821	195928	motif_7	4.57E-07	TGTGTATGCC	GTCGCTGGGCATGCTGCTG	ACAGAAACTG	Chlre2_kg.scaffold_24000220	SBP

Protein ID (JGI v.3.1)	TranscriptID (JGI v.4)	MotifID	Motif p-value	Left_flank	Site	Right_flank	Gene model name	TF family
148821	195928	motif_9	1.45E-06	ACAAAGCGTC	GCCGGCCGAGCTGGCCGCGC	AAACCATCTT	Chlre2_kg.scaffold_24000220	SBP
153371	153371	motif_1	3.85E-08	GCCCGCGCCG	AGCAGCTGCTGCGGC	CGCACACGAC	Chlre2_kg.scaffold_60000060	SET
153371	153371	motif_2	1.07E-08	TCTGACAACA	AATAAATAATGT	TGCGAGAGCA	Chlre2_kg.scaffold_60000060	SET
153371	153371	motif_6	1.03E-05	TGGCTGCAGC	ATTTGACGAT	GATGAGGACG	Chlre2_kg.scaffold_60000060	SET
153371	153371	motif_7	6.53E-06	GCGGGGGCGG	GGCAGAGGCCGTGGTCGGG	GGCGGGGCCG	Chlre2_kg.scaffold_60000060	SET
153371	153371	motif_8	8.36E-08	GGAGCTTAAT	CTCACCACCCTTGTGCAGT	CGCTCAAAAA	Chlre2_kg.scaffold_60000060	SET
153371	153371	motif_9	1.42E-07	CGCAGCGACA	GGCTGCGCAGCATCAGGCGC	AACAGGCATC	Chlre2_kg.scaffold_60000060	SET
17550	17550	motif_1	8.78E-06	CCTGCGATAG	CGGCTCCGCTGCCGC	GAACAGTAGG	fgenesh1_pg.C_scaffold_36000015	C2H2
17550	17550	motif_10	3.02E-12	CCCCTCCCAG	CAACAAGTTATCTTTATTA	AAGCAGCTTT	fgenesh1_pg.C_scaffold_36000015	C2H2
17550	17550	motif_2	9.75E-07	TCACAACGTT	AAAAAATGATTT	GGACAGATAA	fgenesh1_pg.C_scaffold_36000015	C2H2
17550	17550	motif_3	2.74E-06	AGTTCGCACG	GGCAAGTGAAGAAA	TGCTGAACAC	fgenesh1_pg.C_scaffold_36000015	C2H2
17550	17550	motif_4	4.19E-11	TTGGTTTAGT	TCAGTTTGGGCTGGTAATTA	CAATCCACCC	fgenesh1_pg.C_scaffold_36000015	C2H2
17550	17550	motif_6	5.12E-06	TCGACCCCCT	ATTTGCCCAT	GCCATCCCCC	fgenesh1_pg.C_scaffold_36000015	C2H2
17550	17550	motif_7	2.94E-07	GCGAGCCGGG	GGCAGAGGCGGTGGTGGAG	GAGGCATCTT	fgenesh1_pg.C_scaffold_36000015	C2H2
