The CO₂ concentrating mechanism and nitrogen starvation effects under photoautotrophic culture conditions of *Chlamydomonas reinhardtii* studied by high throughput DNA sequencing technology

by

Wei Fang

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Genetics

Program of Study Committee:
Martin H. Spalding, Major Professor
Peng Liu
David J. Oliver
Steven R. Rodermel
Eve S. Wurtele

Iowa State University

Ames, Iowa

2015

Copyright © Wei Fang, 2015. All rights reserved.
TABLE OF CONTENTS

ACKNOWLEDGMENTS .................................................................................................................. ii

CHAPTER 1 GENERAL INTRODUCTION .................................................................................. 1
  Literature Review and Introduction ....................................................................................... 1
  Figures .................................................................................................................................. 25
  References ........................................................................................................................... 26

CHAPTER 2 TRANSCRIPTOME-WIDE CHANGES IN CHLAMYDOMONAS GENE
  EXPRESSION REGULATED BY CARBON DIOXIDE AND THE CO₂ CONCENTRATING
  MECHANISM REGULATOR CIA5/CCM1 .............................................................................. 34
    Abstract ............................................................................................................................. 35
    Introduction ....................................................................................................................... 36
    Results ............................................................................................................................... 42
    Discussion ......................................................................................................................... 58
    Materials and Methods .................................................................................................... 72
    Acknowledgments ........................................................................................................... 85
    Figures and Tables ........................................................................................................... 86
    References ......................................................................................................................... 94

CHAPTER 3 FURTHER INVESTIGATIONS OF THE DIFFERENCE BETWEEN LOW
  AND VERY LOW CO₂ ACCLIMATION STATES ................................................................. 99
    Abstract ............................................................................................................................. 100
    Introduction ....................................................................................................................... 102
    Results ............................................................................................................................... 108
    Discussion ......................................................................................................................... 115
    Material and Methods ...................................................................................................... 120
    Figures ............................................................................................................................... 124
    References ......................................................................................................................... 128

CHAPTER 4 TRANSCRIPTOME STUDY OF CHLAMYDOMONAS REINHARDTII IN
  NITROGEN STARVED CONDITION AND THE ADP-GLUCOSE PYROPHOSPHORYLASE
  REGULATORY SUBUNIT MUTANTS FOR PHOTOAUTOTROPHIC CULTURES .................. 132
    Abstract ............................................................................................................................. 133
    Introduction ....................................................................................................................... 135
    Results ............................................................................................................................... 139
    Discussion ......................................................................................................................... 157
    Material and Methods ...................................................................................................... 166
    Figures and Tables ........................................................................................................... 170
    References ......................................................................................................................... 185

CHAPTER 5 GENERAL SUMMARY ......................................................................................... 190
First I would like to thank my advisor, Dr. Martin Spalding, he kindly guided me and also inspired me through my entire PhD project, and it is my great honor to have a mentor with such wisdom and working spirit in my early scientific career. Secondly, my Statistics minor advisor and my committee member, Dr. Peng Liu, opened another door for me, which leaded to my research focuses on the high-throughput sequencing technology. I would also like to thank my committee members, Dr. David Oliver, Dr. Steven Rodermel, and Dr. Eve Wurtele, for their guidance and support throughout the course of this research.

In addition, I would also like to thank my colleagues: Dr. Yingjun Wang, Dr. Deqiang Duanmu, Dr. David Wright, Dr. Dan Stessman, Dr. Han Gao, Soujanya Akella, Bo Chen, Alfredo Kono who created a friendly and relaxing atmosphere in Dr. Spalding’s lab and provided countless helps for my work and my life. I am also very grateful for generous contributions from our collaborators: Dr. Yaqing Si and Dr. Eric Fritz in the statistics and animal science department at Iowa State University; Dr. Sorel Fitz-Gibbon, Dr. Stephen Douglass, Dr. David Casero, Dr. Sabeeha Merchant, and Dr. Matteo Pellegrini at University of California, Los Angeles. My PhD research projects would not be possible without their efforts.

Finally, I would also like to thank my parents, my family and my girlfriend for their encouragement, support, and love.
CHAPTER 1
GENERAL INTRODUCTION

Literature Review and Introduction

The unicellular green alga *Chlamydomonas reinhardtii* (hereafter, simply *Chlamydomonas*) has been known for several decades as a unicellular model organism. With the complete genomic sequence information and a variety of known mutants, *Chlamydomonas* has been used to make substantial breakthroughs in a wide range of biological research fields, like photosynthesis, chloroplast genetics, flagellar function, nutrient stress responses, *et cetera* (Hemschemeier and Happe 2011; Terashima et al. 2011; Dutcher 2014; Irihimovitch and Yehudai-Resheff 2008). In recent years, much research has focused on its metabolic processes, especially regarding carbon metabolism, because of the potential biomass, biofuel and renewable bioproduct production ability in industrial applications (Skjanes et al. 2013). There are several advantages of choosing *Chlamydomonas* as model organism to study algal carbon metabolism, and they are listed as the following:

As a tiny unicellular, photosynthetic organism, *Chlamydomonas* is easily cultured under ordinary lab conditions. The diameter of an ordinary cell is about 10 µm, and a pair of flagella makes the cell mobile in the aquatic environment. The major habitats of *Chlamydomonas* are wet soil or fresh water, and under illuminated conditions it is capable of photoautotrophic growth. With continuous light source and sufficient inorganic carbon supply, the optimal doubling time can be as short as 7 hours for wild
type strains when cultured in liquid media. Agar based solid media can also be used and the colonies can be easily visualized and characterized (Harris et al. 2009).

*Chlamydomonas* can be cultured with an external organic carbon source: acetate, even without a CO$_2$ supply. As early as 1953, it was noted that *Chlamydomonas* is able to uptake acetate as both a carbon and an energy source (Sager and Granick, 1953). Although it still has not been studied in detail, the assimilation of acetate has been widely accepted as occurring through the glyoxylate cycle. The acetate is first incorporated into acetyl-coenzyme A (acetyl-CoA), and this process can be done through: direct conversion by acetyl-CoA synthetase (ACS); or in 2 steps by acetate kinase (ACK) converting acetate to acetate-phosphate, and followed by phosphate acetyltransferase (PAT) catalysis to form acetyl-CoA (reviewed by Spalding 2009). Acetyl-CoA can either enter the citric acid cycle to generate energy and/or into the glyoxylate cycle to provide intermediates for the synthesis of carbohydrates, proteins and other cellular components, which makes *Chlamydomonas* viable not only in photoautotrophic and mixotrophic (photoheterotrophic) conditions, but also in dark heterotrophic conditions when acetate is present in the medium. This property is especially useful when studying photosynthesis or Calvin cycle deficient mutants, since the same types of mutants are unable to maintain viability for most other photosynthetic species.

The life cycle of *Chlamydomonas* is a typical haplontic life cycle: the vegetative cell with haploid genome is the dominate stage (Harris et al. 2009). So the haploid genome makes genetic studies relatively easy and straight forward. Gametogenesis
occurs under nitrogen starvation conditions: different vegetative cells with distinct mating types: “mt+” and “mt-“, will develop into 2 mating types of gametes, accordingly. Zygotes are formed by mating-induced fusion between the mt+ and the mt- gametes. The diploid zygote undergoes meiosis, then the genomic recombination, and it will generate 4 haploid zoospores, to complete the sexual cycle and return to haploid vegetative cells.

*Chlamydomonas* has raised more attention in recent years, when the entire genome was successfully sequenced and annotated; annotations are constantly updated as well. With about 112 Mbp total genome size, 17 linkage groups were identified and recognized as chromosomes (Merchant et al. 2007). As of late 2014, the most updated genome assembly version 5.5, and the transcriptome annotation is available on the Phytozome portal (http://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Creinhardtii), with easy navigation and searching tools. Roughly 17 thousand gene models are annotated currently (late 2014). An ongoing project is also building a mutant database which aims to generate two mutants for 80% of all genes, and through 9pm US Pacific time on June 23rd, 2014, about 1,700 mutant strains were already available for distribution (https://dpb.carnegiescience.edu/labs/jonikas-lab/Chlamydomonas-mutant-library). These and additional mutants will be available through the Chlamydomonas Resource Center (http://www.chlamy.org/).
The carbon dioxide concentrating mechanism in *Chlamydomonas reinhardtii*:

Despite the power of natural selection and long-time evolution, the photosynthetic process has weaknesses due to the changing environment on earth. Recent atmospheric CO$_2$ concentration normally ranged about 300 to 500 ppm (0.03 to 0.05 %), which is significantly less than that in ancient geological period. The enzyme responsible for the first CO$_2$ assimilation reaction, Ribulose-1,5-bisphosphate carboxylase/oxygenase (Rubisco), does not function efficiently and has a slow catalytic turnover rate (Laing et al. 1975; Lilley and Walker 1975). Converting 1 molecule of CO$_2$ and one molecule of Ribulose-1,5-bisphosphate (RuBP) to 2 molecules of 3-phosphoglycerate (3-PGA) thus becomes a bottleneck reaction in the Calvin Cycle. In addition, the photorespiration pathway that evolved due to the inadequate specificity of Rubisco, which can also oxygenate RuBP to form one molecule of phosphoglycolate and one of 3-PGA, is estimated to consume as much as one third of photosynthetic energy with little biological benefit (Hesketh, 1967; Kebeish et al., 2007). Many photosynthetic species have evolved mechanisms to elevate the substrate CO$_2$ concentration internally to overcome these difficulties (Raven et al. 2008).

As a model organism, the CO$_2$ concentrating mechanism (CCM) of *Chlamydomonas* has been studied for decades. Early research suggested this green alga could have distinct physiological states depending on the varied environmental CO$_2$ levels (Spalding and Jeffery 1989). Since then, numerous genes in *Chlamydomonas* have been identified to respond to limiting CO$_2$ conditions, which shaped a general picture of
this acclimation system. These genes include putative transporters or channels that
directly mediate the influx of inorganic carbon (Ci: \( \text{CO}_2 \), \( \text{HCO}_3^- \), and \( \text{CO}_3^{2-} \)), carbonic
anhydrases (CAs) that facilitate the conversion between Ci species \( \text{CO}_2 \) and \( \text{HCO}_3^- \),
proteins that assist the construction of internal compartments to hold Rubisco, and the
regulatory genes that control the induction and activation of the CCM. With the full
functionality of the CCM, *Chlamydomonas* cells are able to adapt to various
environmental \( \text{CO}_2 \) concentrations and maintain a sufficient photosynthetic rate for
growth (Wang et al. 2011).

Because the *Chlamydomonas* CCM is an inducible system, its regulation has
drawn great attention as molecular biology tools advanced. Two major regulatory genes
have been discovered and characterized based on their relationship to the CCM. A zinc-
finger type transcription regulator was revealed by the *cia5* mutant, which fails to
acclimate to limiting \( \text{CO}_2 \) conditions (Moroney et al., 1989). The responsible gene,
known as *CIA5* or *CCM1*, was cloned by complementation of *cia5* from an indexed library
(Xiang et al., 2001) or from a tagged allele of *cia5* (Fukuzawa et al., 2001), respectively.
Expressions of most putative Ci transporters and induced CAs require a functional CIA5;
but interestingly, the expression of CIA5 itself does not respond to changing \( \text{CO}_2 \) levels
(Fukuzawa et al., 2001; Xiang et al., 2001; Miura et al., 2004). Another regulator, LCR1, is
a Myb domain containing transcription factor, and it appears to regulate the expression
of at least three limiting \( \text{CO}_2 \) induced genes, including *CAH1*, *LCI1*, and *LCI6*. LCR1 itself is
also induced by limiting \( \text{CO}_2 \) conditions, and this induction requires functional CIA5
(Yoshioka et al., 2004). Because CIA5 is extensively related to the induction or regulation
of CCM related genes, even including the downstream regulator LCR1, CIA5 is often considered as the master regulator of the CCM. Whether a gene is regulated by both varied CO₂ conditions and CIA5 has become one of the key characters for people to identify CCM related genes (Miura et al., 2004; Fang et al. 2012).

The discovery and characterization of *Chlamydomonas* CCM related genes has been actively carried out over the past 30 plus years. Ci species need to be transported into and accumulated in specific cellular compartments, and they need to be converted to the form of CO₂ for Rubisco. The barriers and the compartments these Ci species can travel through include: extra cellular matrix (cell wall), plasma membrane, cytosol, chloroplast membrane, chloroplast stroma, thylakoid lumen (acidic pH for converting HCO₃⁻ to CO₂), and pyrenoid (where the majority of Rubisco located). The following paragraphs will discuss the major studied genes that are involve in the CCM through this locational order, and describe their regulation by limiting CO₂ conditions and CIA5.

The cell wall of *Chlamydomonas* is an extracellular matrix composed of carbohydrates and proteins. Two CAs found in the cell wall are believed to facilitate the inter-conversion of the Ci species before entering the plasma membrane. The *CAH1* gene encoding an alpha type CA was one of the earliest identified limiting CO₂ induced genes (Moroney et. al, 1985). Although its role in the CCM seems to be important due to its highly elevated expression under limiting CO₂ conditions and its response to CIA5 regulation (Moroney et. al, 1985; Fukuzawa et. al, 1990; Miura et al. 2004; Fang et al. 2012), a *CAH1* mutant did not show any significant growth phenotype compared to wild
type (Van and Spalding, 1999). The CAH2 gene also encodes an alpha CA that has an amino acid sequence very similar to that of CAH1, but its expression is repressed under limiting CO₂ conditions. The potential function of CAH2 is still unclear.

LCI1 is a plasma membrane protein that is considered to be involved in Ci influx in *Chlamydomonas*. It is highly induced under limiting CO₂ conditions (Burow et al. 1996), and its expression is also controlled by the 2 identified CCM regulator proteins, CIA5 and LCR1 (Miura et al. 2004; Yoshioka et al. 2004). The lcr1 mutant has a high CO₂ requiring phenotype, presumed to result at least partly from the lack of LCI1 expression. Artificially introduced overexpression of LCI1 in the lcr1 mutant shows increased Ci accumulation and photosynthetic affinity (Ohnishi et al. 2010). This evidence coupled with its localization (Ohnishi et al. 2010) makes the LCI1 protein a conclusive candidate for Ci transport in the CCM.

Two Rhesus-like proteins (RHP1 and RHP2), which have similarity to the Rh proteins in the human red blood membrane, are predicted to have 12 transmembrane domains (Soupene et al. 2002, 2004; Kustu and Inwood 2006). The localization of RHP1 has been identified to the plasma membrane (Yoshihara et al. 2008), and the expression of RHP1 is highly induced in high CO₂ conditions compare to limiting CO₂ conditions. Strains with RNAi knockdown of RHP1 show growth defects under high CO₂ condition, so RHP1 was proposed as a channel to further facilitate the diffusion of CO₂ molecules across the plasma membrane under CO2 enrichment conditions (Soupene et al. 2004).
The high light induced gene, *HLA3*, was identified as a putative ATP-binding cassette (ABC) type transporter in the multi-drug-resistance-related protein subfamily (Im and Grossman 2002), and it is also regulated by limiting CO\(_2\) conditions and the CCM master regulator CIA5 (Miura et al. 2004; Xiang et al. 2001; Fukuzawa et al. 2001; Fang et al. 2012). Multiple lines of evidence suggest that the localization of HLA3 protein is on the plasma membrane (Im and Grossman 2002; Gao et al., in review). RNAi knockdown of the *HLA3* gene combined with either a *LCIB* gene (encoding a chloroplast stroma protein) mutation or co-knockdown of *LCIA* (encoding a chloroplast membrane protein) shows defects in both Ci assimilation and growth under high pH conditions, where the dominant Ci species is HCO\(_3^-\) (Duanmu et al. 2009a). The transcript level of *HLA3* gene is kept low when the external Ci source is abundant or the functional CCM regulator CIA5 is absent (Fang et al. 2012). When the *HLA3* gene is artificially activated in high CO\(_2\) acclimated wild type cells, or activated in *LCIA* overexpressing strains in high CO\(_2\) acclimated wild type cells, the Ci accumulation and the Ci dependent photosynthetic O\(_2\) evolution rates are significantly increased at very low Ci concentrations (Gao et al. in review). Thus, the putative role of HLA3 is considered to be mediation of plasma membrane HCO\(_3^-\) transport, and possibly working in tandem with LCIA.

The gene *LCIA* (*NAR1.2*) encodes a chloroplast membrane protein that belongs to the Formate/Nitrite transporter (FNT) family (Rexach et al., 2000; Galvan et al., 2002). Other than its proposed involvement in nitrite transport, as other NAR subfamily members, the expression pattern of *LCIA* is significantly regulated by limiting CO\(_2\) conditions and the master CCM regulator CIA5 (Miura et al. 2004; Fang et al. 2012). LCIA
has six transmembrane domains and it is demonstrated to be located on the chloroplast envelope (Wang and Spalding 2014). Defective Ci uptake in co-knockdown strains of LCIA and another putative Ci transporter gene HLA3 further enhanced the evidence supporting the hypothesized Ci transporter role of LCIA in the CCM (Duamnu et al. 2009). When the environmental Ci concentration is very limited, a highly reduced CO$_2$ dependent photosynthesis was observed in the LCIA-LCIB double mutants; and the similarly decreased photosynthesis in the LCIA single mutant under high pH conditions implies that LCIA is required for the active HCO$_3^-$ uptake, presumably acting in tandem with HLA3 to transport HCO$_3^-$ across both the plasma membrane (HLA3-mediated) and the chloroplast envelope (LCIA-mediated) (Wang and Spalding 2014).

The ycf10 gene encodes a chloroplast envelope protein that has an amino acid sequence similar to that of the plastid-encoded CemA protein from plants and the pxcA gene product from cyanobacteria. Ci uptake activity is decreased in ycf10 gene knockdown strains. It has been speculated that ycf10 may be involved in the light-induced Na$^+$ dependent proton extrusion process, as for pxcA in cyanobacteria (Rolland et al., 1997). No dramatic regulation of ycf10 gene expression was detected either by limiting CO$_2$ conditions or by CIA5 (Fang et al., 2012), since it is plastid-encoded.

Two other limiting CO$_2$ inducible chloroplast envelope proteins, CCP1 and CCP2, share 96% identical amino acid sequences. They are related to the mitochondrial carrier protein superfamily, whose members often function in transporting various metabolites across the mitochondria inner membrane (Spalding and Jeffery, 1989; Ramazanov et al.,
The expression of CCP1 and CCP2 is highly induced by limiting CO₂ conditions, and this induction is regulated by CIA5 (Fang et al. 2012). Knockdown strains of CCP1 and CCP2 did not affect Ci uptake significantly, but still affected the growth of Chlamydomonas cells in limiting CO₂ conditions (Pollock et al. 2004). Mixed evidences leave the exact functions of these 2 proteins still in doubt, although they are still candidates for involvement in the suggested CCM model (Wang et al. 2011).

Another set of limiting CO₂ induced genes, LCIB and three related genes, LCIC, LCID, and LCIE, are predicted to be soluble chloroplast proteins (Wang and Spalding, 2006; Moroney and Ynalvez, 2007; Spalding, 2008). LCIB mutants fail to accumulate internal Ci in low CO₂ conditions and are thus unable to grow in common atmospheric, low CO₂ concentration (0.04%) (Spalding et al., 1983; Wang and Spalding, 2006). However, LCIB mutants have revealed 3 distinctive acclimation states: “high CO₂ (>0.5%)” where the environmental Ci is abundant; “Low CO₂ (<0.5% and >0.02%)” where resemble the current atmospheric CO₂ level; and “Very Low CO₂ (<0.02%)” where the environmental CO₂ is very limiting. Both of the LCIB allelic mutants pmp1 and ad1 have an “air dier” phenotype, in that they fail to grow in Low CO₂ conditions but are able to grow just slightly more slowly than wild type under very low CO₂ conditions (Wang and Spalding, 2006). The expression of LCIB is not entirely repressed in High CO₂ (5%), and it is up-regulated in limiting CO₂ conditions, including both Low and Very Low CO₂. Also, this limiting CO₂ up-regulated expression for LCIB is controlled by CIA5 (Fang et al. 2012). The localization of LCIB protein is dependent on the CO₂ acclimation state of Chlamydomonas as well: either distributed through the stroma under either high CO₂ or
low CO₂ conditions, or concentrates around the pyrenoid structure under very low CO₂ condition (Yamano et al., 2010; Wang and Spalding, 2013; 2014).

Efforts have been made to further understand the “air dier” phenotype. A mutant that has a defective thylakoid lumen CA gene, CAH3, suppresses the LCIB mutant phenotype: the failed growth under low CO₂ condition can be partially recovered by mutating the CAH3 gene in pmp1 strains, although these recovered strains are not able to grow in very low CO₂ condition like the pmp1 strain (Duanmu et al. 2009b). The transcript level of CAH3 gene is responsive to CIA5 regulation and mildly up-regulated in limiting CO₂ conditions (Fang et al. 2012). Evidence also suggests CAH3 is post-translationally regulated through phosphorylation under varied CO₂ conditions: when wild-type cells were acclimated to limiting CO₂ conditions, the CAH3 activity increased about 5–6 fold without significant protein abundance change, the localization of CAH3 was also changed from an association with PSII in the stroma thylakoids to become concentrated in the thylakoids tubules inside the pyrenoid (Blanco-Rivero et al. 2012).

The spherical pyrenoid structure is observable under the microscope within the cup shaped single chloroplast of Chlamydomonas. The pyrenoid structure does not have a lipid membrane boundary, and it is often surrounded by a starch sheath. The most important characteristic of the pyrenoid is that the majority of Rubisco is concentrated in this spherical structure. Considering the alkaline condition in the chloroplast stroma and the fact that Rubisco can only utilize one Ci species, CO₂, concentrating Rubisco in a compartment makes the directional conversion from HCO₃⁻ to CO₂ in the Rubisco
location feasible. CAH3 is a thylakoid lumen protein, and our current hypothesis is that HCO$_3^-$ actively accumulated in the chloroplast stroma enters via an unidentified HCO$_3^-$ channel into the thylakoid lumen tubules within the pyrenoid (Blanco-Rivero et al. 2012), where it is dehydrated to CO$_2$ by CAH3 utilizing the acidic condition inside the thylakoid lumen. This released CO$_2$ directly supplies the CO$_2$ substrate for Rubisco at high concentration in the pyrenoid (Yamano et al. 2010). Because of the high cross-membrane diffusion rate of CO$_2$, the internal Ci pool is believed to accumulate in the form of HCO$_3^-$, and LCIB has been proposed to re-capture the un-used CO$_2$ leaked from the pyrenoid to the chloroplast stroma (Duanmu et al., 2009b). It has also been suggested that this recapture of CO$_2$ by LCIB may not only function as a recycling system, but may also facilitate active CO$_2$ influx into the stromal HCO$_3^-$ pool when the extracellular CO$_2$ source is sufficient (Wang and Spalding 2014).

Other than cah3, many “air dier” phenotype suppressor strains have been generated from LCIB mutants. As reported by Duanmu and Spalding (2011), insertional mutagenesis generated several strains with confirmed, heritable suppressor phenotypes. As mentioned above, two of these suppressors have mutations disrupting the CAH3 gene. Also in this report, the suppressor phenotype of the su8 strain was shown to be linked with the insertion marker, but apparently causes a large, possibly complex disruption on chromosome 16. The insert sequences are not linked with the phenotype in su1 and either su4/su5 strains, which makes the mutation identification challenging. By the combination of genetics and high-throughput sequencing analyses, the su1 mutation has been identified on chromosome 16 in the LCI15 gene, which encodes a

Although su4 and su5 mutations appear to be dominant, numerous crosses between the su4 and su5 strains did not generate any complemented progeny, suggesting that they are very closely linked or possibly allelic. Rough genetic mapping of su5 showed the mutation has 87% linkage with the marker IDA7, and it is noted that the latest genome assembly version 5.5 (phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Creinhardtii) has located this marker on chromosome 16, not on chromosome 1, as described in the version 4 genome assembly (http://genome.jgi-psf.org/Chlamydomonas/Chlamydomonas.info.html) and in (Duanmu and Spalding 2011).

Although it has remained largely a hypothesis, we now have a dynamic model of how Ci species are uptaken and accumulated in Chlamydomonas under different acclimation states, as illustrated in Figure 1 (Wang et al. 2011). In high CO₂ environment, the CCM system is not induced, therefore some putative CO₂ channel proteins like RHP1 may be involved in facilitating the CO₂ diffusion through membranes, and the major inflow of Ci is in the form of CO₂ diffusion, because HCO₃⁻ has much lower permeability than CO₂ across bilayer lipids. Since it is expressed at a lower level, LCIB, perhaps together with LCIC or CAH6, may still accumulate some CO₂ by converting it to HCO₃⁻ in the chloroplast stroma.
When the CO$_2$ source is limited, as in Low CO$_2$, the expression of major HCO$_3^-$ transporters are induced, including HLA3, LCI1, and LCIA. But considering that $LCIB$ mutants cannot survive in Low CO$_2$ environment, it is possible that these HCO$_3^-$ transporters are not fully activated at the post translational level, and that the elevated LCIB presence is still the major driving force of Ci uptake (Wang and Spalding 2014). If we assume the function of LCIB is converting CO$_2$ in the stroma to HCO$_3^-$, this process may serve 2 physiological functions: trapping the CO$_2$ diffusing from the environment to the chloroplast, or recapturing the CO$_2$ generated by CAH3-catalyzed dehydration of HCO$_3^-$ from the stromal pool but not captured by Rubisco. Thus, the ultimate purpose of LCIB appears to be to maintain an internal HCO$_3^-$ pool. Functional CAH3 will keep converting the HCO$_3^-$ (may be partially accumulated through not fully activated HCO$_3^-$ transporters) to CO$_2$, and when the LCIB is defected, the converted CO$_2$ will freely diffuse out of the cell because the recruiting mechanism is absent, and the mechanism of utilizing LCIB as a driven force to actively trapping and accumulating Ci internally is absent, therefore causing an “air dier” phenotype. In the $LCIB$-$CAH3$ double mutant, the leaky consumption of the HCO$_3^-$ is also reduced, and this could be a part of the reason why the double mutant can survive; however, the Ci uptake mechanism in Low CO$_2$ without functional LCIB is still a mystery for the double mutant. If the environmental CO$_2$ concentration is further decreased as in the Very Low CO$_2$ condition, all the major HCO$_3^-$ transporters will be fully activated, and they take over the role of LCIB as the major contributor of the internal HCO$_3^-$ pool accumulation. Under this condition, the source of CO$_2$ that can be used by Rubisco is mainly coming from the conversion of
HCO$_3^-$, so the cah3 mutants will not be viable. Evidence suggests that these HCO$_3^-$ transporters may be only fully functional in Very Low CO$_2$ (Gao et al. in review; Wang and Spalding, 2014). The identification of one of the suppressors of the air dier phenotype as a LCI15 mutant (Akella et al. unpublished) may indicate that the functional activity of these HCO$_3^-$ transporters is regulated by other CCM proteins, possibly through a negative post translational control when energy-consuming HCO$_3^-$ transport is not yet required.

Even with all the knowledge we have today, a clear picture of _Chlamydomonas_ CCM is still far from complete. Key factors like the Ci transport across the thylakoid membrane are still missing, and the explanations to distinguish the low and very low CO$_2$ acclimation states remain as hypotheses. However, with quickly developing technologies, new strategies are vastly emerging. In my thesis, I show additional perspectives contributed by utilizing high-throughput sequencing technologies to provide additional insight.

**Carbon assimilations beyond the CCM**

Photosynthesis and the photosynthetic carbon assimilation beyond the CCM in _Chlamydomonas_ highly resemble the processes in higher plants (reviewed by Spalding 2009). In the Calvin Cycle, after the CO$_2$ is fixed, 3PGA is the immediate product. 3PGA is chemically reduced to glyceraldehyde 3-phosphate (G3P, also called 3-phosphoglyceraldehyde), which, together with its isomerization partner
dihydroxyacetone phosphate (DHAP), represents a major metabolite pool (triose phosphate) from which surplus product can exit from the Calvin Cycle to feed other metabolic pathways. Triose phosphate can be exported from the chloroplast to support biosynthesis in the cytosol or converted into pyruvate and thus enter the fatty acid synthesis pathway and ultimately lipid through the formation of acetyl CoA. G3P and DHAP also can be joined together to produce fructose 6-phosphate (F6P), which can be used in the synthesis of starch. In recent years, the carbon storage pathways have drawn attention, because of algae’s potential industrial production value. In general, organic carbon accumulation will mainly turn to 2 directions: starch synthesis or glycerolipid synthesis. Here, I will very briefly discuss some of the recently generated mutants that are related to starch synthesis, and the current understanding of the triacylglyceride (TAG) lipid accumulation process under nitrogen starved conditions.

The starch content is considered to be synthesized and accumulated in the chloroplast in *Chlamydomonas*, and this accumulation is often occurred in low nutrient stressed conditions (for example the nitrogen starvation condition). When this accumulation happens, the pyrenoid of *Chlamydomonas* is usually surrounded by a sheath of closely packed starch granules (Badger et al. 1998). As several starch synthesis related mutants were generated, people have gained more understanding of the starch synthesis process in *Chlamydomonas*.

ADP glucose in Chlamydomonas is generated by ADP glucose pyrophosphorylase (AGPase) from glucose 1-phosphate and ATP as the immediate building block of starch
synthesis. This reaction is a required, controlling step of carbon flux to starch. Mutants in two different loci have been generated with defects in this step, and starch synthesis is impaired in these mutants as revealed by simple iodine vapor spray tests. The \textit{sta1} mutant was the first mutant generated with defective starch accumulation in \textit{Chlamydomonas}, and this mutant has been reported to accumulate less than 5% of the starch accumulated by wild type (Ball et al. 1991). AGPase in this mutant is less sensitive to 3-PGA activation, but retains normal enzyme catalytic activity. Subsequent research suggested the mutation occurs in the regulatory subunit of the AGPase (Van den Koornhuyse et al. 1996). The lack of amylose and altered amylopectin chain size in the \textit{sta1} mutant is also seen under nitrogen starvation (Libessart et al. 1995). Another mutant with a defective AGPase catalytic subunit was generated as \textit{sta6}, which confirmed the important function of AGPase in Chlamydomonas starch synthesis (Zabawinski et al. 2001).

Different starch accumulation contents were further revealed by other mutants. A \textit{sta5} mutant was also generated at the same study as \textit{sta1}, and it exhibits a 90% decreased total starch amount and a low amylose phenotype (Ball et al. 1991). A mutation in the \textit{STA2} locus impaired the GBSSI enzyme that is responsible for amylose synthesis, which showed a lack of amylose phenotype (Wattebled et al. 2002). On the other hand, 2 mutants: \textit{sta3} and \textit{sta4}, have higher amylose content than amylopectin (Ral et al. 2006; Dauvillee et al. 2006). It was reported that both \textit{sta3} and \textit{sta4} mutants have significantly less starch accumulation in nitrogen starved conditions, but not in well
supplied nutrient conditions (Ral et al. 2006; Dauvillee et al. 2006). STA3 was identified as encoding SSIII, and the STA4 locus encodes the plastidial phosphorylase, PHOB.

The sta7 mutant contains a disrupted isoamylase gene (Dauvillee et al. 2000; Posewitz et al. 2004) and also has significantly decreased levels of starch, but it accumulates a soluble glycogen-like product. Thus the function of STA7 is considered to be polysaccharide debranching. The sta1, sta6 and sta7 mutants have starchless phenotypes (little or no starch accumulation), therefore they are considered as candidates for genetic engineering to alter the carbon flow partitioning.

As mentioned above, many industrial applications (biofuel production, for example) favor carbon flow to the TAG and other lipid accumulations. The knowledge of lipid metabolism in Chlamydomonas is still mainly based on gene orthology through comparisons to seed plants or yeast. In recent days, the lipid accumulation process has drawn substantial attention from researchers, and several detailed observations have been performed in the effort to understand lipid accumulation, especially the TAG lipid metabolism of Chlamydomonas.

