Green Evolution and Dynamic Adaptations Revealed by Genomes of the Marine Picoeukaryotes *Micromonas*

Alexandra Z. Worden[1,*], Jae-Hyeok Lee[2,†], Thomas Mock[3,†‡], Pierre Rouzé[4,†], Melinda P. Simmons[1,†], Andrea L. Aerts[5], Andrew E. Allen[6], Marie L. Cuvelier[1,7], Evelyne Derelle[8], Meredith V. Everett[7], Elodie Foulon[9], Jane Grimwood[5,10], Heidrun Gundlach[11], Bernard Henrissat[12], Carolyn Napoli[13], Sarah M. McDonald[1], Micaela S. Parker[3], Stephane Rombauts[4], Aasf Salamov[5], Peter Von Dassow[9], Jonathan H. Badger[6], Pedro M. Coutinho[11], Elif Demir[1], Inna Dubchak[5], Chelle Gentemann[14], Wenche Eikrem[15], Jill E. Gready[16], Uwe John[17], William Lanier[18], Erika A. Lindquist[5], Susan Lucas[5], Klaus F. X. Mayer[10], Herve Moreau[8], Fabrice Not[9], Robert Otillar[5], Olivier Panaud[19], Jasmyn Pangilinan[5], Ian Paulsen[20], Benoit Piegu[19], Aaron Poliakov[5], Steven Robbens[4], Jeremy Schmutz[5,10], Eve Toulza[21], Tania Wyss[22], Alexander Zelensky[23], Kemin Zhou[5], E. Virginia Armbrust[3], Debashish Bhattacharya[18], Ursula W. Goodenough[2], Yves Van de Peer[4], Igor V. Grigoriev[5]

[1]Monterey Bay Aquarium Research Institute, Moss Landing, CA 95039 USA; [2]Washington University in St. Louis, St. Louis, MO 63130, USA; [3]University of Washington, Seattle, WA, USA; [4]Department of Plant Systems Biology, Flanders Institute for Biotechnology (VIB) and Department of Molecular Genetics, Ghent University, 9052 Gent, Belgium; [5]U.S. Department of Energy (DOE) Joint Genome Institute (JGI), Walnut Creek, CA 94598, USA; [6]J. Craig Venter Institute, San Diego, CA 92121,USA; [7]Rosenstiel School of Marine and Atmospheric Science, University of Miami, Miami, FL 33149, USA; [8]Observatoire Océanologique, CNRS-Université Pierre et Marie Curie, 66651 Banyuls sur mer, France FR; [9]Station Biologique de Roscoff, Centre National de la Recherche Scientifique et Université Pierre et Marie Curie, Roscoff Cx, France; [10]MIPS/IBIS Institute for Bioinformatics and System Biology, German Research Center for Environmental Health (GmbH), 85764 Neuherberg, Germany; [11]Architecture et Fonction des Macromolécules Biologiques, Universities of Aix-Marseille I & II Marseille, France; [12]Alfred Wegener Institute for Polar and Marine Research, Am Handelshafen Bremerhaven, DE; [13]Biology Institute, University of Arizona, Tucson, AZ 85719, USA, US; [14]Remote Sensing Systems, Santa Rosa, CA 95401, USA; [15]Avdeling for Marinbiologi og Limnologi, University of Oslo, Oslo, Norway; [16]Division of Molecular Bioscience, College of Medicine, Biology and the Environment, Australian National University, Canberra, Australia; [17]Department of Biology, University of Iowa, Iowa City, IA 52242, USA; AU; [18]Laboratoire Genome et Development des Plantes Université de Perpignan, 66860 Perpignan, FR.; [19]Department of Chemistry and Biomolecular Sciences, Macquarie University, NSW 2109, AU; [20]Ecosystèmes lagunaires, Université Montpellier II, F-34095 Montpellier Cedex 05, FR; [21]Department of Biology, University of Miami, Miami, FL 33149, USA; [22]Department of Genetics, Erasmus Medical Center, Rotterdam, The Netherlands; [23]Stanford Human Genome Center, Stanford University School of Medicine, Palo Alto, CA 94304, USA.

APRIL 2009

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231

DISCLAIMER

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor The Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or The Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or The Regents of the University of California.

Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes *Micromonas*

A.Z. Worden[1]†, J-H. Lee[2§], T. Mock[3§*], P. Rouzé [4§], M.P. Simmons[1§], A.L. Aerts[5], A.E. Allen[6], M.L. Cuvelier[1,7], E. Derelle[8], M.V. Everett[7], E. Foulon[9], J. Grimwood[5,23], H. Gundlach[10], B. Henrissat[11], C. Napoli[13], S.M. McDonald[1], M.S. Parker[3], S. Rombauts[4], A. Salamov[5], P. Von Dassow[9], J.H. Badger[6], P.M. Coutinho[11], E. Demir[1], I. Dubchak[5], C. Gentemann[14], W. Eikrem[15], J.E. Gready[16], U. John[12], W. Lanier[17], E.A. Lindquist[5], S. Lucas[5], K.F.X. Mayer[10], H. Moreau[8], F. Not[9], R. Otillar[5], O. Panaud[18], J. Pangilinan[5], I. Paulsen[19], B. Piegu[18], A. Poliakov[5], S. Robbens[4], J. Schmutz[5,23], E. Toulza[20], T. Wyss[21], A. Zelensky[22], K. Zhou[5], E.V. Armbrust[3], D. Bhattacharya[17], U.W. Goodenough[2], Y. Van de Peer[4] and I.V. Grigoriev[5]

*Present address: University of East Anglia, Norwich, NR47TJ, UK.
†To whom correspondence should be addressed. E-mail: azworden@mbari.org
[§]These authors contributed equally to this work (alphabetical order)

[1]Monterey Bay Aquarium Research Institute, Moss Landing, CA 95039 USA; [2]Washington University in St. Louis, St. Louis, MO 63130, USA; [3]University of Washington, Seattle, WA, USA; [4]Department of Plant Systems Biology, Flanders Institute for Biotechnology (VIB) and Department of Molecular Genetics, Ghent University, 9052 Gent, Belgium; [5]U.S. Department of Energy (DOE) Joint Genome Institute (JGI), Walnut Creek, CA 94598, USA; [6]J. Craig Venter Institute, San Diego, CA 92121,USA; [7]Rosenstiel School of Marine and Atmospheric Science, University of Miami, Miami, FL 33149, USA; [8]Observatoire Océanologique, CNRS-Université Pierre et Marie Curie, 66651 Banyuls sur mer, France FR; [9]Station Biologique de Roscoff, Centre National de la Recherche Scientifique et Université Pierre et Marie Curie, Roscoff Cx, France; [10]MIPS/IBIS Institute for Bioinformatics and System Biology, German Research Center for Environmental Health (GmbH), 85764 Neuherberg, Germany; [11]Architecture et Fonction des Macromolécules Biologiques, Universities of Aix-Marseille I & II Marseille, France; [12]Alfred Wegener Institute for Polar and Marine Research, Am Handelshafen Bremerhaven, DE; [13]Biology Institute, University of Arizona, Tucson, AZ 85719, USA, US; [14]Remote Sensing Systems, Santa Rosa, CA 95401, USA; [15]Avdeling for Marinbiologi og Limnologi, University of Oslo, Oslo, Norway; [16]Division of Molecular Bioscience, College of Medicine, Biology and the Environment, Australian National University, Canberra, Australia; [17]Department of Biology, University of Iowa, Iowa City, IA 52242, USA; AU; [18]Laboratoire Genome et Development des Plantes Université de Perpignan, 66860 Perpignan, FR.; [19]Department of Chemistry and Biomolecular Sciences, Macquarie University, NSW 2109, AU; [20]Ecosystèmes lagunaires, Université Montpellier II, F-34095 Montpellier Cedex 05, FR; [21]Department of Biology, University of Miami, Miami, FL 33149, USA; [22]Department of Genetics, Erasmus Medical Center, Rotterdam, The Netherlands; [23]Stanford Human Genome Center, Stanford University School of Medicine, Palo Alto, CA 94304, USA.

Abstract

Picoeukaryotes are a taxonomically diverse group of organisms less than two micrometers in diameter. Photosynthetic marine picoeukaryotes in the genus *Micromonas* thrive from tropical to polar ecosystems and serve as sentinel organisms for biogeochemical fluxes of modern oceans during climate change. These broadly distributed primary producers belongs to an anciently diverged sister clade to land plants. Although *Micromonas* isolates have high 18S rDNA identity, we found that genomes from two isolates shared only 90% of their predicted genes. Their independent evolutionary paths were emphasized by riboswitch arrangements as well as intronic repeat elements discovered in just one isolate and in metagenomic data, but not other genomes. Divergence appears to have been facilitated by selection and acquisition processes that actively shape the 'unique' gene pools of each differently than core genes. Analyses of the *Micromonas* genomes offer valuable insights into ecological differentiation and the dynamic nature of early plant evolution.

Ancestral green algae were of fundamental importance to the eukaryotic 'greening' that shaped the geochemistry of our planet. This process began over a billion years ago when a cyanobacterium was captured by a heterotrophic protist and incorporated as an endosymbiont, giving rise to the first alga (*1*). The extant Prasinophytae retain characteristics believed present in the last common ancestor of green algae (chlorophytes) and land plants (streptophytes, including charophyte algae) (*2*). Most prasinophytes within the monophyletic marine order Mamiellales (Fig. 1a, fig. S1), such as *Micromonas*, are tiny (≤2 μm diameter) and known as picoeukaryotes. *Micromonas* is a motile unicell, with a single chloroplast and mitochondrion (Fig. 1a inset), first reported as a dominant phytoplankter in the 1950s (*3*) and now recognized as having a global distribution (Fig. 1b) (*4*).

Today's oceans contain a polyphyletic diversity of algae, some with plastids that share ancestry with land-plants (green algae) and others ('chromalveolates') derived from red algae through secondary or tertiary (eukaryotic-eukaryotic) endosymbioses (*5, 6*). Unlike most episodic chromalveolate bloomers and the freshwater green alga *Chlamydomonas* (*7*), the Mamiellales have reduced genomes, as first shown for *Ostreococcus* (*8, 9*). *Ostreococcus* has a narrower environmental distribution than *Micromonas* (Fig. 1b.) and a small genome (12-13 Mb containing only ~8000 genes). Open-ocean bacteria, including SAR11 and *Prochlorococcus* (*10, 11*), show similar patterns of cell-size and genome minimization. Conditions favoring picophytoplankton growth, such as increased stratification, less mixing, and reduced nutrient concentrations in ocean surface waters are predicted climate-change outcomes, and thus picoeukaryote dynamics may be useful ecosystem indicators.

We sequenced the nuclear genomes of *Micromonas* isolates RCC299 and CCMP1545 (Table 1, fig. S2 and S3) (*12*). These isolates are from distant ocean provinces and fall into distinct

phylogenetic clades that can co-occur (*12, 13*) (Fig. 1) but are generally considered a single species (*Micromonas pusilla*). TEM revealed no morphological differences (*12*) and 18S rDNA identity was high (97%). Surprisingly, only 90% of their 10,056 (RCC299) and 10,575 (CCMP1545) predicted genes (table S1) were shared (Fig. 2a). In contrast, *Ostreococcus lucimarinus* and *O. tauri* share 97% of cataloged genes (*12*), and yeast genera can share ~95% of homologs (*14*). The divergence we observed between the *Micromonas* isolates supports their classification as distinct species.

Synteny, GC-content and codon usage pointed to a shared evolutionary history for RCC299 and CCMP1545, but underscored their genomic divergence (text S1). Each genome contained a region with 14% lower than average GC-content, composing 7% (RCC299) and 8% (CCMP1545) of the genome (figs. S3 to S4), and where transcriptional activity was higher (text S1). Similar regions in *Ostreococcus* (*8, 9*) comprise smaller genome proportions. DNA alignment between RCC299 and CCMP1545 low GC-region(s) was poor, protein colinearity absent and codon usage different, in contrast to normal GC-chromosomes (figs. S4 to S6).

Two major evolutionary themes emerged from our analyses. First, the common ancestor of the Mamiellales had already undergone genomic reduction, highlighted by their organellar genomes (text S2, fig. S7, tables S2 to S4). Second, *Micromonas* appeared to be less derived than *Ostreococcus*, rendering insights into genetic composition of the proto-prasinophyte (the common ancestor of plants and prasinophytes) and specialization in extant species. Most 'core' nucleus-encoded genes (common to the 4 Mamiellales genomes) were found to have known functions (Fig. 2a, b) in key pathways (text S3 to S6, table S5 to S9, fig. S8) such as photosynthesis, and included selenoproteins (text S3, table S10). A significant proportion of genes grouped with land plants (Fig. 2c). Core genes branching with chromalveolates (Fig. 2c,

mostly diatoms and brown algae) presumably reflected losses (or extensive divergence) in other green lineage organisms and red algae or perhaps horizontal gene transfer (HGT).

The proto-prasinophyte features we discovered in *Micromonas*, included transcription factors likely belonging to the "basal green toolkit" (text S7, figs. S9 to S11, table S11). For example, early-branching land plants encode most higher-plant transcription factor families except for the YABBY family (*15*), which was therefore posited to be evolutionarily associated with leaf development. However, we found YABBY in *Micromonas*, although it is absent from *Chlamydomonas* and *Ostreococcus*, indicating it was part of the basal toolkit (fig. S11). We also found diversified homeodomains (fig. S12, table S12), relevant to the evolution of green regulatory networks.

Although prasinophytes are often considered asexual, our observations indicated that the proto-prasinophyte was sexual. First, meiotic-specific and non-meiotic representatives of the *RECA-RAD51*, *TOP6A/SPO11* and *MUTS* gene families were found (text S5, table S13). Second, the low GC-regions showed features of sex chromosomes, including RWP-RK transcription factor family genes (text S7, table S14). Third, numerous Mamiellales genes encoded hydroxyproline-rich glycoproteins (HRGP; text S6, table S15, fig. S13), which are cell-wall components in *Chlamydomonas* and plants (*16*). Like the many carbohydrate-active enzymes (text S6, table S17), this was unexpected because cell walls have not been observed in *Micromonas* or *Ostreococcus* (e.g., Fig. 1a inset) (*4*). In *Chlamydomonas*, one HRGP gene set is expressed only after sexual fusion to produce a thick, adhesive zygote wall (*17*). *Micromonas* may behave similarly. Collectively, these data indicate the occurrence of sexual differentiation and formation of a resistant life-cycle stage.

Fourteen percent of genes were 'shared' between RCC299 and CCMP1545, but not with

Ostreococcus (Fig. 2, text S3 and S8, table S18, fig. S14). Shared enzymes for the synthesis and remodelling of peptidoglycan in the plastid provided new insight into the evolutionary history of the ancestral cyanobacterial endosymbiont (text S6) (*18, 19*). The shared genes also showed more rapid evolutionary rates than core genes (fig. S15) indicating that they escaped constraints acting on the Mamiellales core but still likely play important roles given their presence in both isolates. Moreover, a larger proportion of 'unique' (mutually exclusive between RCC299 and CCMP1545) genes branched with opisthokont or bacterial lineages (Fig. 2c), consistent with acquisition by horizontal gene transfer. Many were of unknown function (Fig. 2b), but may provide useful indicator information. Following early genome reduction, fundamentally different selection/acquisition processes acting on the unique genes appear to have promoted differentiation.

Marked differences in nutrient transport were seen compared with other green-lineage organisms. Between the *Micromonas* species, 52 of the 59 transporter gene families common to land plants were present as well as several transporter gene families found in marine chromalveolates but not in other green-lineage members (text S9, table S19). Both *Micromonas* spp. had more transporter families represented and higher numbers of transporters than *Ostreococcus*, although CCMP1545 was missing specific transporter gene families including some related to nitrogen uptake (text S9, table S19). These differences possibly reflected environmental parameters, since RCC299 is from highly oligotrophic waters where nutrient scavenging is essential.

We explored other genomic features related to competition and mortality that influence community structure (text S10 to S13, figs. S16 to S18). Two types of carbon-concentrating mechanisms (CCM) were identified (text S12, figs. S17 and S18), that can alleviate CO_2

limitation during blooms. The more unusual *Micromonas* CCM, a C_4-like carbon fixation pathway, includes a novel NADP-dependent Malic-Enzyme (NADP-ME) targeted to the plastid lumen, a localization that likely reduces CO_2 leakage (text S12). Since C_4-like pathways have now been identified in the four Mamiellales genomes and in diatoms (text S12) they may represent a fairly basic necessity rather than a rare form of optimization in a few taxa. Both *Micromonas* species appeared to have more robust defenses against heavy metal toxicity and reactive oxygen species (text S13, table S20) than *Ostreococcus.* The larger *Micromonas* genome sizes may thus facilitate broader physiological response capabilities than its smaller relative.

We found few (CCMP1545, table S21) or no (RCC299) recognizably functional Transposable Elements (TEs). Most eukaryotes, including *Ostreococcus* (*9*), contain many TEs, and TE content is positively correlated with genome size above a ~10 Mb threshold (*20, 21*). Any relic or degenerate TEs in *Micromonas* had low similarity to known TEs, and structural features of class II elements were not found. GC bias was thought responsible for the high proportion of TEs in the low GC-region(s) of *Ostreococcus* and for loss of synteny in these regions (*9*). However, the low GC-region(s) of *Micromonas*, although rearranged (fig. S5), had few simple repeats, contained only potential relic TEs, and showed high transcriptional activity (text S1, theoretically facilitating TE insertion), suggesting TE activity/propagation is actively hindered.

We discovered intronic repeat sequences in CCMP1545 that were absent from RCC299 and other published genomes (text S14, tables S22 and S23, figs. S19 to S22). These abundant "Introner Elements" (IEs) were located within introns, extending nearly to donor and acceptor sites (Fig. 3, figs. S21), and lacked known TE characteristics (*22*). RCC299 genes generally had fewer introns than IE-bearing CCMP1545 homologs (e.g., Fig. 3), and CCMP1545 had higher

intron frequency overall (Table 1). The 9,904 IEs fell into four heterogeneously distributed subfamilies (fig. S22, table S22) comprising 9% of the genome. We also found IEs in Sargasso Sea metagenome data (*23*) with flanking coding domains of high similarity to CCMP1545 but lower similarity to RCC299. *Micromonas* 18S rDNA sequences in the same metagenome data belong to uncultured clade M_IV (Fig 1a) (*13*). Given the extent of genome reduction, the abundance of IE suggests they are functional or resistance to purging.

Putative RNA interference (RNAi) components also differed between the *Micromonas* species (text S4, table S6). Only RCC299 had an Argonaute-encoding gene. A version of Argonaute is also found in *Chlamydomonas* and plants, but not *Ostreococcus*. DEAD Box and SDE3 gene analyses provided circumstantial evidence for a diverged RCC299 RNA helicase. Argonaute can act to combat TE invasion (*24*), which is notable given that RCC299 had no recognizable TEs or IEs.

Both *Micromonas* spp. had putative thiamine pyrophosphate (TPP) riboswitches, untranslated mRNAs that regulate gene expression by metabolite binding (*25, 26*). These were not associated with homologous genes nor with known thiamine-biosynthesis related genes, except for *NMT1* (Table 2, text S15). CCMP1545 riboswitches were located at both gene ends (e.g., Fig. 4a), an arrangement never before seen, and formed two divergent groups, 5′ riboswitches shared with *Ostreococcus* and 3′ riboswitches shared with RCC299 (Fig. 4b). A conserved 3′ riboswitch was shared between *FOLR*-like (RCC299) and *SSSF-P* (CCMP1545), even though these genes were not held in common, yet *Ostreococcus* also had *SSSF-P* and a 5′ riboswitch (Fig. 4a). Only one of the seven *Micromonas* riboswitches was associated with a multi-exon gene (*FOLR*-like). Thus it appears that the putative riboswitches in *Micromonas* act akin to bacterial riboswitches and lack the spliceosomal functions that evolved in other eukaryotes (*26*).

Deficiencies in the thiamine-biosynthesis pathway (*27, 28*) were notable (text S15). However, comparison with other lineages indicated the *Micromonas* riboswitch-containing genes represent ancient thiamine pathway components. We identified TPP riboswitches associated with *SSSF-P* in SAR11 bacteria, which also lack classical thiamine-biosynthesis genes (*10*), and with *SSSF-F* in *Chlamydomonas* and *Volvox*. The functional importance of the gene-riboswitch associations is supported by the same gene-riboswitch pairings being found in disparate lineages (text S15).

The *Micromonas* genomes reveal features of the ancestral algae that initiated the billion-year trajectory of the green lineage and the greening of Earth. Their divergence, combined with acquisition strategies consistent with horizontal gene transfer, highlight the dynamic nature of marine protistan evolution and provide a springboard for unraveling functional aspects of phytoplankton populations. The challenge now is to identify biogeochemically important features within this natural diversity and apply them in assessing ecological transformations caused by environmental change.

Supporting Online Material

Materials and Methods

SOM text

Tables S1 to S25

References

Figs. S1 to S22

References and Notes

1. J. D. Hackett *et al.*, *Mol. Biol. Evol.* **24**, 1702 (2007).
2. L. A. Lewis, R. M. McCourt, *Am. J. Bot.* **91**, 1535 (2004).
3. E. W. Knight-Jones, P. R. Walne, *Nature* **167**, 445 (1951).
4. A. Z. Worden, F. Not, in *Microbial Ecology of the Oceans,* D. L. Kirchman, Ed. (Wiley, Hoboken, 2008), pp. 594.
5. A. Reyes-Prieto, A. P. Weber, D. Bhattacharya, *Annu. Rev. Genet.* **41**, 147 (2007).
6. C. E. Lane, J. M. Archibald, *Trends Ecol. Evol.* **23**, 268 (2008).

7. S. S. Merchant *et al.*, *Science* **318**, 245 (2007).
8. E. Derelle *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 11647 (2006).
9. B. Palenik *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 7705 (2007).
10. S. J. Giovannoni *et al.*, *Science* **309**, 1242 (2005).
11. G. Rocap *et al.*, *Nature* **424**, 1042 (2003).
12. Materials and methods and supplemental (SOM) text, table and figures are available as supporting materials at *Science* online.
13. A. Z. Worden, *Aquat. Microbial Ecol.* **43**, 165 (2006).
14. C. Hall, S. Brachat, F. S. Dietrich, *Eukaryot Cell* **4**, 1102 (2005).
15. S. K. Floyd, J. L. Bowman, *Int. J. Plant Sci.* **168**, 1 (2007).
16. J. H. Lee, S. Waffenschmidt, L. Small, U. Goodenough, *Plant Physiol*. **144**, 1813 (2007).
17. U. Goodenough, H. Lin, J. H. Lee, *Semin. Cell Dev. Biol*. **18**, 350 (2007).
18. T. Cavalier-Smith, *Trends Plant Sci.* **5**, 174 (2000).
19. G. I. McFadden, *Curr. Opin. Plant Biol.* **2**, 513 (1999).
20. B. Gaut, J. Ross-Ibarra, *Science* **320**, 484 (2008).
21. M. Lynch, *The Origins of Genome Architecture,* (Sinauer, Stamford, 2007), pp. 389.
22. T. Wicker *et al.*, *Nature Rev. Genet.* **8**, 973 (2007).
23. J. Venter *et al.*, *Science* **304**, 66 (2004).
24. A. A. Aravin, G. J. Hannon, J. Brennecke, *Science* **318**, 761 (2007).
25. A. Wachter *et al.*, *Plant Cell* **19**, 3437 (2007).
26. M. T. Cheah, A. Wachter, N. Sudarsan, R. R. Breaker, *Nature* **447**, 497 (2007).
27. M. T. Croft, M. Moulin, M. E. Webb, A. G. Smith, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 20770 (2007).
28. D. DellaPenna, R. L. Last, *Science* **320**, 479 (2008).
29. We thank the CCMP and RCC for providing isolates, in particular, F. LeGall, A. Houdan and D. Vaulot. We also thank R. Gausling, C. Perle, Q. Ren, D. Root, L. Stal, J. Van Wye, T. Weissman, R.M. Welsh and U. Wollenzien. F. Partensky, N. Simon, S. Ball facilitated chloroplast and starch (performed by P. Deschamps, published elsewhere) annotations, C. Rancurel and B. Cantarel assisted with CAZymes (CNRS funding). We are grateful to S. Giovannoni for thoughtful criticism of the manuscript and overall enthusiasm. Genome sequencing was performed under the DOE Biological and Environmental Research Program contracts DE-AC02-05CH11231, DE-AC52-07NA27344, DE-AC02-06NA25396 and DEFC02-99ER62873. UWG and J-HL were funded by NSF MCB 0326829. Funding carrying the project from inception to completion was provided by a Young Investigator in Marine Microbiology award to AZW from the Gordon and Betty Moore Foundation, with additional funds from NSF MCB 0429359 and the Lucille and David Packard Foundation.

AZW coordinated the project and annotation; AZW and UWG wrote the manuscript with input and sections from J-HL, TM, PR and MPS (joint second authors listed in alphabetical order) while YvP and DB performed intellectually-based editing to which S Rombauts and MSP contributed; IVG coordinated the sequencing and analysis at JGI. ALA, AEA, MLC, E Derelle, MVE, EF, JG, HG, BH, CN, SMM, MSP, S Rombauts, AS, PvD also made significant contributions (listed in alphabetical order). JHB and AEA constructed the phylogenomic analysis tool. AZW, EVA, KFXM, UWG and YvP supervised analyses; AZW conceived the study with input from DB, HM, and EVA. All others contributed as members of the *Micromonas* genome consortium or JGI sequencing

and are listed in alphabetical order. RCC299 and CCMP1545 assemblies and annotations are available at www.jgi.doe.gov/MicomonasRCC299 and www.jgi.doe.gov/MpusillaCCMP1545, respectively. Genome assemblies together with predicted gene models and annotations were deposited at DDBJ/EMBL/GenBank under the project accessions ACCO00000000 and ACCP00000000 for RCC299 and CCMP1545, respectively.

Table 1. Characteristics of the *Micromonas* genomes.

Characteristic	CCMP1545	RCC299
Genome size (Mb)	21.9	20.9
G+C (%)	65	64
Number of genes	10,575	10,056
Gene size (bp)	1,557	1,587
Multiexon genes (%)	50	37
Introns ($gene^{-1}$)	0.90	0.57
Intron length (bp)	187	163

Table 2. Genes with associated TPP riboswitches in RCC299, CCMP1545 and *Ostreococcus* (both *O. tauri* and *O. lucimarinus*). The position of the riboswitch, relative to the gene, is indicated in the column entitled "Riboswitch." Abbreviations: DC, domain containing; NF, not found by BLASTP or TBLASTN. See text S15 for gene descriptions.

	RCC299			**CCMP1545**			***Ostreococcus***		
Gene name	ProtID	Riboswitch		ProtID	Riboswitch		Presence	Riboswitch	
		5′	3′		5′	3′		5′	3′
NMT1	102273	no	yes	58387	no	no	NF	-	-
FOLR-like	106264	no	yes	NF	-	-	NF	-	-
EFG-DC	56895	no	yes	NF	-	-	NF	-	-
SSSF-F	NF	-	-	48760	yes	yes	yes	yes	no
SSSF-P	NF	-	-	60112	yes	yes	yes	yes	no

Figure legends

Figure 1. *Micromonas* phylogeny and distribution. **(A)** A consensus neighbor-joining (NJ) distance 18S rRNA gene tree illustrating the distinct *Micromonas* clades *(12)*. Bootstrap values represent percent of 1000 replicates (NJ), and, where provided, the second value represents the maximum likelihood bootstrap percentages. The genome isolates sequenced in this work are highlighted (yellow). The previously sequenced *Ostreococcus tauri* and *O. lucimarinus* neighbor each other in clade O_I. Relationship to plants and other photosynthetic lineages is shown in fig S1. **Inset,** *Micromonas* thin section showing nucleus (n) chloroplast (c), flagellum (f) and mucronate extension (the thin tip at the end of the flagellum, arrow). **(B)** Mean sea surface temperature (SST) for 2006, using Global High-Resolution SST (GHRSST) blended infrared and microwave SSTs, and locations where *Micromonas* (solid pins and circles around the isolates used in this work) and *Ostreococcus* (dashed lines) 18S rDNA sequences have been recovered. *Micromonas* appeared in all temperature regimes.

Figure 2. Comparison of Mamiellales gene complements. **(A)** Venn diagram comparing RCC299 and CCMP1545, *O. tauri* and *O. lucimarinus* gene complements. Circle sizes roughly represent relative numbers of genes in each genome. **(B)** Proportions of genes within EuKaryotic Orthologous Groups (KOGs) and without KOG placement for the gene pools: **unique**, genes in one *Micromonas* species only and not the other Mamiellales (proportions shown are for RCC299, see fig. S14 for CCMP1545); **shared**, genes in both *Micromonas* species but neither *Ostreococcus*; and **core**, found in the 4 Mamiellales genomes. **(C)** Phylogenomic profiling for core, shared and unique genes as percentage of gene pool affiliated (≥50% bootstrap values) with different lineages.

Figure 3. Depiction of *Micromonas* orthologs with and without Introner Elements (IE). Single-exon (horizontal green bars represent exons) RCC299 (Prot. ID 84234, chromosome 8)

corresponds to a multi-exon gene in CCMP1545 (Prot. ID 70142, scaffold 11). Different IE elements are shown (red, orange) within introns (thin green lines). Diagonally oriented green lines show syntenic relationships by connecting exons with >70% nucleotide identity (minimum 100 bp). Purple (RCC299, reversed orientation) and blue (CCMP1545) curves/peaks represent 16-mer frequencies.

Figure 4. Thiamine pyrophosphate (TPP) riboswitch arrangements. **(A)** High nucleotide identity of 3′ riboswitch sequences (yellow profiles) associated with *FOLR*-like (pink, RCC299 only) and *SSSF-P* (light blue, CCMP1545) and identity between CCMP1545 and *Ostreococcus* 5′ riboswitches (white profiles) associated with *SSSF-P* homologs (light blue). Plant and bacterial riboswitches are often located in 3′ UTRs (*25*) and fungal riboswitches in 5′ UTRs. CCMP1545 has them in both positions. The downstream gene (purple) is a putative dihydrouridine synthase conserved in the four Mamiellales genomes. **(B)** Secondary structure of *FOLR*-like-associated riboswitch showing the positions that are conserved among a range of organisms, particularly plants (yellow background), and a conserved position in all known plant riboswitches but not conserved in *Micromonas* (pink boxed 'U'). Nucleotides adjacent to P2, P4 and P5 regions reflect differences in the CCMP1545 *SSSF-P* 3′ riboswitch (light blue) and CCMP1545 *SSSF-F* 5′ riboswitch (brown). Differences in the more variable P1 and P3 are not marked to maintain figure simplicity.

Figure 1, Worden et al.

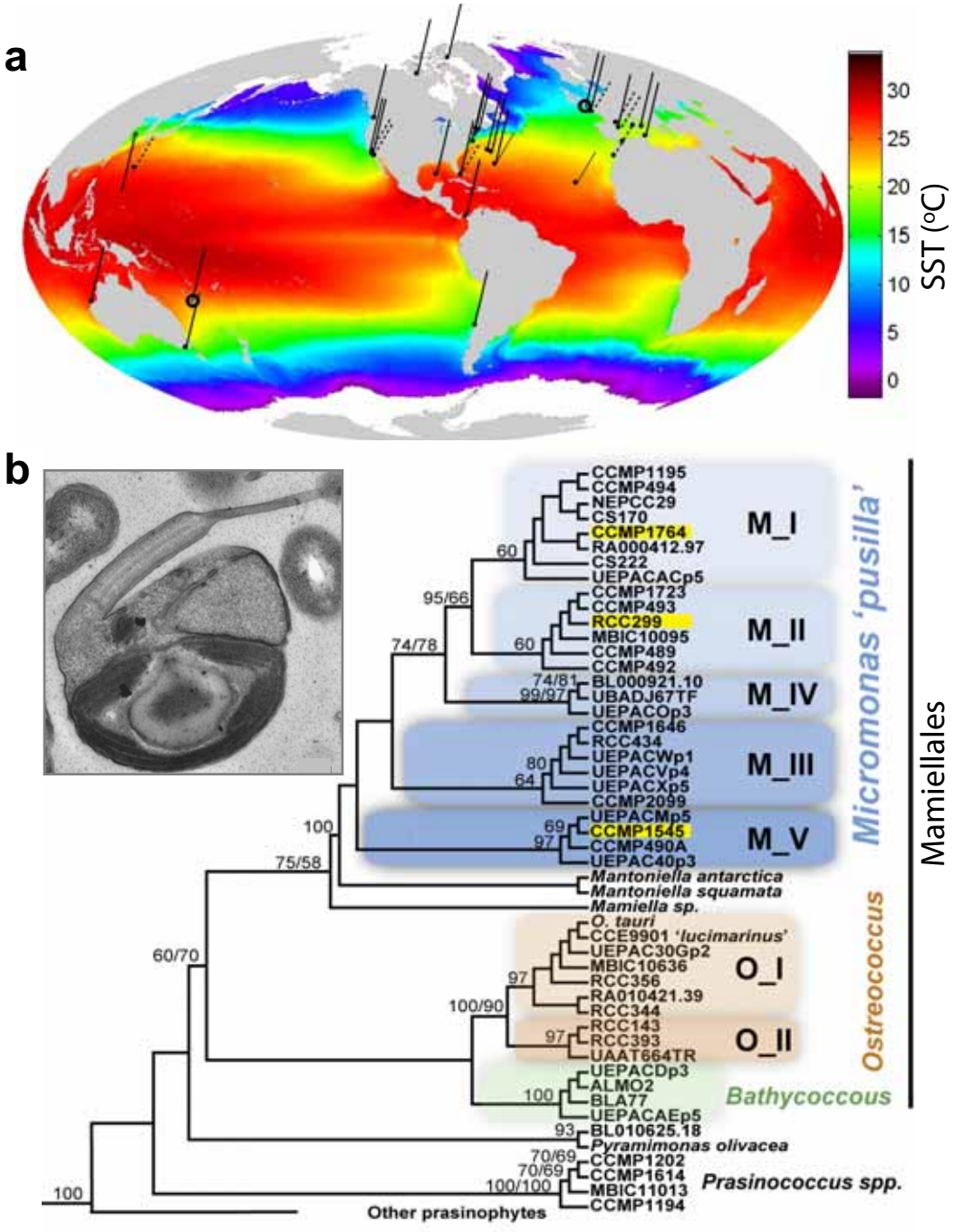


Figure 2, Worden et al.

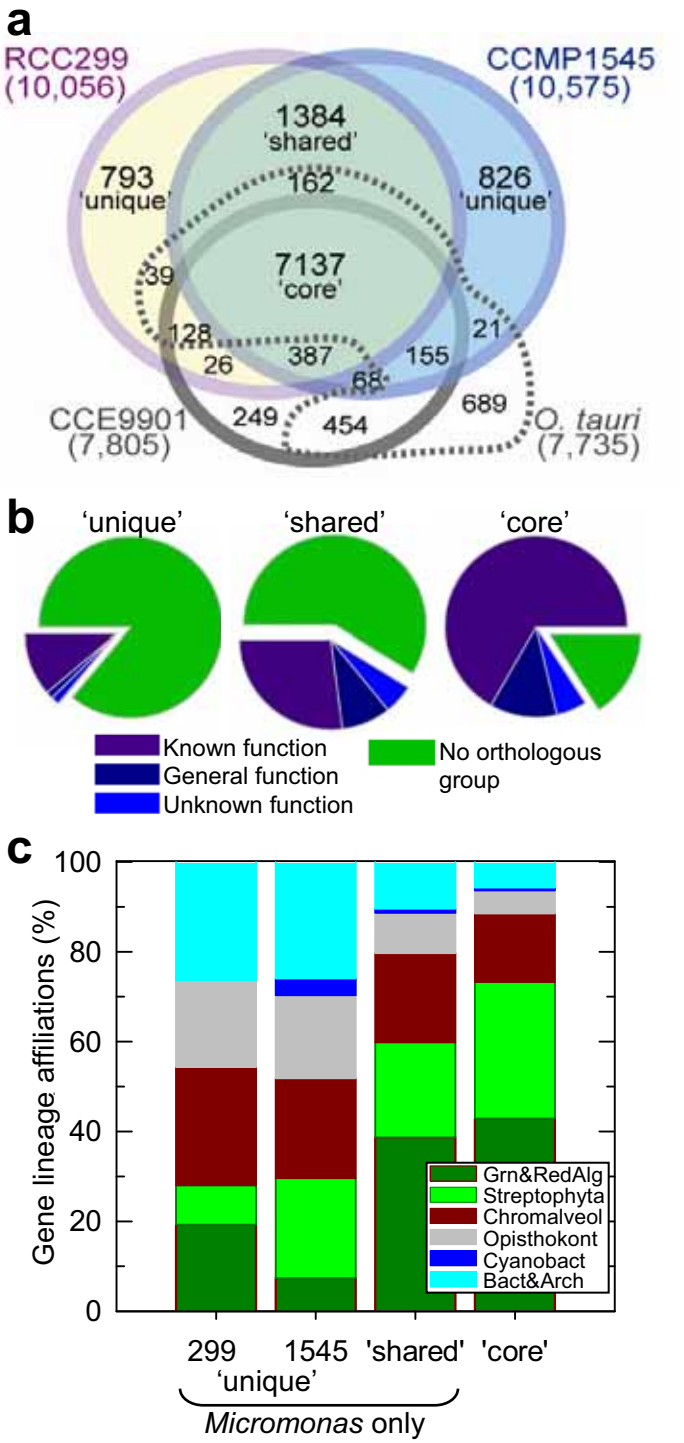


Figure 3, Worden et al.

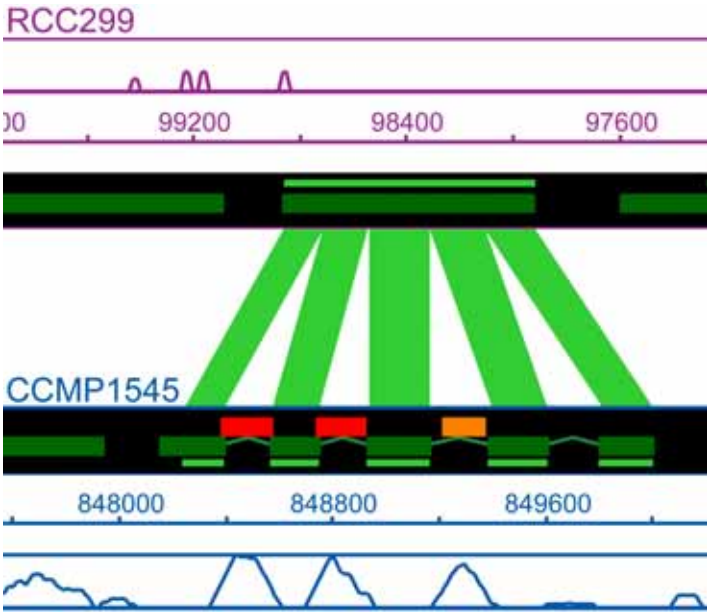
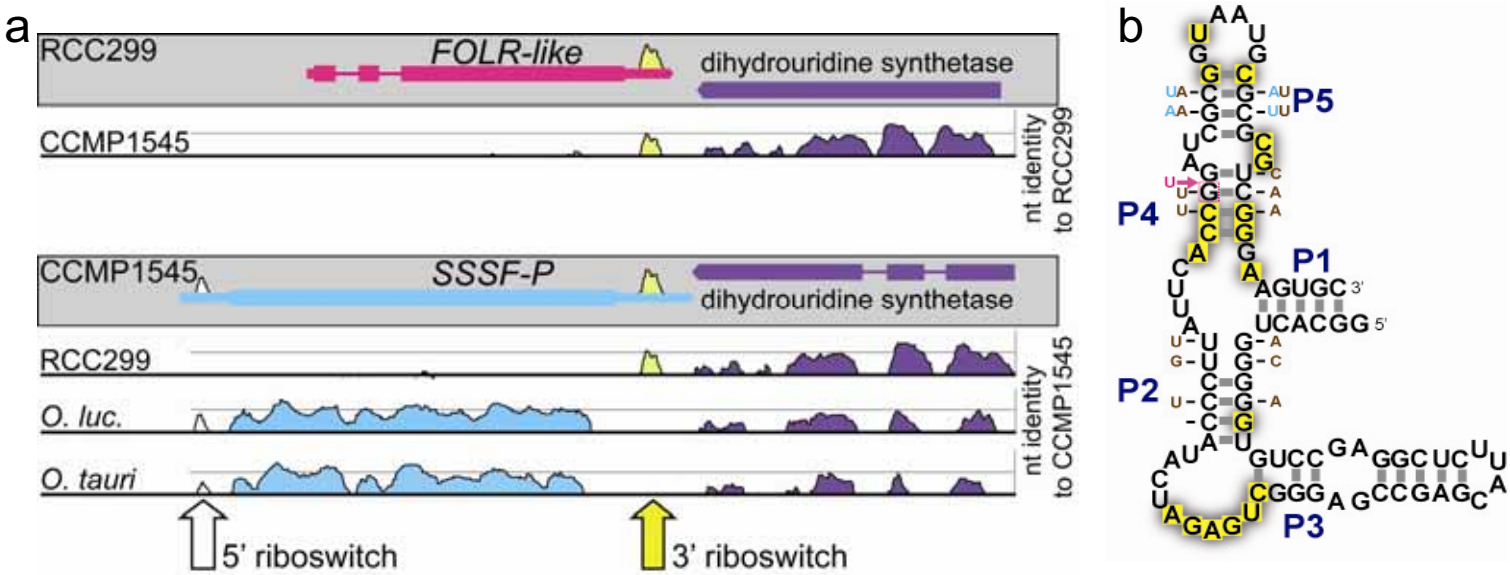


Figure 4, Worden et al.



Supporting Online Material for

Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes *Micromonas*

Worden Alexandra Z,* et al.

*To whom correspondence should be addressed. E-mail: azworden@mbari.org

This PDF file includes
The *Micromonas* Genomes: SUPPLEMENTAL MATERIAL

TABLE OF CONTENTS

MATERIALS and METHODS (pp 4-10)

SUPPORTING TEXT (pp 11-29)
Note that supporting texts are meant to accompany the main text sections in which they are referenced.

1.) Genome structure, gene rearrangements and expression analysis p 11
2.) Mitochondrial and Chloroplast genomes p 12
3.) Gene pools (core, shared, unique) and selenoproteins p 13
4.) Chromatin and RNAi p 14
5.) Sex p 14
6.) Carbohydrate active enzymes (CAZy) and HRGPs p 18
7.) Transcription factors p 19
8.) Flagella related genes p 21
9.) Nutrient acquisition and transport p 22
10.) Forces of mortality p 22
11.) Polyketide synthetases p 23
12.) Carbon concentrating mechanisms (CCMs) p 24
13.) Mechanisms for alleviating oxidative stress p 26
14.) Introner Elements p 26
15.) TPP riboswitches and thiamine biosynthesis p 27

SUPPLEMENTARY TABLES (pp 30-84)

1.) Summary of gene predictions, functional annotations and comparisons p 30
2.) Mitochondrial encoded genes p 31
3.) Chloroplast encoded genes p 32
4.) Percent 'missing positions' in chloroplast genomes p 33
5.) Carbon metabolism pathway genes p 34
6.) Chromatin and RNAi associated genes p 35
7.) Core cell cycle genes p 40
8.) Photosynthesis related genes p 41
9.) Chlorophyll and carotenoid biosynthesis related genes p 43
10.) Selenoproteins in *Micromonas* and other green lineages p 45
11.) Transcription factor families p 47
12.) Homeodomain transcription factors p 51
13.) Sex – meiosis p 52
14.) RWP-RK family genes p 53
15.) HRGP genes p 54
16.) Gene models used in fig. S13 p 55
17.) Carbohydrate active enzyme (CAZy) encoding genes p 56
18.) Flagella related genes p 68

19.) Analysis of membrane transporter families p 71
20.) Reactive oxygen scavenging (ROS) related genes p 74
21.) Transposable elements in CCMP1545 p 76
22.) Introner elements in CCMP1545 p 77
23.) Repeat Elements in CCMP1545 and RCC299 p 81
24.) Sequence read statistics p 83
25.) Assembly statistics p 83

SUPPORTING REFERENCES (pp 84-87)

SUPPORTING FIGURES (figs. S1 to S22 follow references)

MATERIALS AND METHODS

Culturing and strain purification

The two *Micromonas* isolates, CCMP1545 and RCC299, were obtained from the Center for Culture of Marine Phytoplankton (CCMP) and Roscoff Culture Collection (RCC), respectively. CCMP1545 was originally isolated by M. Parke in 1950 at approximately 50°36' N, 04°17' W (English Channel/North Atlantic waters near Plymouth, England) in 1950 and was axenic upon receipt. RCC299 (also known as NOUM17) was isolated from surface waters in 1998 at 166°20' E, 22°20' S (South Pacific) by S. Boulben. This isolate was rendered clonal and axenic through a series of plating and antibiotic treatments. Subsequently, the axenic RCC299 was then deposited at the CCMP (CCMP ID: CCMP2709) and redeposited at the RCC (RCC ID: RCC827), although it has not been maintained in an axenic state at the latter. Strains were grown in K or L1 media (*1*), with CCMP1545 performing well in L1 made with Sargasso Sea water as a base seawater source and RCC299 growing well in K made with artificial seawater as a base (http://www.mbari.org/phyto-genome/resources). Standard growth conditions are defined as 21°C at approximately 200 μEin m^{2} sec^{-1} photosynthetically active radiation (PAR). Two approaches were used to verify that cultures were contamination free: 1) inoculation into Test Media (http://www.mbari.org/phyto-genome/resources), with incubation for days, weeks and months in the dark, and 2) DAPI staining (as in (*2*)) followed by visual inspection by epifluorescence microscopy.

Pulse Field Gel Electrophoresis

After flow-cytometric titration, RCC299 and CCMP1545 cells were harvested by centrifugation at 8000 x *g* for 20 min and embedded in low melting point agarose. The embedded cells were digested by proteinase K (1 mg ml^{-1}, Sigma) at 37°C for 24h and then analysed by PFGE as described previously (*3*). From 1 x 10^{7} to 5 x 10^{7} cells were loaded per lane in a 0.8% agarose gel (Type D-5, Euromedex France); the electrophoresis parameters were 6 V cm^{-1}, 0.5X TBE, 120° switching angle, 14°C and switch times of 60-120s for 24h. The gel was stained with ethidium bromide to visualise chromosomal bands and to compare karyotypes.

Electron Microcopy

Thin sections for transmission electron microscopy were prepared following the protocol of Eikrem and Moestrup 1998 (*4*). Our image(s) of *Micromonas* show features previously recognized as 'characteristic,' i.e. *Micromonas* is characterized by having one flagellum, and unlike most prasinophytes it lacks scales. The shape of the cell, the length of the flagellum and its mucronate extension (Fig. 1a, arrow) may vary between cells, but consistent morphological differences between RCC299 and RCC834 (which is CCMP1545, as provided by the RCC, and shown in Fig. 1a inset) have not been identified.

DNA, genome sequencing and assembly

DNA isolation was performed using a modification of a previously published CTAB extraction procedure (*5*). The initial data sets of sequence reads for RCC299 and CCMP1545 were derived from 3 whole-genome shotgun (WGS) libraries (insert sizes of 1-3 KB, 6-8 KB, and 35-40 KB). These were screened for vector using cross_match, trimmed for vector and quality (*6*), filtered for reads shorter than 100 bp (table S24), and finally assembled using JAZZ, a WGS assembler developed at the JGI (*6, 7*).

The genome sizes and sequence depth were initially estimated to be 23-24.5 MB and 8.0x, respectively. Initial assemblies were filtered to remove short (<1 KB) and redundant scaffolds (<5 KB, where >80% matched a scaffold that was >5 KB). Statistics of final assemblies are shown in table S25. To estimate the completeness of the assembly, sets of 28,460 (RCC299) and 29,928 (CCMP1545) ESTs (see below) were BLAT-aligned to the unassembled trimmed data sets of the corresponding genomes, as well as the assemblies (table S25).

To perform finishing on RCC299, initial read layouts from the JAZZ WGS assembly were converted into the JGI Phred/Phrap/Consed pipeline (*8*). Following manual inspection of the assembled sequences, finishing was performed by resequencing plasmid subclones and by walking on plasmid subclones or fosmids using custom primers. All finishing reactions were performed with 4:1 BigDye to dGTP BigDye terminator chemistry (Applied Biosystems). Repeats in the sequence were resolved by transposon-hopping 8 kb plasmid clones. Fosmid clones were shotgun sequenced and finished to fill large gaps, resolve large repeats or to resolve chromosome duplications and extend into chromosome telomere regions. The finished genome consists of 20,989,326 bp of finished sequence with an estimated error rate of less than 1 error in 100,000 bp. All 17 chromosomes are contiguous, telomere to telomere, and without gaps. There are 4 regions of large tandem duplications still unresolved.

