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

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

a Centre National de la Recherche Scientifique, Laboratoire Information Génomique et Structurale UPR2589, Institut de Microbiologie de la Méditerranée IFR88, 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 flaggella-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 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 CO₂ (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 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

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

<table>
<thead>
<tr>
<th>Features</th>
<th>NC64A</th>
<th>C. reinhardtii</th>
<th>Micromonas</th>
<th>CCMP1545</th>
<th>Micromonas RCC299</th>
<th>O. tauri</th>
<th>O. lucimarinus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nuclear genome size (Mb)</td>
<td>46.2</td>
<td>121</td>
<td>21.9</td>
<td>20.9</td>
<td>12.6</td>
<td>13.2</td>
<td></td>
</tr>
<tr>
<td>Number of chromosomes</td>
<td>12a</td>
<td>17</td>
<td>19</td>
<td>17</td>
<td>20</td>
<td>21</td>
<td></td>
</tr>
<tr>
<td>GC content total (%)</td>
<td>67</td>
<td>64</td>
<td>65</td>
<td>64</td>
<td>59</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>Gene count</td>
<td>9,791</td>
<td>15,143</td>
<td>10,575</td>
<td>10,056</td>
<td>7,892</td>
<td>7,651</td>
<td></td>
</tr>
<tr>
<td>Avg. protein length (aa)</td>
<td>456</td>
<td>444</td>
<td>439</td>
<td>473</td>
<td>387</td>
<td>399</td>
<td></td>
</tr>
<tr>
<td>Avg. gene density (kb/gene)</td>
<td>4.7</td>
<td>5.0</td>
<td>2.1</td>
<td>2.2</td>
<td>1.6</td>
<td>1.7</td>
<td></td>
</tr>
<tr>
<td>Avg. number of exons per gene</td>
<td>7.3</td>
<td>8.3</td>
<td>1.9</td>
<td>1.6</td>
<td>1.6</td>
<td>1.3</td>
<td></td>
</tr>
<tr>
<td>Avg. exon length (nt)</td>
<td>170</td>
<td>190</td>
<td>731</td>
<td>958</td>
<td>750</td>
<td>970</td>
<td></td>
</tr>
<tr>
<td>Avg. intron length (nt)</td>
<td>209</td>
<td>373</td>
<td>187</td>
<td>163</td>
<td>126</td>
<td>187</td>
<td></td>
</tr>
<tr>
<td>Avg. coding sequence (%)</td>
<td>29</td>
<td>17</td>
<td>64</td>
<td>68</td>
<td>73</td>
<td>69</td>
<td></td>
</tr>
</tbody>
</table>

aa, amino acids; nt, nucleotides.

*a*Estimation based upon pulse field gel electrophoresis analysis.
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).

**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 archaea. The corresponding genes may have been exchanged by lateral gene transfer between the nuclear genome of land plants

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 α = 0.05 for exon density (P = 0.488).
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, 2008), 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.

**Dynamics of Chlorella Protein Families**

Twenty-eight PFAM protein families showed a biased distribution 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 Mamiellalean 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* (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 (ODA-IC: 2/2), and outer dynein arm light chain proteins (ODA-ALC: 4/5). Putative orthologs were identified in *Micromonas* for each of these proteins (except ODA-DC1), but none were found in the *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

![Figure 4](image-url)  
**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 'f' mark shows organisms known to build motile flagella: Crei, *C. reinhardtii*; NC64A, *Chlorella* sp NC64A; Otau, *O. tauri*; Oluc, *O. lucimarinus*; M. CCP, *M. pusilla* CCMP1545; M. RCC, *Micromonas* sp RCC299.  
(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.
protein complexes called IFT particles. *Chlorella* encodes putative orthologs to the proteins IFT52, IFT57, and IFT88 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, IFT57 and IFT88, 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. reinhardtii* 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*, *Dictostelium*, eu-bacteria, 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; χ² 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 assignment 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 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 nonphotocatalytic 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 evolve those genes and enzymes within the common ancestor of chlorophytes and streptophytes. 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 nonphotocatalytic 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.

