10-1-2007

The *Chlamydomonas* Genome Reveals the Evolution of Key Animal and Plant Functions

Sabeeha S. Merchant
University of California - Los Angeles

Simon E. Prochnik
Joint Genome Institute

Olivier Vallon
Université Paris 6

Elizabeth H. Harris
Duke University

Steven J. Karpowicz
University of California - Los Angeles

See next page for additional authors

Accepted version. *Science*, Vol. 318, No. 5848 (October 2007): 245-250. DOI. © 2007 American Association for the Advancement of Science. Used with permission.

Authors

Sabeeha S. Merchant, Simon E. Prochnik, Olivier Vallon, Elizabeth H. Harris, Steven J. Karpowicz, George B. Witman, Astrid Terry, Asaf Salamov, Lillian K. Fritz-Laylin, Laurence Maréchal-Drouard, Wallace F. Marshall, Liang-Hu Qu, David R. Nelson, Anton A. Sanderfoot, Martin H. Spalding, Vladimir V. Kapitonov, Qinghu Ren, Patrick Ferris, Erika Lindquist, Harris Shapiro, Susan M. Lucas, Jane Grimwood, Jeremy Schmutz, Pierre Cardol, Heriberto Cerutti, Guillaume Chanfreau, Chun-Long Chen, Valerie Cognat, Martin T. Croft, Rachel Dent, Susan Dutcher, Emilio Fernández, Hideya Fukuzawa, David González-Ballester, Diego González-Halphen, Armin Hallmann, Marc Hanikenne, Michael Hippler, William Inwood, Kamel Jabbari, Ming Kalanon, Richard Kuras, Paul A. Lefebvre, Stéphane D. Lemaire, Alexey V. Lobanov, Martin Lohr, Andrea Manuell, Iris Meier, Laurens Mets, Maria Mittag, Telsa Mittelmeier, James V. Moroney, Jeffrey Moseley, Carolyn Napoli, Aurora M. Nedelcu, Krishna Niyogi, Sergey V. Novoselov, Ian T. Paulsen, Greg Pazour, Saul Purton, Jean-Philippe Ral, Diego Mauricio Riaño-Pachón, Wayne Riekhof, Linda Rymarquis, Michael Schroda, David Stern, James Umen, Robert Willows, Nedra Wilson, Sara Lana Zimmer, Jens Allmer, Janneke Balk, Katerina Bisova, Chong-Jian Chen, Mark Elias, Karla Gendler, Charles Hauser, Mary Rose Lamb, Heidi Ledford, Joanne C. Long, Jun Minagawa, M. Dudley Page, Junmin Pan, Wirulda Pootakham, Sanja Roje, Annkatrin Rose, Eric Stahlberg, Aimee M. Terauchi, Pinfen Yang, Steven Ball, Chris Bowler, Carol L. Dieckmann, Vadim N. Gladyshev, Pamela Green, Richard Jorgensen, Stephen Mayfield, Bernd Mueller-Roeber, Sathish Rajamani, Richard T. Sayre, Peter Brokstein, Inna Dubchak, David Goodstein, Leila Hornick, Y. Wayne Huang, Jinal Jhaveri, Yigong Luo, Diego Martinez, Wing Chi Abby Ngau, Bobby Otillar, Alexander Poliakov, Aaron Porter, Lukasz Szajkowski, Gregory Werner, Kemin Zhou, Igor V. Grigoriev, Daniel S. Rokhsar, and Arthur R. Grossman

The *Chlamydomonas* Genome Reveals the Evolution of Key Animal and Plant Functions

Sabeeha S. Merchant
Department of Chemistry and Biochemistry, University of California Los Angeles
Los Angeles, CA

Simon E. Prochnik
U.S. Department of Energy, Joint Genome Institute
Walnut Creek, CA

Olivier Vallon
CNRS, UMR 7141, CNRS/Université Paris 6, Institut de Biologie Physico-Chimique
Paris, France

Elizabeth H. Harris
Department of Biology, Duke University
Durham, NC

Steven J. Karpowicz
Department of Chemistry and Biochemistry, University of California Los Angeles
Los Angeles, CA

Science, Vol. 318, No. 5848 (October 2007): pg. 245-250. DOI.

George B. Witman

epartment of Cell Biology, University of Massachusetts Medical School
Worcester, MA

Astrid Terry

U.S. Department of Energy, Joint Genome Institute
Walnut Creek, CA

Asaf Salamov

U.S. Department of Energy, Joint Genome Institute
Walnut Creek, CA

Lillian K. Fritz-Laylin

Department of Molecular and Cell Biology, University of California at Berkeley
Berkeley, CA

Laurence Maréchal-Drouard

Institut de Biologie Moléculaire des Plantes, CNRS
Strasbourg Cedex, France

Wallace F. Marshall

Department of Biochemistry and Biophysics, University of California at San Francisco
San Francisco, CA

Liang-Hu Qu

Biotechnology Research Center, Zhongshan University
Guangzhou, China

David R. Nelson

Department of Molecular Sciences and Center of Excellence in Genomics and Bioinformatics, University of Tennessee
Memphis, TN

Anton A. Sanderfoot

Department of Plant Biology, University of Minnesota
St. Paul MN

Martin H. Spalding

Department of Genetics, Development, and Cell Biology, Iowa State University
Ames, IA

Vladimir V. Kapitonov

Genetic Information Research Institute
Mountain View, CA

Qinghu Ren

The Institute for Genomic Research
Rockville, MD

Patrick Ferris

Plant Biology Laboratory, Salk Institute
La Jolla, CA

Erika Lindquist

U.S. Department of Energy, Joint Genome Institute
Walnut Creek, CA

Harris Shapiro

U.S. Department of Energy, Joint Genome Institute
Walnut Creek, CA

Susan M. Lucas

U.S. Department of Energy, Joint Genome Institute
Walnut Creek, CA

Jane Grimwood

Stanford Human Genome Center, Stanford University School of Medicine
Palo Alto, CA

Jeremy Schmutz

Stanford Human Genome Center, Stanford University School of Medicine
Palo Alto, CA

Pierre Cardol

CNRS, UMR 7141, CNRS/Université Paris 6, Institut de Biologie Physico-Chimique
Paris, France
Plant Biology Institute, Department of Life Sciences, University of Liège
Liège, Belgium

Heriberto Cerutti

University of Nebraska-Lincoln, School of Biological Sciences–Plant Science Initiative
Lincoln, NE

Guillaume Chanfreau

Department of Chemistry and Biochemistry, University of California Los Angeles
Los Angeles, CA

Chun-Long Chen

Biotechnology Research Center, Zhongshan University
Guangzhou, China

Valérie Cognat

Institut de Biologie Moléculaire des Plantes, CNRS
Strasbourg Cedex, France

Martin T. Croft

Department of Plant Sciences, University of Cambridge
Cambridge, UK

Rachel Dent

Department of Plant and Microbial Biology, University of California at Berkeley
Berkeley, CA

Susan Dutcher

Department of Genetics, Washington University School of Medicine
St. Louis, MO

Emilio Fernández

Departamento de Bioquímica y Biología Molecular, Facultad de Ciencias, Universidad de Córdoba, Campus de Rabanales
Córdoba, Spain

Patrick Ferris

Plant Biology Laboratory, Salk Institute
La Jolla, CA

Hideya Fukuzawa

Graduate School of Biostudies, Kyoto University
Kyoto, Japan

David González-Ballester

Department of Plant Biology, Carnegie Institution
Stanford, CA

Diego González-Halphen

Departamento de Genética Molecular, Instituto de Fisiología Celular, Universidad Nacional Autónoma de México
México 04510 DF, Mexico

Armin Hallmann

Department of Cellular and Developmental Biology of Plants, University of Bielefeld
D-33615 Bielefeld, Germany

Marc Hanikenne

Plant Biology Institute, Department of Life Sciences, University of Liège
B-4000 Liège, Belgium

Michael Hippler

Department of Biology, Institute of Plant Biochemistry and Biotechnology, University of Münster
48143 Münster, Germany

William Inwood

Department of Plant and Microbial Biology, University of California at Berkeley
Berkeley, CA

Kamel Jabbari

CNRS UMR 8186, Département de Biologie, Ecole Normale Supérieure
75230 Paris, France

Ming Kalanon

Plant Cell Biology Research Centre, The School of Botany, The University of Melbourne, Parkville
Melbourne, VIC 3010, Australia

Richard Kuras

CNRS, UMR 7141, CNRS/Université Paris 6, Institut de Biologie Physico-Chimique
75005 Paris, France

Paul A. Lefebvre

Department of Plant Biology, University of Minnesota
St. Paul MN

Stéphane D. Lemaire

Institut de Biotechnologie des Plantes, UMR 8618, CNRS/Université Paris-Sud
Orsay, France

Alexey V. Lobanov

Department of Biochemistry, N151 Beadle Center, University of Nebraska
Lincoln, NE

Martin Lohr

Institut für Allgemeine Botanik, Johannes Gutenberg-Universität
55099 Mainz, Germany

Andrea Manuell
Department of Cell Biology and Skaggs Institute for Chemical Biology, Scripps Research Institute
La Jolla, CA

Iris Meier
PCMB and Plant Biotechnology Center, Ohio State University
Columbus, OH

Laurens Mets
Molecular Genetics and Cell Biology, University of Chicago
Chicago, IL

Maria Mittag
Institut für Allgemeine Botanik und Pflanzenphysiologie, Friedrich-Schiller-Universität Jena
07743 Jena, Germany

Telsa Mittelmeier
Department of Molecular and Cellular Biology, University of Arizona
Tucson, AZ

James V. Moroney
Department of Biological Science, Louisiana State University
Baton Rouge, LA

Jeffrey Moseley
Department of Plant Biology, Carnegie Institution
Stanford, CA

Carolyn Napoli

Department of Plant Sciences, University of Arizona
Tucson, AZ

Aurora M. Nedelcu

Department of Biology, University of New Brunswick, Fredericton
NB, Canada E3B 6E1

Krishna Niyogi

Department of Plant and Microbial Biology, University of California at Berkeley
Berkeley, CA

Sergey V. Novoselov

Department of Biochemistry, N151 Beadle Center, University of Nebraska
Lincoln, NE

Ian T. Paulsen

The Institute for Genomic Research
Rockville, MD

Greg Pazour

Department of Physiology, University of Massachusetts Medical School
Worcester, MA

Saul Purton

Department of Biology, University College London
London WC1E 6BT, UK

Science, Vol. 318, No. 5848 (October 2007): pg. 245-250. DOI. This article is © American Association for the Advancement of Science and permission has been granted for this version to appear in e-Publications@Marquette. American Association for the Advancement of Science does not grant permission for this article to be further copied/distributed or hosted elsewhere without the express permission from American Association for the Advancement of Science.

Jean-Philippe Ral

Unité de Glycobiologie Structurale et Fonctionnelle, UMR8576 CNRS/USTL, IFR 118, Université des Sciences et Technologies de Lille
Cedex, France

Diego Mauricio Riaño-Pachón

Universität Potsdam, Institut für Biochemie und Biologie
D-14476 Golm, Germany

Wayne Riekhof

Department of Medicine, National Jewish Medical and Research Center
Denver, CO

Linda Rymarquis

Delaware Biotechnology Institute, University of Delaware
Newark, DE

Michael Schroda

Institute of Biology II/Plant Biochemistry
79104 Freiburg, Germany

David Stern

Boyce Thompson Institute for Plant Research at Cornell University
Ithaca, NY

James Umen

Plant Biology Laboratory, Salk Institute
La Jolla, CA

Robert Willows

Department of Chemistry and Biomolecular Sciences, Macquarie University
Sydney 2109, Australia

Nedra Wilson

Department of Anatomy and Cell Biology, Oklahoma State University, Center for Health Sciences
Tulsa, OK

Sara Lana Zimmer

Boyce Thompson Institute for Plant Research at Cornell University
Ithaca, NY

Jens Allmer

Izmir Ekonomi Universitesi
35330 Balcova-Izmir Turkey

Janneke Balk

Department of Plant Sciences, University of Cambridge
Cambridge CB2 3EA, UK

Katerina Bisova

Institute of Microbiology, Czech Academy of Sciences
Czech Republic

Chong-Jian Chen

Biotechnology Research Center, Zhongshan University
Guangzhou 510275, China

Marek Elias

Department of Plant Physiology, Faculty of Sciences, Charles University
128 44 Prague 2, Czech Republic

Karla Gendler

Department of Plant Sciences, University of Arizona
Tucson, AZ

Charles Hauser

Bioinformatics Program, St. Edward's University
Austin, TX

Mary Rose Lamb

Department of Biology, University of Puget Sound
Tacoma, WA

Heidi Ledford

Department of Plant and Microbial Biology, University of California at Berkeley
Berkeley, CA

Joanne C. Long

Department of Chemistry and Biochemistry, University of California Los Angeles
Los Angeles, CA

Jun Minagawa

Institute of Low-Temperature Science, Hokkaido University
060-0819, Japan

M. Dudley Page
Department of Chemistry and Biochemistry, University of California Los Angeles
Los Angeles, CA

Junmin Pan
Department of Biology, Tsinghua University
Beijing, China

Wirulda Pootakham
Department of Plant Biology, Carnegie Institution
Stanford, CA

Sanja Roje
Institute of Biological Chemistry, Washington State University
Pullman, WA

Annkatrin Rose
Appalachian State University
Boone, NC

Eric Stahlberg
PCMB and Plant Biotechnology Center, Ohio State University
Columbus, OH

Aimee M. Terauchi
Department of Chemistry and Biochemistry, University of California Los Angeles
Los Angeles, CA

Pinfen Yang
Department of Biology, Marquette University
Milwaukee, WI

Steven Ball
UMR8576 CNRS, Laboratory of Biological Chemistry
59655 Villeneuve d'Ascq, France

Chris Bowler
CNRS UMR 8186, Département de Biologie, Ecole Normale Supérieure
75230 Paris, France
Cell Signaling Laboratory, Stazione Zoologica
I 80121 Naples, Italy

Carol L. Dieckmann
Department of Molecular and Cellular Biology, University of Arizona
Tucson, AZ

Vadim N. Gladyshev
Department of Biochemistry, N151 Beadle Center, University of Nebraska
Lincoln, NE

Pamela Green
Delaware Biotechnology Institute, University of Delaware
Newark, DE

Richard Jorgensen
Department of Plant Sciences, University of Arizona
Tucson, AZ

Stephen Mayfield
Department of Cell Biology and Skaggs Institute for Chemical Biology, Scripps Research Institute
La Jolla, CA

Bernd Mueller-Roeber

Universität Potsdam, Institut für Biochemie und Biologie
D-14476 Golm, Germany

Sathish Rajamani

Graduate Program in Biophysics, Ohio State University
Columbus, OH

Richard T. Sayre

PCMB and Plant Biotechnology Center, Ohio State University
Columbus, OH

Peter Brokstein

U.S. Department of Energy, Joint Genome Institute
Walnut Creek, CA

Inna Dubchak

U.S. Department of Energy, Joint Genome Institute
Walnut Creek, CA

David Goodstein

U.S. Department of Energy, Joint Genome Institute
Walnut Creek, CA

Leila Hornick

U.S. Department of Energy, Joint Genome Institute
Walnut Creek, CA

Y. Wayne Huang

U.S. Department of Energy, Joint Genome Institute
Walnut Creek, CA

Jinal Jhaveri
U.S. Department of Energy, Joint Genome Institute
Walnut Creek, CA

Yigong Luo
U.S. Department of Energy, Joint Genome Institute
Walnut Creek, CA

Diego Martínez
U.S. Department of Energy, Joint Genome Institute
Walnut Creek, CA

Wing Chi Abby Ngau
U.S. Department of Energy, Joint Genome Institute
Walnut Creek, CA

Bobby Otillar
U.S. Department of Energy, Joint Genome Institute
Walnut Creek, CA

Alexander Poliakov
U.S. Department of Energy, Joint Genome Institute
Walnut Creek, CA

Aaron Porter
U.S. Department of Energy, Joint Genome Institute
Walnut Creek, CA

Lukasz Szajkowski
U.S. Department of Energy, Joint Genome Institute
Walnut Creek, CA

Gregory Werner
*U.S. Department of Energy, Joint Genome Institute
Walnut Creek, CA*

Kemin Zhou
*U.S. Department of Energy, Joint Genome Institute
Walnut Creek, CA*

Igor V. Grigoriev
*U.S. Department of Energy, Joint Genome Institute
Walnut Creek, CA*

Daniel S. Rokhsar
*U.S. Department of Energy, Joint Genome Institute
Walnut Creek, CA
Department of Molecular and Cell Biology, University of
California at Berkeley
Berkeley, CA*

Arthur R. Grossman
*Department of Plant Biology, Carnegie Institution
Stanford, CA*

Abstract:

Chlamydomonas reinhardtii is a unicellular green alga whose lineage diverged from land plants over 1 billion years ago. It is a model system for studying chloroplast-based photosynthesis, as well as the structure, assembly, and function of eukaryotic flagella (cilia), which were inherited from the common ancestor of plants and animals, but lost in land plants. We sequenced the ~120-megabase nuclear genome of *Chlamydomonas* and performed comparative phylogenomic analyses, identifying genes encoding uncharacterized proteins that are likely associated with the function and

biogenesis of chloroplasts or eukaryotic flagella. Analyses of the *Chlamydomonas* genome advance our understanding of the ancestral eukaryotic cell, reveal previously unknown genes associated with photosynthetic and flagellar functions, and establish links between ciliopathy and the composition and function of flagella.

Chlamydomonas reinhardtii is a ~10-μm, unicellular, soil-dwelling green alga with multiple mitochondria, two anterior flagella for motility and mating, and a chloroplast that houses the photosynthetic apparatus and critical metabolic pathways (Fig. 1 and fig. S1) (1). *Chlamydomonas* is used to study eukaryotic photosynthesis because, unlike angiosperms (flowering plants), it grows in the dark on an organic carbon source while maintaining a functional photosynthetic apparatus (2). It also is a model for elucidating eukaryotic flagella and basal body functions and the pathological effects of their dysfunction (3, 4). More recently, *Chlamydomonas* research has been developed for bioremediation purposes and the generation of biofuels (5, 6).

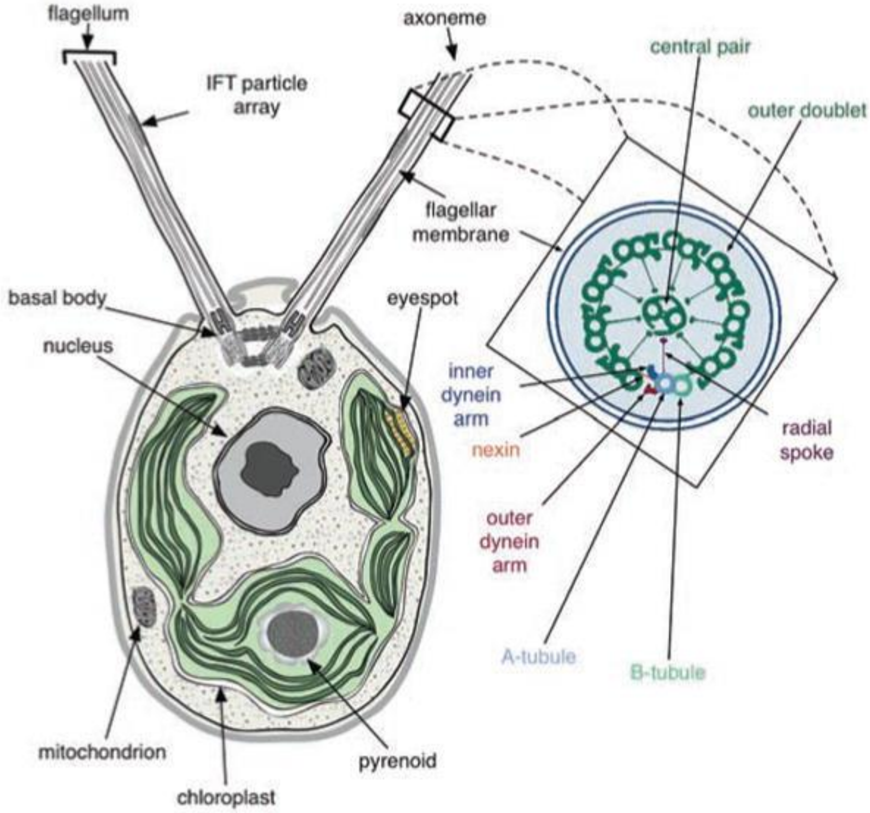


Fig. 1 A schematic of a *Chlamydomonas* cell (from transmission electron micrographs) showing the anterior flagella rooted in basal bodies, with intraflagellar transport (IFT) particle arrays between the axoneme and flagellar membrane, the basal cup-shaped chloroplast, central nucleus and other organelles. An expanded cross section of the flagellar axoneme, as redrawn from (48), shows the nine outer doublets and the central pair (9+2) microtubules; axoneme substructures are color-coded and labeled (see inset).

The Chlorophytes (green algae, including *Chlamydomonas* and *Ostreococcus)* diverged from the Streptophytes (land plants and their close relatives) (Fig. 2) over a billion years ago. These lineages are part of the green plant lineage (Viridiplantae), which previously diverged from opisthokonts (animals, fungi, and Choanozoa) (7). Many *Chlamydomonas* genes can be traced to the green plant or plant-animal common ancestor by comparative genomic analyses. Specifically, many *Chlamydomonas* and angiosperm genes are derived from ancestral green plant genes, including those associated with photosynthesis and plastid function; these are also present in *Ostreococcus* spp. and the moss *Physcomitrella patens* (Fig. 2). Genes shared by *Chlamydomonas* and animals are derived from the last plant-animal common ancestor and many of these have been lost in angiosperms, notably those encoding proteins of the eukaryotic flagellum (or cilium) and the associated basal body (or centriole) (8). *Chlamydomonas* also displays extensive metabolic flexibility under the control of regulatory genes that allow it to inhabit distinct environmental niches and to survive fluctuations in nutrient availability (9).

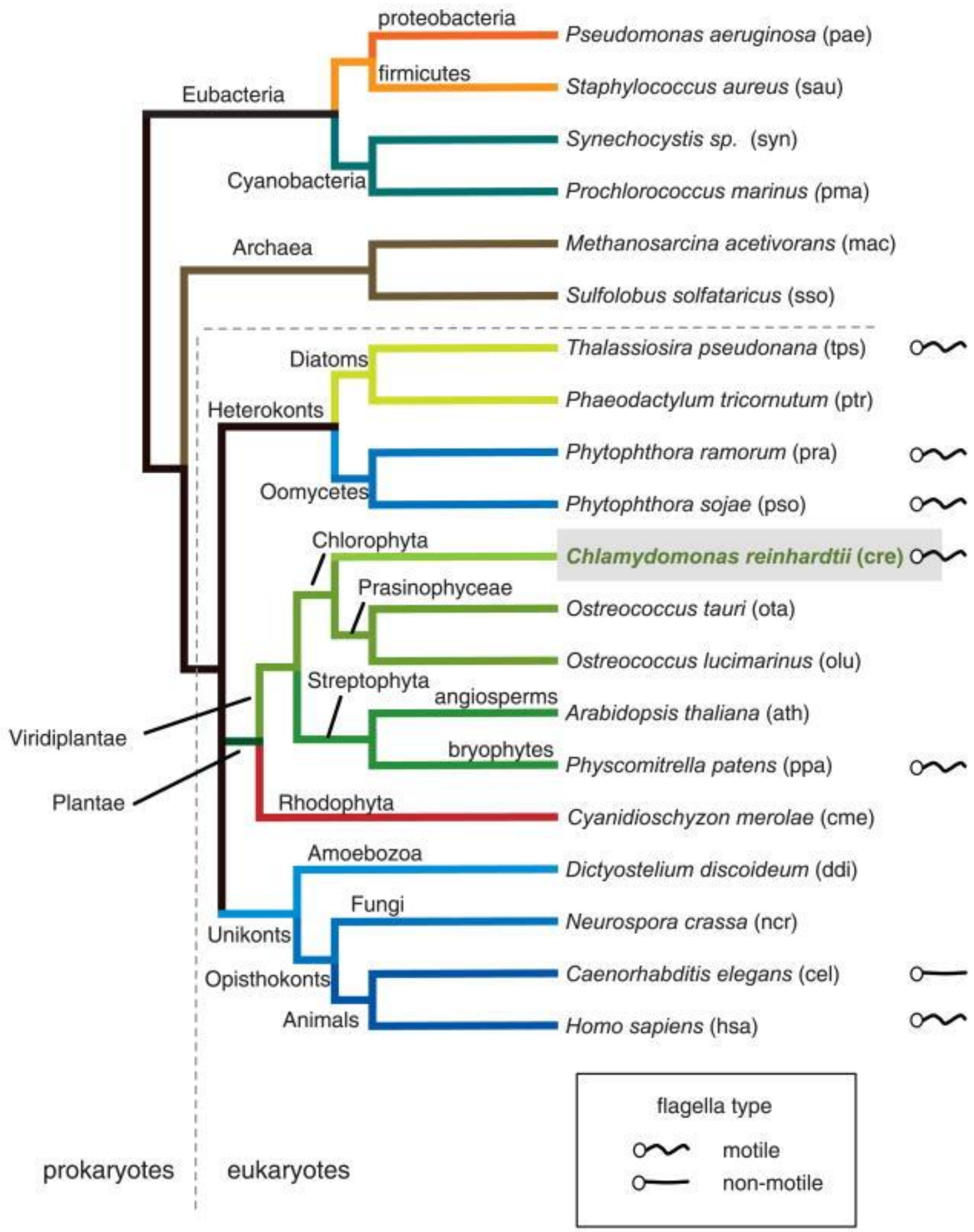
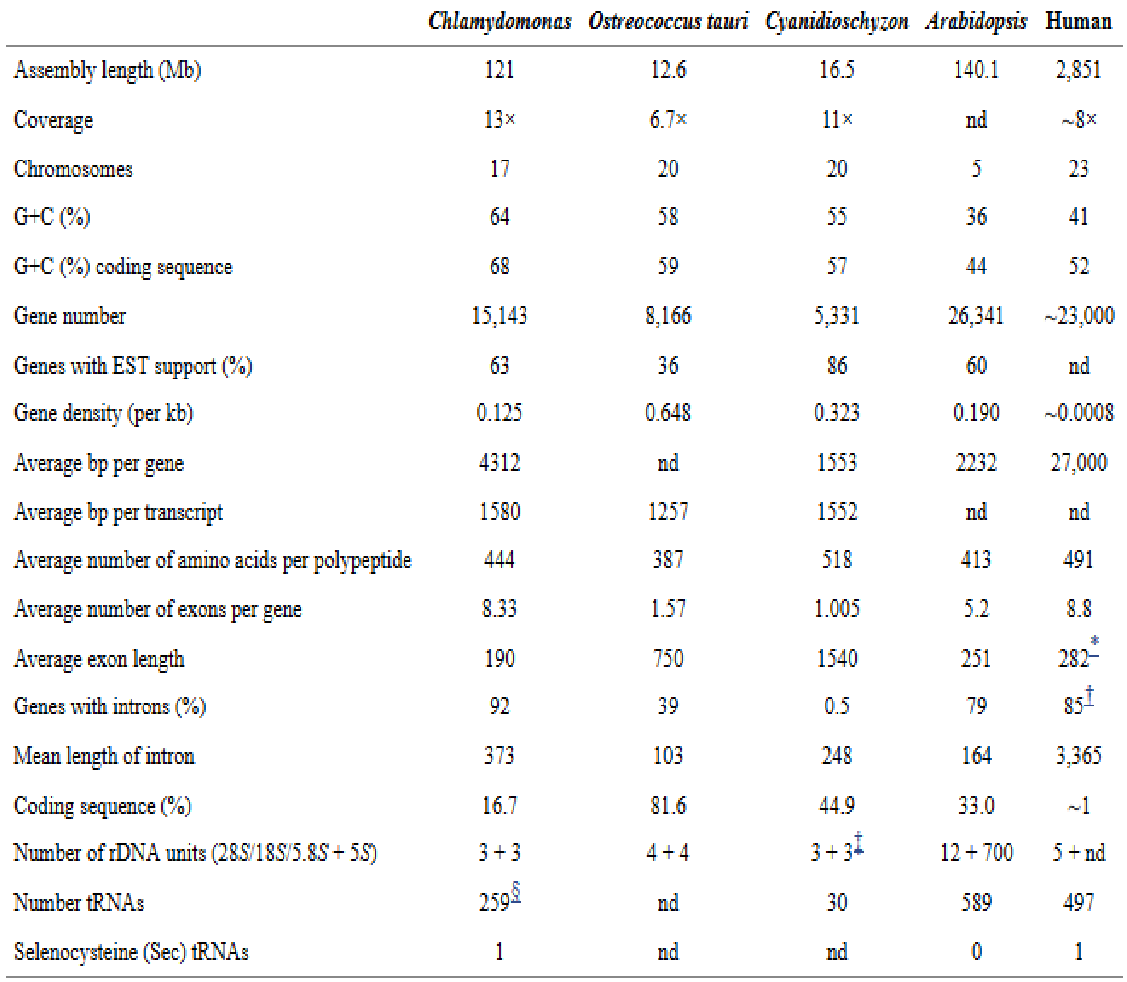

proteobacteria
Pseudomonas aeruginosa (pae)
firmicutes
Staphylococcus aureus (sau)
Eubacteria
Synechocystis sp. (syn)
Cyanobacteria
Prochlorococcus marinus (pma)
Archaea
Methanosarcina acetivorans (mac)
Sulfolobus solfataricus (sso)
Thalassiosira pseudonana (tps)
Diatoms
Phaeodactylum tricornutum (ptr)
Heterokonts
Phytophthora ramorum (pra)
Oomycetes
Phytophthora sojae (pso)
Chlorophyta
Chlamydomonas reinhardtii (cre)
Prasinophyceae
Ostreococcus tauri (ota)
Ostreococcus lucimarinus (olu)
Streptophyta
angiosperms
Arabidopsis thaliana (ath)
Viridiplantae
bryophytes
Physcomitrella patens (ppa)
Plantae
Rhodophyta
Cyanidioschyzon merolae (cme)
Amoebozoa
Dictyostelium discoideum (ddi)
Fungi
Neurospora crassa (ncr)
Unikonts
Opisthokonts
Caenorhabditis elegans (cel)
Animals
Homo sapiens (hsa)
flagella type
motile
non-motile
prokaryotes
eukaryotes

Fig. 2 Evolutionary relationships of 20 species with sequenced genomes (54, 55) used for the comparative analyses in this study include cyanobacteria and nonphotosynthetic eubacteria, Archaea and eukaryotes from the oomycetes, diatoms, rhodophytes, plants, amoebae and opisthokonts. Endosymbiosis of a cyanobacterium by a eukaryotic protist gave rise to the green (green branches) and red (red branches) plant lineages, respectively. The presence of motile or nonmotile flagella is indicated at the right of the cladogram.