As noted above, the CCMP1545 genome was also sequenced using a WGS approach, and the resulting data assembled using (*9*). The genome was sequenced to 8.5x coverage with a total of 336,513 sequence reads (84.8% from paired plasmids and 15.2% from paired fosmids) and assembled into 21 scaffolds formed from a total of 213 contigs (and unresolved tandem repeat arrays). This draft genome was improved by manual inspection of all scaffolds to correct errors and misassemblies, and by manually extending all scaffolds to capture telomere signatures using previously unassembled fosmid paired reads. The improved draft assembly consists of 21,958,260 base pairs. There are 21 scaffolds, 19 representing complete chromosomes (telomere to telomere). The 2 additional small scaffolds presumably belong inside 2 of the larger gaps within assembled chromosomes.

Chloroplast and mitochondrial genomes were also sequenced and assembled. For RCC299, the finished mitochondrial genome was circular and complete at 47,425 bp with no detectable errors, and an overall GC content of 50%. The chloroplast genome was 72,585 bp and although of finished standard, was linear and hence considered incomplete. For CCMP1545, organellar genomes were assembled into an additional 6 scaffolds. The CCMP1545 draft mitochondrial genome was 41,691 bp with 5 captured gaps and GC content of 34.5%, while the draft chloroplast genome consisted of 5 unordered scaffolds.

RNA and EST libraries

For RCC299, mid exponential growth cells in standard conditions were used. For CCMP1545, the bulk of material was also from mid exponential, standard growth conditions; some material was also pooled with material from high light exposed cells. Cells were harvested by centrifugal pelleting using two sequential spins generally at 6000 x *g*. Cell pellets were then frozen at -80°C until extraction. Total RNA was isolated using the Qiagen RNeasy kit in conjunction with Qiashredder columns. Samples were quantified using a Nanodrop spectrophotometer and quality assessed on RNA chips (Agilent Bioanalyzer). Poly A+ RNA was isolated from total RNA using the Absolutely mRNA Purification kit (Stratagene, La Jolla, CA) and manufacturer's instructions. cDNA synthesis and cloning used a modified procedure based on the "SuperScript plasmid system with Gateway technology for cDNA synthesis and cloning"

(Invitrogen). 1-2 μg of poly A+ RNA, reverse transcriptase (SuperScript II, Invitrogen) and oligo dT-NotI primer (5'- GACTAGTTCTA GATCGCGAGCGGCCGCCCTTTTTTTTTTTTTTT -3') were used to synthesize first strand cDNA. Second strand synthesis was performed with *E. coli* DNA ligase, polymerase I, and RNaseH followed by end repair using T4 DNA polymerase. The SalI adaptor (5'- TCGACC CACGCGTCCG and 5'- CGGACGCGTGGG) was ligated to the cDNA, digested with NotI (NEB), and subsequently size selected by gel electrophoresis (1.1% agarose). Size ranges of cDNA were cut out of the gel (0.6-2 kb and >2 kb) and directionally ligated into the SalI and NotI digested vector pCMVsport6 (Invitrogen). The ligation was transformed into ElectroMAX T1 DH10B cells (Invitrogen).

Library quality was first assessed by randomly selecting 24 clones and PCR amplifying the cDNA inserts with the primers M13-F (GTAAAACGACGGCCAGT) and M13-R (AGGAAACAGCTATGACCAT). The number of clones without inserts was determined and 384 clones for each library were picked, inoculated into 384 well plates and grown for 18 h at 37°C. Each clone was amplified using RCA; the 5′ and 3′ ends of each insert were then sequenced using vector specific primers (FW: 5′- ATTTAGGTGACACTA TAGAA and RV 5′ – TAATACGACTCACTATAGGG) and Big Dye chemistry (Applied Biosystems).

The JGI EST Pipeline began with the cleanup of DNA sequences derived from the 5′ and 3′ end reads from the cDNA clone libraries. Phred software (*10, 11*) was used to call bases and generate quality scores. Vector, linker, adapter, poly-A/T, and other artifact sequences were removed using the Cross_match software (*10, 11*) and an internally developed short pattern finder. Low quality regions of the read were identified using 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 the EST. ESTs shorter than 150 bp were removed from the data set, as were those containing common contaminants such as *E. coli*, common vectors and sequencing standards.

EST clustering was performed *ab-initio*, based on alignments between each pair of trimmed, high quality ESTs. Pair-wise EST alignments were generated using the Malign software (Chapman, et. al., unpublished), a modified version of the Smith-Waterman algorithm (*12*). ESTs sharing an alignment of at least 98% identity and 150 bp overlap were assigned to the same cluster. These were considered relatively strict clustering cutoffs, and were intended to avoid placing divergent members of gene families in the same cluster. However, note that in these genomes this clustering effort largely failed and resulted in significant over clustering of ESTs derived from adjacent expressed genes (specifically, convergent overlapping pairs, COPs). ESTs that did not share alignments were also assigned to the same cluster if they were derived from the same cDNA clone.

EST cluster consensus sequences were generated using Phrap (*10, 11*) on the ESTs comprising each cluster. All alignments generated by Malign were restricted such that they would always extend to within a few bases of the ends of both ESTs. Therefore, each cluster was more like a ‘tiling path’ across the gene, which matches well with the genome-based assumptions underlying the Phrap algorithm. Additional improvements were made to the Phrap assemblies by using the ‘forcelevel 4’ option, which decreased the chances of generating multiple consensi for a single cluster where the consensi differed only by sequencing errors. In manual annotation efforts, only paired reads were considered, as clustering produced misleading results for the reasons noted above. Overall, 28,450 and 29,928 ESTs were sequenced for RCC299 and CCMP1545, respectively.

Gene modeling, automated assignments and manual curation

The genomes of *Micromonas* RCC299 and CCMP1545 were annotated using several gene predictors. First, the assembly was masked for repeats. We then employed *ab initio* Fgenesh (*13*) trained on manually curated *Micromonas* genes, homology-based Fgenesh+ (*13*) and GeneWise (*14*) and the EuGene hybrid approach that combines different types of evidence. EuGene3.4 was trained specifically for *Micromonas* on a set of manually curated gene-models. The *ab initio* part was first optimized and subsequently trained to weight the contribution of EST, protein homology and alignments with other genomes. The predictions used the independent EST libraries from both RCC299 and CCMP1545. The TAIR7, SWISSPROT and *Ostreococcus* proteome were used as sources of protein homology, and the *Ostreococcus* genomes and Sargasso Sea environmental sequences (*15*) were also used as sources of genomic DNA. Fgenesh was trained on over 5,000 *Micromonas* RCC299 genes including reliable homology-based models and putatively full-length (FL) genes assembled from ESTs. However, due to EST overclustering caused by transcripts from genes that formed COPs, cluster derived models were generally demoted. Fgenesh showed 75% sensitivity and 80% specificity of predictions on a test set. The same parameters were used for both genomes. Homology based gene predictors were seeded with BLASTX alignments of proteins from NCBI's non-redundant protein set. Putative full-length genes were derived from clustered ESTs available for each of the genomes and directly mapped to genomic sequences. *Micromonas* ESTs and consensus sequences derived from EST clusters were used to predict additional gene models and extend CDS predicted by above mentioned methods into full-length genes, which introduced errors in regions of overlapping genes except in the case of EuGene models. These were often corrected manually. A single representative model per each locus was chosen based on a similarity measure that combines a modified score of alignment with proteins from other organisms and correlation with available ESTs. This model set, the "filtered models" (FM), was further manually curated by the user community to provide the 'Gene Catalog,' an improved annotation over the FM. In the case of CCMP1545, Introner Elements (IE) seemed to often disrupt the success of modeling algorithms, none of which were trained with 'knowledge' of IE. This resulted in single genes being modeled as two separate genes (with IE in the intergenic space), or sometimes IE in CDS themselves, or in overly long introns (see also text S14).

Predicted gene models were functionally annotated by sequence similarity to annotated genes from the NCBI non-redundant set and specialized databases using BLAST and hardware accelerated double-affine Smith-Waterman alignments (timelogic.com). For example, predicted genes were annotated using Gene Ontology (*16*), eukaryotic orthologous groups (KOGs, (*17*)), and KEGG metabolic pathways (*18*) as summarized in table S1. Functional and structural domains were predicted in protein sequences using InterPro program (*19*).

Whole-Genome Alignments for chromosome-scale synteny between both *Micromonas* species were analyzed with i-ADHoRe, which identifies runs of collinear predicted proteins between genomic regions (*20*). We used a gap size of 10 genes, a Q-value of 0.9, and a minimum of 5 homologs to define a collinear block.

Manual annotation employed various prediction and phylogenetic approaches and a range of expertise. The user community generally selected EuGene models as the 'best performing' (a qualitative assessment). In terms of model modifications, only paired EST reads were considered since automatic clustering had resulted in overclustering and incorrect predictions of UTRs as well as frequent "missing" of genes on the opposite strand that had overlap. ChloroP (http://www.cbs.dtu.dk/services/ChloroP/) and TargetP

(http://www.cbs.dtu.dk/services/TargetP/) were used to identify transit peptides for plastid targeting. Leader sequences for mitochondria targeting were identified by several programs: Predator (http://urgi.infobiogen.fr/predotar/predotar.html), TargetP, Mitoprot (http://ihg.gsf.de/ihg/mitoprot.html) and PSORT (http://wolfpsort.seq.cbrc.jp/). For mitochondrial sequences, results were regarded as significant only if at least two programs predicted a mitochondrial targeting sequence and if the selected leader sequence showed similarity to a leader sequence from a protein that is, by function and precursor sequence, most likely targeted to either the plastids or mitochondria. Lumenal-targeting was analyzed by searching for TAT- and SEC- specific motifs (e.g. RR for TAT), transit peptides (Hydrophobicity Plots (http://www.bmm.icnet.uk/~offman01/hydro.html)), and cleavage sites for the lumenal-processing peptidase (e.g. AXA for TAT). In many cases both BLASTP and/or TBLASTN were used to verify presence/absence in each *Micromonas* species as well as *Ostreococcus* and other organisms. Alignments were performed using ClustalW (e.g. EBI) as well as KALIGN (*21*) with manual improvement as necessary. A series of different phylogenetic analyses were used, often RAxML (*22*) or PhyML (*23*).

Chromatin-associated and RNAi-associated protein encoding sequences displayed at The Chromatin Database (www.chromdb.org) were used as queries to search catalog protein sequences. A succession of protein queries consisted of *O. lucimarinus*, *C. reinhardtii*, and *A. thaliana* gene sets. BLASTP e-values were set initially at e^{-40} and dropped down to e^{-10} for successive BLASTP searches to identify divergent proteins. Preliminary evolutionary assignments were made using a specialized ChromDB local BLAST program. Select cases (SDE-3 RNA helicases and Argonaute-like proteins) were examined using edited multiple sequence alignments produced by MUSCLE (*24*) to test phylogenetic relatedness using the neighbor-joining program of Mega 3.1 (*25*) and 1000 replicates for bootstrap analyses (pairwise deletion).

Identification of *Micromonas* selenoproteins was performed by using known *Ostreococcus* sequences as well as homologs from *C. reinhardtii, M. musculus, Drosophila* and *H. sapiens* as queries against the *Micromonas* genomes, using the BLASTP algorithm to identify putative homologs. If no hits were found by BLASTP then the TBLASTN algorithm was used, and if again no hits were found, it was concluded the gene sequence was not present. Selenoproteins were then confirmed in a two-step process: 1) the Sec codon was located and, where possible, confirmed with EST evidence, and the models were adjusted to include the Sec codon as a coding element rather than a stop codon; 2) the 3′ UTR, or if no UTR was identified then the downstream region from the particular gene's stop codon, was run through SECISearch (*26*), utilizing both default and loose conditions. SECIS elements identified were included in annotation. If no SECIS element was found this was also noted.

Membrane transport protein analysis used the complete "catalog" protein sequence datasets from both *Micromonas* genomes (as of 14 February 2008) and analysis using the TransAAP pipeline (*27*) to determine their predicted complement of membrane transport proteins. This approach combined BLAST searches against a curated membrane transport protein database (Transport DB), as well as HMM seaches and COG-based searches against membrane transporter protein families. Membrane transporters were assigned to protein families based on sequence similarities, and the numbers of different types of transporters were compared between *Micromonas* and other marine eukaryotes analyzed in TransportDB (*27*). The final analysis herein involved a selection of genomes, both published and unpublished to represent the different lineages indicated: *Micromonas* RCC299 and CCMP1545, *O. lucimarinus* (CCE9901), *O. tauri*,

P. tricornutum (chromalveolate), *T. pseudonana* (chromalveolate), *P. patens* ("lower plant"), *A. thaliana* (plant), *O. sativa* (plant), *T. thermophila* SB210 (bacteria), and *C. reinhardtii* (green alga). When statements of presence/absence in *Micromonas* or *Ostreococcus* are made, they were manually verified with TBLASTN using BLAST on the JGI browser website. Manual verification of the analysis was not always made for the other non-Mamiellales genomes.

To test whether putative C-Type Lectin Domain (CTLD) containing proteins identified using JGI motif-based searches were actually CTLD homologs, we used Hidden Markov Model-based profiles constructed from Metazoan CTLD sequence alignments and validated previously (*28*) to analyze the protein sequences in question. Additionally, all predicted protein sequences (unfiltered sets) and unmasked genomic DNA were scanned using the same HMM to detect CTLDs. HMMER (*29*) and GeneWise (*30*) software packages were used for analysis of protein and DNA sequences, respectively.

TPP riboswitch searches of both the RCC299 and CCMP1545 genomes were made using the Infernal software, together with the RFAM TPP entry RF00059 (*31*). Mfold and manual investigation were used to verify secondary structure. Putative TPP riboswitches in the SAR11 bacterial genomes were identified by eye, and tested in mfold.

Transposon and Introner identification

Kmer analysis was performed using an enhanced suffix tree index created for each genome using vmatch (http://www.vmatch.de/). For each 16mer from a non-overlapping sliding window along the chromosomes, the frequency (sum of forward and reverse hits) was determined by a query against the index.

Repeat elements were derived from the kmer data by merging 16mers with a frequency ≥10, allowing any number of 16mers with frequencies ≥5 and a maximum of three successive 16mers with frequencies < 5 within one element. The high frequency kmer derived elements were clustered by vmatch (single linkage clustering, 80 % identity limit) yielding 9 families with member numbers between 160 and 850. In CCMP1545, most were located within introns and corresponded to IEs (table S3, text SX). The Apollo synteny viewer (http://apollo.berkeleybop.org/current/index.html) was used for the visualization of syntenic relationships (Fig. 3) with customized color codes. Relic repeat elements were identified by homology to mips-REdat, a plant repeat database (http://mips.gsf.de/proj/plant/webapp/recat/index.jsp). Repeat sequence distributions for both genomes are shown in table S23.

To identify and annotate TE (as shown in table S21), the genome sequences were 'shredded' into 100 kbp overlapping segments. Each segment was manually inspected by dot plot comparison of these segments against themselves using dotter (*32*). This allowed identification of putative structural features of TE (Long Terminal Repeats or Terminal Inverted Repeats). All regions with potential TE features were then searched against the whole genome using BLASTN in order to determine whether the sequence was repeated. In addition, repbase (v11.09; (*33*)) was used to create a database of TE-related peptides (1180 sequences). Using this resource, a TBLASTN search was conducted against the whole genome of each *Micromonas* species to detect TE-related conserved protein domains. All the predicted peptides and ORFs of *Micromonas* were then used to perform homology searches against the repbase nucleotide bank (~6000 sequences). For each putative TE-related sequence found in *Micromonas*, a reverse BLAST search against the nrpep database was used to verify whether the match corresponded to a TE.

Gene pools and phylogenomic analyses

To generate 'core', 'shared' and 'unique' gene data sets, catalog gene sequences from each of the four genomes (RCC299, CCMP1545, *O. tauri* and *O. lucimarinus*) were run by BLASTP against a combined set of all predicted models (in case some models had been excluded from the catalog, e.g. due to gene overlap; using all models helped prevent this from being an issue) from each of these 4 genomes; hits were considered homologs at an e-value threshold of e^{-05}, or lower. Data presented here reflects use of the RCC299 gene catalog (as of 14 February 2008) as the query against the other 3 Mamiellales genomes. Note that 'unique' genes do not represent genes unique to biology, but rather those genes found only in one *Micromonas* species and not the other three Mamiellales genomes. These could also be referred to as "niche defining" genes, but given that the inferences are made on a total of 4 genomes, and that it may be the interaction of these genes with other core components, rather than the genes themselves, that determine their function and role in niche partitioning, we elected not to use this term. It should be noted that modeling algorithms did not handle IE well, leading to disruption of gene models that may have influenced this analysis; again, the use of all models as opposed to just catalog models was designed to help ameliorate this problem.

For phylogenomic analyses summarized in Fig. 2, trees were created using APIS (Automated Phylogenetic Inference System; Badger, unpublished), an automated system for creation and summarizing of phylogenetic trees for each protein encoded by a genome. The homologs used by APIS for each phylogenetic tree were obtained using WU-BLAST (Gish, 2004) to compare query proteins against an extended version of ComboDB (Wu, unpublished) that contained taxonomic, genomic, protein, and coding DNA information for 46 eukaryotic, 52 archaeal, 687 bacterial, and 1928 viral complete (or nearly complete) genomes (as of June 1st, 2008). The full-length sequences of these homologs were retrieved from the database and aligned using MUSCLE (*24*). Bootstrapped neighbor-joining trees were produced using QuickTree (*34*). The inferred tree was then midpoint rooted prior to analysis, allowing automatic determination of the taxonomic classification of the organisms with protein sequences in the same clade as the query protein sequence. Scripts were then written (Ruby programming language) to identify trees containing clades in which sequences from the Mamiellales clade (all 4 genomes in the case of the 'core'), or the two *Micromonas* genomes (in the case *Micromonas* 'shared' genes), or only a single *Micromonas* species, clustered with sequences from members of a particular target group (e.g. 'bacteria and archaea' or 'Streptophyta'), without including sequences from any other taxonomic groups. The bootstrap value of the node connecting the *Micromonas* sequence(s) to the target was noted in order to identify particularly robust groupings. If the *Micromonas* sequences lay on the opposite sides of the tree root from the target sequences, the bootstrap value of the target clade itself was used instead. The lengths of the branch leading to the *Micromonas* sequence (or to the clade containing the *Micromonas* sequences) were analyzed using the statistical environment "R" (R Development Core Team, 2008). The "density" function of "R" was used to create kernel density estimations (plots similar to histograms, but continuous (*35*)) to see if the distribution of branch lengths in any two categories of interest was the same or different. For the distributions shown in Figure 2c, 5188 (2811) genes within the 'core' set returned trees.

Shown in parentheses is the number of these that could be positively linked to a specific lineage with bootstrap ≥ 50; for 'shared' these numbers were 419 (212), for RCC299 'unique'they were 100 (57), and for CCMP1545 they were 'unique' 52 (27).

SUPPORTING TEXT

1) Genome structure, gene rearrangements and expression analysis. RCC299 is one of very few fully finished eukaryotic genomes, with 17 chromosomes assembled telomere to telomere. CCMP1545 has 19 chromosomes (sequential numbering from largest to smallest, fig. S3) assembled into 21 scaffolds, 2 presumably belonging to gap regions within assembled chromosomes. We concluded, during sequencing, that we 'hit' the telomere, and hence the end of the chromosome, by reaching the identified telomere signature:

```
>telomere.fasta
AAACCCTAAACCCTAAACCCTAAACCCTAAACCCTAAACCCTAAACCCTA
AACCCTAAACCCTAAACCCTAAACCCTAAACCCTAAACCCTAAACCCTAA
ACCCTAAACCCTAAACCCTAAACCCTAAACCCTAAACCCTAAACCCTAAA
CCCTAAACCCTAAACCCTAAACCCTAAACCCTAAACCCTAAACCCTAAAC
CCTAAACCCTAAACCCTAAACCCTAAACCCTAAACCCTAAACCCTAAACC
CTAAACCCTAAACCCTAAACCCTAAACCCTAAACCCTAAACCCTAAACCC
TAAACCCTAAACCCTAAACCCTAAACCCTAAACCCTAAACCCTAAACCCT
AAACCCTAAACCCTAAACCCTAAACCCTAAACCCTAAACCCTAAACCCTA
AACCCTAAACCCTAAACCCTAAACCCTAAACCCTAAACCCTAAACCCTAA
ACCCTAAACCCTAAACCCTAAACC
```

The chromosomal structure seen in assemblies is well supported by PFGE analysis of both species (fig. S2). RCC299 has 64% GC and a 21 Mb genome, while CCMP1545 has 65% GC and a 22 Mb genome. We were able to link across centromere repeat sequences during assembly; given that the largest insert libraries were 40 kb, centromeres appeared to be smaller than 40 kb.

For most chromosomes, 75-76% of RCC299 and CCMP1545 DNA sequences aligned with each other (fig. S4). Synteny was also apparent (fig. S5), although less evident than between the two sequenced *Ostreococcus* species (*36*). In RCC299, the low GC-region spanned 1.5 Mb (in chromosome 1) and was 50% GC, while in CCMP1545 it spanned 1.7 Mb (in chromosome 2) and was 51% GC (fig. S3). Similar regions in *O. tauri* and *O. lucimarinus* (*36, 37*) comprise only 4% and 3%, respectively, of the genome. For these low-GC-regions only 37-40% DNA aligned with DNA of the other *Micromonas* species. Approximately half the genes encoded in both the low GC- and normal GC-regions (for which trees could be constructed) appeared to have clear green-lineage affiliations, making inter-phylum horizontal gene transfer (HGT) an unlikely explanation for the observed differences between these two types of genomic regions.

In addition to the low GC-regions of chromosomes 1 (RCC299) and 2 (CCMP1545), the smallest chromosome of each *Micromonas* genome also had below-average % GC content: chromosome 17 of RCC299 is 55% GC while scaffold 19 of CCMP1545 is 54% GC. Most predicted genes on these smallest chromosomes were poorly characterized (no obvious database homologs) and specific to the respective species; consequently, synteny was lacking. For these smallest chromosomes only 3-5% of the entire sequence could be aligned to the other genome (fig. S4). In contrast, for the *Ostreococcus* genomes, 92% of all chromosomes, 78% of Chr2 of each strain (the low GC-region(s)), and 15% of Chr19/18 aligned to each other. Comparing the *Micromonas* genomes to *Ostreococcus* showed 28-35% of chromosomal sequence, on average, aligned with *Ostreococcus* chromosomes. Thus, the *Ostreococcus* genomes, while distant from *Micromonas*, bore much higher identity to each other than RCC299 and CCMP1545 did to each other, even though some level of synteny was preserved (fig. S5).

Codon usage was more distinct between RCC299 and CCMP1545 than between the two *Ostreococcus* species. Principle components analysis of codon usage also showed greater divergence between the *Micromonas* genomes than seen between the *Ostreococcus* genomes (fig. S6). Furthermore, the smallest chromosomes (chromosome 17 of RCC299 and scaffold 19 of CCMP1545) and low GC-region(s) of RCC299 chromosome 1 and CCMP1545 scaffold 2 showed significant bias in codon usage compared to the corresponding normal GC-sets (fig. S6). Finally, codon usage deviated significantly between the normal GC-chromosome sets and the low GC-region(s) of each *Micromonas* species (fig. S6).

In randomly selected normal GC-fragments from RCC299, 40 ± 3% of the genes bore transcriptional support (non-normalized EST library) under standard growth conditions; by contrast, low GC-fragments showed twice as much transcriptional activity, with 82 ± 4% of the genes having EST support. To determine these percentages, the RCC299 genome was divided into 131 equal-sized contiguous fragments (160,188 bp). A pseudo-random sequence of numbers, created in Matlab, was used to select 8 genome fragments for manual analysis along with 2 additional non-randomly selected fragments, from the low GC-region of chromosome 1. A cursory (non-quantitative) analysis of CCMP1545 indicated that its low GC-region also had a higher percentage of models with EST support than did other regions. Notably, data from RCC299 low GC-fragments (as generated above) indicated that 66% of genes in this region formed COPs.

2.) Mitochondrial and Chloroplast genomes. The GC content (~35%) of the mitochondrial (mt) genomes of both *Micromonas*, as well as their reduced mt genome size (RCC299: 47,425 bp; CCMP1545, not fully assembled: 41,691 bp), were comparable to other sequenced Prasinophytae (*38*), including *O. tauri* (44,237 bp, 38% GC) and *Nephroselmis olivacae* (45,223 bp, 33% GC), but were much smaller than the mt genomes of higher plants, e.g. *A. thaliana* (366,924 bp) and *Marchantia polymorpha* (186,609 bp). RCC299 clearly contained a duplicated region within its mt genome; in CCMP1545, some genes seemed to be duplicated, but because the mt genome is incomplete, final determination of duplicated region(s) was not possible. The presence of such duplicated regions was not known in green algae until recently described in the two sequenced *Ostreococcus* species. This duplication is not an invariant feature of prasinophytes since the *N. olivacae* mt genome lacks any kind of duplication.

The gene content of both *Micromonas* mt genomes was almost identical (table S2, fig. S7a, b): the complete RCC299 mt genome contained 63 genes (unique ORFs were not taken into account, and duplicated genes were counted only once), of which 34 are protein coding genes, 3 rRNAs and 26 tRNAs; the incomplete CCMP1545 mt genome had 60 genes (33 protein coding genes, 3 rRNAs and 24 tRNAs). The *rps11* gene and two tRNAs were not found in CCMP1545, which could be due to a CCMP1545-specific loss or to incomplete sequencing. A comparison of gene content in the mt genomes of the two *Micromonas* species with other members of the Prasinophytae established an almost identical gene repertoire. One gene, *atp1*, previously identified in other Prasinophytae, seems to have been lost from both *Micromonas* species. Furthermore, *rrn5* and *rnpB* were not detected. Notably, nuclear-encoded mitochondrial-targeted *RECA* genes, thought responsible for allowing plants to maintain large mt genomes, were not found (*39*) (see text S3).

The chloroplast (cp) genomes also had highly similar gene content between RCC299 and CCMP1545 (table S3), as well as to *Ostreococcus*. The complete cp genome of RCC299 was 72,585 bp, smaller than a number of other green lineage cp genomes. A comparison of amino

acid positions present (from all encoded genes) demonstrated the degree of reduction in these cp genomes compared with many other photosynthetic taxa (table S4). Note that the CCMP1545 cp genome was not fully assembled. We also performed a phylogenetic reconstruction using 6 cp genome-encoded proteins in order to establish the position of the Mamiellales with respect to other green and red lineage organisms (fig. S1). In order to root the tree we used gene sequences of the early-diverging glaucophyte *Cyanophora paradoxa* (*40*). The chromalveolate plastids (e.g. as represented in fig. S1) are derived from red algae through a secondary (eukaryotic) endosymbiosis (Yoon et al. 2002). While most of the methods used in this analysis were as described in the materials and methods section, bayesian phylogenetic inference was also used (*41*).

3.) Gene pools (core, shared, unique) and selenoproteins. We analyzed Mamiellales nucleus-encoded 'gene pools' (Fig. 2a) to investigate shared and differentiated features of the protein encoding gene complement (see methods). The majority of the Mamiellales 'core' genes (7137 genes, or ~71% of predicted genes in RCC299) fell within known eukaryotic orthologous groups (KOG; Fig. 2b). The most heavily represented KOGs entailed protein synthesis/turnover and signal transduction. Core pathways were similar to other green-lineage organisms, including the oxidative pentose pathway, Calvin and TCA cycle components (table S5, fig. S8). In addition, other core components included some aspects of light harvesting (although expanded relative to *Ostreococcus* (*42*)), much of photosynthesis (table S8) and pigment biosynthesis (table S9), as well as some aspects of starch metabolism (*43*). Some chromatin and RNAi associated genes fell within the core (text S4, table S6). Cell-cycle genes also fell within this group (table S7) as well as some carbohydrate active enzymes (CAZy, text S6). Differences generally entailed gene copy numbers.

The *Micromonas* core pool included selenoproteins (text S7, table S13a, b). Of 29 selenoproteins previously identified in *Ostreococcus* (*44*), RCC299 contained 17, plus 3 non-selenoprotein homologs; CCMP1545 contained 15, plus 4 non-selenoprotein homologs. Both *Micromonas* species also contained the transcriptional machinery necessary for inclusion of selenocysteine in proteins. The selenoproteome of the Global Ocean Sampling (GOS) expedition dataset was recently analyzed, revealing 3,600 selenoprotein sequences (*45*). While the focus was on bacterial and archaeal sequences, eukaryotic selenoproteins were also detected, including a number found in both *Micromonas* species, such as protein disulfide isormerase, *SELM*, *SELT* and thioredoxin disulfide reductase. One conclusion of this study was that water salinity and temperature appear to influence the utilization of Sec. However, RCC299 and CCMP1545 had similar selenoprotein profiles, suggesting there may also be other influential driving forces. The variations in selenoprotein content (fewer in *Micromonas*) among the Mamiellales may represent a speciation force, as proposed for *Ostreococcus* (*37*), since the distribution of selenoproteins and non-seleno homologs across the included taxa was diverse even within a genus. This proposal is still highly speculative and requires experimental verification.

'Unique' (793, RCC299; 826, CCMP1545) and 'shared' (1384) pools revealed a variety of categories in which *Micromonas* was enriched over *Ostreococcus*, or in some cases one *Micromonas* enriched over the other (fig. S14). Both *Micromonas* species were enriched over *Ostreococcus* in several areas (fig. S14), including secondary metabolite synthesis and transport as well as amino acid transport and metabolism, which may relate to environmental factors. Both *Micromonas* species also had genes encoding a flagellum (text S8, table S21), as well as more chromatin-associated genes than *Ostreococcus* (table S6). 8% of genes identified in RCC299

were not found in CCMP1545, *O. tauri* or *O. lucimarinus* (herein termed 'unique', Fig. 2). CCMP1545 had the same percentage of 'unique' genes (see fig. S14).

4.) Chromatin and RNAi. Both *Micromonas* species had the basic complement of proteins necessary for assembling histones into nucleosomes, e.g. histone chaperones such as NAP1 homologs (*46*), chromatin assembly factors CAC1, CAC2 and CAC3 (*47*), the HIR class of chaperones (*48*), FACT complex subunits SSRP1/POB3 (*49*), ASF1 (*50*), and most of the nucleosome remodeling proteins that disassemble and move nucleosomes during transcription and DNA replication (table S6). The latter group includes the SWI/SNF superfamily of ATP-dependent nucleosome remodelers that influence chromatin structure through the disruption of histone-DNA interactions to slide and reposition nucleosomes (*51*).

In yeast, the INO80 protein is the ATPase component of the multi-subunit complex bearing its name (*52*), a complex that functions in double-stranded DNA break repair (*53*). While highly conserved among plants, animals and fungi, INO80 homologs, as well as the actin-related protein subunits ARP5 and ARP8, are not found in the genomes of sequenced chlorophytes and prasinophytes (*37*), and this trend persisted for both *Micromonas* species. However, the closely related ATPase SWR1, which also functions in DNA repair, and its two actin-related protein subunits, ARP4 and ARP6, were present in CCMP1545 and RCC299 (as well as *Ostreococcus* and chlorophytes, although only ARP4 was reported in *O. tauri*). In addition, two other proteins, homologs of RVB (*54*), which are subunits of both the INO80 and SWR1 complexes, were present. However, a significant loss is SNF5, a component of the SWI/SNF complex (*55*). Both *Micromonas* and *Ostreococcus* lacked putative homologs of the DNMT1 class of CpG maintenance methyltransferases, as well as DNMT3 homologs responsible for *de novo* CpG methylation.

H1 (Histone linker protein 1) and CARM1 were present in *Micromonas* but not *Ostreococcus* (table S5). The *Micromonas* genomes also encoded a protein arginine methyltransferase with a number of specific substrates including histone H3 R17 and R26 (*56*), but neither *Micromonas* nor *Ostreococcus* species appeared to have PRMT1-like proteins. PRMT1 has a number of overlapping functions with CARM1 and PRMT5 (*57*); thus it is significant that predicted CARM1 proteins were found in both species of *Micromonas* but not *Ostreococcus*. The analysis of DEAD Box and SDE3 genes provided circumstantial evidence that RCC299, but not CCMP1545, also has diverged RNA helicases akin to *Arabidopsis* SDE3 (*58*). Finally the highly conserved Argonaute-encoding gene in RCC299 (Prot. ID 113410) contained PIWI, PAZ, and DUF1785 (pfam08699, often found in Argonaute) domains, was not found in CCMP1545 or either *Ostreococcus*. Additional features of chromatin proteins shared by both *Micromonas* and *Ostreococcus* are detailed in table S6.

5.) Sex. Three meiotic recombination gene families, in particular, the *TOP6A*/*SPO11*, *RECA-RAD51-DMC1* and *MUTS* homolog families, support the capacity for a sexual cycle in prasinophytes (table S13). SPO11 causes the double-strand DNA break at the initiation of meiotic recombination, and both the *Micromonas* and *Ostreococcus* genomes contained homologs of the plant meiosis-specific SPO11-2. In addition, they each contain a *TOP6A* gene that shared homology with the non-meiotic *TOP6A/SPO11-3* known to be required for normal growth and development in plants (*59-62*). SPO11 is evolutionarily related to the TOP6A component of the Type II DNA Topoisomerase 6, but has lost the topoisomerase function due to the absence of a corresponding TOP6B subunit in eukaryotes (*63*). Two to three SPO11/TOP6A

paralogs are found in plants and some other eukaryotes. One, TOP6A (SPO11-3), appears to be a topoisomerase because it interacts with a TOP6B protein in *Arabidopsis* (*60, 61*). The other two SPO11/TOP6A homologs, SPO11-1 and SPO11-2, are only required during meiosis in plants (*64, 65*). Animals and yeast have only the meiotic-specific homolog SPO11-1. The ancestral eukaryote may have contained 3 SPO11/TOP6A homologs (the meiosis-specific SPO11-1 and SPO11-2, and the non-meiotic TOP6A and TOP6B proteins) (*66*). Similar to diatoms (many of which are known to undergo a sexual cycle) and to red algae, *Micromonas* and *Ostreococcus* both appeared to have lost SPO11-1 but retained SPO11-2 as well as the non-meiotic TOP6A and TOP6B. The functional differentiation between SPO11-2 and SPO11-1 in plants is still not known; however, *Arabidopsis* mutants with disrupted SPO11-2 display strong meiotic defects but normal vegetative development and no increased sensitivity to DNA-damaging treatments (*65*). Therefore, the presence of SPO11-2 orthologues in the Mamiellales was consistent with the retention of meiosis in these organisms.

A second family of recombinases lends support to the existance of sexuality in these organisms. All 7 members of the RAD51 family (table S13) known in plants and animals (*67-70*) were identified in RCC299 and CCMP1545. The RECA/RAD51 family of proteins is involved in homology searching and strand exchange during homologous recombination. The RAD51 family is a set of ancient eukaryotic paralogs of the bacterial RECA DNA recombination proteins, and includes genes involved in DNA repair during vegetative growth as well as genes predominantly involved in meiotic recombination (*39, 67*). RAD51 and DMC1 act as classical recombinases and are thus functional homologs of RECA, while the other 5 members of the family (the so-called "RAD51 paralogs": RAD51B, RAD51C, RAD51D, XRCC2 and XRCC3) are required for efficient recombination in various contexts *in vivo*, although their exact role is not known. For example, in *Arabidopsis*, RAD51B, RAD51D and XRCC2 mutants are fertile but have reduced capacity for repair of certain types of DNA damage, while RAD51C and XRCC3, in addition to their role in DNA repair, are also required for meiotic recombination (*68-71*). Interestingly, *O. tauri* did not appear to contain an ortholog of RAD51B, suggesting that there is enough redundancy in this family of proteins to allow some to be lost, at least for vegetative growth. In plants, animals and fungi, the DMC1 recombinases are strictly meiosis-specific proteins with no known essential roles outside gametogenesis. The *Micromonas* and *Ostreococcus* predicted DMC1 and XRCC3 had strong homology to the *Arabidopsis* DMC1 and XRCC3, respectively. However, the putative XRCC2 was poorly conserved and a clear ortholog was not identified in either *Ostreococcus* genome. The observation that *Micromonas* seems to have retained all of the RAD51 paralogs found in plants and animals, including those not needed in meiosis (RAD51B) and those that are only known to be involved in meiosis (DMC1), supported the hypothesis that the Mamiellales have retained the capacity for meiotic recombination.

Finally, members of the MSH family, eukaryotic proteins related to the bacterial MUTS proteins involved in mismatch recognition and repair, were identified. There are several MSH members in plants, yeast, and animals. Likewise, the *Micromonas* genomes each contained several members. MSH4 and MSH5 are partners in promoting crossover formation during meiotic recombination (*72*), following an interference-sensitive pathway that also involves MER3. In *Arabidopsis*, *MSH4* is specifically expressed in meiosis, is not required for normal vegetative growth and development, but is required for fertility (*73*). The presence of clear orthologs of *MSH4* and *MSH5* in *Micromonas* suggests the capacity for interference-sensitive crossover formation and resolution during meiotic recombination. Curiously, *MSH4* orthologs were not clearly identified in the *Ostreococcus* genomes although other *MSH*s were found,

including an apparent *MSH5* ortholog. This could indicate important differences between *Micromonas* and *Ostreococcus* lifestyles, requiring further investigation.

A suite of other genes involved in meiotic recombination was also identified (table S13). RAD50 and MRE11 form a complex that functions to repair double-strand DNA breaks during meiotic recombination and also in DNA repair and telomere maintenance (*74*). Both of these genes were found in the Mamiellales genomes. In yeast and animals, RAD50 and MRE11 also interact with a third partner, NBS1/XRS2 which is less conserved and has not yet been unambiguously identified in plants (*69*). We were not able to identify a homolog of NBS1/XRS2 in the Mamiellales.

MND1 and HOP2 (also called MEU13 and TBIP) function during strand invasion in meiotic recombination to ensure homologous pairing and facilitate loading of the RAD51-DMC1 complex (*75, 76*). Predicted genes encoding clear homologs of these proteins exist in the prasinophyte genomes. Under normal conditions, plants in which the *MND1* gene is disrupted do not display any apparent defects in vegetative growth and development but display meiotic defects (*77, 78*). One report notes that *MND1*-disrupted plants are defective in DNA repair in response to gamma irradiation and that gamma irradiation induces MND1 expression in non-meiotic tissues (*77*), but a later study reports that mnd1 mutants were not different from wild-type plants in response to hydroxyurea-treatment or gamma irradiation (*78*). Thus the conservation of MNDd1 and HOP2 in prasinophytes is consistent with the capacity for meiotic recombination in these cells, but these proteins may instead be retained for a non-meiotic function (e.g. DNA repair).

In yeast, animals and plants, resolution of crossovers can occur through interference-sensitive and interference-insensitive pathways. The interference-sensitive pathway is characterized by MER3, MLH1, HOP2, MND1, MSH4 and MSH5, whereas the interference-insensitive pathway is characterized by the endonuclease MUS81 and its partner MMS4 (EME1) (*73*). The *Micromonas* and *Ostreococcus* genomes both contain clear homologs of proteins in the interference-sensitive pathway (e.g. MER3). They also contain a clear homolog of the MUS81 endonuclease, but its partner MMS4 is not well-conserved: a potential highly divergent homolog was identified in RCC299 only (table S13 - located at Ch11: 492755-494269), but the homology was weak (only 10.4% identity with the corresponding *Arabidopsis* predicted protein, although the e-value, $1.2e^{-10}$, was significant). The Parting Dancers (PTD) gene, which plays an unidentified role in crossover resolution in plants, was not found. Nevertheless, these data suggest *Micromonas* may have the capacity for both types of crossover resolution, although the fact that the MMS4 homolog appears to be evolving rapidly may indicate that it is diverging in function.

Homologs of *BRCA2* were identified in both the *Micromonas* and *Ostreococcus* genomes. In *Arabidopsis*, BRCA2 protein plays an essential role in meiosis, interacting with DMC1 and RAD51 (*79*). Furthermore, two paralogs of *RAD54* were identified in each of the prasinophyte genomes. RAD54 is known to play an important role in recombinational DNA repair and meiotic recombination, interacting with RAD51 proteins (*80*). Hence its presence further supports the likelihood of these processes in *Micromonas*.

Two families thought to be essential to meiotic recombination were not found in any of the Mamielalles genomes. 1) RAD52 facilitates RAD51 binding to single-strand DNA in yeast and animals but was not found in the prasinophytes. However, this protein also has not yet been identified in plants (*69*) and so it may have been lost from some eukaryotic lineages. 2) The

SWI1/AM1 protein regulates meiotic commitment in plants, but homologs were not identified in any of the Mamiellales genomes.

HOP1/ASY1 is involved in chromosome synapsis in yeast, animals and plants (*81*). Weak but significant homologs of the *HOP1/ASY1* (e^{-20} – e^{-21}) were identified in the *Ostreococcus* genomes, but these predicted genes shared homology to only part of the corresponding *Arabidopsis* ASY1 protein, and no homologs were found in *Micromonas*. STAG3 is a SSC3 homolog specifically involved in meiotic sister-chromatid arm cohesin in animals. The *Arabidopsis* homolog, AtSCC3, is involved in both meiotic and mitotic chromatid cohesin (*82*). A STAG3/SCC3 homolog was identified in all prasinophyte genomes. Other related proteins, such as Zip1, so far only identified in yeast, and Shugoshin (SGO1 and SGO2), found in yeasts, animals and plants (*83*) could not be identified.

Essential non-meiotic proteins related to core meiotic proteins were also identified. As noted above, both TOP6A (evolutionarily related to SPO11) and TOP6B topoisomerase components were found in both *Micromonas* species. In plants, absence of these leads to a dwarf phenotype and impairs endo-reduplication (*59-62*). Genes for TOP6A and TOP6B are found in many eukaryotes and they were perhaps present in the ancestral eukaryotic cell (*66*). Plants also have several RECA paralogs, some chloroplast-targeted and some mitochondrion-targeted. In contrast, animals lack RECA, and this is proposed to have led to the more extensive mt genome size reduction and variability in animals (*39*). The Mamiellales contained a homolog of the chloroplast-targeted *RECA*, but not the mitochondrial-targeted *RECA* present in plants. We also found plant-like *MSH1* in all four genomes, suggesting that these organisms may have the capacity for genomic and phenotypic regulation of the mt genome. MSH1 is the mitochondrial-targeted MUTS homolog and MUTS homologs are generally involved in mismatch repair (*84*). While MSH4 and MSH5 have evolved distinct functions in meiosis (see above), MSH1, so far identified only in yeast and plants (not in vertebrates, nematodes, or insects), appears to have evolved a distinct function in mt genomes (*85, 86*). The MSH1-like homologs identified herein were similar to the distinct plant MSH1 involved in "sub-stoichiometric shifting". During this process creation and suppression of sub-genomic DNA molecules through recombinatorial mechanisms allows changes in the relative copy number of portions of the mt genome (*86*).

In addition to evidence for meiosis, sex in the Mamiellales is also suggested by similarities to *Chlamydomonas* sex. In *Chlamydomonas*, sex determination is regulated by a gene called *MID* (*MInus Dominance*). The *minus* (*MT-*) but not the *plus* (*MT+*) mating-type locus carries a *MID* gene that is expressed early in gametogenesis (*87*); *plus* cells carrying a *MID* transgene differentiate as *minus* (*88*); and mutations/deletions of *MID* cause *minus* cells to differentiate as *plus* (*87, 88*). In several related volvocacean algae, *MID* is also confined to genomes of one mating type (*89, 90*). The MID protein is a member of a large family of presumed but poorly-characterized transcription factors found primarily in plants that share the motif RWP-RK (*102*).

We identified 3 RWP-RK motif subfamilies in *Micromonas* and *Ostreococcus* (table S14). One subfamily encoded Mid-like proteins (MLPs) similar to MID. The MID/MLP proteins were markedly shorter (mean length 180 aa for algal MLPs, 156 aa for MID) than those in the other two subfamilies (mean lengths 979 and 534 aa for algal members), and were in turn subdivided into 3 clades: MID, MLPa, and MLPb. Both *Ostreococcus* genomes carried an *MLPa* gene; *Micromonas* RCC299 carried an *MLPb* gene; and *Micromonas* CCMP1545 lacked any *MID/MLP* genes. This pattern would be expected if the first 3 genomes were derived from "*minus*-equivalent" strains and the fourth from a "*plus*-equivalent" strain; that is, the lack of an

MLP in CCMP1545 would be expected for an "opposite" mating type (although, if CCMP1545 and RCC299 are separate species as the data suggest, they would not be expected to mate).

Interestingly, the two *Ostreococcus* MLPa genes reside in chromosome 2, and the *Micromonas* RCC299 MLPb gene resides in chromosome 1. Chromosome 2 of *Ostreococcus* has been hypothesized to represent a "sex chromosome" based on its distinctive low GC content, high transposon endowment (*36*), and intra-strain gene rearrangements (Palenik et al., 2007); the latter 2 features are shared with the *MT* loci of *Chlamydomonas* (*91*) and the fungus *Cryptococcus* (*92, 93*). Each *Micromonas* species has a chromosome with similar characteristics (text S1, including the MLPb-bearing chromosome 1 of RCC299), and hence is also a sex-chromosome candidate

Sex represents a potent alternative to death because diploid spores are often resistant to environmental fluctuations and digestion by predators. Combined with the maximization of adaptivity (via recombinant offspring), the likelihood of sex in *Micromonas* is an important ecological consideration, and its elucidation could result in a valuable laboratory-based genetic system for the Mamiellales.

6.) Carbohydrate active enzymes (CAZymes) and HRGPs. Hydroxyrich glycoproteins (HRGPs) are known in both algae and land plants and can be distinguished by long runs of proline interspersed with a subset of additional amino acids, where the sequences are usually organized as quasi-repetitive modules (*94, 95*). HRGPs comprise up to 10% of land-plant cell walls and are apparently the sole constituents of *Chlamydomonas* walls (*96*).

The predicted Mamiellales HRGP-encoding genes (fig. S13, table S15-16) encoded "chimeric" proteins with both globular- and P-rich shaft domains. Certain shaft repeat modules (e.g. SP2-SP6, XP3) were also found in other lineages. In *Chlamydomonas*, HRGP globular-domain homologs are only found in other volvocine algae. In contrast, one *Micromonas* HRGP globular domain was homologous to ADAM (fig. S13b), a metalloproteinase active in cleaving extracellular portions of transmembrane proteins (*97*) and found in animal, fungal and red-lineage, but not other green-lineage, genomes. Another globular domain found in *Micromonas* was homologous to vinculin, an adhesion protein in mammals and other opisthokonts.

The glycan components of cell-walls vary from one species to another but usually contain terminal β-L-arabinosyl decorations and sometimes even linear arabinan side chains (*98*). In *Arabidopsis* at least two enzymes from the "carbohydrate-active enzyme" (CAZy) family GT77 are involved in the β-L-arabinosylation of cell wall components (*99*), and GT77 was one of the most abundant families of glycosyltransferases found in *Micromonas* (table S17). This family was also abundant in *Ostreococcus* and in the fragmented gene models of *C. reinhardtii*.

Higher plants are extremely rich in enzymes that build, modify and cleave glycosidic bonds (*100*), whereas the two *Micromonas* genomes and *Ostreococcus* contained a considerably smaller number of genes encoding the collective suite of these enzymes – CAZymes (see www.cazy.org; table S17). For instance, plants have about 10 times more genes encoding glycosidases than found in *Micromonas* or *Ostreococcus*.

A closer inspection (table S17c) revealed that this 'overview' gene count masks the fact that *Micromonas* contained some CAZymes absent from plants. For example, the *Micromonas* genomes contained a complete set of CAZymes required for the synthesis and remodelling of peptidoglycan: a candidate UDP-GlcNAc: N-acetylmuramyl-(pentapeptide) PP-undecaprenol N-acetylglucosaminyltransferase (CAZy family GT28), a candidate bifunctional b-glycosyltransferase/penicillin-binding transpeptidase (CAZy family GT51) and a candidate

peptidoglycan lytic transglycosylase (CAZy family GH103), none of which were found in *Ostreococcus* or plants. (As a side note, we commonly use penicillin while rendering *Micromonas* cultures axenic and have not observed *Micromonas* senstitivity to this antibiotic.) The finding of these CAZymes was particularly interesting given that we also found genes encoding monogalactosyldiacylglycerol synthase (MGDG) and UDP-GlcNAc: N-acetylmuramyl-(pentapeptide) PP-undecaprenol N-acetylglucosaminyltransferase (MURG). It has been hypothesized that plant MGDGs evolved from MURG when plastidic peptidoglycan synthesis became dispensible (*101, 102*). Both *Micromonas* genomes contained genes for MURG and for MGDG, strongly suggesting that MGDG appeared before peptidoglycan synthesis was lost, most likely via an early duplication of the ancestral cyanobacterial *MURG* gene. Again, the genes responsible for encoding these proteins are missing from *Ostreococcus*, suggesting that *Micromonas*, but not *Ostreococcus*, is still capable of producing peptidoglycan. Given the posited pressure for genome reduction in *Micromonas*, the presence of the genes for peptidoglycan synthesis suggests that they are functionally significant.

A second example documenting a more complex situation than revealed by the CAZyme overview statistics, is that genes encoding enzymes involved in starch metabolism (CAZy family GH13, GH77, GT5, GT35) were as abundant in *Micromonas* as in higher plants, see also (*43*), the exception being that β-amylases (CAZy family GH14) are approximately 5 times more abundant in plants.