The storage TAG synthesis starts with the central energy containing molecule acetyl-CoA. After going through the reactions catalyzed by acetyl-CoA carboxylase (ACCase) and fatty acid synthesis complex (FAS), fatty acyl-CoA molecules are generated. And under the effect of acyltransferases like GPAT and LPAT (glycerol 3-phosphate acyltransferase; and lysophosphatidyl acyltransferase), the fatty acyl chain in the acyl CoA molecules are assembled as esters on the glycerol 3-phosphate backbone. The
phosphate group on the phosphatidic acid (PtdOH) is removed by phosphatidate phosphatase to form diacylglycerol (DAG), and another acyl CoA will assemble the last fatty acid chain on the TAG by diacylglycerol acyltransferases (DAGATs). Other than the de novo synthesis route, membrane lipids can also be used as a source of fatty acid chain when the cells are accumulating TAG. The membrane lipids can be turned into PtdOH or DAG through lipase activities, or transfer the fatty acid chain into DAG to produce TAG by phospholipid diacylglycerol acyltransferase (PDAT). Many of the genes that encode the enzymes discussed above were discovered in the Chlamydomonas genome through orthology studies in comparison with genes from Arabidopsis or yeast (Merchant et al. 2012). However, the molecular level or biochemical level data generally are still lacking to confirm their actual functions in Chlamydomonas.

In Chlamydomonas, many stress responses are triggered when environmental nitrogen is limited. One of these responses is to divert carbon flow into storage materials by increasing partitioning to starch and TAG. Many research teams have reported that Chlamydomonas produces lipid bodies filled with TAG (Wang et al. 2009; Moellering and Benning 2010; Work et al. 2010; Li et al. 2010; Fan et al. 2011; Siaut et al. 2011). Because TAG is a great precursor for biodiesel production, this TAG lipid accumulation suggests the potential biofuel production value of Chlamydomonas, or at least of using Chlamydomonas as the model organism to study this process. Because starch is a major energy and organic carbon sink in the cell, many people have focused on those strains that lack starch production. Many starchless mutants were observed for TAG accumulation, and the result is very promising: cytoplasmic lipid bodies full of
TAG are large and visible under the optical microscope, when the starchless strains are nitrogen starved; in some studies, over feeding these starchless strains with external acetate nutrient when nitrogen is limited, further increased the TAG abundance, and in some cases the cell can even float after centrifugation (Goodson et al. 2011; Goodenough et al. 2014).

Some efforts have been made to understand lipid body formation and mechanisms of TAG accumulation in Chlamydomonas. A major lipid droplet protein (MLDP) has been identified in isolated liposomes as the most abundant protein in the lipid body proteome, and a betaine lipid biosynthetic enzyme BAT1 was discovered as the second most abundant protein. RNAi knockdown of MLDP gene expression resulted in larger lipid bodies, however the TAG content in the cell was not observed to increase (Moellering and Benning 2010). The MLDP protein seems to be algae specific and has little homology to known functional domains, its exact function is still unclear.

Once the effective transcriptome sequencing technology evolved, people began utilizing transcriptome data to find information regarding TAG accumulation in Chlamydomonas. Although the process is complicated, many gene expression behaviors have been identified or confirmed in transcriptome experiments, which have provided evidence to support functions predicted from orthology studies. However, currently all Chlamydomonas transcriptome studies of TAG accumulation in response to nitrogen starvation in either wild type or starchless strains have focused on mixotrophic growth conditions (acetate and light), which is not favored for industrial biofuel production.
Illumina high-throughput sequencing technology

Understanding the genetic code has always been an important aspect of life science research. With the development of new technologies in chemistry and engineering, the traditional Sanger sequencing strategy advanced to a relatively large scale automated process at the end of last century. The completion of human genome project brought vast excitement, and the modern era of biological research has opened a new page. Numerous species, including *Chlamydomonas* (Merchant et al. 2007), have been whole genome sequenced, and the availability of genome sequence information has become one of the most important advantages for people working on a specific model organism. Due to the high demand of genome sequence information, although Sanger’s sequencing still holds the golden standard of accuracy (Liu et al. 2012), its chain-termination strategy is a bottleneck for the high-throughput production of sequence information demanded for modern research.

The development of Illumina’s sequencing technology can be traced back to 1996, initiated by Dr. Pascal Mayer and Dr. Laurent Farinelli at Glaxo-Welcome's Geneva Biomedical Research Institute (GBRI) (http://en.wikipedia.org/wiki/Manteia_Predictive_Medicine). The strategy utilizes reversible dye-terminator technology and an engineered high density DNA cluster surface to perform sequencing while synthesizing the complementary strand of sample DNA molecules. Mixed DNA molecule samples are first prepared with enzyme digestion
and attachment of known adaptor sequences to generate a sample library. Short DNA molecules that are complemented with the adaptor sequences are chemically attached to a surface with very high density, and the glass based slides with these short DNA coated surfaces become the platform where the sequencing reactions occur. With hybridization between the adaptor tails and the short DNA molecules coated on the glass slide, DNA sample libraries are attached to the sequencing slide surface. Single DNA sample molecules are then “bridge” amplified with polymerase, so that local clonal DNA colony clusters are formed. To determine the sequence, four types of reversible terminator nucleotides are added and non-incorporated nucleotides are washed away. A microscopic camera records the fluorescence signal from the labeled nucleotides of each DNA cluster, then the fluorescent dye blocking the 3' termini is chemically removed to allow the next sequencing round. The DNA chains are extended one nucleotide at a time and image acquisitions are performed at a delayed moment after each extension reaction, and very large arrays of DNA colonies are captured by sequential images taken from the camera. The work flow of Illumina sequencing technology has been reviewed by Metzker in (2009), and on the web site of Illumina Inc. (http://technology.illumina.com/technology/next-generation-sequencing/sequencing-technology.html).

The major advantage of Illumina sequencing is the relatively high accuracy and the capability of generating very high throughput data: for short DNA fragment read lengths of 75 to 100 base pairs, up to about 35 Gb of sequences can be generated in few days according to Illumina’s claim in 2010, and this capacity has likely increased since.
RNA samples can be easily converted to DNA libraries in the sample preparation step, making this technology especially useful to apply to transcriptome studies. After validation of its high accuracy and reproducibility (Croucher et al., 2009), the method has been widely used to replace the previously popular microarray based transcriptome quantification technologies. Because the transcriptome is sequenced at the same time as being quantified in Illumina’s sequencing process, requirements of prior knowledge about high quality transcriptome sequence and the manufacturing of high quality microarray chips can be bypassed. This is particularly advantageous when working with non-model organisms, or even less popular model organisms where genome sequences are not available. By implementing successful transcriptome applications, Illumina dominated the Next-Generation Sequencing market by holding about 70% market shares in January 2014 (Zimmerman, 2014, MIT Technology Review).

With the advancement of high throughput sequencing technologies, biologists are facing great challenges as well. The huge amount of data generated by sequencing requires that scientists possess knowledge of computer science, computational biology, statistics, and bioinformatics. And people originally with computer science or statistics backgrounds often lack the comprehensive understanding required to effectively address biological questions. So, even though it is important, collaboration between scientists from different backgrounds is still a challenging barrier. However, growing efforts are devoted to these fast developing cross-disciplinary fields, to facilitate the applications of advanced sequencing technologies. The research presented in my thesis is focused on biological questions but utilized high throughput sequencing technologies.
to gain insight into these biological questions. These cases are also examples of successful collaborations, and as the author I have received valuable experiences from these collaborations.

Thesis organization:

This thesis first includes a publication regarding to the transcriptome sequencing study of the wild type and CIA5/CCM1 mutant under different environmental CO$_2$ conditions (Chapter 2). In Chapter 3, efforts to explain the multiple environmental CO$_2$ acclimation states are presented, which include: a genome sequencing effort has been made to identify the previously studied “air-dier” phenotype (LCIB mutation) suppressor strains su4/5 (Duanmu and Spalding 2011); and the transcript level study of 5 major CCM genes $LCIA$, $LCIB$, $LCI1$, $HLA3$ and $LCIB$ in a time course under the induction of variant CO$_2$ gradients ranged from Low CO$_2$ to Very Low CO$_2$ concentrations. The carbon assimilations beyond Ci uptake are explored in Chapter 4, which presented a study about the transcriptome changes of wild type strains and $sta1$ mutant strains under limited nitrogen environments. A general summary of conclusions is presented in the final chapter.
Figure legends

Figure 1: The illustration of a schematic model for the Chlamydomonas CCM (Wang et al. 2011). See the text for a full explanation of our current understanding of CCM function.
References


inorganic carbon uptake in the green alga *Chlamydomonas reinhardtii*. *Plant Cell*, 22(9), 3105-3117.


through substrate supply in the monocellular alga *Chlamydomonas reinhardtii*. *Journal of Biological Chemistry*, 271(27), 16281-16287.


CHAPTER 2

TRANSCRIPTOME-WIDE CHANGES IN CHLAMYDOMONAS GENE EXPRESSION REGULATED BY CARBON DIOXIDE AND THE CO₂ CONCENTRATING MECHANISM REGULATOR CIA5/CCM1

This chapter is a research article published in Plant Cell. 2012 May;24(5):1876-1893

Wei Fang⁹, Yaqing Siᵇ, Stephen Douglassᶜᵉ, David Caseroᶜᵉ, Sabeeha S. Merchantᵈᵉ, Matteo Pellegriniᶜᵉ, Istvan Ladungaᶠ, Peng Liuᵇ, Martin H. Spaldingᵃ¹

ᵃDepartment of Genetics, Development and Cell Biology, Iowa State University  
bDepartment of Statistics, Iowa State University  
cDepartment of Molecular, Cell, and Developmental Biology, University of California, Los Angeles  
dDepartment of Chemistry and Biochemistry, University of California, Los Angeles  
ᵉInstitute of Genomics and Proteomics, University of California, Los Angeles  
fDepartment of Statistics, University of Nebraska, Lincoln, Nebraska 68588-0665

Author Contributions:

W.F., P.L., and M.H.S. designed the research; W.F. performed the research; Y.S. and P.L. contributed new analytical tools; all authors analyzed data, and all authors contributed to the writing of the paper.

Corresponding author: Martin H. Spaldingᵃ¹ (mspaldin@iastate.edu)

Distribution of Materials: The authors responsible for distributing materials integral to the findings presented in this article in accordance with the policy described in the instructions for Authors (www.plantcell.org) are: Martin H. Spalding (mspaldin@iastate.edu) and Wei Fang (fangwei@iastate.edu)
Abstract

We used RNA-Seq to query the *Chlamydomonas reinhardtii* transcriptome for regulation by CO₂ and by the transcription regulator CIA5 (CCM1). Both CO₂ and CIA5 are known to play roles in acclimation to low CO₂ and in induction of an essential CO₂-concentrating mechanism (CCM), but less is known about their interaction and impact on the whole transcriptome. Our comparison of the transcriptome of a wild type vs. a cia5 mutant strain under three different CO₂ conditions: high CO₂ (5%); low CO₂ (0.03-0.05%); and very-low CO₂ (<0.02%) provided an entrée into global changes in the gene expression patterns occurring in response to the interaction between CO₂ and CIA5. We observed a massive impact of CIA5 and CO₂ on the transcriptome, affecting almost 25% of all *Chlamydomonas* genes, and we discovered an array of gene clusters with distinctive expression patterns that provide insight into the regulatory interaction between CIA5 and CO₂. Several individual clusters respond primarily to either CIA5 or CO₂, providing access to genes regulated by one factor but decoupled from the other. Three distinct clusters clearly associated with CCM-related genes may represent a rich source of candidates for new CCM components, including a small cluster of genes encoding putative inorganic carbon transporters.
Introduction

The photosynthetic conversion of inorganic carbon (Ci) into organic form is responsible for the abundance of biomass on earth. In this process, ribulose 1,5-bisphosphate carboxylase/oxygenase (Rubisco) catalyzes the initial incorporation of CO$_2$ via the carboxylation of ribulose 1,5-bisphosphate (RuBP) by CO$_2$ (reviewed by Andersson 2008). Although critically important, the catalytic activity of Rubisco is slow compared to many other enzymes and also cannot discriminate completely between CO$_2$ and O$_2$; the oxygenation of RuBP is competitive with the carboxylation reaction.

Under present atmospheric conditions, CO$_2$ assimilation rates often are limited by the CO$_2$ concentration, and in many photosynthetic species, ranging from cyanobacteria and algae to C4 vascular plants, an active CO$_2$ concentrating mechanism (CCM) has evolved to help offset the deficiencies of Rubisco (Raven et al. 2008). CCMs are especially prevalent in aquatic photosynthetic organisms.

*Chlamydomonas reinhardtii*, a unicellular green alga that serves as a reference organism, also exhibits acclimations to varied CO$_2$ levels (reviewed by Spalding 2009). *Chlamydomonas* must overcome the 10,000-fold slower diffusion of CO$_2$ in water relative to air. Thus active transport and accumulation of Ci, either as CO$_2$ or as HCO$_3^-$, plays a critical role in the *Chlamydomonas* CCM (Moroney and Ynalvez 2007; Spalding 2008). Internal accumulation of Ci occurs against a large concentration gradient, so accumulation must occur as HCO$_3^-$, because its permeability across lipid membranes is 1000-fold lower than that of CO$_2$. However, Rubisco uses CO$_2$ as substrate, so, along
with Ci transporters, carbonic anhydrases (CAs), which catalyze interconversion of CO₂ and HCO₃⁻, also play important roles in the CCM, (Spalding et al.1983a; Coleman and Grossman1984; Moroney et al. 2011).

The *Chlamydomonas* CCM is induced by low CO₂ concentrations, the discovery of CCM-related genes has been based on identifying genes with elevated expression under limiting CO₂ (lower than 0.05%) compared to high CO₂ (1-5% CO₂) (Spalding and Jeffery 1989; Chen et al. 1997; Somanchi and Moroney 1999; Miura et al. 2004; Yamano and Fukuzawa 2009). Many CAs and putative transporters or other LCI (Low CO₂ Inducible) genes have been discovered by this criterion and have been hypothesized to relate to the CCM of *Chlamydomonas* (Miura et al. 2004; Yamano and Fukuzawa 2009).

The detailed regulatory mechanisms of the CCM remain unclear, but two important transcription regulators have been identified and characterized based on their relationship to the CCM. A zinc-finger type transcription regulator, CIA5 (or CCM1), was identified by complementation of the cia5 mutant (Moroney et al. 1989), which is unable to acclimate to limiting CO₂ conditions, and, independently, by cloning of a tagged allele of *cia5, ccm1* (Xiang et al. 2001; Fukuzawa et al. 2001). Expression of most putative Ci transporters and induced CAs requires CIA5, even though the expression of CIA5 itself does not depend on the CO₂ level, so post-translational activation of CIA5 in low CO₂ apparently is required for CIA5 to regulate these genes (Xiang et al. 2001; Fukuzawa et al. 2001; Miura et al. 2004). Another transcription regulator, LCR1, has a Myb domain and appears to regulate the expression of at least 3 limiting CO₂ induced
genes, *CAH1, LCI1* and *LCI6*. *LCR1* itself also is induced by limiting CO$_2$, and this induction requires CIA5 (Yoshioka et al. 2004). Because of the extensive connection of CIA5 to regulation of the CCM-related genes, including LCR1, CIA5 is often called the “master regulator” of the CCM.

Regarding the mechanism of Ci transport and accumulation in the CCM, the first barrier to Ci uptake is the plasma membrane. Two CIA5-regulated genes encoding candidate transporters have been implicated in Ci transport across the plasma membrane: *HLA3* encodes a putative ATP-binding cassette (ABC) type transporter, is induced under low CO$_2$ conditions, and knockdown of its expression impairs photosynthesis, Ci uptake and growth in alkaline conditions (Duanmu et al. 2009a). *LCI1* encodes a plasma membrane protein reported to increase Ci uptake in LCR1 mutants when expressed transgenically (Ohnishi et al. 2010). Two Rhesus-like proteins, RHP1 and RHP2, also are predicted to be plasma membrane located (Yoshihara et al. 2008). The RHP1 protein has been proposed as a CO$_2$ channel to facilitate CO$_2$ influx under high CO$_2$ conditions (Soupene et al. 2002, 2004), and its expression is reportedly up-regulated in high CO$_2$.

Some chloroplast envelope proteins also are candidates to transport Ci into the stroma. The *LCIA* (*NAR1.2*) gene, which encodes a Formate/Nitrite Transporter (FNT) family protein targeted to the chloroplast envelope, is induced in low CO$_2$ and requires CIA5 for expression (Galvan et al. 2002; Miura et al. 2004). *LCIA* has been reported to increase HCO$_3^-$ transport when transfected into *Xenopus* oocytes (Mariscal et al. 2006),
and its product has been implicated in Ci transport in *HLA3-LCIA* co-knockdown *Chlamydomonas* strains (Duanmu et al. 2009a). RNAi knockdown of *CCP1* and *CCP2*, which encode nearly identical, low-CO$_2$ inducible chloroplast envelope proteins (Spalding and Jeffery 1989; Ramazanov et al. 1993; Chen et al. 1997) resulted in poor growth under low CO$_2$ conditions, although no direct evidence for a defect in Ci transport or photosynthesis was demonstrated (Pollock et al. 2005).

The combined transport of HCO$_3^-$ across the plasma membrane and the chloroplast envelope results in the accumulation of HCO$_3^-$ in the chloroplast stroma. Since Rubisco, located in the pyrenoid, cannot use HCO$_3^-$, a specific CA, CAH3, dehydrates the accumulated HCO$_3^-$ to CO$_2$ in the thylakoid lumen, taking advantage of the acidic lumen environment to drive nearly complete conversion of HCO$_3^-$ to CO$_2$ (Spalding 2008; Moroney et al. 2011). This essential role of CAH3 also mandates the transport or facilitated diffusion of HCO$_3^-$ across the thylakoid membrane, but this has not yet been demonstrated.

Another set of low-CO$_2$ induced genes, *LCIB* and three related genes, *LCIC*, *LCID* and *LICE*, also have been implicated in Ci transport and accumulation even though they are predicted to be soluble chloroplast proteins (Wang and Spalding 2006; Moroney and Ynalvez 2007; Spalding 2008). *LCIB* mutants fail to accumulate internal Ci in low CO$_2$ conditions and are thus unable to grow in air levels of CO$_2$ (Spalding et al. 1983b; Wang and Spalding 2006). Notably, *LCIB* mutants have revealed the existence of a third acclimation state at very low CO$_2$ concentrations (<0.02%): both *LCIB* allelic mutants
pmp1 and ad1 die under low CO₂ conditions (<0.05% and >0.02%), but are able to grow slowly under very low CO₂ (<0.02%) conditions (Wang and Spalding 2006).

Because transmembrane domains are not evident, these LCIB family proteins cannot be stand-alone Ci transporters. It has been suggested that they might serve as Ci transport regulators or as Ci transport complex subunits (Wang and Spalding 2006), but LCIB either distributes through the stroma or concentrates around the pyrenoid (Duanmu et al. 2009a; Yamano et al. 2010), making interaction with Ci transporters unlikely. Mutants defective in the thylakoid lumen CA, CAH3, suppress the LCIB mutation phenotype (Duanmu et al. 2009a), suggesting a role for LCIB and LCIC, with which LCIB forms a heteromeric complex (Yamano et al. 2010), in preventing the leakage of CO₂ from the stroma. CAH6, a putative chloroplast stromal CA, also may be involved in CO₂-to-HCO₃⁻ conversion in the stroma to reduce diffusive loss of CO₂ from the chloroplast (Mitra et al. 2004).

Even though the Chlamydomonas CCM has been extensively studied in recent years, we still know little about the limiting-CO₂ acclimation process, and the potential for discovery of new genes involved in this process is very high. The acclimation to limiting CO₂ and induction of the CCM in Chlamydomonas appear to be regulated by the “master regulator”, CIA5 (or CCM1) (Miura et al. 2004). The cia5 mutant appears to completely lack induction of the CCM, although it is viable under high CO₂ conditions and grows more slowly than WT in air levels of CO₂. Also, most identified low-CO₂ induced genes remain uninduced when cia5 is exposed to low CO₂ (Moroney et al. 1989;
Spalding et al. 2002). Aside from being a critical upstream regulator of the CCM and other low CO₂ acclimation responses and likely requiring post-translational activation in low CO₂, the details of CIA5 function remain undiscovered. CIA5 has been proposed to be a transcription regulator (Xiang et al. 2001; Fukuzawa et al. 2001), but we know very little about sequences recognized by its putative DNA binding domain or the genes it directly regulates downstream.

To better understand the CCM and low-CO₂ acclimation of *Chlamydomonas* in general, as well as the function of CIA5, we conducted RNA-Seq experiments employing the Illumina Genome Analyzer II because of its superiority over the traditional microarray methods (González-Ballester et al., 2010; Castruita et al. 2011; Wang et al., 2010) using two *Chlamydomonas* strains: the 137c wild type (cc125); and *cia5* (cc2702), a mutant in the 137c background with a point mutation in CIA5. In order to also gain insight into the multiple acclimation states, the strains were grown at three different CO₂ concentrations as quantified below: high CO₂ (H-CO₂); low CO₂ (L-CO₂); and very low CO₂ (VL-CO₂). Our transcriptome comparison identified a massive impact of CIA5 and CO₂ on the transcriptome and revealed an array of gene clusters with distinctive expression patterns that provide insight into the regulatory interaction between CIA5 and CO₂. Individual gene clusters responded primarily to CIA5, to CO₂ or to an interaction between the two. This study of transcriptome-wide gene expression patterns provides insight into the massive impact of these two factors and their interaction on *Chlamydomonas* gene expression in addition to identifying compelling new candidates for CCM functional components.
Results

Identification of differentially expressed genes

This transcriptome study was designed employing three CO$_2$ acclimation states: H-CO$_2$ (5% CO$_2$); L-CO$_2$ (0.033-0.041%); and VL-CO$_2$ (0.011-0.015%); and two strains (genotypes): the cia5 mutant and its original wild type (WT) progenitor, 137c. Processing of RNA samples on the Illumina Genome Analyzer system yielded more than 12 million reads mapped to the transcriptome for each sample, and more than 90% of these were uniquely mapped to the *Chlamydomonas* genome (Supplemental Table 1 online). We detected expression for 15,649 of 15,818 filtered Augustus 5.0 gene models (>99% coverage). Since Augustus 5.0 predictions were based on the *Chlamydomonas* Version 4 genome assembly, we also acquired annotation information from the filtered Version 4 model set available from the JGI database as user annotation references and sources for the common gene names.

As an aid to examining gene expression level distributions, we calculated the Reads Per Kilobase of exon model per Million of aligned reads (RPKM) values as normalized expression estimates for each gene model in each sample. The shape of distributions for the average RPKM values are very similar among the six conditions, as are the 5th, 50th and 95th percentiles of these distributions (Supplemental Figure 1 online). Also, the calculated correlation coefficients, based on the log-transformed RPKM values after eliminating genes with zero count in either of the two replicates,
between the two biological replicates for each condition range from 0.935 to 0.983, indicating high correlation between replicates.

To evaluate the reliability of our RNA-Seq results, we performed quantitative Polymerase Chain Reaction (qPCR) on eight previously studied genes (CAH1, CAH3, CAH6, CIA5/CCM1, HLA3, LCIB, LCIE, and RHP1) using the same RNA samples as those used for RNA-Seq. These genes were selected to represent a wide range of expression levels and expression patterns under the conditions used. For all eight genes, the expression patterns from RNA-Seq and qPCR agree very well visually and also are highly correlated, with correlation coefficients ranging from 0.92 to 0.995 (Supplemental Figure 2 online).

After validating our RNA-Seq results with qPCR, we applied a generalized linear model analysis based on a negative binomial distribution, and conducted an “overall test” to determine which genes vary in expression among any of the six treatment groups, where a treatment group is defined by a strain-by-induction condition combination (see Methods for details). While controlling the false discovery rate (FDR) at 2.5% using Benjamini and Hochberg’s method (Benjamini and Hochberg 1995), we identified 3678 genes as differentially expressed (DE) among the six treatment groups (Supplemental Data Set 1 online). This number is similar in scale to the 5,884 differentially expressed genes at 30, 60 or 180 minutes after CO₂ deprivation in wild type *Chlamydomonas* cells reported in the companion publication (Brueggeman et al., 2012).
The overall test identified genes with differential expression in any of the six treatment groups. The transcript levels of these genes might be affected by: 1) the CO$_2$ concentration; 2) the presence/absence of functional CIA5; and/or 3) the interaction of CO$_2$ concentration and the presence/absence of functional CIA5. In order to provide more detailed information about how the CO$_2$ level or the presence/absence of CIA5 affects gene expression, we utilized a C/S impact model. Under this model, we separately tested for a "CO$_2$ effect" (due to varied CO$_2$ levels, "C-effect"), a "strain effect" (due to varied genotypes, "S-effect"), and an interaction effect between CO$_2$ levels and genotype ("CS-effect") using the generalized linear model ("C/S impact test") as described in the Methods section. When we control the FDR level at 2.5%, this C/S impact test identifies most of the DE genes, with only 165 of the 3678 DE genes identified from the overall test failing to show significance for any one of the 3 possible effects. Among the other 3513 genes, 2230 exhibit significant C-effect, 2787 exhibit significant S-effect, and 372 exhibit significant CS-effect (Supplemental Data Set 1 online).

To facilitate a closer comparison of previously reported LCI genes with our results, we also conducted a pair-wise comparison of our expression data for the WT in H-CO$_2$, L-CO$_2$ and VL-CO$_2$ conditions using the DESeq package (Anders and Huber, 2010), which was reported to be one of the best methods for identifying DE genes between two treatment groups (Kvam et al., 2011). When we controlled FDR at level 2.5% using Benjamini and Hochberg’s method (Benjamini and Hochberg, 1995), we identified 345 genes differentially expressed for the L-CO$_2$ vs. H-CO$_2$ pair-wise comparison and 696
genes differentially expressed for the VL-CO$_2$ vs. H-CO$_2$ pair-wise comparison (Supplemental Data Set 2 online). Surprisingly, no genes were identified as differentially expressed for the VL-CO$_2$ vs. L-CO$_2$ pair-wise comparison.

Reproducibility across laboratories

The companion study by Brueggeman et al. (2012) focused exclusively on the effects of CO$_2$ deprivation on gene expression. Their focus on the time course for induction from 0 to 3 h nicely complements our present study, which compares the impact of CO$_2$ deprivation and CIA5 on gene expression following a 4-h induction in limiting CO$_2$. Their findings support many of our observations and conclusions regarding transcriptome changes associated with CO$_2$ deprivation. Nonetheless, the two studies were conducted completely independently and involved significant differences in experimental conditions (e.g., light and temperature) and in the Chlamydomonas strains used. Therefore, the differences found in the patterns and the magnitude of gene expression between these two studies are not unexpected (i.e., considerably lower inter-laboratory reproducibility than intra-laboratory reproducibility is expected). Further details about the reproducibility of our expression estimates and their correlation with results from our companion paper can be found at the end of Methods and in Supplemental Figures 3, 4 and 5 online.
Clusters of genes with similar expression patterns

We applied a model-based clustering algorithm to identify distinct gene expression profiles among identified DE genes and chose a total of 16 clusters to maintain as few tight clusters as possible while including most of the distinct expression patterns (Figure 1). Each gray line in Figure 1 represents the expression pattern for an individual gene, and the single black line indicates the average behavior for all genes in that cluster.

When sorted by cluster, the C/S impact test results (Supplemental Figure 6 online) confirmed many of the visually observed patterns in the clusters. For example, gene expression patterns in clusters 1, 2, 3, 10, 11 and 13, the “CIA5 clusters”, appear to be affected mainly by the presence/absence of CIA5 but only minimally by CO₂. In agreement with this, 1060 (~76%) of the 1396 genes in these 6 CIA5 clusters exhibit only a significant S-effect (no C-effect or CS-effect).

In contrast, the expression patterns in clusters 4, 7 and 9, the “CO₂ clusters”, appear to respond to variation in CO₂, with little apparent difference between the genotypes. Accordingly, 415 of the 764 genes in these 3 CO₂ clusters exhibit only a significant C-effect (no S-effect or CS-effect). Like those in the CO₂ clusters, genes in clusters 6 and 12 also exhibit visually parallel changes in response to changing CO₂ between the genotypes, but also show a slightly larger expression shift between the genotypes (Figure 1). Many genes in these two “pseudo-CO₂” clusters exhibit C+S effects.
or only C-effect, but the larger proportion of genes exhibiting S-effect distinguishes them from the CO\textsubscript{2} clusters (Supplemental Figure 6 online).

Clusters 8, 14, and 15, the “CCM clusters”, exhibit a pattern of induction or up-regulation under limiting CO\textsubscript{2} and repression by the absence of CIA5. These three CCM clusters also contain a considerable number of genes exhibiting significant C+S-effects (both C and S effects) and C+S+CS-effects (all C, S, and CS effects), as well as a large number of genes exhibiting only S-effects and few or no genes exhibiting only C-effects (Supplemental Figure 6 online).

Genes in clusters 5 and 16 show the mildest changes over the 6 strain-by-treatment combinations (Supplemental Figure 6 online), and exhibit a mix of genes with all 3 effects (C-effect, S-effect and CS-effect). These two clusters also include the largest proportion of genes that were detected by the overall test but not the individual test for any one of the 3 possible effects from the C/S impact model.

**Functional implications of the gene expression clusters**

In addition to the distribution of genes into clusters based on similar expression patterns, we used two complementary methods to examine the DE genes within each cluster for commonalities of function: the Algal Functional Annotation Tool (Lopez et al. 2011) and manual curation. In employing the Algal Functional Annotation Tool, we used the Gene Ontology terms (GO terms) based on orthology to *Arabidopsis thaliana* to
overcome the limitation of available annotations for *Chlamydomonas*. We compiled all GO terms that showed statistical significance (*p*<0.01) in at least one gene cluster, and generated a summary heat map to visualize an overview of the resulting functional information by clusters (Supplemental Figure 7 online). A detailed list of GO terms identified for the clusters can be found in Supplemental Data Set 3 online. In the heat map, GO terms were subjected to hierarchical clustering so that gene clusters with common significant ontology terms are placed close to each other in the tree. Although details of the identified GO terms and associated genes corresponding to the heat map are found in Supplemental Data Set 3 online, Supplemental Figure 7 online illustrates that very few of the significant GO terms (20 out of 210) overlap among any of the 16 cluster entries, which suggests that the genes separated into clusters based on distinctive expression patterns also tend to be involved in varied biological processes, providing independent support for our clustering results.

In addition to tabulating the significant GO terms associated with each gene cluster, the total number of unique genes represented within all of the significant GO categories for each cluster was determined. For example, Table 1 indicates that 22 GO terms were identified by the Algal Functional Annotation Tool to be associated with cluster 1, but these 22 GO terms represent only 3 unique genes, since each of the 3 genes is associated with multiple GO terms. This example is not unique; in many of the clusters, the significant GO category hits represented only a small number of individual genes, even if the number of GO category hits was high. On the other hand, 60 unique genes (32.1% of the genes) in cluster 5 were included among the significant GO term
hits. However, only in 6 of the clusters, (4, 5, 6, 9, 10 and 11) were at least 8% of the genes in the cluster identified among the GO hits.

Because of the relative paucity of functional annotation in the *Chlamydomonas* genome, the Algal Functional Annotation Tool was unable to provide much functional information for more than a few gene clusters. Therefore, we also employed manual curation to place DE genes into 8 broad functional categories (Supplemental Data Set 4 online). Not surprisingly, the most abundant manual functional category of DE genes in all 16 clusters was “unknown”, which is represented by the difference between 100% and the sum of all other functional categories for each cluster in Figure 2 and accounts for 38-62% of the genes in each. Among the 16 clusters, the most abundant of the 8 manual categories after “unknown”, are: “metabolism”, “signaling”, and “gene expression and regulation”. The relative distribution of DE genes among these 8 manually curated categories is illustrated for each gene cluster in Figure 2, and a compilation of the primary functional category in each cluster identified as including the largest proportion of genes (excluding the “unknown” category) is summarized in Table 1.

No single functional category among the genes in each major cluster group (CIA5, CO2 and CCM clusters) was consistently apparent by either method. However, those CIA5 clusters (1, 3 and 10) with higher gene expression in WT all have signaling as the primary functional category, whereas those clusters (2, 11 and 13) with higher gene expression in *cia5* have metabolism as the primary functional category. Furthermore,
the functional categories of signaling and gene expression together accounted for more than half of the genes in CIA5 clusters 1, 3 and 10, excluding the unknown category. Of these CIA5 clusters, only clusters 10 and 11 contained more than 8% of the genes identified as GO category hits, but these GO hits agreed with the primary functional category of signaling for cluster 10, in that they fell in the general areas of intracellular trafficking, proteolysis and regulation of processes, and of metabolism for cluster 11, in that they fell in the general area of metabolic processes (Supplemental Data Set 3 online).

