9-2010

The *Chlorella variabilis* NC64A Genome Reveals Adaptation to Photosymbiosis, Coevolution with Viruses, and Cryptic Sex

Guillaume Blanc
Aix-Marseille Université, guillaume.blanc@igs.cnrs-mrs.fr

Garry Duncan
Nebraska Wesleyan University, gduncan@nebrwesleyan.edu

Irina Agarkova
University of Nebraska-Lincoln, iagarkova2@unl.edu

Mark Borodovsky
Georgia Institute of Technology - Main Campus

James Gurnon
University of Nebraska-Lincoln, jgurnon2@unl.edu

See next page for additional authors

The Plant Cell, Vol. 22: 2943–2955, September 2010, www.plantcell.org © 2010 American Society of Plant Biologists

RESEARCH ARTICLES

The *Chlorella variabilis* NC64A Genome Reveals Adaptation to Photosymbiosis, Coevolution with Viruses, and Cryptic Sex [C][W]

Guillaume Blanc,[a,1] Garry Duncan,[b] Irina Agarkova,[c] Mark Borodovsky,[d] James Gurnon,[c] Alan Kuo,[e] Erika Lindquist,[e] Susan Lucas,[e] Jasmyn Pangilinan,[e] Juergen Polle,[f] Asaf Salamov,[e] Astrid Terry,[e] Takashi Yamada,[g] David D. Dunigan,[c] Igor V. Grigoriev,[e] Jean-Michel Claverie,[a] and James L. Van Etten[c]

[a] Centre National de la Recherche Scientifique, Laboratoire Information Génomique et Structurale UPR2589, Aix-Marseille Université, Institut de Microbiologie de la Méditerranée, 13009 Marseille, France
[b] Biology Department, Nebraska Wesleyan University, Lincoln, Nebraska 68504-2796
[c] Department of Plant Pathology, University of Nebraska, Lincoln, Nebraska 68583-0722
[d] Wallace H. Coulter Department of Biomedical Engineering, School of Computational Science and Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332
[e] Department of Energy Joint Genome Institute, Walnut Creek, California 94598
[f] Brooklyn College of the City University of New York, Department of Biology, Brooklyn, New York 11210-2889
[g] Department of Molecular Biotechnology, Graduate School of Advanced Science of Matter, Hiroshima University, Higashi-Hiroshima 739-8530, Japan

***Chlorella variabilis* NC64A, a unicellular photosynthetic green alga *(Trebouxiophyceae*), is an intracellular photobiont of *Paramecium bursaria* and a model system for studying virus/algal interactions. We sequenced its 46-Mb nuclear genome, revealing an expansion of protein families that could have participated in adaptation to symbiosis. NC64A exhibits variations in GC content across its genome that correlate with global expression level, average intron size, and codon usage bias. Although *Chlorella* species have been assumed to be asexual and nonmotile, the NC64A genome encodes all the known meiosis-specific proteins and a subset of proteins found in flagella. We hypothesize that *Chlorella* might have retained a flagella-derived structure that could be involved in sexual reproduction. Furthermore, a survey of phytohormone pathways in chlorophyte algae identified algal orthologs of *Arabidopsis thaliana* genes involved in hormone biosynthesis and signaling, suggesting that these functions were established prior to the evolution of land plants. We show that the ability of *Chlorella* to produce chitinous cell walls likely resulted from the capture of metabolic genes by horizontal gene transfer from algal viruses, prokaryotes, or fungi. Analysis of the NC64A genome substantially advances our understanding of the green lineage evolution, including the genomic interplay with viruses and symbiosis between eukaryotes.**

INTRODUCTION

Green algae (phylum Chlorophyta) are a highly diverse group of photosynthetic eukaryotes from which the terrestrial plant lineage emerged >1 billion years ago (Heckman et al., 2001). During the evolutionary history of Earth, they have become major players in global energy/biomass production and biogeochemical recycling (Grossman, 2005). Algae originally included in the genus *Chlorella* are among the most widely distributed and frequently encountered algae in freshwaters (Fott and Novakova, 1969). They exist in aqueous environments as well as on land. They are typically small (~2 to 10 μm in diameter), unicellular, coccoid, nonmotile, and contain a single chloroplast. Some have a rigid cell wall, and they are reported to lack a sexual cycle (Takeda, 1991). Accessions of *Chlorella* were extensively used as model systems in early research on photosynthesis (Benson, 2002).

Over a hundred algal isolates were originally assigned to the genus *Chlorella*, but their taxonomy classification has long remained unreliable because of their lack of conspicuous morphological characters. Recent molecular analyses now separate them into two classes of chlorophytes, the Trebouxiophyceae, which contains the true *Chlorella*, and the Chlorophyceae (Takeda, 1988; Huss et al., 1999). Here, we report on the genome of *Chlorella* sp NC64A (NC64A), recently renamed *Chlorella variabilis* NC64A (Ryo et al., 2010), that is a bona fide member of the true *Chlorella* genus, belonging to the class Trebouxiophyceae (see Supplemental Figure 1 online). The true *Chlorella* species, including NC64A, are characterized by glucosamine as a major component of their rigid cell walls (Takeda, 1991; Chuchird et al., 2001). The Trebouxiophyceae contain most of the known green algal

[1] Address correspondence to guillaume.blanc@igs.cnrs-mrs.fr.
The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantcell.org) is: Guillaume Blanc (guillaume.blanc@igs.cnrs-mrs.fr).
[C]Some figures in this article are displayed in color online but in black and white in the print edition.
[W]Online version contains Web-only data.
www.plantcell.org/cgi/doi/10.1105/tpc.110.076406

endosymbionts (Friedl and Bhattacharya, 2002), living in lichens, unicellular eukaryotes, plants, and animals (e.g., mussels, hydra, etc). Most *Chlorella* species are naturally free-living; however, NC64A is a hereditary photosynthetic endosymbiont (i.e., photobiont) of the unicellular protozoan *Paramecium bursaria* (Karakashian and Karakashian, 1965). This symbiosis is facultative in lab conditions since both the paramecium and NC64A can be cultivated separately. NC64A is also a host for a family of large double-stranded DNA viruses that are found in freshwater throughout the world; the genomes of six of these viruses have been sequenced (Wilson et al., 2009). Like other microalgae, there is an increasing interest in using *Chlorella* in a variety of biotechnological applications, such as biofuels (Schenk et al., 2008), sequestering CO_2 (Chelf et al., 1993), producing molecules of high economic value, or removing heavy metals from wastewaters (Rajamani et al., 2007). The sequence of the NC644A genome presented here will help in the optimization of these various processes, while further documenting the evolution of the green lineage.

RESULTS AND DISCUSSION

Global Genome Structure

The 46.2-Mb NC64A nuclear genome was sequenced at 9× coverage using the whole-genome shotgun Sanger sequencing approach. The genome size of NC64A is intermediate compared with those of Mamiellale (12.6 to 21.9 Mb) and *Chlamydomonas reinhardtii* (121 Mb) (Table 1). Sequence assembly yielded 413 scaffolds with lengths >1 kb (see Supplemental Table 1 online). Eighty-nine percent of the genome assembly is contained in 30 scaffolds with lengths ranging from 494 kb to 3.12 Mb (Figure 1). Mapping of 7624 clustered EST sequences onto the genome sequences suggests that the assembly contains >97% of the gene complement. The NC64A karyotype resolved by pulse field gel electrophoresis analysis revealed 12 chromosomes ranging in size from ~1.1 to 8.6 Mb (see Supplemental Figure 2 online).

The nuclear genome sequences have the highest average GC content (67.2% GC) reported so far in sequenced eukaryotic genomes (Table 1). However, several genomic segments present in scaffolds, ranging from 40 to 625 kb, have conspicuously lower GC contents (55 to 65% GC) than the rest of the genome (Figure 1). These low-GC regions represent 15.6% of the total genome size (6.20 Mb). They have a significantly higher frequency of genes with EST support than does the rest of the genome (Kruskal-Wallis test P value = $P_{KWT} < 0.0001$), suggesting that they correspond to regions of higher transcriptional activity (Figure 2A). In addition, genes located in low-GC regions exhibit significantly shorter introns (Figure 2B) and a less biased codon usage (Figure 2C) relative to the high-GC regions ($P_{KWT} < 0.0001$). Low-GC regions are also enriched in repeated sequences (most prominent in the 60 to 65% GC range; Figure 2E), but the trend is only marginally significant ($P_{KWT} = 0.024$). Although the median exon density is slightly smaller for low-GC regions (Figure 2D), the difference from that found in the high-GC regions is not statistically significant ($P_{KWT} > 0.05$). The majority (1100) of the 1384 NC64A proteins encoded in low-GC regions have their best BLASTP match to homologs in chlorophytes and land plants (see Supplemental Figure 3 online). This suggests that the low-GC regions did not result from an invasion of horizontally transferred foreign DNA sequences.

Low-GC genomic regions also exist in the prasinophytes *Micromonas* and *Ostreococcus*, where their origin and function are still unclear (Palenik et al., 2007; Worden et al., 2009). As in NC64A, the *Micromonas* low-GC chromosomes exhibit higher transcription levels than do the normal-GC chromosomes (Worden et al., 2009). These features common to *Micromonas*, *Ostreococcus*, and *Chlorella* suggest that variation in GC content is a characteristic of many chlorophytes. However, the nature of genes present in the NC64A low-GC regions does not suggest a specific function or a mechanism by which their compositional shift evolved. However, we noticed that the NC64A low-GC regions exhibited a significant underrepresentation of genes involved in transcription, chromatin structure and dynamics, and extracellular structures (see Supplemental Table 2 online).

Repeated Sequences

Only a few algal repetitive sequences are available in public databases. This prevented us from performing an exhaustive search for repetitive sequences based on a set of reference sequences. Therefore, we used a de novo identification approach where repeated sequences are defined as any sequence

Table 1. Comparison of NC64A Genome Statistics to Those of Sequenced Chlorophyte Genomes

Features	NC64A	*C. reinhardtii*	*Micromonas* CCMP1545	*Micromonas* RCC299	*O. tauri*	*O. lucimarinus*
Nuclear genome size (Mb)	46.2	121	21.9	20.9	12.6	13.2
Number of chromosomes	12[a]	17	19	17	20	21
GC content total (%)	67	64	65	64	59	60
Gene count	9,791	15,143	10,575	10,056	7,892	7.651
Avg. protein length (aa)	456	444	439	473	387	399
Avg. gene density (kb/gene)	4.7	5.0	2.1	2.2	1.6	1.7
Avg. number of exons per gene	7.3	8.3	1.9	1.6	1.6	1.3
Avg. exon length (nt)	170	190	731	958	750	970
Avg. intron length (nt)	209	373	187	163	126	187
Avg. coding sequence (%)	29	17	64	68	73	69

aa, amino acids; nt, nucleotides.

[a]Estimation based upon pulse field gel electrophoresis analysis.

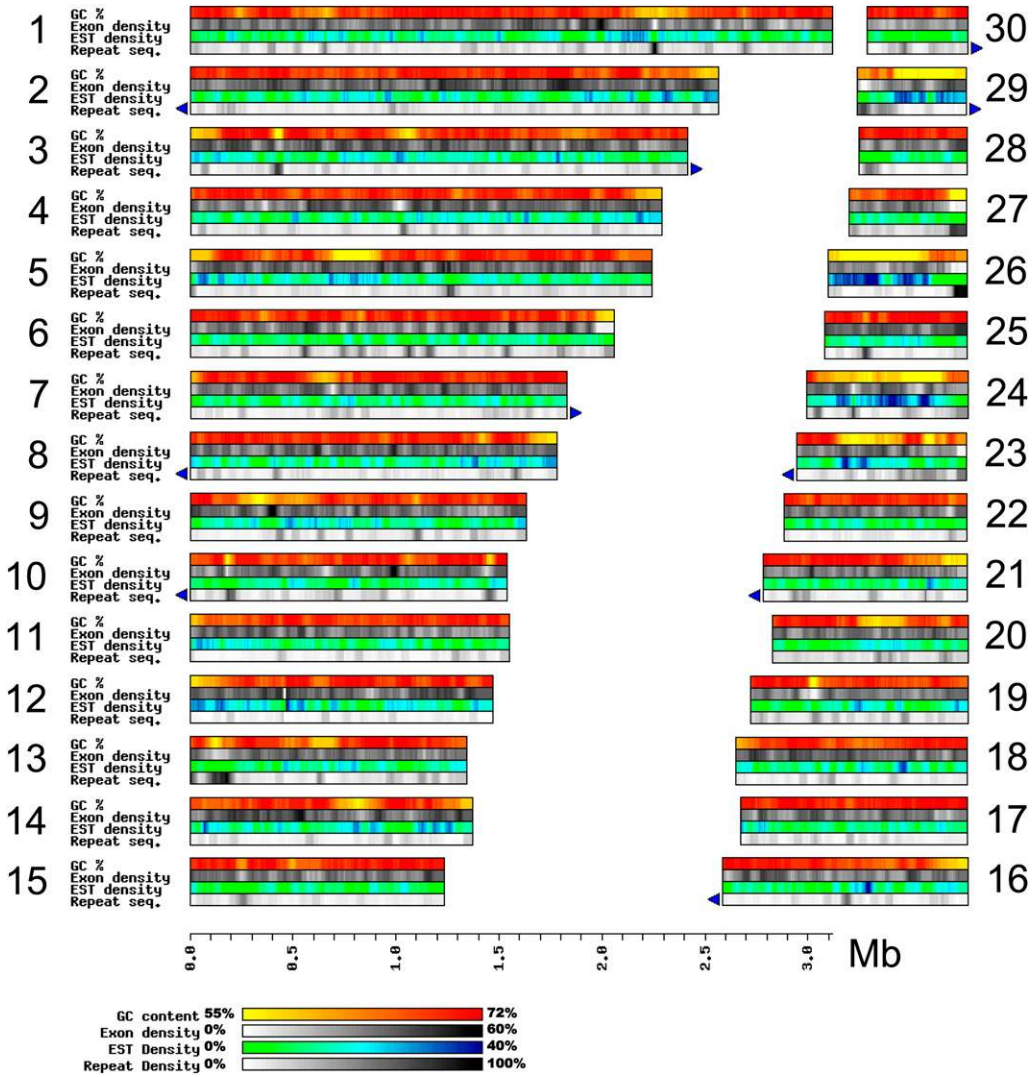


Figure 1. General Characteristics of the *Chlorella* sp NC64A Genome Assembly.

This figure represents the 30 major scaffolds, which contain 89% of the total genome. GC percentage, exon density, EST density, and repeat density were calculated in 40-kb sliding windows with a step of 5 kb. Density was calculated as the percentage of nucleotide in the window covered by the relevant feature (i.e., exon, EST, or repeat sequence). Blue triangles represent telomeric repeat arrays.

with more than one copy in the genome (as detected by BLASTN with an E-value < 1e-5), regardless of its size and nature (transposable element, simple repeat, duplicated gene, or low complexity sequences). The cumulative lengths of such repeated sequences represent 5.53 Mb (12%) of the genome (see Supplemental Table 3 online), which makes NC64A relatively repeat-poor compared with land plants (repeat content ranges from 20 to 30% in *Arabidopsis thaliana* to >90% in large genomes such as wheat [*Triticum aestivum*]). The content in repeated sequences is probably slightly underestimated because repeats frequently flanked sequence gaps. Half of the repeated sequences (51.6%) have no resemblance to known repeat families (see Supplemental Table 3 online). About 10% (536 kb or 1.2% of the genome) contain open reading frames with deduced protein sequences similar to proteins in public databases (excluding transposable element related proteins) and therefore correspond to highly similar gene duplicates or gene fragments (at the nucleotide level). An additional 40.2% could be classified in known repetitive sequence families based on TBLASTX sequence similarity searches (E-value < 1e-15) against the Repbase database. NC64A has the major classes of known transposable elements (see Supplemental Table 3 online): long terminal repeat (LTR) retrotransposons (Gypsy-like elements and TY1/Copia-like elements), non-LTR retrotransposons (RandI, L1, RTE, and GilM elements form the most prominent families), endogenous-retrovirus-like sequences, and DNA transposons (Novosib-like). The NC64A telomeric repeat unit is identical to that of flowering plants [i.e., $(TTTAGGG)_n$]. Eighteen scaffolds exhibit telomeric repeat arrays at a terminus and represent ends of chromosomes (Figure 1).

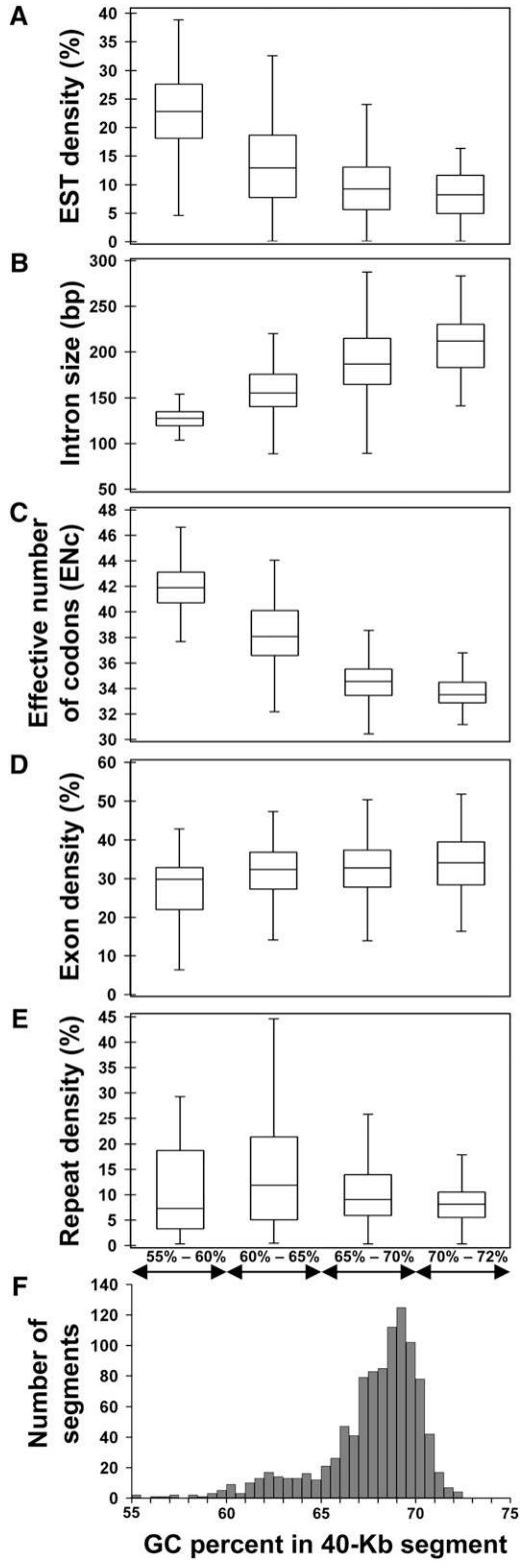


Figure 2. Features of Low-GC Regions in *Chlorella* sp NC64A.

Nonoverlapping 40-kb segments of the NC64A genome assembly were classified into four GC content classes. The distributions of genomic segments in each of the GC content classes are depicted by box plots for the following features: EST density (as defined in the Figure 1 legend) **(A)**, average size of introns supported by EST data **(B)**, mean effective number of codons (ENc) per gene **(C)**, exon density **(D)**, and repeat density **(E)**. **(F)** shows the distribution of genomic segments as a function of their GC content. The bottom and top of boxes represent the 1st and 3rd quartiles, Q1 and Q3, respectively, and the band near the middle of boxes represents the median. The extremities of the lines appearing below and above the boxes represent the lowest value still within 1.5 IQR (interquartile range = Q3 to Q1) of the lower quartile Q1, and the highest value still within 1.5 IQR of the upper quartile Q3. We applied the Kruskal-Wallis statistical test to each genomic feature to test the null hypothesis of equivalence between the distributions of values in the four GC bins. Distributions of EST density, intron size, and Enc were found to be significantly different between the four GC bins ($P < 0.0001$), whereas for repeat density, the difference was only marginally significant ($P = 0.024$). The null hypothesis of equivalence of distributions could not be rejected at $\alpha = 0.05$ for exon density ($P = 0.468$).

