Activation of the Carbon Concentrating Mechanism by CO2 Deprivation Coincides with Massive Transcriptional Restructuring in *Chlamydomonas reinhardtii*

Andrew J. Brueggeman  
*University of Nebraska-Lincoln*

Dayananda S. Gangadharaiah  
*University of Nebraska-Lincoln*

Matyas F. Cserhati  
*University of Nebraska-Lincoln*, mcserhati2@unl.edu

David Casero  
*University of California, Los Angeles*

Donald P. Weeks  
*University of Nebraska–Lincoln*, dweeks1@unl.edu

See next page for additional authors

Follow this and additional works at: [http://digitalcommons.unl.edu/biochemfacpub](http://digitalcommons.unl.edu/biochemfacpub)

Part of the [Biochemistry, Biophysics, and Structural Biology Commons](http://digitalcommons.unl.edu/biochemfacpub)

---

*Brueggeman, Andrew J.; Gangadharaiah, Dayananda S.; Cserhati, Matyas F.; Casero, David; Weeks, Donald P.; and Ladunga, Istvan, "Activation of the Carbon Concentrating Mechanism by CO2 Deprivation Coincides with Massive Transcriptional Restructuring in *Chlamydomonas reinhardtii*" (2012). Biochemistry -- Faculty Publications. Paper 106. [http://digitalcommons.unl.edu/biochemfacpub/106](http://digitalcommons.unl.edu/biochemfacpub/106)*

This Article is brought to you for free and open access by the Biochemistry, Department of at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in Biochemistry -- Faculty Publications by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.
Activation of the Carbon Concentrating Mechanism by CO₂ Deprivation Coincides with Massive Transcriptional Restructuring in Chlamydomonas reinhardtii

Andrew J. Brueggeman,a Dayananda S. Gangadharaiah,a Matyas F. Cserhati,b David Casero,c,d Donald P. Weeks,a and Istvan Ladunga\textsuperscript{a,\textsuperscript{1}}

\textsuperscript{a}Department of Biochemistry, University of Nebraska, Lincoln, Nebraska 68588-0665
\textsuperscript{b}Department of Statistics, University of Nebraska, Lincoln, Nebraska 68588-0665
\textsuperscript{c}Department of Molecular, Cell, and Developmental Biology, University of California, Los Angeles, California 90095
\textsuperscript{d}Institute of Genomics and Proteomics, University of California, Los Angeles, California 90095

A CO₂-concentrating mechanism (CCM) is essential for the growth of most eukaryotic algae under ambient (392 ppm) and very low (<100 ppm) CO₂ concentrations. In this study, we used replicated deep mRNA sequencing and regulatory network reconstruction to capture a remarkable scope of changes in gene expression that occurs when Chlamydomonas reinhardtii cells are shifted from high to very low levels of CO₂ (≤100 ppm). CCM induction 30 to 180 min post-CO₂ deprivation coincides with statistically significant changes in the expression of an astonishing 38% (5884) of the 15,501 nonoverlapping C. reinhardtii genes. Of these genes, 1088 genes were induced and 3828 genes were downregulated by a log\(_2\) factor of 2. The latter indicate a global reduction in photosynthesis, protein synthesis, and energy-related biochemical pathways. The magnitude of transcriptional rearrangement and its major patterns are robust as analyzed by three different statistical methods. De novo DNA motif discovery revealed new putative binding sites for Myeloid oncogene family transcription factors potentially involved in activating low CO₂-induced genes. The \((CA)_n\) repeat (9 ≤ n ≤ 29) is present in 29% of upregulated genes but almost absent from promoters of downregulated genes. These discoveries open many avenues for new research.

INTRODUCTION

In nature, Chlamydomonas reinhardtii and other eukaryotic algae depend on a CO₂-concentrating mechanism (CCM) to supply sufficient inorganic carbon (C\(_i\); CO₂ or bicarbonate) for photosynthesis-fueled cell growth and proliferation. Mutant cells lacking key components of the CCM molecular machinery or its regulatory system (Moroney and Ynalvez, 2007; Duanmu et al., 2009; Yamano and Fukuzawa, 2009) do not grow or grow poorly unless supplied with high concentrations of CO₂ (e.g., >10,000 ppm) that are well above the ambient level of ~392 ppm. Because the diffusion rate of CO₂ in aqueous environments is ~10,000 times slower than in air, most natural populations of microalgae exist in CO₂-limited conditions. This is especially true for dense algal populations growing under abundant sunlight. Under such conditions, CO₂ concentrations can become very low (<100 ppm) and cells induce the CCM to maximal levels. CO₂ starvation induces the transcription of numerous genes encoding proteins closely associated with the CCM and its activities (Moroney and Ynalvez, 2007; Duanmu et al., 2009; Yamano and Fukuzawa, 2009). Indeed, C. reinhardtii and most other eukaryotic algae have developed a finely tuned regulatory system that suppresses expression of CCM-related genes under conditions of replete CO₂ (i.e., >0.1% CO₂) and activates expression of these genes when CO₂ becomes limiting (Moroney and Ynalvez, 2007; Duanmu et al., 2009; Yamano and Fukuzawa, 2009; also reviewed in the companion article Fang et al., 2012).

Previous studies using RNA gel blot analyses and microarray analyses have revealed a number of CCM-associated genes and other CO₂-responsive genes whose transcription is tied to the physiological changes that accompany cell acclimation to CO₂ stress conditions (Im and Grossman, 2002; Im et al., 2003; Miura et al., 2004; Wang et al., 2005; Yamano et al., 2008; Yamano and Fukuzawa, 2009).

Here, we report an extensive global analysis of the massive transcriptional changes evoked by the deprivation of C\(_i\) in C. reinhardtii. We measured these transcriptional events using replicated deep RNA sequencing (RNA-Seq) on the Illumina platform. The highly reproducible RNA-Seq experiments not only confirm earlier observations based on array analyses quoted above but also extend the list of differentially expressed genes from a few hundred to over 4000.

We report the discovery of an extensive system of head-to-head (HTH; also called divergent) gene pairs, many of them sharing bidirectional or connected promoters. HTH conformation and bidirectional or shared promoters frequently perform the highly accurate coregulation of gene pairs encoding subunits of the same protein complex or two proteins of similar or related functions. Here, we focus on those HTH, coregulated, gene pairs that are most relevant to the CCM. Advanced
RESULTS AND DISCUSSION

Overview

C\textsubscript{i} deprivation is a major stress that evokes a dramatic transcriptional response in algae. Using EST-based macroarrays, the Fukuzawa laboratory (Miura et al., 2004; Kohinata et al., 2008; Yamano et al., 2008) and Grossman and Weeks laboratories (Wang et al., 2005) pioneered the transcriptional profiling of C\textsubscript{i} deprivation. In initiating our studies, our hypothesis was that revolutionary progress in sequencing technology and statistical methodology would allow us to discover a large number of activated or repressed biological processes and individual genes that may have escaped detection using EST arrays. To test this hypothesis, we performed deep RNA-Seq using the Illumina Genome Analyzer II platform at the Joint Genome Institute (JGI) of the Department of Energy. In total, the eight samples collected at four time points (0, 30, 60, and 180 min after C\textsubscript{i} deprivation) produced 98.3 million uniquely mapped sequencing reads (12.3 million 71-base-long reads per sample; see Methods). When no more than two mismatches were allowed in the anchor regions, ~38% of the reads did not map uniquely or contained more than two base errors due to sequencing errors, genomic variability, alternative splicing (Labadorf et al., 2010), a number of recently duplicated genes (Villand et al., 1997), and repetitive DNA elements (Merchant et al., 2007). Even with this conservative approach, RNA-Seq represents a major advance from micro- and macroarrays: It provides an unprecedented high coverage of transcripts, eliminates cross-hybridization effects, does not rely on the commercial availability of arrays, and is more robust against errors in predicted exon structures (Margulies et al., 2005). The significantly increased performance of RNA-Seq has been shown specifically for \textit{C. reinhardtii} (González-Ballester et al., 2010).

The high technological reproducibility of the RNA-Seq measurements performed at the Department of Energy’s JGI is shown by the strong correlations of transcript levels between biological replicates (0.958, 0.965, 0.939, and 0.973 for 0, 30, 60, and 180 min time points after carbon deprivation, respectively). These high Pearson correlation coefficients indicate reproducible and multiplicative (linear) biases and that the nonlinear bias is miniscule. Note that linear, multiplicative, and reproducible bias does not alter fold change values by multiplying the transcript levels both in the numerator and the denominator. Such biases include sequencing reads that match imperfectly to the genome or the transcriptome (Li et al., 2010), amplification, and sequencing biases (Dohm et al., 2008). Additive effects, such as unusual exons, may reduce the extent of differential expression. These effects are due to imperfect gene models, such as those predicted by the Augustus method (Stanke and Waack, 2003) and alternative splicing (Labadorf et al., 2010). Such additive effects remain our primary concern. Recently duplicated genes pose further challenges in mapping the 71-base-long sequencing reads to the transcriptome because these reads contain erroneous base calls, particularly at their 3' ends. Such gene pairs include major effectors of CO\textsubscript{2} concentration, such as four carbonic anhydrases (CAH\textsubscript{s}), CAH\textsubscript{1}-CAH\textsubscript{2} and CAH\textsubscript{4}-CAH\textsubscript{5}, that are recent duplicates. The pair CAH\textsubscript{4}-CAH\textsubscript{5}, for example, contains exons that are over 90% identical (Villand et al., 1997).