A third example of how overview gene counts can be misleading is that, in contrast to plants, the two *Micromonas* species encoded a secreted protein with multiple family 1 carbohydrate-binding modules (CBM1). This type of module is found almost exclusively in fungi, with the exception of a few proteins from Rhodophyta and Stramenopiles such as *Phytophthora* (www.cazy.org). In fungi these protein modules are found attached to cellulases and hemicellulases and target these enzymes to cellulose (*103*). In *Micromonas*, their role is likely to be different as these modules were not attached to a catalytic module. The target ligand unfortunately could not be predicted.

Several families of CAZymes common in plants were completely absent or extremely reduced in *Micromonas* and *Ostreococcus* (table 17c). In cases where experimental work has been conducted, the functions of the plant-encoded proteins have been assigned to the synthesis and remodelling of the plant cell wall polysaccharides (*104*). By extension, the differences between higher plants and the *Micromonas* species offer an indirect predictive tool to identify CAZyme families involved in plant cell wall polysaccharide biosynthesis.

Finally, sucrose metabolism appeared to be absent from the *Micromonas* (and *Ostreococcus*), genomes as deduced from the complete lack of invertases (CAZy families GH32 and GH100), and there was no evidence for chitin metabolism (CAZy families GH18, GH19, GH20).

7.) Transcription factors. Among the 65 TF gene families found in land plants, about half (31) are only found in the green lineage (fig. S9, S10, table S11). We propose that 10 of these were components of the "basal green toolkit" since they were also found in *Micromonas* (several are lacking in *Chlamydomonas* and *Ostreococcus*), and that their absence in certain green radiations is a consequence of gene-family loss. For example, a pattern similar to that seen for YABBY (main text), was also seen for the ULT TF family (fig. S9, S11, table S11), and the Mamiellales and land plants share TFs not found in *Chlamydomonas* (e.g. WOX, GRF, ULT) (fig. S12). Similarly, *Chlamydomonas* TFs formed clades with land plants that were separated from prasinophyte homologs (e.g. Alfin, HSF, SBP).

Several TF gene families were chosen for cladistic analysis because their members are important players in land plant development and physiology. The phylogenies revealed distinctive patterns of ancestry in the 3 algal lineages represented by *Chlamydomonas*, *Micromonas* and *Ostreococcus*. Alfins (fig. S10) from land plants contain a conserved N-terminal domain plus a C-terminal PHDfinger (PHDf) domain as the DNA-binding motif. One of two alfin members from the Volvocales also possesses a C-terminal PHDf domain. Four alfin genes from *Micromonas* and *Ostreococcus* did not contain a C-terminal PHDf domain (and may not function as transcription factors).

The AP2 domain is found in 4 subfamilies of land-plant transcription factors (AP2, ANT, ERF/DREB, and RAV), and is thought to be derived from HNH endonucleases (*105*). Some AP2 and ANT subfamilies contain 2 copies of AP2/ERF domains, while others have only one. Land-plant ANT members have distinctive 8-aa insertions in the first repeat; similar ~12-aa insertions identify ANT homologs from green-algal lineages (fig. S13). Two-copy members are designated as AP2 families, although the AP2 members from land-plant and green algae do not form a monophyletic clade. Except for RAV, 3 subfamily members were found in all 3 green-algal lineages. Subfamily expansions in land plants have been mainly in ANT and ERF/DREB. Algal sequences are more diversified in domain sequences and copy numbers, even between the 2 *Micromonas* species (up to 4 domain copy numbers could be found in an ORF, indicated as 1R-4R, fig. S10).

Overall, our findings contribute additional features to the ancestral TF toolkit identified by Floyd and colleagues (*106*) in lower land-plants (the bryophyte *Physcomitrella* and the lycophyte *Selaginella moellendorfii*) encoding most of the higher land plant TF families. Certain differences in TF distributions (such as YABBY and ULF) may prove to have arisen by gain (e.g. by HGT), or by independent domain shuffling and assembly, in *Micromonas*. Many of these TFs presumably served different functions in ancestral lineages before being recruited to mediate more complex developmental pathways. These data, together with other features of the *Micromonas* genomes, indicate that the most recent common green ancestor was a flagellated (see also text S8, table S18) proto-prasinophyte, and that over the past >1 billion years, 2 independent radiations led to the modern chlorophytes/ prasinophytes and the charophytes/land plants.

Pairwise comparisons of related Mamiellales genomes allowed analysis of the stability of TF gene numbers (table S11). Families with less than 5 members were manually analyzed, utilizing BLASTP and TBLASTN searches for homologs, retrieval of orthologous information via VISTA tracks provided on the JGI genome site, and collecting members by IPR code searches of the final protein dataset. 27 out of 43 families in *Micromonas* have less than 5 members, and 5 out of the 27 families showed changes in the number of members between the 2 *Micromonas* species, all of which resulted from clade-specific expansion of members and none from the loss of a specific clade. None of the 5 cases involved tandem duplication. In contrast, clade-level loss or emergence was evident in large gene families (e.g. homeodomain (HD) [table S12] and RWP-RK [table S14]). Comparisons among algal lineages (*Chlamydomonas*, *Ostreococcus*, *Micromonas*) revealed numerous examples of loss or emergence of new members with either diverged sequences (5 cases out of 10 small gene families) or distinctive domain architectures (e.g. RR-Dof, 3-4 repeats of AP2/EREBP). RR-Dof exemplified a new clade arisen via domain shuffling. However, the simplest explanation for these patterns was again that each clade was present in the common ancestor and lost in subsequent radiations.

Homeodomain (HD) TFs are thought to have facilitated the evolution of multicellularity (*107, 108*). Of the 9 HD classes identified in *Micromonas*, 2 were unique (one to RCC299 and CCMP1545, the other only in CCMP1545). We found evidence of HD diversification *via* domain shuffling and class-specific expansions. For example, all the Mamiellales had members of the non-TALE WUSCHEL homeobox (WOX) gene family (fig. S15b), an important mediator of embryo development in monocots and dicots that is also found in moss (*109*) but not in *Chlamydomonas*. This class appeared to have undergone an expansion in CCMP1545 but not the other Mamiellales. All green algal genomes investigated contain 3 TALE homeoprotein classes, as is also true for *Cyanidioshyzon merolae* and known Rhodophyta (*110*). The *Micromonas* genomes had single representatives for each of these classes (fig. S12). Land plants appear to lack the BELL-rel1 group. The *Micromonas KNOX* members (Prot. ID 62285, RCC299 and Prot. ID 8801, CCMP1545) had a KNOX1 domain that was conserved in all KNOX class members (*111*). Only one TALE-encoding sequence within the *Micromonas* genomes had EST support (CCMP1545 Prot ID 8801). The fact that TALE HD classes have not expanded, even though non-TALE members have, was consistent with the postulated critical role of the former in the sexual cycle (*110*). The *Micromonas* strains also shared one HD class, GSP1, unique to green algae. Three apparently prasinophyte specific HD classes, one HOXDDT-related, one OCP3-related and one PHDf-containing (PHDf2), were also identified.

8.) Flagella related genes. Flagellar genes within the two *Micromonas* strains (table S18) were examined using known *C. reinhardtii* flagellar genes (*112*), detailed at http://labs.umassmed.edu/chlamyfp/index.php as a query database. To increase certainty of the resulting gene calls in CCMP1545 and RCC299, the candidate *Micromonas* genes were subsequently BLASTed against the ExPasy database. TBLASTN analysis of known *C. reinhardtii* genes was also performed against the *O. tauri* and *O. lucimarinus* genomes, a genus that lacks flagella (*113, 114*).

We identified most of the major protein-encoding genes known to be involved in flagellar structure and maintenance in *C. reinhardtii*. However, both *Micromonas* appear to lack Tektin, a protein essential for microtubule structural integrity (*115*). The conserved consensus sequence (RPNVELCRD) common to all Tektins was not found by BLASTP or TBLASTN in either species. BLASTP, TBLASTN and BLASTX using the *Chlamydomonas* sequence also did not yield hits. Interestingly, the two genes annotated as Tektin in *Chlamydomonas* do not contain the conserved consensus sequence (which seems to be restricted to animal tektins). In contrast to the overall flagellar gene complement identified in *Micromonas*, which is a motile organism, the flagella-less *Ostreococcus* lack the genes involved in intraflagellar transport and radial-spoke formation. However, both *Ostreococcus* genomes (as well as *Micromonas*) contained some genes thought to be associated with the flagellum in *C. reinhardtii*, e.g. a mating-related CALK protein kinase and the cGMP-dependent protein kinase, see also (*116*). Several other genes identified in *Ostreococcus* that could appear to be flagellar related (e.g. α- and β-tubulin) are also known to be components of the cytoskeleton.

Both *Micromonas* and *Ostreococcus* species have homologs of the phototropin blue light receptor genes and a mastigoneme-like gene associated with the *Chlamydomonas* flagellum. Mastigonemes are hair-like structures found on the flagella of the Heterokonta (Chromalveolata) and cryptophyte algae (*117*) but not visible, nor thought present, in the prasinophytes.

9.) Nutrient acquisition and transport. We compared transporter profiles to those of *Physcomitrella*, *Arabidopsis*, and *Oryza sativa* (for this analysis these species represented "land-plants") as well as diatoms, *Ostreococcus* and other organisms (table S19). CCMP1545 has undergone specific losses compared to the other green- and red-lineage genomes, including several transporters that may affect nitrogen utilization. Bacterial-like transporters in the BCCT family (betaine/carnitine/choline transporters), specific for compounds containing a quaternary nitrogen atom, were absent from CCMP1545 and the land plants although present in other Mamiellales and red-lineage genomes. Amino-acid polyamine organo-cation (APC) family members - amino acid permeases and proton-dependent oligonucleotide transporters (POT) - potentially related to acquisition of nitrogen and other nutrients were also missing, although present in land plants and diatoms. A nucleobase-cation symporter-2 (NCS2) was also present in RCC299, but not CCMP1545 (or *Ostreococcus*) that has been shown to mediate uptake of purine compounds in other organisms (*118*) and is down-regulated upon ammonium exposure in fungi (*119*).

Two families traditionally associated with calcium transport were also present in RCC299 and *Ostreococcus* but not CCMP1545: 1) the transient receptor potential Ca^{2+} channels (TRP-CC) was missing from CCMP1545 and land-plant genomes although found in red-lineage genomes; 2) annexins, which have been considered calcium-dependent phospholipid-binding proteins but are also linked with inhibition of exocytosis and endoctyosis, signal transduction, organization of the extracellular matrix, resistance to reactive oxygen species and DNA replication, were found in the other Mamiellales, land plants, diatoms and bacteria but not CCMP1545 or *Chlamydomonas* (see also table S19).

Other examples of differential distributions of transporters include a nucleobase-cation symporter-2 (NCS2) found in RCC299, land-plants and *Chlamydomonas* but not in CCMP1545 or *Ostreococcus*. Both *Micromonas* species also had oligopeptide transporters (OPTs), some of which are known to transport phytochelatins (*120*), but are absent from *Ostreococcus*, *Chlamydomonas*, *Physcomitrella* (*109*) and diatoms. Some families identified in *Micromonas* were not found in other green-lineage genomes but were present in marine chromalveolates (the diatoms), e.g. the metazoan-like neurotransmitter-sodium symporter (NSS) and the phosphate-Na^+ symporter (PNaS). Losses in *Ostreococcus* (as well as CCMP1545) may relate to nutrient rich environments inhabited by the strains with sequenced genomes, ribotypes of which are not seen in the open ocean (*114*). An open-ocean strain of *Ostreococcus* is now being sequenced which should help support, or dismiss, this hypothesis.

10.) Forces of mortality. Because these picoplankton are important primary producers, but too small to sink on their own accord, the specific forces that dictate mortality are important to their role in global carbon cycling. Perhaps the best studied agent of mortality in *Micromonas* is their viruses, which have been shown to contribute to the demise of blooms (*121*). Interestingly, phylogenomic analyses revealed virally derived gene clusters within the nuclear genomes of RCC299 and CCMP1545, although the specific affiliations observed (*O. tauri* and *E. huxleyi* viruses) are likely a function of there being relatively few sequences available for viruses of marine protists.

C-type (calcium-dependent) lectins represent a diverse protein family that plays important roles in cellular interactions, innate immunity and glycoprotein turnover (*122*). In marine systems, it has been proposed that lectins on phytoplankton prey cells may bind to target carbohydrate ligands on protistan predators, increasing the feeding efficiency and selectivity of

grazing (*123, 124*). Therefore, identification of C-Type Lectin Domain (CTLD) containing proteins in *Ostreococcus* and *Micromonas* could provide valuable insights on environmental interactions and potentially new approaches for experimental investigation of grazing in the natural environment. However, in-depth analysis indicated that this categorization of the auto-predicted KOG4297 (C-type lectin domains) maybe incorrect. A number of putative proteins fell within KOG4297, but we could not validate these as CTLDs in *Micromonas* RCC299 or CCMP1545, *T. pseudonana*, or either of the *Ostreococcus* genomes. Furthermore, using independent searches we did not detect homologs of the CTLD-like domains of bacterial proteins with phylogenetic relationships to animal CTLDs, or with the animal link domains, which are thought to have evolved from canonical CTLDs by loss of the long loop region. In order to verify our methods we also searched *Monosiga brevicolis* (a predatory choanoflagellate) for CTLDs (e.g. Q7YZH9), and detected strongly conserved CTLDs (Ca-binding, QPN motif (*122*)) as well as a CCP domain.
Based on our analysis, if KOG4297 does represent CTLDs, they are not only very highly diverged but also unlikely to bind Ca^{2+}/carbohydrate in the same way as C-type lectins. Some may be Cys-rich domains from which CTLDs evolved, but they appear to be only possible distant homologs. Validation of such relationships would require greater knowledge of the early stages of CTLD evolution than currently available.

11.) Polyketide Synthetases. Polyketides are a structurally diverse class of natural products derived from the polymerization of acetyl and propionyl subunits in a process similar to fatty acid synthesis. Such compounds are of pharmaceutical and biomedical interest because many have potent biological effects as antibiotics, anti-tumor compounds, natural insecticides and immunosuppressive agents (*125*). Numerous functions in nature have been proposed for these secondary metabolites, ranging from chemical defense against predation to fatty acid elongation to complex cell communication. The presence of polyketides in bacteria, fungi and streptophytes has been known for decades, but their occurrence in protists has only recently been confirmed (*126, 127*).

The polyketide synthase (PKS) enzymes are large multi-domain complexes that structurally and functionally resemble fatty acid synthase (FAS) enzymes involved in lipid metabolism. FAS and PKS catalyze the sequential condensation of acyl units onto a growing carbon chain and both enzymes possess a similar set of functional domains: ketoacyl synthase (KS), acyl transferase (AT), ketoacyl reductase (KR), dehydratase (DH), enoyl reductase (ER), phosphopantetheine attachment site (PP) (or acyl carrier protein [ACP]), and thioesterase (TE). Whereas FAS is dependent upon the presence of the complete set of aforementioned functional units, the minimal structure of PKS requires only the ACP, KS and AT domains for the condensation reaction. The other domains (when present) catalyze the stepwise reduction of the initial carbonyl units (*128*).

Polyketide synthases are generally classified into three major structural sub-groups. Type I PKSs are large, highly modular proteins, whereas Types II are aggregates of monofunctional proteins. These enzymes include several modules, some of which are responsible for chain elongation while others catalyze the associated reduction steps. In most bacteria, each module directs one round of chain extension and post-condensation modification to generate non-aromatic polyketides. In fungi and some bacteria, each module/enzyme of Type I PKS is used iteratively, yielding either aromatic or non-aromatic compounds. By comparison, Type II PKSs are multi-protein complexes whereby the individual enzymes are used iteratively for each cycle of chain extension. These Type II complexes are found exclusively in bacteria for synthesis of

aromatic polyketides. The Type III PKSs, also known as chalcone synthases, are homodimeric and function iteratively as condensing enzymes. Their distribution was believed to be essentially restricted to streptophytes, within which they employ unusual starter units to act directly on acyl-CoA thioesters, independently of PP. Recently, microbial genome sequencing has revealed additional Type III PKSs in bacteria, most of which are of unknown function (*129*). Others, studying bacterial Type I PKS evolution, concluded that FAS and PKS passed through a long joint evolutionary process with the modular PKS type arising from bacterial FAS and primary iterative PKS (*130*).

The *Micromonas* species each contained 2 *PKS* genes, while 3 are found in *Ostreococcus*. Those in RCC299 are 14,181 aa and 6,849 aa in size, while those in CCMP1545 are 19,361 aa and 7,931aa, respectively. It is interesting that prasinophytes have PKS type I and not type III given that PKS III was formally described as the "plant PKS form" although recently also found in bacteria (*129*). John et al (2008) also identified candidate *PKS* sequences in the genomes of *C. reinhardtii* as well as *O. tauri* (*127*) and *O. lucimarinus* (*37*). The presence of Type I *PKS* genes among green algae was surprising because such sequences were not found in the genomes of streptophytes and red algae (*127*)). To date no other member of the green lineage exhibits PKS type I, and functions for these putative *PKS* genes cannot yet be assigned without further biochemical analyses. The different numbers of *PKS* genes between *Micromonas* and *Ostreococcus* and the different domain structures between the two *Micromonas* speciess suggests that different products would result from these enzymes (fig. S16a).

We further analyzed the β-ketoacyl synthase (abbreviated as KS) domain, the most conserved domain within Type I PKS genes (*131*). This domain has the greatest potential for revealing divergent homologs and thus provides an informative basis for comparative and phylogenetic analyses. Phylogenetic analyses produced the same clade structuring as previous publications (*131*), with the bacterial clades, the metazoan FAS, and Ascomycota KS clades (both reducing and non-reducing) being bootstrap (BP) supported (fig. S16b). The KS sequences from *Micromonas* fell into the Chlorophyta Clade 1, a well-supported monophyletic clade. Whereas the *Ostreococcus* Type I PKS sequences fall into two clades - Chlorophyta Clade 1 and Clade 2 - *Micromonas* KS domains fell into Clade 1 exclusively. Chlorophyta Clade 1 (sequences from *Chlamydomonas*, *Ostreococcus*, and *Micromonas*) formed a relatively tight but mixed group, not reflecting the presumed long independent evolutionary history of these three taxa (*132, 133*).

The view that Type III PKS was the characteristic PKS for plants is now challenged by the finding of Type I PKS in all five available complete chlorophyte complete genomes. Taking into account the grouping of PKS sequences of *Chlamydomonas*, *Ostreococcus,* and *Micromonas*, and the fact that those species diverged long ago, this would put the gene divergence event at the base of chlorophyte evolution. However, we could not identify Type I PKS genes in other lineages, i.e. from the Rhodophyta *Cyanydioschyzon merolae* and *Galdieria sulphuraria*, the only two red algae sequenced thus far. A simple explanation is that these genes were lost during evolution of the rhodophytes. Moreover, these two rhodophyte species are unicellular organisms that live in acidic hot springs, and thus may not be good representatives of this lineage overall.

12.) Carbon concentrating mechanisms (CCMs). Algae have developed CCMs, presumably largely due to the fact that the central enzyme in CO_2 uptake (RUBISCO) has low CO_2 affinity (*134*). In addition, RUBISCO is less than half-saturated in seawater CO_2 levels, and thus can be rate-limiting under bloom conditions. Phytoplankton appear to have developed several mechanisms to alleviate this issue, the best known being a CCM driven by transport of

bicarbonate and carbonic anhydrases. Both *Micromonas* species had several bicarbonate transporters and at least one α-carbonic anhydrase (αCA) with targeting signals for the chloroplast lumen (table S5, fig. S17-18). This predicted CCM type showed strong similarity to *Chlamydomonas* and *Ostreococcus* but not to diatoms (fig. S17). A C_4-like CCM was also found in *Micromonas*, where RCC299 and CCMP1545 differ in the copy number and subcellular targeting of proteins; for example, CCMP1545 had a cytosolic βCA not found in RCC299 (Table S5). C_4-photosynthesis is an important feature in some higher plants in tropical environments. A CCM based on carbonic anhydrases was reported in *Micromonas* a decade ago and proposed to provide a competitive advantage given absence of such activity in other small phytoplankton (*135*).

Micromonas, in contrast to *Ostreococcus*, had lumenal-targeted αCAs (fig. S17) while the βCAs from *Ostreococcus* were targeted to the stroma of the plastid. Lumenal-targeted αCAs are well known from *C. reinhardtii* and higher plants (*136*) and are essential for growth under ambient CO_2 concentrations (*136*). In addition, it has been experimentally shown in *Chlamydomonas* that the lumenal-targeted CA is necessary for proper CCM function at low C_i by providing an ample supply of CO_2 for RUBISCO (*137*). Cytosolic CAs in CCMP1545 and in both *Ostreococcus* species might be able to capture leaking CO_2 for re-import of bicarbonate into the plastid.

Differences between the Mamiellales were most pronounced with respect to the types of CAs. The δCAs found in *Ostreococcus* do not appear to be present in *Micromonas*; however the distinction of this class of CA has recently come into question (*138*). Lumenal-targeting of the CAs and the NADP-dependent Malic-Enzymes (NADP-ME) in *Micromonas* may provide advantages over decarboxylation in the stroma, by reducing leakage caused by passive diffusion from the stroma into the cytosol, akin to the experimentally verified *Chlamydomonas* lumenal-targeted CA necessary for CCM function at low C_i (*137*). In *Ostreococcus*, decarboxylation by NADP-ME is thought to occur in the chloroplast stroma (*36*). Although CO_2 generated in the plastid stroma may be partially lost via diffusion, CAs located inside the cytosol could prevent such losses. This type of loss would be less likely to impact diatoms (fig. S17) since they have 4 membranes surrounding their chloroplasts and some also have a girdle lamella (Thoms et al. 2001). The prasinophytes have only 2 membranes and no girdle lamella, allowing CO_2 to more easily diffuse out of the plastid. Their small size also increases losses by diffusion due to shorter distances between localization of CO_2 generation and the cell exterior and, as noted above, the CO_2 affinity of RUBISCO in some green algae is low (*134*) compared to red algae.

Each of the *Micromonas* strains also has a single phosphoenolpyruvatcarboxylase (*PEPC*) gene copy. Both proteins show indication of cytosolic localization. This carbon asimilation step appears to produce oxalacetate (OAA), which is transported into the plastid by specific transporters. OAA becomes reduced inside the plastid by malate dehydrogenase. Malate is decarboxylated by NADP-ME and in both *Micromonas* this gene had a Tat signal peptide similar to findings for αCAs (figs. S17-18). Three different 2-oxoglutarate/malate translocators were identified in each species, with at least one in each apparently chloroplast- targeted. Interestingly, only one homolog seemed to be plastid-targeted in RCC299, but 2 were in CCMP1545. This decarboxylation reaction produces pyruvate, which is phosphorylated by a pyruvatephosphate dikinase (PPDK) to form phosphoenolpyruvate (PEP), which is exported into the cytosol and can then be used again for fixation of HCO_3 by the PEPC. The CCM findings are particularly interesting given that, although chromalveolates (diatoms) and prasinophytes have highly differentiated evolutionary histories (*139*), the invention of multiple CCMs as predicted here

presumably reflects a requirement for baseline survival and not a unique competitive advantage harbored by a few taxa.

13.) Mechanisms for alleviating oxidative stress. Photosynthetic organisms use a variety of mechanisms to temper the accumulation of Reactive Oxygen Species (ROS), including scavengers of both ROS and heavy metals, because ROS production is often facilitated by the proximity of electron transport and oxygen production in the photosynthetic apparatus.

Interestingly, both *Micromonas* genomes contain genes for phytochelatin synthase, although *Ostreococcus* does not. Phytochelatins are peptides found in higher plants and *Chlamydomonas* that bind heavy metals and have been implicated in mitigating copper toxicity in coastal settings; they are also active in reactive oxygen species (ROS) scavenging. While from an ecological perspective it is not surprising to find phytochelatins in *Micromonas*, it is surprising given that, as found for *Ostreococcus* (which does not have phytochelatins), there has been an apparent reduction in use of iron co-factors (also used in mitigation).

A suite of other known ROS scavengers were identified in both *Micromonas* genomes (table S20). For example, superoxide dismutases (SODs) catalyse the conversion of superoxide radicals to molecular oxygen and hydrogen peroxide. Four major groups of SODs are known and distinguished by their metal co-factors: Fe, Mn, Cu/Zn, and Ni. These SOD metalloforms are not equally distributed among marine phytoplankton, and phylogenetic evidence coupled with subcellular localizations suggest different selective pressures (*140*). Within the Mamiellales, the SODs were similar. Each of the four sequenced genomes contained only one of the Fe/Mn family of SODs and these likely do not bind iron, based on analysis of alignments and two critical conserved residues which distinguish Fe-binding from Mn-binding SODs (*140*). Both *Micromonas* genomes also had at least 3 Cu/Zn SODs. This suggests that although the *Micromonas* species from the sequenced clades have been found in some coastal settings, they may be prepared for life in oligotrophic waters (where a high requirement for iron can be detrimental). In contrast, the predominantly estuarine diatom *T. pseudonana*, which likely had a freshwater ancestor (*141*), can produce both Fe and Mn binding SODs but not Cu/Zn SODs, possibly due to its evolution in largely iron-rich waters (*142, 143*).

14.) Introner Elements. We have named the repetitive DNA elements discovered within introns of CCMP1545 genes "Introner Elements" (IE), specifically, IE1 (ca. 210 bp, 6,987 identified), IE2 (ca. 130 bp, 1525 identified), IE3 (ca. 145 bp, 958 identified) and IE4 (ca. 190 bp, 434 identified). Notably, the four classes of IEs were not equally abundant nor were they uniformly distributed over the genome (table S22a, figs. S19-22). The low GC-region (most of scaffold 2) as well as the smallest chromosome of CCMP1545, which also has a lower than average %GC content, were both almost devoid of IE. Between the 4 classes there was no clear distributional bias (fig. S22) of one over the other (other than the large numerical differences), although specific regions (i.e., scaffold 19 and much of scaffold 2) varied tremendously in representation of IE.

Almost 10,000 IE were found in the CCMP1545 genome (table S22). Surprisingly, no IE homologs were found in the RCC299 genome (see also table S23, fig. S20 on repeat sequences). It should be noted that scrutiny of IE positioning revealed that a significant fraction of automatically generated gene models harboured IE in CDS. In such cases, manual examination led to correction (primarily done when EST support was available) of faulty gene models, resulting in the relocation of the particular IE into an intron (see table S22b). To determine the

extent of colinearity of genes and introners, we also manually investigated a subset of the cases where IE were identified on the opposite strand from a gene-coding sequence, finding they should actually have been placed in (introns of) the opposite-strand transcripts (e.g. based on directional EST data).

Introners had several peculiar features (Fig. 3, fig. S21). They were internal to coding sequences and colinear with genes (i.e. found on the coding strand), suggesting that a transcription-linked mechanism was employed during propagation. The 5′-most sub-sequence was the most conserved, up to 70 bp downstream of the donor site, even in degenerate members of the family, but the 3′ side showed an interesting array of short modules repeated three times in tandem. These modules comprised two conserved motifs separated by a variable sequence, with the consensus CTTCAA$N_{(6\text{-}13)}$(C,T)(C,T)G<u>A</u>CG. Interestingly, the second motif was very likely a branch-point motif, as shown by its excellent complementarity to the U2snRNA branch site, 5′-guGUAGUAu-3′ (fig. S21). IE2 also had the same branch-point modules, but they were repeated only twice (fig. S21b). In addition, they had long pyrimidine tracts, T_5XC_3 and CT_7, upstream and downstream, respectively, of these tandem branch-point repeats. This typology showed a bias towards highly efficient splicing, and suggests that the first CTTCAA motif in branch-point modules may play a role in splicing as well, perhaps serving as an intron-splicing enhancer. IE did not appear to be positioned in a manner that could cause gene inactivation. Clear sequence consensus could not be found on either of the exon borders of IE, although the proximal sequences did show a tendency for higher GC content.

An exhaustive search of public nucleic acid databases returned no IE hits except for Sargasso Sea metagenomic data (*15, 144*). Flanking sequences from these Sargasso Sea IEs displayed high identity to flanking sequences of CCMP1545, and had higher identity to CCMP1545 than to RCC299. CCMP1545 belongs to clade M_V, but M_V 18S rRNA gene sequences were not found in the Sargasso Sea metagenomic data, possibly because the sequencing depth of that study was sufficiently low that such a sequence was ‘missed.’ That said, an 18S rRNA gene sequence from *Micromonas* clade M_IV was detected, as were *Ostreococcus* 18S rRNA gene sequences. Therefore, M_IV may also harbor IE.

Overall, IE compose a large (9%) percentage of the CCMP1545 genome, and are largely responsible for its 5% larger size than the genome of RCC299: artificial IE removal renders the CCMP1545 genome 4% smaller than that of RCC299. Given the lack of known RNAi components in CCMP1545 (see below), one hypothesis is that IE may function as interfering RNAs in this species.

15.) TPP Riboswitches and thiamine biosynthesis. A putatively archaic mode of gene regulation comes in the form of riboswitches, untranslated mRNA regions that regulate gene expression in a process involving metabolite binding. In bacteria these switches modulate premature transcriptional termination or translational initiation whereas their primary role in fungi and plants is control of alternative splicing in several thiamine-biosynthetic genes (*145-147*). The most widespread class (*145*) responds to thiamine pyrophosphate (TPP) and is thought to date to the ancient RNA world, akin to the deep roots attributed to the importance of biologically active thiamine (vitamin B_1) (*148*). While several thiamine pathway elements common to *Chlamydomonas* (*147, 149*) and higher plants were not found in the Mamiellales, the identification of putative TPP aptamers associated with anomalous genes suggests that these genes may represent ancient components of thiamine-biosynthesis pathways.

Nevertheless, the Mamiellales appear to be thiamine auxotrophs. Examples of thiamine related *Chlamydomonas* genes missing from both *Micromonas* and *Ostreococcus* include *THIC* and plant-like *THI4*, which synthesize the pyrimidine and thiazole precursors respectively, again indicating that thiamine synthesis is not possible. Other classical genes involved in the thiamine pathway (*TH1*, *THIM* and a *TENA*/*THI4* superfamily member) were present in CCMP1545 and *Ostreococcus* but absent from RCC299. In CCMP1545, these genes formed a block of genes co-localized with the TPP-riboswitch containing *SSSF-F*, which had similarity to a putative pantothenate:Na^+ symporter (vitamin B_5) *PANF*; this block was bordered by unrelated genes on the 5′ and 3′ sides. Orthologs of these same bordering genes were co-localized in RCC299 as well, but the entire block of thiamine genes, and *SSSF-F*, were absent from the genome. A number of other shared and differentiated thiamine-related features were also identified. Shared features related to thiamine utilization included genes with potential homology to the folate/thiamine transporter family *THT1* (RCC299, Prot. ID 99797; CCMP1545, Prot. ID 54517) and a homolog (*TPK1*) of *THI80* in yeast, which converts thiamine to its biologically active form, thiamine pyrophosphate (TPP). Thus, thiamine could potentially be transported by THT1 and phosphorylated to TPP by TPK1. Other options appeared available for use of precursors derived from thiamine degradation products, at least in CCMP1545 and *Ostreococcus*. For these Mameillales, *TENA/THI4* superfamily members had highest identity (apart from to each other) to bacterial *TENA* genes involved in thiamine salvage in bacteria (*150*) and yeast (*151*).

The absence of *TH1* from RCC299 and the absence of classical thiazole precursor synthesis in both *Micromonas* species made the role of the *NMT1* gene in RCC299 and CCMP1545 unclear. Its presence, and the presence of *PDX1* and *PDX2*, indicate that the *Micromonas* species produce vitamin B_6, with NMT1 putatively synthesizing the pyrimidine precursor (HMP-P) for thiamine synthesis from vitamin B_6, as it does in fungi. The TPP riboswitch associated with RCC299 *NMT1* also implicates NMT1 in the vitamin B_1 pathway. In yeast, *NMT1* is multi-exon and has a TPP riboswitch with an experimentally confirmed role in alternative splicing (*146*), a role unlikely in RCC299 because *NMT1* had only one exon. In RCC299 there were additional possibilities for transport including a putative SLC19 family protein (Prot. ID 104570). SLC19 is seen in many metazoa, but never before in green- or red-lineage organisms. The RCC299 version was implicated in thiamine transport, given higher relatedness to SLC19A2 and SLC19A3 (thiamine transporters) than to SLC19A1 (folate transporter), all belonging to SLC19 (*152*). Furthermore, the RCC299 *FOLR*-like gene with an associated TPP riboswitch was distant enough from other folate receptors that it could also play a role as a thiamine receptor, although it has only one TMS and appeared unlikely to mediate thiamine transport by itself. Like the putative RCC299 *SLC19* gene, *FOLR*-like was affiliated with metazoan sequences and present in *Bigelowiella natans*, a member of the eukaryotic supergroup Rhizaria which resulted from a secondary endosymbiosis event involving a green-lineage organism (*139*). Homologs for these genes were not found in CCMP1545. The RCC299 EFG domain-containing gene of unknown function (not found in CCMP1545) has similarity to metazoan delta proteins and a predicted signal peptide; in metazoa these Ca^+ binding domains are associated with membrane-bound and extracellular proteins. We did not find riboswitches associated with the *FOLR*-like gene or EFG domain-containing genes in other organisms, although the search was not exhaustive.

In the case of the 3′ riboswitch conserved between CCMP1545 and RCC299, for which the associated genes do not share homolgy (i.e. *FOLR*-like, RCC299; *SSSF-P*, CCMP1545), the downstream gene (DuS) is conserved in all 4 Mamiellales (Fig. 4). In RCC299, the *FOLR*-like gene is weakly affiliated with the metazoan folate receptor (*FOLR*) and harbors the conserved 3′

riboswitch. But this gene is not found in the CCMP1545 genome. Instead, in CCMP1545, the 3′ riboswitch is associated with a sodium/substrate symporter family gene (SSSF) herein referred to as *SSSF-P*, a gene distantly related to the proline symporter *PUTP* and present in *Ostreococcus* as well. In addition to finding a 3′ riboswitch in *SSSF-P* in CCMP1545, we also identified a 5′ riboswitch that is found as well in *Ostreococcus* and affiliated with the same gene in both *Ostreococcus* strains (Fig. 4).

The TPP riboswitch-gene affiliations were identified in other lineages as well. *SSSF-P* homologs, with putative TPP riboswitches were present in two SAR11 bacterial clade genomes (HTCC1062 and HTCC1002); they were located at the 5′ side of this gene and have 98% identity to each other in the SAR11 genomes. The SAR11 riboswitches form P2 and P3 regions and have the conserved CUGAGA motif, but have less classical secondary structures in the P4 and P5 regions. We also identified unreported putative riboswitches in *Chlamydomonas* and *Volvox*. Previously reports on *Chlamydomonas* identified aptamers associated with classical thiamine metabolism genes (*THIC* and *THI4*) (*147*). The newly identified riboswitches were, as in CCMP1545, affiliated with a gene (*SSSF-F*) with similarity to a putative pantothenate:Na^+ symporter (vitamin B_5; *PANF*). Although the *FOLR*-like gene (RCC299), as well as an *SLC19*-like gene (RCC299), may point to thiamine-related avenues in RCC299, our search did not reveal these RCC299 riboswitch-gene associations elsewhere. However, it is likely significant that *SSSF-F* and *SSSF-P* have associated TPP riboswitches in other organisms and that in CCMP1545 and *Ostreococcus* these genes are co-located with thiamine-related genes. The similarities seen across disparate lineages indicate that the riboswitch associations have functional significance. Consequently we hypothesize these genes are ancient thiamine-pathway genes, in the case of *SSSF-F* and *SSSF-P* perhaps transporting pyrimidine and thiazole.

SUPPORTING TABLES

table S1a. Summary of predicted genes and their characteristics in *Micromonas* RCC299 and CCMP1545 (JGI autostatistics for gene catalog as of 14 February 2008).

	RCC299	CCMP1545
Total genes	10,056	10,575
Supported by ESTs	3,820 (38%)	4,896 (46%)
Supported by homology	6,653 (66%)	7,551 (71%)
Single exon genes	6,368 (63%)	5,264 (50%)
Exons per gene	1.6	1.9
Exons per multi-exon gene	2.5	2.8
Gene length (bp)	1,587	1,557
Transcript length (bp)	1,497	1,390
Protein length (aa)	4,72.84	439
Exon length (bp)	958	731
Intron length (bp)	163	187

table S1b. Functional annotation (JGI pipeline, not manual) of predicted genes in the *Micromonas* RCC299 and CCMP1545 catalogs.

Type	Annotated genes		Distinct functions	
	RCC299	CCMP1545	RCC299	CCMP1545
KOG	6554	7086	3025	3028
EC	1908	1806	627	592
GO	4911	4787	1888	1843
InterPro	6582	6745	4820	4762

table S1c. *Micromonas* genome characteristics in a phylogenomic context. 'Red' refers to chromalveolates. Abbreviations: 1545, *Micromonas* CCMP1545; RCC299, *Micromonas* RCC299; *Oluc*, *Ostreococcus lucimarinus*; *Crei*, *Chlamydomonas reinhardtii*; *Atha, Arabidopsis thaliana*; *Tpse*, *Thalassiosira pseudonana*; *Mbre*, *Monosiga brevicolis*; *Dmel*, *Drosophila melangaster*; *Psti*, *Pichia stipitis*; *Scer, Saccharomyces cerevisiae*.

	Micromonas		Other Green lineage			Red	Choan/Meta		Fungi	
	1545	299	*Oluc*	*Crei*	*Atha*	*Tpse*	*Mbre*	*Dmel*	*Psti*	*Scer*
Genome size (Mb)	21.9	20.9	13.2	121	140	32	42	180	15.4	12.1
G+C (%)	65	64	60	63	36	47	55	42	40	38
Number of genes	10,575	10,056	7,651	15,142	26,341	11,776	9,196	14,601	5,841	5,807
Gene size (bp)	1,557	1,587	1,284	4,312	2,232	1,745	3,004	5,247	1,627	1,455
Multiexon genes (%)	50	37	20	92	79	61	89	82	28	5
Introns (gene^{-1})	0.90	0.57	0.27	7.33	4.2	1.54	6.59	4.9	0.44	0.04
Intron length (bp)	187	163	187	373	164	125	174	1,192	135	256
kb gene^{-1}	2.1	2.2	1.7	8	5.26	2.72	4.57	13.2	2.64	2.08

table S2. Genes encoded in the mitochondrial genomes of RCC299 and of CCMP1545. Note that the CCMP1545 mt genome is not fully assembled and therefore may incorrectly be portrayed as "missing" some genes. Abbreviations: 299, RCC299; 1545, CCMP1545; *located in duplicated block. Note that duplicated genes were counted only once.

Gene Products	**Gene Count**		**Gene Symbol**	
	299	1545	RCC299	CCMP1545
Small subunit ribosomal proteins	11	10	*rps 2,3,4,7,8,10,11,12,13,14,19*	*rps 2,3,4,7,8,10,12,13,14,19*
Large subunit ribosomal proteins	4	4	*rpl 5,6,14,16*	*rpl 5,6,14,16*
NADH dehydrogenase	10	10	*nad 1*,2,3,4,4L,5,6*,7,9,10*	*nad 1,2,3,4,4L,5,6,7,9,10*
ATP synthase	4	4	*atp 4*, 6*,8*,9*	*atp 4*, 6,8,9*
Cytochrome c oxidase	3	3	*cox 1*,2,3*	*cox 1*,2,3*
Ubiquinol: cytochrome c oxidoreductase	1	1	*cob*	*cob**
Sec-independent translocation pathway	1	1	*mtt2*	*mtt2*
Ribosomal RNAs	3	3	*rns*, rnl**	*rns, rnl*
Transfer RNAs (*trn*)	26	24	*Y*(gua), *P*(ugg), *R*(ucu), *G*(ucc), *L*(uaa), *S*(uga), *Me*(cau)*, *K*(uuu), L(uag), L(gag), T(ggu), *S*(gcu), *D*(guc), *V*(uac), *C*(gca), Q(uug), A(ugc), *Mf*(gau)*, *E*(uuc)*, *W*(cca)*, *R*(acg)*, *I*(gau)*, *N*(guu)*, *F*(gaa), *H*(gug), *G*(gcc)*	*Y*(gua), *P*(ugg), *G*(ucc), *L*(uaa), *S*(uga), *Me*(cau), *K*(uuu), *L*(uag), *T*(ggu), *H*(guc) *S*(gcu), *D*(guc), *V*(uac), *C*(gca), *Q*(uug), *Mf*(gau), *E*(uuc), *W*(cca), *R*(acg), *I*(gau), *N*(guu), *F*(gaa), *G*(gcc), *A*(ugc)
unknown ORF	1	NF	*orf169*	

table S3. Genes encoded in the chloroplast genomes of RCC299 and CCMP1545 as well as *O. tauri* (note the CCMP1545 chloroplast genome is incomplete, i.e. not fully sequenced or assembled). Abbreviations: *Otau*, *O. tauri*; *genes in duplicated block.

Gene symbol	Gene product	RCC299	CCMP1545	*Otau*
rpl 2, 5, 14, 16, 20, 23, 32, 36 *rps 2, 3, 4, 7, 8, 9, 11, 12, 14, 18, 19*	Ribosomal proteins	19	15	19
rpo A, B, C1, C2 *tufA, infA*	Transcription Translation	8	7	8
psa A, B, I, J, M, ycf3 *psb A*, B, C, D, E, F, I, J,K, L, N , T, Z* *pet A, B, G* *atp A, B, E, F, H, I* *rbcL*	Photosynthesis gene	31	21	31
*rrl *, rrs*, rrf**	rRNA	3	3	3
trnR (acg), *trnI* (gau), *trnA* (ugc), *trnC* (gca), *trnN* (guu), *trnK* (uuu), *trnM* (cau), *trnG* (ucc), *trnV*, *trnH*, *trnT* (ugu), *trnR* (ucu), *trnL* (uag), *trnE* (uuc), *trnM* (cau), *trnL* (gag), *trnY* (gua), *trnL* (uaa), *trnQ* (uug), *trnD* (guc), *trnP* (ugg), *trnW* (cca), *trnS* (gcu), *trnS* (uga), *trnF* (gaa), *trnM* (cau)	tRNA	26	15	27

table S4. Percent missing data from 10,129 combined chloroplast gene amino acids (aa). 10,129 was the maximum number of aa based on inclusion of large-genome plastids from the red algae (e.g., *Porphyra*) and the basal greens (e.g., *Chaetosphaeridium*). The red, green, glaucophyte, and chromalveolate algae appear in red, green, magenta, and brown text, respectively. Percent missing reflects percentage of the 10,129 that were 'absent' at the aa level, largely as a result of gene loss to the nucleus. In the case of CCMP1545 the high percentage missing likely reflects the fact that the cp genome sequence was incomplete).

Taxon	% aa missing
Cyanidium	0.2
Cyanidioschyzon	2.1
Galdieria	1.3
Gracilaria	1.2
Porphyra purpurea	0.0
Porphyra yezoensis	4.4
Adiantum	0.7
Anthoceros	0.3
Bigelowiella	23.6
Chara	6.7
Chaetosphaeridium	0.0
Chlamydomonas	25.2
Chlorella	15.5
Chlorokybus	14.4
Ginkgo	12.1
Huperzia	5.8
Leptosira	6.6
Marchantia	1.1
Mesostigma	2.6
Oltmannsiellopsis	7.9
Ostreococcus	10.4
Physcomitrella	3.1
Pseudendoclonium	2.7
Psilotum	0.8
Scenedesmu	18.0
Staurastrum	5.4
Stigeoclonium	9.1
Zygnema	6.6
Nephroselm	2.4
***Micromonas* CCMP1545**	**28.1**
***Micromonas* RCC299**	**14.4**
Cyanophora	5.4
Emiliania	4.9
Thalasiossira	3.5
Odontella	3.4
Guillardia	2.7
Rhodomonas	2.7

table S5. Carbon metabolism pathway genes identified in RCC299 and CCMP1545. TBLASTN was used in cases where BLASTP did not reveal a homolog. Abbreviations: -, not found, including by TBLASTN.

		RCC299		CCMP1545	
Enzyme	**Name**	**Prot. ID**	**Localization**	**Prot. ID**	**Localization**
bicarbonate transporter	*AE1*	96291	mitochon.	47080	mitochon.
bicarbonate transporter	*AE2*	104739	unknown	49415	unknown
bicarbonate transporter	*AE3*	86990	unknown	25768	unknown
α carbonic anhydrase	*CA1*	96952	plastid	9513	unknown
δ carbonic anhydrase	*CA2a*	99052	uknown	-	-
β carbonic anhydrase	*CA2b*	-	-	48071	unknown
NADP malic enzyme	*ME1*	62430	chloroplast	1435	chloroplast
NADP malic enzyme	*ME2*	92999	not known	43783	chloroplast
NADP malic enzyme	*ME3*	97726	mitochon.	45850	mitochon.
NADP malic enzyme	*ME4*	81102	cytosol	21535	cytosol
malate dehydrogenase	*MDH1*	104869	chloroplast	57848	chloroplast
malate dehydrogenase	*MDH2*	75917	chloroplast	37587	unknown
malate dehydrogenase	*MDH3*	94811	mitochon.	45503	unknown
malate dehydrogenase	*MDH4*	99233	cytosol	45425	unknown
malate dehydrogenase	*MDH5*	97181	unknown	11694	unknown
malate dehydrogenase	*MDH6*	94075	unknown	49609	cytosol
malate dehydrogenase	*MDH7*	-	-	63993	unknown
malate dehydrogenase	*MDH8*	-	-	32142	unknown
putative malate transporter	*CITT1*	104913	unknown	16532	unknown
putative malate transporter	*CITT2*	99724	unknown	46041	chloroplast
putative malate transporter	*CITT3*	104955	chloroplast	45892	unknown
pyruvate orthophosphate dikinase	*PPDK1*	96907	chloroplast	27547	chloroplast
pyruvate orthophosphate dikinase	*PPDK2*	-	-	31359	unknown
phosphoenolpyruvate carboxykinase	*PEPCK*	-	-	-	-
phosphoenolpyruvate carboxylase	*PEPC1*	104763	cytosol	31189	cytosol

table S6. Chromatin and RNAi associated proteins. Summary of predicted plant proteins from www.chromdb.org; Abbreviations: ✓, presence of predicted proteins; 0, similar sequence were not found in the current genome assemblies (also highlighted in yellow); ?, has sequence similarity but preliminary phylogenetic analyses do not support an inference; Angio, Angiosperms; 299, RCC299; 1545, CCMP1545; *Oluc*, *O. lucimarinus*; *Otau*, *O. tauri*.