17550	17550	motif_8	1.11E-11	AATCATCGCT	CTCATCAGCCTTGTACCAT	TATCCCCCAG	fgenesh1_pg.C_scaffold_36000015	C2H2
17550	17550	motif_9	1.90E-07	ACAGTAGGTA	GGCGGTGGCGGCGGCGGCGC	TGCCCTGCTT	fgenesh1_pg.C_scaffold_36000015	C2H2
183777	183777	motif_1	4.74E-06	GTCAGCTCCG	CCCAACTGCTGCTGC	GTTTATGTTC	estExt_fgenesh2_kg.C_10285	bHLH
183777	183777	motif_2	3.02E-07	GATGCATAAA	AATTAAGAATGT	TTATGCATTG	estExt_fgenesh2_kg.C_10285	bHLH
183777	183777	motif_3	5.94E-06	GGAAAATTTT	GCAAAATTAAGATA	GCGTACGCGT	estExt_fgenesh2_kg.C_10285	bHLH
183777	183777	motif_5	1.87E-06	CACACGCCCG	TTTCCTGTTG	TATCTTTTCC	estExt_fgenesh2_kg.C_10285	bHLH
183777	183777	motif_6	4.68E-07	TCATCCCATA	ATTTGCGCAT	TTTGCCCACT	estExt_fgenesh2_kg.C_10285	bHLH
183777	183777	motif_7	2.77E-06	GCACTGCAAC	GTGGCTGCTGGTGGTGGGG	GACAAGGCAG	estExt_fgenesh2_kg.C_10285	bHLH
184359	205561	motif_1	2.91E-06	GGTGCTGGTA	CACAGCTGCTGGCGA	GCCCCCGGGT	estExt_fgenesh2_kg.C_230055	MYB-related
184359	205561	motif_4	1.35E-08	TCCGTCCGAC	TGAGTTTTGGCAGGTTTTTA	ATTGTGAGCC	estExt_fgenesh2_kg.C_230055	MYB-related
184359	205561	motif_7	4.57E-07	GTCAGCACAA	GGCGCTGTTGAAGGTGCAT	ACAGGTGCTT	estExt_fgenesh2_kg.C_230055	MYB-related
184359	205561	motif_8	9.00E-08	GCTGAATACC	CTCAGCAGCTGTAAATCAT	TTGCACCAGC	estExt_fgenesh2_kg.C_230055	MYB-related
184359	205561	motif_9	1.90E-07	GCTCCCTTGA	GGCGGCGGCGCTATCCACGC	ATGCTTTTTT	estExt_fgenesh2_kg.C_230055	MYB-related
184386	184386	motif_1	3.20E-05	TTCCGGCGCG	TTGCGGTGCTGGTGC	TCCGTCGCGA	estExt_fgenesh2_kg.C_280024	HMG
184386	184386	motif_2	1.60E-06	CAGGTAGCCG	TTTTAATAAACT	GGGTGCAGCC	estExt_fgenesh2_kg.C_280024	HMG