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

<table>
<thead>
<tr>
<th>Arabidopsis Enzyme Namea</th>
<th>Chlorella sp NC64Ae</th>
<th>C. reinhardtii</th>
<th>O. tauri</th>
<th>O. lucimarinus</th>
<th>Micromonas sp RC299</th>
<th>Micromonas sp CCMP1545</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abscisic acid pathway</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Abscisic-aldehyde oxidase 58208 (30%)</td>
<td>138368 (37%)</td>
<td>XP_001695565 (32%)</td>
<td>CAL 58065 (42%)</td>
<td>XP_001421564 (41%)</td>
<td>ACO 64017 (44%)</td>
<td>EHE 54518 (43%)</td>
</tr>
<tr>
<td>9-cis-epoxycarotenoid dioxygenase (NCED5) NP_174302</td>
<td>138731 (57%)</td>
<td>XP_001701701 (56%)</td>
<td>CAL 58064 (46%)</td>
<td>XP_001421704 (41%)</td>
<td>ACO 63977 (41%)</td>
<td>EHE 54773 (40%)</td>
</tr>
<tr>
<td>Zeaxanthin epoxidase (ABA1) NP_201504</td>
<td>35609 (42%)</td>
<td>CAL 58064 (46%)</td>
<td>XP_001421704 (41%)</td>
<td>ACO 63977 (41%)</td>
<td>EHE 54773 (40%)</td>
<td></td>
</tr>
<tr>
<td>Zeaxanthin epoxidase (NPQ1) NP_172331</td>
<td>143922 (50%)</td>
<td>XP_001700885 (66%)</td>
<td>CAL 51621 (68%)</td>
<td>XP_001417229 (68%)</td>
<td>ACO 63109 (68%)</td>
<td>EHE 57631 (67%)</td>
</tr>
<tr>
<td>Cytokinin pathway</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Isopentenyl-transferase 9 (ATIPT9) NP_851043</td>
<td>55198 (36%)</td>
<td>CAL 53743 (35%)</td>
<td>XP_001418572 (36%)</td>
<td>ACO 61527 (42%)</td>
<td>EHE 53639 (41%)</td>
<td></td>
</tr>
<tr>
<td>Cytokinlin receptor (HOG) BAH19670</td>
<td>37522 (81%)</td>
<td>XP_001693339 (77%)</td>
<td>CAL 55423 (75%)</td>
<td>XP_001419579 (74%)</td>
<td>ACO 62742 (73%)</td>
<td>EHE 58817 (69%)</td>
</tr>
<tr>
<td>Brassinosteroid pathway</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Iahhosterol oxidase (STE1) NP_186907</td>
<td>37407 (45%)</td>
<td>XP_001701457 (50%)</td>
<td>CAL 5207 (32%)</td>
<td>XP_001416556 (33%)</td>
<td>ACO 69953 (51%)</td>
<td>EHE 53090 (51%)</td>
</tr>
<tr>
<td>7-Hydrocholesterol reductase (SW53) NP_001076993</td>
<td>18410 (37%)</td>
<td>XP_001696675 (34%)</td>
<td>CAL 5207 (32%)</td>
<td>XP_001416556 (33%)</td>
<td>ACO 66602 (34%)</td>
<td>EHE 52455 (34%)</td>
</tr>
<tr>
<td>Steroid reductase (DWF5) NP_181340</td>
<td>18410 (37%)</td>
<td>XP_001696675 (34%)</td>
<td>CAL 5207 (32%)</td>
<td>XP_001416556 (33%)</td>
<td>ACO 66602 (34%)</td>
<td>EHE 52455 (34%)</td>
</tr>
<tr>
<td>Jasmonic acid pathway</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12-Oxophytodienoate reductase (OPR1) NP_177794</td>
<td>52565 (48%)</td>
<td>XP_001699402 (51%)</td>
<td>CAL 53100 (44%)</td>
<td>XP_001417042 (52%)</td>
<td>ACO 65308 (52%)</td>
<td>EHE 60148 (51%)</td>
</tr>
<tr>
<td>3-Hydroxyacyl-CoA dehydrogenase/enoyl-CoA hydratase (MPF2) NP_187324</td>
<td>52565 (54%)</td>
<td>XP_001696661 (45%)</td>
<td>CAL 53100 (44%)</td>
<td>XP_001417042 (52%)</td>
<td>ACO 65308 (52%)</td>
<td>EHE 60148 (51%)</td>
</tr>
<tr>
<td>Polymine (sperridine) pathway</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arg decarboxylase (ADC1) NP_179243</td>
<td>25497 (40%)</td>
<td>CAL 5207 (32%)</td>
<td>XP_001416556 (33%)</td>
<td>ACO 66602 (34%)</td>
<td>EHE 52455 (34%)</td>
<td></td>
</tr>
<tr>
<td>Agmatine iminohydrolase (ATAI) NP_196434</td>
<td>133066 (54%)</td>
<td>XP_001692868 (53%)</td>
<td>CAL 5207 (32%)</td>
<td>XP_001416556 (33%)</td>
<td>ACO 66602 (34%)</td>
<td>EHE 52455 (34%)</td>
</tr>
<tr>
<td>N-Carbamoyl-purinesine amidohydrolase (NLPG1) NP_850101</td>
<td>18182 (57%)</td>
<td>XP_001692986 (53%)</td>
<td>CAL 5207 (32%)</td>
<td>XP_001416556 (33%)</td>
<td>ACO 66602 (34%)</td>
<td>EHE 52455 (34%)</td>
</tr>
<tr>
<td>Spermidine synthase 2 (SPDS2) NP_177188</td>
<td>26108 (53%)</td>
<td>XP_001702843 (56%)</td>
<td>ABO 98745 (56%)</td>
<td>XP_001420452 (58%)</td>
<td>ACO 70332 (55%)</td>
<td>EHE 54321 (56%)</td>
</tr>
<tr>
<td>Orn decarboxylase NP_001083827f</td>
<td>133981 (50%)</td>
<td>XP_001688872 (46%)</td>
<td>CAL 51811 (45%)</td>
<td>XP_001417323 (45%)</td>
<td>ACO 63617 (46%)</td>
<td>EHE 58717 (46%)</td>
</tr>
<tr>
<td>Auxin pathway</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Auxin receptor ABP1 NP_192207</td>
<td>17596 (48%)</td>
<td>26559 (40%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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 used as query in BLAST search.