Similarly, the CO2 cluster 9 with increased transcript abundance at higher CO2 concentrations has gene expression as the primary functional category, whereas those clusters (4 and 7) with increased gene expression at lower CO2 concentration have metabolism as the primary functional category. Furthermore, the functional categories of gene expression and signaling combined accounted for more than half of the genes in CO2 cluster 9, excluding the unknown category. Of the 3 CO2 clusters, only 4 and 9 each contained more than 8% of the genes identified as GO category hits, and in both cases these GO hits agreed with the primary functional category identified. In CO2 cluster 9, which had gene expression as its primary functional category, the GO category hits fell in the general areas of RNA processes and nitrogen metabolism (Supplemental Data Set 3 online). For CO2 cluster 4, which had metabolism identified as its primary functional category, the GO category hits fell in the general area of catabolic processes (Supplemental Data Set 3 online), and a large proportion of the cluster 4 genes in the
manually curated metabolism category were putative catabolic genes (Supplemental Data Set 4 online).

Among the CCM clusters, the various functional categories appeared to be relatively evenly dispersed, with the exception of cluster 15. Although cluster 15 had the somewhat common category of metabolism as its primary functional category, it is notable in having, among all the clusters, the highest proportion (~14%) of manually curated genes in the transport functional category.

**Key CO₂ assimilation related genes and pathways**

In addition to the segregation of genes into broad functional categories, we also analyzed the distribution of specific groups of genes among the gene expression pattern clusters, such as previously reported low-CO₂ inducible (LCI) genes, Calvin cycle genes, photorespiratory pathway genes, and carbonic anhydrase genes.

Of 2274 genes exhibiting a C-effect and/or a CS-effect (Supplemental Data Set 1 online), and thus indicating a statistically significant response to CO₂ concentration, 1350 were up-regulated in WT L-CO₂ vs. H-CO₂, and 418 had a fold change of 2 or greater. This selection of genes is the most comparable to classic LCI genes reported previously (Chen et al. 1996; Somanchi et al. 1999; Miura et al. 2004; Wang et al. 2005; Wang and Spalding, 2006; Yamano and Fukuzawa 2009). We selected 106 of these previously reported LCI genes for a direct comparison with genes identified as having a
C-effect or CS-effect. Among these 106 previously reported LCI genes, 49 exhibit a C-effect or CS-effect, and 45 of these were up-regulated in either L-CO$_2$ or VL-CO$_2$ conditions compared with H-CO$_2$ in the WT strain (Supplemental Data Set 5 online). We also used a recently proposed statistical method implemented in the Bioconductor package DESeq (Anders and Huber, 2010) to perform a direct, pair-wise comparison of gene expression for H-CO$_2$ vs. either L-CO$_2$ or VL-CO$_2$, which identified a highly overlapping but slightly different list of 40 previously reported LCI genes as up-regulated in our experiment. In combination with the C/S impact model, the DESeq analysis supports 53 of the previously reported LCI genes as up-regulated in either L-CO$_2$ or VL-CO$_2$ (Supplemental Data Set 5 online).

Data included in our companion paper (Brueggeman et al. 2012) demonstrate a low-CO$_2$ up-regulation of 40 of the 106 previously identified LCI genes, and 35 of these overlap with the 53 genes identified here as being up-regulated by L-CO$_2$ or VL-CO$_2$. In combination, our data and the data from Brueggeman et al. (2012) provide support for up-regulation of 60 of the 106 previously reported LCI genes. In addition to showing down-regulation for 3 of the same 4 previously reported LCI genes that our data identify as down-regulated, our companion paper identifies an additional 2 previously reported LCI genes that exhibit down-regulation under their experimental conditions (Supplemental Data Set 5 online).

From our list of 49 previously reported LCI genes showing a C-effect or CS-effect, 36 genes fall into the CCM clusters (8, 14, and 15), and an additional 7 genes in the CCM
clusters were identified as up-regulated in L-CO$_2$ or VL-CO$_2$ based on the pair-wise \textit{DESeq} analysis (Supplemental Data Set 5 online). To explore this relationship further, we selected 10 intensively studied, CIA5-regulated, LCI genes (\textit{CAH1, CAH3, CCP1, CCP2, HLA3, LCIA, LCIB, LCIC, LCI1} and \textit{LCR1}; highlighted in Supplemental Data Set 5 online) implicated as functionally involved in the CCM (Spalding 2008; Wang et al. 2011), and found all to be contained in CCM clusters 8, 14 and 15, and all except \textit{CAH3} were identified as differentially expressed by our companion paper, Brueggeman et al. (2012). Figure 3 provides a schematic model of the proposed \textit{Chlamydomonas} CCM, including Ci uptake and accumulation processes (modified from Wang et al. 2011), including identified and proposed locations of the various CAs (Moroney et al. 2011). A major proportion of the CCM/CA genes in this model were included in CCM clusters 15, 14 or 8, providing additional validation of our clustering results.

Although a large proportion of the previously reported LCI genes identified here as DE genes were found to be associated with CCM clusters 8, 14 and 15, a substantial number also were associated with CO$_2$ clusters 7 and 9. Of the 53 previously reported LCI genes supported by our data, 7 fell into CO$_2$ cluster 7, 4 genes fell into CO$_2$ cluster 9, and 1 each into CIA5 cluster 11 cluster 12 (Supplemental Data Set 5 online). Six previously reported LCI genes that were not supported as LCI genes by our data (i.e., no C-effect or CS-effect and not differentially expressed based on the \textit{DESeq} analysis) were identified as DE genes but fell mostly into CIA5 clusters 3, 10, 11 and 13.
By visual inspection, the gene expression pattern in CCM cluster 15 shows very low expression in H-CO$_2$ and induction in VL-CO$_2$ and L-CO$_2$ conditions for WT, and very low expression in any CO$_2$ conditions for the cia5 mutant. CCM clusters 8 and 14, on the other hand, show only up-regulation of expression under VL-CO$_2$ and L-CO$_2$ conditions for WT, relative to the modest expression in H-CO$_2$ conditions, and almost equally low expression under any CO$_2$ conditions for the cia5 mutant. Thus the patterns for CCM clusters 8, 14, and 15 progress from mild up-regulation of expression to high-level induction, respectively. Only 35 genes showed the high-level induction and were grouped in CCM cluster 15, so every gene in this cluster is listed in Table 2, and all genes in clusters 8, 14 and 15 are listed in Supplemental Data Set 6 online.

Based on our manual functional curation, CCM cluster 15 contains a relatively large proportion of genes in the manually annotated transport functional category (Figure 2) and includes essentially all the genes for which there is either compelling evidence for a Ci transport role for the gene product in the CCM (LCIA, LCI1, and HLA3) or a strong argument for the gene product as a good candidate for Ci transport (CCP1). As with all the clusters, a large proportion (15/35) of the genes in CCM cluster 15 falls into the “unknown” functional category. However, it is notable that 5 of the 15 genes of unknown function in cluster 15 are putative transmembrane proteins.

Within CCM cluster 15, all 17 genes with significant C-effects also exhibit S-effects (includes CAH1, CAH4, CAH5, LCI1, LCR1, CCP1, HLA3, and LCIA), and 8 of these (includes CAH1, CAH4, CAH5, and CCP1) also have a significant CS-effect. Seven genes
that did not exhibit any significant C/S impact effects have expression levels in the lowest 3% of genes, with a mean RPKM lower than 0.023 for all 7 genes. Of the remaining 11 genes in CCM cluster 15 with significant expression levels, all exhibit only S-effect except one (protein ID 520458; also shows CS-effect).

CAs catalyze the reversible hydration of CO₂ to HCO₃⁻, and serve critical roles for the CCM (Moroney et al. 2011) (Figure 3). Among the 9 identified alpha and beta CA genes (Table 3), CAH1, CAH4 and CAH5 fell into CCM cluster 15 and have all 3 significant C+S+CS-effects, as described above. These three CA genes are strongly induced in low CO₂ and thus may be directly involved in the Ci transport and accumulation process of the CCM, or at least in the acclimation to low CO₂. CAH3, the thylakoid lumen CA required for dehydration of stromal HCO₃⁻ (Moroney et al. 2011), exhibits both significant C+S-effects and was placed in cluster 8, which contains genes whose visual expression patterns indicate modest up-regulation in response to limiting CO₂. CA genes CAH8 and CAH9 showed mainly S-effects and fell into CIA5 clusters 11, and 13, respectively, and CAH6 showed both C+S-effects and fell into cluster 12. CAH2 and CAH7 were not identified as DE genes.

We also scrutinized the genes encoding enzymes of the Calvin cycle and the photorespiratory pathway (Spalding 2009), since these important carbon metabolism pathways are expected to respond to CO₂ concentration (Figure 4). Eight of the 15 genes involved in the Calvin Cycle were differentially expressed in our experiment. Two fructose bisphosphate aldolase genes FBA1 and FBA3, the sedoheptulose
bisphosphatase gene \textit{SEBP1}, and one of the two Rubisco small subunit genes, \textit{RBCS1}, were found in CIA5 clusters 11 and 13, both of which show increased gene expression in the \textit{cia5} mutant but relatively little effect of CO\textsubscript{2} concentration. Two critical kinase encoding genes, \textit{PGK1} and \textit{PRK1}, and the fructose bisphosphatase gene \textit{FBP1} were included in cluster 5, which shows a pattern of mildly increasing gene expression with increasing CO\textsubscript{2} concentration, as well as mildly increased gene expression in the \textit{cia5} mutant. The \textit{RPI1} gene was in CO\textsubscript{2} cluster 9, which shows significantly increased expression under higher CO\textsubscript{2} concentration but only modest expression increase in the \textit{cia5} mutant. Thus, aside from \textit{PGK1} and \textit{RPI1}, in which both C-effect and S-effect were detected, all DE genes from the Calvin cycle show only S-effects and increased expression in \textit{cia5}.

On the other hand, the expression of photorespiratory pathway genes was strongly affected by the CO\textsubscript{2} concentration; many of the genes were up-regulated in L-CO\textsubscript{2} and VL-CO\textsubscript{2} (Figure 4). Accordingly, the photorespiratory genes, \textit{AAT1}, \textit{GLYK}, \textit{GYD1}, \textit{HPR1}, \textit{SGA1}, and all glycine decarboxylase complex subunit genes, except \textit{GCSH} and \textit{DLDH1}, fell into CCM cluster 8, even though the Algal Functional Annotation Tool only identified 2 genes, \textit{AAT1} and \textit{HPR}, among the GO hits for photorespiration in cluster 8 (Supplemental Data Set 3 online). These cluster 8 photorespiratory genes, which encode enzymes spanning the entire pathway from glycolate to phosphoglycerate, appear to be regulated by both CIA5 and CO\textsubscript{2}; accordingly, all exhibited C-effects and S-effects, and all, except \textit{AAT1} and \textit{HPR1}, exhibited CS-effects.
Of those photorespiratory pathway genes not in CCM cluster 8, \textit{GCSH} and \textit{SHMT3} were found in CIA5 clusters 10 and 13, respectively, with S-effects only, \textit{SHMT1} was captured in CO$_2$ cluster 7 with only significant C-effect, and \textit{AGT1} was found in cluster 12 with a significant C+S-effect. Some photorespiratory genes, such as the three phosphoglycolate phosphatase genes, \textit{PGP1}, \textit{PGP2} and \textit{PGP3}, were not identified as being differentially expressed in our experiment even though phosphoglycolate phosphatase activity was reported to increase in response to limiting CO$_2$ (Marek and Spalding 1991; Tural and Moroney 2005). One isoform of alanine-glyoxylate transaminase (\textit{AGT2}), one isoform of serine hydroxymethyltransferase (\textit{SHMT2}) and the glycine decarboxylase complex subunit \textit{DLDH1} also were not identified as DE genes under the conditions used.
Identification of DE genes

In this investigation our primary objective was to gain insight into the transcriptome-wide changes in the patterns of gene expression that occur in response to the interaction between CO$_2$ concentration and the transcription regulator CIA5. An additional benefit expected was the identification of candidate genes that may play significant roles in the CCM. To address these objectives, we analyzed the gene expression profiles of two genotypes, WT (137c) and cia5, under 3 different CO$_2$ concentrations by utilizing an overall test to identify 3678 genes that showed differential expression in at least one of the 6 treatments (2 genotypes X 3 conditions). This identification of over 3600 DE genes, which represents almost 20% of the *Chlamydomonas* transcriptome, revealed massive changes in gene expression in response to the combination of CO$_2$ concentration changes and the presence/absence of CIA5.

Further detailed analysis of the 3678 DE genes was performed using 2 additional methods: 1) C/S impact tests for C-effects, S-effects and CS-effects, for each gene; and 2) a cluster analysis of the gene expression patterns across the 6 conditions. Whereas cluster analysis grouped DE genes with similar expression patterns, the C/S impact test provided quantitative evaluations of individual environmental induction and strain effects. The majority of genes identified as DE genes by the overall test showed one or more significant C/S impact effects when tested for C-, S- and CS-effects. Only about 5%
(165 out of 3678) of the DE genes identified by the overall test were not identified as having significant individual effects in the C/S impact test, possibly due to different power of detection inherent in the overall test and the C/S impact test.

Cluster analysis, in combination with identification of individual C-effects, S-effects and CS-effects, revealed clusters of genes regulated primarily by CIA5 (predominantly S-effects; CIA5 clusters), regulated primarily by CO₂ (predominantly C-effects; CO₂ clusters) and regulated by interaction of CO₂ and CIA5 (predominately CS-effects and combinations of C-effects, S-effects, and CS-effects). The delineation of these clusters directly addressed our overall objective of gaining insight into the patterns of gene expression in response to interaction between CO₂ and CIA5, as well as revealing specific genes regulated by CO₂, by CIA5 and by the interaction of CO₂ and CIA5. Based on reports of induction or up-regulation of CCM-related genes in low CO₂, genes functionally involved in the *Chlamydomonas* CCM were expected to be among the third general group of genes, those regulated by both CO₂ and CIA5.

**Comparison with previously reported LCI genes**

Although our major objective was to discover a spectrum of gene expression patterns in response to the interaction between CO₂ and CIA5, we also performed direct pairwise comparisons in the WT strain between H-CO₂ and either L-CO₂ or VL-CO₂ in order to provide a more detailed analysis of the differential expression of genes in
response to low or limiting CO\textsubscript{2}. Because of the historical connection between low-CO\textsubscript{2} up-regulated genes and the CCM, we included these \textit{DESeq} analyses to enrich the comparisons between previously reported LCI genes and DE genes identified in this study.

Because of the interest in the CCM specifically, at least 106 genes have been reported as LCI genes, many of which also reportedly require CIA5 for differential expression (Chen et al. 1996; Somanchi et al. 1999; Miura et al. 2004; Wang et al. 2005; Wang and Spalding, 2006; Yamano and Fukuzawa 2009). Some of these LCI genes, such as \textit{CAH3}, \textit{LCI1}, \textit{LCIA}, \textit{LCIB} and \textit{HLA3} reportedly play important roles in the \textit{Chlamydomonas} CCM (Galvan et al. 2002; Wang and Spalding 2006; Duanmu et al. 2009a; 2009b; Ohnishi et al. 2010), and the function of others, such as \textit{CCP1}, \textit{CCP2}, and \textit{LCIC}, in the CCM also has been implicated (Pollock et al. 2005; Wang and Spalding 2006; Yamano et al. 2010). However, only 53 of 106 previously reported LCI genes were supported as L-CO\textsubscript{2} or VL-CO\textsubscript{2} up-regulated DE genes in this study. This discrepancy is not unexpected, because the different strains and different light, CO\textsubscript{2} concentration and other environmental conditions used among the various studies almost certainly will result in variations in the genes responding, and because DE gene identification may be impacted by a shifting in the population distribution among the cell division cycle (CDC) phases in response to a shift from H-CO\textsubscript{2} to L-CO\textsubscript{2} conditions (Dillard et al. 2011). In addition, the 4 h induction time used here may not identify genes that are differentially expressed only earlier or later than 4 h. Indeed, the companion paper by Brueggeman et al. (2012) documents significant changes in gene expression during a 3 h time course
following CO₂ depletion but also reports the lack of induction of several previously reported LCI genes. Only 40 previously reported LCI genes were supported by Brueggeman et al. (2012) as low-CO₂ up-regulated. In combination with those supported by our data, 60 previously reported LCI genes are supported as up-regulated under the conditions used in the two studies combined.

Many previously reported LCI genes have not been further characterized or confirmed beyond initial observations, which in many cases used no statistical procedure to control false discovery rates (FDR) (Miura et al. 2004; Yamano and Fukuzawa 2009). The greater sensitivity of RNA-Seq and our more reliable statistical approach provide significant advantages over previous studies. Therefore, in addition to the impact of environmental and strain differences on the absence of some LCI genes from our list of DE genes, some previously reported LCI genes may not represent bona fide LCI genes. Our data indicate that 4 previously reported LCI genes are actually down-regulated by L-CO₂ or VL-CO₂, and, in addition to supporting the down-regulation of 3 of these 4 genes, our companion paper (Brueggeman et al. 2012) identified 2 more previously reported LCI genes as down-regulated by low CO₂.

The results reported here and in our companion paper (Brueggeman et al. 2012) complement and extend past reports of differential expression by supporting 60 previously reported LCI genes, directly contradicting 6 others, and leaving the remaining 40 as not clearly supported under the conditions used in the two studies. In addition,
the two companion studies identified a large number of additional genes as regulated by CO₂, CIA5 or both.

**CCM clusters**

Of 57 previously reported LCI genes identified in this study as DE genes, 43 fell into the CCM clusters, 8, 14 and 15, all of which exhibited expression patterns expected for “classic” LCI genes; i.e., high expression for WT in L-CO₂ and VL-CO₂ but lower expression in H-CO₂ and consistently lower expression in cia5 under all CO₂ concentrations. In addition to those in the CCM clusters, 11 of the previously reported LCI genes, including 4 that were down-regulated by low CO₂, fell into CO₂ clusters 7 and 9, which exhibit little impact of the presence/absence of CIA5, suggesting that the genes included are not likely to be involved in the CCM. This illustrates an important value of sorting gene expression patterns into clusters, which provide richer insight into the identification of likely functional CCM genes than provided by the LCI approach alone.

The CCM cluster 15 contains only 35 of the 3678 differentially expressed genes but may be a rich source of candidate functional CCM genes. Of the 35 genes in cluster 15, eight have RPKM expression levels lower than 0.05 across all conditions, making them unlikely candidates for a significant role in the CCM. The remaining 27 genes in CCM cluster 15, which includes all the genes that encode transport proteins strongly implicated or suspected as Ci transporters, as well as the CCM regulatory gene LCR1, the
LCIB-like gene, LCI'E, and the well-studied CA genes, CAH1, CAH4 and CAH5, must be good candidates for a functional or regulatory role in the Chlamydomonas CCM. Considering the burden of transporting one of the highest flux inorganic nutrients, as well as the selection against wasting energy on transport when CO₂ is abundant, such an expression pattern is not surprising for genes encoding Ci transporters and other conditionally critical CCM components.

Among the remaining genes of CCM cluster 15, four encoding stress-induced light harvesting chlorophyll proteins, LHCSR2 and LHCSR3, and DnaJ-like, putative chaperonins, DNJ15 and DNJ31, may represent general stress-response elements, and a few genes, such as 522486 (putative guanylate cyclase), 519249 (putative protein kinase), and 516770 (putative PRLI interacting factor), encode potential signaling elements. However, the most intriguing group of cluster 15 genes may be the six unknown or little-known transmembrane-protein-encoding genes, including 524386, 512353, 510680, 522781, 516290, and 523507, since their expression patterns parallel those of all likely transporter-encoding genes so far identified. Therefore these genes rank high as possible undiscovered Ci transporters.

Key CO₂ assimilation related genes and pathways outside the CCM

CAs are expected to play important roles in microalgae because of the poor solubility and diffusion rate of CO₂ in water and the critical importance of
interconversion of these Ci forms internally. Of the 9 confirmed α-CA and β-CA genes (Moroney et al. 2011), 7 were identified as DE genes in our experiment (Table 3 and Figure 3), which is not unexpected, given the importance of Ci uptake and accumulation in the CCM and the well-known differential expression of the 3 CCM cluster 15 CA genes. S-effect appeared to influence more CA genes than C-effect and CS effect, which reinforces our thoughts about the extent of CIA5 influence. Also, this study confirms CAs as important functional targets for further study regarding the dynamics of the CCM.

The expression of a number of Calvin cycle genes, including one Rubisco small subunit gene, was impacted by the cia5 mutation. DE genes encoding Calvin cycle enzymes were found in clusters 5, 9, 11 and 13, all of which exhibit increased expression in cia5 relative to WT, but which vary in their responses to changes in CO₂ concentration. Reinforcing the implication that the Calvin cycle DE genes respond primarily to CIA5, all these genes exhibited a significant S-effect, with only two, PGK1 (cluster 5) and RPI1 (CO₂ cluster 9) exhibiting a significant C-effect. Although we have no clear explanation for a CIA5 role in regulation of several Calvin cycle genes, the increased expression of these genes in the cia5 mutant argues for a role of CIA5 in carbon assimilation independent of the CO₂ concentration. This observation may reflect a role for CIA5 as a repressor of Calvin cycle genes under as yet unidentified conditions, where the absence of CIA5 activity might result in a modest increase in expression of these Calvin cycle genes.
Photorespiration results from the low specificity of Rubisco and competes with CO₂ assimilation via the Calvin cycle, so lower CO₂ concentrations increase the Rubisco oxidase reaction relative to the carboxylase reaction and increase the demand for photorespiratory enzymes. The apparent regulation of a number of these genes by CIA5 and CO₂ is consistent with previous reports of low-CO₂ induced expression of photorespiratory pathway genes (Marek and Spalding 1991, Tural and Moroney 2005), but, contrary to previous reports, we did not see differential expression of any PGP genes. This discrepancy might be explained if PGP activity is regulated at the post transcriptional level, or if the change in PGP gene expression occurs in a time frame not captured by our 4 h induction time point. Unlike the DE Calvin cycle genes, most of the DE photorespiratory pathway genes exhibited significant C-, S- and CS-effects, rather than just significant S-effects, demonstrating that many of the photorespiratory pathway DE genes are regulated by both CO₂ and CIA5.

Not all photorespiratory genes appear to be regulated by CIA5, but the apparent regulation of a large fraction of them by this protein argues that CIA5 plays a significant role in regulation of the photorespiratory pathway. Notably, those photorespiratory pathway genes that exhibit significant S-effect show a differential expression opposite to that of the Calvin cycle genes, i.e., they have increased expression in WT relative to cia5. This expression pattern across the genotypes is consistent with CIA5 acting as an inducer of photorespiratory pathway genes in contrast to its putative role as a mild repressor of several Calvin cycle genes.
CIA5 clusters and the impact of CIA5

CIA5 appears to serve much broader and more extensive roles than indicated by the phenotype of cia5, which grows similar to WT either heterotrophically or mixotrophically in acetate or photoautotrophically in H-CO₂, and even grows slowly in L-CO₂ (Spalding 2008). Most genes in CIA5 clusters 1, 2, 3, 10, 11 and 13 show clear regulation by CIA5 but little regulation by CO₂ concentration, indicating that low-CO₂ activation of CIA5 is not always required for function of CIA5. More than 76% of the 1396 genes in these 6 clusters exhibit only a significant S-effect (no C-effect or CS-effect), and, including genes that also show a significant C-effect or CS-effect, almost 95% exhibit a significant S-effect. Furthermore, of 3678 identified DE genes, over 62% show a significant S-effect, including those that also exhibit a significant C-effect and/or CS-effect, and more than half of those genes regulated directly or indirectly by CIA5 show only a significant S-effect. Thus, almost 15% of all Chlamydomonas genes are regulated in some way by CIA5, and almost 7.5% of all genes are regulated by CIA5 independent of any changes in the CO₂ concentration.

CIA5 is very likely involved in the upstream regulation of multiple physiological processes, and, although its own transcript abundance does not change, the presence/absence of CIA5 (and its potential activation/inactivation) may have a major impact on the expression of genes encoding many secondary regulatory genes, including those encoding transcription factors and signal transduction components involved in
regulation of a number of processes. Consistent with this expectation, manual curation and, in most cases, the Functional Annotation Tool, identified signaling or signaling plus gene expression as key functional categories for the 3 CIA5 clusters with increased expression in the WT, arguing that when the presence of CIA5 increases transcript abundance for specific genes, it appears to do so as an upstream activator of positive signaling pathways and/or other gene expression activators. On the other hand, both manual curation and the Functional Annotation Tool pointed to metabolism as the primary functional category for the 3 CIA5 clusters in which the presence of CIA5 resulted in decreased expression of specific genes (decreased expression in the WT). Thus, when the presence of CIA5 coincides with the repression of specific genes, CIA5 appears to act as an upstream repressor of specific metabolic functions. Notably, half of the Calvin cycle genes identified as differentially expressed fall into these 3 CIA5 clusters, and most of the others fall in cluster 5, which has somewhat similar characteristics, including decreased gene expression in WT and metabolism as primary functional category.

**CO₂ clusters and the impact of CO₂**

Regulation of gene expression by CO₂ concentration also appears to be more extensive than expected. Most genes in the CO₂ clusters 4, 7 and 9 show clear CO₂ concentration regulation, but little or no apparent effect of CIA5. Almost 55% of the 764 genes in these 3 clusters exhibit only a significant C-effect (no S-effect or CS-effect), and,
when including those genes that also show a significant S-effect or CS-effect, over 90% exhibit a significant C-effect. Furthermore, of 3678 identified DE genes, over 60% show a significant C-effect, including those that also exhibit a significant S-effect and/or CS-effect, and approximately 30% of those genes regulated directly or indirectly by CO₂ concentration show only a significant C-effect. Thus CO₂ concentration significantly affects the expression of over 60% of the DE genes and more than 14% of all genes detected in this experiment. This means that about 14% of all *Chlamydomonas* genes are regulated by CO₂ concentration, most of which (over 10% of the genes) also exhibit some form of CIA5 regulation. However, almost 4% of all genes are regulated by CO₂ apparently independent of CIA5.

Manual curation and the Functional Annotation Tool both identified gene expression as the key functional category for the CO₂ cluster (cluster 9) with decreased expression of genes in L-CO₂ or VL-CO₂ relative to H-CO₂, which is consistent with either limiting CO₂ acting as an upstream repressor or elevated CO₂ acting as an upstream activator, respectively, of genes involved in regulation of gene expression. On the other hand, manual curation and, in one case, the Functional Annotation Tool, identified metabolism as the primary functional category for the 2 CO₂ clusters (clusters 4 and 7) with increased expression of genes in L-CO₂ or VL-CO₂ relative to H-CO₂, which is consistent with either limiting CO₂ acting as an upstream activator or elevated CO₂ acting as an upstream repressor, respectively, of genes involved in specific metabolic functions. Notably, both functional characterizations of cluster 4 revealed enrichment in putative catabolic genes, suggesting that low CO₂ concentrations cause starvation and
stimulate the expression of genes involved in degrading and remobilizing existing molecules.

More generally, the abundance of metabolism as a primary functional category in 9 of the 16 gene clusters, including CIA5 clusters 2, 11 and 13, CO₂ clusters 4 and 7, and CCM clusters 8 and 15 suggests that a major impact of changes in CO₂ and CIA5 is on the expression of genes encoding metabolic enzymes. This seems reasonable, since large changes in metabolism may well accompany substantial changes in CO₂ concentration. These conjectures also are strongly supported by Brueggeman et al. (2012) who report a marked decrease in expression of numerous genes involved in anabolic processes following CO₂ deprivation. Alternatively, the high frequency of metabolism as a primary functional category also may reflect an annotation bias; metabolism-related genes may be easier to annotate, resulting in their disproportionate representation among the manually annotated genes.

**Multiple acclimation states**

Although there is compelling evidence demonstrating a distinction between the VL-CO₂ and L-CO₂ acclimation states of *Chlamydomonas*, none of our transcriptome analyses were able to identify any genes differentially expressed between the VL-CO₂ and L-CO₂ induction conditions for WT. Since a large number of DE genes were identified for L-CO₂ or VL-CO₂ vs. H-CO₂ conditions, the apparent absence of DE genes in the VL-
CO₂ vs. L-CO₂ comparison likely reflects at least a paucity of DE genes distinguishing these two acclimation states under the conditions used. Based on our data, we suggest at least two possible conclusions regarding the distinction between the L-CO₂ and VL-CO₂ acclimation states: 1) these two acclimation states are controlled at levels beyond transcript abundance, or 2) differential expression of genes distinguishing these two acclimation states is evident only earlier or later than the 4 h acclimation time used in our experiment. On the other hand, if we assume a very limited number of genes are differentially expressed in VL-CO₂ vs. L-CO₂ under our experimental conditions, our one-time test of over fifteen thousand genes may have elevated the problem of multiple testing and reduced our power of detection. To test this assumption, future experiments could increase the number of biological replicates and/or sequence to a greater depth of coverage.

**Summary**

This transcriptome study resulted in a number of new insights regarding the global regulation of genes by CO₂ concentration, by CIA5 and by the combination or interaction of CO₂ and CIA5. Gene expression patterns were classified into distinct clusters, many of which could be characterized as responding primarily to CIA5 or CO₂ based on the C/S impact test and visual inspection of the expression patterns. Regulation by CIA5 independent of CO₂ demonstrates that low-CO₂ activation of CIA5 is not essential for its function. Three distinct gene expression clusters with response to
both CIA5 and CO₂ were clearly associated with CCM-related genes and may prove to
represent a rich source of candidates for new CCM components, especially cluster 15,
which may contain a significant number of putative CCM-related transporter genes.
These expression pattern clusters should also represent a more robust source of insight
than the LCI gene approach with regard to the role of CIA5 and CO₂ regulation on
limiting CO₂ acclimation responses in general and on the function of the CCM
specifically. An example of this is the indication, based on expression patterns observed
with Calvin cycle genes and photorespiratory genes, that CIA5 may act as an upstream
activator of photorespiratory genes and a mild upstream repressor of Calvin cycle genes.
Thus this study of transcriptome-wide patterns of gene expression related to CO₂ and
CIA5 provides insight into the massive impact of these two factors and their interaction
on gene expression in *Chlamydomonas*, in addition to identifying compelling new
candidates for functional CCM components and highlighting new questions to be
addressed in subsequent work.
Materials and Methods

Chlamydomonas strains and culture conditions

*Chlamydomonas reinhardtii* wild-type strain cc125 was obtained from the Chlamydomonas Resource Center (University of Minnesota, Minneapolis). The cia5 mutant (strain cc2702) was a gift from Dr. Donald P. Weeks (University of Nebraska-Lincoln). Media and growth conditions for *Chlamydomonas* were described previously (Wang and Spalding, 2006). All strains were maintained on CO\(_2\) minimal plates in high CO\(_2\) (air enriched with 5% CO\(_2\)) chambers at room temperature, under continuous illumination (50 \(\mu\)mol photons m\(^{-2}\) s\(^{-1}\)). Liquid cultures were grown on a gyratory shaker (speed 200 rpm) under white light (approximately 100 \(\mu\)mol photons m\(^{-2}\) s\(^{-1}\)) at room temperature.

**Induction and RNA isolation**

Liquid cultures of all strains were grown under H-CO\(_2\) (5% v/v CO\(_2\)) to a concentration of 1.0 to 2.0 million cells per mL. Cell cultures were then equally distributed into 9 flasks, with 3 flasks each aerated with H-CO\(_2\) (5%), L-CO\(_2\) (nominally 300~500 ppm; actually 330 to 410 ppm), or VL-CO\(_2\) (nominally 100~200 ppm; actually 110 to 152 ppm). Selection of gas flow lines and position on the shaker were completely random. After 4 h induction, cultures aerated with the same CO\(_2\) concentrations were combined and centrifuged to collect cells. Two biological replicates of each strain and of
each induction condition were processed for RNA isolation as previously described (Wang and Spalding, 2006), and crude RNA samples were cleaned with DNase I and RNeasy MinElute Cleanup kit (Qiagen Company; Catalog number 79254 and 74204).

**Sequencing and alignment**

The DNA Facility at Iowa State University processed the cleaned RNA samples, prepared the libraries and generated sequences (one sample per lane on the flow cell) on a Genome Analyzer II system (Illumina). Not less than 12 million reads were obtained for each sample (details can be found in Supplemental Table 1 online). Raw and processed sequence files are available at the NCBI Gene Expression Omnibus (accession GSE33927).

Raw reads were aligned to the version 4.0 assembly of the *Chlamydomonas reinhardtii* genome (http://genome.jgi-psf.org/chlamy/chlamy.info.html) using gmap (Wu and Watanabe, 2005), which tries to align every single read to the genome as an mRNA (possibly spliced) read without previous knowledge of the genome annotation or sequencing coverage. Alignment files were processed to retrieve unique alignments with an alignment score higher than $r-15$, where $r$ is the read length. This choice allows us to a) keep two-blocks alignments (reads that span two different exons) where only one block is reported by gmap (that typically fails to provide alignment blocks smaller than 16 because of its indexing strategy) and b) keep alignments for trimmed reads with
sequencing errors at the 3’ end. Only high-similarity (less than three mismatches) and intron-like alignments (defined here as those with up to one gap smaller than 5kb) were used for expression estimation.