Protein ID (JGI v.3.1)	TranscriptID (JGI v.4)	MotifID	Motif p-value	Left_flank	Site	Right_flank	Gene model name	TF family
184386	184386	motif_6	6.73E-05	TCCCGTTTGA	ATTTGCCGGT	CTGAAATCTT	estExt_fgenesh2_kg.C_280024	HMG
184386	184386	motif_7	1.37E-06	AAAAGGTGCA	GCCACACGGGAAGCTGCTG	GCGTCGACCC	estExt_fgenesh2_kg.C_280024	HMG
184386	184386	motif_8	2.66E-08	CCCCGGTGGG	CTCTTCATGTTTGTGACGT	CCCAGGCGAC	estExt_fgenesh2_kg.C_280024	HMG
190458	190458	motif_1	1.30E-05	GGCCAAGACA	CGCCGTTGCTGCCAA	CGGTCGGGAG	estExt_fgenesh2_pg.C_190061	Sigma70-like
190458	190458	motif_6	1.26E-06	AGTGTTGAGC	ATTTGAGCAT	TTCCCGCAGT	estExt_fgenesh2_pg.C_190061	Sigma70-like
190458	190458	motif_7	5.63E-06	ACGTCCGAGG	AGAACTGCTGGAGCTACTG	TAGGGATCAG	estExt_fgenesh2_pg.C_190061	Sigma70-like
190458	190458	motif_8	2.05E-07	CCTTTAGTTG	CGTATCAGTCGTGGGCAAT	TGTAGAAGAC	estExt_fgenesh2_pg.C_190061	Sigma70-like
190458	190458	motif_9	1.45E-06	GGCCTCTGCC	GGCTAGGTGGGCACACACGC	TTCTGACACG	estExt_fgenesh2_pg.C_190061	Sigma70-like
194168	194168	motif_1	2.96E-08	AACAATCCCC	CGCCGCTGCTGCTGC	TGCCGCTGCT	estExt_fgenesh2_pg.C_570011	Orphans
194168	194168	motif_3	3.34E-07	ATCTCAGTTG	AGGAAATTAAAATA	AGGCCCTGGA	estExt_fgenesh2_pg.C_570011	Orphans
194168	194168	motif_6	1.50E-05	CTGCTCAAGG	GTTTGGGCAT	CTCAAACTCG	estExt_fgenesh2_pg.C_570011	Orphans
194168	194168	motif_7	2.07E-07	TGCTGCTGCT	GCCGCTGCTGCTGCTGCTG	CTGCTGCTGC	estExt_fgenesh2_pg.C_570011	Orphans
194168	194168	motif_8	1.25E-08	GAGGAGAGAT	GTCTTCAGGCGTGTGTCAT	CCGCTACTGC	estExt_fgenesh2_pg.C_570011	Orphans
194168	194168	motif_9	7.45E-07	TGCTGCTGCT	GCCGCCGGCGGCGCAGGTGC	GGGCCGTGGA	estExt_fgenesh2_pg.C_570011	Orphans
194555	194555	motif_1	1.59E-06	GTGTGGCGGC	CGCCGCCGCTGCCGC	CGAGCGCTCC	estExt_fgenesh2_pg.C_690037	C3H
194555	194555	motif_3	9.12E-08	TATAACGCGC	AACAAGCTAAAACA	GTATAGTCTG	estExt_fgenesh2_pg.C_690037	C3H
194555	194555	motif_4	3.60E-07	TCGATTGACA	CCACTTTGGTATAAAATGGA	GGGGTTTTAA	estExt_fgenesh2_pg.C_690037	C3H
194555	194555	motif_6	2.48E-06	CGATGCAATT	ATTTGCGAAT	CTTCCTTCCT	estExt_fgenesh2_pg.C_690037	C3H
194555	194555	motif_7	1.29E-05	GTCCATGTCG	GGGACTCGGGGTGGGCGTG	GCTGGTGGCC	estExt_fgenesh2_pg.C_690037	C3H
195840	195840	motif_1	1.04E-06	AGGACGCGGA	TGCTGCTGCTGCTGC	TATTACTCAT	FER_158808	RWP-RK
195840	195840	motif_7	4.27E-09	GGCAGCCAGC	GGCACTGGTGTTGGTGGTG	GTGTTGGTGG	FER_158808	RWP-RK
195840	195840	motif_9	8.93E-10	CCATCGGCCT	GGCGGCGCCGCCGCACGCGC	TGCTCCCGCA	FER_158808	RWP-RK
195860	195860	motif_1	1.39E-05	GTTCGATAGC	AGGAGCTGCTCGTGG	CATCCGCCTG	FER_estExt_fgenesh2_pg.C_260253	RWP-RK
195860	195860	motif_2	5.16E-07	GCCCGATTGA	TATTAGTAAAGT	GTTACTCTTG	FER_estExt_fgenesh2_pg.C_260253	RWP-RK
195860	195860	motif_3	1.05E-06	GACCTGACTT	AGCAAGCGAAAATT	AGGTTAAGCT	FER_estExt_fgenesh2_pg.C_260253	RWP-RK
195860	195860	motif_4	1.11E-08	GTGTCGTCTT	TGACTTTGAACTGTATTTAA	GGTTAATGTG	FER_estExt_fgenesh2_pg.C_260253	RWP-RK
195860	195860	motif_6	2.42E-05	CAGGCGCAGA	GTTTGTTGAT	TACACAGTTG	FER_estExt_fgenesh2_pg.C_260253	RWP-RK
195860	195860	motif_7	1.65E-06	CACGTGTGAG	GCAGGTGGGCAAGGTGCAG	TGCTGCCTGG	FER_estExt_fgenesh2_pg.C_260253	RWP-RK
195860	195860	motif_8	2.74E-09	GTGATGATCA	CTGATCAGGCGTGAGCCAT	GATCGGATGC	FER_estExt_fgenesh2_pg.C_260253	RWP-RK