Chromatin-associated Proteins	Angio	299	1545	*Oluc*	*Otau*
Histones and Histone Linker Proteins					
Linker Histone Domain Proteins					
Histone H1 linker protein	✓	✓	✓	0	0
Single myb histone protein group	✓	✓	✓	✓	✓
High Mobility Group Family A	✓	0	0	0	0
Core and Histone Variants					
Histone H2A	✓	✓	✓	✓	✓
Histone H2B	✓	✓	✓	✓	✓
Histone H3	✓	✓	✓	✓	✓
Histone H4	✓	✓	✓	✓	✓
Nucleosome Organization: Assembly and Displacement					
Nucleosome Assembly (Chaperones)					
Nucleosome/chromatin assembly complex proteins (NAP1 homologs)	✓	✓	✓	✓	✓
NAP1 Class	✓	✓	✓	✓	✓
SET translocation (myeloid leukemia-associated)	✓	0	0	0	0
ACF1 homologs	✓	✓	✓	✓	✓
CAC1 homologs	✓	✓	✓	✓	✓
CAC2 homologs	✓	✓	✓	✓	✓
HIRA protein group	✓	✓	✓	✓	✓
ASF1 homologs	✓	✓	✓	✓	✓
POB3 and SSRP homologs; FACT complex proteins	✓	✓	✓	✓	✓
Histone chaperone for HTZ1p/H2A-H2B dimer	✓	0	0	0	0
RNA Polymerase Transcription Elongation Factors					
PAF1 Complex Components					
PAF1 complex protein (PAF1 homologs)	✓	✓	✓	✓	✓
PAF1 complex protein (LEO1 homologs)	✓	✓	✓	✓	✓
PAF1 complex protein (CTR9 homologs)	✓	✓	✓	✓	✓
PAF1 complex protein (RTF1 homologs)	✓	✓	✓	✓	✓
PAF1 complex protein (CDC73 homologs)	✓	✓	✓	✓	✓
Other RNA polymerase II elongation factors					
TATA binding protein associated factor 5 protein	✓	✓	✓	✓	✓
Spt4 homologs	✓	✓	✓	✓	✓
Spt5 homologs	✓	✓	✓	✓	✓
Spt6 homologs	✓	✓	✓	✓	✓
Spt16 homologs	✓	✓	✓	✓	✓
Spt2 homologs	✓	0	0	0	0
Elf1 homologs	✓	0	0	0	0

Chromatin Remodeling Complexes					
ATP-Dependent Nucleosome Remodeling					
SNF2 Superclass					
TAFII17nd/MOT class	✓	✓	✓	✓	✓
SWI2/SNF2 class	✓	✓	✓	✓	✓
FUN3nd/SWRI/INO80 classes	✓	✓	✓	✓	✓
FUN30/ETL	✓	✓	✓	✓	✓
SWR1	✓	✓	✓	✓	✓
INO80	✓	0	0	0	0
Chromodomain class	✓	✓	✓	✓	✓
HOMSA_CHDL and ARATH_CHR10 subclass	✓	0	0	0	0
CHD1 subclass	✓	0	0	0	0
Mi2-Kismet subclass	✓	✓	✓	0	1
DDM/LSH class	✓	✓	✓	✓	✓
ISWI class	✓	✓	✓	✓	✓
MOM group	✓	0	0	0	0
RIS1 Superclass					
RIS1 class	✓	?	?	0	0
CHR36/39 class	✓	✓	✓	✓	✓
RAD5 class	✓	0	0	0	0
SMARCA3 class	✓	✓	0	0	0
RAD16 class	✓	✓	✓	✓	✓
RAD26 Superclass					
RAD26 class	✓	✓	✓	✓	✓
HARP class	✓	✓	✓	✓	✓
RAD54 class	✓	✓	✓	✓	✓
ATRX class	✓	0	0	0	0
DRD1 class	✓	0	0	0	0
SWI3 and RSC8 homologs	✓	✓	✓	✓	✓
SWP73 and RSC6 homologs	✓	✓	✓	✓	✓
SNF5 homologs	✓	0	0	0	0
Actin-related proteins					
ARP1 Class	0	0	0	0	0
ARP2 Class	✓	0	0	✓	✓
ARP3 Class	✓	0	0	✓	✓
ARP4 Class	✓	✓	✓	✓	✓
ARP7 Class	✓	✓	✓	✓	✓
ARP5 Class	✓	0	0	0	0
ARP6 Class	✓	✓	✓	✓	0
Plant ARP8 Class	✓	0	0	0	0
ARP8_9/ARATH_ARP9 Superclass	✓	0	0	0	0
DNA-dependent ATPase and helicase RuvB (RUVB-like proteins)	✓	✓	✓	✓	✓
SWR complex proteins (SWC4 homlogs)	✓	✓	✓	✓	✓

Other Chromatin Remodeling Complexes and Associated Proteins					
NURF complex component (RBBP4/CAF1 homologs)	✓	✓	✓	✓	✓
Proteasomal ATPases; Group A	✓	✓	✓	✓	✓
Proteasomal ATPases; Group B	✓	✓	✓	✓	✓
Esc-E(z)-E(Pc) complex of polycomb protein	✓	✓	✓	✓	✓
Enhancer of Polycomb-like protein group [E(Pc) homologs]	✓	✓	✓	✓	✓
VRN2, EMF2, FIS2 homologs	✓	0	0	?	?
Polycomb group (esc homologs)	✓	✓	✓	✓	✓
Histone Modifications					
Reversible methylation					
Methylation					
KMT_H3K4	✓	✓	✓	✓	✓
KMT_H3K9 [Su(var)3-9]	✓	?	?	0	0
KMT_H3K27 (Enhancer of zeste Group)	✓	✓	✓	✓	✓
KMT_H3K36 (SET2/SETD2/ASH) Group)	✓	✓	✓	✓	✓
ARATH_ATXR5 Group	✓	0	0	0	0
S-ET interrupted and unclassified	✓	✓	✓	✓	✓
Protein arginine methyltransferases					
PRMT1	✓	0	0	0	0
CARM1	✓	✓	✓	0	0
PRMT7	✓	✓	✓	✓	✓
PRMT5	✓	✓	✓	✓	✓
PRMT3	✓	✓	✓	✓	✓
PRMT6	✓	✓	✓	✓	✓
Demethylation					
HDMA : Histone demethylases (AOF2/ LSD1 homologs)	✓	✓	✓	✓	✓
ADP Ribosylation	✓	0	0	0	0
Ubiquination	✓	✓	✓	✓	✓
Reversible acetylation					
Acetylation					
Histone acetyltransferases (GNAT superfamily)	✓	✓	✓	✓	✓
Histone acetyltransferases (MYST family)	✓	✓	✓	✓	✓
Histone acetyltransferases [CREBBP (CBP) family]	✓	✓	✓	✓	✓
Histone acetyltransferases (TAFI homologs)	✓	✓	✓	✓	✓
Deacetylation					
Histone deacetylases (RPD3/HDA1 superfamily)	✓	✓	✓	✓	✓
SRT : Histone deacetylases (SIR2 family)	✓	✓	✓	✓	✓
HDT : Histone deacetylases (plant-specific HD2 family)	✓	0	0	0	0
Histone Modification-Associated Proteins and Complexes					
COMPASS (SET1C) complex protein (SWD1 homologs)	✓	✓	✓	✓	✓

SET1 complex component (ASH2/BRE2 homologs)	✓	0	0	0	0
YEATS domain-containing family group	✓	✓	✓	✓	✓
SIN3 complex components (SAP18 homologs)	✓	0	0	0	0
Histone deacetylase complex protein (SIN3 homologs)	✓	✓	✓	✓	✓
Histone acetyltransferase complex component (ADA2 homologs)	✓	✓	✓	✓	✓
COMPASS (SET1C) complex protein (SWD3 homologs)	✓	✓	?	✓	✓
COMPASS (SET1C) complex protein (SWD2 homologs)	✓	✓	✓	✓	✓
Modified-Histone Binding Proteins					
Bromodomain Proteins	✓	✓	✓	✓	✓
Bromodomain-containing proteins containing AAA ATPase domains	✓	✓	✓	✓	✓
Bromodomain-containing proteins containing WD40 domain repeats	✓	✓	✓	✓	✓
Diverse group of chromodomain-containing proteins	✓	✓	✓	✓	✓
Inhibitor of Growth protein group (ING1-5 homologs)	✓	✓	✓	✓	✓
MRG domain-containing proteins	✓	0	0	0	0
DNA Modifying Proteins					
DNA methyltransferases					
DNMT Class (CpG/CpnpG)	✓	0	0	0	0
DNMT2 class (CpNpN;RNA-guided)	✓	✓	✓	✓	✓
DRM class	✓	0	0	0	0
SNF2_N/helicase domain proteins	0	✓	✓	✓	✓
ROS/Demeter DNA glycosylases	✓	✓	✓	✓	✓
Variant in Methylation (VIM)	✓	0	0	0	0
Non-histone DNA Binding Proteins					
HMGB : High Mobility Group Family B	✓	✓	✓	✓	✓
BAH-PHD domain-containing protein group	✓	✓	✓	✓	✓
Methyl binding domain proteins	✓	✓	✓	0	0
ARID/BRIGHT DNA binding domain group	✓	✓	✓	✓	✓
DEK_C domain proteins	✓	✓	✓	✓	✓
RNAi Components					
Argonaute gene family	✓	✓	0	0	0
Double stranded RNA-Binding protein group	✓	0	0	0	0
Suppressor of gene silencing	✓	0	0	0	0
RNA Polymerase IV Small Subunit	✓	0	0	0	0
RNA Polymerase IV Large Subunit	✓	0	0	0	0
RNA-dependent RNA polymerases	✓	0	0	0	0
RNA helicases	✓	?	0	?	?
HUA Enhancer	✓	0	0	0	0
Dicer-Like group	✓	0	0	0	0
Enhanced RNA interference	✓	0	0	0	0
Chromosome Dynamics					

Condensin Complex Components	✓	✓	✓	✓	✓
SMC protein group	✓	✓	✓	✓	✓
Non-SMC subunits	✓	✓	✓	✓	✓
Barren domain-containing	✓	✓	✓	✓	✓

table S7. Comparative analysis of core cell cycle genes in a selection of green lineage organisms. enes were identified using *Ostreococcus* orthologs as bait for BLAST searches on both *Micromonas* JGI portals. High scoring matches were found in all cases using BLASTP so that use of TBLASTN was not necessary. Abbreviations: CN, gene copy number; *Otau*, *O. tauri*; *Oluc*, *O. lucimarinus*; *Crei*, *C. reinhardtii*; *Atha*, *A. thaliana*; nd, not determined. Deflines for genes are as follows: *CDK*, Cyclin Dependant-Kinase; *CYC*, Cyclin; *CKS*, *CDK* Subunits; *RB*, Retinoblastoma protein; *E2F*, Transcription Factors; *DP*, Dimerization Protein; *DEL*, DP- and E2F-like protein; *CDC*, Cell Division Cycle; *APC*, Anaphase Promoting Complex; *CCS*: Cell Cycle Switch.

Gene	**RCC299**		**CCMP1545**		***Otau*[a] CN**	***Oluc*[b] CN**	***Crei*[d] CN**	***Atha*[a] CN**
	CN	Protein ID[c]	CN	Protein ID[c]				
CDKA	1	55646	1	69828	1	1	1	1
CDKB	1	105013	1	70812	1	1	1	4
CDKC	1	77851	1	14515	1	1	1	2
CDKD	1	96962	1	3997	1	1	1	3
CYCA	2	107361; 113653	2	31383; 54394	2	2	1	10
CYCB	1	95175	1	36034	1	1	1	9
CYCD	2	61685; 104783	2	3819; 49242	2	2	3	10
CYCT	1	107365	1	64909	1	1	1	?
CYCH	1	64801	1	11155	1	1	?	1
CKS	1	78882	1	57520	1	1	1	2
RB	1	64476	1	48829	1	1	1	1
E2F	1	55361	1	32031	1	1	1	3
DP	1	98766	1	4731	1	1	1	2
DEL	1	84006	1	48504	1	1	1	3
WEE1	1	74444	1	54624	1	1	1	1
CDC25	1	61065	1	59751	1	1	0	0
APC1	1	85693	1	33905	1	1	nd	1
APC2	1	96683	1	35295	1	1	nd	1
APC3	1	56651	1	18029	1	1	nd	2
APC4	1	58493	1	56629	1	1	nd	1
APC5	1	98655	1	51550	1	1	nd	1
APC6/CDC16	1	51833	1	55310	1	1	nd	1
APC7	1	83212	1	57873	1	1	nd	1
APC8/CDC23	1	78382	1	15779	1	1	nd	1
APC10	1	59781	1	38111	1	1	nd	1
APC11	1	86792	1	9662	1	1	nd	1
CDC26	0	/	0	/	1	0	nd	1
CDC20, CDH1, AMA1	1	97422	1	40257	1	1	nd	5
CDH1-CCS52	1	83216	1	45511	1	2 (chr 21)	nd	3

[a]From Robbens S., Khadaroo B., Derelle E., Ferraz C., Inze D., Van de Peer Y., Moreau H. (2005) Genome wide-analysis of core cell cycle genes in the unicellular green alga *Ostreococcus tauri*. Mol. Biol. Evol. 22, 589-597; [b]From JGI web site: http://genome.jgi-psf.org/Ost9901_3/Ost9901_3.home.html; [c]From JGI web site: http://genome.jgi-psf.org/Chlre3/Chlre3.home.html; [d]From Bisova K., Krylov D.M., Umen J.G. (2005) Genome-wide annotation and expression profiling of cell cycle regulatory genes in *Chlamydomonas reinhardtii*. Plant. Physiol. 137, 475-491.

table S8. A comparison of annotated *O. tauri* photosynthesis related genes to the *Micromonas* RCC299 and CCMP1545 genomes. BLASTP using previously published *O. tauri* annotations was used to find genes and TBLASTN in cases where BLASTP was unsuccessful. Note that this table was prepared primarily using annotated *O. tauri* genes as the query. Abbreviations: NF, not found; C, chloroplast encoded; N, nucleus encoded; cp, chloroplast. X demarks location where a particular gene is encoded.

Gene name	Defline	RCC299			CCMP1545			*Otau*
		C	N	Prot. ID	C	N	Prot. ID	
Photosystem I								
psaA	P700 apoprotein subunit Ia	x			x			Y
psaB	P700 apoprotein subunit Ib	x			x			Y
psaC	Photosystem I subunit VII	x			x			Y
PSAD	Photosystem I subunit II, cp precursor		x	84657		x	36482	Y
psaE	Photosystem I subunit IV, cp precursor		x	90731		x	32586	Y
psaF	Photosystem I subunit III, cp precursor		x	90596		x	26202	Y
psaG	Photosystem I subunit V		x	93147		x	52059	Y
psaH	Photosystem I subunit VI, cp precursor		x	108180		x	70928	Y
psaI	Photosystem I subunit VIII	x						Y
psaJ	Photosystem I subunit IX	x						Y
psaK	Photosystem I subunit X, cp precursor		x	104961		x	53824	Y
psaL	Photosystem I subunit XI, cp precursor		x	95054		x	47719	Y
psaM	Photosystem I subunit XII	x			x			Y
psaN	Photosystem I subunit N, cp precursor		x	93625		x	58597	Y
ycf4	Photosystem I assembly protein ycf4		x	104909		x	28946	Y
ycf3	Photosystem I assembly protein ycf3	x			x			Y
Photosystem II								
psbA1	Photosystem II D1 protein	x			x			Y
psbA2	Photosystem II D1 protein	x			x			Y
psbB	Photosystem II CP47 protein	x			x			Y
psbC	Photosystem II CP43 protein	x			x			Y
psbD	Photosystem II D2 protein	x			x			Y
psbE	Cytochrome b_{559} α subunit	x			x			Y
psbF	Cytochrome b_{559} β subunit	x			x			Y
psbH	Photosystem II psbH protein	x			x			Y
psbI	Photosystem II psbI protein	x			x			Y
psbJ	Photosystem II psbJ protein	x			x			Y
psbK	Photosystem II psbK protein	x			x			Y
psbL	Photosystem II psbL protein	x			x			Y
PSBM	Photosystem II PSBM protein		x	96751		x	42681	Y
psbN	Photosystem II psbN protein	x			x			Y
PSBO	Photosystem II manganese-stabilizing polypeptide		x	92317		x	48994	Y
PSBP	Photosystem II oxygen-evolving complex 23 kDa protein		x	104990		x	34167	Y
PSBQ	Oxygen-evolving enhancer protein 3 (OEE3), cp precursor		x	106554		x	49738	Y
PSBR	Photosystem II PSBR protein, cp precursor		x	61954		x	56305	Y
PSBS	Photosystem II PSBS protein		x	108747		x	51750	Y
psbT	Photosystem II psbT protein	x			x			Y
PSBW 1	Photosystem II 13kD protein, cp precursor		x	82984		x	19538	Y
PSBW 2	Photosystem II 13kD protein, cp precursor		x	106364		x	51380	Y
PSBX	Photosystem II PSBX protein, cp precursor		x	NF		x	51772	Y
PSBY	Photosystem II PSBY protein		x	112658		x	63313	Y

psbZ	Photosystem II 11 kD protein	x			x		Y
ALB3.1	ALBINO3-like protein, cp precursor		x	96326		52221	Y
ALB3.2	ALBINO3-like protein, cp precursor		x	51662		31798	Y
Cytochrome *b_6/f* complex							
petA	Cytochrome *f*	x			x		Y
petB	Cytochrome *b_6*	x			x		Y
PETC	Cytochrome *b_6/f* complex iron-sulfur subunit cp precursor		x	88922		x	36185	Y
PETD	Cytochrome *b_6-f* complex subunit IV		x	106769		x	44239	Y
petG	Cytochrome *b_6-f* complex subunit V	x			x		Y
petL	Cytochrome *b_6-f* complex subunit VI			NF		NF	NF
PETM	Putative Cytochrome *b_6/f* complex subunit PETM		x	109111		x	60361	NF
PETN	Cytochrome *b_6-f* complex subunit VIII		x	94459		x	63318	Y
Soluble electron carriers and putative carriers							
PETE	Plastocyanin, chloroplast precursor		x	60322		x	59039	Y
FDX4	*Chlamydomonas FDX4-like*		x	107389		x	27711	Y
FDX6	*Chlamydomonas FDX6-like*		x	108124		x	58790	Y
PETF	Ferredoxin, chloroplast precursor		x	96644		x	30100	Y
FDX3	*Chlamydomonas FDX3-like*		x	104777		x	37020	Y
PETJ	Cytochrome *c_{553}*, chloroplast precursor		x	104960		x	53843	Y
PETH	Ferredoxin-NADP oxidoreductase		x	106827		x	54402	Y
ATP synthase							
atpA	ATP synthase CF1 α chain	x			x		Y
atpB	ATP synthase CF1 β chain	x			x		Y
ATPC	ATP synthase CF1 γ chain, cp precursor		x	104789		x	48500	Y
ATPD	ATP synthase CF1 δ chain, cp precursor		x	93509		x	57603	Y
atpE	ATP synthase CF1 ε chain	x			x		Y
atpF	ATP synthase CF0 subunit B	x			x		Y
ATPG	ATP synthase CF0 subunit B', cp precursor		x	104717		x	49458	Y
atpH	ATP synthase CF0 subunit C	x			x		Y
atpI	ATP synthase CF0 subunit A	x			x		Y

table S9. Chlorophyll and carotenoid biosynthesis related genes. BLASTP using previously published *O. tauri* annotations was used to find genes and TBLASTN in cases where BLASTP was unsuccessful. Note that common genes within this pathway that are not listed under gene name below are not necessarily missing, but have not been annotated. Abbreviations: *Otau*, *O. tauri*; NF, not found; Y, yes present.

Gene Name	Defline	RCC299	CCMP1545	*Otau*
		Prot. ID	Prot. ID	
Chlorophyll pathway				
HEMA	Glutamyl-tRNA reductase	62986	48234	Y
HEML	Glutamate-1-semialdehyde 2,1-aminomutase, cp precursor	95160	29561	Y
HEMB	δ-aminolevulinic acid dehydratase, cp precursor	93011	45299	Y
HEMC	Porphobilinogen deaminase, cp precursor	78011	31842	Y
HEMD	Urophorphyrin III synthase	60121	50233	Y
HEME1	Uroporphyrinogen decarboxylase, cp precursor	96206	49908	Y
HEME2	Uroporphyrinogen decarboxylase, cp precursor	104963	23171	Y
HEME3	Uroporphyrinogen decarboxylase, cp precursor	77925	37342	Y
HEME4	Uroporphyrinogen decarboxylase, cp precursor	54920	64427	Y
HEMF1	Coproporphyrinogen III oxidase, cp precursor	104790	52031	Y
HEMF2	Coproporphyrinogen III oxidase, cp precursor (putative)	80311	23563	Y
HEMG	Protoporphyrinogen IX oxidase	60613	35610	Y
HEMH	Ferrochelatase II, chloroplast precursor	96687	51216	Y
CHLH1	Magnesium-chelatase subunit CHLH, cpprecursor	104801	26233	Y
CHLH2	Magnesium-chelatase subunit CHLH like protein, cp precursor	80464	16188	Y
CHLI1	Magnesium-chelatase subunit CHLI, cp precursor	107341	49430	Y
CHLI2	Magnesium-chelatase subunit CHLI like protein, cp precursor	105016	22176	Y
CHLM	Magnesium-protoporphyrin IX methyltransferase, cp precursor	96236	57377	Y
CHLB	Light-independent protochlorophyllide reductase subunit B	NF	NF	NF
CHLL	Light-independent protochlorophyllide reductase iron-sulfur ATP-binding protein	NF	NF	NF
CHLN	Light-dependent protochlorophyllide oxido-reductase, cp precursor	NF	NF	NF
DVR	3,8-divinyl protochlorophyllide a 8-vinyl reductase	62770	40801	Y
PORA	Light-dependent protochlorophyllide oxido-reductase , cp precursor	93411	33005	Y
PORB/C	Light-dependent protochlorophyllide oxido-reductase, cp precursor	96428	17564	Y
CHLG	Chlorophyll synthetase	59614	33362	NF
CAO	Chlorophyll *a* oxygenase (chlorophyll *b* synthase)	104843	60555	Y
UPM1	Uroporphyrin III methylase	87830	22906	Y
HY1	Heme oxygenase	108005	46050	Y
Carotenoid pathway				
IPI	Isopentenyl-diphosphate δ-isomerase I (IPP isomerase I)	54661	22570	Y
GGPS	Geranyl pyrophosphate synthase	104883	50196	Y
PSY1	phytoene synthase/geranylgeranyl-diphosphate geranylgeranyl transferase	85839	32317	Y
PSY2	Putative phytoene synthase / geranylgeranyl-diphosphate geranylgeranyl transferase	NF	22804	Y

PDS	phytoene desaturase/phytoene dehydrogenase	104873	49039	Y
ZDS	ζ-carotene desaturase	NF	NF	NF
CRTR	β-carotene hydroxylase	59945	38006	Y
ZEP1	zeaxanthin epoxidase	82567	19991	Y
ZEP2	zeaxanthin epoxidase	NF	24859	Y
VDE	violaxanthin deepoxidase	104842	60541	Y
NXS	Neoxanthin synthase	NF	NF	NF
CCD	epoxycarotenoid cleavage enzyme	62157	23105	Y

table S10a. Selenoprotein distribution in *Micromonas* RCC299 and CCMP1545. Searches were made by BLASTP and then by TBLASTN as needed using published sequences for *Ostreococcus* selenoproteins, as well as homologs from *C. reinhardtii, M. musculus, Drosophila* and *H. sapiens.* Abbreviations: Y, yes SECIS element found; N, no SECIS element found; NF, gene not found.

Gene Name	Defline	RCC299	Seleno in RCC299?	CCMP1545	Seleno in CCMP1545?
PDI	Protein disulfide isomerase. Putative ER precursor	112673	Y	70202	Y
DSBA	DSBA oxidoreductase	112684	N	70882	Y
DSBA2	DSBA oxidoreductase	113700	Y	70883	Y
PRDX	Peroxiredoxin, bacteria like	112683	Y	70211	Y
MSRA	Peptide-methionine-(S)-S-oxide reductase	64541	N	34655	N
MSRA	Peptide-methionine-(S)-S-oxide reductase	62480	N	53038	N
	Putative Methyltransferase	112687	Y	70215	Y
SELH	Selenoprotein H	113704	Y	70887	Y
SELM	Selenoprotein M	112677	Y	70203	Y
SELU	Selenoprotein U	NF	-	NF	-
SEP15	Selenoprotein Sep15	NF	-	NF	-
	Trx-Fold Selenoprotein	NF	-	NF	-
SELW	Selenoprotein W	112681	Y	70205	Y
SELT	Selenoprotein T	112678	Y	70888	Y
	Thioredoxin-disulfide reductase	113706	Y	70201	N
SELS	Selenoprotein S	NF	-	NF	-
SELO	Selenoprotein O	113708	Y	70889	Y
SELK	Selenoprotein K	NF	-	NF	-
GPX1	Glutathione peroxidase	112672	Y	70206	Y
GPX2	Glutathione peroxidase	112674	Y	70207	Y
GPX3	Glutathione peroxidase	112671	Y	70208	Y
GPX4	Glutathione peroxidase	112675	Y	70209	Y
GPX5	Glutathione peroxidase	112676	Y	70210	Y
	Un-named Selenoprotein	112679	Y	NF	-
Selenoprotein machinery					
SBP2	SECIS-binding protein 2	108533		70891	
SELB	Selenocysteine-specific elongation factor	55375		13765	
SELD	Selenide, water dikinase	112686		70214	

table S10b. Selenoprotein homologs from *Ostreococcus* and *Chlamydomonas*, from (*44*) as well as independent searches of *Chlamydomonas*. Abbreviations: *Crei*, *C. reinhardtii*; *Otau*, *O. tauri*; *Oluc*, *O. lucimarinus*; Y, gene present; NF, gene not found; *There are 4 of these genes in *C. reinhardtii*, one of which is a selenoprotein; however those with the highest similarity to the *Micromonas* protein encoding genes do not appear to encode selenoproteins; **There were 3 methyltransferases; ***Appears to be *SELO* but no SECIS element could be detected; ****Two copies.

Gene Name	Defline	*Crei*	Seleno in *Crei*?	*Otau*	Seleno in *Otau?*	*Oluc*	Seleno in *Oluc?*
PDI	Protein disulfide isomerase. Putative ER precursor	NF	-	Y	Y	Y	Y
DSBA	DSBA oxidoreductase	NF	-	Y	Y	Y	Y
DSBA2	DSBA oxidoreductase	Y	Y	Y	Y	Y	Y
PRDX	Peroxiredoxin, bacteria like	Y	Y	Y	Y	Y	Y
MSRA	Peptide-methionine-(S)-S-oxide reductase	Y*	N*	Y	Y	Y	Y
MSRA	Peptide-methionine-(S)-S-oxide reductase	Y*	N*	NF	-	NF	-
	Putative Methyltransferase	Y	Y**	Y	Y	Y	Y
SELH	Selenoprotein H	Y	Y	Y	Y	Y	Y
SELM	Selenoprotein M	Y	Y	Y	Y	Y	Y
SELU	Selenoprotein U	Y	Y	Y	Y	Y	Y
SEP15	Selenoprotein Sep15	NF	-	Y	Y	Y	Y
	Trx-Fold Selenoprotein	NF	-	Y	Y	Y	Y
SELW	Selenoprotein W	Y	Y	Y	Y	Y	Y****
SELT	Selenoprotein T	Y	Y	Y	Y	Y	Y
	Thioredoxin-disulfide reductase	Y	Y	Y	Y	Y	Y
SELS	Selenoprotein S	NF	-	Y	Y	Y	Y
SELO	Selenoprotein O	Y	N***	Y	Y	Y	Y
SELK	Selenoprotein K	Y	Y	Y	Y	Y	Y
GPX1	Glutathione peroxidase	Y	N	Y	Y	Y	Y
GPX2	Glutathione peroxidase	Y	N	Y	Y	Y	Y
GPX3	Glutathione peroxidase	Y	Y	Y	Y	Y	Y
GPX4	Glutathione peroxidase	Y	Y	Y	Y	Y	Y
GPX5	Glutathione peroxidase	NF		Y	Y	Y	Y
	Un-named Selenoprotein	NF	-	Y	Y	Y	Y
Selenoprotein machinery							
SBP2	SECIS-binding protein 2	Y		Y		Y	
SELB	Selenocysteine-specific elongation factor	Y		Y		Y	
SELD	Selenide, water dikinase	Y		Y		Y	

table S11. Transcription factor families and their representation in a selection of green lineage organisms. Coloring indicates the following: green, Green Plant-specific (GPS); yellow, Land Plant-specific (LPS). **bold font**, numbers confirmed using BLASTP. Abbreviations: *, transcriptional-regulators; **, non-searchable IPR# in JGI site; n.a., not available; n.d. not determined; ZF, zinc-finger; *Atha*, *A. thaliana*; *Osat*, *O. sativa*; *Ptri*, *Populus trichocarpa*; *Smoe*: *Selaginella moellendorffii*; *Ppat*, *P. patens*; *Crei*, *C. reinhardtii*; *Otau*, *O. tauri*; *Oluc*, *O. lucimarinus*; 1545, *Micromonas* CCMP1545; RCC299, *Micromonas* RCC299.

Domain	InterPro ID	*Atha*	*Osat*	*Ptri*	*Smoe*	*Ppat*	*Crei*	*Otau*	*Oluc*	1545	299
Alfin	n.a.	7	13	9	3	7	**2**	**1**	**1**	**1**	**1**
AP2/EREBP	IPR001471	146	182	212	112	156	**12**	**9**	**10**	**13**	**12**
ARF	IPR010525	23	41	37	14	13	**0**	**0**	**0**	**0**	**0**
ARID*	IPR001606	10	7	13	14	8	**2**	**1**	**1**	**2**	**3**
AS2	IPR004883	42	39	57	29	29	**0**	**0**	**0**	**0**	**0**
AUX/IAA*	IPR003311	29	46	33	7	2	**0**	**0**	**0**	**0**	**0**
B3	IPR003340	60	57	108	30	37	**1**	**0**	**0**	**1**	**1**
BBR-BPC	IPR010409	7	7	16	2	0	**0**	**0**	**0**	**0**	**0**
BES1	IPR008540	8	6	12	9	6	**0**	**0**	**0**	**0**	**0**
bHLH	IPR001092	127	184	148	102	98	**4**	**2**	**2**	**2**	**2**
BTP	IPR006565	6	8	7	2	1	**1**	**0**	**0**	**1**	**1**
bZIP	IPR004827	72	109	85	51	47	11	9	14	16	13
C2C2 superfamily											
C2C2-CCT	IPR010402	37	54	39	24	23	**6**	**4**	**5**	**3**	**3**
C2C2-DOF	IPR003851	36	36	42	43	21	**1**	**2**	**2**	**1**	**1**
C2C2-GATA	IPR000679	26	23	32	16	11	**12**	**6**	**11**	**8**	**7**
C2C2-YABBY	IPR006780	**6**	**8**	13	**0**	**0**	**0**	**0**	**0**	**1**	**1**
C2H2	IPR007087	134	113	81	131	49	26	20	19	37	36
C3H	IPR000571	59	90	78	73	37	15	19	23	22	31
CAMTA	IPR005559	6	8	7	8	1	**0**	**1**	**1**	**1**	**1**
CCAAT	IPR003958	35	56	32	26	20	6	5	5	5	5
CPP	IPR005172	8	16	13	8	6	**1**	**2**	**2**	**2**	**1**
CSP	IPR002059	4	3	7	6	3	**3**	**4**	**3**	**4**	**4**

E2F/DP	IPR003316	8	9	10	8	11	**3**	**3**	**3**	**4**	**4**
EIL	IPR006957	6	12	6	9	2	**0**	**0**	**0**	**0**	**0**
FHA	IPR000253	16	19	19	18	15	13	11	10	12	12
GeBP	IPR007592	21	15	7	0	0	**0**	**0**	**0**	**0**	**0**
GIF	IPR007726	3	3	5	3	4	**1**	**1**	**1**	**1**	**1**
GRAS	IPR005202	33	58	96	84	39	**0**	**0**	**0**	**0**	**0**
GRF	IPR014977**	9	18	9	5	2	**0**	**0**	**0**	**1**	**1**
HB	IPR001356	**104**	**90**	106	**34**	**42**	**5**	**8**	**8**	**14**	**9**
HMG	IPR000910	11	19	12	16	9	11	7	7	11	10
HRT-like	n.a.	2	1	1	2	7	**0**	**0**	**0**	**0**	**0**
HSF	IPR000232	23	36	31	14	8	**2**	**1**	**1**	**4**	**6**
JUMONJI-C*	IPR003347	17	17	23	30	17	9	13	20	24	23
JUMONJI-N*	IPR003349	14	27	11	12	5	**4**	**1**	**1**	**3**	**3**
LFY	IPR002910	1	1	1	2	2	**0**	**0**	**0**	**0**	**0**
LUG*	n.a.	2	11	6	2	1	**0**	**0**	**0**	**0**	**0**
MADS	IPR002100	104	83	111	33	22	**2**	**1**	**1**	**1**	**1**
MBF1*	IPR013729**	3	3	3	2	3	**1**	**1**	**1**	**1**	**1**
MYB-superfamily											
MYB-R2R3	IPR001005	49	84	84	n.d.	31	14	17	n.d.	27	25
MYB-related	IPR001005	150	138	216	n.d.	64	14	10	n.d.	11	10
MYB-G2-like	n.a.	43	56	67	26	46	**5**	**3**	**3**	**4**	**4**
MYB-SHAQKYF	IPR006447	101	221	49	36	10	**5**	**5**	**5**	**6**	**5**
NAC/NAM	IPR003441	107	149	172	38	32	**0**	**0**	**0**	**0**	**0**
NZZ/SPL	n.a.	1	1	2	3	3	**0**	**0**	**0**	**0**	**0**
PBF-2	IPR009044**	2	2	2	5	0	**1**	**1**	**1**	**3**	**2**
PcG	IPR001214	34	34	45	112	31	21	25	27	46	43
PHD*	IPR001965	56	79	86	115	68	13	17	20	32	24

PLATZ	IPR006734	10	20	20	18	13	**1**	**1**	**1**	**1**	**1**
RR	IPR001789	76	181	106	55	81	10	8	10	13	12
RWP-RK	IPR003035	**14**	**13**	18	23	**6**	**16**	**4**	**4**	**6**	**6**
S1Fa-like	IPR006779	3	4	2	0	1	**0**	**0**	**0**	**0**	**0**
SAP	n.a.	1	0	1	2	0	**0**	**0**	**0**	**0**	**0**
SBP	IPR004333	16	28	29	22	14	**20**	**0**	**0**	**3**	**3**
SRS	IPR007818	10	6	10	6	2	**0**	**0**	**0**	**0**	**0**
TAZ	IPR000197	9	10	7	6	5	**2**	**1**	**1**	**1**	**1**
TCP	IPR005333	23	24	34	9	6	**0**	**0**	**0**	**0**	**0**
Trihelix	n.a.	26	23	47	>8	28	**0**	**0**	**0**	**0**	**0**
TUB/TLP	IPR000007	11	21	11	15	6	**2**	**1**	**1**	**1**	**1**
ULT	n.a.	2	2	3	**0**	**0**	**0**	**1**	**1**	**1**	**1**
VOZ	n.a.	2	2	4	**0**	2	**0**	**0**	**0**	**0**	**0**
WRKY	IPR003657	72	113	104	36	37	**1**	**3**	**3**	**2**	**2**
ZF-HD	IPR006456	16	15	25	14	8	**0**	**0**	**0**	**0**	**0**
ZF-LIM	IPR001781	13	13	21	9	11	**1**	**3**	**3**	**3**	**4**
ZIM	IPR010399	18	29	22	19	16	**0**	**0**	**0**	**0**	**0**
Total # of TFs		65	64	65	60	59	40	39	39	44	44
Green plant-specific		31	30	31	26	26	7	6	6	10	10
Non green-specific		34	34	34	34	33	33	33	33	34	34

References
DATF: http://datf.cbi.pku.edu.cn; Guo et al., 2005
PlnTFDB: http://plntfdb.bio.uni-potsdam.de/v2.0/; Riano-Pachon, et al., 2007

Domain Name/Domain description

Alfin	Alfin homology w/ or w/o a C-term ZF (C4 plus HC3)
AP2/EREBP	ERF domain found in AP1/EREBP/RAV factors
ARF	Auxin response factor witn C-term B3
ARID*	AT-rich interaction domain/BRIGHT DNA binding domain
AS2	Protein of unknown function DUF260, LOB domain
AUX/IAA*	AUX/IAA family
B3	Found in RAV/ARF/ABI3 factors
BBR-BPC	DUF1004, GAGA-binding factor
BES1	A role in BR-regulated gene expression
bHLH	Helix-loop-helix dimerization domain
BTP	Bromodomain transcription factor

bZIP	bZIP transcription factor
C2C2-CCT	Found in C2C2-CO-like family, subfamilies with ZF or RR
C2C2-DOF	ZF-DOF domain w/ or w/o RR
C2C2-GATA	GATA zinc finger
C2C2-YABBY	ZF-YABBY domain plus HMG-YABBY domain
C2H2	Zinc finger, C2H2 type
C3H	Zinc finger C-x8-C-x5-C-x3-H type (and similar)
CAMTA	CG-1 domain
CCAAT	Subfamilies DR1/HAP2/HAP3/HAP5
CPP	Tesmin/TSO1-like CXC domain
CSP	Cold-shock protein, DNA-binding
E2F/DP	E2F/DP family winged-helix DNA-binding domain
EIL	Ethylene insensitive 3-like protein
FHA	FHA domain
GeBP	Protein of unknown function, DUF573
GIF	SSXT protein (N-terminal region)
GRAS	GRAS family transcription factor
GRF	Growth regulating factor
HB	Homeobox domain
HMG	HMG (high mobility group) box
HRT-like	Unusual ZF, CX8-9CX10CX2H
HSF	HSF-type DNA-binding
JUMONJI-C*	C terminal domain, found in histone demethylases
JUMONJI-N*	N terminal domain, accompanied by jumanjiC
LFY	Floricaula / Leafy protein
LUG*	LUFS/Leunig homology domain
MADS	SRF-type transcription factor
MBF1*	Multiprotein bridging factor 1, HTH motif
MYB-R2R3	two or more Myb domains
MYB-related	Single MYB
MYB-G2-like	Subfamily of SHAQKYF, w/ or w/o RR
MYB-SHAQKYF	MYB with SHAQKYF motif, excluding G2-like
NAC/NAM	No apical meristem (NAM)
NZZ/SPL	Nozzle/Sporocyteless
PBF-2	Found in Whirly family, SS DNA binding protein
PcG	SET domain
PHD*	PHD-finger
PLATZ	Protein of unknown function, DUF597
RR	Response regulator receiver domain
RWP-RK	NIN/MID like protein
S1Fa-like	DNA binding protein S1FA
SAP	Related to IPR011044
SBP	SBP domain
SRS	Domain of unknown function (DUF702)
TAZ	TAZ ZF
TCP	TCP family transcription factor
Trihelix	MYB+Helix+MYB
TUB/TLP	Tub family
ULT	SAND domain at N-terminus + ZF at C-terminus
VOZ	VOZ domain for dimerization and DNA binding
WRKY	WRKY DNA -binding domain
ZF-HD	ZF-HD protein, dimerization domain
ZF-LIM	LIM domain
ZIM	ZIM motif

table S12. Homeodomain gene classification and profiles in the green lineage. Abbreviations: *Atha*, *A. thaliana*; *Osat*, *O. sativa* J; *Smoe*, *S. moellendorffii*; *Ppat*, *P. patens*; *Crei*, *C. reinhardtii*; *Otau*, *O. tauri*; *Oluc*, *O. lucimarinus*; 1545, CCMP1545; 299, RCC299; *Cmer*, *C. merolae*; *, based on v1 catalog; ** *Micromonas*-specific class included.

Sub-family		**angiosperms**		**lyco-phytes**	**mosses**	**chloro-phyceae**	**Mamiellales**				**bangio-phytes**
Species		***Atha***	***Osat***	***Smoe****	***Ppat***	***Crei***	***Otau***	***Oluc***	**1545**	**299**	***Cmer***
TALE super-class	KNOX	8	13	7	5	1	1	1	1	1	
	BELL	13	13	2	3						
	GSP1					1	1	1	1	1	
	algal					1	1	1	1	1	3
non-TALE super-class	HLZ1	31	13	4	12						
	HLZ2	10	14	2	5						
	HLZ3	5	5	4	5						
	HLZ4	16	12		4						
	WOX	15	14	9	3		1	1	5	1	
	HOX-DDT	2	2	2	1	1					
	OCP3	1	1	1	1	1	1	1	1	1	
	PHDf	2	2	1	1						
	PHDf2						1	1	1	1	
	LD	1	1	1	1						
	Other	2		1	1		2	2	4**	3**	3
Total		106		34	42	5	8	8	14	9	6

table S13. Meiosis related genes in *Micromonas* and comparison to *Ostreococcus.* Abbreviations: *Otau*, *O. tauri*; *Oluc*, *O. lucimarinus*; NF, not found by BLASTP or TBLASTN; *Important non-meiotic paralog of meiotic gene; **Model missing significant portions of homology.

Gene name	RCC299	CCMP1545	*Otau*	*Oluc*
Double strand break formation				
SPO11	97333	70176	20970	27014
*TOP6A**	104711	64682	32686	12516
*TOP6B**	99582	32362	11079	49589
Double strand break processing: MRX complex				
RAD50	106725	46506	30141	28736
MRE11	61359	70239	11562	43144
NBS1-XRS2	NF	NF	NF	NF
Rad51/recA family: homology searching and strand exchange				
*RECA**	59802	61635	19698	26240
RAD51A	112646	58104	18636	33041
RAD51B	64031	56117	not found	50387
RAD51C	112654	49939	20285	17626
RAD51D	96417	60599	13084	17589
DMC1	112647	70177	35537	17346
putative *XRCC3*	61644	51479	no clear ortholog	no clear ortholog
putative *XRCC2*	57387	70178	37766	93957
Loading of Rad51/DMC1, strand invasion, cross-over formation, mismatch repair				
RAD54 homolog 1	60933	58384	18619	16586
RAD54B	99013	57638	1479	16860
BRCA2	62958	62866	12787**	26389
RAD52	NF	NF	NF	NF
MLH1	96433	44648	29370	51248
MER3	57057	56409	17102	31874
MSH4	61120	70180	NF	NF
MSH5	112660	70184	19426**	6754**
*MSH1**	81056	51959	27845	34085
MND1	98019	34787	19307	33704
HOP2/MEU13/TBPIP	108157	39167	32332	37149
MUS81	96848	49893	32052	92534
MMS4/EME1	103060?	NF	NF	NF
Synaptonemal complex proteins, chiasma formation, post-meiotic segregation				
HOP1/ASY1	NF	NF	32765	31598
putative *STAG3*	54972	64715	23222	13918
Shugoshin/*SGO1*/*SGO2*	NF	NF	NF	NF
ZIP1	NF	NF	NF	NF
PMS1	65816	4434	17943	432
Other meiosis-related genes				
SWI1/AM1	NF	NF	NF	NF
PTD	NF	NF	NF	NF

table S14. The number of RWP-RK genes in analyzed green lineage organisms. RWP are algae-specific. Abbreviations: *Atha*, *A. thaliana*; *Osat*, *O. sativa*; *Ppat*, *P. patens*; *Crei*, *C. reinhardtii*; *Otau*, *O. tauri*; *Oluc*, *O. lucimarinus*; *on idiomorphic chromosomes or loci.

	MID/MLPs			NIT2/NLPs			RKDs			RWP			
Species	**MID**	**MLPa**	**MLPb**	**NLPlp**	**NLPpr**	**NIT2**	**LP-a**	**LP-b**	**RKDpr**	**a**	**b**	**c**	**Total**
Atha				9			4	1					14
Osat	1	1	1	5			4	1					13
Ppat			4	4				1					9
Crei	1*					1				11	3		16
Otau		1*			1				1*			1	4
Oluc		1*			1				1*			1	4
CCMP1545					3				1		1	1	6
RCC299			1*		2				1		1	1	6

table S15a. Search for HRGP candidates in green plant species (with shaft length ≥40). Published peptide gene models from *A. thaliana*, *P. patens, C. reinhardtii*, *O. tauri*, *Micromonas* RCC299 and CCMP1545 and *C. merolae* (as an outgroup) were subjected to SEG filtration (for low-complexity regions) and proline-content cutoff (30%) within low-complexity regions. Known *Arabidopsis* HRGPs were used to decide parameters for SEG and Pro-cutoff. (SEG parameters are set for 3.0/3.1; window size is 50 for land plants and *Chlamydomonas* and 30 for Mamiellales and *C. merolae*). Abbreviations: *Atha*, *A. thaliana*; *Ppat*, *P. patens*; *Crei*, *C. reinhardtii*; *Otau*, *O. tauri*; *Cmer*, *C. merolae*.

	Atha	***Ppat***	***Crei***	***Otau***	**RCC299 / CCMP1545**	***Cmer***
# of HRGP candidates	238	98	209	48	124 / 118	8
≥ 200 aa	41	8	7	0	2 / 5	2
≥ 100 aa	70	20	26	4	17 / 26	1

table S15b. Protein ID numbers for HRGP candidates reported in table S15a in RCC299, CCMP1545 as well as *O. tauri* (for which no previous reports on HRGPs had been made). For those in the *Micromonas* species, discarded gene models and the models with obvious non-HRGP motifs were excluded below.

	Protein ID numbers
RCC299 (102)	50376, 51916, 54844, 55401, 55531, 55539, 55684, 54016, 56042, 56758, 56778, 57200, 57391, 57584, 57684, 57699, 57825, 57919, 57959, 58084, 58188, 58437, 58473, 58576, 58581, 58618, 58759, 59043, 59125, 59230, 59232, 59235, 59528, 59639, 59666, 59798, 59810, 59886, 60160, 60603, 60945, 60962, 61072, 61337, 61487, 61550, 61575, 61732, 61923, 62037, 62162, 62604, 62866, 62867, 62897, 62947, 63160, 64081, 64086, 64139, 65015, 82791, 96442, 97407, 98644, 98844, 98903, 98966, 98990, 99146, 99194, 99459, 99833, 99853, 99936, 99988, 100163, 100979, 101034, 101192, 101569, 102272, 102449, 102575, 102654, 102909, 103279, 103527, 103632, 104180, 104310, 104590, 104601, 105416, 105988, 106402, 107639, 108073, 108625, 109368, 109388, 109610
CCMP1545 (111)	15496, 37791, 37953, 38431, 39086, 39363, 39637, 40570, 40794, 41353, 41372, 41654, 41758, 42194, 42270, 42435, 42493, 42647, 43505, 46642, 46816, 46822, 47070, 47301, 47578, 47677, 47857, 48042, 48104, 48223, 48266, 48454, 48918, 49461, 49842, 50433, 50505, 50565, 50694, 50959, 51162, 51206, 51486, 51544, 51730, 51975, 52067, 52092, 52297, 52382, 52492, 52612, 52677, 53306, 53340, 53461, 54072, 54330, 54364, 54374, 54533, 54739, 55968, 56527, 56533, 56638, 56695, 56835, 57056, 57098, 57111, 57113, 57203, 57241, 57418, 57551, 57645, 57766, 57995, 58134, 58355, 58671, 58708, 58712, 58805, 59505, 59731, 59817, 60066, 60121, 60131, 60420, 60482, 60564, 60863, 60893, 61736, 61760, 62805, 63086, 63216, 63656, 64110, 66506, 66636, 66789, 66818, 66823, 66830, 67304, 70231
***O. tauri* (47)**	4035, 10800, 11011, 11679, 11728, 11991, 12432, 12477, 14208, 21155, 23661, 23844, 24218, 24272, 24585, 28562, 29981, 29991, 30415, 30442, 30532, 30666, 31169, 31469, 32142, 32533, 32819, 32887, 32982, 33741, 33839, 34163, 34552, 34659, 34888, 35163, 35210, 35676, 35767, 35797, 35814, 36165, 36944, 37053, 37206, 37637, 37936

table S16. Gene models used to generate HRGP structures shown in figure S13. Abbreviations: Otau, *O. tauri*; Oluc, *O. lucimarinus*.

Motifs	RCC299 / CCMP1545	Proline content (%)	*Ostreococcus*	Proline content (%)
SP2-SP6	61337 / ----	73	Otau_37936	69
	102272 / 59505	~65	Otau_36165/Oluc_93602	~73
XP2	62948 / 48223	~63	Otau_11679/Oluc_16112	64
XP3/P*X/PXX	101034 / 41372	~54	Otau_32982/Oluc_31804	32
	103527 / 70231	45	Otau_37637	~33
***P**	---- / 42435	~27	Otau_30532	46

table S17a. RCC299 CAZymes. These genes were identified within the catalog gene models and bulk annotated in the CAZy pipeline, which also indicates whether a model has “issues” or not. Abbreviations: *, Only 5' or 3' end of the model needs editing; **, Both 5' and 3' end of the model needs editing; √, No issues with the model; ?, Unchecked model; !, model likely needs other types of editing but EST did not provide support to improve.