To avoid mappings to the duplicated genes, rigorous procedures (with one or two mismatches in the anchor regions that connect two exons) are necessary. However, this rigor also drastically reduces the coverage of all genes due to both sequencing errors and polymorphisms. Reduced coverage reduces the number of significantly differentially expressed genes. Therefore, we performed the mapping with both one and two allowed mismatches in the anchor region, as implemented in the tophat program (Trapnell et al., 2009). With one allowed mismatch, fewer but more accurate transcript levels were obtained than with two mismatches. For example, our data, as expected from earlier studies, demonstrated that CAH\textsubscript{1} is strongly upregulated at 3 h ASVL\textsubscript{CO\textsubscript{2}}. However, in our initial analyses allowing two mismatches, CAH\textsubscript{2}, which had earlier been reported not to respond to CO\textsubscript{2}, was falsely classified as upregulated. When reanalyzed using only one mismatch per read, the vast majority of reads in the 60- and 180-min time points were mapped to the CAH\textsubscript{1} gene, with few being attributable to CAH\textsubscript{2}. Quantitative RT-PCR (qRT-PCR) confirmed the results of the more rigorous alignments (see Supplemental Figure 1 online).

The lists of differentially expressed genes may be influenced by the choice of statistical methodology. Therefore, we analyzed our data using three different computational tools, edgeR, DESeq, and baySeq. Because differential expression of a large number (~16k) of genes is estimated using very few replicates (samples), statistical tools derived from large-sample asymptotic theory do not work. In particular, small sample size affects the correction for overdispersion (greater variability than expected based on Poissonian or other simple models), modeling the empirical distributions, and calculating statistical significance. To solve these issues, edgeR (Robinson et al., 2010) shrinks genewise dispersion estimates toward a constant value using an empirical Bayesian model and performs Fisher’s exact test. DESeq (Anders and Huber, 2010) uses nonparametric regression models to fit the negative binomial variance as a function of the mean, assuming a locally linear relationship between overdispersion and mean expression levels. baySeq (Hardcastle and Kelly, 2010) is free of this assumption and uses a fully empirical Bayesian approach to estimate the posterior probabilities. We compared the numbers of overlapping differentially
expressed genes reported by the edgeR, DESeq, and baySeq packages at 180 versus 0 min ASVLCO2 (see Supplemental Figure 2 online). Because the exact test implemented in edgeR calculates lower false discovery rate (FDR) q-values than the other two methods (Lopez et al., 2011), at FDR $\leq 0.01$, edgeR, DESeq, and baySeq reported 4222, 2364, and 3248 differentially expressed genes, respectively. The lists of differentially expressed genes are more consistent when the FDR threshold is elevated to 0.05, a still conservative level. All three methods reported differential expression for as many as 3141 genes (see Supplemental Figure 2 online). An additional 702 genes were jointly reported by both edgeR and baySeq, and a further 95 genes were called jointly by edgeR and DESeq. Because of the high overlaps with other methods, and its wider acceptance, below we limit our discussions to the results obtained by the edgeR tool.

Our results reproduced the observed induction of major CCM-associated genes published by Miura et al. (2004), Wang et al. (2005), Yamano et al. (2008), Yamano and Fukuzawa (2009), as well as in the companion article (Fang et al., 2012). In addition, we report a large number of genes that have not been associated with the CCM previously. Some of the notable similarities and differences in gene sets of our studies and those presented in our companion article (Fang et al., 2012) are discussed throughout this section (with special attention to the potential causes of observed differences provided near the end of this section).

Major Transcriptional Changes

Our results greatly extend many aspects of earlier observations of differentially expressed genes following CO2 deprivation. This is indicated by relatively similar lists of induced genes published previously (Miura et al., 2004; Wang et al., 2005; Yamano et al., 2008; Yamano and Fukuzawa, 2009) and by us (see Supplemental Data Set 1 online). In addition, RNA-Seq and modern statistical methodologies allowed us to discover an unexpectedly high 5864 genes that are differentially regulated at either 30, 60, or 180 min ASVLCO2 relative to the 0 min control (FDR $\leq 0.01$ and abs(log2(fold change)) $\geq 1$) or 3828 genes (FDR $\leq 0.001$ and abs(log2(fold change)) $\geq 2$).

Robust temporal expression patterns emerged under our conditions for imposition of CO2 deprivation. We found that the transcriptional response becomes widespread only after 30 min and increases (or decreases) for many, but not all, genes. The relatively slow onset of significant transcript changes is likely coupled to the relatively slow decline in CO2 concentrations employed in our experiments (Figure 1). At 30 min after deprivation, we found only 37 upregulated and five repressed genes relative to the 0 min control (FDR $\leq 0.001$ and abs(log2(fold change)) $\geq 2$; or in absolute, nonlogarithmic scale, a fourfold increase or decrease) (see Supplemental Data Sets 2 and 3 online). At an hour ASVLCO2, 409 genes are upregulated and 1663 genes are repressed (see Supplemental Data Sets 2 and 3 online). At 3 h ASVLCO2, 981 genes are induced and 1188 genes are repressed. These numbers are approximately doubled at the more typical thresholds (FDR q $\leq 0.01$ and abs(log2(fold change)) $\geq 1$; see Supplemental Data Sets 2 and 3 online). To measure transcript levels by a different method, we performed qRT-PCR analyses on the same RNA samples that were submitted for Illumina sequencing. Three different genes induced by CO2 deprivation (Low CO2 Induced A (LCIA), CAH5, and LCIB) displayed similar expression patterns between RNA-Seq and qRT-PCR, while a fourth gene, CAH2, previously reported as not responding or responding negatively to CO2 depletion (Moroney et al., 2011), showed moderate decreases in transcript levels using both RNA-Seq and RT-PCR measurements (see Supplemental Figure 1 online). Also to be noted (as described above) is the strong correlation of transcript levels in data obtained from biological replicates used for RNA-Seq analyses. Together, these observations confirm the high technological reproducibility of deep RNA-Seq as well as the reproducibility of our biological samples.

Gene Sets Affected by CO2 Deprivation

For brevity, unless otherwise noted, we limit our discussion in this section to transcript differences in cells maintained in high CO2 to transcripts in cells 3 h ASVLCO2. High-level overviews of the massive transcriptional response were obtained using GO categories (Ashburner et al., 2000). PLAZA, an online platform for plant comparative genomics (Proost et al., 2009), assigns GO categories to $\sim$7700 C. reinhardtii genes. GO categories, like pathways or genes colocalted within a chromosomal band, allow examination of transcriptional changes at the level of gene sets as opposed to individual genes. These gene sets may be enriched in upregulated or downregulated genes. To avoid the subjectivity of interpretations, the statistical significance of enrichment is calculated using gene set enrichment analyses (GSEAs; Subramanian et al., 2005; see Methods). GSEA is capable of handling high-throughput processing of large databases of gene sets, such as GO or the KEGG Database of Biochemical Pathways (Okuda et al., 2008).

The massive downregulation of several fundamental metabolic processes is highly statistically significant as shown by our GSEAs using GO annotations. We found significant repression of translation, ribosomal activities, RNA processing, intracellular protein transport, transport from the endoplasmic reticulum to the Golgi apparatus, nucleic acid binding, ATP synthesis, protein kinase activity, photosynthesis, oxidation reduction processes, protein folding, and unfolded protein binding, etc. (see Supplemental Data Set 4 online). Decreased metabolic activity may reflect the stress/survival mode of metabolism, which is necessary to cope with the stress of CO2 starvation. Upregulated GO categories are less abundant than downregulated ones, biased due to the limited annotations of the CCM- and other plant-specific biological processes in contrast with the basic metabolic processes present in most eukaryotes. Induced GO categories include nucleosome, carbon use, calcium ion transport, and proteolysis-related gene sets (see Supplemental Data Set 4 online). Unexpectedly, five flagella-related categories also are upregulated. GO analyses of several other biochemical pathways provided little additional information, possibly due to limited annotations.

To complement GO categories, we also studied other major functional and subcellular categories (Figure 2). One of these categories is a set of 595 plant-specific (greencut2) genes, which are conserved through diverse representatives of the
plant kingdom but have no known relatives in animals, fungi, or prokaryotes (Karpowicz et al., 2011). As many as 193 greencut2 genes were repressed, while only 30 were activated (FDR $q$-value < $10^{-256}$ as calculated by the permutation test implemented in GSEA; Figure 2). To examine whether carbon deprivation reduces photosynthetic activity, we analyzed the 393 genes that code for proteins localized in the chloroplast. Of these nuclear or chloroplast genes, 120 were downregulated and only 18 were induced (FDR $q$ < $10^{-256}$).

Using DNA arrays containing oligonucleotides derived from ESTs, the Fukuzawa laboratory (Miura et al., 2004; Yamano et al., 2008; Yamano and Fukuzawa, 2009) and the Grossman and Weeks laboratories (Wang et al., 2005) published several lists of genes differentially regulated by $C_1$ deprivation. In our much more extensive analyses, we reproduced the induction of 40 previously reported genes (38%; see Supplemental Data Set 1 online). $LCI1$ (Formate/nitrite transporter1.2 [NAR1.2]), a putative plasma membrane anion transporter (Ohnishi et al., 2010), is among the most dramatically induced genes (fold change: $2^{12} \approx 4000$). Fourteen previously designated low CO$_2$-induced genes decreased their expression, including chloroplast geranylgeranyl pyrophosphate synthase (geranylgeranyl diphosphate synthase (GGPS) or $LCI14$), $LCI21$, $LCI25$, and Light Harvesting Complex1 (LHCSR1) (at FDR ≤ 0.01 and log$_2$(fold change) ≤ −1).