Protein ID (JGI v.3.1)	TranscriptID (JGI v.4)	MotifID	Motif p-value	Left_flank	Site	Right_flank	Gene model name	TF family
195860	195860	motif_9	4.58E-08	CGCCTTCGTG	GGCGGGGCAGGCACAGGCGC	AGAGTTTGTT	FER_estExt_fgenesh2_pg.C_260253	RWP-RK
195891	195891	motif_1	6.85E-07	ACAGTTGGCC	AAGCGCTGCTGCCGC	TCGCCCGCAA	FER_158168	RWP-RK
195891	195891	motif_10	2.17E-12	CGCAGCAGCT	CAAGAATTTATGATTATAA	TTCTAAGATA	FER_158168	RWP-RK
195891	195891	motif_2	2.02E-06	GCGAATTCGT	ATTAAGTGATGT	TGAACTCGAA	FER_158168	RWP-RK
195891	195891	motif_3	3.34E-07	TAATTCTAAG	ATAAATCTAAAACA	AATCCCAGCG	FER_158168	RWP-RK
195891	195891	motif_4	3.23E-07	CGAAGAAAAG	TTAGCTTGAGCTGATAAACT	GAGGAGGCCA	FER_158168	RWP-RK
195891	195891	motif_6	4.72E-05	CAACCATGAC	GTATGCGCAT	GCAACGGCAT	FER_158168	RWP-RK
195891	195891	motif_7	1.46E-05	TGCCGAAAGA	GGAACTGGAGAATGTCGAC	GTAACTGGCC	FER_158168	RWP-RK
195891	195891	motif_9	1.61E-06	CAGCCTGTCG	GGCTTCGCGTCAGCAGGCGC	CGGCCAGCAG	FER_158168	RWP-RK
93505	93505	motif_1	1.69E-07	GTGTACAGCA	AGCAGCTGCTGGGGC	CGGGCACAGC	fgenesh1_est.C_scaffold_5000041	SBP
93505	93505	motif_3	3.16E-07	AGGTGCGTTG	AGCAATCCAAAACA	GGCTGGGCTG	fgenesh1_est.C_scaffold_5000041	SBP
93505	93505	motif_5	4.51E-07	TAAACACTCC	TTTCCTTTTG	TCTCCGCAGG	fgenesh1_est.C_scaffold_5000041	SBP
93505	93505	motif_6	7.92E-05	AGCTCGCGTT	ACTTGGGAAT	GCGTTGCGTG	fgenesh1_est.C_scaffold_5000041	SBP
93505	93505	motif_7	2.34E-06	CGTGCGGCGT	GGCGCTGGGGCATGTCCTC	TTTGGCACGG	fgenesh1_est.C_scaffold_5000041	SBP
93505	93505	motif_8	4.11E-08	GTCCTGCAGC	CGCATCATCCGTGTGCGGT	TCGATGAGGT	fgenesh1_est.C_scaffold_5000041	SBP
93505	93505	motif_9	1.58E-07	CGTCGCACCC	GGCTACGGCGCCAGCCATGC	CGGGAACGGC	fgenesh1_est.C_scaffold_5000041	SBP
95196	95196	motif_1	4.79E-09	CCCACGCTGG	AGCAGCTGCTGCCGC	AGTCGGTGGA	fgenesh1_est.C_scaffold_28000028	bZIP
95196	95196	motif_3	4.77E-09	CAGGGGGGAC	AGCAAGTAAAAACA	TCAACGGTGG	fgenesh1_est.C_scaffold_28000028	bZIP
95196	95196	motif_4	6.26E-08	CGGTCTGGGC	TGAAATTGGGATTGGATTAA	ACTACAAACT	fgenesh1_est.C_scaffold_28000028	bZIP
95196	95196	motif_5	4.51E-07	AACTTCCTCT	TTTTCTTTTG	TGCTCTTGCA	fgenesh1_est.C_scaffold_28000028	bZIP
95196	95196	motif_7	1.03E-06	TGGACGCGGT	GGCGGACCTCGAGGTGCTG	GGTGAGTGCG	fgenesh1_est.C_scaffold_28000028	bZIP
95196	95196	motif_9	1.35E-08	TGACGCCGAT	GGCGGCGGCGGCGGCGGCGC	CGGCTCCGAC	fgenesh1_est.C_scaffold_28000028	bZIP