Genome sequencing and assembly

The 121-megabase (Mb) draft sequence (10) of the *Chlamydomonas* nuclear genome was generated at 13× coverage by whole-genome, shotgun end-sequencing of plasmid and fosmid libraries, followed by assembly into ~1500 scaffolds (1). Half of the assembled genome is contained in 25 scaffolds, each longer than 1.63 Mb. The genome is unusually GC-rich (64%) (Table 1), which required modification of standard sequencing protocols. Alignments of expressed sequence tags (ESTs) to the genome suggest that the draft assembly is 95% complete (1).



	Chlamydomonas	*Ostreococcus tauri*	*Cyanidioschyzon*	*Arabidopsis*	Human
Assembly length (Mb)	121	12.6	16.5	140.1	2,851
Coverage	13×	6.7×	11×	nd	~8×
Chromosomes	17	20	20	5	23
G+C (%)	64	58	55	36	41
G+C (%) coding sequence	68	59	57	44	52
Gene number	15,143	8,166	5,331	26,341	~23,000
Genes with EST support (%)	63	36	86	60	nd
Gene density (per kb)	0.125	0.648	0.323	0.190	~0.0008
Average bp per gene	4312	nd	1553	2232	27,000
Average bp per transcript	1580	1257	1552	nd	nd
Average number of amino acids per polypeptide	444	387	518	413	491
Average number of exons per gene	8.33	1.57	1.005	5.2	8.8
Average exon length	190	750	1540	251	282[*]
Genes with introns (%)	92	39	0.5	79	85[†]
Mean length of intron	373	103	248	164	3,365
Coding sequence (%)	16.7	81.6	44.9	33.0	~1
Number of rDNA units (28*S*/18*S*/5.8*S* + 5*S*)	3 + 3	4 + 4	3 + 3[‡]	12 + 700	5 + nd
Number tRNAs	259[§]	nd	30	589	497
Selenocysteine (Sec) tRNAs	1	nd	nd	0	1

Table 1 Comparison of *Chlamydomonas* genome statistics to those of selected sequenced genomes. nd, Not determined. [Source for all but *Chlamydomonas* (1)].

Science, Vol. 318, No. 5848 (October 2007): pg. 245-250. DOI.

[*] National Center for Biotechnology Information (NIH) NCBI 36 from Ensembl build 38.

[†] [Source (56)].

[‡] Three regions contain 5*S* rDNA exclusively, and three regions contain 28*S*-18*S*-5.8*S* rDNAs exclusively.

[§] 65 tRNAs that were included in SINE elements were removed from the tRNA-scanSE predictions.

The *Chlamydomonas* nuclear genome comprises 17 linkage groups (figs. S2 to S18) presumably corresponding to 17 chromosomes, consistent with electron microscopy of meiotic synaptonemal complexes (11). Seventy-four scaffolds, representing 78% of the draft genome, have been aligned with linkage groups (Fig. 3 and figs. S2 to S18). Sequenced ESTs from a field isolate (1) of *Chlamydomonas*, fertile with the standard laboratory strain, identified 8775 polymorphisms, resulting in a marker density of 1 per 13 kb (12, 13). By comparing physical marker locations on scaffolds with genetic recombination distances, we estimated 100 kb per centimorgan (cM) on average.

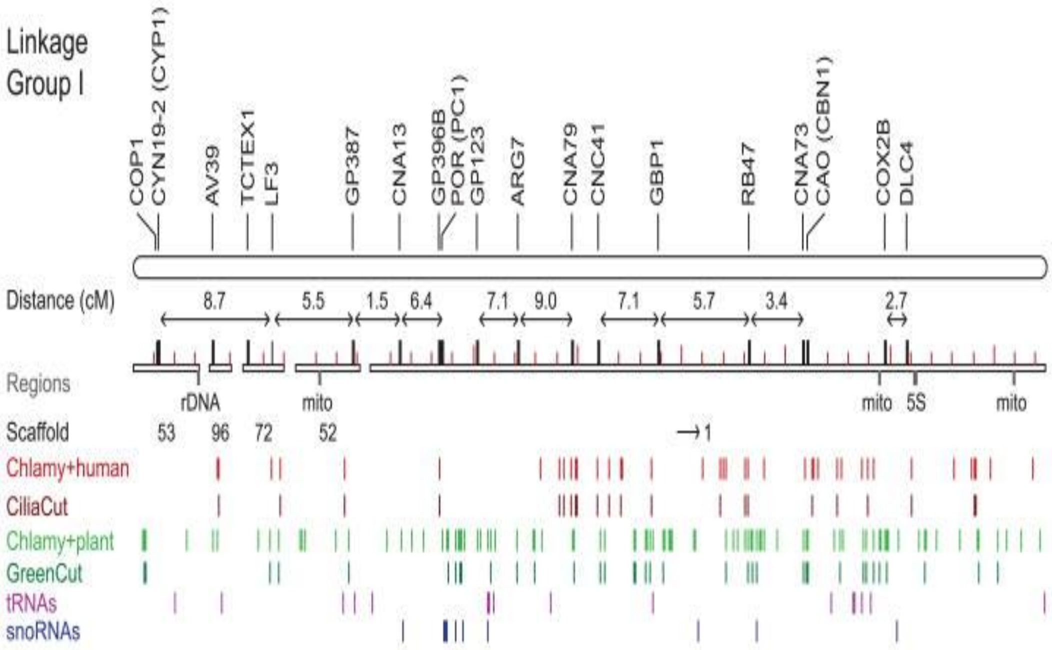


Fig. 3 Linkage group I depicted as a long horizontal rod, with genetically mapped scaffolds shown as open rectangles below (the scaffold number is under each scaffold,

and arrows indicate the orientation of the scaffold where it is known; other scaffolds were placed in their most likely orientation on the basis of genetic map distances. The scale of each map is determined by molecular lengths of the mapped scaffolds. Short and long red ticks are drawn on scaffolds every 0.2 Mb and 1.0 Mb, respectively. We assumed small 50 kb gaps between scaffolds. Genetic distances between markers (centimorgans), where they are known, are shown by two-headed arrows above the scaffold, with the gene symbol and any synonyms in parentheses shown at the top. Genomic regions are labeled below the scaffolds: 5*S*, rDNA, mito (insertion of mitochondrial DNA). *Chlamydomonas* genes with homologs in other organisms/lineages ("Cuts" as defined in the text and Fig. 5) are shown as tracks of vertical bars: light red, genes shared between *Chlamydomonas* and humans, but not occurring in nonciliated organisms; dark red, genes in CiliaCut; light green, genes shared between *Chlamydomonas* and *Arabidopsis*, but not in nonphotosynthetic organisms; dark green, genes in GreenCut; magenta, predicted tRNAs, including those that represent SINE sequences; dark blue, small nucleolar RNAs (snoRNAs).

The *Chlamydomonas* genome has approximately uniform densities of genes, simple sequence repeats, and transposable elements. Several AT-rich islands coincide with gene- and transposable element–poor regions (figs. S2 to S18). As in most eukaryotes, the ribosomal RNA (rRNA) genes are arranged in tandem arrays. They are located on linkage groups I, VII, and XV, although assembly has only been completed on the outermost copies. We identified 259 transfer RNAs (tRNAs) (1) (table S1), 61 classes of simple repeats, ~100 families of transposable elements (1), and 64 tRNA-related short interspersed elements (SINEs) (tables S2 and S3), which is unusual for a microorganism. We also identified tRNAs clusters and a number of recent tRNA duplications (fig. S19), as well as clusters of genes associated with specific biological functions (fig. S20). Few chloroplast and mitochondrial genome fragments were detected in the nuclear genome ("cp" and "mito" in Fig. 3, and figs. S2 to S18).

Protein coding genes and structure

Ab initio and homology-based gene prediction, integrated with EST evidence, was used to create a reference set of 15,143 protein-coding gene predictions (1) (tables S4, S5, and S6). More than 300,000 ESTs were generated from diverse environmental conditions; 8631 gene models (56%) are supported by mRNA or EST evidence (14), and 35% have been edited for gene structure and/or annotated by manual curation, as of June 2007. Protein-coding genes have, on average, 8.3 exons per gene and are intron-rich relative to other unicellular eukaryotes and land plants (15) (fig. S21); only 8% lack introns (Table 1) (1). The average *Chlamydomonas* intron is longer (373 bp) than that of many eukaryotes (16), and the average intron number and size are more similar to those of multicellular organisms than those of protists (fig. S21) (1, 17). Only 1.5% of the introns are short (<100 bp), and we did not observe the bimodal intron size distribution typical of most eukaryotes (fig. S21A). Furthermore, 30% of the intron length is due to repeat sequences (1), which suggests that *Chlamydomonas* introns are subject to creation or invasion by transposable elements.

Gene families

We identified 1226 gene families in *Chlamydomonas* encoding two or more proteins (1); of these, 26 families have 10 or more members (table S7). The genes of 317 of the 798 two-gene families are arranged in tandem, which suggests extensive tandem gene duplications. Gene families contain similar proportions of the total gene complement of *Chlamydomonas*, human, and *Arabidopsis*. As in *Arabidopsis*, *Chlamydomonas* has large families of kinases and cytochrome P-450s, but the largest one is the class III guanylyl and adenylyl cyclase family. With 51 members, the *Chlamydomonas* family is larger than that in any other organism (18). Although these cyclases are not found in plants, in animals they catalyze the synthesis of cGMP and cAMP (18), which serve as second messengers in various signal transduction pathways. Cyclic nucleotides are critical for mating processes, as well as flagellar function and regulation in *Chlamydomonas* (19–21), and may be vital for acclimation to changing nutrient conditions (22, 23). *Chlamydomonas* also encodes diverse families of proteins critical for nutrient acquisition (23, 24).

Transporters

The transporter complement in *Chlamydomonas* suggests that it has retained the diversity present in the common plant-animal ancestor. *Chlamydomonas* is predicted to have 486 membrane transporters (figs. S22 and S23) (1) that fall into the broad classes of 61 ion channels, 124 primary (active) adenosine triphosphate (ATP)–dependent transporters and 293 secondary transporters; eight are unclassified. The 69-member ATP-binding cassette (ABC) and 26-member P-type adenosine triphosphatase (ATPase) families are large, as in *Arabidopsis*, and overall, the complement of transporters in *Chlamydomonas* resembles that of both *Ostreococcus* spp. and land plants (fig. S22). Furthermore, a number of plant transporters not found in animals are encoded on the *Chlamydomonas* genome (fig. S22 and table S8).

We also found copies of genes encoding animal-associated transporter classes, including some with activities related to flagellar function (e.g., the voltage-gated ion channel superfamily) (25) (fig. S22 and table S8). A number of these transporters redistribute intracellular Ca^{2+} in response to environmental signals such as light. Changing Ca^{2+} levels may modulate the activity of the flagella, which are structures found in animals but not in vascular plants (see below).

The *Chlamydomonas* genome also encodes a diversity of substrate-specific transporters that are important for acclimation of the organism to the fluctuating, often nutrient-poor, conditions of soil environments (24). Of the eight sulfate transporters, four are in the H^+/SO_4^{2-} family (characteristic of the plant lineage), three are in the Na^+/SO_4^{2-} family (not found in plants but present in opisthokonts), and one is a bacterial ABC-type SO_4^{2-} transporter (associated with the plastid envelope). The 12-member PiT phosphate transporter and 6-member KUP potassium channel families are larger than in other unicellular eukaryotes, and the former underwent a lineage-specific expansion. *Chlamydomonas* has 11 AMT ammonium transporters, which is only surpassed by the number in rice.

Phylogenomics and the origins of *Chlamydomonas* genes

To explore the evolutionary history of *Chlamydomonas*, we initially compared the *Chlamydomonas* proteome to a representative animal (human) and angiosperm (*Arabidopsis*) proteome (1). We plotted the best matches, calculated on the basis of BLASTP (Basic Local Alignment Search Tool for searching protein collections) scores, of every *Chlamydomonas* protein to the *Arabidopsis* and human proteomes (Fig. 4A). Most *Chlamydomonas* proteins exhibit slightly more similarity to *Arabidopsis* than to human proteins. Many *Chlamydomonas* proteins with greater similarity to animal homologs are present in the flagellar and basal body proteomes (Fig. 4A and below). This is consistent with the maintenance of flagella and basal bodies as cilia and centrioles, respectively, in animals (8), and their loss in angiosperms.

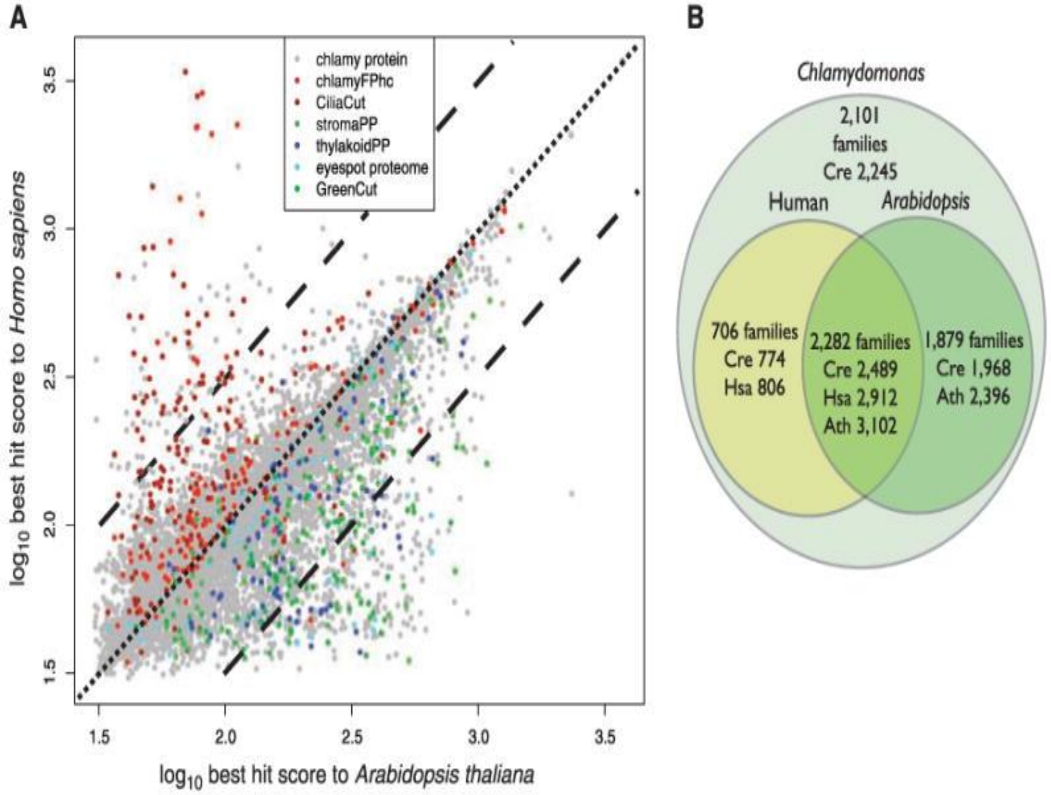


Fig. 4 (**A**) Scatter plot of best BLASTP hit score of *Chlamydomonas* proteins to *Arabidopsis* proteins versus best BLASTP hit score of *Chlamydomonas* proteins to human proteins. Functional or genomic groupings are colored [see inset key in (A)]: *Chlamydomonas* flagellar proteome (42) high confidence set (chlamyFPhc); CiliaCut; *Arabidopsis* stroma plastid proteome (stromaPP); *Arabidopsis* thylakoid plastid proteome (thylakoidPP); eyespot proteome; GreenCut; remaining proteins are gray. (**B**) *Chlamydomonas* protein paralogs were grouped into families together with their homologs from human and *Arabidopsis*. The outer circle represents the proteins in *Chlamydomonas*, 7476 (out of 15,143 total), that fall into 6968 families. Another 7937 proteins cannot be placed in families. Counts of families (and the numbers of proteins from each species in them) with proteins from *Chlamydomonas* and human only, *Chlamydomonas* and *Arabidopsis* only, and *Chlamydomonas* and human and *Arabidopsis*, are shown in the inner circles and the overlap between the two inner circles, respectively. Cre, *Chlamydomonas*; Hsa, human; Ath, *Arabidopsis*.

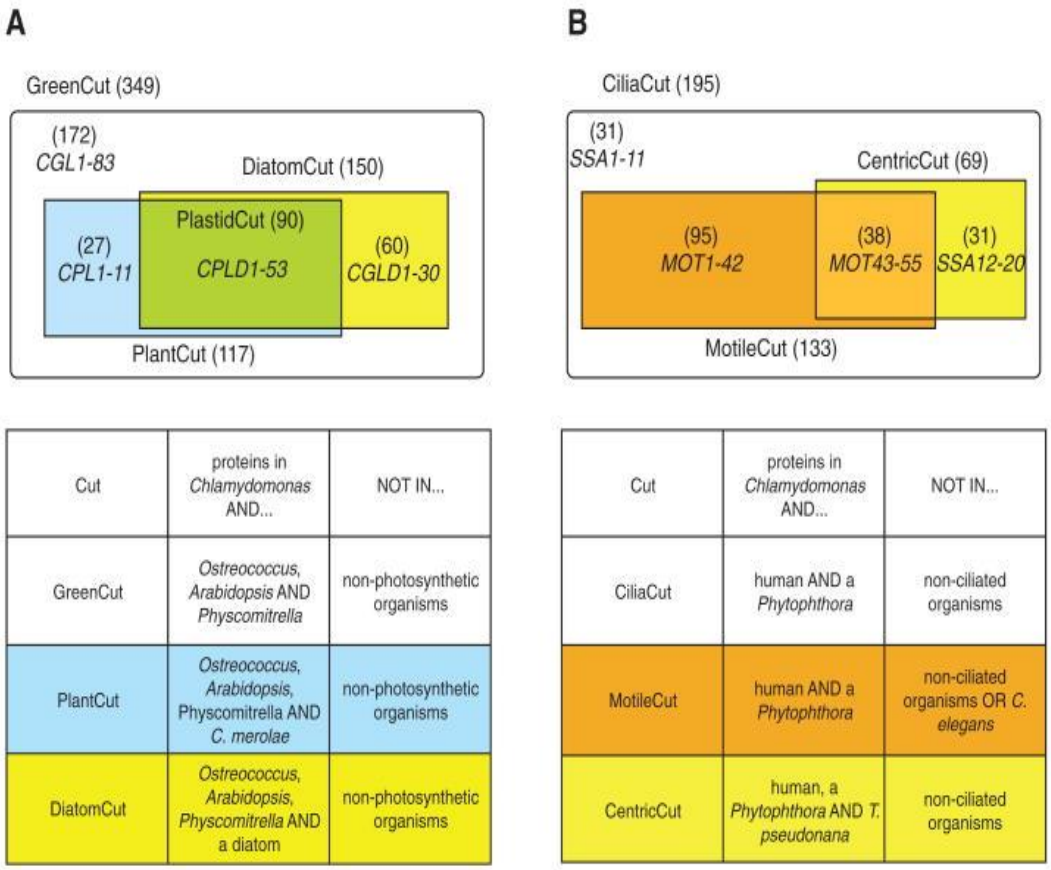
A mutual best-hit analysis of *Chlamydomonas* proteins against proteins from organisms across the tree of life (1) identified 6968 protein families of orthologs, co-orthologs (in the case of recent gene duplications), and paralogs (1). Of the *Chlamydomonas* proteins, 2489 were homologous to proteins from both *Arabidopsis* and humans (Fig. 4B). *Chlamydomonas* and humans shared 706 protein families (774 and 806 proteins, respectively), but these were not shared with *Arabidopsis*. These genes were either lost or diverged beyond recognition in green plants (table S9), and are enriched for sequences encoding cilia and centriole proteins (8, 26). Conversely, 1879 protein families are found in both *Chlamydomonas* and *Arabidopsis* (1968 and 2396 proteins, respectively), but lack human homologs. *Chlamydomonas* proteins with homology to plant, but not animal, proteins were either (i) present in the common plant-animal ancestor and retained in *Chlamydomonas* and angiosperms, but lost or diverged in animals; (ii) horizontally transferred into *Chlamydomonas*; or (iii) arose in the plant lineage after divergence of animals (but before the divergence of *Chlamydomonas*). This set is enriched for proteins that function in chloroplasts (table S9 and below).

The plastid and plant lineages

The plastids of green plants and red algae are primary plastids, i.e., direct descendants from the primary cyanobacterial endosymbiont (27). Diatoms, brown algae, and chlorophyll a– and c–containing algae are also photosynthetic, but their photosynthetic organelles were acquired via a secondary endosymbiosis (28, 29). Because of shared ancestry, nucleus-encoded plastid-localized proteins derived from the cyanobacterial endosymbiont are closely related to each other and to cyanobacterial proteins.

We searched the 6968 families that contain *Chlamydomonas* proteins for those that also contained proteins from *Ostreococcus*, *Arabidopsis* and moss, but that did not contain proteins from nonphotosynthetic organisms. The search identified 349 families, which we named the GreenCut (Fig. 5A, table S10 and table SA); each of these families has a single *Chlamydomonas* protein. On the basis of manual curation of GreenCut proteins of known function (1) (table S11), we estimated ~5 to 8% false-positives and ~14% false-negatives (1). By comparing GreenCut proteins to those of the red alga *Cyanidioschyzon merolae*, which diverged before the split of green algae from land plants (Fig. 2), we identified the subset of proteins present across the plant kingdom; we named this subset the

PlantCut (Fig. 5A, table S10 and table SA). GreenCut protein families that also included representatives from the diatoms *Thalassiosira pseudonana* (30) or *Phaeodactylum tricornutum* (31) were placed in the DiatomCut (Fig. 5A and table S10 and table SA). Given the phylogenetic position of diatoms and their secondary endosymbiosis-derived plastids, we hypothesize that protein families present in both the PlantCut and DiatomCut should contain only those GreenCut proteins associated with plastid function. This subset is referred to as the PlastidCut (Fig. 5A).



Cut	proteins in *Chlamydomonas* AND...	NOT IN...
GreenCut	*Ostreococcus, Arabidopsis* AND *Physcomitrella*	non-photosynthetic organisms
PlantCut	*Ostreococcus, Arabidopsis,* Physcomitrella AND *C. merolae*	non-photosynthetic organisms
DiatomCut	*Ostreococcus, Arabidopsis, Physcomitrella* AND a diatom	non-photosynthetic organisms

Cut	proteins in *Chlamydomonas* AND...	NOT IN...
CiliaCut	human AND a *Phytophthora*	non-ciliated organisms
MotileCut	human AND a *Phytophthora*	non-ciliated organisms OR *C. elegans*
CentricCut	human, a *Phytophthora* AND *T. pseudonana*	non-ciliated organisms

Fig. 5 Summary of genomic comparisons to photosynthetic and ciliated organisms. (**A**) GreenCut: The GreenCut comprises 349 *Chlamydomonas* proteins with homologs in representatives of the green lineage of the Plantae (*Chlamydomonas*, *Physcomitrella*, and *Ostreococcus tauri* and *O. lucimarinus*), but not in nonphotosynthetic organisms. Genes encoding proteins of unknown function that were not previously annotated were given names on the basis of their occurrence in various cuts. CGL refers to conserved only in the green lineage. The GreenCut protein families, which also include members from the red alga *Cyanidioschyzon* within the Plantae, were assigned to the PlantCut (blue plus green rectangles). CPL refers to conserved in

the Plantae. GreenCut proteins also present in at least one diatom (*Thalassiosira* and *Phaeodactylum*) were assigned to the DiatomCut (yellow plus green rectangle). CGLD refers to conserved in the green lineage and diatoms. Proteins present in all of the eukaryotic plastid-containing organisms in this analysis were assigned to the PlastidCut (green rectangle). CPLD refers to conserved in the Plantae and diatoms. The criteria used for the groupings associated with the GreenCut are given in the lower table. (**B**) 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). Unnamed proteins in this group were named MOT (motility). Proteins with homologs in *Caenorhabditis* are associated with nonmotile cilia (white and yellow areas). Proteins in this group that were not already named were named SSA. 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).

The GreenCut contains proteins of the photosynthetic apparatus, including those involved in plastid and thylakoid membrane biogenesis, photosynthetic electron transport, carbon fixation, antioxidant generation, and a range of other primary metabolic processes (table S11 and table SA). Although light-harvesting chlorophyll-binding proteins are poorly represented (1), we identified specialized chlorophyll-binding proteins, as well as a photosynthesis-specific kinase, involved in state transitions. Numerous GreenCut entries are enzymes of plastid-localized metabolic pathways (lipid, amino acid, starch, nucleotide, and pigment biosynthesis) or are unique to plants or highly divergent from animal counterparts. Although tRNA synthetases are conserved between kingdoms, those in the GreenCut represent organellar isoforms that are often targeted to both plastids and mitochondria in plants (32). GreenCut proteins that do not function in the plastids tend to be green lineage–specific or highly diverged from animal counterparts. For example, the *Chlamydomonas* GreenCut protein TOM20 (1), an outer mitochondrial membrane receptor involved in protein import, evolved convergently from a different ancestral protein in plants than in fungi and animals (33).

Of the 214 proteins in the GreenCut without known function, 101 have no motifs or homologies from which function can be inferred, and we can predict only a general function for the others (table S12). Given that 85% of the known proteins in the GreenCut are localized to chloroplasts (table S13), we predict that the set of unknowns contains many novel, conserved proteins that function in chloroplast metabolism and regulation.

The most reducing and oxidizing biological molecules are generated in chloroplasts via the activity of photosystem I and photosystem II, respectively. The flow of electrons through the photosystems causes damage to cellular constituents as a consequence of the accumulation of reactive oxygen species. Therefore, regulation of these molecules is important. Accordingly, plastids house more redox regulators than do mitochondria. Thioredoxins are critical redox-state regulators, and we identified novel thioredoxins in the GreenCut (table S12). These novel thioredoxins have noncanonical active sites or are fused to domains of inferred function (e.g., a vitamin K–binding domain) in plastid metabolism (fig. S1). These findings reveal the potential for identifying unique redox signaling pathways with selectivity and midpoint potentials associated with specific thioredoxin redox sensors (1).

Chlamydomonas has a structure called the eyespot (Fig. 1) which can sense light and trigger phototactic responses. The eyespot is composed of several layers of pigment granules, similar to plastoglobules in plants, and thylakoid membrane, which are directly apposed to the chloroplast envelope and a region of the plasma membrane carrying rhodopsin-family photoreceptors. The pigment granules or plastoglobules contain many proteins with unknown function, many of which are present in the GreenCut, and are likely critical to plastid metabolism; these include SOUL domain, AKC (see below), and PLAP (plastid- and lipid-associated protein) protein families (34–36). SOUL domain proteins of the GreenCut (SOUL4 and SOUL5) have homologs in the *Arabidopsis* plastoglobule proteome (34, 35), and at least one (SOUL3) is associated with the eyespot. The SOUL domain, originally identified in proteins encoded by highly expressed genes in the retina and pineal gland, can bind heme (37, 38). This domain may be important as a heme carrier and/or in maintaining heme in a bound, non-phototoxic form until it associates with proteins or may function in signaling circadian cues.

We also identified plant-specific AKCs (ABC1 kinase in the chloroplast, AKC1 to 4 in the GreenCut), one of which (designated

EYE3) is required for eyespot assembly (39). These AKCs are distinct from the mitochondrial ABC1 kinase that regulates ubiquinone production (40). Protein phosphatases present in the GreenCut and plastoglobules may turn off signaling initiated by the AKCs.

The PLAPs (PLAP1 to 4 in the GreenCut), also called plastoglobulins, are also associated with the eyespot or plastoglobule. These proteins were originally identified by their abundance in carotenoid-rich fibrils and chromoplast plastoglobules and may be structural or organizational components of this plastid subcompartment. Other GreenCut proteins associated with plastoglobules (34, 36) include short-chain dehydrogenases, an aldo-keto isomerase, various methyltransferases with unspecified substrates, esterases and lipases, and a protein with a pantothenate kinase motif.

