Organellar Genomes and Carbohydrate Active Enzymes from Draft Nuclear Genome of a Prasinophyte Alga, *Pyramimonas parkeae*

By

Anchittha Satjarak

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy
(Botany)

at the
UNIVERSITY OF WISCONSIN-MADISON
2017

Date of final oral examination: 2/14/2017

This dissertation is approved by the following members of the Final Oral Committee:

Linda Graham, Professor, Botany
Donna Fernandez, Professor, Botany
Kenneth Sytsma, Professor, Botany
Chris Hittinger, Assistant Professor, Genetics
Eunsoo Kim, Assistant Curator, American Museum of National History

Acknowledgements

My dissertation work was profoundly personal yet collaborative product. With this in mind, I would like to acknowledge people who, without their help and guidance, my research would be diminished. I gratefully acknowledge my graduate committee – Dr. Linda Graham, Dr. Donna Fernandez, Dr. Kenneth Sytsma, Dr. Chris Hittinger, and Dr. Eunsoo Kim – for their guidance and assistance. In particular, I sincerely express my deep gratitude to my dear advisor Prof. Dr. Linda Graham for accepting me into the Department of Botany, providing me funding for *Pyramimonas parkeae* research through NSF, giving me thought provoking comments and advice, and supporting me academically and mentally throughout my time at UW-Madison.

My thanks go to the Graham Lab's current and formal members – Dr. James Graham, Dr. Lee Wilcox, Dr. Shahrizim Zulkifly, Dr. Christopher Cardona-Correa, Jennifer Knack, Michael Braus, Michael Piotrowski, and Elizabeth Phillippi for the helping me with lab support and their company. I also thank Marie Trest who helped me with algal culture and culturing techniques.

I would also like to specially thank Dr. Eunsoo Kim from the American Museum and National History and the Kim Lab's members – Dr. John Burns and Amber Paasch – who contributed whole genome sequences of *Cymbomonas tetramitiformis* for constructing its organellar genomes and helped with its manuscripts preparation.

Many thanks also go to the generous financial support from many sources: Development and Promotion of Science and Technology Talents Project (Royal Government of Thailand scholarship), Office of the Civil Service Commission (Royal Thai Fellowship), NSF DEB111-9944 (awarded to Dr. Linda Graham), Raper tralvel award, Newcomb award, Davis award, and ABLS and AFS funds.

I appreciate the assistance with software installation and computational scripts modification from Steve Goldstein, Isaac Knoflicek, and Lindsay Traeger. I thank Laboratory of Genetics for access to computational infrastructure and support. Also, I thank Sarah Friedrich for the help with figures.

I also thank my fellow graduate students at UW-Madison and other schools. I would like to mention Neeranooch Malangpoo, Chuanchom Aumnate, Paradee Thammapichai, Orathai Viriyavatsharagon, Saranya Peerakietkhajorn, Suchawan Pornsukarom, Kanokwan Sansanaphongpricha, Arranee Chotigo, Wittawat Sanrang, Sucheewin Chotchatchawankul, Chadthip Rodtassana, Bhanubhatra Jittiang, Kewalin Sounburee, Tippapha Pisithkul, Pichamon Limcharoenchat Pattama Hannok, Saowanee Jiwlawat, Matthew Pace, Cathy Hu, Nelly Martin, Alejandro Zuluaga, Brian Sidoti, Lauren Moscoe, Alfonso Doucette, Lindsay Traeger, and Soo-Hyun Kim. The support during my time in graduate school is immeasurable.

Lastly, I thank my family for being my biggest supporters. My parents, Suvit and Chumnong, taught me the value of education and hard work. My sister, who is my best friend, always gives me encouragement. Without the love and support from my family, I may not have got where I am today.

Thank you.

Abstract

Prasinophytes form a paraphyletic assemblage of early diverging green algae, which have the potential to reveal the traits of the last common ancestor of the main two green lineages in Viridiplantae: 1) the clade consisting of chlorophyte algae and 2) the clade consisting of streptophyte algae and embryophytes. In this study, we sequenced the whole genome of an alga from prasinophyte clade I–*Pyramimonas parkeae* strain NIES254–and constructed its chloroplast, mitochondrial, and draft nuclear genomes. We investigated intra-specific variability of both chloroplast and mitochondrial genomes, inter-specific variability of mitochondrial genomes, and utilized the information from both organellar genomes to explore the relationships among prasinophytes. The results showed that both mitochondrial and chloroplast genomes of *P. parkeae* exhibited variability at the intra-specific level. Similarly, variability at the inter-specific level was observed among prasinophyte mitochondrial genomes. Our phylogenetic analyses suggested that the information from prasinophyte chloroplast and/or mitochondrial genomes was sufficient to resolve monophyly of the known prasinophyte clades. However, these data were not sufficient to resolve deeper level relationships among prasinophyte clades. The draft nuclear genome of *P. parkeae*, along with existing transcriptomic sequence, was used to investigate carbohydrate active enzymes and make comparisons with those of other Viridiplantae whose genomes have been fully sequenced. We found that the *P. parkeae* nuclear genome encoded carbohydrate active protein families similar to those previously observed for other prasinophytes, green algae, and early-diverging embryophytes for which full nuclear genomic sequence is publically available. Sequences homologous to genes related to biosynthesis of starch and cell wall carbohydrates were identified in the *P. parkeae* genome, indicating molecular traits common to Viridiplantae. Sequences clustering with bacterial genes that encode cellulose synthases (Bcs) were found in the *P. parkeae* genome and transcriptome, and these sequences included regions coding for domains common to bacterial and plant cellulose synthases. These new sequences were incorporated into phylogenies aimed at illuminating the evolutionary history of cellulose production by Viridiplantae. Genomic sequences related to biosynthesis of xyloglucans, pectin, and starch likewise shed light on the origin of key Viridiplantae traits.