Table appendix D-6. Motifs found in the promoter regions of the two co-expressed MYB TFs

Protein ID	MotifID	Motif p-value	Left flank	Motif site	Righ flank	Motif regular expression	MAST position p-value
188627	1	2.24E-12	AAATCGATTC	AAGTGCATAAAAAGCATACA	TGTATGCACT	AA[AG]T[GT][CG]A[GT]AAA[AC]AGCAT[AG]CA	4.8 e-11
184359	1	6.17E-12	GCGTCAACTG	AAATTGAGAAACAGCATGCA	CACACGCTTT	AA[AG]T[GT][CG]A[GT]AAA[AC]AGCAT[AG]CA	8.7 e-11
184359	2	8.20E-07	AGCATACACC	AAAATCTAT	TCGCCTTGAA	AAAATC[GT]AT	> 0.0001
188627	2	2.01E-06	TGACAAAGAG	AAAATCGAT	TCAAGTGCAT	AAAATC[GT]AT	> 0.0001
188627	3	2.29E-08	GGCCGCATGT	TCATTTGCATCA	TCGCGCTCAT	TCATTTGCA[CT]CA	8.2 e-08
184359	3	5.29E-08	CAGCTGTAAA	TCATTTGCACCA	GCATACACCA	TCATTTGCA[CT]CA	1.5 e-07
188627	4	9.05E-12	GGCCTTGGAC	ATAACTCGTTAAAACCCAAA	GGCCGCCCCA	AT[AG]AC[AT][AC][GT][GT]TA[AC]AACC[AC]A[AC]A	2.3 e-10
184359	4	1.77E-11	AGTTTCGTGC	ATGACAATGTACAACCAACA	GGCGGGAAGG	AT[AG]AC[AT][AC][GT][GT]TA[AC]AACC[AC]A[AC]A	3.3 e-10
184359	5	4.26E-09	AGGACATTGA	ATTACCAAGGAGAA	TGCCAATAAT	ATTACCA[AT]G[GT]AG[AC]A	7.2 e-05
188627	5	6.92E-09	CGTGGCGTTC	ATTACCATGTAGCA	GCCGCCGAAA	ATTACCA[AT]G[GT]AG[AC]A	3.3 e-08
184359	6	1.24E-06	CACACACGCT	TTCATAACT	TGGGTTACCG	TTCA[CT]AACT	> 0.0001
188627	6	2.87E-06	CCCTGCTTCG	TTCACAACT	CCACATCCGC	TTCA[CT]AACT	> 0.0001
184359	7	1.81E-05	ACTTCCGCCT	AATACAA	TCCGCCTCCC	AA[GT]ACAA	> 0.0001
188627	7	4.44E-05	GCGCCAGTTG	AAGACAA	CTTTGCTAAA	AA[GT]ACAA	> 0.0001
184359	8	1.58E-08	GTAATATTTG	CAGTAATAGGCAC	AATCACTTAT	CAG[GT]AATAGG[CT]AC	3.1 e-05
188627	8	2.57E-08	CCATCCCTGC	CAGGAATAGGTAC	GCGGGCTGCC	CAG[GT]AATAGG[CT]AC	4.9 e -08
184359	9	2.63E-06	TCGTCCGCTA	TGAAGTCCA	TCCTTCATCG	TGAAGTCCA	> 0.0001
188627	9	2.63E-06	CTGATACTTC	TGAAGTCCA	CTTCTACTCA	TGAAGTCCA	> 0.0001
184359	10	3.05E-05	GTGTAAGGGC	CATTGAA	CGACTGCATT	CATTGAA	> 0.0001
188627	10	3.05E-05	GGAACGCCGT	CATTGAA	GCGAACACAT	CATTGAA	> 0.0001

Publications

Manuscripts published or submitted

Winck F.V., Riaño-Pachón D.M., Sommer F., Rupprecht J., Mueller-Roeber B. (2011) The nuclear proteome of the green alga *Chlamydomonas reinhardtii. Proteomics*, in revision.