In sum, the eyespot or plastoglobules contain proteins that likely function in the synthesis, degradation, trafficking, and integration of pigments and lipophilic cofactors into the metabolic machinery of the cell and, most notably, into the photosynthetic apparatus, where they are in high demand. The numerous proteins in the GreenCut associated with the eyespot/plastoglobules may reflect the diverse repertoire of compounds, such as quinones, tocopherols, carotenoids, and tetrapyrroles (fig. S1B), required by photosynthetic organisms.

The 90 proteins in the PlastidCut (Fig. 5A) are likely to function in basic plastid processes because they are conserved in all plastid-containing eukaryotes. Sixty-one of these have unknown functions, with genes for most (except CPLD6 and CPLD29) expressed in chloroplast-containing cells, as assessed from EST representation in *Chlamydomonas* and *Physcomitrella*. For *Arabidopsis* homologs, expression (41) indicates that the genes represented in the PlastidCut tend to be expressed in leaves or all tissue, similar to genes that function in photosynthesis or primary chloroplast metabolism. Greater than 70% of previously unknown PlastidCut proteins have homologs in cyanobacteria, which suggests a critical, conserved, plastid-associated function.

Flagellar and basal body gene complement

Chlamydomonas uses a pair of anterior flagella to swim and sense environmental conditions (Fig. 1). Each flagellum is rooted in a basal body, which also functions as a centriole during cell division. The flagellar axoneme has the nine outer doublet microtubules plus a central pair (9+2) (Fig. 1) characteristic of motile cilia (cilia and eukaryotic flagella are essentially identical organelles). In addition to motile cilia, animals contain nonmotile cilia that function as a sensory organelle and typically lack outer and inner dynein arms, radial spokes, and central microtubules (Fig. 1), all of which are involved in the generation and regulation of motility. Both types of cilia have sensory functions and share conserved sensing and signaling components.

The loss of flagella in angiosperms, most fungi, and slime molds allowed us to identify cilia-specific genes through searches for proteins retained only in flagellate organisms (8, 26). We searched the 6968 *Chlamydomonas* protein families (see above) for those that also contained proteins from human and a *Phytophthora* spp., but not from aciliates, and identified 186 protein families that we named the CiliaCut; these families contain 195 *Chlamydomonas* (Fig. 5B and table SB) and 194 human proteins. One hundred and sixteen of the *Chlamydomonas* proteins had been computationally identified (8, 26), and 45 were identified in this study (1).

The *Chlamydomonas* CiliaCut proteins of unknown function that are missing from *Caenorhabditis*, which has only nonmotile sensory cilia (26), were designated MOT (motile flagella), whereas proteins of unknown function shared with *Caenorhabditis* were designated SSA (sensory, structural and assembly) (Fig. 5B). Thirty-five percent of CiliaCut proteins are in the *Chlamydomonas* flagellar proteome (42), double the number known from previous studies, and 27 of 101 previously identified flagellar proteins (42) are present in the CiliaCut. The CiliaCut contained δ-tubulin, which is required for basal body assembly (43), and a previously undescribed dynein light chain. Some flagellar proteins were not found by this analysis because they have orthologs in plants and fungi, whereas others are absent because they lack human orthologs. Most dynein heavy chains are missing, most likely due to the difficulty of identifying members of large gene families with a mutual best hit approach (1).

We manually curated 125 CiliaCut proteins (fig. S24) and identified large subsets as flagellar structural components (16%), mediating protein-protein interactions (26%), signaling (11%), GTP-binding (6%) and trafficking (6%). These results are consistent with proteomic analysis of the flagellum (42) and highlight the importance of signaling even in motile flagella.

The 62 CiliaCut proteins that *Chlamydomonas* shares with *Caenorhabditis* are predicted to have structural, sensory, or assembly roles in the cilium. As expected, the 133 CiliaCut proteins missing from *Caenorhabditis* (Fig. 5B) (1), designated the MotileCut, include a number of proteins associated with motility (42) (table S14). This data set also contains 31 proteins of unknown function found in the flagellar and basal body proteomes, 36 known but uncharacterized proteins, and 55 novel proteins (designated MOT1 to MOT55); these flagellar proteins are all predicted to be involved specifically in motility.

A comparison of CiliaCut proteins with proteins encoded by the *Physcomitrella* genome indicates that *Physcomitrella* has lost five of the outer dynein arm proteins (Fig. 1, table S14). However, *Physcomitrella* contains inner dynein arm subunits IDA4 and DHC2, as well as subunits of the central microtubules, the radial spokes, and the dynein regulatory complex (table S14). From this we conclude that *Physcomitrella* sperm flagella have a "9+2" axoneme containing inner dynein arms, central microtubules, and radial spokes, but lack the outer dynein arms. Although the structure of the *Physcomitrella* sperm flagellum is not known, sperm flagella of the bryalean moss *Aulacomnium palustre* have just such an axoneme (44).

In contrast, the motile flagella of centric diatoms lack the central pair of microtubules (45, 46). Orthologs of 69 of the 195 CiliaCut proteins (named CentricCut, Fig. 5B) were predicted to be present in the centric diatom *Thalassiosira*. As expected, *Thalassiosira* lacks all central pair proteins. However, it also lacks all radial spoke and inner dynein arm proteins, but has most of the outer dynein arm proteins. The contrasting patterns of loss of axonemal structures predicted for *Physcomitrella* and *Thalassiosira* suggest that the central pair and radial spokes function as a unit with the inner arms, but are dispensable for the generation of motility by the outer arms.

Intraflagellar transport (IFT), which is conserved in ciliated organisms except malaria parasites (47), is essential for flagellar growth (48). The IFT machinery consists of at least 16 proteins in two complexes (A and B) that are moved in anterograde and retrograde directions by the molecular motors kinesin-2 and cytoplasmic dynein 1b, respectively (Fig. 1). Our analysis of *Thalassiosira* reveals that it has components of the anterograde motor and complex B, but has lost the retrograde motor and complex A (table S14). This is intriguing, as retrograde IFT is essential for flagellar maintenance in *Chlamydomonas* (49) and is important for recycling IFT components (50). In addition, both *Physcomitrella* and *Thalassiosira* have lost the Bardet-Biedl syndrome (BBS) genes. BBS gene products are associated with the basal body in *Chlamydomonas* and mammals (8, 51) and sensory cilia in *Caenorhabditis* (52), where they may be involved in IFT (53).

We searched the CiliaCut proteins for proteins shared with *Ostreococcus* spp., a green alga lacking a flagellate stage. The *Ostreococcus* spp. retain 46 (24%) of the 195 CiliaCut proteins but, consistent with loss of the flagellum, are missing genes encoding the IFT-particle proteins and motors, the inner and outer dynein arm proteins, the radial spoke and central pair proteins, and 32 out of 39 flagella-associated proteins (FAPs) (table S14). They have also lost many genes encoding basal body proteins, including all BBS proteins (table S14), which suggests that *Ostreococcus* also lack basal bodies. However, *Ostreococcus* spp. have retained many other CiliaCut proteins (table S14), which suggests either that they recently lost their flagella, or that they retained flagellar proteins for other cellular functions.

Conclusions

This analysis of the *Chlamydomonas* genome sheds light on the nature of the last common ancestor of plants and animals and identifies many cilia- and plastid-related genes. The gene complement also provides insights into life in the soil environment where extreme competition for nutrients likely drove expansion of transporter gene families, as well as sensory flagellar and eyespot functions (e.g., facilitating nutrient acquisition and optimization of the light environment). As more of the ecology and physiology of *Chlamydomonas* and other unicellular algae are explored, additional direct links between gene content and functions associated with the soil life-style will be unmasked with increased potential for biotechnological exploitation of these functions.

Acknowledgments

We thank R. Howson for help with drawing figures, E. Begovic and S. Nicholls for comments on the manuscript. SM is supported by the grants NIH GM42143, DOE DE-FG02-04ER15529 USDA 2004-35318-1495. SP and DSR are funded by USDA and DOE, Joint Genome Institute. ARG is supported by USDA 2003-35100-13235, DOE DE-AC36-99GO10337 and the NSF-funded *Chlamydomonas* Genome Project, MCB 0235878. SJK was supported in part by a Ruth L. Kirschstein National Research Service Award GM07185. The authors declare they have no conflicts of interest. Genome assembly together with predicted gene models and and annotations were deposited at DDBJ/EMBL/GenBank under the project accession ABCN00000000. Since manual curation continues, some models or anotations are changing and the latest set of gene models and annotations is available from www.jgi.doe.gov/chlamy. The most recent set, which includes a number of changes compared with the frozen set used for this analysis, was submitted as the first version, ABCN01000000.

References and Notes

1. Materials and methods and supplemental online (SOM) text are available as supporting material on *Science* Online.

2. Harris EH. Annu Rev Plant Physiol Plant Mol Biol. 2001;52:363. [PubMed]

3. Keller LC, Romijn EP, Zamora I, Yates JR, 3rd, Marshall WF. Curr Biol. 2005;15:1090. [PubMed]

4. Pazour GJ, Agrin N, Walker BL, Witman GB. J Med Genet. 2006;43:62. [PMC free article] [PubMed]

5. Vilchez C, Garbayo I, Markvicheva E, Galvan F, Leon R. Bioresour Technol. 2001;78:55. [PubMed]

6. Ghirardi ML, et al. Annu Rev Plant Biol. 2007;58:71. [PubMed]

7. Yoon HS, Hackett JD, Ciniglia C, Pinto G, Bhattacharya D. Mol Biol Evol. 2004;21:809. [PubMed]

8. Li JB, et al. Cell. 2004;117:541. [PubMed]

9. Grossman AR, et al. Curr Opin Plant Biol. 2007;10:190. [PubMed]

10. Chlamydomonas reinhardtii v 3.0. DOE Joint Genome Institute; www.jgi.doe.gov/chlamy.

11. Storms R, Hastings PJ. Exp Cell Res. 1977;104:39. [PubMed]

12. Kathir P, et al. Eukaryot Cell. 2003;2:362. [PMC free article] [PubMed]

13. Rymarquis LA, Handley JM, Thomas M, Stern DB. Plant Physiol. 2005;137:557. [PMC free article] [PubMed]

14. Jain M, et al. Nucleic Acids Res. 2007;35:2074. [PMC free article] [PubMed]

15. Yuan Q, et al. Plant Physiol. 2005;138:18. [PMC free article] [PubMed]

16. Yandell M, et al. PLoS Comput Biol. 2006;2:e15. [PMC free article] [PubMed]

17. Palenik B, et al. Proc Natl Acad Sci USA. 2007;104:7705. [PMC free article] [PubMed]

18. Schaap P. Front Biosci. 2005;10:1485. [PubMed]

19. Hasegawa E, Hayashi H, Asakura S, Kamiya R. Cell Motil Cytoskeleton. 1987;8:302. [PubMed]

20. Pasquale SM, Goodenough UW. J Cell Biol. 1987;105:2279. [PMC free article] [PubMed]

21. Gaillard AR, Fox LA, Rhea JM, Craige B, Sale WS. Mol Biol Cell. 2006;17:2626. [PMC free article] [PubMed]

22. Gonzalez-Ballester D, de Montaigu A, Higuera JJ, Galvan A, Fernandez E. Plant Physiol. 2005;137:522. [PMC free article] [PubMed]

23. Pollock SV, Pootakham W, Shibagaki N, Moseley JL, Grossman AR. Photosynth Res. 2005;86:475. [PubMed]

24. Grossman A, Takahashi H. Annu Rev Plant Physiol Plant Mol Biol. 2001;52:163. [PubMed]

25. Somlo S, Ehrlich B. Curr Biol. 2001;11(9):R356. [PubMed]

26. Avidor-Reiss T, et al. Cell. 2004;117:527. [PubMed]

27. Gray MW. Curr Opin Genet Dev. 1999;9:678. [PubMed]

28. Bhattacharya D, Yoon HS, Hackett JD. Bioessays. 2004;26:50. [PubMed]

29. Keeling P. Protist. 2004;155:3. [PubMed]

30. Armbrust EV, et al. Science. 2004;306:79. [PubMed]

31. Phaeodactylum tricornutum, v2.0. DOE Joint Genome Institute; www.jgi.doe.gov/phaeodactylum.

32. Duchêne AM, et al. Proc Natl Acad Sci USA. 2005;102:16484. [PMC free article] [PubMed]

33. Perry AJ, Hulett JM, Likic VA, Lithgow T, Gooley PR. Curr Biol. 2006;16:221. [PubMed]

34. Ytterberg AJ, Peltier JB, van Wijk KJ. Plant Physiol. 2006;140:984. [PMC free article] [PubMed]

35. Schmidt M, et al. Plant Cell. 2006;18:1908. [PMC free article] [PubMed]

36. Vidi PA, et al. J Biol Chem. 2006;281:11225. [PubMed]

37. Zylka MJ, Reppert SM. Brain Res Mol Brain Res. 1999;74:175. [PubMed]

38. Sato E, et al. Biochemistry. 2004;43:14189. [PubMed]

39. Lamb MR, Dutcher SK, Worley CK, Dieckmann CL. Genetics. 1999;153:721. [PMC free article] [PubMed]

40. Do TQ, Hsu AY, Jonassen T, Lee PT, Clarke CF. J Biol Chem. 2001;276:18161. [PubMed]

41. Zimmermann P, Hirsch-Hoffmann M, Hennig L, Gruissem W. Plant Physiol. 2004;136:2621. [PMC free article] [PubMed]

42. Pazour GJ, Agrin N, Leszyk J, Witman GB. J Cell Biol. 2005;170:103. [PMC free article] [PubMed]

43. O'Toole ET, Giddings TH, McIntosh JR, Dutcher SK. Mol Biol Cell. 2003;14:2999. [PMC free article] [PubMed]

44. Bernhard DL, Renzaglia KS. Bryologist. 1995;98:52.

45. Manton I, Kowallik K, von Stosch HA. J Cell Sci. 1970;6:131. [PubMed]

46. Heath IB, Darley WM. J Phycol. 1972;18:51.

47. Briggs LJ, Davidge JA, Wickstead B, Ginger ML, Gull K. Curr Biol. 2004;14:R611. [PubMed]

48. Rosenbaum JL, Witman GB. Nat Rev Mol Cell Biol. 2002;3:813. [PubMed]

49. Pazour GJ, Dickert BL, Witman GB. J Cell Biol. 1999;144:473. [PMC free article] [PubMed]

50. Qin H, Diener DR, Geimer S, Cole DG, Rosenbaum JL. J Cell Biol. 2004;164:255. [PMC free article] [PubMed]

51. Ansley SJ, et al. Nature. 2003;425:628. [PubMed]

52. Blacque OE, et al. Genes Dev. 2004;18:1630. [PMC free article] [PubMed]

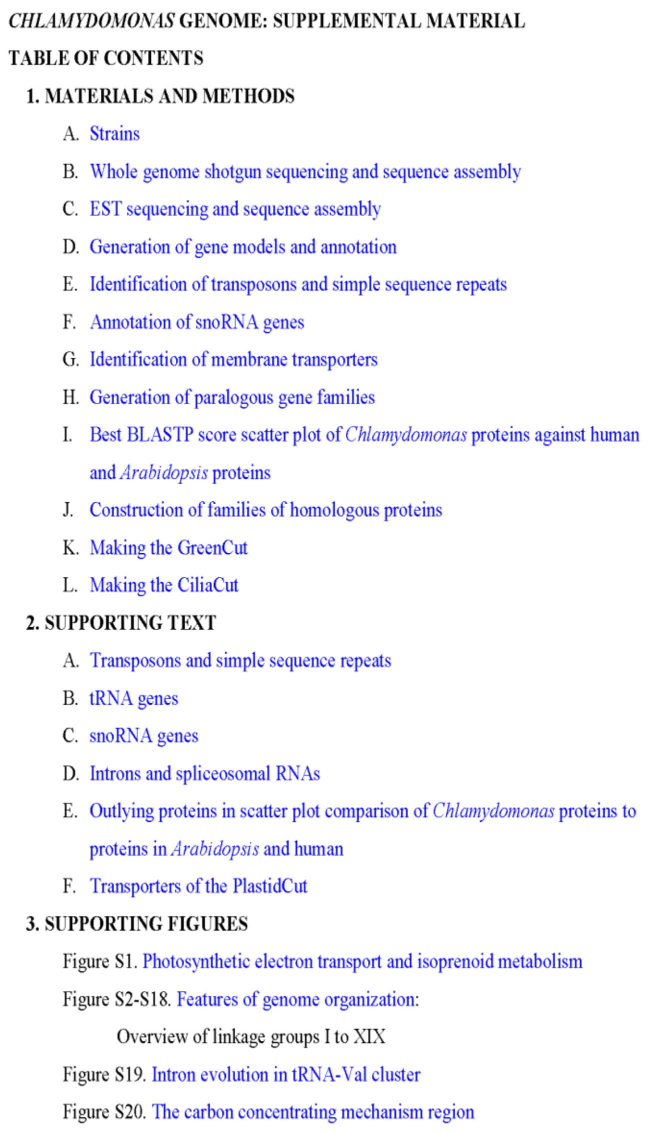
53. Ou G, et al. Mol Biol Cell. 2007;18:1554. [PMC free article] [PubMed]

54. Ciccarelli FD, et al. Science. 2006;311:1283. [PubMed]

55. Keeling PJ, et al. Trends Ecol Evol. 2005;20:670. [PubMed]

56. Eichinger L, et al. Nature. 2005;435:43. [PMC free article] [PubMed]

Supplementary Material



CHLAMYDOMONAS GENOME: SUPPLEMENTAL MATERIAL

TABLE OF CONTENTS

1. MATERIALS AND METHODS

A. Strains: High quality genomic DNA was prepared from strain CC-503 *cw92 mt+*, a cell wall-deficient mutant isolated from strain 137c, which contains *nit1* and *nit2* mutations. A BAC library was prepared from the same strain (*1*). Most of the cDNA libraries were derived from wild-type strain CC-1690 21 gr *mt+* and most of the ESTs were sequenced at Stanford (*2, 3*). Strains CC-503 and CC-1690 were derived from the same original field isolate collected in Massachusetts in 1945, but their parent strains have been cultured separately since the mid-1950s. CC-2290 or S1D2 mt^{-}, which was used for generating some ESTs at the DOE - Joint Genome Institute (JGI) (see below), was collected in the 1980s in Minnesota (*4*). These strains are available from the *Chlamydomonas* Resource Center (*5*). ESTs from the Kazusa DNA Research Institute, Institute of Applied Microbiology, Tokyo, were from strain C-9. This strain also derives from the 1945 field isolate, and is listed in the *Chlamydomonas* Resource Center collection as strain CC-408 (*6*).

B. Whole genome shotgun sequencing and sequence assembly: The initial data set was derived from whole-genome shotgun sequencing (*7*) of 11 libraries supplemented with BAC end sequences. We used nine plasmid libraries, six with an insert size of 2-3 kb, three with an insert size of 6-8 kb and two fosmid libraries with an insert size of 35-40 kb. The reads from the different libraries were as follows: 2,153,471 reads from the 2-3 kb insert libraries comprising 1,683 Mb of raw sequence, 894,846 reads from the 6-8 kb insert libraries comprising 887 Mb of raw sequence, and 184,542 reads from the 35-40 kb insert libraries comprising 184 Mb of raw sequence (including BAC end sequence). The reads were screened for vector sequence with cross_match (*8*) and trimmed for vector and low quality sequences. Reads shorter than 100 bases after trimming were excluded from the assembly. This reduced the data set to 1,903,662 reads from the 2-3 kb insert libraries comprising 807 Mb of raw sequence, 830,326 reads from the 6-8 kb insert libraries comprising 544 Mb of raw sequence, and 153,719 reads from the 35-40 kb insert libraries comprising 49 Mb of raw sequence.

The high GC content of the *Chlamydomonas* genome caused reduced cloning efficiency and premature termination of sequencing reactions, resulting in uneven shotgun sequence coverage across the genome and reduced read lengths. To overcome this difficulty, DMSO (5% final) was added to both the amplification and sequencing reactions. In addition, the RCA Finishing Kit (Amersham Biosciences, Piscataway, NJ) improved amplification of GC-rich sequences and reduced band compression and the formation of secondary structures that resulted in sequencing errors.

The trimmed read sequence data were assembled with release 1.0.3 of Jazz, a whole genome shotgun assembler developed at the DOE Joint Genome Institute (*9*). A word size of 14 was used for seeding alignments between reads, with a minimum of 15 shared words required before an alignment between two reads would be attempted. To reduce the number of collapsed repeats, words present in the sequence data in more than 65 copies were excluded from the set used to seed alignments. A mismatch penalty of −30.0 was used, which generally allows assembly of > 97% identical sequences. The genome size and sequence coverage were estimated to be 130 Mb and 13.0X, respectively. The initial assembly contained 125.5 Mb of scaffold sequence, of which 15.5 Mb (12.4%) represented gaps. There were 7,091 scaffolds, with a scaffold N/L50 of 26/1.63 Mb, and a contig N/L50 of 658/41.7 kb. Short scaffolds (<1 kb length) were removed.

The assembly was next filtered for redundant scaffolds that matched larger scaffolds (<5 kb length where >80% matched a scaffold of >5 kb length). Mitochondrion and chloroplast genome sequences, available prior to the nuclear assembly, were used to identify scaffolds comprising organelle sequence. Finally, scaffolds that showed homology to prokaryotic and non-cellular contaminants [i.e. viroids, viruses, other unclassified, top-level categories at NCBI (*10*)] were identified and removed. After filtering, 121.0 Mb of scaffold sequence remained, of which 15.3 Mb (12.7%) represented gaps.

The filtered assembly (v3.0) contained 1,557 scaffolds, with a scaffold N/L50 of 25/1.63 Mb, and a contig N/L50 of 608/44.5 kb. The sequence coverage was 12.8X ± 0.3X. To estimate the completeness of the assembly, a set of 168,110 ESTs was aligned with BLAT (*11*) to both the entire set of unassembled trimmed reads prior to running through Jazz (pre-assembled) and the assembled sequence; 159,136 ESTs (94.7%) were more than 80% covered by the unassembled data, 160,841 (95.7%) were more than 50% covered and 161,241 (95.9%) were more than 20% covered. By way of comparison, 159,084 ESTs (94.6%) matched the assembled sequence, showing that the assembly covers approximately 95% of the pre-assembled reads.

Whole genome alignment with WU-BLASTN (*12*) of the *Chlamydomonas* v3.0 assembly to the genome sequence of *Ralstonia eutropha* JMP134 (*13*) and *Populus trichocarpa* (*14*) revealed 299 *Chlamydomonas* scaffolds with regions identical to *Ralstonia* or *Populus* genomic sequence. 291 of these scaffolds (each ≤40 kb and assembled from ≤22 sequence reads, and together totaling 1.9 Mb of sequence) were manually removed. A new assembly with the remaining 1,226 scaffolds (assembly v3.1) was generated and is available for download on the JGI *Chlamydomonas* genome browser (*15*).

Of the 74 scaffolds that could be mapped to linkage groups only two show evidence of misassembly (i.e. contain segments that map to two different linkage groups). The approximate positions of the breakpoints are known: the segment of scaffold_6 with coordinates from 1 to ~1.23 Mb maps to LG V and the segment from ~1.44 Mb to 2.94 Mb maps to LG VII; the segment of scaffold_14 from 1 to ~0.874 Mb maps to LG III and the segment ~0.879 Mb to 2.12 Mb maps to LG XVIII.

The Stanford Human Genome Center has been finishing the genome of *Chlamydomonas* since April 2005 with the goal of releasing a finished reference sequence in 2007. The finishing process has been complicated by extreme variations in GC content, sequence hairpins and the presence of many small tandem repeats. Experiments performed to improve the quality of the genome sequence include: resequencing using dGTP chemistry, custom primer walks using a variety of different chemistries and conditions, transposon sequencing and the generation of small insert shatter libraries. In addition, a BAC library (with a mean insert size of 174 kb) provided by Andreas Gnirke from Exelixis (South San Francisco, CA, USA) has been end-sequenced; this library has been used to make further scaffold joins across the genome, reducing the scaffold number (>25 kb) from 168 to 91.

C. EST sequencing and sequence assembly: *E. coli* colonies harboring cDNA clones from *Chlamydomonas* strain S1D2 were plated onto solid agarose medium at a density of approximately 1,000 colonies per plate. The bacteria were grown at 37°C for 18 h and individual colonies were picked robotically and inoculated into LB medium with an appropriate antibiotic in a 384 well plate format. Plasmid DNA was amplified by a rolling circle mechanism (Templiphi, GE Healthcare, Piscataway, NJ) and purified. The insert of each clone was sequenced from both ends with primers complementary to flanking vector sequences (Forward: 5'-ATTTAGGTGACACTATAGAA: Reverse: 5'-TAATACGACTCACTATAGGG) using Big Dye terminator chemistry; the products of the sequencing reactions were resolved by an ABI 3730 sequenator (ABI, Foster City, CA), yielding a total of 34,403 reads. Detailed sequencing protocols can be found in (*16, 17*).

The JGI EST Assembly Pipeline was run on a combined set of 196,594 sequences comprising the 34,403 S1D2 sequences together with ~160,000 sequences from NCBI mRNA and EST databases (*18*) and ~ 2,000 other sequences from various libraries. The pipeline began with the cleanup of 5'and 3' end reads from individual cDNA clones. The Phred program (*8, 19*) was used to call the bases and generate quality scores. Vector, linker, adapter, poly-A/T, and other artifact sequences were removed with the cross_match software, and an internally-developed algorithm that identifies short patterns. Low quality sequence reads were identified with internally-developed software, which masks regions with a combined quality score of less than 15. The longest high quality region of each read was used as an individual EST. ESTs shorter than 150 bases and those representing common contaminants, including *E. coli* genomic sequence, vector sequences, and sequencing standards are removed from the data set. EST clustering was performed ab initio, on the basis of alignments between pairs of trimmed, high quality ESTs. Pairwise EST alignments were generated with the Malign software (*20*), which is a modified version of the Smith-Waterman algorithm (*21*) that has been developed at the JGI for use in whole genome shotgun assembly. ESTs with 150 bp overlaps that align at ≥98% identity were assigned to the same cluster. These were relatively strict clustering cutoffs intended to avoid placing divergent members of gene families into the same cluster. However, this could separate splice variants into different clusters. Optionally, ESTs that do not share alignments were assigned to the same cluster if they were derived from the same cDNA clone. EST cluster consensus sequences were generated by running the Phrap program on the ESTs of each cluster. All alignments generated by Malign are required to extend to within a few bases of the ends of both

ESTs. Therefore, each cluster resembles a 'tiling path' across the gene that matches well with the genome-based assumptions underlying the Phrap algorithm. Additional improvements of the Phrap assemblies were achieved by using the 'forcelevel 4' option, which decreases the chances of generating multiple consensus sequences for a single cluster, where the differences in the consensus sequences may only represent sequencing errors. EST clustering generated 38,869 clusters containing 40,219 consensus sequences.

D. Generation of gene models and annotation: The genome assembly was annotated using the JGI Annotation Pipeline, which combines several gene prediction, annotation and analysis tools. First, the genome assembly was masked using RepeatMasker (*22*) and a custom repeat library (see below). Next, the EST (*3*) and full-length cDNAs were clustered into 32,960 consensus sequences (see above) and aligned to the scaffolds with BLAT (*11*). Model organism protein sequences from the non-redundant (NR) set of proteins from the National Center for Biotechnology Information (Genbank) (*18*) were aligned to the scaffolds with BLASTX (*23*). Gene models and associated transcripts/proteins were predicted or mapped using data from 5,476 putative full-length cDNAs derived from available mRNA, EST and ACEG sequences, and methods such as Genewise (*24*) and *ab initio* approaches such as Fgenesh and Fgenesh+ (*25*). Fgenesh was trained on 495 known genes and reliable homology-based models. The clustered ESTs/cDNAs were used to extend and correct predicted gene models where the exons overlapped and splice junctions were not consistent in comparing EST sequences to gene models. The use of EST information often added 5' and/or 3' UTRs to the models. With gene structure in place, function was assigned to models based on Smith-Waterman (*21*) homology to annotated genes from NR (*18*), KEGG (*26-28*) and KOG (*29*) databases. InterproScan (*30*) was used to identify predicted domains and the Gene Ontology (GO) (*31*) was used to identify function and/or subcellular location. Of the gene models present in the gene catalog (see below), 3,137 models from version 2 of the genome assembly (chlre.v2.0) were mapped forward (Table S4).

Although multiple models with overlapping sequences were generated for each locus, a single model was chosen for the gene catalog set. Model selection was based on maximizing protein sequence relationship and EST support for splice sites, ORFs and model completeness (i.e. inclusion of 5' methionine, 3' stop codon, and UTRs). After a first automatic filtering, the catalog was refined by the annotators, including through generation of *ad hoc* gene models. The catalog was frozen on July 6, 2006, yielding 15,143 gene models, at 14,673 loci ("Frozen Gene Catalog"). All analyses discussed in this paper were carried out on this set. 9,461 (62%) predicted proteins from the Frozen Gene Catalog appear to be full-length, on the basis of the presence of start and stop codons. 4,369 (29%) also have both 5' and 3' UTRs. Furthermore, the majority of predicted genes are supported by EST (56%) or BLASTP (*23*) homology (63%) evidence (Table S5). Of the 6,298 predicted proteins without homology, 30% are *ab initio* fgenesh models with no apparent support and 59% have some support on the basis of EST or distant sequence relationships (E-value > 1E-5). Of the latter group 309 (4.9%) were annotated by users. An analysis based on Smith-Waterman alignments (E-value < 1E-5) (Table S6) yielded 9,435 (62%) gene models with homology to proteins in the COG database (*29, 32*) and/or with Gene Ontology annotations (*31*). Of the predicted gene models 35% have a manually assigned gene function. Furthermore, as of June 2007, 5,141 had been manually-annotated in an attempt to improve the gene set prior to submission to DDBJ/EMBL/GenBank (ABCN01000000). This resulted in an overall decrease in the number of gene models from 15,413 to 14,662. Annotation is on-going and data are available at the JGI genome portal (*15*). Periodic updates will be submitted to DDBJ/EMBL/GenBank (*33*).