Yamano and Fukuzawa (2009) hypothesized that thylakoid membrane proteins may be involved in the regulation of CCM. We successfully reproduced the earlier reported expression patterns of genes coding for low CO$_2$-inducible chloroplast envelope proteins, including oxygen-evolving proteins, plastid division protein, a number of carbonic anhydrases, photosystem II stability/
assembly factor photosystem II stability/assembly factor (HCF136), and peptidyl-prolyl cis-trans isomerase (see Supplemental Data Sets 2 and 3 online).

**A Shift of C. reinhardtii Cells from High to Very Low CO2 Triggers Transcription of Several Known Genes Encoding CCM-Associated Proteins and Several New CO2-Responsive Genes**

Examination of transcript abundance in this RNA-Seq study for the 106 genes identified in earlier analyses of induced gene transcription triggered by shifts from high to low CO2 conditions (see Supplemental Data Set 1 online) confirms that most are highly induced within an hour ASVLCO2. Among the most highly induced genes at 3 h are several well-recognized CCM-associated or CO2-responsive genes, including LCIA, encoding a putative bicarbonate transporter (4000-fold); LCI1, a gene regulated by the CO2-responsive LCR1 transcription factor (see below) and encoding a transmembrane protein possibly involved in Ci transport (3000-fold); CCP1, a C, Accumulating5 (CIA5)/CCM1-regulated gene encoding a chloroplast envelope protein (2000-fold); CAH1, encoding a periplasmic carbonic anhydrase (660-fold); LCIE (188-fold); CCP2, encoding a protein closely related to CCP1 (119-fold); LCI1D, encoding a protein highly similar to LCIB (see below) (89-fold); LCR1, encoding a Myeloid oncogene (Myb)-like transcription factor that regulates expression of CAH1 and LCI1 (52-fold); High Light Acclimation3 (HLA3), encoding a plasma membrane–localized bicarbonate transporter (40-fold); and LCIB, encoding a putative chloroplast CO2-scavenging protein (26-fold) (see Supplemental Data Sets 1 and 2 online). Using cluster analysis of gene expression, we found that expression patterns of these genes during the 3-h time course employed in this study fall primarily into the patterns displayed in clusters A to D (Figure 3). In these clusters, mRNA levels markedly increase by an hour ASVLCO2 and remain high at 3 h.

Strikingly, our deep RNA-Seq analyses increased the number of low CO2-induced (LCI) genes to 1875 (see Supplemental Data Set 2 online), well beyond the 106 previously recognized LCI genes listed in Supplemental Data Set 1 online. The annotation of C. reinhardtii genes has recently been improved (Castruita et al., 2011; Lopez et al., 2011). Still, the function of a number of proteins encoded by differentially regulated genes remains unknown, including 56 out of the 100 most highly induced genes (see Supplemental Data Set 2 online).

**A General, but Transient, Decrease in Transcription of Most Genes after a Shift to Very Low CO2 Conditions**

Past studies of genes whose transcription changes in response to decreased CO2 levels have focused primarily on those genes whose products are components of the CCM and whose transcription is steadily increased in the first few hours ASVLCO2. Expression patterns typical of such genes are shown in clusters A to D of Figure 3. However, 2399 genes are first downregulated but then induced by 3 h ASVLCO2, forming a distinctive check mark (√) signature of transcript levels. Accordingly, the groups
of genes in boxes F to I all display a significant decrease in transcript numbers at 30 and/or 60 min with a subsequent increase at 180 min. Indeed, if one examines the data of Supplemental Data Set 3 online that displays 2400 genes downregulated at 180 min ASVLCO2, nearly all of the last 1740 genes display a striking decrease in transcript levels at 60 min, but a marked recovery of transcript levels at 180 min (i.e., the check mark expression pattern; see Supplemental Figure 3 online). This observation suggests that, in general, there is a rapid decrease in gene transcription immediately ASVLCO2. Thus, our studies have revealed that *C. reinhardtii* cells not only rapidly sense a decrease in CO2 concentrations and respond, as expected, by upregulating CCM-associated genes, but also transition into stress/survival mode by downregulating thousands of genes not directly related to the CCM. This observation opens a new area of study related to transcriptional regulatory networks that orchestrate the rapid and appropriate responses of *C. reinhardtii* cells to changes in CO2 abundance.

**Transient Gene Induction following CO2 Reduction**

Antithetic to the check mark pattern of expression discussed above are patterns for 139 genes that are rapidly induced (by a factor of 2 or higher; FDR q ≤ 0.01) during the first hour ASVLCO2 but markedly decrease in transcript levels 2 h later. This caret (∧) expression pattern (see Supplemental Figure 4 online) is observed for the low CO2-induced bestrophin-like protein (516308) regulated by the transcriptional activator CIA5 and, interestingly, for 20 flagellar genes as well. Further experiments will determine if among these genes displaying this transient induction pattern are those encoding transient regulators that control acclimations to long-term restricted CO2 availability. To assess the functions of such transiently expressed genes, mutations or artificial alterations of gene expression (e.g., using RNA interference techniques or gene replacement techniques) will be needed.

**Coregulation of Functionally Related Gene Pairs in a HTH Conformation**

We discovered an extensive system of HTH (also called divergent) gene pairs, many of which are apparently regulated by bidirectional promoters. In general, when the distance between two start codons is small (in *C. reinhardtii*, <300 to 500 bp), coregulation is most likely due to a single bidirectional promoter that regulates transcription in both directions. Bidirectional
promoters often perform usually tight coregulation of the upstream Crick strand and the downstream Watson strand genes (Trinklein et al., 2004). Bidirectional promoters facilitate the stoichiometric production of subunits of a particular protein complex or members of a particular biochemical or signaling pathway. Such mechanisms have been described in the human pilot-ENCODE project (Birney et al., 2007), yeast (Li et al., 2011b), mouse (Li et al., 2011a), flies, and other organisms. In C. reinhardtii, however, publications are scant regarding the 8852 genes that evolved into the HTH conformation. The few reported coregulated HTH gene pairs include the Argona1 and Dicer-like1 genes (Casas-Mollano et al., 2008) and two mitochondrial, β-type carbonic anhydrases CAH4 and CAH5 (then termed ca1 and ca2) reported by Villand et al. (1997). CAH4 and CAH5, like CAH1 and CAH2, recently formed inverted repeats, share very high sequence identity even in introns and promoter regions, and are upregulated under CO2 starvation.

Strong transcriptional correlations within HTH gene pairs may indicate fundamental regulatory mechanisms in C. reinhardtii. Here, we focus on CCM-related HTH gene pairs. Coregulation, frequently by a postulated bidirectional promoter, in many of the 4276 HTH gene pairs in C. reinhardtii is indicated by two observations. First, the median correlation coefficient between transcript levels of the HTH gene pairs is as high as 0.674 compared with 0.522 in all other gene pairs (P < 10⁻²⁵⁶, Wilcoxon-Mann-Whitney test; Figure 4; see Supplemental Figure 5 online), and for 25% of HTH gene pairs, this correlation exceeds 0.917. For other conformations, the correlation in the top 25% of non-HTH gene pairs is 0.8650 (for the whole distributions, P < 10⁻²⁵⁶, Wilcoxon-Mann-Whitney test; see Supplemental Figure 5 and Supplemental Data Set 5 online). The high correlation of the top 25% even in the non-HTH genes indicates the existence of other mechanisms of coregulation, including locus control regions, microRNAs, and the binding of similar transcription factors (reviewed in Ladunga, 2010). The second indication for the bidirectional regulation is the short distance between the start codons of HTH gene pairs (median length: 831 ± 3228 bp so). By contrast, intergenic regions of tail-to-tail, head-to-tail, and tail-to-head gene pairs tend to be more than twice as long (median: 1712 ± 5595 bp, P < 10⁻²⁵⁶, Wilcoxon-Mann-Whitney test; see Supplemental Figure 5 online). Short intergenic regions of tail-to-tail, head-to-tail, and tail-to-head gene pairs do not necessarily indicate coregulation.

The correlation coefficient for the transcript levels exceeds 0.9 in 1188 pairs and 0.8 in 1711 pairs (Figure 4; see Supplemental Figure 5 and Supplemental Data Set 5 online). Such correlation exists for both induced and repressed gene pairs.

Several CCM and Related Pathway Genes Exist in HTH Conformation

Among the 1845 significantly induced nonoverlapping genes [FDR q ≤ 0.01, log2(fold) ≥ 1], 212 genes evolved into HTH conformations, in which both genes are upregulated, and an additional 1116 HTH pairs, in which one gene is upregulated. The former include S2 genes previously recognized as regulated by CIA5/CCM1 and associated with CCM (see Supplemental Data Set 1 online). In three pairs, CCP1/LCIE, CCP2/LCID, and CAH4/CAH5, each gene has been reported as CCM related. The sequence similarity, tight chromosomal linkage, and highly coordinated expression of LCID/CCP2 and LCIE/CCP1 (see Supplemental Figure 6 online) suggest recent gene pair duplications. Another nine known CCM-related genes form pairs with genes that were not reported as CCM related before: LC11/Flagellar Associated Protein292 (FAP292), LCI6/513965, CAH1/CLR18, LHC519/525344, LCI7/525344, LCI7/523113, Early Light Inducible4 (ELI4)/519915, LCI31/517052, and GGPS/512500. Transcript levels in seven of these nine pairs are correlated with r > 0.6.

