CHARACTERIZING GENE EXPRESSION CHANGES AND GENE FAMILY EVOLUTION IN THE CONTEXT OF STRESS RESPONSE IN MICRO-ALGAL SPECIES

By

Guangxi Wu

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Cell and Molecular Biology – Doctor of Philosophy

2013

ABSTRACT

CHARACTERIZING GENE EXPRESSION CHANGES AND GENE FAMILY EVOLUTION IN THE CONTEXT OF STRESS RESPONSE IN MICRO-ALGAL SPECIES

By

Guangxi Wu

Algae are a large group of organisms that are of ecological and economical importance, as they play an important role in the food web and the biogeochemical cycle, and some of them are found to be suitable for biofuel production. Under stress conditions, they accumulate a large amount of storage lipid. However, the mechanistic details of this phenomenon are not yet well understood. Towards deciphering this phenomenon, I focused on characterizing gene expression changes and gene family evolution in the context of stress response in two micro-algal species. I first analyzed the global gene expression changes of the green alga model organism *Chlamydomonas reinhardtii* under normal and nitrogen deprived conditions. I found that global gene expression changes significantly under nitrogen deprivation; lipid metabolism was associated with up-regulated genes while DNA replication and photosynthesis were down-regulated. Second, I outlined global gene family evolution in the green algal lineage, particularly in the extent of duplicate retention and loss, and stress response evolution among duplicates. I found that stress responsive genes tend to be lineage-specifically retained and functional gains occur frequently in gene duplicates, shaping the stress response gene repertoire in a species-specific manner. Finally, I assembled and annotated the genome of the stramenopile *Nannochloropsis oceanica* CCMP1779, a phylogenetically distant species to *C. reinhardtii*

considered for biofuel production. I found that *N. oceanica* responds similarly to nitrogen deprivation as *C. reinhardtii*, suggesting that general metabolic change is conserved across distantly related species. In addition, we compared the *N. oceanica* genome to *Nannochloropsis gaditana* to reveal its uniqueness in gene repertoire. We found that it is significantly different from *N. gaditana* and this might reflect physiological and biochemical differences. Overall, I reveal that though major metabolic changes under stress in micro-algae across diverse phylogeny are similar, the particular genes involved in stress response could be significantly different as they have been shaped by lineage-specific evolution. Thus, the species-specific mechanism in stress response cannot be deciphered by studying one model organism. Therefore, it would be beneficial to explore more micro-algal species, especially the ones considered for biofuel production, to discover their uniqueness in stress response towards identifying and engineering the ideal alga for biofuel production.

ACKNOWLEDGEMENTS

First, I would like to acknowledge my advisor Dr. Shin-Han Shiu for his continuous support and excellent mentorship during my years in his laboratory as a graduate student. My achievements are not possible without his efforts. I would also like to thank my former advisor Dr. Thomas Whittam (1954-2008) for jumpstarting my interest in bioinformatics and molecular evolution.

Next, I would like to thank my guidance committee members for sharing their insightful perspective in their respective field and helpful opinions about my dissertation project. Specifically I thank Dr. Christoph Benning for nurturing my interest in stress response in algae; Dr. C. Robin Buell for discussions in genomics and bioinformatics; Dr. Barbara Sears for sharing her knowledge in *Chlamydomonas reinhardtii* and scientific writing; and Dr. Barry Williams for advice in evolutionary biology.

Last but not least, I would like to thank my coworkers, collaborators, the CMB program, friends, and family for their endless support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

INTRODUCTION

Algae are a group of heterogeneous organisms that are of ecological importance in the food web, in symbiotic relationships with other organisms, and in the biogeochemical cycle as a major generator of atmospheric oxygen and organic carbon (Round, 1984; Graham et al., 2009). They are also a major contributor to biodiversity with an estimated number of 10 million species (Graham et al., 2009). Moreover, algae are useful to humans as food sources for human and aquaculture, sources of industrially useful products, pollutant removal agents, and model organisms for scientific research (Graham et al., 2009). Recently, scientists have started exploring the possibilities of using microalgae to generate biofuels (Hu et al., 2008).

Microalgae live in diverse habitats that have a wide range of salinity, temperatures, pH, and nutrient availabilities (Graham et al., 2009). Some microalgae are considered suitable for biofuel production due to their high oil content; when grown under various stress conditions, they redirect metabolism towards the accumulation of neutral lipids such as triacylglycerol (TAG) (Hu et al., 2008). Thus, studying stress responses in microalgae is of biological and economical importance. Here, I focus on two single-celled microalgal species, the green alga *Chlamydomonas reinhardtii* and the stramenopile *Nannochloropsis oceanica* CCMP1779. *C. reinhardtii* is a soil-dwelling alga with one chloroplast, multiple mitochondria, and two anterior flagella; it is in the green algal lineage that is more closely related to the ancestor of land plants than any other protists (Merchant et al., 2007). *C. reinhardtii* has been used as model organism to study photosynthesis and flagella and basal body functions, and its genome assembly is available (Merchant et al., 2007). *N. oceanica* is a marine coccoid with a plastid derived from secondary symbiosis (Reyes-Prieto et al., 2007; Graham et al., 2009); it is under the class of eustigmatophyceae in heterokontophyta, which also includes diatoms and brown algae (Hoek, 1995). In addition, it is considered for biofuel production due to its high oil content (Gouveia and Oliveira, 2009).

In microalgae, nitrogen (N) deprivation is among the most critical stress conditions that affects lipid metabolism (Hu et al., 2008), and in *C. reinhardtii*, the formation of lipid droplets high in TAG content after N deprivation has been well documented (Wang et al., 2009b; Moellering and Benning, 2010). Though *C. reinhardtii* is not directly considered for biofuel production, investigating its stress response mechanism could provide basic insights into other algal species, especially green algae. To investigate the mechanistic details of N deprivation response on the gene expression level, the transcriptomes under normal and N deprived conditions were determined. Microarrays have been used to characterize gene expression changes in *C. reinhardtii* under stress conditions (Ledford et al., 2004; Jamers et al., 2006; Ledford et al., 2007; Mus et al., 2007; Nguyen et al., 2008; Simon et al., 2008; Yamano et al., 2008; Mustroph et al., 2010); however, they do not cover all genes in the genome and have other limitations, such as reliance on existing knowledge, high background noise due to cross-hybridization, and limited dynamic range of detection (Wang et al., 2009a). In this study, I applied RNA-seq approaches to characterize the global gene expression changes in *C. reinhardtii* under N deprivation. RNA-seq revealed the major metabolic changes after stress induction. However, inferring metabolic fluxes from gene expression has its caveats as gene expression changes do not directly translate into metabolic changes. Therefore, subsequent labeling experiments were performed by the Shachar-Hill lab to investigate the meaningfulness of such inference.

After determining the global gene expression changes under N deprivation in *C. reinhardtii*, I next investigated how gene duplicate retention and subsequent functional gain at the expression level have shaped the gene repertoire involved in stress response in green algae. Gene duplication is an important source of generating raw genetic material for evolution to act upon. The most common fate of gene duplicates is rapid loss following the duplication events although a significant number of duplicates are retained (Lynch et al., 2001; Moore and Purugganan, 2003; Moore and Purugganan, 2005). The subsequent gene family expansion due to duplicate retention could contribute to organismal and regulatory complexity (Lespinet et al., 2002; Vogel and Chothia, 2006). Among retained duplicates, functional bias exists, for example: functional categories related to stress response tend to be associated with retained duplicates in a wide range of organisms (Lespinet et al., 2002; Wapinski et al., 2007; Hanada et al., 2008). In green algae, despite a number of studies on individual gene families (Porter et al., 1996; Hallmann, 2006; Mariscal et al., 2006; Hua et al., 2011; Herrera-Valencia et al., 2012) and one cross-species analysis focused on a limited number of functional categories (Blanc et al., 2012), there is not yet a global analysis of all gene families. Here, I applied phylogenetic approaches to investigate the gene family evolution in *C. reinhardtii* and eight other species with genome sequences available on www.jgi.org. Further, I examined the functional evolution of gene duplicates in the context of stress response by reconstructing the ancestral gene response state in *C. reinhardtii* using above mentioned N deprivation and five other stress related RNA-seq datasets. Such a study revealed how gene duplication and subsequent gain-of-function events contributed to the diversity of stress responsive gene repertoires in the green algal lineage, ultimately leading to the tremendous diversity in phenotypes, in aspects such as fatty acid composition, the interaction of growth and lipid content, the time course of lipid accumulation, and the threshold of stress required to stimulate lipid formation (Hu et al., 2008; Adams et al., 2013).

Though *C. reinhardtii* has traditionally been used as a model organism for microalgae (Merchant et al., 2007), it has caveats: it is not a direct candidate for biofuel production and has limited tools for molecular manipulation. For example, targeted inactivation of genes by homologous recombination is not available, and RNA interference is not efficient in generating loss-of-function *C. reinhardtii* mutants. Therefore, there is a need to develop a new microalgal model organism. The genus of *Nannochloropsis* in the supergroup of stramenopila is directly considered for biofuel production and various species of this genus have been studied for their lipid composition and accumulation (Sukenik and Carmeli, 1990; Schneider and Roessler, 1994; Tonon et al., 2002; Danielewicz et al., 2011). Moreover, a number of studies have focused on the biomass production under different conditions (Hu and Gao, 2003; Xu et al., 2004; Gouveia and Oliveira, 2009; Rodolfi et al., 2009; Simionato et al., 2011; Srinivas and Ochs, 2012). Among them, one study considered a *Nannochloropsis* species suitable for biofuel production due to a high oil content (28.7% of dry weight); its oil quantity increases greatly when grown under N deficient conditions (Gouveia and Oliveira, 2009). Recently, high-efficiency homologous recombination was achieved in *N. oceanica* (Kilian et al., 2011), opening up possibilities of developing an alternative model organism to interrogate the relationships between stress response mechanism and lipid metabolism in microalgae. Here, I focus on the publicly available *N. oceanica* CCMP1779, which was chosen for its growth in culture, antibiotic sensitivity, and ease of nuclear transformation. I assembled its genome and used transcriptomes under normal and N deprived conditions to facilitate its gene annotation. A consortium of scientists manually annotated selected pathways in order to better understand the biology of *N. oceanica*. In addition, a protocol for transformation was provided by the Benning lab. Next, I performed comparative genomics analysis between *N. oceanica* and the recently sequenced genome of *Nannochloropsis gaditana* (Radakovits et al., 2012) to reveal the difference in gene repertoire within the *Nannochloropsis* genus. Overall, draft genome assembly and extensive annotation, coupled with a transformation protocol, are valuable resources for the research community to further explore the biology and the industrial potential of this species.

In conclusion, my research examined the mechanistic and evolutionary details of stress response in two microalgal species towards the ultimate goal of deciphering the mechanisms of stress response and subsequent lipid accumulation. My work sheds light on the impact of model organism usage in this research field while providing an attractive new alternative. In addition, my results will aid the identification of genes for further molecular characterization and manipulation to optimize lipid production.

REFERENCES

REFERENCES

Adams C, Godfrey V, Wahlen B, Seefeldt L, Bugbee B (2013) Understanding precision nitrogen stress to optimize the growth and lipid content tradeoff in oleaginous green microalgae. Bioresour Technol **131**: 188–194

Blanc G, Agarkova I, Grimwood J, Kuo A, Brueggeman A, Dunigan DD, Gurnon J, Ladunga I, Lindquist E, Lucas S, et al (2012) The genome of the polar eukaryotic microalga *Coccomyxa subellipsoidea* reveals traits of cold adaptation. Genome Biol **13**: R39

Danielewicz MA, Anderson LA, Franz AK (2011) Triacylglycerol profiling of marine microalgae by mass spectrometry. J Lipid Res **52**: 2101–2108

Gouveia L, Oliveira AC (2009) Microalgae as a raw material for biofuels production. J Ind Microbiol Biotechnol **36**: 269–274

Graham L, Graham J, Wilcox L (2009) Algae, 2nd ed. Pearson

Hallmann A (2006) The pherophorins: common, versatile building blocks in the evolution of extracellular matrix architecture in *Volvocales*. Plant J **45**: 292–307

Hanada K, Zou C, Lehti-Shiu MD, Shinozaki K, Shiu S-H (2008) Importance of Lineage-Specific Expansion of Plant Tandem Duplicates in the Adaptive Response to Environmental Stimuli. Plant Physiol **148**: 993–1003

Herrera-Valencia VA, Macario-González LA, Casais-Molina ML, Beltran-Aguilar AG, Peraza-Echeverría S (2012) In Silico Cloning and Characterization of the Glycerol-3-Phosphate Dehydrogenase (GPDH) Gene Family in the Green Microalga *Chlamydomonas reinhardtii*. Curr Microbiol **64**: 477–485

Hoek CVD (1995) Algae: An Introduction to Phycology. Cambridge University Press

Hu H, Gao K (2003) Optimization of growth and fatty acid composition of a unicellular marine picoplankton, *Nannochloropsis* sp., with enriched carbon sources. Biotechnol Lett **25**: 421–425

Hu Q, Sommerfeld M, Jarvis E, Ghirardi M, Posewitz M, Seibert M, Darzins A (2008) Microalgal triacylglycerols as feedstocks for biofuel production: perspectives and advances. Plant J Cell Mol Biol **54**: 621–639

Hua Z, Zou C, Shiu S-H, Vierstra RD (2011) Phylogenetic Comparison of F-Box (FBX) Gene Superfamily within the Plant Kingdom Reveals Divergent Evolutionary Histories Indicative of Genomic Drift. PLoS ONE **6**: e16219

Jamers A, Van der Ven K, Moens L, Robbens J, Potters G, Guisez Y, Blust R, De Coen W (2006) Effect of copper exposure on gene expression profiles in *Chlamydomonas reinhardtii* based on microarray analysis. Aquat Toxicol Amst Neth **80**: 249–260

Kilian O, Benemann CSE, Niyogi KK, Vick B (2011) High-efficiency homologous recombination in the oil-producing alga *Nannochloropsis* sp. Proc Natl Acad Sci **108**: 21265–21269

Ledford HK, Baroli I, Shin JW, Fischer BB, Eggen RIL, Niyogi KK (2004) Comparative profiling of lipid-soluble antioxidants and transcripts reveals two phases of photo-oxidative stress in a xanthophyll-deficient mutant of *Chlamydomonas reinhardtii*. Mol Genet Genomics MGG **272**: 470–479

Ledford HK, Chin BL, Niyogi KK (2007) Acclimation to singlet oxygen stress in *Chlamydomonas reinhardtii*. Eukaryot Cell **6**: 919–930

Lespinet O, Wolf YI, Koonin EV, Aravind L (2002) The Role of Lineage-Specific Gene Family Expansion in the Evolution of Eukaryotes. Genome Res **12**: 1048–1059

Lynch M, O'Hely M, Walsh B, Force A (2001) The Probability of Preservation of a Newly Arisen Gene Duplicate. Genetics **159**: 1789–1804

Mariscal V, Moulin P, Orsel M, Miller AJ, Fernández E, Galván A (2006) Differential Regulation of the *Chlamydomonas* Nar1 Gene Family by Carbon and Nitrogen. Protist **157**: 421–433

Merchant SS, Prochnik SE, Vallon O, Harris EH, Karpowicz SJ, Witman GB, Terry A, Salamov A, Fritz-Laylin LK, Maréchal-Drouard L, et al (2007) The *Chlamydomonas* Genome Reveals the Evolution of Key Animal and Plant Functions. Science **318**: 245–250

Moellering ER, Benning C (2010) RNA Interference Silencing of a Major Lipid Droplet Protein Affects Lipid Droplet Size in *Chlamydomonas reinhardtii*. Eukaryot Cell **9**: 97–106

Moore RC, Purugganan MD (2005) The evolutionary dynamics of plant duplicate genes. Curr Opin Plant Biol **8**: 122–128

Moore RC, Purugganan MD (2003) The early stages of duplicate gene evolution. Proc Natl Acad Sci **100**: 15682–15687

Mus F, Dubini A, Seibert M, Posewitz MC, Grossman AR (2007) Anaerobic acclimation in *Chlamydomonas reinhardtii*: anoxic gene expression, hydrogenase induction, and metabolic pathways. J Biol Chem **282**: 25475–25486

Mustroph A, Lee SC, Oosumi T, Zanetti ME, Yang H, Ma K, Yaghoubi-Masihi A, Fukao T, Bailey-Serres J (2010) Cross-kingdom comparison of transcriptomic adjustments to low-oxygen stress highlights conserved and plant-specific responses. Plant Physiol **152**: 1484–1500

Nguyen AV, Thomas-Hall SR, Malnoë A, Timmins M, Mussgnug JH, Rupprecht J, Kruse O, Hankamer B, Schenk PM (2008) Transcriptome for photobiological hydrogen production induced by sulfur deprivation in the green alga *Chlamydomonas reinhardtii*. Eukaryot Cell **7**: 1965–1979

Porter ME, Knott JA, Myster SH, Farlow SJ (1996) The Dynein Gene Family in *Chlamydomonas reinhardtii*. Genetics **144**: 569–585

Radakovits R, Jinkerson RE, Fuerstenberg SI, Tae H, Settlage RE, Boore JL, Posewitz MC (2012) Draft genome sequence and genetic transformation of the oleaginous alga *Nannochloropis gaditana*. Nat Commun **3**: 686

Reyes-Prieto A, Weber APM, Bhattacharya D (2007) The Origin and Establishment of the Plastid in Algae and Plants. Annu Rev Genet **41**: 147–168

Rodolfi L, Chini Zittelli G, Bassi N, Padovani G, Biondi N, Bonini G, Tredici MR (2009) Microalgae for oil: strain selection, induction of lipid synthesis and outdoor mass cultivation in a low-cost photobioreactor. Biotechnol Bioeng **102**: 100–112

Round FE (1984) The Ecology of Algae. Cambridge University Press

Schneider JC, Roessler P (1994) Radiolabeling Studies of Lipids and Fatty Acids in *Nannochloropsis* (eustigmatophyceae), an Oleaginous Marine Alga1. J Phycol **30**: 594–598

Simionato D, Sforza E, Corteggiani Carpinelli E, Bertucco A, Giacometti GM, Morosinotto T (2011) Acclimation of *Nannochloropsis gaditana* to different illumination regimes: Effects on lipids accumulation. Bioresour Technol **102**: 6026–6032

Simon DF, Descombes P, Zerges W, Wilkinson KJ (2008) Global expression profiling of *Chlamydomonas reinhardtii* exposed to trace levels of free cadmium. Environ Toxicol Chem SETAC **27**: 1668–1675

Srinivas R, Ochs C (2012) Effect of UV-A Irradiance on Lipid Accumulation in *Nannochloropsis oculata*. Photochem Photobiol **88**: 684–689

Sukenik A, Carmeli Y (1990) Lipid Synthesis and Fatty Acid Composition in *Nannochloropsis* Sp. (eustigmatophyceae) Grown in a Light-Dark Cycle1. J Phycol **26**: 463–469

Tonon T, Harvey D, Larson TR, Graham IA (2002) Long chain polyunsaturated fatty acid production and partitioning to triacylglycerols in four microalgae. Phytochemistry **61**: 15–24

Vogel C, Chothia C (2006) Protein Family Expansions and Biological Complexity. PLoS Comput Biol **2**: e48

Wang Z, Gerstein M, Snyder M (2009a) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet **10**: 57–63

Wang ZT, Ullrich N, Joo S, Waffenschmidt S, Goodenough U (2009b) Algal Lipid Bodies: Stress Induction, Purification, and Biochemical Characterization in Wild-Type and Starchless *Chlamydomonas reinhardtii*. Eukaryot Cell **8**: 1856–1868

Wapinski I, Pfeffer A, Friedman N, Regev A (2007) Natural history and evolutionary principles of gene duplication in fungi. Nature **449**: 54–61

Xu F, Cai Z-L, Cong W, Ouyang F (2004) Growth and fatty acid composition of *Nannochloropsis* sp. grown mixotrophically in fed-batch culture. Biotechnol Lett **26**: 1319–1322

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

CONTRIBUTION

In Miller et al., I conducted sequencing read processing, differential RNA-seq analysis, and global functional analysis using GeneOntology, and contributed to manuscript preparation. In Vieler et al., I contributed to experimental design, conducted hybrid genome assembly, gene annotation, transcript assembly and differential expression analysis, global functional analysis, comparison of *Nannochloropsis* genomes, and manuscript preparation.

CHAPTER 1

CHANGES IN TRANSCRIPT ABUNDANCE IN *CHLAMYDOMONAS REINHARDTII* FOLLOWING NITROGEN-DEPRIVATION PREDICT DIVERSION OF METABOLISM

Note: The contents of this chapter have been published in Plant Physiology.

Rachel Miller*, Guangxi Wu*, Rahul R. Desphande, Astrid Vieler, Katrin Gaertner, Xiaobo Li, Eric R. Moellering, Simone Zäuner, Adam Cornish, Bensheng Liu, Blair Bullard, Barbara B. Sears, Min-Hao Kuo, Eric L. Hegg, Yair Shachar-Hill, Shin-Han Shiu, and Christoph Benning (2010) Changes in Transcript Abundance in *Chlamydomonas reinhardtii* following Nitrogen-deprivation Predict Diversion of Metabolism. Plant Physiol 154:1737-1752.

*These authors contributed equally.

ABSTRACT

Like many microalgae, *Chlamydomonas reinhardtii* forms lipid droplets rich in triacylglycerols when nutrient deprived. To begin studying the mechanisms underlying this process, nitrogen (N) deprivation was used to induce triacylglycerol accumulation and changes in developmental programs such as gametogenesis. Comparative global analysis of transcripts under induced and non-induced conditions was applied as a first approach to studying molecular changes that promote or accompany triacylglycerol accumulation in cells encountering a new nutrient environment. Towards this goal, high-throughput sequencing technology was employed to generate large numbers of expressed sequence tags of eight biologically independent libraries, four for each condition, N replete and N deprived, allowing a statistically sound comparison of expression levels under the two tested conditions. As expected, N deprivation activated a subset of control genes involved in gametogenesis while down-regulating protein biosynthesis. Genes for components of photosynthesis were also down-regulated, with the exception of the *PSBS* gene. N deprivation led to a marked redirection of metabolism: the primary carbon source, acetate, was no longer converted to cell building blocks by the glyoxylate cycle and gluconeogenesis but funneled directly into fatty acid biosynthesis. Additional fatty acids may be produced by membrane remodeling, a process that is suggested by the changes observed in transcript abundance of putative lipase genes. Inferences on metabolism based on transcriptional analysis are indirect, but biochemical experiments supported some of these deductions. The data provided here represent a rich source for the exploration of the mechanism of oil accumulation in microalgae.

INTRODUCTION

The search for sustainable sources of biofuels has led to renewed interest in microalgae as a potential feedstock and, consequently, a flurry of research has recently been initiated in microalgae (Wijffels and Barbosa, 2010). Microalgae accumulate large quantities of oils in the form of triacylglycerols (TAGs) when nutrient deprived, and a thorough analysis of the underlying molecular mechanism is currently in its infancy (Hu et al., 2008). At this time, *Chlamydomonas reinhardtii* is the premier microalgal molecular model for this analysis. As such, the formation of lipid droplets following nitrogen (N) deprivation has recently been documented in detail (Wang et al., 2009; Moellering and Benning, 2010). Although *C. reinhardtii* is not under direct consideration for the production of biomass as a biofuel feedstock, the analysis of its metabolism and physiology is expected to provide basic insights into mechanisms of TAG accumulation relevant to other microalgae at least of the green algal phylum.

The genome of *C. reinhardtii* is available (Merchant et al., 2007), and its annotation is currently at version 4 (http://genome.jgi-psf.org/chlamy/chlamy.home.html). At this time, a number of microarrays have been used to interrogate changes in response to environmental factors (Ledford et al., 2004; Jamers et al., 2006; Ledford et al., 2007; Mus et al., 2007; Nguyen et al., 2008; Simon et al., 2008; Yamano et al., 2008; Mustroph et al., 2010). These microarrays could not cover all genes in the genome, but more recently, massively parallel cDNA sequencing approaches were applied to *C. reinhardtii*, overcoming the shortcomings of microarrays (González-Ballester et al., 2010). Likewise, we have chosen a cDNA sequencing-based approach using 454 and Illumina technologies in parallel that allow the generation of large numbers of ESTs of varying abundance, which can be counted to obtain a measure of gene expression (Weber et al., 2007).

The goal of this study was to determine major changes in gene expression following N deprivation, the nutrient condition established in our previous analysis of lipid droplet formation and TAG accumulation in *C. reinhardtii* (Moellering and Benning, 2010). Comparison of the transcript levels of induced, N-deprived *C. reinhardtii* cultures with those of un-induced, N-replete cultures was expected to reflect the metabolic changes leading to TAG accumulation. Of course, making inferences about metabolism based on gene expression levels has its caveats, as gene expression does not necessarily directly translate into metabolic fluxes. To interrogate the meaningfulness of some of the transcript-level changes we observed with regard to metabolism, we also performed labeling experiments using acetate as the precursor. Acetate is a typical carbon source provided to *C. reinhardtii* for photoheterotrophic growth enabling short doubling times, and it is readily incorporated into fatty acids, the main constituents of TAGs. Keeping in mind that these are clearly conditions optimized for an experimental laboratory system, we nevertheless expect to be able to make basic inferences that will be relevant to a broader understanding of the induction of TAG biosynthesis and lipid droplet accumulation in green algae.

RESULTS

Defining Conditions for N Deprivation of *C. reinhardtii*

Ideally, one would like to use finely spaced time-course experiments to distinguish rapid versus long-term changes in gene expression following N deprivation. However, because our resources were limited, we decided to focus on two conditions, N replete and N deprived. Independent biological replicates allowed for statistically sound interpretations of the data. To determine the time point for N deprivation most likely to provide an accurate snapshot of readjustment of transcript steady-state levels following N deprivation, we first used northern-blot hybridization to compare the expression of genes known or expected to be regulated in *C. reinhardtii* following N deprivation. An ammonium transporter, AMT4, which has been previously shown to be activated by N deprivation (Mamedov et al., 2005), and two putative diacylglycerol acyltransferases, tentatively designated DGTT2 and DGTT3 (protein identifiers 184281 and 400751, genome version 4), were monitored to test the various conditions. RNA was isolated from cells grown in standard (10 mm NH_4^+) and low-N (0.5 mm NH_4^+) Tris-acetate phosphate (TAP) medium to impose N deprivation as well as from cells grown to mid-log phase in standard TAP medium, then transferred to either standard or no-N (0 mm NH_4^+) TAP, with samples taken at 24 and 48 h, to accomplish more drastic N deprivation (Figure 1.1A). For standardization, equal amounts of RNA were loaded and the 18S rRNA abundance was examined. Although *C. reinhardtii* ribosomes turn over following N deprivation (Siersma and Chiang, 1971; Martin et al., 1976), their abundance drops no lower than 50% (see below). *AMT4* mRNA was absent from the un-induced cells and present at a high level in N-limited or N-deprived cells under the conditions tested. *DGTT2* mRNA was present at low levels in all conditions. *DGTT3* mRNA was present at low levels and increased slightly following N deprivation. N deprivation for 48 h showed the greatest difference in RNA levels compared with the N-replete cultures. Based on this basic analysis and our previous time-course study of lipid droplet formation and TAG accumulation (Moellering and Benning, 2010), a 48-h period of N deprivation was chosen to compare global transcript levels in N-replete and N-deprived cells.

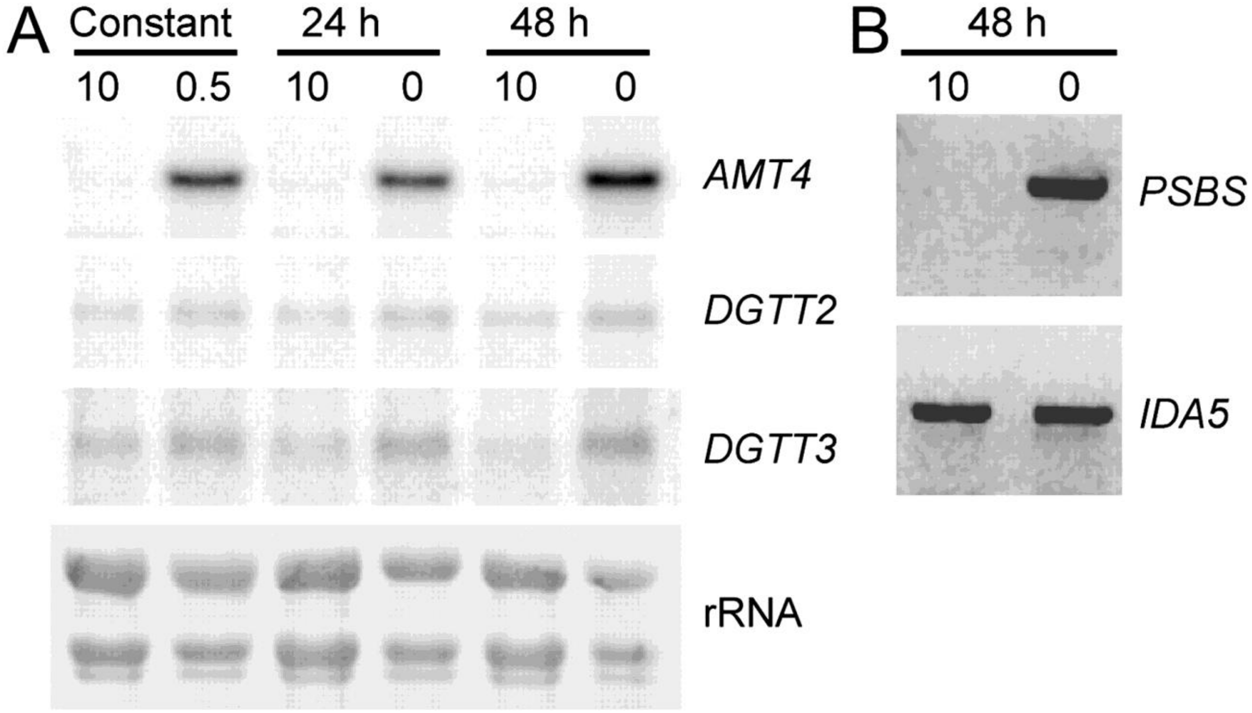


Figure 1.1. Transcript levels of specific genes. Cultures were grown in TAP medium that was N replete (10 mm NH_4^+; 10), continual N limited (0.5 mm NH_4^+; 0.5), or N deprived (0 mm NH_4^+; 0) for 24 or 48 h. The expression levels of *AMT4*, *DGTT2*, and *DGTT3* were measured by RNA-DNA hybridization, and rRNA was visualized as a loading control. B, Cultures were grown for 48 h in either N-replete or N-deprived conditions. The levels of *PSBS1* and *PSBS2* transcripts were measured using RT-PCR, and the constitutive *IDA5* gene served as a control.

Global Characteristics of *C. reinhardtii* Gene Expression following N Deprivation

To determine the differential expression of genes in *C. reinhardtii* under N-replete and N-deprived conditions, two sequencing approaches, 454 and Illumina, were applied. Read length is longer for 454, but the number of reads per run is lower. As shown in Table 1.1, 60- to 85-fold more sequence tags were generated with Illumina than with 454 sequencing. Among the 454 reads, 78% to 80% mapped to the *C. reinhardtii* genome. For the Illumina data, we mapped in three different ways, with varying stringency depending on whether 3′-end read quality and exon-spanning reads were considered. Without filtering reads, a substantially smaller proportion of Illumina reads (63%–68%) were mapped compared with 454 reads. Trimming low-quality 3′ regions of reads resulted in a further 2.7% decrease in the number of mapped reads. Despite the large number of unmapped Illumina reads, out of 16,710 *C. reinhardtii* gene models, 15,505 (92%) had one or more reads. In contrast, only 6,372 gene models (38.1%) were supported by the 454 transcriptome data set. In addition, nearly all genes covered by 454 were also covered by Illumina. Therefore, our sequencing data covered most annotated genes, enabling us to interrogate differential expression under normal conditions and following N deprivation. In addition, as expected, Illumina data provided a better coverage of the gene space than 454 sequences.

Table 1.1 Summary of expression tags generated using two different sequencing methods.

Sequencing methods	**454**		**Illumina**					
Treatment[a]	R	D	R1	R2	R3	D1	D2	D3
Total[b]	2.51E5	2.15E5	1.69E7	1.83E7	1.78E7	1.77E7	1.79E7	1.52E7
Mapped[c]	2.01E5	1.68E5	1.09E7	1.24E7	1.15E7	1.07E7	1.13E7	9.86E6
Genic[d]	1.08E5	1.32E5	8.96E6	9.43E6	1.02E7	8.33E6	8.82E6	7.81E6
Intergenic[e]	9.30E4	3.61E4	1.93E6	2.97E6	1.31E6	2.38E6	2.48E6	2.05E6

[a] Treatment types: (R) N-replete and (D) N-deprived media. For Illumina sequencing, there are three replicates for each treatment.

[b] Number of sequencing reads after filtering out low quality reads based on 454 and Illumina base-calling methods.

[c] Number of sequencing reads after mapping to *C. reinhardtii* v4.0 genome

[d] Number of mapped sequencing reads overlapping with *C. reinhardtii* v4.0 filtered gene models

[e] Number of mapped sequencing reads not overlapping with any *C. reinhardtii* v4.0 filtered gene models

To determine differential gene expression following N deprivation, we modeled count data with a moderated negative binomial distribution (see “Materials and Methods”). Using thresholds of 5% or less false discovery rates and 2-fold or greater change for the Illumina data set, 2,128 and 1,875 genes were categorized as up- and down-regulated, respectively, following N deprivation. To see if fold changes inferred based on 454 and Illumina data sets were consistent, we determined the statistical correlation in fold change between these two data sets (Figure 1.2) and found that it was rather weak (Pearson’s correlation coefficient, $r^2 = 0.10$, $P < 2.2 \times 10^{-16}$). There was an apparent anomaly, as 4,313 genes (out of 6,369 genes with one or more reads from both data sets) had a high degree of up- and down-regulation, which was observed with the 454 but not the Illumina data set (Figure 1.2A). Most of these genes with extreme responses based on 454 had very low counts (less than 10 reads combined in both conditions or zero read in one of the conditions; Figure 1.2A, red data points). As a result, high and likely inaccurate fold change values were assigned to those genes. In fact, if we only considered 2,056 genes with 10 or more reads combined and one or more reads in both conditions, the correlation between Illumina and 454 data was substantially improved ($r^2 = 0.57$, $P < 2.2 \times 10^{-16}$; Figure 1.2A, blue data points).

One important consideration in identifying differentially regulated genes is that there is a considerable transcript length bias in Illumina data. A longer transcript tends to have more reads than a shorter transcript expressed at the same level (Oshlack and Wakefield, 2009; Bullard et al., 2010). Consistent with earlier findings, we observed a significant correlation between the number of reads assigned to a protein sequence and its length (Spearman’s rank $\rho = 0.33$, $P <$ 2.2e-16; Figure 1.2B). Because of this length bias, longer transcripts may have more significant differences in differential gene expression studies and in some cases will lead to false-positive differential expression calls (Oshlack and Wakefield, 2009; Bullard et al., 2010). However, unlike previously published studies, we did not find a significant correlation between the percentage of genes differentially expressed and sequence length ($\rho = 0.09$, $P = 0.7$; Figure 1.2C). This finding indicates that, although length bias remains an issue, our differential expression call may not be as significantly affected as reported previously.

When using Tophat 1.0.8 for read mapping, 64% - 73% of all reads are mapped to the genome. When using Tophat 2.0.0 to re-run read mapping, 70%- 78% of all reads are mapped to the genome. When comparing the differential expression calling between the previous results and the results based on the second Tophat run (Figure 1.2D), a high correlation was observed among differentially expressed genes (Pearson's correlation coefficient, $r^2 = 0.89$, $P < 2.2 \times 10^{-16}$), suggests that the differential expression calling results from two mapping runs are highly similar.

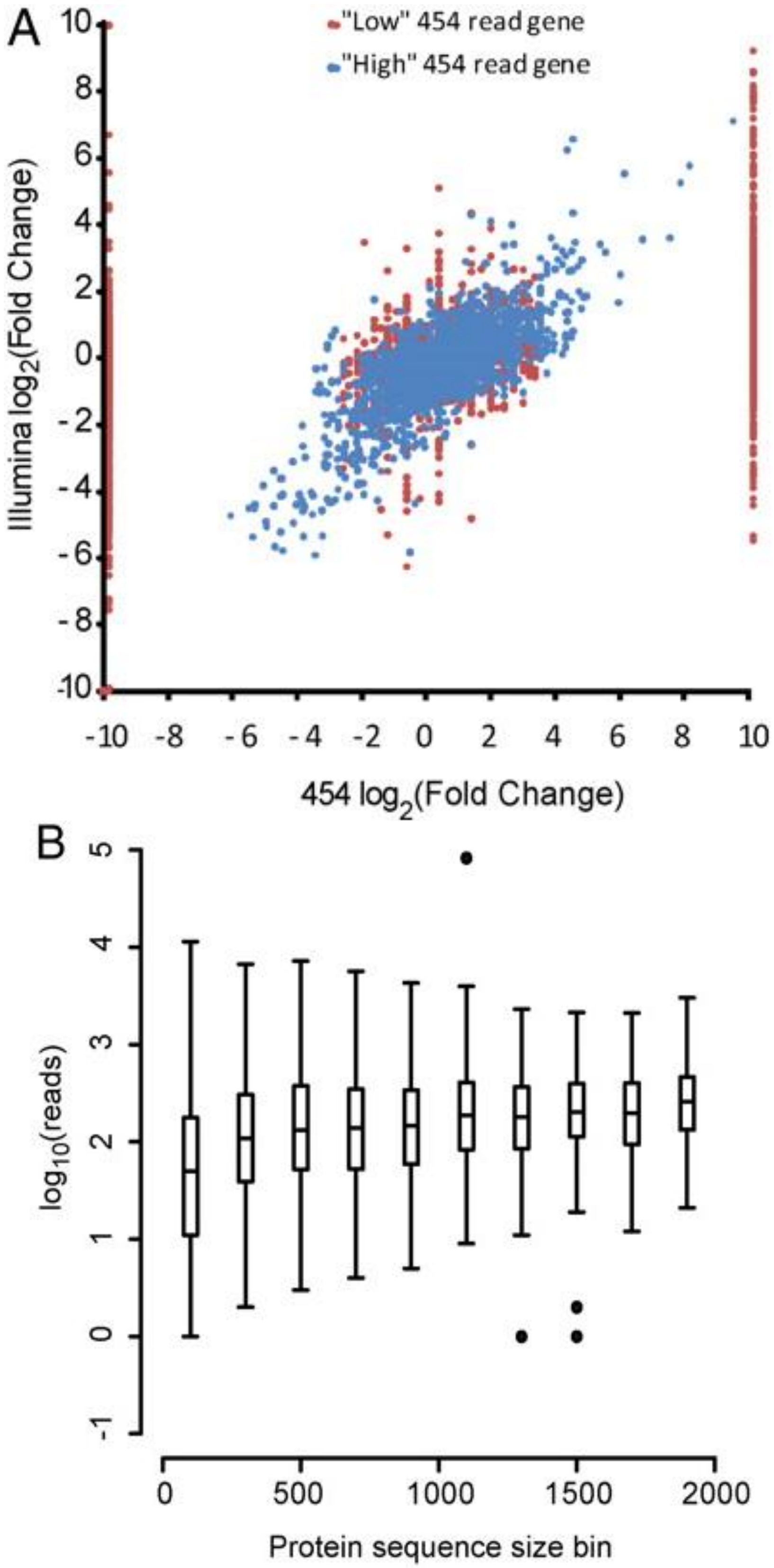


Figure 1.2. Fold change correlation between Illumina and 454 data sets and impacts of Illumina length bias on differential expression call.