Algae- and Land Plant–Specific Protein Families

We predicted and annotated 9791 protein genes in the NC64A genome, a number comparable to that of the *Micromonas* species (Table 1). Like *Chlamydomonas*, the NC64A protein genes are intron rich with 7.3 exons per gene on average, but the average NC64A intron length is shorter than in *Chlamydomonas*. An overview of the NC64A gene repertoire is provided in Supplemental Results, Supplemental Table 1, and Supplemental Figures 4 to 6 online. Comparison of the numbers of PFAM protein domains revealed 27 protein families that are present in all completely sequenced chlorophyte algae (NC64A, *C. reinhardtii*, *Micromonas* sp RCC299 and CCMP1545, *Ostreococcus lucimarinus*, and *Ostreococcus tauri*) but absent in three representative and completely sequenced land plants (*Physcomitrella patens*, *Arabidopsis*, and *Oryza sativa*) (see Supplemental Table 4 online). Most of these algal genes probably existed in the last common ancestor shared with terrestrial plants since all of them have homologs in other eukaryotes. This would imply that they were subsequently lost in the branch leading to land plants. Many of these protein families are involved in basic metabolism, such as respiration (cytochrome c/c1 heme lyase), amino acid synthesis (asparaginase), carbohydrate metabolism (including ACN9 protein and iron-containing alcohol dehydrogenase), protein synthesis (including the selenocystein aminotransferase and posttranslational modification enzyme PAM), and DNA or RNA metabolism (DNA binding protein HU and helicase family) (see Supplemental Table 4 online). The six chlorophytes contain a 3′5′-cyclic nucleotide phosphodiesterase gene that modulates the levels of the secondary messenger 3′:5′-cyclic nucleotides in signal transduction pathways (Beavo, 1995). Chlorophyte-specific protein families also included the formate/nitrite transporter, type I polyketide synthase, and pyruvate decarboxylase (fatty acid metabolism).

By contrast, 184 protein domain families were present in all three land plants but absent in chlorophytes, including NC64A (see Supplemental Data Set 1 online); 102 of them have homologs in eukaryotes (excluding viridiplantae) and may have existed in the common ancestor with green algae and subsequently been

lost in the Chlorophyta lineage. Furthermore, 12 protein domain families are exclusively found in land plants and bacteria or archea. The corresponding genes may have been exchanged by lateral gene transfer between the nuclear genome of land plants and the genomes of prokaryotes or organelles. The remaining 70 protein domains have no recognizable homologs outside of land plants. Many of the 184 land plant protein domain families are involved in development, cell signaling, stress and hormonal response, transcriptional regulation, defense, and polysaccharide and cell wall metabolism (see Supplemental Data Set 1 online). Thus, in addition to the higher number of gene duplications that are characteristic of land plants (Flagel and Wendel, 2009), some of these proteins were probably important in the rise of multicellularity and terrestrial colonization in the Streptophyte lineage. For example, land plant–specific protein families involved in auxin signaling have presumably played a significant role in the emergence of organs, establishment of a complex developmental program, and adaptation to changing environment (Galván-Ampudia and Offringa, 2007). They include the auxin/indole-3-acetic acid transcriptional regulator family, auxin response factor transcription factor family and dormancy-associated and protein products of the ARG7 auxin-responsive gene family. We also found protein families involved in resistance to drought (e.g., dehydrin and Di19 proteins) that are specific to land plants; these families have perhaps been important in the adaptation to water-limiting conditions during the colonization of land (Bateman et al., 1998). Unlike chlorophyte algae, terrestrial plants have proteins involved in polysaccharide metabolism, lignin metabolism (e.g., Phe ammonia lyase and caffeic acid 3-*O*-methyltransferase) and cell wall metabolism (e.g., pectate lyase and pectinesterase), some of which probably contributed in the stiffening and consolidation of cell walls to withstand the weight of land plants subjected to gravity.

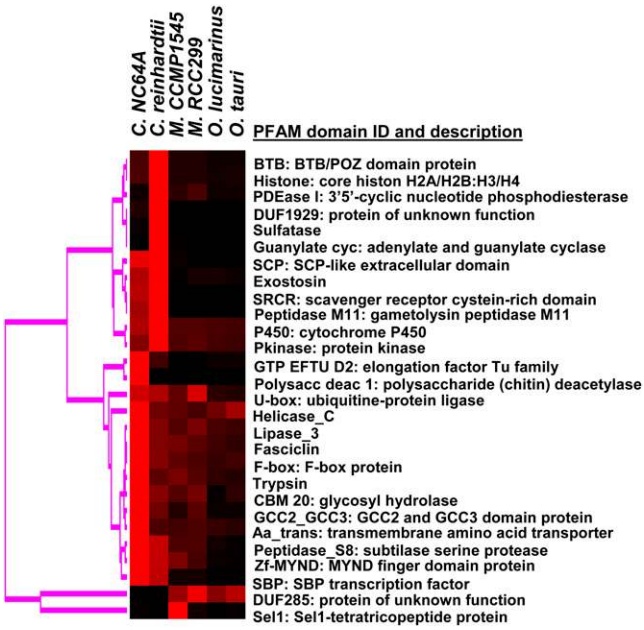


Figure 3. Heat Map of PFAM Protein Families with Significantly Biased Distribution among Chlorophyte Algae.

PFAM protein families that have either significantly expanded or shrunk in one or more sequenced chlorophytes (χ^2 test, $\alpha = 0.05$ after Bonferroni correction). Full red and black indicate 100 and 0%, respectively, of the total number of proteins in the PFAM family for the six algae. Real counts and description of PFAM protein families are given in Supplemental Table 5 online. The leftmost graph represents the hierarchical clustering of the PFAM domains by the average linkage methods using correlation coefficients between profiles.

Dynamics of *Chlorella* Protein Families

Twenty-eight PFAM protein families showed a biased distribution of proteins among the six chlorophyte algae (Figure 3; see Supplemental Table 5 online). Some PFAM domains were specifically overrepresented in NC64A compared with the five other chlorophytes. A subset of those PFAM domains was also found in excess in organisms that have intracellular or symbiotic life styles. We therefore hypothesize that the corresponding proteins in NC64A could also play a role in the mutualistic symbiosis with the protozoan *P. bursaria.* These PFAM domains include several families of proteins containing protein–protein interaction domains (F-box and MYND) and adhesion domains (fasciclin). Although the functions of domains may differ, proteins containing protein–protein interaction domains generally exist in excess in intracellular bacteria and symbiotic eukaryotes compared with their free-living relatives. For example, in intracellular bacteria, ankyrin proteins and tetratricopeptide repeat proteins are implicated in host–pathogen interactions (Petri et al., 2000), linked to the cytoplasmic incompatibility phenotype of the eukaryotic host (Tram et al., 2003; Iturbe-Ormaetxe et al., 2005) and directly secreted into the host (Wu et al., 2004). Protein families that contain ankyrin and WD40 domains are also prominent in the plant symbiont *Laccaria bicolor* (Martin et al., 2008), although there is no direct evidence that these proteins are involved in symbiosis.

NC64A also has an excess of proteins with Cys-rich GCC2_GCC3 PFAM motifs (Figure 3; see Supplemental Table 5 online), which are found in a wide variety of extracellular proteins. The symbiont *L. bicolor* secretes Cys-rich proteins (albeit not of the GCC2_GCC3 type) into their host, some of which are upregulated in symbiotic tissues and implicated in the establishment of symbiosis (Martin et al., 2008).

We found a significant increase in the number of amino acid transporters (Aa_trans domain) in NC64A (35 proteins). Fourteen of them have ESTs, indicating they are expressed. Some of these transporters may be expressed when in a symbiotic environment (note: the ESTs in this study were from NC64A cells not engaged in symbiosis). This observation is consistent with previous studies, which suggest that *Chlorella* symbionts, including NC64A, possess an efficient system for importing amino acids from the *P. bursaria* host and can use amino acids as a source of nitrogen instead of nitrate (Kato et al., 2006). As a complement to amino acid transporters, NC64A contains many trypsin-like proteases that may be involved in degrading peptides into amino acids.

We also found an increased number of proteins with a class 3 lipase signature (Lipase_3 domain) (Figure 3; see Supplemental Table 5 online). A previous study reported that algal symbionts,

such as zooxanthellae, translocate photosynthetic carbon into their animal hosts in the form of intact lipids, glycerol, and fatty acids (Battey and Patton, 1984); this process is mobilized by lipases. The hypothesis that NC64A feeds its *P. bursaria* host by translocating lipid molecules remains to be confirmed experimentally.

Protein families, such as protein kinases (Pkinase domain), guanylate and adenylate cyclases (Guanylate_cyc), and 3′,5′-cyclic nucleotide phosphodiesterases (PDEase_I domain), are more prevalent in *Chlamydomonas* compared with *Chlorella* and the mamiellalean algae, suggesting that cellular signaling is more complex in *Chlamydomonas*. *Chlamydomonas* also has a complex set of arylsulfatase-like proteins (sulfatase domain), some of which are secreted in response to sulfur deprivation (Pollock et al., 2005); by contrast, NC64A has only one homolog, and the four Mamiellale species have none, indicating that they adapt differently to low sulfur environments.

Evidence of Sexual Reproduction in *Chlorella*

Although *Chlorella* species have long been assumed to be asexual, NC64A encodes all of the known meiosis-specific proteins inventoried by Schurko and Logsdon (2008) and Malik et al. (2008), namely, dosage suppressor of MCk1 DMC1, homologous-pairing proteins HOP1 and HOP2, meiotic recombination protein MER3, meiotic nuclear division protein MND1, and mutS homolog protein MSH4. These genes are also found in most of the other sequenced chlorophyte algal species (see Supplemental Table 6 online). *Chlorella* species, including symbionts of *P. bursaria*, have been observed only in the haploid phase (Pickett-Heaps, 1975; Gerashchenko et al., 2001; Kadono et al., 2004), but the presence of meiosis genes suggests that NC64A also has a diploid phase and that its sexual reproductive cycle might have been overlooked, like the cryptic sex recently identified in *Ostreococcus* species (Grimsley et al., 2010). In addition, we found 19 NC64A homologs of the *Chlamydomonas* gametolysin proteins that promote the disassembly of the gametic cell walls and allow gamete fusion as well as an NC64A ortholog (id:137637) to the *Chlamydomonas* GCS1 protein essential for cell fusion (Goodenough et al., 2007). These results suggest that meiosis and sexual reproduction are part of the NC64A life cycle.

Conserved Flagella Proteins in *Chlorella*

NC64A, *O. lucimarinus*, and *O. tauri* are thought to be nonmotile green algae because flagella have never been observed in these organisms. Conversely, *Micromonas* sp CCMP1545, *Micromonas pusilla* RCC299, and *C. reinhardtii* are motile green algae: both *Micromonas* species have one flagellum and *Chlamydomonas* has two flagella. A proteomic study (Pazour et al., 2005) identified 360 *Chlamydomonas* flagellar proteins with high confidence. We used the reciprocal best BLASTP hit (RBH) criterion between the *Chlamydomonas* proteome and that of other sequenced chlorophytes to identify orthologs to the *Chlamydomonas* flagella proteins. Unexpectedly, we identified many putative (RBH) orthologs to the *Chlamydomonas* flagellar proteins in the NC64A genome (103 out of 360 *Chlamydomonas* flagella proteins = 29%; see Supplemental Data Set 2 online), of which 48% (50/103) and 53% (55/103) were also found in *O. tauri* and *O. lucimarinus*, respectively, while 85% (88/103) were also found in both *Micromonas* sp CCMP1545 and *M. pusilla* RCC299.

Proteins normally involved in the axonemal outer dynein arm, a structure responsible for movement of flagella, are included among putative *Chlorella* orthologs (Figure 4; see Supplemental Data Set 2 online): outer dynein arm docking complex proteins (ODA-DC proteins; numbers of proteins in NC64A/*Chlamydomonas* = 3/3), outer dynein arm heavy chain proteins (ODA-DHC: 3/3), outer dynein arm intermediate chain proteins (OAD-IC: 2/2), and outer dynein arm light chain proteins (ODA-LC: 4/5). Putative orthologs were identified in *Micromonas* for each of these proteins (except ODA-DC1), but none were found in the

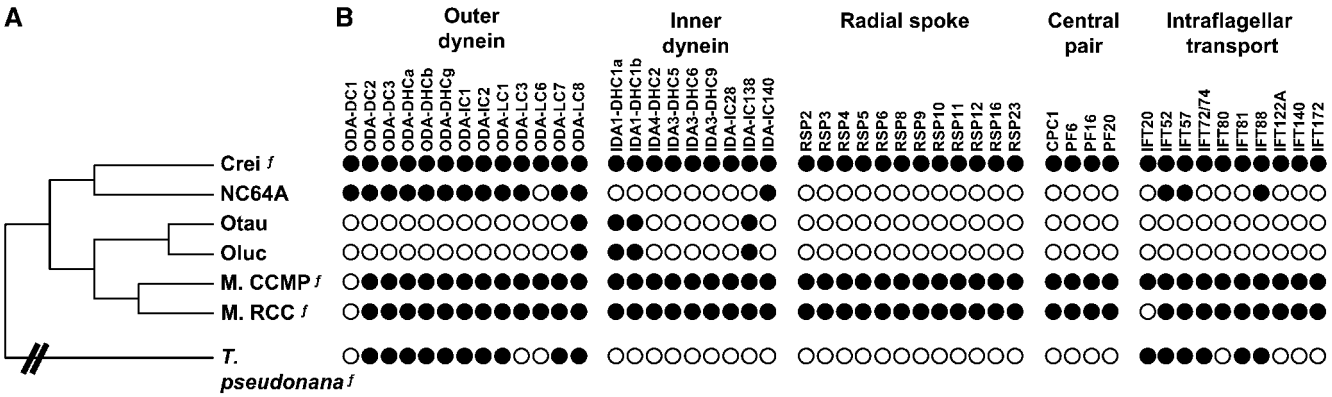


Figure 4. Distribution of Selected Flagellar Proteins across Chlorophytes.

(A) Cladogram showing the likely evolutionary relationships of sequenced chlorophytes and *T. pseudonana* based on the 18S phylogenetic tree shown in Supplemental Figure 1 online. The *ƒ* mark shows organisms known to build motile flagella. Crei, *C. reinhardtii*; NC64A, *Chlorella* sp NC64A; Otau, *O. tauri*; Oluc, *O. lucimarinus*; M. CCMP, *M. pusilla* CCMP1545; M. RCC, *Micromonas* sp RC299.

(B) Presence (dot) or absence (circle) of putative (Reciprocal Best Hit) orthologs to *Chlamydomonas* outer dynein proteins, inner dynein proteins, radial spoke proteins, central pair proteins, and IFT proteins.

Ostreococcus species (except ODA-LC8). In *Chlamydomonas*, the assembly and maintenance of flagella depend on a process called intraflagellar transport (IFT) (Cole, 2003). The IFT system consists of a motor complex associated with groups of large protein complexes called IFT particles. *Chlorella* encodes putative orthologs to the proteins ITF52, ITF57, and ITF88 involved in the IFT particle (Figure 4) as well as putative orthologs to the kinesin-2 motor protein FLA8 (Joint Genome Initiative [JGI] 37158) and the kinesin-associated protein KAP (JGI 139946). Surprisingly, two of the proteins identified in *Chlorella*, namely, ITF57 and ITF88, were until now exclusively found in organisms that have flagella (Wickstead and Gull, 2007) except *Plasmodium falciparum* that is known to build its flagella throughout an IFT-independent mechanism.

C. reinhardii has 249 flagellar proteins that exhibit no RBH with NC64A. *Micromonas* spp retained many of them (101/249 [41%] RBHs in both *Micromonas* species). By contrast, only 18/249 (7%) and 19/249 (8%) putative orthologs were identified in *O. tauri* and *O. lucimarinus*, respectively. Overall, 89 proteins were present in all motile sequenced chlorophyte algae but absent in NC64A and the *Ostreococcus* species. This flagella-specific set includes most proteins known to function in inner-arm dynein complexes (including inner-arm dyneins), the central pair complex, the IFT particle, and all proteins of the *Chlamydomonas* radial spoke (Figure 4; see Supplemental Data Set 2 online).

The conservation of a substantial subset of the *C. reinhardtii* flagella proteins in NC64A is intriguing. In particular, our results suggest that NC64A has retained an almost complete set of outer-arm dynein proteins (heavy, intermediate, and light chains and docking complex) that are found only in eukaryotes that exhibit motile cilia/flagella at some point in their life cycle (Wickstead and Gull, 2007). Merchant et al. (2007) identified 195 *C. reinhardtii* proteins that have homologs in two motile ciliates (*Homo sapiens* and *Phytophthora* spp) but not in a group of reference aciliates (*Arabidopsis*, *Neurospora*, *Cyanidioschyzon*, *Dictyostelium*, eubacteria, and archaea). This protein set, designated the CiliaCut, is thought to contain proteins involved in flagellar function. In agreement with the results obtained above, 63 proteins of the CiliaCut (63/195 = 32%) had putative orthologs (RBH) in NC64A (see Supplemental Figure 7 online). Merchant et al. (2007) further subdivided the CiliaCut on the basis of whether or not a homolog was present in *Caenorhabditis elegans*, which has only nonmotile sensory cilia, and *Thalassiosira pseudonana*, which builds unusual motile flagella during gametogenesis. The 62 CiliaCut proteins with homologs in *C. elegans* were predicted to have structural, sensory, or assembly roles and designated the SSA, whereas the 69 CiliaCut proteins with homologs in *T. pseudonana* were designated the CentricCut. Interestingly, two-thirds of the CiliaCut proteins with putative orthologs in NC64A (42/63 = 67%) were classified in the CentricCut. This distribution was found to be significantly nonrandom (P value < 2E-7; χ^2 test). By contrast, we found no significant association of the NC64A orthologs with the SSA subset. Thus, the pattern of conservation of putative flagellar proteins in NC64A is most similar to that of *T. pseudonana*, which like NC64A, lacks the genes encoding the radial spoke, central pair, and inner dynein proteins (Figure 4) (Merchant et al., 2007; Wickstead and Gull, 2007).

Altogether, these results lead to two hypotheses that should be verified experimentally: (1) the conserved flagella proteins might have acquired other biological roles when the flagellar apparatus was lost, which allowed the corresponding genes (i.e., encoding the retained flagella proteins) to remain under selective pressure; (2) given that NC64A is probably capable of sexual reproduction as suggested above, we speculate that *Chlorella* retained the ability to form rudimentary, possibly motile, flagella or flagellum-derived structures, similar to those of *T. pseudonana*. If true, we hypothesize that this inferred structure might serve in the recognition of the mating partner and initiate cell fusion, producing an as yet unidentified zygote.

Phytohormones in Algae

Phytohormones regulate much of the growth and development in land plants, and they are involved in the plant's response to infection. Most types of land plant hormones have been biochemically detected in green algae, including chlorophytes (Tarakhovskaya et al., 2007). Some of those hormones appear to play the same roles as in land plants (e.g., cytokinin [Stirk et al., 2002] and auxin [de-Bashan et al., 2008]), but little is known about algal hormone biosynthesis (Bajguz, 2009). Hormone biosynthetic pathways in land plants are associated with plastids. Since chlorophyte algae contain plastids, we anticipated finding orthologs to the enzymes that synthesize hormones in land plants, as well as to their hormone receptors. We did not attempt to compile an exhaustive search of all chlorophyte hormone pathway steps or their receptors. Extensive gene duplication in the *Arabidopsis* genome used as reference prevented us from identifying clear algal orthologs of some enzymes involved in hormone synthesis. Instead, we looked for the presence of one or more clear orthologous enzymes for some key steps in plant hormone pathways and receptors. Orthology assignation was performed by combining information from reciprocal best hit analysis, phylogenetic tree reconstruction, and protein domain organization (see Supplemental Results online).