Table of contents

CHAPTER 1: INTRODUCTION TO THE PRASINOPHYTE *PYRAMIMONAS PARKEAE*

This thesis focuses on organellar and nuclear genomes of *Pyramimonas parkeae*, which is classified within an early-diverging, paraphyletic group of green algae known as prasinophytes (*prasinos* is Greek for "green"). Long recognized to model Earth's earliest green algae, prasinophytes display early-evolved traits of Viridiplantae, the monophyletic lineage containing all green algae and land plants. Consequently, the study of prasinophytes is regarded as important in understanding the origin of plant traits upon which humans have come to depend; examples include chloroplastic starch as a cellular carbohydrate storage and cellulose-rich cell walls.

The following survey of *P. parkeae* taxonomy, morphology, evolutionary significance, and genomics is designed to provide an introduction to three subsequent thesis chapters. The first chapter focuses on comparative chloroplast genomics of *P. parkeae* strains, the second compares the mitochondrial genomes of *P. parkeae* and a closely-related prasinophyte genus, and the third addresses carbohydrate biosynthesis-related genes encoded in the nuclear genome of one strain of *P. parkeae*. In each case, the new information about *P. parkeae* is discussed within the overall context of Viridiplantae trait evolution.

Taxonomy and morphology of Pyramimonas parkeae

The species *Pyramimonas parkeae* was first described, on the basis of ultrastructural features, by Richard E. Norris and Barbara R. Pearson in 1975. This unicellular green alga is characterized by four flagella that emerge from an apical depression surrounded by four cytoplasmic lobes. The single chloroplast has four lobes that extend into the apical cytoplasmic lobes and a single pyrenoid whose surface is associated with starch biosynthesis (Figure 1). These lobes underlie the generic name *Pyramimonas*: *pyramis* is Greek for pyramid; *monas* is Greek for unit. The species name honors the eminent British phycologist (algae expert) Mary Parke (1908-1989), who helped to pioneer the use of electron microscopy to study microalgae. By contrast to most other Viridiplantae, but in common with many other prasinophytes, *Pyramimonas parkeae* lacks a cellulosic cell wall; instead, the body surface and flagella are covered by non-cellulosic scales consisting mainly of pectin-like polysaccharides and 2-keto sugar acids that also occur in plant cell walls. Chemically similar scales occur on the surfaces of related prasinophyte species (reviewed by Becker et al. 1994), and also on the surfaces of flagellate reproductive cells of more complex ulvophyceaean green algae and streptophyte green algae (Charophyceae) (e.g. Graham and McBride 1979), which likely inherited the ability to produce scales from prasinophyte-like ancestors. However, prasinophyte scales are not regarded to be precursors of the cellulose-rich cell walls that characterize most members of the Viridiplantae (Melkonian and Robenek 1981).

Another significant characteristic of *P. parkeae* is a vesicle-enclosed structure of a type generally known as the extrusome and more specifically as the ejectisome (Norris and Pearson 1975), because barb-shaped vesicle contents are discharged from cells in response to biological, chemical or physical stimuli. Ejectisome extrusion is thought to function as both a defensive and escape mechanism: release of ejectisomes may repel predators and at the same time propel cells in the opposite direction. At the ultrastructural level, the development and mature structure of ejectisomes present in *Pyramimonas* and some other prasinophytes are intriguingly similar to ejectisomes produced by many cryptomonads and the related plastidless protists known as katablepharids (Kugrens et al. 1994, Lee and Kugrens 1991). In light of phylogenomic evidence for close relationship of cryptomonads to the ancestry of Viridiplantae (Burki et al. 2016), this structural similarity makes sense. Although the chemical composition of such ejectisomes is not completely clear, and analogous cellular structures have not been identified in Viridiplantae other than prasinophytes, their presence may represent an ancestral feature of earliest Viridiplantae.