Counts per gene were estimated by requiring complete overlap between each alignment and the transcript genomic coordinates and after normalization of transcript sequence coverage by read length for each sample. This strategy attempts to remove the impact that different read lengths could have on the final results, but almost identical results were found with alternative methods (see the end of Methods). Augustus 5.0 gene models (http://augustus.gobics.de/predictions/chlamydomonas/) were used as the reference annotation in this work. The original annotation was filtered to keep only the first prediction for each locus (“_t1” transcripts). Total number of sequences, percent of uniquely aligned reads and total gene counts per sample are provided in Supplemental Table 1 online. Gene counts were used for differential expression analysis. Expression estimates for each sample are provided in units of RPKM (reads per kilobase of exon model per million of aligned reads; Mortazavi et al. 2008).

Normalization and Statistical analysis

The primary statistical analysis was performed in the R statistical programming language (version 2.10.1, from http://www.r-project.org/). A generalized linear model (GLM) with negative binomial distribution and logarithm link function was fitted to the
counts of reads separately for each gene while the sequencing depths were used as the offsets for normalization purpose. Fixed factors in the model included CO$_2$ condition, strain and their interaction. An overall test was conducted to identify genes with differential expression in any of the 6 treatment groups. This test was performed by comparing the full model with 6 separate means and a reduced model with the same mean for all 6 groups using quasi-likelihood based approach. The set of $p$-values from the overall test for each gene was adjusted for false discovery rate (FDR) control as described by Benjamini and Hochberg (1995). For each of the genes identified as differentially expressed by the overall test while FDR was controlled at 2.5%, we further tested (C/S impact test) for CO$_2$ effect (C-effect), strain effect (S-effect), and CO$_2$ and strain interaction effect (CS-effect), respectively. These C/S impact tests were conducted by comparing the full model with the appropriate reduced models using quasi-likelihood based $F$-test with the $R$ function “drop1”. FDR was controlled for each set of $p$-values using Benjamini and Hochberg’s method.

The DE genes identified by the overall test were clustered by a model-based clustering method implemented in an $R$ package, $MBCluster.Seq$ (Si et al., 2011), assuming the observed counts following negative binomial distributions. Results were evaluated for variations in the number of clusters from 10 to 30. To maintain the clusters as few and tight as possible while including most of the patterns, 16 was chosen to be the total number of clusters for further analysis by visual inspection of the clustering results.
Pair-wise comparisons between VL-CO$_2$ and L-CO$_2$ states, VL-CO$_2$ and H-CO$_2$ states or L-CO$_2$ and H-CO$_2$ states were performed using the Bioconductor package \textit{DESeq} (Anders and Huber, 2010), which performs variance stabilization by borrowing information across genes (Anders and Huber, 2010). The set of $p$-values for each test was adjusted for false discovery rate (FDR) control as described in Benjamini and Hochberg (1995).

\textbf{Quantitative real time PCR analysis}

SYBR green one step quantitative PCR system was used for qPCR analysis (Quanta BiosciencesTM, catalog number: 95086-50). All experiments were performed on a Bio-rad iCycler iQ Real Time PCR detection system (catalog number: 170-8740) using primers described in Supplemental Table 2 online. The RNA samples used as templates for qPCR were the same as those used for RNA-Seq. The CBLP gene was used as internal control for normalization of qPCR data. Pearson correlations were calculated for each gene across 6 strain-treatment conditions between RNA-Seq and qPCR methods, based on average log$_2$ fold change of 2 biological replicates.

\textbf{Functional categorization for DE genes using the Pathways Tool}

The Algal Functional Annotation Tool (Lopez et al. 2011) was used to investigate the biological processes associated with each cluster. To address the limitation of
annotations availability for *Chlamydomonas reinhardtii*, we used the Gene Ontology terms (GO terms) based on orthology to *Arabidopsis thaliana* as the framework for GO term selection. After inputting the gene list for each cluster separately, *p*-values were generated for each GO term for each cluster entry. GO terms lacking any hits were assigned a *p*-value of “1”, and other terms not statistically significant (*p* >0.01) for any cluster entry were excluded in generating the summary heat map (Supplemental Figure 7 online).

**Manual curation for DE genes**

The principles we followed when manually curating these genes were: 1) manual annotation was used if available; 2) if there was no manual annotation, the automated annotation domain information based on the Augustus 5.0 gene model was used to guide curation of the gene; 3) if there was no manual annotation and no identified domains from the automated annotation, the gene was marked as “unknown”, or as “unknown transmembrane”, if one or more transmembrane regions were predicted; 4) any domain information provided by automated annotation of the gene models was used to place the genes into broad functional categories. The categories used included general biological pathways and general protein functions, such as “metabolism” or “signaling”. From this process, we placed all genes into 8 general categories reflective of putative function: 1) signaling (including protein kinases, cyclic nucleotide synthesis and metabolism, etc.); 2) gene expression and regulation (including transcription,
translation, RNA processing, chromatin structure and dynamics, etc.); 3) transport (including Ci transport, ion transport, metabolite transport, etc.); 4) metabolism (including amino acid, photosynthesis, photorespiration, carbohydrate, acetate, lipid, macronutrients, etc.); 5) stress and cell death (including oxidative stress, autophagy, programmed cell death, etc.); 6) cell structure and function (including cell wall, cytoskeleton, vesicular trafficking, protein trafficking, cell division, cell motility, etc.); 7) protein modification and regulation (including proteases, protein modifications other than kinases, etc.); 8) unknown.

**An alternative RNA-Seq data evaluation for reproducibility and comparative analysis.**

Here we provide additional data regarding the quality and reproducibility of our RNA-Seq libraries, along with a comparative analysis with the results from our companion paper (Brueggeman et al. 2012). To this end, we have used a different, simplified pipeline for all data in order to highlight the biological differences and remove any potential differences due to the computational methods. Specifically, we performed the steps below:

In order to remove any dependency with the various parameters involved in gapped-alignment algorithms, we first obtained non-gapped alignments to Augustus 5 transcript sequences using *bowtie*. The potential issues introduced by this approach (slightly worse alignment rates, missing annotations in the genome with sequence
similarity to the annotated genome) are not expected to make a difference in the expression and differential expression results for most genes.

Trimming sequences: in order to remove the impact that different read length and error rates could have in comparing different RNA-Seq libraries, the results below correspond to trimmed libraries (60bp) showing very similar error rate profiles (around 1% at the 3’ end, data not shown). The rate of unique alignment to Augustus 5 transcripts is in the range of 70-80% for all libraries (See Supplemental Table 1 online).

Unique hits from bowtie were compiled to build the count matrix per gene per sample, for both our libraries and those from our companion paper. The previous matrix was normalized to compute expression estimates in unit of RPKMs (reads per kilobase of exon model per million of aligned reads). This normalization was performed after imputation of missing values (0’s were imputed a value of 1 count, in order to regularize fold changes and differential expression estimates), and filtering those genes with no counts in any sequencing lane (absent or non-mappable genes).

Some remarks that can be made from these normalized values follow:

**High-expression tails:** as normalized RPKMs values provide relative expression measures, one of the main sources of ambiguity when comparing two different RNA-Seq datasets comes from the distribution of the high-expression genes. Very small changes in the high-expression tail of the distribution changes significantly the estimates for moderately and slightly expressed genes. As an example, we can compute the fraction
of the total RPKMs corresponding to the top 100 highly expressed transcripts. For those
conditions that are similar in both manuscripts, the numbers are:

Fang/Spalding dataset High CO2 #1 = 0.4117

Fang/Spalding dataset High CO2 #2 = 0.4650

Fang/Spalding dataset Very Low CO2 #1 = 0.3849

Fang/Spalding dataset Very Low CO2 #2 = 0.4445

Brueggeman/Ladunga dataset 0 hours #1 = 0.4528

Brueggeman/Ladunga dataset 0 hours #2 = 0.5357

Brueggeman/Ladunga dataset 3 hours #1 = 0.6786

Brueggeman/Ladunga dataset 3 hours #2 = 0.7031

For instance, the last number means that 70% of the RPKMs come from the top 100
genes in the 3 hours sample (second replicate) of the Brueggeman et al. 2012 dataset.
This observation, in its turn, will clearly impact the mean-variance distributions so that
the number of reported differentially expressed genes is potentially different. The
biological interpretation of this difference in the expression distributions can be found
in both manuscripts, and can be easily understood from the differences in experimental
conditions and/or genetic background.
Reproducibility. A high correlation between fold-change estimates from both RNA-Seq and PCR is shown and discussed in this manuscript for a number of relevant genes. Regarding the reproducibility of our RNA-Seq expression estimates for different replicates, Supplemental Figure 3 online shows mean-difference (MD) scatter plots for replicates of the same condition in our manuscript. The x-axis values are geometric means of the expression of the two replicates, while the y-axis shows the fold change between replicates. To assist on the visualization of these graphs, shown are line plots with the 90th percentile (red), mean (green) and 10th percentile (cyan) of the fold changes. This means that above the red line fall the 10% higher fold changes between replicates and below the cyan is the 10% higher negative fold changes between replicates. These plots show both that the replicates fold change distributions are centered around 0 as expected (green lines), and that the expression estimates from both replicates are consistent for a majority of the transcriptome. Very similar results were found for the experiments from our companion paper (Brueggemen et al. 2012). It is worth mentioning that the highest reproducibility is observed for the mutant samples, most likely due to a lower sensitivity to fluctuations in CO$_2$ levels.

**Fold-change comparative plots:** fold changes between control and experiment were obtained from the mean expression estimates across replicates for each condition. We focus here on those comparisons that are common to both manuscripts (very low/high CO$_2$ for our dataset and 0 versus 3 hours on the Brueggeman et al. 2012 dataset). Supplemental Figure 4 online compares both fold change estimates. The left panel shows a scatter plot along with results from a linear fit. The black line is a guide to
the eye, representing an ideal linear relationship (slope=1, no offset, expected if both experiments were completely equivalent). A linear fit provides the results highlighted in red font, with slope 1.13 and offset = -0.56. The correlation between both fold changes distribution is 0.60. Together, these results indicate that both datasets have the same whole-transcriptome trends in fold expression. The same data is shown in the right panel of Supplemental Figure 4 online as a smoothed density plot. It is clear that fold changes for a majority of genes are in close agreement between both datasets, in accordance with the discussion in the manuscripts regarding the similarity between the reported regulated genes.

**Fold-change comparative plots for different analysis pipelines:** Supplemental Figure 5 online shows scatter plots of log$_2$ fold changes estimated for data in this manuscript with two different analysis pipelines. The x-axis corresponds to the fold change estimates presented in the main manuscript while the y-axis plots estimates from the pipeline introduced above. Red lines represent a perfect linear relationship. The agreement between both estimates is apparent for the whole fold change dynamic range.

**Accession Numbers**

All sequence data from this article can be found in the EMBL/GenBank data libraries under accession numbers listed in Supplemental Data Set 2 online. Accession numbers
for specifically discussed genes are: AAT1, EDP08011; AGT1, EDO97315; AGT2, EDO96807; CAH1, EDP04241; CAH3, EDP00852; CAH4, EDO96058; CAH5, EDP07024; CAH6, EDO96552; CAH7, EDO99006; CAH8, EDO99999; CAH9, EDP07163; CCM1/CIA5, EDP07542; CCP1, EDP04147; CCP2, EDP04238; DLDH1, EDP01871; DNJ15, EDP03107; DNJ31, EDO98634; FBA1, EDO98285; FBA3, EDO97897; FBP1, EDP05318; GCSH, EDP08614; GLYK, EDP03009; GYD1, EDP01639; HLA3, EDP07736; HPR1, EDP05213; LCI1, EDP06069; LCI6, EDP02960; LCIA/NAR1.2, EDP04946; LCIB, EDP07837; LCIC, EDP04956; LCID, EDP04142; LCIE, EDP04243; LCR1, BAD13492; LHCSR2, EDP01013; LHCSR3, EDP01087; PGK1, EDO98586; PGP1, EDP06184; PGP2, EDP05829; PGP3, EDP08194; PRK1, EDP02974; RBCS1, EDO96904; RHP1, EDP01722; RHP2, EDP01723; RPI1, EDP04506; SEBP1, EDP04487; SGA1, EDO97196; SHMT1, EDO97448; SHMT2, EDO97351; SHMT3, EDP00905. *Chlamydomonas* strains (available from the *Chlamydomonas* Stock Center - http://chlamycollection.org/) used in this work: the 137c wild type (strain cc125); *cia5* mutant (strain cc2702).

**Supplemental Data**

The following materials are available in the online version of this article.

**Supplemental Figure 1.** Gene expression level distributions for each treatment condition.

**Supplemental Figure 2.** Validation of RNA-Seq by qPCR.
**Supplemental Figure 3.** Mean-difference scatter plots for biological replicates.

**Supplemental Figure 4.** Comparison of log2 fold change estimates between different datasets.

**Supplemental Figure 5.** Comparison of log2 fold change estimates between different analysis pipelines.

**Supplemental Figure 6.** Distribution of C/S impact test results by cluster.

**Supplemental Figure 7.** Heat map for GO category hits based on the Algal Functional Annotation Tool.

**Supplemental Table 1.** Alignment Statistics for the transcriptome sequencing experiment.

**Supplemental Table 2.** List of qPCR primers.

**Supplemental Data Set 1.** Overall and C/S Impact Test.

**Supplemental Data Set 2.** DESeq summary.

**Supplemental Data Set 3.** Gene Ontology analysis.

**Supplemental Data Set 4.** Manual curation of genes.

**Supplemental Data Set 5.** Previously reported LCI genes.

**Supplemental Data Set 6.** Genes in CCM clusters 8, 14 and 15.
ACKNOWLEDGMENTS

This work was supported by the U.S. Department of Agriculture National Research Initiative Competitive Grants (2007-35318-18433), the National Science Foundation (MCB-0952323), and the Department of Energy Advanced Research Projects Agency-Energy Program (DEAR0000010) to M.H.S., the Institute of Genomics and Proteomics (Department of Energy Cooperative Agreement DE-FC02-02ER63421 to David Eisenberg) and National Institutes of Health (R24GM092473) to S.M. and M.P.
Figures and Tables

Figure Legends

Figure 1. Clustering. The expression patterns of 3678 selected DE genes for 16 clusters.
The horizontal axis indicates each strain and CO2 induction condition: WV = wild type under VL-CO2 induction; WL = wild type under L-CO2 induction; WH = wild type under H-CO2 induction; MV = cia5 under VL-CO2 induction; ML = cia5 under L-CO2 induction; MH = cia5 under H-CO2 induction. The vertical axis indicates the log2 fold change calculated between each condition and the average across all six conditions. Each gray line symbolizes the expression pattern of one gene, and the bold back line illustrates the average expression pattern of all genes in each cluster.

Figure 2. Distribution of genes in manual functional categories within each cluster. The functional categories (protein modification and regulation, cell structure and function, stress and cell death, metabolism, transport, gene expression and regulation, and signaling) were determined manually based on a combination of existing annotation and automated identification of functional domains. Percentages indicate the sum of genes in each functional category (indicated by color). The difference between the summed percentage and 100% represents the “unknown” category, which is not included.

Figure 3. Gene cluster distribution within a hypothetical CCM/Ci transport model. The cluster label for each gene is indicated by a yellow octagon containing a cluster number. In this schematic representation of the CCM, most of the putative CCM elements and
many previously identified CO$_2$-responsive genes, including the putative Ci transporters *HLA3, LCI1, LCIA, CCP1* and *CCP2* and CCM related carbonic anhydrases *CAH1* and *CAH3*, and other identified CCM genes like *LCIB* and *LCIC* are indicated as having expression patterns categorized in the “CCM clusters (clusters 8, 14, and 15)”. This figure is modified from Wang et al. (2011), with the addition of the low-CO$_2$ inducible chloroplast membrane proteins, CCP1 and CCP2, (Chen et al. 1997), and carbonic anhydrase locations (Moroney et al. 2011). Transport proteins with demonstrated CCM function are shown in dark blue, known transport proteins with suspected CCM function are shown in purple, predicted but unknown transport proteins are shown in light blue, known soluble enzymes are shown in red and the soluble LCIB/LCIC complex is shown in orange and yellow.

**Figure 4. Genes involved in the Calvin Cycle and photorespiration.**

(A) Schematic of the Calvin Cycle and photorespiration pathway in *Chlamydomonas* (Spalding 2009). Yellow boxes indicate genes involved in each reaction, the red numbers in parentheses indicate the cluster in which each gene was found.

(B) Summary of detailed information about each gene. C = cluster number; S.E. = Significant Effect which are individual effects having a q-value <0.025 by the C/S impact test, where C = CO$_2$ effect, S = strain effect, and CS = CO$_2$ and strain interaction effect.
Figures:

Figure 1: Clustering.
Figure 2: Distribution of Genes in Manual Functional Categories within Each Cluster.

Figure 3: Gene Cluster Distribution within a Hypothetical CCM/Ci Transport Model.
Figure 4: Genes Involved in the Calvin Cycle and Photorespiration.
### Table 1. Gene Ontology Categories of Clusters

Summary of the identified GO terms and primary functional categories within the 16 clusters using the Algal Functional Annotation tool and manual curation, respectively.

<table>
<thead>
<tr>
<th>Clusters</th>
<th>Total Genes</th>
<th>Cluster Group</th>
<th>GO Terms</th>
<th>Unique Genes</th>
<th>% Unique Genes</th>
<th>Main GO Terms</th>
<th>Primary Functional Category</th>
<th>% of Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>124</td>
<td>CIA5</td>
<td>22</td>
<td>3</td>
<td>2.4%</td>
<td>N/S</td>
<td>Signaling</td>
<td>16%</td>
</tr>
<tr>
<td>2</td>
<td>95</td>
<td>CIA5</td>
<td>9</td>
<td>4</td>
<td>4.2%</td>
<td>N/S</td>
<td>Metabolism</td>
<td>19%</td>
</tr>
<tr>
<td>3</td>
<td>387</td>
<td>CIA5</td>
<td>25</td>
<td>8</td>
<td>2.1%</td>
<td>N/S</td>
<td>Signaling</td>
<td>14%</td>
</tr>
<tr>
<td>4</td>
<td>150</td>
<td>CO₂</td>
<td>45</td>
<td>15</td>
<td>10.0%</td>
<td>Catabolic processes</td>
<td>Metabolism</td>
<td>24%</td>
</tr>
<tr>
<td>5</td>
<td>187</td>
<td>-</td>
<td>33</td>
<td>60</td>
<td>32.1%</td>
<td>Biosynthetic processes</td>
<td>Metabolism</td>
<td>29%</td>
</tr>
<tr>
<td>6</td>
<td>240</td>
<td>-</td>
<td>10</td>
<td>30</td>
<td>12.5%</td>
<td>RNA modification; protein localization</td>
<td>Gene expression-regulation</td>
<td>23%</td>
</tr>
<tr>
<td>7</td>
<td>243</td>
<td>CO₂</td>
<td>4</td>
<td>7</td>
<td>2.9%</td>
<td>N/S</td>
<td>Metabolism</td>
<td>16%</td>
</tr>
<tr>
<td>8</td>
<td>360</td>
<td>CCM</td>
<td>6</td>
<td>8</td>
<td>2.2%</td>
<td>N/S</td>
<td>Metabolism</td>
<td>12%</td>
</tr>
<tr>
<td>9</td>
<td>371</td>
<td>CO₂</td>
<td>23</td>
<td>49</td>
<td>13.2%</td>
<td>RNA processes, nitrogen metabolism</td>
<td>Gene expression-regulation</td>
<td>28%</td>
</tr>
<tr>
<td>10</td>
<td>460</td>
<td>CIA5</td>
<td>34</td>
<td>63</td>
<td>13.7%</td>
<td>Intracellular trafficking, proteolysis, regulation of processes</td>
<td>Signaling</td>
<td>19%</td>
</tr>
<tr>
<td>11</td>
<td>183</td>
<td>CIA5</td>
<td>10</td>
<td>15</td>
<td>8.2%</td>
<td>Small molecule metabolic processes</td>
<td>Metabolism</td>
<td>22%</td>
</tr>
<tr>
<td>12</td>
<td>463</td>
<td>-</td>
<td>1</td>
<td>3</td>
<td>0.7%</td>
<td>N/S</td>
<td>Signaling</td>
<td>13%</td>
</tr>
<tr>
<td>13</td>
<td>147</td>
<td>CIA5</td>
<td>2</td>
<td>3</td>
<td>2.0%</td>
<td>N/S</td>
<td>Metabolism</td>
<td>32%</td>
</tr>
<tr>
<td>14</td>
<td>138</td>
<td>CCM</td>
<td>6</td>
<td>3</td>
<td>2.2%</td>
<td>N/S</td>
<td>Signaling</td>
<td>14%</td>
</tr>
<tr>
<td>15</td>
<td>35</td>
<td>CCM</td>
<td>3</td>
<td>1</td>
<td>2.9%</td>
<td>N/S</td>
<td>Metabolism</td>
<td>20%</td>
</tr>
<tr>
<td>16</td>
<td>95</td>
<td>-</td>
<td>1</td>
<td>3</td>
<td>3.2%</td>
<td>N/S</td>
<td>Metabolism</td>
<td>20%</td>
</tr>
<tr>
<td>Name</td>
<td>ProteinID</td>
<td>Description</td>
<td>Average RPKM</td>
<td>Significant Effects</td>
<td>Primary Functional Category</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------</td>
<td>-----------</td>
<td>-------------------------------------------------------</td>
<td>--------------</td>
<td>---------------------</td>
<td>-----------------------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAH1</td>
<td>522126</td>
<td>Carbonic anhydrase, periplasmic, α type</td>
<td>1223.8</td>
<td>C+S+CS</td>
<td>Metabolism</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAH4</td>
<td>522732</td>
<td>Mitochondrial carbonic anhydrase, β type</td>
<td>226.5</td>
<td>C+S+CS</td>
<td>Metabolism</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAH5</td>
<td>522733</td>
<td>Mitochondrial carbonic anhydrase, β type</td>
<td>157.3</td>
<td>C+S+CS</td>
<td>Metabolism</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CCL28</td>
<td>510019</td>
<td>Low-CO2-inducible chloroplast envelope protein</td>
<td>200.7</td>
<td>C+S+CS</td>
<td>Transport</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CYC6</td>
<td>516039</td>
<td>Cytochrome c6</td>
<td>1.2</td>
<td>S</td>
<td>Metabolism</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DNI15</td>
<td>514023</td>
<td>DnaJ-like protein</td>
<td>12.8</td>
<td>S</td>
<td>Gene Expression</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DNI31</td>
<td>518238</td>
<td>DnaJ-like protein</td>
<td>13.1</td>
<td>S</td>
<td>Gene Expression</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>*HFO7</td>
<td>523344</td>
<td>Histone H4</td>
<td>&lt;0.1</td>
<td></td>
<td>Gene Expression</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HLA3</td>
<td>518934</td>
<td>ABC transporter</td>
<td>289.5</td>
<td>C+S</td>
<td>Transport</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>*KIR1</td>
<td>526069</td>
<td>Ketoacid isomerase-like protein</td>
<td>&lt;0.1</td>
<td></td>
<td>Unknown</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LCI1</td>
<td>520703</td>
<td>Low-CO2-inducible membrane protein</td>
<td>224.0</td>
<td>C+S</td>
<td>Transport</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LCI23</td>
<td>523507</td>
<td>Low-CO2-inducible protein</td>
<td>75.8</td>
<td>C+S</td>
<td>Unknown(TM)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LCIE</td>
<td>522129</td>
<td>LCIB-like gene</td>
<td>1.6</td>
<td>S</td>
<td>Metabolism</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LCR1</td>
<td>519760</td>
<td>Low-CO2 response regulator</td>
<td>79.0</td>
<td>C+S</td>
<td>Gene Expression</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LHCSR2</td>
<td>525378</td>
<td>Stress-related chlorophyll a/b binding protein 2</td>
<td>45.7</td>
<td>C+S+CS</td>
<td>Metabolism</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LHCSR3</td>
<td>525376</td>
<td>Stress-related chlorophyll a/b binding protein 3</td>
<td>51.7</td>
<td>C+S+CS</td>
<td>Metabolism</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NAR1.2</td>
<td>524076</td>
<td>Anion transporter; LCIA</td>
<td>209.7</td>
<td>C+S</td>
<td>Transport</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>516770</td>
<td></td>
<td>PRLI-interacting factor L</td>
<td>44.9</td>
<td>S</td>
<td>Signaling</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>509757</td>
<td></td>
<td>Acetyltransferase</td>
<td>27.5</td>
<td>C+S</td>
<td>Unknown</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>519249</td>
<td></td>
<td>Serine/threonine protein kinase</td>
<td>53.5</td>
<td>S</td>
<td>Signaling</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>522781</td>
<td></td>
<td></td>
<td>4.85</td>
<td>S</td>
<td>Unknown(TM)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>516290</td>
<td></td>
<td></td>
<td>51.5</td>
<td>C+S</td>
<td>Transport</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>510680</td>
<td></td>
<td></td>
<td>44.6</td>
<td>C+S</td>
<td>Unknown(TM)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>520458</td>
<td></td>
<td></td>
<td>20.7</td>
<td>S+CS</td>
<td>Unknown</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>512353</td>
<td></td>
<td></td>
<td>16.8</td>
<td>C+S+CS</td>
<td>Unknown(TM)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>522486</td>
<td></td>
<td></td>
<td>8.7</td>
<td>C+S</td>
<td>Signaling</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>524386</td>
<td></td>
<td></td>
<td>6.4</td>
<td>S</td>
<td>Unknown(TM)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>524387</td>
<td></td>
<td></td>
<td>2.3</td>
<td>S</td>
<td>Unknown</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>*</td>
<td>512735</td>
<td></td>
<td>&lt;0.1</td>
<td>S</td>
<td>Unknown</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>*</td>
<td>519540</td>
<td></td>
<td>&lt;0.1</td>
<td></td>
<td>Unknown</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>*</td>
<td>522103</td>
<td></td>
<td>&lt;0.1</td>
<td></td>
<td>Unknown</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>*</td>
<td>510710</td>
<td></td>
<td>&lt;0.1</td>
<td></td>
<td>Signaling</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>*</td>
<td>518019</td>
<td></td>
<td>&lt;0.1</td>
<td></td>
<td>Unknown</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>*</td>
<td>511100</td>
<td></td>
<td>&lt;0.1</td>
<td></td>
<td>Unknown</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* ProteinID = Augustus 5.0 gene model protein ID; Average RPKM = average RPKM across all 6 treatment conditions; Significant Effect = individual effect having a q-value <0.025 by C/S impact test, where “C” means CO2 effect, “S” means strain effect, and “CS” means interaction effect; General Category = result from our manual curation, where “unknown(TM)” = gene of unknown function containing at least one putative transmembrane domain.
* Genes have average expression level lower than 0.05 RPKM.
<table>
<thead>
<tr>
<th>Name</th>
<th>ProteinID</th>
<th>Description and Subcellular Location</th>
<th>q-values</th>
<th>Cluster</th>
<th>Significant Effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAH1</td>
<td>522126</td>
<td>α-CA, periplasm</td>
<td>3.4E-04</td>
<td>15</td>
<td>C+S+CS</td>
</tr>
<tr>
<td>CAH2</td>
<td>522125</td>
<td>α-CA, periplasm</td>
<td>3.2E-01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAH3</td>
<td>526413</td>
<td>α-CA, thylakoid lumen</td>
<td>1.5E-02</td>
<td>8</td>
<td>C+S</td>
</tr>
<tr>
<td>CAH4</td>
<td>522732</td>
<td>β-CA, mitochondria</td>
<td>2.7E-03</td>
<td>15</td>
<td>C+S+CS</td>
</tr>
<tr>
<td>CAH5</td>
<td>522733</td>
<td>β-CA, mitochondria</td>
<td>3.4E-03</td>
<td>15</td>
<td>C+S+CS</td>
</tr>
<tr>
<td>CAH6</td>
<td>512520</td>
<td>β-CA, chloroplast stroma</td>
<td>5.4E-03</td>
<td>12</td>
<td>C+S</td>
</tr>
<tr>
<td>CAH7</td>
<td>515107</td>
<td>β-CA, unknown</td>
<td>2.5E-02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAH8</td>
<td>526207</td>
<td>β-CA, plasma membrane</td>
<td>1.6E-02</td>
<td>11</td>
<td>S</td>
</tr>
<tr>
<td>CAH9</td>
<td>522626</td>
<td>β-CA, cytoplasm</td>
<td>2.3E-02</td>
<td>13</td>
<td>S</td>
</tr>
</tbody>
</table>

* ProteinID = Augustus 5.0 gene model protein ID; q-values = q-values calculated by overall test; Significant Effect = individual effect having a q-value < 0.025 by C/S impact test, where “C” means CO₂ effect, “S” means strain effect, and “CS” means interaction effect.
References


CHAPTER 3

FURTHER INVESTIGATIONS OF THE DIFFERENCE BETWEEN LOW AND VERY LOW CO\textsubscript{2} ACCLIMATION STATES:

DEEP SEQUENCING TO IDENTIFY SUPPRESSOR MUTATIONS OF THE “AIR-DIER” PHENOTYPE AND DETAILED TRANSCRIPT ABUNDANCE PROFILING OF MAJOR CHLAMYDOMONAS REINHARDTII CCM GENES IN VARIOUS LIMITING CO\textsubscript{2} CONDITIONS

I would like to acknowledge the guidance, helps and contributions received:

The \textit{su4 su5} strains used in this study and the rough genetic mapping results were generously provided by Dr. Deqiang Duanmu. The procedure of high through-put sequencing genomic study was kindly guided by Dr. Eric Fritz who is working in Dr James Reecy’s lab in the animal science department of Iowa State University. The experimental design of this comparative genomic study was guided by Dr. Martin Spalding and Dr. James Reecy. Guidance and information about coverage depth and SNPs/INDELs impact factors are provided by Dr. Sorel Fitz-Gibbon who works in Dr. Matteo Pellegrini’s lab in University of California, Los Angeles. The qPCR quantification of the 5 key CCM genes was helped by undergraduate assistant Mr. Yuchen Zhao.
Abstract

The green algae species *Chlamydomonas reinhardtii* can acclimate to varied environmental CO$_2$ concentration through an inducible CO$_2$ concentration mechanism (CCM). The chloroplast stroma protein gene LCIB mutant strains exhibits an “air dier” phenotype: they are viable only under High (5% by volume) or Very Low CO$_2$ (less than 200 ppm), but fail to grow under Low CO$_2$ (300 to 500 ppm) conditions. This demonstrated the acclimations are different in these conditions. Some second-site mutant suppressor strains of the “air dier” phenotype suppressor were generated previously in an effort to identify genes involved in the multiple acclimations and the CCM. Two of these suppressor strains, *su4* and *su5*, are putatively allelic and their mutations are not linked with the insertional mutagenesis marker. In this study we employed high through-put sequencing to examine the Single Nucleotide Polymorphisms/ Insertions and Deletions (SNPs/INDELs) mutations in the *su4/su5* strains, and generated a candidate list of possible locations of the suppressor mutations in these 2 strains. In addition to the suppressor identifications, we also carefully monitored the transcript levels of 5 major CCM genes *LCIA, LCIB, LCI1, CAH1*, and *HLA3* in detailed 12 hour time courses after switching the environmental CO$_2$ from High to a variety of limiting CO$_2$ conditions (6 gradients spanning from 100 ppm to 350 ppm, which included both Low and Very Low CO$_2$ conditions). We observed the transcript levels of *HLA3* were induced later than other 4 tested genes in both the wild type strain and the *pmp1* mutant strain (“air dier”, with a nonsense point mutation in *LCIB*). The patterns of transcript abundance of these genes exhibited moderate variations in the
150 ppm and 200 ppm CO₂ conditions for the wild type strain but not in the *pmp1* mutant.
Introduction

*Chlamydomonas reinhardtii* is a unicellular green algae species capable of acclimating to varied environmental CO$_2$ concentrations. A CO$_2$ concentrating mechanism (CCM) is induced when the environmental CO$_2$ concentration is at or below current atmospheric level, in order to compensate for the low specificity and the low efficiency of the Ribulose-1,5-bisphosphate carboxylase (Rubisco) (Wang et al. 2011). Numerous investigations have focused on discovering the genes and gene products involved in the CCM of *Chlamydomonas*.