E. Identification of transposons and simple sequence repeats: Censor (*34*) was used to identify occurrences of known transposon sequences. These sequences were clustered into families of transposons and retrotransposons and consensus sequences were manually curated. This process identified many new transposon families. The newly identified transposons were annotated and deposited in Repbase (*35*). The genome also contains an extensive range of simple sequence repeats that were identified with Censor (*34*). These have been compiled in a library (similar to the library associated with RepeatMasker).

F. Annotation of snoRNA genes: The snoRNA genes were identified using snoRMP (snoRNA Mining Platform), which is based on the SnoScan (*36*) and SnoGPS (*37*) algorithms, combined with secondary structure prediction and comparative genomic analysis. These approaches predict snoRNA function and have been used successfully for snoRNA gene identification in yeast, plants, mammals and other genomes (*38, 39*).

G. Identification of membrane transporters: To identify membrane-associated transport systems, the complete, predicted proteome was searched against a curated database of transport proteins (*40*) using BLASTP (*23*). All query proteins with significant hits (E-value < 0.001) were collected and searched against the NCBI non-redundant protein and PFAM databases (*41*). Transmembrane protein topology was predicted by TMHMM (*42*) and a web-based interface was implemented to facilitate annotation processes, which incorporate (i) number of hits to the transporter database, (ii) the BLAST and HMM search E-value and score, (iii) the number of predicted transmembrane segments, and (iv) description of top hits to the non-redundant protein database. Detailed transporter profiles and abbreviations for transporter families can be found in (*40, 43*) and at the website TransportDB (*44*). The MPT and IISP transporter families were not included as complete data on these two families in all eukaryotes is not available.

H. Generation of paralogous gene families: We constructed *Chlamydomonas* gene families to investigate both the size and functions of proteins associated with these families. Protein sequences were compared by an all-against-all WU-BLASTP (*12*). The bit score was parsed from the BLAST output and used as the basis for Markov Clustering (MCL) (*45*) with an inflation index of 2.0. PFAM domains were assigned to members of families by RPSBLAST (*23*) (expect score < 1E-10). In the absence of PFAM domain homology, gene families were annotated with InterproScan (*29*). A correlation of >0.5 between nucleotides in the EST and nucleotides in the gene model was taken as evidence for expression of the gene. Sequences from each family were blasted to the NR data base (*18*) to determine homology. For comparison, the same analysis was performed for human, *Arabidopsis*, *Dictyostelium*, *Ostreococcus* spp., and *Neurospora crassa*. *Chlamydomonas* sequences with homology to transposable elements or which contain fragments from transposable elements, exhibit overlapping exonic regions, and do not have support for being expressed are unlikely to represent bonafide *Chlamydomonas* protein-coding genes and were not analyzed further.

In addition to the 51-member type III adenylyl/guanylyl cyclase domain-containing family, there is another family of three proteins with cyclase domains linked to heme NO-binding domains, as well as a pair of cyclases that is in a separate family type. This brings the total number of potential cyclases encoded on the *Chlamydomonas* genome to 56.

I. Best BLASTP score scatter plot of *Chlamydomonas* proteins against human and *Arabidopsis* proteins: The BLASTP scores of every *Chlamydomonas* protein against every human protein and Arabidopsis protein were taken from the BLAST analysis that we performed as part of the construction of homologous protein families (below). A scatter plot was generated with the coordinates of every point determined by the best blast score of the *Chlamydomonas* protein to *Arabidopsis* proteins on the x-axis and to human proteins on the y-axis.

J. Construction of families of homologous proteins: As a pre-requisite to comparing gene content of *Chlamydomonas* to other organisms at the whole-genome scale, we constructed families of homologous proteins from all sequences from *Chlamydomonas* and a wide phylogenetic range of prokaryotic and eukaryotic organisms (Fig. 2). Where several closely-related genome sequences were available, we chose manually- or well-annotated species to represent clades of interest. The shared ancestry (homology) of family members enabled us to infer shared function, allowing functional annotations to be transferred among family members. To create protein families, we first blasted [WU-BLASTP 2.0MP-WashU (20- Apr-2005) (macosx-10.3-g5-ILP32F64 2005-04-21T15:44:27)] (*12*) all protein sequences in *Chlamydomonas* to all protein sequences in the red alga (*Cyanidioschyzon,* strain 10D) (*46*), green algae *Ostreococcus tauri* (assembly v2.0) and *O. lucimarinus* (assembly v2.0) (*47-49*), the land plants *Arabidopsis thaliana* (*50*), and *Physcomitrella patens* (assembly v.1) (*51*), the cyanobacteria *Synechocystis* sp. strain PCC6803 (GenBank Accession: BA000022) and *Prochlorococcus marinus* strain MIT9313 (*52*), bacteria including *Pseudomonas aeruginosa* (strain PA01) (GenBank Accession: AE004091.1) and *Staphylococcus aureus* (subsp. aureus, strain N315) (GenBank Accessions: BA000018.1 AP003139.1), the Archaea *Methanosarcina acetivorans* strain C2A (*53*) and *Sulfolobus solfataricus* strain P2 (*54*), the oomycetes *Phytophthora ramorum* (v1) (*55*) and *P. soja*e (assembly v1) (*56*), the diatoms *Thalassiosira pseudonana* (assembly v3.0) (*57*) and *Phaeodactylum tricornutum* (assembly v2.0) (*58*), the amoeba *Dictyostelium discoideum* (*59, 60*), the fungus *Neurospora crassa* (assembly v7.0; annotation v3.0) (*61*), and the metazoans human (*61-63*) and *Caenorhabditis elegans* (*62*). The blast score of each pair of proteins was extracted and used as a measure of evolutionary distance. Assignment of orthology was determined by mutual best hit between two proteins, using this metric. In creating individual protein families, we first generated all possible ortholog pairs consisting of one *Chlamydomonas* protein and a protein from another organism. Next, paralogs were added to each pair of proteins. A paralog from a given organism was added if its p-dist (defined as 1 – the fraction of identical aligning amino acids in the proteins) was less than a certain fraction of the p-dist between the two orthologs in the pair. The fractions were chosen to be 0.5 for pairs of organisms involving *Chlamydomonas* and a eukaryote and 0.1 for *Chlamydomonas* and a prokaryote. Two considerations led to the choice of these values. In order to assign function correctly, we wanted to include only 'in-paralogs' (paralogs that had duplicated after speciation) (*63*). Secondly, we determined empirically that higher (less stringent) values led to the generation of unwieldy protein families with >22,000 members that could not be analyzed further. In a last step, all pair-wise families of two orthologs plus paralogs were merged if they contained the same *Chlamydomonas* proteins. This created 6,968 families of homologous proteins. Each individual family consists of one or more *Chlamydomonas* paralog(s), mutual best hits to proteins of other species (orthologs) and any paralogs in each of those species. The set of protein families was used in subsequent 'cuts' for analysis of proteins associated with chloroplast or ciliary function (see below). To accomplish this, we built a software tool that allowed us to search for protein families containing any desired combination of species. We call the search results a 'cut' as it represents a phylogenetic slice through the collection of protein families.

The random nature of gene duplication and subsequent divergence and loss that leads to large gene families means that it is sometimes impossible to precisely assign orthology and paralogy between genes. As a result, mutual best hit relationships between sequences may not exist, preventing family construction, or may not be between correct proteins, leading to inclusion of non-homologous proteins in families. This problem was particularly evident in the large family containing the Light Harvesting Complex Proteins (LHCP), for which only two members were included, and the axonemal dynein proteins, for which only two of 14 members in *Chlamydomonas* were included. Furthermore, a cytoplasmic dynein sequence from a diatom was included in the IDA4 inner dynein arm family, probably because the flagella-less diatom is missing genuine inner or outer dynein arms, and its cytoplasmic dynein therefore represents the mutual best hit.

K. Making the 'GreenCut': Having constructed families of homologous proteins, centered on *Chlamydomonas* proteins, we used our search tool (see above) to identify protein families in which all members were present in species in the green lineage of the Plantae, which includes *Chlamydomonas*, the prasinophyte algae *Ostreococcus* spp. (*47*) the angiosperm *Arabidopsis*, and the bryophyte *Physcomitrella* (*50, 51*), but not present in nonphotosynthetic organisms. We refer to this as the 'GreenCut' (Supplemental File 1).

Estimation of false negative frequency: The algorithm was designed to generate a conservative list of proteins, which might result in loss of some proteins that are specific to the green lineage or chloroplast function. We used the components of the photosynthetic apparatus to gauge the effectiveness of the method in recovering proteins expected to be unique to green chloroplasts. Since the cytochrome b_6f complex and the ATP synthase function are also in respiratory membranes in bacteria, we considered only the photosystems, their unique donors and acceptors (plastocyanin, ferredoxin, FNR) and Calvin Cycle enzymes that function only in photosynthetic carbon metabolism (Rubisco and phosphoribulokinase). Using only nucleus-encoded proteins, we generated an "expect inventory" of PsbO, P, Q, R, S, W, X, Y, PsaD, E, F, G, H, K, L, O, plastocyanin, ferredoxin, FNR, RbcS and phosphoribulokinase. Of these 21 proteins, 18 appear in the GreenCut, which gives a potential false negative frequency of ~14%.

Estimate of false positive frequency: There are 135 encoded proteins in the Knowns (K) and Known by Inference (KI) categories. Each of the K and KI proteins was assigned to a subcellular compartment based primarily on annotation of their *Arabidopsis* homologs (TAIR database), but also based on experimental evidence in the literature for *Chlamydomonas* or other photosynthetic organisms (tomato, spinach and tobacco) (**Table S13**). At least 85% (115/135) of the proteins were assigned to the chloroplast, with 9%

(12 out of 135) in other intracellular compartments and the remaining 8 proteins having an undetermined localization. The proteins we regard as false positives are RAD9/At3g05480, ERD2B/At1g19970 (KDEL receptor), SEC12/At5g50550, CYN23b/At1g26940 (ER cyclophilin), CGL28/At1g53650 (RNA binding protein), EFL1/At2g21340 and MER/At3g27730, which represent 5% of the total number of proteins. If CGL22/At2g03670, AMI2/At1g08980, SNE1/At5g28840 and CCD1/At3g63520 are included as false positives (some of these proteins appear to function in processes with plant specific peculiarities), the number increases to 8%. The high percentage of chloroplast localized proteins, as well as proteins that have functions unique to plants, gives an indication of the validity of the method, providing a basis for assessing functions of the unknown proteins. In fact, for one protein, PRMT3403/At3g12270, its presence in a cluster with moss and algae prompted a re-evaluation of the group and an assignment of function as the ribosomal protein arginine methyl transferase, resulting in the movement of the protein from the UP to the KI category. Phylogenetic analysis now places PRMT3403 and At3g12270 together in a green lineage-specific clade.

L. Making the 'CiliaCut': Having made families of putatively homologous proteins (see above), we searched the families for those in which all members were from ciliated organisms; the collection of proteins in these families is designated 'CiliaCut'. To make the CiliaCut, we searched the complete set of homologous protein families for families with members in human, *Chlamydomonas* and at least one *Phytophthora*, but not in the non-ciliated organisms *Arabidopsis*, *Neurospora*, *Cyanidioschyzon*, *Dictyostelium* or eubacteria and archaea. *Phytophthora* are ciliated protists that diverged from animals and plants a relatively short time before animals and plants diverged from each other. Despite this deep divergence, both the core motility machinery and signal transduction pathways are likely to be associated with *Phytophthora* flagella; *Phytophthora* spp. have motile flagellate zoospores that chemotax to their host plants (*64*), implying that their flagella also contain signal transduction components. Therefore, the proteins required for these core pathways should be present in the CiliaCut dataset, and their inclusion adds specificity to the CiliaCut.

There were fourteen *Chlamydomonas* genes in the CiliaCut families that appeared to contain transposons. These were removed from the analyses. The remaining CiliaCut proteins were classified based on the function of characterized orthologous family members, PFAM domain predictions, published information, protein domain searches, and previous comparative genomics (*65, 66*), proteomics (*67, 68*), tissue-specific gene expression studies (*69*), and the ciliome database (*70*).

Estimation of sensitivity and specificity in the 'CiliaCut': There is no simple way to assess how many of the genes in the CiliaCut are genuinely cilia-related and how many of the genuinely cilia-related genes are missing (analagous to the analysis performed for the GreenCut). Nonetheless, we made two attempts to address this issue. First, we compared the CiliaCut proteins to those in the *Chlamydomonas* Flagellar Proteome (chlamyFP) (*67*) and second, we compared the CiliaCut proteins to a curated list of proteins known to be involved in flagellar function.

We assumed that the high confidence proteins from the chlamyFP were very likely to be genuine. 35% (68 out of 195) of CiliaCut proteins are in the chlamyFP high confidence set, whereas only 15% (104 of 687) and 17% (32 of 187) of the proteins in the studies of Li (*66*) and Avidor-Reiss (*65*), respectively, are present in chlamyFP. This represents a greater than two-fold increase in specificity in the CiliaCut relative to previous work, presumably reflecting the inclusion of distantly related flagellate organisms as well as the inclusion of additional information based on the completion of genome sequences.

We also examined the known flagellar proteins identified prior to the generation of chlamyFP. We made a list of 13 randomly-chosen proteins known to be flagella-specific, including only one protein from each protein family; this avoids under-clustering of members of large gene families (see above). One of these genes (tektin) was not present in the CiliaCut, nor is it present in the genomes of 2 species of *Phytophthora.* Presence in at least one *Phytophthora* was required for inclusion in the CiliaCut. Of the remaining 12 proteins, 6 (50%) are in the CiliaCut. Similarly, 44% of the CiliaCut genes are upregulated following deflagellation (*71*) and 58% of these upregulated genes are in CiliaCut. These analyses suggest that the CiliaCut is 50-60% complete.

2. SUPPORTING TEXT

A. Transposons and simple sequence repeats: Known and novel families of transposons were identified and curated (see above). Most remarkable is the presence of SINEs (Tables S2 and S3), small interspersed transposable elements ancestrally related to tRNAs, which rely on LINEs (long interspersed transposable elements) for their propagation. There are 5 families (>200 copies) of SINEs, two of which have precisely kept the tRNA structure and intron position (see section B, immediately below). This is the first example of SINE families described in a unicellular organism.

The repeat landscape of the *Chlamydomonas* genome is dominated by GC-rich, simple sequence runs and transposons, totalling 2.1% and 8.9% of the genomic sequence respectively. The transposons include ~100 families of transposable elements represented by 147 consensus sequences (a unique transposon family is defined as less than 75% identical to transposons in other families). There are also many non-autonomous transposable elements that do not encode proteins. The most thoroughly studied transposon in *Chlamydomonas* is Gulliver (*GUL*) (*72*), whose pattern has been used as a feature of various *Chlamydomonas* field isolates to determine their ancestry. *GUL*, which is present at 14 positions on the genome, is scattered among different scaffolds. Genetic mapping of the *GUL* transposons is consistent with their locations on the physical map.

B. tRNA genes: Most of the 259 *Chlamydomonas* tRNAs (Table S1) are clustered on the genome and appear to result from recent gene duplications (Fig. S19A). The tRNA number in *Chlamydomonas* compares with 390 in *Dictyostelium discoideum*, 272 in *Saccharomyces cerevisiae*, 284 in *Drosophila melanogaster*, 496 in *Homo sapiens*, and 630 in *Arabidopsis thaliana*. However, prediction tools such as tRNAscan-SE (*73*) lead to an inflated number of tRNAs because of the highly conserved tRNA SINE retrotransposon elements (see above). SINE elements have evolved from tRNAs and can be abundant in eukaryotic genomes (*74*). The *Chlamydomonas* genome contains 40 SINEX-3 elements with 5 different anticodons that resemble 34 tRNA-Arg-CCG, 1 tRNA-Arg-ACG, 3 tRNA-Trp-CCA, 1 tRNA-Gly-CCC and 1 tRNA-Gln-CTG (Table S2). There are also 29 tRNA-related SINE elements that resemble 11 tRNA-Asp-ATC and 18 tRNA-Asp-GTC (Table S3). In all cases the SINE and authentic tRNA sequences are highly similar, and all SINE retrotransposon elements have an intron of 11-13 nucleotides between positions 37 and 38 of the tRNA sequence. Furthermore, many SINE-tRNA sequences end with a genome-encoded CCA, which is also present on some authentic *Chlamydomonas* tRNAs (see below). It is possible, as suggested for mammals, that these SINEs are important for transcriptional control, especially related to stress responses (*74, 75*).

There are a number of interesting features associated with *Chlamydomonas* tRNAs. A surprisingly large fraction (60%) of *Chlamydomonas* tRNAs contain introns as compared to human (7%), *Drosophila melanogaster* (5%) and *Saccharomyces cerevisiae* (22%). As in the SINE elements, the introns are located at position 37/38, but the size of the intron is extremely variable, ranging from 8-57 nucleotides. Seven of the tRNAs have the 3' terminal CCA encoded on the genome; a sequence normally added post-transcriptionally, after exonucleolytic trimming of the precursor tRNAs. The presence of a CCA in the genomic tRNA sequence is common in some bacteria and archaea but, to our knowledge, has rarely been described in eukaryotes (*76*). As in bacteria, the *Chlamydomonas* genome encodes RNAse PH and RNAse Z homologs, which in *Bacillus subtilis* are responsible for trimming CCA-containing and CCA-free tRNAs, respectively (*77*).

In some organisms, tRNAs are clustered on the genome. In *Dictyostelium* about 20% of the tRNA genes occur as pairs or triplets separated by 5-20 kb. *Arabidopsis* contains large families of tandemly arrayed tRNA that are on the same DNA strand (*78*). In *Chlamydomonas*, tRNA gene clustering is even more striking, with 160 tRNAs (approximately 60% of the total) associated on the same or opposite DNA strands, and separated by spacers that can be as short as 3-7 nt. As an example of clustered and duplicated tRNAs, we analyzed 12 tRNA-Val genes on scaffold 20 (Fig. S19A); 5 of these have an anticodon AAC and a genome-encoded CCA terminal-sequence while 7 have an anticodon CAC. These genes are grouped in two repeat units contained within a 35 kb genomic region. One of the repeat units contains 3 sets, each with 2 tRNAs; this represents duplications in which the tRNAs have remained within ~2 kb on the genome. The second repeat unit contains 2 sets, each with 3 tRNAs. These tRNA-Val sets are on opposite strands and separated on the genome by ~8.5 kb, but the positions and orientations of the genes within each set are essentially identical. Individual genes from each of the putative gene pairs (genes 7 and 12, 8 and 11, 9 and 10 in Fig. S19A) have anticodons that are identical and introns that are identical, or nearly identical, suggesting a duplication of one entire set. The duplication is likely to have occurred recently on the basis of the near sequence identity between the analogous introns and the neighbor-joining tree made from the intron sequences (Fig. S19B).

C. snoRNA genes: The snoRNA genes are crucial to the biosynthesis of ribosomal RNAs, mediating important steps in folding, site-specific nucleotide modification and precursor cleavage via sequence-specific interactions. The box C/D and box H/ACA snoRNAs guide methylation and conversion of uridine to pseudouridine in their targets, respectively. The *Chlamydomonas* draft genome contains 315 snoRNA genes encoding 124 families, with 71 of the box C/D type and 53 of the box H/ACA type. The box C/D snoRNAs were predicted to guide methylation at 91 sites on rRNAs (31 on 18S, 1 on 5.8S, and 59 on 28S), and 3 sites on U6 snRNAs. Among the 91 rRNA methylation sites, there are 71 analogous sites in other organisms, although 20 are likely *Chlamydomonas* specific. Box H/ACA snoRNAs were predicted to guide pseudouridylation at 63 sites on rRNAs (28 on 18S and 35 on 28S), and 2 sites on U6 snRNA. Among the 63 rRNA pseudouridylation sites, there are 42 analogous sites in other organisms.

About 50% of the *Chlamydomonas* snoRNA genes are present as a single copy on the genome; the rest exist in families of 2 to13 paralogs. Most (71%) snoRNA genes are arranged on the genome in 70 gene clusters, each with 2-6 genes; 52 of these clusters are intron-encoded. Out of the 315 snoRNA genes, 94 were initially predicted to lie between protein-coding genes. After examination of EST and homology data, only 28 were confirmed as intergenic (13 loci). The remaining snoRNAs are found in introns. The polycistronic arrangement of snoRNAs in *Chlamydomonas* is similar to that of rice, although such an arrangement is not observed in vertebrates.

D. Introns and spliceosomal RNAs: Most eukaryotes have a characteristic population of introns with a mode size of between ~60 and 110 nucleotides, although longer introns are common in the human and other large genomes because of repetitive elements embedded in the introns. Surprisingly, the intron size for *Chlamydomonas* gene models, generated as described above, averages 373 nucleotides, which is considerably larger than that of many other eukaryotes (Fig. S21A). Furthermore, the peak intron size in the 60-110 nucleotide range, a feature of the typical bimodal distribution observed for many eukaryotes (Fig. S21A), is missing. These observations are not an annotation artifact as an almost identical peak value for intron length was obtained in the analysis of EST-derived ACEGs.

We calculated the proportion of nucleotides in introns that overlap predicted repeat sequence (see above). 30% of intron sequence consists of repeats, nearly three times the proportion for the whole genome of 11%. This suggests invasion by repeats as a possible mechanism of intron expansion.

Chlamydomonas introns show classical 3' and 5' splice site consensus sequences (CAG^ and G^GTG, respectively), but the classical sequence surrounding the branchpoint (CTNAY) is often difficult to recognize. This suggests that canonical base-pairing between the U2 snRNA and the branchpoint sequence contributes only marginally to the assembly of the spliceosome onto most pre-mRNAs. Similarly, the U1 consensus AAACUUACCU sequence that binds the 5' splice site of introns is not a perfect match to the consensus splice site in *Chlamydomonas* introns (ACG^**GU**GCG).

Altogether, 30 loci were identified that encode the 5 spliceosomal snRNAs. Two of the five U1 genes, four of the six U2 and one of the two U4 genes (all transcribed by Pol II) show EST coverage, with various degrees of truncation at the 5' end. In general, the snRNA-encoding sequences are found within introns of protein coding genes (supported by EST or homology-based analyses). An alternative transcription start gives rise to a transcript extending several hundred base pairs beyond the mature 3' end of the snRNA. The snRNAs are polyadenylated and spliced, using the same canonical exon/intron boundaries as the "host" gene. These observations are consistent with the highly unusual notion that *Chlamydomonas* snRNAs are transcribed as long precursors that are spliced and polyadenylated before maturation. Polyadenylation has been shown for *Dictyostelium* snRNAs (*79*) but splicing of a snRNA precursor has not been described.

E. Outlying proteins in scatter plot comparison of *Chlamydomonas* proteins to proteins in *Arabidopsis* and human: As expected, proteins from the high confidence *Chlamydomonas* Flagellar Proteome (chlamyFP set) (*67*) and CiliaCut (Fig. 4A, red and purple points, respectively) are shifted toward the human axis and conversely, many proteins associated with thylakoid, stroma, eyespot proteomes, and GreenCut (dark blue, green, light blue and dark green points, respectively) lie closer to the *Arabidopsis* axis. Two high confidence chlamyFP points represent proteins with general enzymatic functions and activities that may not be strictly related to flagella function or biogenesis. There is one dark red point outlier from the CiliaCut which closely aligns with a homolog in *Arabidopsis*. There are also two outliers in the thylakoid proteome (Fig. 4A) that are more similar to human than to *Arabidopsis* proteins. In both proteomics sets, the outliers might represent contaminants present in the preparations used to generate the proteomic database.

In analogous analyses, we generated scatter plots of the best blast scores between *Chlamydomonas* proteins and proteins of other photosynthetic organisms (*Arabidopsis*, *Ostreococcus tauri* and *Thalassiosira pseudonana*) (Fig. S25). As expected, these plots show significantly fewer outlying proteins and reveal a closer overall similarity of *Chlamydomonas* proteins to those of *Arabidopsis* than to those of either *O. tauri* or *T. pseudonana*.

F. Transporters of the PlastidCut: Three transporters in the PlastidCut, CPLD21-CPLD23, are predicted to be sugar nucleotide transporters, consistent with the key role of plastids in sugar metabolism. More proteins, including exchangers/carriers that are involved in transporting the substrates and products of plastid metabolism such as phosphate, phosphate-esterified carbon compounds and organic acids, are conserved if we consider only the green lineage. A novel plastid transporter, TIM22B, was also identified in this analysis. This plastid-localized protein has evolved from the expansion of a family of mitochondrial pre-protein translocases (*80*) and is an interesting candidate for functional analysis because it may be involved in the movement of peptide substrates with bound ligands, such as FeS clusters or other minerals that are metabolized in the plastid.

3. SUPPORTING FIGURES

Fig. S1. Photosynthetic electron transport and isoprenoid metabolism: (**A**) 'Z' scheme of photosynthesis, showing photosystems (PS) II and I which are complexes of Psb and Psa polypeptides, respectively, and the cytochrome b_6f complex; Fd, ferredoxin; Trx, thioredoxin; redrawn from (*81*); (**B**) summary of isoprenoid metabolism with enzymes of the pathway mentioned in the text (purple), and end-products (orange); adapted from (*82*). The chloroplast is the site of synthesis of heme, chlorophyll, quinones (phylloquinone, plastoquinones), tocopherols (Vitamin E), and carotenoids, each derived from a common pool of isoprenoid pathway precursors and many having functions in light harvesting, photoprotection (e.g. antioxidants), and as cofactors for electron transfer reactions (*82, 83*). We noted many proteins in the UP categories of the GreenCut are predicted to function in isoprenoid metabolism based on their similarity to known enzymes in these pathways (see Table S12).

Supplemental Figure 1

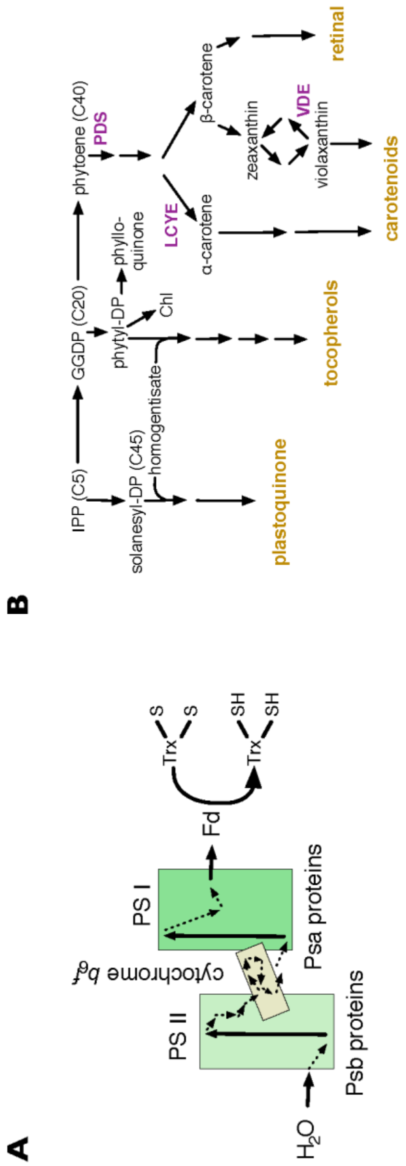


Fig. S2-S18. Features of genome organization: Each Linkage Group is depicted as a long horizontal rod, with genetically-mapped scaffolds shown as open rectangles below (the scaffold number is under each scaffold and arrows indicate orientation where determined; the reverse strand is assumed where orientation is not known). The scale of each map is determined by molecular lengths of the mapped scaffolds. Short and long red ticks are drawn on scaffolds every 0.2 Mb and 1.0 Mb, respectively. We assumed small 50 kb gaps between scaffolds, except where there is genetic evidence of a larger gap (e.g. see Linkage Group X). Genetic distances between markers (cM), where they are known, are shown by two-headed arrows above the scaffold. Genomic regions are labeled below the scaffolds: 5S, rDNA, mito (insertion of mitochondrial DNA), T (telomere), Cp (chloroplast DNA insertion). *Chlamydomonas* genes with homologs in other organisms/lineages ("Cuts" are defined in the text and Fig. 5) are shown as tracks of vertical bars: light red, genes shared between *Chlamydomonas* and humans, but not occurring in non-ciliated organisms; dark red, genes in "CiliaCut"; light green, genes shared between *Chlamydomonas* and *Arabidopsis*, but not in non-photosynthetic organisms; dark green, genes in "GreenCut" ; magenta, predicted tRNAs, including those that represent SINE sequences; dark blue, snoRNAs. Below, on separate axes, are features of the genomic sequence (in 25 kb windows): %GC (grey), gene density (red), transposable element (TE) density (blue), and simple repeat (Rep) density (teal). The %GC graph includes horizontal lines denoting 25, 50 and 75% GC. The other three graphs show a mean (solid horizontal line) and +/– SD (dashed horizontal line) for the scaffold, and are scaled to the densest region on any of the mapped scaffolds, which are as follows: gene density, 12 per 25 kb window; TE density, 44 per 25 kb window; repeat density, 46 per 25 kb window.