Authors contributions: FVW and BMR conceived and designed the study. FVW designed the experiments. FVW and JR prepared the cell cultures and established experimental conditions. FVW performed the isolation of nuclei and preparation of protein extracts. FVW and FS performed the preparation of peptides and performed analysis on LC/MS/MS. FVW and DMRP performed the protein identification and build the website. FVW, BMR and DMRP wrote the manuscript. All authors contributed to final corrections of the manuscript.

Winck, F.V., Kwasniewski, M., Wienkoop, S., Mueller-Roeber, B. (2011) An optimized method for the isolation of nuclei from *Chlamydomonas reinhardtii* (CHLOROPHYCEAE), *J. Phycol.,* 47, 333–340.

Authors contributions: BMR and FVW conceived and designed the study. FVW and MK designed the experiments. FVW performed the experiments of isolation of nuclei, microscopic analysis, protein separation and immunochemical detections. FVW and SW performed the LC/MS/MS analysis. FVW performed the MALDI-ToF analysis. FVW performed the protein identification. BMR and FVW wrote the manuscript.

Martins-de-Souza D, Oliveira BM, Castro-Dias E, **Winck FV**, Horiuchi RS, Baldasso PA, Caetano HT, Pires NK, Marangoni S, Novello JC (2008) The untiring search for the most complete proteome representation: reviewing the methods, *Briefings in functional genomics & proteomics*, **7**, 312-321.

Authors contributions: DMS, BMO, FVW, SM and JCN participated in the design of the review and final corrections of the manuscript. DMS and FVW wrote the conclusions. FVW wrote the sessions about low expressed proteins, subcellular proteome and

biofluid analysis. BMO and ECD wrote the session about narrow range and non-equilibrium pH electrophoresis. ECD and HRS wrote the session of proteome pre-fractionation. PAB, HTC and NKP wrote the session fluorimetric detection methods.

Manuscripts in preparation

Winck FV, Arvidsson S, Riaño-Pachón DM, Koseska A, Donner S, Rupprecht J, Mueller-Roeber B. Deciphering gene regulatory network of carbon concentrating mechanism in *Chlamydomonas reinhardtii*.

Previous publications

Vilca-Quispe A, Ponce-Soto LA, **Winck FV**, Marangoni S. Isolation and characterization of a new serine protease with thrombin-like activity (TLBm) from the venom of the snake *Bothrops marajoensis*. *Toxicon*. 2010 Apr 1; 55(4):745-53. PMID: 19931298.

Lira CB, Siqueira Neto JL, Giardini MA, **Winck FV**, Ramos CH, Cano MI. LaRbp38: a Leishmania amazonensis protein that binds nuclear and kinetoplast DNAs. *Biochem Biophys Res Commun*. 2007 Jul 6; 358(3):854-60. PMID: 17506988.

Silva JA, Damico DC, Baldasso PA, Mattioli MA, **Winck FV**, Fraceto LF, Novello JC, Marangoni S. Isolation and biochemical characterization of a galactoside binding lectin from *Bauhinia variegata candida* (BvcL) seeds. *Protein J*. 2007 Apr; 26(3):193-201. PMID: 17203390.

Martins-de-Souza D, Astua-Monge G, Coletta-Filho HD, **Winck FV**, Baldasso PA, de Oliveira BM, Marangoni S, Machado MA, Novello JC, Smolka MB. Absence of classical heat shock response in the citrus pathogen *Xylella fastidiosa*. *Curr Microbiol*. 2007 Feb; 54(2):119-23. PMID: 17211542.

Purcino RP, Medina CL, Martins de Souza D, **Winck FV**, Machado EC, Novello JC, Machado MA, Mazzafera P. *Xylella fastidiosa* disturbs nitrogen metabolism and causes a stress response in sweet orange *Citrus sinensis* cv. Pera. *J Exp Bot*. 2007; 58(11):2733-44. PMID: 17609534.