Because of the subcellular enrichment of CO$_2$ is against the random diffusion process and also because of the cell structural barriers, considerable efforts have been focused on discovering the inorganic carbon transporters. There are many putative transporters or channels that mediate the active uptake of inorganic carbon (Ci: CO$_2$, HCO$_3^-$, and CO$_3^{2-}$) species. Current candidate transporters include: plasma membrane transporters LCI1, HLA3; and chloroplast transporters LCIA, ycf10, CCP1/2.

LCI1 is a plasma membrane protein that is highly induced under limiting CO$_2$ conditions (Burow et al. 1996; Ohnishi et al. 2010). Artificially introduced overexpression of *LCI1* in the high CO$_2$ requiring *lcr1* mutant strain (lacks expression of *LCI1, CAH1* and *LCI6*) shows increased Ci accumulation and increased photosynthetic affinity (Ohnishi et al. 2010).
The putative ATP-binding cassette (ABC) type transporter gene HLA3 is regulated by light and highly induced under limiting CO2 conditions (Im and Grossman 2002; Miura et al. 2004; Xiang et al. 2001; Fukuzawa et al. 2001; Fang et al. 2012). Multiple lines of evidence confirm that the HLA3 protein is localized on the plasma membrane (Im and Grossman 2002; Gao et al., in revision), and is considered to be a putative plasma membrane HCO3- transporter, also possibly working in tandem with LCIA (a chloroplast membrane protein putatively involved in HCO3- transport) (Duanmu et al. 2009a; Gao et al., in revision). An HLA3 gene knockdown combined with either an LCIB gene (encodes a chloroplast stroma protein putatively involved in active CO2 uptake) mutation or co-knockdown of LCIA gene shows defected Ci assimilation and growth in alkaline pH when the dominant environmental Ci species is HCO3- (Duanmu et al. 2009a). The artificial overexpression of HLA3 in the non-induced high CO2 condition, either alone or in combination with LCIA overexpression increases both Ci accumulation and Ci dependent photosynthetic O2 evolution rates (Gao et al., in revision).

The expression pattern of LCIA gene is significantly regulated by limiting CO2 conditions (Miura et al. 2004; Fang et al. 2012). LCIA protein is predicted to have six transmembrane domains and has been confirmed to localize on the chloroplast envelope (Wang and Spalding 2014). Defective Ci uptake in co-knockdown strains of LCIA and HLA3 further support the putative Ci transporter role of LCIA in the CCM (Duanmu et al. 2009a). When the Ci concentration is very limited, the LCIA-LCIB double mutants exhibit a highly reduced CO2 dependent photosynthesis; and the LCIA single mutant has similarly decreased photosynthesis under high pH conditions. These
observations imply that LCIA is involved in active HCO$_3^-$ uptake across the chloroplast envelope.

Other candidates Ci transporters on the chloroplast envelope include the plastid-encoded ycf10 (Rolland et al., 1997), and two very similar proteins CCP1 and CCP2 (Spalding and Jeffery, 1989; Ramazanov et al., 1993; Chen et al., 1997). Mixed evidences seem to suggest they may also be involved in CCM- associated Ci transport (Rolland et al., 1997; Pollock et al. 2004; Wang et al. 2011; Fang et al. 2012).

Carbonic anhydrase (CA) enzymes catalyze the conversion between Ci species, and some of these CAs putatively play important roles in the CCM. One alpha type CA, CAH1, is located in the extracellular matrix, and is highly induced under limiting CO$_2$ (Moroney et. al, 1985). A CAH1 mutant did not show a significant growth phenotype compared to wild type (Van and Spalding, 1999), however the expression pattern of this gene seems to be linked to CCM functionality (Moroney et. al, 1985; Fukuzawa et. al, 1990; Miura et al. 2004; Fang et al. 2012), and it may play a role in very low CO$_2$ conditions. The CAH3 gene encodes a thylakoid lumen CA, whose transcript level is mildly up-regulated in limiting CO$_2$ conditions (Fang et al. 2012). CAH3 is also regulated through phosphorylation under varied CO$_2$ conditions: during acclimation to limiting CO$_2$ conditions, the enzyme activity increased about 5–6 fold while maintaining similar protein abundance, and the localization of CAH3 shifted from stroma thylakoids, in which it was associated with photosystem II, to become concentrated in the thylakoid tubules inside the pyrenoid (Blanco-Rivero et al. 2012).
The soluble chloroplast stromal protein LCIB is also considered to play an important role in the *Chlamydomonas* CCM (Spalding et al., 1983; Wang and Spalding, 2006; 2014a; 2014b; Moroney and Ynalvez, 2007; Spalding, 2008; Yamano et al., 2010), For wild type *Chlamydomonas*, acclimation to limiting CO$_2$ environments seems to be different at different CO$_2$ levels (Vance and Spalding, 2005). Mutants with defects in *LCIB* revealed that *Chlamydomonas* has 3 acclimation states in response to the varied environmental CO$_2$ levels. Two mutant strains, *pmp1* and *ad1*, have either a nonsense point mutation or a deletion mutation, respectively, at the *LCIB* locus (Wang and Spalding 2006). These *LCIB* mutants exhibit an “air dier” phenotype, which means they can survive and grow either under High CO$_2$ condition (5% by volume) or Very low CO$_2$ conditions (less than 200 ppm), but fail to be viable under atmospheric or “Low CO$_2$“ conditions (300 to 500 ppm). Duanmu and Spalding (2011) applied insertional mutagenesis to identify several second-site suppressor mutations in *LCIB* mutant backgrounds. These strains exhibit partial or complete suppression of the “air-dier” phenotype, which indicates the mutated loci may also be involved in the CCM and perhaps also the multiple acclimation states of *Chlamydomonas*. Two of the identified suppressor loci, *su6* and *su7*, were demonstrated to be mutated in the thylakoid lumen carbonic anhydrase, *CAH3*. The *CAH3* mutations only partially recovered the viability in Low CO$_2$ conditions, and were unable to survive in Very low CO$_2$ condition (Duanmu et al. 2009b). The *su8* strain resulted in nearly wild-type growth and was identified as having a single, linked insertion on chromosome 16 according to version 4 genome assembly (http://genome.jgi-psf.org/Chlre4/Chlre4.home.html) that caused a large and
probably complex chromosomal disruption, including multiple DNA deletions totaling about 100 kb. The *su1*, *su4* and *su5* strains were also independently generated mutants with close to wild type growth, but the inserted DNA fragment was not linked to the phenotype in any of these suppressed lines. To identify the affected locus in *su1*, a high throughput DNA sequencing approach was applied, and a mutation in a limiting CO\textsubscript{2} induced gene *LCI15*, which encodes a PRLI-interacting factor like protein, appears to be responsible (Akella et al. unpublished).

The *su4* and *su5* strains were not able to complement each other in more than 100 genetic crosses, so even though they also were identified as dominant mutations, they must be closely linked or affecting a single gene (Duanmu and Spalding, 2011). However, the inserted DNA fragments are not linked with the phenotype, which makes identification of the causal mutated locus impossible to identify through a traditional PCR approach. On the other hand, genotyping by high throughput DNA sequencing has become an affordable alternative (Lin et al. 2013) that may be applicable in this case, because the apparent allelism of the *su4* and *su5* mutations should help identify the causal locus. In this study, we used the Illumina MiSeq system to sequence the genomes of the *su4* and *su5* strains and their parental strains. Through comparative genomic studies, we generated a list of candidate loci for *su4* and *su5* mutations.

In addition to the suppressor study, we also made an effort to understand the multiple acclimation states associated with LCIB function through investigating detailed changes in gene expression levels for key genes early in the acclimation response. Many
putative Ci transporters and CAs have been uncovered, however our attempt to identify differences in transcript expression levels between Low CO$_2$ and Very Low CO$_2$ conditions was inconclusive (Fang et al. 2012). Therefore, we performed a detailed transcript abundance experiment for 5 major CCM related genes. These genes include the periplasmic carbonic anhydrase gene $CAH1$; 3 putative Ci transporter genes $LCI1$, $LCIA$, and $HLA3$; and the chloroplast stromal protein encoding gene $LCIB$. In this study, we utilized bioreactors to control conditions precisely and quantitative real time PCR (qPCR) to carefully quantify the transcript abundance. We monitored the transcript levels of these 5 major CCM genes through a 24 h time course under 6 CO$_2$ concentrations within the Very Low CO$_2$ and Low CO$_2$ ranges and in 2 different genomic backgrounds.
Result

Strain lineage used for the suppressor study

The original su4 and su5 strains were generated by insertional mutagenesis (plasmid pSI103; Sizova et al. 2001) of pmp1 (CC-1860), which has a point nonsense mutation in the LCIB gene locus (Wang and Spalding, 2006; Duanmu and Spalding 2011). We used progeny and cc498, respectively, from a cross between the original su4/5 strains (pmp-su4 and pmp-su5) and another lcib mutant, ad1. The su4/5 mutants with ad1 background were deposited as ad-su4 (CC-4497) and ad-su5 (CC-4498) in the Chlamydomonas Resource Center, which lists these ad-su4/5 strains as generated from a cross with ad1-1 mt- (CC-4156), but, according to Dr. Duanmu’s record (unpublished), the ad-su4/5 strains were generated from a cross with ad1-2 mt+ (CC-4158).

Considering that the mating type for pmp1-su4/5 are both mt- (Duanmu and Spalding, 2011), and the mating type of ad1-1 strain is also mt-, we are reasonably certain that the ad-su4/5 strains were actually generated by crossing pmp1-su4/5 and ad1-2.

Therefore, in this genotyping study, we used the ad1-2 strain and the pmp1 strain as the parental backgrounds to investigate the su4/5 mutations in the ad-su4/5 strains.

Experimental design of the suppressor study

The strategy for comparative genotyping was straight forward: sequence the genome of the parental strains and the mutants; identify all variations in comparison
with the reference genome; and identify variations of Single Nucleotide Polymorphisms/Insertions and Deletions (SNPs/INDELs) that are unique to the mutant strains. Because \textit{su4} and \textit{su5} mutations are allelic or at least very closely linked, we focused only on variations located near to each other on the \textit{ad-su4} and \textit{ad-su5} genomes as potential candidates for causal mutations of the “air dier” suppressor phenotype. Because the \textit{su4} and \textit{su5} mutations are both dominant, we considered two most likely scenarios regarding the mechanisms of these mutations: first case is a loss of function mutation in the coding region of a subunit of a multimeric protein, in which multiple numbers of this subunit are required to assembly the functional enzyme; so, even though functional, complemented subunits are translated, the defective subunit may still be assembled and “poison” the function of the whole enzyme complex. In the second case, a mutation might occur in a negative regulation region of a gene (the product of this putative gene can positively affect the Ci uptake under Low CO2 condition and acting upstream to the \textit{LCIB} gene), allowing expression of this gene even in the presence of a wild-type cis regulatory elements \textit{su4/5} allele on the complementary chromosome in diploid vegetative cells. This could result in a dominant loss of function mutation. In this study, our investigation is focused on mainly on the first scenario.

\textbf{Results of the suppressor study}

Using Illumina MiSeq system, we sequenced the genomic DNA of \textit{pmp1}, \textit{ad1-2}, \textit{ad-su4}, and \textit{ad-su5}. We have achieved 40 – 50X sequence data coverage of the
approximately 111 Mb genomes for each strain. After alignment of all the sequences, we used the available reference genome assembly version 5 to identify single nucleotide polymorphism (SNPs) and small sized insertions and deletions (INDELs) for all 4 strains. Additionally, the sequencing errors called “artifacts” caused by Illumina’s sequencing technology may also cause falsely discovered SNPs/INDELs specific to particular sequence arrangements on the DNA fragments (Howe et al. 2012). Therefore, we first cleaned up all of our SNPs/INDELs data (http://stormo.wustl.edu/SNPlibrary/index.html) by eliminating the same SNPs/INDELs found in any of 16 unrelated strains also sequenced by the Illumina’s system (Lin et al. 2013). After this filtering procedure, roughly 80% of the SNPs/INDELs were excluded from the candidate list (Figure 1 A &B).

The genetic lineage suggested the genomic backgrounds of $\text{ad-su4/5}$ are derived from $\text{pmp1}$ and $\text{ad1-2}$, so we also eliminated all SNPs/INDELs for $\text{ad-su4/5}$ that were exactly the same as those in either $\text{pmp1}$ or $\text{ad1-2}$. More than half of the remaining SNPs/INDELs were filtered out by this step (Figure 1 A &B). In the third step, we used the SnpEff software (Cingolani et al. 2012) to predict SNPs/INDELs that have a predicted impact on amino acid coding, since we focused on coding regions in this study. After selecting SNPs/INDELs with “higher” impact (altered amino acid coding), roughly 200 loci were left as our candidates (Figure 1 B & C).
Relation to genetic mapping results

The su5 mutation was roughly mapped genetically by Duanmu and Spalding (2011), where it was reported that the IDA7 marker (Kathir et al. 2003) was linked with the su5 mutation in 87% of the progeny. Since the IDA7 marker was placed on chromosome 1 in version 4 Chlamydomonas genome assembly, the reported conclusion was that su5 and, by implication, su4 was also located on chromosome 1. However, in the latest version 5 genome assembly, the locus of the IDA7 marker has been placed on chromosome 16 instead. Therefore, we focused our candidate mutation search on chromosome 16. After filtering, 2 SNPs/INDELs for su4 and 9 for su5 were left, but the physical distance between ad-su4 and ad-su5 called SNPs/INDELs are all above 500 kb (Supplemental Table1, tab chromosome 16).

Unexpected result on LCIB locus for su4 and su5

Although it is unrelated to the goal of this experiment, we examined the alignment at the LCIB locus for the 4 strains sequenced, and we discovered unexpected genotypes for ad-su4 and ad-su5 at that locus. The ad1 strain was originally generated from the wall-less strain cw10 (CC-849, mt-) by insertion of a zeocin resistance gene Ble, and a 36 kb region, which includes the LCIB gene locus, was deleted, causing the “air-dier” phenotype (Wang and Spalding, 2006). The ad1-2 strain was generated by crossing the ad1 strain with CC-620 (wild type walled), and selected for zeocin
resistance and an “air-dier” phenotype. Since \( ad-su4 \) and \( ad-su5 \) were maintained on zeocin containing media, and they were generated by crossing \( pmp1-su4/5 \) to \( ad1 \) strains, we assumed the \( ad1 \) version of the \( LCIB \) locus (\( LCIB \) deleted) in the \( ad-su4/5 \) strains should be retained, including the 36 kb deletion. However, our sequence alignment at the \( LCIB \) locus revealed the genotypes of both \( ad-su4 \) and \( ad-su5 \) actually resembled that from \( pmp1 \) (point mutation) rather than from \( ad1 \) (deletion). According to the Phytozome genome assembly (www.phytozome.net, version 5), \( LCIB \) is located on chromosome 10, position from 4608866 to 4611901. Figure 2 illustrates the lack of alignment coverage in \( ad1 \) from position 4601819 to 4641143 on chromosome 10. On the other hand, \( pmp1 \), \( ad-su4 \), and \( ad-su5 \) all showed alignment coverage in that region, and they shared the same point mutation which converts the 35\(^{th}\) amino acid codon from TAC (Tyrosine) to TAA (stop codon) at position 4609045. Since the \( LCIB \) gene is still mutated and both \( ad-su4/5 \) strains exhibited a wild-type growth phenotype (i.e., not air dier), the su4/5 suppressor mutations are still present in the \( ad-su4/5 \) strains.

Therefore, it appears that the \( ad1 \) 36 kb deletion has been repaired in these strains using the \( pmp1 \) allele of the \( LCIB \) locus and flanking sequences as a template, probably during the original crosses between \( pmp1-su4/5 \) and \( ad1-2 \), although the mechanism by which the zeocin resistance \( Ble^r \) gene remains integrated in the genome of the \( ad-su4/5 \) strains during this template-based repair is still unclear.
Detailed analysis of key CCM gene expression

We determined the time courses for induction or up-regulation of 5 key CCM-related genes during growth of *Chlamydomonas* following transfer of cultures from High CO$_2$ to various levels of limiting CO$_2$ conditions in bioreactors. We selected 5 CCM-related genes, *LCIA*, *LCIB*, *LCI1*, *CAH1*, and *HLA3*, all strongly induced or up-regulated by CO$_2$ and under regulation of the CCM master regulator CIA5 (Fang et al. 2012). In contrast with the single 4 h time point sampling in our previous transcriptome study (Fang et al. 2012), the mRNA transcript levels for these genes were carefully quantified by qPCR in detailed time courses at 6 limiting CO$_2$ concentration gradients. Two strains were used: cc125 (wild type) and *pmp1* (point mutation on *LCIB* gene, “air dier” phenotype). Following culture of cells at 5% CO$_2$ (High CO$_2$ condition), the inflow gas was switched to one of 6 limiting CO$_2$ concentrations: 100, 150, 200, 250, 300, and 350 ppm CO$_2$. Samples were taken at 8 time points: 0, 1, 2, 3, 4, 6, 8, and 12 hours after switching the inflow gas to limiting CO$_2$. For each strain and condition, 3 independently grown samples were used as biological replicates.

The transcript quantification results are summarized graphically in Figure 3. By visual inspection, the most obvious observations are the earlier up-regulation of *LCIB* and the delayed up-regulation of *HLA3* in wild type cc125 strains. Most genes other than *HLA3* exhibited peak or near-peak up-regulation at 2 h, but *HLA3* achieved a similar induction level at 3 h under limiting CO$_2$ concentrations 100, 250, 300, and 350 ppm. The 150 and 200 ppm CO$_2$ concentrations lie between the typical Low CO$_2$ and Very Low
CO₂ ranges (Wang and Spalding, 2006), and almost all genes exhibited peak expression at 2 hours in these two CO₂ concentrations, including *HLA3*. The apparent delay in *HLA3* induction was not observed across all CO₂ concentrations in the *pmp1* mutant, which contains a defective LCIB protein and exhibits an “air-dier” phenotype.

All 5 studied genes were induced or up-regulated under the various limiting CO₂ conditions tested, and their transcript abundances are all maximized or nearly maximized by the 4 h time point. This expression time course validates the sampling time we used in our previous transcriptome study (Fang et al. 2012). Notably, the delayed *HLA3* gene induction pattern and the faster induction pattern of *LCIB* apparent in Figure 3 under both Low and Very low CO₂ conditions was also observed in a previous report using only 100ppm CO₂ (Brueggeman, et al. 2012). In Figure 4, we constructed hierarchical trees based the time needed during the first 4 hours for transcript expression to reach its maximum. *HLA3* was distinguished from all other genes in cc125, although in the *LC11* was similar to *HLA3* in the *pmp1* “air-dier” strain. *LCIB* expression was distinguished from all other genes in both strains.
Discussion

The focus of the comparative genomic experiments was to identify genes responsible for suppression of the “air-dier” phenotype by the 2 independently generated but apparently allelic mutations, su4 and su5. These suppressor genes may be linked to the multiple CO₂ acclimation states of Chlamydomonas, just like the LCIB, CAH3, and LCI15 genes identified previously (Wang and Spalding, 2006; Duanmu et al., 2009b; Akella et al., unpublished). The strategy to use deep sequencing to identify candidate suppressors was very straightforward, since a rough mapping of su5 was already reported, and the su4 and su5 mutations are apparently allelic mutations (Duanmu et al. 2011), and, in addition, the parental strains (although not direct parents, but strains covering all the genetic backgrounds) were available, and Chlamydomonas strains are haploid. Based on initial sequence data, the number of called SNPs/INDELs was exceedingly high. This high rate of detected random mutations also was observed in Lin et al. (2013), which raised a serious issue regarding false positive discoveries in these kinds of experiments. After filtering out the same SNPs/INDELs called from this collection of unrelated strains, we had only about 20% of the original SNPs/INDELs left. These numbers gave us an indication of how many systematic sequencing errors can affect the SNPs detection procedure. In this study, we did not further investigate these systematic “artifacts” errors, but subtracting the SNPs/INDELs list from unrelated strains seems to be a necessary step to reduce the false positives caused by sequencing “artifacts”.

Chlamydomonas genome assembly work has been constantly updated, and the apparent miss-assembly of mapping marker IDA7 may still be a source of confusion as we try to pinpoint a gene or a mutation. The IDA7 marker was assembled to chromosome 1 in version 4 assembly, but was later placed on chromosome 16 in version 5 assembly. A similar miss-assembly was also observed in Lin et al. (2013), therefore we consider more genetic mapping should be performed, which means our assumption of su4/5 mutation are located on chromosome 16 should be re-evaluated as well.

The unexpected genotype at the LCIB locus initially made us skeptical about our strain lineage information, especially in combination with the disagreement regarding lineage between the Chlamydomonas Resource Center and the depositor of the strains used. However, we did double examine the phenotype and mating type, they are all concordant to with lineage documentations. Based on the observed evidence, we hypothesize that during the genomic recombination process, the large deletion region in the ad1 strains was repaired, perhaps through a homology-based repair mechanism, with the zeocin resistance gene re-captured from the deletion gap in the ad1 strain when the cells were selected for zeocin resistance. The exact location of the inserted BleR gene may be investigated further to explain why and how this apparent deletion repair occurred. However, since no functional LCIB gene is present in ad-su4/5, this unexpected genotype should not affect the search for su-4 and su-5 loci.

Since there is no linkage between the su4/5 mutations and the inserted fragment, the generated unique SNPs/INDELs listed for the ad-su4/5 strains were used
as a reference (Supplemental Table file 1). Our strategy was based on the assumption that the $su4/5$ mutations were most likely to have occurred in the coding region, which obviously ignored the possibility that the $su4/5$ mutations are in a regulatory region. If we consider this possible scenario, then the candidate SNPs/INDELs list would be significantly increased (in Supplemental Table1, on chromosome 16 the numbers of SNP/INDELs detected outside the coding regions in $su4$ and $su5$ are 44 and 79, respectively). Therefore, in order to successfully identify the actual loci of the $su4/5$ mutations, we believe further genetic mapping may be necessary.

In addition to the suppressor identification attempts, we also sought evidence from changes in transcript levels of key CCM genes to explain the varied acclimations to Low and Very Low CO$_2$ conditions, since such data are currently not available in detailed time courses and CO$_2$ concentration gradients. When we previously examined the transcriptome of cc125 (137c) using 4-hour time points, no significant variations were found in gene expression between the Low and Very Low CO$_2$ conditions (Fang et al. 2012). Therefore, we looked more closely at the regulation of key CCM genes over the range of CO$_2$ concentrations spanning the Very Low to Low CO$_2$ acclimation states. In these experiments, we used controlled bioreactors, induction using multiple different CO$_2$ concentrations, and sampled over a 12 h time course. We observed variations in the gene expression patterns between $LCIB$ (faster up-regulation) and $HLA3$ (slower up-regulation) and the other 3 genes, and also variations in these patterns between the wild type strain and the “air-dier” strain, $pmp1$. 
With regard to the earlier up-regulation of *LCIB*, it should be noted that LCIB is critical for low CO\textsubscript{2} acclimation, the next concentration tier below high CO\textsubscript{2}, and maintenance of its constitutive, low level expression in high CO\textsubscript{2} has been suggested to allow cells to respond immediately to decreased CO\textsubscript{2}, avoiding the delay before *de novo* synthesis of the full suite of CCM proteins, most of which play more significant roles in very-low CO\textsubscript{2} (Wang et al., in review). Similarly, a more rapid up-regulation of *LCIB* gene expression upon transfer from high CO\textsubscript{2} to limiting CO\textsubscript{2} may be consistent with the distinct functions of this CCM gene product associated with different CO\textsubscript{2} acclimation states (Wang and Spalding 2014b; Wang et al. in review). On the other hand, the reason behind the slower up-regulation of HLA3 relative to the other 4 genes is not easily explained.

The variations in gene expression patterns under varied induction CO\textsubscript{2} concentrations associated with a mutation in the *LCIB* gene suggests either that the defect in the LCIB mutants has a pleotropic effect on gene expression or the *LCIB* gene product is directly or indirectly involved in the regulatory process itself. The former seems more likely, but further molecular or biochemical investigations are needed to fully understand the cause of these observed transcript variations.

It is interesting to note that the observed differences in gene expression patterns are consistent with potential different up stream regulation for these 5 major CCM genes. However, the observed variations in transcript levels were not dramatically different, so it is possible that post-transcription level regulation is a more important
factor regulating the functional changes associated with multiple acclimation states. If so, the observed differences in the transcript abundance patterns may represent either minor parts of the overall regulation or are indirectly caused by the post-transcriptional changes. The observation data was collected in the Supplemental data package 2, and the remaining RNA samples can be used to investigate transcript levels of other genes in the future.
Material and Methods

Strains and maintaining conditions

The cc125 (137c), pmp1, ad1-2 strains were maintained on the CO₂ minimum agar media, and the ad-su4 and ad-su5 strains were maintained on the CO₂ minimum agar plates that contain 10 µg/ml zeocin, which was described in (Wang and Spalding, 2006). All strains are available at the Chlamydomonas Resource Center, the University of Minnesota, St. Paul, MN. The culture collection numbers are: CC-125 wild type for the 137c strain; cc-1860 for the pmp1 strain; CC-4158 for the ad1-2 strain; CC-4497 for the ad-su4 strain; and CC-4498 for the ad-su5 strain.

DNA and RNA isolations

Genomic DNA samples were isolated and purified as described previously (Van and Spalding 1999). All samples for the transcript abundance study were collected directly from the Bioreactors using a syringe. Samples were immediately transferred to 50 ml tubes and centrifuged at 4,000 x g for 3 minutes, and the supernatant was poured off and then transferred to liquid nitrogen. This sampling process was generally completed within 7 minutes; it was designed to minimize changes within the RNA. Frozen cells were stored at −80°C before RNA extraction. The RNA samples were extracted by following the protocols of RNeasy Plant Mini Kit (Qiagen, Cat. No.74904)
without grinding, and DNase cleaned by the Turbo DNA-free kit (Ambion, Cat. No. AM1907).

**DNA sequencing and analyses**

The genomic DNA samples were pair-end sequenced by the Illumina MiSeq system for 500 cycles, 250 cycles from each end. The library preparation and the sequencing process were done at the DNA Facility of Iowa State University. The reference genome assembly version 5 file (Creinhardtii_236.fa.gz) was obtained from [www.phytozome.net](http://www.phytozome.net). The alignment was processed by the software “BWA-0.7.4”, alignment result SAM (Sequence Alignment/Map) files were sorted and transferred to BAM (Binary Sequence Alignment/Map) files by “Samtools-0.1.19” and “Picard tools-1.92”. The SNPs/INDELs were called by “GenomeAnalysisTK-2.5-2 (GATK)” package, and the mutation impact factors for *ad-su4* and *ad-su5* samples were annotated by “SnpEff-LGPLv3”. All the result VCF (Variant Call Format) files were stored in the Supplemental data file package 1. Visualization of the alignments was achieved by the “Integrative Genomics Viewer-2.3 (IGV)” software.

**Bioreactor and the liquid culture conditions**

Plate cultures were first transferred in to liquid CO2 minimum media (Geraghty et al. 1990) and grown in a 5% CO2 chamber at 25°C and 250 rpm shaking speed for
about 4 days. About 160 million cells in the liquid culture (less than 50 mL) were used to inoculate the bioreactors (add fresh CO₂ minimum medium to the 670 mL total volume). Air containing 5% CO₂ were supplied to the bioreactors by the rate of 0.01 L/min initially. After at least 20 hours, when the cell density of the bioreactor liquid culture reach 1 to 3 million cells per mL, the 0 hour time point samples were taken, and immediately switched to the air supply with 100, 150, 200, 250, 300, and 350 ppm CO₂ concentrations (ordered from Airgas USA LLC., with ±4% analytical uncertainty) at the same 0.01 L/min flow rate. The setups of the bioreactor system are the same as described in (Bigelow et. al., 2014), except the light source strength was controlled at 100 µEm⁻² s⁻¹ from both directions. The following samples were taken by 1, 2, 3, 4, 6, 8, and 12 hours after switching the influx gas sources.

**Quantitative real time PCR analysis**

The qPCR quantifications were performed exactly like described in Fang et al. (2012), except the primer sequences used for the LCIA gene are: 5’ ATAGGAGTAGTTGGGTAG 3’ and 5’ CTGCTGAAGATTGTGTAT 3’; and for the LCI1 gene are: 5’ GACACCCAGGAGGGCAT TG 3’ and 5’ TGATGATGGAGGGGAAGACG 3’. The original data files and the Ct value summary table files are packed in the Supplemental data file package 2. As reported in Fang et al. (2012), the CBLP gene was used as internal control for normalization, and the normalized Ct (Ctᵣ) values were calculated as CtᵣCBLP - Ctᵣgene. The percentage of maximum expression for each gene at each time point under every
condition (Ct\%\textsubscript{r}) was calculated by \( \text{Ct}_{\%} = \text{Ct}_{n}/(\text{maximum Ct}_{i} - \text{minimum Ct}_{i}) \times 100 \) in that condition. The hierarchical tree presented in Figure 4 was constructed in R platform: distance data were calculated by the Euclidean method based on the first 4 hours Ct\%\textsubscript{r} values, and the hierarchical tree was constructed by the UPGMA method.

Supplemental data and tables are available at:

https://iastate.box.com/s/qe8644jqntisq12a9k9b
**Figures**

**Figure legends**

**Figure 1: Summary of the SNPs/INDELs screening process (data summarized in Supplemental Table 1, tab “SNPs and Indels number”).** (A), bar graph of SNPs/INDELs detected in \textit{ad-su4} and \textit{ad-su5} across the 17 Chlamydomonas chromosomes and multiple scaffolds, including the total SNPs/INDELs number: before any subtraction; after subtracting the unrelated SNPs/INDELs from the WUSTL database (Lin et al. 2013); after further subtraction of SNPs/INDELs from \textit{pmp1} and \textit{ad1-2}; and selected SNPs/INDELs having higher impact factor to alter amino acid coding; (B), a table summarizing the SNPs/INDELs at each step in the screening process; (C), the remaining higher impact SNPs/INDELs summarized on a bar graph across all 17 chromosomes. All SNPs/INDELs on scaffolds were filtered out in the screening.

**Figure 2: Alignment coverage map of the region around the \textit{LCIB} locus in \textit{ad1-2}, \textit{pmp1}, \textit{ad-su4} and \textit{ad-su5} strains.** Sequence coverage is illustrated by Integrated Genome Viewer (IGV, version 2.3). The gray histogram bars illustrate the counts of alignment coverage at each position on the reference genome. The deleted region in \textit{ad1} did not have coverage, and the point nonsense mutation in the \textit{LCIB} gene is indicated by the green arrow. Note the deleted region in the \textit{ad1-2} mutant does have alignment coverage in the \textit{ad-su4} and \textit{ad-su5} sequence files, and that the \textit{LCIB} gene loci of \textit{ad-su4} and \textit{ad-su5} contain mutations exactly like the \textit{pmp1} strain.
Figure 3: Patterns of transcript abundance for 5 key CCM genes *LCIA, LCIB, LCI1, CAH1,* and *HLA3* over a 12 h time course following a change from 5% CO$_2$ to the indicated CO$_2$ concentration. The RNA transcript levels were examined by qPCR using *CBLP* as internal control, and the Ct value changing range between minimum and maximum expression was used to normalize the scale of the transcript abundance. Higher relative percentage values represent closer to the maximum transcript abundance calculated in terms of the log$_2$ fold change values.

Figure 4: Distance tree describing similarities in transcript expression patterns between studied genes in two different genotypes. Transcript levels over the first 4 hours were used to calculate the similarity distances. Relative Ct values were converted to 0 to 100 % maximum expression levels, so the analysis considered how the expression of each gene reached its maximum. Euclidean distances were calculated between each gene using expression data from all limiting CO$_2$ conditions for each strain. Hierarchical trees were constructed by UPGMA method.
Figures:

**Figure 1**

A graph showing the number of SNPs/INDELs across different chromosomes, with different colors representing different datasets.

B Table showing the number of SNPs/INDELs before and after subtraction of unrelated genome and further subtraction of ad1-2 & pmp1.

<table>
<thead>
<tr>
<th></th>
<th>su4</th>
<th>su5</th>
<th>su4</th>
<th>su5</th>
<th>su4</th>
<th>su5</th>
<th>su4</th>
<th>su5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Originally</td>
<td>37,621</td>
<td>38,115</td>
<td>4,825</td>
<td>4,964</td>
<td>1,878</td>
<td>1,983</td>
<td>201</td>
<td>196</td>
</tr>
<tr>
<td>called SNPs/INDELs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>after subtraction</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>of unrelated genome</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Further subtract</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ad1-2 &amp; pmp1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AA sequences altered</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

C Graph showing the remained SNPs/INDELs for su4 & su5 across different chromosomes.