RCC299-Glycoside Hydrolases (GH)			Model Notes
ProtID	**Protein Defline**	**Description**	
51444	GH Family 2	related to β-galactosidases	*
58752	GH Family 3	related to β-N-acetylhexosaminidases	√
69097	GH Family 5	related to β-mannosidases / β-mannanases	*
83090	GH Family 5	related to β-mannosidases / β-mannanases	√
102881	GH Family 5	candidate β-glycosidase	√
108703	GH Family 13	too short to be annotated	*
68279	GH Family 13	candidate α-glycosidase	√
59406	GH Family 13	candidate α-glycosidase	√
62796	GH Family 13	candidate α-glycosidase	√
54903	GH Family 13	related to α-glycosidases	√
105010	GH Family 13	candidate isoamylase	√
106402	GH Family 13	candidate isoamylase	√
97635	GH Family 13	candidate limit dextrinase	√
96256	GH Family 13	candidate α-amylase	*
97885	GH Family 13	candidate α-amylase	√
96388	GH Family 13	candidate α-amylase	*
96728	GH Family 13	candidate α-amylase	*
55966	GH Family 13	candidate α-amylase	*
84396	GH Family 13	candidate 1,4-α-glucan branching enzyme	√
92897	GH Family 13	candidate 1,4-α-glucan branching enzyme	*
104965	GH Family 13	candidate 1,4-α-glucan branching enzyme	√
96665	GH Family 14	candidate β-amylase	**
108038	GH Family 14	candidate β-amylase	*
58948	GH Family 16	related to β-glycosidase	√
104347	GH Family 18	distantly related to chitinases	√

109150	GH Family 31	candidate α-glycosidase	*
77598	GH Family 31	candidate α-glycosidase	*
84462	GH Family 31	candidate α-glycosidase	√
84555	GH Family 31	candidate α-glucosidase	√
65917	GH Family 36	candidate α-galactosidase or raffinose synthase or stachyose synthase	*
82031	GH Family 37	candidate trehalase	√
65443	GH Family 38	candidate α-mannosidase	*
62591	GH Family 38	distantly related to α-mannosidases	√
109216	GH Family 47	candidate α-mannosidase	**
78864	GH Family 47	candidate α-mannosidase	*
97702	GH Family 47	candidate α-mannosidase	*
89545	GH Family 77	candidate 4-α-glucanotransferase	√
86873	GH Family 77	candidate 4-α-glucanotransferase with N-terminal CBM20 module	√
59291	GH Family 89	candidate α-N-acetylglucosaminidase	√
58717	GH Family 103	related to peptidoglycan lytic transglycosylases	√
55788	Unknown protein	related to α-glucosidases	√
RCC299-Glycosyltransferases (GT)			
79960	GT Family 1	C-terminal subunit of α-glycosyltransferase	√
72673	GT Family 1	N-terminal subunit of α-glycosyltransferase	*
79533	GT Family 2	distantly related to β-glycosyltransferases	√
57182	GT Family 2	candidate β-glycosyltransferase	√
51621	GT Family 2	candidate β-glycosyltransferase	√
72071	GT Family 2	related to β-glycosyltransferases	*
97002	GT Family 2	candidate GDP-Man: dolichyl-phosphate β-mannosyltransferase	√
97997	GT Family 2	candidate polysaccharide-forming β-glycosyltransferase, related to mannan synthases	√
100449	GT Family 4	related to α-glycosyltransferases	√
56201	GT Family 4	related to α-glycosyltransferases	√
85686	GT Family 4	related to α-glycosyltransferases	√

100701	GT Family 4	related to α-glycosyltransferases	*
92755	GT Family 4	related to α-mannosyltransferases	√
97051	GT Family 4	related to α-N-acetylglucosaminyltransferases	√
56980	GT Family 4	distantly related to glycosyltransferases	√
98317	GT Family 4	related to α-mannosyltransferases	√
83849	GT Family 4	candidate digalactosyldiacylglycerol synthase	√
86997	GT Family 4	candidate digalactosyldiacylglycerol synthase	√
80692	GT Family 4	related to UDP-sulfoquinovose:diacylglycerol sulfoquinovosyltransferases	√
105006	GT Family 5	candidate glycogen synthase	*
88745	GT Family 5	candidate glycogen synthase	*
104736	GT Family 5	candidate glycogen synthase	√
83804	GT Family 5	candidate glycogen synthase	√
66536	GT Family 5	candidate glycogen synthase	*
104862	GT Family 5	candidate glycogen synthase	√
104780	GT Family 5	candidate glycogen synthase	√
83891	GT Family 5	candidate glycogen synthase	√
104863	GT Family 5	candidate glycogen synthase	√
71804	GT Family 7	related to β-glycosidases	**
64091	GT Family 7	related to β-glycosyltransferases	√
68567	GT Family 13	candidate β-N-acetylglucosaminyltransferase	*
55432	GT Family 13	candidate β-glycosyltransferase	√
83473	GT Family 20	candidate bifunctional trehalose-6-phosphate synthase/trehalose-6-phosphate phosphatase	√
54984	GT Family 20	candidate bifunctional trehalose-6-phosphate synthase/trehalose-6-phosphate phosphatase	√
63004	GT Family 21	candidate β-glycosyltransferase	√
63738	GT Family 22	related to Dol-P-Man: α-mannosyltransferases	√
81043	GT Family 22	related to Dol-P-Man: α-mannosyltransferases	√
79984	GT Family 22	related to Dol-P-Man: α-mannosyltransferases	√
57621	GT Family 22	related to Dol-P-Man: α-mannosyltransferases	√

55318	GT Family 24	candidate UDP-glucose:glycoprotein α-glucosyltransferase	√
65017	GT Family 25	distantly related to β-glycosyltransferases	√
64958	GT Family 25	distantly related to β-glycosyltransferases	√
87019	GT Family 28	related to β-glycosyltransferases	√
77908	GT Family 28	candidate monogalactosyldiacylglycerol synthase	√
55907	GT Family 32	candidate α-glycosyltransferase	√
68238	GT Family 33	candidate β-glycosyltransferase	√
104976	GT Family 35	candidate α-glucan phosphorylase	*
87288	GT Family 35	candidate α-glucan phosphorylase	√
104969	GT Family 35	candidate α-glucan phosphorylase	√
78825	GT Family 41	distantly related to O-linked N-acetylglucosamine transferase	√
54821	GT Family 47	candidate β-glycosyltransferase	√
50255	GT Family 47	candidate β-glycosyltransferase	√
99882	GT Family 47	related to β-glycosyltransferases	√
79168	GT Family 48	candidate β-1,3-glucan synthase	**
78941	GT Family 48	candidate β-1,3-glucan synthase	**
51435	GT Family 50	candidate dolichyl-phosphate-sugar a-glycosyltransferase	√
60258	GT Family 51	candidate bifunctional family GT51 β-glycosyltransferase/PBP transpeptidase (candidate murein polymerase)	√
104553	GT Family 57	candidate Dol-P-sugar: α-glycosyltransferase	*
67658	GT Family 57	candidate Dol-P-sugar: α-glycosyltransferase	**
58254	GT Family 57	candidate Dol-P-sugar: α-glycosyltransferase	√
68602	GT Family 58	related to Dol-P-Man: α-mannosyltransferases	√
54895	GT Family 60	related to glycosyltransferases	√
99531	GT Family 60	distantly related to glycosyltransferases	√
70089	GT Family 64	candidate α-glycosyltransferase	**
108264	GT Family 66	candidate oligosaccharyl transferase STT3 subunit	*
94195	GT Family 66	candidate oligosaccharyl transferase STT3 subunit	√
51871	GT Family 71	related to α-mannosyltransferases	√
98290	GT Family 76	related to Dol-P-Man: a-mannosyltransferases	√

68853	GT Family 77	candidate glycosyltransferase	*
66770	GT Family 77	candidate glycosyltransferase	*
107680	GT Family 77	candidate arabinosyltransferase	√
99546	GT Family 77	candidate glycosyltransferase	√
62282	GT Family 77	candidate glycosyltransferase	√
61291	GT Family 77	candidate glycosyltransferase	√
60906	GT Family 77	distantly related to glycosyltransferases	√
80159	GT Family 77	candidate α-glycosyltransferases	*
101746	GT Family 77	candidate α-glycosyltransferase	√
64841	GT Family 77	candidate α-glycosyltransferase	*
109240	Unknown	too short for reliable annotation	!
109532	Unknown	distantly related to glycosyltransferases	√
55865	Unknown	distantly related to glycosyltransferases	√
50172	Unknown	distantly related to glycosyltransferases	√
101129	Unknown	distantly related to glycosyltransferases	√
62693	Unknown	distantly related to glycosyltransferases	√
108755	Unknown	distantly related to glycosyltransferases	√
RCC299-Proteins with Carbohydrate Binding Modules (CBM) not found above			
99058	CBM Family 1	CBM Family 1 protein (secreted protein with five CBM1 modules)	√
76823	CBM Family 20	candidate α-glucan-binding	√
76840	CBM Family 20	candidate α-glucan-binding	√
108362	CBM Family 20	candidate α-glucan-binding	√
59249	CBM Family 20	candidate α-glucan-binding	√
109264	CBM Family 20	candidate α-glucan-binding	√
108056	CBM Family 20	candidate α-glucan-binding	√
105911	CBM Family 20	candidate α-glucan-binding	√
106222	CBM Family 20	candidate α-glucan-binding	√
103766	CBM Family 20	candidate α-glucan-binding	√
104865	CBM Family 20	candidate α-glucan-binding	√
59003	CBM Family 20	candidate α-glucan-binding	√

61329	CBM Family 43	distantly related to β-1,3-glucan binding	√
81096	CBM Family 45	candidate α-glucan, water dikinase	√
96442	CBM Family 45	candidate α-glucan, water dikinase	√
107497	CBM Family 48	candidate α-glucan-binding	√
89788	CBM Family 48	candidate α-glucan-binding	√
55897	CBM Family 48	candidate α-glucan-binding	√

table S17b. CCMP1545 CAZymes. These genes were wre identified within the catalog gene models and bulk annotated in the CAZy pipeline, which also indicates whether a model has “issues” or not. Abbreviations: *, Only 5' or 3' end of the model needs editing; **, Both 5' and 3' end of the model needs editing; √, No issues with the model; ?, Unchecked model; !, model likely needs other types of editing but EST did not provide support to improve.

CCMP1545-Glycoside hydrolases (GH)			Model Notes
Prot ID	Protein Defline	Description	
58880	GH Family 3	related to β-N-acetylhexosaminidases	*
4622	GH Family 5	related to β-mannosidases / β-mannanases	**
3588	GH Family 5	related to β-mannosidases / β-mannanases	**
3730	GH Family 5	related to β-mannosidases / β-mannanases	**
31367	GH Family 5	related to β-mannosidases / β-mannanases	**
63839	GH Family 13	candidate α-glycosidase	√
64215	GH Family 13	related to α-glycosidases	√
34447	GH Family 13	candidate α-glycosidase	√
46298	GH Family 13	candidate α-glycosidase	√
45924	GH Family 13	candidate isoamylase	√
34146	GH Family 13	candidate isoamylase	*
24397	GH Family 13	candidate limit dextrinase	**
26488	GH Family 13	candidate α-amylase	*
26376	GH Family 13	candidate α-amylase	*
28161	GH Family 13	candidate α-amylase	*
3621	GH Family 13	candidate α-amylase	*
43705	GH Family 13	candidate α-amylase	*
44337	GH Family 13	candidate 1,4-α-glucan branching enzyme	√
19221	GH Family 13	candidate 1,4-α-glucan branching enzyme	√
49461	GH Family 13	candidate 1,4-α-glucan branching enzyme	√
19268	GH Family 14	candidate β-amylase	*
21290	GH Family 14	candidate β-amylase	*
3233	GH Family 18	distantly related to chitinases	**

31321	GH Family 31	candidate α-glycosidase	√
34641	GH Family 31	candidate α-glycosidase	√
10168	GH Family 31	candidate α-glycosidase	√
27057	GH Family 31	candidate α-glucosidase	*
51844	GH Family 36	candidate α-galactosidase or raffinose synthase or stachyose synthase	!
1967	GH Family 37	candidate trehalase	!
3052	GH Family 47	candidate α-mannosidase	**
2646	GH Family 47	candidate α-mannosidase	*
1532	GH Family 47	candidate α-mannosidase	*
25076	GH Family 77	candidate α-mannosidase	√
20445	GH Family 77	candidate α-mannosidase	√
58912	GH Family 103	candidate α-mannosidase	√
48400	Unknown	candidate α-mannosidase	√
CCMP1545-Glycosyltransferases (GT)			
8344	GT Family 1	C-terminal subunit of α-glycosyltransferase	√
26128	GT Family 1	N-terminal subunit of α-glycosyltransferase	√
6279	GT Family 2	candidate β-glycosyltransferase	**
49491	GT Family 2	distantly related to β-glycosyltransferases	*
54974	GT Family 2	candidate β-glycosyltransferase	√
64253	GT Family 2	candidate β-glycosyltransferase	√
49578	GT Family 2	candidate GDP-Man: dolichyl-phosphate β-mannosyltransferase	**
23128	GT Family 2	candidate polysaccharide-forming β-glycosyltransferase; related to mannan synthases	*
3914	GT Family 4	related to α-glycosyltransferases	√
10566	GT Family 4	distantly related to glycosyltransferases	√
36870	GT Family 4	candidate α-glycosyltransferase	√
62793	GT Family 4	related to α-glycosyltransferases	√
24693	GT Family 4	related to α-mannosyltransferases	?

12476	GT Family 4	related to α-N-acetylglucosaminyltransferases	√
62070	GT Family 4	related to α-mannosyltransferases	√
20694	GT Family 4	candidate digalactosyldiacylglycerol synthase	√
46248	GT Family 4	candidate digalactosyldiacylglycerol synthase	√
56959	GT Family 4	related to UDP-sulfoquinovose:diacylglycerol sulfoquinovosyltransferase	√
20747	GT Family 5	candidate starch synthase	*
44605	GT Family 5	candidate starch synthase	*
46038	GT Family 5	candidate starch synthase	*
49698	GT Family 5	candidate starch synthase	*
1440	GT Family 5	candidate starch synthase	**
29784	GT Family 5	candidate starch synthase	*
47271	GT Family 5	candidate starch synthase	√
33450	GT Family 5	candidate starch synthase	√
48851	GT Family 5	candidate starch synthase	*
38613	GT Family 7	candidate β-glycosyltransferase	*
58523	GT Family 7	candidate β-glycosyltransferase	**
32201	GT Family 13	candidate β-N-acetylglucosaminyltransferase	**
12696	GT Family 13	candidate β-glycosyltransferase	**
23871	GT Family 17	candidate β-glycosyltransferase	√
12941	GT Family 20	candidate bifunctional trehalose-6-phosphate synthase/trehalose-6-phosphate phosphatase	*
25187	GT Family 20	candidate bifunctional trehalose-6-phosphate synthase/trehalose-6-phosphate phosphatase	*
51114	GT Family 21	candidate β-glycosyltransferase	*
64397	GT Family 22	related to Dol-P-Man: α-mannosyltransferases	?
22710	GT Family 22	related to Dol-P-Man: α-mannosyltransferases	√
57017	GT Family 22	related to Dol-P-Man: α-mannosyltransferases	√
52108	GT Family 22	related to Dol-P-Man: α-mannosyltransferases	√

31914	GT Family 24	candidate UDP-glucose:glyco α-glucosyltransferase	√
54865	GT Family 25	distantly related to β-glycosyltransferases	√
19208	GT Family 25	distantly related to β-glycosyltransferases	√
51652	GT Family 25	distantly related to β-glycosyltransferases	√
4541	GT Family 28	related to β-glycosyltransferases	√
45255	GT Family 28	candidate monogalactosyldiacylglycerol synthase	√
3776	GT Family 31	candidate β-glycosyltransferase	*
57280	GT Family 32	candidate α-glycosyltransferase	*
2692	GT Family 33	candidate β-glycosyltransferase	*
35182	GT Family 35	candidate α-glucan phosphorylase	√
35264	GT Family 35	candidate α-glucan phosphorylase	√
49649	GT Family 35	candidate α-glucan phosphorylase	√
46435	GT Family 41	distantly related to O-linked N-acetylglucosamine transferase	√
53679	GT Family 43	candidate β-glycosyltransferase	**
1409	GT Family 51	candidate bifunctional family GT51 β-glycosyltransferase/PBP transpeptidase	*
35078	GT Family 57	related to Dol-P-Glc: α-glucosyltransferases	√
32727	GT Family 57	related to Dol-P-Glc: α-glucosyltransferases	√
3334	GT Family 58	related to Dol-P-Man: α-mannosyltransferases	*
56425	GT Family 60	related to α-glycosyltransferases	*
16819	GT Family 64	candidate α-glycosyltransferase	*
691	GT Family 66	candidate oligosaccharyltransferase	*
32514	GT Family 66	candidate oligosaccharyltransferase	√
42346	GT Family 76	related to Dol-P-Man: α-mannosyltransferases	!
29478	GT Family 77	candidate glycosyltransferase	*
36030	GT Family 77	candidate glycosyltransferase	√
67258	GT Family 77	candidate glycosyltransferase	*
62376	GT Family 77	candidate glycosyltransferase	*

51877	GT Family 77	candidate glycosyltransferase	√
60390	GT Family 77	related to glycosyltransferases	**
47386	GT Family 77	distantly related to glycosyltransferases	√
59967	Unknown	distantly related to glycosyltransferases	√
56194	Unknown	distantly related to glycosyltransferases	√
54830	Unknown	distantly related to glycosyltransferases	√
CCMP1545-Proteins with Carbohydrate Binding Modules (CBM) not found above			
70125	CBM Family 20	candidate α-glucan-binding	*
41138	CBM Family 20	candidate α-glucan-binding	√
63161	CBM Family 20	candidate α-glucan-binding	√
60721	CBM Family 20	candidate α-glucan-binding	√
57937	CBM Family 20	candidate α-glucan-binding	√
48556	CBM Family 20	candidate α-glucan-binding	√
52920	CBM Family 41	candidate α-glucan-binding	√
47308	CBM Family 45	candidate α-glucan, water dikinase	√
48104	CBM Family 45	candidate α-glucan, water dikinase	√
30666	CBM Family 48	candidate 5'-AMP-activated kinase with CBM48 a-glucan-binding domain	*
32976	CBM Family 48	candidate α-glucan-binding	√
CCMP1545-Category unknown			
60822	Unknown		

table S17c. Comparison of carbohydrate active anzymes (CAZy) to other organisms. Manual checking (with TBLASTN) was used only when there was a discrepancy between RCC299 and CCMP1545, where one showed zero members while the other showed one or more. Abbreviations: Otau, *O. tauri*; Osat, *O. sativa*; Atha, *A. thaliana*.

Glycoside Hydrolases Family

	1	2	3	5	9	10	13	14	16	17	18	19	20	27	28	29	31	32	33	35	36	37	38	43	47	51	63	77	79	81	85	89	95	100	103	TOTAL
RCC299	0	1	1	3	0	0	16	2	1	0	1	0	0	0	0	0	4	0	0	0	1	1	2	0	3	0	0	2	0	0	0	1	0	0	1	40
CCMP1545	0	1	1	4	0	0	15	2	0	0	1	0	0	0	0	0	4	0	0	0	1	1	0	0	3	0	0	2	0	0	0	0	0	0	1	36
Otau	1	1	0	5	0	0	8	2	1	0	1	0	0	0	0	0	4	0	0	0	1	1	0	0	2	0	0	2	0	0	0	0	0	0	0	29
Osat	35	2	16	19	25	11	18	11	31	67	34	16	5	5	42	2	7	11	1	15	6	1	4	2	4	8	1	2	5	1	2	1	1	8	0	419
A.tha	48	2	15	13	25	12	10	9	33	51	10	14	3	4	69	1	5	8	1	18	6	1	4	2	5	2	2	2	3	2	2	1	1	9	0	393

Glycosyltransferases Family

	1	2	4	5	7	8	10	13	14	16	17	19	20	21	22	24	25	28	29	30	31	32	33	34	35	37	39	41	43	47	48	50	51	57	58	59	60	61	64	65	66	68	71	75	76	77	TOTAL
RCC299	1	6	11	9	2	0	0	2	0	0	0	0	2	1	4	1	2	2	0	0	0	1	1	0	3	0	0	1	0	3	2	1	1	3	1	0	2	0	1	0	2	0	1	0	1	10	77
CCMP1545	2	6	10	9	2	0	0	2	0	0	1	0	2	1	4	1	3	2	0	0	1	1	1	0	3	0	0	1	1	0	0	1	1	2	1	0	1	0	1	0	2	0	0	0	1	7	70
Otau	0	6	9	6	1	0	0	0	0	0	0	0	2	0	4	1	0	1	0	0	1	1	1	0	3	0	0	2	0	0	2	1	0	0	1	0	1	0	1	0	1	0	1	0	1	7	54
Osat	202	47	25	11	0	39	3	1	12	1	4	1	11	1	4	1	0	4	5	1	40	3	2	6	2	18	0	3	10	35	11	1	0	2	1	1	0	25	3	1	2	1	0	3	1	16	560
Atha	121	42	24	6	0	42	3	1	11	1	6	1	11	1	3	1	0	4	3	1	33	6	1	8	2	10	0	2	4	39	12	1	0	2	1	1	0	7	3	1	2	3	0	5	1	19	445

Polysaccharide Family

	1	4	TOTAL
RCC299	0	0	0
CCMP1545	0	0	0
Otau	0	0	0
Osat	12	4	16
Atha	27	7	34

Carbohydrate Esterase Family

	1	6	8	11	13	TOTAL
RCC299	1	0	0	1	0	2
CCMP1545	1	0	0	1	0	2
Otau	1	0	0	1	1	3
Osat	1	4	37	1	10	53
Atha	1	2	67	5	12	87

Carbohydrate-Binding Module Family

	1	13	18	20	22	41	43	45	48	49	TOTAL
RCC299	5	0	0	12	0	0	1	3	10	0	31
CCMP1545	6	0	0	6	0	2	0	3	8	0	25
Otau	0	0	0	3	0	1	0	1	5	0	10
Osat	0	3	14	5	12	0	52	4	13	4	107
Atha	0	5	10	3	15	0	60	6	12	3	114

Expensins

		TOTAL
RCC299	1	1
CCMP1545	2	2
Otau	1	1
Osat	63	63
Atha	34	34

Color legend:

- probably involved in cell wall synthesis/remodelling
- proteins targetting chitin : antifungal defence ?
- proteins clearly more abundant in Micromonas than in higher plants
- each of the two Micromonas has one GT1 that occurs as two subunits
- invertases

table S18. Flagella related genes in RCC299 and CCMP1545 and comparison to those identified within *O. tauri* and *O. lucimarinus*. Gene categories are taken from the *Chlamydomonas* Flagellar Proteome found at http://labs.umassmed.edu/chlamyfp/index.php. The lack of many flagellar-related genes in *Ostreococcus* is expected as it is not motile, while *Micromonas* is motile. Many of the *Micromonas* genes identified here fall within the 'shared' *Micromonas* pool of genes identified in Figure 2. Abbreviations: MP, motor protein; HC, Heavy Chain; NF, not found by TBLASTN; *Otau*, *O. tauri*; *Oluc*, *O. lucimarinus*.

Description	RCC299	CCMP1545	*Otau*	*Oluc*
Tubulin				
Alpha-1 Tubulin	92378	27021	Ostta4:8661	Ost9901_3:33239
Alpha-2 tubulin	94939	49722	Ostta4:23001	Ost9901_3:26788
Beta-1 tubulin	104730	55542	Ostta4:14981	Ost9901_3:12139
Beta-2 tubulin	NF	NF	Ostta4:14914	Ost9901_3:28827
Intraflagellar Transport (IFT)				
Kinesin-II MP	58932	43314	NF	NF
Kinesin II MP FLA8	58933	54062	NF	NF
Kinesin II associated Protein	96325	729	NF	NF
Cytoplasmic Dynein HC 1b	96530	46683	NF	NF
Dynein 1b Light Intermediate Chain	64335	36452	NF	NF
IFT 20	NF	57918	NF	NF
IFT 52	55001	32015	NF	NF
IFT 57	77398	13268	NF	NF
IFT 72/74	96354	67097	NF	NF
IFT 80	88460	44033	NF	NF
IFT 81	103474	28664	NF	NF
IFT 88	NF	NF	NF	NF
IFT 122	97668	46242	NF	NF
IFT 140	113468	42853	NF	NF
IFT 172	104776	36991	NF	NF
Outer Dynein Arm (ODA)				
ODA Heavy Chain alpha	104912	26271	NF	NF
ODA Heavy Chain beta	105018	48719	NF	NF
ODA Heavy Chain gamma	105012	59769	NF	NF
ODA Intermediate Chain 1	83829	50559	NF	NF
ODA Intermediate Chain 2	96653	49656	NF	NF
ODA Light Chain 1	80279	14094	NF	NF
ODA Light Chain 2	104800	49621	NF	NF
ODA Light Chain 3	55326	33499	NF	NF
ODA Light Chain 4	55326	24914	NF	NF
ODA Light Chain 5	86994	51433	NF	NF
ODA Light Chain 6	61118	34374	NF	NF
ODA Light Chain 7a	96385	23376	NF	NF
ODA Light Chain 7b	61322	59459	NF	NF
ODA Light Chain 8	113471	49456	Ostta4:30069	Ost9901_3:39788
ODA Docking Complex 1	105761	57860	NF	NF
ODA Docking Complex 2	59973	50387	NF	NF
ODA Docking Complex 3	51234	49419	NF	NF
ODA Protein ODA5	NF	NF	NF	NF
ODA5-associated adenylate kinase	88986	7468	NF	NF
Inner Dynein Arms (IDA)				
IDA Heavy Chain 1-alpha	78637	35522	Ostta4:17641	Ost9901_3:40832
IDA Heavy Chain 1-beta	104997	45975	Ostta4:21159	Ost9901_3:43542
IDA Heavy Chain 2	104998	45980	NF	NF
IDA Heavy Chain 3	96693	70829	NF	NF

IDA Heavy Chain 4	64231	45818	NF	NF
IDA Heavy Chain 5	56610	55899	NF	NF
IDA Heavy Chain 6	104978	45916	NF	NF
IDA Heavy Chain 7	NF	35233	NF	NF
IDA Heavy Chain 8	NF	NF	NF	NF
IDA Heavy Chain 9	NF	NF	NF	NF
IDA Heavy Chain 11	NF	NF	NF	NF
IDA Intermediate Chain IC138	55196	25391	NF	NF
IDA Intermediate Chain IC140	113460	51332	NF	NF
IDA Intermediate Chain Actin	90942	49663	Ostta4:29599	Ost9901_3:51545
IDA Light Chain p28	107419	18544	NF	NF
IDA Light Chain Tctex1	64859	22007	NF	NF
IDA Light Chain Tctex2b	97536	19963	NF	NF
Caltractin / Centrin	90289	49514	Ostta4:28265	Ost9901_3:31762
Dynein Regulatory Complex				
Dynein Regulatory Complex Protein	113289	49429	NF	NF
Radial Spoke				
Radial Spoke Protein 1	NF	NF	NF	NF
Radial Spoke Protein 2	NF	NF	NF	NF
Radial Spoke Protein 3	60075	12661	NF	NF
Radial Spoke Protein 4	NF	34926	NF	NF
Radial Spoke Protein 5	NF	NF	NF	NF
Radial Spoke Protein 6	NF	NF	NF	NF
Radial Spoke Protein 7	86323	34566	NF	NF
Radial Spoke Protein 8	96753	42622	NF	NF
Radial Spoke Protein 9	104734	38431	NF	NF
Radial Spoke Protein 10	80926	70831	NF	NF
Radial Spoke Protein 11	98177	63054	NF	NF
Radial Spoke Protein 12	96416	41216	NF	NF
Radial Spoke Protein 15	NF	NF	NF	NF
Radial Spoke Protein 16	62518	19233	NF	NF
Radial Spoke Protein 17	NF	NF	NF	NF
Radial Spoke Protein 23	98338	15249	NF	NF
Central Pair				
Kinesin-Like Protein 1	97339	39045	NF	NF
Central Pair Protein PF16	105015	35126	NF	NF
Central Pair Associated WD-Repeat Protein	96259	70832	NF	NF
Phosphatase 1	104871	26683	Ostta4:29521	Ost9901_3:27323
Central Pair Protein PF6	61791	42615	NF	NF
Central Pair Complex 1	102066	39998	NF	NF
Flagellar Membrane				
Gliding motility related CaM kinase	60958	58417	NF	NF
Flagella Membrane Glycoprotein 1A	NF	NF	NF	NF
Flagella Membrane Glycoprotein 1B	NF	NF	NF	NF
Mastigoneme	95172	36039	NF	NF
Basal Body				
Tubulin Gamma	84777	27658	Ostta4:25029	Ost9901_3:119480
Tubulin Delta	64707	20062	NF	NF
Tubulin Epsiolon	58680	58956	NF	NF

SF-assemblin	79164	16280	Ostta4:18452	Ost9901_3:12738
Bardet-Biedl Synd. 1	85433	36840	NF	NF
Bardet-Biedl Synd. 2	97220	29081	NF	NF
Bardet-Biedl Synd. 3	63795	44981	NF	NF
Bardet-Biedl Synd. 4	99142	50814	NF	NF
Bardet-Biedl Synd. 5	87458	34174	NF	NF
Bardet-Biedl Synd. 7	100192	16331	NF	NF
Bardet-Biedl Synd. 8	105855	46012	NF	NF
Bardet-Biedl Synd. 9	55060	55647	NF	NF
Similar to oral-facial-digital 1	59168	60630	NF	NF
Variable Flagellar Number 3	102797	47536	NF	NF
Basal Body Protein BLD10	55365	70912	Ostta4:9321	Ost9901_3:28747
Axoneme				
Calmodulin	104708	49475	Ostta4:24922	Ost9901_3:39965
Deflagellation Inducible Protein	96664	40878	Ostta4:36928	Ost9901_3:27600
Heat Shock 70 kDa Protein	104823	30210	Ostta4:22076	Ost9901_3:28169
Coiled-Coil Flagellar Protein	100280	67048	Ostta4:36810	Ost9901_3:25488
Flagellar Protofilament Ribbon Protein	109496	46589	NF	NF
Nucleoside-diphosphokinase regulatory subunit p72	96444	33490	Ostta4:24398	Ost9901_3:27168
Protein Phosphatase 2a	97654	30915	Ostta4:33846	Ost9901_3:32622
Profilin	79195	16052	NF	NF
Tektin	NF	NF	NF	NF
Mating Related				
Putative CALK protein kinase	113993	58004	Ostta4:4957	Ost9901_3:6235
cGMP-dependent protein kinase	64355	53209	Ostta4:1251	Ost9901_3:12211
Methionine Synthase	NF	NF	NF	NF
Protein Kinase Regulated by Mating	63856 58552	63856	NF	NF
Length Control				
Long Flagella Protein 1	NF	NF	NF	NF
Long Flagella Protein 3	NF	NF	NF	NF
Long Flagella Protein 4	96377	40264	NF	NF
Glycogen Synthase Kinase 3	104950	30445	Ostta4:28157	Ost9901_3:49296
Uncategorized				
Katanin p80 subunit	104810	56940	NF	NF
Katanin p60 subunit	97315	16445	NF	NF
Microtubule-associated protein EB1	95506	68917	NF	NF
Flagellar Autotomy Protein FA2 Protein Kinase	112655	58165	Ostta4:5403	Ost9901_3:43364
Flagellar Autotomy Protein	61668	70916	Ostta4:29659	Ost9901_3:40751
Novel Actin-Like Protein	NF	NF	NF	NF

table S19. Analysis of transporters in RCC299 and CCMP1545 compared to the *Ostreococcus* genomes. Annotations can be found on the JGI portals. Additional comparisons are most easily viewed at TransporterDB (http://www.membranetransport.org/). Abbreviations: NF, not found by BLASTP or TBLASTN; 0, zero found, but result not confirmed using TBLASTN; *, more may be present but potential hit did not provide enough evidence to clearly support; blue text, numbers were modified from that in TransporterDB after manual investigation; ?, could not find category in current release of TransporterDB.

Transporter Family	***Micromonas***		***Ostreococcus***	
	1545	299	*Oluc*	*Otau*
ATP dependent				
The ATP-binding Cassette (ABC) Superfamily	51	55	40	42
The Arsenite-Antimonite (ArsAB) Efflux Family	1	1	1	1
The Chloroplast Envelope Protein Translocase (CEPT or Tic-Toc) Family	15	14	0	1
The H+- or Na+-translocating F-type, V-type and A-type ATPase (F-ATPase) Superfamily	23	24	23	23
The H+-translocating Pyrophosphatase (H+-PPase) Family	3	3	3	5
The Type II (General) Secretory Pathway (IISP) Family	16	16	8	11
The Mitochondrial Protein Translocase (MPT) Family	11	12	11	9
The P-type ATPase (P-ATPase) Superfamily	14	15	14	14
Ion Channels				
The Ammonia Transporter Channel (AMT) Family	5	6	4	4
The Annexin (Annexin) Family	NF	1	1	1
The Anion Channel-forming Bestrophin (Bestrophin) Family	0	0	0	0
The Intracellular Chloride Channel (CLIC) Family	3	1*	0	0
The Copper Transporter (CTR) Family	2	2	1	1
The Glutamate-gated Ion Channel (GIC) Family of Neurotransmitter Receptors	NF	2	NF	1
The Inward Rectifier K+ Channel (IRK-C) Family	3	2	NF	0
The Neurotransmitter Receptor, Cys loop, Ligand-gated Ion Channel (LIC) Family	NF	1	NF	NF
The Major Intrinsic Protein (MIP) Family	0	0	1	0
The CorA Metal Ion Transporter (MIT) Family	5	4	2	1
The Small Conductance Mechanosensitive Ion Channel (MscS) Family	6	6	3	3
The Non-selective Cation Channel-2 (NSCC2) Family	1	1	1	1
The Polycystin Cation Channel (PCC) Family	1	3	NF	NF
The Presenilin ER Ca2+ Leak Channel (Presenilin) Family	1	1	0	0
The Chloroplast Envelope Anion Channel-forming Tic110 (Tic110) Family	1	1	1	1
The Transient Receptor Potential Ca2+ Channel (TRP-CC) Family	NF	1	0*	1
The Urea Transporter (UT) Family	0	0	0	0
The Voltage-gated Ion Channel (VIC) Superfamily	19	27	16	16
Secondary Transporter				
The ATP:ADP Antiporter (AAA) Family	1	1	1	1
The Amino Acid/Auxin Permease (AAAP) Family	7	8	5	4
The Anion Exchanger (AE) Family	4	6	2	3
The Auxin Efflux Carrier (AEC) Family	4	4	2	1
The Amino Acid-Polyamine-Organocation (APC) Family	NF	1	1	1
The Arsenite-Antimonite (ArsB) Efflux Family	1	1	1	1

The Bile Acid:Na+ Symporter (BASS) Family	3	6	5	5
The Betaine/Carnitine/Choline Transporter (BCCT) Family	NF	1	2	2
The Ca2+:Cation Antiporter (CaCA) Family	7	8	4	4
The Cation-Chloride Cotransporter (CCC) Family	1	1	NF	NF
The Cation Diffusion Facilitator (CDF) Family	2	5	2	2
The Chromate Ion Transporter (CHR) Family	1	1	1	1
The Chloride Carrier/Channel (ClC) Family	6	6	4	4
The Monovalent Cation:Proton Antiporter-1 (CPA1) Family	3	5	5	6
The Monovalent Cation:Proton Antiporter-2 (CPA2) Family	5	6	3	4
The Choline Transporter Like (CTL) Family	4	5	0	0
The Dicarboxylate/Amino Acid:Cation (Na+ or H+) Symporter (DAACS) Family	1	1	0	0
The Divalent Anion:Na+ Symporter (DASS) Family	5	5	5	6
The Drug/Metabolite Transporter (DMT) Superfamily	33	56	37	36
The Equilibrative Nucleoside Transporter (ENT) Family	2	2	1	1
The Folate-Biopterin Transporter (FBT) Family	3	3	3	3
The Formate-Nitrite Transporter (FNT) Family	1	1	1	1
The Glycerol Uptake (GUP) Family	1	1	NF	NF
The Hydroxy/Aromatic Amino Acid Permease (HAAAP) Family	2	2	3	3
The Lysosomal Cystine Transporter (LCT) Family	1	1	0	0
The Lactate Permease (LctP) Family	0	0	0	0
Mitochondrial tRNA Import Complex (M-RIC) (Formerly 9.C.8)	1	1	?	?
The Mitochondrial Carrier (MC) Family	40	42	41	39
The Chloroplast Maltose Exporter (MEX) Family	1	1	0	0
The Major Facilitator Superfamily (MFS)	44	53	36	39
The Multidrug/Oligosaccharidyl-lipid/Polysaccharide (MOP) Flippase Superfamily	13	15	12	10
The Mitochondrial Tricarboxylate Carrier (MTC) Family	0	0	0	0
The Nucleobase:Cation Symporter-1 (NCS1) Family	2	3	1	1
The Nucleobase:Cation Symporter-2 (NCS2) Family	NF	1	NF	NF
The NhaA Na+:H+ Antiporter (NhaA) Family	2	1	1	1
The NhaC Na+:H+ Antiporter (NhaC) Family	0	0	0	0
The NhaD Na+:H+ Antiporter (NhaD) Family	1	1	0	0
The Ni2+-Co2+ Transporter (NiCoT) Family	0	1	1	1
The Metal Ion (Mn2+-iron) Transporter (Nramp) Family	1	1	1	1
The Neurotransmitter:Sodium Symporter (NSS) Family	1	1	0	0
The Oligopeptide Transporter (OPT) Family	1	1	NF	NF
The Cytochrome Oxidase Biogenesis (Oxa1) Family	4	3	4	3
The Inorganic Phosphate Transporter (PiT) Family	1	1	2	2
The Phosphate:Na+ Symporter (PNaS) Family	2	4	0	0
The Proton-dependent Oligopeptide Transporter (POT) Family	NF	1	1	1

The Reduced Folate Carrier (RFC) Family	NF	1	NF	NF
The Resistance-Nodulation-Cell Division (RND) Superfamily	6	8	3	6
The Silicon Transporter (Sit) Family	0	0	0	0
The Solute:Sodium Symporter (SSS) Family	1	2	3	3
The Sulfate Permease (SulP) Family	3	4	2	3
The Twin Arginine Targeting (Tat) Family	3	3	3	3
The Telurite-resistance/Dicarboxylate Transporter (TDT) Family	0	0	0	0
The Threonine/Serine Exporter (ThrE) Family	0	0	0	0
The Vacuolar Iron Transporter (VIT) Family	1	1	0	0
The Zinc (Zn2+)-Iron (Fe2+) Permease (ZIP) Family	6	6	7	5
Unclassified				
The ATP Exporter (ATP-E) Family	1	1	0	0
The HlyC/CorC (HCC) Family	2	2	0	0
The Iron/Lead Transporter (ILT) Superfamily	0	0	0	0
The Mg2+ Transporter-E (MgtE) Family	2	2	0	1
The NIPA Mg2+ Uptake Permease (NIPA) Family	1	1	0	0
The Peroxisomal Protein Importer (PPI) Family	3	4	2	2
The Integral Membrane Peroxisomal Protein Importer-2 (PPI2) Family	1	1	0	0
The Tellurium Ion Resistance (TerC) Family	1	1	0	0
The Putative 4-Toluene Sulfonate Uptake Permease (TSUP) Family	2	3	0	0
The YggT or Fanciful K+ Uptake-B (FkuB; YggT) Family	3	3	0	0

table S20. Gene models putatively encoding enzymes involved in photorespiration or scavenging of reactive oxygen species (ROS) in RCC299 and CCMP1545. Abbreviations: NF, not found by BLASTP or TBLASTN.

Enzyme name	Gene Name	RCC299 ProtID	CCMP1545 ProtID
Photorespiration			
phosphoglycolate phosphatase	*PGP*	96626	36042
		70469	
glycolate oxidase	*GOX*	57273	45056
		98069	10655
putative serine-pyruvate aminotransferase and/or alanine glyoxylate transaminase	*SPT, AGT*	104778	23625
	SPT2, AGT2	59863	NF
glycine cleavage system (glycine decarboxylase) T-protein	*GDCT*	59804	50571
glycine cleavage system (glycine decarboxylase) P-protein	*GDCP*	104877	24398
glycine cleavage system (glycine decarboxylase) H-protein	*GDCH*	104779	37647
Dihydrolipoamide dehydrogenase also known as glycine decarboxylase L-protein, glycine cleavage system L-protein	*DLDH, GCSL*	104984	28757
	DLDH, GCSL2	104967	34732
Serine/glycine hydroxymethyltransferase	*SHMT, GHMT*	104794	29249
	SHMT, GHMT2	96092	49729
		63428	49634
hydroxypyruvate-like domains reductase		109401	49752
		62952	51051
		55222	14394
		90135	31180
putative glycerate kinase	*GLYK*	96229	44766
glutamine synthetase	*GSII, GLN*	112708	55724
		107969	30309
		58286	4228
Scavengers of ROIs			
Cu/Zn superoxide dismutase		63146	51251
		91758	36200
		108979	22091
Mn superoxide dismutase	*MSD*	60680	49539
L-ascorbate peroxidase		70664	5269
		71118	34503
		109237	48214
putative peroxidase		109115	8305

glutathione peroxidase	*GPX1*	112672	70206
	GPX2	112674	70207
	GPX3	112671	70208
	GPX4	112675	70209
	GPX5	112676	70210
peroxiredoxin	*PRDX*	112683	70211
putative peroxiredoxin			19224
glutathione synthase		108131	18089
glutathione S-transferase and/or glutaredoxin domains		85553	26627
		59720	
		56266	52891
		55730	59893
		63691	
		81517	15637
		55589	4004
		62817	
		107847	60783
		83445	59116
		84542	44340
		88190	5265
		103187	59973
		77911	49421
		64089	31490
		85068	9447
		108742	
		61069	
		70483	
glutathione reductase		91929	45738
monodehydroascorbate reductase	*MDR*	61108	36082
putative phytochelatin synthase		61050	6746

table S21. Genomic localization of transposable elements identified in CCMP1545 (none were identified in RCC299).

Element	Type	Scaffold	Start	End	State
Microline	LINE	19	74998	77148	complete
Microline	LINE	19	77267	77797	partial
Microline	LINE	19	84738	87305	complete
Microline	LINE	19	106622	108889	complete
Microline	LINE	19	112251	114872	complete
Microline	LINE	19	152825	153126	partial

table S22a. Distribution and relative frequency of Introner Elements (IE) in CCMP1545 scaffolds. The %GC reflects the GC content of the particular scaffold; the two low GC-chromosomes are highlighted in yellow. Numbers of IE falling into the 4 different classes are given and, in parentheses, the percentage of that IE class type. In most cases the "All" category represents average values for the 19 scaffolds. The IE relative frequency (IE Mb^{-1}) was computed for each scaffold and for the whole genome, with the highest (pink) and lowest (green) values being located on scaffold 1 and 19, respectively. See also table S22b for manually curated examples.

Scaffold	Length (bp)	%GC	IE1	IE2	IE3	IE4	IE in scaff	IE Mb^{-1}
scaffold_01	2211167	66.3	2686 (71)	572 (15)	369 (9.5)	170 (4.5)	3787	1713
scaffold_02	2171923	51.4	315 (58)	79 (14.5)	103 (19)	48 (9)	545	251
scaffold_03	1965373	67.1	218 (71)	77 (25)	10 (3)	4 (1)	309	157
scaffold_04	1602726	66.7	131 (70)	38 (20)	14 (7.5)	3 (2)	186	116
scaffold_05	1532348	67.3	242 (74)	48 (15)	29 (9)	9 (3)	328	214
scaffold_06	1224724	67.1	254 (76)	58 (17)	18 (5)	4 (1)	334	273
scaffold_07	1183541	67.9	278 (71)	60 (15)	43 (11)	13 (3)	394	333
scaffold_08	1177029	67.3	241 (76)	44 (14)	22 (7)	10 (3)	317	269
scaffold_09	1106798	67.1	152 (66)	46 (20)	21 (9)	10 (4)	229	207
scaffold_10	1116513	66.6	239 (72)	58 (17)	33 (10)	4 (2)	334	299
scaffold_11	952308	67.1	247 (72)	44 (13)	35 (15)	15 (4)	341	358
scaffold_12	950943	66.6	266 (72)	55 (15)	36 (10)	14 (4)	371	390
scaffold_13	892804	66.8	294 (73)	60 (15)	39 (10)	12 (3)	405	454
scaffold_14	880324	66.6	318 (68)	87 (18.5)	43 (9)	22 (5)	470	534
scaffold_15	793361	66.7	196 (72)	54 (20)	15 (5.5)	7 (3)	272	343
scaffold_16	777317	66.8	295 (74.5)	48 (12)	41 (10)	12 (3)	396	509
scaffold_17	640895	67.5	309 (74)	49 (12)	39 (9)	23 (5.5)	420	655
scaffold_18	518050	65.8	306 (66)	48 (10)	58 (12)	54 (12)	466	900
scaffold_19	245704	49.3	0	0	0	0	0	0
All	**21,943,848**	**65.2**	**6987**	**1525**	**958**	**434**	**9904**	**451**

table S22b. Genome coordinates and protein IDs (of the gene sequence containing the particular introner) of five example IE from each of the four IE categories (IE1, IE2, IE3, and IE4) found in CCMP1545. Coordinates given are relative to the plus strand, regardless of which strand the IE was located on. EST? indicates whether the gene model and IE has EST support (yes) or not (no)FastA sequences are provided below the table.

IE Name	Prot. ID	Strand	EST?	Left Coordinate	Right Coordinate
IE1.1	59716	minus	Yes	1008904	1009118
IE1.2	47641	minus	Yes	158413	158617
IE1.3	59239	plus	Yes	163455	163644
IE1.4	65615	plus	Yes	1014862	1015076
IE1.5	42614	plus	Yes	872832	873048
IE2.1	49634	plus	Yes	178558	178660
IE2.2	21889	minus	No	610573	610707
IE2.3	42614	plus	Yes	870986	871080
IE2.4	59720	minus	Yes	1015865	1015970
IE2.5	65296	minus	Yes	186492	186604
IE3.1	31241	minus	Yes	38139	38311
IE3.2	64010	minus	Yes	45491	45667
IE3.3	36039	minus	Yes	115521	115693
IE3.4	52727	minus	Yes	568166	568338
IE3.5	42577	minus	Yes	777693	777865
IE4.1	9381	minus	No	93616	93885
IE4.2	55024	plus	Yes	108986	109212
IE4.3	38389	minus	No	348046	348243
IE4.4	70978	minus	No	105609	105819
IE4.5	55027	plus	Yes	115771	115914

IE1

>IE1.1
GCGCGTTCTCTCTCAAACTGGTCCCCATACGACCGCGTCGGCGTGGTGCACGCCGATCCTTAAGGA
CTTTTCTTCCCGTCGCATCTCTCCGCCTACCCCACGGTTTCAATCCCGACACACCGCGATGCCTTTC
AACTCCGCTTCTGACGCCTTTGAACTCCACCCCGACGTCGCTTCGTACGGACCCTCGACCCTCAG
>IE1.2
GTGAGTTGACACACTGGTCCCCATACGACCCCGTCGGCGTGGTGAACGCCGTTTCTTAAGGACTTT
GCCCGTCGTTTCTCTCCGCCCACCCCACGGTTTCAATCCCCGCCCGCGACGCCTTTCAACTCCGCT
TCTGACGCCTTTGAACTCCACCCCGACGTTCGCTCGTATGGACCCTCGACCCTCAG
>IE1.3
GTGCGTTCTATACACTGGTCCCCATACGACCCCAATGGCGAGGTGGACGCCGATCCTTAAGGACTTT

GCCCGTCGTATCTCTCCGCCCACACCTCGCTTTCAATCCCCGCCTTCGACGCCTTTCAACTCCCAAC
TGACGCCTTTCAACTCCACCCCGACGTTCGCTCGTATAG
>IE1.4
GCGCGTTCTGTCTCACACTGGTCCCCGTACGACCGCGTCGGCGTGTTGAACGCCGATCCTTAAGGA
CTTTCTCTCCCGGCGTGTCTCTCCGTCCATCACCCCTCGCTTTCAATCCCCCCGCCCTCGACGCCTT
TCAACTCCATCTGACGCCTTTGAACCTCACCCCGACGTTCGCTTGTACGGACCCTCGACCCTCAG
>IE1.5
GCGCGTTCTATACACTGGTCCCCATACGACCGCGTTCTCGTGGTGAACGCCGATCCTTAAGGACTTT
TCCCGGCGATTCTCTCCGCCCATCCCTCGCTTTCAATCCCCGACCTCGACGCCTTTCAACTCCATCT
GACGCCTTTGAACTCCACCCCGACATCGCCTCAG

IE2
>IE2.1
GCGCGTCGCGTCGCGCCGTCTCGCGCCCGCGTCCCTCGGTGGTTTCAACGTTTGATCGCGTTCCCT
TTCAACTGATGACCGACGCATCGCCCTCCTTCTACAG
>IE2.2
GTGCGTCTGACCGCTCCCCATACGACCCCGTTCGCGTTTCGCGCGTCGTTTCTGAAGCCCTTTTTCT
TCACCCGCGCTTTCCGCTTTCAATATTTGATCGCGTCCCCTTTCAACTGACCGATGAACGACCATCA
G
>IE2.3
GTGCGTTCAGGGTGACAAAAAGTTAGTTTTTCACCCGTATTGCCCGGTTTCATCAACATTTGATCGCG
TCCCCTTTCAACAAATGACCGGTGAACTTTTTTTGTACGGCGGAATGGCCCTCATCATGCAG
>IE2.4
GTGCGTTCTATACAAAAGTTTTTCACCCATCGTCCGGTTTCAACGTTTGATCGCGTCCCCTTTCAACT
GACTGGTGAACTATTTTTGTATGGAATGGCCCTAAAAG
>IE2.5
GTGCGTTCTACGTGTATACAAACGTTTTTCACCCACCGCTCGGTTTCAACACTTGATCGCGTCCCCTT
TCAACTGACCGATGAACATTTTTTGTATGGAACGACCCTCATCAG

IE3
>IE3.1
CACCCCGACGACGCGGTGAGACTGCTTCCCATACGACCCCGTTCGCGTGGTGCACGCCATTCCTTA
AGGACTTTTCCCGTCGTCACTCTTCACCCGCGCTTCCCTTTCAACGTTTGACCGGTAAGACGTTCGA
CTGACCGATCGCTTCACCCACGCAGAATCAACAACATCAT
>IE3.2
CCGCGTCCGCGCGCGGTGAGACTGCTTCCCATACGACCCCGTTCGCGTGGTGCACGCCGTTCCTTA
AGGACTTTTCCCGTCTTCACTCTTCACCCGCGCTCCCCTTTCAACGTTCGTTTGACCGGTAAGACGTT
CGACTGACCGATCGCTTCACCCACGCAGGTCGCCCGCCTCGAG
>IE3.3
GGCGCGGCGGCGGCGGTGAGACTGCTTCCCATACGACCCCGTTCGCGTGGTGAACGCCGTTCCTT
AAGGACTTTTCCCGTCGTCACTCTTCACCCGCGTTTCCCTTTCAACGCTTGACCGGTAAGACGTTCG
ACTGACCGATCGCTTCACCCACGCAGGCGACGAGCGCGGCG
>IE3.4
AAGGCGTCCGCGGGGGTGAGACTGCTTCCCATACGACCCTGTTCGCGTGGTGCACGTCGTTCCTTG
AGGACTTTTCCCGTCGTCACTCTTCACCCGCGCTTCCCCTCCAACGTTTGACCGGTAAGACGTTCGA
CTGACCGATCGCTTCACCCACGCAGGCGCCGACGCCCGGG
>IE3.5
GACGACAACGCCGAGGTGAGACTGCTTCCCATACGACCCCGTTCGCGTGGCGCGCGCCGTTCCTTA
AGGACTTTTCCCGTCGACACTCTTCACCCGCGCTTCCCTTTCAACGTTTGACCGGTACGACGTTCGA
CTGACCGATCGCTTCACCCACGCAGTCGCTGTGCGCGCTC

IE4
>IE4.1
GTGAGACTGGTTCCCATACGACCCCGTTCTCGTGTTGATCGTCGTTTCTTAAGGAGTTCTGTGAGAC
CGCTTCCCGTACGACCTACCCCGTTCGTTCGCGCGGTGAACGCCGTTTCTTAAGGCGTACTTTCCTT

TCCTTTCGCGCGCTGTGAACCTCACGTTCGTTTTGACACGTGGGAATGATATCCACCAATCACATGC
ACGCGCGACTGACACGTGTCTTCCCTCGGCCTATCACAGgatg
>IE4.2
GTGAGACTGGAAGATGAGATCACAGACTGCGACTGCTCCCCGTACGACCACGTTCGCGTGTTGATC
GTCGTTTCTTAAGGCGTTCGTTATTTCGCGCGCTATGAACCTCACGTTTTGACACGTGGGATTTATCT
CCACCAATCATATGTGTACGCGACTGACACGTGTCTTCCCTCGGCCAATGAATGATATCGCAG
>IE4.3
GTGAGACTGCTTCCCGTACGACTCCGTTCGCGTGTTGATCGTCGTTTCTTAAGGAGTTCTTTCCTTTC
GCGCGCTATGATATGAACCTCACTTTTTGACAATCGGGAATGATATACACCAATCACATGCACACGTG
ACTGACACGTGTTTTCCCTCGGCCTATCACAG
>IE4.4
GTGAGACTGCTTCCCGTACGACCCCGTTCGCGCGTTGATCGTCGTTTCTTAAGGAGTTCGTTTTTTC
CTTTCGCGCGCTATGTCATGAATCTCACGTTCGTTTTGACACGTGTGAAGTATATCCACCAATAACAT
CCTCGCGTGACTGACACGTGTCTTCCCTCGGCCTATCGCGTCGCAG
>IE4.5
GTGAGACGGCAGACTGCTTCCCATACGACTCCGTTCGCGCGTCGATCGTCGTTTCTTAAGGAGTTCA
GGAGTTCTTTCCTTTCGCGCGCTATGTTATGAACCTCTCGTTTTGACACGTGGGAATGATAT

table S23a. Repeat sequences in the *Micromonas* CCMP1545 genome, many of which are IE.