Figure 1.2 (cont’d)

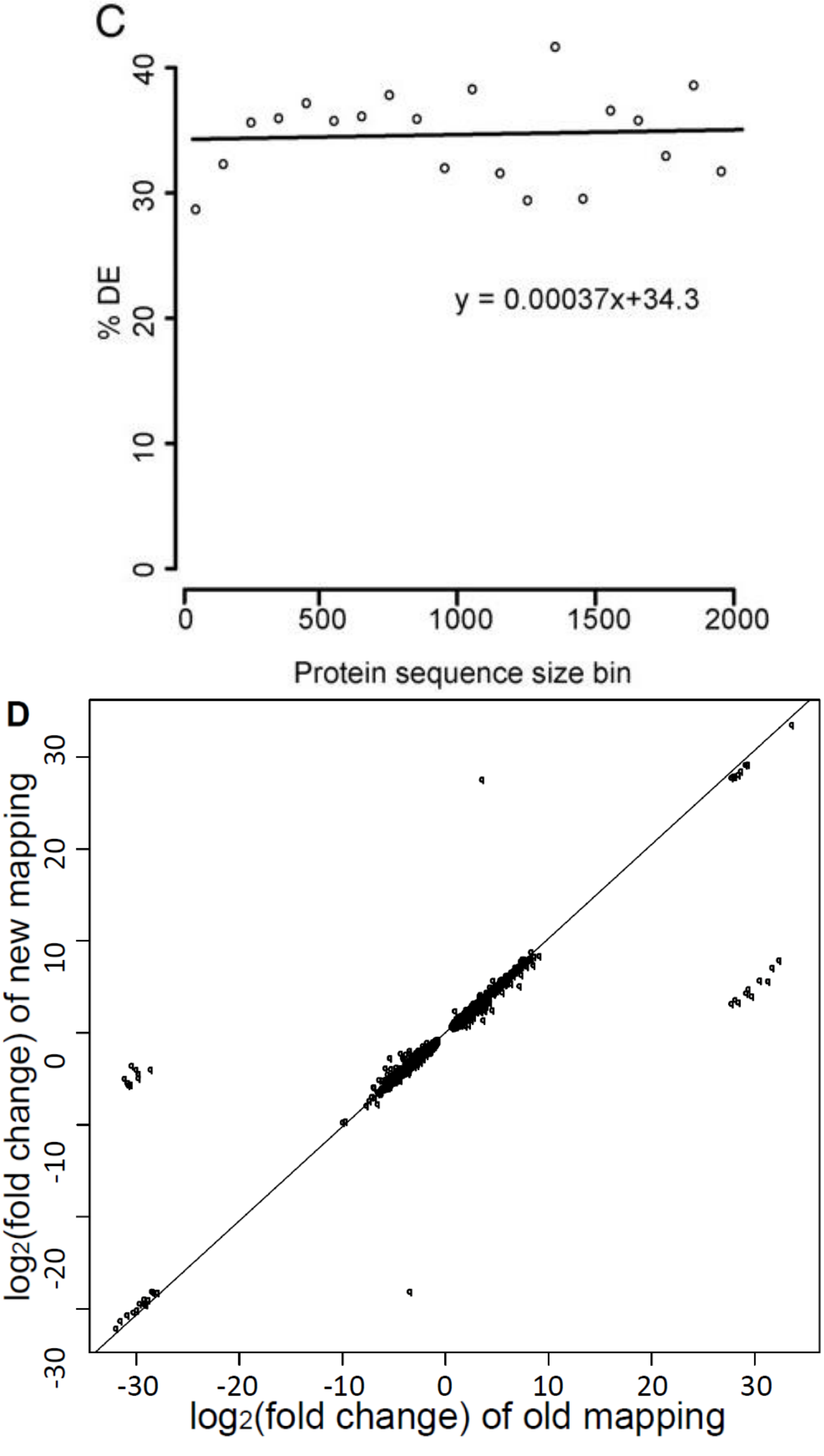


Figure 1.2. (cont’d)

A, Only genes with one or more 454 and Illumina reads under either N-replete (+N) or N-deprived (−N) conditions were plotted. Fold change is determined by the number of reads following N deprivation divided by the number of reads under N-replete conditions for each gene. For genes with 2^{10}-fold or greater or 2^{-10}-fold or lesser changes, the fold change values were set to 10. Blue circles (“high” 454 read genes) indicate genes with 10 or greater 454 reads (+N and –N combined) and one or more 454 reads in both +N and –N. Red circles (“low” 454 read genes) indicate genes that did not satisfy one or both of the above criteria.

B, Each box plot depicts the numbers of reads for protein-coding genes (log base 10) in a protein sequence size bin (0–2,000 amino acids, bin size of 100 amino acids). All proteins of 2,000 or more amino acids are classified as 2,000 amino acids. Outliers are shown in black circles.

C, Percentage of genes that are regarded as differentially expressed (DE) in each protein sequence size bin. The line indicates the linear fit, and the equation for the line is shown as well.

D, Fold change of Illumina new mapping and old mapping, only genes that are differentially expressed are included. For interpretation of the references to color in this and all other figures, the reader is referred to the electronic version of this dissertation.

Approximately 7% to 14% of the Illumina reads mapped to the “intergenic regions” (Table 1.1). We assembled Illumina reads into 42,574 transcribed fragments (transfrags). Among them, 17,095 transfrags did not map with, or within the vicinity of (1,855 bases, 99th percentile intron length), current gene models. With the same conservative criterion, transfrags were joined into 1,828 “intergenic transcriptional units.” Most importantly, 287 of these intergenic transcriptional units were up-regulated and 176 were down-regulated following N deprivation. These transfrags are un-annotated genes that require further analysis to establish their authenticity.

Gene Ontology (GO) annotation was used to coarsely identify major categories of genes involved in particular biological processes to assess trends in their transcriptional regulation following N deprivation. We found multiple GO categories with significant enrichment in their numbers of differentially regulated genes (Table 1.2). Particularly, genes associated with lipid metabolism tend to be up-regulated, while those involved in photosynthesis and DNA replication initiation tend to be down-regulated.

Table 1.2. Gene Ontology categories significantly enriched in differentially regulated *C. reinhardtii* genes

GO	Annotation	GO R[a]	No GO R[b]	GO U[c]	No GO U[d]	Reg[e]	*p*[f]	*q*[g]
GO:0006270 bp[h]	DNA replication initiation	9	926	0	6633	down	6.48E-9	4.51E-6
GO:0015995 bp	chlorophyll biosynthetic process	8	927	0	6633	down	5.29E-8	2.76E-5
GO:0033014 bp	tetrapyrrole biosynthetic process	6	929	0	6633	down	3.51E-6	7.32E-4
GO:0009765 bp	photosynthesis, light harvesting	22	913	14	6619	down	5.72E-12	5.98E-9
GO:0015979 bp	photosynthesis	23	912	0	6633	down	1.02E-21	2.13E-18
GO:0005576 cc[i]	extracellular region	11	924	14	6619	down	8.28E-5	1.44E-2
GO:0009522 cc	photosystem I	4	931	0	6633	down	2.32E-4	3.23E-2
GO:0009538 cc	photosystem I reaction center	5	930	0	6633	down	2.85E-5	5.42E-3
GO:0009654 cc	oxygen evolving complex	6	929	0	6633	down	3.51E-6	7.32E-4
GO:0019898 cc	extrinsic to membrane	4	931	0	6633	down	2.32E-4	3.23E-2
GO:0003755 mf[j]	peptidyl-prolyl cis-trans isomerase activity	21	914	32	6601	down	4.39E-7	1.83E-4
GO:0004600 mf	cyclophilin	19	916	30	6603	down	2.29E-6	5.98E-4
GO:0016851 mf	magnesium chelatase activity	4	931	0	6633	down	2.32E-4	3.23E-2
GO:0030051 mf	FK506-sensitive peptidyl-prolyl cis-trans isomerase	19	916	30	6603	down	2.29E-6	5.98E-4
GO:0042027 mf	cyclophilin-type peptidyl-prolyl cis-trans isomerase activity	19	916	30	6603	down	2.29E-6	5.98E-4
GO:0006006 bp	glucose metabolic process	5	944	1	6618	up	1.65E-4	3.83E-2
GO:0006468 bp	protein amino acid phosphorylation	117	832	501	6118	up	2.37E-6	2.47E-3

Table 1.2 (cont'd)

GO:0006629 bp	lipid metabolic process	18	931	38	6581	up	9.92E-5	2.59E-2
GO:0004672 mf	protein kinase activity	113	836	479	6140	up	2.08E-6	2.47E-3
GO:0004713 mf	protein-tyrosine kinase activity	85	864	342	6277	up	8.04E-6	4.20E-3
GO:0004674 mf	protein serine/threonine kinase activity	92	857	391	6228	up	2.50E-5	7.46E-3

[a] GO R indicates number of significantly up or down (R) regulated genes with the GO annotation in question. [b] No GO R, number of significantly up or down regulated genes without the GO annotation. [c] GO U, number of genes without significant expression change with the GO annotation . [d] No GO U, number of genes with no significant expression change that do not have the GO annotation. [e] Reg, direction of regulation (nitrogen deprived compared to nitrogen replete). [f] Fisher's exact test *p* value. [g] *q* value is calculated using R package qvalue. [h] bp, biological process. [i] cc, cellular component. [j] mf, molecular function.

Induction of Gametogenesis and Sexual Reproduction

Because N deprivation triggers gametogenesis (Martin and Goodenough, 1975; Kurvari et al., 1998), we examined several genes known to be involved in mating-type plus (*mt*$^+$) gamete differentiation or sexual fusion in *C. reinhardtii* as internal controls for the induction state of the cells following N deprivation. Following N deprivation, cells had substantial increases in the abundance of transcripts of four of the six genes considered. These genes encode FUS1, which is a glycoprotein required for sex recognition, SAG1 (the *mt*$^+$agglutinin gene), peptidase M gametolysin, which releases the gametes from the cell wall, and NSG13, which is a protein of unidentified function known to be expressed in gametes, as summarized by (Harris, 2009). A second gametolysin gene (encoding peptidase M11) and *GSP1*, which encodes a gamete-specific transcription factor, did not show increased expression following N deprivation, but perhaps that is because only a single time point was examined.

Effects on Genes of N Metabolism and Protein Biosynthesis

Many genes involved in N import and assimilation are known to be induced following N deprivation (Schnell and Lefebvre, 1993; González-Ballester et al., 2004; Harris and Stern, 2009). Our analysis revealed greater than 2-fold up-regulation for several genes, including those that encode NO_3^- NO_2^- transporters and reductases, as well as transport systems for NH_4^+ and organic N sources. Of the genes involved in assimilation of NH_4^+ by the Gln synthetase-Glu synthase cycle, only *GLN3* was up-regulated. Similarly, most genes involved in amino acid biosynthesis did not show a greater than 2-fold change. Thus, transcript abundance suggests that following N deprivation, pathways for the acquisition of new N sources are strongly up-regulated, whereas biosynthetic pathways that utilize the assimilated N remain relatively unaffected.

Decades ago, N deprivation of *C. reinhardtii* was found to result in degradation and resynthesis of both cytoplasmic and chloroplast ribosomes (Siersma and Chiang, 1971; Martin et al., 1976). Both the rRNA and proteins of the ribosomes were turned over under the conditions of N deprivation that also induce gamete differentiation. Hence, we expected that the mRNAs for the ribosomal proteins might show different steady-state levels in the comparison of logarithmically growing cells and cells that have been N deprived for 48 h. Indeed, following N deprivation, the abundance of transcripts encoding proteins of the chloroplast ribosomes consistently decreased to 30% to 50% of their levels of expression in logarithmically growing cells.

A subset of the cytosolic 80S ribosomal protein genes has been identified in the version 4.0 genome data set. Among those that have been annotated, most are encoded by single-copy genes, although a few have two copies (e.g. *L7*, *L10*, *L13*, and *L23*). As these gene products are assembled into ribosomes, the respective genes have high levels of constitutive expression. The abundance of the transcripts in vegetative cells and N-deprived cells was fairly similar.

The RPL22 ribosomal protein of the cytosolic ribosomes is encoded by a multigene family in *C. reinhardtii*. Of the 37 *RPL22* genes in version 4.0 of the genome data set, 13 appeared not to be expressed under either condition tested and six had barely detectable levels of transcripts. Of the 18 remaining genes, two gave rise to the most predominant transcripts, and their transcripts did not change markedly in abundance. Four of the 18 genes were moderately expressed, and their transcript levels doubled in the N-deprived cells. Six of the 18 genes had markedly lower levels of transcripts following N deprivation, while six others showed relatively constant levels of expression. The *RPL22* genes are scattered among at least six chromosomes, and no correlation was found between location and level of gene expression.

General Changes in Primary Metabolism

Changes in transcript abundance of genes encoding enzymes of primary metabolism are depicted in Figure 1.3. Transcripts encoding key enzymes of the glyoxylate cycle, gluconeogenesis, and the photosynthetic carbon fixation cycle markedly decrease following N deprivation. Transcript abundance for the glyoxylate cycle enzymes isocitrate lyase and malate synthase decreased more than 16-fold. In addition, mRNA abundance of the cytosolic (predicted) phosphoenolpyruvate carboxykinase, which catalyzes the committed reaction of gluconeogenesis, dropped to 25% of the levels in N-replete cells, as did transcripts encoding enzymes involved in carbon fixation and reduction, ribulose-bisphosphate carboxylase, sedoheptulose 1,7-bisphosphate aldolase, and sedoheptulose-bisphosphatase. In contrast, there was a considerable increase in the transcript abundance of the cytosolic enzyme pyruvate phosphate dikinase. This is a key enzyme in the C4 photosynthetic pathway and is regulated by light. It has also been associated with suppressed phospho*enol*pyruvate carboxykinase activity (Magne et al., 1997) and salt stress (Fisslthaler et al., 1995). Recently, this enzyme has been shown to play an important role in N remobilization (Taylor et al., 2010). Likewise, the transcript abundances for enzymes of the pentose phosphate cycle predicted to be localized in the cytosol, Glc-6-P 1-dehydrogenase and phosphogluconate dehydrogenase (decarboxylating), were increased under those conditions. The mRNA encoding for one of the pyruvate decarboxylase subunits represented in the data set was also increased in abundance following N deprivation. The pyruvate decarboxylase complex converts pyruvate to acetyl-CoA, which is a precursor of fatty acid biosynthesis. Genes for other enzymes of the glycolytic pathway, such as pyruvate kinase, did not show very drastic changes in response to the N deprivation.

To verify whether the changes observed in RNA abundance actually reflect changes in the activity of glyoxylate and gluconeogenic pathways, cells were grown in the presence of [U-^{13}C] acetate. As the cells take up acetate as a carbon source, the distribution of the ^{13}C in the cellular metabolites gives an insight into the activity of the pathways leading to them. The intracellular amino acids as well as the sugar units of the carbohydrates and RNAs were analyzed with gas chromatography-mass spectrometry (GC-MS) as described in “Materials and Methods” (Figure 1.4). The natural abundance refers to the naturally occurring distribution of ^{13}C in the molecule. The mass isotopomers M0, M1, M2, etc. refer to molecules having, respectively, zero, one, two, etc. atoms of ^{13}C.

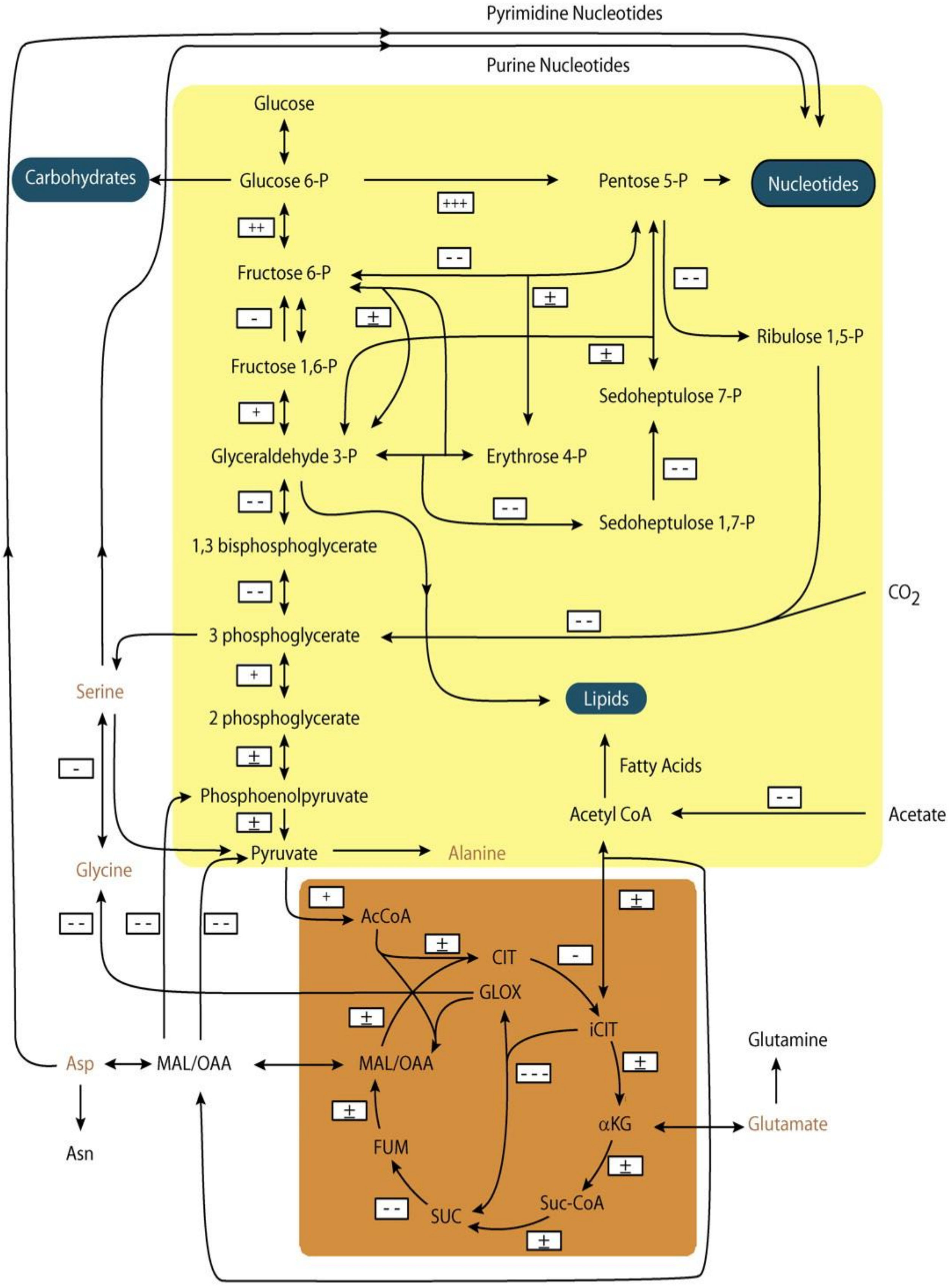


Figure 1.3. Regulation of genes involved in primary metabolism.

Figure 1.3. (cont'd)

The figure indicates the central metabolic pathways of *C. reinhardtii* and gives the differential regulation of gene expression following N deprivation. Symbols represent $\log_2$ fold change as follows: +++, greater than 5; ++, greater than 2 and less than 5; +, greater than 1; ±, less than 1 and greater than −1; −, less than −1; −−, less than −2 and greater than −5; −−−, less than −5.

Cells grown in N-replete medium showed a higher degree of labeling in Ser and Gly than did N-deprived cells (Figure 1.4). The fully labeled fraction (M3) accounted for almost 80% of the total Ser in the cells grown in N-replete medium. Hence, most of the Ser was derived from the gluconeogenic pathway, which incorporates the labeling of acetate into the glycolytic intermediate 3-phosphoglycerate, a precursor of Ser. There was also about 10% of M2 Ser. This probably derived from the reaction catalyzed by the reversible Ser hydroxylmethyltransferase, which favors the production of Ser from Gly (Mattingly et al., 1976). The Gly in this reaction would mostly be fully labeled (M2; Fig. 4), largely from the glyoxylate cycle, hence giving rise to M2 Ser. In cells grown in N-deprived medium, we observed a markedly lower incorporation of the ^{13}C atoms into the amino acids. Almost 80% of Gly was unlabeled (M0), indicating a very low activity of the glyoxylate cycle. Similarly, N-deprived cells had reduced label in carbohydrates and Rib. Since these molecules were formed essentially by the gluconeogenic pathway during growth in the medium employed, the N-deprived cells appeared to have much lower gluconeogenic activity. Thus, these biochemical data corroborated the transcript abundance data (Figure 1.3) that suggested a down-regulation of the glyoxylate and gluconeogenic pathways in N-deprived cells.

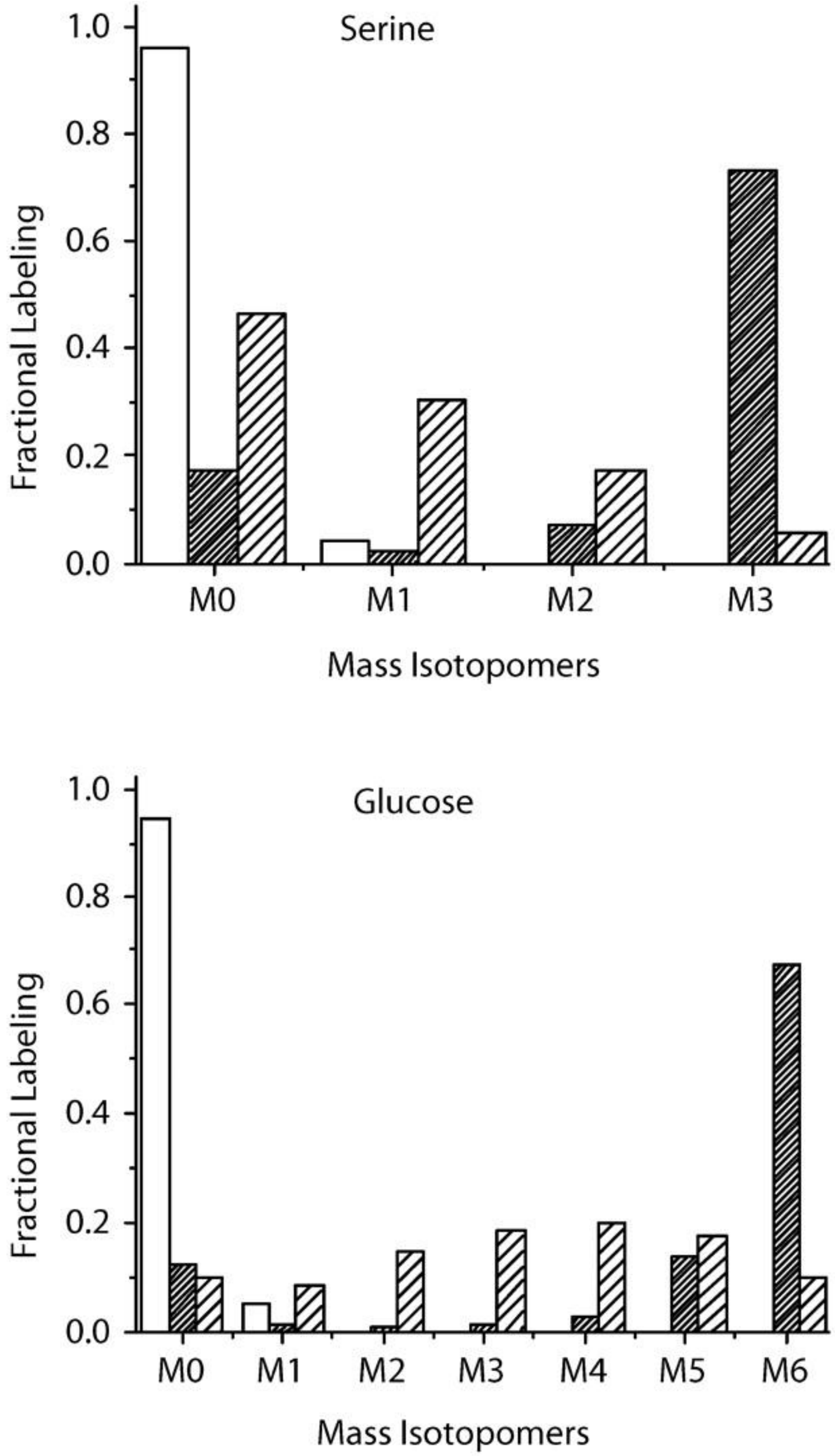
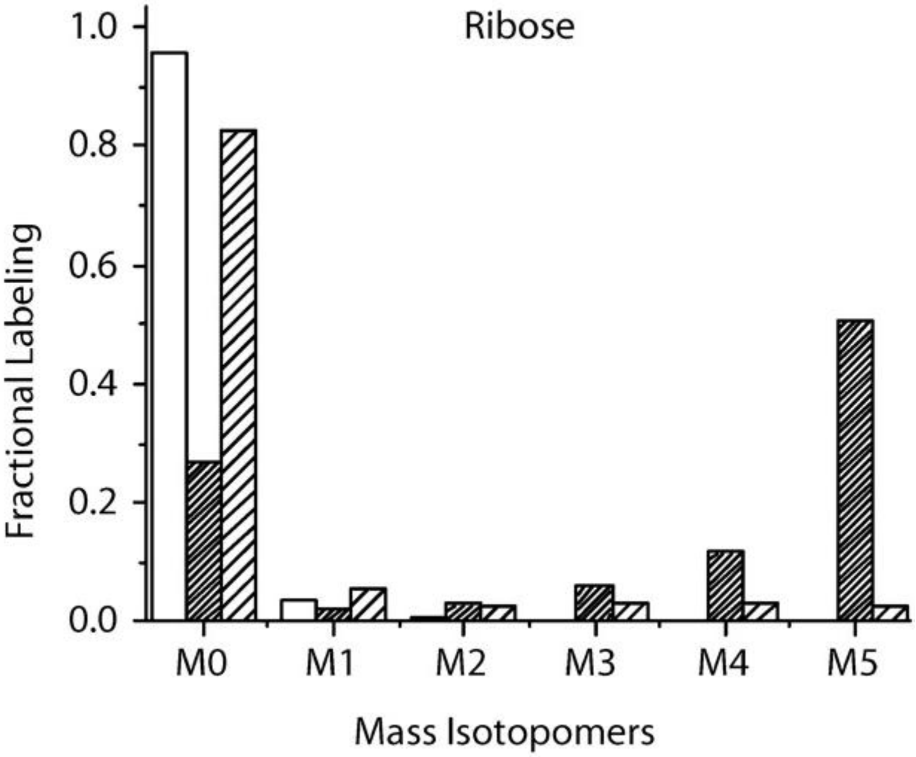


Figure 1.4. Changes in labeling patterns reflecting changes in gene expression.

Figure 1.4 (cont'd)



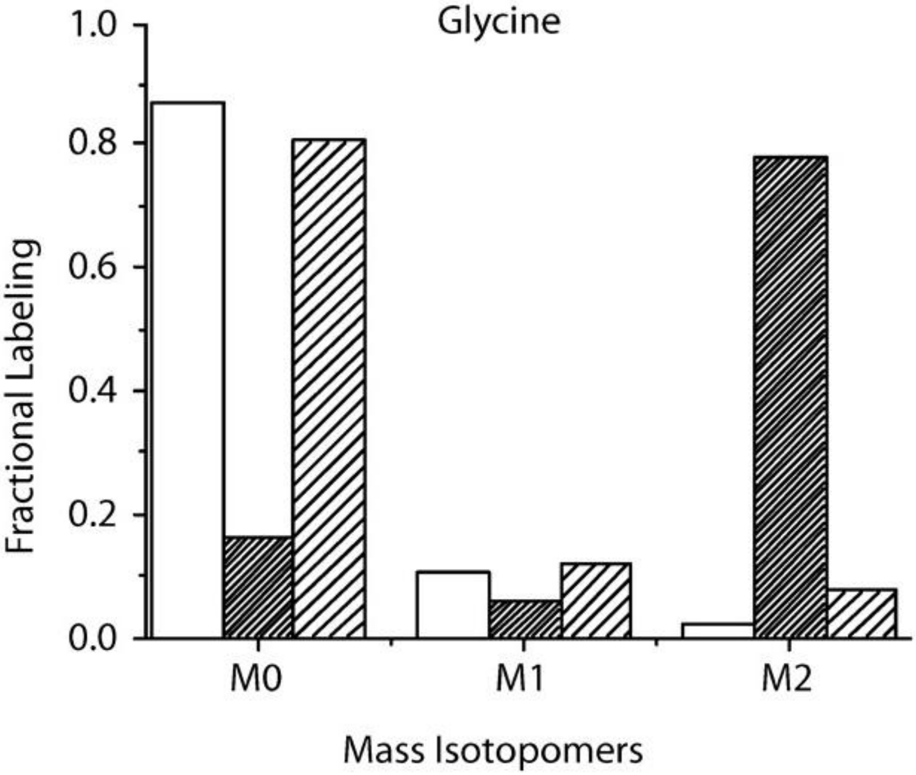


Figure 1.4 (cont'd)

Labeling of metabolites after 24 h of N deprivation compared with N-replete cells is shown. The metabolites give an indication of the pathway activity. Intracellular Ser and Gly were extracted by quick quenching and extraction with cold methanol. Rib and Glc were obtained from the acid hydrolysis of RNA and cellular polysaccharides, respectively. The labeling was analyzed by GC-MS after derivatization. Natural labeling refers to the naturally occurring distribution of ^{13}C in the molecule. The mass isotopomers M0, M1, M2, etc. refer to molecules having, respectively, zero, one, two, etc. atoms of ^{13}C. White bars, Natural labeling; fine cross-hatched bars, N-replete cells; coarse cross-hatched bars, N-deprived cells.

No appreciable change in transcripts for genes encoding components of the mitochondrial respiratory pathway was noted following N deprivation. However, an 11-fold increase in the transcript abundance of an alternative oxidase gene (*AOX1*) was observed, while the *AOX2* transcript was down-regulated 4-fold. These findings were consistent with previous observations of changes in gene expression of *AOX1* and *AOX2* (Baurain et al., 2003).

The candidate genes for peroxisomal β-oxidation showed an overall decrease in their transcript levels following N deprivation, with acyl-CoA oxidase and 3-oxoacyl-CoA thiolase (*ATO1*) transcript abundance decreasing most drastically (greater than 3-fold). The only exception was an enoyl-CoA oxidase/isomerase candidate gene (*ECH1*), which showed increased transcript levels (greater than 2-fold). An apparent down-regulation of fatty acid oxidation is in line with the accumulation of TAGs under these conditions.

Reduced Transcript Abundance for Most Photosynthetic Genes

In *C. reinhardtii*, photosynthetic efficiency decreases following N deprivation, at least partially due to a reduction in the abundance of light-harvesting complexes (Plumley and Schmidt, 1989; Peltier and Schmidt, 1991) and selective degradation of the cytochrome b_6f complex (Bulté and Wollman, 1992; Majeran et al., 2000). Likewise, the abundance of transcripts encoding photosynthesis-related proteins was substantially reduced following N deprivation. This regulation was not restricted to light-harvesting complexes and cytochromes but extended to the two photosystems as well. Following N deprivation, the steady-state level of all nucleus-encoded PSI genes decreased by at least 6-fold, while the abundance of transcripts from genes encoding the corresponding light-harvesting proteins was decreased even further, resulting in a 19- to 43-fold decrease relative to N-replete conditions. Only four of the cytochrome subunits are encoded by the nuclear genome, and three of them showed a considerable down-regulation (6-fold) following N deprivation. In contrast, the transcript levels of *PETO* were weakly increased (2-fold). This observation supports the hypothesis that this protein may have a regulatory role as opposed to being a functional cytochrome b_6f subunit (Hamel et al., 2000), because the PETO protein is only loosely bound to the complex and its function is not required for the oxidoreductase activity. Expression of all nuclear genes encoding PSII components also decreased following N deprivation, although the two least abundant transcripts decreased only slightly. The PSII light-harvesting complex encoding transcripts showed a comparable change in abundance. Most of the transcript levels were reduced, while the weakly expressed *LHCB7* gene showed no alteration in transcript levels.

The only two genes of the light-harvesting complex of PSII not following that pattern were *PSBS1* and *PSBS2*. Their transcript levels were strongly increased following N deprivation (119- and 52-fold, respectively). This result was confirmed by reverse transcription (RT)-PCR (Figreu 1.1B).

Specific Changes in Gene Expression Related to General Lipid Metabolism

N deprivation has been demonstrated to lead to the accumulation of TAG in specialized organelles as well as to structural changes and breakdown of the intracellular membrane systems such as the thylakoids and the endoplasmic reticulum (ER; Martin et al., 1976; Moellering and Benning, 2010). Therefore, we expected this to be reflected in the expression of genes encoding enzymes of lipid metabolic pathways. However, changes in transcript levels of genes encoding fatty acid metabolism were modest (Figure 1.5). A 2-fold increase in transcript levels for ketoacyl-acyl carrier protein (ACP) synthetase was observed. This enzyme is part of the fatty acid synthase II complex that catalyzes the acyl-ACP-dependent elongation steps from C4 to C14 in higher plants. The gene for acyl-ACP thioesterase (FAT1) also showed elevated transcript levels following N deprivation (about 4-fold). Its reaction terminates fatty acid synthesis by cleaving the acyl chain from ACP. This reaction competes with the direct transacylation of ACP by glycerol-3-phosphate acyltransferases for the formation of phosphatidate. An increase in FAT1 activity, therefore, could be indicative of increased fatty acid export from the chloroplast to the ER, where TAG assembly occurs, as acyl-ACPs have to be hydrolyzed prior to export (Pollard and Ohlrogge, 1999).

A strong increase in transcript levels was observed for the gene encoding the committing step of TAG synthesis. Out of the five putative diacylglycerol acyltransferases genes identified in the version 4.0 genome data set, only four were expressed under either or both growth conditions. One of these genes (*DGTT1*, PID 285889) was almost completely suppressed under N-replete conditions but showed a large increase in transcript abundance following N deprivation. However, its overall transcript abundance was too low to be detected by northern blot compared with other genes encoding putative diacylglycerol acyltransferases, which were much less differentially expressed, consistent with the initial RNA-DNA hybridization analysis (Figure 1.1A).

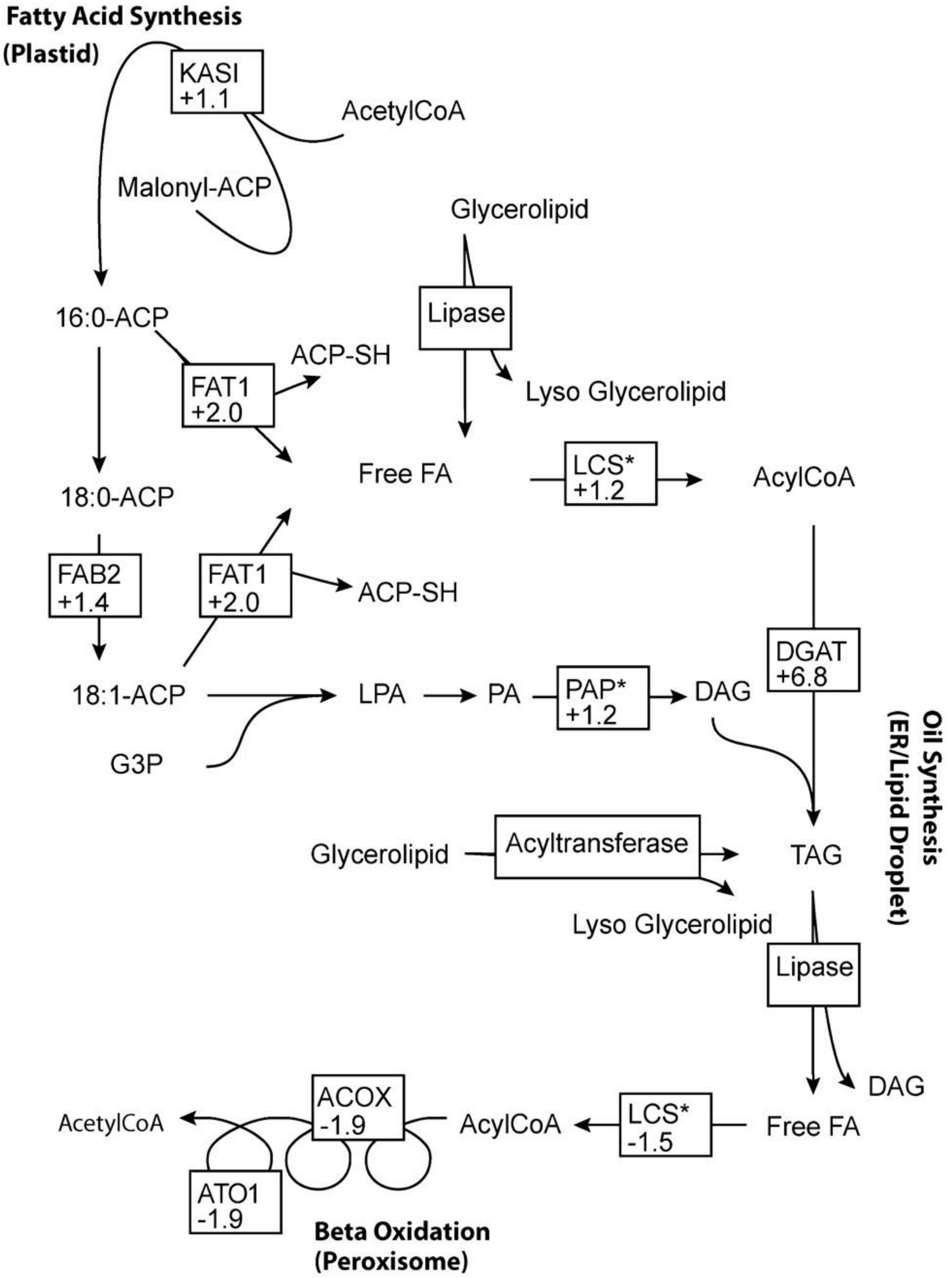


Figure 1.5. Selected changes in glycerolipid metabolism transcript abundance.

Figure 1.5 (cont’d)

Numbers indicate $\log_2$ fold change of transcript abundance following N deprivation. Enzymes labeled with an asterisk cannot be unequivocally assigned to a specific step in the metabolic pathway and are hypothetical.

Phosphatidic acid phosphatase takes part in the Kennedy pathway of glycerolipid and TAG synthesis (Figure 1.5). Both of the two candidate genes for phosphatidic acid phosphatase in *C. reinhardtii* annotated in the version 4.0 data set showed increased transcript levels following N deprivation. Both are part of the PAP2 family, which is thought to have a broad substrate specificity (Carman and Han, 2006). The increase in the expression of the presumed phosphatidic acid phosphatase genes is consistent with the notion that DAG is generated from phosphatidic acid for further TAG biosynthesis following N deprivation.

Out of a total of 16 putative membrane-bound desaturase- and hydroxylase-encoding genes found in *C. reinhardtii*, only three showed a change in transcript abundance that is greater than 2-fold. Transcript abundance for microsomal Δ12-desaturase was more than 3-fold higher following N deprivation, as was that for the plastidic acyl-ACP-Δ9-desaturase, which introduces the first double bond in an acyl chain. Other microsomal desaturase-encoding transcripts, such as that encoding FAD13, an ω13/Δ5-desaturase, were also slightly increased in abundance, whereas the plastid desaturase-encoding genes were not affected.

Of all lipid-related genes, those encoding putative lipases showed the strongest differences in transcript abundance between the two conditions tested. By searching for “lipase,” “phospholipase,” or “patatin” through the version 4.0 genome sequence data, 130 proteins containing the GXSXG motif common to hydrolases were identified. Among the respective genes, 35 (27%) showed increased and 11 (8.5%) showed decreased transcript levels by 2-fold or more following N deprivation. In addition, many potential lipases may be considered constitutively expressed. Seventy-four out of 130 (57%) lipase candidates were expressed at slightly higher levels following N deprivation. Some of these genes may encode lipases that are important for the turnover or replacement of membranes during cell growth or gamete fusion.

Changes in RNA Abundance for Transcription Factors

The Plant Transcription Factor Database (Pérez-Rodríguez et al., 2010) was used to identify 386 genes encoding putative transcription factors and transcriptional regulators in the *C. reinhardtii* transcript data set, which could be sorted into 53 families. Of the 368 genes, 83 showed a 2-fold or greater change in transcript abundance following N deprivation, with 46 being up-regulated and 37 being down-regulated.

To date, only a few of the putative transcription factors identified in the *C. reinhardtii* genome have a known function. Transcript abundance for the gene encoding NIT2, a transcription factor regulating nitrate metabolism, was increased 6-fold, while that for NAB1, a transcription factor regulating light-harvesting proteins, was decreased 16-fold, consistent with previously described physiological changes in response to N deprivation (Mussgnug et al., 2005; Camargo et al., 2007). The transcript level for the GSP1 mt^{+} gamete-specific transcription factor was decreased 3-fold at the 48-h sample point. When looking at the changes in RNA abundance of putative transcription factor genes, no obvious trends emerged. However, transcripts falling into the AP2-EREBP and bHLH families were generally more abundant following N deprivation, while those of the FHA family were generally decreased.

DISSCUSSION

Microalgae such as *C. reinhardtii* undergo drastic changes in metabolism and ultimately development when N deprived. Some of the most remarkable changes involve gametogenesis (Harris, 2009) and metabolic changes that lead to the accumulation of TAGs (Hu et al., 2008). The former aspect has been studied since the 1970s. However, focus on the latter has been largely motivated by the renewed interest in microalgae as biofuel feedstocks (Wijffels and Barbosa, 2010). As sequencing technology has become increasingly fast and affordable, comparison of transcriptomic changes under different experimental conditions by massive parallel sequencing of cDNA libraries is a viable first approach toward identifying genes that define changes in response to N deprivation or other nutrient stresses (González-Ballester et al., 2010). With the goal of gaining a better understanding of the factors underlying or even controlling the process of TAG accumulation following N deprivation, the focus has to be on metabolism and genes that encode enzymes of relevant pathways or regulatory factors.

The validity of making inferences on metabolism from transcriptome data in this study has been verified in different ways. First, specific genes known to be induced following N deprivation, such as genes involved in gametogenesis or ammonium transport, were found to be expressed as described previously. Second, major metabolic changes predicted by transcript analysis, such as the redirection of acetate from the glyoxylate cycle and gluconeogenesis to fatty acid biosyntheses following N deprivation (Figure 1.3), were corroborated by labeling experiments (Figure 1.4).

By and large, gross changes in transcript abundance in response to N deprivation follow expected themes: genes encoding enzymes directly involved in N metabolism or N compound uptake have to be induced, protein biosynthesis is reduced to adjust to the decreased availability of amino acids, and photosynthesis is down-regulated to adjust to the altered metabolic state of the cell. In cyanobacteria, N deprivation led to the degradation of the highly abundant phycobili light-harvesting proteins so that they could be used as an N source for protein synthesis (Collier and Grossman, 1992). In *C. reinhardtii*, there is evidence that the cytochrome subunits are degraded not in response to the low concentrations of N per se but rather to the changed energy content of the cell. Thus, one possible advantage for the cells to decrease photosynthesis following N deprivation is to prevent the accumulation of reactive oxygen species (Bulté and Wollman, 1992). At the same time, genes encoding proteins of the respiratory chain in mitochondria were only moderately affected following N deprivation except for those encoding alternative oxidases, which showed elevated transcript abundance. These enzymes are induced under a number of stress conditions that affect the redox environment of the cell, but these effects can be quite indirect and are often difficult to causally connect to the applied stress, in this case N deprivation. It should also be pointed out here that only the expression of nuclear genes is probed in this study; the expression levels of genes for organelle-encoded proteins relevant to respiration or photosynthesis have not been examined.