**Figure 2**

A screenshot from a genome analysis software showing deletion regions in ad1 strains. Point C to A mutation shared by pmp1, ad-su4, and ad-su5.
Figure 3

Vertical axis: percentage of maximum expression relative to Ct value changing ranges
Horizontal axis: Hours after induction

Figure 4

Similarity distance
References


CHAPTER 4

TRANSCRIPTOME STUDY OF *CHLAMYDOMonas REINHARDTII* IN NITROGEN STARVED CONDITION AND THE ADP-GLUCOSE PYROPHOSPHORYLASE REGULATORY SUBUNIT MUTANTS FOR PHOTOAUTOTROPHIC CULTURES

I would like to acknowledge the guidance, helps, and contributions received:

The experiment was designed by Dr. Marcia Almeida-deMacedo (who worked in Dr. Eve Wurtele’s research group at Iowa State University) under the guidance by Dr. Peng Liu and Dr. Spalding; the bioreactor growth of *Chlamydomonas* culture, cell density measurements, chlorophyll content and dry weight measurements were done be Dr. Dan Stessman; the RNA sequencing reads alignment was done by Dr. Sorel Fitz-Gibbon who works in Dr. Matteo Pellegrini’s lab in University of California, Los Angeles; the preliminary data of differential gene expression was done by Dr. Almeida-deMacedo; the usage of the EdgeR package and the final statistical analysis was done under the guidance of Dr. Liu.
Abstract

Lipid accumulation in microalgae species has drawn people’s attention for its potential as a biofuel production source. This lipid accumulation process is commonly triggered by nitrogen starvation conditions (N-starvation) in the model green algae *Chlamydomonas reinhardtii*, and starchless mutants with defects in the key starch biosynthesis enzyme ADP-glucose pyrophosphorylase exhibit even more elevated lipid levels. Several studies have been conducted that explore the transcriptome response of N-starvation for *Chlamydomonas* and its starchless mutant *sta6*, but in all reported studies mixotrophic culturing conditions were used, which involve the provision in the medium of organic, fixed carbon in the form of acetate. In contrast, we characterized the transcriptome of *Chlamydomonas* in response to N-starvation conditions in two different *sta1*-derived, ADP-glucose pyrophosphorylase mutants and their corresponding, paired wild-type strains under strictly photoautotrophic conditions. The use in this study of 2 wild type-*sta1* mutant pairs to conduct the analysis provides a strategy to filter genes that are responsive to the genotype as well as to the N-starvation conditions. By comparing with previous work based on mixotrophic growth, we found many responses that were similar to those of mixotrophic cultures, including a general down regulation of photosynthetic genes and an up-regulation of the nitrogen assimilation genes, and several de novo fatty acid synthesis genes that are negatively regulated in both wild type and *sta1* mutants. Many dark reaction genes on the other hand, maintained their expression levels better in contrast to the general down regulations triggered by the N-starvation for the mixotrophic cultures. Also, many
glyoxylate cycle genes are up-regulated in the starchless mutant autotrophic cultures very similarly to those in mixotrophic cultures. Although a substantial number of differentially expressed genes identified under autotrophic conditions differed from those reported for mixotrophic conditions, only a few differences in key genes or pathways highlighted in previous studies were discovered when we are comparing with the previous work. This suggests that major metabolic consequences of N-starvation conditions are likely to be similar in autotrophic and mixotrophic conditions when comparing the ADP-glucose pyrophosphorylase mutants to the wild types.
Introduction

The increased consumption of fossil fuel is becoming one of the major obstacles for continuous, sustainable growth in the modern era. The fast growth rate and the photosynthetic capability of some microalgal species have caught people’s attention lately, and these microalgae are considered by many as ideal candidates for the production of advanced renewable biofuels.

*Chlamydomonas reinhardtii* (hereafter, *Chlamydomonas*) is a photosynthetic unicellular green alga which has been used as a genetic and physiological model in lab research for a variety of studies for decades. The whole genome sequence for *Chlamydomonas* is available along with actively updated annotations (http://www.phytozome.net/search.php for the version 5 assembly). Similarly available from various *Chlamydomonas* experiments are resources for transcriptomics, proteomics, and metabolomics (Bölling and Fiehn, 2005; May et al., 2008; Atteia et al., 2009; Rolland et al., 2009; Mühlhaus et al., 2011; Urzica et al., 2012; Wang et al., 2012; Blaby et al. 2013; Miller et al. 2010; Goodenough et al. 2014). Photosynthetic carbon assimilation processes of *Chlamydomonas* have been intensively studied in the past (Bishop 1973; Raven 2010; Terashima et al. 2011; Fang et al. 2012; Wang et al. 2011; Heinnickel and Grossman 2013). In addition to the autotrophic growth through photosynthesis, *Chlamydomonas* is also capable of utilizing extra-cellular acetate as carbon and/or energy sources (Sager and Granick, 1953); therefore the growth of *Chlamydomonas* can also be heterotrophic or mixotrophic, and the assimilation of
acetate is widely accepted through the glyoxylate cycle (reviewed by Spalding 2009). The growth nutrient requirements were well characterized for the three trophic conditions (Kropat et al., 2011; Catalanotti et al., 2012; Johnson and Alric, 2012), as well.

Like many other microalgae species, conditions when the nitrogen source is limited in the environment (N-starvation) elicit responses for *Chlamydomonas* that include accumulation of reduced carbon into the forms of polysaccharide (including starch or chrysolaminaran) and triacylglycerols (TAG) (Varum and Myklestad, 1984; Granum et al., 2002; Guschina and Harwood, 2006; Hu et al., 2008; Rodolfi et al., 2009; Wang et al., 2009; Work et al., 2010; Siaut et al., 2011). In the accumulation of TAG, lipid droplets can be formed internally, and they are visible under the microscope (Wang et al., 2009; Moellering and Benning, 2010), so the possibility of using *Chlamydomonas* as the model organism to study the lipid accumulation process and development of microalgae biofuel is widely considered.

Because of the attention on the lipid accumulation process of *Chlamydomonas*, the lipid accumulation characteristics of several previously identified starchless mutant strains (Work et al. 2010; Siaut et al. 2011; Blaby et al. 2013) were explored, bearing the logic that more carbon flow should be diverted to lipid synthesis and accumulation when starch synthesis is blocked. Starch synthesis in *Chlamydomonas* is mediated by the enzyme ADP-glucose pyrophosphorylase, which catalyzes the formation of ADP-glucose (building blocks of starch) from glucose 1-phosphate. Two of the reported starchless mutants: *sta1*, and *sta6*, which have mutations in the regulatory and functional subunits
of the ADP-glucose pyrophosphorylase enzyme, respectively, exhibit dramatically
decreased or nearly abolished starch synthesis (Ball et al. 1991; Van den Koornhuyse et

Transcriptomic and metabolic changes associated with the sta6 mutation under
N-starvation conditions have been carefully studied (Blaby et al. 2013; Goodenough et
al. 2014). Mixotrophic cultures of sta6 exhibited significantly increased internal TAG
accumulation compare to the wild type or sta6-complemented strains (Blaby et al.
2013), and with additional acetate provided in the medium during the course of N-
starvation, the increased accumulation of lipid droplets in the cell can even cause the
cell to float after centrifugation (Goodenough et al. 2014). However, in spite of the
markedly higher accumulation of TAG in sta6, fatty acid and lipid metabolism genes
were reported as mostly having similar expression levels between the wild type and sta6
mutant (Blaby et al. 2013). Only a few genes are exceptions, including the DGTT2 gene
that encodes a type-2 diacylglycerol acyltransferase, which exhibited increased
transcript abundance mostly at the beginning of the exposure to N-starved conditions;
the DGAT3 gene that encodes a putative soluble diacylglycerol acyltransferase, and the
MLDP1 gene that encodes a major lipid droplet protein, both of which exhibited
induction at the end of the N-starvation treatment. However, some inconsistencies
were reported in terms of the transcriptome data when using different strains for the N-
starvation and sta6 mutation studies (Goodenough et al. 2014).
On the other hand, considering a major purpose for biofuel production is to convert inorganic carbon back to the organic form, which is reversing the fossil fuel consumption/carbon release process, supplying the carbon source as acetate is energetically unsustainable and economically limiting. An autotrophic growth condition is most preferred for this goal. Fortunately, lipid accumulation behaviors under the N-starvation condition also have been observed in photoautotrophic growth cultures (Msanne et al. 2012; Li et al. 2010).

In recognition of the need to also understand the effect of N-starvation and starch biosynthesis mutations on lipid accumulation under photoautotrophic conditions, we applied an RNA sequencing based transcriptome approach to characterize gene expression patterns in response to the N-starvation conditions for 4 strains: sta1 as a starchless strain with significantly lowered ADP-glucose pyrophosphorylase activity; the 137c (cc125) strain as the corresponding wild type background strain for sta1 (Ball et al. 1991); 21gr as a high growth (compare to 137c) wild type strain; 21st1 as a starchless derivative of 21gr, generated by crossing 21gr and sta1. In addition to considering the benefit of 21gr strain’s higher growth rate, conducting this research with more than one genomic background may also help us filter the differently regulated genes that are directly related to the mutated enzyme but not caused by the varied genome background. In addition, our results are also directly compared to the previous transcriptomic studies of N-starvation and/or starchless mutations (Miller et al. 2010; Boyle et al. 2012; Blaby et al. 2013).
Results

Experimental design

To study the possible role of STA1 and the impact of disrupted starch synthesis during N-starved conditions, we established our experiment in an autotrophic bioreactor environment for algal culturing to facilitate careful control of environmental conditions and to maximize reproducibility. Algae strains were grown photoautotrophically under constant light with 5% CO₂ as the source of inorganic carbon; acetate was not added to the culture medium. Cultures for every strain we tested were duplicated in different bioreactors under the same conditions. Once culture densities reached log growth phase, media were refreshed in each bioreactor with either the same culture medium used initially or with medium lacking all nitrogen nutrients. Samples were taken 0, 24, or 48 hours after refreshing the media.

The sta1 mutant strain was derived from 137c wild type background using X-ray-mediated mutagenesis (Ball et al. 1991), so these two strains are included as one pair in the study. Additionally, a high growth wild type Chlamydomonas strain, 21gr, was included, since it grows to about 2 times the stationary phase cell density as that of 137c (data not shown). A starchless progeny, 21st1, from a genetic cross between sta1 and 21gr, was selected to provide a similar sta1 mutation to pair with 21gr. Therefore, 4 strains, 2 nitrogen conditions, and 3 time points were utilized in the experiment, and these 24 different treatment combinations were repeated to generate 2 biological replicates.
After sequencing the RNA samples, we successfully aligned sequences to the latest *Chlamydomonas* transcriptome, Augustus u111.6 (JGI version 5.5). The relative expression levels for each gene were calculated as FPKM values (http://cufflinks.cbcb.umd.edu/index.html), and the uniquely aligned read counts were fitted by a negative binomial model using the EdgeR package on the R statistical analysis platform (Robinson et al. 2010). Different contrasts were tested and the resulting FDR values were calculated based on p-value sets.

**Comparison of N-starvation responses between wild type strains and corresponding sta1 mutants by multiple contrast testing**

It has been reported that under the N-starved conditions, starchless strains, such as *sta6* and *sta1*, with defective ADP-glucose pyrophosphorylase accumulate more TAG (Wang et al., 2009; Li et al., 2010; Work et al., 2010). Other than the direct blocking of organic carbon flow into starch biosynthesis, we wished to examine whether gene expression in *sta1* mutant strains also is altered to accommodate this metabolic change or even to facilitate carbon flow toward lipid synthesis pathways. We conducted a multiple contrast test utilizing our experimental design to screen for genes that meet these criteria: firstly, the conditions of N-starvation and N-abundance were contrasted for both 24 h and 48 h samples for each strain; secondly, we contrasted the strains according to their genetic lineages, selecting genes that were expressed differentially between 137c and *sta1*, and between 21gr and 21st1, respectively; and lastly, we
identified common genes that exhibited lower than 0.01 FDR in both strain pair
(137c/\textit{sta1} and 21gr/21st1) comparisons. By applying this sophisticated test, we were
specifically seeking unique N-starvation responsive genes that also respond specifically
to the \textit{sta1} mutation, regardless of the genetic background (137c or 21gr).

This very rigorous selection process only yielded a list of 7 genes, and their
expression patterns are illustrated in Figure 1. The \textit{GSM1} gene, which encodes a
gametogenesis related protein contributed by the minus mating type (Lee et al. 2008),
exhibited induced expression in \textit{sta1} mutated strains, but only mild changes in wild type
strains; and a predicted GARP complex subunit gene, \textit{VPS53}, shared the same
expression pattern as \textit{GSM1}. Interestingly, the expression levels of these 2 genes were
reported to be up-regulated in wild type compare to \textit{sta6}, another ADP-glucose
pyrophosphorylase mutant, in N-starvation and heterotrophic growth conditions (Blaby
et al. 2013), which is opposite to the response observed in our experiment. Similar to
\textit{GSM1} and \textit{VPS53}, an unannotated gene \textit{Cre08.g358571} is induced in \textit{sta1} mutants, and
a similar induction was observed in \textit{sta6} as well (Blaby et al. 2013). The \textit{RBD2}
rubredoxin-like gene was highly down-regulated in wild type strains under N-starvation,
and a green lineage gene, \textit{CGL55}, exhibited a nearly opposite expression pattern. There
is no annotation information for the other 2 genes, \textit{Cre12.g512300}, and \textit{Cre12.g537250},
whose expressions were generally down regulated following the N-starvation stress.
**Genes with common N-starvation responses between wild type and sta1 mutants**

As only a limited number of genes behaved differently in the sta1 mutants under N-starved conditions were detected, we also examined the genes that have similar expression patterns in both mutants and wild type strains. We conducted contrast tests between N-starvation and N-abundance conditions for each strain based on both 24 h and 48 h sampling time points. We compared our N-starvation results with those of a previous study that reported differentially expressed genes between the N-starved and N-replete conditions for a mixotrophically (light plus acetate) grown wild type strain at only 48 h sampling time point (Miller et al. 2010). The numbers of genes identified as differentially expressed (FDR less than 0.01) are summarized in Figure 2, which includes the 4 strains we tested and the results from Miller et al. (2010).

From Figure 2, we observed considerable discrepancies between different strains and autotrophic vs. mixotrophic growth conditions. In order to focus on the genes with common N-starvation responses in all strains under autotrophic growth, we further analyzed 475 genes with significant differential expression (< 0.01 FDR) that are shared within the 4 strains we tested. The expression patterns of these genes are shown on a heatmap based on log2 fold change of the FPKM expression levels (Figure 3). The hierarchical clustering on this heatmap revealed at least 4 distinct expression patterns across 24 treatments. The log2-fold expression changing values and the clustering information are summarized in Supplemental Table 1.
To understand the biological influences associated with the changing expressions of these genes, we examined the individual clusters using the web based “Algal Functional Annotation Tool” (Lopez et al. 2011) to identify related MapMan (Thimm et al. 2004) gene ontology terms. The results are summarized in Table 1.

Cluster 1 included 210 genes that exhibit decreased expression levels after switching the cultures to N-starved conditions. It was somewhat unexpected that many lipid metabolism genes were included in this cluster, and expression patterns for many of these genes are inconsistent with the previous mixotrophic study (Miller et al. 2010). A acyl-ACP d9 desaturase gene FAB2 showed up regulation in Miller et al. (2010). And several other down regulated genes did not exhibit significant changes in the Miller et al. (2010) study, which includes: FAD5, a MGDG specific palmitate delta-7 desaturase; KAS1/2, which encode parts of the fatty acid synthase; BCC1, which encodes a component of the Acetyl-CoA Carboxylase; DGAT3, which encodes a putative soluble diacylglycerol acyltransferase; SAS1, which encodes the S-adenosylmethionine synthetase; and PL5B1, which encodes a glycerol-3-phosphate acyltransferase. These down regulation patterns may hint at differences in some lipid metabolism processes between the autotrophic and mixotrophic cultures.

Many of the photosynthesis light reaction genes are also included in cluster 1. Based on categorizations in previous work (Blaby et al. 2013), this cluster includes 19 genes encoding Photosystem I/II and light harvesting proteins. This observation is consistent with observations under mixotrophic growth conditions, which confirms that
investment of resources in photosynthetic light harvesting and electron transport is decreased under the stress of N-starvation.

Cluster 2 genes show a mild increase in expression during N-starved conditions. Many nitrogen metabolism genes are included in this cluster, which is consistent with previous work (Miller et al. 2010). When nitrogen availability is scarce, organic nitrogen uptake related genes are mildly up-regulated, which include ammonium transport genes AMT1D, AMT1E.a, urea transport genes, UAPA2, DUR1 and DUR3A, and the putative amino acid transporter gene, AOT7. Another notable group in this cluster includes several genes related to amino acid synthesis, which includes those encoding arginine synthesis enzymes AGK1, AGS1, and ARG1; aspartate aminotransferase, AST4; and a glutamine synthetase gene, GLN1. We interpret this response a reflecting the expected up-regulation of these genes to maximize utilization of limited nitrogen resources. GPD1, encoding glycerol-3-phosphate dehydrogenase, is also included in cluster 2, which might reflect a higher demand for the formation of glycerol under N-starved vs. N-replete conditions. Nucleotide metabolism genes CMPL1/CMPS1 and CYG46/47 which encode carbamoyl phosphate synthase and guanylate cyclase, respectively, are regulated in the same way. A potassium ion transporter gene, KUP1, and a few putative potassium channel genes were also captured by cluster 2, suggesting a higher demand for potassium under N-starved vs. N-replete conditions, although the implications of this are unclear.
Cluster 3 captures a group of genes that show unexpected variations between the N-starvation and N-replete conditions: they are up-regulated in the 48 hour time point only in the N-replete conditions for the 4 strains tested. The ontology analysis revealed this cluster is enriched in several sulfur assimilation genes, such as periplasmic arylsulfatase genes ARS1/2, sulfur transporter genes SLT1/2 and SULTR2, sulfur permease genes SULP1/2, ATP-sulfurylase genes ATS1/2, and ferredoxin-sulfite reductase gene SIR1. After healthy growth, the accumulated cell density at stationary phase might result in depletion of some nutrients, inducing the expression of genes related to uptake of specifically depleted nutrients from the medium. The specificity of this response for genes in the sulfur assimilation pathway suggests a specific depletion of sulfate under the conditions used, although this was not confirmed by analysis of the medium.

Genes in cluster 4 exhibit a high elevation of transcript abundance under the N-starved conditions, TAG formation gene DGTT1 was captured in this cluster, which is consistent with previous reports using mixotrophic cultures (Miller et al. 2010). In addition to cluster 2 genes which have a similar but milder up-regulation, more nitrogen assimilation genes, including AMT1A (ammonium transporter), LAO1 (periplasmic L-amino acid oxidase), NII1 (Nitrite reductase), and NIT1 / NIA1 (nitrate reductase), are included in this cluster.

Nitrogen starvation is apparently a substantial stress for all strains. The cultures look pale after 48 hours in nitrogen limited medium. We determined the chlorophyll
content (Figure 4), which confirmed our observation and is consistent with the down-regulations of photosynthetic light harvesting and light reaction genes captured in cluster 1.

Genes N-starvation response for wild type strains in various metabolic pathways

In order to further understand the N-starvation and any effects distinguishing autotrophic vs. mixotrophic conditions in a variety of biological pathways, a group of contrast tests focusing on the wild type strains were conducted to directly compare our results to previous transcriptome studies (Miller et al., 2010; Boyle et al. 2012; Blaby et al. 2013). Miller et al. (2010) considered both N-starvation and N-replete conditions, and their transcriptome data were collected at the 48 h time point only. So, for direct comparison with their data, we statistically tested the contrasts between N-starvation and N-replete conditions in both wild type strains 137c and 21gr at only 48 h sampling time, and the gene expression changes were also calculated as “log2 Fold Change” (log2FC) for the N-starved sample compared to the N-replete sample. The other 2 studies: Boyle et al. (2012) and Blaby et al. (2013) also included transcriptome data, but their experiments focused on the time course and did not include N-replete samples for reference. Therefore, their 0 h samples are used in this comparison, so we also tested the statistical contrasts between the 0 h sample and the 48 h sample for both wild type strains under N-starved treatment, with the log2FC values calculated as the 48 h sample compared to the 0 h sample. Boyle et al. (2012) has no biological replicates, and their
transcriptome results were presented only for genes involved in lipid metabolism; Blaby et al. (2013) focused on comparisons with sta6 mutant strains, so the specific statistical analyses between 0 h and 48 h sample on the wild type strain were not included. Because full expression level data sets were not presented in either Boyle et al. (2012) or Blaby et al. (2013), we only considered the available information and the genes log2FC values in the comparison of our data with theirs. It is noted that the medium was changed during the active growth phase of Chlamydomonas cells (cell density about 2 to 5 million cells per mL), and the additional 48 h culturing time may have a considerable effect on these genes expression levels, so using the 48 h N-replete culture as control (Miller et al. 2010) is more likely to reveal the gene expression changes affected only by the nitrogen depletion.

The formation of TAG storage lipid is one of the main interests in this study, so the genes involve in TAG synthesis are included in this comparative study, and the results are summarized in Table 2, and a graphic illustration of the roles of these genes in the TAG synthesis pathway is also presented (Figure 5). Important changes in gene expression (> 2 fold change, meaning log2FC < -1 or > 1) captured in our test are also indicated in Figure 5.

As indicated in Table 2 and Figure 5, we observed some down regulations of many de novo fatty acid synthesis genes, including components of acetyl-CoA carboxylase and fatty acid synthase, in response to N-starvation either comparing N-starved vs. N-replete cells at 48 h. This observed decrease in expression of de novo fatty
acid synthesis genes contrasts with Miller et al. (2010): who saw no significant decrease in these genes comparing N-starved vs. N-replete mixotrophic cells, and the Acyl-ACP thiolase gene \( \text{FAT1} \), and one of the 3-ketoacyl-ACP-synthase genes \( \text{KAS1} \) were even up-regulated. Therefore, this general down-regulation appears unique for autotrophic cells. On the other hand, the majority of the TAG synthesis genes downstream from \textit{de novo} fatty acid synthesis generally were comparable in the N-starved and N-replete conditions for autotrophic cells, except for a significant down regulation of \( \text{DGAT3} \) in both 137c and 21gr and an up-regulation of \( \text{DGTT1} \) and \( \text{DGTT5} \) in 21gr only. These observations for expression are consistent with data for mixotrophic cells (Miller et al. 2010), except for our observed down-regulation of \( \text{DGAT3} \), and their reported up-regulation of \( \text{DGTT1}, \text{PDAT1}, \text{PAP1} \) and \( \text{PGA4} \).

When we were comparing the 48 h N-starved vs. 0 h controls, the down regulation of \textit{de novo} fatty acid synthesis genes appeared to be more significant for both 137c and 21gr strains. The results of previous studies comparing 48 h N-starved vs. 0 h N-starved mixotrophic cells reported either down-regulation of ACCase genes only (Boyle et al. 2012) or no down-regulation at all (Blaby et al. 2013). General up-regulation of diacylglycerol acyltransferase (DGAT) genes was reported by Blaby et al. (2013) in the 48 h vs 0 h comparisons for mixotrophic cells. However, we only observed up-regulation of \( \text{DGTT1} \) and \( \text{DGTT5} \) for both 137c and 21gr, which is consistent with the observations of Boyle et al. (2012) for mixotrophic cells.
We observed a general down-regulation of photosynthesis light reaction and light absorption genes in the autotrophic cells in response to N-starvation. Without acetate in the medium, the only energy supply and carbon assimilation is through photosynthesis, and the observed down-regulation of a majority of the light reaction related genes is consistent with the data on mixotrophic cells from Miller et al. (2010) (supplemental data table 4). This same general response was also observed in the starchless strains we tested (cluster 1 in figure 3), which indicates that the stress of N-starvation affects the expression of photosynthesis light absorption and electron transfer similarly in cells from a wide variety of strains.

Interestingly, the Calvin cycle genes show considerable variation in regulation compare to observations by Miller et al. (2010) (Table 3 and Figure 6). Up-regulation of GAP1 (glyceraldehyde 3-phosphate dehydrogenase), and down-regulation of FBA1 (Fructose-1,6-bisphosphate aldolase), GAP3 (Glyceraldehyde 3-phosphate dehydrogenase), SEBP1 (Sedoheptulose-1,7-bisphosphatase), and RBCS2A (Rubisco small subunit 2) agree with Miller et al. (2010), especially for the 137c strain. Other genes reported as down-regulated by Miller et al. (2010) did not exhibit statistically significant down-regulation, and TPIC (Triose phosphate isomerase) and TAL2 (Transaldolase) even have mildly increased average FPKM transcript levels for the 21gr strain. We suspect the down-regulation of Rubisco expression is due to a need to scavenge nitrogen for protein synthesis, since Rubisco is the most abundant protein in the cell. In general, expression of Calvin cycle enzyme genes is better maintained in the autotrophic cultures, and we suspect that the absence of an organic carbon source
(acetate) may be responsible for this result. Moreover, any down-regulation of Calvin cycle genes is much more obvious when comparing the 48 h culture to the 0 h control, but aging effects on cultures may also have an impact on this comparison, in addition to any impact of N-starvation itself. It is also noted that the two Fructose-1,6-bisphosphate aldolase genes \textit{FBA1} and \textit{FBA3} were regulated differently in our study (although they were uniformly down regulated in Miller et al. 2010) by the N-starvation, so as the two Glyceraldehyde 3-phosphate dehydrogenase genes \textit{GAP1} and \textit{GAP3} (they were regulated in different directions in Miller et al. 2010, which agrees to our autotrophic experiment results). The subcellular compartment targeting predictions for \textit{FBA1} and \textit{GAP1} are still ambiguous (reviewed by Spalding et al. 2009, table 8.1 page 272), so it is possible these genes are functioning in the cytosol and participated in the gluconeogenic process, however further molecular level studies are needed to confirm this speculation.

The glyoxylate cycle is a major acetate assimilation pathway for \textit{Chlamydomonas}. As indicated in Table 4, Miller et al. (2010) observed a general down regulation of \textit{ACS2} and \textit{ACS3} (acetyl CoA synthetase), \textit{CIS2} (citrate synthase), \textit{ICL1} (isocitrate lyase), \textit{MAS1} (malate synthase), and \textit{ACH1} (aconitate hydratase) in mixotrophic cells. Although we also observed a down-regulation of \textit{ACS2}, \textit{MAS1} and, to a lesser extent, \textit{ICL1} (Table 4), similar to the Calvin cycle genes, we generally observed less down-regulation with the autotrophically grown cells. In fact, \textit{CIS2} and \textit{ACH1} were actually up-regulated in the N-starved samples. Figure 7 illustrates the positions of these down-regulated genes in acetate assimilation and the glyoxylate cycle, showing that the incorporation of acetyl-
CoA into organic acids may not be affected significantly by the N-starvation condition. Since autotrophic cells do not receive an external supply of acetate, it is possible that under N-starvation, acetyl-CoA generated from lipid catabolism is diverted back into the energy and lipid production paths.

In wild type strains, gluconeogenesis and starch synthesis generally represent carbon flux toward products other than lipids. Our data for genes in these pathways are presented relative to those from studies under mixotrophic conditions in Table 5. We observed in autotrophic cells down-regulation of \textit{PCK1} (phosphoenolpyruvate carboxykinase), \textit{FBA1} and \textit{GAP3} (although only in one line) and up-regulation of \textit{GAP1} and \textit{PGM2} (phosphoglycerate mutase) similar to those reported by Miller et al. (2010) for mixotrophic cells (the subcellular compartment locations of \textit{FBA1} and \textit{GAP1} are ambiguous). As Figure gluconeogenesis illustrates, the \textit{PCK1} gene product catalyzes the conversion from oxaloacetate (OAA) to phosphoenolpyruvate (PEP), and the \textit{PGM} genes products catalyze the conversion between 2-phosphoglycerate and 3-phosphoglycerate. The observed down regulation of \textit{PCK1} may suggest the gluconeogenesis were slowed down by the N-starvation stress, and N-starvation triggered starch accumulation (Ball et al. 1990) is a result of less usage of sugar due to the generally affected metabolism.

With regard to starch biosynthesis, the up-regulation of \textit{PGI1} (encodes the enzyme that catalyzes the conversion between fructose-6-bisphosphate and glucose-6-phosphate) reported by Miller et al. (2010) for mixotrophic cells was not observed in our study of autotrophic cells, however this up-regulation also was not evident in Blaby et
al. (2013) or in our data when comparing 48 h to 0 h N-starved samples. We observed significantly lower expression of **STA6** in the 137c strain and of **STA2** in both strains, whereas Miller et al. (2010) reported down-regulation of **STA6** but the genome assembly used by them did not include **STA2**. Additionally, **SSS2** exhibited down-regulation in Miller et al. (2010) and in our results, although our results are not supported statistically; this situation is similar for the observed up-regulation of **SSS5** in Miller et al. (2010). **STA6** encodes the catalytic subunit of AGPase, the enzyme that catalyzes the first committed step of starch synthesis which generally continues during N-starvation, so the observed down-regulation may reflect a decreased starch synthesis by 48 h of N-starvation or may simply reflect the stress of generally lowered metabolism under N-starved conditions.

Since nitrogen sources are absent during N-starvation, it is not surprising that many genes involved in nitrogen uptake exhibited elevated transcript levels in our study of autotrophic cells, which agrees well with the results of Miller et al. (2010) using mixotrophic cells (Supplemental data table 4). Regardless of the carbon source, which organic or inorganic, *Chlamydomonas* cells apparently try to scavenge any available nitrogen in the medium.
Comparison of \textit{sta1} and wild type strains under autotrophic N-starvation conditions

One of the goals of this project was to analyze the variations in gene expression patterns between wild type strains and starchless strains under N-starved conditions. We are also specifically interested in the difference between the mixotrophic and more industrially relevant autotrophic culture conditions. Therefore, we tested different transcript expression levels by contrasting samples both from 137c to \textit{sta1}, and from 21gr to \textit{21st1} under N-starved autotrophic conditions. This comparison is analogous to previously reported work of Blaby et al. (2013), in which the differential expression comparison contrasted the \textit{sta6} strain with the cc4349 wild type and with \textit{STA6} complemented strains. Using an FDR threshold at 0.01, 2335 differentially expressed genes were identified in the contrast between 137c and \textit{sta1}, and 656 differentially expressed genes were selected from the contrast between 21gr and \textit{21st1}. Of these, only 306 genes were differentially expressed in both genotype pairs. Blaby et al. (2013) studied the \textit{sta6} mutant, which affects a different subunit of ADP-glucose-pyrophosphorylase than \textit{sta1}, and a total of 990 genes were differentially expressed when comparing \textit{sta6} to the parental or complemented strains.

Only 88 genes show differential expression in all three starchless vs. wild-type comparisons (Figure 9), perhaps because of the variation between mixotrophic and autotrophic growth conditions, even though \textit{STA1} and \textit{STA6} encode subunits of a same enzyme. However, even when comparing two starchless vs. wild type pairs in which both contained the same starchless mutation and were grown under autotrophic
conditions, only 306 common, differentially expressed genes were identified. Genes with statistically significant expression differences in our starchless vs. wild type comparisons with autotrophic cells and in the Blaby et al. (2013) study with mixoptrophic cells are summarized in Supplemental Table 2.

Two genes involved in the gluconeogenesis pathway were reported to be up-regulated in the \textit{sta6} mutants (Blaby et al. 2013): \textit{PCK1} (phosphoenolpyruvate carboxykinase; catalyzes the formation of phosphoenolpyruvate from oxaloacetate), and \textit{FBP1} (fructose-1,6-bisphosphatase; catalyzes the dephosphorylation of fructose-1,6-bisphosphate to fructose-6-phosphate). \textit{PCK1} is significantly up-regulated in both \textit{sta1} mutants and the \textit{sta6} mutant, and is included among the 88 genes differentially expressed in all three starchless vs. wild type pairs. \textit{FBP1}, on the other hand, is up-regulated in the \textit{sta6} mutant under mixotrophic conditions, but under autotrophic conditions only showed statistically significant transcript variation in the 21gr vs. 21st1 comparison, not in the 137c vs. \textit{sta1} comparison.