Evolutionary significance of Pyramimonas parkeae

Pyraminomas parkeae is classified within Pyramimonadales (clade I) of the paraphyletic green algal group informally known as Prasinophyceae, which includes approximately eight additional clades: Mamiellophyceae (clade II), Nephroselmidophyceae (clade III), Chlorodendrales (clade IV), Pycnococcaceae (clade V), Prasinococcales (clade VI, also known as Palmophyllophyceae class nov.), clade VII, clade VIII, and clade IX (Leliaert et al. 2016) (Figure 2). Clade numbers more or less reflect the order of discovery, rather than phylogenetic position. Whole genomes have been published only for several representatives of clade II, otherwise, only organellar genomes are known for various prasinophyte species.

As noted, prasinophytes are characterized by a collection of traits considered plesiomorphic for green algae; these features include organic body scales, perennation structures known as cysts, and occurrence in at least some Pyramimonadales of phagomixotrophy, a particle-feeding process that may have been the mechanism by which Viridiplantae acquired chloroplasts (Burns et al. 2015). Phylogenetic analyses using chloroplast, mitochondrial, and nuclear molecular data are congruent with morphology in resolving prasinophyte clades as the earliest-diverging extant green algae. Prasinophytes are early diverging green algae, which are phylogenetically close to the divergence of the main two green lineages – the clade consisting of chlorophyte algae (Trebouxiophyceae + Ulvophyceae + Chlorophyceae) (see Figure 2) and the clade consisting of streptophytes (land plants and their closest green algal relatives) (Leliaert et al. 2016). Therefore, prasinophytes are key to understanding the nature of the common ancestor of the Viridiplantae.

In order to better understand early Viridiplantae diversification, genetic components of *P. parkeae* were investigated because this species has been the subject of previous ultrastructural investigation and some molecular analyses. A culture of NIES254 obtained from the National Institute for Environmental Studies (Tsukuba, Japan) microbial culture collection was employed because no previous genomic work had been published for this *P. parkeae* strain. Shotgun sequencing was conducted and assembled sequence data were used to answer specific questions regarding organelle genomes and genomic aspects of carbohydrate metabolism. Chapter 2 of this thesis describes the *P. parkeae* NIES254 chloroplast genome, Chapter 3 describes the *P. parkea*e NIES254 mitochondrial genome, and Chapter 4 describes *P. parkeae* NIES254 nuclear genes associated with carbohydrate biosynthesis and deconstruction. Additional, future nuclear genomic work could be conducted with the shotgun sequence data in hand, but are outside the scope of this thesis project.

Chloroplast and mitochondrial genomes of P. parkeae *NIES254*

Chloroplast and mitochondrial genomes have generally been considered to be reasonable sources of molecular data for phylogenetic analyses. Genes in these organelle genomes are considered to be conserved because they encode crucial proteins participating in photosynthesis and cellular respiration. By contrast, nuclear genomic sequences display considerable variation within and between species.

Even so, previous investigators have noted that prasinophyte green algal chloroplast and mitochondrial genomes display higher than expected variation in organization, gene content, and gene order (Turmel et al. 1999, Robbens et al. 2007, Turmel et al. 2009, Worden et al. 2009, Turmel et al. 2010, Vaulot et al. 2012, Pombert et al. 2013, Turmel et al. 2013, Lemieux et al. 2014, Satjarak et al. 2016, Satjarak and Graham 2017). Surprising variation in chloroplast and mitochondrial genomes has been observed at the intra-specific level in the clade II prasinophyte *Ostreococcus tauri* (Blanc-Mathieu et al. 2013). To investigate whether intra-specific variation is likewise present in *P. parkeae*, chloroplast and mitochondrial genomes of *P. parkeae* NIES254 were assembled and compared to the chloroplast genome of *P. parkeae* CCMP726 (Turmel et al. 2009) and mitochondrial genome of *P. parkeae* SCCAP K-0007 (Hrdá et al. 2016). Our results showed that intra-specific variation occurs in both chloroplast and mitochondrial genomes of *P. parkeae*.

In addition, availability of the new *P. parkeae* NIES 254 chloroplast and mitochondrial genomes provided the opportunity to investigate whether the additional data could help to better resolve prasinophyte relationships. Results of phylogenetic analyses indicated that information from mitochondrial and/or chloroplast genomes was sufficient to indicate monophyly of known clades, but not relationships among prasinophyte clades.

Pyramimonas parkeae *genes encoding carbohydrate active enzymes*

The Carbohydrate Active enZymes (CAZymes) are enzymes that are involved in synthesis and breakdown of polysaccharides (http://www.cazy.org/, Cantarel et al. 2009). These enzymes are classified into four classes based on their enzymatic activities: Glycosyl Transferases (GTs), Glycoside Hydrolases (GHs), Polysaccharide Lyases (PLs), Carbohydrate Esterases (CEs), and a group of non-enzymatic carbohydrate binding modules (CBMs). GTs are responsible for synthesis of glycosides by catalyzing the formation of glycosyl bonds between a donor sugar substrate and another molecule. In contrast, GHs and PLs are responsible for the breakdown of these glycosidic linkages. GHs break down carbohydrate molecules by hydrolyzing the bonds between subunit sugars while PLs specifically cleave uronic acid-containing polysaccharide chains by β-elimination. CEs de-acetylate polysaccharide side-chains and are thought to modify the cross-linking of hemicellulose with lignin. CBMs allow for specific binding to different carbohydrate biopolymers, thereby facilitating precise biopolymer modification before addition to the cell wall (Cantarel et al. 2009).