A Role for PSBS following N Deprivation?

One particular surprise was the strong up-regulation of *PSBS* following N deprivation in *C. reinhardtii*. In Arabidopsis (*Arabidopsis thaliana*), PSBS has been shown to play a critical role in nonphotochemical quenching (Li et al., 2000; Li et al., 2002), but previous studies in *C. reinhardtii* have not detected either of the two PSBS proteins in the thylakoids, and nonphotochemical quenching was shown to be independent of these proteins (Bonente et al., 2008). Only rarely have ESTs been found for *PSBS* transcripts, with the exception of the cDNA stress collection II, which contains RNAs from different stress treatments, including the switch from ammonium to nitrate (Shrager et al., 2003; Bonente et al., 2008). Our results indicate that *PSBS* expression in *C. reinhardtii* is induced by ammonium deprivation, as was observed previously for the expression of the gene for alternative oxidase, *AOX1* (Baurain et al., 2003).

Recycling of Membrane Lipids or de Novo Synthesis of TAGs?

The elevation of synthesis and export of fatty acids from the chloroplast following N deprivation could indicate that TAG is assembled from fatty acids that are synthesized de novo. This step would require the activation of the fatty acids by a long-chain acyl-CoA synthetase. In fact, increased abundance of RNA encoding a putative long-chain acyl-CoA synthetase was observed, and the respective protein has been identified in the lipid droplet proteome (Moellering and Benning, 2010). However, another enzyme that could contribute to the changing spectrum of fatty acids is a putative phospholipid/glycerol acyltransferase, for which the transcript level decreased during N deprivation. Long-chain acyl-CoA synthetases are likely to play a key role in determining the fate of fatty acids in the cell (Shockey et al., 2002). Regulation of the respective genes could be a major factor in controlling the flux of fatty acids toward glycerolipid synthesis and their degradation by β-oxidation.

Major intracellular changes occur following N deprivation, and these are likely accompanied by remodeling of membranes. Thus, fatty acids in membrane lipids might be recycled into TAGs. Consistently, some of the transcripts whose abundance changes the most were those encoding putative lipases. In general, lipases belong to a family of enzymes that deesterify carboxyl esters, such as TAGs and phospholipids. As TAGs accumulate following N deprivation, TAG lipases would be expected to be down-regulated. However, classifying lipases with selective substrate specificity based solely on their primary sequences is challenging, a fact that needs to be taken into consideration when interpreting our data set. A TAG lipase typically contains a Ser-Asp/Glu-His catalytic triad, with the Ser catalytic center located in a GXSXG motif (Brady et al., 1990; Winkler et al., 1990). Some recently characterized TAG lipases in animals, yeast, and plants contain a patatin-like, iPLA2 family Ser-Asp catalytic dyad (Zimmermann et al., 2004; Athenstaedt and Daum, 2005; Eastmond, 2006; Kurat et al., 2006). Genes encoding lipases specific for membrane lipids would be expected to be up-regulated, as they might mobilize fatty acids from membrane lipids into TAGs. Moreover, signaling pathways involving lipid products generated by lipases, such as diacyglycerol, may also control steady-state TAG levels (Kanoh et al., 1993). Further biochemical characterization of some of the most regulated lipase candidate genes will be necessary to determine their role in TAG accumulation following N deprivation.

On the other hand, we recognize that lipase expression or activities may also be controlled at the posttranscriptional level, including translational regulation and posttranslational modifications of the encoded proteins. In mammals, a hormone-sensitive lipase is phosphorylated by protein kinase A upon cAMP elevation and consequently exhibits better accessibility to lipid droplets (Holm et al., 2000). Some of the *C. reinhardtii* TAG lipases may have a similar regulatory pattern and hence not show significant transcriptional changes when cells are N deprived. Reverse genetic studies on the lipase candidates and forward genetic screens for mutants with TAG deficiency phenotypes will disclose the bona fide TAG lipases and other lipases that impact TAG metabolism.

It should also be noted that during the analysis of the lipid gene data set, the annotation ambiguities of several fatty acid desaturases became obvious in the version 4.0 genomic sequence data set: the *C. reinhardtii* genome harbors four presumed paralogs of FAD5 (named FAD5a–FAD5d). Based on our current prediction analysis (Emanuelsson et al., 1999) and previous reports (Riekhof et al., 2005; Harris and Stern, 2009), FAD5a and FAD5b are presumed to be targeted to the chloroplast, whereas FAD5c and FAD5d are likely located in the ER membrane. However, experimental corroboration is still needed.

CONCLUSION

Our interpretation of this data set places emphasis on TAG metabolism and potential regulatory factors, which are undoubtedly not yet completely identified. We expect that others will be able to mine this data set, taking into account different biological processes pertaining to N deprivation. Cross-querying this data set with a lipid droplet proteomics data set (Moellering and Benning, 2010) should further narrow the possible candidates relevant for TAG accumulation. Likewise, a meta-analysis of this and other data sets, including those from other species, could facilitate the identification of genes most likely involved in TAG accumulation, as was recently done for low-oxygen stress (Mustroph et al., 2010). Thus, this study represents only a first step of many toward gaining a molecular understanding of TAG accumulation and other cellular changes triggered by N deprivation in *C. reinhardtii*.

MATERIALS AND METHODS

Strains and Growth Conditions

The *C. reinhardtii* strain used was dw15.1 (cw15, nit1, mt^{+}), kindly provided by Arthur Grossman. The cells were grown in liquid cultures under continuous light (approximately 80 μmol photons m$^{−2}$ s$^{−1}$). For N-replete growth, TAP medium (Harris, 2009a) with 10 mm NH_4^{+} (TAP + N) was used. For preliminary experiments, N deprivation was applied by two methods: continuous growth in TAP with 0.5 mm NH_4^{+} or growth in TAP + N to 5 × 10^{6} cells mL$^{−1}$, followed by transfer to TAP with no NH_4^{+} (TAP − N) for an additional 24 or 48 h. For further experiments, N deprivation was defined as growth in TAP + N to 5 × 10^{6} cells mL$^{−1}$, followed by transfer to TAP − N for 48 h.

For labeling studies, the cells were grown in 500-mL shaker flasks with a culture volume of 50 mL with continuous shaking. For the N deprivation experiment, cells were first grown in TAP medium with unlabeled acetate with at least five cell doublings to mid logarithmic phase to reach a biomass equivalent to 0.3 to 0.4 g cell dry weight L−1. The cells were divided up and transferred to TAP medium containing [U-13C]acetate (Isotec), either TAP + N or TAP − N.

Sequencing Read Processing

To generate material for high-throughput sequencing, cells were grown in 100 mL of TAP + N to 5 × 10^{6} cells mL$^{−1}$. The cultures were split in half, and cells were collected by centrifugation, with one pellet being re-suspended in 50 mL of TAP + N and the other in 50 mL of TAP − N. After 48 h, the total RNA was harvested using a Qiagen RNeasy Plant Mini kit. The RNA samples were treated with Qiagen RNase-free DNase I during extraction.

For 454 sequencing, full-length cDNA pools were generated with the Clontech SMART cDNA library construction kit. cDNA was synthesized using a modified cDNA synthesis primer (5′-TAGAGACCGAGGCGGCCGACATGTTTTGTTTTTTTTTCTTTTTTTTTTVN-3′). Full-length cDNAs were amplified by PCR and pooled to increase their concentration. An *Sfi*I digest was performed, followed by size fractionation. Fractions with the highest intensity and size distribution were pooled and purified. The resulting cDNA pools were then submitted to the Michigan State University-Research Technologies Service Facility for sequencing on a 454 GSFLX Titanium Sequencer (454 Life Sciences). For Illumina sequencing, total RNA was submitted directly to the Michigan State University-Research Technologies Service Facility for sequencing on an Illumina Genome Analyzer II (Illumina).

Default parameters were used to pass reads using 454 and Illumina quality-control tools. The filtered sequence data were deposited in the National Center for Biotechnology Information Short Read Archive with the reference series number GSE24367 and subseries numbers GSE24365 and GSE24366 for the Illumina and the 454 data sets, respectively. The filtered 454 sequencing reads were mapped to the *C. reinhardtii* version 4.0 assembly from the Joint Genome Institute with GMAP (Wu and Watanabe, 2005). In GMAP, the maximum intron length was set at 980 bp, which is at the 95th percentile of annotated *C. reinhardtii* intron lengths. The Illumina reads were mapped with Bowtie (Langmead et al., 2009) using parameters as follows: two or fewer mismatches, sum of Phred quality values at all mismatched positions at 70 or less, and excluding reads mapped to one or more locations. Because the sequence qualities of Illumina reads degrade quickly toward the 3′ end, an alternative mapping data set was generated with reads trimmed from the 3′ end (until the 3′-end-most position with Phred-equivalent score was 20 or greater). Trimmed reads of less than 30 bp were excluded from further analysis. In addition to sequence quality issues, some reads may span two exons and would not be mapped by Bowtie correctly. We used TopHat (Trapnell et al., 2009) to identify these exon-spanning reads to generate another set of read mapping. The information from TopHat was used for assembling mapped reads into transfrags with Cufflinks (Trapnell et al., 2010). In Cufflinks, the maximum intron length was set at 1,855 bp (99th percentile of all the intron lengths), 5% minimum isoform fraction, and 5% pre-mRNA fraction. Transfrags within 1,855 bp of an existing *C. reinhardtii* version 4.0 gene model were regarded as potential missing exons of annotated genes. The rest were regarded as intergenic exons, and adjacent transfrags less than 1,855 bp apart were joined into “transcriptional units.”

Northern-Blot Analysis and RT-PCR

Total RNA was harvested from N-replete or N-deprived cells as described above, and 4 μg of each total RNA was separated on a 1% formaldehyde gel and transferred to a Hybond-N$^+$ nylon membrane (GE Healthcare). Probes were synthesized from cDNA and labeled with ^{32}P using the Amersham Megaprime labeling kit (GE Healthcare). The blots were hybridized with the labeled probes in Ambion ULTRAhybe (Applied Biosystems/Ambion) at 42°C overnight. The blots were washed twice for 5 min with low-stringency buffer (1× SSC and 0.1% SDS) at

60°C and then twice for 5 min with high-stringency buffer (0.1× SSC and 0.1% SDS) at 60°C. The blots were exposed to a Molecular Dynamics phosphor screen (GE Healthcare) overnight and visualized with a Storm 820 phosphor imager (GE Healthcare). Probes were synthesized from cDNA for AMT4 (5′-GTATTGCGTCCGATCTGC-3′ and 5′-CGTGGAAATGCTGTAGGG-3′), DGTT2 (5′-TAAAGCACCGACAAATGTGC-3′ and 5′-CATGATCTGGCATTCTGTGG-3′), and DGTT3 (5′-GGTGGTGCTCTCCTACTGGA-3′ and 5′-CCATGTACATCTCGGCAATG-3′).

For RT-PCR, RNA was extracted from N-replete and N-deprived cultures using TRIzol reagent (Invitrogen) and subjected to DNase treatment with the Turbo DNA-free kit from Ambion. A total of 1 μg of DNA-free RNA was used for cDNA synthesis with the Invitrogen Moloney murine leukemia virus reverse transcriptase. A total of 0.5 μg of oligo(dT)$_{12–18}$ primer (Invitrogen) and 0.5 μg of random hexamer primers (Promega) were added to the RNA, and the volume was adjusted to 20 μL final volume. After heating the samples at 70°C for 10 min, they were incubated on ice for 5 min. Twenty units of RNase inhibitor (Applied Biosystems), 20 nmol of deoxyribonucleotide triphosphates (Invitrogen), 4 μL of first-strand buffer, and 0.2 μmol of dithiothreitol were added to the reaction. The reaction mixture was incubated at 37°C for 10 min for primer annealing. A total of 200 units of Moloney murine leukemia virus reverse transcriptase was added, and the reaction was incubated at 37°C for 1 h followed by deactivation at 70°C for 10 min. A total of 1 μL of a 1:10 dilution of the respective cDNA was used as template for a PCR using GoTaq polymerase (Promega). The reaction mixture (25 μL) contained 1× buffer, 5 nmol of deoxyribonucleotide triphosphates, 12.5 pmol of each primer, and 1 unit of polymerase. PCR cycle conditions were 3 min of initial denaturation at 94°C, following 40 cycles of 30 s of denaturation, 30 s of annealing at 60°C, and 3 min of elongation at 72°C. Final elongation was performed at 72°C for 10 min. The *PSBS*-specific primers (5′-ATGGCCATGACTCTGTCGAC-3′ and 5′-TTAGGCGGACTCCTCGTCC-3′) amplify both *PSBS1* and *PSBS2*. The *IDA5* gene (5′-GCCAGGTCTCTGCTCTGGTG-3′ and 5′-TACTCGGACTTGGCGATCCA-3′) served as a control.

Analysis of Differential Gene Expression

Differential expression between *C. reinhardtii* cultured in N-replete and N-depleted medium was determined using the numbers of mapped reads overlapped with annotated *C. reinhardtii* genes as inputs to EdgeR (Robinson et al., 2010). In the Joint Genome Institute database, multiple sets of *C. reinhardtii* version 4 gene models are available. We used the "filtered" gene models, which contain the best gene model for each locus. Genes were regarded as differentially expressed if they have 2-fold or greater change between N-replete and N-deprived samples and 5% or less false discovery rate. Differential expressed genes were regarded as up-regulated if their expression levels in N-deprived samples were significantly higher than those in N-replete samples. Conversely, down-regulated genes were those with significantly lower levels of expression following N deprivation.

In addition to EdgeR, we used three other methods to evaluate differential expression: Fisher's exact test (Bloom et al., 2009), likelihood ratio test (Marioni et al., 2008), and a method based on intensity ratio and average intensity (MARS; Wang et al., 2010). All three methods were implemented in the DEGexp package (Wang et al., 2010). We found that among 4,004 differentially expressed genes called by EdgeR, 99.7% to 100% were regarded as differentially expressed by the other three methods. On the other hand, EdgeR calls overlap with 96.6%, 94.7%, and 95.8% of calls by Fisher's exact test, likelihood ratio test, and MARS, respectively. Our findings indicate that EdgeR is more conservative than the other methods, but the overall differential expression calls are highly similar among methods. We should note that methods other than EdgeR did not explicitly consider variance between replicates and, as a result, will likely have a higher false-positive differential expression call rate than that of EdgeR. Therefore, in all subsequent analyses, we used only EdgeR-based differential expression calls.

GO annotation for the *C. reinhardtii* version 4.0 genome was acquired from the Joint Genome Institute. Enrichment of differentially regulated genes in each GO category was determined using Fisher's exact test. To account for multiple testing, the *P* values from Fisher's exact tests were adjusted (Storey, 2003) and a false discovery rate of 5% was used as the threshold for enriched GO terms.

GC-MS Analysis

To quantify ^{13}C-labeling patterns such as mass-isotopomer distributions and fractional ^{13}C enrichment, samples were analyzed using GC-MS using an HP 6890 GC apparatus (Hewlett-Packard) equipped with DB-5MS column (5% phenyl-methyl-siloxan-diphenylpolysiloxan; 30 m × 0.251 mm × 0.25 μm; Agilent) and a quadrupole mass spectrometer (MS 5975; Agilent). Electron ionization was carried out at 70 eV. The obtained mass spectrometric data were corrected for the natural abundance of the elements to give fractional ^{13}C labeling.

Sampling, Extraction, and Analysis of Intracellular Amino Acid

Cells from TAP + N and TAP − N cultures were harvested after 24 h. This time point was chosen for these labeling experiments because at the required cell concentration (approximately 0.2 g cell dry weight L^{-1}) for metabolite extraction, acetate depletion occurs in cells grown in TAP + N at later time points due to the high initial inoculum.

The harvested cells (approximately 25 mg cell dry weight) were centrifuged at 3,000*g* for 1 min, and the supernatant was removed and quenched with 5 mL of cold 100% methanol (Winder et al., 2008). The metabolites were harvested by vortexing the cells. A second extraction was performed with 5 mL of chloroform:methanol (1:2). The extracts were then pooled. Water was slowly added to the pooled extracts for phase separation. The polar metabolites, which include the amino acids, were present in the aqueous phase. The aqueous phase was then dried under N_2 and converted to its *t*-butyldimethylsilyl derivative using *N*-methyl-*N*-(*t*-butyldimethylsilyl)trifluoroacetamide (Mawhinney et al., 1986). The GC and MS conditions for this analysis were as described previously (Deshpande et al., 2009).

Extraction and Analysis of Rib

Rib for analysis was obtained from the RNA as described (Boren et al., 2003). RNA was extracted from cells (0.1–0.2 g L^{-1}) at the same time point as the intracellular amino acids using the Tri reagent as described in the protocol (Molecular Research Center). The RNA was acid hydrolyzed to its monomers and dried under N. It was further analyzed using GC-MS by derivatizing it to its per-*O*-trimethylsilyl-*O*-ethyl oxime (MacLeod et al., 2001). The ions 481 to 486 (mass-to-charge ratio), corresponding to the whole carbon backbone of the Rib molecule (C_1–C_5), were monitored using single ion monitoring of the MS data.

Extraction and Analysis of Carbohydrate

The carbohydrates in the cells were acid hydrolyzed by 2 n HCl at 102°C, and the monomeric compounds were analyzed by GC-MS after the sample was dried under N_2. The sample was then converted to its di-*O*-isopropylidene acetate derivative for analysis by GC-MS (Hachey et al., 1999). The ions 287 to 293 (mass-to-charge ratio), corresponding to the whole carbon backbone of Glc (C_1–C_6), were monitored.

Sequence data from this article can be found in the National Center for Biotechnology Information Gene Expression Omnibus under accession numbers GSE24367, GSE2466, and GSE2465.

ACKNOWLEDGEMENT

We thank Titus Brown, Gaurav Moghe, and Cheng Zou for help with analysis of Illumina data, James Bullard for help with comparing differential expression call methods, and the Joint Genome Institute for providing C. reinhardtii sequences and gene annotation.

REFERENCES

REFERENCES

Athenstaedt K, Daum G (2005) Tgl4p and Tgl5p, Two Triacylglycerol Lipases of the Yeast *Saccharomyces cerevisiae* Are Localized to Lipid Particles. J Biol Chem **280**: 37301–37309

Baurain D, Dinant M, Coosemans N, Matagne RF (2003) Regulation of the Alternative Oxidase Aox1 Gene in *Chlamydomonas reinhardtii*. Role of the Nitrogen Source on the Expression of a Reporter Gene under the Control of theAox1 Promoter. Plant Physiol **131**: 1418–1430

Bloom JS, Khan Z, Kruglyak L, Singh M, Caudy AA (2009) Measuring differential gene expression by short read sequencing: quantitative comparison to 2-channel gene expression microarrays. BMC Genomics **10**: 221

Bonente G, Passarini F, Cazzaniga S, Mancone C, Buia MC, Tripodi M, Bassi R, Caffarri S (2008) The Occurrence of the *psbS* Gene Product in *Chlamydomonas reinhardtii* and in Other Photosynthetic Organisms and Its Correlation with Energy Quenching [†]. Photochem Photobiol **84**: 1359–1370

Boren J, Lee W-NP, Bassilian S, Centelles JJ, Lim S, Ahmed S, Boros LG, Cascante M (2003) The Stable Isotope-based Dynamic Metabolic Profile of Butyrate-induced HT29 Cell Differentiation. J Biol Chem **278**: 28395–28402

Brady L, Brzozowski AM, Derewenda ZS, Dodson E, Dodson G, Tolley S, Turkenburg JP, Christiansen L, Huge-Jensen B, Norskov L, et al (1990) A serine protease triad forms the catalytic centre of a triacylglycerol lipase. Nature **343**: 767–770

Bullard JH, Purdom E, Hansen KD, Dudoit S (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics **11**: 94

Bulté L, Wollman F-A (1992) Evidence for a selective destabilization of an integral membrane protein, the cytochrome b6/f complex, during gametogenesis in *Chlamydomonas reinhardtii*. Eur J Biochem **204**: 327–336

Camargo A, Llamas Á, Schnell RA, Higuera JJ, González-Ballester D, Lefebvre PA, Fernández E, Galván A (2007) Nitrate Signaling by the Regulatory Gene NIT2 in *Chlamydomonas*. Plant Cell Online **19**: 3491–3503

Carman GM, Han G-S (2006) Roles of phosphatidate phosphatase enzymes in lipid metabolism. Trends Biochem Sci **31**: 694–699

Collier JL, Grossman AR (1992) Chlorosis induced by nutrient deprivation in *Synechococcus* sp. strain PCC 7942: not all bleaching is the same. J Bacteriol **174**: 4718–4726

Deshpande R, Yang TH, Heinzle E (2009) Towards a metabolic and isotopic steady state in CHO batch cultures for reliable isotope-based metabolic profiling. Biotechnol J **4**: 247–263

Eastmond PJ (2006) SUGAR-DEPENDENT1 Encodes a Patatin Domain Triacylglycerol Lipase That Initiates Storage Oil Breakdown in Germinating *Arabidopsis* Seeds. Plant Cell Online **18**: 665–675

Emanuelsson O, Nielsen H, von Heijne G (1999) ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. Protein Sci Publ Protein Soc **8**: 978–984

Fisslthaler B, Meyer G, Bohnert HJ, Schmitt JM (1995) Age-dependent induction of pyruvate, orthophosphate dikinase in *Mesembryanthemum crystallinum* L. Planta **196**: 492–500

González-Ballester D, Camargo A, Fernández E (2004) Ammonium transporter genes in *Chlamydomonas*: the nitrate-specific regulatory gene Nit2 is involved in Amt1;1 expression. Plant Mol Biol **56**: 863–878

González-Ballester D, Casero D, Cokus S, Pellegrini M, Merchant SS, Grossman AR (2010) RNA-Seq Analysis of Sulfur-Deprived *Chlamydomonas* Cells Reveals Aspects of Acclimation Critical for Cell Survival. Plant Cell Online **22**: 2058–2084

Hachey DL, Parsons WR, McKay S, Haymond MW (1999) Quantitation of monosaccharide isotopic enrichment in physiologic fluids by electron ionization or negative chemical ionization GC/MS using di-O-isopropylidene derivatives. Anal Chem **71**: 4734–4739

Hamel P, Olive J, Pierre Y, Wollman F-A, Vitry C de (2000) A New Subunit of Cytochromeb 6 f Complex Undergoes Reversible Phosphorylation upon State Transition. J Biol Chem **275**: 17072–17079

Harris EH (2009) The *Chlamydomonas* sourcebook Vol. 1: Introduction to *Chlamydomonas* and Its Laboratory Use. Elsevier, Amsterdam

Harris EH, Stern DB (2009) The *Chlamydomonas* sourcebook Vol. 2 Vol. 2. Elsevier, Amsterdam

Holm C, Østerlund T, Laurell H, Contreras JA (2000) Molecular Mechanisms Regulating Hormone-Sensitive Lipase and Lipolysis. Annu Rev Nutr **20**: 365–393

Hu Q, Sommerfeld M, Jarvis E, Ghirardi M, Posewitz M, Seibert M, Darzins A (2008) Microalgal triacylglycerols as feedstocks for biofuel production: perspectives and advances. Plant J Cell Mol Biol **54**: 621–639

Jamers A, Van der Ven K, Moens L, Robbens J, Potters G, Guisez Y, Blust R, De Coen W (2006) Effect of copper exposure on gene expression profiles in *Chlamydomonas reinhardtii* based on microarray analysis. Aquat Toxicol Amst Neth **80**: 249–260

Kanoh H, Sakane F, Imai S-I, Wada I (1993) Diacylglycerol kinase and phosphatidic acid phosphatase—enzymes metabolizing lipid second messengers. Cell Signal **5**: 495–503

Kurat CF, Natter K, Petschnigg J, Wolinski H, Scheuringer K, Scholz H, Zimmermann R, Leber R, Zechner R, Kohlwein SD (2006) Obese Yeast: Triglyceride Lipolysis Is Functionally Conserved from Mammals to Yeast. J Biol Chem **281**: 491–500

Kurvari V, Grishin NV, Snell WJ (1998) A Gamete-specific, Sex-limited Homeodomain Protein in *Chlamydomonas*. J Cell Biol **143**: 1971–1980

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

Ledford HK, Baroli I, Shin JW, Fischer BB, Eggen RIL, Niyogi KK (2004) Comparative profiling of lipid-soluble antioxidants and transcripts reveals two phases of photo-oxidative stress in a xanthophyll-deficient mutant of *Chlamydomonas reinhardtii*. Mol Genet Genomics MGG **272**: 470–479

Ledford HK, Chin BL, Niyogi KK (2007) Acclimation to singlet oxygen stress in *Chlamydomonas reinhardtii*. Eukaryot Cell **6**: 919–930

Li X-P, Björkman O, Shih C, Grossman AR, Rosenquist M, Jansson S, Niyogi KK (2000) A pigment-binding protein essential for regulation of photosynthetic light harvesting. Nature **403**: 391–395

Li X-P, Gilmore AM, Niyogi KK (2002) Molecular and Global Time-resolved Analysis of a psbSGene Dosage Effect on pH- and Xanthophyll Cycle-dependent Nonphotochemical Quenching in Photosystem II. J Biol Chem **277**: 33590–33597

MacLeod JK, Flanigan IL, Williams JF, Collins JG (2001) Mass spectrometric studies of the path of carbon in photosynthesis: positional isotopic analysis of 13C-labelled C4 to C7 sugar phosphates. J Mass Spectrom **36**: 500–508

Magne Ø, Driscoll BT, Finan TM (1997) Increased pyruvate orthophosphate dikinase activity results in an alternative gluconeogenic pathway in *Rhizobium (Sinorhizobium) meliloti*. Microbiology **143**: 1639–1648

Majeran W, Wollman F-A, Vallon O (2000) Evidence for a Role of ClpP in the Degradation of the Chloroplast Cytochrome b6f Complex. Plant Cell Online **12**: 137–149

Mamedov TG, Moellering ER, Chollet R (2005) Identification and expression analysis of two inorganic C- and N-responsive genes encoding novel and distinct molecular forms of eukaryotic phosphoenolpyruvate carboxylase in the green microalga *Chlamydomonas reinhardtii*. Plant J **42**: 832–843

Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y (2008) RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. Genome Res **18**: 1509–1517

Martin NC, Chiang K-S, Goodenough UW (1976) Turnover of chloroplast and cytoplasmic ribosomes during gametogenesis in *Chlamydomonas reinhardi*. Dev Biol **51**: 190–201

Martin NC, Goodenough UW (1975) Gametic differentiation in *Chlamydomonas reinhardtii*. I. Production of gametes and their fine structure. J Cell Biol **67**: 587–605

Mattingly SJ, Dipersio JR, Higgins ML, Shockman GD (1976) Unbalanced growth and macromolecular synthesis in *Streptococcus* mutans FA-1. Infect Immun **13**: 941–948

Mawhinney TP, Robinett RSR, Atalay A, Madson MA (1986) Analysis of amino acids as their tert.-butyldimethylsilyl derivatives by gas—liquid chromatography and mass spectrometry. J Chromatogr A **358**: 231–242

Merchant SS, Prochnik SE, Vallon O, Harris EH, Karpowicz SJ, Witman GB, Terry A, Salamov A, Fritz-Laylin LK, Maréchal-Drouard L, et al (2007) The *Chlamydomonas* Genome Reveals the Evolution of Key Animal and Plant Functions. Science **318**: 245–250

Moellering ER, Benning C (2010) RNA Interference Silencing of a Major Lipid Droplet Protein Affects Lipid Droplet Size in *Chlamydomonas reinhardtii*. Eukaryot Cell **9**: 97–106

Mus F, Dubini A, Seibert M, Posewitz MC, Grossman AR (2007) Anaerobic acclimation in *Chlamydomonas reinhardtii*: anoxic gene expression, hydrogenase induction, and metabolic pathways. J Biol Chem **282**: 25475–25486

Mussgnug JH, Wobbe L, Elles I, Claus C, Hamilton M, Fink A, Kahmann U, Kapazoglou A, Mullineaux CW, Hippler M, et al (2005) NAB1 Is an RNA Binding Protein Involved in the Light-Regulated Differential Expression of the Light-Harvesting Antenna of *Chlamydomonas reinhardtii*. Plant Cell Online **17**: 3409–3421

Mustroph A, Lee SC, Oosumi T, Zanetti ME, Yang H, Ma K, Yaghoubi-Masihi A, Fukao T, Bailey-Serres J (2010) Cross-kingdom comparison of transcriptomic adjustments to low-oxygen stress highlights conserved and plant-specific responses. Plant Physiol **152**: 1484–1500

Nguyen AV, Thomas-Hall SR, Malnoë A, Timmins M, Mussgnug JH, Rupprecht J, Kruse O, Hankamer B, Schenk PM (2008) Transcriptome for photobiological hydrogen production induced by sulfur deprivation in the green alga *Chlamydomonas reinhardtii*. Eukaryot Cell **7**: 1965–1979

Oshlack A, Wakefield MJ (2009) Transcript length bias in RNA-seq data confounds systems biology. Biol Direct **4**: 14

Peltier G, Schmidt GW (1991) Chlororespiration: an adaptation to nitrogen deficiency in *Chlamydomonas reinhardtii*. Proc Natl Acad Sci **88**: 4791–4795

Pérez-Rodríguez P, Riaño-Pachón DM, Corrêa LGG, Rensing SA, Kersten B, Mueller-Roeber B (2010) PlnTFDB: updated content and new features of the plant transcription factor database. Nucleic Acids Res **38**: D822–D827

Plumley FG, Schmidt GW (1989) Nitrogen-dependent regulation of photosynthetic gene expression. Proc Natl Acad Sci **86**: 2678–2682

Pollard M, Ohlrogge J (1999) Testing Models of Fatty Acid Transfer and Lipid Synthesis in Spinach Leaf Using in Vivo Oxygen-18 Labeling. Plant Physiol **121**: 1217–1226

Riekhof WR, Sears BB, Benning C (2005) Annotation of Genes Involved in Glycerolipid Biosynthesis in *Chlamydomonas reinhardtii*: Discovery of the Betaine Lipid Synthase BTA1Cr. Eukaryot Cell **4**: 242–252

Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics **26**: 139–140

Schnell RA, Lefebvre PA (1993) Isolation of the *Chlamydomonas* regulatory gene NIT2 by transposon tagging. Genetics **134**: 737–747

Shockey JM, Fulda MS, Browse JA (2002) Arabidopsis Contains Nine Long-Chain Acyl-Coenzyme A Synthetase Genes That Participate in Fatty Acid and Glycerolipid Metabolism. Plant Physiol **129**: 1710–1722

Shrager J, Hauser C, Chang C-W, Harris EH, Davies J, McDermott J, Tamse R, Zhang Z, Grossman AR (2003) *Chlamydomonas reinhardtii* Genome Project. A Guide to the Generation and Use of the cDNA Information. Plant Physiol **131**: 401–408

Siersma PW, Chiang K-S (1971) Conservation and degradation of cytoplasmic and chloroplast ribosomes in *Chlamydomonas reinhardtii*. J Mol Biol **58**: 167–185

Simon DF, Descombes P, Zerges W, Wilkinson KJ (2008) Global expression profiling of *Chlamydomonas reinhardtii* exposed to trace levels of free cadmium. Environ Toxicol Chem SETAC **27**: 1668–1675

Storey JD (2003) The positive false discovery rate: a Bayesian interpretation and the q-value. Ann Stat **31**: 2013–2035

Taylor L, Nunes-Nesi A, Parsley K, Leiss A, Leach G, Coates S, Wingler A, Fernie AR, Hibberd JM (2010) Cytosolic pyruvate,orthophosphate dikinase functions in nitrogen remobilization during leaf senescence and limits individual seed growth and nitrogen content. Plant J **62**: 641–652

Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics **25**: 1105–1111

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol **28**: 511–515

Wang L, Feng Z, Wang X, Wang X, Zhang X (2010) DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. Bioinformatics **26**: 136–138

Wang ZT, Ullrich N, Joo S, Waffenschmidt S, Goodenough U (2009) Algal Lipid Bodies: Stress Induction, Purification, and Biochemical Characterization in Wild-Type and Starchless *Chlamydomonas reinhardtii*. Eukaryot Cell **8**: 1856–1868

Weber APM, Weber KL, Carr K, Wilkerson C, Ohlrogge JB (2007) Sampling the *Arabidopsis* Transcriptome with Massively Parallel Pyrosequencing. Plant Physiol **144**: 32–42

Wijffels RH, Barbosa MJ (2010) An Outlook on Microalgal Biofuels. Science **329**: 796–799

Winder CL, Dunn WB, Schuler S, Broadhurst D, Jarvis R, Stephens GM, Goodacre R (2008) Global metabolic profiling of *Escherichia coli* cultures: an evaluation of methods for quenching and extraction of intracellular metabolites. Anal Chem **80**: 2939–2948

Winkler FK, D'Arcy A, Hunziker W (1990) Structure of human pancreatic lipase. Nature **343**: 771–774

Wu TD, Watanabe CK (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics **21**: 1859–1875

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

Zimmermann R, Strauss JG, Haemmerle G, Schoiswohl G, Birner-Gruenberger R, Riederer M, Lass A, Neuberger G, Eisenhaber F, Hermetter A, et al (2004) Fat Mobilization in Adipose Tissue Is Promoted by Adipose Triglyceride Lipase. Science **306**: 1383–1386

CHAPTER 2

GREEN ALGAL RETAINED DUPLICATE GENES TEND TO BE STRESS RESPONSE GENES AND EXPERIENCE FREQUENT RESPONSE GAINS

Note: The contents of this chapter are in a manuscript in preparation.

Guangxi Wu, David E. Hufnagel, Alisandra Denton, Shin-Han Shiu. Green algal retained duplicate genes tend to be stress response genes and experience frequent response gains

ABSTRACT

Green algae are being explored for biofuel production particularly because of their high oil content under stress conditions. In addition, diversity in stress modulated oil content increase exists among different species in response to stress. To investigate how gene families and gene expression have evolved in the context of stress response that ultimately leads to the oil content differences, we characterized the expansion patterns of gene families in nine green algal species, and examined evolution of stress response among gene duplicates in *Chlamydomonas reinhardtii.*

Substantial variation in domain family sizes exists among green algal species. Lineage-specific expansion of families occurred throughout the green algal lineage but inferred gene losses occurred more often than gene gains, suggesting a continuous reduction of algal gene repertoire. Retained duplicates tend to be involved in stress response, similar to land plant species. However, stress response genes tend to be pseudogenized also. When comparing ancestral and extant gene stress response state, we found that response gains occur in 13% duplicate gene branches, much higher than 6% in *Arabidopsis thaliana*.

The frequent gains of stress response among green algal duplicates potentially reflect a high rate of innovation, resulting in a species-specific gene repertoire contributed to adaptive response to stress. Particularly, this species-specificity in gene family evolution and its relationship to stress response are likely related to the oil content diversity among species, and could be further explored towards identifying suitable green algal species for oil production.

INTRODUCTION

Green algae are a group of photosynthetic organisms that are more closely related to land plants than to other major eukaryotic groups (Graham et al., 2009). A number of micro-green-algal species are suitable for biofuel production, and the lipid content of these algae increases significantly under various stress conditions (Hu et al., 2008). For example, in the green algal model *Chlamydomonas reinhardtii*, lipid droplets rich in triacylglycerol (TAG) form after nitrogen (N) deprivation (Wang et al., 2009; Moellering and Benning, 2010). Other stress conditions, such as salt stress and sulfur deprivation, also lead to increased TAG content in *C. reinhardtii* (Siaut et al., 2011; Cakmak et al., 2012). In addition, stress response and subsequent lipid accumulation exhibit tremendous diversity in green algae (Hu et al., 2008). For example, in response to N deficiency several green algae showed substantial differences in the tradeoff between growth and lipid content, the lipid accumulation time-course, and the stress level required to stimulate lipid formation (Adams et al., 2013). Thus, a better understanding of the mechanistic details of stress response in green algae will not only contribute to our knowledge about adaptation to stressful environments but also will have the potential for improving microalgal biofuel production.

To better understand how green algae respond to stress on a genomic level, we focused on the retention and loss of gene family members. This is because gene duplications lead to raw materials for evolution to act on (Ohno, 1970; Force et al., 1999) and they are a common feature of most eukaryotic species (Lynch et al., 2001). Genes involved in stress response tend to be retained in various eukaryotes (Lespinet et al., 2002; Hanada et al., 2008). Thus a thorough look at the retention and loss patterns among green algal gene families may provide an evolutionary perspective on the connection between stress and green algal oil production. In general, the majority of gene duplicates are rapidly lost following the duplication event, but a significant number of duplicates are retained (Lynch et al., 2001; Moore and Purugganan, 2003; Moore and Purugganan, 2005), contributing to organismal and regulatory complexity (Lespinet et al., 2002; Vogel and Chothia, 2006). Gene retention occurs in a lineage-specific manner in a wide range of organisms (Jordan et al., 2001; Lespinet et al., 2002; Hanada et al., 2008) and a significant functional bias exists. In fungi, stress-related genes tend to undergo many duplications and losses (Wapinski et al., 2007). Similarly, in *Arabidopsis thaliana*, stress responsive genes tend be retained (Hanada et al., 2008) but also tend to be pseudogenized (Zou et al., 2009a). In a study of duplicates in yeasts, nematode, fruit fly, and *A. thaliana*, genes involved in response to environmental stress are prone to be retained in a lineage-specific manner (Lespinet et al., 2002). There is not yet a global study summarizing the gene gain and loss patterns of known gene families in green algae. In addition, it is not known if duplicate retention is correlated with their stress responsiveness in green algae.

After duplication, gene duplicates may acquire a novel function that contributes to adaptation (Ohno, 1970). Such neo-functionalization can be an important source of inter-specific differences in stress response that could lay the foundation for diversity in stress-induced oil production in algae. An earlier study in *A. thaliana* showed that, although the predominant fate of duplicate was loss of stress response, in around 6% of the cases there was evidence of stress responsive gain (Zou et al., 2009b). Examining evolution in gene expression among duplicates could thus further elucidate how gene duplication and subsequent functional innovation shaped the gene repertoire involved in stress response, and likely contributing to the diversity in stress response in green algae. There is no global study on functional evolution of duplicate genes in green algae.

In this study, we integrated genomic and transcriptomic data to find out how gene families and gene expression have evolved in the green algal lineage in the context of stress. We first examined the variation of domain family composition in nine green algal species compared to land plants. The nine green algal species included are *Micromonas pusilla* RCC299, *Micromonas pusilla* CCMP1545, *Ostreococcus* sp.RCC809, *Ostreococcus tauri*, *Ostreococcus lucimarinus*, *Chlorella* NC64A, *Coccomyxa* sp.C169, *C. reinhardtii,* and *Volvox carteri* (see Methods). Then we investigated how gene gain and loss events occurred in the green algal lineage and examined the functional bias in retained duplicates in *C. reinhardtii* using phylogenetic approaches. We also examined the pseudogenization of gene duplicates and their functional bias. Finally, we characterized the evolution of gene expression after duplication events to find out how gene function evolved in the context of stress response. Our study reveals the evolutionary trajectory of stress responsive gene families in the green algal lineage, and because lipid production is sharply induced by stress (Hu et al., 2008), our results could help to pinpoint candidate genes for further study on lipid metabolism.

RESULTS AND DISCUSSION

Variation in domain family sizes among land plants and green algae

Our goal here was to define the overall gain and loss of duplicates in gene families over the course of green algal evolution. Gene families can be defined in two ways. In an earlier study, a protein sequence similarity based approach was used to identify protein families using full length protein sequences in green algae (Blanc et al., 2012). Here, we used another approach by defining a family as proteins having the same domain, using the Pfam database (Finn et al., 2010); protein domains are well-defined regions of a protein that can perform a specific function and form a structural unit (Graur, 2000). This approach was adopted because the sequence similarity approach could group non-homologous genes/regions into the same family (A is related to B, B to C, but not A and C). On the other hand, our approach only includes proteins with known domains. On average 65.7% green algal genes have one or more recognized domains, ranging from 56% to 76% among different species, while land plants have slightly higher percentage (**Figure 2.1A**). This could reflect the fact that overall green algae are less well studied than land plants and might have unknown algal-specific domains. It could also indicate a discrepancy in annotation quality between different green algal species. Given that genes with known domains are regarded as gene models with higher confidence (Cantarel et al., 2008), using protein domains to establish family alleviates the annotation quality issue.