By contrast, gene expression is negatively correlated in 1048 HTH gene pairs, which presents an enigmatic mechanism of transcription regulation (see Supplemental Data Set 6 online). Of these, the transcript levels of 287 pairs had a correlation of −0.8 or lower (see Supplemental Figure 5 and Supplemental Data Set 6 online). Apparently, the extent of negative correlation does not depend on the distance between the two start codons, in sharp contrast with the extent of positive correlations in transcript levels. The most likely explanation for negatively correlated HTH genes is the steric occlusion between overlapping promoter regions, where the direct and indirect binding of transcriptional regulators and the RNA polymerase II to one promoter region excludes binding to the other promoter region. This phenomenon is termed as promoter occlusion or promoter interference (Nakajima et al., 1993). Alternatively, in a minority of negatively correlated HTH gene pairs, switch motifs may account for negatively correlated transcription. We postulated that switch motif(s) might also account for the induction of one gene and the repression of the other. It is also possible that insulator motifs separate the two adjacent but unidirectional and oppositely oriented promoter regions. Using Multiple Expectation Maximization for Motif Elicitation (MEME) and MAST (Bailey et al., 2009), we found two novel putative switch motifs (Figure 5).
Putative switch motif 1 is a degenerate but 21-bp-long motif, while putative switch motif 2 is a well-conserved, 15-bp-long motif (Figure 5). Putative switch motif 1 is present in 44 and switch motif 2 is found in nine intergenic regions between negatively correlated HTH genes. Neither switch motif was found in promoter regions of positively correlated HTH genes.

**Regulation of Specific Sets of CCM Genes**

**Carbonic Anhydrases**

There are three known *C. reinhardtii* α-carbonic anhydrase genes, six known β-carbonic anhydrases, and three putative γ-carbonic anhydrases (Moroney et al., 2011) (see Supplemental Data Set 7 online). Four of the nine α- and β-CA genes (CAH1, CAH4/CAH5, and CAH7) were activated (twofold or more) (Figure 2; see Supplemental Data Sets 2 and 7 online) and none were repressed significantly (FDR ≤ 0.01). Transcripts encoding CAH1, an α-type, periplasmic CA and CAH4/CAH5, recently duplicated, mitochondrial, β-type CAs were induced by over 29 (512-fold), similar to RT-PCR observations by Ynalvez et al. (2008). These results are confirmed by the companion article (Fang et al., 2012), including the activation of CAH1 and CAH4/CAH5. As a notable difference to the observations of Ynalvez et al. (2008) and Fang et al. (2012), we found CAH3 slightly upregulated. In the cia5 mutant, Fang et al. found that CAH5, the most highly upregulated CA gene in wild-type cells, is actually repressed. Moreover, in cia5, CAH1, CAH4, CAH6, and CAH8 were weakly induced (Ynalvez et al., 2008), suggesting the possible existence of a minor CCM regulatory mechanism independent of CIA5.

Levels of CAH2, encoding another periplasmic α-type carbonic anhydrase (Rawat and Moroney, 1991) either remained stable or slightly declined in previous RNA gel blot and macroarray transcriptome analyses (Miura et al., 2004; Wang et al., 2005; Yamano et al., 2008; Ynalvez et al., 2008; Yamano and Fukuzawa, 2009). These observations were corroborated in our transcriptome analysis and further confirmed by qRT-PCR analyses (see Supplemental Figure 1 online). These observations strengthen the conclusion that although CAH2 is possibly a recent duplication of the CAH1 gene (or vice versa), it apparently does not have a role in CCM under limiting CO2 conditions (Fujitaka et al., 1990).

Because of its possible existence, the effect of its possible existence on the plasma membrane, a recently discovered CA, CAH8, has been considered to be a potential contributor to uptake in *C. reinhardtii* (Ynalvez et al., 2008; Moroney et al., 2011). The lack of significant induction of the CAH8 gene by both the Moroney group (Ynalvez et al., 2008) and this study (see Supplemental Data Set 7 online) would suggest that if this is so, CAH8 either contributes little to enhanced Ci uptake during induction of the CCM or that the CAH8 enzyme is present constitutively in nonlimiting quantities. Quantitative measurements of CAH8 concentrations and activities and/or the availability of CAH8 mutants or knockdown lines are needed to differentiate between these alternatives.

Three putative γ-CA genes are not induced by CO2 starvation (see Supplemental Data Set 7 online). These putative CAs appear to be strongly coupled with the Complex I mitochondrial electron transport chain (Klodmann et al., 2010) and therefore may contribute little to the changes in cellular metabolism triggered by CO2 deprivation.

**LCIB and LCIC**

Although no longer considered as directly involved in Ci transport (because of the lack of transmembrane domains), the soluble chloroplast stromal protein, LCIB (Wang and Spalding, 2006), and the complex of LCIC with its close family member, LCIC (Yamano et al., 2010), have been implicated as involved in CO2 retention in the chloroplast or in assisting the Ci transport system within the chloroplast. The novel CO2-requiring phenotypes of mutants containing defective LCIB genes (Van et al., 2001), the unexpected close functional relationship between LCIB and CAH3 (Duanmu et al., 2009), and the close physical association of LCIB (and supposedly the LCIB/LCIC complex) with the ribulose-1,5-bisphosphate carboxylase/oxygenase-rich pyrenoid body when cells are exposed to CO2-limiting conditions (Yamano et al., 2010; Wang et al., 2011) have strongly implicated these proteins as key players in the *C. reinhardtii* CCM. Consistent with this involvement is an observed 20- to 30-fold increase in LCIB and LCIC transcript levels in cells moved from a CO2-replete to a CO2-depleted condition (see Supplemental Data Set 1 online).

**HLA3 and Other Putative Ci Transporters**

Transcripts encoding HLA3, an ATP-energized HCO3- transporter (Duanmu et al., 2009), increase at 60 min by 22-fold and at 180 min by 40-fold (see Supplemental Data Set 2 online). Increases of LCIA (NAR1.2) transcription are a remarkable 3300-fold at 60 min and 4000-fold at 180 min ASVLO2. The list of potential Ci transporters also includes the CCP1 gene (induced 2000-fold) and the CCP2 gene (induced 120-fold), two long-recognized CO2-responsive genes whose products have been implicated in potential Ci transport across the chloroplast envelope. Because these four are among the most highly induced
genes (see Supplemental Data Set 2 online), CCM assembly or augmentation in response to CO₂ deprivation likely requires the de novo assembly of a number of C₂ transporters at several critical subcellular locations. Transporters in the plasma membrane include HLÀ3 (Duanmu et al., 2009) and LCI1 (Ohnishi et al., 2010); in the chloroplast envelope, they include LCIÀ (Duanmu et al., 2009), CCP1, and CCP2 (Pollock et al., 2004). Because evidence exists for a critical role of CAH3 in converting HCO₃⁻ to CO₂ in the lumen of the thylakoid membranes of the chloroplast (Moroney et al., 2011) and the diffusion of the resulting CO₂ to ribulose-1,5-bis-phosphate carboxylase/oxygenase in the pyrenoid body, an argument has been made for the existence of a transporter or channel to shuttle bicarbonate from the alkaline chloroplast stroma to the acidic lumen of the thylakoid membrane. A future quest for this hypothetical HCO₃⁻ transporter/channel among the strongly induced genes encoding transmembrane proteins may prove rewarding.

In our RNA-Seq study, GGPS (LC114) (part of the −15-element chlorophyll a biosynthesis pathway in higher plants) (Robinson et al., 2010), LC125 (encoding a protein homologous to stress induced one-helix protein in Arabidopsis thaliana), and LHCSTR1 (encoding a stress-related chlorophyll a/b binding protein) are all downregulated ASVLCO₂. Because both GGPS and LHCSTR1 are induced by the transcriptional regulator CIA5/CCM1 (Wang et al., 2005), their downregulation may be due to other transcription factors or non-coding RNAs.

Mitochondrial Functions

Our transcriptome analyses suggest that carbon deprivation suppressed most genes involved in classical mitochondrial functions. Transcripts from 24 of the 73 annotated mitochondrial proteins were suppressed and only six were increased (see Supplemental Data Sets 2 and 3 online; Figure 2). On the whole, the decreased transcription of genes producing mitochondrial components is reflective of the decreased metabolic activity of cells exposed to CO₂-limiting conditions.

Regulation of the Response to Very Low CO₂

Despite considerable efforts, most regulators of the CCM remain unknown or incompletely characterized. Indeed, very few C. reinhardtii transcription factor binding sites have been determined experimentally (Kuco et al., 2003; Yoshioka et al., 2004; Sommer et al., 2010; Kropat et al., 2011). Inferences from other photosynthetic organisms are also limited due to the scarcity of known transcription factor binding sites (Yilmaz et al., 2011). Previous studies and our companion article (Fang et al., 2012) have focused on CIA5/CCM1, the master regulator of the transcriptional response to low CO₂, and the putative transcription factor LCR1, itself regulated by CIA5 (Kuco, Yoshioka et al. 2003). The CIA5/CCM1 orthologs in C. reinhardtii and Volvox carteri share 51% identity at the amino acid level (Yamano et al., 2011). This conservation extends to two zinc-finger domains (a mutation that gives rise to the CCM defective phenotype of cia5), a multifunctional protein–protein interaction domain characteristic of transcriptional regulators (Plevin et al., 2005), putative nuclear export and localization signals, and a sumoylation site. We reported earlier (Wang et al., 2005) that regardless of the ambient CO₂ levels, the CIA5/CCM1 gene is constitutively expressed and the protein is also constitutively localized in the nucleus. Given that there is no evidence for CIA5/CCM1 binding to DNA (Kohinata et al., 2008), our hypothesis is that the atypical zinc-finger domains (Kohinata et al., 2008) of CIA5/CCM1 do not bind directly to DNA but regulate other transcription factors as a transcriptional activator via direct or indirect interactions. We postulate that, once activated, these latter factors may bind to cis-regulatory elements in the promoter regions of CO₂-responsive genes to enhance transcription.