In plants, CAZymes are particularly important for synthesis of the cell wall – a rigid layer of polysaccharides lying outside the plasma membrane that represents the majority of plant biomass and is of considerable economic importance. For this reason, many studies have been done to understand the metabolism of plant cell wall components – cellulose, hemicellulose, and other polysaccharides (e.g., Geisler-Lee et al. 2006, Popper et al. 2011, Kumar and Turner 2015).

Cellulose is the main component of the plant primary cell wall. In land plants and some streptophyte algae, this polysaccharide is synthesized by cellulose synthesizing complexes (CSCs) that occur in rosette assemblages at the cell membrane. Cellulose synthase (CesA) is the protein that is the main component of CSCs. CesA contains catalytic regions for synthesizing cellulose microfibrils and protein domains important for CesA-CesA interactions involved in forming the rosette structure (Kumar and Turner 2015).

In addition to CesA, chlorophyte algae and at least some embryophytes are known to also produce a bacterial-type cellulose synthase (Bcs). This protein forms CSC at the cell membrane, but bacterial-type CSCs occur in linear (not rosette) arrays (Tsekos 1999, Romling 2002, Harholt et al. 2012, Ulvskov et al. 2013, Mikkelsen et al. 2014). The presence of both proteins and their phylogenies suggest that Viridiplantae might have acquired the Bcs protein from cyanobacterial endosymbiont ancestral to green plastids (Nobles et al. 2001). Some investigators suggest that Bcs later diverged into derivative CesA and Bcs protein lineages sometimes during the evolution of streptophytes (Mikkelsen et al. 2014). If this evolutionary scenario is correct, the last common ancestor of Viridiplantae should contain a genomic sequence similar to Bcs.

To date, five prasinophyte genomes, all classified as prasinophyte clade II, have been fully and partially sequenced – *Ostreococcus tauri*, *O. lucimarinus*, *Bathycoccus prasinos*, and two *Micromonas spp.* (Derelle et al. 2006, Palenik et al. 2007, Worden et al. 2009, Moreau et al. 2012). All contain a nuclear sequence similar to Bcs, though the functions of potential protein products have not as yet been investigated. These protein sequences have not generally been included in studies of the evolutionary diversification of cellulose synthase proteins (e.g., Mikkelsen et al. 2014) because prasinophyte algae have not been demonstrated to produce cell wall cellulose microfibrils. The present study sought to find evidence for Bcs protein-encoding genes in the nuclear genome of *P. parkeae* and if found to incorporate them into a phylogenetic analysis.

In addition, the availability of *P. parkeae* nuclear genome sequences and publically accessible transcriptomic data provide the opportunity to investigate other proteins involved in plant cell wall production, e.g. xyloglucan and pectin. These sequence data sets also provide an opportunity to investigate proteins that are involved in metabolism of starch – an important feature of Viridiplantae.

Summary

Whole genomic DNA was obtained for the wall-less clade I prasinophyte *P. parkeae* NIES254 and used to assemble chloroplast, mitochondrial, and draft nuclear genome. Comparative analyses of organelle genomes from NIES254 and other clade I strains showed that *P. parkeae* organelle genomes exhibit surprisingly high variation at the intra-specific level and though useful for demonstrating clade monophyly, do not help to resolve higher-level prasinophyte relationships. In addition, to shed light on the evolution of Viridiplantae cell walls and starch, carbohydrate active enzymes were inferred from *P. parkeae* NIES254 draft nuclear genome sequence. Sequences homologous to genes that encode plant proteins having known functions in the production of cell wall components and storage compounds were described.