Ponce-Soto LA, Baldasso PA, Romero-Vargas FF, **Winck FV**, Novello JC, Marangoni S. Biochemical, pharmacological and structural characterization of two PLA2 isoforms Cdr-12 and Cdr-13 from *Crotalus durissus* ruruima snake venom. *Protein* J. 2007 Jan; 26(1):39-49. PMID: 17203396.

Pereira DR, Martins D, **Winck FV**, Smolka MB, Nishimura NF, Rabelo-Gonçalves EM, Hara NH, Marangoni S, Zeitune JM, Novello JC. Comparative analysis of two-dimensional electrophoresis maps (2-DE) of *Helicobacter pylori* from Brazilian patients with chronic gastritis and duodenal ulcer: a preliminary report. *Rev Inst Med Trop Sao Paulo*. 2006; May-Jun; 48(3):175-7. PMID: 16847509.

Damico DC, Lilla S, de Nucci G, Ponce-Soto LA, **Winck FV**, Novello JC, Marangoni S. Biochemical and enzymatic characterization of two basic Asp49 phospholipase A2 isoforms from *Lachesis muta muta* (Surucucu) venom. *Biochim Biophys Acta*. 2005 Oct 30; 1726(1):75-86. PMID: 16005152.

Smolka MB, Martins-de-Souza D, **Winck FV**, Santoro CE, Castellari RR, Ferrari F, Brum IJ, Galembeck E, Della Coletta Filho H, Machado MA, Marangoni S, Novello JC. Proteome analysis of the plant pathogen *Xylella fastidiosa* reveals major cellular and extracellular proteins and a peculiar codon bias distribution. *Proteomics*. 2003 Feb; 3(2):224-37. PMID: 12601815.

Curriculum vitae

Flavia Vischi Winck

E-mail: flaviavw@gmail.com
Birthday: 25th April, 1981. Campinas, São Paulo.
Nationality: Brazilian.

Education

University of Potsdam, Department of Molecular Biology, Potsdam, Germany.
Doctoral work in Natural Sciences (Aug 2007 – present).
Topic: Molecular Biology.
Thesis title: Nuclear proteomics and transcription factor profiling in *Chlamydomonas reinhardtii.*
Supervisor: Prof. Dr. Bernd Mueller-Roeber

Universidade Estadual de Campinas, Department of Biochemistry, Campinas, Brazil.
Master in Functional and Molecular Biology. (March 2005 – July 2007).
Topic: Functional and Molecular Biology.
Thesis title: Study of the proteins HrpF and AvrXacE2 in pathogenicity of *Xanthomonas axonopodis* pv. Citri
Supervisor: Prof. Dr. Marcos Antônio Machado.

Universidade Estadual de Campinas, Campinas, Brazil.
Licenciate in Biology (January 2000 - January 2005).

Escola Técnica Estadual "Conselheiro Antônio Prado" (ETECAP), Campinas, Brazil.
Biochemistry technician (January 1996 - January 1998).

Fellowships received

German Federal Ministry of Education and Research (BMBF), Germany.
Doctoral work, Project: Golm- FORschungseinheiten zur SYSstembiologie (GoFORSYS), (2007.-.2011).

Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), Brazil.
Master´s work, (2005 - 2007). Project: Study of the proteins HrpF and AvrXacE2 in the pathogenicity of *Xanthomonas axonopodis* pv. Citri

European Union ALFA Program "Polylife", Germany.

International course on Biopolymers in Materials and Life Sciences, (September, 2006).

Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Brazil.
Undergraduate´s research, (2003 - 2005). Project: Identification of proteins related to the pathogenicity of *Xanthomonas axonopodis* pv citri.

External talks (last two years)

Humboldt University, Havel Spree Colloquium. December 4th, 2009, Germany.
Talk entitled: "Exploring the nuclear proteome of the green alga *Chlamydomonas reinhardtii*"

Universidade de São Paulo campus "Luiz de Queiroz", March 27th, 2009, Brazil.
Seminar entitled: "Systems biology applied to plant studies."