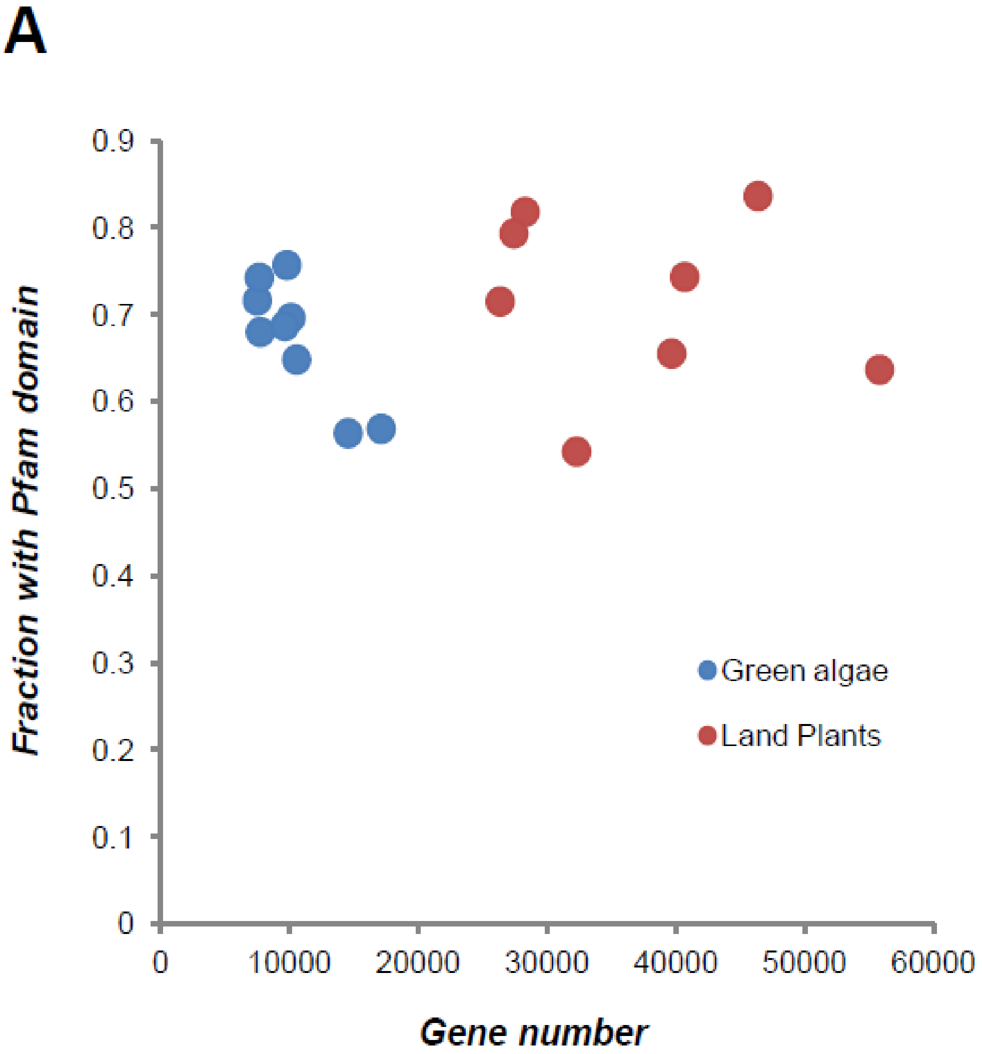


Figure 2.1. Domain families in green algae and land plants.

Figure 2.1 (cont'd)

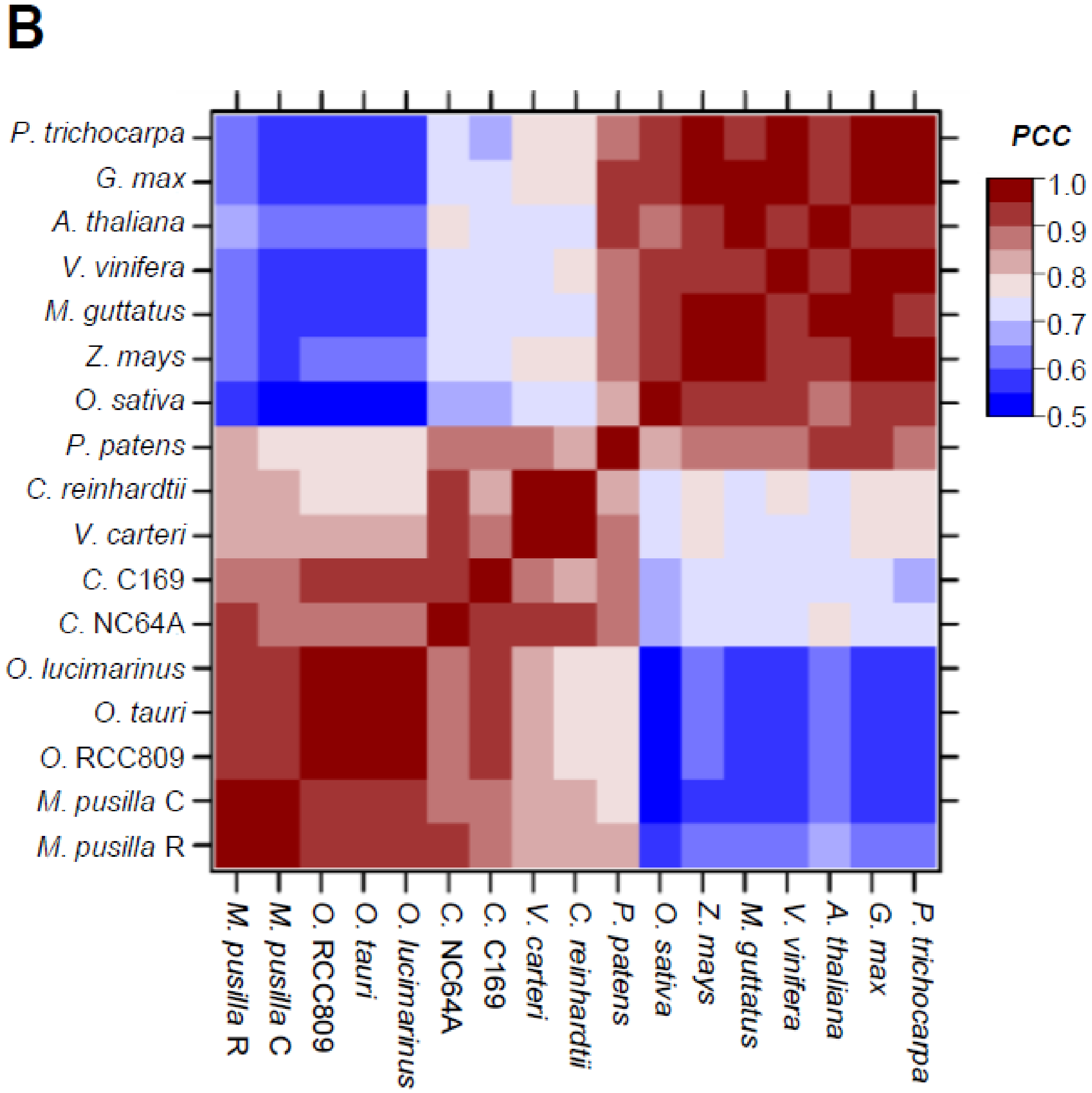


Figure 2.1 (cont'd)

(A). The coverage of Pfam domain annotation in green algal and land plant genomes. Blue dots indicate green algal genomes, and red dots indicate land plant genomes. X axis indicates the numbers of genes in each species, and Y axis indicates the proportion of genes with ≥1 Pfam domains in each species.

(B). Correlation of domain family size profiles among green algae and land plants. The X and Y axis indicate the green algal and land plant species included in this study. The color scale indicates the range of PCC of domain family sizes between two species. PCC stands for Pearson's Correlation Coefficient. Only abbreviations of species names are given here. For full species names, please refer to methods.

Among 5,441 domain families found in nine green algae and eight land plants (see Methods), 3,897 (71.6%) are shared between green algae and land plants, 759 (13.9%) are green alga specific, and 785 (14.4%) are land plant specific. As expected, more closely related species tend to have more highly correlated domain family sizes and there are clearly two distinct clusters, one for green algae and the other for land plants (**Figure 2.1B**). Nonetheless, there is large variation in family sizes among green algal species. This is likely the results of lineage-specific gene gain and loss events. Thus we next asked how lineage-specific gene gains and losses have contributed to the extant domain family sizes in the green algal lineage.

Gene gains and losses throughout green algal lineage

The variation in domain family sizes among green algae suggests extensive lineage-specific evolution of gene families. To find out how gene gain and loss events over time have shaped the domain family size differences among green algal species, we conducted phylogenetic analysis on domain families present in green algae. To address the concern for gene annotation quality in green algae, we examined current annotation and found that only a small number of recognizable protein domains are represented in intergenic regions (11% of domains found in annotated genes in *V. carteri*, 2-6% in other species, See **methods**). Therefore, we used domain sequences in annotated genes for further analysis. Domain sequences from two land plants, *A. thaliana* and *Physcomitrella patens*, were included as outgroups. Among 4,656 domain families containing green algal genes, 4,207 with at least four sequences in green algae and the two land plants were further analyzed and a phylogenetic tree was built for each domain family. After reconciliation of the domain trees with a species tree (Moreau et al., 2012), orthologous groups (OGs) among the green algal and the land plant species were established for inferring gene gain and loss events (**Figure S2.1B**). We have also generated another species tree based on 18s rRNA sequences (**Figure S2.5**) and the two trees are largely similar, except in the *Ostreococcus* lineage. This ambiguity is likely due to the short branch length in this lineage that is difficult to resolve. This tree-based approach is shown to be consistent with similarity-based approach to identify OGs in land plants (Hanada et al., 2008). Overall, gene gain and loss events were frequent in every branch of the phylogenetic tree (**Figure 2.2**). Interestingly, in the green algal lineage, gene loss occurred more frequently than gene gain on every branch, suggesting extensive net gene loss since the green algae-land plant common ancestor.

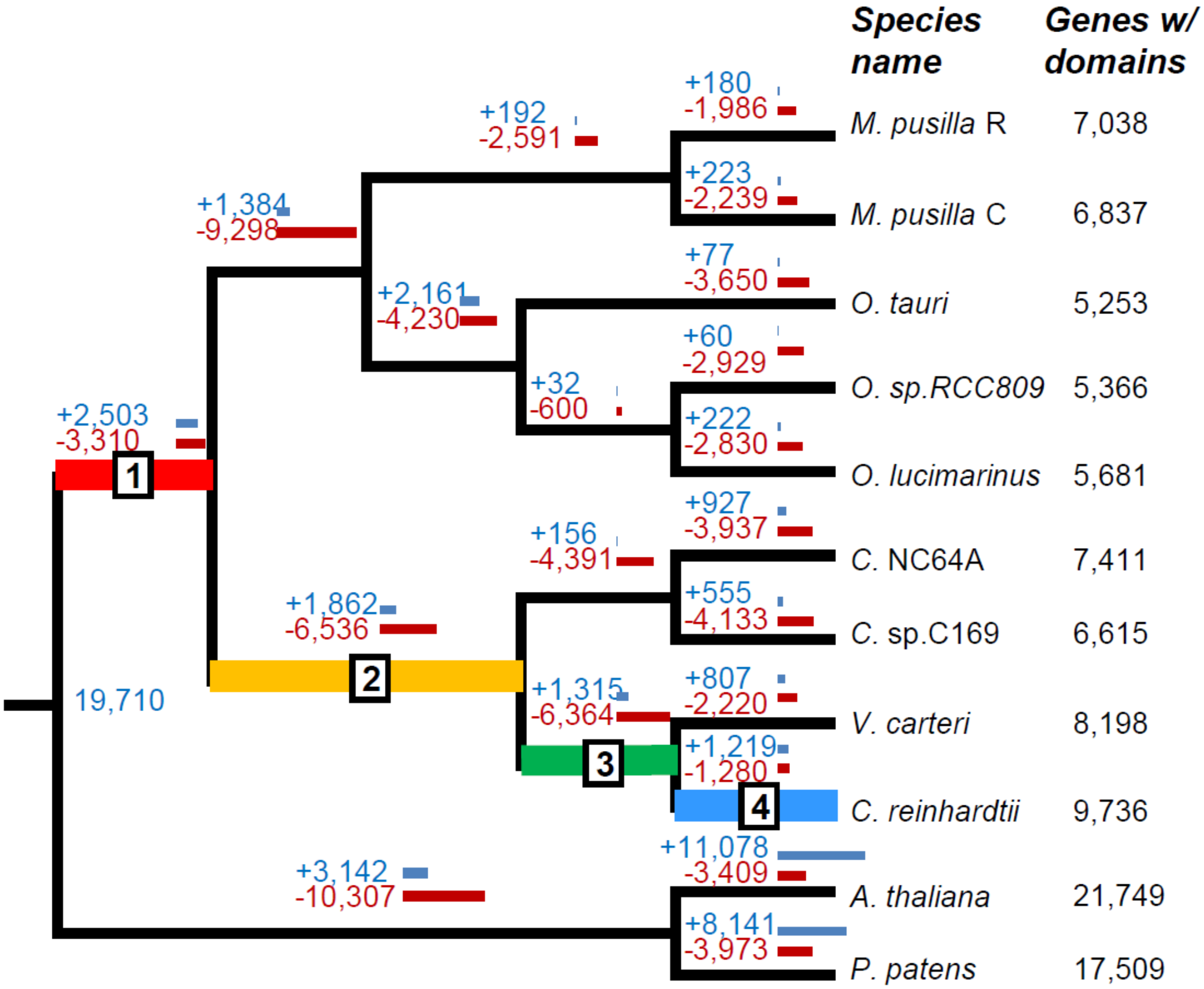


Figure 2.2. Gene gains and losses across land plants and green algae. Blue and red bars indicate number of the gene gain and loss events on each branch, respectively. Species names and numbers of annotated genes with domain in each species are shown on the right. Branch 1, 2, 3, 4 indicate four different time periods in the *C. reinhardtii* lineage evolution. Refer to methods for full species names.

Another reason for an excess of loss events could be because there were errors in phylogenetic inferences. To ascertain that our phylogenetic inference of orthologous groups was robust, we conducted two tests. First, we randomly picked 1,500 gene families for bootstrap studies. Among 50,690 branches, 46.5% have a bootstrap value of ≥ 80% (**Figure S2.3A**). We found that 89.4% of the Volvocales branches (*C. reinhardtii* and *V. carteri*) have a bootstrap value of ≥ 80% but only 47.1% in the *Ostreococcus* branches (*O. sp.*RCC809 and *O. lucimarinus*). Our results showed that shorter branches in general have lower bootstrap values than longer branches (**Figure S2.3B**). Second, we examined 129 domain families with only one copy in each of the 9 green algal and 2 land plant species assuming that this group of domain families has not undergone gene gain and loss. Thus, their gene trees should be identical to the species tree in topology. Because the alternative scenario is possible (gene gains and losses have occurred) in these families, this is a conservative estimate of the orthology inference accuracy. Our results showed a varying degree of consistency in branching between gene trees and the species tree on different branches on the species tree, ranging from 37.2% to 98.3% (**Figure S2.4**). Longer branches on the species tree, indicating longer evolutionary distance, correspond with higher consistency with gene trees. We should emphasize that in branches with high consistency, for example the branch leading to the split of *Micromonas* and *Ostreococcus* with 84% consistent trees, there was a comparable gene loss to gain ratio compared to those with low consistency, indicating that green algae have experienced extensive gene loss.

Several other observations are consistent with the extensive gene loss phenomenon we have inferred in green algal species. First, a substantial number of gene families were lost on lineages leading to *C. reinhardtii* and *O. lucimarinus* after the split from land plants (Cock et al., 2010). Second, in *Ostreococcus,* downsizing of many gene families and gene losses were observed (Derelle et al., 2006; Palenik et al., 2007). Third, a comparative analysis including *Micromonas* and *Ostreococcus* revealed that the common ancestor of Mamiellales had already experienced genome reduction (Worden et al., 2009). Therefore, although lineage-specific gene gains took place, the lineage-specific losses contributed more significantly to the species-specific gene repertoire in green algae.

Pseudogenes in green algae

We found that genes are frequently lost throughout the green algal lineage. Some of the lost genes may still be present in the genome in the form of pseudogenes. Pseudogenes are defined as defunct genomic regions that are fragments of functional genes some with in-frame stops and/or frameshifts. To investigate pseudogenes in green algal genomes, we identified them using a modified pipeline (Zou et al., 2009a). A total of 18,352 pseudogenes were identified in all nine green algal species. In general, pseudogenes are less abundant in green algae than in land plants, even when normalized against genome sizes or total number of genes (**Figure 2.3A**). Among green algae and plants, the number of pseudogenes increases as the genome size and number of genes increase (**Figure 2.3B and C,** Pearson's correlation coefficient of 0.95 and 0.91). Larger domain families also tend to have more pseudogenes in green algae (**Figure 2.3D**), although the correlation is weak (Pearson's correlation coefficient of 0.20). This is possibly due to their small genome sizes and evolutionary pressure towards a more compact genome (Derelle et al., 2006; Palenik et al., 2007).

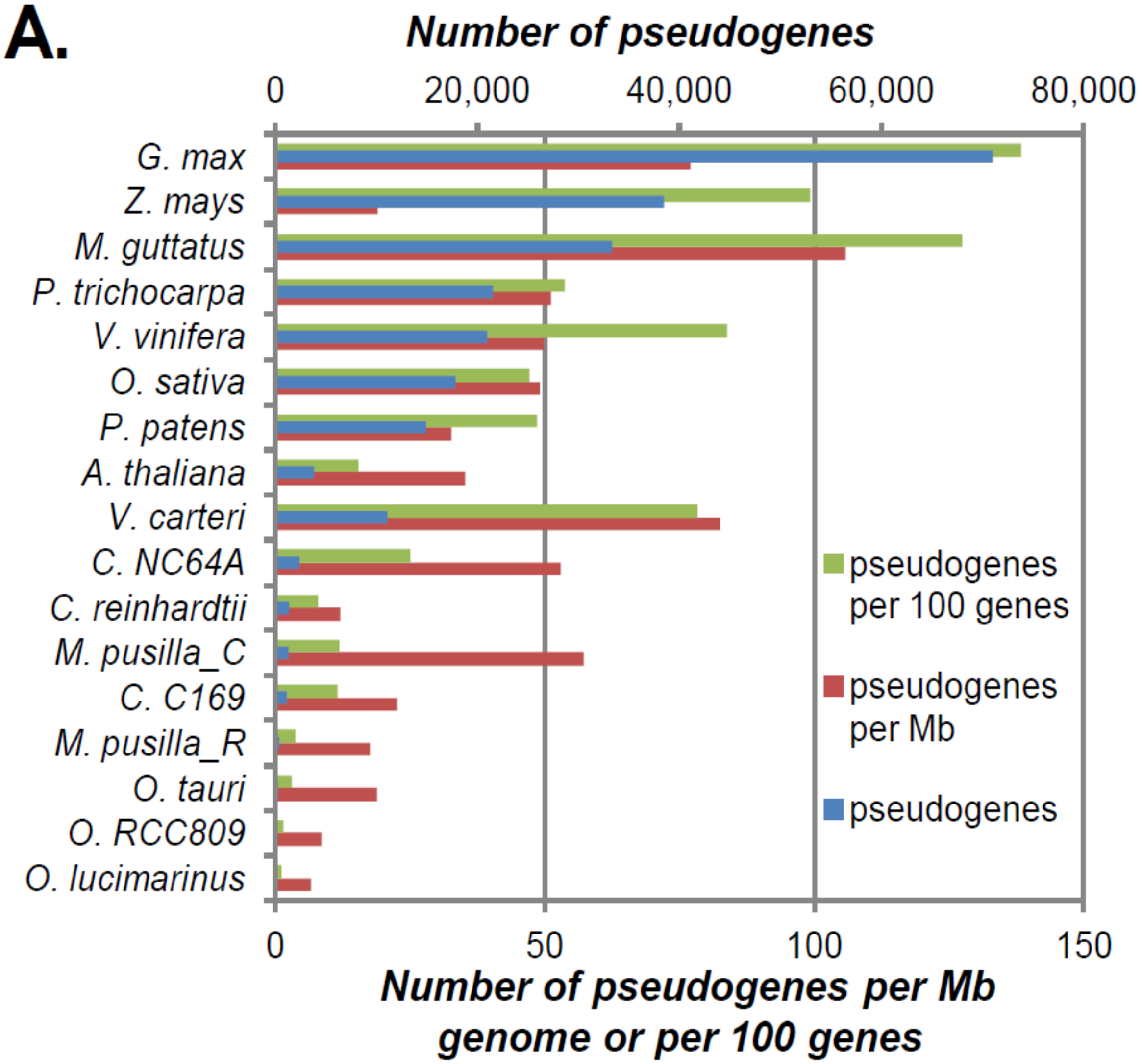


Figure 2.3. **Pseudogene counts in green algae and land plants.**

Figure 2.3 (cont'd)

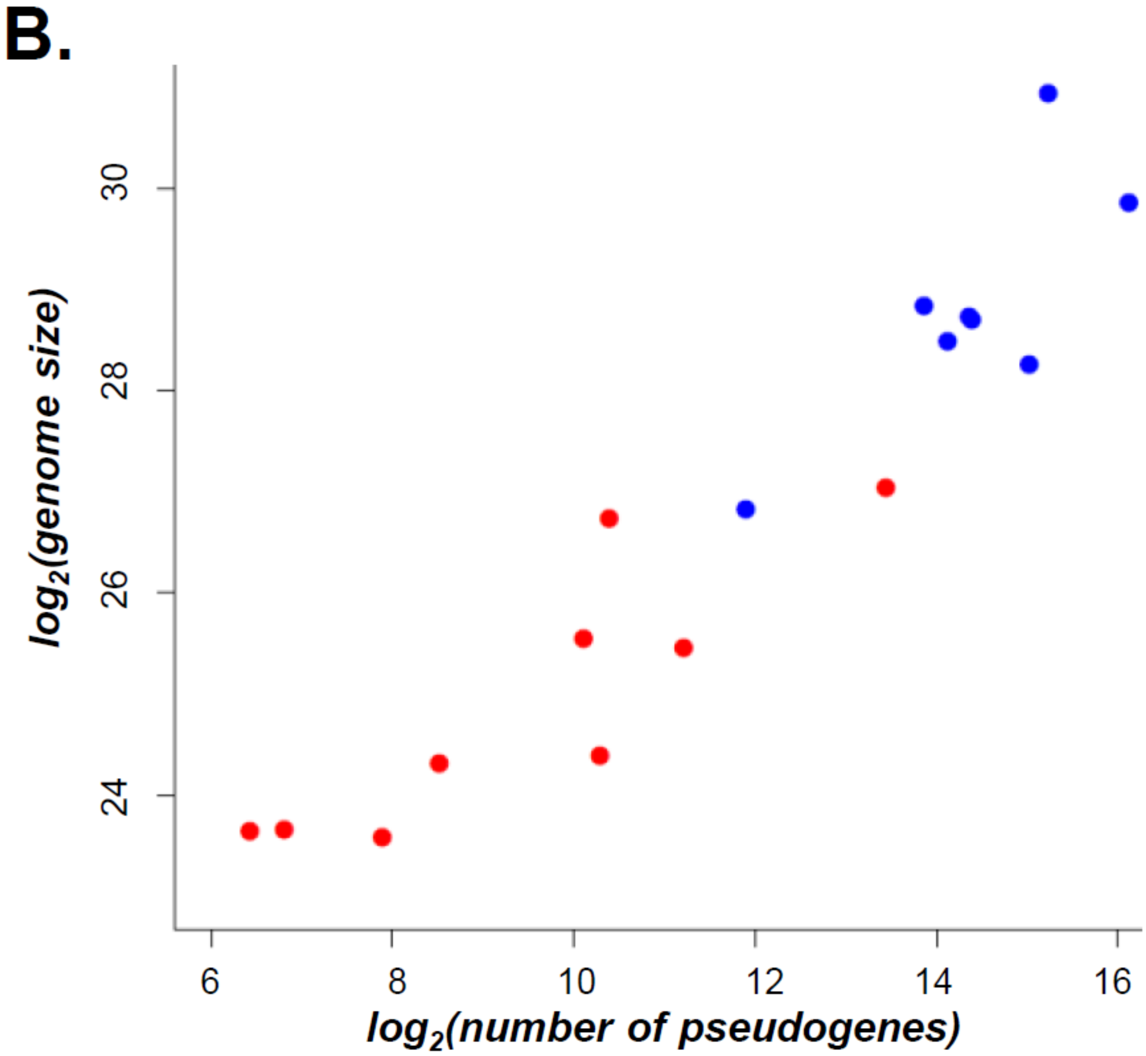


Figure 2.3 (cont'd)

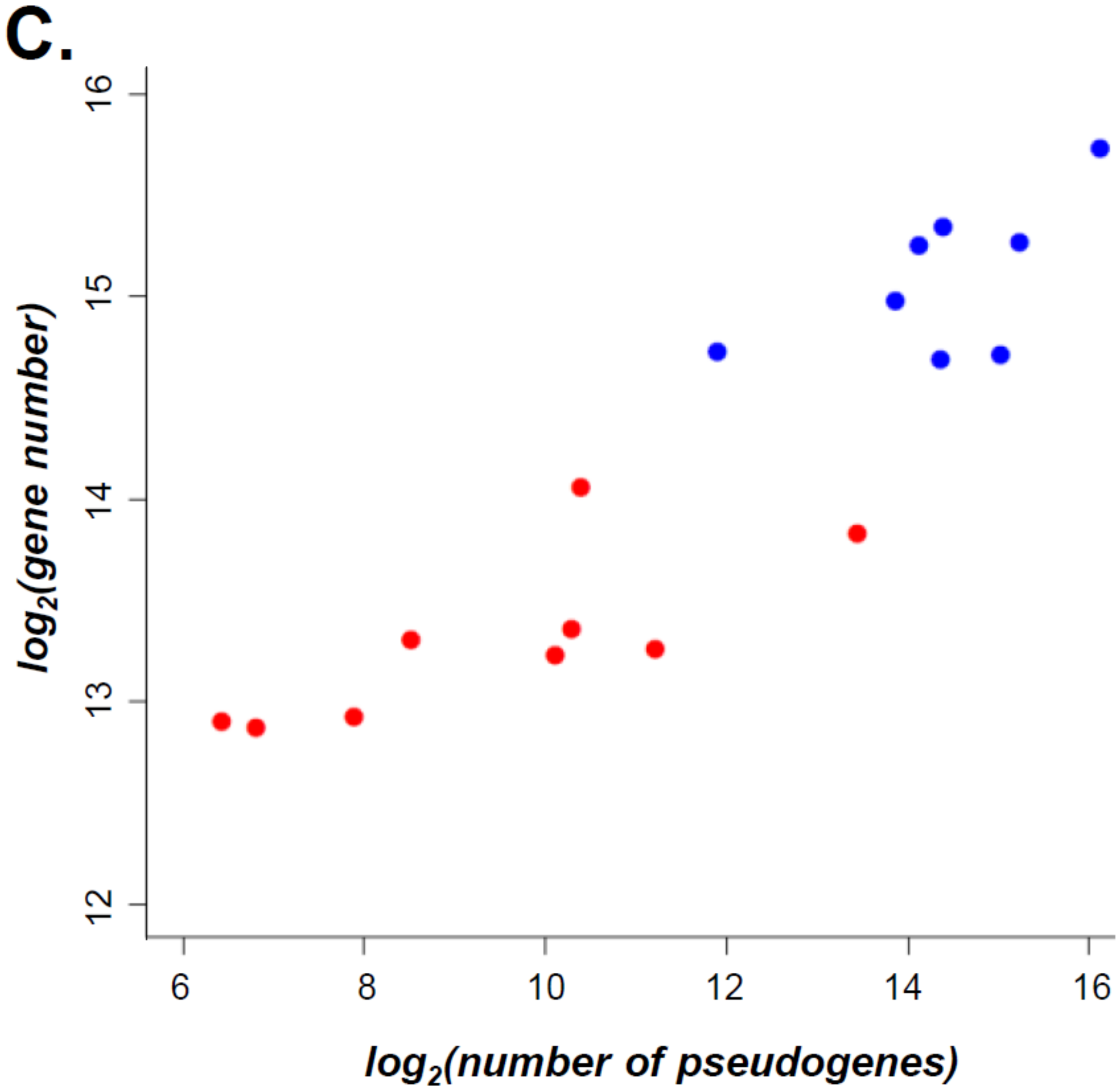


Figure 2.3 (cont’d)

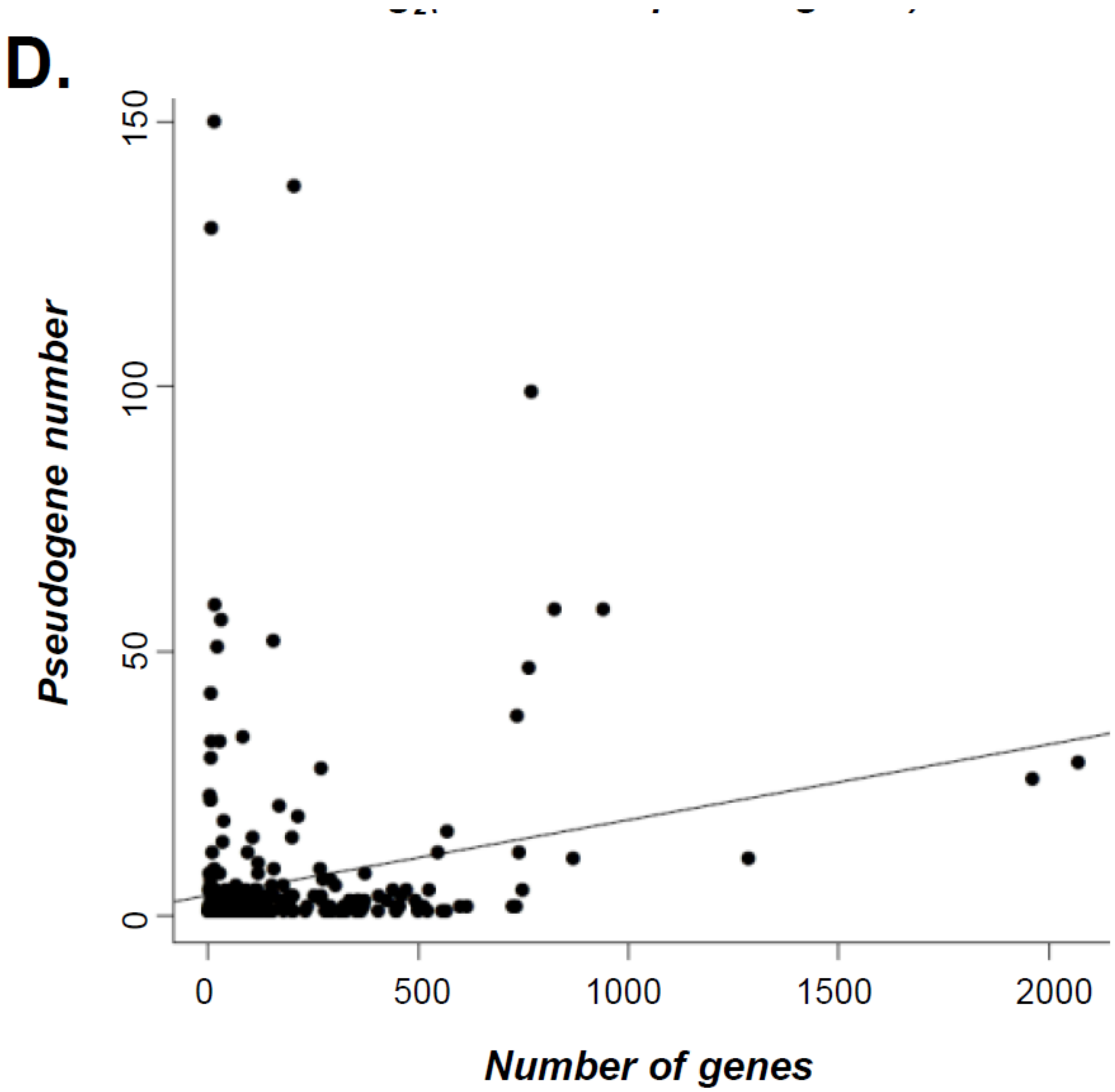


Figure 2.3 (cont'd)

(A). Pseudogenes identified in green algal and land plant genomes. Three numbers are given for each species, including total pseudogene counts, pseudogene counts per Mb genome, and pseudogene counts per 100 genes.

(B). Number of pseudogenes and genome sizes in green algae and land plants. Red dots indicate green algae and blue dots indicate land platns.

(C). Number of pseudogenes and genes in green algae and land plants. Red dots indicate green algae and blue dots indicate land platns.

(D). Number of genes and pseudogenes in each domain family in green algae. Each dot indicates a domain family. X-axis indicates the number of genes in each domain family. Y-axis indicates number of pseudogenes in each domain family.

Functions of retained *C. reinhardtii* genes duplicated after the *C. reinhardtii-V. carteri* split

Although gene losses appear to be more frequent, there are abundant retained duplicates throughout the green algal lineage. However, it is not clear if there is a functional bias among retained duplicates. To determine such a bias, we focused on *C. reinhardtii* since it is a green algal model organism with a relatively well-annotated gene set (Merchant et al., 2007). After annotating the *C. reinhardtii* proteome with GO categories based on sequence similarity (see **Methods**), 5,725 of 17,114 proteins (33.5%) are in ≥1 GO categories. During reconciliation of domain and species trees, the branches in the species tree where duplications took place were inferred as well. Thus we can examine functional biases of retained duplicates in each of the branches leading to *C. reinhardtii*. This information allows us to ask if functions of retained genes were consistent over the course of *C. reinhardtii* evolution. Focusing on the *C. reinhardtii* lineage after its split from the *V. carteri* lineage, 1,817 duplication events (involving 2,682 retained duplicates) took place in the *C. reinhardtii* lineage. Among 13 categories enriched in retained duplicates, they can be classified into the following three types.

The first type of functional categories belong to those involved in stress response (**Figure 2.4A, Table S2.3**), similar to land plants and several other eukaryotes (Lespinet et al., 2002; Hanada et al., 2008). This result is further corroborated by the results from stress expression datasets (detailed in a later section), suggesting that these retained duplicates might have contributed to the species-specific stress response in green algae. One example is the heat shock protein Hsp20 family (PF00011): three duplication events took place in *C. reinhardtii* after the split from *V. carteri*, creating four *C. reinhardtii*-specific duplicates that are responsive to stress (at least one in six conditions mentioned below). The second type of retained duplicate enriched categories is transport including ion, phosphate, and transmembrane transport. One potential explanation is that functional divergence among duplicated transport genes allowed the regulation, affinity, and subcellular location of transporters to be fine tuned. For example, the potassium transporter family (PF02795) in *C. reinhardtii* experienced five duplication events after the split from *V. carteri* and resulted in six *C. reinhardtii*-specific duplicates, all of which are responsive to stress (at least one in six conditions mentioned below). These duplicates might have contributed to the species-specificity of ion transport and stress response in *C. reinhardtii*. In other green algal species, one example of such fine-tuning is the nitrate and ammonium transporters in *Micromonas* (McDonald et al., 2010); however, it is not limited to nutrient management: channelrhodopsins, ion-channels involved in light perception and phototaxis, are also found to be diverse even within the *Chlamydomonas* genus (Hou et al., 2012). The third type of enriched categories is signaling. For example, nine retained duplicates are involved in the synthesis of cyclic nucleotides that are secondary messengers important for the regulation of flagella formation (Hasegawa et al., 1987; Gaillard et al., 2006) and generally activation of ion channels (Ward et al., 2009). Other examples of enriched signaling categories include protein phosphorylation and signal transduction (**Figure 2.4A**, **Table S2.3**). Nucleosome assembly could also be involved in signaling as nucleosome is proposed to be a signaling module in addition to its function of DNA packaging (Turner, 2012). Together, these enrichments indicate duplicates involved in complex environmental interactions and signaling system tend to be retained, potentially because duplicates in these categories provide capacity for responding to new environments and for new routes of regulation.

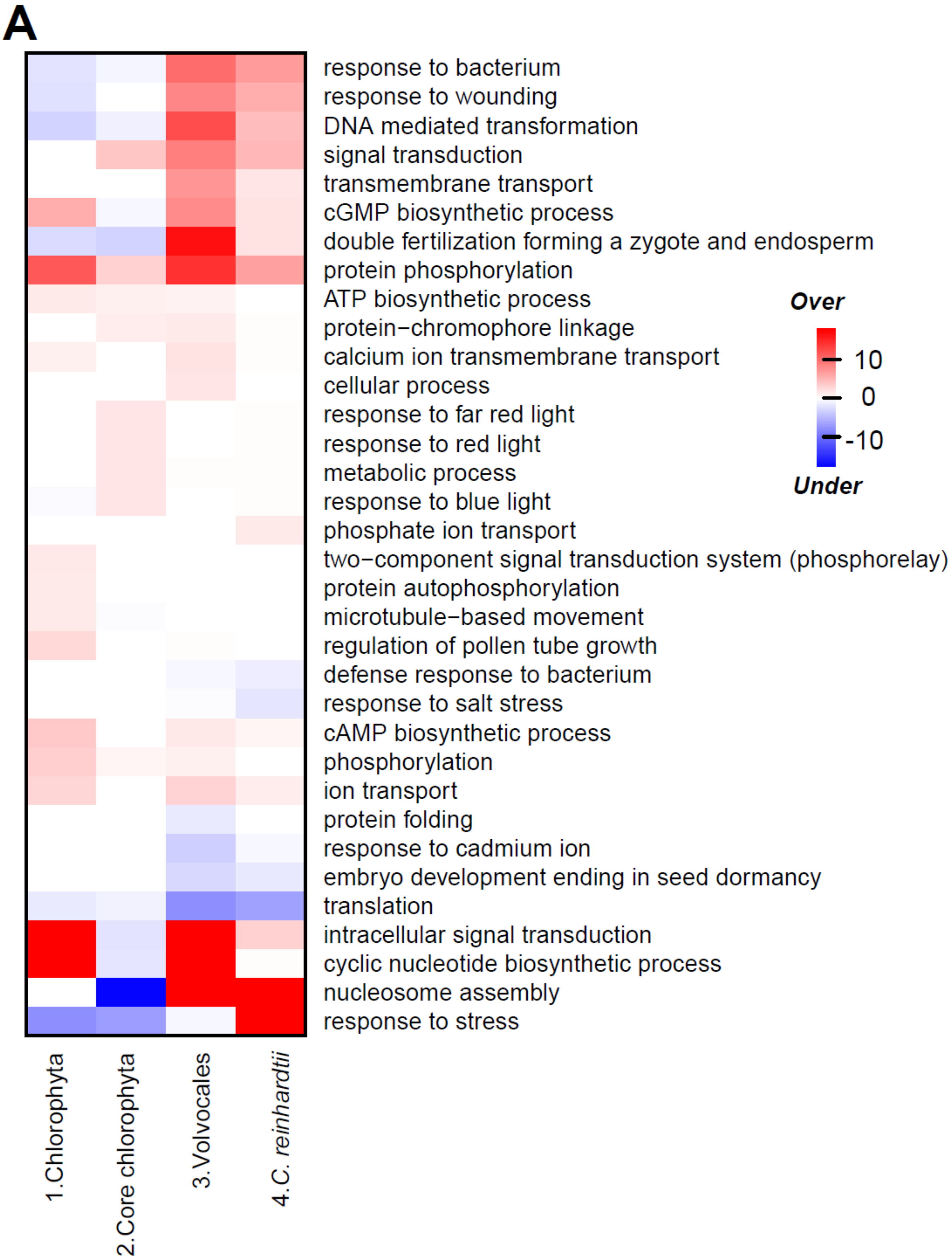
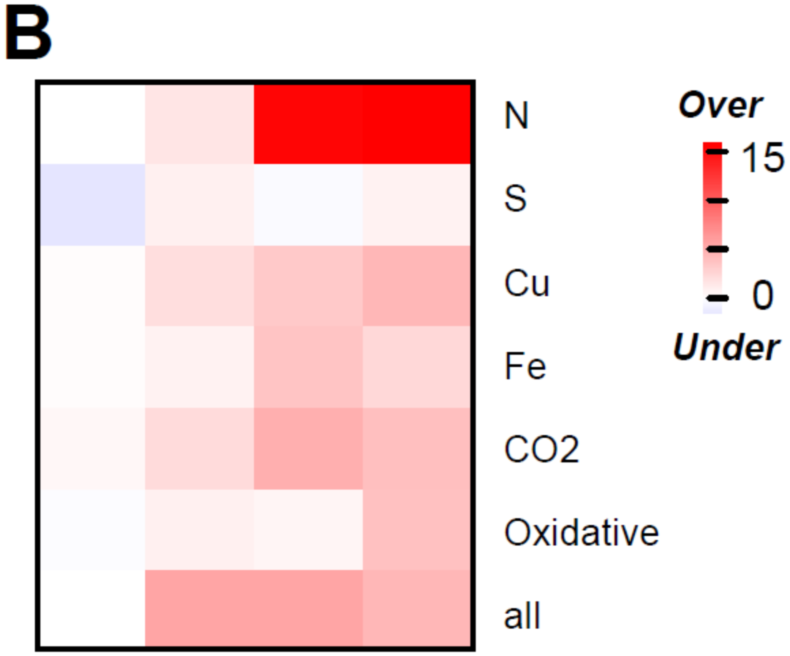


Figure 2.4. Functions and stress responsiveness of retained duplicates in the *C. reinhardtii* lineage.

Figure 2.4 (cont'd)



(A) Functions of retained duplicates. Columns indicate four phylogenetic branches as in **Figure 2.2** with the most recent on the far right. Rows indicate GO biological processes terms with over or under-represented numbers of retained duplicate genes in at least one branch. Values shown are -$\log_{10}$(p) for over-represented and $\log_{10}$(p) for under-represented categories, respectively. **(B)** Stress responsiveness of retained duplicates**.** Stress conditions shown are N (nitrogen deprivation), S (sulfur deprivation), Cu (copper deprivation), Fe (iron deprivation), CO2 (CO_2 deprivation), oxidative (oxidative stress), and all (all six conditions combined). Values shown are -$\log_{10}$(p) for over-representation and $\log_{10}$(p) for under-representation. Branches are the same as **(A)**.