Fig S2. Overview of linkage group I

Fig S3. Overview of linkage group II.

Fig. S4. Overview of linkage group III.

Fig. S5. Overview of linkage group IV.

Fig. S6. Overview of linkage group V.

Fig. S7. Overview of linkage group VI.

Fig. S8. Overview of linkage group VII.

Fig. S9. Overview of linkage group VIII.

Fig. S10. Overview of linkage group IX.

Fig. S11. Overview of linkage group X.

Fig. S12. Overview of linkage group XI.

Fig. S13. Overview of linkage group XII+XIII.

Fig. S14. Overview of linkage group XIV.

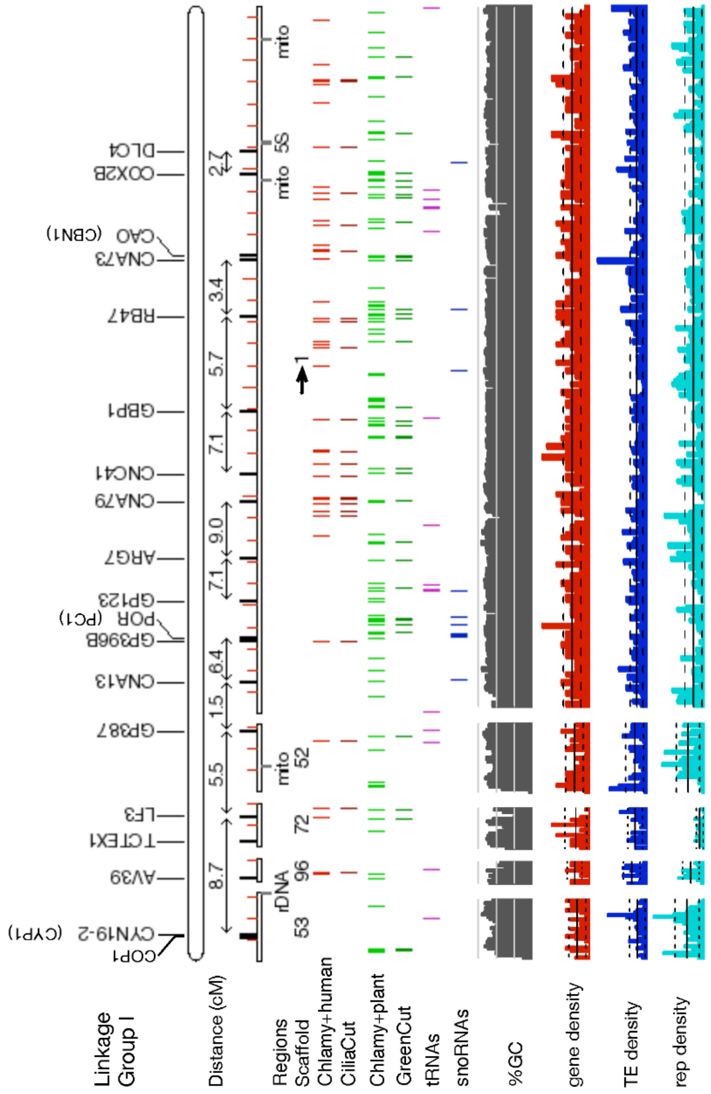
Fig. S15. Overview of linkage group XV.

Fig. S16. Overview of linkage group XVI+XVII.

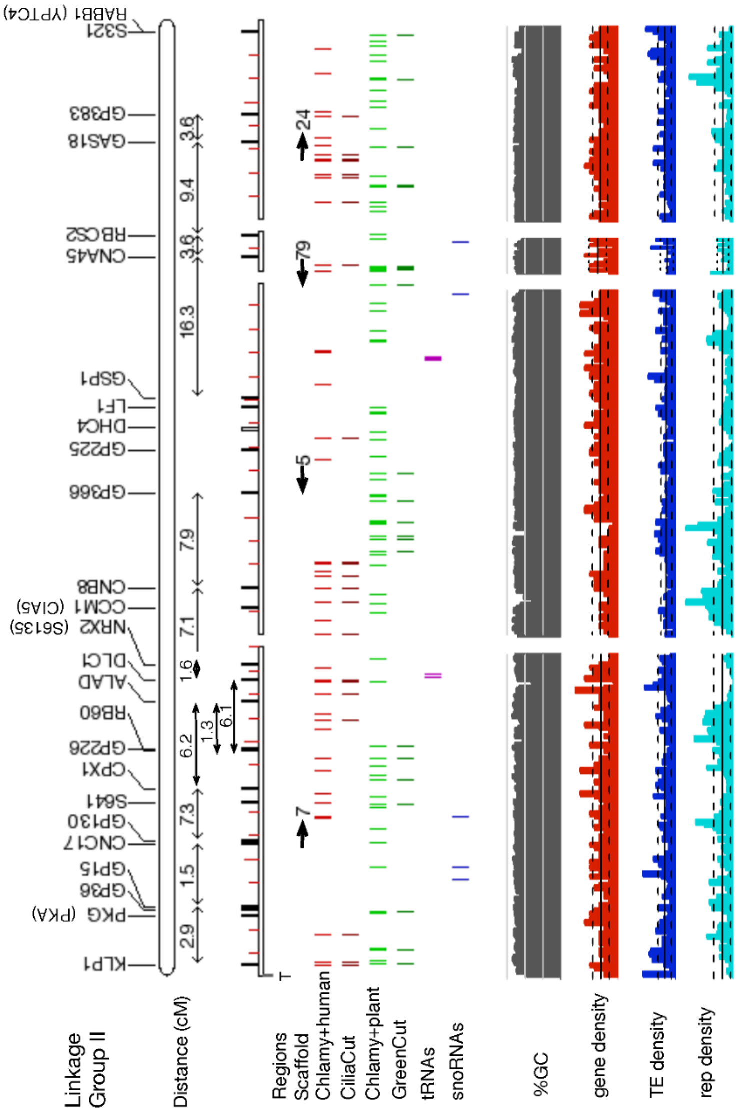
Fig. S17. Overview of linkage group XVIII.

Fig. S18. Overview of linkage group XIX.

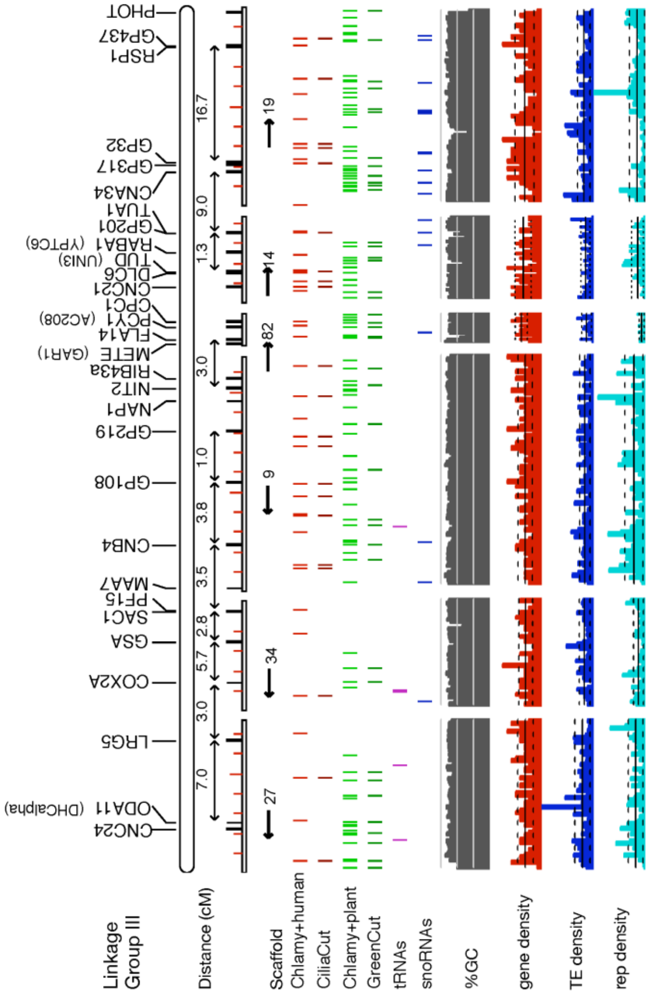
Supplemental Fig 2



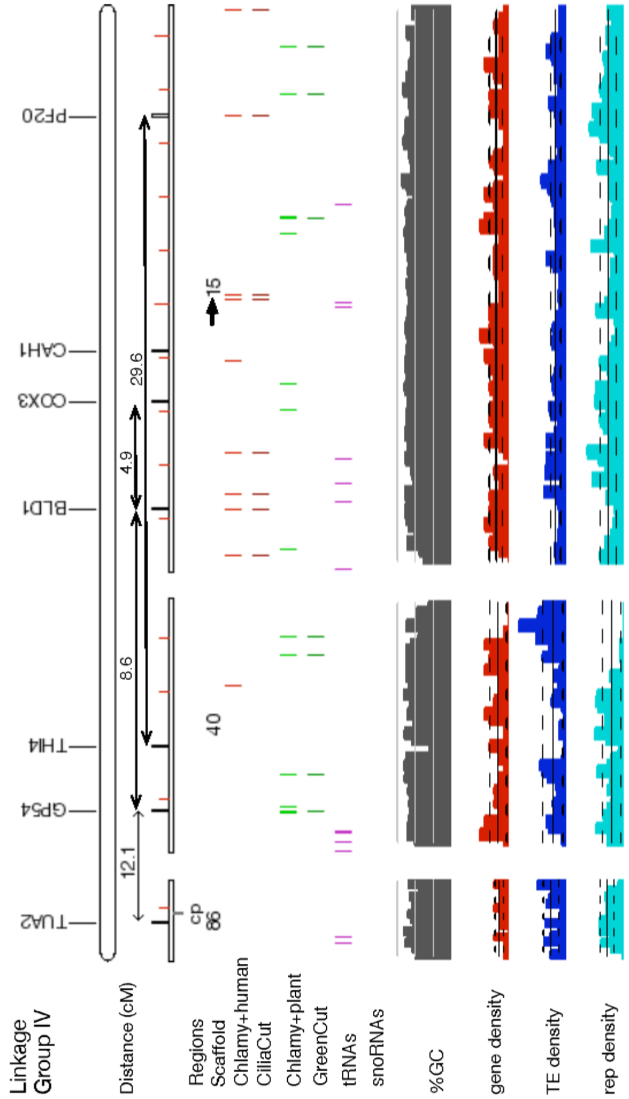
Supplemental Fig 3



Supplemental Fig 4

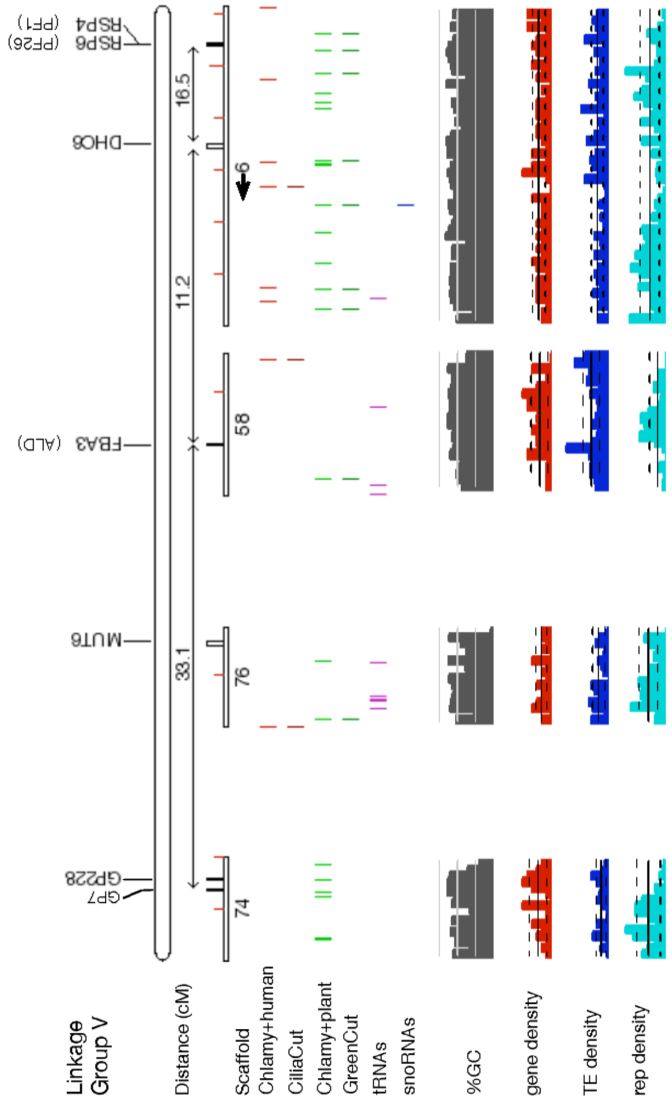


Supplemental Fig 5

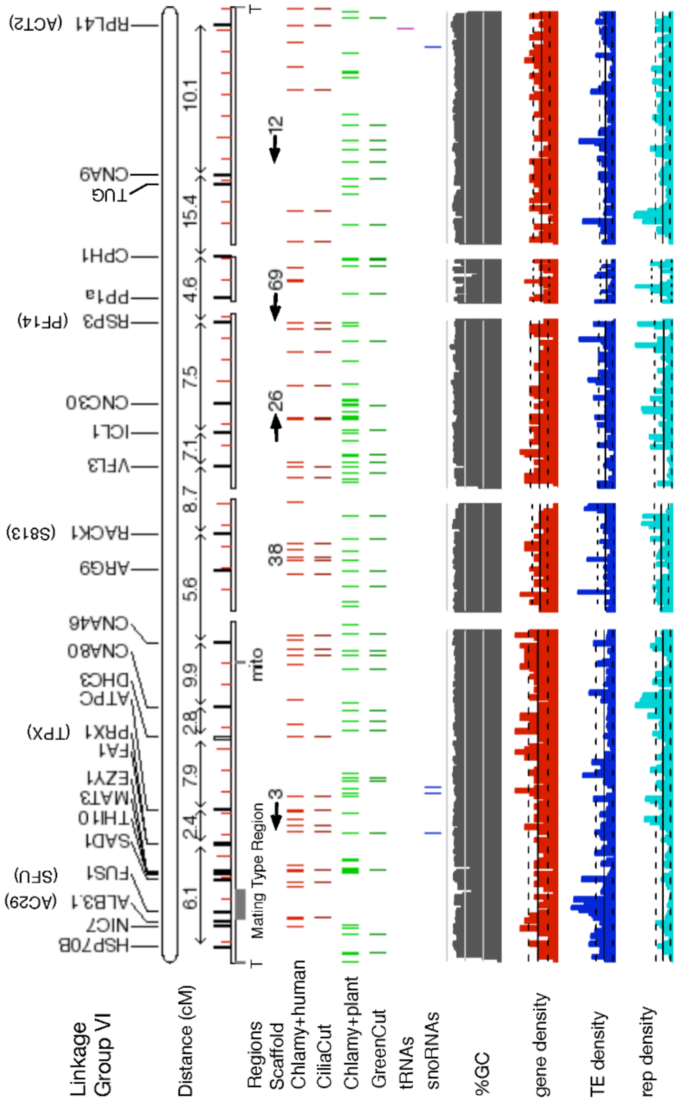


Science, Vol. 318, No. 5848 (October 2007): pg. 245-250. DOI.

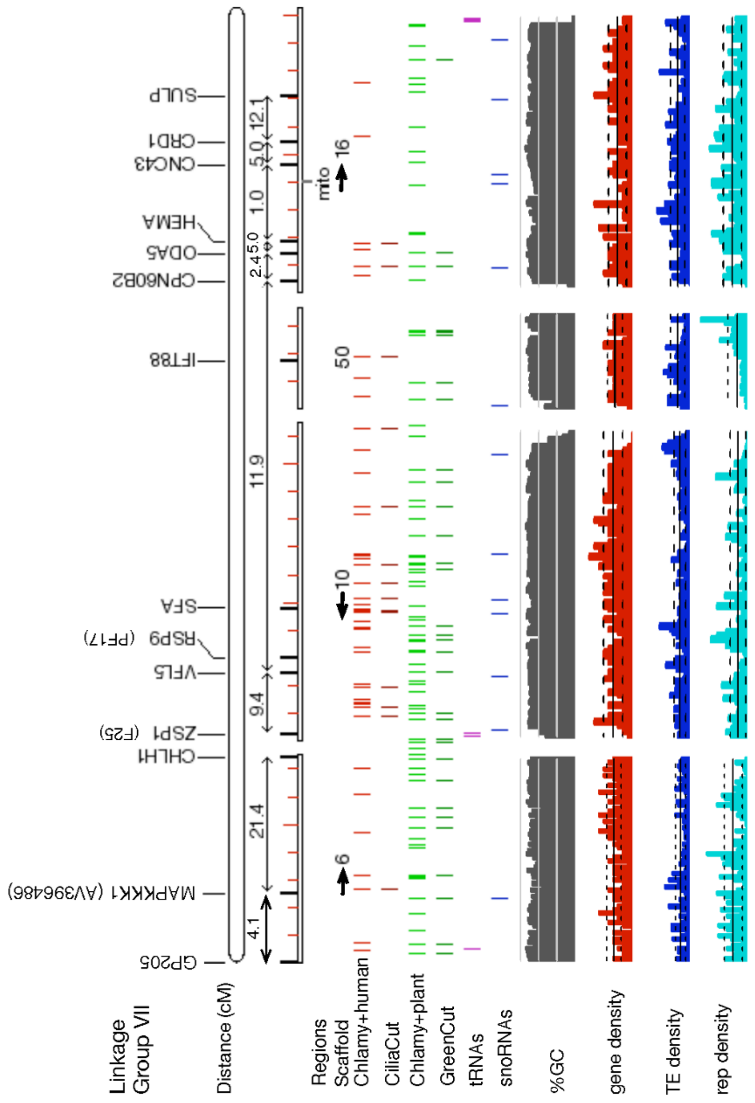
Supplemental Fig 6



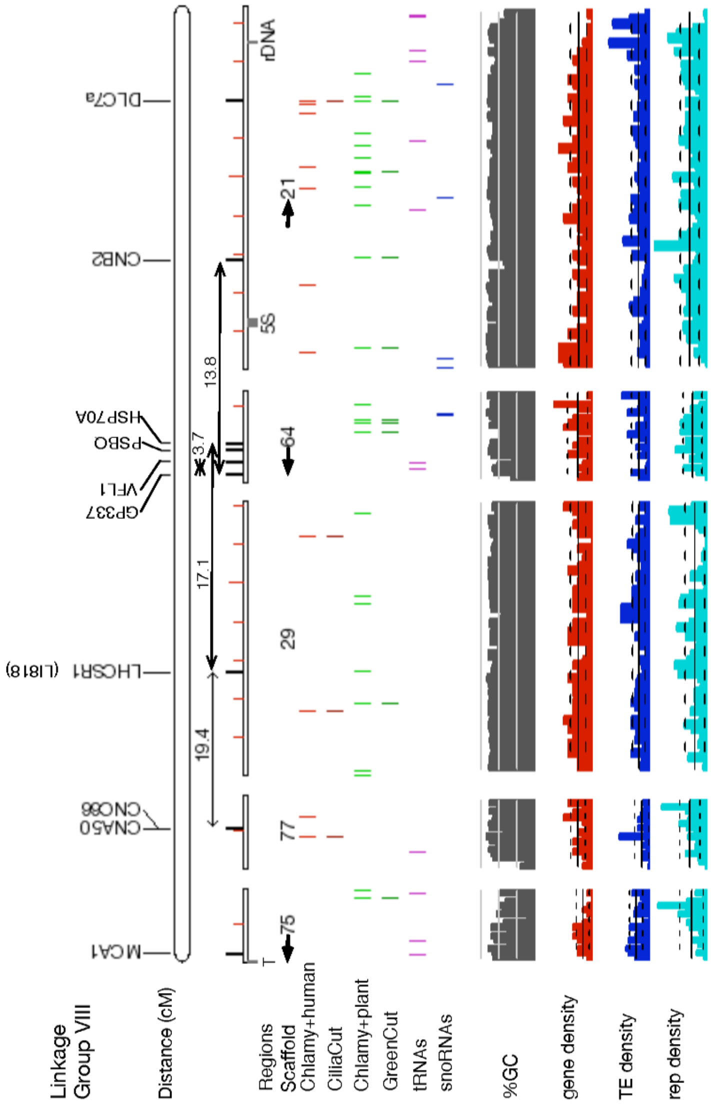
Supplemental Fig 7



Supplemental Fig 8

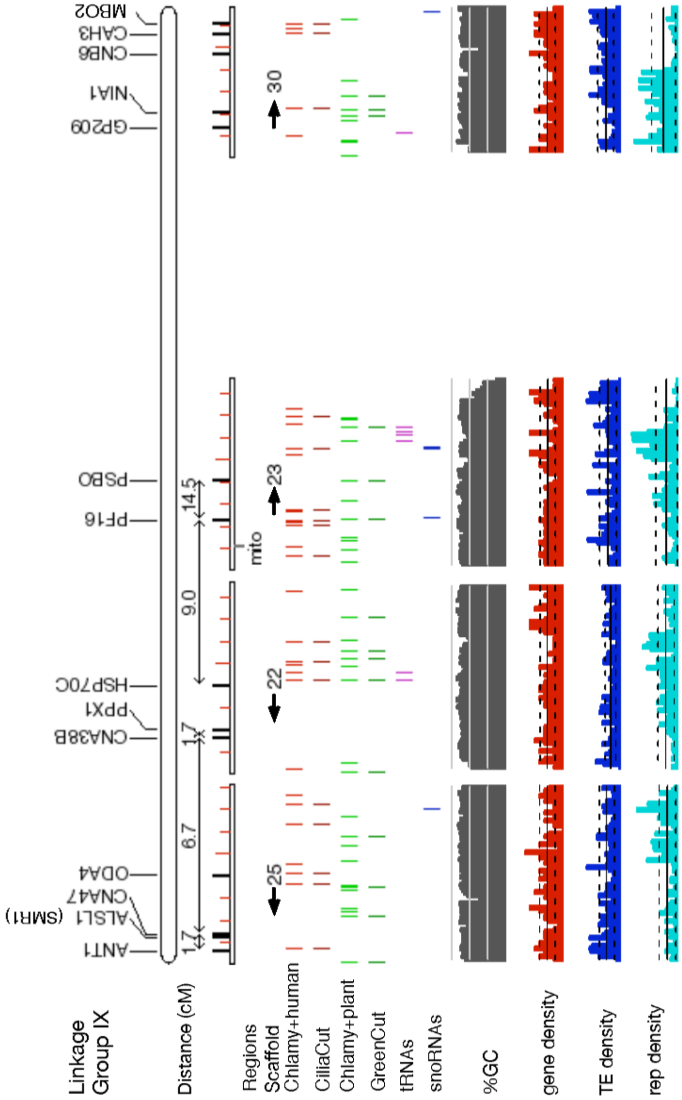


Supplemental Fig 9



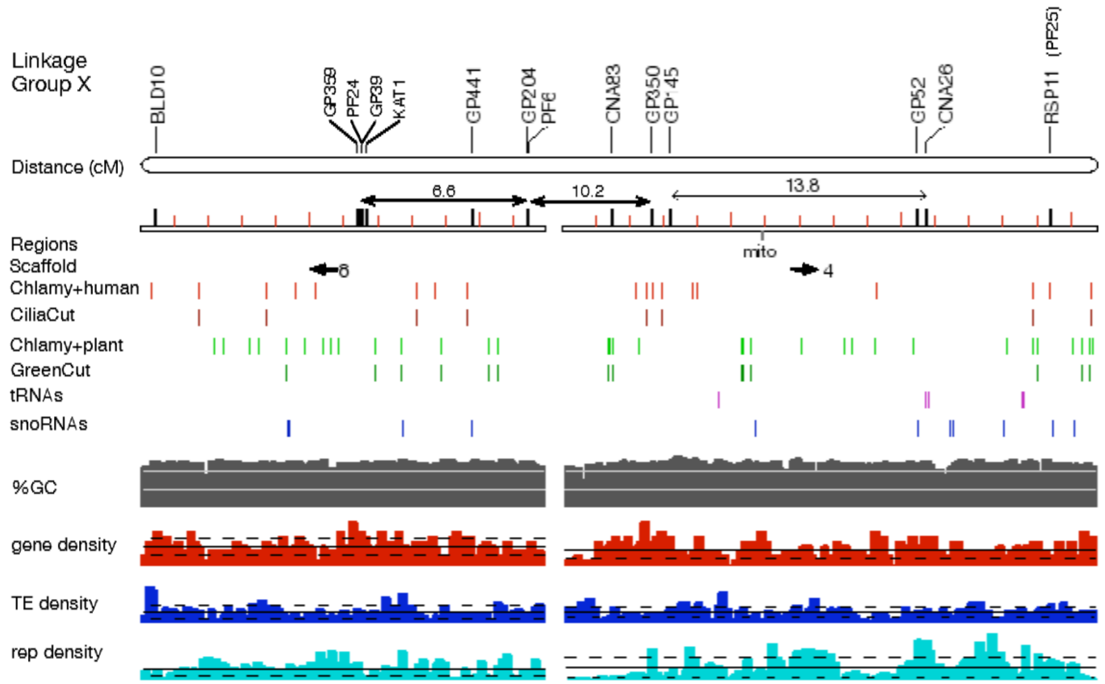
Science, Vol. 318, No. 5848 (October 2007): pg. 245-250. DOI. This article is © American Association for the Advancement of Science and permission has been granted for this version to appear in e-Publications@Marquette. American Association for the Advancement of Science does not grant permission for this article to be further copied/distributed or hosted elsewhere without the express permission from American Association for the Advancement of Science.

Supplemental Fig 10

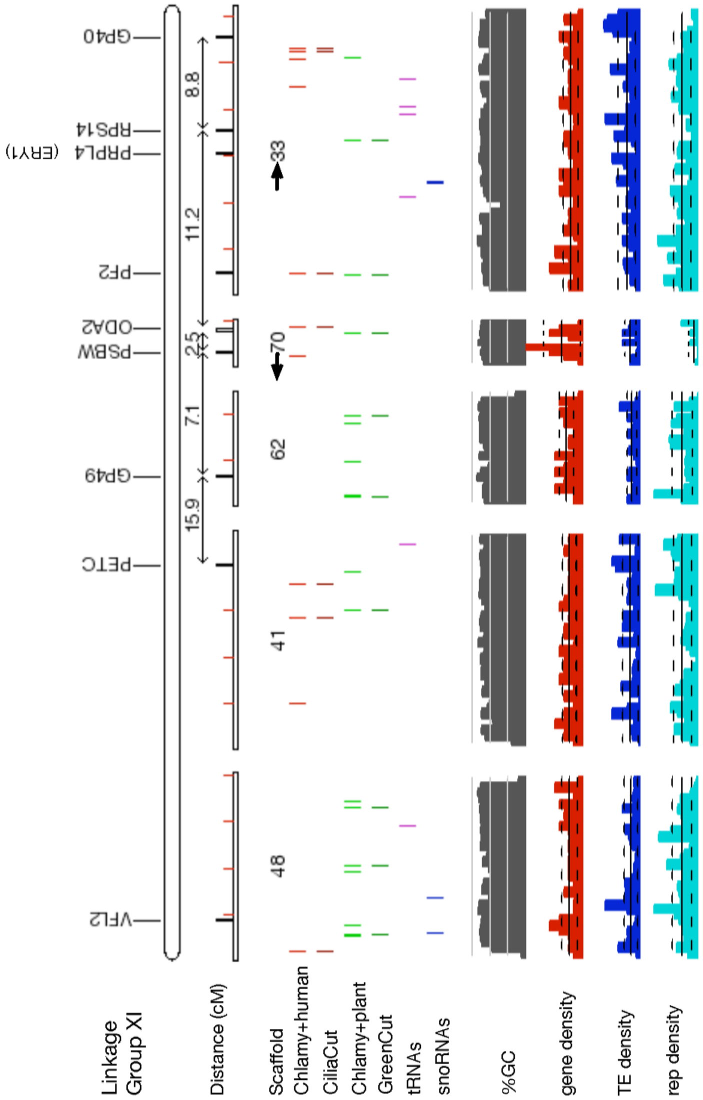


Science, Vol. 318, No. 5848 (October 2007): pg. 245-250. DOI. This article is © American Association for the Advancement of Science and permission has been granted for this version to appear in e-Publications@Marquette. American Association for the Advancement of Science does not grant permission for this article to be further copied/distributed or hosted elsewhere without the express permission from American Association for the Advancement of Science.