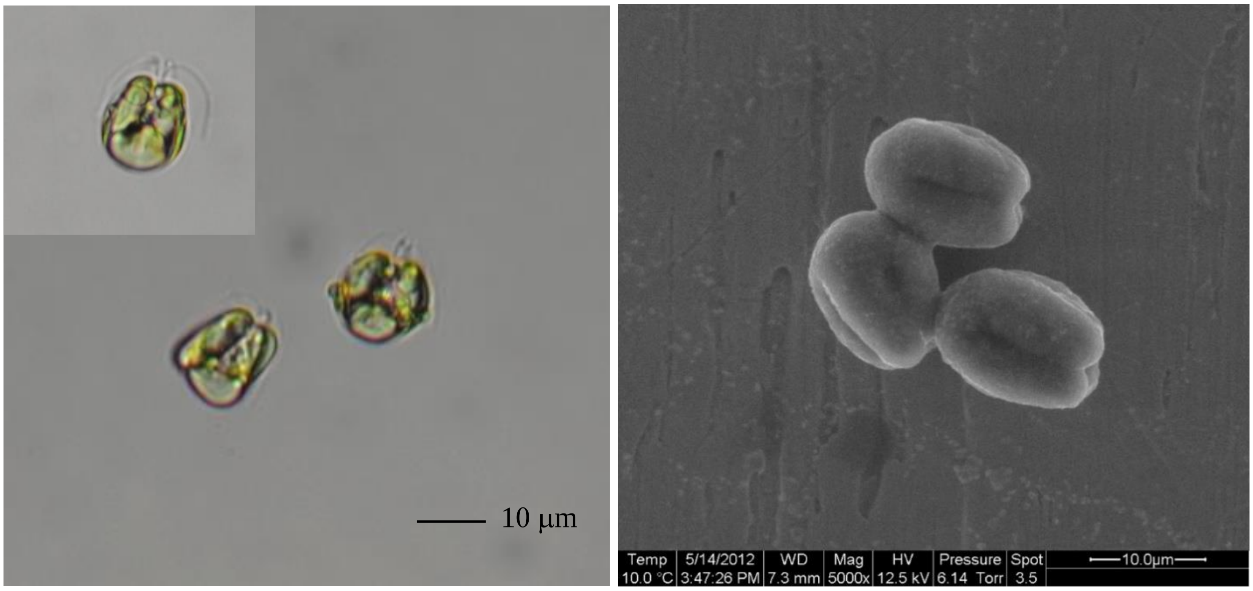


Figure. 1 *Pyramimonas parkeae*. Left, light microscopy reveals four-lobed cells having four flagella emerging from an apical depression. Right, electron microscopy likewise shows four-lobed cellular structure though flagella were lost during preparation.

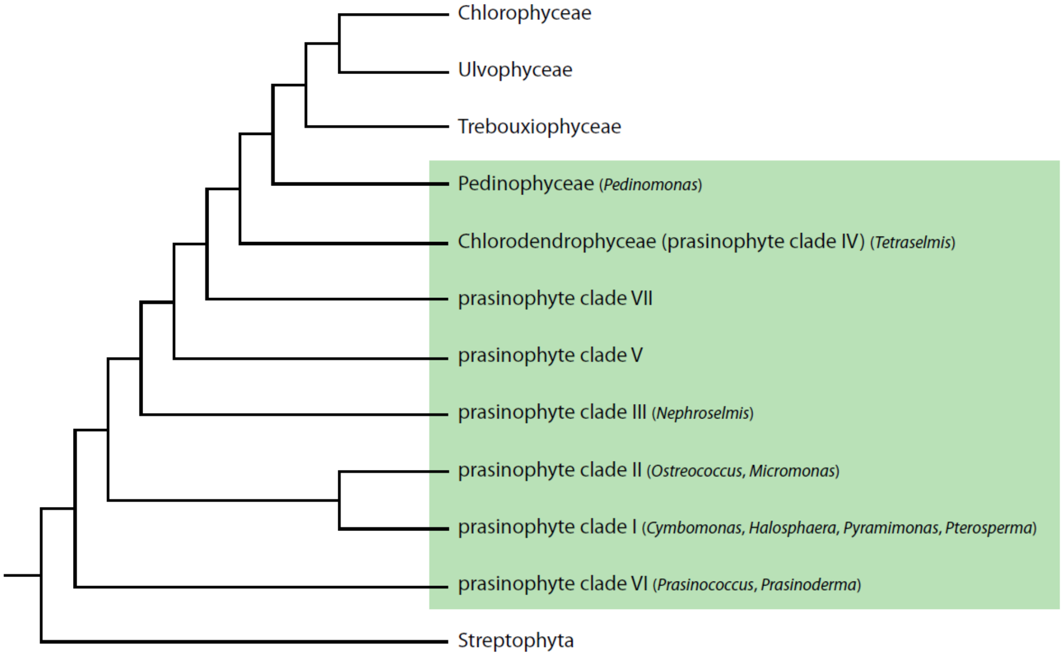

Chlorophyceae
Ulvophyceae
Trebouxiophyceae
Pedinophyceae (*Pedinomonas*)
Chlorodendrophyceae (prasinophyte clade IV) (*Tetraselmis*)
prasinophyte clade VII
prasinophyte clade V
prasinophyte clade III (*Nephroselmis*)
prasinophyte clade II (*Ostreococcus, Micromonas*)
prasinophyte clade I (*Cymbomonas, Halosphaera, Pyramimonas, Pterosperma*)
prasinophyte clade VI (*Prasinococcus, Prasinoderma*)
Streptophyta

Figure 2. Relationship of prasinophytes and core chlorophytes and streptophytes. Organisms in the green box have traditionally been considered to be prasinophytes (Graham et al. 2016, original diagram based on information in Lemieux et al. 2014).

References

Becker, B., Marin, B. & Melkonian, M. 1994. Structure, composition, and biogenesis of prasinophyte cell coverings. *Protoplasma* 181(1-4):233-244.