Consistency of functional biases over the course of *C. reinhardtii* evolution

The above discussion was on the *C. reinhardtii* lineage after its split from the *V. carteri* lineage. In the process of green algal evolution, some duplicates may be preferentially retained throughout while others may be period-specific due to the ever-changing environment. In addition, different kinds of genes may have different longevities thus resulting in differences in enriched categories over time. To distinguish these possibilities, the timing of duplication of each retained gene was pinpointed to an internal or external branch that led to *C. reinhardtii*. The functions of ancestral genes at the time of duplication were assumed to be the same as those of their descendants. First we focus on categories that are branch specific. The branches are numbered as in **Figure 2.2 and 2.4A**. Stress response is enriched in just branch 4, while nucleosome assembly is enriched in 3 and 4. The expansion of stress related gene families in just the youngest branch may be because environmental condition would constantly change in the evolutionary history of the *C. reinhardtii* lineage. Thus previously retained duplicates conferring selective advantage may not be adaptive due to exposure of green algae to ever-changing environments.

In contrast to stress, we found that signaling categories are consistently enriched for retained duplicates over time. Further, retained duplicates in branch 3 and 4 tend to play markedly similar roles as branch 1, particularly those in ion transport, signaling and cyclic nucleotide synthesis. Retained duplicates in branch 2 were associated with several types of light response and signal transduction. (**Figure 2.4A**, **Table S2.3**). This result suggests that gene families involved in signaling and to a lesser degree transport are constantly expanded in a lineage-specific manner, and more generally indicates consistent innovation in interaction with the environment. Further, these enrichments are consistent with other eukaryotes (Lespinet et al., 2002; Hanada et al., 2008). Three categories involved in double fertilization, pollen tube growth, and embryo development are not relevant to the single-cellular *C. reinhardtii*. They are likely annotated by the sequence similarity to plant proteins by Blast2go (Conesa et al., 2005).

Functional categories enriched in conserved genes and genes associated with pseudogenes

In addition to retained duplicates, we defined "conserved genes" as those with the same copy number (1 to 3) in among green algal species. The categories enriched in conserved genes in *C. reinhardtii* were few and largely involved in housekeeping functions, including GO terms such as translation and two other ribosome-related terms. Other terms enriched in conserved genes included tetrapyrrole synthesis, and photosystem I reaction center (**Table S2.2**). In addition to functional bias in retention and conservation, we would also like to examine functional bias in pseudogenization. To find out if there is such bias, we tested GO enrichment in genes associated with pseudogenes, defined as genes that are the closest relatives to pseudogenized duplicates. We found that stress response is the only biological process enriched in genes associated with pseudogenes, indicating that, in addition to their high rate of retention, these genes have frequently undergone gene loss, a finding similar to a pseudogene study in *A. thaliana* (Zou et al., 2009a). This coupling between a high birth as well as high death rate is likely because genes involved in responding to specific stress conditions could become unnecessary when that condition does not persist.

Retained duplicates and their stress responsiveness

Multiple stress conditions lead to increased oil content in microalgae (Hu et al., 2008). In addition, we showed that stress response categories were enriched in retained duplicates in the *C. reinhardtii* lineage. However, GO annotation in *C. reinhardtii* is established solely based on computational approaches and has only 33.5% of genes annotated. To address these issues, we asked if retained genes tend to be stress responsive compared to singletons under six conditions: deficiency in N, sulfur, iron, copper, and CO2, and oxidative stress (González-Ballester et al., 2010; Miller et al., 2010; Castruita et al., 2011; Fang et al., 2012; Page et al., 2012; Urzica et al., 2012). In all stress conditions except sulfur deprivation, stress responsiveness tends to be over-represented among retained genes (p-value $\leq$0.01), (**Figure 2.4B, Table S2.1**). When combining all six RNA-seq datasets and defining stress responsiveness of a gene as being responsive in $\geq$1 conditions, retained duplicates still tend to be stress responsive (**Figure 2.4B, Table S2.1**). This result corroborates our conclusion that retained duplicates tend to be in the stress response functional categories.

According to functional category analysis, stress response was not enriched among retained duplicates except in branch 4, which is the most recent branch (**Figure 2.4A**). Consistent with this finding, analysis of stress expression data also revealed that retained genes tend to be stress responsive in more conditions if they were duplicated more recently (**Figure 2.4B**), with 5 conditions in branch 4 and gradually declining to none in branch 1. This is also consistent with our finding that the closest relatives of pseudogenes tend to be involved in stress response, again reinforcing the idea that, although stress response genes tend to be retained at a higher rate compared to genome average, they tend to be more short-lived.

Stress response evolution post duplication in the *C. reinhardtii* lineage

Retained duplicates tend to be stress responsive (**Figure 2.4 A and B**), suggesting that gene duplication might provide a source for innovations in response to stress. This possibility then leads to the question how often innovation, defined as gain of stress responsiveness from a non-responsive ancestral gene, occurred in green algae. To answer this question, we first integrated phylogenetic data and stress related expression datasets to infer the ancestral response state (**Figure S2.1C**). With ancestral states, the gain and loss events of duplicated genes can then be distinguished (Zou et al., 2009b). Stress response states are defined as U (up-regulated by ≥2 fold, false discovery rate ≤ 5%), D (down-regulated by ≥2 fold, false discovery rate ≤ 5%), and N (not significantly changed). Only the ancestral nodes leading immediately to extant genes were included in our analysis to avoid complications from predicting responses in nested branches (Zou et al., 2009b).

We define four “evolutionary events” (retention, gain, loss, or switch) based on a comparison of stress response states between an extant gene and its most immediate ancestral gene node for each of the six stress conditions. We found that 6,330 comparisons (10.9%) were relevant because they involved either extant and/or ancestral stress responsive genes. Among the six stress conditions, the median number of events involving retention of the ancestral stress response is 48% (32% U -> U, and 16% D -> D). We also found that 35% were response loss events (23% U -> N, and 12% D -> N). Meanwhile, comparatively fewer events (13%) involved functional gain (9% N -> U, and 4% N -> D) and even fewer events (2%) involved functional switch (1% D -> U, and 1% U -> D) (**Figure 2.5A**). To find out if younger duplicates tend to retain their ancestral response state, we analyzed the relative frequencies of all four stress response evolution scenarios against time, using synonymous substitution rate (Ks) as a proxy for time (**Figure 2.5B-E**). Regardless of the Ks value, it is generally true that the rates of stress response evolution scenarios are retention > loss > gain >> switch. However, the relative abundance of each scenario changed overtime (**Figure 2.5B-E**). When $0 \leq Ks \leq 1.2$, the rate of retention decreased from 63.8% to 45.5% (median of all conditions), and it remains relatively stable afterwards (**Figure 2.5B**). On the contrary, the rate of loss increased from 22.2% to 37.5% when $0 \leq Ks \leq 1.2$, and remains relatively stable after that (**Figure 2.5C**). These results show that younger duplicate genes tend to retain the ancestral stress response state, similar to *A. thaliana* (Zou et al., 2009b). The rate of functional gain peaked at Ks = 0.9, and decreased thereafter, as similar pattern was observed in *A. thaliana* (Zou et al., 2009b).

Nonetheless, compared to the results of a similar study in *A. thaliana* (Zou et al., 2009b), *C. reinhardtii* has comparable rates in response retention, loss, and switch, but a much higher rate of functional gain (13% versus 6%). Note that we examined only six abiotic conditions in *C. reinhardtii* compared to the 16 conditions encompassing biotic and abiotic stress environments in the *A. thaliana* study (Zou et al., 2009b). Considering that the relative abundance of the evolutionary scenarios is similar across divergent conditions (Zou et al., 2009b), having more data will likely not contribute to significant changes in the gain rate estimate in either direction. We should also emphasize that, regardless of the Ks value, *C. reinhardtii* consistently has a higher rate of functional gain (**Figure 2.5D**) when compared to *A. thaliana* (Zou et al., 2009b). Taken together, our findings indicate that innovation occurs more often in *C. reinhardtii* than in *A. thaliana* in the context of stress response. The excess gain events could be due to the fact that *C. reinhardtii* is a single cellular organism and has a shorter life cycle while encounters more diverse environmental conditions, as one would expect that a shorter life cycle would lead to a higher number of mutations per unit time compared to species with a longer generation time, thus providing more raw material for adaptation to occur upon.

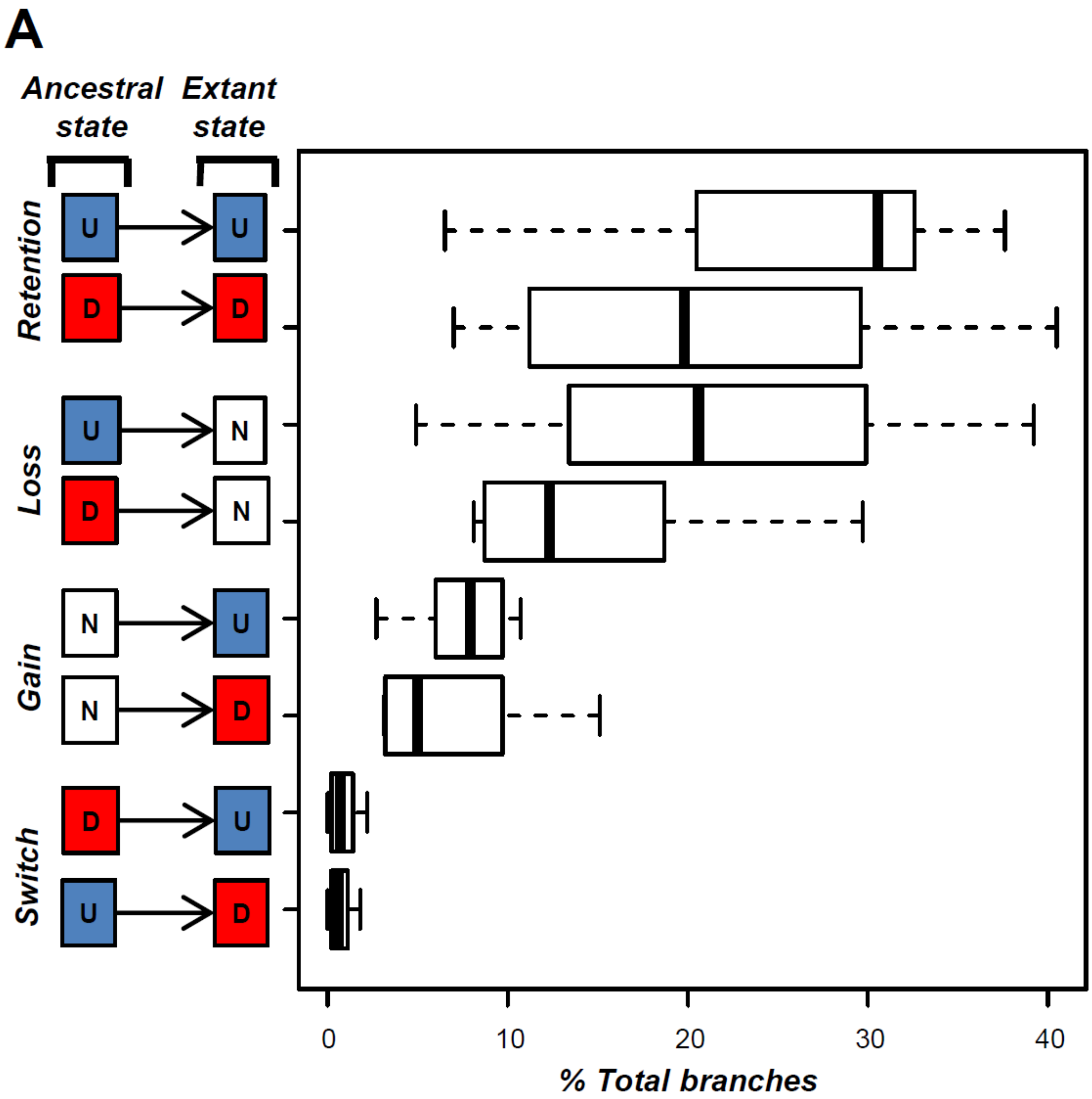


Figure 2.5. Stress response evolution scenarios of *C. reinhardtii* duplicates compared to their ancestral genes.

Figure 2.5 (cont'd)

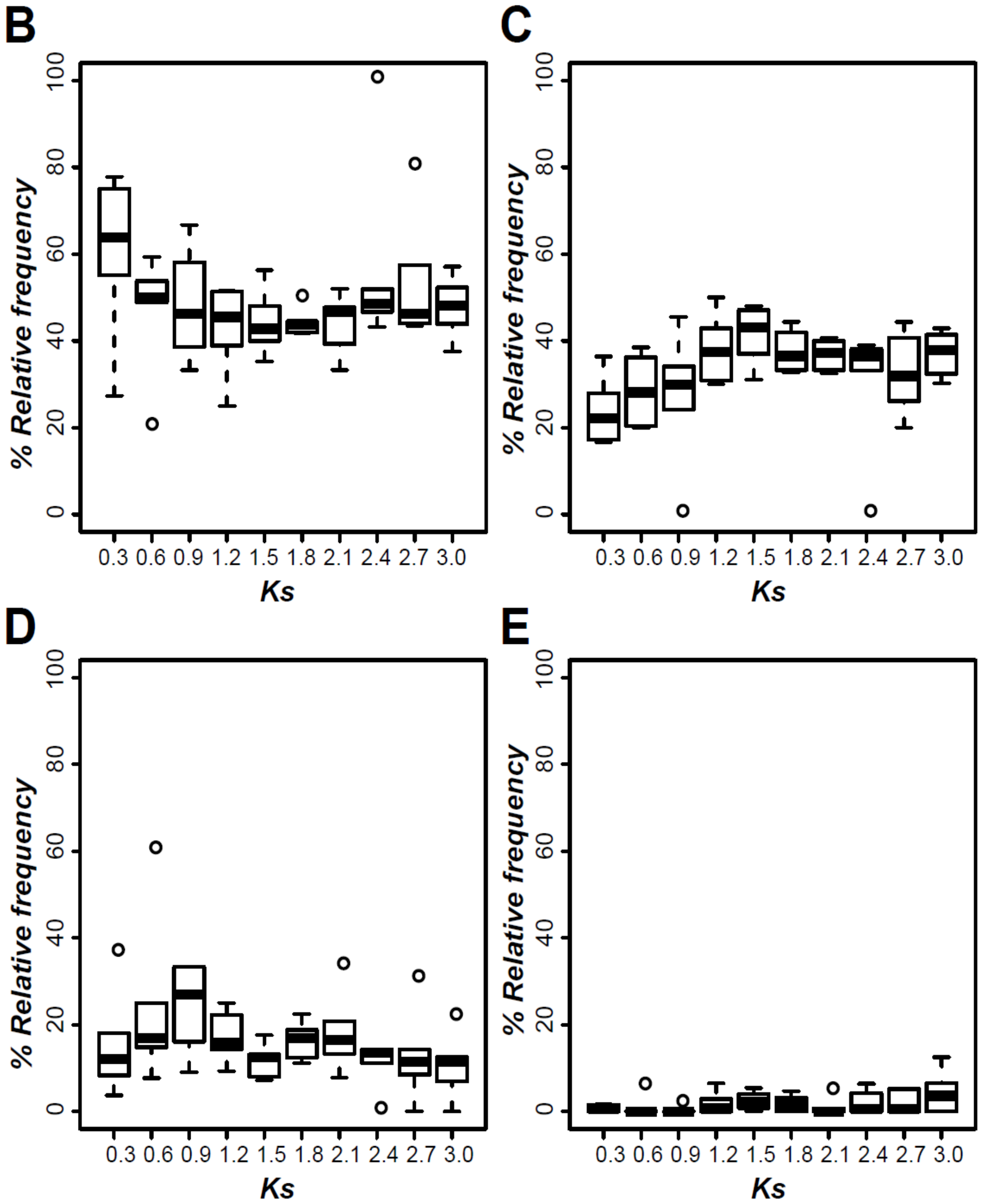


Figure 2.5 (cont’d)

(A). There are four scenarios as indicated on the far left column. The ancestral state column shows the reconstructed ancestral response state and the extant state column shows the extant gene response state. ‘U’, ‘D’, and ‘N’ indicate up-regulated, down-regulated, and not-regulated under a stress condition, respectively. On the right, the preponderance (in percent total) of each scenario for a given stress condition is shown. The percent of each scenario for six conditions are summarized in a boxplot for each ancestral-extant combination.

(B-E). stress response evolution among *C. reinhardtii* paralogs using Ks as a proxy for time. X-axis indicates Ks values, and Y-axis indicates the relative frequency of each stress evolution scenario. **(B).** stress response retention. **(C)**. loss. **(D)**. gain. **(E)**. switch.

CONCLUSION

Our analysis of gene family evolution, functional evolution and pseudogenization in the green algal lineage complement previous studies in other eukaryotes, reinforcing that the association between lineage-specific evolution and stress response is a common feature of eukaryotes. This association is likely due to the selective pressure under ever-changing environment. In this scenario, stress gene duplicates were frequently under positive selection. In addition, the high rate of innovation in acquiring abilities to respond to stress in *C. reinhardtii* duplicates contributes to a highly diverse stress responsive gene repertoire that can potentially be adaptive.

The model organism *C. reinhardtii* is used to study the mechanism of stress response and lipid accumulation in green algae although it is not a direct candidate for biofuel production (Miller et al., 2010). As the general metabolic changes under stress might be similar across divergent micro-algal species (Vieler et al., 2012), the particular genes involved in stress response could be quite different as they are shaped by lineage-specific family expansion and subsequent gain-of-function events. Such species-specificity cannot be deciphered when focusing on one model organism. Therefore, in addition to focusing on the well-established model organism of *C. reinhardtii* to discover the general biology of stress response in green algae, it is necessary to investigate diverse green algal species considered for biofuel production to discover their uniqueness in the relationships between stress response and lipid accumulation, towards the ultimate goal of finding the perfect alga for biofuel production.

METHODS

Protein domains in green algal species

Genome and protein sequences of nine green algal species (see Introduction) and eight land plant species were obtained from DOE Joint Genome Institute (www.jgi.doe.gov) and Phytozome (www.phytozome.net, version 7.0). The land plant species are *Populus trichocarpa*, *Glycine max*, *A. thaliana*, *Vitis vinifera*, *Mimulus guttatus*, *Zea mays*, *Oryza sativa*, and *P. patens*. HMMER (Finn et al., 2011) was used with trusted cutoff to scan algal and plant protein sequences for Pfam domains (Finn et al., 2010). Fisher's exact test was used to test the enrichment of GO categories in conserved domain families in green algae and expanded domain families in green algae using GO annotation on Pfam domains (ftp.sanger.ac.uk).

To assess the completeness of gene annotation in green algae, we identified intergenic regions with coding potential and domain presence. Green algal protein sequences were aligned to green algal genome sequences with BLAST (tblastn, E-value threshold ≤ 1e-5, (Altschul et al., 1990)). Matches with ≥30 amino acids long and ≥40% identity were kept and the matching genomic sequences were translated into peptide sequences using Genewise (Birney et al., 2004). After filtering out sequences with frameshifts and identical sequences, the rest were consolidated by concatenating sequences with ≥5 amino acid overlaps. Protein domains in the concatenated sequences were identified with HMMER using trusted cutoff. Domains overlapping with domains identified in annotated genes or pseudogenes (see later section) were eliminated. Overlapping domain sequences were merged and identical sequences removed (**Figure S2.1A**).

Identification of missing domains in green algal genomes with variable annotation qualities

The robustness of our domain family analysis is fundamentally dependent on the quality of gene annotation in green algae. For example, an un-annotated domain would lead to a false prediction of gene loss. Most of the green algal genome annotations are automatically generated using computational approaches with various degrees of manual intervention. Thus, the quality of the annotation is likely highly variable and some genes and thus protein domains may not be annotated. Given our goal is to evaluate the gain and loss patterns of algal domain families, these potentially missing domains are false negatives that can have a significant impact on our subsequent studies. Thus, to identify domains in the genomes that are missed by current annotation, we aligned all the protein sequences of nine green algal species to their genomes to identify all sequences with coding potential.

A total of 810,578 matches were identified. The matching genomic sequences were translated into peptide sequences using Genewise (Birney et al., 2004). After removing redundant sequences that were identical in their entirety and merging overlapping sequences with identical amino acids in the overlapping region, 474,350 sequences remained. Overlapping sequences that were not identical in the overlapping regions were not removed since they might contain different domains. We identified 238,446 domains from these non-redundant sequences.

After removing identical domain sequences, merging overlapping domain sequences, removing domain sequences overlapping with domains in annotated genes and pseudogenes identified using a published pipeline (Zou et al., 2009a), 6,432 domain sequences remained and are referred to as domains in un-annotated regions, of which one-third are in *V. Carteri* (**Figure S2.2B**). They are likely to be domains residing partially or completely in intronic or intergenic regions. Out of the 1,985 *V. carteri* domains, 862 are retrotransposon related, while a total of 223 domains in all eight other species are retrotransposon related. Most of the domains in un-annotated regions are shorter than the average of annotated domains (**Figure S2.2A**), and the number is very small compared to the annotated domains in each species (**Figure S2.2B**). For these reasons, we conclude that despite the automated nature of the current green algal annotations they contain most of the known Pfam domains in green algae and further analysis was done only using the annotated domains.

Defining green algal orthologous groups (OGs) and lineage-specific expansions (LSEs)

To build a phylogeny for each protein domain X, sequences of domain X were extracted from full length protein sequences of the green algae and land plants and aligned using MAFFT (Katoh et al., 2002). Using the alignments generated, the phylogeny of each domain family was inferred using RAxML (Stamatakis, 2006) with parameters -f d -m PROTGAMMAJTT. Domain family trees were reconciled with a previously published species tree (Moreau et al., 2012) using NOTUNG (Chen et al., 2000) to identify orthologous groups. We have also used the 18s rRNA sequences of two land plants and all nine green algal species to built a species phylogeny using RAxML (Stamatakis, 2006) with parameters -f a -x 12345 -p 12345 -# 1000 -m GTRGAMMA.

Given the topology is highly similar, only the previously published phylogeny was used. For large domain families that RAxML run didn't finish in 160 hours, a neighbor joining tree was built with PHYLIP (Felsenstein, 2005) and the distance trees were broken down to smaller sub-clusters with ≥ 4 genes and ≤ 300 genes, and distance to root ≥ 0.05. Each sub-cluster was regarded as a "sub-family". Domain sequences of sub-family members were used to identify orthologous groups with the same approach as above (**Figure S2.1B**). To test the robustness of our orthology inference, bootstrap values were acquired on 1,500 randomly chosen domain families using RAxML (-f a -# 400 -m PROTGAMMAJTT -x 12345). Bootstrap values and branch length (from a species tree (Moreau et al., 2012)) are plotted on **Figure S2.3B**.

To functionally annotate the *C. reinhardtii* genome, protein sequences were first aligned to the nr protein database (www.ncbi.nlm.nih.gov) to identify putative homologs with an e-value cut-off of 1e-5. GO annotation was then inferred using blast2go based on genes with GO entries in the nr database (Conesa et al., 2005). Fisher's exact test was used to test the enrichment of retained duplicates in GO categories and a significantly enriched category had false discovery rate ≥5%.

Pseudogenes in green algae

Pseudogenes of green algae were predicted using a previously defined pipeline (Zou et al., 2009a) with some modifications: 1) we used the whole genomes for the BLAST search and filtered out hits overlapping with genes as opposed to using the intergenic sequences only; 2) we made the pseudoexon merging recursive so that multiple pseudogenes can be derived from one pseudoexon cluster provided that the merged pseudogenes do not overlap with each other; 3) RepeatMasker (ver. 3.3.0) was run after the pseudogene pipeline on pseudogene sequences using Viridiplantae repeats (Cutoff = 300, Divergence = 30) and pseudogenes with hits other than "Simple_repeat", "Low_complexity" and "Satellite" were removed from the dataset; 4) to control for false positives due to proteins being split between contigs, we removed pseudogenes that were within the 95th percentile genomic intron length from the end of a contig if they do not have disabling mutations, defined as a frame shift or premature stop codon. A "high confidence pseudogene" has one or more disabling mutations and has the following conditions satisfied: the distance between the pseudogene and the ends of the contig larger than the distance between the matching region and the end of the protein plus 95th percentile genomic intron length on both ends.

Inferring ancestral stress response state

To identify genes responsive to various stress conditions, *C. reinhardtii* RNA-seq datasets of N (Miller et al., 2010), S (González-Ballester et al., 2010), Fe (Page et al., 2012), Cu (Castruita et al., 2011), and CO2 deficiency (Fang et al., 2012), and oxidative stress (Urzica et al., 2012), were obtained from the Sequence Read Archive at NCBI (http://www.ncbi.nlm.nih.gov/sra). Reads were aligned to *C. reinhardtii* genome using Tophat (Trapnell et al., 2009), with following options: -i 13 -I 8712 -g 1. For each dataset, differential expression was determined using EdgeR (Robinson et al., 2010) with a threshold of fold change

≥ 2 and false discovery rate ≤ 5%. Domain family phylogenies and extant stress response states were combined to infer the ancestral stress response states of *C. reinhardtii* genes using BayesTraits (Pagel, 1999). Three discrete functional states were defined as 1) up-regulation (u, by ≥ 2 fold and FDR ≤ 0.05), 2) down-regulation (d, by ≥ 2 fold and FDR ≤ 0.05), and 3) no-regulation (n). The ancestral states were estimated using multistate, maximum likelihood and most recent common ancestor (MRCA) options for each family phylogeny and each stress condition. Only ancestral states with a posterior probability > 0.5 were used for subsequent analysis. BayesTraits cannot be used in cases that all genes in one tree had same state, and we assumed in that case all ancestral genes had the same state as the extant ones. Ancestral gene response states were compared to extant gene response states to infer innovations and losses in stress response (**Figure S2.1C**).

ACKNOWLEDGEMENTS

This project was supported by National Science Foundation grants IOS-1126998 and MCB-0929100 to SHS, Michigan State University College of Natural Science Dissertation Continuation Fellowship to GW, and Deutsche Forschungsgemeinschaft grant IRTG-1525 to AKD. We would also like to thank Andreas P.M. Weber for proof-reading the manuscript.

APPENDIX

APPENDIX

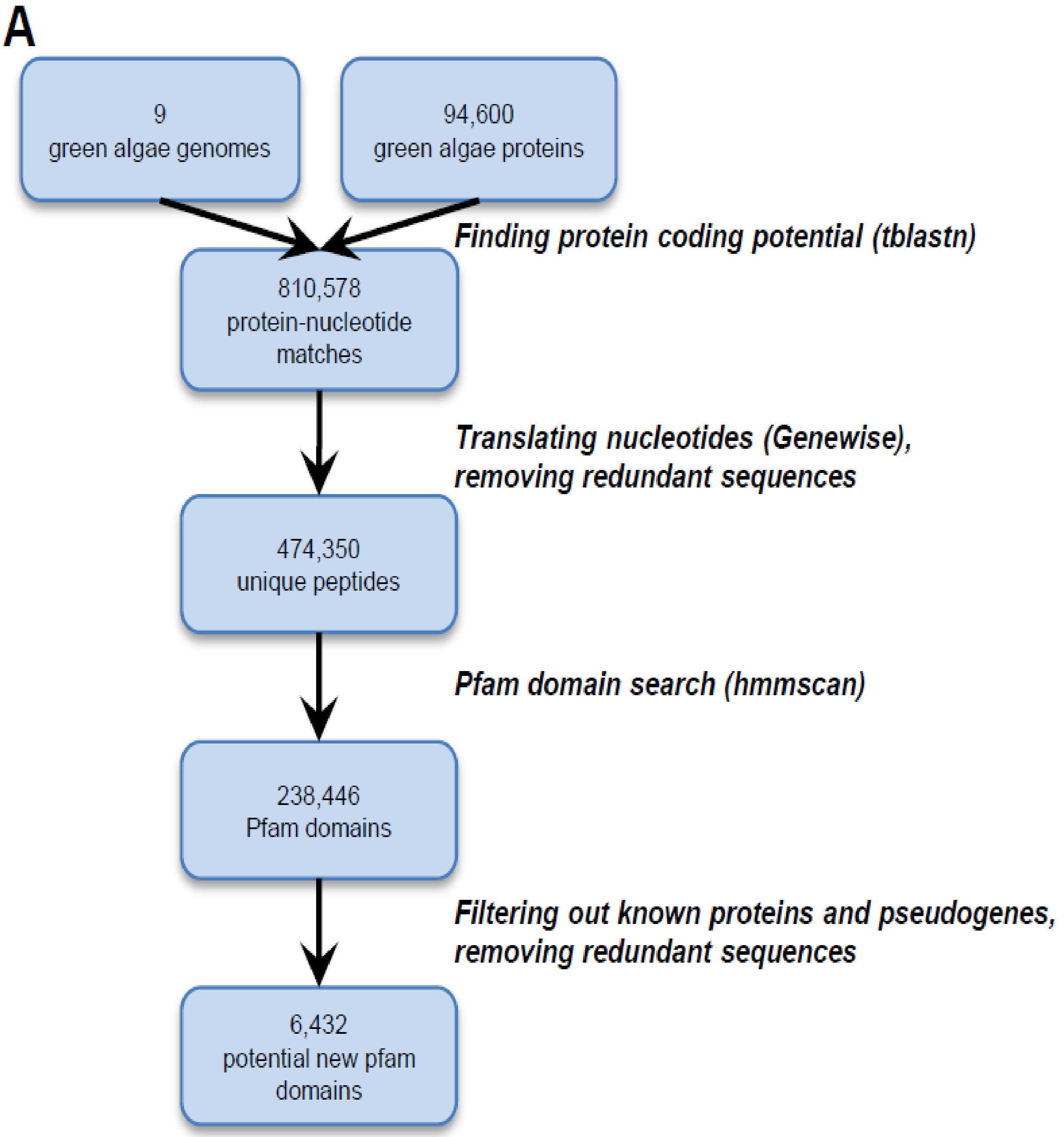

A
9
green algae genomes
94,600
green algae proteins
Finding protein coding potential (tblastn)
810,578
protein-nucleotide matches
Translating nucleotides (Genewise),
removing redundant sequences
474,350
unique peptides
Pfam domain search (hmmscan)
238,446
Pfam domains
Filtering out known proteins and pseudogenes,
removing redundant sequences
6,432
potential new pfam domains

Figure S2.1. Analysis pipelines. (A). Pipeline used to identify missing domains. **(B)**. Pipeline for identifying retained duplicates in green algal lineage. **(C)**. Pipeline for ancestral stress response state inference in *C. reinhardtii*.

Figure S2.1 (cont'd)

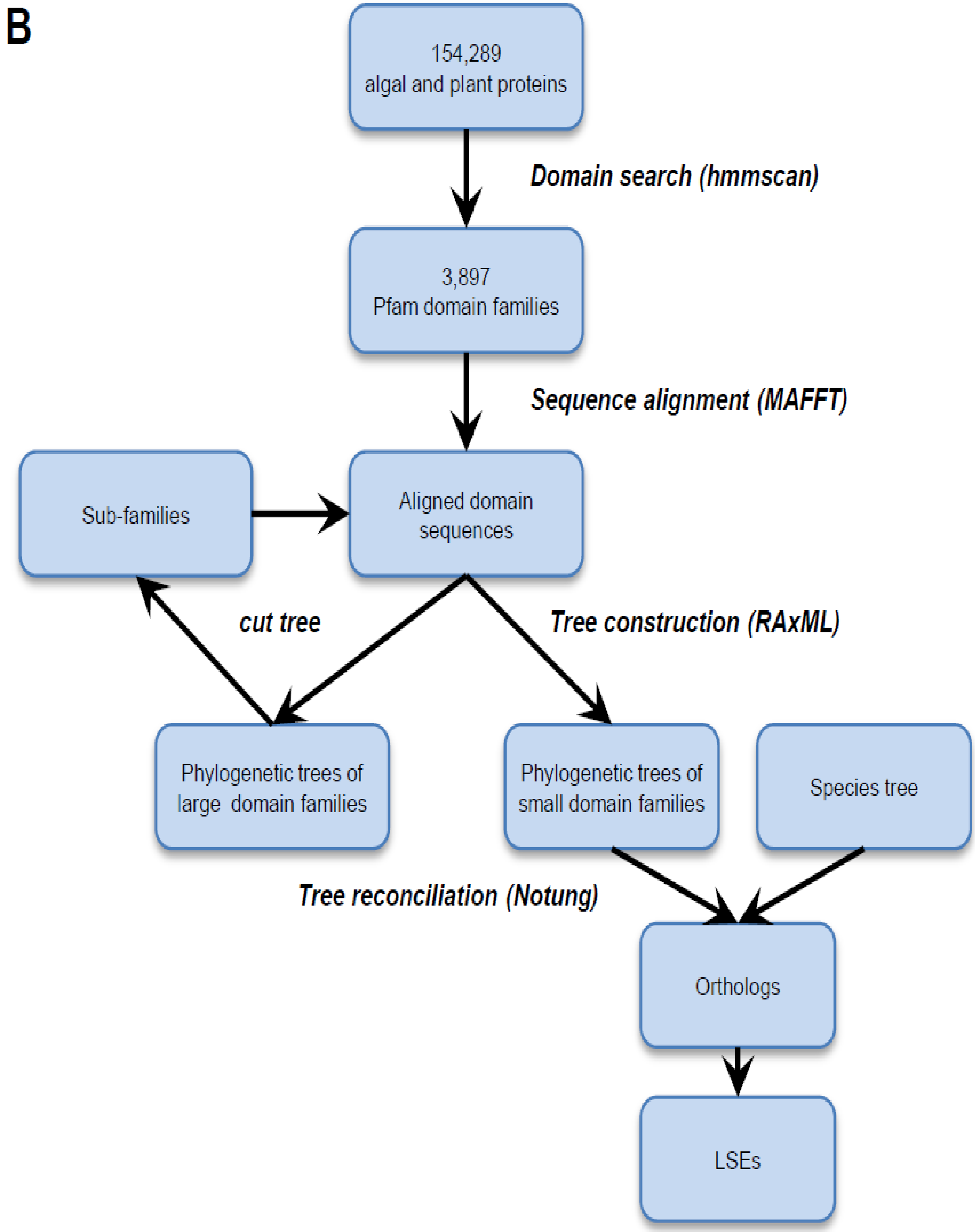
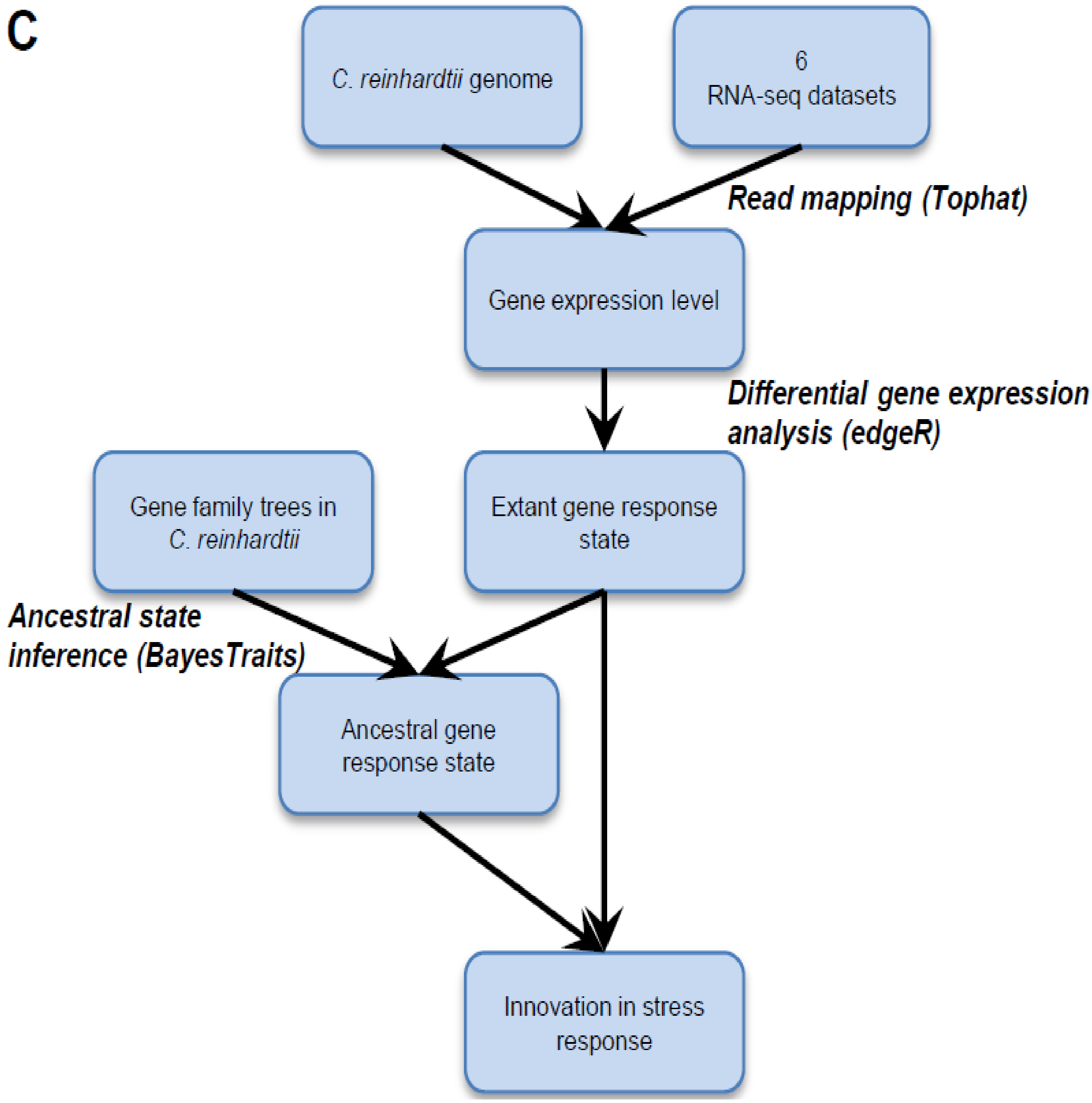

B
154,289
algal and plant proteins
Domain search (hmmscan)
3,897
Pfam domain families
Sequence alignment (MAFFT)
Sub-families
Aligned domain sequences
cut tree
Tree construction (RAxML)
Phylogenetic trees of large domain families
Phylogenetic trees of small domain families
Species tree
Tree reconciliation (Notung)
Orthologs
LSEs

Figure S2.1 (cont'd)


C
C. reinhardtii genome
6
RNA-seq datasets
Read mapping (Tophat)
Gene expression level
Differential gene expression analysis (edgeR)
Gene family trees in C. reinhardtii
Extant gene response state
Ancestral state inference (BayesTraits)
Ancestral gene response state
Innovation in stress response

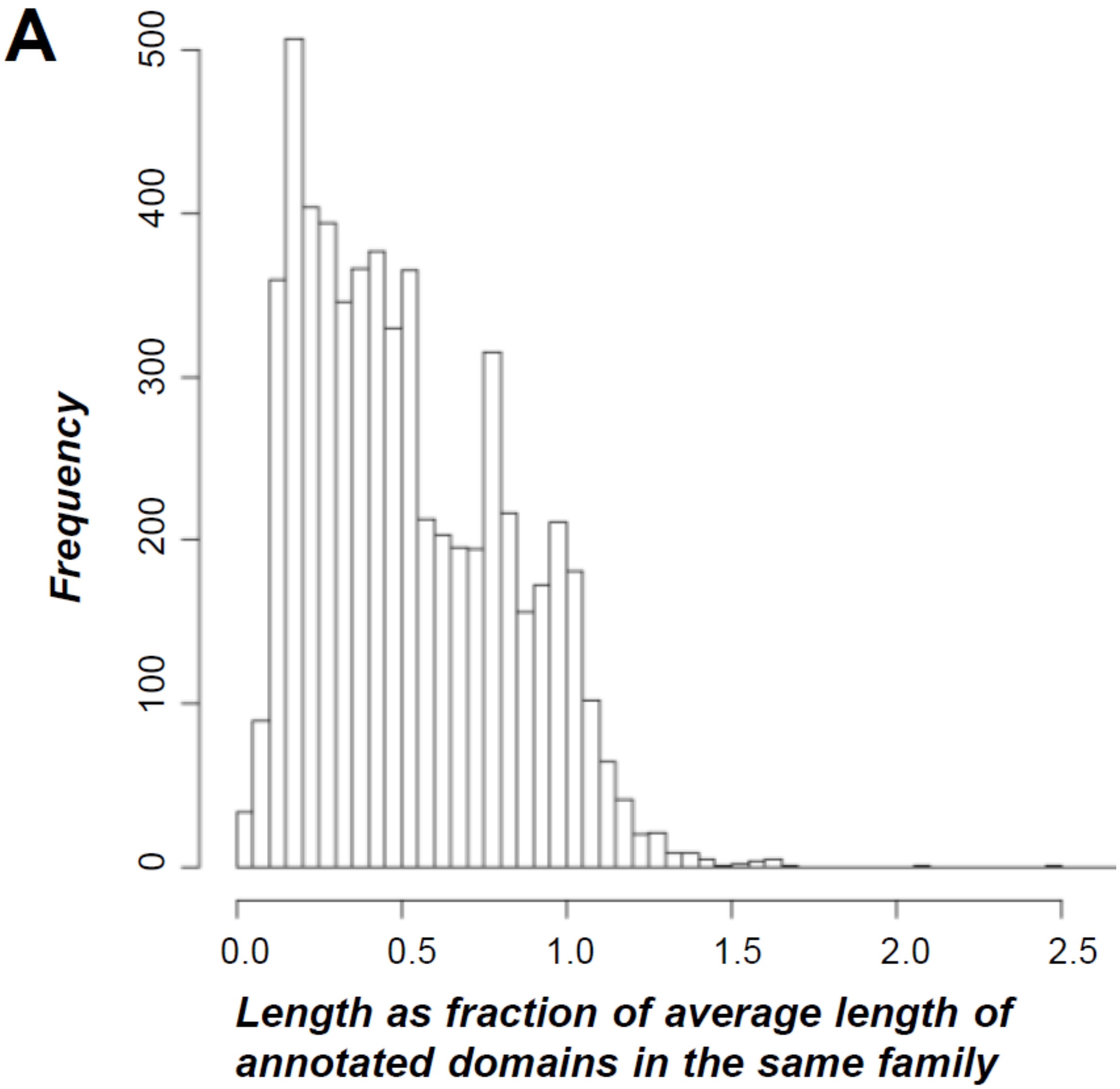


Figure S2.2. Distribution of domains in un-annotated regions (DURs) in green algae.