We explored the NC64A genome as well as five other chlorophyte genomes and found probable orthologs to *Arabidopsis* enzymes involved in the synthesis of a variety of plant hormones, including abscisic acid, cytokinin, brassinosteroid, and polyamines (Table 2; see Supplemental Results online). The sequenced chlorophyte algae did not exhibit homologs (BLASTP and TBLASTN analyses E-value cutoff = 1e-5) to *Arabidopsis* enzymes involved in the gibberellin biosynthetic pathway (gibberellin biosynthetic proteins GA1, GA2, and GA3; gibberellin oxidase proteins GA20OX1, GA2OX1, and GA3OX1) or the ethylene biosynthetic pathway (1-aminocyclopropane-1-carboxylate synthase and 1-aminocyclopropane-1-carboxylate oxidase [ACO]). We did find putative orthologs to some of the known *Arabidopsis* hormone receptors, including those for abscisic acid (chelatase H subunit [CHLH]), auxin (Auxin Binding Protein1 [ABP1]), and cytokinin (high osmolarity glycerol protein [HOG]) (Table 2). A recent survey of genomic data also reported the existence of orthologs of some of the components of the auxin signaling systems, including ABP1, in chlorophyte algae (Lau et al., 2009). In *Arabidopsis*, the auxin signaling cascade alternative to ABP1 involves the TIR1/AFB family of F-box proteins, auxin response factor, and

Table 2. Accession Numbers of Putative Chlorophyte Orthologs to *Arabidopsis* Proteins Involved in Phytohormone Biosynthesis or Reception

Arabidopsis Enzyme Name[a]	*Chlorella* sp NC64A[b]	*C. reinhardtii*	*O. tauri*	*O. lucimarinus*	*Micromonas* sp RC299	*Micromonas* sp CCMP1545
Abscisic acid pathway						
Abscisic-aldehyde oxidase (AAO3) NP_180283	58208 (30%)					
9-*cis*-epoxycarotenoid dioxygenase (NCED5) NP_174302	138368 (37%)	XP_ 001695565 (32%)				
Zeaxanthin epoxidase (ABA1) NP_201504	138731 (57%)	XP_ 001701701 (58%)	CAL 58065 (42%)	XP_ 001421564 (41%)	ACO 64017 (44%)	EEH 54518 (43%)
Violaxanthin deepoxidase (NPQ1) NP_172331	35609 (42%)		CAL 58064 (46%)	XP_ 001421704 (41%)	ACO 63977 (41%)	EEH 54773 (40%)
Abscisic acid receptor (CHLH) NP_001078578	143922 (50%)	XP_ 001700895 (66%)	CAL 51621 (68%)	XP_ 001417229 (68%)	ACO 63109 (68%)	EEH 57631 (67%)
Cytokinin pathway						
Isopentenyl-transferase 9 (ATIPT9) NP_851043	55198 (36%)		CAL 53743 (35%)	XP_ 001418572 (36%)	ACO61527 (42%)	EEH 53639 (41%)
Cytokinin receptor (HOG) BAH19670	37522 (81%)	XP_ 001693339 (77%)	CAL 55423 (75%)	XP_ 001419579 (74%)	ACO67241 (73%)	EEH 58817 (69%)
Brassinosteroid pathway						
lathosterol oxidase (STE1) NP_186907	37407 (45%)	XP_ 001701457 (50%)				
7-Hydrocholesterol reductase (DWF5) NP_001077693					ACO 69953 (51%)	EEH 53090 (51%)
Steroid reductase (DET2) NP_181340	18410 (37%)	XP_ 001696975 (34%)	CAL 52707 (32%)	XP_ 001416556 (33%)	ACO 66602 (34%)	EEH 52455 (34%)
Jasmonic acid pathway						
12-Oxophytodienoate reductase (OPR1) NP_177794	52565 (48%)	XP_ 001699402 (51%)				
3-Hydroxyacyl-CoA dehydrogenase/enoyl-CoA hydratase (MFP2) NP_187342	52565 (54%)	XP_ 001696661 (45%)	CAL 53100 (44%)	XP_ 001417042 (52%)	ACO 65308 (52%)	EEH 60148 (51%)
Polyamine (spermidine) pathway						
Arg decarboxylase (ADC1) NP_179243	25497 (40%)					EEH 59440 (38%)
Agmatine iminohydrolase (ATAIH) NP_196434	133066 (54%)					
N-Carbamoyl-putrescine amidohydrolase (NLP1) NP_850101	18182 (57%)	XP_ 001692986 (53%) XP_ 001690094 (53%)				
Spermidine synthase 2 (SPDS2) NP_177188	26108 (53%)	XP_ 001702843	ABO 98745 (56%)	XP_ 001420452 (58%)	ACO 70332 (55%)	EEH 54321 (56%)
Orn decarboxylase NP_001063827[c]	133981 (50%)	XP_ 001698872 (46%)	CAL 51811 (45%)	XP_ 001417323 (45%)	ACO 63617 (46%)	EEH 58717 (46%)
Auxin pathway						
Auxin receptor ABP1 NP_192207	17596 (48%), 26559 (40%)					

The percentages of sequence identity in the best high-scoring pair (BLASTP) between proteins and their putative orthologous *Arabidopsis* protein are shown in parentheses.

[a]*Arabidopsis* accession number of protein used as query in BLAST search.

[b]At the JGI portal site (http://genomeportal.jgi-psf.org/), select *Chlorella* NC64A.

[c]This enzyme is not found in *Arabidopsis*; accession number is for *O. sativa*.

auxin/indole-3-acetic acid proteins (Lau et al., 2008). None of these proteins were found to have a significant match with the sequenced chlorophytes, suggesting that this signaling cascade is absent in these organisms (Lau et al., 2009). By contrast, all major components of this pathway were identified in the moss *P. patens*, which implies that their origin goes back to at least the early evolution of land plants (Rensing et al., 2008).

The presence of putative chlorophyte orthologs to *Arabidopsis* proteins involved in phytohormone biosynthesis and perception does not necessarily imply that these green algae can produce, sense, and respond to hormones through pathways analogous to those in land plants. To our knowledge, some of the identified enzymes have no other role than hormone biosynthesis in land plants (e.g., ATP/ADP isopentenyltransferase AtIPT, Sterol 1 protein STE1, and DWARF5), while others are also involved in the production of molecules with no hormonal function (e.g., abscisic acid 1 protein [ABA1] and nonphotochemical quenching protein [NPQ1] involved both in the xanthophyll cycle and in the synthesis of ABA precursors). However, the presence of putative orthologs to the *Arabidopsis* auxin receptor ABP1 in NC64A is congruent with earlier studies demonstrating that auxin induces cell division in *Chlorella pyrenoidosa* (Vance, 1987) and cell enlargement in *Chlorella vulgaris* (Yin, 1937), two species closely related to NC64A (see Supplemental Figure 1 online). Our

analysis suggests that at least some of the genes specifically involved in phytohormone biosynthesis and perception in land plants were established prior to their evolution. Unicellular ancestors of streptophytes and chlorophytes were perhaps able to communicate with each other before the emergence of multicellular land plants. We suggest that the existence of these features likely facilitated the evolution of multicellularity.

Cell Wall Metabolism and Interplay with *Chlorella* Viruses

With 233 predicted enzymes involved in carbohydrate metabolism, NC64A appears much better equipped for synthesizing and modifying polysaccharides than the other sequenced chlorophytes that have between 92 (*O. tauri*) and 168 (*C. reinhardtii*) of such predicted enzymes (see Supplemental Data Set 3 online) (Cantarel et al., 2009). However, we did not find homologs of the *Arabidopsis* proteins involved in the synthesis of cellulose (cellulose synthase CesA) or hemicellulose (hemicellulose synthase CLS), the major components of the primary cell wall of land plants. Instead, experimental evidence suggests that the cell wall of *Chlorella* species, including NC64A, contain glucosamine polymers such as chitin and chitosan (Kapaun and Reisser, 1995; Sun et al., 1999). We found two NC64A paralogs for chitin synthase and, remarkably, 25 paralogs for chitin deacetylase, which converts chitin into chitosan. Both NC64A chitin syntase proteins contain conserved amino acids essential for the catalytic activity of the *Saccharomyces cerevisiae* enzyme (i.e., Asp-441, Asp-562, Gln-601, Arg-604, Trp-605, Asn-797, Asp-800, Trp-803, and Thr-805; Yabe et al., 1998) (see Supplemental Figure 8A online). We also identified putative proteins involved in the degradation of these polysaccharides: two chitinase genes (plant and prokaryotic types [glycosyl hydrolase families GH19 and GH18, respectively]) and four chitosanase genes. The prokaryotic type chitinase protein exhibits protein domains that are homologous to the PF-ChiA chitinase and cellulose binding domains found in the chitinase of archaeon *Pyrococcus furiosus*. It also exhibits the conserved amino acid sequence (DXDXE motif) that plays an important role in the catalytic mechanism of family 18 chitinases (Watanabe et al., 1994) (see Supplemental Figure 8B online). The four NC64A chitosanases contain the three catalytic residues Glu-36, Asp-40, and Thr-45 of *Streptomyces* sp N174 chitosanase (Lacombe-Harvey et al., 2009) (see Supplemental Figure 8C online).

Chitin is a natural component of fungal cell walls and of the exoskeleton of arthropods but is not normally present in green algae. The origin of chitin and its derivatives in the *Chlorella* genus has long been an enigma. Except for the plant-type chitinase gene, which is found in land plants (but not in chlorophytes apart from *Chlorella*), the four gene classes involved in forming and remodeling chitin cell walls (i.e., chitin synthase, chitin deacetylase, chitinase, and chitosanase) are absent in all the other fully sequenced Viridiplantae species. By contrast, homologs for each of these families exist in genomes of *Chlorella* viruses. The viral genes are presumably involved in degradation of the *Chlorella* cell wall (chitinase and chitosanase) (Kang et al., 2005) and production of chitinous fibers on the external surface of virus-infected cells (chitin synthase and chitin deacetylase) (Kawasaki et al., 2002). Phylogenetic analysis suggests that the *Chlorella* ancestor exchanged the bacterial-type chitinase and chitin-deacetylase genes with the chloroviruses (Figure 5). The fact that these genes are absent in the other Viridiplantae species studied to date argues in favor of the capture of the viral genes by *Chlorella*. Alternatively, capture of the *Chlorella* genes by chloroviruses would imply that *Chlorella* genes were vertically inherited from the Viridiplantae ancestor and that these genes were independently lost in many lineages of the Viridiplantae, a very improbable scenario. Another scenario would imply a first

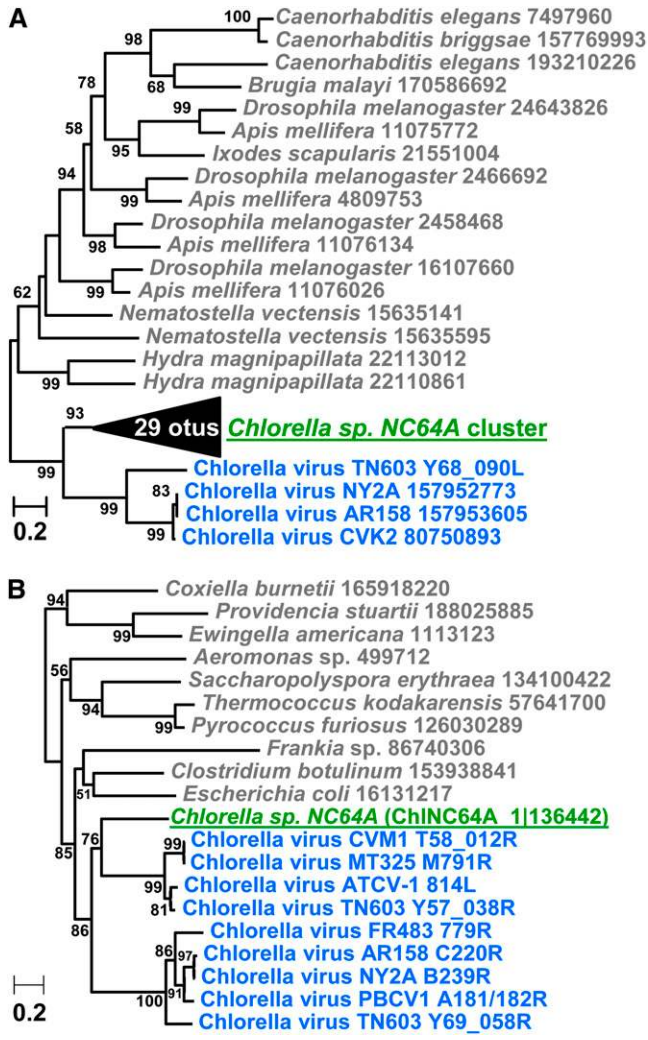


Figure 5. Maximum Likelihood Phylogenetic Tree of the Chitin Deacetylase and Chitinase Proteins.

For both protein families, we used the WAG+I+G model of substitutions. Approximate likelihood ratio test values >50% are indicated beside branches. Phylogenetic trees are midpoint rooted. Alignments used to generate these trees are available as Supplemental Data Sets 4 and 5 online.

(A) Phylogenetic tree of chitin deacetylases. The multiple sequence alignment contained 134 gap-free sites.

(B) Phylogenetic tree of chitinases. The multiple sequence alignment contained 228 gap-free sites.

[See online article for color version of this figure.]

capture of the genes by HGT from prokaryotes or fungi to *Chlorella*, after which a *Chlorella* virus picked up the two genes from *Chlorella*. Phylogenetic reconstructions of the chitosanase and chitin synthase proteins indicate that the corresponding *Chlorella* and *Chlorella* virus genes are phylogenetically related, but no direct gene exchange occurred between *Chlorella* and the known *Chlorella* viruses (see Supplemental Figures 9 and 10 online). Collectively, our results are congruent with the hypothesis that components of the *Chlorella* chitin metabolism were acquired horizontally from viruses or distantly related chitin-producing cellular organisms rather than from a Viridiplantae ancestor.

Conclusion

The first sequence of a trebouxiophycean genome unveiled important features of the evolution and genomic organization of the green phylum. For instance, the existence of genomic regions displaying large differences in GC content, correlating with differences in their expression levels, now appear to be a characteristic feature of many chlorophyte genomes. Understanding the role and mechanism by which this compositional shift is established and maintained is one of the next challenges in phycology.

We presented evidence suggesting that *Chlorella* could have acquired components of its chitin biosynthetic pathway by HGT from a chlorovirus or a microorganism. A similar evolutionary scenario was also evoked for the eukaryotic microalga *Emiliania huxleyi* that exchanged seven genes of the sphingolipid biosynthesis pathway with its large DNA virus, EhV (Monier et al., 2009), though the direction of gene transfer is unknown. Thus, the large DNA viruses predominantly associated with microalgae and marine protists might have played a much larger role in the evolution of their hosts than previously recognized. Conversely to the traditional view of viruses as gene pickpockets, large DNA viruses might have a propensity to enhance the metabolic capabilities of their host by donating genes (Villarreal, 2004). In the case of *Chlorella*, the acquisition of a chitinous cell wall may have conferred a protective barrier against other viral and bacterial parasites lacking the chitinase/chitosanase enzymes required to penetrate and/or escape the algal cell. This might have increased the fitness of *Chlorella* compared with its ancestors unable to synthesize chitin. This HGT might be the key event that promoted the radiation and success of the *Chlorella* genus (i.e., *Chlorella* may have achieved a cosmopolitan distribution because most of its previous parasites failed to penetrate its newly acquired chitinous cell wall).

Our results illustrate the role that comparative genomics can play in uncovering unsuspected biological functions; here, the identification of genes involved in meiosis, gamete fusion, and flagella. This led us to hypothesize that *Chlorella* retained the capability of sexual reproduction despite the fact that no sexual life cycle has been described in this genus. These findings naturally pose the question of the maintenance of sexual reproduction in an organism capable of rapid clonal population growth. In *C. reinhardtii*, mating between two haploid partners is induced by stress conditions (e.g., lack of nitrogen), producing a zygote resistant to freezing and desiccation (Goodenough et al., 2007). There is some recent evidence that viruses may have played a role in the success of sexual reproduction. Sexual reproduction can confer a selective advantage to the host in the arms race against its parasites (the so-called Red Queen hypothesis) by increasing the efficiency with which selection can fix beneficial mutations that result in virus resistance (Morin, 2008). A more direct viral pressure is illustrated by the haptophyte microalga *E. huxleyi* escaping infection by the phycodnavirus EhV by switching from its virus-sensitive diploid stage to a morphologically distinct haploid stage immune to the virus (the Cheshire Cat escape strategy) (Frada et al., 2008). *Ostreococcus* and *Chlorella* species are normally haploid but contain meiosis-related genes. They are both infected by phycodnaviruses (OsV and *Chlorella* viruses, respectively) that are phylogenetically related to EhV (Wu et al., 2009). By analogy to the EhV–*E. huxleyi* model, it is tempting to speculate that these microalgae have a virus-resistant diploid phase that might only become detectable after viruses have decimated the haploid population.

The presence of putative chlorophyte orthologs for land plant proteins functioning in critical hormone metabolic steps and as hormone receptors opens the possibility that phytohormone biosynthesis and perception could also be present in chlorophyte algae, although perhaps in a rudimentary form compared with land plants. Consequently, it has been suggested that green algae would be a model organism for the study of plant hormones (and receptors) because they are unicellular and can be grown axenically in the laboratory (Stirk et al., 2002). A fuller understanding of the role of plant hormone molecules in green algae as well as of their synthesis and perception would possibly lead to the selection and improvement of better algal strains that could benefit agricultural practices in developing countries (Stirk et al., 2002), result in better production of biodiesel, and improve the quality and quantity of nutrient supplements (proteins, vitamins, etc.). While bioinformatics/genomics can provide strong clues, enzyme and receptor functions remain to be experimentally tested to verify these many predictions.

METHODS

A detailed description of methods is provided in Supplemental Methods online.

Genome Sequencing and Assembly

The NC64A genome was sequenced using the whole-genome sequencing strategy. The data were assembled using release 2.10.11 of Jazz, a whole-genome sequencing assembler developed at the JGI (Aparicio et al., 2002). After excluding redundant and short scaffolds from the initial assembly, there was 46.4 Mb of scaffold sequence, of which 4.0 Mb (8.5%) were gaps. The filtered assembly contained 431 scaffolds, with a scaffold N/L50 of 12/1.5 Mb (the number of scaffolds/length of the shortest scaffold, respectively, such that the sum of scaffolds of equal length or longer is at least 50% of the total length of all scaffolds), and a contig N/L50 of 441/27.6 kb. The sequence depth derived from the assembly was 8.95 $\pm$ 0.15.

Pulse Field Gel Electrophoresis

Pulse field gel electrophoresis studies were performed according to Agarkova et al. (2006). Chromosomal DNAs were separated in a

CHEF-DR II (Bio-Rad) unit in a 0.8% agarose gel. Electrophoresis conditions and running buffer were selected to resolve the target chromosome sizes. The exact conditions are described in the figure legends.

EST Sequencing and Assembly

Chlorella sp NC64A cells were grown to log phase (1.5×10^7 cells/mL). NC64A poly(A)$^+$ RNA was isolated from total RNA using the Absolutely mRNA Purification kit (Stratagene). One to two micrograms of poly(A)$^+$ RNA, reverse transcriptase SuperScript II (Invitrogen), and oligo(dT)-NotI primer were used to synthesize first-strand cDNA. Second-strand synthesis was performed with *Escherichia coli* DNA ligase, polymerase I, and RNaseH followed by end repair using T4 DNA polymerase. The cDNA inserts were directionally ligated into the *Sal*I- and *Not*I-digested vector pCMVsport6 (Invitrogen). Subcloned inserts were then sequenced with Big Dye terminator chemistry (Applied Biosystems). A total of 38,400 ESTs were generated. The ESTs were processed through the JGI EST pipeline. A total of 23,828 ESTs remained after trimming vector sequences and removing short sequences. EST clusters were assembled using CAP3 (Huang and Madan, 1999) to form consensus sequences. Clustering and assembly of all 23,828 ESTs resulted in 7499 consensus sequences.

Genome Annotation and Sequence Analysis

The genome assembly v1.0 of NC64A was annotated using the JGI annotation pipeline, which combines several gene predictors and filtering steps (see Supplemental Methods online). Phylogenetic analyses were performed on the phylogeny.fr web tool (Dereeper et al., 2008). De novo identification of repeated sequences was performed by aligning the genome against itself using the BLASTN program (E-value < 1e-15). Individual repeat elements were organized into families with the RECON program using default settings (Bao and Eddy, 2002). RECON constructed 2980 repetitive sequence families from 10,723 individual repeat elements. Second, identification of known repetitive sequences was performed by aligning the prototypic sequences contained in Repbase v12.10 (Jurka et al., 2005) using TBLASTX. The results of the two methods were combined.

Accession Numbers

Assembly and annotations of *Chlorella* sp NC64A are available from JGI Genome Portal at http://genome.jgi-psf.org/NC64A and can also be found in the GenBank/EMBL data libraries under accession number ADIC00000000.

Supplemental Data

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

Supplemental Figure 1. Phylogenetic Position of *Chlorella variabilis* NC64A among Chlorophyte Algae.

Supplemental Figure 2. Pulsed Field Gel Electrophoresis of *Chlorella* sp NC64A Chromosomes.

Supplemental Figure 3. Taxonomic Distribution of Best Matches for Proteins Encoded in Low-GC Regions.

Supplemental Figure 4. Gene Duplication in Selected Viridiplantae.

Supplemental Figure 5. Sequence Motifs at Intron Splice Sites.

Supplemental Figure 6. Taxonomic Distribution of Best Matches for Representative Chlorophyte Algae.

Supplemental Figure 7. NC64A Putative Orthologs to *Chlamydomonas* CiliaCut Proteins.

Supplemental Figure 8. Alignment of NC64A Proteins with Their Reference Proteins Involved in Chitin Metabolism.

Supplemental Figure 9. Maximum Likelihood Phylogenetic Tree of Chitosanase Proteins.

Supplemental Figure 10. Maximum Likelihood Phylogenetic Tree of Chitin Synthase Proteins.

Supplemental Table 1. NC64A Nuclear Genome Assembly Statistics.

Supplemental Table 2. Eukaryotic Ortholog Groups (KOG) Functional Categories among Low-GC and Normal-GC Regions.

Supplemental Table 3. Repeated Sequences in the NC64A Genome.

Supplemental Table 4. Chlorophyte Algae–Specific PFAM Protein Domains.

Supplemental Table 5. PFAM Domains with Biased Distribution in Chlorophyte Green Algae.