Blanc-Mathieu, R., Sanchez-Ferandin, S., Eyre-Walker, A. & Piganeau, G. 2013. Organellar inheritance in the Green Lineage: insights from *Ostreococcus tauri*. *Genome Biol. Evol.* 5(8):1503-1511.

Burki, F., Kaplan, M., Tikhonenkov, D.V., Zlatogursky, V., Minh, B.Q., Radaykina, L.V., Smirnov, A., Mylnikov, A.P. & Keeling, P.J. 2016. Untangling the early diversification of eukaryotes: a phylogenomic study of the evolutionary origins of Centrohelida, Haptophyta and Cryptista. *Proc. R. Soc. B* 283(1823):20152802

Burns, J.A., Paasch, A., Narechania, A. & Kim, E., 2015. Comparative genomics of a bacterivorous green alga reveals evolutionary causalities and consequences of phago-mixotrophic mode of nutrition. *Genome Biol. Evol.* 7(11):3047-3061.

Cantarel, B.L., Coutinho, P.M., Rancurel, C., Bernard, T., Lombard, V. & Henrissat, B. 2009. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. *Nucleic Acid Res.* 37(suppl 1):D233-D238.

Derelle, E., Ferraz, C., Rombauts, S., Rouzé, P., Worden, A.Z., Robbens, S., Partensky, F., Degroeve, S., Echeynié, S., Cooke, R. & Saeys, Y. 2006. Genome analysis of the smallest free-living eukaryote *Ostreococcus tauri* unveils many unique features. *Proc. Natl. Acad. Sci. U.S.A.* 103(31):11647-11652.

Geisler-Lee, J., Geisler, M., Coutinho, P.M., Segerman, B., Nishikubo, N., Takahashi, J., Aspeborg, H., Djerbi, S., Master, E., Andersson-Gunnerås, S. & Sundberg, B. 2006. Poplar carbohydrate-active enzymes. Gene identification and expression analyses. *Plant Physiol.* 140(3):946-962.

Graham, L.E., Graham, J.M., Wilcox, L.W., Cook, M.E. 2016. Algae. 3rd ed.

LJLM Press, Madison, WI, 595 pp.

Graham, L.E. & McBride, G.E. 1979. The occurrence and phylogenetic significance of a multilayered structure in *Coleochaete spermatozoids*. *Am. J. Bot.* 66(8):887-894.

Harholt, J., Sørensen, I., Fangel, J., Roberts, A., Willats, W.G., Scheller, H.V., Petersen, B.L., Banks, J.A. & Ulvskov, P. 2012. The glycosyltransferase repertoire of the spikemoss *Selaginella moellendorffii* and a comparative study of its cell wall. *PloS ONE* 7(5):e35846.

Hrdá, Š., Hroudová, M., Vlček, Č. & Hampl, V. 2016. Mitochondrial Genome of Prasinophyte Alga *Pyramimonas parkeae*. *J. Euk. Microbiol.* 0:1-10.

Kugrens, P., Lee, R.E. & Corliss, J.O. 1994. Ultrastructure, biogenesis, and functions of extrusive organelles in selected non-ciliate protists. *Protoplasma* 181(1-4):164-190.

Kumar, M. & Turner, S. 2015. Plant cellulose synthesis: CESA proteins crossing kingdoms. *Phytochemistry* 112:91-99.

Lee, R.E. & Kugrens, P. 1991. *Katablepharis ovalis,* a colorless flagellate with interesting cytological characteristics. *J. Phycol.* 27(4):505-513.

Leliaert, F., Tronholm, A., Lemieux, C., Turmel, M., DePriest, M.S., Bhattacharya, D., Karol, K.G., Fredericq, S., Zechman, F.W. and Lopez-Bautista, J.M. 2016. Chloroplast phylogenomic analyses reveal the deepest-branching lineage of the Chlorophyta, Palmophyllophyceae class. nov. *Sci.Rep*. 6:25367

Lemieux, C., Otis, C. & Turmel, M. 2014. Six newly sequenced chloroplast genomes from prasinophyte green algae provide insights into the relationships among prasinophyte lineages and the diversity of streamlined genome architecture in picoplanktonic species. *BMC genomics* 15(1):1.

Melkonian, M. & Robenek, H. 1981. Comparative ultrastructure of underlayer scales in four species of the green flagellate *Pyramimonas*: a freeze-fracture and thin section study. *Phycologia 20*(4):365-376.

Mikkelsen, M.D., Harholt, J., Ulvskov, P., Johansen, I.E., Fangel, J.U., Doblin, M.S., Bacic, A. & Willats, W.G. 2014. Evidence for land plant cell wall biosynthetic mechanisms in charophyte green algae. *Ann. Bot.* 114(6):1217-1236.

Moreau, H., Verhelst, B., Couloux, A., Derelle, E., Rombauts, S., Grimsley, N., Van Bel, M., Poulain, J., Katinka, M., Hohmann-Marriott, M.F. and Piganeau, G., 2012. Gene functionalities and genome structure in *Bathycoccus prasinos* reflect cellular specializations at the base of the green lineage. *Genome Biol.*13(8):R74.