To examine this hypothesis, we searched for potential transcription factor binding sites in the putative promoter regions of two gene sets. The first gene set is the union of low CO₂-induced genes identified from previous publications and/or from our study. The second set is the potential CIA5/CCM1-regulated genes identified by the comparative transcriptomic studies of the cia5 mutant and wild-type C. reinhardtii cells in our companion article (Fang et al., 2012). For these and other sets of promoter regions, we performed DNA sequence motif discovery (see Methods). We found several relatively conserved DNA sequence motifs that, as discussed below, are candidates for transcription factor binding sites. Because none of the binding sites is present in more than a quarter of the low CO₂-induced genes, the CCM is apparently regulated by multiple transcription factors, cofactors, and possibly other regulators. From its high level in the regulatory hierarchy, CIA5/CCM1 appears to regulate many or most subordinate regulators, but not all of them.

LCR1, whose binding (enhancer) sites are essential for the induction of the CAH1 gene, is a prime example of a CIA5/CCM1-regulated transcription factor. Using gel mobility shift assays, the Fukuzawa laboratory identified two enhancer elements (EE’s) in the promoter region of CAH1 (Yamano et al., 2011). These elements, EE-1 (5’-AGATTTCACCGGTTGGAAGGAGGT-3’) and EE-2 (5’-CGACTTACGAAAGGAGAT-3’), share the DNA sequence GANTTNC (referred to here as Motif 1A). Later, this motif was shown to be the binding site for LCR1 (Miura et al., 2004). LCR1 is an outlier of the 1R-Myb subfamily; its Myb domain is only ≈20% identical to its relatives. We confirmed the extremely strong induction of the gene coding for the transcription factor LCR1 following CO₂ deprivation (see Supplemental Data Set 2 online). We found for the GANTTNC motif (and its reverse complement) in the upstream 500-bp regions upstream of the start codon of the 893 genes that were upregulated in our experiments using very stringent upregulation criteria: log₂FC ≥ 1 and FDR q ≤ 0.0001 3 h after CO₂ deprivation. A total of 476 promoter regions (53%) matched this low information content motif or its reverse complement (see Supplemental Table 1 online). Surprisingly, however, 697 of the 1227 downregulated genes (78%) and 7141 of the 13,380 genes (53%) that were not significantly differentially regulated by the above criteria also matched Motif 1A. To increase the specificity of the low information content Motif 1A, we searched for its supersets EE-1 and EE-2 in the promoter regions of upregulated genes but found no exact matches other than the previously identified single copy in the CAH1 promoter region. Therefore, we performed de novo motif discovery using a strategy described in detail earlier (Ladunga, 2010). Although
no chromatin immunoprecipitation, protein binding array, or protein–protein interaction data are available for C. reinhardtii, diverse algorithms for promoter sequence analysis, each with different advantages, knockout mutants, and expression data allowed us to obtain insight into the regulatory network of CCM-associated genes (see Methods).

**Putative Novel cis-Regulatory elements: (CA)n Repeats**

From the promoter regions of all upregulated genes in our experiments, the MEME tool (Bailey et al., 2009) reported new motifs. The most remarkable is a (CA)$_n$ repeat (Motif 2), where $n$ ranges from 9 to over 20. (CA)$_n$ is present in 29% of the promoters of upregulated genes ASVLCO$_2$, but it is almost absent from the promoters of downregulated genes. Indications of potential transcription activating mechanisms associated with (CA)$_n$ repeats can be found in mammalian and yeast systems. In the human genome, 19.4 CA repeats occur per mega base pair, representing the most common simple-sequence repeat motif (Waterston et al., 2002). Among all dinucleotides, the CA dinucleotides are the most stable regardless of their environment, and the stability of the DNA structure enhances the stability of the chromatin as well (Fujii et al., 2007). This stable chromatin structure allows the binding of numerous transcriptional activators and splicing regulators. Heterogeneous nuclear RiboNuclearProtein L (hnRNPL) in mammals binds to an intronic polymorphic CA repeat region in the human endothelial nitric oxide synthase gene (Hui et al., 2005). hnRNPL has a role in determining alternative splicing and splicing efficiency. A homologous nuclear ribonucleoprotein exists in C. reinhardtii as well (Johnston et al., 1999). In humans, the length of the (CA)$_n$ repeats positively correlates with the expression of the interferon $\gamma$ gene (Pravica et al., 2000) and the integrin $\alpha2$ gene. Similarly, alleles with longer (CA)$_n$ repeats in the upstream regulatory region of the integrin $\alpha2$ gene enhance the specific binding of poly(ADP-ribose)polymerase-1, a molecular nick sensor, and KU80 polypeptide (Ku80)/70, two components of transcription co-activator complexes (Chelli et al., 2010). C. reinhardtii genes encode for proteins remotely similar to Poly(ADP-ribose)polymerase-1 (PARP-1), including the chloroplast tscA maturation factor (Protein ID: XP_001692722), conserved in V. carteri, Chlorella vulgaris NC64A, Ostreococcus taui, Physcomitrella patens, and Microcosm pulsatil. Other candidates include two Myb2-like transcription factors (Protein IDs 511319 [XP_001690083] and 519163 [XP_001699726]), each of them harboring as many as 11 to 12 SANT DNA binding domains conserved across many eukaryotes, even in mammals. Myb2 transcription factors, associated with calmodulins, coregulate salt and dehydration response in Arabidopsis (Yoo et al., 2009) and numerous other biological processes. LCR1, despite its robust induction, is a less likely candidate because it shares only the Myb domain with RAP1. In summary, the presence of (CA)$_n$ repeats in 29% of the promoters of upregulated genes and their complete absence from the promoters of downregulated genes (see Supplemental Data Set 3 online) combined with knowledge of the activator role these elements in the human integrin $\alpha2$ and interferon $\gamma$ genes discussed above, suggest that the (CA)$_n$ motif in C. reinhardtii may also serve as binding sites for transcriptional activator proteins.

A fourth member of the Myb family of transcription factors is also the induced Myb11 (Protein ID 511940, XP_002945989). Its Arabidopsis ortholog regulates the biosynthesis of flavonoids that absorb potentially damaging UV-B radiation (Stracke et al., 2010a, 2010b). In Arabidopsis, intense light conditions induce Myb11, even under normal CO$_2$ conditions (Stracke et al., 2010b). In C. reinhardtii, because relatively intense light appears to be necessary for the activation of carbon-concentrating mechanisms (Yamano et al., 2008), Myb11 and other light-induced Myb transcription factors may be necessary for low CO$_2$ response.

The discovery of these new, putative, transcription factors should facilitate the discovery of the transcription factor network(s) that regulate several CCM-associated genes. The development of synthetic gene promoters containing these and other putative transcription factor binding sites (motifs) and the isolation of the transcription factors that bind these sites are but two approaches by which the C. reinhardtii community can begin to understand the detailed mechanisms by which algal cells sense and respond to external changes, such as altered CO$_2$ abundance.

**Transcription Factors That May Bind to the (CA)n Motif**

Searching against all known motifs in the TRANSFAC Professional Database (Matys et al., 2006), we found that (CA)$_n$ and similar sequences are bound both in vivo and in vitro by yeast Ras-related protein1 (Rap1), another Myb family repressor-activator protein. In yeast, the general transcription factors Rap1 and autonomously replicating sequence-binding factor1 (Abf1) can create local nucleosome-free regions by evicting nucleosomal histones, thereby facilitating the binding of more specific transcription factors (Casabruna et al., 2011). Rap1 was observed to bind to 5’-CACACCCACACACC-3’ motifs as well, and even low affinity binding sites for Abf1 and Rap1 play a role in determining nucleosome occupancy (Ganapathi et al., 2007). Rap1 is also implicated in silencing at telomeres and silent mating type loci in yeast (Iglesias et al., 2011). A BLAST search could not detect any homologs of Rap1 in the C. reinhardtii proteome. However, more sensitive hidden Markov model searches as implemented in the hmmsearch tool (Johnson et al., 2010) against the PFAM Database of protein domains (Punta et al., 2012) indicated that, at the very least, 55 C. reinhardtii proteins carry some combinations of the Myb-, Sant-, BRCT domain of the BRAC1 oncogene, or homeodomain-related domains that are characteristic of Rap1. Of these, 35 genes are slightly upregulated at 180 min after CO$_2$ deprivation. Based on domain similarity with Rap1 and the upregulation of their genes, our best candidate for (CA)$_n$ binding is the predicted protein 515479 (XP_001692722), conserved in V. carteri, Chlorella vulgaris NC64A, Ostreococcus taui, Physcomitrella patens, and Microcosm pulsatil. Other candidates include two Myb2-like transcription factors (Protein IDs 511319 [XP_001690083] and 519163 [XP_001699726]), each of them harboring as many as 11 to 12 SANT DNA binding domains conserved across many eukaryotes, even in mammals. Myb2 transcription factors, associated with calmodulins, coregulate salt and dehydration response in Arabidopsis (Yoo et al., 2009) and numerous other biological processes. LCR1, despite its robust induction, is a less likely candidate because it shares only the Myb domain with RAP1. In summary, the presence of (CA)$_n$ repeats in 29% of the promoters of upregulated genes and their complete absence from the promoters of downregulated genes (see Supplemental Data Set 3 online) combined with knowledge of the activator role these elements in the human integrin $\alpha2$ and interferon $\gamma$ genes discussed above, suggest that the (CA)$_n$ motif in C. reinhardtii may also serve as binding sites for transcriptional activator proteins.