Figure S2.2 (cont'd)

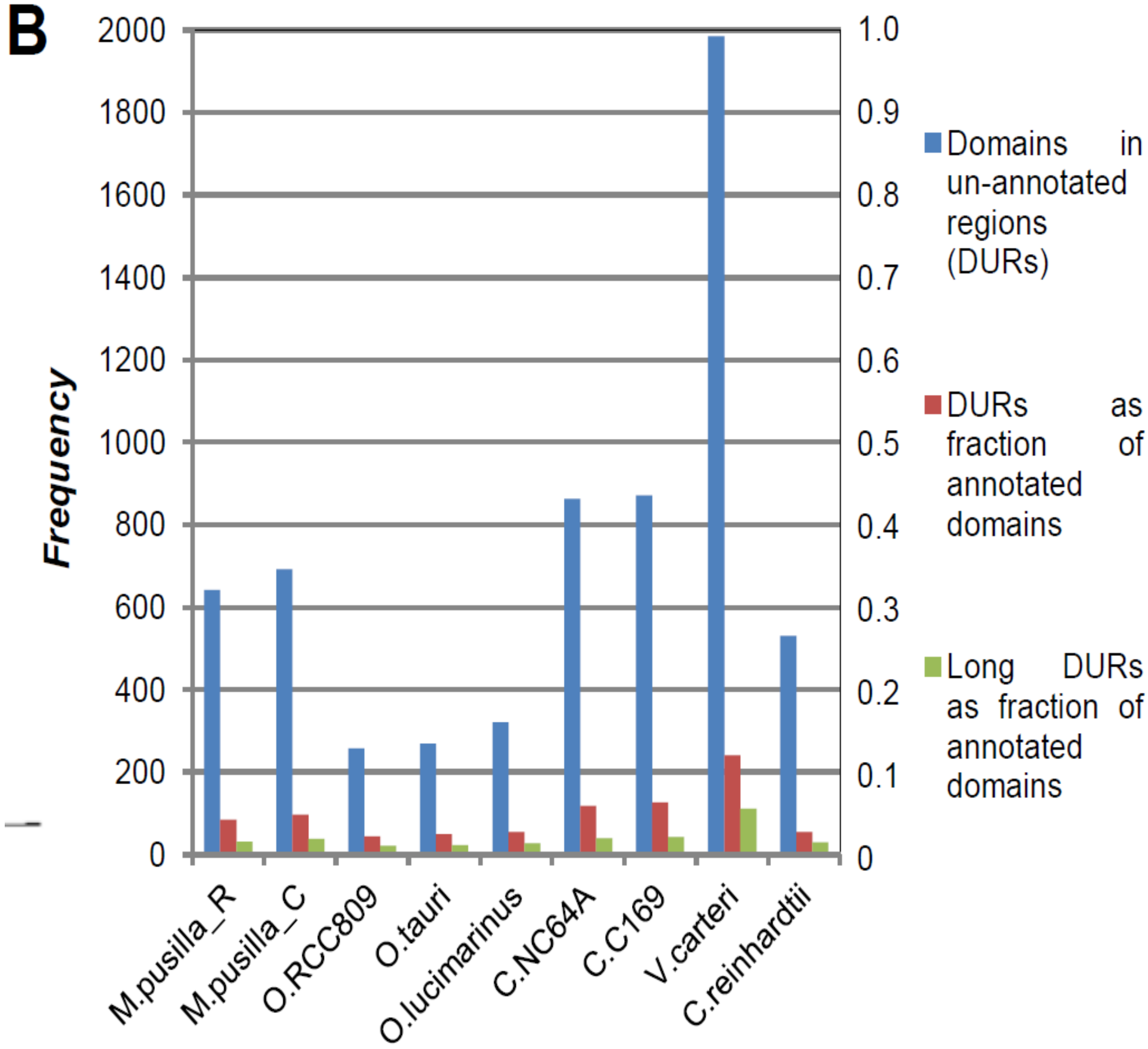


Figure S2.2 (cont’d)

(A). DUR distribution by length. X-axis shows the length of DURs as fraction of the average length of annotated domains in the same family. Y-axis indicates frequency.

(B). Preponderance of DURs among green algal species. Blue: the number of DURs in each green algal species. Red: the number of DURs relative to the number of annotated domains. Green: the number of DURs that are longer than half of the average lengths of annotated domains in the same family relative to the number of annotated domains in each species.

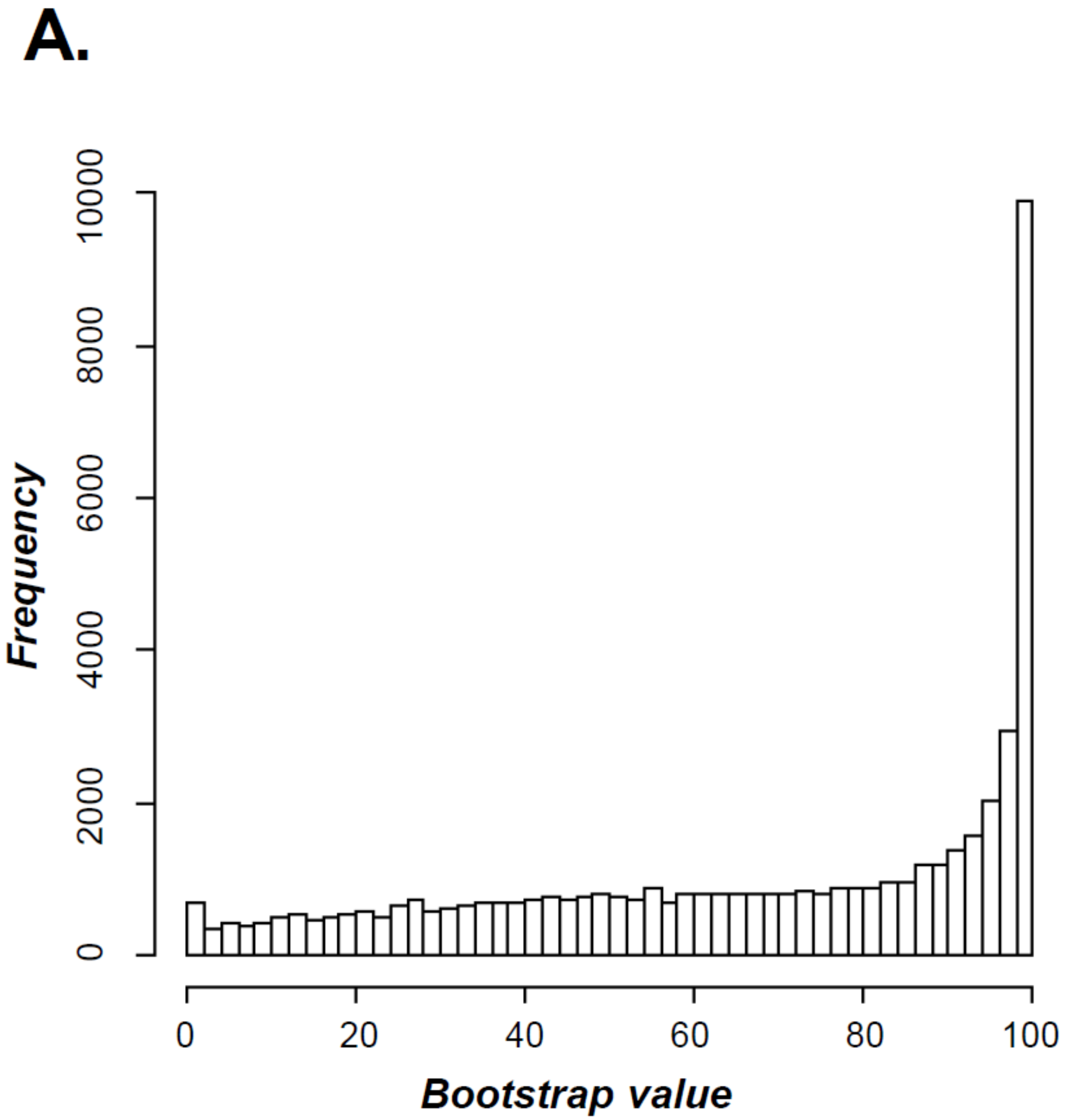


Figure S2.3. Bootstrap values of various branches.

Figure S2.3 (cont’d)

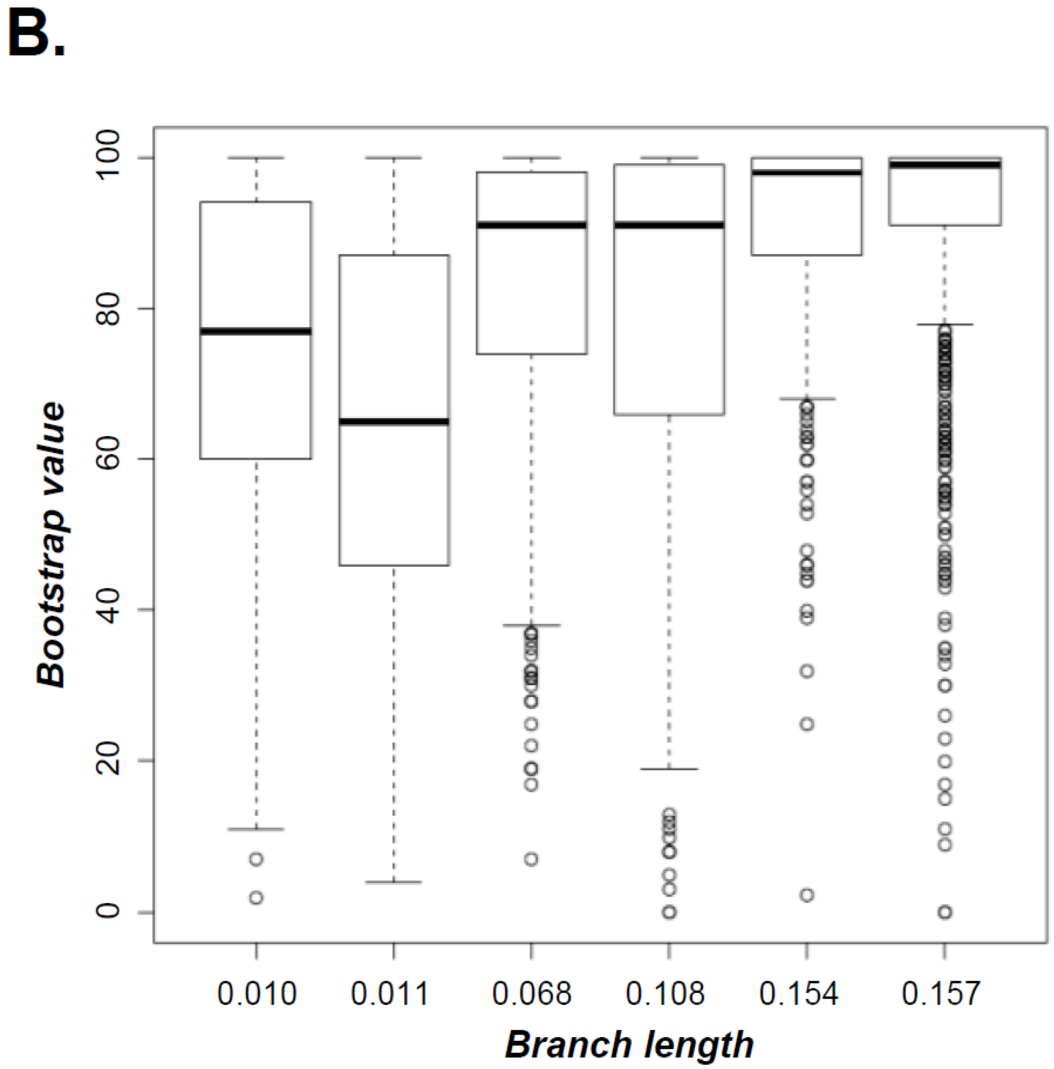


Figure S2.3 (cont'd)

(A). Bootstrap value distribution of all branches in phylogenetic trees of 1,500 domain families in green algae and land plants. X-axis indicates bootstrap values and Y-axis indicates number of branches in each bootstrap value bin.

(B). Distributions of bootstrap values in the domain phylogenetic trees and branch length of the species tree. X-axis indicates the branch length of each banch on the species tree. Y-axis indicates the bootstrap value distribution of each branch.

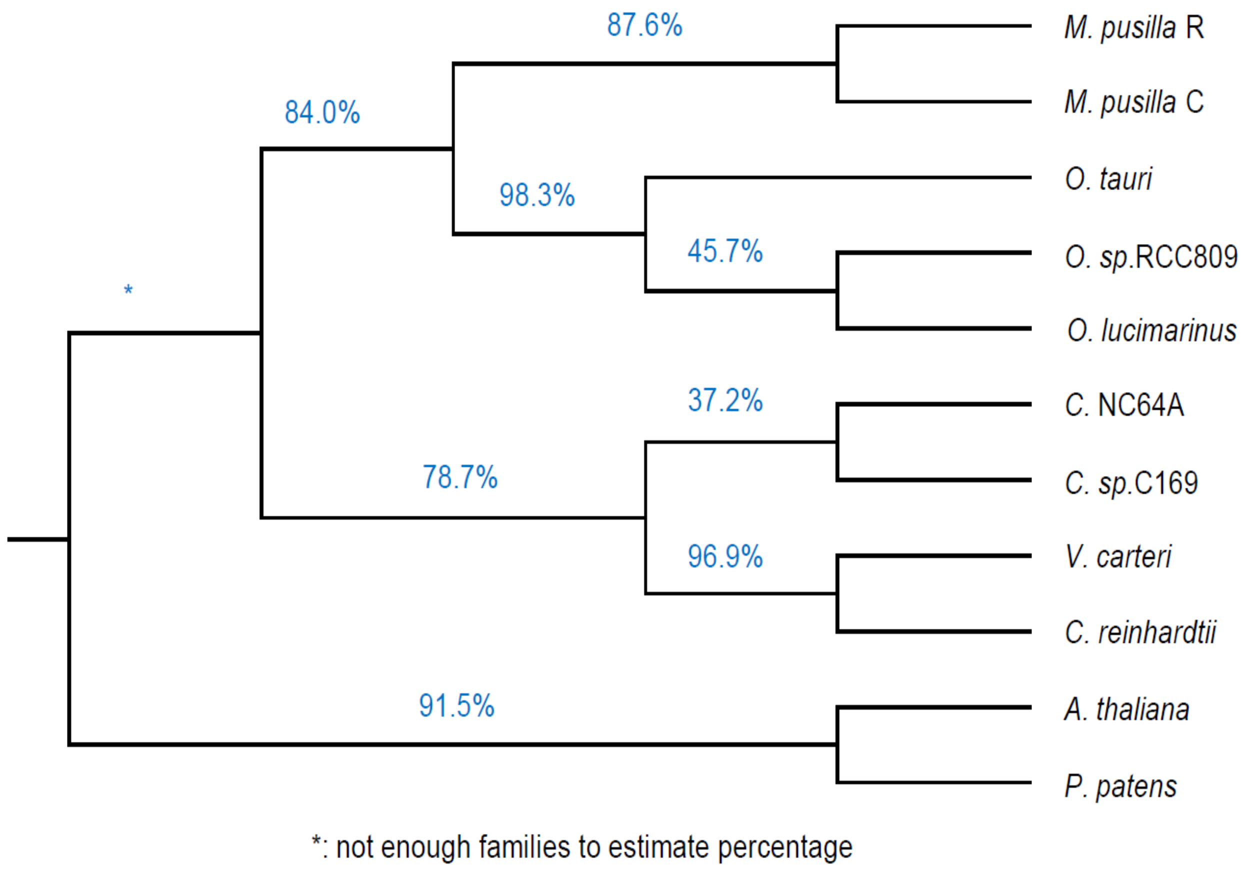


Figure S2.4. **Consistency between domain and species tree.** The species tree topology is shown. The number on each branch indicates the percentage of domain family trees with the same branching pattern as the species tree on the branch in question.

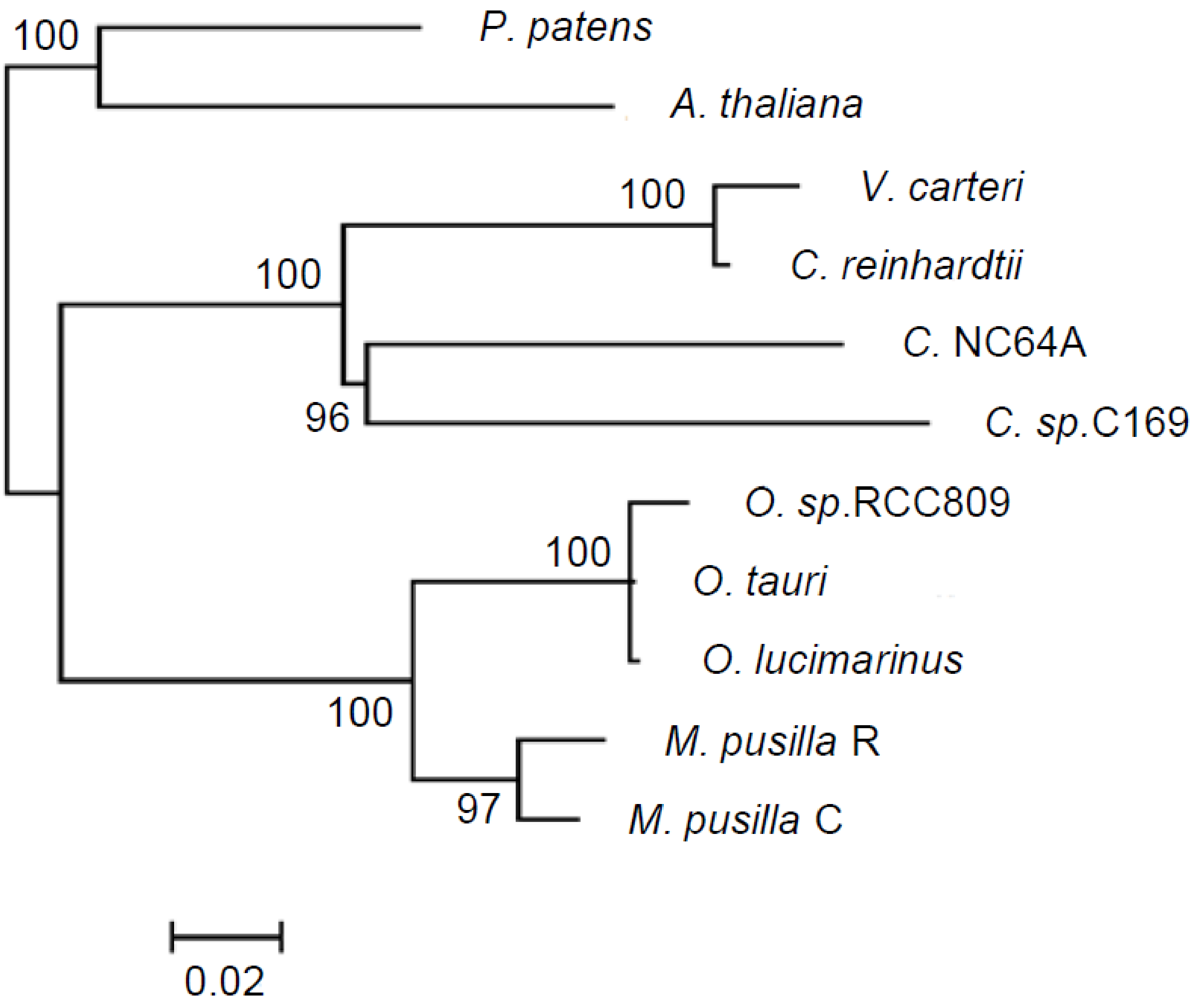


Figure S2.5. **Green algal species tree based on 18s rRNA sequences.** Numbers indicate the bootstrap values.

Table S2.1. *C. reinhardtii* retained duplicate genes tend to be stress responsive

Stress conditions	**DR**[a]	**DN**[b]	**NDR**[c]	**NDN**[d]	**p**[e]
N deprivation	654	1156	3934	10838	8.71E-17
S deprivation	383	1324	3023	11443	1.40E-01
Cu deprivation	48	1703	196	14367	3.72E-05
Fe deprivation	41	1722	203	14439	3.42E-03
CO2 deprivation	106	1679	567	14024	8.50E-05
Oxidative stress	438	1438	2919	12028	1.26E-04
all	1014	908	7169	7858	3.37E-05

[a] DR indicates retained duplicates responsive to stress; [b] DN, retained duplicates not responsive to stress; [c] NDR, genes that are not retained duplicates (singletons) responsive to stress; [d] NDN, genes that are not retained duplicates not responsive to stress; [e] p, p-value of Fisher's exact test.

Table S2.2. Gene Ontology categories significantly enriched in conserved genes or genes most closely related to pseudogenes in *C. reinhardtii*

GO	**GO**	**GO**	**NG**	**NG**	***p*[e]**	**FDR[f]**	**Annotation**
	G[a]	**N[b]**	**OG[c]**	**ON[d]**			
Conserved genes							
GO:0006412 bp	27	66	131	4451	6.09E-19	9.29E-16	translation
GO:0003735 mf	27	67	131	4450	8.31E-19	9.29E-16	structural constituent of ribosome
GO:0005840 cc	26	63	132	4454	2.44E-18	1.82E-15	ribosome
GO:0005622 cc	23	108	135	4409	3.28E-11	1.83E-08	intracellular
GO:0009538 cc	4	1	154	4516	6.12E-06	2.74E-03	photosystem I reaction center
GO:0033014 bp	4	2	154	4515	1.79E-05	6.66E-03	tetrapyrrole biosynthetic process
Genes associated with pseudogenes							
GO:0006950 bp	34	106	72	5513	3.77E-30	1.36E-26	response to stress
GO:0005524 mf	36	594	70	5025	1.76E-10	3.19E-07	ATP binding

[a]GOG indicates number of genes with a particularly GO annoation in the specified group (conserved genes or genes most closely related to pseudogenes). [b]GON, number of genes with GO not in specified group. [c]NGOG, number of genes without GO in specified group. [d]NGON, number of genes without GO not in specified group. [e]p value is calculated using Fisher's exact test. [f]FDR value is calculated using R package qvalue. bp, mf, and cc indicate biological process, molecular function, and cellular components, respectively.

Table S2.3. Gene Ontology biological processes significantly enriched in *C. reinhardtii* retained duplicates

GO	Annotation	GO D[a]	GO N[b]	NGO D[c]	NGO N[d]	*p*[e]	FDR[f]
***C. reinhardtii* lineage (4)**							
GO:0006950	response to stress	109	31	770	4815	6.23E-63	2.25E-59
GO:0006334	nucleosome assembly	94	32	785	4814	2.80E-51	3.38E-48
GO:0009617	response to bacterium	23	15	856	4831	2.26E-10	1.02E-07
GO:0006468	protein phosphorylation	57	108	822	4738	4.97E-10	1.80E-07
GO:0009611	response to wounding	22	18	857	4828	6.96E-09	1.94E-06
GO:0007165	signal transduction	27	33	852	4813	3.94E-08	1.02E-05
GO:0009294	DNA mediated transformation	20	18	859	4828	9.18E-08	2.08E-05
GO:0035556	intracellular signal transduction	25	39	854	4807	3.19E-06	5.24E-04
GO:0009567	double fertilization forming a zygote and endosperm	15	19	864	4827	5.78E-05	8.71E-03
GO:0006182	cGMP biosynthetic process	9	6	870	4840	9.50E-05	1.23E-02
GO:0055085	transmembrane transport	28	61	851	4785	1.41E-04	1.70E-02
GO:0006817	phosphate ion transport	6	2	873	4844	2.73E-04	2.99E-02
GO:0006811	ion transport	10	11	869	4835	4.84E-04	4.61E-02
Volvocales lineage (3)							
GO:0006334	nucleosome assembly	90	36	1094	4505	3.09E-35	5.58E-32
GO:0035556	intracellular signal transduction	52	12	1132	4529	2.50E-25	3.01E-22
GO:0009190	cyclic nucleotide biosynthetic process	41	4	1143	4537	3.09E-24	2.24E-21
GO:0009567	double fertilization forming a zygote and endosperm	32	2	1152	4539	3.25E-20	1.31E-17
GO:0006468	protein phosphorylation	83	82	1101	4459	9.12E-18	3.00E-15
GO:0009294	DNA mediated transformation	31	7	1153	4534	1.21E-15	3.13E-13
GO:0009617	response to bacterium	29	9	1155	4532	2.47E-13	5.25E-11
GO:0007165	signal transduction	37	23	1147	4518	4.64E-12	8.82E-10
GO:0009611	response to wounding	28	12	1156	4529	2.17E-11	3.74E-09

Table S2.3 (cont'd)

GO:0006182	cGMP biosynthetic process	15	0	1169	4541	5.05E-11	7.94E-09
GO:0055085	transmembrane transport	45	44	1139	4497	2.80E-10	4.05E-08
GO:0006811	ion transport	14	7	1170	4534	6.51E-06	6.93E-04
GO:0070588	calcium ion transmembrane transport	7	1	1177	4540	1.05E-04	8.34E-03
GO:0009987	cellular process	52	103	1132	4438	1.70E-04	1.31E-02
GO:0006171	cAMP biosynthetic process	8	3	1176	4538	2.96E-04	2.19E-02
GO:0018298	protein-chromophore linkage	6	1	1178	4540	4.46E-04	2.94E-02
Core chlorophyta lineage (2)							
GO:0007165	signal transduction	39	21	1792	3873	1.69E-07	8.74E-05
GO:0006468	protein phosphorylation	83	82	1748	3812	1.01E-06	4.56E-04
GO:0008152	metabolic process	86	102	1745	3792	6.40E-05	1.35E-02
GO:0009637	response to blue light	13	3	1818	3891	6.98E-05	1.35E-02
GO:0010218	response to far red light	14	4	1817	3890	8.51E-05	1.40E-02
GO:0010114	response to red light	15	5	1816	3889	9.58E-05	1.44E-02
GO:0018298	protein-chromophore linkage	7	0	1824	3894	3.40E-04	4.55E-02
Chlorophyta lineage (1)							
GO:0035556	intracellular signal transduction	61	3	1538	4123	1.15E-30	4.16E-27
GO:0009190	cyclic nucleotide biosynthetic process	44	1	1555	4125	9.07E-24	1.09E-20
GO:0006468	protein phosphorylation	94	71	1505	4055	3.48E-15	2.10E-12
GO:0006182	cGMP biosynthetic process	15	0	1584	4126	4.68E-09	1.41E-06
GO:0006171	cAMP biosynthetic process	11	0	1588	4126	7.87E-07	1.90E-04
GO:0016310	phosphorylation	59	63	1540	4063	1.76E-06	3.98E-04
GO:0006811	ion transport	16	5	1583	4121	5.85E-06	1.11E-03
GO:0080092	regulation of pollen tube growth	9	0	1590	4126	1.02E-05	1.75E-03
GO:0000160	two-component signal transduction system (phosphorelay)	8	1	1591	4125	2.48E-04	2.89E-02

Table S2.3 (cont'd)

GO:0006754	ATP biosynthetic process	9	2	1590	4124	3.15E-04	3.56E-02
GO:0046777	protein autophosphorylation	11	4	1588	4122	3.33E-04	3.61E-02
GO:0007018	microtubule-based movement	19	14	1580	4112	3.44E-04	3.61E-02

[a]GOD indicates number of retained duplicates with GO. [b]GON, number of genes that are not retained duplicates with GO. [c]NGOD, number of retained duplicates without GO. [d]NGON, number of genes that are not retained duplicates without GO. [e]p value is calculated using Fisher's exact test. [f]FDR value is calculated using R package qvalue. Numbers in parenthesis indicate branches as shown on Figure 2.2.

REFERENCES

REFERENCES

Adams C, Godfrey V, Wahlen B, Seefeldt L, Bugbee B (2013) Understanding precision nitrogen stress to optimize the growth and lipid content tradeoff in oleaginous green microalgae. Bioresour Technol **131**: 188–194

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol **215**: 403–410

Birney E, Clamp M, Durbin R (2004) GeneWise and Genomewise. Genome Res **14**: 988–995

Blanc G, Agarkova I, Grimwood J, Kuo A, Brueggeman A, Dunigan DD, Gurnon J, Ladunga I, Lindquist E, Lucas S, et al (2012) The genome of the polar eukaryotic microalga *Coccomyxa subellipsoidea* reveals traits of cold adaptation. Genome Biol **13**: R39

Cakmak T, Angun P, Demiray YE, Ozkan AD, Elibol Z, Tekinay T (2012) Differential effects of nitrogen and sulfur deprivation on growth and biodiesel feedstock production of *Chlamydomonas reinhardtii*. Biotechnol Bioeng **109**: 1947–1957

Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, Holt C, Alvarado AS, Yandell M (2008) MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res **18**: 188–196

Castruita M, Casero D, Karpowicz SJ, Kropat J, Vieler A, Hsieh SI, Yan W, Cokus S, Loo JA, Benning C, et al (2011) Systems Biology Approach in *Chlamydomonas* Reveals Connections between Copper Nutrition and Multiple Metabolic Steps. Plant Cell Online **23**: 1273–1292

Chen K, Durand D, Farach-Colton M (2000) NOTUNG: A Program for Dating Gene Duplications and Optimizing Gene Family Trees. J Comput Biol **7**: 429–447

Cock JM, Sterck L, Rouzé P, Scornet D, Allen AE, Amoutzias G, Anthouard V, Artiguenave F, Aury J-M, Badger JH, et al (2010) The *Ectocarpus* genome and the independent evolution of multicellularity in brown algae. Nature **465**: 617–621

Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics **21**: 3674–3676

Derelle E, Ferraz C, Rombauts S, Rouzé P, Worden AZ, Robbens S, Partensky F, Degroeve S, Echeynié S, Cooke R, et al (2006) Genome analysis of the smallest free-living eukaryote *Ostreococcus tauri* unveils many unique features. Proc Natl Acad Sci **103**: 11647–11652

Fang W, Si Y, Douglass S, Casero D, Merchant SS, Pellegrini M, Ladunga I, Liu P, Spalding MH (2012) Transcriptome-Wide Changes in Chlamydomonas reinhardtii Gene Expression Regulated by Carbon Dioxide and the CO2-Concentrating Mechanism Regulator CIA5/CCM1. Plant Cell Online **24**: 1876–1893

Felsenstein J (2005) PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

Finn RD, Clements J, Eddy SR (2011) HMMER web server: interactive sequence similarity searching. Nucleic Acids Res **39**: W29–W37

Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, et al (2010) The Pfam protein families database. Nucleic Acids Res **38**: D211–D222

Force A, Lynch M, Pickett FB, Amores A, Yan Y, Postlethwait J (1999) Preservation of Duplicate Genes by Complementary, Degenerative Mutations. Genetics **151**: 1531–1545

Gaillard AR, Fox LA, Rhea JM, Craige B, Sale WS (2006) Disruption of the A-Kinase Anchoring Domain in Flagellar Radial Spoke Protein 3 Results in Unregulated Axonemal cAMP-dependent Protein Kinase Activity and Abnormal Flagellar Motility. Mol Biol Cell **17**: 2626–2635

González-Ballester D, Casero D, Cokus S, Pellegrini M, Merchant SS, Grossman AR (2010) RNA-Seq Analysis of Sulfur-Deprived *Chlamydomonas* Cells Reveals Aspects of Acclimation Critical for Cell Survival. Plant Cell Online **22**: 2058–2084

Graham L, Graham J, Wilcox L (2009) Algae, 2nd ed. Pearson

Graur D (2000) Fundamentals of molecular evolution. Sinauer Associates, Sunderland, Mass.

Hanada K, Zou C, Lehti-Shiu MD, Shinozaki K, Shiu S-H (2008) Importance of Lineage-Specific Expansion of Plant Tandem Duplicates in the Adaptive Response to Environmental Stimuli. Plant Physiol **148**: 993–1003

Hasegawa E, Hayashi H, Asakura S, Kamiya R (1987) Stimulation of in vitro motility of *Chlamydomonas* axonemes by inhibition of cAMP-dependent phosphorylation. Cell Motil Cytoskeleton **8**: 302–311

Hou S-Y, Govorunova EG, Ntefidou M, Lane CE, Spudich EN, Sineshchekov OA, Spudich JL (2012) Diversity of *Chlamydomonas* Channelrhodopsins. Photochem Photobiol **88**: 119–128

Hu Q, Sommerfeld M, Jarvis E, Ghirardi M, Posewitz M, Seibert M, Darzins A (2008) Microalgal triacylglycerols as feedstocks for biofuel production: perspectives and advances. Plant J Cell Mol Biol **54**: 621–639

Jordan IK, Makarova KS, Spouge JL, Wolf YI, Koonin EV (2001) Lineage-specific gene expansions in bacterial and archaeal genomes. Genome Res **11**: 555–565

Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res **30**: 3059–3066

Lespinet O, Wolf YI, Koonin EV, Aravind L (2002) The Role of Lineage-Specific Gene Family Expansion in the Evolution of Eukaryotes. Genome Res **12**: 1048–1059

Lynch M, O'Hely M, Walsh B, Force A (2001) The Probability of Preservation of a Newly Arisen Gene Duplicate. Genetics **159**: 1789–1804

McDonald SM, Plant JN, Worden AZ (2010) The Mixed Lineage Nature of Nitrogen Transport and Assimilation in Marine Eukaryotic Phytoplankton: A Case Study of Micromonas. Mol Biol Evol **27**: 2268–2283

Merchant SS, Prochnik SE, Vallon O, Harris EH, Karpowicz SJ, Witman GB, Terry A, Salamov A, Fritz-Laylin LK, Maréchal-Drouard L, et al (2007) The *Chlamydomonas* Genome Reveals the Evolution of Key Animal and Plant Functions. Science **318**: 245–250

Miller R, Wu G, Deshpande RR, Vieler A, Gärtner K, Li X, Moellering ER, Zäuner S, Cornish AJ, Liu B, et al (2010) Changes in Transcript Abundance in *Chlamydomonas reinhardtii* following Nitrogen Deprivation Predict Diversion of Metabolism. Plant Physiol **154**: 1737–1752

Moellering ER, Benning C (2010) RNA Interference Silencing of a Major Lipid Droplet Protein Affects Lipid Droplet Size in *Chlamydomonas reinhardtii*. Eukaryot Cell **9**: 97–106

Moore RC, Purugganan MD (2005) The evolutionary dynamics of plant duplicate genes. Curr Opin Plant Biol **8**: 122–128

Moore RC, Purugganan MD (2003) The early stages of duplicate gene evolution. Proc Natl Acad Sci **100**: 15682–15687

Moreau H, Verhelst B, Couloux A, Derelle E, Rombauts S, Grimsley N, Bel MV, Poulain J, Katinka M, Hohmann-Marriott MF, et al (2012) Gene functionalities and genome structure in *Bathycoccus prasinos* reflect cellular specializations at the base of the green lineage. Genome Biol **13**: R74

Ohno S (1970) Evolution by gene duplication. Springer-Verlag

Page MD, Allen MD, Kropat J, Urzica EI, Karpowicz SJ, Hsieh SI, Loo JA, Merchant SS (2012) Fe Sparing and Fe Recycling Contribute to Increased Superoxide Dismutase Capacity in Iron-Starved *Chlamydomonas reinhardtii*. Plant Cell Online **24**: 2649–2665

Pagel M (1999) The Maximum Likelihood Approach to Reconstructing Ancestral Character States of Discrete Characters on Phylogenies. Syst Biol **48**: 612–622

Palenik B, Grimwood J, Aerts A, Rouzé P, Salamov A, Putnam N, Dupont C, Jorgensen R, Derelle E, Rombauts S, et al (2007) The tiny eukaryote *Ostreococcus* provides genomic insights into the paradox of plankton speciation. Proc Natl Acad Sci **104**: 7705–7710

Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics **26**: 139–140

Siaut M, Cuine S, Cagnon C, Fessler B, Nguyen M, Carrier P, Beyly A, Beisson F, Triantaphylides C, Li-Beisson Y, et al (2011) Oil accumulation in the model green alga *Chlamydomonas reinhardtii*: characterization, variability between common laboratory strains and relationship with starch reserves. BMC Biotechnol **11**: 7

Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics **22**: 2688–2690

Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics **25**: 1105–1111

Turner BM (2012) The adjustable nucleosome: an epigenetic signaling module. Trends Genet **28**: 436–444

Urzica EI, Adler LN, Page MD, Linster CL, Arbing MA, Casero D, Pellegrini M, Merchant SS, Clarke SG (2012) Impact of Oxidative Stress on Ascorbate Biosynthesis in *Chlamydomonas* via Regulation of the VTC2 Gene Encoding a GDP-l-galactose Phosphorylase. J Biol Chem **287**: 14234–14245

Vieler A, Wu G, Tsai C-H, Bullard B, Cornish AJ, Harvey C, Reca I-B, Thornburg C, Achawanantakun R, Buehl CJ, et al (2012) Genome, Functional Gene Annotation, and Nuclear Transformation of the Heterokont Oleaginous Alga *Nannochloropsis oceanica* CCMP1779. PLoS Genet **8**: e1003064

Vogel C, Chothia C (2006) Protein Family Expansions and Biological Complexity. PLoS Comput Biol **2**: e48

Wang ZT, Ullrich N, Joo S, Waffenschmidt S, Goodenough U (2009) Algal Lipid Bodies: Stress Induction, Purification, and Biochemical Characterization in Wild-Type and Starchless *Chlamydomonas reinhardtii*. Eukaryot Cell **8**: 1856–1868

Wapinski I, Pfeffer A, Friedman N, Regev A (2007) Natural history and evolutionary principles of gene duplication in fungi. Nature **449**: 54–61

Ward JM, Mäser P, Schroeder JI (2009) Plant Ion Channels: Gene Families, Physiology, and Functional Genomics Analyses. Annu Rev Physiol **71**: 59–82

Worden AZ, Lee J-H, Mock T, Rouzé P, Simmons MP, Aerts AL, Allen AE, Cuvelier ML, Derelle E, Everett MV, et al (2009) Green Evolution and Dynamic Adaptations Revealed by Genomes of the Marine Picoeukaryotes *Micromonas*. Science **324**: 268–272

Zou C, Lehti-Shiu MD, Thibaud-Nissen F, Prakash T, Buell CR, Shiu S-H (2009a) Evolutionary and Expression Signatures of Pseudogenes in *Arabidopsis* and Rice. Plant Physiol **151**: 3–15

Zou C, Lehti-Shiu MD, Thomashow M, Shiu S-H (2009b) Evolution of Stress-Regulated Gene Expression in Duplicate Genes of *Arabidopsis thaliana*. PLoS Genet **5**: e1000581

CHAPTER 3

GENOME AND FUNCTIONAL GENE ANNOTATION OF THE HETEROKONT OLEAGINOUS ALGA *NANNOCHLOROPSIS OCEANICA* CCMP1779

Note: The contents of this chapter have been published in PLoS Genetics.

Astrid Vieler*, Guangxi Wu*, Chia-Hong Tsai, Blair Bullard, Adam J. Cornish, Christopher Harvey, Ida-Barbara Reca, Chelsea Thornburg, Rujira Achawanantakun, Christopher J. Buehl, Michael S. Campbell, David Cavalier, Kevin L. Childs, Teresa J. Clark, Rahul Deshpande, Erika Erickson, Ann Armenia Ferguson, Witawas Handee, Que Kong, Xiaobo Li, Bensheng Liu, Steven Lundback, Cheng Peng, Rebecca Roston, Sanjaya, Jeffrey P. Simpson, Allan TerBush, Jaruswan Warakanont, Simone Zäuner, Eva M. Farre, Eric L. Hegg, Ning Jiang, Min-Hao Kuo, Yan Lu, Krishna K. Niyogi, John Ohlrogge, Katherine W. Osteryoung, Yair Shachar-Hill, Barbara B. Sears, Yanni Sun, Hideki Takahashi, Mark Yandell, Shin-Han Shiu, Christoph Benning (2012) Genome, Functional Gene Annotation, and Nuclear Transformation of the Heterokont Oleaginous Alga *Nannochloropsis oceanica* CCMP1779. PLoS Genet 8(11): e1003064. doi:10.1371/journal.pgen.1003064

*These authors contributed equally.

ABSTRACT

Unicellular marine algae have promise for providing sustainable and scalable biofuel feedstocks, although no single species has emerged as a preferred organism. Moreover, adequate molecular and genetic resources prerequisite for the rational engineering of marine algal feedstocks are lacking for most candidate species. Heterokonts of the genus Nannochloropsis naturally have high cellular oil content and are already in use for industrial production of high-value lipid products. First success in applying reverse genetics by targeted gene replacement makes *Nannochloropsis oceanica* an attractive model to investigate the cell and molecular biology and biochemistry of this fascinating organism group. Here, we present the assembly of the 28.7 Mb genome of *N. oceanica* CCMP1779. RNA sequencing data from nitrogen-replete and nitrogen-depleted growth conditions support a total of 11,973 genes, of which in addition to automatic annotation some were manually inspected to predict the biochemical repertoire for this organism. Among others, more than 100 genes putatively related to lipid metabolism, 114 predicted transcription factors, and 109 transcriptional regulators were annotated. Comparison of the *N. oceanica* CCMP1779 gene repertoire with the recently published *N. gaditana* genome identified 2,649 genes likely specific to *N. oceanica* CCMP1779. Many of these *N. oceanica*–specific genes have putative orthologs in other species or are supported by transcriptional evidence. However, because similarity-based annotations are limited, functions of most of these species-specific genes remain unknown. Aside from the genome sequence and its analysis, protocols for the transformation of *N. oceanica* CCMP1779 are provided. The availability of genomic and transcriptomic data for *N. oceanica* CCMP1779, along with efficient

transformation protocols, provides a blueprint for future detailed gene functional analysis and genetic engineering of *Nannochloropsis* species by a growing academic community focused on this genus.