Supplemental Table 6. Meiosis-Specific Protein GenBank Identification (gi) Numbers and Percentage of Protein Sequence Identity with Reference *Arabidopsis* Proteins.

Supplemental Data Set 1. Land Plant–Specific PFAM Protein Domains.

Supplemental Data Set 2. Putative Orthologs to *Chlamydomonas* Flagellar Proteins in Sequenced Chlorophytes.

Supplemental Data Set 3. Carbohydrate-Active (CAZy) Enzymes in Chlorophyte Green Algae.

Supplemental Data Set 4. Multiple Sequence Alignment of Chitin Deacetylase Proteins Used in Figure 5A.

Supplemental Data Set 5. Multiple Sequence Alignment of Chitinase Proteins Used in Figure 5B.

Supplemental Data Set 6. Multiple Sequence Alignment of 18S Genes Used in Supplemental Figure 1.

Supplemental Data Set 7. Multiple Sequence Alignment of Chitosanase Proteins used in Supplemental Figure 9.

Supplemental Data Set 8. Multiple Sequence Alignment of Chitin Synthase Proteins Used in Supplemental Figure 10.

Supplemental Methods.

Supplemental Results.

Supplemental References.

ACKNOWLEDGMENTS

We thank Marek Elias for helpful discussion on flagella proteins. We also thank Magali Lescot, Dmitry Brogun, Timo Greiner, Ming Kang, Gentry L. Lewis, Suzanne Rose, Eliza Wiech, and Giane M. Yanai-Balser for their contribution in the annotation. Genome sequencing conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract DE-AC02-05CH11231. This work was partially supported by Marseille-Nice Génopole, the PACA-Bioinfo platform, the French fund "Infrastructures en Biologie Santé et Agronomie," and Public Health Service Grant GM32441 from the National Institute of General Medical Sciences (to J.L.V.E.). G.D. was partially funded by National Institutes of Health Grant P20 RR016469.

Received May 6, 2010; revised July 15, 2010; accepted September 1, 2010; published September 17, 2010.

REFERENCES

Agarkova, I.V., Dunigan, D.D., and Van Etten, J.L. (2006). Virion-associated restriction endonucleases of chloroviruses. J. Virol. **80:** 8114–8123.

Aparicio, S., et al. (2002). Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. Science **297:** 1301–1310.

Bajguz, A. (2009). Brassinosteroid enhanced the level of abscisic acid in *Chlorella vulgaris* subjected to short-term heat stress. J. Plant Physiol. **166:** 882–886.

Bao, Z., and Eddy, S.R. (2002). Automated de novo identification of repeat sequence families in sequenced genomes. Genome Res. **12:** 1269–1276.

Bateman, R.M., Crane, P.R., DiMichele, W.A., Kenrick, P.R., Rowe, N.P., Speck, T., and Stein, W.E. (1998). Early evolution of land plants: phylogeny, physiology, and ecology of the primary terrestrial radiation. Annu. Rev. Ecol. Syst. **29:** 263–292.

Battey, J.F., and Patton, J.S. (1984). A reevaluation of the role of glycerol in carbon translocation in zooxanthellae-coelenterate symbiosis. Mar. Biol. **79:** 27–38.

Beavo, J.A. (1995). Cyclic nucleotide phosphodiesterases: Functional implications of multiple isoforms. Physiol. Rev. **75:** 725–748.

Benson, A.A. (2002). Following the path of carbon in photosynthesis: A personal story. Photosynth. Res. **73:** 29–49.

Cantarel, B.L., Coutinho, P.M., Rancurel, C., Bernard, T., Lombard, V., and Henrissat, B. (2009). The Carbohydrate-Active EnZymes database (CAZy): An expert resource for Glycogenomics. Nucleic Acids Res. **37** (Database issue): D233–D238.

Caturegli, P., Asanovich, K.M., Walls, J.J., Bakken, J.S., Madigan, J.E., Popov, V.L., Dumler, J.S., and Dumler, J.S. (2000). ankA: An *Ehrlichia phagocytophila* group gene encoding a cytoplasmic protein antigen with ankyrin repeats. Infect. Immun. **68:** 5277–5283.

Chelf, P., Brown, L.M., and Wyman, C.E. (1993). Aquatic biomass resources and carbon dioxide trapping. Biomass Bioenergy **4:** 175–183.

Chuchird, N., Hiramatsu, S., Sugimoto, I., Fujie, M., Usami, S., and Yamada, T. (2001). Digestion of chlorella cells by chlorovirus-encoded polysaccharide degrading enzymes. Microbes Environ. **16:** 206–212.

Cole, D.G. (2003). The intraflagellar transport machinery of *Chlamydomonas reinhardtii*. Traffic **4:** 435–442.

de-Bashan, L.E., Antoun, H., and Bashan, Y. (2008). Involvement of indole-3-acetic acid produced by the growth-promoting bacterium Azospirillum spp. in promoting growth of *Chlorella vulgaris*. J. Phycol. **44:** 938–947.

Dereeper, A., Guignon, V., Blanc, G., Audic, S., Buffet, S., Chevenet, F., Dufayard, J.F., Guindon, S., Lefort, V., Lescot, M., Claverie, J.M., and Gascuel, O. (2008). Phylogeny.fr: Robust phylogenetic analysis for the non-specialist. Nucleic Acids Res. **36** (Web Server issue): W465–W469.

Flagel, L.E., and Wendel, J.F. (2009). Gene duplication and evolutionary novelty in plants. New Phytol. **183:** 557–564.

Fott, B., and Novakova, M. (1969). A monograph of the genus Chlorella. The fresh water species. Stud. Phycol. **11:** 1–73.

Frada, M., Probert, I., Allen, M.J., Wilson, W.H., and de Vargas, C. (2008). The "Cheshire Cat" escape strategy of the coccolithophore *Emiliania huxleyi* in response to viral infection. Proc. Natl. Acad. Sci. USA **105:** 15944–15949.

Friedl, T., and Bhattacharya, D. (2002). Origin and evolution of green lichen algae. In Symbiosis: Mechanisms and Model Systems, J. Seckbach, ed (Dordrecht, The Netherlands: Springer), pp. 341–357.

Galván-Ampudia, C.S., and Offringa, R. (2007). Plant evolution: AGC kinases tell the auxin tale. Trends Plant Sci. **12:** 541–547.

Gerashchenko, B.I., Kosaka, T., and Hosoya, H. (2001). Growth kinetics of algal populations exsymbiotic from *Paramecium bursaria* by flow cytometry measurements. Cytometry **44:** 257–263.

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

Grimsley, N., Péquin, B., Bachy, C., Moreau, H., and Piganeau, G. (2010). Cryptic sex in the smallest eukaryotic marine green alga. Mol. Biol. Evol. **27:** 47–54.

Grossman, A.R. (2005). Paths toward algal genomics. Plant Physiol. **137:** 410–427.

Heckman, D.S., Geiser, D.M., Eidell, B.R., Stauffer, R.L., Kardos, N. L., and Hedges, S.B. (2001). Molecular evidence for the early colonization of land by fungi and plants. Science **293:** 1129–1133.

Huang, X., and Madan, A. (1999). CAP3: A DNA sequence assembly program. Genome Res. **9:** 868–877.

Huss, V.A.R., Frank, C., Hartmann, E.C., Hirmer, M., Kloboucek, A., Seidel, B.M., Wenzeler, P., and Kessler, E. (1999). Biochemical taxonomy and molecular phylogeny of the genus Chlorella sensu lato (Chlorophyta). J. Phycol. **35:** 587–598.

Iturbe-Ormaetxe, I., Burke, G.R., Riegler, M., and O'Neill, S.L. (2005). Distribution, expression, and motif variability of ankyrin domain genes in *Wolbachia pipientis*. J. Bacteriol. **187:** 5136–5145.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. Cytogenet. Genome Res. **110:** 462–467.

Kadono, T., Kawano, T., Hosoya, H., and Kosaka, T. (2004). Flow cytometric studies of the host-regulated cell cycle in algae symbiotic with green paramecium. Protoplasma **223:** 133–141.

Kang, M., Dunigan, D.D., and VAN Etten, J.L. (2005). Chlorovirus: A genus of Phycodnaviridae that infects certain chlorella-like green algae. Mol. Plant Pathol. **6:** 213–224.

Kapaun, E., and Reisser, W. (1995). A chitin-like glycan in the cell wall of a Chlorella sp. (Chlorococcales, Chlorophyceae). Planta **197:** 577–582.

Karakashian, S.J., and Karakashian, M.W. (1965). Evolution and symbiosis in the genus Chlorella and related algae. Evolution **19:** 368–377.

Kato, Y., Ueno, S., and Imamura, N. (2006). Studies on the nitrogen utilization of endosymbiotic algae isolated from Japanese *Paramecium bursaria*. Plant Sci. **170:** 481–486.

Kawasaki, T., Tanaka, M., Fujie, M., Usami, S., Sakai, K., and Yamada, T. (2002). Chitin synthesis in chlorovirus CVK2-infected chlorella cells. Virology **302:** 123–131.

Lacombe-Harvey, M.E., Fukamizo, T., Gagnon, J., Ghinet, M.G., Dennhart, N., Letzel, T., and Brzezinski, R. (2009). Accessory active site residues of Streptomyces sp. N174 chitosanase: Variations on a common theme in the lysozyme superfamily. FEBS J. **276:** 857–869.

Lau, S., Jürgens, G., and De Smet, I. (2008). The evolving complexity of the auxin pathway. Plant Cell **20:** 1738–1746.

Lau, S., Shao, N., Bock, R., Jürgens, G., and De Smet, I. (2009). Auxin signaling in algal lineages: Fact or myth? Trends Plant Sci. **14:** 182–188.

Malik, S.B., Pightling, A.W., Stefaniak, L.M., Schurko, A.M., and Logsdon, J.M., Jr. (2008). An expanded inventory of conserved meiotic genes provides evidence for sex in *Trichomonas vaginalis*. PLoS ONE **3:** e2879.

Martin, F., et al. (2008). The genome of *Laccaria bicolor* provides insights into mycorrhizal symbiosis. Nature **452:** 88–92.

Merchant, S.S., et al. (2007). The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. Science **318:** 245–250.

Monier, A., Pagarete, A., de Vargas, C., Allen, M.J., Read, B., Claverie, J.-M., and Ogata, H. (2009). Horizontal gene transfer of an entire metabolic pathway between a eukaryotic alga and its DNA virus. Genome Res. **19:** 1441–1449.

Morin, P.J. (2008). Sex as an algal antiviral strategy. Proc. Natl. Acad. Sci. USA **105:** 15639–15640.

Palenik, B., et al. (2007). The tiny eukaryote Ostreococcus provides genomic insights into the paradox of plankton speciation. Proc. Natl. Acad. Sci. USA **104:** 7705–7710.

Pazour, G.J., Agrin, N., Leszyk, J., and Witman, G.B. (2005). Proteomic analysis of a eukaryotic cilium. J. Cell Biol. **170:** 103–113.

Pickett-Heaps, J.D. (1975). Green Algae: Structure, Reproduction, and Evolution in Selected Genera. (Sunderland, MA: Sinauer Associates).

Pollock, S.V., Pootakham, W., Shibagaki, N., Moseley, J.L., and Grossman, A.R. (2005). Insights into the acclimation of *Chlamydomonas reinhardtii* to sulfur deprivation. Photosynth. Res. **86:** 475–489.

Rajamani, S., Siripornadulsil, S., Falcao, V., Torres, M., Colepicolo, P., and Sayre, R. (2007). Phycoremediation of heavy metals using transgenic microalgae. Adv. Exp. Med. Biol. **616:** 99–109.

Rensing, S.A., et al. (2008). The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants. Science **319:** 64–69.

Ryo, H., Mitsunori, I., and Nobutaka, I. (2010). *Chlorella variabilis* and *Micractinium reisseri* sp. nov. (Chlorellaceae, Trebouxiophyceae): Redescription of the endosymbiotic green algae of *Paramecium bursaria* (Peniculia, Oligohymenophorea) in the 120th year. Phycological Res. **58:** 188–201.

Schenk, P.M., Thomas-Hall, S.R., Stephens, E., Marx, U.C., Mussgnug, J.H., Posten, C., Kruse, O., and Hankamer, B. (2008). Second generation biofuels: High-efficiency microalgae for biodiesel production. Bioenergy Res. **1:** 20–43.

Schurko, A.M., and Logsdon, J.M., Jr. (2008). Using a meiosis detection toolkit to investigate ancient asexual "scandals" and the evolution of sex. Bioessays **30:** 579–589.

Stirk, W.A., Ördög, V., Van Staden, J., and Jäger, K. (2002). Cytokinin and auxin-like activity in Cyanophyta and microalgae. J. Appl. Phycol. **14:** 215–221.

Sun, L., Adams, B., Gurnon, J.R., Ye, Y., and Van Etten, J.L. (1999). Characterization of two chitinase genes and one chitosanase gene encoded by *Chlorella virus* PBCV-1. Virology **263:** 376–387.

Takeda, H. (1988). Classification of Chlorella strains by cell wall sugar composition. Phytochemistry **27:** 3823–3826.

Takeda, H. (1991). Sugar composition of the cell wall and the taxonomy of Chlorella (Chlorophyceae). J. Phycol. **27:** 224–232.

Tarakhovskaya, E.R., Maslov, Y.I., and Shishova, M.F. (2007). Phytohormones in algae. Russ. J. Plant Physiol. **54:** 163–170.

Tram, U., Ferree, P.M., and Sullivan, W. (2003). Identification of Wolbachia—Host interacting factors through cytological analysis. Microbes Infect. **5:** 999–1011.

Vance, B.D. (1987). Phytohormone effects on cell division in *Chlorella pyrenoidosa* chick (TX-7-11–05) (chlorellaceae). J. Plant Growth Regul. **5:** 169–173.

Villarreal, L.P.P. (2004). Viruses and the Evolution of Life. (Washington DC: ASM Press).

Watanabe, T., Uchida, M., Kobori, K., and Tanaka, H. (1994). Site-directed mutagenesis of the Asp-197 and Asp-202 residues in chitinase A1 of *Bacillus circulans* WL-12. Biosci. Biotechnol. Biochem. **58:** 2283–2285.

Wickstead, B., and Gull, K. (2007). Dyneins across eukaryotes: A comparative genomic analysis. Traffic **8:** 1708–1721.

Wilson, W.H., Van Etten, J.L., and Allen, M.J. (2009). The Phycodnaviridae: The story of how tiny giants rule the world. In Lesser Known Large dsDNA Viruses, J.L. Van Etten, eds (Berlin-Heidelberg, Germany: Springer-Verlag), pp. 1–42.

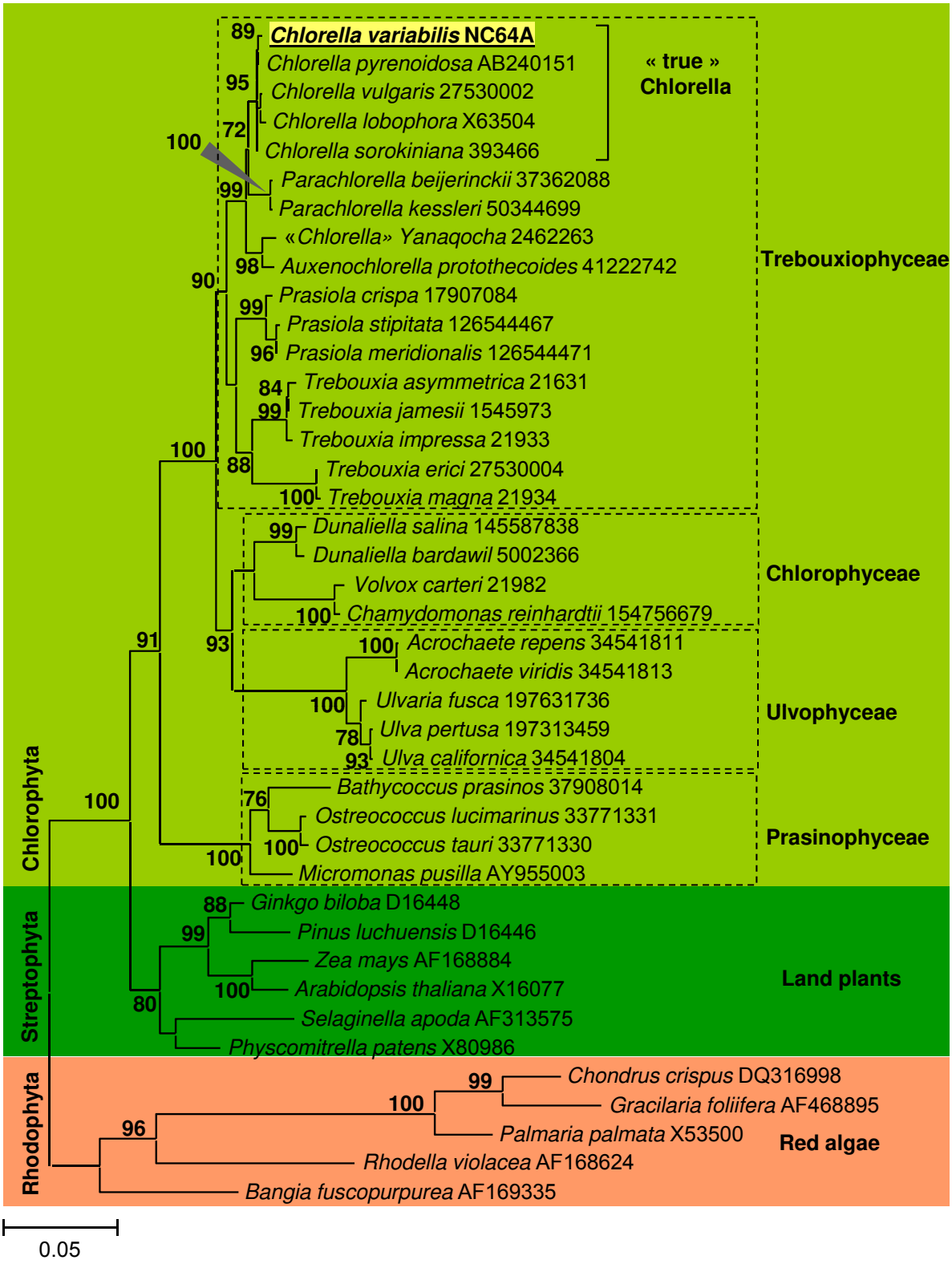
Worden, A.Z., et al. (2009). Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes Micromonas. Science **324:** 268–272.

Wu, G.A., Jun, S.R., Sims, G.E., and Kim, S.H. (2009). Whole-proteome phylogeny of large dsDNA virus families by an alignment-free method. Proc. Natl. Acad. Sci. USA **106:** 12826–12831.

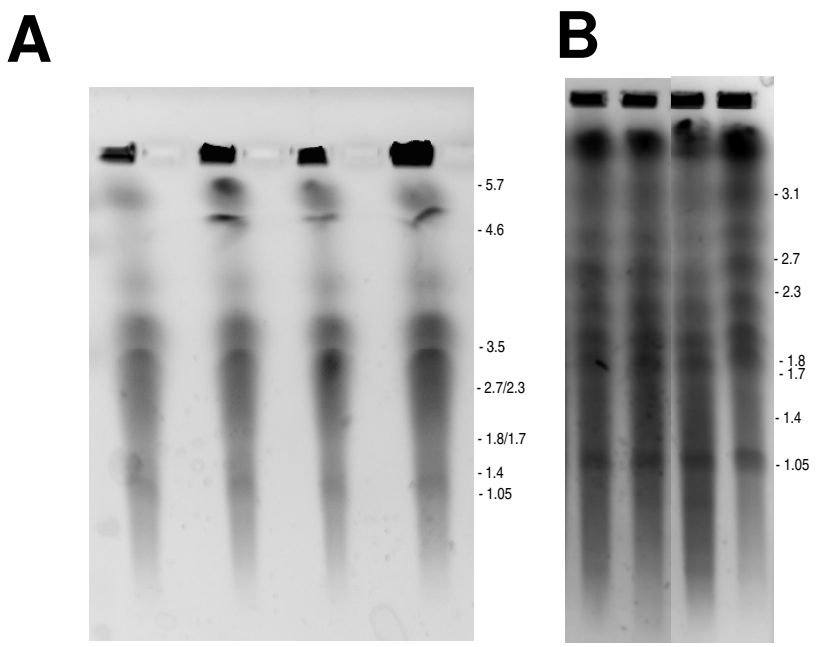
Wu, M., et al. (2004). Phylogenomics of the reproductive parasite *Wolbachia pipientis* wMel: A streamlined genome overrun by mobile genetic elements. PLoS Biol. **2:** E69.

Yabe, T., Yamada-Okabe, T., Nakajima, T., Sudoh, M., Arisawa, M., and Yamada-Okabe, H. (1998). Mutational analysis of chitin synthase 2 of *Saccharomyces cerevisiae*. Identification of additional amino acid residues involved in its catalytic activity. Eur. J. Biochem. **258:** 941–947.