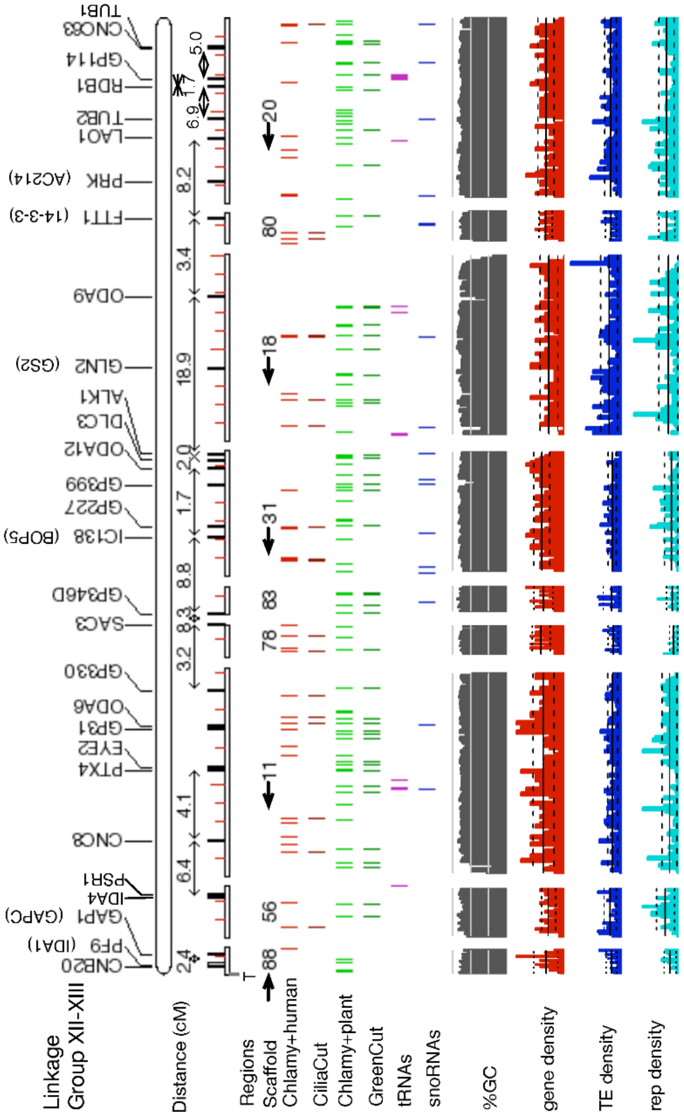
Supplemental Fig 11



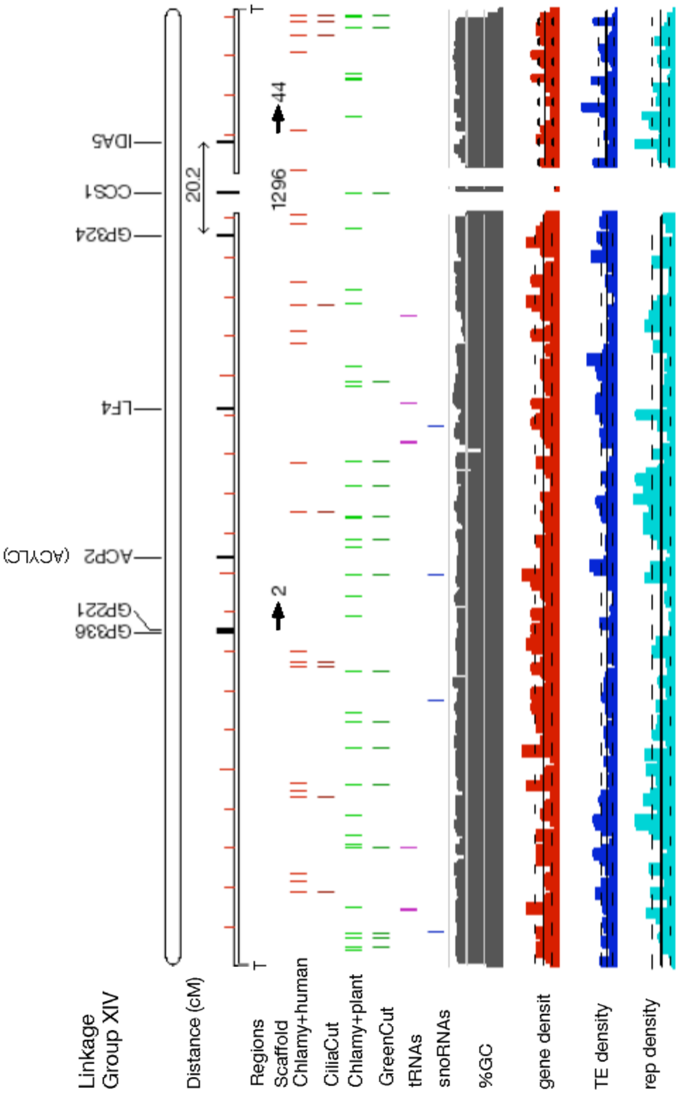
Supplemental Fig 12



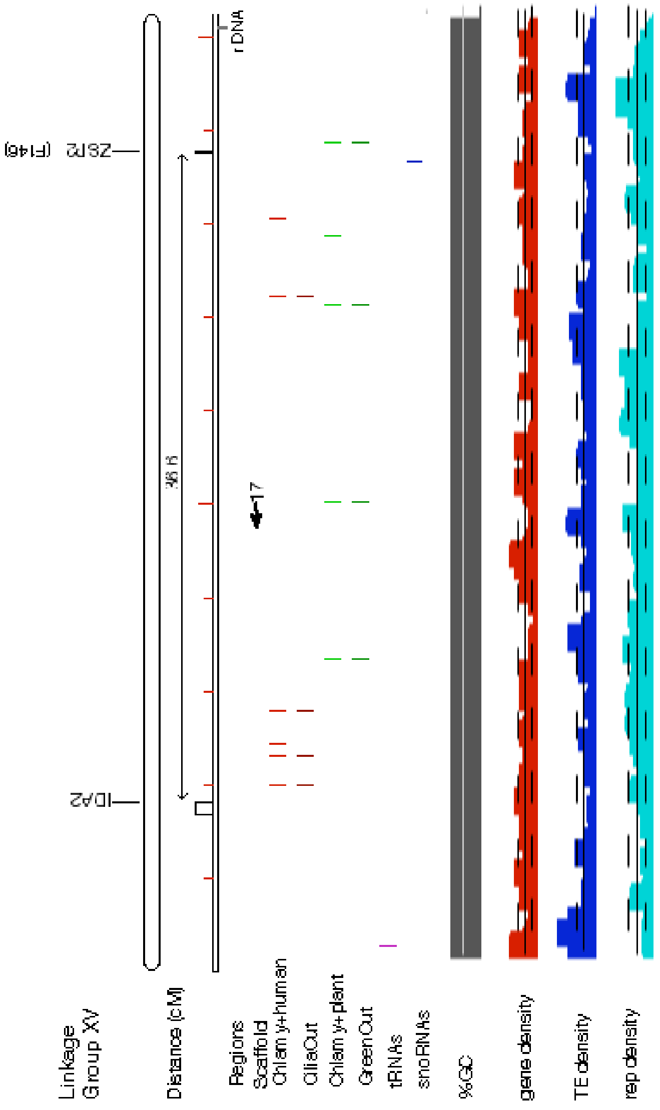
Supplemental Fig 13



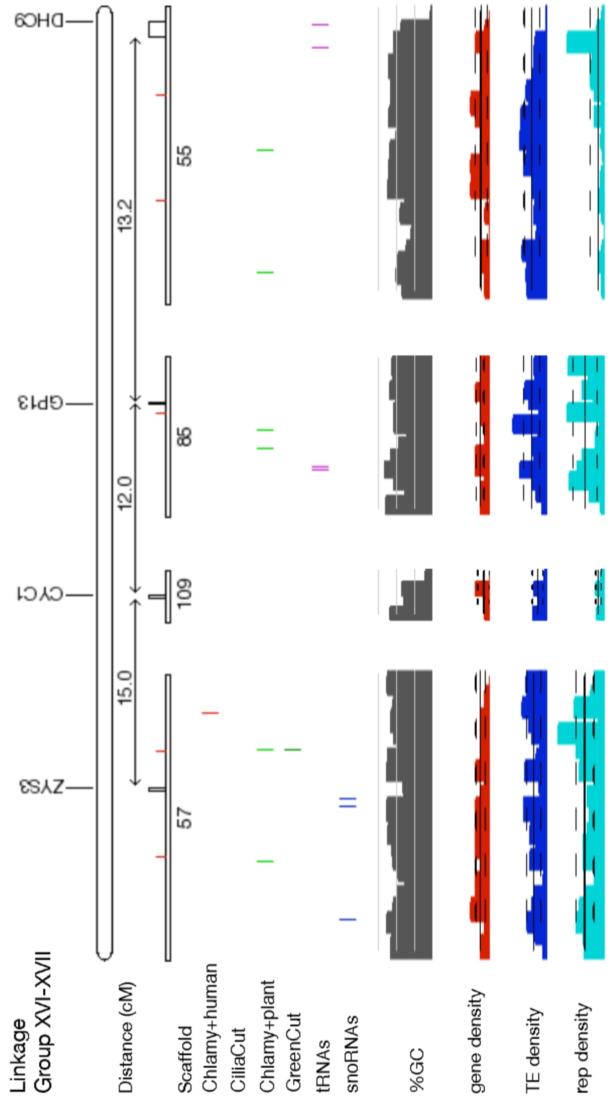
Supplemental Fig 14



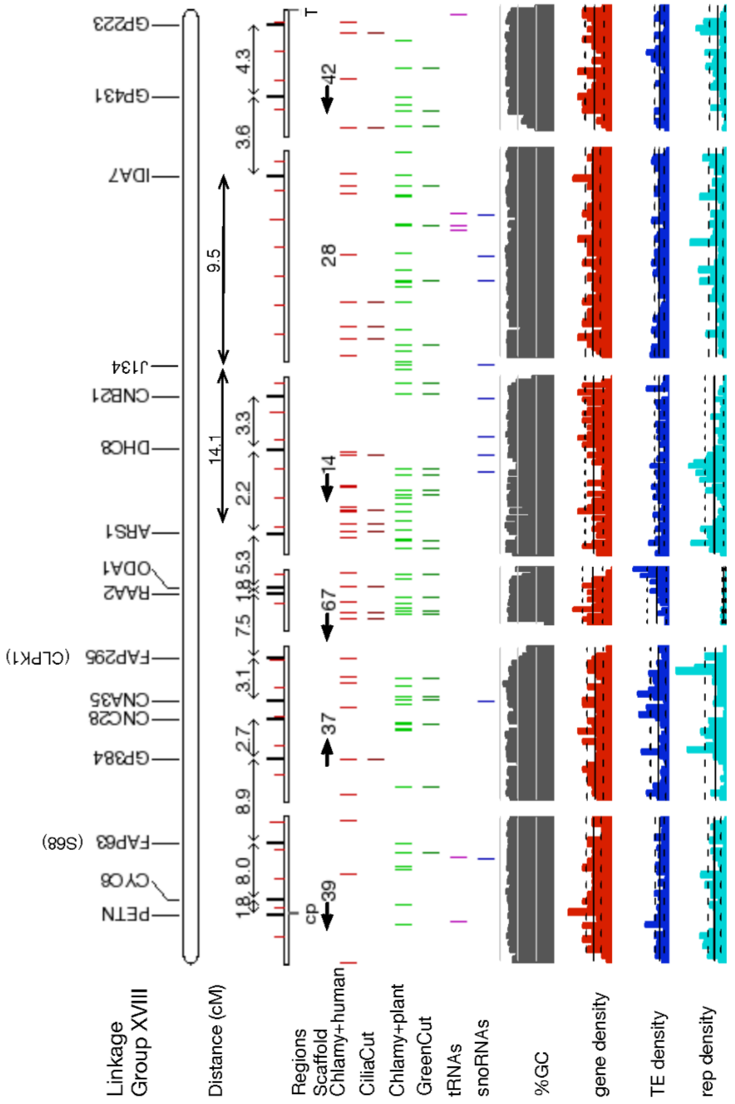
Supplemental Fig 15



Supplemental Fig 16



Supplemental Fig 17



Science, Vol. 318, No. 5848 (October 2007): pg. 245-250. DOI.

Supplemental Fig 18

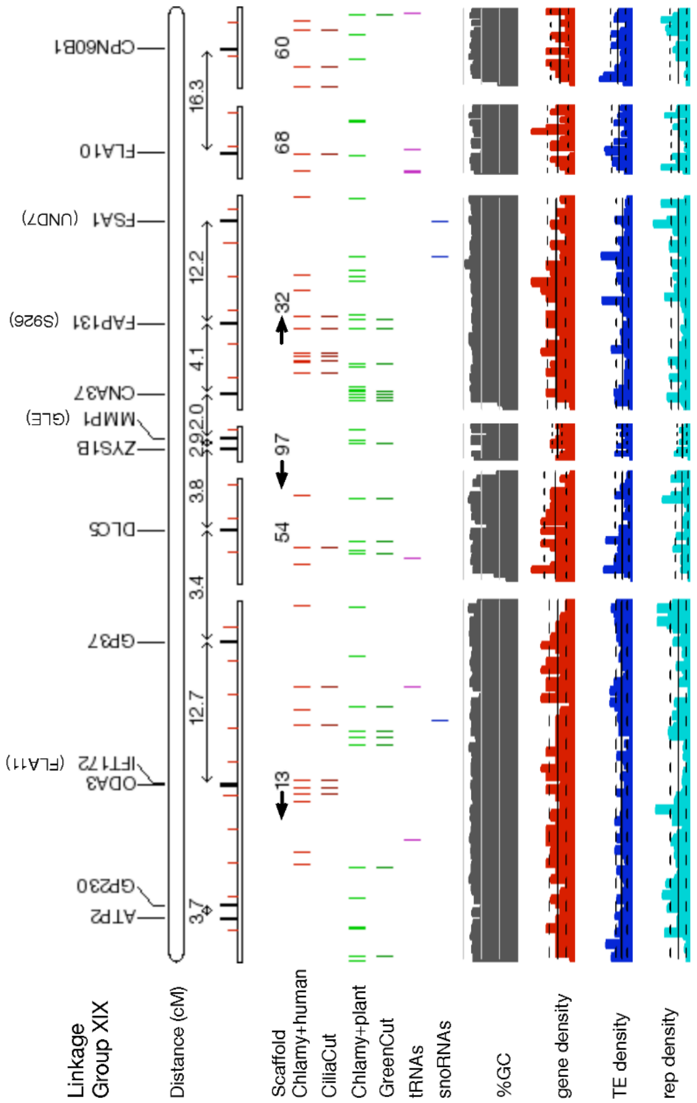


Fig. S19. Intron evolution in tRNA-Val cluster: (**A**) The 12 tRNAs, numbered consecutively, on scaffold 20:1350500-1386900 (LG XII-XIII) are depicted as arrows that indicate orientation on the chromosome, and color indicating those tRNAs that share sequence similarity (especially in the introns; see Fig. S19B). The spacing in bp between the tRNAs is indicated by the numbers above the intergenic regions. The anticodon is shown below each gene, and the asterisk within the arrow indicates that the tRNA has a genome-encoded CCA. (**B**) A neighbor-joining tree of the tRNA intron sequences with sequence differences between introns of the paired genes highlighted in bold black.

Supplemental Fig 19

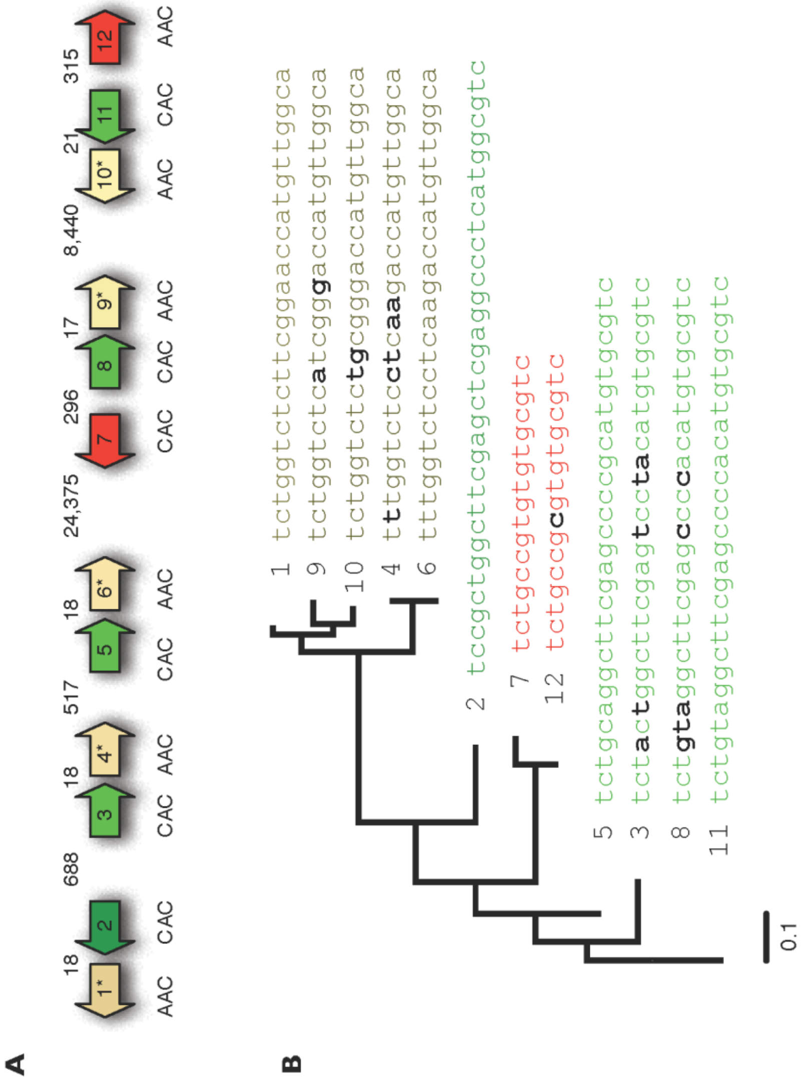
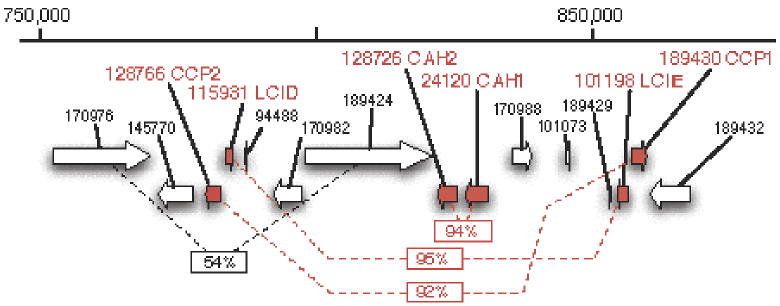

A
AAC
1*
18
2
CAC
688
3
CAC
18
4*
AAC
517
5
CAC
18
6*
AAC
24,375
7
CAC
296
8
CAC
17
9*
AAC
8,440
10*
AAC
21
11
CAC
315
12
AAC
B
1 tctggtctcttcggaaccatgttggca
9 tctggtctcatcgggaccatgttggca
10 tctggtctctgcgggaccatgttggca
4 tttggtctcctcaagaccatgttggca
6 tttggtctcctcaagaccatgttggca
2 tccgctggcttcgagctcgaggccctcatggcgtc
7 tctgccgtgtgtgcgtc
12 tctgccgcgtgtgcgtc
5 tctgcaggcttcgagcccgcatgtgcgtc
3 tctactggcttcgagtcctacatgtgcgtc
8 tctgtaggcttcgagccccacatgtgcgtc
11 tctgtaggcttcgagccccacatgtgcgtc
0.1

Fig. S20. The carbon concentrating mechanism region: The ~100 kb region of the genome (scaffold 15) that contains several genes associated with the carbon concentrating mechanism (CCM). Arrows are used to depict the different genes and their lengths and orientations and each gene is labeled with a JGI Chlre.v3.0 protein ID and gene name (where one has been assigned). Coordinates (bp) on scaffold 15 are shown along the line at the top. The red arrows depict the six CCM genes (*CCP2*, *LCID*, *CAH2*, *CAH1*, *LCIE* and *CCP1*), which were identified from both sequence and experimental data. The arrangement of the genes suggests three recent duplications. Neighboring and intervening genes are shown as open arrows. On the lower portion, red dashed lines connect the duplicated CCM sequences, with % nucleotide identity shown in boxes. One additional gene pair of unknown function in this region shows significant paralogy (black dashed lines connecting 170976 & 189424).

Supplemental Fig 20



Science, Vol. 318, No. 5848 (October 2007): pg. 245-250. DOI. This article is © American Association for the Advancement of Science and permission has been granted for this version to appear in e-Publications@Marquette. American Association for the Advancement of Science does not grant permission for this article to be further copied/distributed or hosted elsewhere without the express permission from American Association for the Advancement of Science.

Fig. S21. Comparison of *Chlamydomonas* intron characteristics to those of other eukaryotes: Introns were collected from the genomes of the organisms listed (see Fig. 2), and graphs were plotted of (**A**) the log lengths of the introns against frequency in the genome, or (**B**) the average length for introns in each of the organisms against the average number of introns.

Supplemental Fig 21

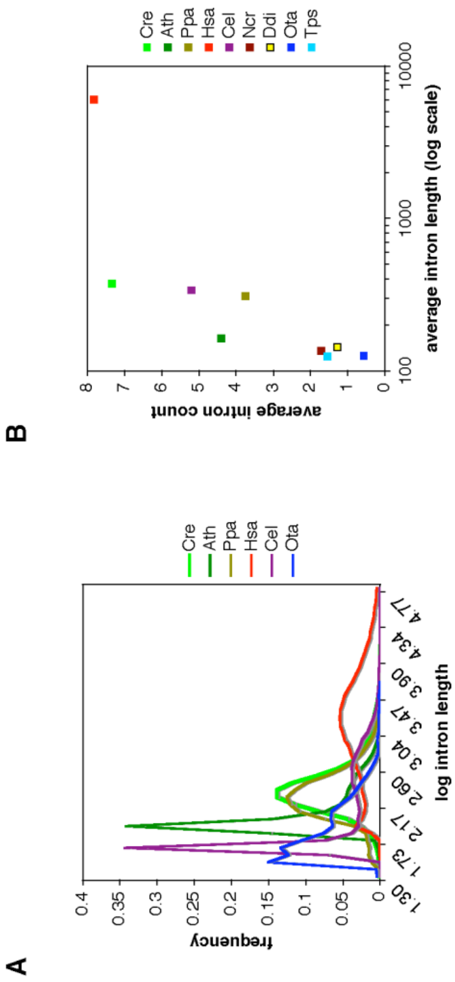


Fig. S22. Summary of transporter families: Transporter families (described along the top of the figure; the abbreviations can be found at (*44*)) that are present in organisms or groups of organisms listed on the left are colored with a red box. The criterion used for identification of the transporters is described in the **MATERIALS AND METHODS** section of this text. Families of transporters present in *Chlamydomonas* are highlighted with a horizontal green bar. Transporter families and organisms were automatically clustered hierarchically to generate the order in which they are displayed, and then grouped by coarse phylogenetic (vertical) and transporter superfamily (horizontal) membership. The analysis has been performed for transporter families present in animals (*H. sapiens*, *C. elegans*, *D. melanogaster*, *A. gambiae*), various single cell eukaryotes (sing euk: *E. histolytica* HM1:IMSS, *C. parvum* genotype 2 isolate, *E. cuniculi*, *T. parva*, *P. falciparum* 3D7, *P. vivax*, *T. thermophila* SB210, *T. brucei* TREU927/4 GUTat10.1, *L. major Friedlin*, *T. cruzi* CL Brener TC3, *T. whippelii* TW08/27, *T. whipplei* Twist), fungi (*S. pombe*, *A. oryzae*, *C. posadasii* C735, *A. nidulans* FGSC-A26, *A. fumigatus* Af293, *N. crassa* 74-OR23-IVA, *C. neoformans*, *S. cerevisiae* S288C), amoeba (*D. discoideum*), land plants (*O. sativa*, *A. thaliana*), *Ostreococcus* spp. (ostreo), *Chlamydomonas* (chlamy), the red alga *C. merolae 10D* and the diatom *Thalassiosira* (red alg+diat), and 220 bacteria. The color shows the proportion of species within the group that have genes for members of the indicated transporter family: black (family absent in all species); bright red (family present in all species); intermediate red color (family present in some species).

Supplemental Fig 22

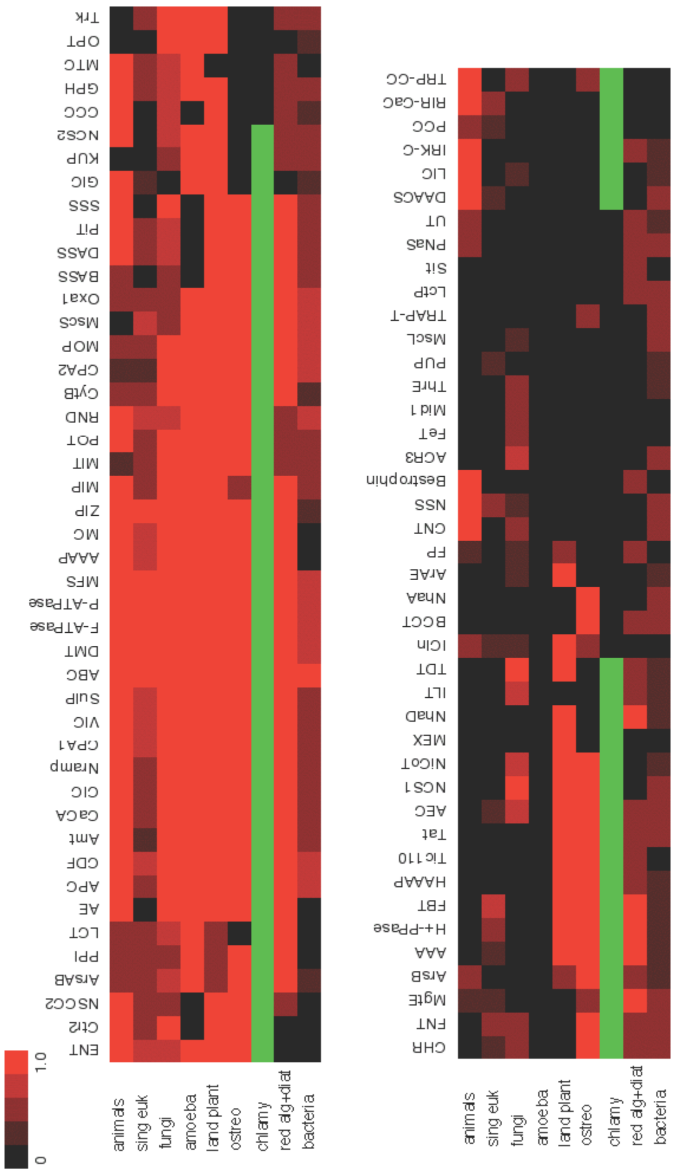


Fig. S23. Complete repertoire of transporter families: Details of clustering of transporter families across bacteria and eukaryotes are shown (summarized in Fig S23). Organisms are in rows; transporter families in columns. Euclidean distance clustering was performed in both dimensions. Red indicates presence of a transporter family; black, absence.

Supplemental Fig 23

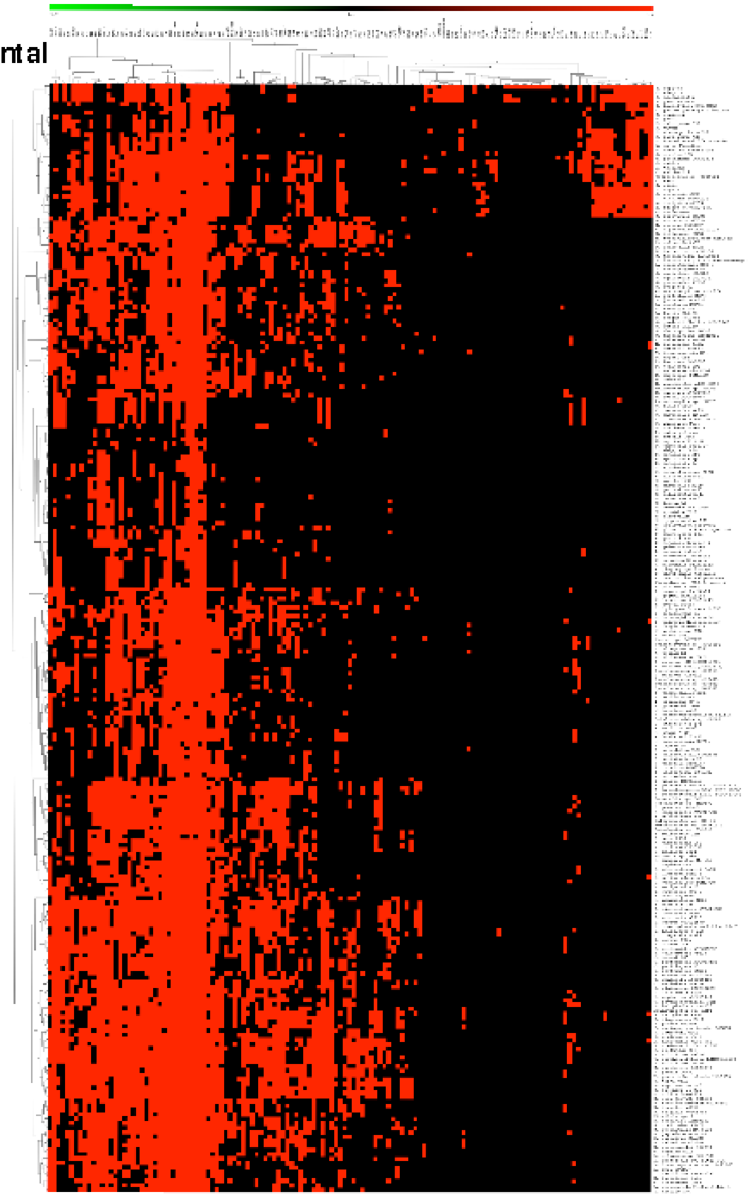


Fig. S24. Classification of CiliaCut proteins: Functional classification of CiliaCut proteins by manual annotation. Classification was based on the published function of characterized protein family members (if any), and/or the molecular function of predicted PFAM domains. 125 (67%) of the CiliaCut proteins were successfully classified; the remaining 80 either were not associated with functional information or the functional information available was ambiguous and is not included.