AUTHOR SUMMARY

Algae are a highly diverse group of organisms that have become the focus of renewed interest due to their potential for producing biofuel feedstocks, nutraceuticals, and biomaterials. Their high photosynthetic yields and ability to grow in areas unsuitable for agriculture provide a potential sustainable alternative to using traditional agricultural crops for biofuels. Because none of the algae currently in use have a history of domestication, and bioengineering of algae is still in its infancy, there is a need to develop algal strains adapted to cultivation for industrial large-scale production of desired compounds. Model organisms ranging from mice to baker's yeast have been instrumental in providing insights into fundamental biological structures and functions. The algal field needs versatile models to develop a fundamental understanding of photosynthetic production of biomass and valuable compounds in unicellular, marine, oleaginous algal species. To contribute to the development of such an algal model system for basic discovery, we sequenced the genome and two sets of transcriptomes of *N. oceanica* CCMP1779, assembled the genomic sequence, identified putative genes, and began to interpret the function of selected genes. This species was chosen because it is readily transformable with foreign DNA and grows well in culture.

INTRODUCTION

The search for sustainable sources of liquid transportation fuels has led to renewed interest in microalgae as potential feedstocks and rising research activity focused on the basic biology of algae. Microalgae can accumulate large quantities of oils (triacylglycerols) and carbohydrates, particularly when nutrient-deprived (Dismukes et al., 2008; Wijffels and Barbosa, 2010). Recent estimates taking into account different locations predict that microalgal photosynthesis can produce between 40,000 and 50,000 L ha^{-1}year^{-1}, which is 5-to-6 times the yield observed for oil palm (Weyer et al., 2010). To realize this potential, it will be necessary to understand photosynthetic growth and metabolism of specific model algae. Even though genomic information and basic molecular tools are available for a range of organisms such as the diatoms *Phaeodactylum tricornutum* (Siaut et al., 2007; Bowler et al., 2008), the brown alga *Ectocarpus siliculosus* (Cock et al., 2010) or the tiny chlorophyte *Ostreococcus tauri* (Derelle et al., 2006), the mechanistic study of microalgal gene functions is currently lagging behind models such as Arabidopsis. Of all algae, *Chlamydomonas reinhardtii* is currently the most thoroughly studied based on the number of entries in the Public Library of Medicine (http://www.ncbi.nlm.nih.gov/pubmed/). Despite its proven versatility, Chlamydomonas is still somewhat limited with regard to available tools for its molecular analysis. For example, efficient targeted inactivation of genes by gene disruption technology is not available and loss-of-function mutants can be difficult to obtain by RNA interference and related techniques. The recent achievement of homologous gene replacement in *Nannochloropsis* *oceanica* (Kilian et al., 2011) opens up potential opportunities to develop this alga into an alternate model organism representing marine, oleaginous microalgae.

Nannochloropsis is classified under the class Eustigmatophyceae of the Heterokontophyta (Hoek, 1995), a diverse algal group that includes brown algae and diatoms. The plastid of this alga is surrounded by four membranes derived from a secondary endosymbiotic event (Reyes-Prieto et al., 2007). Strains from this genus have been investigated for their lipid composition and lipid accumulation, e.g. (Sukenik and Carmeli, 1990; Schneider and Roessler, 1994; Tonon et al., 2002; Danielewicz et al., 2011). In addition, the biomass production by strains of *Nannochloropsis* grown under different conditions has been increasingly studied in recent years, e.g. (Hu and Gao, 2003; Xu et al., 2004; Rodolfi et al., 2009; Simionato et al., 2011; Srinivas and Ochs, 2012). Given the potential of this alga as an industrial feedstock and the progress made in developing homologous gene replacement, several research groups have set out to sequence the genome of different *Nannochloropsis* strains and draft genomes of *N. oceanica* (Pan et al., 2011) and *Nannochloropsis gaditana* (Radakovits et al., 2012) have recently become available.

Here, we focus on the publicly available strain *N. oceanica* CCMP1779, which we chose based on its growth in culture, its sensitivity to antibiotics, and ease of integrating transformation markers into its nuclear genome. We sequenced its genomic DNA and two sets of cDNAs obtained from two different growth conditions to aid in the annotation of genes. Its genome has been tentatively compared to that of *N. gaditana*. In addition a team of scientists has begun to manually annotate and examine the gene repertoire for specific pathways and processes to better understand the biology of this alga.

RESULTS AND DISCUSSION

Strain selection—antibiotic sensitivity, growth and introduction of selectable markers

Out of 20 axenic *Nannochloropsis* strains obtained from the Provasoli-Guillard National Center for Marine Algae and Microbiota (NCMA, formerly CCMP), strains of the *N. salina* (CCMP369), *N. gaditana* (CCMP1775 and 536) and *N. granulata* (CCMP529), as well as two not further specified strains (CCMP1779 and CCMP531) were selected based on uniformly dispersed, robust growth in enriched artificial sea water (16 g/L marine salt content) in batch culture as well as on agar-solidified medium. Both unspecified *Nannochloropsis sp.* strains cluster with strains of the *N. oceanica* species in a rooted tree (Saitou and Nei, 1987) based on 26 published 18S rRNA nucleotide sequences (Figure 3.1) using *Eustigmatos vischeri* (Eustigmatophyceae) as an out-group (Andersen et al., 1998). For this reason, these strains are hereafter referred to as *N. oceanica*. Because of poor growth under the conditions we have used, *N. oculata* and the fresh water species *N. limnetica* were not further analyzed.

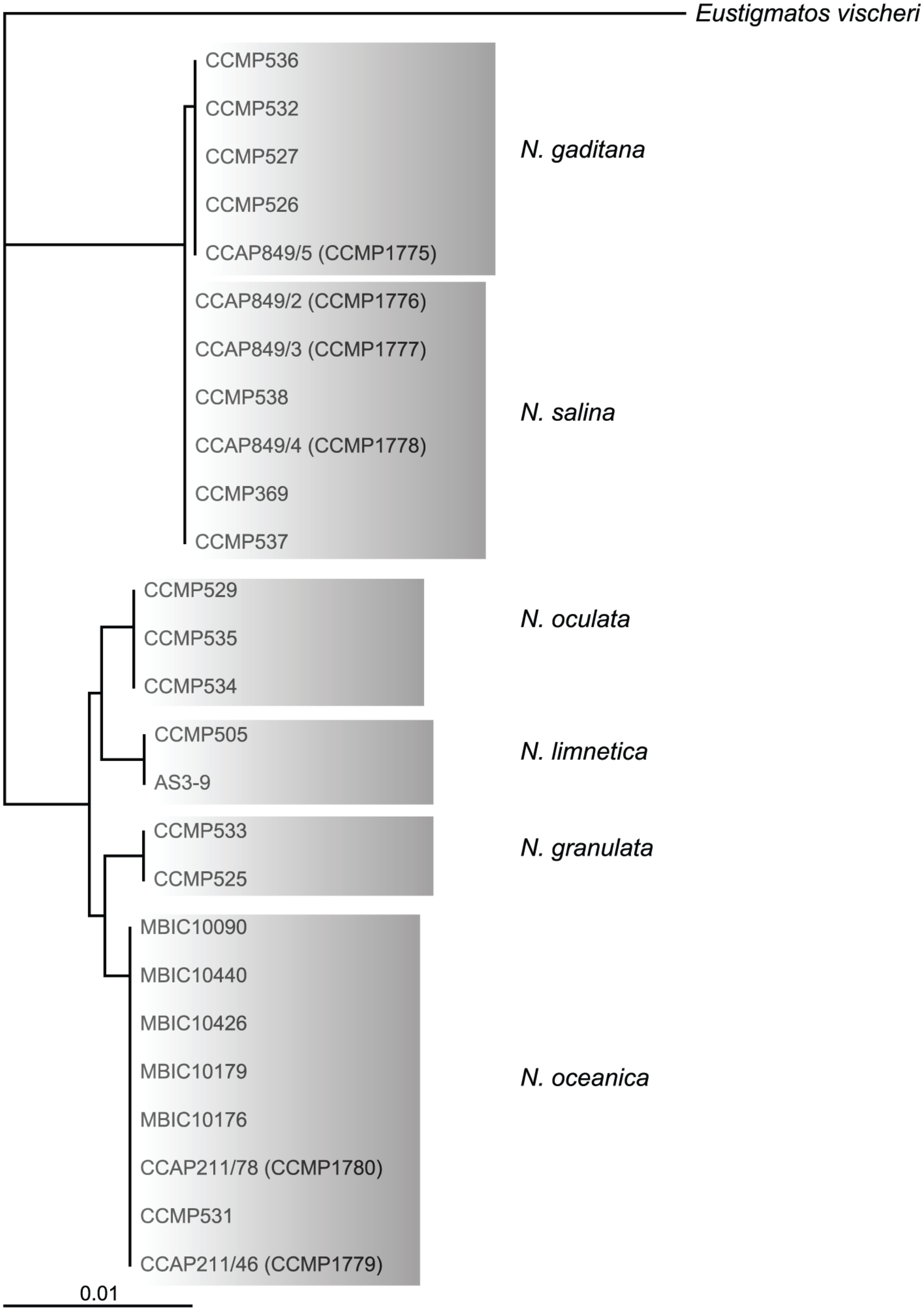


Figure 3.1. Rooted neighbor joining tree of 18s rRNA sequences of different *Nannochloropsis* species using *Eustigmatos vischeri* as an outgroup.

Figure 3.1 (cont'd)

Labels refer to strain identification numbers from the respective culture collections, if applicable the synonym is given as 2nd name. CCMP, Provasoli Guillard Culture Collection for Marine Phytoplankton, USA; CCAP, Culture Collection of Algae and Protozoa, UK; MBIC, Marine Biotechnology Institute Culture Collection, Japan, AS3-9 from (Fawley and Fawley, 2007).

The use of antibiotics is essential for eliminating contaminants from cultures and genes conferring resistance to antibiotics are frequently used as markers for the introduction and genomic insertion of foreign DNA. Therefore, we tested the *Nannochloropsis* strains for their sensitivity to a range of antibiotics. Cells were plated at high density on agar-solidified medium containing the antibiotics at high density to determine the appropriate dosage (Table 3.1). Zeocin (5 µg/mL), and Hygromycin B (25 µg/mL) were chosen for use in subsequent selection marker studies, because of the consistent inhibition of growth at low concentrations by these antibiotics. Sensitivity to Paromomycin and Hygromycin B varied among the *Nannochloropsis* strains; Paromomycin had promise as a selective agent for the two *N. oceanica* strains (CCMP1779 and CCMP531), which were also the most sensitive to Hygromycin B. Of those four antibiotics, plasmids with genes that confer resistance to Zeocin, Hygromycin B, or Paromomycin are readily available and commonly used for transformation of Chlamydomonas as reviewed in (Harris, 2009). Sensitivity to antibiotics is often determined by its rate of entry into the respective cells, which may be determined by the cell membrane and its transporters and the physical barrier provided by the cell wall. Differences in cell wall composition or thickness allowing more efficient cell entry of antibiotics are possible explanations for increased sensitivity in *N. oceanica* strains. Since efficient uptake of antibiotics or other supplemented molecules (such as metabolic substrates, inhibitors or nucleic acids) is a desirable trait for a laboratory model organism, we focused on *N. oceanica*.

Table 3.1. Comparison of the effect of selected antibiotics on different *Nannochloropsis* species. Shown are the lethal doses in µg/mL determined by plating dilutions of cell suspensions on half salinity f/2 agar plates. '>' indicates the highest concentration of the respective antibiotic tested and no detectable impact on cell growth observed. All of the *Nannochloropsis* strains listed here were found to be resistant to the following antibiotics with the respective concentrations in µg/mL given in parenthesis: Rifampicin (10), Benomyl (5), Nystatin (5), Spectinomycin (100), Ampicillin (200), Chloramphenicol (100).

Species	Strain	Zeocin	Paromomycin	Hygromycin B	Spectinomycin
N. oceanica	ccmp1779	5	5	25	>100
	ccmp531	5	10	50	>100
N. granulata	ccmp529	5	>200	100	>100
N. salina	ccmp369	5	>200	>100	>100
N. gaditana	ccmp1775	5	>100	100	>100
N. gaditana	ccmp536	5	>100	>100	>100

All *Nannochloropsis* strains were resistant to low concentrations of Rifampicin (10 µg/mL), Benomyl (5 µg/mL), Nystatin (5 µg/mL), and higher concentrations of Spectinomycin (100 µg/ml), Ampicillin (200 µg/ml), and Chloramphenicol (100 µg/mL). Hence these antibiotics can be useful for selecting against bacterial and other possible contaminants in *Nannochloropsis* cultures.

Basic growth characteristics of *N. oceanica* CMP1779 were determined. The growth curves were fitted to a sigmoidal curve and the averaged exponential growth rate k, maximum cell density a_{max} and time of half maximum cell density x_c were determined (Table 3.2). Under photoautotrophic conditions in enriched sea water the exponential growth rate of the population, k, reached an average of 0.66±0.17 d^{-1} and cultures grew to a cell density of approximately 6×10^{7} cells mL^{-1} (a_{max}). The addition of vitamins did not enhance growth in liquid culture, whereas the addition of an external carbon source drastically increased final cell densities in stationary phase, reaching up to 8.7×10^{7} or 1.5×10^{8} cells mL^{-1} when the medium was supplemented with 30 mM glucose or fructose, respectively. The intrinsic growth rate did not increase, indicating a positive effect of sugars on cell division only during the later log phase and/or early stationary phase when self-shading limited growth in the photoautotrophic culture.

Table 3.2. Growth parameters of *N. oceanica* CCMP1779 in f/2 medium using different supplements. V= f/2 Vitamine mix, Gl = Glucose, Fr=Fructose, curves have been determined in triplicates based on cell density and fitted to a sigmoidal logistic function type 1 individually using OriginPro software (y=a/1+exp(-k*(x*x_c))). Parameters a (Amplitude, here: max. cell density in cell/ml), x_c (time of ½a in d) and k (coefficient, intrinsic growth rate d^{-1}) are arithmetic means with standard deviation.

	f/2		**f/2+V**		**f/2+Gl**		**f/2+Fr**	
a	6.3E7	+/-8.5E6	6.4E7	+/-8.3E6	8.7E+07	+/-1.7E7	1.5E8	+/-7.4E6
x_c	13.65	+/-1.50	11.47	+/-0.79	11.01	+/-0.90	12.86	+/-0.12
k	0.66	+/-0.17	0.63	+/-0.15	0.61	+/-0.07	0.45	+/-0.01

Introduction of foreign DNA and stable integration into the genome are crucial for many reverse-genetics approaches. Recently, efficient protocols using an electroporation approach have been published for *N. oceanica* sp. and *N. gaditana* (Kilian et al., 2011; Radakovits et al., 2012). We tested the strain CCMP1779 for nuclear transformation using an endogenous promoter region of a structural lipid droplet surface protein (Vieler et al., 2012) driving the *aphVII* gene that confers resistance to Hygromycin B. Transformation was performed by electroporation without prior enzymatic treatments (Li and Tsai, 2009), and selection on 50 µg/mL Hygromycin B, and resulted in a transformation rate of $1.25\times10^{-06}\pm0.6\times10^{-06}$ per µg plasmid DNA (Table 3.3). This equals a more than 10-fold increase in transformation events compared to plasmid pHyg3 (Berthold et al., 2002) that was engineered for *C. reinhardtii*. The insertion of the transgene into the genome was confirmed for selected clones of both constructs by Southern hybridization (Figure 3.2).

Table 3.3: Number of resistant colonies achieved by electroporation of *N. oceanica* CCMP1779 cells in the presence of linearized pHyg3, pSelect100 plasmids per µg linearized plasmid DNA and transformation rates. Arithmetic means are given from three (pSelect100) or four (pHyg3 and no plasmid control) independent experiments with standard deviation. All transformation reactions contained denatured salmon sperm DNA in 10-fold excess compared to plasmid DNA.

	# resistant colonies per µg plasmid DNA		transformation rate (integrations per cell)	
pHyg3	7.68	+/-6.38	1/13,029,316	+/-6.38E-08
pSelect100	125	+/-6.01	1/802,998	+/-6.01E-08
no plasmid [1]	5.5	+/-8.41	1/160,000,000	+/-8.20E-09

[1] for the no plasmid transformation, the average number of resistant colonies per 100 000 000 cells and the rate of spontaneous resistance occurring is shown.

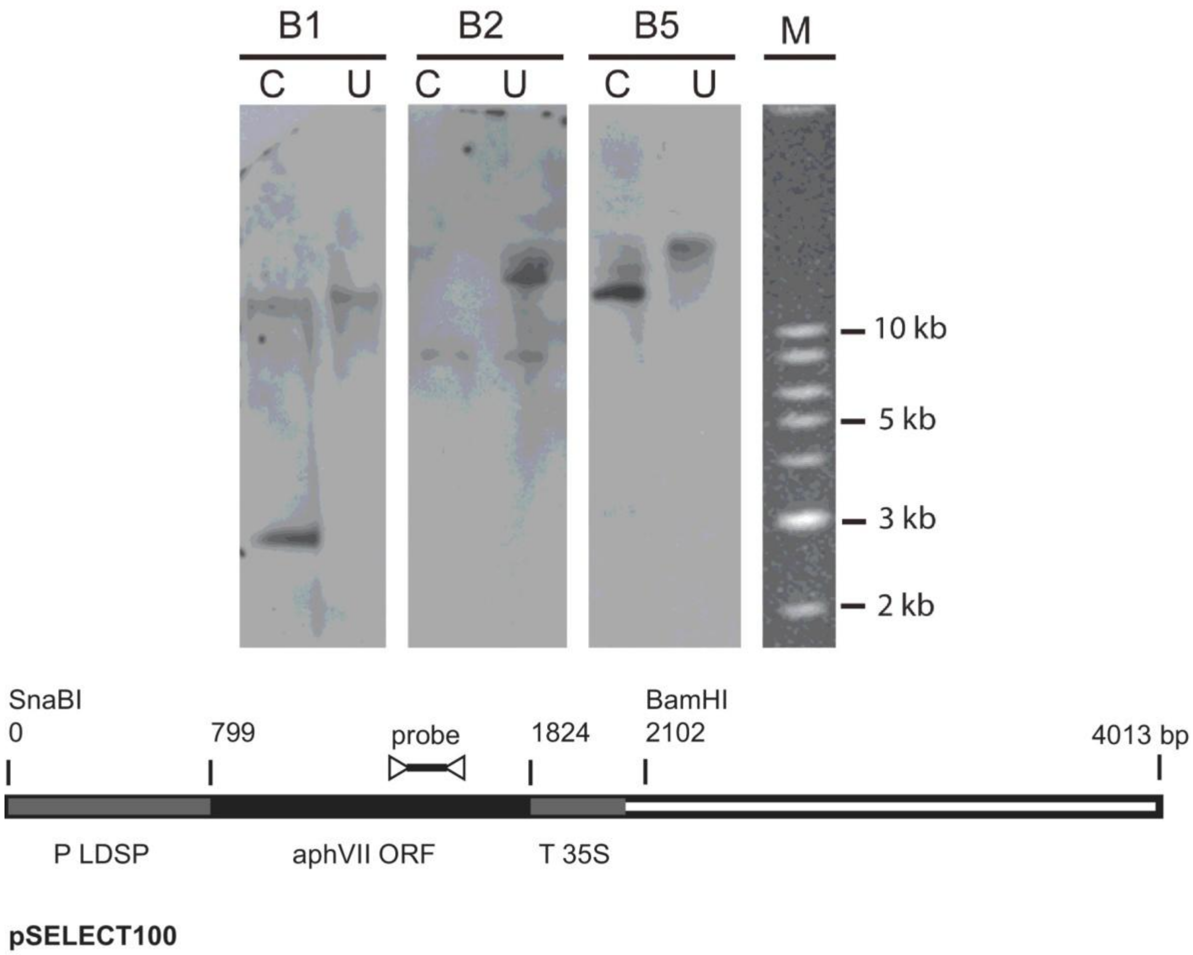


Figure 3.2. Nuclear Transformation. Southern Hybridization of CCMP 1779 transgenic clones transformed pSelect100 plasmid. C, DNA digested with BamHI restriction endonuclease, U, DNA probed undigested. Lower panel depicts a schematic map of the SnaBI linearized plasmids with the basic features indicated. P LDSP, Promoter region of LDSP (NannoCCMP1779_4188), ORF aphVII, open reading frame of, T 35S, terminator sequence of 35S.

Genome sequencing strategy, assembly, and annotation

The *N. oceanica* CCMP1779 genome was sequenced with 454 and Illumina technology. Both types of reads were used to generate a hybrid assembly with 3,731 contigs, an assembly size of 28.7 Mb and an N50 of 24,152 bp (see Materials and Methods; Figure 3.3, NCBI/SRA SRP013753). The coverage of the hybrid assembly was calculated to be ~116-fold (30-fold for 454, and 86-fold for Illumina data). In addition to genomic sequences, we conducted RNA-sequencing (RNA-seq) and generated a *de novo* assembly of 65,321 transcripts. Using these transcripts, we assessed the parameter choice for genome assembly (see Materials and Methods). RNA-seq reads were also mapped to the final genome assembly and assembled into 35,756 transcripts to facilitate structural annotation.

Genome annotation was carried out using the MAKER pipeline (Cantarel et al., 2008). In addition to *ab initio* gene predictions, transcripts from RNA-seq and protein sequences from six other heterokonts (see Materials and Methods for species) were incorporated to generate a draft gene annotation with evidence-based quality values (AED, Annotation Edit Distance) (Eilbeck et al., 2009). Basic information about predicted genes and the genome is shown in Table 3.4. The final annotation set contains 11,973 protein-coding genes: 6,362 gene models with transcript and/or protein similarity support and an additional 5,611 *ab initio* predictions (NCBI/GEO GSE36959). Protein domain search results showed that the percentage of proteins with InterPro domains in CCMP1779 is comparable to but slightly lower than that of the other six sequenced heterokonts (Figure 3.4C). We also found 83.4% of the proteins from the CEGMA database that contain highly conserved eukaryotic proteins (Parra et al., 2007). For comparison, the representation of CEGMA proteins in the green alga *Chlamydomonas*, the parasitic protozoan

Toxoplasma gondii, and the heterokont *Ectocarpus siliculosus* are 88.9%, 66.2% (Parra et al., 2007), and 85.8% (Cock et al., 2010), respectively. These findings demonstrate that our annotation is of similar quality as that for the other eukaryotes, particularly heterokont genome annotations.

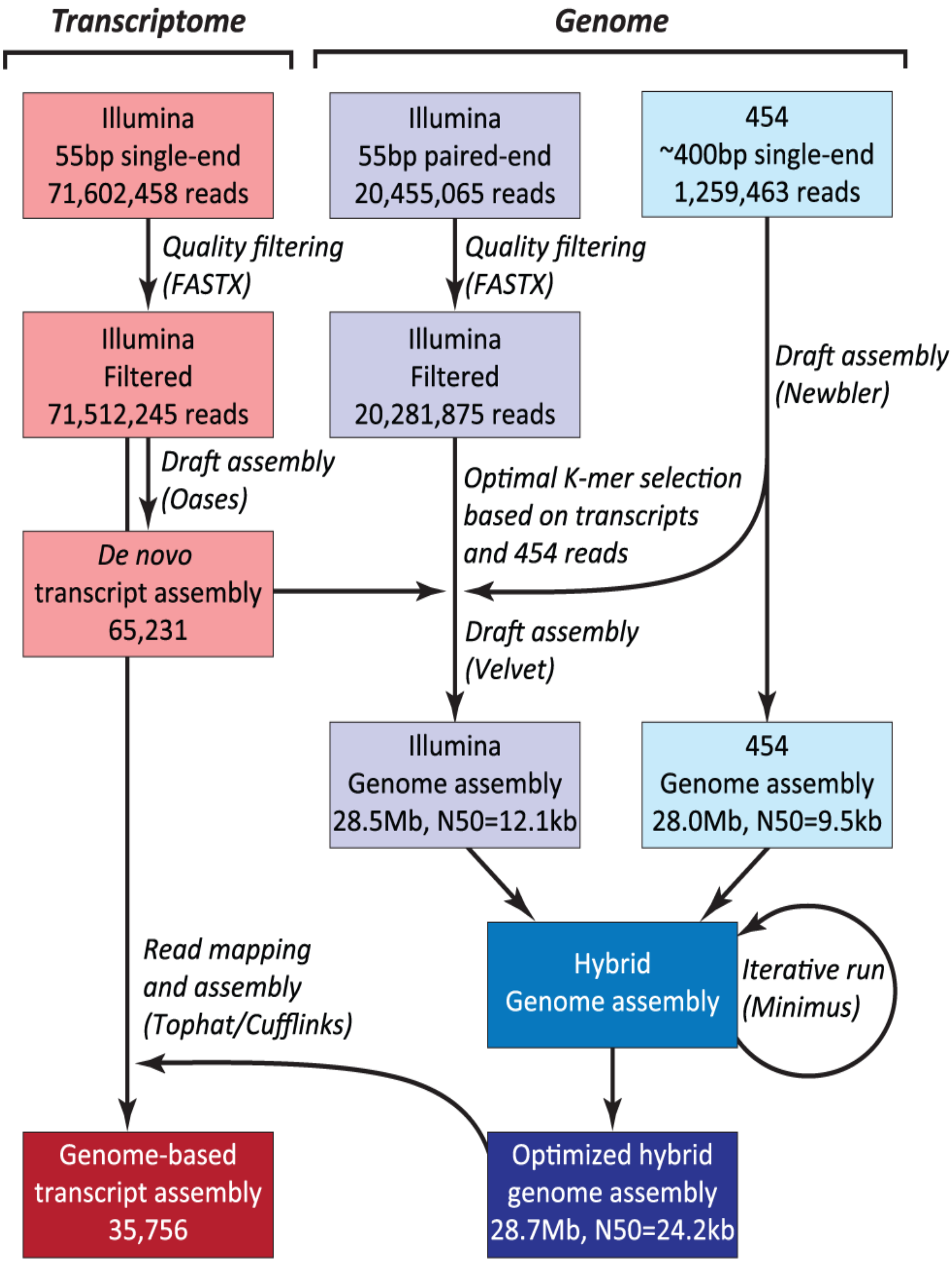


Figure 3.3. Hybrid assembly strategy using Illumina and 454 reads.

Figure 3.3 (cont’d)

N50 is used in reference to average contig length. The definition of N50 is the length N for which 50% of all bases in the sequence assembly are in a sequence of length L < N. Kb means kilobase.

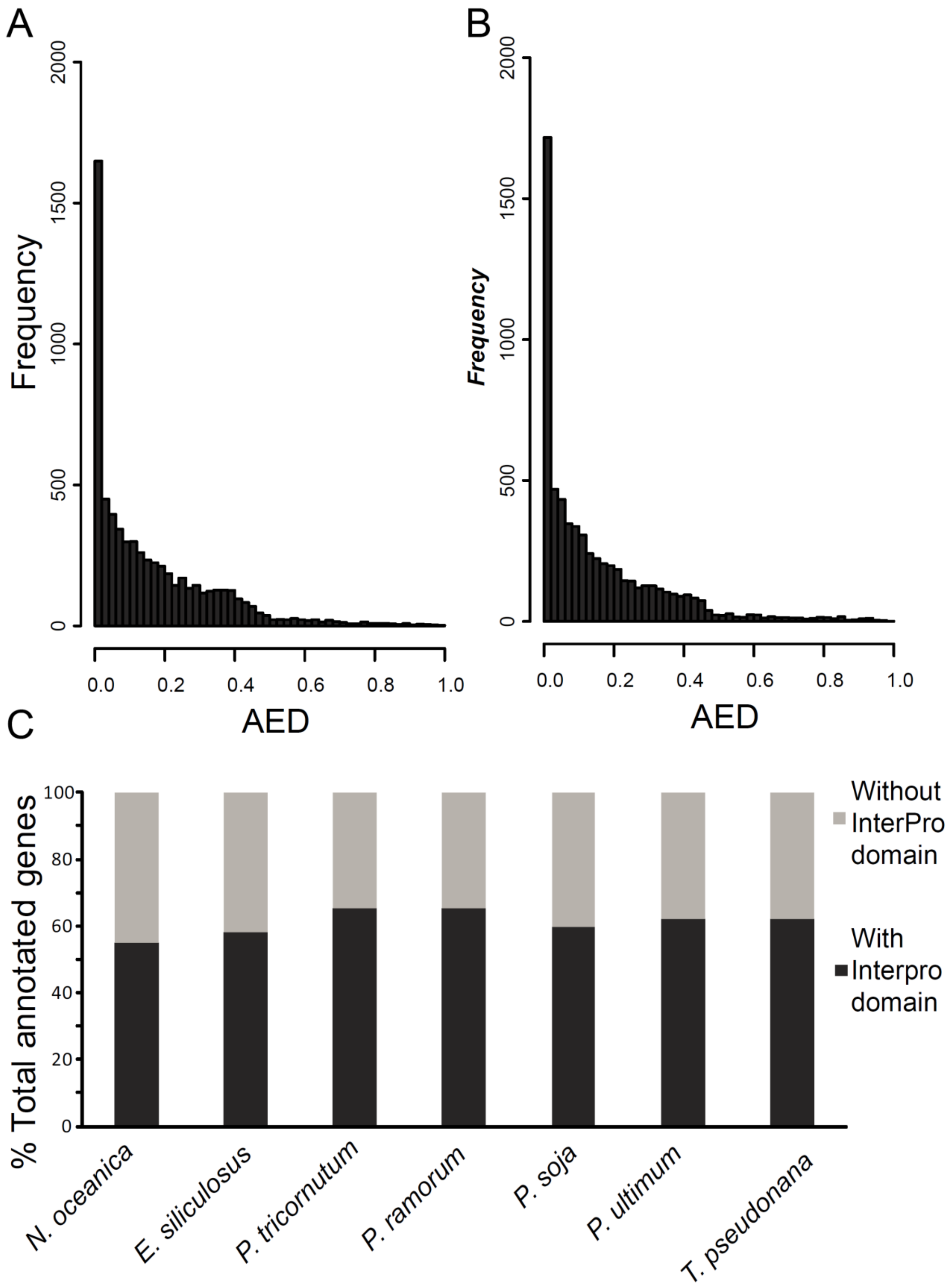


Figure 3.4. Gene Annotation.

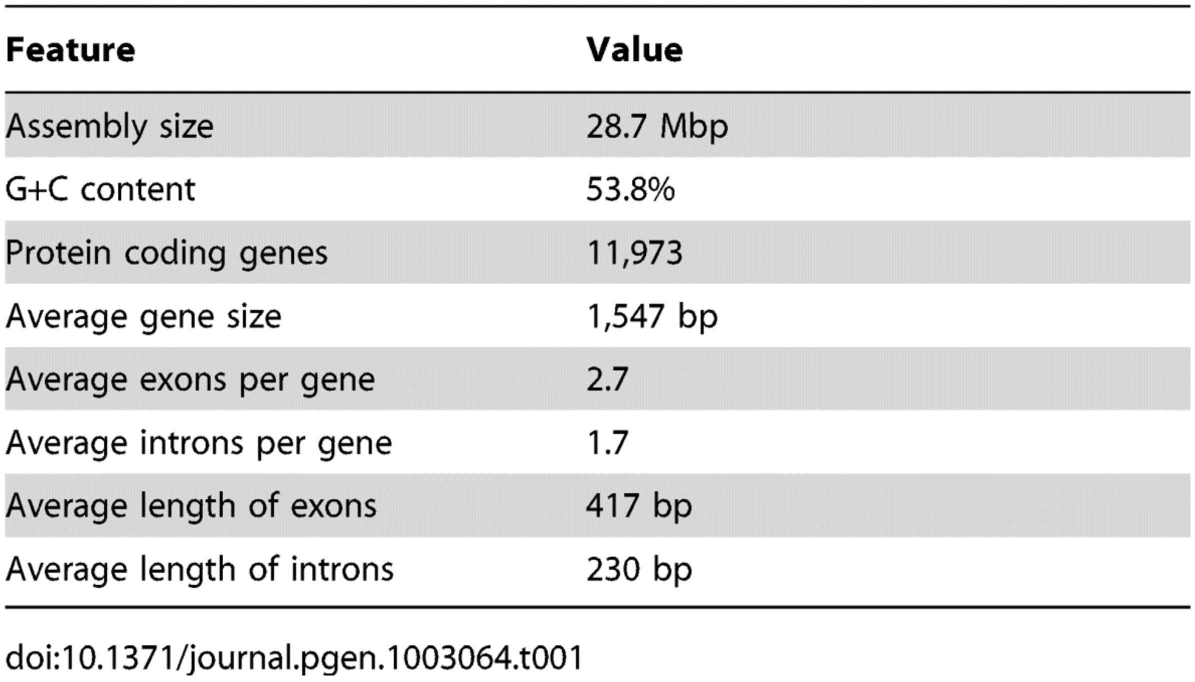
Figure 3.4 (cont'd)

(A) Annotation Edit Distance (AED) distribution of gene models in the first annotation set after eliminating entries with AED = 1.

(B). AED distribution of gene models in the second annotation after eliminating entries with AED = 1.

(C) Proportion of gene models with protein domain hits in different heterokonts (abbreviated as indicated in Materials and Methods).

Table 3.4: Genome summary.



Feature	Value
Assembly size	28.7 Mbp
G+C content	53.8%
Protein coding genes	11,973
Average gene size	1,547 bp
Average exons per gene	2.7
Average introns per gene	1.7
Average length of exons	417 bp
Average length of introns	230 bp

doi:10.1371/journal.pgen.1003064.t001

Functional annotation based on protein domains, functional category assignments, and expression

To generate functional annotation, we first identified protein domains in annotated genes. Of the 12,012 identified protein models in our first annotation run, 4,847 did not have a significant match in the NCBI (http://www.ncbi.nlm.nih.gov) non-redundant protein database (version 4, January 20, 2012). One potential explanation for this relatively high number of putatively unique genes is that related sequences are not annotated in heterokonts. In addition, we cannot rule out the possibility of false positive gene prediction. Of the 7,165 (59.6%) protein models with matches, 721 protein sequences could not be mapped by Blast2GO (Conesa et al.,

2005) to retrieve GO (Gene Ontology) terms and annotation to select reliable functions. Manual examination of a random selection of these proteins revealed that they matched mostly uncharacterized proteins, usually from other heterokont genomes such as *E. siliculosus* or *Albugo laibachii*. A total of 26,573 GO terms were assigned after augmented annex annotation (Myhre et al., 2006) and merging primary GO annotations with the InterPro Scan results (Zdobnov and Apweiler, 2001) (Figure 3.5). A total of 5,981 (49.8%) CCMP1779 genes had GO annotations.

Our RNA-seq runs were conducted with RNA samples obtained from cells grown under nitrogen (N)-replete and N-deprived conditions that typically differ in the biosynthesis of storage lipids among other metabolic functions, (see e.g. Miller et al., 2010). To assess whether expression of genes in certain functional categories were particularly influenced by these conditions, we determined the enrichment of GO terms in up- and down-regulated genes. At 1% significance level (Fisher's Exact Test), genes with 7 and 27 GO terms were significantly enriched in up- and down-regulated genes, respectively (Table 3.5). In particular, genes associated with photosynthesis and DNA replication tended to be down-regulated following N deprivation, but also genes for central carbon metabolism were affected, such as gluconeogenesis and glycolysis. We previously observed similar effects for *Chlamydomonas* (Miller et al., 2010) which is evolutionarily distant from *Nannochloropsis*.

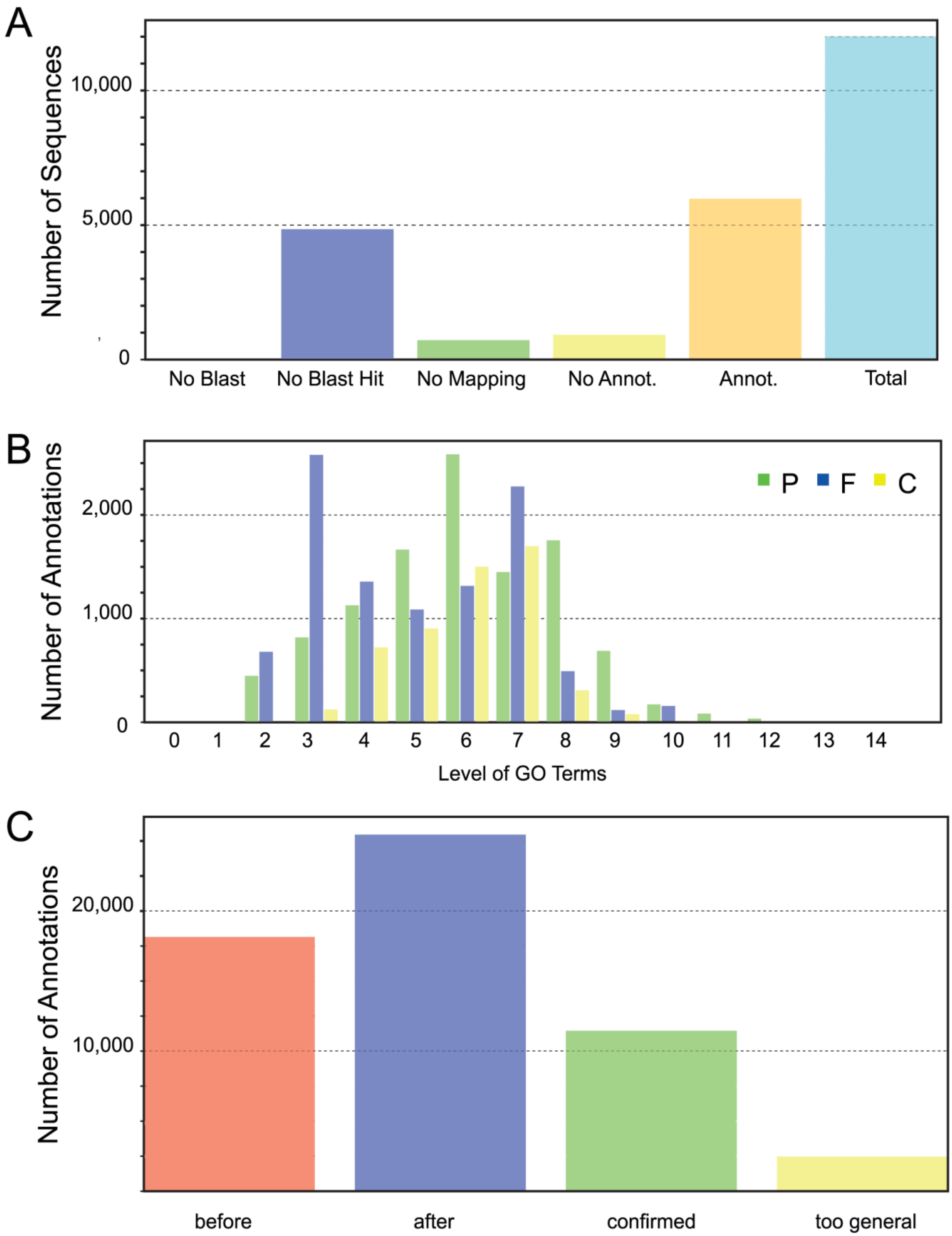


Figure 3.5. Gene Ontology.

Figure 3.5 (cont'd)

(A) Overview of Blast2Go functional annotation results. No Blast, Number of sequences without blast search performed; No Blast Hit, Number of sequences without blastp hits at the given threshold (e-value<10^{-5}); No Mapping, Number of blast hits that did not map to the Blast2GO database; No Annot., Number of mapped hits that did not retrieve GO annotations from the Blast2GO database; Annot., Number of sequences that did retrieve one or more GO annotations from the Blast2GO database; Total, Total number of analyzed sequences.

(B) The distribution of GO annotations by GO level shows the respective number of added GO annotations in relation to their GO level for each category (P biological process, F molecular function, C cellular component).

(C) Results distribution after implementation of InterProScan results. Before, Total number of added GO terms after Blast2GO annotation; after, Total number of GO annotations after implementation of InterProScan results; confirmed, Number of initial GO annotations confirmed by InterProScan result; too general, Number of GO annotations removed after InterProScan because of a lack of specificity.

Table 3.5. Enriched GO categories in up- and down-regulated genes during N-deprived versus N-replete conditions based on RNAseq data.