Yin, H.C. (1937). Effect of auxin on *Chlorella vulgaris*. Proc. Natl. Acad. Sci. USA **23:** 174–176.



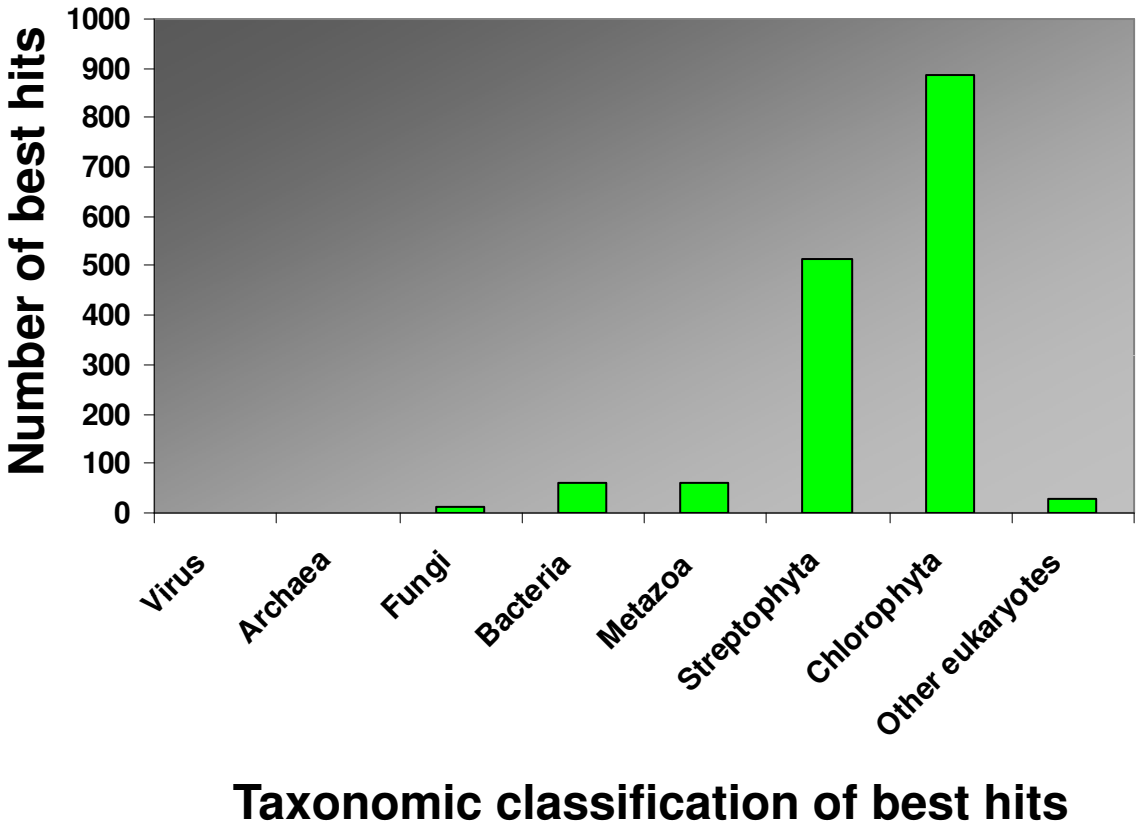
Supplemental Figure 1: Phylogenetic position of *Chlorella* variabilis NC64A among chlorophyte algae. Maximum likelihood (ML) phylogenetic tree based on the analysis of 18S gene sequences and the HKY+Γ+*I* substitution model. The multiple-sequence alignment contained 1509 gap-free sites (provided in Supplemental data set 6 online). Selection of the best substitution model for ML tree searches was performed with the MODELTEST program. Approximate likelihood ratio test support values for branches are given beside branches (only values >50%). Genbank identification (gi) numbers of sequences are given beside species names.



Supplemental Figure 2: Pulsed field gel electrophoresis of *Chlorella* sp. NC64A chromosomes

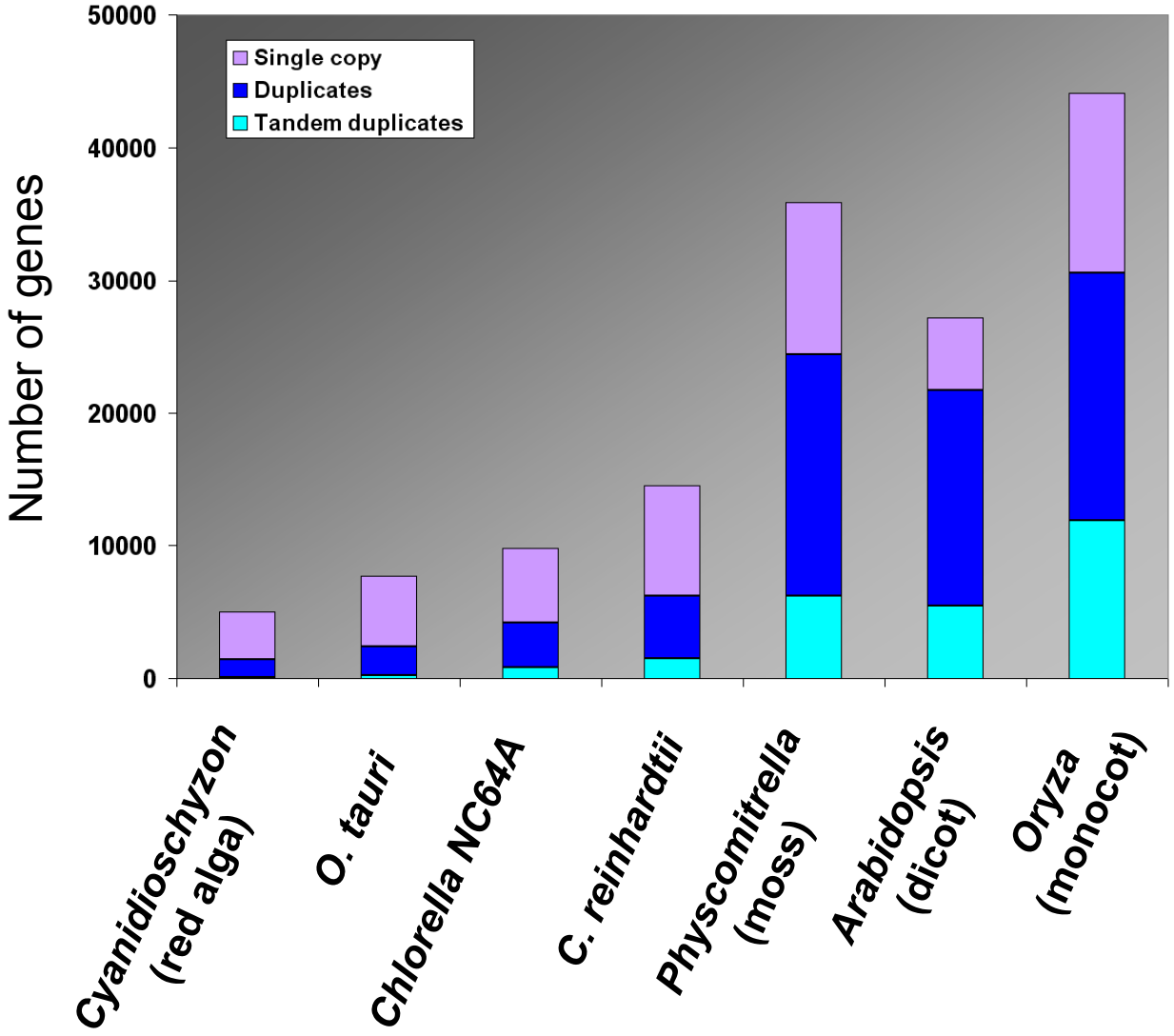
Electrophoresis conditions were as follows: **(A)** 0.8% agarose gel in 0.5× TBE buffer with pulse ramped from 0.45K sec to 1.2K sec for 72 h at 3 V/cm; **(B)** 0.7% agarose gel in 0.5× TBE buffer with pulse ramped from 4.5K sec to 3.0K sec at 1.4 V/cm for 72 hrs; then from 3.6K sec to 3.0K sec at 1.4 V/cm for 24 hrs; then 2.7K sec at 1.8 V/cm for 24 hrs; then from 3.9K sec to 0.8K sec at 2.1 V/cm for 48 hrs;

PFGE revealed 12 bands ranging in size from 1.1 Mb to >6.4 Mb [1.1 Mb. 1.9 Mb, 2.1 Mb, 2.4 Mb, 2.8 Mb, 3.2 Mb, 3.8 Mb, 4.2 Mb, 4.7 Mb, 5.0 Mb, 6.4 Mb, and >6.4 Mb (these sizes were calculated as the average of 18 gel runs with different pulse conditions)]. The top band represents compressed chromosomal DNA beyond the resolution limits. Summation of the sizes of all bands (except for the top band with compressed DNA) gives a genome size of ~ 37.6 Mb that is ~ 8.6 Mb less than the actual size, based on the genome sequence assembly (46.2 Mb). This suggests that the top unresolved band (A2 and A3) may be the 12th chromosome of 8.6 Mb.

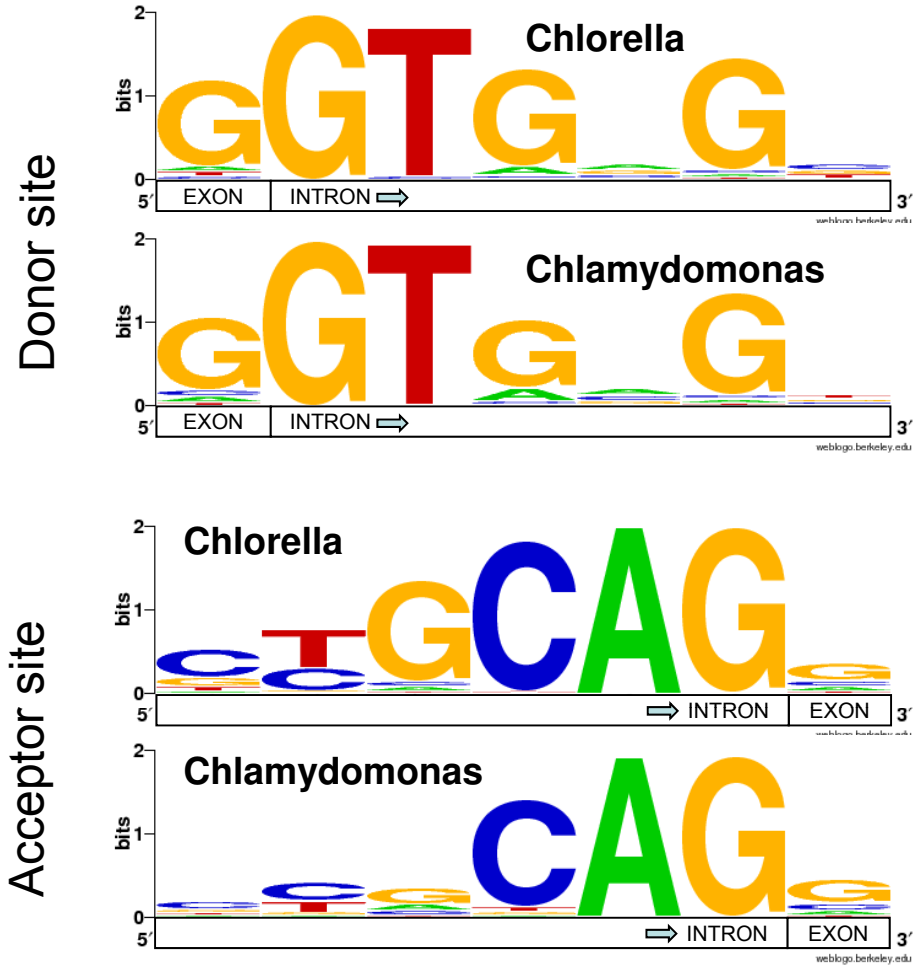


Supplemental Figure 3: Taxonomic distribution of best matches for proteins encoded in low-GC regions

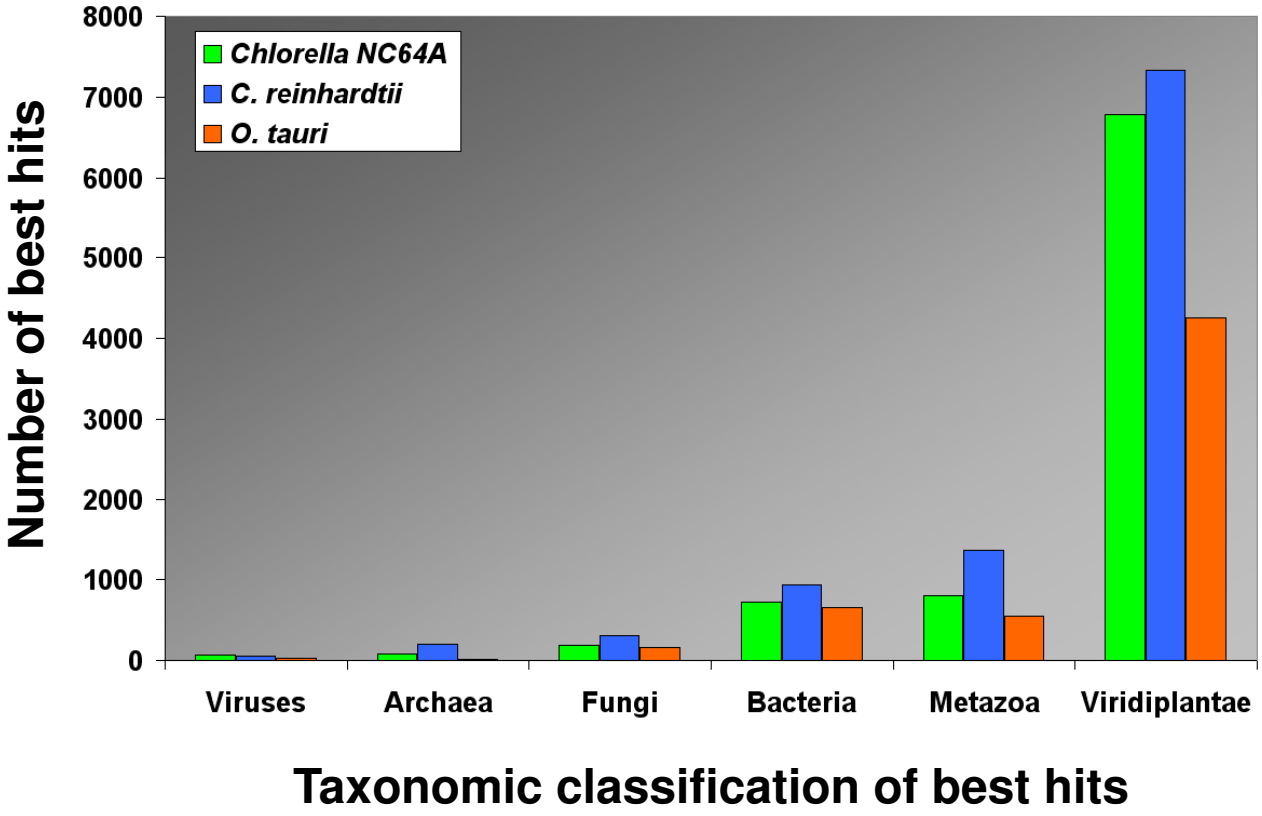
1,384 NC64A proteins encoded in the low-GC regions of Chlorella variabilis NC64A were aligned against the NCBI-NR database using the BLASTP program. Only the best BLAST hits with E-value < 1e-5 were recorded.



Supplemental Figure 4: Gene duplication in selected Viridiplantae
Genes were considered as duplicate when their translation products matched in BLASTP searches with E-value < 1e-5. Duplicate genes were considered as resulting from tandem duplication if they lay less than ten genes apart in the genome.



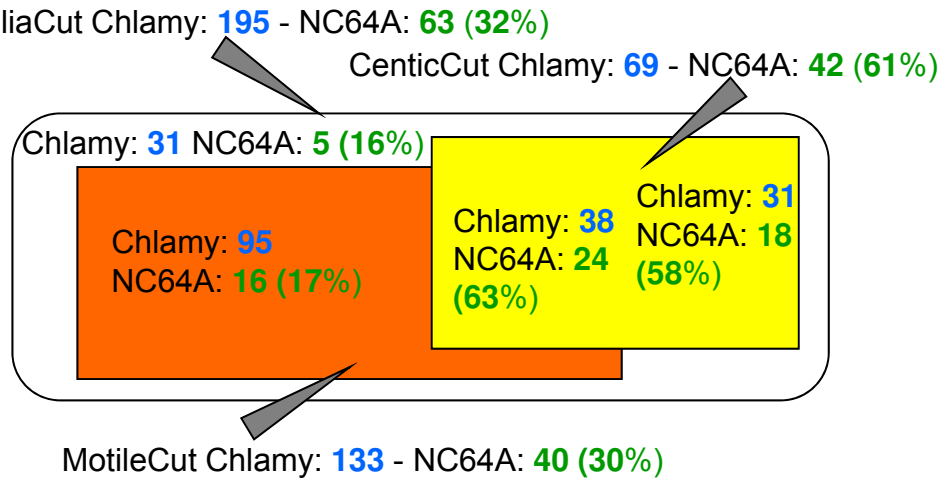
Supplemental Figure 5: Sequences motifs for intron splice sites
Different font sizes indicate the probability of a particular nucleotide at the respective motif position. Logos were calculated from intron sequences whose borders were confirmed by EST data.



Supplemental Figure 6: Taxonomic distribution of best matches for representative Chlorophytes

The full proteome complements of the three green algae were aligned against the NCBI-NR database using the BLASTP program. Only the best BLAST hits with E-value < 1e-5 were recorded.

A



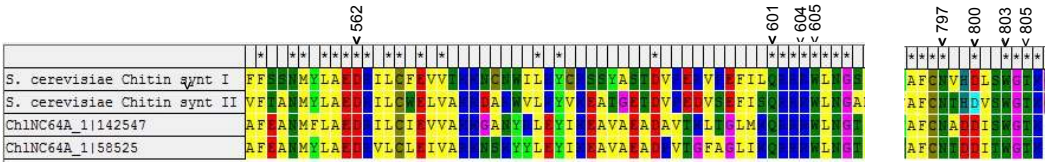
B

	MotileCut	SSA (non MotileCut)	P-value
Chlamydomonas CiliaCut proteins	133	62	
Chlamydomonas CiliaCut proteins with putative orthogues in NC64A	40 (30%)	23 (37%)	0.42

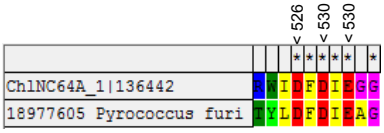
	CentricCut	non CentricCut	P-value
Chlamydomonas CiliaCut proteins	69	126	
Chlamydomonas CiliaCut proteins with putative orthogues in NC64A	42 (61%)	21 (17%)	2.07E-07

Supplemental Figure 7: NC64A putative orthologues to Chlamydomonas CiliaCut proteins
(A) CiliaCut: The CiliaCut contains 195 Chlamydomonas proteins with homologs in human and species of Phytophthora, but not in nonciliated organisms. This group was subdivided on the basis of whether or not a homolog was present in Caenorhabditis, which has only nonmotile sensory cilia. The 133 CiliaCut proteins without homologs in Caenorhabditis were designated the MotileCut (orange rectangle). Proteins with homologs in *Caenorhabditis* are associated with nonmotile cilia (white and yellow areas) and were named SSA (for structural, sensory or assembly). The CentricCut (yellow plus light orange box) is made up of 69 CiliaCut homologs present in the centric diatom Thalassiosira. These proteins can be divided into those also in the MotileCut (38 proteins; light orange box) or those not present in the MotileCut (31 proteins; yellow box). Blue numbers indicate the number of Chlamydomonas proteins in each of the subdivisions of the CiliaCut, whereas green numbers indicates the number and proportion (in parentheses) of Chlamydomonas CiliaCut proteins with putative orthologues in NC64A . Modified from figure 5 in Merchant et al. (2007). (B) Distribution of Chlamydomonas proteins in the CiliaCut sub-categories. P-values were calculated with the Chi-square test and are associated to the null hypothesis that the Chlamydomonas CiliaCut proteins with putative orthologues in NC64A are randomly distributed in the CiliaCut sub-categories (relative to the distribution of all CiliaCut proteins)