Three key enzymes in the glyoxylate cycle and acetate assimilation, \textit{ACS3}, \textit{ICL1}, and \textit{MAS1}, are up-regulated in the \textit{sta1} mutants under autotrophic conditions, as well as in \textit{sta6} under mixotrophic conditions (Blaby et al. 2013). The consistency between autotrophic and mixotrophic growth conditions suggests that the regulation of these enzymes is not directly linked to the availability of external acetate, but that these genes rather respond to a generally altered carbon flow under nitrogen limited stress conditions, an altered flow that includes glyoxylate cycle activity and acetate
metabolism. The 88 differentially expressed genes common to all three starchless vs. wild type comparisons includes 2 genes encoding components of the light harvesting complex of photosystem II: \textit{LHCBM2} and \textit{LHCBM6}, which exhibit similar up-regulation in both the \textit{sta1} and \textit{sta6} mutants under both mixotrophic and autotrophic conditions.

In addition to the 88 common differential expression genes, the expression levels of another 218 genes are significantly varied in \textit{sta1} mutants under autotrophic, nitrogen-limited conditions, but were not detected as differentially expressed in \textit{sta6} mutants under mixotrophic growth. Another acetate assimilation gene \textit{ACS2}, encoding an acetyl-coA synthase is included within this group, as well as the \textit{SQD1} gene, which encoding the UDP-sulfoquinovose synthase involved in thylakoid membrane lipid metabolism. Both of these genes are up-regulated in both \textit{sta1} mutant strains when nitrogen is limited.

It is also notable that 2 TAG synthesis gene, \textit{DGTT2} and \textit{DGAT3}, and the major lipid droplet protein gene \textit{MLDP} were identified as having elevated expressions in \textit{sta6} mutants compare to the wild type strains under mixotrophic conditions (Blaby et al. 2013), but were not identified as differentially expressed between \textit{sta1} mutants and their corresponding wild type strains in our experiments with autotrophic cultures. However, the type 1 diacylglycerol acyltransferase gene, \textit{DGTT1}, and the glycerol-3-phosphate dehydrogenase gene, \textit{GPD2}, were significantly up-regulated under N-starved conditions in all 4 strains, and there are no significant variations between wild type strains and \textit{sta1} mutants in our tests under autotrophic conditions, which is comparable
to the mixotrophic experiments results in Blaby et al (2013). On the other hand, $DGAT3$ was down-regulated when nitrogen is limited in the media for our autotrophic cultures, and we did not observe significant different expressions between the wild type and starchless strains (Supplemental Table 3), which is varied from the observed up regulation in sta6 compare to the wild type (Blaby et al. 2013) when cells were cultured mixotrophically. It is possible the regulation of the putative type 3 diacylglycerol acyltransferase gene DGAT3 is varied from other diacylglycerol acyltransferase genes, however the detailed information is still limited at this time.

We also compared expression of genes involved starch synthesis and compared them with results reported in Blaby et al. (2013) (Table 6 starch synthesis starchless compare). In both the mixotrophic sta6 and the autotrophic sta1 comparisons, other than the mutated genes themselves, the only statistically significant change captured was up-regulation of $STA2$ for the sta1 and 137c strains comparison, although other strain comparisons exhibited similar but not statistically significant up-relation of $STA2$. It is notable that $STA2$ was down-regulated in wild type strains when comparing N-starved with N-replete conditions (Table 5), so it is unclear why the transcript level of this gene is elevated when ADP-glucose pyrophosphorylase is not functional.
Nitrogen is one of the essential elements for all life forms on earth, so it is intuitive that an N-starved environment is unfavorable for active cell growth. From our studies, it is clearly shown that *Chlamydomonas* up-regulates genes associated with nitrogen assimilation, which indicates an effort to scavenge as much of the remaining nitrogen as possible. At the same time, photosynthetic light reactions are severely affected, which can be understood as a mechanism to generally slow down metabolism and also to minimize nitrogen investment in the active growth metabolic processes. Both of which should assist with survival under the nutrient limited situations. We also observed this inhibition of growth by comparison of the cell dry weights per unit culture volume at the end of the experiment (Table 7), which confirmed decreased biomass accumulation in the nutrient starved cultures. On one hand, the ability of *Chlamydomonas* to acclimate to the environment when key nutrient components are scarce is clearly evident, but, on the other hand, cell growth was definitely decreased, which is counterproductive if one wishes to utilize *Chlamydomonas* for biofuel production.

Most of the previous *Chlamydomonas* research about N-starved conditions was conducted under mixotrophic growth conditions, where acetate is provided as an external, organic carbon source. Although acetate is a welcome nutrient for cell growth, the cost of acetate for industrial algal biomass production is clearly not favored either economically or energetically, and the use of acetate to produce biofuel is counter to
the motivation for biofuels, which is to reverse the CO₂ emission process. Therefore, one of the goals for this study is to investigate the differences when *Chlamydomonas* are cultured photoautotrophically, in comparison with observations reported from mixotrophic conditions (Miller et al. 2010; Blaby et al. 2013).

Considering the interest in biofuel production potential, we first carefully focused on the genes involved in the TAG synthesis pathway (Table 2 and Figure 5). As presented in the Results section, many *de novo* fatty acid synthesis genes are down-regulated in N-starved autotrophic cultures in contrast to the data from previous work with mixotrophic cultures (Miller et al. 2010; Blaby et al. 2013). Unlike mixotrophic cultures that can utilize acetate as an energy and carbon source, autotrophic cultures can only harvest energy through photosynthetic light reactions. When nitrogen is limited, it appears that the production of major light reaction proteins becomes hindered to minimize nitrogen investment. The cells also might be shifting from an “active growth state” to a “preserving and storage” state. In this state, if external acetate is provided, it can be easily be converted to acetyl-CoA (acetate kinase genes *ACK1/2* and phosphate acetyltransferase gene *PAT1* did not exhibit significantly lowered expression in either autotrophic and mixotrophic experiments, see Table 4), the resources for *de novo* fatty acid synthesis are still widely available, which may explain why transcript level variations of *de novo* fatty acid synthesis genes occur between autotrophic and mixotrophic culture.
External CO\textsubscript{2} is the preferred carbon source for biofuel production, since CO\textsubscript{2} and sunlight are relatively inexpensive, and the fixation of CO\textsubscript{2} is carried out by the Calvin cycle pathway in \textit{Chlamydomonas}. Calvin cycle enzymes, especially Rubisco, contribute as major protein components in photosynthetic species. When an external organic carbon source (acetate) is available, the functional necessity of the Calvin cycle becomes less critical for survival compared to pure autotrophic growth. Therefore, it is probably economical under mixotrophic growth conditions for cells to lower the expression of Calvin cycle enzymes to preserve limited nitrogen resources. We observed down regulations of some Calvin cycle genes, including \textit{FBA1}, \textit{GAP3}, and the highly expressed \textit{RBCS2A} (~30,000 FPKM maximum for 21gr), under autotrophic conditions, but the transcript levels for the genes involved in this carbon fixation pathway were much better maintained under autotrophic conditions in contrast to the decreased expression reported for mixotrophic experiments (Table 3, Miller et al. 2010). We found that the \textit{RBSC1} gene (> 5,000 FPKM maximum for 21gr) exhibited even slightly elevated average FPKM values in N-starved conditions relative to N-replete conditions. If we consider that direct CO\textsubscript{2} assimilation is a critical goal for biofuel production, providing acetate to cultures seems to be a less efficient strategy, because genes of the major CO\textsubscript{2} assimilation pathway were obviously affected more compared to those in acetate-free, autotrophic cultures. On the other hand, if we are looking for potential high value lipid production practices, the cost of supplying extra acetate might be very beneficial.

Although a detailed portrait of acetate uptake and metabolism is still missing, it is widely accepted that acetate probably is assimilated via the glyoxylate cycle following
its conversion to acetyl-CoA. While the acetate assimilation path through acetate phosphate is better maintained, the major glyoxylate cycle genes like *MAS1*, and *ICL1* are down regulated under N-starved conditions; and similar down regulations were also observed for these genes in both mixotrophic or autotrophic experiments (Table 4). Additionally, the phosphoenolpyruvate carboxykinase gene *PCK1* (catalyzing the conversion from oxaloacetate to phosphoenolpyruvate, which is the first committed step of gluconeogenesis through oxaloacetate, see Table 5 and starch and Figure 8) is also down regulated in both autotrophic and mixotrophic experiments. This may reflect a re-balancing of carbon flow toward fatty acid synthesis and away from carbohydrate synthesis through organic acids when mixotrophic cultures switch to “storage state”, and simply preserving energy and carbon in autotrophic cultures. When we further examined expression of several genes putatively involved in the fatty acid beta-oxidation process (Table 8), although in our case the changes are milder and not statistically significant, we observed similar general down-regulation of these genes, except *ECH1*, in both autotrophic and mixotrophic cultures. We believe this observation is consistent with our hypothesis that the algae cells are not breaking down fatty acid to supply carbon for carbohydrate accumulation.

One of the diversions of carbon flow into the storage path is into starch synthesis and accumulation, so we also included the starchless mutant *sta1* and its genetic variant *21st1* in our transcriptome study. Other than genes mentioned in the Results section, many of the genes detected as differentially expressed in starchless mutants compare to the wild type strains under N-starved conditions did not have annotations defining their
functional importance directly related to metabolic processes; and the observed differential gene expressions between wild type and sta1 mutant for our autotrophic experiments are highly comparable to the mixotrophic sta6 mutant experiment (Blaby et al. 2013).

When we compared our results to those of Blaby et al. (2013), we also carefully examined transcript levels of genes involved in the starch synthesis process. The only interesting observation was the elevated expression of the STA2 gene in both sta1 and sta6 mutants compare to wild type strains (Table 6). The STA2 gene was previously identified as very important for long chain amylose synthesis (Wattebled et al. 2002), and it was down regulated in autotrophic, N-starved conditions (Table 5). On the other hand, STA2 expression was maintained in strains lacking functional ADP-glucose-pyrophosphorylase (Table 5 and Table 6 compared to Blaby et al. 2013). Perhaps amylose synthesis is affected both by nutrient conditions and by starch accumulation, but further experiments will be needed to reveal the relationship between these factors.

Several glyoxylate cycle genes and the gluconeogenic gene PCK1 exhibited higher transcript levels in the ADP-glucose-pyrophosphorylase mutant sta6 under N-starvation stress (Blaby et al. 2013), which is to be expected, but this observation was made with sta1 mutants in autotrophic cultures as well, which was somewhat unexpected since we normally associate these pathways with acetate assimilation. The up-regulation of these genes in autotrophic growth suggests that glyoxylate cycle flux may come from acetyl-
CoA generated internally from, e.g., lipid catabolism, since exogenously supplied acetate is absent. Additionally, although increased expression of genes associated with acetyl-CoA metabolism occurs in starchless mutants in mixotrophic growth conditions, acetate consumption from the medium was similar between wild type and *sta6* (Blaby et al. 2013), arguing against increased acetate flux in the mutant, even though previous work showed that the amount of external acetate supplied during mixotrophic growth positively correlated with lipid droplet sizes (Goodson et al., 2011; Ramanan et al., 2013). The direct connection between defects in starch synthesis and elevated glyoxylate cycle gene expression in both autotrophic and mixotrophic cultures suggests that these cells may be catabolizing their lipid reserves for carbon and energy in the absence of starch as a normal source for carbon and energy under limited photosynthesis conditions (i.e., in the wild-type strains in the same N-starvation experiments). Interestingly, this hypothesis also suggests that *Chlamydomonas* is catabolizing lipid even while accumulating net TAG under N-starved conditions. An alternative explanation may be the proposed hypothesis (Blaby et al. 2013) regarding downstream signaling by trehalose formation in the ADP-Glucose pyrophosphorylase mutants (Bölling and Fiehn, 2005), which may not be mutually exclusive if trehalose accumulation signals the need for lipid catabolism. These hypotheses should be investigated further.

Under nitrogen starvation conditions, expression profiles of many genes are altered (Miller et al. 2010). It would be very interesting to observe whether different N-starvation responses occur in starchless mutants. The *sta1* mutant was generated from
137c wild type by X-ray mutagenesis, and the 21st1 mutant was generated by crossing sta1 with 21gr wild type. Therefore, many other genotypic variations should exist between the mutants and their reference wild type strains in addition to the mutation in the STA1 locus. That this genomic background caused variation can readily be observed from the large amount of non-overlapping DE genes generated when the sta1/WT differential expression was compared between the genome pairs. Therefore, we also added the criteria that the selected N-starvation response genes must be shared in both 137c/sta1 and 21gr/21st1 comparisons. In this way we were hoping to filter out genes not affected by STA1 but rather by other genomic variations. After the application of these screening criteria, we only identified 7 genes that responded to the N-starvation stress differently in the starchless strains, and we did not find the expression patterns and the current annotations for these 7 genes as clearly related to specific metabolic functions. Since ADP-glucose pyrophosphorylase is not recognized as a regulatory protein, the sta1 mutation may not dramatically affect the N-starvation responses in spite of its clear impact on starch synthesis and carbohydrate metabolism.

During careful examination of the gene expression profile for several biological processes, we identified many genes that behaved differently between the starchless strains and the wild type strains, like the ACS2/3, MAS1, and ICL1 genes involved in the glyoxylate cycle pathway. This suggests that the criteria we applied to the statistical analysis may have been too strict. It is also notable that the differentially expressed genes exceeded 3,300 (more than 4,000 if using version 4 gene model) in the Miller et al. (2010) study for 1 wild type strain, whereas only about 1,100 to 1,600 differentially
expressed genes were detected for each strain we tested. A major difference that can affect the detection power of statistical analyses is the biological replication number, which was three replications in Miller et al. (2010), but only two in our study. Further investigations may provide additional evidence to indicate whether additional replications would reveal more genes or whether the variation in the number of identified differentially expressed genes is due to biological factors (culturing conditions and strains).

The experimental design for our experiments consisted of 4 different genotypes, 2 culture conditions (N-starved and N-replete), and 3 time points. Such a complete and balanced design provided an opportunity to conduct complicated tests. For example, the first test, which identified genes with altered N-starvation responses in the wild type and mutant strains, regardless of the genetic backgrounds, yielded only 7 genes meeting these stringent criteria; if we do not consider the N-replete reference (i.e., only compare the 48 h N-starved condition with the 0 h condition), the number of differentially expressed genes shared by 137c/sta1 and 21gr/21st1 under N-starvation will be 306 (Figure 2), which may include genes that change their expression because of the mutation alone. As sequencing technology advances, many statistical analysis tools are being developed to handle complex designs. The EdgeR package we used in this study is virtually capable to conduct any combination of contrast tests for complex designs, and because the number of samples is relatively large, in theory the dispersion estimation proposed by the EdgeR package can improve on its accuracy (Robinson et al. 2010). We
highly recommend the generalized linear model approach to the simple pairwise comparison when the biological question and the design involve multiple factors.

*Chlamydomonas* and many other algae species accumulate lipids under N-starved conditions, but an ample nitrogen supply is very important for producing energy harvesting and carbon fixation proteins in photosynthesis. This creates a dilemma with regard to the use of *Chlamydomonas* or other algae as bio-factories to produce TAG for biofuel or renewable bio-products industries: on one hand, nitrogen stress conditions result in decreased photosynthesis, probably because of the lack of nitrogen for synthesis of photosynthetic components, which also likely limits *de novo* fatty acid synthesis in photoautotrophically grown cells; on the other hand, over-accumulation of TAG can be achieved very effectively under N-starved conditions by supplying expensive external acetate as both an energy and carbon source (Goodenough et al. 2014), which bypasses the need for photosynthesis. However, the provision of organic carbon to support TAG accumulation defies the CO$_2$ re-capturing purpose of advanced biofuel production. However, the availability and utilization of starchless mutants blocking carbon flow to carbohydrate storage demonstrates the power of bioengineering, so perhaps with the quickly advancing modern biotechnology one day algae can be engineered to maintain high levels of fatty acid synthesis without the provision of acetate.
Materials and Methods

Strains used

*Chlamydomonas reinhardtii* 137c wild type strain was obtained from the *Chlamydomonas* resource center (University of Minnesota, Minneapolis, MN) as culture collection cc125 (mating type minus). The 21gr wild type strain (mating type plus) was obtain from the *Chlamydomonas* resource center (University of Minnesota, Minneapolis, MN) as culture collection cc1690. And the sta1 mutant strain (mating type minus) was obtained from Susan Dutcher (Washington University), who repeatedly backcrossed the original sta1 strain with CC125. The 21st1 strain was generated by crossing the 21gr strain with the sta1 stain for one generation, and we selected one of the progenies that showed starchless phenotype by the iodine vapor spray examinations mentioned in previous study (Koornhuyse et al., 1996 sta1 paper). All strains were maintained on the CO\(_2\) minimal plates in 5% CO\(_2\) conditions as described before (Fang et al. 2012).

Growth conditions and measurements, RNA isolation

Maintained cells were first transferred into liquid media and grow on a shaker for 4 days, and about 60 to 65 million cells were then used to inoculate the total volume of 670 mL in the bioreactor. The MMU medium that has urea as the nitrogen source was used as the normal culturing medium, which contains 10 mM urea, 1.22 mM, K\(_2\)HPO\(_4\), 0.76 mM KH\(_2\)PO\(_4\), 0.405 mM MgSO\(_4\)-7H\(_2\)O, 0.34 mM CaCl\(_2\)-2H\(_2\)O, 1 ml/L of Hutner’s trace elements, and buffered with 20 mM MOPS titrated with Tris to a pH = 7.3 (Bigelow et.
al., 2014). After the inoculation, the culture was grown in the bioreactor until the OD750 reached 3.5 for the 137c strain culture, and 5.0 for other 3 strains cultures. The cell density at these selected OD750 values reached roughly 18 to 20 million cells per mL for all strains. After reaching the designated OD density values (40 to 85 hours after inoculation), 0 hour time point samples were taken, and the cell were centrifuged at 1000g for 7 minutes, washed with MMU (either +N or –N by not including the urea content) once, and resuspended with fresh MMU +N/-N medium accordingly as mentioned above. By 24 or 48 hours after refreshing the medium, culture samples were taken for the RNA isolation procedures. Chlorophyll content, dry weight per mL of liquid culture, liquid culturing conditions, and the bioreactor culturing conditions are the same as previously indicated (Bigelow et. al., 2014).

**RNA isolation and sequencing**

Cell culture samples were centrifuged at 4000 g for 5 minus, and the cell pellets were immediately frozen by liquid nitrogen and temporarily stored in -80°C freezer. TRIzol® Reagent (Life Technologies) was used to fully dissolve the cell sample, and then the aqueous phase containing RNA was isolated after chloroform extraction according to the manual. Crude RNA samples were isolated by the PureLink® RNA Mini Kit (Life Technologies), and cleaned with DNase I and the RNeasy MinElute Cleanup kit (Qiagen). Cleaned RNA samples were paired end sequenced on the Illumina Hi-Seq sequencing system at BGI (Beijing Genomics Institute) Hong Kong division. Raw sequencing data files
are stored at the link: “https://iastate.box.com/s/0thb4ks6cocrhqlwxkjaw” (retrieving password: “spalding”).

**Sequencing alignment and FPKM calculation**

Raw RNA sequences were aligned to the *Chlamydomonas* version 5 genome assembly and the gene reads count were based on the Augustus annotation version u11.6 presented at the phytozome 10 website (http://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Creinhardtii), the alignment were generated by the same criteria like previously reported (Fang et. al., 2012). The gene expression counts were calculated by the cufflinks tool available at “(http://cufflinks.cbcb.umd.edu/manual.html) as fragment counts for pair-end sequencing data. The Fragments Per Kilobase of transcript per Million mapped reads (FPKM) were calculated also by using the cufflinks tool, which provided a cross comparable gene expression value.

**Statistical analysis for differentially expressed genes and clustering analysis**

Differential expression was analyzed by the EdgeR package (release 3.0 at http://www.bioconductor.org/packages/release/bioc/html/edgeR.html) on the R platform (version 3.1.1 at http://www.r-project.org/). The raw fragment count was used in the analysis, and the tagwised dispersion estimation and the generalized linear model approach was applied. Multiple contrast tests were performed accordingly, and the results are controlled by the false discovery rate FDR (Benjamini and Hochberg, 1995) to
be smaller than 1%. The hierarchical clustering analyses were based on the log₂ FPKM values.

All the supplemental tables are accessible by the following address:

https://iastate.box.com/s/zyxgvdb5ye6b1vrp7p5d
Figures and Tables

Figure legends

Figure 1: Heatmap of 7 genes that have significantly varied N-starvation responses.
Color coding is shown in the upper left corner, and log$_2$-fold change values were calculated based on FPKM. Gene names are indicated on the right, and conditions are indicated along the bottom by a series of text indicating the strain name (137c, 21gr, 21st1 or sta1), nitrogen condition (“M” for minus nitrogen, N-starvation; “P” for plus nitrogen, N-replete), and sampling time (hours after changing the media) in consecutive order (for example: 137cM24 represents strain 137c culture sampled at 24 hours after changing to nitrogen free medium). The heatmap and hierarchical clustering were generated by the gplot package in R.

Figure 2: Venn diagram summarizing the numbers of genes detected as differentially expressed in our tests for 4 strains (autotrophic culture) and in Miller et al. (2010) study (mixotrophic culture). A set of 297 genes were detected having significantly different expression levels between N-starvation and N-replete conditions are shared by all tests with FDR controlled at 0.01, and 475 genes are shared by the 4 strains we tested. Also, 701 genes included in the Miller et al. (2010) test have gene model discrepancies that prevented the gene IDs from being converted and compared directly.

Figure 3: A heatmap of 475 genes having significant N-starvation responses for all 4 tested strains. Color coding is shown in the upper left corner, and log$_2$-fold change values were calculated based on FPKM. Gene names are indicated on the right, and
conditions are indicated on the bottom by a series of text indicating the strain name (137c, 21gr, 21st1 or sta1), nitrogen condition ("M" for minus nitrogen, N-starvation; “P” for plus nitrogen, N-replete), and sampling time (hours after changing the media) in consecutive order (for example: 137cM24 represents strain 137c culture sampled at 24 hours after changing to nitrogen free medium). The heatmap and hierarchical clustering was generated by the gplot package in R.

**Figure 4: Chlorophyll content.** Chlorophyll contents (µg/10^6 cells) were measured 48 hours after shifting the media. The error bars illustrate the standard error. All 4 strains have significantly lowered chlorophyll content in N-starved medium.

**Figure 5: The TAG synthesis pathway and genes that are involved in each step.** Changes in gene expression levels greater than 2 fold in the 48 h N-starved samples compared to the 48 h N-replete samples are indicated following the gene name: “D” represents down regulation (log2FC < -1), “U” represents up regulation (log2FC > 1), and “n” means variation is less than 2 fold change. The underlined bold formatting like “D” or “U” means the change is statistically significant (FDR <0.01). These labels in parentheses are arranged by order of (137c/21gr), so the genes are labeled with parentheses if larger or equal to 2 fold expression change is detected in either strain.

**Figure 6: Summary of the Calvin cycle CO2 assimilation pathway and genes involved in each step.** Gene expression levels that varied over 2 fold in 48 h N-starved samples compared to 48 h N-replete samples are illustrated following the gene name: “D” represents down regulation (log2FC < -1), “U” represents up regulation (log2FC > 1), and
“n” means variation is no more than 2 fold. The underlined bold formatting like “D” or “U” means the change is statistically significant (FDR <0.01). Labels in the parentheses are arranged in the order of (137c/21gr), so genes are labeled with parentheses if >2 fold change in expression is detected in either strain.

**Figure 7: Summary of the glyoxylate cycle pathway genes involved in each step.** Gene expression levels varied over 2 fold changes in the 48 h N-starved samples compared to the 48 h N-replete samples are indicated following the gene name: “D” represents down regulation (log2FC < -1), “U” represents up regulation (log2FC > 1), and “n” means no more than 2 fold change. The underlined bold formatting, like “D” or “U”, indicates the change is statistically significant. These labels in parentheses are arranged to indicate the strains in the order of (137c/21gr).

**Figure 8: Summary of the gluconeogenesis pathway and genes involved in each step.** Gene expression that varied >2 fold in 48 h N-starved samples relative to 48 h N-replete samples are indicated following the gene name: “D” represents down-regulation (log2FC <-1), “U” represents up-regulation (log2FC >1), and “n” indicates no more than a 2 fold change. The underlined bold formatting, like “D” or “U” means the change is statistically significant. These labels in parentheses are arranged in order of (137c/21gr), so expression changes are indicated for genes in either strain if >2 fold change was detected.

**Figure 9. Numbers of genes differentially expressed under different comparisons.**

Comparisons of “137c vs. sta1” and “21gr vs. 21st1” were conducted to contrast
samples collected under N-starved autotrophic growth conditions; DE gene numbers included in “sta6 vs cc4349 or STA6 mixotrophic” are combined from previous published work (Blaby et al. 2013), which compared sta6 mutants with the cc4349 parental strain or STA6 complemented strains.

**Figures:**

Heatmap of 7 genes that are having different N-starve response between sta1 mutants and wild types:

Figure 1
701 genes have discrepancy issues, their version 4 ID can not be converted to version 5.5 ID.

Figure 2

Heatmap and hierarchical clustering of 475 N-stavation response genes that are shared by 4 strains:

Figure 3
Figure 4

chlorophyll content at 48 hours after switch media

- **N replete**
- **N starved**

Figure 5

Diagram of lipid metabolism:
- Acyl-CoA
- Malonyl-CoA
- Glycerol-3-phosphate
- Lysophosphatidic acid
- Membrane lipids
- Phosphatidic acid

Enzymes and genes involved:
- ACP1
- ACP2
- MCT1 (D/D)
- FAT1
- KAS1 (n/n)
- KAS2 (D/n)
- KAE1 (D/n)
- HAD1
- ENR1 (D/n)
- acyl-CoA
- PLSB1
- DGAT1
- DGAT2
- DGAT3 (D/D)
- DGAT5 (n/n)
- DGTT1
- DGTT2
- DGTT3
- DGTT4
- DGTT5 (n/n)
- KDG5
- KDG6
- KDG7
- PTAH1
- PAP1
- PAP2
- PGA2
- PGA4 (n/n)
Figure 8

Figure 9
Tables:

Table 1:

<table>
<thead>
<tr>
<th>MapMan Ontology level 2 terms</th>
<th>Hits</th>
<th>Score</th>
<th>Summary of expression patterns</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cluster 1: 210 genes</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lipid metabolism, FA synthesis and FA elongation</td>
<td>9</td>
<td>4.4E-10</td>
<td></td>
</tr>
<tr>
<td>PS.lightreactions</td>
<td>9</td>
<td>8.4E-09</td>
<td></td>
</tr>
<tr>
<td>protein synthesis</td>
<td>14</td>
<td>4.4E-06</td>
<td>Decreased transcript levels under N-starved condition, mild decrease in N-abundance condition for late 48 hour cultures</td>
</tr>
<tr>
<td>nucleotide metabolism, phosphotransfer and pyrophosphatases</td>
<td>4</td>
<td>0.0002</td>
<td></td>
</tr>
<tr>
<td>cell cycle</td>
<td>4</td>
<td>0.003</td>
<td></td>
</tr>
<tr>
<td>protein.folding</td>
<td>4</td>
<td>0.003</td>
<td></td>
</tr>
<tr>
<td>hormone metabolism, auxin</td>
<td>2</td>
<td>0.009</td>
<td></td>
</tr>
<tr>
<td><strong>Cluster 2: 120 genes</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>amino acid metabolism, synthesis</td>
<td>4</td>
<td>0.003</td>
<td></td>
</tr>
<tr>
<td>transport,calcium</td>
<td>2</td>
<td>0.005</td>
<td>Decreased expression only in N-abundance conditions.</td>
</tr>
<tr>
<td>transport, potassium</td>
<td>2</td>
<td>0.009</td>
<td></td>
</tr>
<tr>
<td>transport, unspecified cations</td>
<td>2</td>
<td>0.009</td>
<td></td>
</tr>
<tr>
<td><strong>Cluster 3: 99 genes</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S-assimilation, APS</td>
<td>1</td>
<td>0.006</td>
<td>Increased expression in N-abundance condition for late 48 hour cultures</td>
</tr>
<tr>
<td>transport, sulphate</td>
<td>1</td>
<td>0.006</td>
<td></td>
</tr>
<tr>
<td><strong>Cluster 4: 46 genes</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N-metabolism, nitrate metabolism</td>
<td>2</td>
<td>8.0E-05</td>
<td>Increased expression under N-starved conditions.</td>
</tr>
<tr>
<td>transport, misc</td>
<td>1</td>
<td>0.006</td>
<td></td>
</tr>
<tr>
<td>transport, nitrate</td>
<td>1</td>
<td>0.010</td>
<td></td>
</tr>
</tbody>
</table>

MapMan gene ontology analysis for the 4 distinct clusters of 475 genes with significant N-starvation response for all 4 tested strains. Analysis was conducted using the Algal Functional Annotation Tool (Lopez et al. 2012). Level 2 MapMan gene ontology terms with less than 0.01 score are presented.
Table 2:

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>48 hour &quot;-N&quot; vs &quot;+N&quot; log2FC</th>
<th>&quot;-N&quot; 48 vs 0 hour log2FC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FPKM 21gr 137c dw15.1</td>
<td>Fang</td>
<td>Miller</td>
</tr>
<tr>
<td>ACX1</td>
<td>α-Carboxyltransferase</td>
<td>200.9</td>
<td>-1.22</td>
</tr>
<tr>
<td>BCC1</td>
<td>Acetyl-CoA biotin carboxyl carrier</td>
<td>533.7</td>
<td>-1.99</td>
</tr>
<tr>
<td>BCC2</td>
<td>Acetyl-CoA biotin carboxyl carrier</td>
<td>505.4</td>
<td>-1.73</td>
</tr>
<tr>
<td>BCX1</td>
<td>β-Carboxyltransferase</td>
<td>117.3</td>
<td>-0.71</td>
</tr>
<tr>
<td>ACP1</td>
<td>Acyl carrier protein</td>
<td>294.0</td>
<td>-0.30</td>
</tr>
<tr>
<td>ACP2</td>
<td>Acyl carrier protein</td>
<td>2964.3</td>
<td>-1.24</td>
</tr>
<tr>
<td>MCT1</td>
<td>Malonyl-CoA:acyl carrier protein</td>
<td>254.6</td>
<td>-3.36</td>
</tr>
<tr>
<td>FAT1</td>
<td>Acyl-ACP thiase</td>
<td>31.4</td>
<td>-0.82</td>
</tr>
<tr>
<td>KAS1</td>
<td>3-ketoacyl-ACP-synthase</td>
<td>157.3</td>
<td>-0.63</td>
</tr>
<tr>
<td>KAS2</td>
<td>3-ketoacyl-ACP-synthase</td>
<td>62.0</td>
<td>-1.37</td>
</tr>
<tr>
<td>KAR1</td>
<td>3-Ketoacyl-ACP reductase</td>
<td>357.8</td>
<td>-1.22</td>
</tr>
<tr>
<td>HAD1</td>
<td>3-Hydroxyacyl-ACP dehydratase</td>
<td>203.9</td>
<td>0.09</td>
</tr>
<tr>
<td>ENR1</td>
<td>Enoyl-ACP reductase</td>
<td>251.4</td>
<td>-1.24</td>
</tr>
<tr>
<td>PLSB1</td>
<td>Glycerol-3-phosphate acyltransferase</td>
<td>90.3</td>
<td>-0.31</td>
</tr>
<tr>
<td>PDA1</td>
<td>Phospholipid diacylglycerol acyltransferase</td>
<td>3.4</td>
<td>0.12</td>
</tr>
<tr>
<td>PAH1</td>
<td>Phosphatidate phosphatase</td>
<td>3.3</td>
<td>0.27</td>
</tr>
<tr>
<td>PAP1</td>
<td>phosphatidate phosphatase</td>
<td>8.1</td>
<td>0.49</td>
</tr>
<tr>
<td>PAP2</td>
<td>phosphatidate phosphatase</td>
<td>11.9</td>
<td>0.10</td>
</tr>
<tr>
<td>KDG1</td>
<td>Diacylglycerol kinase</td>
<td>4.6</td>
<td>-0.32</td>
</tr>
<tr>
<td>KDG3</td>
<td>Diacylglycerol kinase</td>
<td>10.1</td>
<td>-0.36</td>
</tr>
<tr>
<td>DGAT1</td>
<td>Diacylglycerol acyltransferase, DAGAT type-1</td>
<td>8.2</td>
<td>-0.69</td>
</tr>
<tr>
<td>DGAT3</td>
<td>Diacylglycerol acyltransferase, DAGAT type-3</td>
<td>255.8</td>
<td>-2.87</td>
</tr>
<tr>
<td>DGTT1</td>
<td>Diacylglycerol acyltransferase, DAGAT type-2</td>
<td>22.4</td>
<td>0.84</td>
</tr>
<tr>
<td>DGTT2</td>
<td>Diacylglycerol acyltransferase, DAGAT type-2</td>
<td>18.9</td>
<td>-0.06</td>
</tr>
<tr>
<td>DGTT3</td>
<td>Diacylglycerol acyltransferase, DAGAT type-2</td>
<td>45.6</td>
<td>0.47</td>
</tr>
<tr>
<td>DGTT4</td>
<td>Diacylglycerol acyltransferase, DAGAT type-2</td>
<td>10.4</td>
<td>-0.61</td>
</tr>
<tr>
<td>DGTT5</td>
<td>Diacylglycerol acyltransferase, DAGAT type-2</td>
<td>0.2</td>
<td>-0.30</td>
</tr>
<tr>
<td>PGA2</td>
<td>Lysophosphatidyl lipid acyltransferases</td>
<td>13.1</td>
<td>0.97</td>
</tr>
<tr>
<td>PGA4</td>
<td>Lysophosphatidyl lipid acyltransferases</td>
<td>1.4</td>
<td>0.51</td>
</tr>
</tbody>
</table>

This table illustrates the N-starvation effect on expression of TAG synthesis genes in our study, and compares these changes those observed with previous studies (Miller et al., 2010; Boyle et al. 2012; Blaby et al. 2013). The "FPKM 21gr" column indicates the gene’s highest FPKM value in the 21gr strain for all treatments and time points, which illustrates the gene’s relative expression level under induced conditions. Two groups of tests and comparisons were included: 48 hour "-N" vs "+N" is a comparison between N-starved and N-replete conditions for 48 h samples only, which compares our results with Miller et al. (2010) only; and "-N" 48 vs 0 h is a comparison between 48 h samples and 0 h samples in N-replete conditions only, which compares our results with Boyle et al. (2012) and Blaby et al. (2011). The log2FC values were calculated from the average
FPKM in our study and the average RPKM in other studies: red or green color highlights log2FC values lower than -1 or higher than 1, respectively, which means they are down-regulated or up-regulated in the 48 h N-starved condition. The underlined bold formatting, such as “red” or “green” indicates that the change is statistically significant (FDR <0.01). Indications of “down” or “up” means that in one of the conditions no counts were captured in the RNA sequencing data, so the calculated fold change is meaningless. Genes with names in bold characters are involved in the de novo fatty acid synthesis.