Nobles, D.R., Romanovicz, D.K. & Brown, R.M., 2001. Cellulose in cyanobacteria. Origin of vascular plant cellulose synthase?. *Plant Physiol.* 127(2):529-542.

Palenik, B., Grimwood, J., Aerts, A., Rouzé, P., Salamov, A., Putnam, N., Dupont, C., Jorgensen, R., Derelle, E., Rombauts, S. & Zhou, K. 2007. The tiny eukaryote *Ostreococcus* provides genomic insights into the paradox of plankton speciation. *Proc. Natl. Acad. Sci. U.S.A.* 104(18):7705-7710.

Pearson, B.R. & Norris, R.E. 1975. Fine structure of cell division in *Pyramimonas parkeae* Norris and Pearson (Chlorophyta, Prasinophyceae). *J. Phycol.*11(1):113-124.

Pombert, J.F., Otis, C., Turmel, M. & Lemieux, C. 2013. The mitochondrial genome of the prasinophyte *Prasinoderma coloniale* reveals two trans-spliced group I introns in the large subunit rRNA gene. *PloS ONE* 8(12):e84325.

Popper, Z.A. & Fry, S.C. 2003. Primary cell wall composition of bryophytes and charophytes. *Ann. Bot.* 91(1):1-12.

Robbens, S., Derelle, E., Ferraz, C., Wuyts, J., Moreau, H. & Van de Peer, Y. 2007. The complete chloroplast and mitochondrial DNA sequences of *Ostreococcus tauri*: organelle genomes of the smallest eukaryote are examples of compaction. *Mol. Biol. Evol.* 24(4):956-968.

Römling, U. 2002. Molecular biology of cellulose production in bacteria. *Res. Microbiol.* 153(4):205-212.

Satjarak, A., Paasch, A.E., Graham, L.E. & Kim, E. 2016. Complete chloroplast genome sequence of phagomixotrophic green alga *Cymbomonas tetramitiformis*. *Genome Announc*. 4(3):e00551-16.

Satjarak. A. & Graham, L.E. 2017. Comparative DNA sequences analyses of *Pyramimonas parkeae* (Prasinophyceae) chloroplast genomes. *J. Phycol.* In press.

Tsekos, I. 1999. The sites of cellulose synthesis in algae: diversity and evolution of cellulose-synthesizing enzyme complexes. *J. Phycol.* 35(4):635-655.

Turmel, M., Otis, C. & Lemieux, C. 1999. The complete chloroplast DNA sequence of the green alga *Nephroselmis olivacea*: insights into the architecture of ancestral chloroplast genomes. *Proc. Natl. Acad. Sci. U.S.A.* 96(18):10248-10253.

Turmel, M., Gagnon, M. C., O'Kelly, C. J., Otis, C. & Lemieux, C. 2009. The chloroplast genomes of the green algae *Pyramimonas*, *Monomastix,* and *Pycnococcus* shed new light on the evolutionary history of prasinophytes and the origin of the secondary chloroplasts of euglenids. *Mol. Biol. Evol.* 26(3):631-648.

Turmel, M., Otis, C. & Lemieux, C. 2010. A deviant genetic code in the reduced mitochondrial genome of the picoplanktonic green alga *Pycnococcus provasolii*. *J. Mol. Evol.* 70(2):203-214.

Turmel, M., Otis, C. & Lemieux, C. 2013. Tracing the evolution of streptophyte algae and their mitochondrial genome. *Genome Biol. Evol.* 5(10):1817-1835.

Ulvskov, P., Paiva, D.S., Domozych, D. & Harholt, J. 2013. Classification, naming and evolutionary history of glycosyltransferases from sequenced green and red algal genomes. *PloS ONE* 8(10):e76511.

Vaulot, D., Lepere, C., Toulza, E., De la Iglesia, R., Poulain, J., Gaboyer, F., Moreau, H., Vandepoele, K., Ulloa, O., Gavory, F. & Piganeau, G. 2012. Metagenomes of the picoalga *Bathycoccus* from the Chile coastal upwelling. *PLoS ONE* 7(6):e39648.

Worden, A.Z., Lee, J.H., Mock, T., Rouzé, P., Simmons, M.P., Aerts, A.L., Allen, A.E., Cuvelier, M.L., Derelle, E., Everett, M.V. & Foulon, E. 2009. Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes *Micromonas*. *Science*

324(5924):268-272.