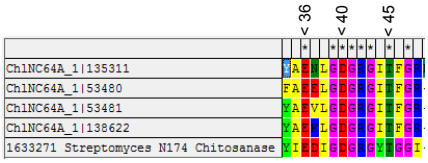
(A) Chitin synthase



(B) Chitinase

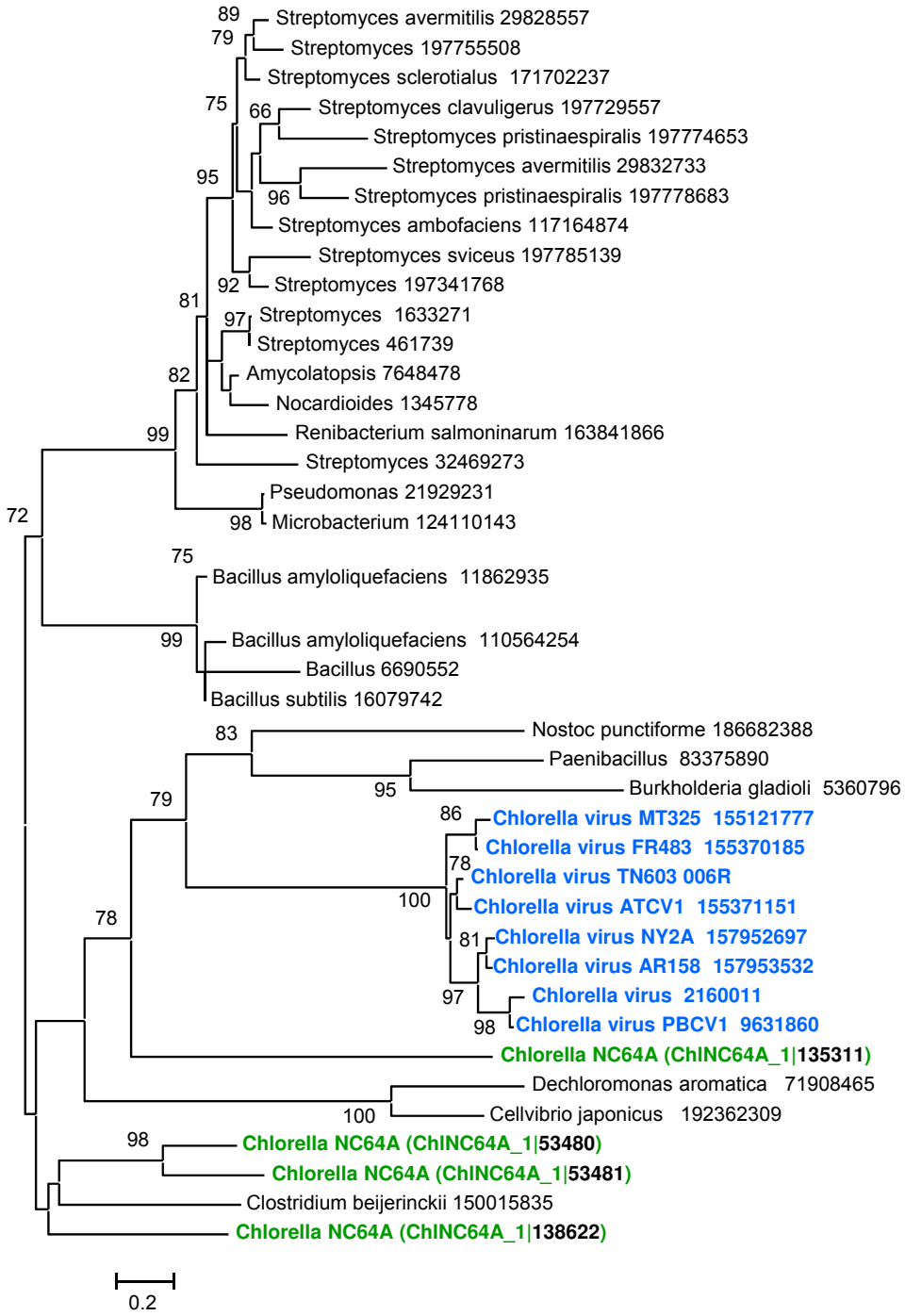


(C) Chitosanase

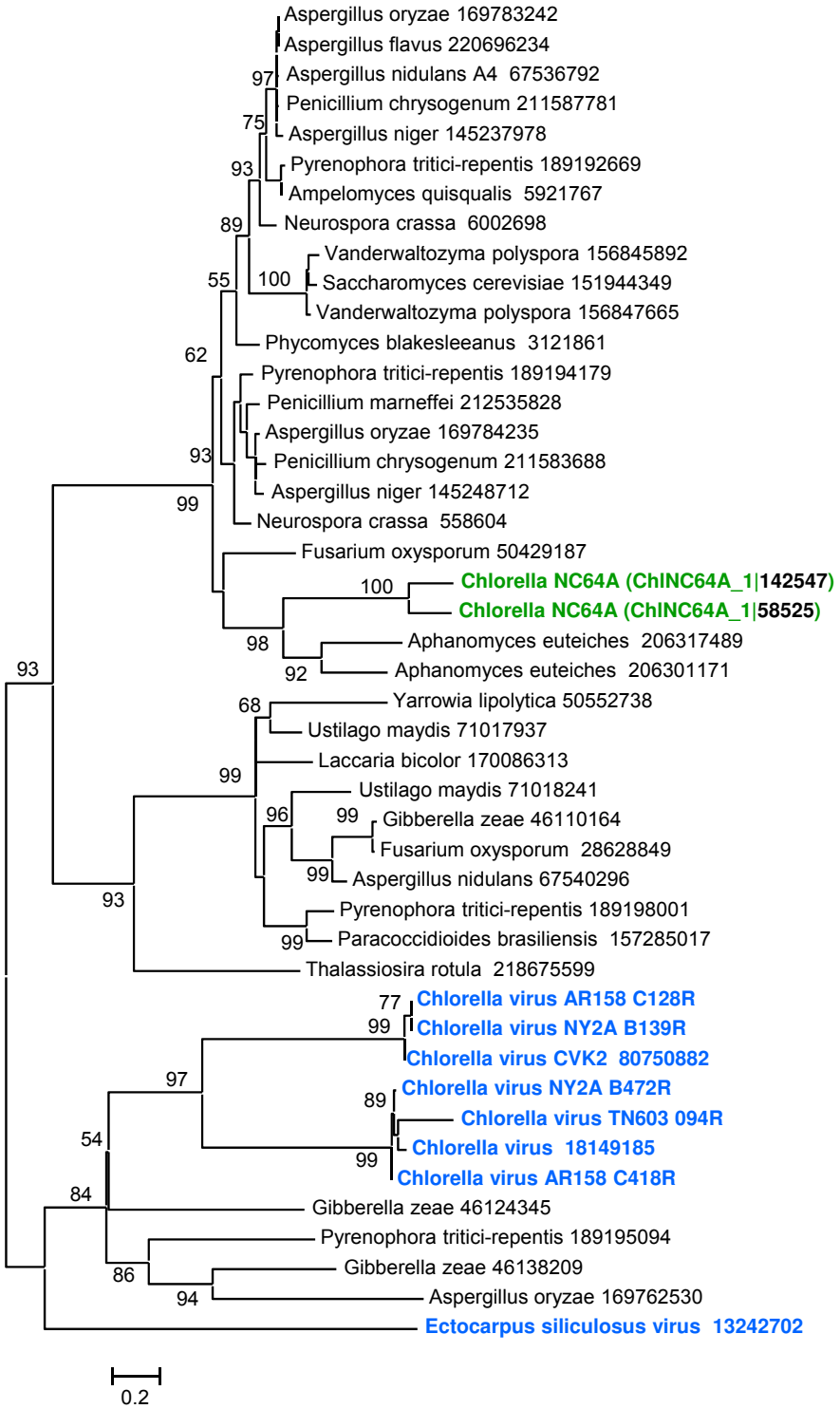


Supplemental Figure 8: Alignment of NC64A proteins with their reference proteins involved in chitin metabolism.

Arrows indicate residues identified as essential for the catalytic activity in the reference protein. (A) Alignment of chitin synthases. Numbering of essential residues refers to the *S. cerevisiae* chitin synthase II. The Genbank id (gi) of the *S. cerevisiae* chitin synthases I and II are 6324137 and 6319512 respectively (B) Alignment of chitinase. Numbering of essential residues refers to the NC64A protein. The gi of the *Pyrococcus furi* reference protein is 1633271. (C) Alignment of chitosanase. Numbering of essential residues refers to the *Streptomyces* sp. NC147 protein (gi: 1633271).



Supplemental Figure 9: ML phylogenetic tree of chitosanase proteins
The multiple-sequence alignment contained 156 gap-free sites (provided in Supplemental data set 7 online). The phylogenetic tree was reconstructed using the WAG+I+G model of substitutions. Approximate likelihood ratio test values >50 are indicated beside branches.



Supplemental Figure 10: ML phylogenetic tree of chitin synthase proteins
The multiple-sequence alignment contained 241 gap-free sites (provided in Supplemental data set 8 online). The phylogenetic tree was reconstructed using the WAG+I+G model of substitutions. Approximate likelihood ratio test values >50% are indicated beside branches.

Supplemental Table 1: Nuclear genome assembly statistics

Features	
Number of scaffolds	413
GC content coding[a]	66%
GC content Intron[a]	68%
Gene count	9,791 (100%)
Complete ORF (with start and stop codons)	8,519 (87%)
Supported by ESTs	4,138 (42%)
Supported by homology (NCBI NR)	8,747 (89%)
Contain Pfam domain	5,537 (57%)
Ave. Gene length (nt)	2,928
Ave. Coding sequence length (nt)	1,368
Repeated sequence content	11.5%

[a] Calculated from genes whose gene structure is confirmed by EST

Supplemental Table 2: EuKaryotic Ortholog Groups (KOG) functional categories among low-GC and normal-GC regions

KOG category	Nb of proteins in category	Percent of proteins in low GC regions	P-value ‡
All KOG categories	7,938	16.21%	
Amino acid transport and metabolism	323	21.36%	0.012
Carbohydrate transport and metabolism	324	19.14%	0.153
Cell cycle control, cell division, chromosome partitioning	170	20.00%	0.180
Cell motility	4	25.00%	0.634
Cell wall/membrane/envelope biogenesis	99	18.18%	0.595
Chromatin structure and dynamics	227	7.93%	0.001*
Coenzyme transport and metabolism	114	20.18%	0.251
Cytoskeleton	271	17.34%	0.614
Defense mechanisms	82	15.85%	0.930
Energy production and conversion	241	14.11%	0.375
Extracellular structures	210	8.10%	0.001*
Function unknown	477	15.93%	0.868
General function prediction only	1,063	14.58%	0.149
Inorganic ion transport and metabolism	188	15.43%	0.770
Intracellular trafficking, secretion, and vesicular transport	315	20.00%	0.068
Lipid transport and metabolism	266	22.56%	0.005
Nuclear structure	56	14.29%	0.696
Nucleotide transport and metabolism	95	25.26%	0.017
Posttranslational modification, protein turnover, chaperones	774	16.54%	0.807
Replication, recombination and repair	243	16.46%	0.917
RNA processing and modification	335	19.40%	0.113
Secondary metabolites biosynthesis, transport and catabolism	140	20.71%	0.148
Signal transduction mechanisms	682	16.57%	0.801
Transcription	843	9.61%	0.000*
Translation, ribosomal structure and biogenesis	396	20.20%	0.031

‡ P-value associated with the null hypothesis that the low-GC and normal-GC regions have an unbiased distribution of proteins in functional category, relative to the total number of proteins. P-values were calculated using the chi-square test.

* statistically significant at $\alpha = 5\%$ after Bonferroni correction

.

Supplemental Table 3: Repeated sequences in the NC64A genome

Class	Family	No. of fragments	Cumulative length (nt)	Percent of total
TOTAL		**29,495**	**5,527,001**	**100.0%**
Unknown		**8,896**	**2,712,218**	**51.6%**
LTR Retrotransposon		**10,824**	**1,160,885**	**21.0%**
	Gypsy	5,797	628,019	11.4%
	Copia	4,048	427,518	7.7%
Non-LTR Retrotransposon		**7,176**	**923,613**	**16.7%**
	RandI	3,164	405,853	7.7%
	L1	1,854	233,943	4.5%
	RTE	645	112,476	2.1%
	GilM	963	109,310	2.1%
Endogenous Retrovirus		**659**	**64,277**	**1.2%**
	ERV1	485	49,403	0.9%
DNA Transposon		**474**	**71,553**	**1.3%**
	Novosib	334	37,568	0.7%
Satellite		**92**	**12,892**	**0.2%**
rRNA - tRNA		**215**	**21,186**	**0.4%**
Gene duplicates*		**1,141**	**536,090**	**10.3%**

* refers to nucleotide sequences with significant BLASTX match with proteins, excluding transposable-element related proteins

Supplemental Table 4: Chlorophyte algae specific PFAM protein domains

PFAM domain	*C. sp. NC64A**	*C. reinhardtii**	*M. sp. CCMP1545**	*M. pusilla **	*O. lucimarinus**	*O. tauri**	Description
Amino acid metabolism							
Asparaginase	**2**	**2**	**1**	**1**	**2**	**1**	**L-Asparaginase type 1 (bacterial)**
Carbohydrate metabolism							
ACN9	**1**	**1**	**1**	**1**	**1**	**1**	**ACN9 protein**
Fe-ADH	**2**	**3**	**2**	**1**	**1**	**1**	**Iron-containing alcohol dehydrogenase**
Fatty acid metabolism							
PYC_OADA	**1**	**1**	**1**	**1**	**1**	**1**	**Conserved domain of pyruvate carboxylase**
Cell adhesion/signal transduction							
CD36	**2**	**1**	**1**	**1**	**1**	**1**	**CD36 family scavenger receptors class B**
Cell envelope							
CotH	**1**	**1**	**2**	**1**	**2**	**2**	**CotH family protein**
Cytoskeleton							
p25-alpha	**3**	**5**	**2**	**1**	**1**	**1**	**p25-alpha microtubule-targeting protein family**
DNA or RNA metabolism							
Bac_DNA_binding	**1**	**1**	**1**	**1**	**1**	**1**	**Histone Like DNA-binding HU-beta protein**
HA	**3**	**1**	**2**	**3**	**4**	**5**	**Helicase domain**
Protein metabolism							
Cu2_monooxygen	**1**	**3**	**1**	**1**	**1**	**2**	**Peptidylglycine alpha-amidating monooxygenase (PAM)**
SLA_LP_auto_ag	**1**	**1**	**1**	**1**	**1**	**1**	**Selenocystein aminotransferase**
Respiration							
Cyto_heme_lyase	**2**	**3**	**2**	**2**	**2**	**2**	**Cytochrome c/c1 heme lyase**
Signal transduction							
PDEase_I	**2**	**28**	**4**	**6**	**2**	**2**	**3'5'-cyclic nucleotide phosphodiesterase**
Transport							
Form_Nir_trans	**3**	**6**	**1**	**2**	**1**	**1**	**Formate/nitrite transporter**
Miscellaneous							
Thioesterase	**1**	**2**	**2**	**2**	**2**	**2**	**Thioesterase domain of type I polyketide synthase**
DUF101	**1**	**1**	**1**	**1**	**1**	**1**	**Protein of unknown function**
DUF1824	**1**	**1**	**1**	**1**	**1**	**1**	**Protein of unknown function**
DUF2009	**1**	**1**	**1**	**1**	**1**	**1**	**Protein of unknown function**
DUF262	**2**	**2**	**2**	**2**	**1**	**1**	**Protein of unknown function**
DUF395	**1**	**3**	**1**	**2**	**3**	**1**	**Protein of unknown function**
GCC2_GCC3	**15**	**5**	**2**	**4**	**1**	**1**	**GCC2 and GCC3 domain**
NIPSNAP	**1**	**2**	**1**	**1**	**1**	**1**	**Protein of unknown function**
Tcp10_C	**1**	**1**	**1**	**1**	**1**	**1**	**T-complex protein 10 C-terminus**
TIG	**1**	**3**	**9**	**9**	**10**	**8**	**IPT/TIG domain**
UPF0079	**1**	**1**	**1**	**1**	**1**	**1**	**Protein of unknown function**
VTC	**2**	**2**	**2**	**2**	**1**	**1**	**Protein of unknown function**
Ycf66_N	**1**	**1**	**1**	**1**	**1**	**1**	**Protein of unknown function**

***** Number of proteins containing a PFAM domain

Supplemental Table 5: PFAM domains with biased distribution in chlorophyte green algae

PFAM domain	***C. sp. NC64A***	***C. reinhardtii***	***M. sp. CCMP1545***	***M. pusilla (RCC299)***	***O. lucimarinus***	***O. tauri***	**P-value**	**Putative function of proteins or domains**
Pkinase	188	432	112	127	99	88	3.E-104*	Protein kinase
Histone	33	149	16	17	11	13	4.6E-77*	Core histone H2A/H2B/H3/H4
Guanylate_cyc	2	76	2	1	0	0	6.0E-73*	Adenylate and Guanylate cyclase
Sel1	3	2	72	6	2	3	3.3E-56*	Sel1 domain containing protein
Polysacc_deac_1	25	0	0	0	0	0	2.7E-25*	Chitin deacetylase
Peptidase_M11	17	34	0	0	0	0	5.1E-24*	Gametolysin peptidase M11
SBP	34	21	3	3	0	0	1.5E-19*	Squamosa promoter binding protein family
zf-MYND	58	42	27	15	4	3	2.4E-19*	MYND finger protein-protein interaction domain
Sulfatase	1	19	0	0	0	0	1.3E-17*	Sulfatase
SRCR	14	24	0	0	0	0	1.3E-16*	Scavenger receptor cysteine-rich domain protein
DUF1929	2	18	0	0	0	0	1.8E-15*	Unknown function
PDEase_I	2	28	4	6	2	2	4.7E-14*	3'5'-cyclic nucleotide phosphodiesterase
BTB	6	28	3	3	2	2	5.3E-14*	BTB/POZ protein-protein interaction domain
Exostosin	16	29	1	3	3	2	8.3E-14*	Exostosin family
Aa_trans	35	8	7	8	6	5	2.8E-11*	Transmembrane amino acid transporter protein
GTP_EFTU_D2	16	3	0	0	1	1	1.1E-10*	GTP-binding translation elongation factor EF-Tu-like
Helicase_C	85	36	31	24	37	53	1.2E-10*	Helicase
Trypsin	37	11	14	10	6	5	1.1E-09*	Trypsin-like protease
SCP	12	14	1	0	0	0	2.5E-09*	SCP-like extracellular protein
p450	25	41	12	12	11	10	8.7E-08*	Cytochrome P450
Peptidase_S8	22	15	6	6	2	1	2.7E-07*	Subtilase serine proteases
F-box	32	14	10	8	6	7	4.8E-07*	F-box protein-protein interaction domain
Lipase_3	32	14	9	13	8	7	7.7E-06*	Lipase (class 3)
GCC2_GCC3	15	5	2	4	1	1	1.3E-05*	GCC2 and GCC3 domain protein
U-box	25	21	12	26	6	5	3.7E-05*	U-box domain protein
DUF285	0	0	10	16	9	12	5.3E-05*	unknown function
CBM_20	18	8	5	7	1	3	1.2E-04*	glycosyl transferase
Fasciclin	23	10	9	7	5	4	1.6E-04*	Fasciclin adhesion domain

* significant at $\alpha = 0.05$ after Bonferroni correction for multiple tests. P-values calculated using the Chi square statistics.

Supplemental Table 6: Meiosis-specific proteins Genbank identification (gi) numbers and percentage of protein sequence identity with reference Arabidopsis proteins (in brakets)

List of species	DMC1[a]	HOP1[b]	HOP2[c]	MER3[d]	MND1[c]	MSH4
Chlorella NC64A (Chlorophyte)[1]	52039 (55%)	142584 (27%)	139916 (28%)	140725 (41%)	132912 (40%)	137861 (37%)
C. reinhardtii (Chlorophyte)*	XP_001700483 (59%)		XP_001695346 (29%)	XP_001698471 (39%)	XP_001695418 (43%)	XP_001699298 (46%)
Micromonas sp. RCC299 (Chlorophyte)	ACO70309 (53%)		ACO62770 (26%)	ACO61833 (36%)	EEH55296 (37%)	XP_002504488 (33%)
O. lucimarinus (Chlorophyte)	XP_001420481 (52%)	XP_001417706 (24%)	XP_001417513 (26%)	XP_001418083 (36%)	XP_001419666 (37%)	
O. tauri (Chlorophyte)*	CAL55792 (54%)	CAL53885 (26%)	CAL51676 (26%)	CAL54173 (36%)	19307 (35%)	
Cyanidioschyzon merolae (Rhodophyte)			CMP311C (26%)		CMG028C (39%)	CMK199C (25%)
Giardia intestinalis (protist)*	AAQ24509 (56%)	XP_001707698 (23%)	XP_001704033 (20%)		XP_001708984 (31%)	
Trichomonas vaginalis (protist)*	XP_001303137 (54%)	ABC61969 (32%)	ABC61980 (27%)	XP_001329476 (32%)	XP_001579664 (33%)	XP_001306678 (26%)
Saccharomyces cerevisiae (yeast)*	NP_011106 (53%)	NP_012193 (26%)	NP_01148 (25%)	NP_011263 (33%)	NP_011332 (25%)	NP_116652 (27%)
A. thaliana (dicot)*	NP_188928	NP_564896	AAO67519	NP_189410	NP_001078469	NP_193469
O. sativa (monocot)*	NP_001065738 (81%)	BAD00095 (65%)	ABF98498 (59%)	CI28521 (55%)	NP_001062766 (72%)	NP_001059660 (61%)
Physcomitrella patens (moss)*	Scaffold 9[2] (76%)	XP_001760173 (50%)	XP_001782602 (41%)	XP_001760306 (57%)	XP_001760266 (65%)	XP_001777754 (57%)
*Homo sapiens**	NP_008999 (61%)	CAI13655 (27%)	NP_037422 (31%)	NP_001017975 (33%)	NP_115493 (27%)	NP_002431 (35%)

* Previously reported as being sexually reproducing species.

a Creates double-stranded DNA breaks (for original citations see Malik et al. 2008).

b A synaptonemal complex protein that binds double-stranded breaks during prophase I of meiosis (for original citations see Malik et al. 2008).

c HOP2 and MND1 dimerize and to assure accurate homologous pairing during prophase I of meiosis (for original citations see Malik et al. 2008).

d Helicase involved in Holiday junction resulution (for original citations see Malik et al. 2008).