Supplemental Fig 24

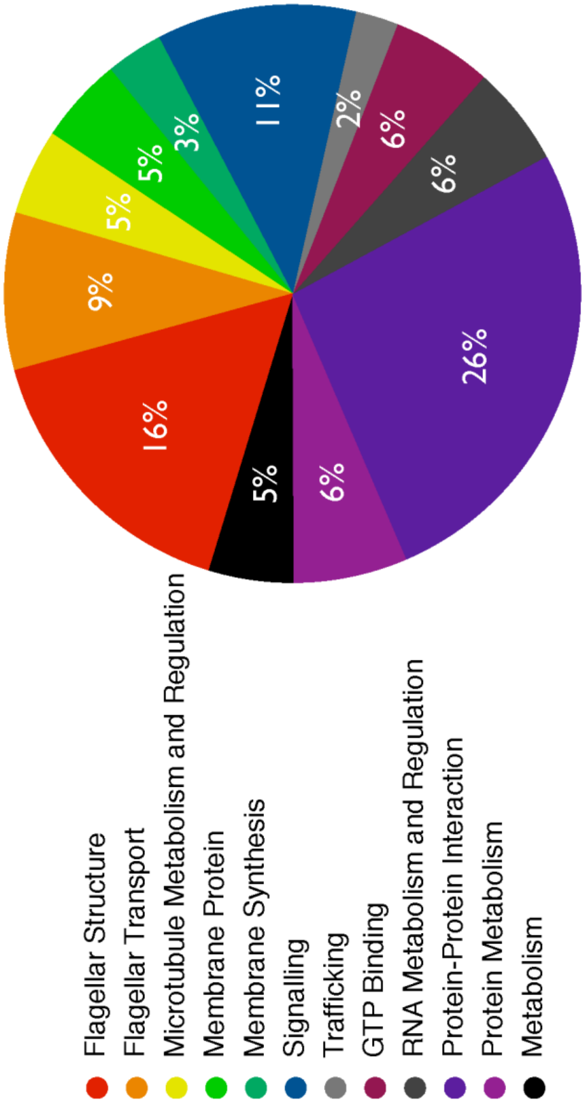
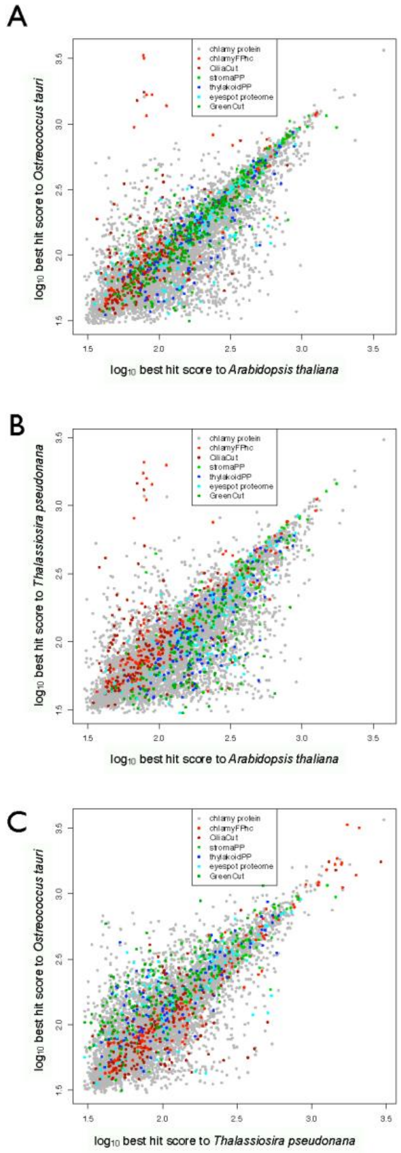


Fig. S25. Best hit scatter plots: Each *Chlamydomonas* protein is plotted by $\log_{10}$ of its best blast hit score to (**A**) *Arabidopsis*, *Ostreococcus tauri*; (**B**) *Arabidopsis*, *Thalassiosira*; (**C**) *Thalassiosira*, *Ostreococcus tauri*. Proteins are grey or colored by membership of functional or comparative genomic grouping: *Chlamydomonas* Flagellar Proteome (*67*) high confidence set (ChlamyFP, red); Stroma Plastid Proteome (stromaPP, green); Thylakoid Plastid Proteome (thylakoidPP, blue); *Chlamydomonas* PS cut7 (cyan); *Chlamydomonas* eyespot proteome (yellow).

Supplemental Fig 25



4. SUPPORTING TABLES

Table S1. Summary of tRNA complement of *Chlamydomonas*: The 259 tRNAs encoded on the *Chlamydomonas* genome are grouped according to how many anticodons encode each amino acid, with total numbers for each amino acid and each anticodon indicated.

Four anticodon amino acids

amino acid	anticodon				total
Ala	AGC	GGC	CGC	TGC	
	13		10	5	28
Gly	ACC	GCC	CCC	TCC	
		17	1	1	19
Pro	AGG	GGG	CGG	TGG	
	13		6	1	20
Thr	AGT	GGT	CGT	TGT	
	6		3	2	11
Val	AAC	GAC	CAC	TAC	
	7		10	1	18

Six anticodon amino acids

amino acid	anticodon						total
Ser	AGA	GGA	CGA	TGA	ACT	GCT	
	5		5	1		8	19
Arg	ACG	GCG	CCG	TCG	TCT	CCT	
	11		3	1	1	2	18
Leu	AAG	GAG	CAG	TAG	TAA	CAA	
	3		10	1	1	2	17

Two anticodon amino acids

amino acid	anticodon		total
Phe	AAA	GAA	
		9	9
Asn	ATT	GTT	
		7	7
Lys	CTT	TTT	

	11	1	12
Asp	GTC	ATC	
	11		11
Tyr	ATA	GTA	
		8	8
Cys	ACA	GCA	
		7	7
Glu	CTC	TTC	
	13	1	14
His	ATG	GTG	
		5	5
Gln	CTG	TTG	
	6	1	7

Other amino acids

amino acid	anticodon			total
Meti	CAT			
	8			8
Mete	CAT			
	6			
Ile	AAT	GAT	TAT	
	7	1	1	9
SeC	TCA			
	1			1
Trp	CCA			
	5			5

Table S2. tRNA-related SINE-3 family elements: Details of the scaffold on which the tRNA-related SINE-3 sequence lies, the class, the amino acid of the tRNA and anticodon sequence, the begin and end coordinates of the intron, the presence (P) or absence (A) of a 3' CCA and the sequence of the tRNA-related portion of the SINE-3 element are shown.

Scaffold	Class	tRNA Type	Anti-codon	Intron Begin	Intron End	CCA	tRNA part of SINE elements
scaffold_7	SINE-Arg	Arg	CCG	2542253	2542265	P	AGGGGGGTCGTCTAAATGGTtAAGACACTCAAGCCGatttcgttaaggcTTCGAGAGAtCCTGGGTTCGAATCCCGGTCACCCCA
scaffold_203	SINE-Arg	Arg	CCG	7724	7712	P	GGGGGGGTCATCTAAATGGTtAAGACACTCAAGCCGatttcgttaaggcTTCGAGAGAtCCTGGGTTCGAATCCCGGTCACCCCA
scaffold_40	SINE-Arg	Arg	ACG	84541	84553	P	GGGGGGGTCGTCTAAATGGTtAAGACACTCAAGACGatttcgttaaggcTTCGAGAGAtCCTGGGTTCGAATCCCGGTCACCCCA
scaffold_121	SINE-Arg	Arg	CCG	47226	47238	P	GGGGGGGTCGTCTAAATGGTtAAGACACTCAAGCCGatttcgttaaggcCTCGAGAGAtCCTGGGTTCGAATCCCGGTCACCCCA
scaffold_124	SINE-Arg	Arg	CCG	4376	4364	P	GGGGGGGTCGTCTAAATGGTtAAGACACTCAAGCCGatttcgttaaggcTTCGAGAGAtCCTGGGTTCGAATCCCGATCACCCCA
scaffold_958	SINE-Arg	Arg	CCG	363	351	P	GGGGGGGTCGTCTAAATGGTtAAGACACTCAAGCCGatttcgttaaggcTTCGAGAGAtCCTGGGTTCGAATCCCGATCACCCCA
scaffold_21	SINE-Arg	Arg	CCG	1832790	1832802	P	GGGGGGGTCGTCTAAATGGTtAAGACACTCAAGCCGatttcgttaaggcTTCGAGAGAtCCTGGGTTCGAATCCCGGTCACCCCA
scaffold_21	SINE-Arg	Arg	CCG	1828284	1828272	P	GGGGGGGTCGTCTAAATGGTtAAGACACTCAAGCCGatttcgttaaggcTTCGAGAGAtCCTGGGTTCGAATCCCGGTCACCCCA
scaffold_40	SINE-	Arg	CCG	41699	41687	P	GGGGGGGTCGTCTAAATGGTtA

	Arg						AGACACTCAAGCCGatttcgttaag gcTTCGAGAGAtCCTGGGTTCGA ATCCCGGTCACCCCA
scaffold_40	SINE- Arg	Arg	CCG	9927	9915	P	GGGGGGGTCGTCTAAATGGTtA AGACACTCAAGCCGatttcgttaag gcTTCGAGAGAtCCTGGGTTCGA ATCCCGGTCACCCCA
scaffold_73	SINE- Arg	Arg	CCG	191771	191783	P	GGGGGGGTCGTCTAAATGGTtA AGACACTCAAGCCGatttcgttaag gcTTCGAGAGAtCCTGGGTTCGA ATCCCGGTCACCCCA
scaffold_113	SINE- Arg	Arg	CCG	6095	6107	P	GGGGGGGTCGTCTAAATGGTtA AGACACTCAAGCCGatttcgttaag gcTTCGAGAGAtCCTGGGTTCGA ATCCCGGTCACCCCA
scaffold_125	SINE- Arg	Arg	CCG	26556	26544	P	GGGGGGGTCGTCTAAATGGTtA AGACACTCAAGCCGatttcgttaag gcTTCGAGAGAtCCTGGGTTCGA ATCCCGGTCACCCCA
scaffold_218	SINE- Arg	Arg	CCG	208	196	P	GGGGGGGTCGTCTAAATGGTtA AGACACTCAAGCCGatttcgttaag gcTTCGAGAGAtCCTGGGTTCGA ATCCCGGTCACCCCA
scaffold_217	SINE- Arg	Arg	CCG	8299	8311	P	GGGGGGGTCGTCTAAATGGTtA AGACACTCAAGCCGatttcgttaag gcTTCGAGAGAtCCTGGGTTCGA ATCCCGGTCACCCCA
scaffold_285	SINE- Arg	Arg	CCG	10820	10808	P	GGGGGGGTCGTCTAAATGGTtA AGACACTCAAGCCGatttcgttaag gcTTCGAGAGAtCCTGGGTTCGA ATCCCGGTCACCCCA
scaffold_545	SINE- Arg	Arg	CCG	8056	8044	P	GGGGGGGTCGTCTAAATGGTtA AGACACTCAAGCCGatttcgttaag gcTTCGAGAGAtCCTGGGTTCGA ATCCCGGTCACCCCA
scaffold_729	SINE- Arg	Arg	CCG	1631	1643	P	GGGGGGGTCGTCTAAATGGTtA AGACACTCAAGCCGatttcgttaag gcTTCGAGAGAtCCTGGGTTCGA ATCCCGGTCACCCCA
scaffold_729	SINE- Arg	Arg	CCG	3577	3589	P	ATCCCGGTCACCCCA
scaffold_7	SINE-	Arg	CCG	2572686	2572674	P	GGGGGGGTCGTCTAAATGGTtA

Science, Vol. 318, No. 5848 (October 2007): pg. 245-250. DOI.

	Arg						AGACACTCAAGCCGatttcgttaag gcTTCGAGAGAtCCTGGGTTCGA ATCCCGGTCACCCCA
scaffold_58	SINE-Arg	Arg	CCG	344247	344235	P	GGGGGGGTCGTCTAAATGGTtA AGACACTCAAGCCGatttcgttaag gcTTCGAGAGAtCCTGGGTTTGA ATCCCGGTCACCCCA
scaffold_121	SINE-Arg	Arg	CCG	13939	13951	P	GGGGGGGTCGTCTAAATGGTtA AGACACTCAAGCCGatttcgttaag gcTTCGAGAGAtCCTGGGTTTGA ATCCCGGTCACCCCA
scaffold_40	SINE-Arg	Arg	CCG	79872	79884	P	GGGGGGGTCGTCTAAATGGTtA AGACACTCAAGCCGatttcgttaag gcTTTGAGAGAtCCTGGGTTCGA ATCCCGGTCACCCCA
scaffold_112	SINE-Arg	Arg	CCG	36173	36161	P	GGGGGGGTCGTCTAAATGGTtA AGACACTCAAGCCGatttcgttaag gcTTTGAGAGAtCCTGGGTTCGA ATCCCGGTCACCCCA
scaffold_1105	SINE-Arg	Arg	CCG	3248	3236	P	GGGGGGGTCGTCTAAATGGTtA AGACACTCAAGCCGatttcgttaag gcTTTGAGAGAtCCTGGGTTCGA ATCCCGGTCACCCCA
scaffold_110	SINE-Arg	Arg	CCG	57194	57181	P	GGGGGGGTCGTCTAAATGGTtA AGACACTCAAGCCGattttcgttaag gcTTTGAGAGAtCCTGGGTTCGA ATCCCGGTCACCCCA
scaffold_112	SINE-Arg	Arg	CCG	40724	40712	P	GGGGGGGTCGTCTAAATGGTtA AGACACTCAAGCCGattttgttaag gcTTTGAGAGAtCCTGGGTTCGA ATCCCGGTCACCCCA
scaffold_87	SINE-Arg	Arg	CCG	68634	68622	P	GGGGGGGTCGTCTAAATGGTtA AGACACTCGAGCCGatttcgttaag gcTTCGAGAGAtCCTGGGTTCGA ATCCCGGTCACCCCA
scaffold_124	SINE-Arg	Arg	CCG	8824	8836	P	GGGGGGGTCGTCTAAATGGTtG AGACACTCAAGCCGatttcgttaag gcTTCGAGAGAtCCTGGGTTCGA ATCCCGGTCACCCCA
scaffold_965	SINE-Arg	Arg	CCG	3317	3329	P	GGGGGGGTCGTCTAAATGGTtG AGACACTCAAGCCGatttcgttaag gcTTCGAGAGAtCCTGGGTTCGA ATCCCGGTCACCCCA
scaffold_60	SINE-	Arg	CCG	451585	451573	P	GGGGGGGTTGTCTAAATGGTtA

	Arg						AGACACTCAAGCCGatttcgttaaggcTTCGAGAGAtCCTGGGTTCGAATCCCGGTCACCCCA
scaffold_270	SINE-Arg	Arg	CCG	2897	2909	P	GGGGGGGTTGTCTAAATGGTtAAGACACTCAAGCCGatttcgttaaggcTTCGAGAGAtCCTGGGTTCGAATCCCGGTCACCCCA
scaffold_52	SINE-Arg	Arg	CCG	566079	566067	P	TGGGGGGTCGTCTAAATGGTtAAGACACTCAAGCCGatttcgttaaggcTTCGAGAGAtCCTGGGTTCGAATCCCGGTCACCCCA
scaffold_1295	SINE-Arg	Arg	CCG	914	902	A	GGGGTcGTCTAAATGGTtAAGACACTCAAGCCGatttcgtcaaggcTTTGAGAGAtCCTGGGTTCGAATCCCAGTCACCCCA
scaffold_18	SINE-Arg	Arg	CCG	96788	96776	A	GGGGTcGTCTAAATGGTtAAGACACTCAAGCCGatttcgtcaaggcTTTGAGAGAtCCTGGGTTCGAATCCCAGTCACCCCA
scaffold_136	SINE-Arg	Trp	CCA	36284	36296	A	GGGGGGGTCGTCTAAATGGTtAAGACACTCAAGCCAatttcgttaaggcTTCGAGAGAtCCTGGGTTCGAATCCCGGTCGCCCCA
scaffold_258	SINE-Arg	Trp	CCA	2370	2358	P	GGGAGGGTCGTCTAAATGGTtAAGACACTCAAGCCAatttcgttaaggcTTCGAGAGAtCCTGGGTTCGAATCCCGGTCACCCCA
scaffold_808	SINE-Arg	Trp	CCA	3681	3693	A	GGGGGGGTCGTCTAAATGGTtAAGACACTCAAGCCAatttcgttaaggcTTCGAGAGAtCCTGGGTTCGAATCCCGGTCGCCCCA
scaffold_99	SINE-Arg	Gly	CCC	112238	112322	P	GGGGGGGTCGTCTAAATGGTtAAGACACTCAAgCCCAtttcgttaaggcTTCGAGAGAtCCTGGGTTCAAATCCCGGTCACCCCA
scaffold_285	SINE-Arg	Gln	CTG	9029	8945	P	GGGGGGGTCGTCTAAATGGTtAAGACACTCAAgCTGAtttcgttaaggcTTCGAGAGAtCCTGGGTTCGAATCCCGGTCACCCCA

Science, Vol. 318, No. 5848 (October 2007): pg. 245-250. DOI.

Table S3. tRNA-related SINE family elements: Details of the scaffold on which the tRNA-related SINE sequence lies, the class, the amino acid of the tRNA and anticodon sequence, the begin and end coordinates of the intron, the presence (P) or absence (A) of a 3' CCA and the sequence of the tRNA-related portion of the SINE-3 element are shown.

Scaffold	Class	tRNA Type	Anti-codon	Intron Begin	Intron End	CCA	tRNA part of SINE elements
scaffold_808	SINE-Asp	Asp	ATC	2922	2972	A	GGGGGGGTAGCTCAGTAGGTaAGAGCACTTCCTTATCAccctgcggacccgggttcaaatctcgtattcggcccgtttcccggcggataAGGTTGAGGtCGTGGGTTCGGATCCCACCCCCCTCA
scaffold_136	SINE-Asp	Asp	ATC	35519	35569	A	GGGGGGGTAGCTCAGTAGGTaAGAGCACTTCCTTATCAccctgcggacccgggttcgaatctcgtattcggcccgtttcccggcagataAGGTTGAGGtCATGGGTTCGGATCCCACCCCCCTCA
scaffold_42	SINE-Asp	Asp	ATC	853996	853946	A	GGGGGGGTAGCTCAGTAGGTaAGAGCACTTCCTTATCAccctgcggacccgggttcgaatctcgtattcggcccgtttcccggcggataAGGTTGAGGtCGTGGGTTCGGATCCCACCCCCCTCA
scaffold_98	SINE-Asp	Asp	ATC	137522	137572	A	GGGGGGGTAGCTCAGTAGGTaAGAGCACTTCCTTATCAccctgcggacccgggttcgaatctcgtattcggcccgtttcccggcggataAGGTTGAGGtCGTGGGTTCGGATCCCACCCCCCTCA
scaffold_986	SINE-Asp	Asp	ATC	868	818	A	GGGGGGGTAGCTCAGTAGGTaAGAGCACTTCCTTATCAccctgcggacccgggttcgaatctcgtattcggcccgtttcccggcggataAGGTTGAGGtCGTGGGTTCGGATCCCACCCCCCTCA
scaffold_20	SINE-Asp	Asp	ATC	685237	685287	A	GGGGGGGTAGCTCAGTAGGTaAGAGCACTTCCTTATCAccctgcggacccgggttcgaatctcgtattcggcccgtttcccggcggataAGGTTGAGGtCGTGGGTTTGGATCCCACCCCCCTCA
scaffold_104	SINE-Asp	Asp	ATC	32583	32633	A	GGGGGGGTAGCTCAGTAGGTaAGAGCACTTCCTTATCAccctgcggacccgggttcgaatctcgtattcggcccgtttcctggcggataAG

							GTTGAGGtCGTGGGTTCGGATCCCACCCCCCTCA
scaffold_55	SINE-Asp	Asp	GTC	536020	535977	A	TCCCCGGTAGCTCAATTGGTAGAGCATGCCGCTGTCAcatggcagacccaggttcgaatcacggattcggccgggttgaggCTGACAAGTATAGaTGCAGGTTCGGATCCTGCCCGGGGAA
scaffold_56	SINE-Asp	Asp	GTC	563038	563081	P	TCCCCGGTAGCTCAATTGGTAGAGCATGCCGCTGTCAcatggcagacccaggttcgaatcgcagattcggccaggttgaggCTGACAAGTATAGaTGCAGGTTCGGATCCTGCCCGGGGAA
scaffold_99	SINE-Asp	Asp	GTC	9414	9457	P	TCCCCGGTAGCTCAATTGGTAGAGCATGCCGCTGTCAcatggcagacccaggttcgaatcgcagattcggccaggttgaggCTGACAAGTATAGaTGCAGGTTCGGATCCTGCCCGGGGAA
scaffold_388	SINE-Asp	Asp	GTC	552	595	P	TCCCCGGTAGCTCAATTGGTAGAGCATGCCGCTGTCAcatggcagacccaggttcgaatcgcagattcggccaggttgaggCTGACAAGTATAGaTGCAGGTTCGGATCCTGCCCGGGGAA
scaffold_2134	SINE-Asp	Asp	GTC	666	624	A	TCCCCGGTAGCTCAATTGGTAGAGCATGCCGCTGTCAcatggcagacccaggttcgaatcgcggattcggccgggttaggCTGACAAGTATAGaTGCAGGTTCGGATCCTGCCCGGGGAA
scaffold_120	SINE-Asp	Asp	GTC	50480	50437	A	TCCCCGGTAGCTCAATTGGTAGAGCATGCCGCTGTCAcatggcagacccaggttcgaatcgcggattcggccgggttgaggCTGACAAGTATAGaTGCAGGTTCGGATCCTGCCCGGGGAA
scaffold_2077	SINE-Asp	Asp	GTC	718	761	A	TCCCCGGTAGCTCAATTGGTAGAGCATGCCGCTGTCAcatggcagacccaggttcgaatcgcggattcggccgggttgaggCTGACAAGTATAGaTGCAGGTTCGGATCCTGCCCGGGGAA
scaffold_51	SINE-Asp	Asp	GTC	33405	33448	A	TCCCCGGTAGCTCAATTGGTAGAGCATGCCGCTGTCAcatggcagacccaggttcgaatctccgattcggccaggttgaggCTGACAAGTATAGaTGCAGGTTCGGATCCTGCCCGGGGAA
scaffold_18	SINE-Asp	Asp	GTC	75512	75555	A	TCCCCGGTAGCTCAATTGGTAGAGCATGCCGCTGTCAcatggcagacccaggttcgaa

							tctccgattcggccaggttgaggCTGACAAGTATAGaTGCAGGTTCGGATCCTGCCCGGGGAA
scaffold_58	SINE-Asp	Asp	GTC	9057	9100	A	TCCCCGGTAGCTCAATTGGTAGAGCATGCCGCTGTCAcatggcagacccaggttcgattcacggattcggccgggttgaggCTGACAAGTATAGaTGCAGGTTCGGATCCTGCCCGGGGAA
scaffold_58	SINE-Asp	Asp	ATC	44296	44346	A	GGGGGGGTAGCTCAGTAGGTaAGAGCACTTCCTTATCAccctgcggacccgggttcgaatctcgtattcggcccgtttcccggcggataAGGTTGAGGtCGTGGGTTCGGATCCCACCCCCCTCA
scaffold_73	SINE-Asp	Asp	ATC	149364	149314	A	GGGGGGGTAGCTCAGTAGGTaAGAGCACTTCCTTATCAccctgcggacccgggttcgaatctcgtattcggcccgtttcccggcggataAGGTTGAGGtCGTGGGTTCGGATCCCACCCCCCTCA
scaffold_55	SINE-Asp	Asp	GTC	491372	491329	P	TCCCCGGTAGCTCAATTGGTAGAGCATGCCGCTGTCAcatggcagacccaggttcgaatcgcagattcggccaggttgaggCTGACAAGTATAGaTGCAGGTTCGGATCCTGCCCGGGGAA
scaffold_59	SINE-Asp	Asp	GTC	308788	308831	A	TCCCCGGTAGCTCAATTGGTAGAGCATGCCGCTGTCAcatggcagacccaggttcgaatcgcagattcggccaggttgaggCTGACAAGTATAGaTGCAGGTTCGGATCCTGCCCGGGGAA
scaffold_110	SINE-Asp	Asp	GTC	47423	47380	P	TCCCCGGTAGCTCAATTGGTAGAGCATGCCGCTGTCAcatggcagacccaggttcgattcacggattcggccgggttgaggCTGACAAGTATAGaTGCAGGTTCGGATTCTGCCCGGGGAA
scaffold_58	SINE-Asp	Asp	ATC	46217	46267	A	GGGGGGGTAGCTCAGTAGGTaAGAGCACTTCCTTATCAccctgcggacccgtgttcgaatctcgtattcggcccgtttcccggcggataAGGTTGAGGtCGTGGGTTCGGATCCCACCCCCCTCA
scaffold_59	SINE-Asp	Asp	GTC	404475	404518	A	TCCCCGGTAGCTCAATTGGTAGAGCATGCCGCTGTCAcatggcagacccaggttcgaatcacggattcggccgggttgaggCTGACAAGTATAGaTGCAGGTTCGGATCCTGCCCGGGGAA
scaffold_18	SINE-	Asp	GTC	1388379	1388336	A	TCCCCGGTAGCTCAATTGGTAGAGCAT

	Asp						GCCGCTGTCAcatggcagacccaggttcgaatcgcagattcggccaggttgaggCTGACAAGTATAGaTGCAGGTTCGGATCCTGCCCGGGGAA
scaffold_59	SINE-Asp	Asp	GTC	406230	406273	A	TCCCCGGTAGCTCAATTGGTAGAGCATGCCGCTGTCAcatggcagacccaggttcgaatcacggattcggccgggttgaggCTGACAAGTATAGaTGCAGGTTCGGATCCTGCCCGGGGAA
scaffold_59	SINE-Asp	Asp	GTC	464906	464949	A	TCCCCGGTAGCTCAATTGGTAGAGCATGCCGCTGTCAcatggcagacccaggttcgaatcacggattcggccgggttgaggCTGACAAGTATAGaTGCAGGTTCGGATCCTGCCCGGGGAA
scaffold_1	SINE-Asp	Asp	ATC	6483869	6483819	A	GGGGGGGTAGCTCAGTAGGTaAGAGCACTTCCTTATCAccctgcggacccgggttcgaatctcatattcggcccgtttcccggcggataAGGTTGAGGtCGTGGGTTCGGATCCCACCCCCCTCA
scaffold_59	SINE-Asp	Asp	GTC	31219	31176	A	TCCCCGGTAGCTCAATTGGTAGAGCATGCCGCTGTCAcatggcagacccaggttcgaatcgcggattcggccgggttgaggCTGACAAGTATAGaTGCAGGTTCGGATCCTGCCCGGGGAA

Table S4. Gene model generation: Gene models in the Frozen Gene Catalog are categorized with respect to the ways in which they were generated. Generation of the model was through homology, *ab initio* predictions, correspondence with ACEGs and ESTs, or mapping of previous models by fgenesh. Some models were generated by users or carried over from assembly v2.0 of the *Chlamydomonas* assembly.

Models	**Number**	**Percentage**
Homology based models	3,022	20
ab initio prediction	6,619	44
Transfers (mapping) of models from chlamy portal version 2.0 to 3.0	3,137	21
ACEGs-based models	439	3
'Known' genes - mapped (not predicted) by fgenesh+	1,112	7
EST based models	201	1
User created models	613	4
Total	15,143	100

Table S5. Support for gene model assignment: The table lists the various methods and tools that support the generation of gene models.

Supporting Evidence	**Number**	**Percentage**
Clustered ESTs support	8,522	56
Swissprot homologs Evalue < 10^{-5}	9,558	63
NR homologs Evalue < 10^{-5}	8,845	58
Pfam domains	6,161	41
Ostreococcous best hits	2,223	15
Cyanidioschyzon best hits	275	2
Greenplants/algae best hits	5,335	35
Fungi/Metazoa best hits	1,729	11
Bacteria, mostly cyanobacteria best hits	1,156	8
ab initio models without support	1,843	12
Manually curated *ab initio* models without support	309	2
Manually assigned name	3914	26

Table S6. Functional assignment of gene models from KOG, GO and KEGG analyses

Functional assignment category	Number	Percentage	Distinct categories
Unique KOG assignments, E-value < 10^{-5}	9,435	62	3,158
Unique Gene Ontology (GO) assignments	6,733	44	3,165
Unique KEGG/EC assignments (60% ID 60% coverage)	2,780	18	798

Table S7. Large protein families: Families of paralogous proteins within each species were made with MCL I=2.0 (*45*); PFAM domains (*41*) were assigned to proteins achieving a score <1e–10 with RPSblast (*23*). Protein families were ranked by size. The table lists the top 20 families based on the number of members in each. Representative PFAM domains are given with PF numbers and descriptions.