Scaffold	Length (bp)	# of Repeat Elements	Repeat bp %	Mean Repeat Element length	Repeat Elements MB^{-1}
CCMP1545					
scaffold_01	2211167	501	4.1	179	231
scaffold_02	2171923	573	6.5	244	266
scaffold_03	1965373	490	4.6	183	251
scaffold_04	1602726	324	3.8	189	203
scaffold_05	1532348	492	6.6	203	324
scaffold_06	1224724	442	7.5	204	366
scaffold_07	1183541	536	9.7	212	456
scaffold_08	1177029	446	8.0	209	382
scaffold_09	1106798	342	5.9	188	312
scaffold_10	1116513	477	8.7	200	436
scaffold_11	952308	452	10.1	211	480
scaffold_12	950943	491	11.4	215	529
scaffold_13	892804	482	12.5	228	548
scaffold_14	880324	553	14.2	223	636
scaffold_15	793361	350	9.6	215	445
scaffold_16	777317	476	13.6	221	616
scaffold_17	640895	458	17.7	247	718
scaffold_18	518050	479	23.0	248	929
scaffold_19	245704	7	0.2	82	29
All	**21,943,848**	**8,371**	**8.2**	**213**	**381**

table S23b. Relic repeat sequences in the genome of *Micromonas* RCC299. Note that these were not related to IE. They were determined by homology to mips-REdat, a plant repeat sequence (http://mips.gsf.de/proj/plant/webapp/recat/index.jsp).

Chromosome	Length (bp)	# of Repeat Elements	Repeat bp %	Mean Repeat Element length	Repeat Elements MB^{-1}
RCC299					
chr_01	2053047	26	0.2	146	13
chr_02	1906540	39	0.2	105	21
chr_03	1759951	26	0.2	166	15
chr_04	1584431	39	0.5	214	25
chr_05	1518631	53	0.5	156	35
chr_06	1431126	39	0.3	94	27
chr_07	1394111	30	0.2	116	22
chr_08	1276783	37	0.5	189	29
chr_09	1260459	30	0.3	123	24
chr_10	1159938	34	0.5	172	29
chr_11	1145872	40	0.7	212	35
chr_12	1087245	34	0.4	135	32
chr_13	1011178	27	0.6	216	27
chr_14	832468	16	0.3	132	19
chr_15	739137	30	0.5	112	41
chr_16	608929	18	0.4	152	30
chr_17	214782	3	0.2	121	14
mitochondrial	47425	0	0.0		0
chloroplast	72585	0	0.0		0
All	**21,104,638**	**521**	**0.4**	**153**	**25**

table S24. Sequence Read Statistics for *Micromonas* RCC299 and CCMP1545.

Sequence Stats	3/8/2006 RCC299 Assembly		1/12/2007 CCMP1545 Assembly	
Insert Size (RCC299)	Untrimmed Reads	Untrimmed Sequence	Untrimmed Reads	Untrimmed Sequence
2-3 KB	145,349	152 MB	149,889	131 MB
6-8 KB	139,583	143 MB	139,392	121 MB
35-40 KB	53,759	55.8 MB	50,302	46.6 MB
Total Untrimmed	338,691	350.8 MB	339,583	298 MB
Insert Size (CCMP1545)	Trimmed Reads	Trimmed Sequence	Trimmed Reads	Trimmed Sequence
2-3 KB	133,758	96.9 MB	137,500	94.2 MB
6-8 KB	125,072	85.4 MB	125,745	86.0 MB
35-40 KB	45,492	25.6 MB	29,444	14.0 MB
Total Trimmed	304,322	208 MB	292,689	194 MB

table S25. Assembly Statistics for *Micromonas* RCC299 and CCMP1545.

Assembly Stats	3/8/2006 RCC299 Assembly		1/12/2007 CCMP1545 Assembly	
	Initial	**Filtered**	**Initial**	**Filtered**
Scaffold Total	521	284	217	69
Scaffold Sequence Total	22.8 MB	22.6 MB	22.1 MB	22.0 MB
Scaffold N50	8	8	8	8
Scaffold L50	1.3 MB	1.3 MB	1.2 MB	1.2 MB
Contig Total	894	629	815	644
Contig Sequence Total	21.5 MB (6.0% gap)	21.2 MB (6.0% gap)	21.7 MB (1.6% gap)	21.6 MB (1.5% gap)
Contig N50	39	38	84	83
Contig L50	154 KB	158 KB	78.6 KB	79.1 KB
Estimated Depth	8.86 ± 0.08		7.96 ± 0.07	
Data Completeness (>20% Covered)	99.0%		98.8%	
Data Completeness (>50% Covered)	98.7%		98.5%	
Data Completeness (>80% Covered)	98.0%		98.2%	
Scaffold Completeness	98.2%		98.1%	

SUPPORTING REFERENCES

1. R. R. L. Guillard, in *Culture of Marine Invertebrate Animals,* W. L. Smith, M. H. Chanley, Eds. (Plenum, New York, 1975), pp.29-60.
2. A. Z. Worden *et al.*, *Environ. Microbiol.* **8**, 21 (2006).
3. J. R. Mead, M. J. Arrowood, W. L. Current, C. R. Sterling, *J. Parasitol.* **74**, 366 (1988).
4. W. Eikrem, Ø. Moestrup, *Phycologia* **37**, 132 (1998).
5. B. Winnepenninckx, T. Backeljau, R. De Wachter, *Trends Genet.* **9**, 407 (1993).
6. J. Chapman, N. Putnam, I. Ho, D. Rokhsar, (unpublished).
7. S. Aparicio *et al.*, *Science* **297**, 1301 (2002).
8. D. Gordon, C. Abaijian, P. Green, *Genome Res.* **8**, 195 (1988).
9. D. B. Jaffe *et al.*, *Genome Res.* **13**, 91 (2003).
10. B. Ewing, P. Green, *Genome Res.* **8**, 186 (1998).
11. B. Ewing, L. Hillier, M. C. Wendl, P. Green, *Genome Res.* **8**, 175 (1998).
12. T. F. Smith, M. S. Waterman, *J. Theor. Biol.* **91**, 379 (1981).
13. A. A. Salamov, V. V. Solovyev, *Genome Res.* **10**, 516 (2000).
14. E. Birney, R. Durbin, *Genome Res.* **10**, 547 (2000).
15. J. Venter *et al.*, *Science* **304**, 66 (2004).
16. M. Ashburner *et al.*, *Nature Genet.* **25**, 25 (2000).
17. E. V. Koonin *et al.*, *Genome Biol.* **5**, R7 (2004).
18. M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, M. Hattori, *Nucleic Acids Res.* **32**, D277 (2004).
19. E. M. Zdobnov, R. Apweiler, *Bioinformatics* **17**, 847 (2001).
20. C. Simillion, K. Vandepoele, Y. Saeys, Y. Van de Peer, *Genome Res.* **14**, 1095 (2004).
21. T. Lassmann, E. L. Sonnhammer, *BMC Bioinformatics* **6**, 298 (2005).
22. A. Stamatakis, *Bioinformatics* **22**, 2688 (2006).
23. S. Guindon, O. Gascuel, *Syst. Biol.* **52**, 696 (2003).
24. R. C. Edgar, *Nucleic Acids Res.* **32**, 1792 (2004).
25. S. Kumar, K. Tamura, M. Nei, *Brief Bioinformatics* **5**, 150 (2004).
26. G. V. Kryukov *et al.*, *Science* **300**, 1439 (2003).
27. Q. Ren, K. Chen, I. T. Paulsen, *Nucleic Acids Res.* **35**, D274 (2007).
28. A. N. Zelensky, J. E. Gready, *Proteins* **52**, 466 (2003).
29. S. R. Eddy, *Bioinformatics* **14**, 755 (1998).
30. E. Birney, M. Clamp, R. Durbin, *Genome Res.* **14**, 988 (2004).
31. S. Griffiths-Jones *et al.*, *Nucleic Acids Res.* **33**, D121 (2005).
32. E. L. L. Sonnhammer, R. Durbin, *Gene* **167,** GC1 (1995).
33. J. Jurka *et al.*, *Cytogentic and Genome Res.* **110**, 462 (2005).
34. K. Howe, A. Bateman, R. Durbin, *Bioinformatics* **18**, 1546 (2002).
35. E. Parzen, *Ann. Math. Stat.* **33**, 1065 (1962).
36. E. Derelle *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 11647 (2006).
37. B. Palenik *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 7705 (2007).
38. S. Robbens *et al.*, *Mol. Biol. Evol.* **24**, 956 (2007).
39. Z. G. Lin, H. Z. Kong, M. Nei, H. Ma, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 10328 (2006).
40. A. Reyes-Prieto, D. Bhattacharya, *Mol Biol Evol* **24**, 2358 (Nov, 2007).
41. F. Ronquist, J. P. Huelsenbeck, *Bioinformatics* **19**, 1572 (2003).

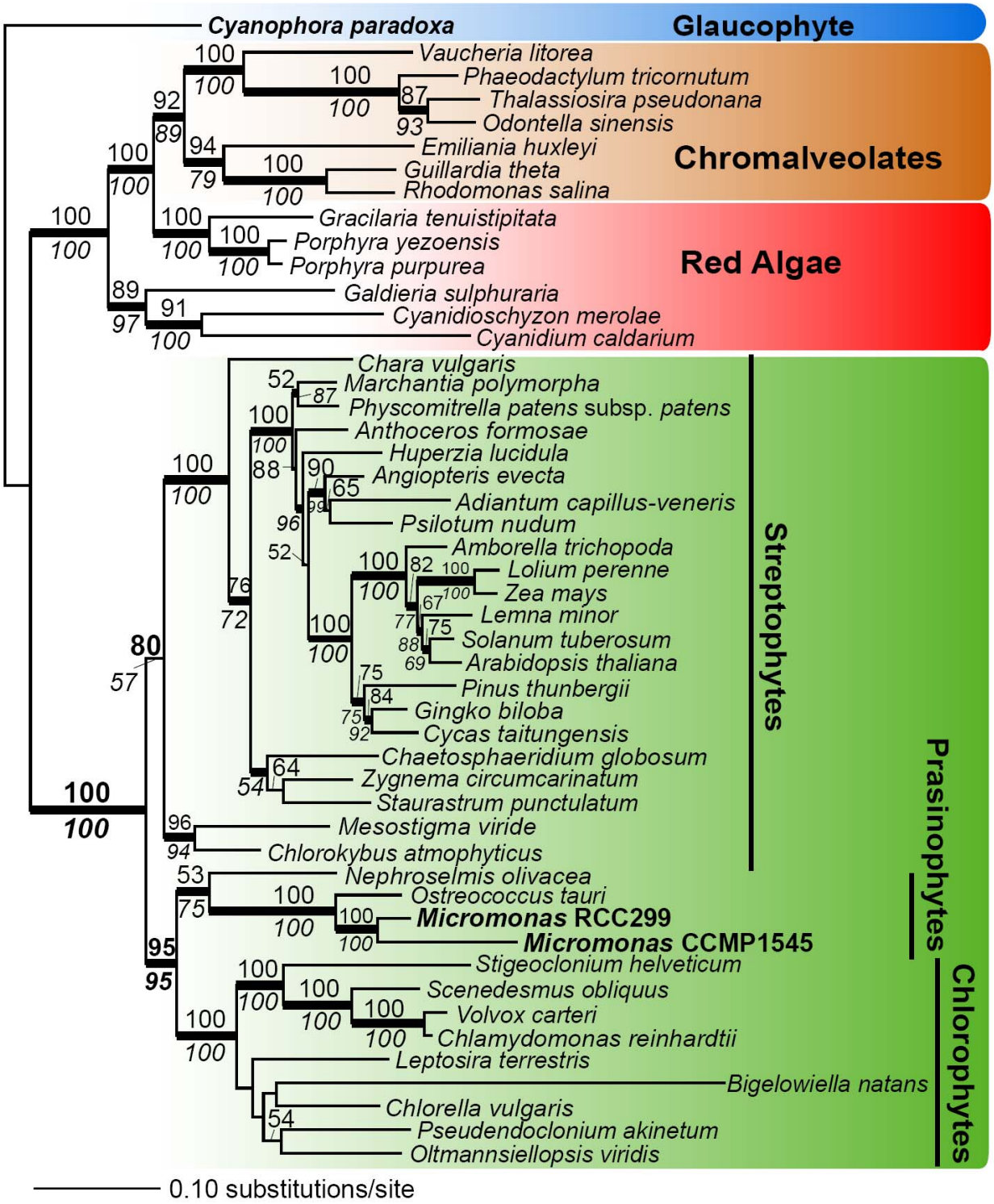
42. C. Six, A. Z. Worden, F. Rodriguez, H. Moreau, F. Partensky, *Mol. Biol. Evol*. **22**, 2217 (2005).
43. P. Deschamps, H. Moreau, A. Z. Worden, D. Dauvillee, S. G. Ball, *Genetics* **178**, 2373 (2008).
44. A. V. Lobanov *et al.*, *Genome Biol.* **8**, R198 (2007).
45. Y. Zhang, V. N. Gladyshev, *PLoS Genet*. **4**, e1000095 (2008).
46. Y. J. Park, K. Luger, *Biochem. Cell Biol.* **84**, 549 (2006).
47. E. Ramirez-Parra, C. Gutierrez, *Trends Plant Sci.* **12**, 570 (2007).
48. D. Ray-Gallet *et al.*, *Mol. Cell.* **9**, 1091 (2002).
49. M. Duroux, A. Houben, K. Ruzicka, J. Friml, K. D. Grasser, *Plant J.* **40**, 660 (2004).
50. F. Mousson, F. Ochsenbein, C. Mann, *Chromosoma* **116**, 79 (Apr, 2007).
51. B. R. Cairns, *Nat. Struct. Mol. Biol.* **14**, 989 (2007).
52. Y. Bao, X. Shen, *Mutat. Res.* **618**, 18 (2007).
53. H. van Attikum, O. Fritsch, S. M. Gasser, *EMBO J.* **26**, 4113 (2007).
54. A. Gribun, K. L. Cheung, J. Huen, J. Ortega, W. A. Houry, *J. Mol. Biol.* **376**, 1320 (2008).
55. J. A. Martens, F. Winston, *Curr. Opin. Genet. Dev.* **13**, 136 (2003).
56. T. Kouzarides, *Cell* **128**, 693 (2007).
57. M. T. Bedford, S. Richard, *Mol. Cell.* **18**, 263 (2005).
58. T. Dalmay, R. Horsefield, T. H. Braunstein, D. C. Baulcombe, *EMBO J.* **20**, 2069 (2001).
59. Y. Yin *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 10191 (2002).
60. F. Hartung *et al.*, *Curr. Biol.* **12**, 1787 (2002).
61. K. Sugimoto-Shirasu *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **102**, 18736 (2005).
62. K. Sugimoto-Shirasu, N. J. Stacey, J. Corsar, K. Roberts, M. C. McCann, *Curr. Biol.* **12**, 1782 (2002).
63. S. Keeney, C. N. Giroux, N. Kleckner, *Cell* **88**, 375 (1997).
64. M. Grelon, D. Vezon, G. Gendrot, G. Pelletier, *EMBO J.* **20**, 589 (2001).
65. N. J. Stacey *et al.*, *Plant J.* **48**, 206 (2006).
66. S. B. Malik, M. A. Ramesh, A. M. Hulstrand, J. M. Logsdon, Jr., *Mol. Biol. Evol.* **24**, 2827 (2007).
67. P. Sung, L. Krejci, S. Van Komen, M. G. Sehorn, *J. Biol. Chem.y* **278**, 42729 (2003).
68. J. Y. Bleuyard, M. E. Gallego, F. Savigny, C. I. White, *Plant J.l* **41**, 533 (2005).
69. J. Y. Bleuyard, M. E. Gallego, C. I. White, *DNA Repair (Amst)* **5**, 1 (2006).
70. J. Thacker, *Cancer Letters* **219**, 125 (2005).
71. W. X. Li *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 10596 (2004).
72. T. Snowden, S. Acharya, C. Butz, M. Berardini, R. Fishel, *Mol.r Cell.* **15**, 437 (2004).
73. J. D. Higgins, S. J. Armstrong, F. C. H. Franklin, G. H. Jones, *Genes & Dev.***18**, 2557 (2004).
74. J. C. Connelly, D. R. Leach, *Trends Biochem. Sci.* **27**, 410 (2002).
75. J. M. Henry *et al.*, *Mol. Cell. Biol.* **26**, 2913 (2006).
76. G. V. Petukhova *et al.*, *Nat. Struct. Mol. Biol.* **12**, 449 (2005).
77. S. Domenichini, C. Raynaud, D. A. Ni, Y. Henry, C. Bergounioux, *DNA Repair (Amst)* **5**, 455 (2006).
78. C. Kerzendorfer *et al.*, *J. Cell. Sci.* **119**, 2486 (2006).
79. E. Dray, N. Siaud, E. Dubois, M. P. Doutriaux, *Plant Physiol.* **140**, 1059 (2006).
80. W. D. Heyer, X. Li, M. Rolfsmeier, X. P. Zhang, *Nucleic Acids Res.* **34**, 4115 (2006).

81. S. J. Armstrong, A. P. Caryl, G. H. Jones, F. C. Franklin, *J. Cell. Sci.* **115**, 3645 (2002).
82. L. Chelysheva *et al.*, *J. Cell. Sci.* **118**, 4621 (2005).
83. T. S. Kitajima, S. A. Kawashima, Y. Watanabe, *Nature* **427**, 510 (2004).
84. R. R. Iyer, A. Pluciennik, V. Burdett, P. L. Modrich, *Chem. Rev.* **106**, 302 (2006).
85. R. V. Abdelnoor *et al.*, *J. Mol. Evol.* **63**, 165 (2006).
86. R. V. Abdelnoor *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 5968 (2003).
87. H. Lin, U. W. Goodenough, *Genetics* **176**, 913 (2007).
88. P. J. Ferris, U. W. Goodenough, *Genetics* **146**, 859 (1997).
89. T. Hamaji *et al.*, *Genetics* **178**, 283 (2008).
90. H. Nozaki, T. Mori, O. Misumi, S. Matsunaga, T. Kuroiwa, *Curr. Biol.* **16**, R1018 (2006).
91. P. J. Ferris, E. V. Armbrust, U. W. Goodenough, *Genetics* **160**, 181 (2002).
92. K. B. Lengeler *et al.*, *Eukaryot. Cell* **1**, 704 (2002).
93. Y. P. Hsueh, A. Idnurm, J. Heitman, *PLoS Genet.* **2**, e184 (2006).
94. P. J. Ferris *et al.*, *Plant Cell* **17**, 597 (2005).
95. J. H. Lee, S. Waffenschmidt, L. Small, U. Goodenough, *Plant Physiol.* **144**, 1813 (2007).
96. G. I. Cassab, *Annu. Re.v Plant Physiol. Plant Mol. Biol.* **49**, 281 (1998).
97. D. F. Seals, S. A. Courtneidge, *Genes Dev.* **17**, 7 (2003).
98. G. J. Seifert, K. Roberts, *Annu. Rev. Plant Biol.* **58**, 137 (2007).
99. J. Egelund *et al.*, *Plant Mol. Biol.* **64**, 439 (2007).
100. P. M. Coutinho, M. Stam, E. Blanc, B. Henrissat, *Trends Plant Sci.* **8**, 563 (2003).
101. G. I. McFadden, *Curr. Opin. Plant Biol.* **2**, 513 (1999).
102. T. Cavalier-Smith, *Trends Plant Sci.* **5**, 174 (2000).
103. A. B. Boraston, D. N. Bolam, H. J. Gilbert, G. J. Davies, *Biochem. J.* **382**, 769 (2004).
104. H. Aspeborg *et al.*, *Plant Physiol.* **137**, 983 (2005).
105. S. R. Wessler, *Trends Plant Sci.* **10**, 54 (2005).
106. S. K. Floyd, J. L. Bowman, *Int.l J. Plant Sci.* **168**, 1 (2007).
107. R. Derelle, P. Lopez, H. Le Guyader, M. Manuel, *Evol. Dev.* **9**, 212 (2007).
108. N. King *et al.*, *Nature* **451**, 783 (2008).
109. S. A. Rensing *et al.*, *Science* **319**, 64 (2008).
110. J. H. Lee, H. Lin, S. Joo, U. Goodenough, *Cell* **133**, 829 (2008).
111. S. Hake *et al.*, *Annu. Rev. Cell. Dev. Biol.* **20**, 125 (2004).
112. G. Pazour, N. Agrin, J. Leszyk, G. Witman, *J. Cell Biol.* **170**, 103 (2005).
113. C. Courties *et al.*, *Nature* **370**, 255 (1994).
114. A. Z. Worden, F. Not, in *Microbial Ecology of the Oceans,* D. L. Kirchman, Ed. (Wiley, Hoboken, 2008), pp. 594.
115. M. Pirner, R. Linck, *J. Biol. Chem.* **269**, 31800 (1994).
116. Q. Wang, J. Pan, W. J. Snell, *Cell* **125**, 549 (2006).
117. C. van den Hoek, D. Mann, H. Jahns, *Algae : An Introduction to Phycology*. (Cambridge Univ. Press, Cambridge, 1995).
118. P. Karatza, P. Panos, E. Georgopoulou, S. Frillingos, *J. Biol. Chem.* **281**, 39881 (2006).
119. A. Pantazopoulou *et al.*, *Fungal Genet. Biol.* **44**, 627 (2007).
120. Y. F. Tsay, C. C. Chiu, C. B. Tsai, C. H. Ho, P. K. Hsu, *FEBS Lett.* **581**, 2290 (2007).
121. C. Brussaard, *J. Eukaryot. Microbiol.* **51**, 125 (2004).
122. A. N. Zelensky, J. E. Gready, *Febs J.* **272**, 6179 (2005).
123. S. A. Wilks, M. A. Sleigh, *Microb. Ecol.* **36**, 165 (1998).

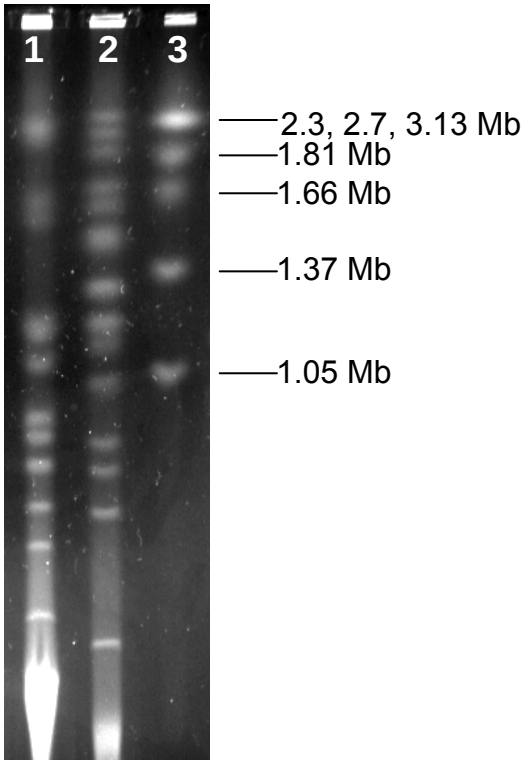
124. E. C. Roberts, M. V. Zubkov, M. Martin-Cereceda, G. Novarino, E. C. Wootton, *FEMS Microbiol. Lett.* **265**, 202 (2006).
125. J. Staunton, K. Weissman, *Nat. Prod. Rep.* **18**, 380 (2001).
126. G. Zhu *et al.*, *Gene* **298**, 79 (2002).
127. U. John *et al.*, *Protist* **159**, 21 (2008).
128. D. A. Hopwood, D. H. Sherman, *Annu. Rev. Genet.* **24**, 37 (1990).
129. F. Gross *et al.*, *Arch. Microbiol.* **185**, 28 (2006).
130. H. Jenke-Kodama, A. Sandmann, R. Muller, E. Dittmann, *Mol. Biol. Evol.* **22**, 2027 (2005).
131. S. Kroken, N. L. Glass, J. W. Taylor, O. C. Yoder, B. G. Turgeon, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 15670 (2003).
132. D. Bhattacharya, L. Medlin, *Plant Physiol.* **116,** 9 (1998).
133. P. J. Keeling *et al.*, *Trends Ecol. Evol.* **20**, 670 (2006).
134. M. Giordano, J. Beardall, J. A. Raven, *Annu. Rev. Plant Biol.* **56,** 99 (2005).
135. M. D. Iglesias-Rodriguez, N. A. Nimer, M. J. Merrett, *New Phytol.* **140**, 685 (1998).
136. J. Karlsson *et al.*, *EMBO J.* **17**, 1208 (1998).
137. D. T. Hanson, L. A. Franklin, G. Samuelsson, M. R. Badger, *Plant Physiol.* **132**, 2267 (2003).
138. M. R. Sawaya *et al.*, *J. Biol. Chem.* **281**, 7546 (2006).
139. C. E. Lane, J. M. Archibald, *Trends Ecol. Evol.* **23**, 268 (2008).
140. F. Wolfe-Simon, D. Grzebyk, O. Schofield, P. G. Falkowski, *J. Phycol.* **41,** 453 (2005).
141. A. J. Alverson, J. K. Jansen, E. C. Theriot, *Mol. Phylogenet. Evol.* **45**, 193 (2007).
142. E. Grill, E. L. Winnacker, M. H. Zenk, *Science* **230**, 674 (1985).
143. L. Wei, B. A. Ahner, *Limnol. Oceanogr.* **50**, 13 (2005).
144. D. B. Rusch *et al.*, *PLoS Biol.* **5**, e77 (2007).
145. A. Wachter *et al.*, *Plant Cell* **19**, 3437 (2007).
146. M. T. Cheah, A. Wachter, N. Sudarsan, R. R. Breaker, *Nature* **447**, 497 (2007).
147. M. T. Croft, M. Moulin, M. E. Webb, A. G. Smith, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 20770 (2007).
148. D. DellaPenna, R. L. Last, *Science* **320**, 479 (2008).
149. M. T. Croft, M. J. Warren, A. G. Smith, *Eukaryot. Cell* **5**, 1175 (2006).
150. A. H. Jenkins, G. Schyns, S. Potot, G. Sun, T. P. Begley, *Nat. Chem. Biol.* **3**, 492 (2007).
151. M. Onozuka, H. Konno, Y. Kawasaki, K. Akaji, K. Nosaka, *FEMS Yeast Res.* **8**, 266 (2008).
152. V. Ganapathy, S. B. Smith, P. D. Prasad, *Pflugers Arch.* **447**, 641 (2004).

SUPPLEMENTARY FIGURES S1-S22 on following pages

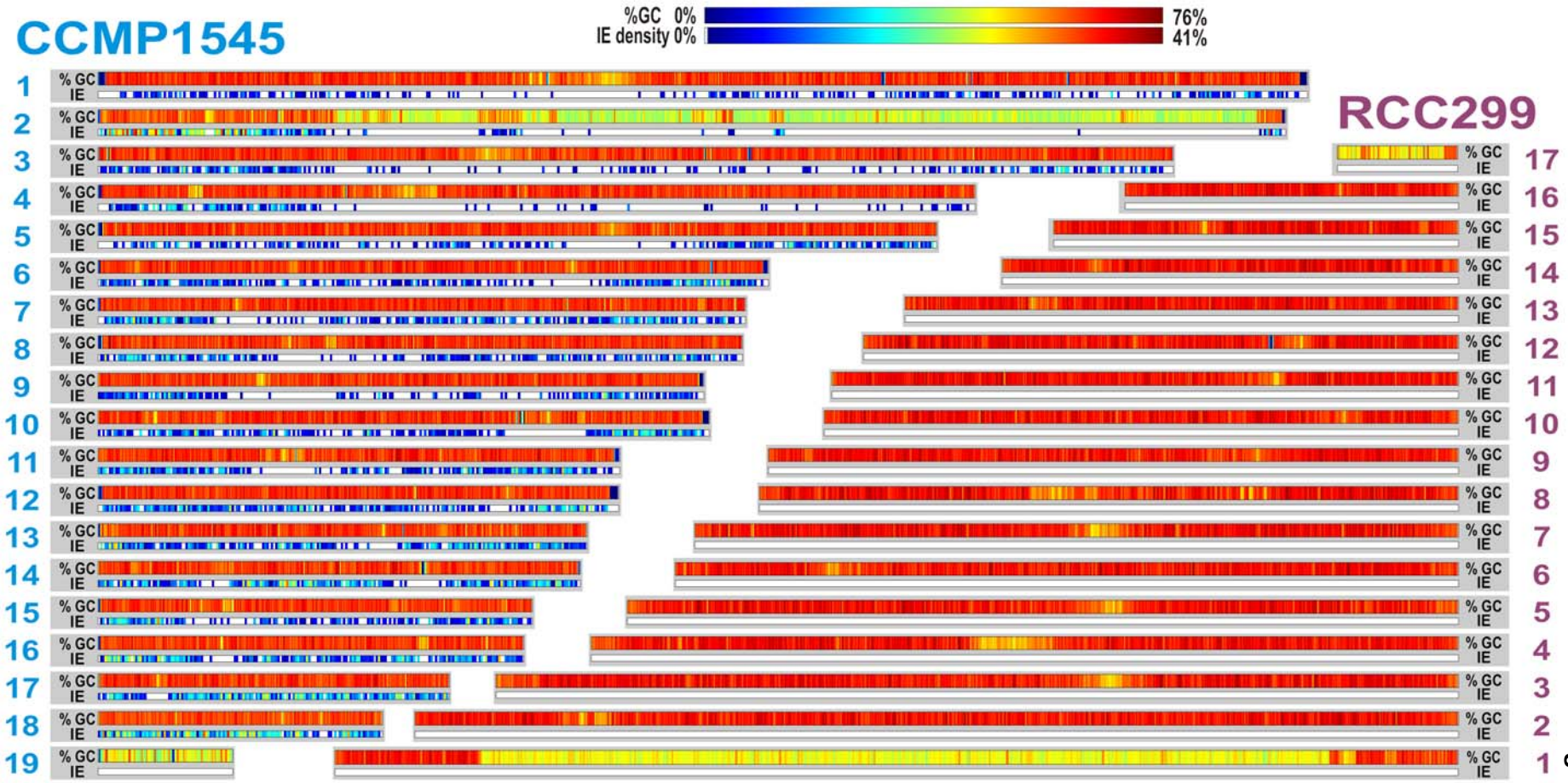
Figure S1, Worden et al.



Supplementary Figure 1. Maximum likelihood tree of of chloroplast genome-encoded proteins. This tree is inferred from a concatenated alignment of 6 conserved plastid-encoded proteins (psaA, psaB, psbA, psbB, psbC, atpA, atpB) (3,697 aa). The results of a bootstrap analysis using RAxML are shown above the branches and PHYML bootstrap values are shown below in italic text. Only bootstrap values 50% or higher are indicated. The thick branches received a posterior probability = 1.0 in a Bayesian phylogenetic inference (MrBayes v.3.1.2). The RAxML, PHYML, and Bayesian analyses used the CPREV model of protein evolution. Branch lengths are proportional to the number of substitutions per site (see scale bar). The *Micromonas* species studied here are shown in large text. This tree was rooted on the branch leading to the early-diverging glaucophyte *Cyanophora paradoxa.*

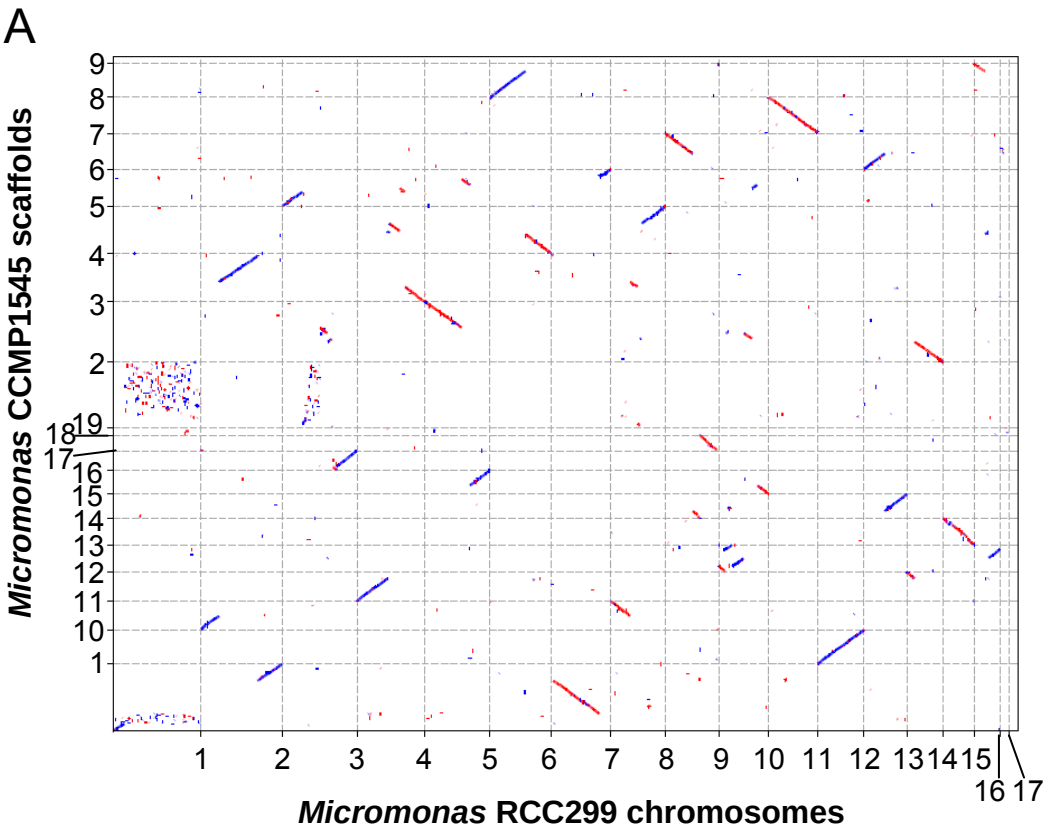


Supplementary Figure 2. Pulsed Field Gel Electrophoresis (PFGE) of the *Micromonas* strains. Lane1: *Micromonas* CCMP1545; lane 2: *Micromonas* RCC299; lane 3: *Hansenula wingei* (yeast, size marker). Chromosome numbers match JGI assemblies as does the estimated genome size. The comparison between the 2 karyotypes shows that even with 2 more chromosomes for CCMP1545, the genome size of the 2 species are similar. This is due to the fact that RCC299 has 13 chromosomes distributed among 1 to 2.1 Mb and CCMP1545 has only 10 chromosomes for the same range. Furthermore, CCMP1545 chromosomes are generally smaller. The gel was stained with ethidium bromide.



Supplementary Figure 3. The chromosomes of *Micromonas* CCMP1545 and RCC299 %GC content. %GC was calculated using a sliding window of 2000 nt. Introner (IE) repeat sequence (see 'genome invasion' section) density is also shown and reflects mapping to the chromosomes and then, based on the coordinates returned, the proportion of sequence being IE, compared to the proportion not being IE, was calculated also using a sliding window (2000 nt).

Figure S4a, Worden et al.
Legend Figure S4



Supplementary Figure 4. **Whole genome DNA alignments between genomes.** (a-c) VISTA dotplots (Couronne et al. 2002) representing whole-genome DNA alignment between pairs of genomes: (a) *Micromonas* RCC299 vs. CCMP1545; (b) *O. lucimarinus* vs. *O. tauri*; (c) RCC299 vs *O. lucimarinus*. Chromosomes and scaffolds are fragments of X or Y axis and aligned regions are shown in blue or red (if inverted). The plots illustrate the much lower synteny between the *Micromonas* species (a) than seen between the *Ostreococcus* species (b). Note the high level of 'reshuffling' between the low GC-regions (Chrom 1 RCC299 and Chrom 2 CCMP1545) as well as for the *Ostreococcus* species. (d) shows the percent of CCMP1545, *O. tauri* and *O. lucimarinus* genomic DNA aligned to RCC299 while (e) shows RCC299, *O. tauri* and *O. lucimarinus* genomic DNA aligned against CCMP1545.

Figure S4b,c, Worden et al.

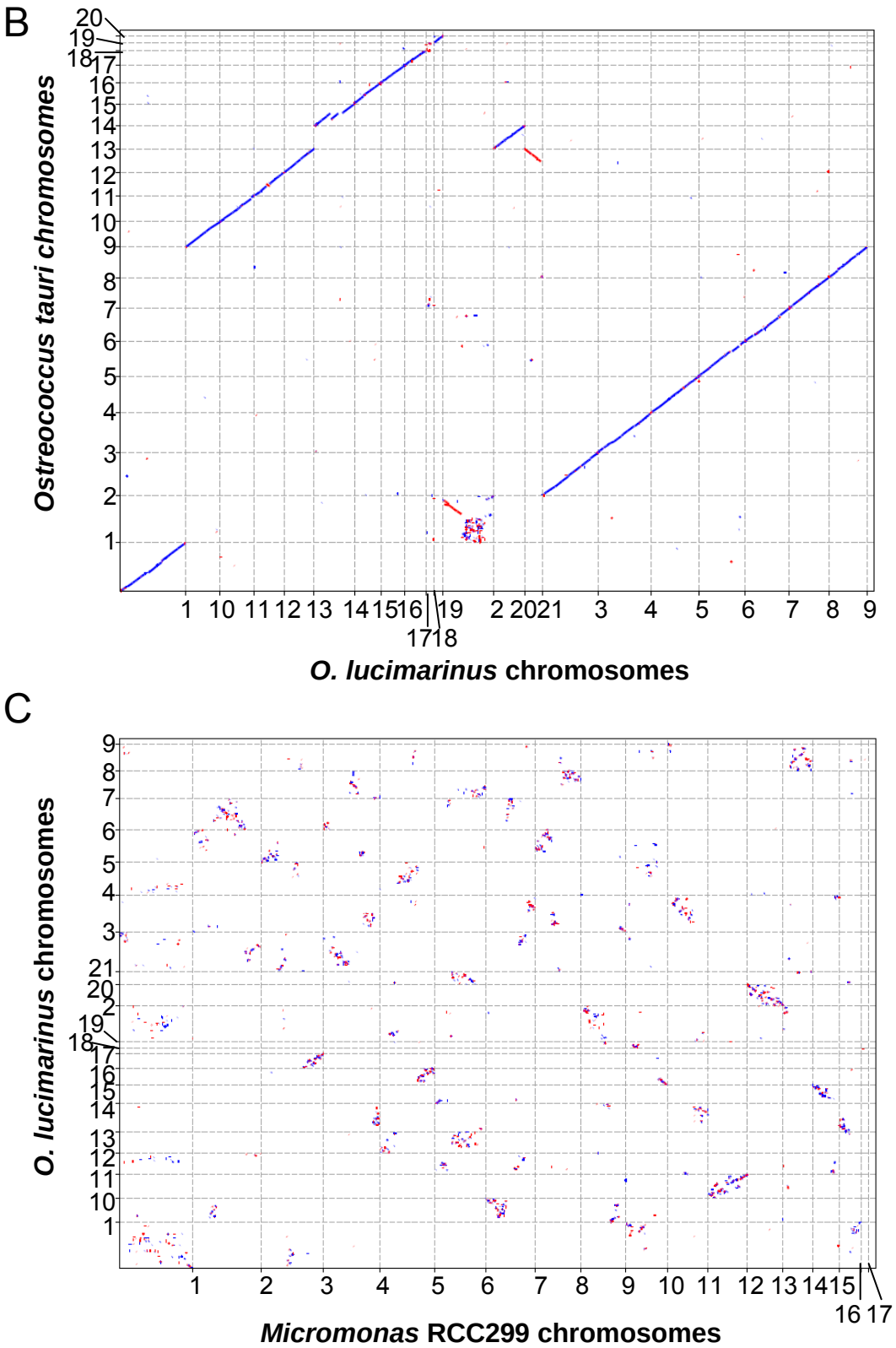
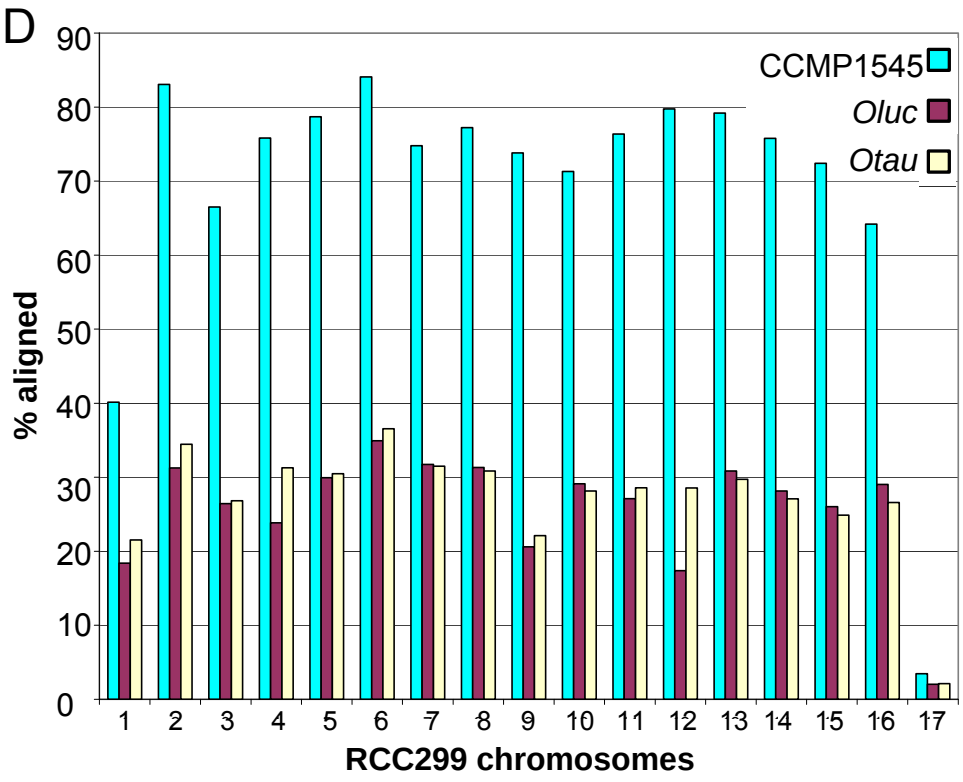
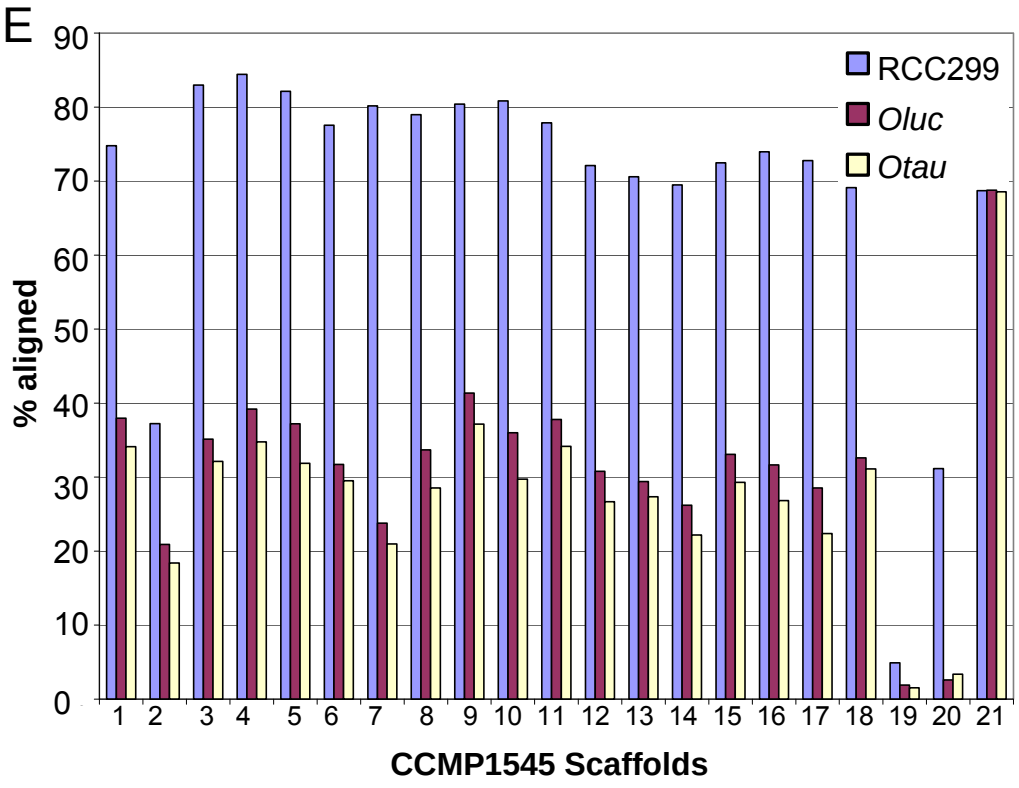
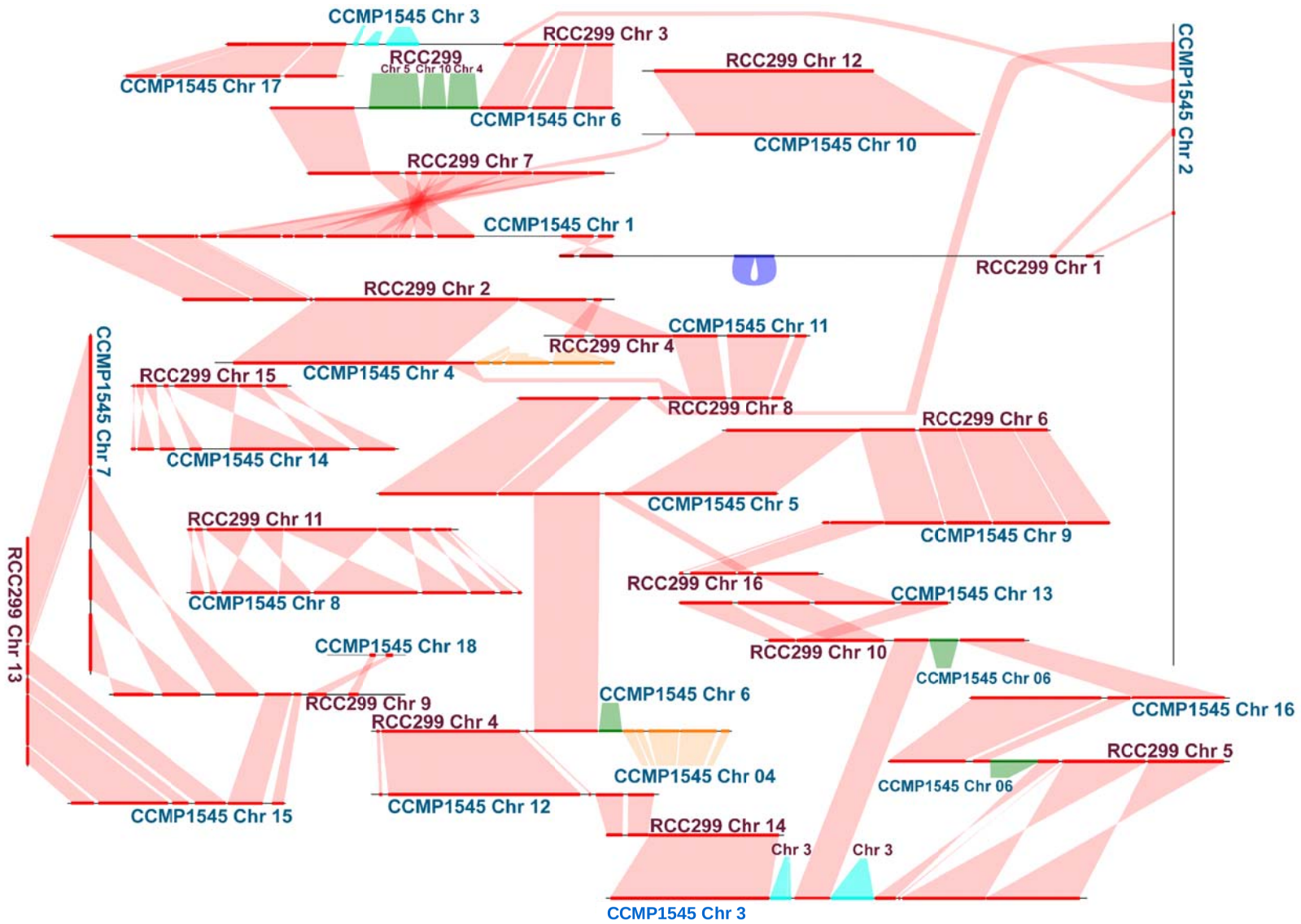


Figure S4d,e, Worden et al.