Transcription Factors That May Bind to the (CA)n Motif

Searching against all known motifs in the TRANSFAC Professional Database (Matys et al., 2006), we found that (CA)$_n$ and similar sequences are bound both in vivo and in vitro by yeast Ras-related protein1 (Rap1), another Myb family repressor-activator protein. In yeast, the general transcription factors Rap1 and autonomously replicating sequence-binding factor1 (Abf1) can create local nucleosome-free regions by evicting nucleosomal histones, thereby facilitating the binding of more specific transcription factors (Casabruna et al., 2011). Rap1 was observed to bind to 5’-CACACCCACACACC-3’ motifs as well, and even low affinity binding sites for Abf1 and Rap1 play a role in determining nucleosome occupancy (Ganapathi et al., 2007). Rap1 is also implicated in silencing at telomeres and silent mating type loci in yeast (Iglesias et al., 2011). A BLAST search could not detect any homologs of Rap1 in the C. reinhardtii proteome. However, more sensitive hidden Markov model searches as implemented in the hmmsearch tool (Johnson et al., 2010) against the PFAM Database of protein domains (Punta et al., 2012) indicated that, at the very least, 55 C. reinhardtii proteins carry some combinations of the Myb-, Sant-, BRCT domain of the BRAC1 oncogene, or homeodomain-related domains that are characteristic of Rap1. Of these, 35 genes are slightly upregulated at 180 min after CO$_2$ deprivation. Based on domain similarity with Rap1 and the upregulation of their genes, our best candidate for (CA)$_n$ binding is the predicted protein 515479 (XP_001692722), conserved in V. carteri, Chlorella vulgaris NC64A, Ostreococcus taui, Physcomitrella patens, and Microcosm pulsatil. Other candidates include two Myb2-like transcription factors (Protein IDs 511319 [XP_001690083] and 519163 [XP_001699726]), each of them harboring as many as 11 to 12 SANT DNA binding domains conserved across many eukaryotes, even in mammals. Myb2 transcription factors, associated with calmodulins, coregulate salt and dehydration response in Arabidopsis (Yoo et al., 2009) and numerous other biological processes. LCR1, despite its robust induction, is a less likely candidate because it shares only the Myb domain with RAP1. In summary, the presence of (CA)$_n$ repeats in 29% of the promoters of upregulated genes and their complete absence from the promoters of downregulated genes (see Supplemental Data Set 3 online) combined with knowledge of the activator role these elements in the human integrin $\alpha2$ and interferon $\gamma$ genes discussed above, suggest that the (CA)$_n$ motif in C. reinhardtii may also serve as binding sites for transcriptional activator proteins.

A fourth member of the Myb family of transcription factors is also the induced Myb11 (Protein ID 511940, XP_002945989). Its Arabidopsis ortholog regulates the biosynthesis of flavonoids that absorb potentially damaging UV-B radiation (Stracke et al., 2010a, 2010b). In Arabidopsis, intense light conditions induce Myb11, even under normal CO$_2$ conditions (Stracke et al., 2010b). In C. reinhardtii, because relatively intense light appears to be necessary for the activation of carbon-concentrating mechanisms (Yamano et al., 2008), Myb11 and other light-induced Myb transcription factors may be necessary for low CO$_2$ response.

The discovery of these new, putative, transcription factors should facilitate the discovery of the transcription factor network(s) that regulate several CCM-associated genes. The development of synthetic gene promoters containing these and other putative transcription factor binding sites (motifs) and the isolation of the transcription factors that bind these sites are but two approaches by which the C. reinhardtii community can begin to understand the detailed mechanisms by which algal cells sense and respond to external changes, such as altered CO$_2$ abundance.
Biological and Experimental Reproducibility

The companion study by Fang et al. (2012) focusing on both the effects of CO₂ deprivation and the role of the CIA5 gene in regulating the C. reinhardtii CCM nicely complement this article. Their findings support many of our prime observations and conclusions in regard to areas in which our studies overlap (i.e., in comparison of transcriptome changes associated with CO₂ deprivation). Nonetheless, the two studies were conducted completely independently and involved a number of significant differences in conditions (detailed below). Thus, it is not unexpected in the comparison of our two data sets there are differences in the patterns and the magnitude of gene expression levels (i.e., it is to be expected that there will be considerably lower interlaboratory reproducibility than intralaboratory reproducibility). Key differences in the two studies include the strains employed, transcript level measurement technologies, the tempo and mode of reducing CO₂ levels, temperature, illumination, and cell densities.

Tempo and Mode of CO₂ Deprivation. In our earlier microarray studies, we rapidly shifted cells from high to ambient CO₂ levels (Wang et al., 2005). In the Fukuzawa laboratory’s EST-based macroarray studies (Miura et al., 2004; Yamano et al., 2008; Yamano and Fukuzawa, 2009), CO₂ levels were shifted gradually. In this study, we initially maintained high CO₂ levels by sparging with 5% CO₂. At 0 time, we switched the input to a mixture of nitrogen and 100 ppm CO₂, which, within the rapidly stirred fermenter, resulted in a slow decline in CO₂ concentration: at 50 min, to ambient levels; at 75 min, to 100 ppm; and finally, at 180 min, to ~20 ppm (Figure 1). By contrast, Fang et al. (2012) measured the effect of CO₂ deprivation 1 h later than we did, 4 h ASVLCO₂. Given the rapid increases and decreases in transcript levels seen for many genes during the 3-h time course of CO₂ starvation (Figure 3), it is expected the extra hour of CO₂ deprivation in the studies of Fang et al. (2012) might well lead to at least some of the differences observed.

Light and Temperature. Our C. reinhardtii cultures grew in a rapidly stirred fermenter at 25°C under white light illumination at 200 μmol photons s⁻¹ m⁻², 2 times higher than in our companion article (Fang et al., 2012). In C. reinhardtii cultures, a light intensity of 200 μmol photons s⁻¹ m⁻² is sufficiently high to potentially cause oxidative damage to pigments, such as chlorophyll (Peers et al., 2009), damage that is much less likely to occur at a light intensity of 100 μmol s⁻¹ m⁻². In addition, Fang et al. maintained cultures at room temperature, whereas our cultures were warmed to a constant 25°C.

Strains of C. reinhardtii May Differ in Splicing and Genomic Polymorphisms. A high proportion of sequencing reads from a nonstandard strain cannot be mapped to the reference transcriptome and genome or to the probes of a microarray, and fewer reads may bias the estimates of differential gene expression. For example, when one or more exons of a multiexon gene are not expressed but the others are, this bias decreases the expression difference that is calculated for a gene model incorrect for the strain.

Quantitative Analysis of Reproducibility

In assessing the potential causes of differences between results in our study and those in our companion article by Fang et al., we feel the above-mentioned variations in strains and culture conditions likely account for the bulk of differences in gene expression levels and patterns. Because previous studies reported almost exclusively genes induced by CO₂ deprivation (Miura et al., 2004; Wang et al., 2005; Yamano et al., 2008, 2010), our analysis will necessarily need to be limited to induced genes. To ensure conservative estimates of differential gene expression, our analyses based on the edgeR tool (Robinson et al., 2010), an FDR threshold of 10⁻⁴ is applied. Our experiments reproduced 166 of the 393 genes (42%) upregulated by CO₂ deprivation in our companion study (Fang et al. 2012) as shown by a Venn diagram (see Supplemental Figure 1 online). We also reproduced 40 genes (38%) upregulated in previous experiments (Miura et al., 2004; Wang et al., 2005; Yamano et al., 2008, 2010), and Fang et al. (2012) reproduced 39 genes (37%). Importantly, 27 of the 106 previously published genes (25%) were reproduced by both Fang et al. (2012) and us. Indeed, the expression patterns of genes most critical to CCM are similar across these three gene sets (see Supplemental Data Set 1 online). In the order of fold change 180 versus 0 min, these genes include carbonic anhydrases 1, 4, 5; NAR1 (LCIA), low carbon–induced genes (LCI) 1, 23, 11, D, B, 15, C, 22, 31, and 19, as well as CCP1 and 2, LHCSR2 and 3, HLA3, guanine deaminase 1 (GUD1), Chloroplast DnaJ-like protein3 (CDJ3), ELH (LCI16), and Alanine Aminotransferase1 (AAT1).