Rank	No. of members	Associated protein domain
1	51	PF00211: adenylyl and guanylyl cyclase catalytic domain
2	44	PF00125: core histone H2A/H2B/H3/H4
3	39	PF00125: core histone H2A/H2B/H3/H4
4	35	PF00125: core histone H2A/H2B/H3/H4
5	35	PF00125: core histone H2A/H2B/H3/H4
6	29	PF00069: protein kinase domain
		PF07714: protein tyrosine kinase
7	22	PF00233: 3'5'-cyclic nucleotide phosphodiesterase
8	20	PF00025: ADP-ribosylation factor family
9	15	PF00069: protein kinase domain
		PF07714: protein tyrosine kinase
10	14	PF03110: SBP domain
11	14	PF00069: protein kinase domain
		PF07714: protein tyrosine kinase
12	14	IPR002290: serine/threonine protein kinase
13	14	PF00071: Ras family
14	14	PF00179: ubiquitin-conjugating enzyme
15	13	PF00067: cytochrome P450
16	12	PF00160: cyclophilin type peptidyl-prolyl cis-trans isomerase

17	12	PF03171: 2OG-Fe(II) oxygenase superfamily
18	12	PF07714: protein tyrosine kinase
19	11	PF00651: BTB/POZ domain
20	11	PF00249: Myb-like DNA-binding domain
21	11	PF01384: phosphate transporter family
22	11	PF00226: DnaJ domain
23	10	PF03016: exostosin family
24	10	PF00240: ubiquitin family
25	10	PF00504: chlorophyll a/b binding protein
26	10	PF00168: C2 domain

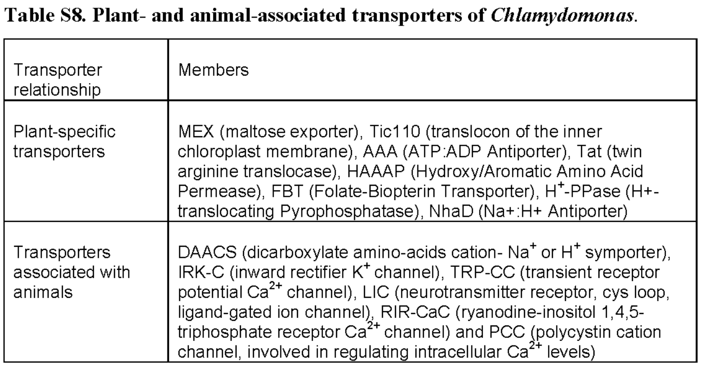


Table S8. Plant- and animal-associated transporters of *Chlamydomonas*.

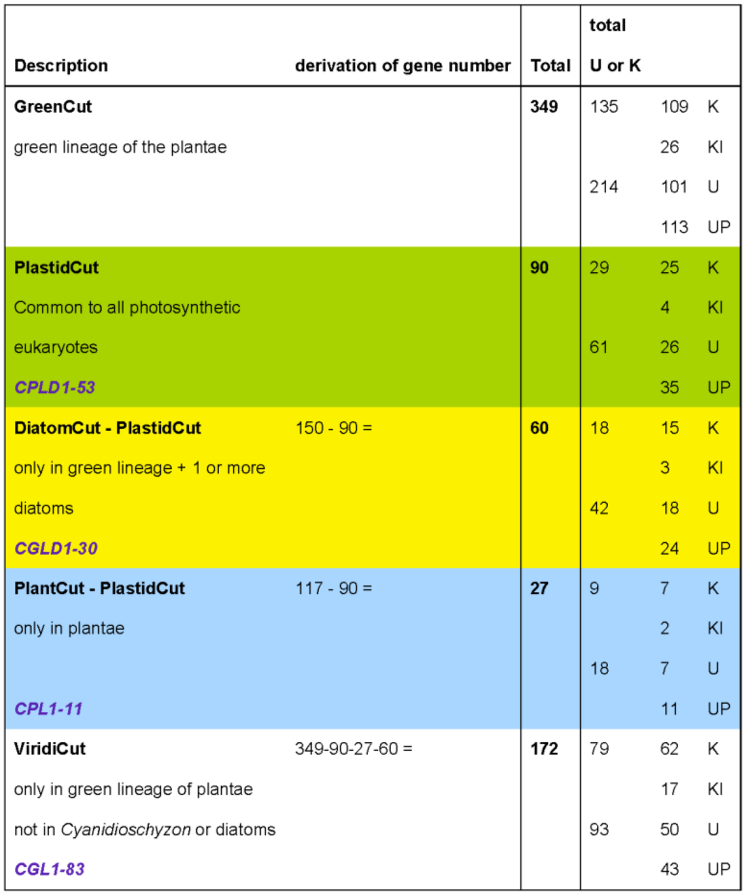
Transporter relationship	Members
Plant-specific transporters	MEX (maltose exporter), Tic110 (translocon of the inner chloroplast membrane), AAA (ATP:ADP Antiporter), Tat (twin arginine translocase), HAAAP (Hydroxy/Aromatic Amino Acid Permease), FBT (Folate-Biopterin Transporter), H^+-PPase (H+-translocating Pyrophosphatase), NhaD (Na+:H+ Antiporter)
Transporters associated with animals	DAACS (dicarboxylate amino-acids cation- Na^+ or H^+ symporter), IRK-C (inward rectifier K^+ channel), TRP-CC (transient receptor potential Ca^{2+} channel), LIC (neurotransmitter receptor, cys loop, ligand-gated ion channel), RIR-CaC (ryanodine-inositol 1,4,5-triphosphate receptor Ca^{2+} channel) and PCC (polycystin cation channel, involved in regulating intracellular Ca^{2+} levels)

Table S9. *Chlamydomonas* protein families similar to those in human or *Arabidospis*: Selected proteins (from scatter plot of **Fig. 4A**), with closer similarity to human (top half) or *Arabidopsis* (bottom half) polypeptides but that are not members of phylogenomic or experimental groupings. Also given are the PFAM descriptions, JGI protein IDs and notes related to their potential functions.

PFAM description	**PFAM or KOG ID**	**JGI v3.0 protein ID (gene name)**	**notes**
Animal-associated proteins			
Tubulin-tyrosine ligase family	PF03133	100760, 146893, 118345, 119250, 126569	Likely associated with flagellar function
Kinesin-associated protein (KAP)	PF05804	182554 (KAP1)	Likely associated with flagellar function
Dynein heavy chain	PF03028	130324 (DHC2)	Associated with flagellar function
Ion transport protein	PF00520	179342, 189093, 192415, 144131, 180826, 144354, 170854, 194450, 194451	Voltage-gated Na^+/Ca^{2+} ion channels; 194450, 194451 are adjacent on the genome; possibly involved in flagellar signaling
Pyridoxal-dependent decarboxylase	PF00278, PF02784	206067 (ODC1), 206062 (ODC2)	
Vitamin B12 dependent methionine synthase; Homocysteine S methyltransferase	PF02965, PF02574	76715 (METH1)	Cobalamin-dependent methionine synthase (METH), which is not found in vascular plants (*84*)
Selenocysteine-specific elongation factor	KOG0461	112829	The selenocysteine specific elongation factor, which is not found in vascular plants
Adenylate and guanylate cyclase catalytic domain	PF00211	193525 (CYG41), 187517 (CYG12)	See text above
Plant-associated proteins			

Ammonium transporter family	PF00909	182688 (AMT1D), 192308 (AMT1A), 183975 (AMT1B)	Similar to ammonium transporter AMT1 in *Arabidopsis*
S1 RNA binding domain UBA/TS-N domain Elongation factor TS	PF00575 PF00627 PF00889	195616 (EFT1)	EF-Ts; Chloroplast small ribosomal subunit protein *PSRP-7* and elongation factor Ts are encoded in this single transcript

Table S10. Proteins in the GreenCut and their division into subgroups: The 349 proteins of the GreenCut were selected based on phylogenetic analyses as described in the Main Text. These were classified as either known (K) or unknown (U) with respect to function. The designation was based on experimental work in the literature for either *Arabidopsis* or *Chlamydomonas* proteins. The modifier I for the K category indicates a function that is known by "inference" (based on a strong sequence identity and full coverage along its length to a protein in a related organism whose function is known). The modifier P for the U category stands for "Predicted" where the gene product is predicted to have a particular enzymatic activity or the sequence contains a structural motif. The distinction between KI and UP may be occasionally blurred because the classifications were made subjectively based on evaluation of the body of literature. Restricting the GreenCut only to those proteins conserved in at least one diatom yielded the DiatomCut with 150 proteins. Restricting the GreenCut only to those proteins conserved in plants yielded the PlantCut with 117 proteins. Restricting the GreenCut only to those proteins conserved in photosynthetic eukaryotes, which include diatoms and plants, yielded the PlastidCut with 90 proteins. The corresponding genes were named according to these groupings unless they had been previously named during manual curation. The name designation *CPL* was given (for conserved in the plant lineage) to genes encoding proteins in the GreenCut that are conserved also in *Cyanidioschyzon* but not in the diatoms, *CPLD* (for conserved in the plant lineage and diatoms) to genes corresponding to proteins in the GreenCut that are conserved in *Cyanidioschyzon* and at least one diatom (PlastidCut), CGLD (for conserved in the green lineage and diatoms) for genes encoding proteins conserved in the GreenCut plus at least one diatom, and CGL (for conserved in the green lineage) for those in the GreenCut that are not present in either *Cyanidioschyzon* or a diatom. This grouping was also designated the ViridiCut. Also see **Fig. 5** and **Supplemental File 1**.



			total		
Description	**derivation of gene number**	**Total**	**U or K**		
GreenCut		**349**	135	109	K
green lineage of the plantae				26	KI
			214	101	U
				113	UP
PlastidCut		**90**	29	25	K
Common to all photosynthetic				4	KI
eukaryotes			61	26	U
CPLD1-53				35	UP
DiatomCut - PlastidCut	150 - 90 =	**60**	18	15	K
only in green lineage + 1 or more				3	KI
diatoms			42	18	U
CGLD1-30				24	UP
PlantCut - PlastidCut	117 - 90 =	**27**	9	7	K
only in plantae				2	KI
			18	7	U
CPL1-11				11	UP
ViridiCut	349-90-27-60 =	**172**	79	62	K
only in green lineage of plantae				17	KI
not in *Cyanidioschyzon* or diatoms			93	50	U
CGL1-83				43	UP

Table S11. Proteins of known function in the GreenCut: Selected chloroplast proteins of known function in the GreenCut are grouped by general function. We excluded proteins of the photosynthetic apparatus, which had been used to estimate the false negative fraction in the GreenCut (see above); these are listed in **Supplemental File 1**. The enzymes LL-diaminopimelate aminotranferase and TGD2 (involved in lipid transfer from the endoplasmic reticulum) are unique to plants, while RPPK2 (phosphoribosyl diphosphate synthase), TAL2 (transaldolase), DLA2, DLD2 (of the pyruvate dehydrogenase complex) and ADCL1 (aminodeoxychorismate lyase) represent plastid-specific isoforms (*85-88*).

Function	**Associated gene products**
Regulation of photosynthesis	PGR5, STT7, RCA2, APE1
Thylakoid membrane biogenesis	CCS1, HCF164, CCB factors, SUFD, EGY1, TAB2, MCA1, CSP41a, THF1
Plastid biogenesis	TOCs, TIC110, TIC40, HSPs, CYNs, FKBPs, CLP subunits, PRORS1
Plastid division	MINE1
Lipid biosynthesis	FAB2, LPAAT, KAS1, DGD1, FAT1, PLSB1
Other carbon metabolism	DLA2, DLD2, TAL2, MDH5, RPI2
Amino acid, nucleotide biosynthesis	CGL37 (shikimate kinase), RPPK2, DPR1, DPA1
Starch biosynthesis	STA6, STA11, STA1, PWD1, SSS2, AMYB1
Pigment, cofactor biosynthesis	CTH1, GUN4, DVR, UROD1, HMOX1, LCYE, ADCL1, CHLD, CAO
Metabolite transporters	LCI20, CEM1, RCP1, TPT3
Anti-oxidant pathways	GSH1, APXs, CDSP32, TRXL/HCF164, SNE1

Table S12. Proteins of unknown function in the GreenCut: Proteins of the GreenCut with unknown functions are tabulated with potential activities associated with these proteins based on annotations of the *Chlamydomonas* genome at (*15*) and the *Arabidopsis* genome (*89*). Note the striking representation of redox-active proteins, proteins that might function in isoprenoid metabolism and proteins from the plastoglobule/eyespot proteomes (see Fig. S1).

Functional Group	***Chlamydomonas* Protein name**	**Description**
SOUL proteins	SOUL4 SOUL5	Related to chicken heme protein identified in retina and pineal gland (which contain light-cued circadian clocks) Also, SOUL3 is found in *Chlamydomonas* eyespot and in *Arabidopsis* plastoglobule
Redox active proteins	TRXL1	Thioredoxin-like protein, unusual active site WCNAC
	TRX10	Thioredoxin-like protein, unusual active site WCPKC
	CITRX	Cytoplasmic in tomato, but highly conserved in the green lineage and diatoms
	CPLD41	Protein disulfide isomerace-like motif + VitK epoxide reductase motif, conserved in cyanobacteria.
	GRX6	Glutaredoxin, CGFS type, probably chloroplastic
	CPLD26	related to pyridoxamine 5' phosphate oxidase
	CPLD32	FAD dependent oxidoreductase
	CPLD49	saccharopine dehydrogenase-like
	CPLD25	short-chain dehydrogenase/reductase
	TEF5	Rieske [2Fe-2S] domain
Isoprenoid pathway	CPLD35	flavin containing amine oxidase related to phytoene desaturase
	VDR1	violaxanthin de-epoxidase related
	CPLD27	coclaurine N-methyl transferase
	CGL2	ubiquinol methyl transferase
	CPLD34	ubiquinol methyl transferase
	AKC1	ABC1 kinases. The mitochondrial homolog regulates UQ

	AKC2 AKC3 AKC4	biosynthesis. A *Chlamydomonas* AKC is the product of the EYE3 locus, required for assembly of the carotenoid pigmented eyespot. ORFs in cyanobacteria with very strong sequence similarity.
	PLAP1 PLAP2 PLAP3 PLAP4	plastid lipid associated protein or Plastoglobulins, conserved in cyanobacteria
Transporters	CPLD21 CPLD22 CPLD23	sugar nucleotide transporters, solute carriers
	ARSA	anion transporter
	CGL51	plastid metabolite exchanger
	CGL7	plastid metabolite exchanger
	CGLD4	ABC transporter
	CGL15	major facilitator superfamily
	MITC4	mitochondrial carrier
	TIM22B	plastid homolog of TIM17/22/23 family
Various metabolic reactions	CPLD3	aldo-keto isomerase
	SNE3	NAD-dependent epimerase/dehydratase
	CGLD13	related to nucleoside diphosphate sugar epimerase, putative chloroplast targeted
	CGL2	methyltransferase
	CGL33A/B	methyl transferase
	CGL75	methyl transferase motif
	CGL77	methyl transferase
	CGLD2	thioesterase
	CGLD24	thioesterase
	CGLD7	esterase / lipase / thioesterase
	CGL69	lipase
	CPLD15	lipase
	CGLD15	related to triacylglycerol lipase
	CGL76	esterase, epoxide hydrolase
	CPLD2	hydrolase
	CGL53	related to carbohydrate hydrolase
	CPLD4	inositol monophosphatase-related

	CGL14	pantothenate kinase motif
	CGL79	carbohydrate kinase motif
	CGLD12	potential galactosyl transferase activity
	CGLD24	related to diacylglycerol acyl transferase
	RIBFL1	related to riboflavin biosynthesis protein RibF
	CGL48	related to lysine decarboxylase domain
biogenesis and nucleic acid transactions	CPLD17	organelle-targeted protein, related to OTU-like cysteine protease family
	CPLD6	metal-dependent CAAX amino terminal protease family
	HEP2	Hsp70 escorting protein 2
	CPLD43	YGGT family
	RNB2	3'-5' Exoribonuclease II
	CPLD16	organelle-targeted, RNA methyl transferase related
	CGL43	RNA binding protein with S1 domain
	CGL72	hemolysin motif and RNA methyltransferase motif
	TPR2	tetratricopeptide repeat protein, organelle-targeted
	CGL71	TPR repeat protein related to YCF37
	CPLD46	DEAD/DEAH-box helicase possibly plastid targeted
	CGLD3	DEAD/DEAH box helicase domain and proline rich domain
	CGLD5A	ethylene response element dna binding domain containing protein
	CGLD5B	AP2-domain transcription factor
	CPL2	transcription factor like protein
	CGLD30	SET domain containing protein, putative histone methyltransferase
	CGL31	pterin carbinolamine dehydratase domain
	CGL49	ARF/SAR superfamily small monomeric GTP binding protein
Regulation	PP2C4	related to protein phosphatase 2C
	PP2C5	related to protein phosphatase 2C
	PP2C6	related to protein phosphatase 2C
	CPL3	related to protein serine / threonine phosphatase
	MAPK2	Mitogen-Activated Protein Kinase Homolog 2
	STPK25	MUT9 related serine/threonine protein kinase
Photosynthesis	CPLD45	possible function in PSII and possible lumen location

Table S13. Subcellular localization of proteins in the GreenCut: The experimental or predicted localization of the proteins in each group (known K, unknown U, which also includes both known inferred, KI, and unknown predicted, UP) is indicated as follows: cp, chloroplast; mito, mitochondrion; other, all other compartments; not known, no data and no prediction. For the known group, the subcellular location is experiment-based for 73% of the proteins. For the unknown group the subcellular location is experiment-based for only 15% of the proteins.

GreenCut			cp	mito	other	unknown
349	135	K+KI	115	3	9	8
	214	U+UP	113	36	19	46

Table S14. CiliaCut proteins: Protein designations, association with flagella, or a specific sub- structure of the flagella, basal body. intraflagellar transport and/or affiliations with specific organisms are given.

Category	Members	Significance
Motililty-associated (MotileCut)	PF16, PF20, KLP1 and hydin	central pair proteins
	RSP3 and RSP9	radial spoke proteins
	DHC2, DHC6 (inner dynein arm components), ODA4, ODA6 (outer dynein arm components), ODA1 (the outer dynein arm docking complex protein), and PF2 (component of the dynein regulatory complex)	
Outer dynein arm proteins lost in moss *Physcomitrella*	ODA4, ODA6, ODA9, DLC1 and DLC4	
DiatomCut	anterograde motor (KAP) and complex B (IFT57, IFT74, IFT81, IFT88)	Intraflagellar transport proteins present in centric diatom *Thalassiosira*
	retrograde motor (represented by D1bLIC) and complex A (represented by IFT140)	Intraflagellar transport proteins lost in centric diatom *Thalassiosira*
Comparison to *Ostreococcus*	ODA1, ODA4, ODA6, ODA9, Tctex1 DHC2, DHC6, RSP3, RSP9, PF16, PF20, KLP1, hydin, KAP, D1bLIC, IFT20, IFT52, IFT57, IFT74, IFT80, IFT81, IFT88, IFT140, IFT172, RIB43a, PKD2, FAPs 9, 21, 22, 32, 36, 43, 46, 47, 50, 60, 61, 66, 69, 73, 74, 75, 81, 94, 100, 111, 116, 118, 122, 134, 146, 155, 156, 161, 184, 198, 240, 251, 253, 259, 263, 264, 247	Flagellar proteins lost in *Ostreococcus*
	MKS1, NPH4, BLD1, BLD2, UNI3, POC11, POC18, FBB5, 9, 11, 15, and all of the BBS proteins (BBS2, 3, 5, 7, 8, 9)	Basal body proteins lost in *Ostreococcus*
	RIB72, PF2, MBO2, DLC1, PACRG1, DIP13, FAPs 14, 44, 45, 52, 57, 59, 67, 82, 106, 250, 267, and POC1	Flagellar proteins retained in *Ostrecoccus*

5. SUPPORTING REFERENCES AND NOTES

1. P. Kathir *et al.*, *Eukaryot Cell* **2**, 362 (2003).
2. J. Shrager *et al.*, *Plant Physiol* **131**, 401 (2003).
3. M. Jain *et al.*, *Nucleic Acids Res* (2007).
4. T. Proschold, E. H. Harris, A. W. Coleman, *Genetics* **170**, 1601 (2005).
5. Chlamydomonas Resource Center, http://www.chlamy.org/libraries.html (2007).
6. E. Asamizu, Y. Nakamura, S. Sato, H. Fukuzawa, S. Tabata, *DNA Res* **6**, 369 (1999).
7. J. L. Weber, E. W. Myers, *Genome Res* **7**, 401 (1997).
8. B. Ewing, L. Hillier, M. C. Wendl, P. Green, *Genome Res* **8**, 175 (1998).
9. S. Aparicio *et al.*, *Science* **297**, 1301 (2002).
10. NCBI, http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Root&id=1&lvl=3&p=mapview&p=has_linkout&p=blast_url&p=genome_blast&lin=f&keep=1&srchmode=3&unlock (2007).
11. W. J. Kent, *Genome Res* **12**, 656 (2002).
12. W. Gish, http://blast.wustl.edu (1996-2004).
13. DOE Joint Genome Institute, http://genome.jgi-psf.org/finished_microbes/raleu/raleu.home.html (2007).
14. DOE Joint Genome Institute, www.jgi.doe.gov/poplar (2007).
15. DOE Joint Genome Institute, www.jgi.doe.gov/chlamy (2007).
16. J. C. Detter *et al.*, *Genomics* **80**, 691 (2002).
17. DOE Joint Genome Institute, http://www.jgi.doe.gov/sequencing/protocols/index.html (2007).
18. K. D. Pruitt, T. Tatusova, D. R. Maglott, *Nucleic Acids Res* **33**, D501 (2005).
19. B. Ewing, P. Green, *Genome Res* **8**, 186 (1998).
20. E. Sobel, H. M. Martinez, *Nucleic Acids Res* **14**, 363 (1986).
21. T. F. Smith, M. S. Waterman, *J Mol Biol* **147**, 195 (1981).
22. A. F. A. Smit, R. Hubley, P. Green, http://www.repeatmasker.org (1996-2004).
23. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *J Mol Biol* **215**, 403 (1990).
24. E. Birney, M. Clamp, R. Durbin, *Genome Res* **14**, 988 (2004).
25. A. A. Salamov, V. V. Solovyev, *Genome Res* **10**, 516 (2000).
26. M. Kanehisa, *Trends Genet* **13**, 375 (1997).
27. M. Kanehisa, S. Goto, *Nucleic Acids Res* **28**, 27 (2000).
28. M. Kanehisa *et al.*, *Nucleic Acids Res* **34**, D354 (2006).
29. R. L. Tatusov *et al.*, *BMC Bioinformatics* **4**, 41 (2003).
30. E. Quevillon *et al.*, *Nucleic Acids Res* **33**, W116 (2005).
31. M. Ashburner *et al.*, *Nat Genet* **25**, 25 (2000).
32. R. L. Tatusov *et al.*, *Nucleic Acids Res* **29**, 22 (2001).
33. NCBI homepage, http://www.ncbi.nlm.nih.gov/ (2007).
34. J. Jurka, P. Klonowski, V. Dagman, P. Pelton, *Comput Chem* **20**, 119 (1996).
35. J. Jurka *et al.*, *Cytogenet Genome Res* **110**, 462 (2005).
36. T. M. Lowe, S. R. Eddy, *Science* **283**, 1168 (1999).
37. P. Schattner *et al.*, *Nucleic Acids Res* **32**, 4281 (2004).

38. C. L. Chen *et al.*, *Nucleic Acids Res* **31**, 2601 (2003).
39. Z. P. Huang *et al.*, *Rna* **11**, 1303 (2005).
40. Q. Ren, K. H. Kang, I. T. Paulsen, *Nucleic Acids Res* **32**, D284 (2004).
41. E. L. Sonnhammer, S. R. Eddy, E. Birney, A. Bateman, R. Durbin, *Nucl Acids Res* **26**, 320 (1998).
42. A. Krogh, B. Larsson, G. von Heijne, E. L. Sonnhammer, *J Mol Biol* **305**, 567 (2001).
43. Q. Ren, I. T. Paulsen, *PLoS Comput Biol.* **1**, 190 (2005).
44. TransportDB, http://www.membranetransport.org/ (2007).
45. A. J. Enright, S. Van Dongen, C. A. Ouzounis, *Nucleic Acids Res* **30**, 1575 (2002).
46. M. Matsuzaki *et al.*, *Nature* **428**, 653 (2004).
47. E. Derelle *et al.*, *Proc Natl Acad Sci U S A* **103**, 11647 (2006).
48. DOE Joint Genome Institute, www.jgi.doe.gov/Olucimarinus (2007).
49. DOE Joint Genome Institute, www.jgi.doe.gov/Otauri (2007).
50. The Arabidopsis Genome Initiative, *Nature* **408**, 796 (2000).
51. DOE Joint Genome Institute, www.jgi.doe.gov/Physcomitrella (2007).
52. DOE Joint Genome Institute, http://genome.jgi-psf.org/finished_microbes /prom9/prom9.home.html (2007).
53. Broad Institute, http://www.broad.mit.edu/annotation/microbes/methanosarcina/ (2007).
54. LBMGE Orsay, http://www-archbac.u-psud.fr/projects/sulfolobus/ (2007).
55. DOE Joint Genome Institute, www.jgi.doe.gov/Pramorum (2007).
56. DOE Joint Genome Institute, www.jgi.doe.gov/Psojae (2007).
57. E. V. Armbrust *et al.*, *Science* **306**, 79 (2004).
58. DOE Joint Genome Institute, www.jgi.doe.gov/phaeodactylum (2007).
59. dictyBase, http://dictyBase.org (2007).
60. L. Eichinger *et al.*, *Nature* **435**, 43 (2005).
61. Broad Institute, http://www.broad.mit.edu/annotation/genome/neurospora /Home.html (2007).
62. EnsEMBL, http://www.ensembl.org/Caenorhabditis_elegans/index.html (2007).
63. M. Remm, C. E. Storm, E. L. Sonnhammer, *J Mol Biol* **314**, 1041 (2001).
64. B. M. Tyler, *Annu Rev Phytopathol* **40**, 137 (2002).
65. T. Avidor-Reiss *et al.*, *Cell* **117**, 527 (2004).
66. J. B. Li *et al.*, *Cell* **117**, 541 (2004).
67. G. J. Pazour, N. Agrin, J. Leszyk, G. B. Witman, *J Cell Biol* **170**, 103 (2005).
68. L. C. Keller, E. P. Romijn, I. Zamora, J. R. Yates, 3rd, W. F. Marshall, *Curr Biol* **15**, 1090 (2005).
69. P. Divina, J. Forejt, *Nucleic Acids Res* **32**, D482 (2004).
70. P. N. Inglis, K. A. Boroevich, M. R. Leroux, *Trends Genet* **22**, 491 (2006).
71. V. Stolc, M. P. Samanta, W. Tongprasit, W. F. Marshall, *Proc Natl Acad Sci U S A* **102**, 3703 (2005).
72. P. J. Ferris, *Genetics* **122**, 363 (1989).
73. T. M. Lowe, S. R. Eddy, *Nucleic Acids Res* **25**, 955 (1997).
74. F. J. Sun, S. Fleurdepine, C. Bousquet-Antonelli, G. Caetano-Anolles, J. M. Deragon, *Trends Genet* **23**, 26 (2007).

75. T. A. Allen, S. Von Kaenel, J. A. Goodrich, J. F. Kugel, *Nat Struct Mol Biol* **11**, 816 (2004).
76. C. Marck, H. Grosjean, *Rna* **8**, 1189 (2002).
77. T. Wen, I. A. Oussenko, O. Pellegrini, D. H. Bechhofer, C. Condon, *Nucleic Acids Res* **33**, 3636 (2005).
78. A. Theologis *et al.*, *Nature* **408**, 816 (2000).
79. A. Hinas *et al.*, *Eukaryot Cell* **5**, 924 (2006).
80. M. W. Murcha *et al.*, *Plant Physiol* **143**, 199 (2007).
81. S. Merchant, M. R. Sawaya, *Plant Cell* **17**, 648 (2005).
82. D. DellaPenna, B. J. Pogson, *Annu Rev Plant Biol* **57**, 711 (2006).
83. M. Lohr, C. S. Im, A. R. Grossman, *Plant Physiol* **138**, 490 (2005).
84. M. T. Croft, A. D. Lawrence, E. Raux-Deery, M. Warren, A. G. Smith, *Nature* **483**, 90 (2005).
85. B. N. Krath, T. A. Eriksen, T. S. Poulsen, B. Hove-Jensen, *Biochim Biophys Acta* **1430**, 403 (1999).
86. I. Lutziger, D. J. Oliver, *FEBS Lett* **484**, 12 (2000).
87. A. O. Hudson, B. K. Singh, T. Leustek, C. Gilvarg, *Plant Physiol* **140**, 292 (2006).
88. G. J. Basset *et al.*, *Plant J* **40**, 453 (2004).
89. TAIR, http://www.arabidopsis.org/ (2007).

About the Authors

Daniel S. Rokhsar : Email: dsrokhsar@lbl.gov (D.S.R.)

Arthur R. Grossman : Email: arthurg@stanford.edu (A.R.G.)