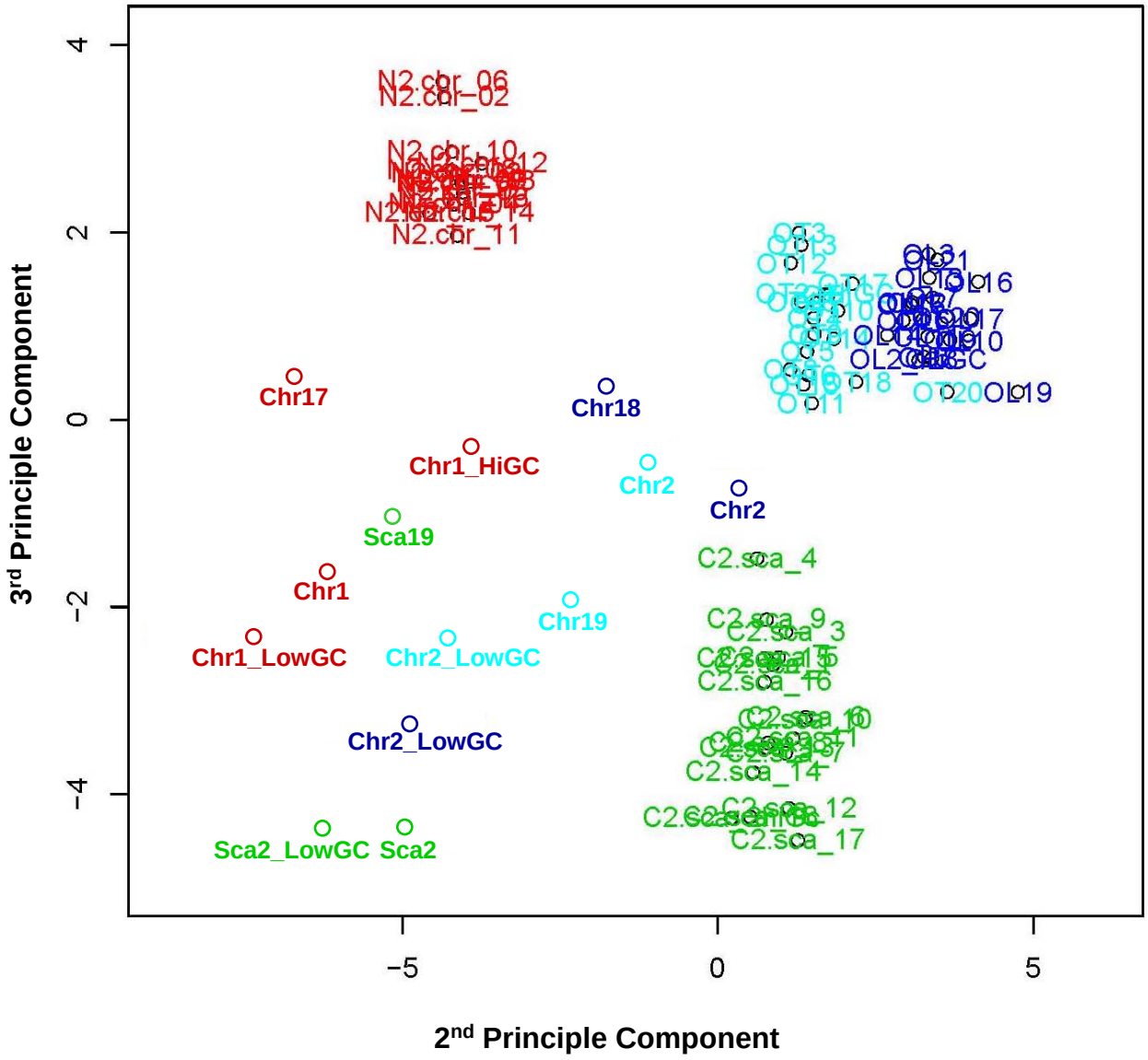






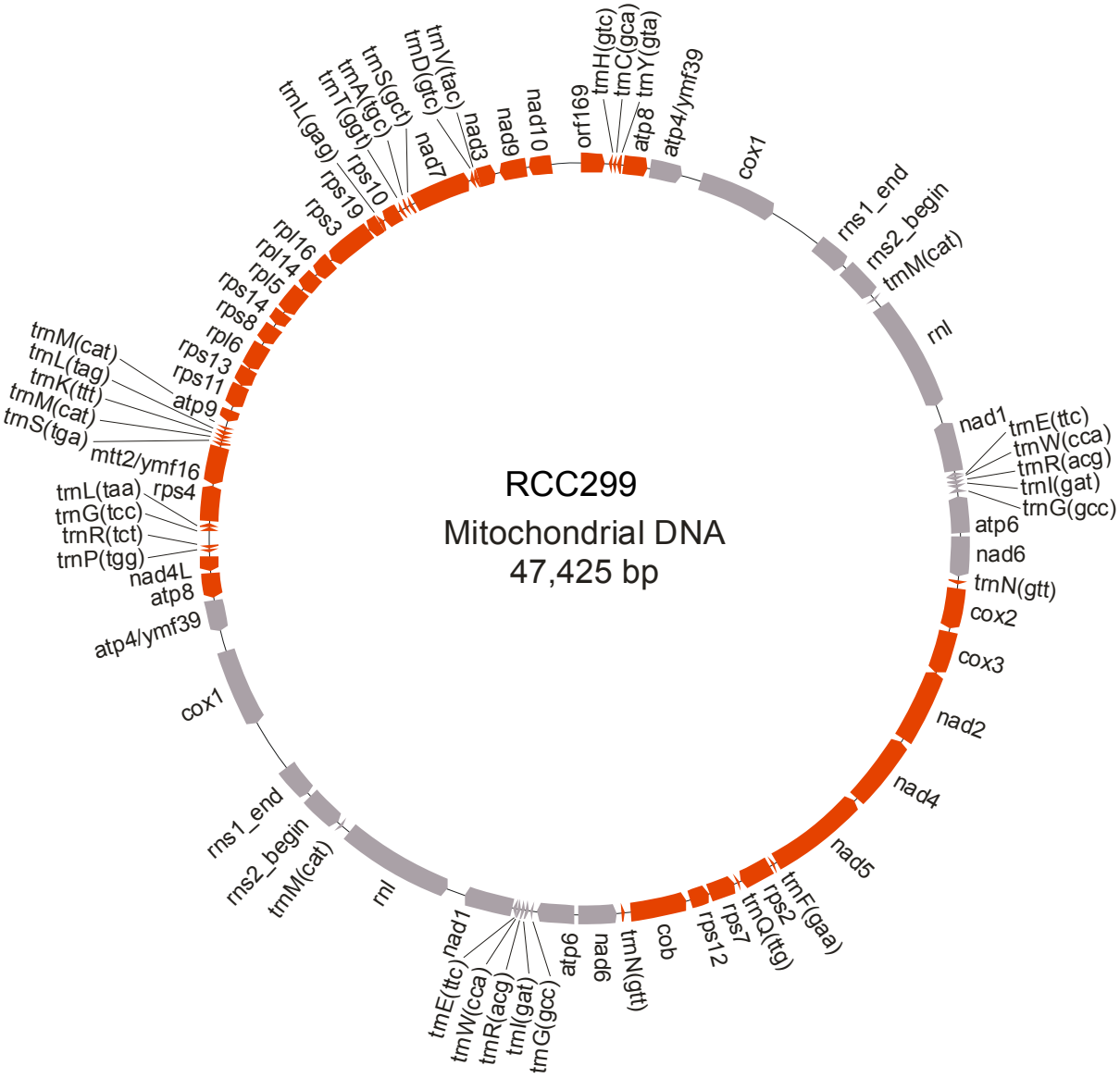
Supplementary Figure 5. Synteny between *Micromonas* RCC299 and *Micromonas* CCMP1545. Depicted areas in red show collinear regions (conserved gene order and content). Blocks of different colors denote different sorts of duplications: blue, an internally duplicated segment; green, a duplicated segment from RCC299 that is collinear with a segment on a different chromosome in CCMP1545; and vice versa in yellow. Chromosomes are shown from the 5' to 3' direction, resulting in "twisting" of some syntenous fragments. Not shown are the smallest two chromosomes within each genome, neither of which bore synteny with any other genome regions in the comparison species.

Figure S6, Worden et al.



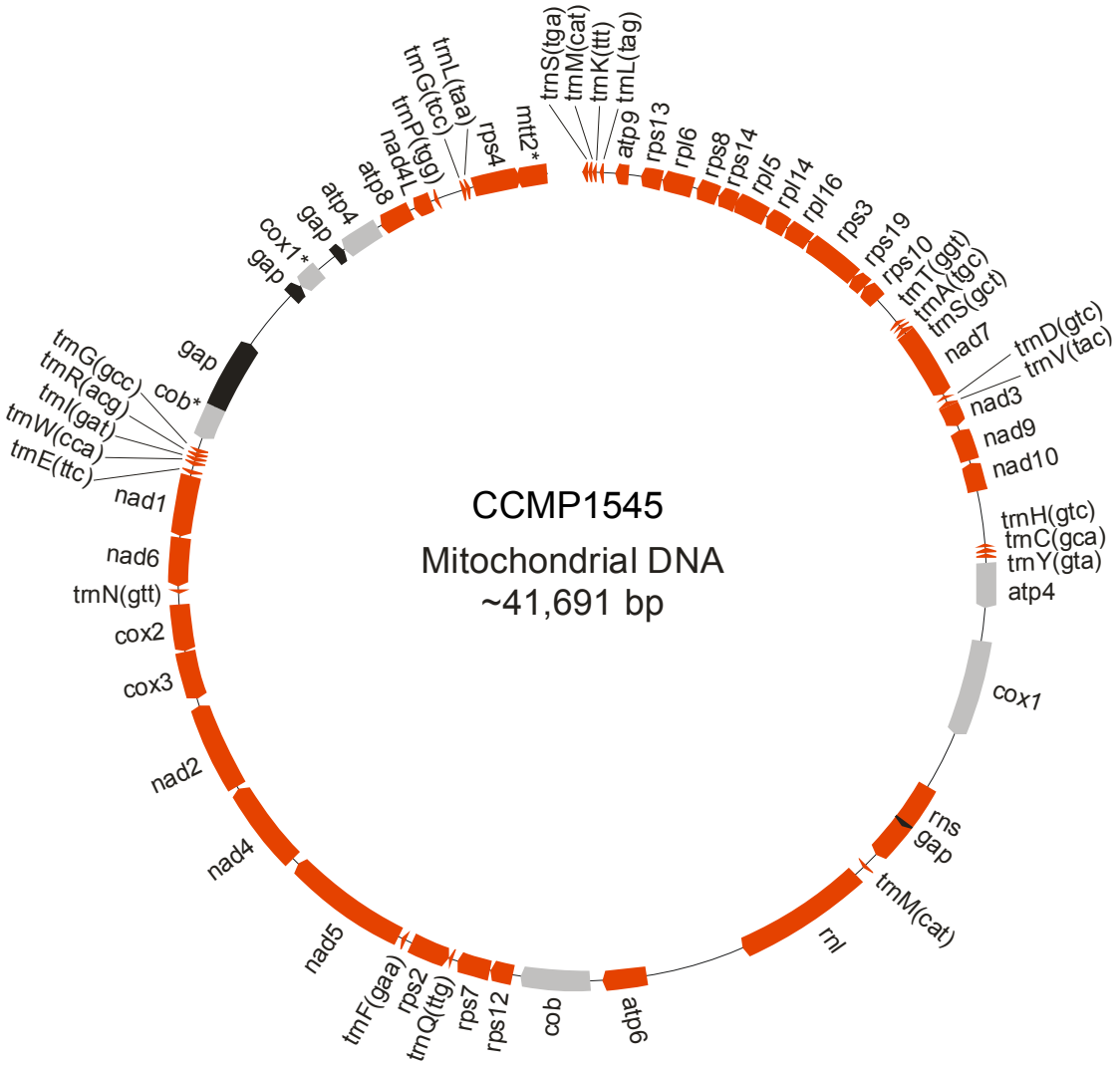
Supplementary Figure 6. Principle components analysis of codon frequencies. Data are shown for *Micromonas* CCMP1545 (green), *Micromonas* RCC299 (red), *O. tauri* (aqua) and *O. lucimarinus* (blue) including normal GC- (most chromosomes, the different colored 'clouds'), low GC-regions (demarked LowGC) and smallest chromosomes (i.e., RCC299, Chr17; CCMP1545, Chr19; *O. tauri*, Chr19; *O. lucimarinus*, Chr18). Numbers at end of text labels indicate the chromosome or scaffold represented.

Figure S7a, Worden et al.



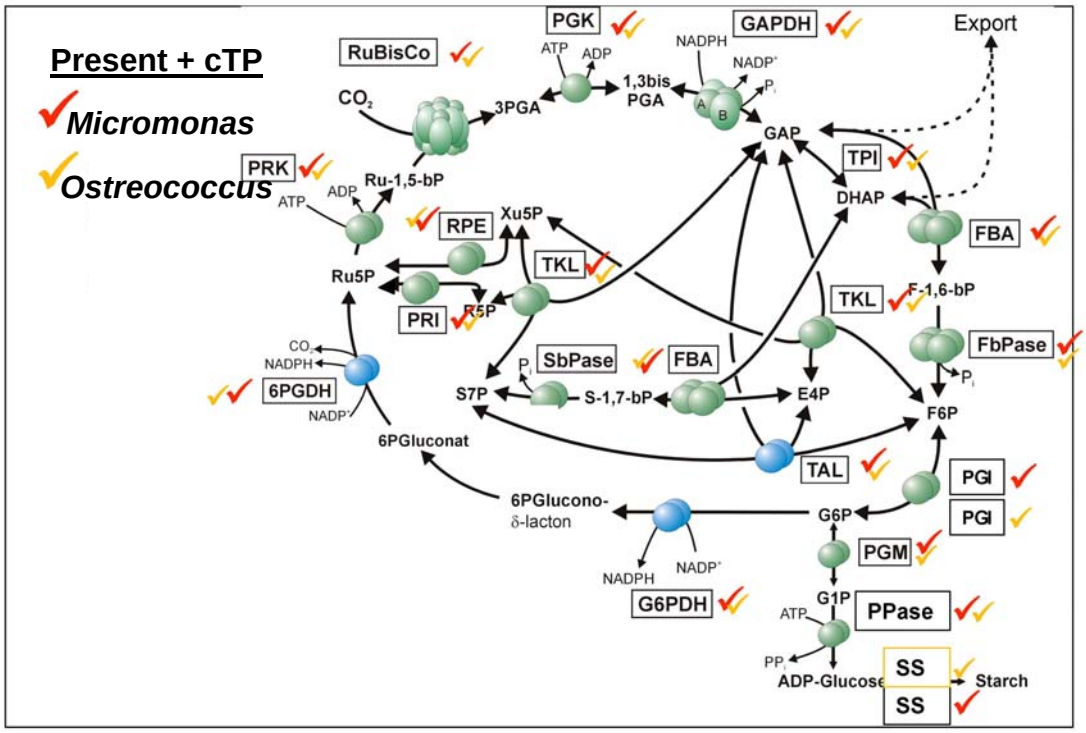
Supplementary Figure 7a. The complete mitochondrial genome of *Micromonas* RCC299. Red blocks: unduplicated genes; grey blocks: duplicated genes.

Figure S7b, Worden et al.



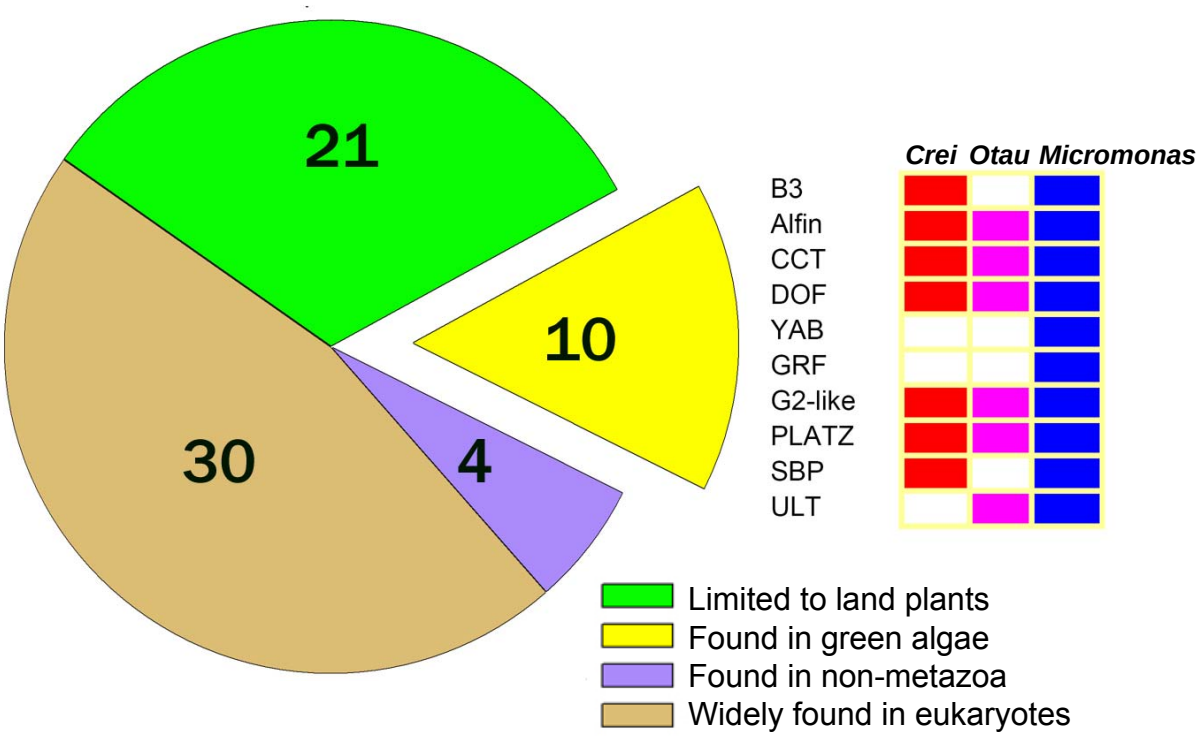
Supplementary Figure 7b. Mitochondrial genome of *Micromonas* CCMP1545. Red blocks: unduplicated genes; grey blocks: duplicated genes; black blocks: gaps; *Incomplete genes. This assembly is probably faulty as there are many gaps and the duplication seen in the completely assembled mitochondrial genome of RCC299 is only partially present.

Figure S8, Worden et al.



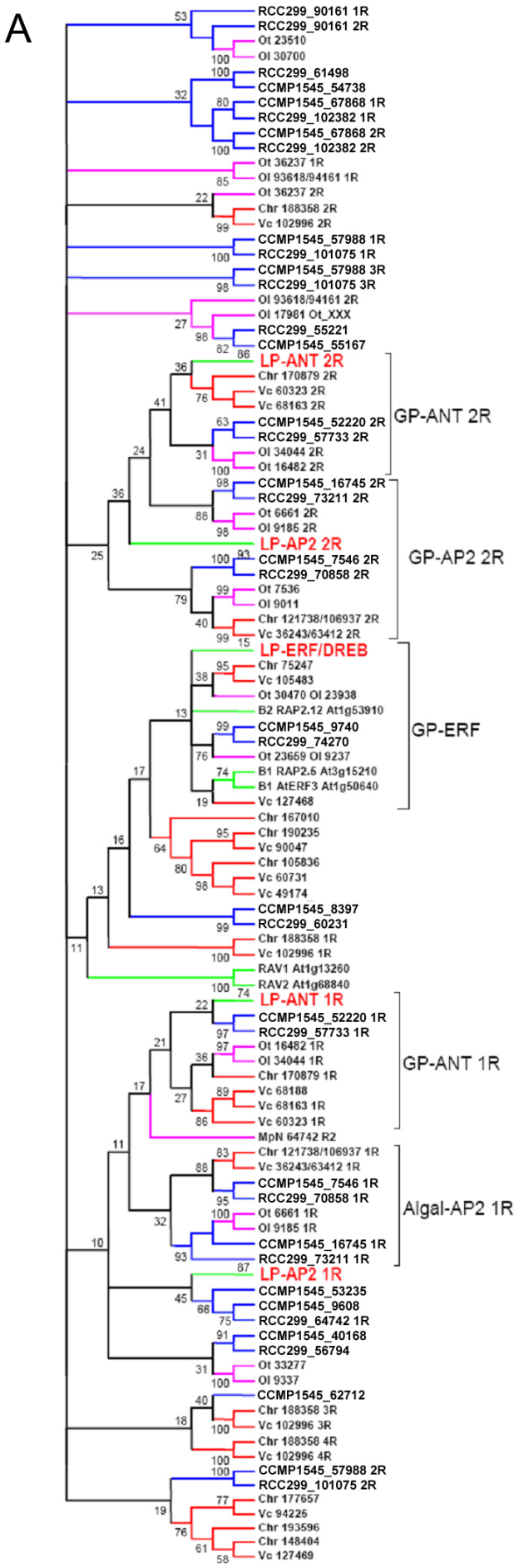
Supplementary Figure 8. Carbon fixation (Calvin Cycle). All known enzymes for a functional active Calvin Cycle were identified in the genomes of RCC299 and CCMP1545 as well as in *Ostreococcus*. Differences between species entailed gene copy numbers and the occurrence of gene fusions, but there were not differences in protein targeting. At least one homolog from each calvin-cycle enzyme has a precurser sequence for plastid targeting. However, the redundancy of these genes was different. For instance, RCC299 has 4 different FBPases encoded in the genome but only 3 were identified in CCMP1545. In contrast to other green algae, such as *C. reinhardtii*, or to diatoms, some of the calvin-cycle genes are characterized by fusion with neighbor genes that are located upstream of the calvin-cycle genes potentially due to genome compaction. For instance, the plastid targeted FBPase in RCC299 (ProtID 56498) is fused with an orf upstream that has a conserved domain (DUF) without known function. Interestingly, the precursor sequence for plastid targeting is upstream of the fused orf. Gene fusion in this case will possibly introduce novel proteins into the plastid. The same fusion construct is present in CCMP1545 and *O. tauri*. Another example for gene fusion is a ribulosephosphate 3-epimerase (RPE) in RCC299 (Prot. ID 96976). In RCC299 this RPE is fused with an upstream sugar/xylulose kinase; the same fusion is found in all four Mamiellales genomes.

Figure S9, Worden et al.



Supplementary Figure 9. **Distribution of 65 plant transcription factors.** We analyzed transcription factor families using information derived from two databases, DATF (http://datf.cbi.pku.edu.cn; Guo et al., 2005) and PlnTFDB (http://plntfdb.bio.uni-potsdam.de/v2.0/; Riano-Pachon, et al., 2007). Almost half of the families (34/65) were found in non-green organisms, including the animal/fungal lineage, while the other half (31/65) were limited to Viridiplantae, including green algae and land-plant species. Genome sequences of 3 green algal lineages showed that at least 10 of these 31 green plant-specific (GPS) families (presence/absence depicted in table on the right) are derived from algal ancestors. It might be expected that the Mamiellales, during genome reduction lost some gene families, especially transcriptional effectors involved in developmental regulation. However, the *Micromonas* genomes contained all 10 shared families. Moreover, the presence of YABBY and ULT families in *Micromonas* is interesting because they have not yet been found in genomes from bryophytes and lycophytes, two early branches in the land-plant phylogeny.

Figure S10, Worden et al.



Supplementary Figure 10. Phylogeny of Transcription Factors. Maximum likelihood-based methods were used to examine phylogenetic relationships between TF family members. RAxML-VI and Multiphyl were used. (a) AP2/ERF phylogeny by RAxML. Topology of a bootstrap consensus tree is shown with branches having less than 10% support condensed (WAG+G model). (b) Alfins, by Multiphyl with JTT+G model. Branches are colored according to lineage as follows: green, land plants; red, volvocales (*Chlamydomonas* and *Volvox*); purple, *Ostreococcus*; and blue, *Micromonas*. Numbers shown after species abbreviations are either for gene locus ID or predicted protein ID. Abbreviations of species names: At, *Arabidopsis thaliana*; Os, *Oryza sativa*; Pp, *Physcomitrella patens*; Ot, *Ostreococcus tauri*; Ol, *Ostreococcus lucimarinus*; Chr, *Chlamydomonas reinhardtii*; Vc, *Volvox cateri*; CCMP1545, *Micromonas* CCMP1545; RCC299, *Micromonas* RCC299; CM, *Cyanidioschyzon merolae*.

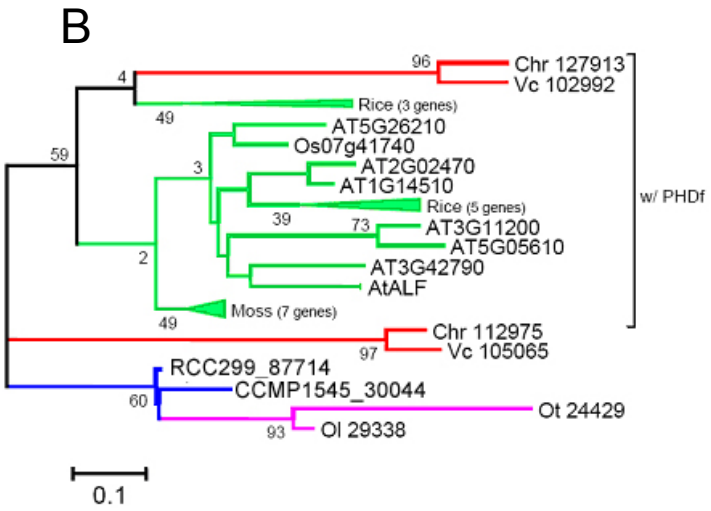
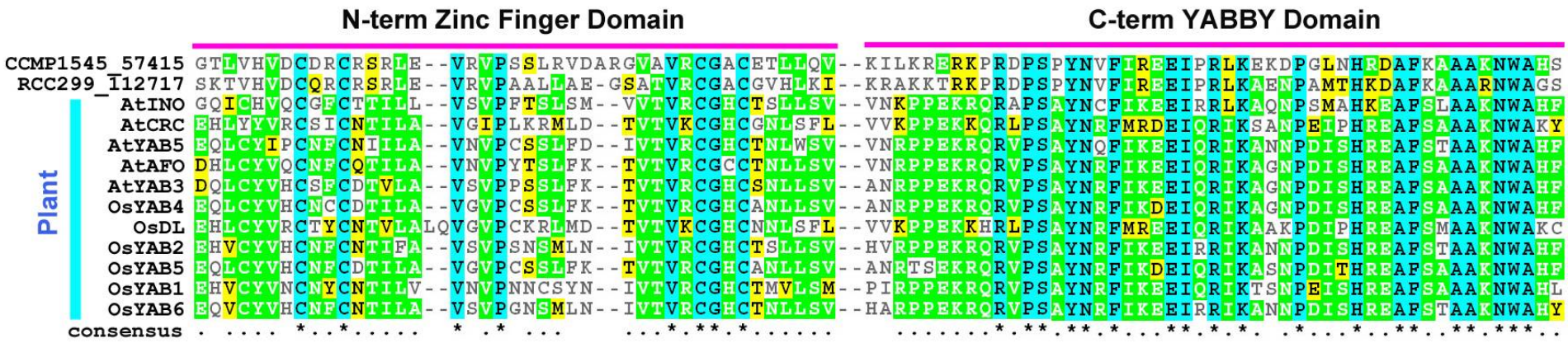


Figure S11a, Worden et al.
Legend for SFigure 11a-c.

A



Supplementary Figure 11. YABBY protein alignment and phylogenetic analysis. (a) Alignment of YABBY proteins. Levels of homology are depicted by different colors (cyan, 100% identical; green, >50% identical; yellow, >50% similar). YABBY family members found in land plant species contain N-terminal zinc-finger domain and C-terminal YABBY domain (related to HMG box domain). Apart form those found in seed plants, the two *Micromonas* YABBY homologs are the only others identified to date that contain domains outside of seed plants. (b) Phylogeny of YABBY proteins. Only full-length sequences are included. *Micromonas* homologs are phylogenetically placed outside of plant-homologs, suggesting that horizontal transfer of YABBY from seed plants to *Micromonas* is unlikely. (c) Phylogeny of YABBY-domain only. The YABBY-like HMG box domain can also be found in *Ostreococcus*, fungi (three Basidiomycete species), and even viruses. Homology between those found in *Micromonas* and fungal homologs may explain the origin of YABBY transcription factor by domain shuffling of the HMG-domain plus zinc-finger domain in the green lineage. However, independent origin of YABBY factors in *Micromonas* cannot be excluded since no additional YABBY members have been identified in other green algae or in non-seed plants. YABBY-like HMG domains from *Ostreococcus* form a clade with a viral protein, not with the *Micromonas* YABBY members. This indicates that the *Ostreococcus* versions are not degenerated versions of other green YABBY members. Sequence information for (a-c) can be found in table S11. Methods for (b-c) are as follows: ML trees were generated using RAxML-VI-HPC (Stamatakis, 2006; http://phylobench.vital-it.ch/raxml-bb/index.php) using the a.a. substitution model JTT. The consensus tree from 100 bootstrap trees is shown; numbers on branches represent bootstrap support values.

Figure S11b, c, Worden et al.

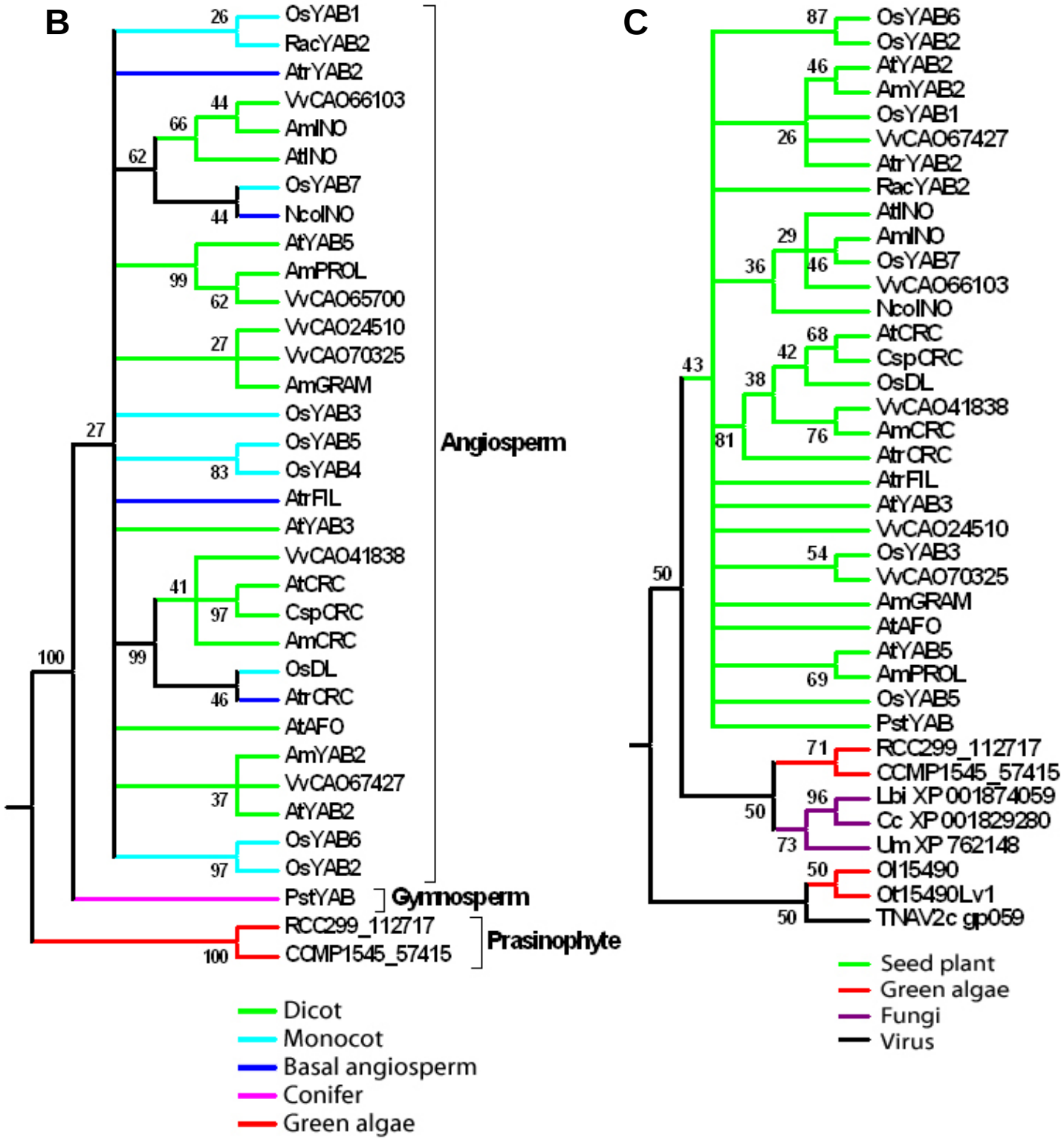
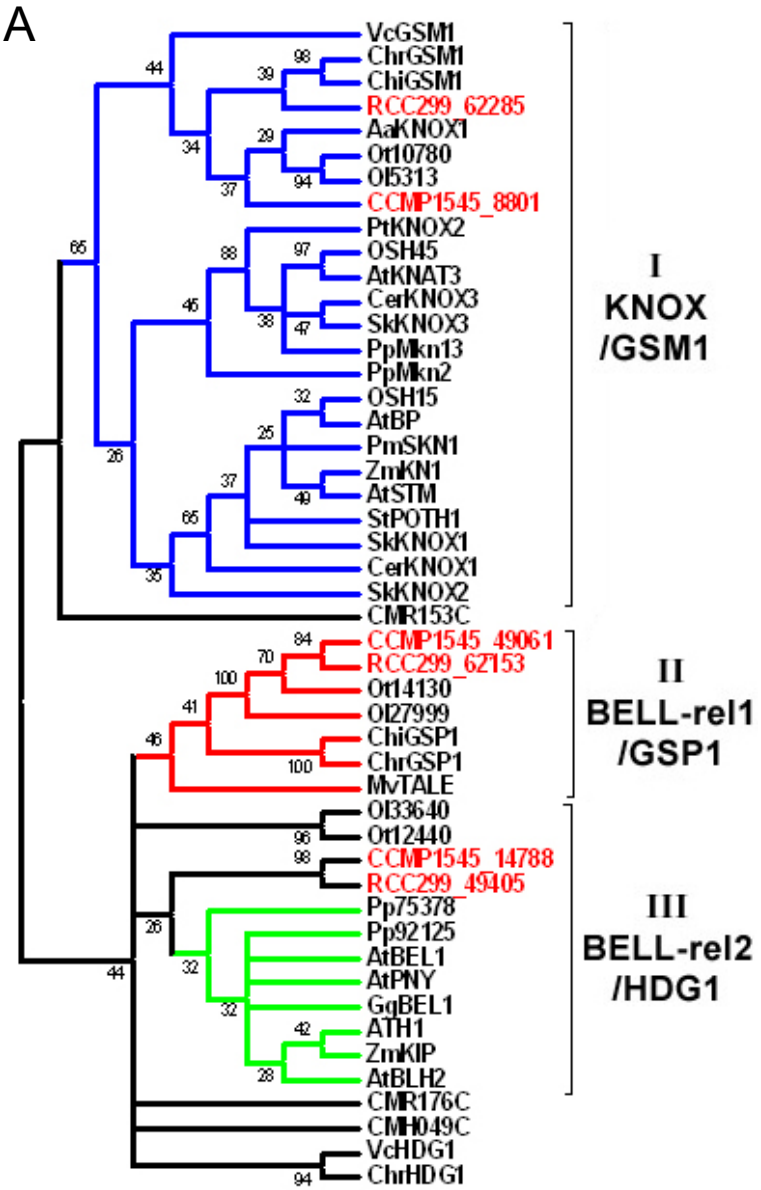
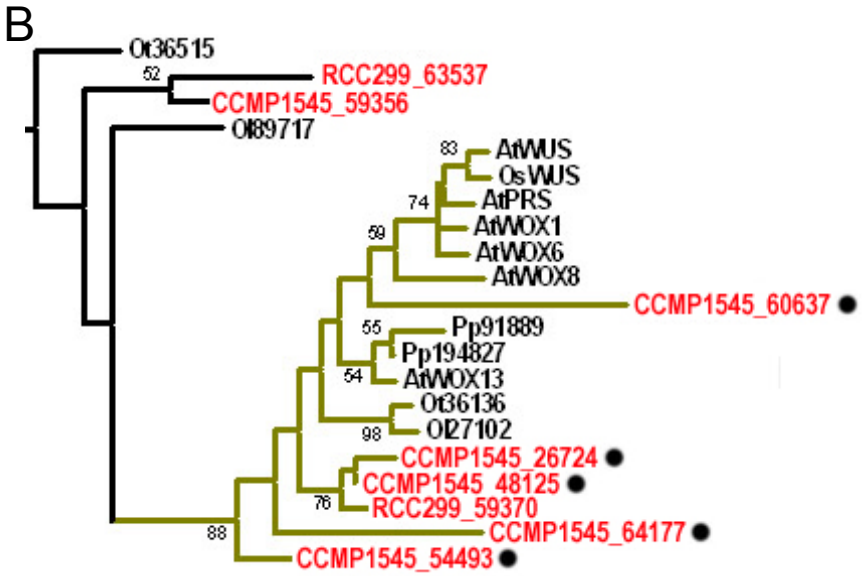
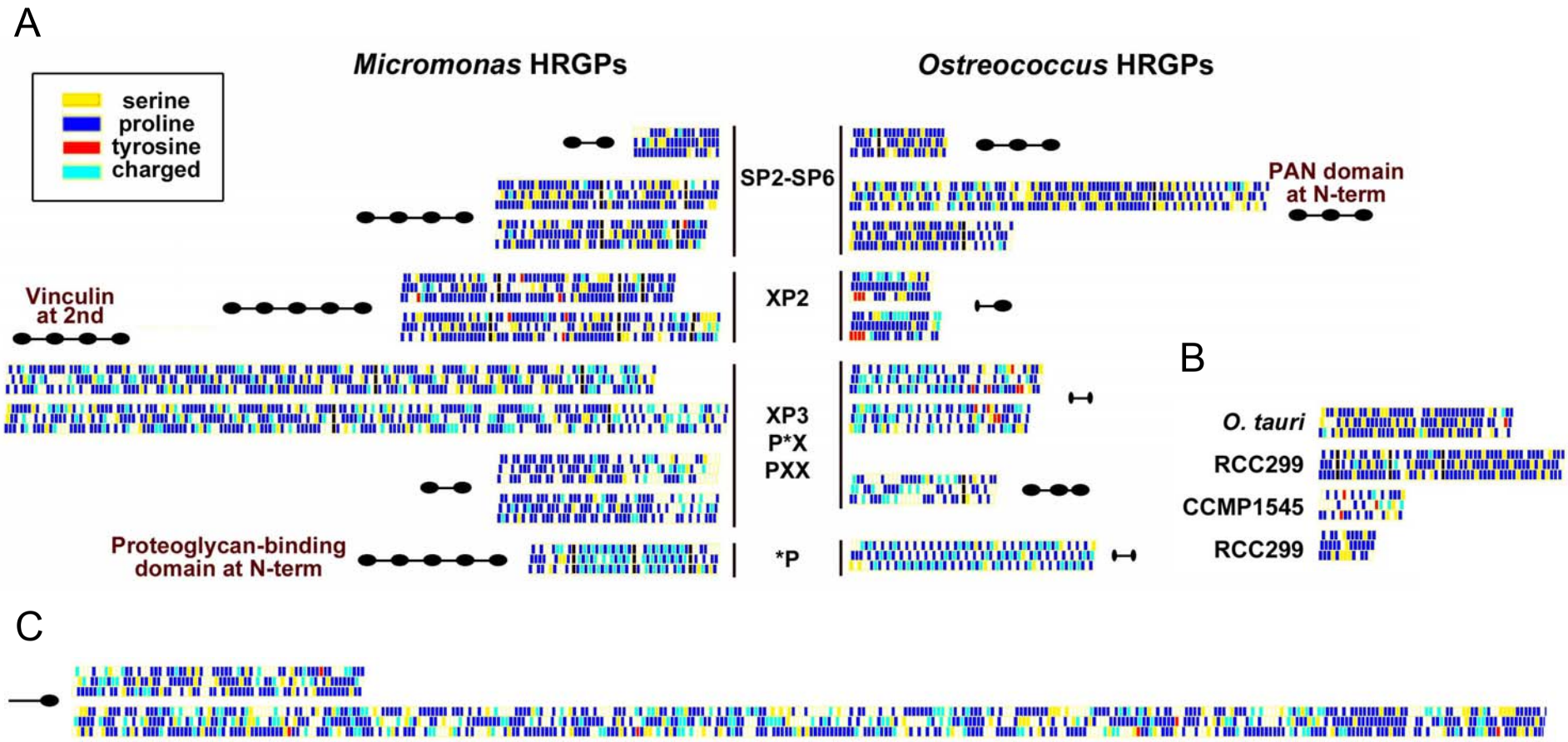


Figure S12, Worden et al.



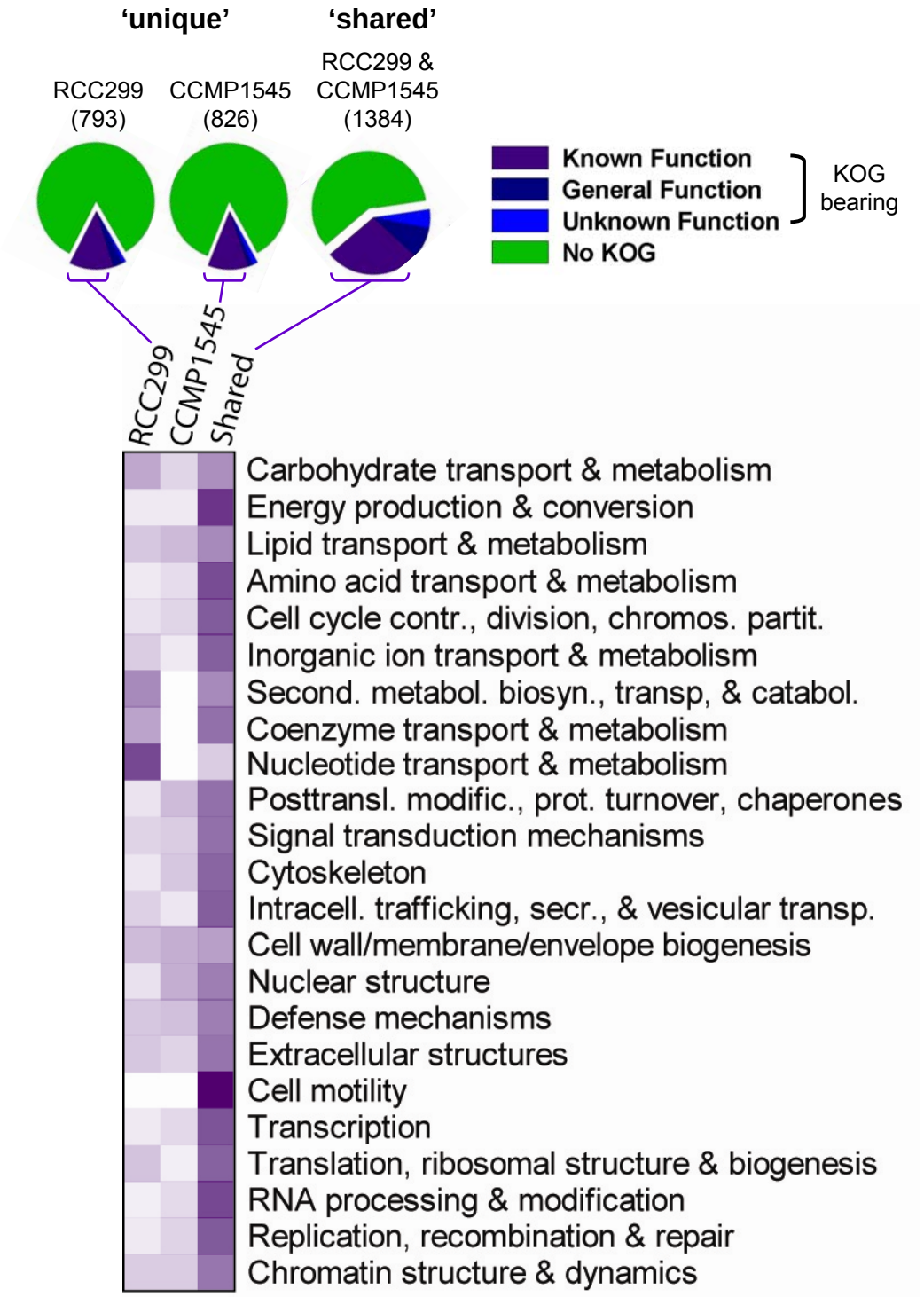


Supplementary Figure 12. Homeodomain proteins of RCC299 and CCMP1545. Those of *Micromonas* are shown in red while those of *Ostreococcus* and other green lineage organisms are shown in black text. (a) ML consensus tree of TALE homeoproteins and from *C. merolae* (included as an outgroup). (b) ML tree of WOX class homeoproteins (taken from the larger tree including all non-TALE classes). WOX class members are colored in branches. 5 proteins from CCMP1545 strain are indicated by dots. Numbers are bootstrap values supporting each branch from 100 replicates. Abbreviations: At, *Arabidopsis thaliana*; Os, *Oryza sativa*; Pp, *Physcomitrella patens*; Ot, *Ostreococcus tauri*; Ol, *Ostreococcus lucimarinus*

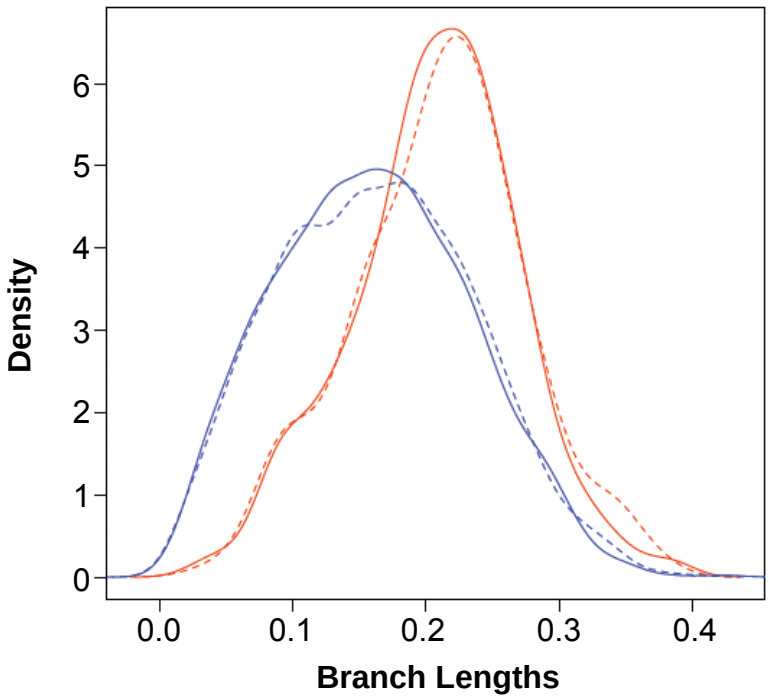


Supplementary Figure 13. Helical wheel diagrams of HRGPs from the sequenced Mamiellales. Overall shaft repeat patterns are retained in orthologs among different lineages (when found), but sequence diversity and length divergence are often extensive, characteristic evolutionary features in this gene family (Lee et al. 2007). Topology of the Mamiellales gene models is shown in cartoons with sticks as shafts and blobs as globular domains. Multiple shafts are connected with a black line between the shafts. (a) Motifs are defined by major repeating units in HRGP shafts, shown at center. For orthologous pairs, shafts from *Micromonas* RCC299 and *O. lucimarinus* are above, and shafts from *Micromonas* CCMP1545 and *O. tauri* are below. Homology information on globular domains is shown above the diagrams. Information on the analyzed gene models is in table S16. (b) N-terminal ADAM-containing HRGP candidates in Mamiellales. The first two do not have orthologous pairs in the related species. Prot. ID numbers: 32533, 103632, 57338, and 55965 from the top. (c) An orthologous pair CCMP1545 (59817, bottom) and RCC299 (62815, top) show extreme divergence in shaft length.