Universidade Estadual de Campinas, March 18th, 2009, Brazil.
Seminar entitled: "Systems biology applied to plant studies: photosynthesis and growth."

Centro de Citricultura Sylvio Moreira (CCSM), March 13th, 2009, Brazil.
Seminar entitled: "Systems biology applied to plant studies: photosynthesis and growth.

Abstracts in congressess (Selected)

WINCK, Flavia Vischi; RIAÑO-PACHÓN, Diego Mauricio; ARVIDSSON, Samuel; SOMMER, Frederik; MUELLER-ROEBER, Bernd. Towards gene regulatory networks in *Chlamydomonas reinhardtii* via a systems biology approach, 3rd Annual Joint Conference on Systems Biology, Regulatory Genomics, and Reverse Engineering Challenges, 2010, New York, USA.

WINCK, Flavia Vischi; KWASNIEWSKI, Miroslaw; RIAÑO-PACHÓN, Diego Mauricio; WIENKOOP, Stephanie; SOMMER, Frederik; MUELLER-ROEBER, Bernd. Towards the nuclear proteome of *Chlamydomonas reinhardtii*. German Symposium on Systems Biology, 2009, Heidelberg, Germany.

WINCK, Flavia Vischi; KWASNIEWSKI, Miroslaw; RIAÑO-PACHÓN, Diego Mauricio; WIENKOOP, Stephanie; MUELLER-ROEBER, Bernd. Towards the nuclear proteome of *Chlamydomonas reinhardtii*. Conference The DNA-proteome: Recent advances towards establishing the protein-DNA interaction space, 2009, Barcelona, Spain.

WINCK, Flavia Vischi; KWASNIEWSKI, Miroslaw; RIAÑO-PACHÓN, Diego Mauricio; MUELLER-ROEBER, Bernd. Towards Chlamydomonas nuclear proteomics: Optimizing the isolation of nuclei and aggregating information. European BioAlpine Convention: Bridging Public and Private Research On Bioinformatics and Proteomics, 2008, Geneva, Switzerland.

WINCK, Flavia Vischi; KWASNIEWSKI, Miroslaw; MUELLER-ROEBER, Bernd. Towards Chlamydomonas nuclear proteomics: Optimizing the isolation of nuclei. The 13th International Chlamydomonas Conference, 2008, *Hyères*-les-Palmiers, France.

WINCK, Flavia Vischi, AMARAL, Alexandre Morais; BAPTISTA, Juliana Cristina; CABRERA, Odalys G; MISSAWA,Silvia Kazue; MARANGONI, Sérgio; NOVELLO, José Camillo; MACHADO, Marcos Antônio. Genome-based proteomics: analysis of protein-protein interactions of *Xanthomonas axonopodis* pv citri. In: XXXV Annual meeting of Biochemistry and Molecular Biology society, 2006, Águas de Lindóia, Brazil.

WINCK, Flavia Vischi; SOUZA, Daniel Martins de; ASTUA-MONGE, Gustavo; NOVELLO, José Camillo; MACHADO, Marcos Antonio; MARANGONI, Sergio. Identification of proteins potentially related with pathogenicity of *Xanthomonas axonopodis* pv. citri (Xac). In: XXXII Annual meeting of Biochemistry and Molecular Biology society, 2003, Caxambu, Brazil.

WINCK, Flavia Vischi; SOUZA, Daniel Martins de; MARANGONI, Sergio; NOVELLO, José Camillo. Fractionation of the *Xylella fastidiosa* proteome using liquid chromatography and 2DE analysis. In: XXXI Annual meeting of Biochemistry and Molecular Biology society, 2002, Caxambu, Brazil.

WINCK, Flavia Vischi; SMOLKA, Markus Bustamante; SOUZA, Daniel Martins de; MARANGONI, Sergio; NOVELLO, José Camillo. Analysis of *Xylella fastidiosa* Proteome Using Narrow pI Range 2D Maps. In: XXX Annual meeting of Biochemistry and Molecular Biology society, 2001, Caxambu, Brazil.