1 Go to the JGI portal site (http://genomeportal.jgi-psf.org/). and select Chlorella sp. NC64A.

2 Identified by TBLASTN alignment of the Arabidopsis protein against the Physcomitrella genome sequence.

*1 **Supplemental methods***

1.1 *Chlorella sp.* NC64A Genomic DNA Preparation

Chlorella sp NC64A was streaked onto a modified Bold's Basal medium (MBBM) plate and a single colony was grown to log phase (1 - 2 X 10^7 cells/ml) in liquid MBBM medium. The cells were harvested by centrifugation for 6 min., 5,000 g 4 C, flash frozen with liquid nitrogen, and stored at -80 C. The cell pellets, containing a collective total of 3 X 10^{10} cells, were then processed using a modification of the standard operating procedure /protocol from JGI for bacterial genomic DNA isolation using CTAB, version number 2.

The cell pellets were thawed in 1X TE buffer, supplemented with 0.5% SDS and 100 μg/ml proteinase K, and incubated at 53 C overnight (16 h). NaCl was added to 0.6 M. Pre-warmed CTAB/NaCl solution was added for a final concentration of 28 mM CTAB/0.65 M NaCl and incubated at 65 C for 1 h. The sample was extracted with chloroform:isoamyl alcohol (24:1) followed by phenol:chloroform:isoamyl alcohol (25:24:1). Nucleic acids were precipitated with isopropanol, washed with 70% EtOH, and dried. The pellets were resuspended in 1X TE buffer, supplemented with RNAse A to 100 μg/ml and incubated at 37 C for 30 min. The total DNA sample was extracted with phenol:chloroform:isoamyl alcohol (25:24:1), precipitated with 0.3 M NaOAc and 2 1/2 vol. EtOH, dried and resuspended in 1X TE buffer.

Total DNA was centrifuged on 40-60% CsCl gradients equilibrated with 1X TE, pH 8.0 buffer containing 1 μg/ml Hoechst 33258 dye to enrich for the nuclear DNA. The upper bands containing chloroplast DNA were removed and the lower bands containing nuclear DNA were collected with a wide-mouth pipet tip. The Hoechst dye was extracted from the DNA twice with an equal volume of CsCl/TE-saturated isopropanol. The samples were diluted with 1X TE and the DNA was precipitated with 0.3 M NaOAc and 2 vol. EtOH at -20 C, washed with 70% EtOH, dried, and resuspended in a total of 800 μl 1X TE, pH 8.0.

The quality of the purified nuclear-enriched genomic NC64A DNA was monitored with a wavelength absorbance scan and electrophoresis on a 0.8% 1X TBE agarose gel compared to varying amounts of lambda phage DNA.

1.2 Pulse Field Gel Electrophoresis

PFGE studies were carried out according to Agarkova et al. (Agarkova et al., 2006). Briefly, 100 ml of actively growing NC64A cells were harvested from 4-day old cultures (1.2-2.0 x 10^7 cells/ml) by centrifugation at 5000 × g for 5 min, washed 3 times with ice cold TE buffer amended with 50 mM EDTA and then re-suspended in 0.5 ml of TE buffer at a concentration 0.6-1.0 × 10^9 cells/ml. The re-suspended cells were mixed with an equal volume of 2% low melting point agarose (Bio-Rad) in TE buffer at 45°C, poured into plug molds (Bio-Rad, Hercules, CA), and placed at 4°C for 15 min to solidify. Agarose blocks were incubated in approximately 2 ml of 1 mg/ml proteinase K in DB solution (250 mM EDTA, pH 9.5; 1% N-lauroylsarcosine) for 24 h. After digestion, samples were washed two times for 30 min with DB solution and cut into small pieces that fit into gel wells. Samples were sealed with 0.8% low melting point agarose at 45°C in electrophoresis buffer. Chromosomal DNAs were

separated in a CHEF-DR II (Bio-Rad) unit in a 0.8 % agarose gel. Electrophoresis conditions and running buffer were selected to resolve the target chromosome sizes. The exact conditions are described in the figure legends. *Hansenula wingei* chromosomes (1.05-3.13 Mb) and *Schizosaccharomyces pombe* chromosomes (3.5-5.7 Mb) (Bio-Rad) were used as DNA size markers. Gels were stained with 1.0 μg/ml ethidium bromide for 30 min and digital images were captured with the ChemiDoc EQ imaging System (Bio-Rad).

1.3 Genome sequencing and assembly

The NC64A genome was sequenced using WGS strategy. Five libraries with insert sizes of 2-3 KB, 6-8 KB, and 35-40 KB were used. The sequenced reads were screened for vector using cross_match software (www.phrap.org), trimmed for vector and quality, and filtered to remove reads shorter than 100 bases, which resulted in the following dataset:

346,070 2-3 KB reads, containing 217 MB of sequence.

318,619 6-8 KB reads, containing 207 MB of sequence.

45,816 35-40 KB reads, containing 23 MB of sequence.

The data was assembled using release 2.10.11 of Jazz, a WGS assembler developed at the JGI (Aparicio et al., 2002). A word size of 13 was used for seeding alignments between reads. The unhashability threshold was set to 40, preventing words present in the data set in more than 40 copies from being used to seed alignments. A mismatch penalty of -30.0 was used, which will tend to assemble together sequences that are more than 97% identical. The genome size and sequence depth were initially estimated to be 47 MB and 8.0, respectively.

After excluding redundant (<5Kb, with 80% of total length contained in scaffolds > 5 KB) and short (<1Kb) scaffolds from the initial assembly, there was 46.2 MB of scaffold sequence, of which 3.9 MB (8.5%) were gaps. The filtered assembly contained 413 scaffolds, with a scaffold N/L50 of 12/1.5 MB, and a contig N/L50 of 441/27.6 KB. The sequence depth derived from the assembly was 8.95 ± 0.15. Mapping of 7,624 clustered EST sequences onto the genome sequences suggests that the assembly contains >97% of the gene complement.

1.4 cDNA library construction and sequencing:

Chlorella sp. NC64A cells were grown to log phase (1.5 x 10^7 cells/ml) and harvested by centrifugation. The cell pellets were immediately flash frozen in liquid nitrogen, disrupted with glass beads and vortexing in the presence of TRIZOL reagent (Invitrogen, Carlsbad, CA), and total RNA was isolated according to manufacturer's instructions. The integrity of the sample was evaluated by spectrophotometry and electrophoresis on a denaturing agarose gel. NC64A poly A+ RNA was isolated from total RNA using the Absolutely mRNA Purification kit and manufacturers instructions (Stratagene, La Jolla, CA). cDNA synthesis and cloning was a modified procedure based on the "SuperScript plasmid system with Gateway technology for cDNA synthesis and cloning" (Invitrogen, Carsbad, CA). 1-2 μg of poly A+ RNA, reverse transcriptase SuperScript II (Invitrogen) and oligo dT-NotI primer (5' GACTAGTTCTAGATCGCGAGCGGCCGCCCT15VN 3') were used to synthesize first strand cDNA. Second strand synthesis was performed with *E. coli* DNA ligase, polymerase I, and RNaseH followed by end repair using T4 DNA polymerase. The SalI adaptor (5' TCGACCCACGCGTCCG and 5' CGGACGCGTGGG) was ligated to

the cDNA, digested with NotI (New England Biolabs, Ipswich, MA), and subsequently size selected by gel electrophoresis (1.1% agarose). Two size ranges of cDNA were cut out of the gel to generate separate size selected cDNA libraries: 0.6kb-2kb (library codes CPBS and CBWF) and >2kb (library code CBWC). The cDNA inserts were directionally ligated into the SalI and NotI digested vector pCMVsport6 (Invitrogen). The ligation was transformed into ElectroMAX T1 DH10B cells (Invitrogen).

Library quality was first assessed by randomly selecting 24 clones and PCR amplifying the cDNA inserts with the primers M13-F (5' GTAAAACGACGGCCAGT) and M13-R (5' AGGAAACAGCTATGACCAT) to determine the fraction of clones without inserts. Colonies from each library were plated onto agarose plates (254 mm plates from Teknova, Hollister, CA) at a density of approximately 1,000 colonies per plate. Plates were grown at 37 C for 18 hr then individual colonies were picked and each used to inoculate a well containing LB media with appropriate antibiotic in a 384 well plate (Nunc, Rochester, NY). Clones in 384 well plates were grown at 37 C for 18 hr. Plasmid DNA for sequencing was produced by rolling circle amplification (Detter et al., 2002) (Templiphi, GE Healthcare, Piscataway, NJ). Subclone inserts were sequenced from both ends using primers complimentary to the flanking vector sequence (Fw: 5' ATTTAGGTGACACTATAGAA Rv: 5' TAATACGACTCACTATAGGG) with Big Dye terminator chemistry and run on ABI 3730 instruments (Applied Biosystems, Foster City, CA). A total of 23,828 ESTs remained after trimming and filtering.

1.5 EST sequence processing and assembly:

A total of 38,400 ESTs including; 6,144 from CPBS, 16,128 from CBWC, and 16,128 from CBWF were processed through the JGI EST pipeline (ESTs were generated in pairs, a 5' and 3' end read from each cDNA clone). To trim vector and adaptor sequences, common sequence patterns at the ends of ESTs were identified and removed using an internally developed tool. Insertless clones were identified if either of the following criteria were met: >200 bases of vector sequence at the 5' end or less than 100 bases of non-vector sequence remained. ESTs were then trimmed for quality using a sliding window trimmer (window = 11 bases). Once the average quality score in the window was below the threshold (Q15) the EST was split and the longest remaining sequence segment was retained as the trimmed EST. EST sequences with less than 100 bases of high quality sequence were removed. ESTs were evaluated for the presence of poly A or poly T tails, which were removed, and the ESTs were reevaluated for length, removing ESTs with less than 100 bases remaining. ESTs consisting of more than 50% low complexity sequence were also removed from the final set of "good ESTs". In the case of resequencing the same EST, the longest high quality EST was retained. Sister ESTs (end pair reads) were categorized as follows: if one EST was insertless or a contaminant then by default the second sister was categorized as the same. However, each sister EST was treated separately for complexity and quality scores. Finally, EST sequences were compared to the Genbank nucleotide database in order to identify contaminants; non-desirable ESTs such as those matching non-cellular and rRNA sequences were removed.

For clustering, ESTs were evaluated with malign, a kmer based alignment tool (Chapman, unpublished), which clusters ESTs based on sequence overlap (kmer = 16, seed length requirement = 32 alignment ID >= 98%). Clusters of ESTs were

further merged based on sister ESTs using double linkage. Double linkage requires that 2 or more matching sister ESTs exist in both clusters to be merged. EST clusters were then each assembled using CAP3 (Huang and Madan, 1999) to form consensus sequences. Clusters may have more than one consensus sequence for various reasons to include; the clone has a long insert, clones are splice variants or consensus sequences are erroneously not assembled. Cluster singlets are clusters of one EST, whereas CAP3 singlets are single ESTs which had joined a cluster but during cluster assembly were isolated into a separate singlet consensus sequence. ESTs from each separate cDNA library were clustered and assembled separately and subsequently the entire set of ESTs for all cDNA libraries were clustered and assembled together. For cluster consensus sequence annotation, the consensus sequences were compared to Swissprot using BLASTx and the hits were reported. Clustering and assembly of all 24,072 filtered ESTs resulted in 7,624 consensus sequences.

1.6 Genome annotation and sequence analysis

The genome assembly v1.0 of NC64A was annotated using the JGI annotation pipeline, which combines several gene predictors:1) putative full length genes derived from 7,624 cluster consensus sequences of 24,072 clustered and assembled NC64A ESTs were mapped to genomic sequence, 2) homology-based gene models were predicted using FGENESH+ (Salamov and Solovyev, 2000) and Genewise (Birney et al., 2004) seeded by BLASTx alignments against sequences from NCBI non-redundant protein set, 3) *ab initio* gene predictor FGENESH (Salamov and Solovyev, 2000) was trained on the set of putative full-length genes and reliable homology-based models. Genewise models were completed using scaffold data to find start and stop codons. An additional 12,784 gene models were predicted using *ab initio* GeneMark-ES (Ter-Hovhannisyan et al., 2008) and combined with the rest of predictions. ESTs and EST clusters were used to extend, verify, and complete the predicted gene models. Because multiple gene models per locus were often generated, a single representative gene model for each locus was chosen based on homology and EST support and used for further analysis. This led to a filtered set of 9,791 gene models with their characteristics support by different lines of evidence summarized in Supplemental Table 1 online.

All predicted gene models were annotated using InterProScan (Zdobnov and Apweiler, 2001) and hardware-accelerated double-affine Smith-Waterman alignments (www.timelogic.com) against SwissProt (www.expasy.org/sprot) and other specialized databases like KEGG (Ogata et al., 1999) and PFAM (Finn et al., 2010). Finally, KEGG hits were used to map EC numbers (http://www.expasy.org/enzyme/), and Interpro hits were used to map GO terms (Ashburner et al., 2000). In addition, predicted proteins were annotated according to KOG classification (Koonin et al., 2004).

1.6.1 Meiosis and sexual reproduction in green algae

Detection of a sexual cycle involving meiosis in eukaryotic microbes in nature or in the laboratory can be quite difficult. Schurko and Logsdon (2008) and Malik et al. (2008) however, devised a meiosis detection inventory/toolkit, which is a collection of meiosis-specific genes/proteins, as a means of partially circumventing the need to visually or experimentally document sexual reproduction and meiosis. They contend that the presence of intact, conserved meiosis-specific genes is a good indicator of meiosis and a sexual cycle. We used their suite of meiosis-specific proteins to search

for the presence of meiosis in *Chlorella* sp. NC64A. We used stringent criteria to identify orthology between protein sequences, which had to meet all three of the following criteria in order to be considered orthologs to well-documented, sexually reproducing land plant species (*Arabidopsis*, *Oryza*, and *Physcomitrella*): (a) the phylogenetic relationships among the sequences had to match with the known species phylogeny; (b) the algal sequences had to contain the same organization of protein domains as the *A. thaliana* sequence; and, (c) when the algal protein sequence was aligned back against the *A. thaliana* database, the best hit had to be the corresponding *A. thaliana* sequence (i.e., reciprocal best blast hit). We hasten to add that in some cases, an absence of an accession number in table 2 only means that we were unable to confidently identify a particular sequence as an ortholog; in fact, highly conserved homologues were sometimes identified. Our analyses included other species of green algae (*Volvox carteri*, *Ostreococcus lucimarinus*, and *Micromonas* sp. RCC299) and a red alga (*Cyanidioschyzon merolae*) whose sexual status is unknown.

1.6.2 Carbohydrate active enzymes in green algae

The prediction of proteins involved in cell wall metabolism was performed using BLAST searches against a reference database of carbohydrate active enzymes (i.e., CAZy; http://www.cazy.org/), as well as HMM searches against protein families (PFAM) involved in polysaccharide metabolisms. For the BLAST searches, we applied family-specific E-value thresholds defined as follows: the reference carbohydrate active protein sequences of the CAZy database were aligned against each other using BLASTP. For each CAZy family, the E-value threshold was defined as the smallest E-value obtained between a member of the family and any carbohydrate active protein sequence not included in the family. We combined the results of BLAST and HMM searches, and assigned the carbohydrate active protein to protein families based on sequence similarities. This approach was applied to the entire proteomes of NC64A. The proteomes of *C. reinhardtii*, *O. lucimarinus*, *O. tauri, Micromonas* sp. CCMP1515 and *Micromonas* sp. RCC299 were reannotated in the same way for comparison purposes. Missing genes were confirmed by TBLASTN alignment against the genomic sequences using land plant protein sequences as query.

1.7 Phylogenetic analyses

Phylogenetic analyses were mainly performed through the phylogeny.fr web platform (Dereeper et al., 2008). The Phylogeny.fr pipeline was set up as follows: homologous sequences were aligned with the MUSCLE program (Edgar, 2004); poorly aligned positions and divergent regions positions were removed from the multiple-alignment using the GBLOCK program (Castresana, 2000). The cleaned multiple-alignment was then passed on to the PHYML program (Guindon and Gascuel, 2003) for phylogenetic reconstruction using the maximum likelihood criterion. Selection of the best fitting substitution model was performed using the ModelTest program for nucleotide sequences (Posada and Crandall, 1998) and ProtTest for amino acid sequences (Abascal et al., 2005). PhyML was run with the aLRT statistical test of branch support (Anisimova and Gascuel, 2006). This test is based on an approximation of the standard Likelihood Ratio Test, and is much faster to compute than the usual bootstrap procedure while branch supports are generally highly correlated between the two methods.

*2 **Supplemental Results***

2.1 Gene content

We predicted and annotated 9,791 protein genes in the NC64A nuclear genome. This number is comparable to the number of genes in *Micromonas* species that exhibit slightly more than 10,000 predicted genes while having genome sizes that are more than two times smaller (20.9 – 21.9 Mb). *C. reinhardtii* (15,143 genes) and land plants (e.g., *A. thaliana* has 26,341 genes) have more genes while the *Ostreococcus* species to be compared with the numbers of genes for *O. tauri* (8,166) Forty-two percent of the nuclear gene models are supported by 24,072 ESTs. Genes are homogeneously distributed across the NC64A genome with no apparent gene island structures like those found in seed plant genomes such as that of maize (Schnable et al., 2009) (Figure 1). The mean gene density (4.7 Kb/gene) is intermediate between two other sequenced green algae, *O. tauri* (1.5 Kb/gene) and *C. reinhardtii* (8 Kb/gene). Although less than their land-plant cousins, which have between 68% (*Physcomitrella*) and 80% (*Arabidopsis*) gene duplicates (including weakly similar genes detected by alignments at the protein level), substantial gene duplications occur in NC64A, representing 43% of the predicted genes (supplemental figure 4 online). It is remarkable that the number of single copy genes in chlrophyte algae and land plants varies relatively little compared to the number of duplicated genes. Altogether these results indicate that emergence of multicellularity and the colonization of land by plants are correlated with an extensive duplication of genes, many of the corresponding original single copies probably preexisted in their green algal ancestor.

The NC64A protein genes are intron-rich with 7.3 exons per gene on average. This figure is smaller than that of *C. reinhardtii* (8.3 exons/gene) and Human (8.8 exons/gene) but higher than land plant species (e.g., *Arabidopsis thaliana*: 5.2 exons/gene; *Physcomitrella patens*: 5.7 exons/gene; *Oryza sativa:* 4.6 exons/gene). The sequence consensus for intron donor and acceptor sites were similar to those of *Chlamydomonas* (Supplemental figure 5 online). Thirty percent of the intron length in *C. reinhardtii* was accounted for by repeat sequences, suggesting that *Chlamydomonas* introns resulted from either creation or invasion by transposable elements (Merchant et al., 2007). In addition, many introns of *Micromonas* CCMP1545 contained introner repeat elements compared to their intronless orthologs in *Micromonas* RCC299 (Worden et al., 2009). Remarkably, *Chlorella* introns appear much less invaded by repeated sequences (i.e., 5.0% of the intron length on average).

A large fraction of the predicted genes (9,021; 92%) is supported by homology with known genes in public databases (BLASTP E-value >1e-5), the majority of which are most similar to chlorophyte algae or streptophyte plant homologues (i.e. *Viridiplantae* in supplemental figure 6 online). NC64A shares 6,948 protein genes (73%) with *C. reinhardtii* (BLASTP e-value < 1e-5), of which 4,712 (48%) form MBH. The average amino-acid identity between mutual best hits is 52.6%, which is lower than the average amino acid identity between monocot and dicot plant species (e.g., ~60% between grapevine and rice (Jaillon et al., 2007)).