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

cThis enzyme is not found in Arabidopsis; accession number is for O. sativa.
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 synthase 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, Asp-605, Asp-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 normally not 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 chitosan 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 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 phyology.

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 ± 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 \times 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 SalI- and NotI-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 \(< 1 \times 10^{-15}\)). 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 Lescoat, Dmitry Brogun, Timo Greiner, Ming Kang, Gentry L. Lewis, Suzanne Rose, Eliza Wiech, and Giane M. Yanai-Balsel 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 Génopest, 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


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

Guillaume Blanc, Garry Duncan, Irina Agarkova, Mark Borodovsky, James Gurnon, Alan Kuo, Erika Lindquist, Susan Lucas, Jasmyln Pangilinan, Juergen Polle, Asaf Salamov, Astrid Terry, Takashi Yamada, David D. Dunigan, Igor V. Grigoriev, Jean-Michel Claverie and James L. Van Etten

*Plant Cell*; originally published online September 17, 2010;

DOI 10.1105/tpc.110.076406

This information is current as of July 8, 2017

Supplemental Data  
/content/suppl/2010/09/08/tpc.110.076406.DC1.html  
/content/suppl/2010/09/09/tpc.110.076406.DC2.html

Permissions  

eTOCs  
Sign up for eTOCs at:  
http://www.plantcell.org/cgi/alerts/ctmain

CiteTrack Alerts  
Sign up for CiteTrack Alerts at:  
http://www.plantcell.org/cgi/alerts/ctmain

Subscription Information  
Subscription Information for *The Plant Cell* and *Plant Physiology* is available at:  
http://www.aspb.org/publications/subscriptions.cfm