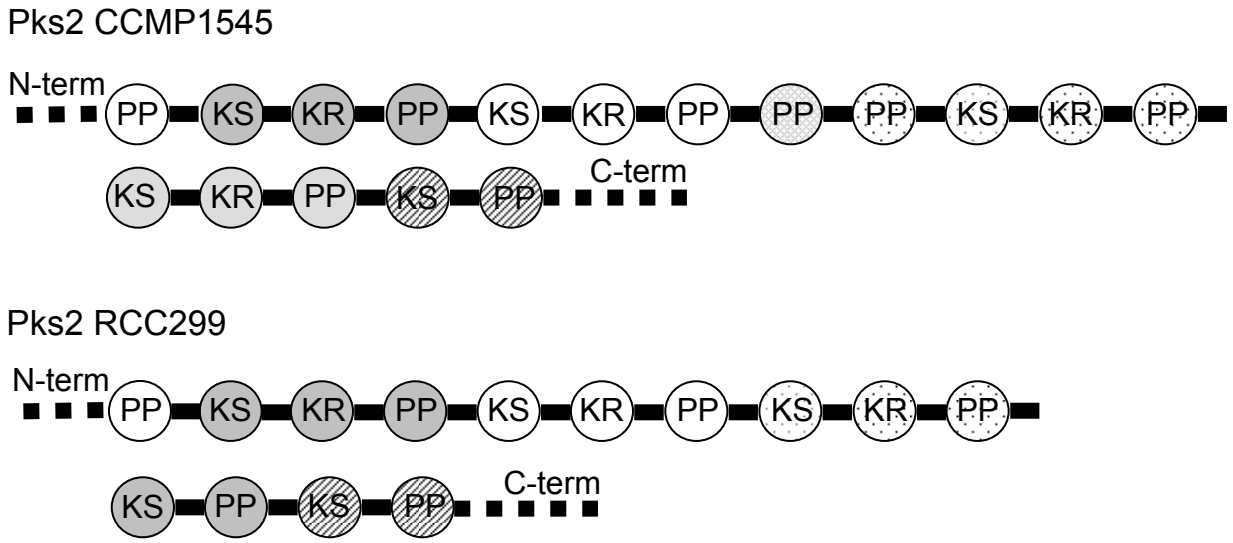
Figure S14, Worden et al.



Supplementary Figure 14. Percentage of genes within known function KOG categories. KOG categories are represented by horizontal heat mapping per specific pools: unique to RCC299 (i.e., not in CCMP1545 or *Ostreococcus* [this does not mean unique to biology!]), unique to CCMP1545 (i.e. not in RCC299 or *Ostreococcus*) and shared by both *Micromonas* (but not in *Ostreococcus*). Numbers of genes falling within known function KOGs are: 89 of 793 and 101 of 826 unique genes in RCC299 and CCMP1545, respectively; 374 of 1384 shared between *Micromonas* (but absent from *Ostreococcus*); and (not shown) 4,943 of 7137 core Mamiellales genes. Catagories such as cell motility, which include flagellum-encoding genes (table S18) are expected to fall fully in the 'shared' (in both *Micromonas* but absent from *Ostreococcus*) since *Micromonas* is motile and *Ostreococcus* is not.

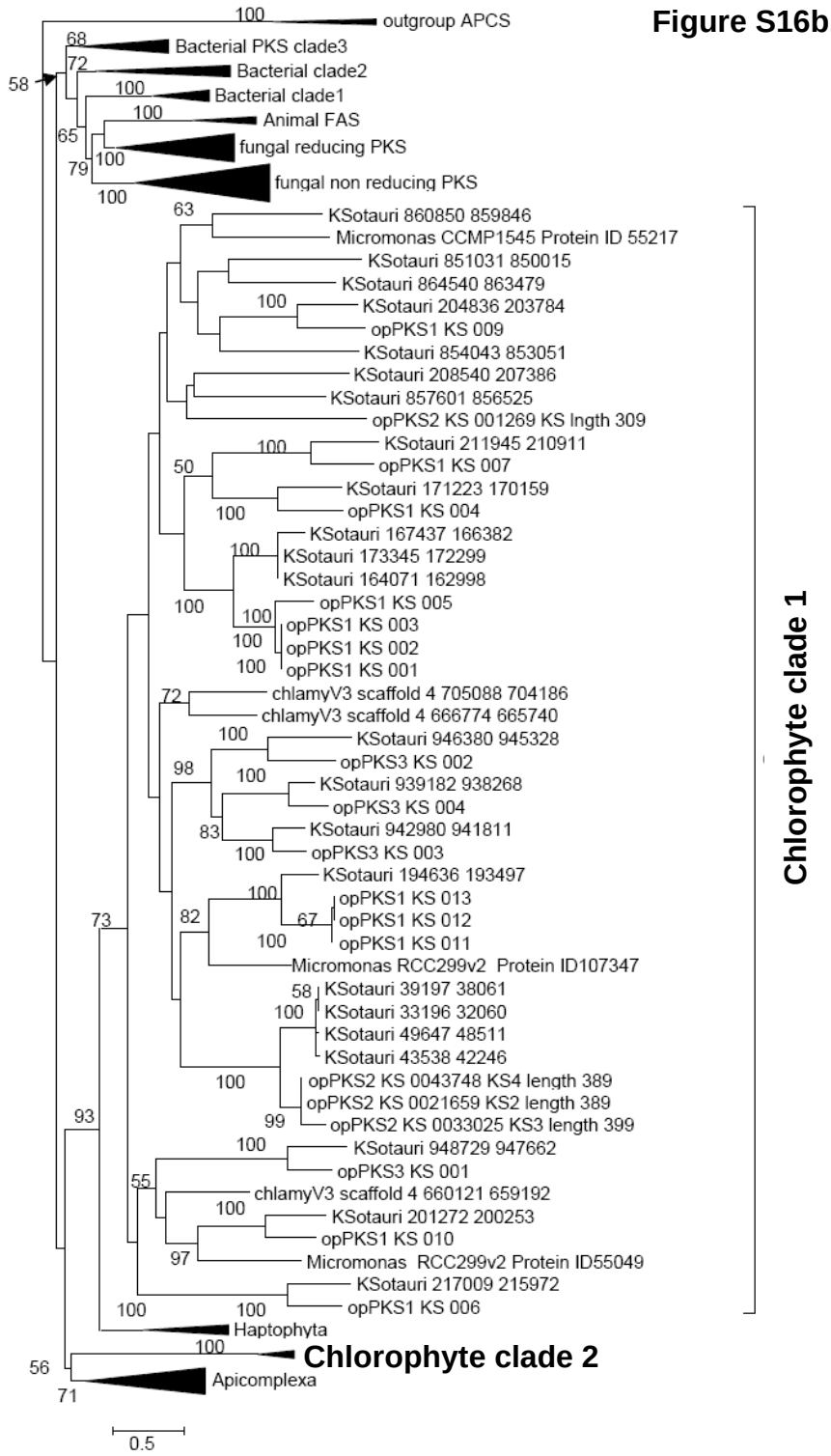


Supplementary Figure 15. Results from phylogenomic profiling of Mamiellales 'core' and *Micromonas* 'shared' genes. Kernal density or frequency distribution plots of branch lengths for *Micromonas* RCC299 (dashed lines) and CCMP1545 (solid lines) in the Mamiellales 'core' (blue) versus *Micromonas* 'shared' (red) gene pools. Note the longer lengths for the 'shared' pool for both species than for 'core' genes, indicated faster evolutionary rates or lower constraints.



Supplementary Figure 16a. Domain structure of Polyketide Synthase in *Micromonas*. Structure is shown for the Pks2 genes in *Micromonas* CCMP1545 and RCC299. Symbols represent the domains of the modular polyketide synthase. Abbreviations are as follows: ketoacyl synthase (KS), ketoacyl reductase (KR), and phosphopantetheine attachment site (PP) or acyl carrier protein [ACP].

Figure S16b, Worden et al.



Supplementary Figure 16b. Phylogenetic analysis of the newly identified PKS sequences in *Micromonas*. Our base data set used the alignment from Kroken et al. (2003) which contains a representative subset of KS domains from bacterial and fungal PKS, metazoan FAS and from oxoacyl-ACP synthases. Newly discovered protistan KS sequences were combined with a subset of sequences from each large clade included in that data set and subsequently re-aligned using kalign (Lassmann and Sonnhammer 2005). The alignment contained 130 sequences and 679 characters (provided upon request). Maximum likelihood phylogenetic trees were calculated with PhyML (Guindon and Gascuel 2003) using a BIO-NJ tree as starting tree, the WAG evolutionary model, with a gamma distribution parameter estimated from the data. Bootstrap analyses were performed with the same settings for 200 replicates.

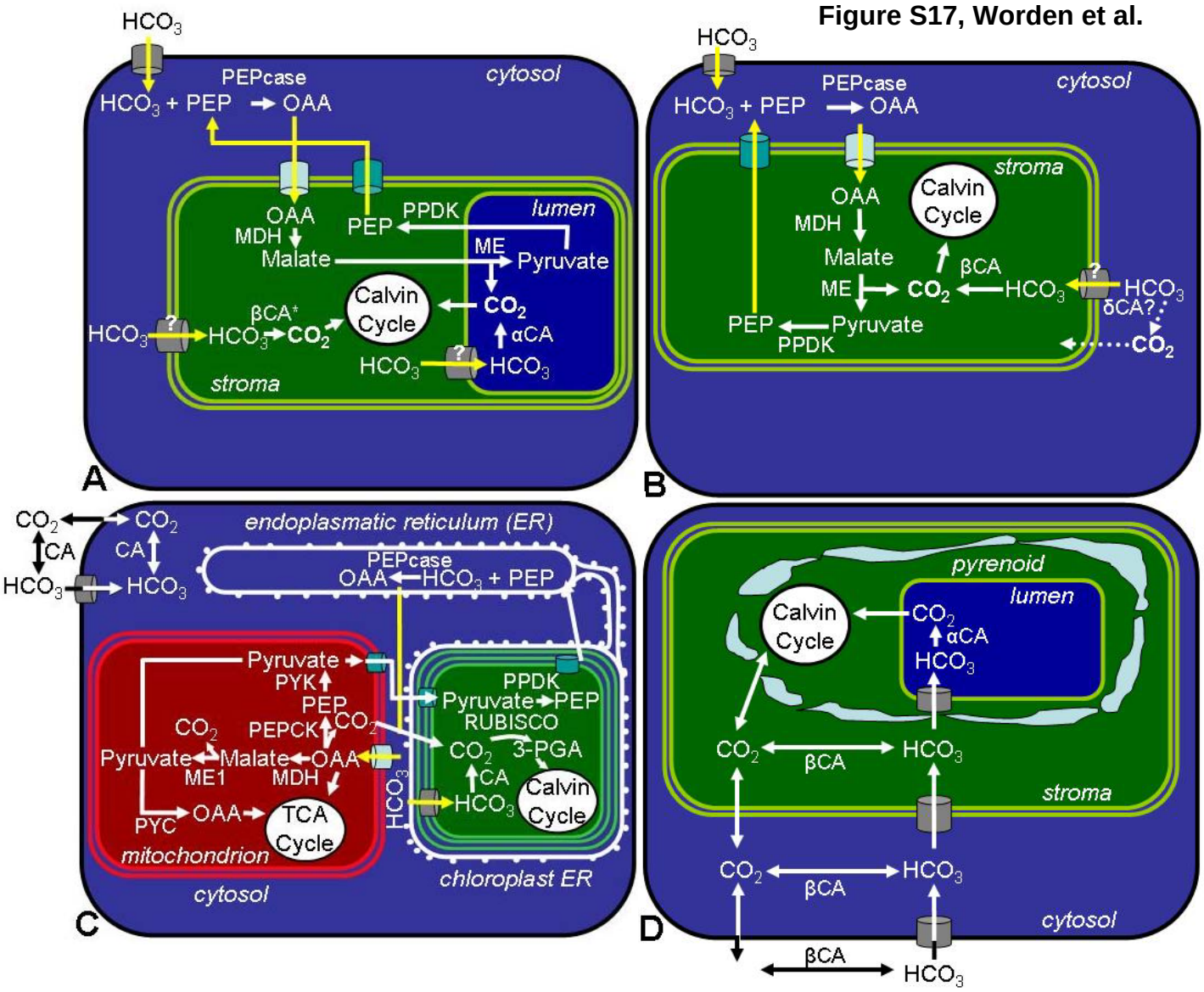
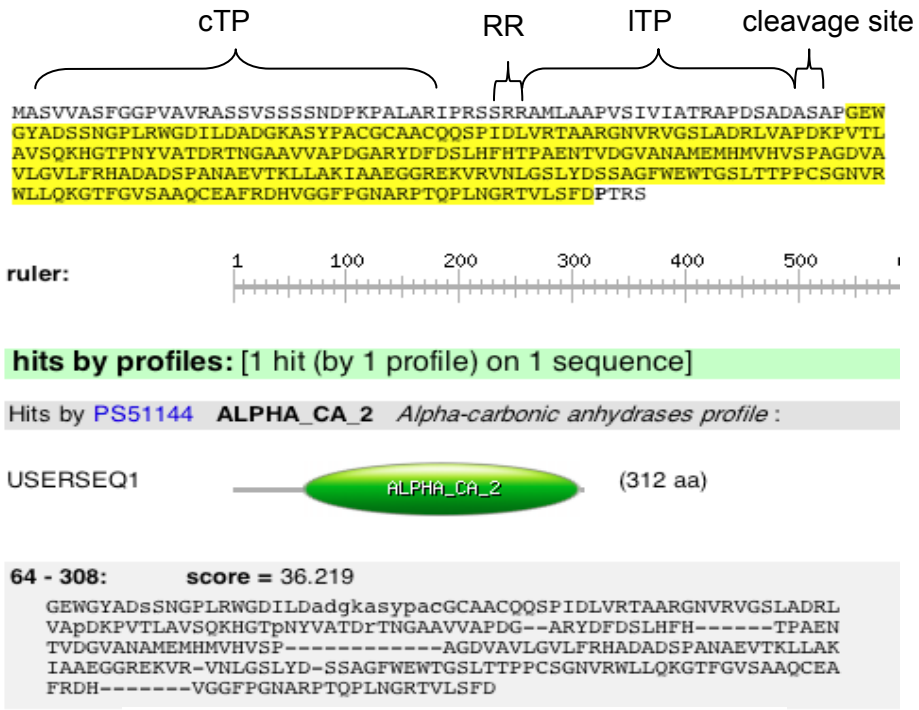
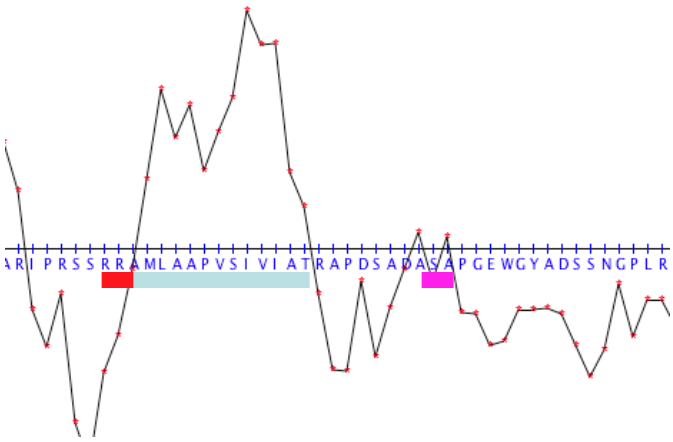


Figure 17. Rethinking carbon-concentrating mechanisms based on complete genome analyses. Proposed models for: (a) *Micromonas* RCC299 and CCMP1545, which appear to use an unusual C_4-like carbon fixation pathway based on decarboxylation of malate inside the thylakoid lumen by a ME. Furthermore, they can accumulate CO_2 by CAs targeted to the chloroplast (cp) lumen and stroma. (b) In contrast, *Ostreococcus* likely accumulates CO_2 inside the cytosol and the cp stroma, performing a C_4-like carbon fixation pathway based on malate inside the stroma, not the lumen. (c) In the diatoms, *T. pseudonana* and *P. tricornutum* plastid targeted translocators for pyruvate and PEP are found. However, cp-targeted OAA and malate transporter are not found; an OAA transporter, ME, and PEPCK targeted to the mitochondria are found instead, suggesting decarboxylation inside the mitochondria. *P. tricornutum* seems to concentrate CO_2 with cytosol and cp stroma-targeted CAs. A putative cp-targeted sodium bicarbonate translocator is also present. (d) *C. reinhardtii* has a CCM based on several CAs in the periplasmatic space, cytosol, cp stroma and lumen. The presence of a pyrenoid in *C. reinhardtii* reduces diffusion losses of CO_2. Abbreviations: PEPCase, phosphoenolpyruvate carboxylase; PPDK, pyruvate-phosphate dikinase; PYC, pyruvate carboxylase; CA, carbonic anhydrase; RUBISCO, ribulose-1,5-bisphosphate carboxylase; MDH, malate dehydrogenase; ME, malic enzyme; PEPCK, phosphenolpyruvate carboxykinase; *only found in CCMP1545.

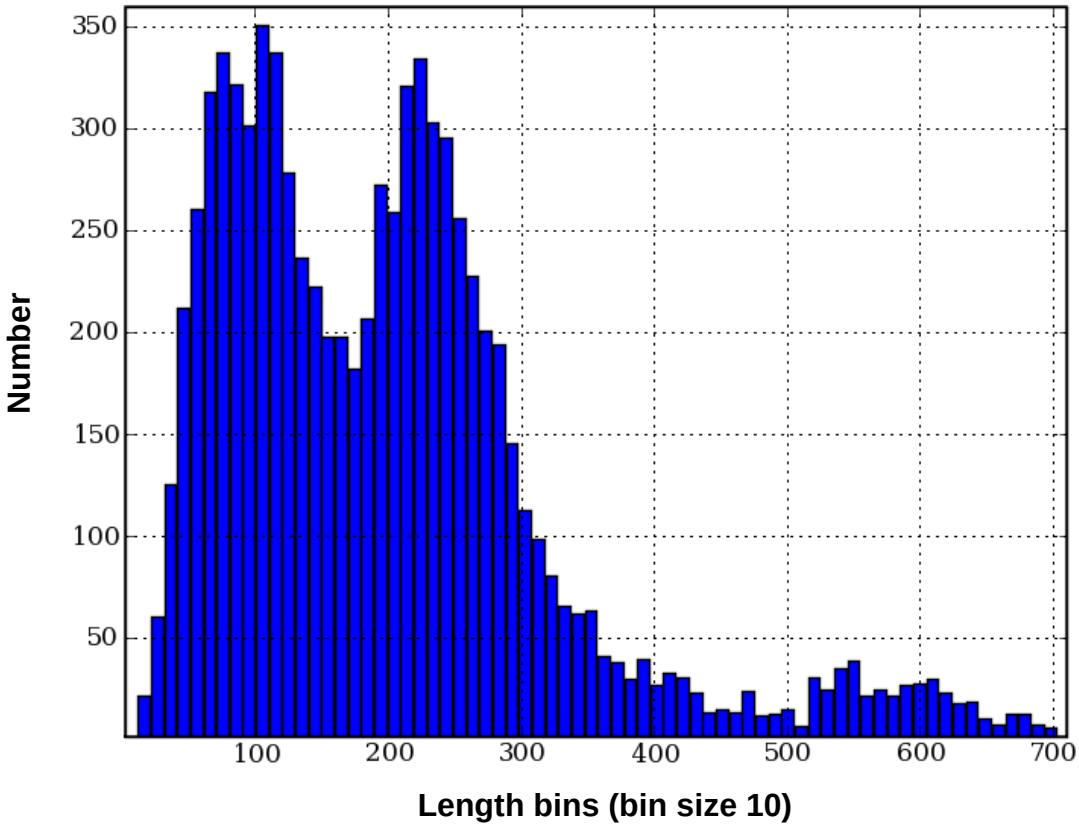
A



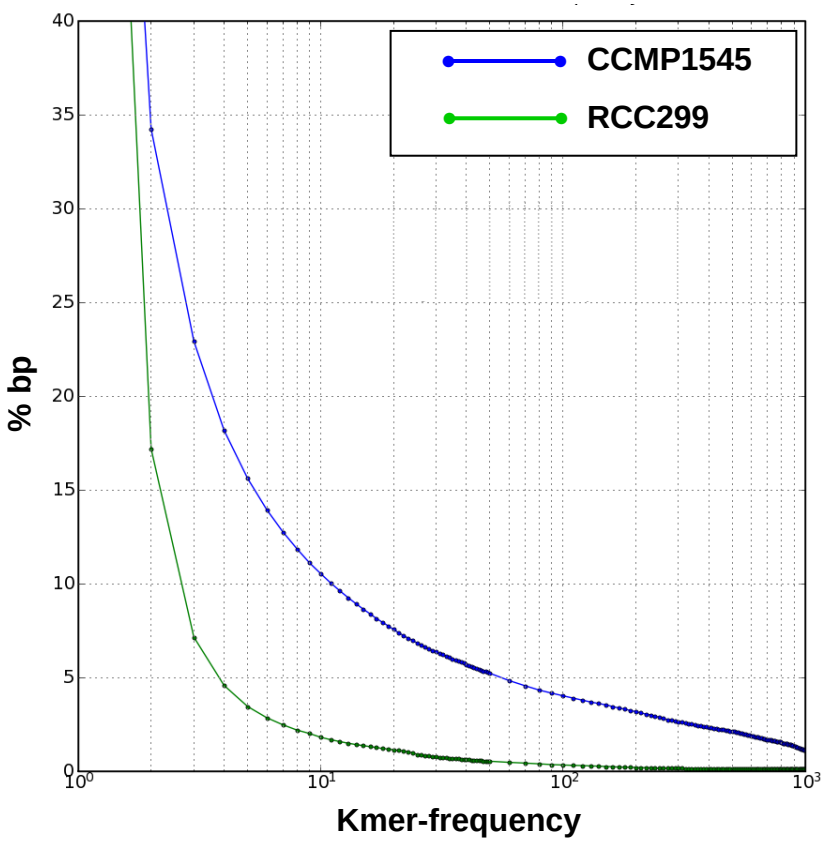
B



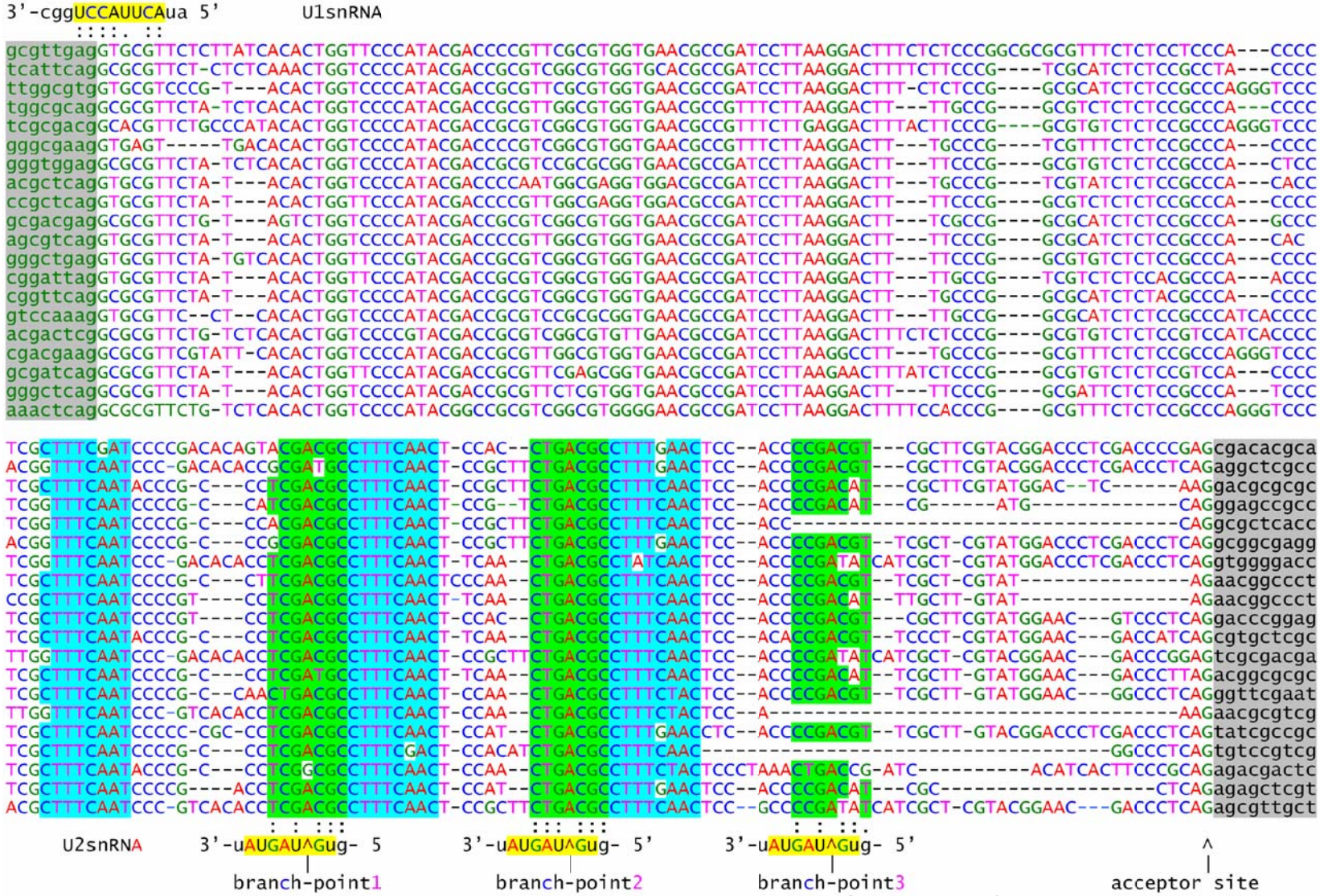
Supplementary Figure 18. Example of lumenal targeting of alpha carbonic anhydrase. (a) Prosite (http://expasy.org/prosite) scan of *Micromonas* RCC299 lumenal targeted α-carbonic anhydrase Prot. ID 96952 (cTP = chloroplast transit peptide, RR = twin-argine motif characteristic of delta-pH dependent TAT transport, ITP = lumenal transit peptide, ASA = cleavage site of lumenal processing peptidase). (b) Hydrophobicity plot (Kyte-Doolittle), (red bar = RR, light blue bar = ITP, pink bar = ASA).



Supplementary Figure 19. Length distribution of Introner Elements. CCMP1545 length peaks at 100, 250 (and ~500-600) bp not found in simple, trf or JGI-masked elements, potentially providing a hint towards insertion mechanism or duplication steps involved.



Supplementary Figure 20. Features of Introner Elements. Fraction of the genome versus 16-mer frequency limits. CCMP1545 has a distinctly higher amount of repetitive sequences than RCC299. All kmers with a frequency ≥ 10 cover 10.5% of the CCMP1545 genome and 1.8% RCC299 genome. Kmer frequencies of ≥100 occur in 4.0% of the CCMP1545 genome, but only 0.3% of the RCC299 genome.



Supplementary Figure 21. Alignment of typical IE1, IE2, IE3 and IE4 Introner Elements. Conserved motifs are highlighted in green (branch-point motif-A) and in blue (motif-B), exon border sequences are shown in grey, U1 and U2snRNA sequences and their putative pairing with IE is also shown as is the location of acceptor site and putative branch-points (arrows). (a) Alignment of typical IE1 sequences. (b) Alignment of typical IE2 sequences. (c) Alignment of typical IE3 sequences. (d) Alignment of typical IE4 sequences. The location of a cryptic acceptor site is indicated, which would fit with the use of branch-point-0. In addition, a nested IE is shown (dark red). Note that this nested IE does not contain any branch-site motif-A, explaining why this IE can not be spliced and 'stays' inserted. (e) Alignment of IE3 and IE4. IE3 sequences are highlighted in blue, IE4 in yellow. The central sequence is not highlighted as it is a degenerate IE copy having features fitting equally to these two classes.

Figure S21a, Worden et al.
Legend for all Figure S21

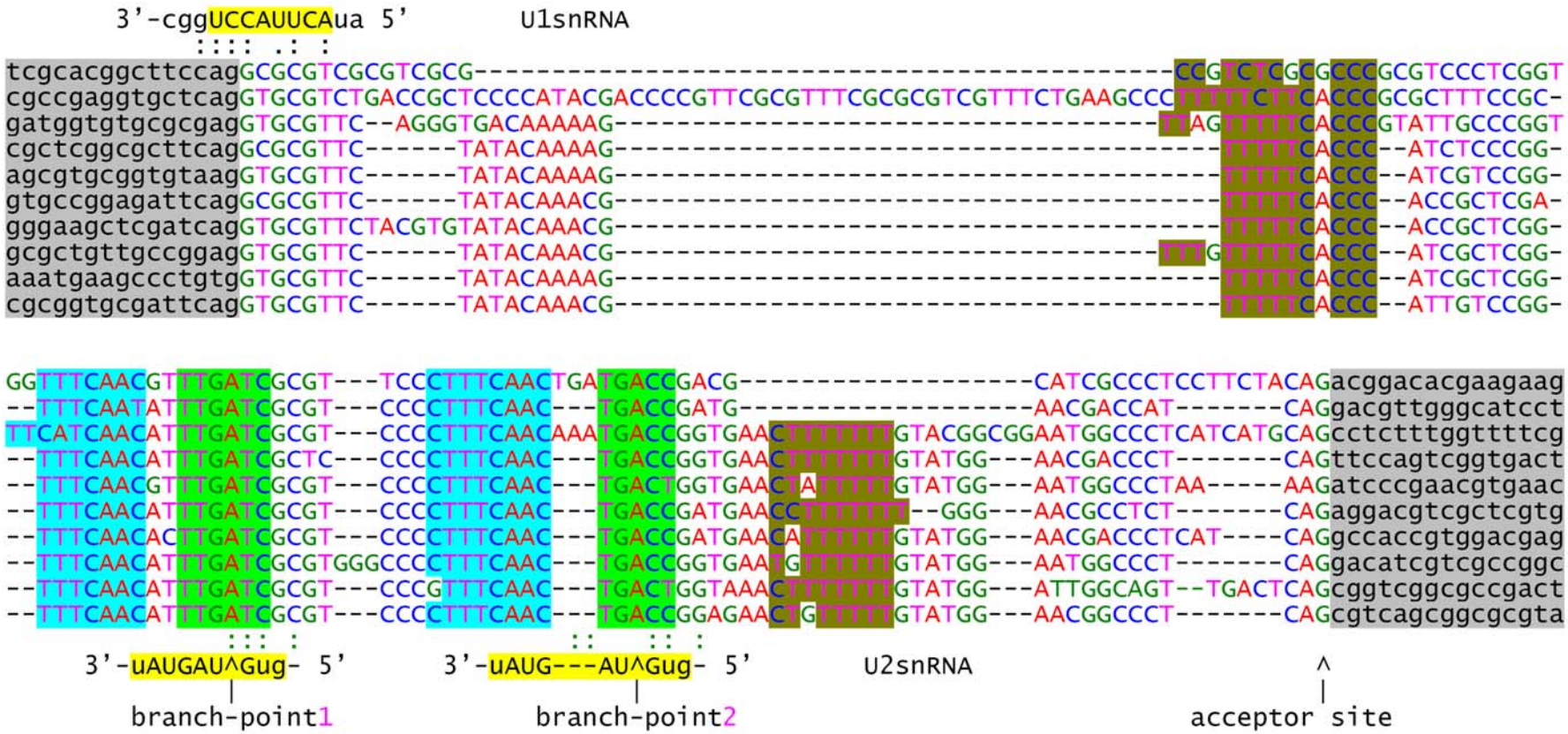
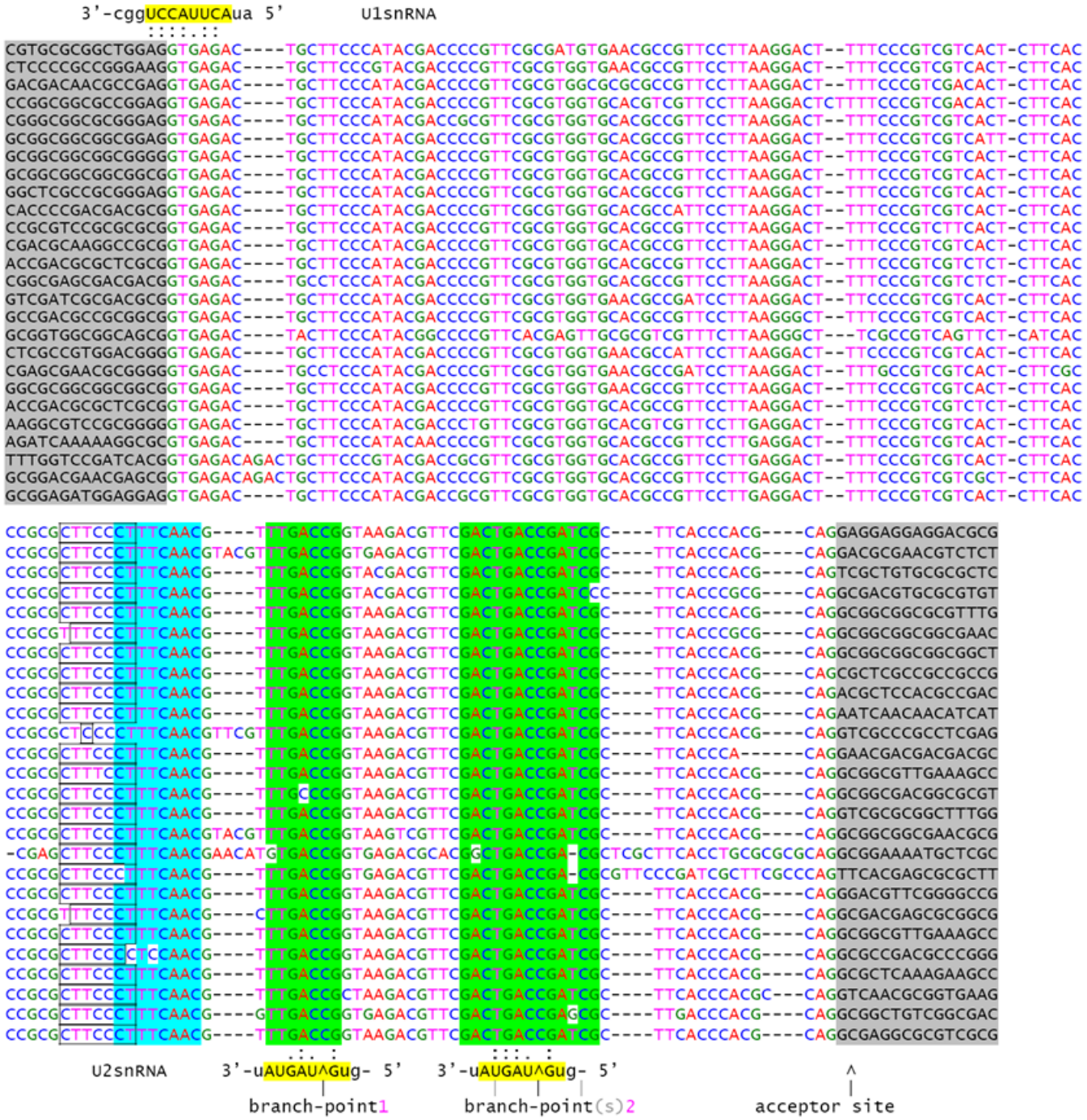


Figure S21b, Worden et al.

Figure S21c, Worden et al.



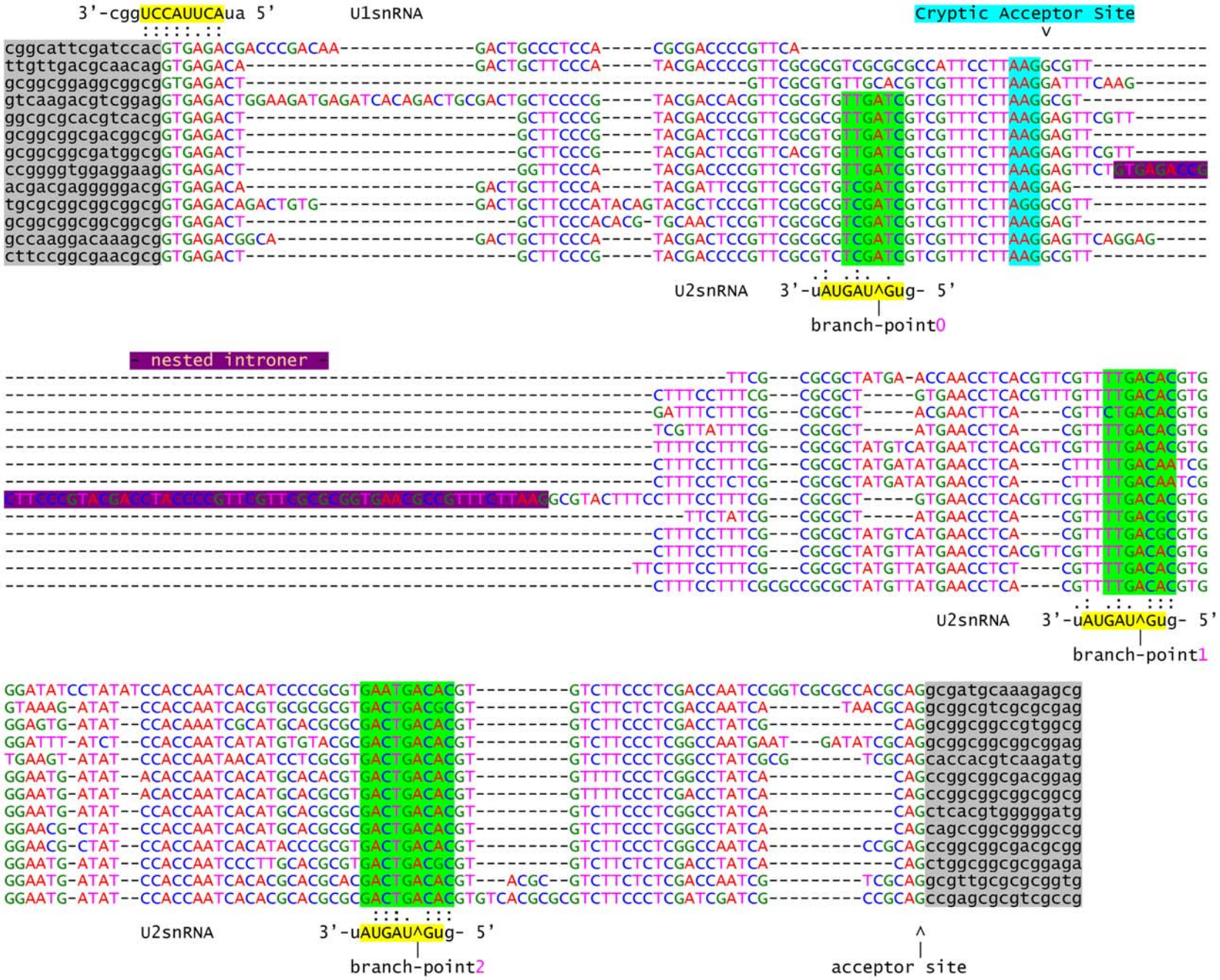


Figure S21d, Worden et al.

```
GTGA---------------------------GACT----------------------GTTCGCGTGTTGCACGTCGTTTCTTAAGGATTTCAAG----------------------------
GTGAGACTGGAAGATGAGATCACAGACTGCGACTGCTCCCCG-----TACGACCACGTTCGCGTGTTGATCGTCGTTTCTTAAGGCGT--------------------------------
GTGA---------------------------GACTGCTTCCCG-----TACGACCCCGTTCGCGCGTTGATCGTCGTTTCTTAAGGAGTTCGTT----------------------------
GTGA---------------------------GACTGCTTCCCG-----TACGACTCCGTTCGCGTGTTGATCGTCGTTTCTTAAGGAGTT-------------------------------
GTGA---------------------------GACTGCTTCCCG-----TACGACTCCGTTCACGTGTTGATCGTCGTTTCTTAAGGAGTTCGTT----------------------------
GTGA---------------------------GACTGGTTCCCA-----TACGACCCCGTTCTCGTGTTGATCGTCGTTTCTTAAGGAGTTCTGTGAGACCGCTTCCCGTACGACCTACCCC
GTGAGACA-----------------------GACTGCTTCCCA-----TACGATTCCGTTCGCGTGTCGATCGTCGTTTCTTAAGGAG--------------------------------
GTGAGACAGACTGTG---------------GACTGCTTCCCATACAGTACGCTCCCGTTCGCGCGTCGATCGTCGTTTCTTAGGGCGTT-------------------------------
GTGA---------------------------GACTGCTTCCCACACG-TGCAACTCCGTTCGCGCGTCGATCGTCGTTTCTTAAGGAGT-------------------------------
GTGAGACGGCA--------------------GACTGCTTCCCA-----TACGACTCCGTTCGCGCGTCGATCGTCGTTTCTTAAGGAGTTCAGGAG---------------------------
GTGA---------------------------GACTGCTTCCCG-----TACGACCCCGTTCGCGTCTCGATCGTCGTTTCTTAAGGCGTT--------------------------------
GTGA---------------------------GACTGCTTCCCG-----TACGACCCCGTTTGCGCGTACAACGCCGTTTCTTAAGGAG---------------------------------
GTGA---------------------------GACTGCTCCCCG-----TACGACCCCGTTCGCGCGTCGTGCGTCGTTTCTCAAGGAC---------------------------------
GTGA---------------------------GACTGCTTCCCA-----TACGACCCCGTTCGCGTGGTGAACGCCGTTCCTTAAGGAC---------------------------------
GTGA---------------------------GACTGCTTCCCG-----TACGACCCCGTTCGCGTGGTGAACGCCGATCCTTAAGGAC---------------------------------
GTGA---------------------------GACTGCTTCCCA-----TACGACCCCGTTCGCGTGTTGATCGCCGTTCCTTAAGGAC---------------------------------
GTGA---------------------------GACTGCTTCCCA-----TACGACCCCGTTCGCGCGTTGATCGTCGTTTCTTAAGGAC---------------------------------
GTGA---------------------------GACTGCTTCCCA-----TACGACCCCGTTCGCGTGTTGATCGTCGTTCCTTAAGGAC---------------------------------

-----------------------------------------GATTTCTTTCG---CGCGCT-----ACGAACTTCA----CGTTCTGACACGTGGGAGTGATATCCACAAATCGCATGC
-----------------------------------------TCGTTATTTCG---CGCGCT-----ATGAACCTCA----CGTTTTGACACGTGGGATTTATCTCCACCAATCATATGT
-----------------------------------------TTTTCCTTTCG---CGCGCTATGTCATGAATCTCACGTTCGTTTTGACACGTGTGAAGTATATCCACCAATAACATCC
-----------------------------------------CTTTCCTTTCG---CGCGCTATGATATGAACCTCA----CTTTTTGACAATCGGGAATGATATACACCAATCACATGC
-----------------------------------------CTTTCCTCTCG---CGCGCTATGATATGAACCTCA----CTTTTTGACAATCGGGAATGATATACACCAATCACATGC
GTTCGTTCGCGCGGTGAACGCCGTTTCTTAAGGCGTACTTTCCTTTCCTTTCG---CGCGCT-----GTGAACCTCACGTTCGTTTTGACACGTGGGAATGATATCCACCAATCACATGC
-----------------------------------------------TTCTATCG---CGCGCT-----ATGAACCTCA----CGTTTTGACGCGTGGGAACGCTATCCACCAATCACATGC
-----------------------------------------CTTTCCTTTCG---CGCGCTATGTCATGAACCTCA----CGTTTTGACGCGTGGGAACGCTATCCACCAATCACATAC
-----------------------------------------CTTTCCTTTCG---CGCGCTATGTTATGAACCTCACGTTCGTTTTGACACGTGGGAATGATATCCACCAATCCCTTGC
---------------------------------------TTCTTTCCTTTCG---CGCGCTATGTTATGAACCTCT----CGTTTTGACACGTGGGAATGATATCCACCAATCACACGC
-----------------------------------------CTTTCCTTTCGCGCCGCGCTATGTTATGAACCTCA----CGTTTTGACACGTGGGAATGATATCCACCAATCACACGC
---------------------------------------------TGTCTTCTC----CAT-----------------------------------------------------------
-----------------------------------------TTCTCCCGTCG--TCACTCT------TCACCCTCACTTTCTTTTCAACGTACG--------TTTGACCGGTGAACCGT
-----------------------------------------TTTCCCCGTCG--TCACTCT------TCACCCGCGCTTCCCTTTCAACGTTTG------------ACCGGTAAGACGT
-----------------------------------------TTTTGCCGTCG--TCACTCT------TCACCCGCGCTTCCCTTTCAACGTTTG------------ACCGGTAAGACGT
-----------------------------------------TTTTCCCGTCG--TCACTCT------TCACCCGCGCTTCCCTTTCAACGTTTG------------ACCGGTAAGACGT
-----------------------------------------TTTTCCCGTCG--TCACTCT------TCACCCGCGCTTCCCTTTCAACGTTTG------------ACCGGTAAGACGT
-----------------------------------------TTTTCACGTCG--TCACTCT------TCACCCGCGCTTCCCTTTCAACGTTTG------------ACCGGTAAGACGT

ACGCGCGACTGACACGT---------GTCTTCCCTCGACCTATCG----------------------CAG
GTACGCGACTGACACGT---------GTCTTCCCTCGGCCAATGAATGATAT--------------CGCAG
TCGCGTGACTGACACGT---------GTCTTCCCTCGGCCTATCGCGT-----------------CGCAG
ACACGTGACTGACACGT---------GTTTTCCCTCGGCCTATCA----------------------CAG
ACGCGTGACTGACACGT---------GTTTTCCCTCGACCTATCA----------------------CAG
ACGCGCGACTGACACGT---------GTCTTCCCTCGGCCTATCA----------------------CAG
ACGCGCGACTGACACGT---------GTCTTCCCTCGGCCTATCA----------------------CAG
CCGCGTGACTGACACGT---------GTCTTCCCTCGGCCAATCA-------------------CCGCAG
ACGCGTGACTGACACGT---------GTCTTCCCTCGACCTATCA----------------------CAG
ACGCACGACTGACACGT-----ACGCGTCTTCTCTCGACCAATCG-------------------TCGCAG
ACGCGCGACTGACACGTGTCACGCGCGTCTTCCCTCGATCGATCG-------------------CCGCAG
-------------------------CG-TTTCTTTA-----------------------------CAG
----TCGACTGACCGACA--------CG--TTCACCCGCCCGCGATCGGTTCACC-------------CAG
----TCGACTGACCGAT--------CG-CTTCACCCCGCCACCCACGATCGACCCTCTTTGACCCTACAG
TCGACTGACCGAT-CGC------------TTCACCCA----------------------------CGCAG
TAGACTGACCGTT-CGC------------TTCACCCG----------------------------CGCAG
TCGACTGACCGAT-CGC------------TTCACCCACGCACGGCGCT-------------------CAG
-CGACTGACCGAT-CGC------------TTCACCCACC----------------------------CAG
```

Figure S21e, Worden et al.

Supplementary Figure 22. Introner Element density in CCMP1545 shown in a multistep mapping process. Numbering to the far left indicates chromosome number. First, IE1 through IE4, respectively, were mapped on chromosomes using repeat-masker. Based on the coordinates returned the proportion of sequence being IE, compared to the proportion not being IE, was calculated using a sliding window (2000 nt) as shown in fig. S1. Here, each gray bar represents a chromosome and the 4 internal bars represent the local densities of the respective IE (IE1 through IE4) . The contrast of the highest density compared to the lowest density is most apparent for Chromosome 2, while Chromosome 19 is totally devoid of IE.