GO	GO R[a]	No GO R[b]	GO U[c]	No GO U[d]	P[e]	Annotation
down-regulated						
GO:0006096 bp[f]	20	645	38	5237	2.15E-06	glycolysis
GO:0015979 bp	10	655	7	5268	2.70E-06	photosynthesis
GO:0015995 bp	9	656	5	5270	3.12E-06	chlorophyll biosynthetic process
GO:0006094 bp	19	646	42	5233	2.07E-05	gluconeogenesis
GO:0006012 bp	8	657	6	5269	3.84E-05	galactose metabolic process
GO:0010007 cc[g]	5	660	1	5274	9.45E-05	magnesium chelatase complex
GO:0016851 mf[h]	5	660	1	5274	9.45E-05	magnesium chelatase activity
GO:0000084 bp	8	657	8	5267	0.000134	S phase of mitotic cell cycle
GO:0015976 bp	12	653	23	5252	0.000256	carbon utilization
GO:0000216 bp	7	658	7	5268	0.000361	M/G1 transition of mitotic cell cycle
GO:0006098 bp	11	654	22	5253	0.000608	pentose-phosphate shunt
GO:0006888 bp	7	658	9	5266	0.000984	ER to Golgi vesicle-mediated transport
GO:0003980 mf	3	662	0	5275	0.001398	UDP-glucose:glycoprotein glucosyltransferase activity
GO:0009773 bp	3	662	0	5275	0.001398	photosynthetic electron transport in photosystem I
GO:0042132 mf	3	662	0	5275	0.001398	fructose 1,6-bisphosphate 1-phosphatase activity
GO:0030604 mf	3	662	0	5275	0.001398	1-deoxy-D-xylulose-5-phosphate reductoisomerase activity
GO:0006000 bp	8	657	14	5261	0.001805	fructose metabolic process
GO:0006271 bp	4	661	2	5273	0.00194	DNA strand elongation involved in DNA replication

Table 3.5 (cont’d)

GO:0005985 bp	6	659	8	5267	0.00262	sucrose metabolic process
GO:0006694 bp	7	658	12	5263	0.00321	steroid biosynthetic process
GO:0030127 cc	4	661	3	5272	0.004127	COPII vesicle coat
GO:0003755 mf	9	656	21	5254	0.004272	peptidyl-prolyl cis-trans isomerase activity
GO:0006270 bp	5	660	6	5269	0.004493	DNA-dependent DNA replication initiation
GO:0007018 bp	10	655	26	5249	0.004883	microtubule-based movement
GO:0070402 mf	3	662	1	5274	0.005123	NADPH binding
GO:0042277 mf	3	662	1	5274	0.005123	peptide binding
GO:0019872 bp	5	660	7	5268	0.007	streptomycin biosynthetic process
up-regulated						
GO:0009308 bp	5	111	1	5823	1.54E-08	amine metabolic process
GO:0008131 mf	5	111	1	5823	1.54E-08	primary amine oxidase activity
GO:0048038 mf	5	111	3	5821	1.39E-07	quinone binding
GO:0005507 mf	5	111	23	5801	0.000179	copper ion binding
GO:0008146 mf	3	113	7	5817	0.000788	sulfotransferase activity
GO:0004190 mf	3	113	13	5811	0.003376	aspartic-type endopeptidase activity
GO:0016887 mf	7	109	94	5730	0.003423	ATPase activity

[a] GO R, Number of significantly up- or down-regulated (R) genes with the GO annotation in question. [b] No GO R, Number of significantly up- or down-regulated genes without the GO annotation. [c] GO U, Number of genes without significant expression change with the GO annotation. [d] No GO U, Number of genes with no significant expression change that do not have the GO annotation. [e] Fisher’s exact test P value. [f] bp, Biological process. [g] cc, Cellular component. [h] mf, Molecular function.

Comparison of *N. oceanica* and *N. gaditana* gene sets

Recently the genome sequence of a related species, *N. gaditana* (*Ng*, Radakovits et al., 2012) has become available providing an opportunity for direct comparison. It was reported that 2,733 genes (30.2% of the total gene models) in the *Ng* genome were exclusive to the species compared to *E. siliculosus* and other distantly related algae (Radakovits et al., 2012). The assembly sizes were ~28.7 Mb for *N. oceanica* CCMP1779 (*No*) and ~29 Mb for *Ng* with a larger protein number identified in *No* (12,012) compared to *Ng* (9,053). To identify unique and conserved gene repertories between the two *Nannochloropsis* species, we first compiled annotated protein coding sequences from both as well as *E. siliculosus* and defined orthologous groups (OGs).

An OG contains a group of genes that were descendants of a single ancestral gene in the most recent common ancestor of both *Nannochloropsis* species. Among 6,395 OGs identified, 5,048 OGs contain genes from both *Nannochloropsis* species. These "shared" OGs contain 5,324 *No* and 5,251 *Ng* genes, respectively. On the other hand, 6,688 *No* and 3,802 *Ng* genes are in single species OGs (which is indicative of gene loss in the other species lineage) or are singleton genes. To evaluate if any of the presumptive *No*-specific genes had a match in the *Ng* genome and, thus, were not truly species-specific, a similarity search was carried out using *No* protein sequences against *Ng* genome sequences. Of the 6,688 presumptive *No*-specific genes, 2,394 had ≥1 significant matches (see Materials and Methods) to the *Ng* genome, while 4,294 remain *No*-specific. Among 3,802 presumptive *Ng*-specific genes, 1,153 of them have ≥1 significant matches to the *No* genome and 2,649 remain *Ng*-specific.

Some of these species-specific genes may be relevant to biological differences between the two species, perhaps related to their distinct life histories. However, they could also be false positive predictions. Using three lines of evidence, we show that some of these species-specific genes are likely authentic. The first is through examining their Annotation Edit Distance (AED), a score that reflects the annotation quality with a range between 0 (perfect match to similar sequences or transcript evidence) and 1 (no match) (Eilbeck et al., 2009). The AED distributions of *No* genes in conserved OGs and those that are species-specific are shown in Figure 3.6. Here conserved OGs refer to OGs with the same number of genes from both *Nannochloropsis* species. Genes in conserved OGs have an average AED score of 0.35, significantly lower than that of species-specific genes (0.73, Kolmogorov-Smirnov Test, *p*<2.2e-16). Given an AED closer to 1 indicating diminishing support, species-specific genes generally have less support based on similarity or transcript evidence compared to conserved genes. Nonetheless, 34.8% of *No*-specific genes have AED<0.5, indicating 50% of the annotated regions overlap with ≥1 similar sequences and/or transcripts. Thus, some of these species-specific genes are likely not spurious.

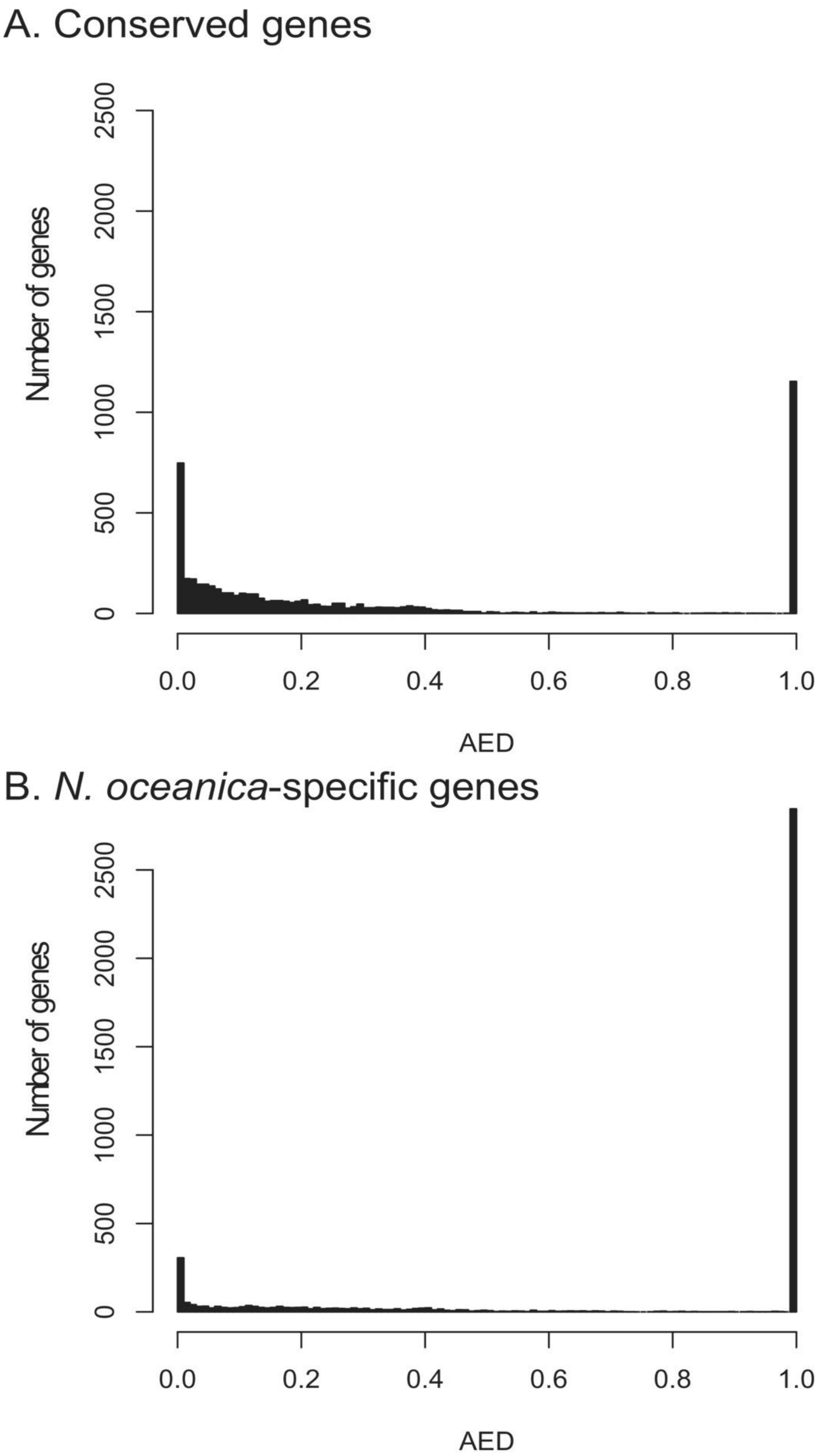


Figure 3.6. AED distributions of conserved and *N. oceanica*-specific genes.

Figure 3.6 (cont'd)

(A) AED distribution of *N. oceanica* genes in a conserved OG. AED stands for annotation edit distance.

(B) AED distribution of *N. oceanica*-specific genes.

The second line of evidence is that a number of *No*-specific and *Ng*-specific genes have putative *E. siliculosus* orthologs. Among 1,040 OGs without ***Ng*** gene, 863 contain both *No* and *E. siliculosus* genes. Similarly, in 307 OGs without *No* gene, 236 have genes from both *Ng* and *E. siliculosus*. These findings indicate that a number of species-specific genes are authentic and the reason they are species-specific is most likely due to gene loss and/or missing annotation in one of the *Nannochloropsis* species. The third line of evidence is that 1,086 *No*- and 253 *Ng*-specific genes have a significant match to annotated proteins from other species that can be used for functional category annotation based on sequence similarity (see previous section on Blast2GO).

We conducted enrichment tests to examine which functional categories tend to be associated with conserved genes or species-specific genes. Here conserved genes are defined as genes that reside in OGs with the same number of genes from both *Nannochloropsis* species. Species-specific genes on the other hand are defined as annotated genes from one species that do not have a protein or genomic match from the second species. We found that conserved genes, as expected, are involved in essential processes including translation, ribosome biogenesis, photosynthesis, and central metabolism (Table 3.6). For species-specific genes, we also identified multiple enriched categories (Table 3.6). However, the degree of enrichment is rather marginal and the test statistics are not particularly robust. This is most likely because there is extremely limited knowledge of gene functions among Heterokont species. One noteworthy enriched GO category (acetyl-CoA carboxylase activity) may reflect subtle differences in fatty acid biosynthesis, which is relevant for the use of the respective organism for the production of biofuel feedstock.

Table 3.6: Enriched GO categories in conserved OGs and *N. oceanica* CCMP1779-specific and *N. gaditana*-specific genes.

GO	**GO G**	**GO No G**	**No GO G**	**No GO No G**	**p**	**Annotation**
Conserved OGs						
GO:0006412 bp	89	21	3288	2582	8.6E-08	translation
GO:0009507 cc	112	33	3265	2570	1.6E-07	chloroplast
GO:0003735 mf	90	24	3287	2579	4.8E-07	structural constituent of ribosome
GO:0005739 cc	157	61	3220	2542	1.9E-06	mitochondrion
GO:0042254 bp	92	31	3285	2572	2.9E-05	ribosome biogenesis
GO:0005840 cc	95	33	3282	2570	3.0E-05	ribosome
GO:0015979 bp	17	0	3360	2603	7.6E-05	photosynthesis
GO:0005743 cc	32	6	3345	2597	4.2E-04	mitochondrial inner membrane
GO:0005737 cc	206	113	3171	2490	3.0E-03	cytoplasm
GO:0006977 bp	11	0	3366	2603	3.5E-03	DNA damage response, signal transduction by p53 class mediator resulting in cell cycle arrest
GO:0006098 bp	27	6	3350	2597	4.1E-03	pentose-phosphate shunt
GO:0055114 bp	285	170	3092	2433	5.9E-03	oxidation-reduction process
GO:0015992 bp	25	6	3352	2597	6.1E-03	proton transport
GO:0051603 bp	9	0	3368	2603	6.4E-03	proteolysis involved in cellular protein catabolic process
GO:0016117 bp	9	0	3368	2603	6.4E-03	carotenoid biosynthetic process
GO:0008565 mf	20	4	3357	2599	7.1E-03	protein transporter activity
GO:0004298 mf	21	4	3356	2599	7.2E-03	threonine-type endopeptidase activity
GO:0051536 mf	37	12	3340	2591	8.5E-03	iron-sulfur cluster binding
GO:0006118 bp	94	46	3283	2557	9.7E-03	electron transport
***N. oceanica* specific genes**						
GO:0004871 mf	15	18	1071	4876	2.7E-04	signal transducer activity
GO:0008131 mf	5	1	1081	4893	1.0E-03	primary amine oxidase activity
GO:0009308 bp	5	1	1081	4893	1.0E-03	amine metabolic process
GO:0046429 mf	4	0	1082	4894	1.1E-03	4-hydroxy-3-methylbut-2-en-1-yl diphosphate synthase activity

Table 3.6 (cont'd)

GO:0030604 mf	3	0	1083	4894	6.0E-03	1-deoxy-D-xylulose-5-phosphate reductoisomerase activity
GO:0004375 mf	3	0	1083	4894	6.0E-03	glycine dehydrogenase (decarboxylating) activity
GO:0005622 cc	42	114	1044	4780	6.0E-03	intracellular
GO:0048038 mf	5	3	1081	4891	6.7E-03	quinone binding
GO:0003989 mf	6	5	1080	4889	7.1E-03	acetyl-CoA carboxylase activity
GO:0003676 mf	64	197	1022	4697	8.4E-03	nucleic acid binding
N. gaditana **specific genes**						
GO:0009507 cc	25	107	228	2648	7.31E-05	chloroplast
GO:0016791 mf	5	2	248	2753	7.39E-05	phosphatase activity
GO:0019843 mf	7	10	246	2745	2.56E-04	rRNA binding
GO:0016730 mf	3	0	250	2755	5.89E-04	oxidoreductase activity, acting on iron-sulfur proteins as donors
GO:0019685 bp	3	0	250	2755	5.89E-04	photosynthesis, dark reaction
GO:0005840 cc	11	38	242	2717	1.95E-03	ribosome
GO:0036068 bp	3	1	250	2754	2.21E-03	light-independent chlorophyll biosynthetic process
GO:0009535 cc	11	42	242	2713	3.76E-03	chloroplast thylakoid membrane
GO:0031347 bp	2	0	251	2755	7.05E-03	regulation of defense response
GO:0009539 cc	2	0	251	2755	7.05E-03	photosystem II reaction center
GO:0030071 bp	2	0	251	2755	7.05E-03	regulation of mitotic metaphase/anaphase transition
GO:0004722 mf	3	3	250	2752	9.71E-03	protein serine/threonine phosphatase activity
GO:0009772 bp	3	3	250	2752	9.71E-03	photosynthetic electron transport in photosystem II
GO:0030076 cc	3	3	250	2752	9.71E-03	light-harvesting complex
GO:0009982 mf	3	3	250	2752	9.71 E-03	pseudouridine synthase activity

Table 3.6 (cont'd)

GO G, Number of genes with the GO annotation in the conserved OGs or species- specific groups of genes. No GO G, Number of genes without the GO annotation in conserved OGs or species- specific groups of genes. GO No G, Number of genes with the GO annotation not in conserved OGs or species- specific groups of genes. No GO No G, Number of genes without the GO annotation not in conserved OGs or species- specific groups of genes. p, Fisher's exact test P value. bp, biological process. mf, molecular function. cc, cellular component.

CONCLUSIONS

The *N. oceanica* CCMP1779 draft genome and its automated annotation reported here provides a starting point for further exploration of the biology and utility of this species. At the same time, a large number of experts in biochemistry and molecular biology manually annotated selected genes mostly in pathways relevant for biofuel production, as well as other cellular and regulatory aspects. However, the current annotation must be considered a work in progress and we would like to encourage the reader to visit the project website at www.bmb.msu.edu/Nannochloropsis.html for further exploration of the data. Comparison of the gene repertoires between *N. oceanica* and *N. gaditana* has indicated that the differences between these two species are comparable in magnitude to those observed between monocotyledonous and dicotyledonous plant species, which diverged from each other 150–200 million years ago. A substantial number of species-specific genes identified may reflect physiological and biochemical differences, which can be explored in future comparative studies. Experimental verification will likely provide insights into adaptations of the respective species to its specific ecological niche, and may also reveal the need for considering species-specific characteristics during genetic engineering for the purpose of biofuels feedstock production. For example possible differences in sets of genes relevant to fatty acid biosynthesis (acetyl-CoA carboxylase) may help us design strategies to maximize oil production in a given strain. Availability of genome sequences of different *Nannochloropsis* species in combination with targeted gene replacement by homologous recombination, which currently has only been documented for an *N. oceanica* strain closely related to CCMP1779 (Kilian et al., 2011), will not only expedite the functional analysis of individual genes in *Nannochloropsis*, but is a prerequisite for future synthetic biology and engineering efforts focused on developing *Nannochloropsis* into a versatile feedstock for different industrial purposes.

MATERIALS AND METHODS

Strains and growth conditions

The *Nannochloropsis sp.* strain used was CCMP1779, available from The Provasoli-Guillard National Center for Culture of Marine Phytoplankton (https://ncma.bigelow.org/). The cells were grown in liquid cultures under continuous light (~80 µmole photons m^{-2} s^{-1}). For N-replete growth, f/2 medium with 2.5 mM nitrate (f/2+N) was used (Vieler et al., 2012). For nitrogen-deprived experiments, N deprivation was applied by growth in f/2+N to 1×10^{7} cells mL^{-1}, followed by transfer to f/2 without nitrogen source to 5×10^{6} cells mL^{-1} for an additional 30 hours.

Nuclear transformation by electroporation

Initial transformation experiments were done with a construct described for nuclear transformation of *C. reinhardtii*, pHyg3 (Berthold et al., 2002), containing a *C. reinhardtii* α-tubulin promoter and the coding sequence of the *Streptomyces hygroscopicus* aph7 gene conferring resistance to Hygromycin B. Subsequently, a plasmid custom made by DNA Cloning Service (http://www.dna-cloning.com) 497pLC-Hpt-SfiI, which contains a 35S promoter region, the aph7 coding sequence and a 35S terminator was digested with restriction endonucleases *Xba*I and *Xho*I to eliminate the promoter region. Additionally, the plasmid contains two *Sfi*I sites to allow directional cloning of further expression cassettes. The native LDSP promoter was amplified from CCMP1779 genomic DNA using the forward and reverse primers 5′-GGCCTAGGTACGTA-GGTCTCTAAGATGGAGTGGATGG-3′ and 5′-TTCAGCTG-TGTTGATGCGGGCTGAGATTGG-3′ and the resulting 790 bp PCR product cloned to the pGEMteasy vector system (Promega, http://www.promega.com) for sequencing resulting in pGEM-pLDSP. The promoter region was released from pGem-pLDSP by *Avr*II and *Pvu*II digest and blunt cloned to the dephosphorylated 497pLC-Hpt-SfiI backbone to result in the selection plasmid pSELECT100.

For transformation cells were harvested at a density of $1–2\times10^{7}$ cells mL^{-1}, washed with ice cold 375 mM sorbitol three times and resuspended in a final volume of 0.2 mL to a concentration of 5×10^{8} cells mL^{-1}. In addition to 2–10 µg *Sna*BI linearized Plasmid DNA, a 10fold excess of salmon sperm DNA (Invitrogen, http://www.invitrogen.com) was supplied into the 2 mm electroporation cuvette. Electroporation was performed using a Bio-Rad (http://www.bio-rad.com) GenePulser II set to 600 Ώ resistance at a field strength of 11 kV cm^{-1}leading to time constants of 20 to 25 ms. After the pulse the cells were resuspended in 5 mL of f/2 media and allowed to recover for 48 h in continuous light with shaking before they were spread on selection agar containing 50 µg ml^{-1} Hygromycin B using warm top agar (f/2 media, 0.05% Phytoblend (Caisson Laboratories, http://www.caissonlabs.com) in 1:1 dilution (vol:vol). Resistant colonies were observed as early as 10–14 days after electroporation; colonies were usually transferred after about 3 weeks.

Southern analysis

For Southern analysis, 10 µg of DNA were digested with *Bam*HI and *Bam*HI/*Xba*I for pHyg3, or *Bam*HI only for the pSELECT100 and separated on an agarose gel (0.9% agarose, 75 Volts, 6 h runtime, 15 cm gel length) before blotting to a Hybond Nylon (Amersham, GE Health Care, http://www.gelifesciences.com) positively charged membrane overnight. Hybridization and detection was performed using the DIG labeling and detection system following the manufacturer's instructions (Roche Applied Sciences, http://www.roche-applied-sciences.com). Hybridization was done in 10 ml ULTRAhyb buffer (Invitrogen) at 68°C for pHyg3 or 42°C for pSelect100. The oligonucleotides for the probe synthesis by PCR were 5′-ACCAACATCTTCGTGGACCT-3′ and 5-‘CTCCTCGAACACCTCGAAGT-3′ for pHyg3 transformed cells and 5′-CGCGCTACTTCGAGCGGAGG-3′ and 5′-GCGCTTCTGCGGGCGATTTG-3′ for pSelect100 transformed cells using the respective plasmid as a template.

DNA and RNA preparation for sequencing and analysis

For preparation of nuclear DNA a 50 mL cell culture (OD_{750} = 0.4 to 0.5) was harvested by centrifugation (4,500× g, 5 min). The cell pellet was lysed in 2× cetyltrimethylammonium bromide (CTAB) buffer (2% CTAB, 100 mM Tris-HCl pH 8.0, 1.4 M NaCl, and 20 mM EDTA) and incubated at 60°C for 60 min. The lysate was mixed with 1 volume of phenol/chloroform and centrifuged (13,000× g min). Transferred the supernatant to a new tube and repeated this step at least once until there was no white interphase. The DNA was precipitated by 1 volume isopropanol and 70% ethanol. High molecular weight DNA was examined by DNA gel electrophoresis.

To generate material for RNA-sequencing, cells were grown in 200 ml f/2+N to 1×10^7 cells mL^{-1}. The cultures were split in half and cells were collected by centrifugation (4,500× g, 5 min), with one pellet being resuspended in 200 mL f/2+N, and the other in 200 mL f/2-N. After 30 hours, the total RNA was isolated using TRIzol Reagent (Invitrogen) according to manufacturer's instructions. The RNA samples were cleaned up using RNeasy columns (Qiagen, http://www.qiagen.com) following the manufacturer's instruction.

Assessment of RNA quantity and quality

The evaluation of RNA quantity and quality was done spectrophotometrically by UV absorbance profile. Additional analysis was performed using an RNA 6000 Nano LabChip Kit for microcapillary electrophoresis (Agilent 2100 Bioanalyzer, http://www.home.agilent.com). This eukaryotic total RNA nano-assay generated information about RNA integrity through electropherograms, gel picture, and RIN value (RNA Integrity Number) (Schroeder et al., 2006).

Genome sequencing and hybrid genome assembly

For genome sequencing, two approaches were employed. First, Illumina GS-II was used to generate 55 bp paired-end reads with a 550 bp library and ~2.3 Gb sequences were generated. The Illumina reads were filtered using FASTX (http://hannonlab.cshl.edu/fastx_toolkit/) with a minimum Phred quality score of 20. Next, Velvet (Zerbino and Birney, 2008) was used to assemble filtered Illumina reads, and a range of ***k***-mer length were tested (31, 33, 35, 37, 39, 41, 43, 45, and 47). To determine an optimal k-mer length, 454 reads longer than 500 bp and *de novo* assembled transcripts were mapped to the genome assemblies using GMAP (Wu and Watanabe, 2005). Based on how well the 454 reads and *de novo* transcripts mapped on to the Illumina assemblies, as well as N50s, numbers of contigs, assembly sizes, and numbers of total reads assembled, ***k***-mer length of = 35 was chosen for generating the final Illumina assembly. Newbler (454 Life Sciences) was used to assemble 454 reads (single-end reads, 449.9 MB sequences) with the “Large Genome Option”.

Illumina and 454 assemblies were combined by iterative Minimus2 (Sommer et al., 2007). Minimus2 was first run with a minimum identity of 98% among and between Illumina and 454 contigs based on and all-against-all contig similarity searches with BLAST (Altschul et al., 1997). If one contig had an alignment ≥200 bp and an identity ≥98% with ≥2 other contigs, only the longest contig among the matching contigs was kept and the rest were set aside before re-running Minimus2. This step was performed because such contigs may represent mis-assembled sequences and will confound Minimus2 as to which contigs it should assemble. A similar procedure was used in the assembly of the *A. laibachii* genome (Kemen et al., 2011). In the next iteration, the contigs set aside beforehand were added back to the assembly and Minimus2 was run again. After another three iterations of Minimus2 run, an optimized assembly was generated.

To assess assembly quality, long 454 reads with high Phred scores were mapped to the genome assembly. First, 454 reads were trimmed from the 3′ end with a minimum Phred score of 20. Then, sequences longer than 200 bp were aligned to the genome using BLAST to determine if a 454 read was broken up in >1 contigs. We also used *de novo* transcript assemblies (see next section) to assess genome assembly quality. The genomic sequence data are deposited in NCBI SRA (SRP013753).

Transcript assembly and differential expression analysis

De novo transcript assemblies were generated from 55 bp directional single-end Illumina reads of N-replete and N-depleted conditions (NCBI/GEO GSE36959) using Oases (http://www.ebi.ac.uk/~zerbino/oases/). First, Oases was run for *k*-mer lengths of 23, 25, 27, 29, 31, 33, 35, and 37, and the results were compiled. To identify a set of high confidence transcripts from the *de novo* assemblies, proteins from six sequenced heterokont genomes, including *Ectocarpus siliculosus* (Cock et al., 2010), *Pythium ultimum* (Lévesque et al., 2010), *Phytophthora sojae* (Tyler et al., 2006)**,** *Phytophthora ramorum* (Tyler et al., 2006)**,** *Thalassiosira Pseudonana* (Armbrust et al., 2004)**,** and *Phaeodactylum tricornutum* (Bowler et al., 2008), were aligned to the *de novo* transcripts and only those with significant matches to known proteins were kept. These transcripts with cross-genome matches were mapped back to the Illumina genome assemblies to evaluate genome assembly quality. In addition to *de novo* transcript assembly, we generated a genome-based transcript assembly.

Transcriptomic reads from N-replete and N-depleted conditions were separately mapped to the hybrid genome assembly using Tophat (Trapnell et al., 2009) (parameters: -I 10 –I 3000 –library-type fr-unstranded –g 1). The mapped reads were assembled into transcripts using Cufflinks (Trapnell et al., 2010) (-I 3000 –library-type fr-secondstrand) and a set of transcripts was generated for each condition.

Genome annotation

The MAKER genome annotation pipeline (Cantarel et al., 2008) was used to annotate the genome. The first run of MAKER was performed using the est2genome option in the absence of a trained gene predictor. Transcripts from both N-replete and N-deprived growth conditions were provided to MAKER along with protein sequences from the above mentioned six sequenced heterokonts. Gene models obtained from the first run were used to train *ab initio* gene prediction programs SNAP (Korf, 2004) and Augustus (Stanke and Waack, 2003). With the trained models, MAKER was rerun. The gene models from the rerun were used for training SNAP and Augustus again. The second round training models were provided to run MAKER for the third time to generate the final annotations. The protein sequences were searched for Pfam domain Hidden Markov Models using HMMER3 (Finn et al., 2010) with trusted cutoffs. CEGMA was run on the genome assembly using default settings (Parra et al., 2007). A total of 11,973 genes (12,012 protein models considering alternative splice forms) were recovered with an average AED score of 0.555. During the course of the study, a new version of MAKER was released. Thus we conducted a second annotation run with the most recent MAKER version, a more recent repeat library, and a larger protein evidence dataset. Given that the AED distributions were highly similar between these two annotation datasets (Figure 3.4A and B) only annotation results from the first set of analysis were used throughout.

InterProScan (Quevillon et al., 2005) was used to identify Pfam protein domains within the predicted protein sets from *N. oceanica* CCMP1779 and six other heterokonts. Protein families were identified by grouping proteins with identical protein domains, and the number of proteins from each species that were classified into each protein family was tallied. Figure 3.4C shows the percentages of proteins that have at least one InterPro domain, and those that have none, of each species.

Functional annotation and determination of differential expression

Blast2GO (Conesa et al., 2005) (http://blast2go.com/b2ghome) was used for functional annotation of predicted protein models with the default settings for the mapping and annotation step. The initial BLAST (Altschul et al., 1997) search was performed with an e-value cut-off of 10^{-5} and a maximum of 20 blast hits. This results in Gene Ontology (GO) annotations of 5,980 *N. oceanica* genes (in 4,012 GOs) and 3,008 *N. gaditana* genes (in 3,205 GOs). Fisher's exact test was used to assess if either the number of conserved or species-specific genes are over-represented in any GO category.

Cuffdiff from the Cufflinks package (Trapnell et al., 2010) was used to analyze the differential gene expression under N-replete and N-deprived growth conditions. Fisher's exact tests were performed to determine the enrichment of each GO category in up- and down-regulated gene clusters and at the 1% significance level based on p-values.

Comparison of *Nannochloropsis* genomes

OrthoMCL (Li et al., 2003) was used to identify Orthologous Groups (OGs) of genes in *N. gaditana*, *N. oceanica* CCMP1779, and *E. siliculosus* (run parameters: percentMatchCutoff = 50, evalueExponentCutoff = −5). BLAST (Altschul et al., 1990) was used to identify significant matches of lineage-specific genes across species. A significant match was defined as identity ≥47.04% (5 percentile in the identity distribution of one-to-one orthologs between *N. gaditana* and *N. oceanica*), Expect value≤10^{-5}, alignment length ≥30 amino acids, and ≥50% of the protein sequence covered in the alignment.

Database tools

To allow easy access to the CCMP1779 genome data, we released a public version of the genome browser along with a basic BLAST tool to search nucleotide and protein databases, accessible at www.bmb.msu.edu/nannochloropsis.html. The genome browser contains EST data aligned to the latest genome assembly as well as alternative gene models in addition to the final models retrieved from the MAKER gene annotation pipeline described above.

ACKNOWLEDGEMENT

We like to acknowledge John Johnston, Michigan State University, for help with the database and server, and Brandon Guelette at Michigan State University for technical assistance with the cell wall analysis. We thank Rebecca Hopkins and Tiffany Liu, Michigan State University, for help with the annotation of light and circadian signaling genes.

REFERENCES

REFERENCES

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol **215**: 403–410

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res **25**: 3389–3402

Andersen RA, Brett RW, Potter D, Sexton JP (1998) Phylogeny of the Eustigmatophyceae Based upon 18S rDNA, with Emphasis on *Nannochloropsis*. Protist **149**: 61–74

Armbrust EV, Berges JA, Bowler C, Green BR, Martinez D, Putnam NH, Zhou S, Allen AE, Apt KE, Bechner M, et al (2004) The Genome of the Diatom *Thalassiosira Pseudonana*: Ecology, Evolution, and Metabolism. Science **306**: 79–86

Berthold P, Schmitt R, Mages W (2002) An Engineered *Streptomyces hygroscopicus* aph 7″ Gene Mediates Dominant Resistance against Hygromycin B in *Chlamydomonas reinhardtii*. Protist **153**: 401–412

Bowler C, Allen AE, Badger JH, Grimwood J, Jabbari K, Kuo A, Maheswari U, Martens C, Maumus F, Otillar RP, et al (2008) The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes. Nature **456**: 239–244

Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, Holt C, Alvarado AS, Yandell M (2008) MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res **18**: 188–196

Cock JM, Sterck L, Rouzé P, Scornet D, Allen AE, Amoutzias G, Anthouard V, Artiguenave F, Aury J-M, Badger JH, et al (2010) The *Ectocarpus* genome and the independent evolution of multicellularity in brown algae. Nature **465**: 617–621

Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics **21**: 3674–3676

Danielewicz MA, Anderson LA, Franz AK (2011) Triacylglycerol profiling of marine microalgae by mass spectrometry. J Lipid Res **52**: 2101–2108

Derelle E, Ferraz C, Rombauts S, Rouzé P, Worden AZ, Robbens S, Partensky F, Degroeve S, Echeynié S, Cooke R, et al (2006) Genome analysis of the smallest free-living eukaryote *Ostreococcus tauri* unveils many unique features. Proc Natl Acad Sci **103**: 11647–11652

Dismukes GC, Carrieri D, Bennette N, Ananyev GM, Posewitz MC (2008) Aquatic phototrophs: efficient alternatives to land-based crops for biofuels. Curr Opin Biotechnol **19**: 235–240

Eilbeck K, Moore B, Holt C, Yandell M (2009) Quantitative measures for the management and comparison of annotated genomes. BMC Bioinformatics **10**: 67

Fawley KP, Fawley MW (2007) Observations on the Diversity and Ecology of Freshwater *Nannochloropsis* (Eustigmatophyceae), with Descriptions of New Taxa. Protist **158**: 325–336

Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, et al (2010) The Pfam protein families database. Nucleic Acids Res **38**: D211–D222

Harris EH (2009) The *Chlamydomonas* Sourcebook: Introduction to *Chlamydomonas* and Its Laboratory Use. Academic Press

Hoek CVD (1995) Algae: An Introduction to Phycology. Cambridge University Press

Hu H, Gao K (2003) Optimization of growth and fatty acid composition of a unicellular marine picoplankton, *Nannochloropsis* sp., with enriched carbon sources. Biotechnol Lett **25**: 421–425

Kemen E, Gardiner A, Schultz-Larsen T, Kemen AC, Balmuth AL, Robert-Seilaniantz A, Bailey K, Holub E, Studholme DJ, MacLean D, et al (2011) Gene Gain and Loss during Evolution of Obligate Parasitism in the White Rust Pathogen of *Arabidopsis thaliana*. PLoS Biol **9**: e1001094

Kilian O, Benemann CSE, Niyogi KK, Vick B (2011) High-efficiency homologous recombination in the oil-producing alga *Nannochloropsis* sp. Proc Natl Acad Sci **108**: 21265–21269

Korf I (2004) Gene finding in novel genomes. BMC Bioinformatics **5**: 59

Lévesque CA, Brouwer H, Cano L, Hamilton JP, Holt C, Huitema E, Raffaele S, Robideau GP, Thines M, Win J, et al (2010) Genome sequence of the necrotrophic plant pathogen *Pythium ultimum* reveals original pathogenicity mechanisms and effector repertoire. Genome Biol **11**: R73

Li L, Stoeckert CJ, Roos DS (2003) OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. Genome Res **13**: 2178–2189

Li S-S, Tsai H-J (2009) Transgenic microalgae as a non-antibiotic bactericide producer to defend against bacterial pathogen infection in the fish digestive tract. Fish Shellfish Immunol **26**: 316–325

Miller R, Wu G, Deshpande RR, Vieler A, Gärtner K, Li X, Moellering ER, Zäuner S, Cornish AJ, Liu B, et al (2010) Changes in Transcript Abundance in *Chlamydomonas reinhardtii* following Nitrogen Deprivation Predict Diversion of Metabolism. Plant Physiol **154**: 1737–1752

Myhre S, Tveit H, Mollestad T, Lægreid A (2006) Additional Gene Ontology structure for improved biological reasoning. Bioinformatics **22**: 2020–2027

Pan K, Qin J, Li S, Dai W, Zhu B, Jin Y, Yu W, Yang G, Li D (2011) Nuclear Monoploidy and Asexual Propagation of *Nannochloropsis Oceanica* (eustigmatophyceae) as Revealed by Its Genome Sequence1. J Phycol **47**: 1425–1432

Parra G, Bradnam K, Korf I (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics **23**: 1061–1067

Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R (2005) InterProScan: protein domains identifier. Nucleic Acids Res **33**: W116–W120

Radakovits R, Jinkerson RE, Fuerstenberg SI, Tae H, Settlage RE, Boore JL, Posewitz MC (2012) Draft genome sequence and genetic transformation of the oleaginous alga *Nannochloropis gaditana*. Nat Commun **3**: 686

Reyes-Prieto A, Weber APM, Bhattacharya D (2007) The Origin and Establishment of the Plastid in Algae and Plants. Annu Rev Genet **41**: 147–168

Rodolfi L, Chini Zittelli G, Bassi N, Padovani G, Biondi N, Bonini G, Tredici MR (2009) Microalgae for oil: strain selection, induction of lipid synthesis and outdoor mass cultivation in a low-cost photobioreactor. Biotechnol Bioeng **102**: 100–112

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol **4**: 406–425

Schneider JC, Roessler P (1994) Radiolabeling Studies of Lipids and Fatty Acids in *Nannochloropsis* (eustigmatophyceae), an Oleaginous Marine Alga1. J Phycol **30**: 594–598

Schroeder A, Mueller O, Stocker S, Salowsky R, Leiber M, Gassmann M, Lightfoot S, Menzel W, Granzow M, Ragg T (2006) The RIN: an RNA integrity number for assigning integrity values to RNA measurements. BMC Mol Biol **7**: 3

Siaut M, Heijde M, Mangogna M, Montsant A, Coesel S, Allen A, Manfredonia A, Falciatore A, Bowler C (2007) Molecular toolbox for studying diatom biology in *Phaeodactylum tricornutum*. Gene **406**: 23–35

Simionato D, Sforza E, Corteggiani Carpinelli E, Bertucco A, Giacometti GM, Morosinotto T (2011) Acclimation of *Nannochloropsis gaditana* to different illumination regimes: Effects on lipids accumulation. Bioresour Technol **102**: 6026–6032

Sommer DD, Delcher AL, Salzberg SL, Pop M (2007) Minimus: a fast, lightweight genome assembler. BMC Bioinformatics **8**: 64

Srinivas R, Ochs C (2012) Effect of UV-A Irradiance on Lipid Accumulation in *Nannochloropsis oculata*. Photochem Photobiol **88**: 684–689

Stanke M, Waack S (2003) Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics **19**: ii215–ii225

Sukenik A, Carmeli Y (1990) Lipid Synthesis and Fatty Acid Composition in *Nannochloropsis* Sp. (eustigmatophyceae) Grown in a Light-Dark Cycle1. J Phycol **26**: 463–469

Tonon T, Harvey D, Larson TR, Graham IA (2002) Long chain polyunsaturated fatty acid production and partitioning to triacylglycerols in four microalgae. Phytochemistry **61**: 15–24

Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics **25**: 1105–1111

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol **28**: 511–515

Tyler BM, Tripathy S, Zhang X, Dehal P, Jiang RHY, Aerts A, Arredondo FD, Baxter L, Bensasson D, Beynon JL, et al (2006) *Phytophthora* Genome Sequences Uncover Evolutionary Origins and Mechanisms of Pathogenesis. Science **313**: 1261–1266

Vieler A, Brubaker SB, Vick B, Benning C (2012) A Lipid Droplet Protein of *Nannochloropsis* with Functions Partially Analogous to Plant Oleosins. Plant Physiol **158**: 1562–1569

Weyer KM, Bush DR, Darzins A, Willson BD (2010) Theoretical Maximum Algal Oil Production. BioEnergy Res **3**: 204–213

Wijffels RH, Barbosa MJ (2010) An Outlook on Microalgal Biofuels. Science **329**: 796–799

Wu TD, Watanabe CK (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics **21**: 1859–1875

Xu F, Cai Z-L, Cong W, Ouyang F (2004) Growth and fatty acid composition of *Nannochloropsis* sp. grown mixotrophically in fed-batch culture. Biotechnol Lett **26**: 1319–1322

Zdobnov EM, Apweiler R (2001) InterProScan – an integration platform for the signature-recognition methods in InterPro. Bioinformatics **17**: 847–848

Zerbino DR, Birney E (2008) Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res **18**: 821–829

CONCLUDING REMARKS

CONCLUDING REMARKS

The general metabolic changes under N deprivation are similar between phylogenetically distant micro-algal species of the model oragnism *C. reinhardtii* and *N. oceanica*; biological processes such as DNA replication and photosynthesis are down-regulated while lipid metabolism is up-regulated (Chapter 1 and Chapter 3). However, I showed that genes involved in stress response tend to be lineage-specifically retained and subsequent functional gains occur frequently after gene duplication (Chapter 2). Therefore, the gene repertoire involved in stress response in each micro-algal species could be highly species-specific, reflecting the phenotypic diversity in stress response and lipid accumulation. This finding prompts the rethinking about the use of a single model organism in the research of micro-algal stress response. An existing model organism could offer insight into the general biology of stress response in microalgae, but characterizing diverse micro-algal species, especially the ones linked to biofuel production, is also necessary to discover their uniqueness in stress response and the ensuing lipid accumulation. For example, obtaining the genome sequence of oil rich *N. oceanica* (Chapter 3) opened up the possibility of developing it into an alternative model organism. Overall, my research serves as one of the first steps towards the final goal of identifying and engineering the ideal alga for biofuel production.