Table 3:

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>21gr</th>
<th>137c</th>
<th>21gr</th>
<th>dw15.1</th>
<th>137c</th>
<th>21gr</th>
<th>cc-4349</th>
</tr>
</thead>
<tbody>
<tr>
<td>FBA1</td>
<td>Fructose-1,6-bisphosphate aldolase</td>
<td>384.1</td>
<td>-4.05</td>
<td>-4.06</td>
<td>-3.29</td>
<td>-0.74</td>
<td>-2.44</td>
<td>-1.77</td>
</tr>
<tr>
<td>FBA3</td>
<td>Fructose-1,6-bisphosphate aldolase</td>
<td>1736.3</td>
<td>-0.77</td>
<td>1.46</td>
<td>-3.04</td>
<td>-0.96</td>
<td>-0.76</td>
<td>-0.73</td>
</tr>
<tr>
<td>FBP1</td>
<td>Fructose-1,6-bisphosphatase</td>
<td>478.0</td>
<td>-0.96</td>
<td>-0.45</td>
<td>-1.07</td>
<td>-1.11</td>
<td>-1.80</td>
<td>-0.43</td>
</tr>
<tr>
<td>GAP1</td>
<td>Glyceraldehyde 3-phosphate dehydrogenase</td>
<td>404.8</td>
<td>1.49</td>
<td>1.39</td>
<td>1.55</td>
<td>0.46</td>
<td>1.47</td>
<td>1.60</td>
</tr>
<tr>
<td>GAP3</td>
<td>Glyceraldehyde 3-phosphate dehydrogenase</td>
<td>2773.9</td>
<td>-2.52</td>
<td>-0.99</td>
<td>-4.98</td>
<td>-2.19</td>
<td>-2.84</td>
<td>-2.09</td>
</tr>
<tr>
<td>PGK1</td>
<td>Phosphoglycerate kinase</td>
<td>1004.8</td>
<td>-1.26</td>
<td>-0.64</td>
<td>-4.21</td>
<td>-3.31</td>
<td>-3.63</td>
<td>-4.44</td>
</tr>
<tr>
<td>PRK1</td>
<td>Phosphoribulokinase</td>
<td>830.8</td>
<td>0.28</td>
<td>0.87</td>
<td>-2.52</td>
<td>-2.52</td>
<td>-3.19</td>
<td>-1.26</td>
</tr>
<tr>
<td>RBCS1</td>
<td>Rubisco small subunit 1</td>
<td>5392.0</td>
<td>0.67</td>
<td>0.32</td>
<td>-3.44</td>
<td>-0.98</td>
<td>-1.79</td>
<td>-0.64</td>
</tr>
<tr>
<td>RBCS2A</td>
<td>Rubisco small subunit 2</td>
<td>29521.3</td>
<td>-3.43</td>
<td>-2.78</td>
<td>-3.33</td>
<td>-2.62</td>
<td>-2.97</td>
<td>-1.77</td>
</tr>
<tr>
<td>RPE1</td>
<td>Ribulose phosphate-3-epimerase</td>
<td>926.0</td>
<td>-0.08</td>
<td>0.82</td>
<td>0.30</td>
<td>0.07</td>
<td>-0.52</td>
<td>0.38</td>
</tr>
<tr>
<td>RPI1</td>
<td>Ribose-5-phosphate isomerase</td>
<td>659.2</td>
<td>-1.26</td>
<td>-0.06</td>
<td>-1.39</td>
<td>-1.82</td>
<td>-2.13</td>
<td>-1.02</td>
</tr>
<tr>
<td>SEBP1</td>
<td>Sedoheptulose-1,7-bisphosphatase</td>
<td>1017.9</td>
<td>-1.83</td>
<td>-1.15</td>
<td>-2.34</td>
<td>-2.53</td>
<td>-3.34</td>
<td>-1.30</td>
</tr>
<tr>
<td>TPI1</td>
<td>Triose phosphate isomerase</td>
<td>137.6</td>
<td>0.23</td>
<td>1.38</td>
<td>-1.36</td>
<td>-0.87</td>
<td>-0.03</td>
<td>-0.25</td>
</tr>
<tr>
<td>TRK1</td>
<td>Transketolase</td>
<td>1337.1</td>
<td>-1.57</td>
<td>0.03</td>
<td>-1.64</td>
<td>-1.67</td>
<td>-1.53</td>
<td>-0.82</td>
</tr>
<tr>
<td>TAL2</td>
<td>Transaldolase</td>
<td>27.0</td>
<td>up</td>
<td>1.03</td>
<td>-1.42</td>
<td>-3.85</td>
<td>-4.57</td>
<td>-2.70</td>
</tr>
</tbody>
</table>

Calvin cycle genes: N-starvation effects in our study in comparison with previous studies (Miller et al, 2010; Blaby et al. 2013). The “FPKM 21gr” column indicates the gene’s highest FPKM value in the 21gr strain for all treatments and time points, which represents the gene’s relative expression level under induced conditions. Two groups of tests and comparisons were included: 48 h “-N” vs “+N” is the comparison between N-starved and N-replete condition for 48 h samples only; and "-N" 48 vs 0 h is a comparison between 48 h samples and 0 h samples in N-starved conditions only. Log2FC values were calculated from the average FPKM in our study and the average RPKM in other studies: red or green color highlight the log2FC value is lower than -1 or higher than 1, which means they are down-regulated or up-regulated in the 48 h N-starved condition. The underlined bold formatting like “red” or “green” means the change is statistically significant (FDR <0.01). And “down” or “up” means in one of the conditions no counts were captured in the RNA sequencing data.
Table 4:

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>48 hour &quot;-N&quot; vs &quot;+N&quot; log2FC</th>
<th>&quot;-N&quot; 48 vs 0 hour log2FC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FPKM</td>
<td>Fang 137c 21gr dw15.1</td>
<td>Fang 137c 21gr cc-4349</td>
</tr>
<tr>
<td>ACK1</td>
<td>Acetate kinase</td>
<td>19.1  -0.15  0.02  -0.88</td>
<td>-0.14  0.44  -0.82</td>
</tr>
<tr>
<td>ACK2</td>
<td>Acetate kinase</td>
<td>30.9  -0.06  0.23  -0.09</td>
<td>-0.16  -0.15  0.53</td>
</tr>
<tr>
<td>ACS1</td>
<td>Acetyl CoA synthetase</td>
<td>18.5  -0.29  -0.13  0.85</td>
<td>-0.36  -0.80  -0.07</td>
</tr>
<tr>
<td>ACS2</td>
<td>Acetyl CoA synthetase</td>
<td>78.1  -2.23  -1.85  -1.49</td>
<td>-0.57  -0.52  -0.73</td>
</tr>
<tr>
<td>ACS3</td>
<td>Acetyl CoA synthetase</td>
<td>46.8  -0.65  0.31  -3.06</td>
<td>-0.85  -0.92  -1.49</td>
</tr>
<tr>
<td>CIS2</td>
<td>Citrate synthase</td>
<td>66.5  1.34  1.00  -3.26</td>
<td>0.94   0.84  -1.25</td>
</tr>
<tr>
<td>ICL1</td>
<td>Isocitrate lyase</td>
<td>39.8  -0.36  -1.40  -5.81</td>
<td>0.45   -1.51 -3.74</td>
</tr>
<tr>
<td>MAS1</td>
<td>Malate synthase</td>
<td>17.7  -1.24  -2.24  -4.30</td>
<td>-0.54  -1.99 -3.77</td>
</tr>
<tr>
<td>MDH2</td>
<td>Malate dehydrogenase</td>
<td>35.4  -0.01  -1.49  -0.61</td>
<td>-0.21  -0.88 -0.92</td>
</tr>
<tr>
<td>PAT1</td>
<td>Phosphate acetyltransferase</td>
<td>19.0  -0.63  0.17  -0.26</td>
<td>0.22   0.35  -0.37</td>
</tr>
<tr>
<td>ACH1</td>
<td>Aconitate hydratase</td>
<td>383.0  1.31  1.44  -1.45</td>
<td>0.36   0.88  -0.11</td>
</tr>
</tbody>
</table>

Glyoxylate cycle genes: N-starvation effects with autotrophic cells compared with previous studies of mixotrophic cells (Miller et al, 2010; Blaby et al. 2013). The “FPKM 21gr” column indicates the gene’s highest FPKM value in the 21gr strain for all treatments and time points, which represents the gene’s relative expression level under induced conditions. Two groups of comparisons are included: 48 h "-N" vs "+N" indicates the comparison between N-starved and N-replete conditions for 48 h samples only; and "-N" 48 vs 0 h is a comparison between 48 h samples and 0 h samples in N-starved conditions only. The log2FC values were calculated from the average FPKM in our study and the average RPKM in other studies: red or green colors highlight log2FC values < -1 or > 1, which means they are down-regulated or up-regulated in the 48 h N-starved condition. The underlined bold formatting, like “red” or “green”, indicates the change is statistically significant.
Table 5: Gluconeogenesis and starch synthesis genes: N-starvation effects in autotrophic cells compared with previous studies with mixotrophic cells (Miller et al, 2010; Blaby et al. 2013). The “FPKM 21gr” column reflects the gene’s highest FPKM value in the 21gr

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>21gr</th>
<th>137c</th>
<th>21gr</th>
<th>dw15.1</th>
<th>137c</th>
<th>21gr</th>
<th>cc-4349</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gluconeogenesis genes</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SDH2</td>
<td>Iron-sulfur subunit of succinate dehydrogenase</td>
<td>139.3</td>
<td>0.19</td>
<td>0.37</td>
<td>-0.30</td>
<td>0.53</td>
<td>0.90</td>
<td>0.03</td>
</tr>
<tr>
<td>SDH3</td>
<td>Succinate dehydrogenase subunit b560</td>
<td>136.0</td>
<td>0.53</td>
<td>0.32</td>
<td>-0.37</td>
<td>0.33</td>
<td>0.40</td>
<td>0.11</td>
</tr>
<tr>
<td>FUM1</td>
<td>Fumarate hydratase</td>
<td>33.4</td>
<td>-0.28</td>
<td>-0.33</td>
<td>-0.84</td>
<td>0.01</td>
<td>-0.16</td>
<td>0.75</td>
</tr>
<tr>
<td>MDH1</td>
<td>Malate dehydrogenase, chloroplastic</td>
<td>170.6</td>
<td>0.70</td>
<td>0.34</td>
<td>-1.39</td>
<td>-0.98</td>
<td>-1.77</td>
<td>-1.45</td>
</tr>
<tr>
<td>MDH4</td>
<td>Malate dehydrogenase, mitochondrial</td>
<td>525.9</td>
<td>0.46</td>
<td>0.13</td>
<td>-0.11</td>
<td>-0.72</td>
<td>-1.32</td>
<td>0.17</td>
</tr>
<tr>
<td>MDH5</td>
<td>Malate dehydrogenase, chloroplastic</td>
<td>146.1</td>
<td>-0.04</td>
<td>0.43</td>
<td>0.34</td>
<td>0.12</td>
<td>-0.48</td>
<td></td>
</tr>
<tr>
<td>FBP1</td>
<td>Fructose-1,6-bisphosphatase</td>
<td>478.0</td>
<td>-0.96</td>
<td>-0.45</td>
<td>-1.07</td>
<td>-1.11</td>
<td>-1.80</td>
<td>-0.54</td>
</tr>
<tr>
<td>FBA1</td>
<td>Fructose-1,6-bisphosphate aldolase</td>
<td>384.1</td>
<td>-4.05</td>
<td>-4.06</td>
<td>-3.29</td>
<td>-0.74</td>
<td>-2.44</td>
<td>-2.01</td>
</tr>
<tr>
<td>FBA3</td>
<td>Fructose-1,6-bisphosphate aldolase</td>
<td>1736.3</td>
<td>-0.77</td>
<td>1.46</td>
<td>-3.04</td>
<td>-0.96</td>
<td>-0.76</td>
<td>-0.59</td>
</tr>
<tr>
<td>GAP1</td>
<td>Glyceraldehyde-3-phosphate dehydrogenase</td>
<td>404.8</td>
<td>-0.96</td>
<td>-0.45</td>
<td>-1.07</td>
<td>-1.11</td>
<td>-1.80</td>
<td>-0.54</td>
</tr>
<tr>
<td>GAP3</td>
<td>Glyceraldehyde-3-phosphate dehydrogenase</td>
<td>2773.9</td>
<td>-0.72</td>
<td>-0.24</td>
<td>-1.06</td>
<td>-1.72</td>
<td>-0.53</td>
<td></td>
</tr>
<tr>
<td>TPI</td>
<td>Triose phosphate isomerase</td>
<td>137.6</td>
<td>0.23</td>
<td>1.38</td>
<td>-1.36</td>
<td>-0.96</td>
<td>-0.76</td>
<td>-0.59</td>
</tr>
<tr>
<td>PCK1</td>
<td>Phosphoenolpyruvate carboxykinase</td>
<td>244.8</td>
<td>-2.77</td>
<td>-1.46</td>
<td>-2.60</td>
<td>-0.81</td>
<td>-0.02</td>
<td>-2.63</td>
</tr>
<tr>
<td>PGH1</td>
<td>Enolase</td>
<td>469.8</td>
<td>0.26</td>
<td>0.21</td>
<td>-0.28</td>
<td>-0.47</td>
<td>-0.80</td>
<td>-0.20</td>
</tr>
<tr>
<td>PGM1</td>
<td>Phosphoglycerate mutase</td>
<td>119.4</td>
<td>0.32</td>
<td>1.14</td>
<td>0.32</td>
<td>0.27</td>
<td>0.52</td>
<td>1.23</td>
</tr>
<tr>
<td>PGM2</td>
<td>Phosphoglycerate mutase</td>
<td>8.5</td>
<td>1.19</td>
<td>-0.66</td>
<td>1.49</td>
<td>-0.38</td>
<td>-1.48</td>
<td>1.58</td>
</tr>
<tr>
<td>PGM3</td>
<td>Phosphoglycerate mutase</td>
<td>13.9</td>
<td>-0.64</td>
<td>0.56</td>
<td>0.02</td>
<td>-1.13</td>
<td>-0.68</td>
<td>2.58</td>
</tr>
<tr>
<td>PGM4</td>
<td>phosphoglycerate mutase</td>
<td>8.1</td>
<td>0.01</td>
<td>0.74</td>
<td>0.56</td>
<td>-0.34</td>
<td>0.27</td>
<td>0.91</td>
</tr>
<tr>
<td>PGM5</td>
<td>phosphoglycerate mutase</td>
<td>2.7</td>
<td>1.64</td>
<td>2.12</td>
<td>-1.57</td>
<td>0.55</td>
<td>0.13</td>
<td>up</td>
</tr>
<tr>
<td>PGI1</td>
<td>Phosphoglucomutase</td>
<td>322.3</td>
<td>-0.40</td>
<td>0.74</td>
<td>3.62</td>
<td>0.09</td>
<td>0.81</td>
<td>-0.10</td>
</tr>
<tr>
<td>PPD2</td>
<td>Pyruvate phosphokinase dikinase</td>
<td>15.6</td>
<td>0.01</td>
<td>0.51</td>
<td>-0.25</td>
<td>-0.09</td>
<td>-0.10</td>
<td></td>
</tr>
<tr>
<td><strong>Starch synthesis genes</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>STA1/AGP1</td>
<td>ADP-glucose pyrophosphorylase large subunit 1</td>
<td>42.9</td>
<td>-0.11</td>
<td>0.11</td>
<td>-0.48</td>
<td>-0.89</td>
<td>-1.67</td>
<td>-0.19</td>
</tr>
<tr>
<td>AGP2</td>
<td>ADP-glucose pyrophosphorylase large subunit 2</td>
<td>250.5</td>
<td>-0.72</td>
<td>-0.24</td>
<td>-1.06</td>
<td>-1.72</td>
<td>-0.53</td>
<td></td>
</tr>
<tr>
<td>STAG/AGP4</td>
<td>ADP-glucose pyrophosphorylase small subunit</td>
<td>12.0</td>
<td>-0.21</td>
<td>-0.44</td>
<td>1.47</td>
<td>1.11</td>
<td>0.76</td>
<td></td>
</tr>
<tr>
<td>STA2/GBS1</td>
<td>Granule-bound starch synthase I</td>
<td>318.6</td>
<td>-3.90</td>
<td>-2.71</td>
<td>-1.39</td>
<td>-0.96</td>
<td>-2.60</td>
<td></td>
</tr>
<tr>
<td>GBS2</td>
<td>Granule bound starch synthase 2</td>
<td>1.8</td>
<td>0.05</td>
<td>0.79</td>
<td>-0.29</td>
<td>-0.20</td>
<td>-0.17</td>
<td></td>
</tr>
<tr>
<td>GPM1</td>
<td>Phosphoglucomutase</td>
<td>159.7</td>
<td>0.55</td>
<td>0.84</td>
<td>-1.09</td>
<td>-1.22</td>
<td>-1.30</td>
<td>-0.16</td>
</tr>
<tr>
<td>PGI1</td>
<td>Phosphoglucomutase</td>
<td>322.3</td>
<td>-0.40</td>
<td>0.74</td>
<td>3.62</td>
<td>0.09</td>
<td>0.81</td>
<td>-0.10</td>
</tr>
<tr>
<td>SSS1</td>
<td>soluble starch synthase</td>
<td>4.0</td>
<td>0.73</td>
<td>1.76</td>
<td>-0.17</td>
<td>-2.05</td>
<td>-2.19</td>
<td></td>
</tr>
<tr>
<td>SSS2</td>
<td>soluble starch synthase</td>
<td>147.7</td>
<td>-2.09</td>
<td>-2.46</td>
<td>-1.53</td>
<td>-1.83</td>
<td>-2.49</td>
<td></td>
</tr>
<tr>
<td>SSS4</td>
<td>soluble starch synthase</td>
<td>50.8</td>
<td>1.18</td>
<td>-0.18</td>
<td>0.39</td>
<td>0.25</td>
<td>0.28</td>
<td>0.91</td>
</tr>
<tr>
<td>SSS5</td>
<td>soluble starch synthase</td>
<td>13.6</td>
<td>0.51</td>
<td>1.14</td>
<td>2.56</td>
<td>1.78</td>
<td>2.71</td>
<td></td>
</tr>
<tr>
<td>STA3/SSS3</td>
<td>soluble starch synthase</td>
<td>39.4</td>
<td>-0.74</td>
<td>-0.57</td>
<td>0.00</td>
<td>-0.15</td>
<td>0.02</td>
<td>-0.42</td>
</tr>
<tr>
<td>SBE1</td>
<td>Starch Branching Enzyme</td>
<td>18.1</td>
<td>-0.07</td>
<td>-0.75</td>
<td>0.48</td>
<td>0.69</td>
<td>0.55</td>
<td>1.95</td>
</tr>
<tr>
<td>SBE2</td>
<td>Starch Branching Enzyme</td>
<td>10.8</td>
<td>-0.11</td>
<td>0.00</td>
<td>-0.71</td>
<td>-0.97</td>
<td>-1.29</td>
<td>0.26</td>
</tr>
<tr>
<td>SBE3</td>
<td>Starch Branching Enzyme</td>
<td>227.0</td>
<td>-0.65</td>
<td>-0.03</td>
<td>0.78</td>
<td>0.34</td>
<td>0.74</td>
<td>1.18</td>
</tr>
</tbody>
</table>

Gluconeogenesis and starch synthesis genes: N-starvation effects in autotrophic cells compared with previous studies with mixotrophic cells (Miller et al, 2010; Blaby et al. 2013). The “FPKM 21gr” column reflects the gene’s highest FPKM value in the 21gr.
strain for all treatments and time points, which represents the gene’s relative expression level under induced conditions. Two groups of comparisons were included: 48 h “-N” vs “+N” is a comparison between N-starved and N-replete conditions for 48 h samples only for our results compared with those of Miller et al. (2010); and "-N" 48 vs 0 h compares 48 h samples and 0 h samples in N-replete conditions only, comparing our results with those of Blaby et al. (2013). Log2FC values were calculated from the average FPKM in our study and the average RPKM in other studies: red or green colors indicate a log2FC value < -1 or > 1, which means they are down-regulated or up-regulated in the 48 h N-starved condition. The underlined bold formatting, like “red” or “green” means the change is statistically significant. The gluconeogenesis gene table also included the genes involved in succinate-malate conversions in mitochondrion and malate-oxaloacetate conversions in both mitochondrion and chloroplasts. The subcellular compartment locations of FBA1 and GAP1 are ambiguous.

Table 6:

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>FPKM 21gr</th>
<th>Log2FC in &quot;-N&quot; at 48 hours, starchless vs WT</th>
</tr>
</thead>
<tbody>
<tr>
<td>STA11</td>
<td>4-alpha-glucanotransferase</td>
<td>42.9</td>
<td>sta1 vs 137c 1.92 0.42 -0.08</td>
</tr>
<tr>
<td>STA1/AGP1</td>
<td>ADP-glucose pyrophosphorylase large subunit</td>
<td>250.5</td>
<td>21st1 vs 21gr -7.81 -7.11 0.16</td>
</tr>
<tr>
<td>STA6/AGP4</td>
<td>ADP-glucose pyrophosphorylase small subunit</td>
<td>395.6</td>
<td>Blaby</td>
</tr>
<tr>
<td>STA2/GBS1</td>
<td>granule-bound starch synthase I</td>
<td>318.6</td>
<td></td>
</tr>
<tr>
<td>GPM1</td>
<td>Phosphoglucomutase</td>
<td>159.7</td>
<td></td>
</tr>
<tr>
<td>PGI1</td>
<td>Phosphoglucomutase</td>
<td>322.3</td>
<td></td>
</tr>
<tr>
<td>SSS4</td>
<td>soluble starch synthase</td>
<td>50.8</td>
<td></td>
</tr>
<tr>
<td>STA3/SSS3</td>
<td>soluble starch synthase</td>
<td>39.4</td>
<td></td>
</tr>
<tr>
<td>SBE1</td>
<td>Starch Branching Enzyme</td>
<td>18.1</td>
<td></td>
</tr>
<tr>
<td>SBE2</td>
<td>Starch Branching Enzyme</td>
<td>10.8</td>
<td></td>
</tr>
<tr>
<td>SBE3</td>
<td>Starch Branching Enzyme</td>
<td>227.0</td>
<td></td>
</tr>
</tbody>
</table>

Starch synthesis starchless compared with wild type. Effect of starchless mutations in N-starved cultures, comparing our results with sta1 autotrophic cultures with sta6 mixotrophic cultures (Blaby et al. 2013). The “FPKM 21gr” column reflects the gene’s highest FPKM value in the 21gr strain for all treatments and time points, which represents the gene’s relative expression level under induced conditions. Two comparisons are included: sta1 vs 137c or 21st1 vs 21gr. Log2FC values compared expression levels of starchless strains with corresponding wild type strains: red or green colors highlights log2FC values < -1 or >1, which means they are down-regulated or up-regulated in 48 h N-starved starchless strains. The underlined bold formatting, like “red” or “green” means the change is statistically significant.
Table 7:
Average dry weight 48 hours after switching medium (g/L of culture)

<table>
<thead>
<tr>
<th>strains</th>
<th>N-starved</th>
<th>standard error</th>
<th>N-replete</th>
<th>standard error</th>
</tr>
</thead>
<tbody>
<tr>
<td>21gr</td>
<td>4.12</td>
<td>0.58</td>
<td>5.27</td>
<td>0.05</td>
</tr>
<tr>
<td>137c</td>
<td>2.49</td>
<td>0.05</td>
<td>3.62</td>
<td>0.20</td>
</tr>
<tr>
<td>sta1</td>
<td>2.58</td>
<td>0.18</td>
<td>4.93</td>
<td>0.08</td>
</tr>
<tr>
<td>21st1</td>
<td>2.77</td>
<td>0.08</td>
<td>5.24</td>
<td>0.23</td>
</tr>
</tbody>
</table>

Dry weights measured 48 h after switching media. A unit volume of liquid culture was collected and dried at 65 °C for 2 days. For all 4 strains, dry biomass was considerably less under the N-starved condition.

Table 8:

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>48 hour &quot;-N&quot; vs &quot;+N&quot; log2FC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FPKM Fang</td>
<td>21gr 137c 21gr dw15.1</td>
</tr>
<tr>
<td>ECH1</td>
<td>Enoyl-CoA hydratase/isomerase</td>
<td>103.6 0.85 0.46 1.12</td>
</tr>
<tr>
<td>HCD1</td>
<td>Enoyl-CoA hydratase/isomerase</td>
<td>67.9 -1.04 -1.01 -1.40</td>
</tr>
<tr>
<td>ATO1</td>
<td>Acetyl-CoA acyltransferase</td>
<td>67.6 -0.34 -1.49 -1.86</td>
</tr>
<tr>
<td>ACO2</td>
<td>acyl-CoA oxidase</td>
<td>7.5 -0.98 -1.54 -1.93</td>
</tr>
<tr>
<td>ACO1</td>
<td>acyl-CoA oxidase</td>
<td>9.3 -0.27 -0.07 -1.11</td>
</tr>
</tbody>
</table>

Beta oxidation genes. N-starvation effect on genes of the beta-oxidation pathway in autotrophic cells compared with those of mixotrophic cells Miller et al (2010). The “FPKM 21gr” column indicates the gene’s highest FPKM value in the 21gr strain for all treatments and time points, which represents the gene’s relative expression level under induced conditions. The test 48 h "-N" vs "+N" compares N-starved and N-replete conditions for 48 h samples only, comparing data from autotrophic cells with those from mixotrophic cells Miller et al (2010). The log2FC values were calculated from the average FPKM in our study and the average RPKM in Miller’s study: red or green colors highlight the log2FC values < -1 or > 1, which means they are down-regulated or up-regulated in the 48 h N-starved condition. The underlined bold formatting, like “red” or “green” means the change is statistically significant.
References


CHAPTER 5

GENERAL SUMMARY

Through the history, advancements in technology always push forward the basic scientific research in a revolutionary way. The high through-put nucleic acid sequencing technology developed in recent years has created numerous new methods for conducting biological researches. Genomic sequencing has become affordable for individual mutants, so the identification of the “Single Nucleotide Polymorphisms/Insertions and Deletions” type mutations and comparative genomics studies can be achieved by relatively small scale efforts; the entire transcriptome can be quantified by sequencing and counting the actual mRNA fragments. These newly developed applications have drawn immediate attentions of life science researchers, especially for those who are working on organisms with limited prior knowledge about their genome and transcriptome. Nowadays, genome sequencings take weeks instead of years; transcriptomes can be quantified even without a reference or high quality microarray chips. The ongoing studies of Chlamydomonas reinhardtii have already been and surely will keep taking the advantage of the second generation sequencing technologies, just like projects presented in this thesis.

The transcriptome study about Chlamydomonas CO₂ concentration mechanism (CCM) provided a general picture of how gene’s transcript levels are affected by the limiting CO₂ condition, and how these regulations are interfered by the mutation of the CCM master regulator CIA5. The clustering analysis revealed groups of genes that fit into
transcript changing patterns that are commonly shared by previously studied CCM genes, which is the up regulations or high inductions under the limiting CO₂ conditions for the wild type strain but not in the CIA5 mutant strain. Genes in these clusters can be considered as future candidate genes involved in the CCM system. We also observed a massive impact of the CIA5 mutation alone: many genes expression patterns were affected in the CIA5 mutant but did show statistically significant responses to the varied CO₂ conditions. Interestingly, the transcriptome variations between Low and Very Low CO₂ acclimated cultures were not significant at the 4 hour sampling time point, which left doubts about genes involved in the multiple acclimation states.

In order to further investigate the multiple CO₂ level acclimation states of Chlamydomonas, a follow up study was carried to identify a pair of putatively allelic mutations su4 and su5, which are causing suppressions of the LCIB mutants “air dier” phenotype. Also by taking the advantage of the high through put sequencing technology, we generated a candidate table of the locations of the su4 and su5 mutations, which can be combined with future genetic mapping studies to reveal these suppressor genes. In addition to the effort of identifying the unknown su4 and su5 genes, we also monitored the transcript expression levels of 5 previously identified CCM genes LCIA, LCIB, LCI1, CAH1, and HLA3. We carefully quantified their RNA transcript levels by quantitative PCR in detailed 12 hour time course, and made a detailed portrait of their transcript changing patterns at varied CO₂ concentration gradients spanning from Low to Very Low CO₂ conditions. We observed moderate variations in these genes transcript changing pattern: the inductions of HLA3 gene were slightly later than the
other 4 tested genes; and the \textit{LCIB} gene reached its maximum expression more rapidly. Detailed molecular level mechanisms causing the different acclimations between Low and Very Low CO\textsubscript{2} conditions are still to be investigated in the future.

Due to the increasing fossil fuel cost and the environmental issues caused by excessive CO\textsubscript{2} emission in the modern industrial age, the potential of advanced biofuel production has drawn attentions. As a fast growing green algae species and a model organism, \textit{Chlamydomonas} is capable of accumulating TAG lipid internally during nitrogen starvations. The transcriptomes of 2 wild type strains 137c and 21gr, and 2 starchless mutant strains (have mutations on the ADP-glucose pyrophosphorylase regulatory subunit) \textit{sta1} and \textit{21st1} were examined to identify the metabolic genes that are affected by nitrogen starvation and the mutations on the ADP-glucose pyrophosphorylase. In contrast to the previous transcriptome studies about nitrogen starvation or starchless mutants that were conducted in the mixotrophic conditions with externally supplied acetate, we completed the experiment in photoautotrophic conditions without acetate, because recycling the atmospheric CO\textsubscript{2} is one of the major purposes for biofuel industry. Comparing to the mixotrophic experiments, we observed similar up regulations of nitrogen assimilation genes and down regulations of light reaction genes. Few glyoxylate cycle genes \textit{ACS2/3, MAS1}, and \textit{ICL1} are up regulated even in the starchless \textit{sta1} mutant under limiting nitrogen condition, and this observation is comparable between the autotrophic and the previously studied \textit{sta6} mixotrophic cultures. On the other hand, some \textit{de novo} fatty acid genes expressions were decreased in the wild type strains we tested comparing to the mixotrophic
nitrogen starved cultures, but the transcript levels of Calvin cycle genes seemed to be better maintained in our nitrogen starved photoautotrophic cultures. This study provided information about carbon flow could be varied in the photoautotrophically grown *Chlamydomonas* cells under nitrogen starvations. The future research should include the evaluations of metabolites levels as well.