2.2 Plant hormones and receptors

Plant hormones have received attention for some time because in seed plants, including agronomic crops, they control processes involved in growth, development

and response to pathogen infection (Bajguz, 2007) (Siewers et al., 2006) (Callis, 2005). More recently there has been acceleration in the identification of plant hormone receptors, some of which have novel mechanisms of action (Chow and McCourt, 2006; Spartz and Gray, 2008). While most categories of plant hormone molecules have been detected in green algae (Tarakhovskaya et al., 2007), some of which appear to play the same roles as in seed plants (Stirk et al., 2002), little is known of algal hormone biosynthesis (Bajguz, 2009).

We used the KEGG pathway database (Ogata et al., 1999) to obtain protein sequences of *A. thaliana* hormone-pathway enzymes. We then searched for protein homologs in the six algal proteomes by using BLASTP. Functional domains within proteins were identified using the RPS-BLAST algorithm available on the NCBI web site (www.ncbi.nih.nlm.gov). All of the *A. thaliana* and green algal accession numbers are given in table 2. Extensive gene duplication in *Arabidopsis* prevented us from identifying additional clear algal orthologs of enzymes involved in hormone synthesis (e.g., auxin synthesis), as well as hormone-receptors (e.g., brassinosteroid receptors BRL2 and BRL3 and the abscisic acid receptors BRI1 and BRI2). Within each hormone category below, the symbol (1) contains content for the hormone pathways, while the symbol (2) contains content for the hormone receptors. We used the same criteria as for the meiotic protein analysis (see above) for inferring orthology between reference *A. thaliana* sequences and *Chlorella* sp. NC64A proteins.

Abscisic acid (ABA): (1) ABA is synthesized in response to a variety of environmental stresses and is involved in the control of a number of downstream responses essential for adaptation to various stresses that affect plant growth and development (Verslues and Zhu, 2005; Mittler, 2006). The synthesis and role of abscisic acid has long been studied in a number of green algal species (Tominaga et al., 1993; Kobayashi et al., 1997; Bajguz, 2009). Under stress, such as low light, zeaxanthin epoxidase (ABA1; EC: 1.14.13.90; KEGG pathway ath00907) converts zeaxanthin to violaxathin. This biochemical step is reversed by violaxanthin de-epoxidase (NPQ1; EC: 1.10.99.3), resulting in the inactivation of the ABA pathway. In a phylogenetic analysis we used the protein zeaxanthin epoxidase from *A. thaliana* and found orthothogs in all seven algal species. Indeed, *Chlorella* sp. NC64A had a 53% identity over a span of 506 amino acids with those from *A. thaliana*. All seven algal sequences contained the Pyr_recox superfamily domain. In a similar analysis for the enzyme violaxanthin de-epoxidase from *A. thaliana* (AT1G08550), we found orthologs for 5/7 green algal species. (There was no evidence of gene duplication encoding either of these two proteins in *Arabidopsis* or any of the green algae.) *Chlorella* sp. NC64A had a 49% identity over a span of 284 amino acids. The five algal species had the VDE superfamily domain. Beyond this biochemical control step, there are several alternative pathways for the production of ABA (KEGG pathway ath00907). We found evidence of homology to *Arabidopsis* proteins for some of these alternative steps, but we could not conclude with confidence that they were orthologs.

(2) The plastidic abscisic acid receptor GUN5-CHLH (Mg-chelatase H subunit) mediates ABA signalling as a positive regulator in seed germination, post-germination growth and stomatal movement (Shen et al., 2006): The *A. thaliana* sequence for GUN5-CHLH was blasted against the seven algal genomes. The blast search and maximum likelihood phylogenic tree suggested that all algal species have orthologs to the *Arabidopsis* sequence. The *Chlorella* GUN5-CHLH gene overlapped with a sequencing gap in the current genome assembly, which gave a truncated

protein (jgi|ChlNC64A_1|143922). The orthologs to *Arabidopsis* have a very high level of identity (64-68%) and similarity (79-82%) extending over a length of >1,200 amino acids, thus extremely highly conserved. The six algal species shared the CobN superfamily and PRK12493 multi-domains with *A. thaliana*.

Auxin pathway: (1) Green algae clearly have homologs to the enzymes involved in the synthesis of auxins, but the large number of gene duplications in *Arabidopsis thaliana* prevented us from identifying clear orthologs. (2) The auxin binding protein1 (ABP1) is presumed to function as a plasma membrane receptor for auxin (Napier et al., 2002): The ABP1 gene is single copy in the *Arabidopsis thaliana* genome. In contrast *Chlorella* sp. NC64A had two orthologs to the *Arabidopsis* sequence, resulting from a gene duplication event that occurred after the separation between green algae (Chlorophytes) and land plants (Streptophytes); however, no clear orthologs were identified in the other six green algal species. Both NC64A sequences had the expected AUXIN/CUPIN binding domains; as well, they had 49% and 51% identity and 68% similarity to the *A. thaliana* sequence.

Brassinosteriod (BR) pathway: (1) Like abscisic acid, brassinosteroids are plant hormones that influence plant growth and development, particularly in response to abiotic stresses and pathogen infection (Krishna, 2003) and have been documented in green algae (Bajguz and Tretyn, 2003). In fact, Bajguz (2009) found that the addition of exogenous brassinosteroid increased the cellular production of abscisic acid in the green alga *Chlorella vulgaris* in response to short-term heat stress, thereby enhancing thermotolerance; the implication of these results is that microalgae could possibly be cultured for industrial purposes, even at sunlit culture temperatures of 45 ^{o}C. The enzymes for brassinosteriod synthesis can be found in KEGG pathways ath00100 and ath00905. One or more green algae contained orthologs to each of three enzymes involved in the synthesis of brassinosteroids: (a) STE1, (b) DWF5 and (c) DET2. Sterol 1 (STE1): *Chlorella* sp. NC64A and three other algal species had sequences that were clearly orthologous to *A. thaliana* STE1, which converts episterol to 5-dehydroepisterol. They all shared the FA hydrolase superfamily with *A. thaliana* and had 45-53% sequence identity. The two *Ostreococcus* species and *Micromonas* have sequences that were considered possible orthologs because they contained the same FA hydrolase superfamily domain and 32-39% identity to *A. thaliana* STE1, but when they were aligned against the *A. thaliana* database, the best hit was not STE1. DWARF 5 (DWF5): It is clear from phylogenetic analysis that *Micromonas* spp. have orthologs to *A. thaliana* DWF5, which converts 5-dehydroepisterol to 24-methylene cholesterol. As well, they have 59% and 51% sequence identity with *A. thaliana* DWF5, respectively, and they both have the ICMT superfamily domain. Finally, when each of the two sequences was blasted to the *A. thaliana* database, DWF5 was the best hit. No orthologs were found in the genomes of *Chlorella*, *Chlamydomonas*, *Volvox* and the *Ostreococcus* species, even using TBLAST alignment. De-etiolated 2 (DET2): DWF4 (EC:1.14.13.-) and DET2 (EC:1.3.99.-) combine to convert campesterol to 6-deoxocathasterone (and 6-oxocampestanol) through a series of interconnected steps (see KEGG pathway ath00905 for details). There are many gene duplications of DWF4 in *A. thaliana* that prevented us from making a clear determination of orthologs in the green algae, although there is clear homology in some species. Six of seven green algae contained an ortholog to *A. thaliana* DET2. All six algal species shared the steroid_dh superfamily domain with *A. thaliana*. When the six algal sequences were blasted to the *A. thaliana* database, the DET2 sequence was the best hit. (2)

Extensive gene duplication in *Arabidopsis* prevented us from identifying the orthologs to the brassinosteroid-receptors BRL2 and BRL3 for which there were clear homology in some of the green algae.

Cytokinin pathway: (1) Cytokinins are involved in the control of cell division, particularly in cell growth (Tarakhovskaya et al., 2007). Miyawaki et al. (2006) found that tRNA isopentenyltranserases (IPT2 and IPT9) are necessary for the synthesis of cZ-type cytokinin in *Arabidopsis* (designated ATIPT2 and ATIPT9; KEGG pathway ath00908). We found evidence of orthology with ATIPT9 in all green algal species except *C. reinhardtii*. *Chlorella* sp. NC64A (jgi|ChlNC64A_1|55198|) had 37% identity for a length of 405 amino acids with ATIPT9. All six algal species with orthologs contained the IPPT superfamily domain. We did not, however, find orthologs of ATIPT2 in any of the green algae. The IPTs 1, 3, 4, 5, 7 and 8, which do not have an associated tRNA, are also involved in cytokinin synthesis; unfortunately, *Arabidopsis* had several-many gene duplications of these genes making it difficult to determine orthology. (2) Cytokinin binding protein-57 (CBP-57) = homology-dependent gene silencing 1 (HOG1): Orthologs of CBP-57 were found in each of the seven algae. The percent identity over a span of 450 amino acids ranged from 71% identity (*Micromonas*) to 77% identity (*Chlorella* sp. NC64A). All 7 green algal species shared the same superfamily/domains with *A. thaliana*, and their best blast hit to the *A thaliana* database was CBP-57.

Ethylene pathway: (1) While its function is not clear in algae, ethylene plays an important role in stressed tissues and maturing fruit of seed plants by inducing senescence, and it is initiated as a defensive response (Tarakhovskaya et al., 2007). Ethylene production has been documented in marine, freshwater and cultured green algae (Maillard et al., 1993; Osborne et al., 1996; Driessche et al., 1997). In seed plants methionine can be converted to ethylene by the following three-step pathway (KEGG pathway: ath00271): in step 1 L-methionine is converted to S-adenosyl-L-methionine by MAT3 (EC: 2.5.1.6), which in step 2 is converted to 1-aminocyclopropane-1-caroxylate by ACS (EC: 4.4.1.14), which in step 3 is converted to ethylene by EFE (EC: 1.14.17.4). For step 1, we found clear phylogenetic evidence for orthology between the *Arabidopsis* MAT3 protein and the seven green algal species (80% identity between the protein sequences of *A. thaliana* and green algae). As well, all seven algal species shared the same S-AdoMet_synt_N superfamily domains as *A. thaliana*. For step 2, the similarity search and phylogenetic analysis indicated that green algae have clear homologs of the *A. thaliana* ACS proteins. However there are also 10 *Arabidopsis* paralogs of ACS4 that presumably arose after the separation between chlorophyte green algae and land plants, so we were unable to demonstrate orthology. The *Arabidopsis* EFE protein (step3) is a member of a large plant multigene family. We identified several homologs in all chlorophyte algal species, but extensive gene duplication in *Arabidopsis* blurred orthologous relationships between *Arabidopsis* and green algae. (2) *A. thaliana* employs at least five families of ethylene receptors: ETR1, ETR2, ERS1, ERS1 and EIN4. While the seven species of green algae clearly had homologs for some of these receptor families, extensive gene duplication in *Arabidopsis* again prevented us from drawing any strong conclusions of orthology.

Polyamine pathway: (1) Polyamines are involved in a wide variety of cellular activities ranging from regulating growth and development (Tarakhovskaya et al., 2007), to involvement in stress responses, to the modulation of ion channels (Kusano et al., 2008). The polyamines spermidine, spermine and homospermidine are synthesized

by short pathways from arginine or ornithine (Kusano et al., 2008). L-arginine is converted to agmatine by arginine decarboxylase [EC: 4.1.1.19]; agmatine is converted to N-carbamoyl-putrescine by agmatine deiminase [EC: 3.5.3.12]; N-carbamoyl-putricine is converted to putrescine by N-carbamoyl-putrescine amidohydrolase [EC: 3.5.1.53]; putrescine is converted to spermidine by spermidine synthase [EC:2.5.1.16]. *Chlorella* sp. NC64A has the complete toolkit of enzymes to synthesize spermidine from L-arginine. The other algae lack arginine decarboxylase. Putrescine can also be synthesized from L-ornithine by ornithine decarboxylase [EC: 4.1.1.17]. All seven algal species have an ortholog of this enzyme. Finally, it is interesting that *Paramecium bursaria* Chlorella virus-1 (PBCV-1) encodes 4 enzymes involved in polyamine biosynthesis including homospermidine synthase (A237R), which converts spermidine + putricine to homospermidine. Homospermidine synthase has not been found in metazoans, land plants or fungi. The level of similarity indicates that this viral enzyme had a bacterial origin (a few archaea encode this enzyme but there is very low sequence identity with PBCV1). Kaiser et al. (1999) demonstrated that this PBCV-1 enzyme is functional. The only other virus known to encode a putative homospermidine synthase is *Ralstonia* phage RSL1 (NCBI hit with A237R). (2) We were unable to find evidence for polyamine receptors in *Arabidopsis*.

3 Supplemental References

Abascal, F., Zardoya, R., and Posada, D. (2005). ProtTest: selection of best-fit models of protein evolution. Bioinformatics **21,** 2104-2105.

Agarkova, I.V., Dunigan, D.D., and Van Etten, J.L. (2006). Virion-associated restriction endonucleases of chloroviruses. Journal of Virology **80,** 8114-8123.

Anisimova, M., and Gascuel, O. (2006). Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. Syst Biol **55,** 539-552.

Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J., Dehal, P., Christoffels, A., Rash, S., Hoon, S., and Smit, A. (2002). Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. Science **297,** 1301-1310.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., and Eppig, J.T. (2000). Gene Ontology: tool for the unification of biology. Nature genetics **25,** 25-29.

Bajguz, A. (2007). Metabolism of brassinosteroids in plants. Plant Physiol Biochem **45,** 95-107.

Bajguz, A. (2009). Brassinosteroid enhanced the level of abscisic acid in Chlorella vulgaris subjected to short-term heat stress. J Plant Physiol **166,** 882-886.

Bajguz, A., and Tretyn, A. (2003). The chemical characteristic and distribution of brassinosteroids in plants. Phytochemistry **62,** 1027-1046.

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

Callis, J. (2005). Plant biology: auxin action. Nature **435,** 436-437.

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol **17,** 540-552.

Chow, B., and McCourt, P. (2006). Plant hormone receptors: perception is everything. Genes & development **20,** 1998.

Dereeper, A., Guignon, V., Blanc, G., Audic, S., Buffet, S., Chevenet, F., Dufayard, J.F., Guindon, S., Lefort, V., Lescot, M., et al. (2008). Phylogeny.fr: robust phylogenetic analysis for the non-specialist. Nucl. Acids Res. **36,** W465-469.

Detter, J.C., Jett, J.M., Lucas, S.M., Dalin, E., Arellano, A.R., Wang, M., Nelson, J.R., Chapman, J., Lou, Y., and Rokhsar, D. (2002). Isothermal strand-displacement amplification applications for high-throughput genomics. Genomics **80,** 691-698.

Driessche, T., Vries, G., and Guisset, J.L. (1997). Differentiation, growth and morphogenesis: Acetabularia as a model system. New Phytologist **135,** 1-20.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucl. Acids Res. **32,** 1792-1797.

Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J.E., Gavin, O.L., Gunasekaran, P., Ceric, G., Forslund, K., et al. (2010). The Pfam protein families database. Nucl. Acids Res. **38,** D211-222.

Guindon, S., and Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol **52,** 696-704.

Huang, X., and Madan, A. (1999). CAP3: A DNA sequence assembly program. Genome Research **9,** 868-877.

Jaillon, O., Aury, J.M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., Choisne, N., Aubourg, S., Vitulo, N., and Jubin, C. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature **449,** 463-467.

Kaiser, A., Vollmert, M., Tholl, D., Graves, M.V., Gurnon, J.R., Xing, W., Lisec, A.D., Nickerson, K.W., and Van Etten, J.L. (1999). Chlorella virus PBCV-1 encodes a functional homospermidine synthase. Virology **263,** 254-262.

Kobayashi, M., Hirai, N., Kurimura, Y., Ohigashi, H., and Tsuji, Y. (1997). Abscisic acid-dependent algal morphogenesis in the unicellular green alga Haematococcus pluvialis. Plant Growth Regulation **22,** 79-85.

Koonin, E., Fedorova, N., Jackson, J., Jacobs, A., Krylov, D., Makarova, K., Mazumder, R., Mekhedov, S., Nikolskaya, A., and Rao, B. (2004). A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. Genome biology **5,** R7.

Krishna, P. (2003). Brassinosteroid-Mediated Stress Responses. J Plant Growth Regul **22,** 289-297.

Kusano, T., Berberich, T., Tateda, C., and Takahashi, Y. (2008). Polyamines: essential factors for growth and survival. Planta **228,** 367-381.

Maillard, P., Thepenier, C., and Gudin, C. (1993). Determination of an ethylene biosynthesis pathway in the unicellular green alga, Haematococcus pluvialis. Relationship between growth and ethylene production. Journal of Applied Phycology **5,** 93-98.

Malik, S.B., Pightling, A.W., Stefaniak, L.M., Schurko, A.M., and Logsdon Jr, J.M. (2008). An Expanded Inventory of Conserved Meiotic Genes Provides Evidence for Sex in Trichomonas vaginalis. PLoS ONE **3,** e2879.

Merchant, S.S., Prochnik, S.E., Vallon, O., Harris, E.H., Karpowicz, S.J., Witman, G.B., Terry, A., Salamov, A., Fritz-Laylin, L.K., and Marechal-Drouard, L. (2007). The Chlamydomonas Genome Reveals the Evolution of Key Animal and Plant Functions. Science **318,** 245.

Mittler, R. (2006). Abiotic stress, the field environment and stress combination. Trends Plant Sci **11,** 15-19.

Miyawaki, K., Tarkowski, P., Matsumoto-Kitano, M., Kato, T., Sato, S., Tarkowska, D., Tabata, S., Sandberg, G., and Kakimoto, T. (2006). Roles of Arabidopsis

ATP/ADP isopentenyltransferases and tRNA isopentenyltransferases in cytokinin biosynthesis. Proc Natl Acad Sci U S A **103,** 16598-16603.

Napier, R.M., David, K.M., and Perrot-Rechenmann, C. (2002). A short history of auxin-binding proteins. Plant Molecular Biology **49,** 339-348.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. (1999). KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res **27,** 29-34.

Osborne, D.J., Walters, J., Milborrow, B.V., Norville, A., and Stange, L.M.C. (1996). Evidence for a non-ACC ethylene biosynthesis pathway in lower plants. Phytochemistry **42,** 51-60.

Posada, D., and Crandall, K.A. (1998). MODELTEST: testing the model of DNA substitution. Bioinformatics **14,** 817-818.

Salamov, A.A., and Solovyev, V.V. (2000). Ab initio gene finding in Drosophila genomic DNA. Genome Research **10,** 516-522.

Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., and Graves, T.A. (2009). The B73 maize genome: complexity, diversity, and dynamics. Science **326,** 1112.

Schurko, A.M., and Logsdon Jr, J.M. (2008). Using a meiosis detection toolkit to investigate ancient asexual" scandals" and the evolution of sex. BioEssays: news and reviews in molecular, cellular and developmental biology **30,** 579-589.

Shen, Y.-Y., Wang, X.-F., Wu, F.-Q., Du, S.-Y., Cao, Z., Shang, Y., Wang, X.-L., Peng, C.-C., Yu, X.-C., Zhu, S.-Y., et al. (2006). The Mg-chelatase H subunit is an abscisic acid receptor. Nature **443,** 823-826.

Siewers, V., Kokkelink, L., Smedsgaard, J., and Tudzynski, P. (2006). Identification of an abscisic acid gene cluster in the grey mold Botrytis cinerea. Applied and Environmental Microbiology **72,** 4619-4626.

Spartz, A.K., and Gray, W.M. (2008). Plant hormone receptors: new perceptions. Genes Dev **22,** 2139-2148.

Stirk, W.A., Ördög, V., Van Staden, J., and Jäger, K. (2002). Cytokinin-and auxin-like activity in Cyanophyta and microalgae. Journal of Applied Phycology **14,** 215-221.

Tarakhovskaya, E.R., Maslov, Y.I., and Shishova, M.F. (2007). Phytohormones in algae. Russian Journal of Plant Physiology **54,** 163-170.

Ter-Hovhannisyan, V., Lomsadze, A., Chernoff, Y.O., and Borodovsky, M. (2008). Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. Genome Research **18,** 1979.

Tominaga, N., Takahata, M., and Tominaga, H. (1993). Effects of NaCl and KNO3 concentrations on the abscisic acid content of Dunaliella sp.(Chlorophyta). Hydrobiologia **267,** 163-168.

Verslues, P., and Zhu, J. (2005). Before and beyond ABA: upstream sensing and internal signals that determine ABA accumulation and response under abiotic stress. Biochemical Society Transactions **33,** 375-379.

Worden, A.Z., Lee, J.-H., Mock, T., Rouze, P., Simmons, M.P., Aerts, A.L., Allen, A.E., Cuvelier, M.L., Derelle, E., Everett, M.V., et al. (2009). Green Evolution and Dynamic Adaptations Revealed by Genomes of the Marine Picoeukaryotes Micromonas. Science **324,** 268-272.

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