Perspectives

The RNA-Seq data presented here relating changes in environmental CO₂ abundance to changes in gene expression levels greatly expand our knowledge and understanding of the CCM compared with data obtained in more limited RNA gel blot and microarray studies conducted earlier. Some general conclusions made in the earlier investigations have been confirmed, such as the expected decline in photosynthetic capacity as CO₂ availability becomes limited. Other observations, such as the marked downregulation of a vast number of genes immediately after CO₂ decrease have not been noted earlier and may provide the foundation for a more detailed examination of how algal cells perceive and respond so rapidly to changes in environmental conditions that demand prompt changes in cellular metabolism and physiology. Likewise, our studies revealed an unexpectedly large number of gene pairs involved in similar cellular activities that are oriented in a HTH conformation, where a shared bidirectional promoter coordinates regulates the response to changes in CO₂ levels. Investigations of common regulatory elements now become a potentially fruitful area of investigation. Our results from a time-course study of changes in the transcriptomes at critical time points after imposition of CO₂ starvation builds on results from the accompanying publication of the Spalding laboratory (Fang et al., 2012) regarding the pivotal regulatory role of CIA5/CCM1 in response to CO₂ deprivation.
Our data confirm that synthesis of many of the key components of the CCM are triggered when CO₂ levels decrease and reveal an even greater number of candidate genes whose function may be essential or helpful to assembly of a fully functional CCM. These genes become potential targets for chromatin immunoprecipitation and deep sequencing (Wang et al., 2011) or for inactivation through RNA interference techniques (Cerutti et al., 2011), insertional gene inactivation (Emnilova et al., 2000), or targeted gene knockout using zinc-finger nucleases or newly developed techniques, such as TAL effector nucleases (Li et al., 2011a). The vast array of genes induced or repressed as a result of CO₂ changes point to a sophisticated set of regulatory networks that must effectively govern multiple cellular responses, an area of study that appears fruitful for future exploration.

METHODS

C₃ Deprivation

Chlamydomonas reinhardtii wild-type strain CC124 was used for analysis. Briefly, cells were grown in 2 liters of Tris Phosphate medium at 25°C and 3% CO₂ to a density of 1 × 10⁶ cells/mL (Harris, 1989) before being transferred to a 3-liter autoclavable glass bioreactor (Applikon Biotechnology) that was connected with EZ control for analysis of temperature, pH, and dissolved oxygen. The bioreactor was illuminated with a light intensity of 200 μmol photons m⁻² s⁻¹, and an input gas containing 5% CO₂ was introduced. Algal cells were allowed to equilibrate with the new environment for 1 h. Following a sampling of the culture, the input gas for the bioreactor was shifted to 100 ppm CO₂, which was monitored in the culture using two CO₂ transmitters (Vaisala; models GMT221 and GMT222). Samples were taken at 15, 30, 60, and 180 min following the shift to 100 ppm CO₂. During the experiment, pH was maintained at 7.2 using 3 M KOH.

RNA Preparation

Cellular samples were taken from the bioreactors in 100-mL volumes and transferred directly to a sterilized 2-liter Erlenmeyer flask submerged in an ice water bath. The flask was agitated within the ice water for 2 min to rapidly decrease the temperature of the algal culture to reduce the degree of transcriptional changes preceding removal from the bioreactor. The cellular samples were then centrifuged at 2000 g for 5 min at 4°C. The supernatant was discarded and the cellular pellet was frozen by immersing the centrifuge tube in liquid nitrogen. Samples were stored at −80°C until RNA extraction.

RNA extraction was performed using TRIzol LS reagent (Life Technologies). Briefly, cellular pellets composed of 5 × 10⁶ cells were re-suspended and incubated for 5 min at room temperature with Trizol LS reagent. Chloroform was added to the samples, agitated for 15 s by tube inversion, and allowed to incubate at room temperature for 15 min. Samples were centrifuged at 12,000 g for 15 min at 4°C. The aqueous phase was recovered and transferred to new tubes. Isopropanol was added to the aqueous phase, mixed, and centrifuged at 12,000 g for 10 min at 4°C to pellet RNA. Isopropanol was removed and the RNA pellet was washed with 75% ethanol. Following further centrifugation (7500 g for 5 min at 4°C), ethanol was removed and the RNA pellets were allowed to air dry for 5 min. RNA pellets were solubilized in 500 μL nuclelease-free water. RNA was further purified by precipitation with lithium chloride. An equal volume (500 μL) of 4 M LiCl was added with mixing, and samples were incubated at −20°C for 60 min. Following this incubation, samples were centrifuged at 16,000 g for 20 min at 4°C. The aqueous phase was removed and discarded and the RNA pellet was washed twice with 75% ethanol. Following air drying, the pellet was resuspended in 50 μL of nuclease-free water. RNA samples were then analyzed using a NanoDrop 2000 spectrophotometer (NanoDrop Products/Thermo Fisher Scientific) to verify RNA quantity and purity.

Preliminary Analysis of RNA Samples

To confirm induction of the carbon-concentrating mechanism, preliminary analysis of RNA samples was performed using qRT-PCR. RNA samples were prepared for analysis using the Plexor Two-Step qRT-PCR system (Promega). qRT-PCR analysis was performed using a 7500 Real-Time PCR System (Life Technologies). The genes LCIA (AB168092), LCIIB (XM_001698292), and mitochondrial carbonic anhydrase (CAH4, XM_001695951) were chosen for analysis as they have been observed to increase in expression during carbon deprivation (Miura et al., 2004; Yamano et al., 2008). CAH2 (XM_54488) was also selected as a control gene reported as displaying a moderate decrease in expression in response to carbon deprivation. CIA5/CCM1 (AF317732) was used as a positive control as it shows constitutive expression during carbon deprivation. Fluorescently labeled primer pairs were designed for each of the aforementioned genes. Quantitative PCR analysis was performed using a 7500 Real-Time PCR System by measuring the threshold cycle (C_t) of each gene. Using the C_t values of CIA5/CCM1 for each RNA sample as a baseline control, the change in Cₜ for each gene could be used to calculate the fold change response of each gene throughout the time course (see Supplemental Figure 1 online).

RNA Sequencing, Mapping, and the Analyses of Gene Expression

From the qRT-PCR data, it was determined that four time points should be analyzed by RNA-Seq. A 15-min ASVLCO₂ sample was omitted from RNA-Seq analysis of this sample showed limited induction of the aforementioned genes. C. reinhardtii equilibrated at ~5% CO₂ was used as the 0 time control, and three time points ASVLCO₂ (30, 60, and 180 min) were also analyzed. To provide for biological replicates, RNA samples from two individual bioreactor runs were analyzed. In total, eight RNA samples were submitted for RNA-Seq. Prior to submission, RNA samples were treated with DNase and resuspended in 8.3 mM Tris-HCl and 4.2 mM EDTA. RNA-Seq was performed at the JGI using an Illumina Genome Analyzer II.

Sequencing reads were mapped to the C. reinhardtii version 4 genome (Department of Energy JGI) as well as to the processed Augustus5 (Stanke and Waack, 2003) exon structure predictions using the tophat and cufflinks software (Trapnell et al., 2009). No more than two mismatches per sequencing read were allowed. Analyses of differential expression including FDR calculations were performed using three independent Bioconductor packages: edgeR (Robinson et al., 2010), DESeq (Anders and Huber, 2010), and baySeq (Hardcastle and Kelly, 2010).

Time series analysis of the transcriptional response was performed by k-means cluster analysis (MacQueen, 1967). This method partitions fold change patterns into k clusters where each fold change time series belongs to the cluster with the nearest mean. The clusters are iteratively refined. Such analyses have been used to identify temporal expression patterns of a large numbers of genes.

Gene Ontology Analyses

Complex functional patterns of the differentially regulated genes emerge at the level of photosynthetic categories, low CO₂-regulated genes, and plant and diatom genes that have no close relatives in other kingdoms or in prokaryotes other than cyanobacteria (Karpowicz et al., 2011). We
extended these analyses to GO (Ashburner et al., 2000), a system for the hierarchical annotation of homologous gene and protein sequences in multiple organisms using a common, controlled vocabulary. GO allows the practical, high-throughput interpretation of experiments including RNA-Seq. To avoid the subjectivity inherent in the ad hoc interpretations for less than obvious patterns, a rigorous method was employed to assess the statistical significance of expression patterns, called GSEA (Subramanian et al., 2005). Briefly, GSEA ranks genes by fold changes and calculates enrichment scores for each set. Then, primarily upregulated gene sets are assigned high positive enrichment scores and primarily downregulated sets are assigned low negative scores. For the statistical significance of these enrichment scores, FDR (Benjamini and Hochberg, 1995) is calculated.

De Novo Motif Discovery of Putative Transcription Factor Binding Sites

Our complex strategy for the discovery and limited confirmation of the transcriptional regulatory network was described earlier (Ladunga, 2010). Even in the almost complete absence of chromatin immunoprecipitation, protein binding array, or protein–protein interaction observations for algae, an array of motif discovery algorithms for promoter sequence analysis, each complementing the others, knockout mutants of transcription factors, and RNA-Seq data allowed us to better understand the CCM regulatory network. A key tool is the MEME package (Bailey et al., 2009) for the identification of statistically overrepresented variable sequence motifs. We searched all promoter regions for all motifs represented as positional weight matrices in the commercial version of the TRANSFAC Database (Matys et al., 2006) using its advanced search tool. Conversely, all identified motif were queried against the TRANSFAC motifs.

Accession Numbers

Sequence data from this article can be found in the GenBank/EMBL data libraries under the following accession numbers (sequences are from C. reinhardtii unless otherwise noted): AAT1, XM_001698466; Saccharomyces cerevisiae Abf1, M29067; Argonaute1, XM_001694788; CAH1, D90206; CAH2, X54488; CAH3, U40871; CAH4, XM_001695951; CAH5, XM_001700718; CAH6, AY463239; CAH7, EU045569; CAH8, EU045570; CAH9, XM_001700857; CAH10, XM_001703185; CAH11, AY538680; CAH12, AY462241; CFP1/CIC1, XM_001692415; CFP2/CIC/P, XM_00169236; CDS, XM_001700205; CIAS/CICM, AF317732; CIIE, XM_001692143; Dicer-like1, EU707797; EU41/LC16, XM_00169432; FAP292, XM_001703500; GGPS/LC14, XM_001694932; GUD1, XM_001695571; HCF1136, XM_001696142; HLA3, XM_001699988; human integrin α2, NM_002203; human interferon γ2, NM_005534; KU80-like domain, XM_001702527; LC11, XM_001703335; LC11, XM_001697911; LC15, XM_001691879; LC19, XM_001696786; LC21, XM_001692906; LC22, XM_001702321; LC23, XM_001695392; LC25, XM_001696602; LC31, XM_001691740; LC6, XM_001693972, AB168091; LC7, XM_001696515; LCIA, AB168092; LCIB, XM_001698292; LCIC, AB168094; LCID, DE671955; LCIE, DQ694007; LCIR1, AB168089; LCIS1, XM_001696086; LHCSR2, XM_001696012; LHCSR3, XM_001696086; Arabidopsis thaliana Myb11, AF062863; Myo2, XM_001690031; human PARP1, NM_001618; Saccharomyces cerevisiae Rap1, M18808; tcsA, XM_001694379. All RNA-Seq data are available at the National Center for Biotechnology Information Sequencing Read Archive at http://trace.ncbi.nlm.nih.gov/Traces/sra/?study=SRP004215.

Supplemental Data

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

Supplemental Figure 1. qRT-PCR Support for RNA Sequencing Results.

Supplemental Figure 2. Overlaps among the Lists of Differentially Expressed Genes Reported by the edgeR, DESeq, and baySeq Tools.

Supplemental Figure 3. Genes Showing a Transient Decrease in Response to CO2 Deprivation.

Supplemental Figure 4. Genes Showing a Transient Increase in Response to CO2 Deprivation.

Supplemental Figure 5. Spacing and Expression of HTH Conformation Genes.

Supplemental Figure 6. Comparison of Transcript Levels from Two Closely Linked Head-To-Head Gene Pairs.

Supplemental Figure 7. Overlaps Among Sets of Genes Induced by CO2 Deprivation.

Supplemental Table 1. Widespread DNA Sequence Motifs as Putative Transcription Factor Binding Sites in the Upstream Regions of Genes.

Supplemental Data Set 1. Comparison of Differential Regulation Observations between RNA-Seq (This Study) and Arrays.

Supplemental Data Set 2. The Statistically Significantly Upregulated Genes for 180 versus 0, 60 versus 0, and 30 versus 0 Min ASVLCO2.

Supplemental Data Set 3. The Statistically Significantly Downregulated Genes for 180 versus 0, 60 versus 0, and 30 versus 0 Min ASVLCO2.


Supplemental Data Set 5. The Most Highly Correlated HTH Gene Pairs Potentially Regulated by Bidirectional Promoters.

Supplemental Data Set 6. Anticorrelated Expression in HTH Gene Pairs with a Pearson Correlation Coefficient $r \leq -0.8$.

Supplemental Data Set 7. Expression Profiles of the 12 Carbonic Anhydrase Genes.

ACKNOWLEDGMENTS

We thank Christa Pennachio, Erika Lindquist, and Feng Chen at JGI for sequencing, Paul Blum and Kemp Horken for help with the biofermenter, Martin H. Spalding, Wei Fang, Sabeeha S. Merchant, and Matteo Pellegrini for the access to their results and for reviewing the article, and Jean-Jack M. Riethoven and Kartik Vedalaveni for assistance. D.P.W. acknowledges support from the Department of Energy JGI (Award JGI 2010 CSP-146), the National Science Foundation (Grants MCB-0952533 and EPSCoR-1004094), and Consortium for Algal Biofuels Commercialization (Department of Energy Award DE-EE0003373). I.L. and D.P.W. received support from an interdisciplinary grant from the University of Nebraska–Lincoln, D.C. is supported by a grant to Sabeeha S. Merchant from the Air Force Office of Science Research (FA 9550-10-1-0095). The views presented in this work are not endorsed by the sponsors.

AUTHOR CONTRIBUTIONS


Received November 22, 2011; revised March 2, 2012; accepted May 6, 2012; published May 25, 2012.
REFERENCES


Supplemental Figure 1. qRT-PCR Support for RNA Sequencing Results.

Original total RNA samples that were used for RNA-Seq analyses during the first three hours after a shift to very low CO₂ conditions were used as substrates for qRT-PCR analyses of transcript levels from the LCIA, LCIB, CAH5 and CAH2 genes. Fold changes in mRNA abundance as determined by the two independent methods are plotted against the duration of exposure to lowered CO₂ concentrations.
Supplemental Figure 2. Overlaps Among the Lists of Differentially Expressed Genes Reported by the edgeR, DESeq and baySeq Tools.

Data representing overlaps are displayed as a Venn diagram in which the numbers represent the number of genes in each overlapping section.
Supplemental Figure 3. Genes Showing a Transient Decrease in Response to CO₂ Deprivation.

Lines represent a cluster of genes displaying a significant decline in expression level at 60 minutes following a shift of *Chlamydomonas* cells to CO₂ depletion conditions, but a marked increase in transcript levels by 180 minutes.
**Supplemental Figure 4. Genes Showing a Transient Increase in Response to CO₂ Deprivation.**

Lines represent a cluster of genes displaying marked increases in gene transcript levels at 60 minutes following a shift of *Chlamydomonas* cells to CO₂ deprivation conditions, but a strong decline in expression by 180 minutes.
Supplemental Figure 5. Spacing and Expression of HTH Conformation Genes.

Genes of HTH Conformation (red lines) tend to be closer to each other and their expression correlates more strongly than those of genes in other conformations (blue lines).

(A) Histogram of gene-wise correlations of transcript levels for 0, 30, 60 and 180 minutes after a shift to very low CO\textsubscript{2} conditions.

(B) Histogram (probability density distribution) of the distances between coding regions. Distances up to 500 base pairs are shown.
Supplemental Figure 6. Comparison of Transcript Levels from Two Closely Linked Head-To-Head (HTH) Gene Pairs.

(A) HTH genes LCID/CCP2 are located in close proximity to the paralogous HTH gene pair, LCIE/CCP1, as the likely result of a recent gene pair duplication.

(B) and (C) Tight co-regulation of the gene pairs LCID/CCP2 (B) and LCIE/CCP1 (C) as measured by RNA-Seq.
Supplemental Figure 7. Overlaps Among Sets of Genes Induced by CO₂ Deprivation.

Genes identified in the current research induced by a factor of 2 or higher (log2(fold change) ≥ 1) and FDR q ≤ 0.0001 (red oval), genes indicated by a general mixed model in the companion publication (Fang et al., 2012, blue oval), and 105 previously published genes (Miura et al., 2004; Wang et al., 2005; Yamano et al., 2008; Yamano and Fukuzawa, 2009).


**Supplemental Table 1.** Widespread DNA Sequence Motifs as Putative Transcription Factor Binding Sites in the Upstream Regions of Genes.

<table>
<thead>
<tr>
<th>No.</th>
<th>Motif or its reverse complement</th>
<th>Identified by</th>
<th>(putative) transcription factor</th>
<th>Count (p-value) of matching promoters</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>upreg</td>
</tr>
<tr>
<td>1A</td>
<td>GANTTNC</td>
<td>Gel mobility shift assay</td>
<td>LCR1</td>
<td>476</td>
</tr>
<tr>
<td>EE-1</td>
<td>AGATTTTCACGGTGGGAAGGAGGT</td>
<td>Gel mobility shift assay</td>
<td>LCR1</td>
<td>1</td>
</tr>
<tr>
<td>EE-2</td>
<td>CGACTTACGAA</td>
<td>Gel mobility shift assay</td>
<td>LCR1</td>
<td>1</td>
</tr>
<tr>
<td>1B</td>
<td>GA[CTA]TTG[AT]</td>
<td>See Methods</td>
<td>LCR1</td>
<td></td>
</tr>
<tr>
<td>1C</td>
<td>See Fig. 6.</td>
<td>See Methods</td>
<td>LCR1</td>
<td>11</td>
</tr>
<tr>
<td>2</td>
<td>(CA)$_n$, $n \geq 9$</td>
<td>grep</td>
<td>Potential homolog of yeast RAP1, human PPAR, etc.</td>
<td>30</td>
</tr>
<tr>
<td>2A</td>
<td>(CA)$_n$, $15 \leq n \leq 25$ with up to 4 mismatches</td>
<td>agrep</td>
<td>See above</td>
<td>34</td>
</tr>
<tr>
<td>3</td>
<td>AGGGGCAGGGG</td>
<td>MEME</td>
<td>unknown</td>
<td></td>
</tr>
<tr>
<td>Switch motif 1</td>
<td>See Fig. 6.</td>
<td>MEME, MAST</td>
<td>na</td>
<td>-</td>
</tr>
<tr>
<td>Switch motif 2</td>
<td>See Fig. 6.</td>
<td>MEME, MAST</td>
<td>na</td>
<td>-</td>
</tr>
</tbody>
</table>