CHAPTER 2: COMPARATIVE DNA SEQUENCE ANALYSES OF *PYRAMIMONAS PARKEAE* (PRASINOPHYCEAE) CHLOROPLAST GENOMES[1]

Anchittha Satjarak[2]

Department of Botany, University of Wisconsin-Madison, 430 Lincoln drive, Madison,
Wisconsin, USA

[2]corresponding author: e-mail: satjarak@wisc.edu, phone: +16082620657, fax: +16082627509

Linda E. Graham

Department of Botany, University of Wisconsin-Madison, 430 Lincoln drive, Madison,
Wisconsin, USA

Running title: Comparative analyses of *Pyramimonas parkeae* chloroplast genomes

Abstract

Prasinophytes form a paraphyletic assemblage of early diverging green algae, which have the potential to reveal the traits of the last common ancestor of the main two green lineages: 1) chlorophyte algae and 2) streptophyte algae. Understanding the genetic composition of prasinophyte algae is fundamental to understanding the diversification and evolutionary process that may have occurred in both green lineages. In this study, we sequenced the chloroplast genome of *Pyramimonas parkeae* NIES254 and compared it with that of *P*. *parkeae* CCMP726, the only other fully sequenced *P*. *parkeae* chloroplast genome. The results reveal that *P*. *parkeae* chloroplast genomes are surprisingly variable. The chloroplast genome of NIES254 is larger than that of CCMP726 by 3,204 bp. NIES254 LSC is 288 bp longer, the SSC is 5,088 bp longer, and the IR is 1,086 bp shorter than that of CCMP726. Similarity values of the two strains are almost zero in four large hot spot regions. Finally, the strains differ in copy number for three protein coding genes: *ycf20*, *psaC*, and *ndhE*. Phylogenetic analyses using 16S and 18 S rDNA and *rbcL* sequences resolved a clade consisting of these two *P*. *parkeae* strains and a clade consisting of these plus other *Pyramimonas* isolates. These results are consistent with past studies indicating that prasinophyte chloroplast genomes display a higher level of variation than is commonly found among land plants. Consequently, prasinophyte chloroplast genomes may less useful for inferring the early history of Viridiplantae than has been the case for land plant diversification.

Key index words: chloroplast DNA variation; chloroplast genome; intraspecific variation; prasinophyte; *Pyramimonas parkeae*

Abbreviations: NIES, Microbial Culture Collection at the National Institute for Environmental Studies; SNPs, single nucleotide polymorphisms.

Introduction

Prasinophytes are early diverging green algae that show heterogeneity in morphology and the potential to elucidate the traits of the last common ancestor of the main two green (Viridiplantae) lineages: 1) chlorophyte algae, the clade representing the majority of present green algal diversity and 2) streptophyte algae, a smaller, paraphyletic algal assemblage known to be closely related to land plants. The independent lineages currently recognized are Pyramimonadales (clade I), Mamiellophyceae (clade II), Nephroselmidophyceae (clade III), Chlorodendrales (clade IV), Pycnococcaceae (clade V), Prasinococcales (clade VI, Palmophyllophyceae class nov., which was recently hypothesized to be closest to the divergence of chlorophyte and streptophyte algae, Leliaert et al. 2016), clade VII, clade VIII, and clade IX (Lemieux et al. 2014).

Chloroplast genome sequences have been sources of data for plant phylogenetics because chloroplasts exhibit uni-parental inheritance and a slow rate of mutation (Wolfe et al. 1987, Allender et al. 2007). Comparative studies of the majority of plant plastid genome architectures show only small variation; gene order and essential gene content are highly conserved in plant plastid genomes (De Las Rivas et al. 2002). However, by comparison to plants, green algal chloroplast genomes seem to evolve in a much less conservative fashion. Green algal plastid genomes show variation in gene order, genome length, and presence of quadripartite structure (Brouard et al. 2010, Turmel et al. 2015, Lemieux et al. 2016, Turmel et al. 2016) Comparisons of 12 chloroplast genomes from 6 prasinophyte clades revealed that prasinophyte chloroplast genomes likewise display variability in organization, gene content, and gene order. (Turmel et al. 1999, Robbens et al. 2007, Turmel et al. 2009, Worden et al. 2009, Lemieux et al. 2014).

The presence of this high variation among chloroplast genomes of prasinophyte clade II (Mamiellophyceae) suggests that variability may also occur at the intra-specific level. Comparison of 13 *Ostreococcus tauri* strains showed intra-specific variation in SNPs, single nucleotide polymorphisms, and presence of large insertion/deletion regions (Blanc-Mathieu et al. 2013). These observations indicate that similar surprising levels of intra-specific variation in chloroplast sequences might occur in other prasinophyte clades, but so far that possibility has not been investigated. We evaluated the level of intra-specific variation in chloroplast sequence in *Pyramimonas parkeae* (R.E. Norris & B.R. Pearson) representing prasinophyte clade I, which has a conserved quadripartite structure and is closely related to divergence of streptophytes.

The complete chloroplast genome sequence of *P. parkeae* CCMP726 was released in 2009 by Turmel et al. For comparative analyses, we assembled the complete *P. parkeae* NIES254 chloroplast genome, a closely related strain obtained from National Institute for Environmental Studies (NIES), Japan. Our comparison showed that the two chloroplast genomes have identical gene content and similar arrangement. However, we found evidence for four large hotspot regions, movement of inverted repeat boundaries, gene copy number difference.

Materials and methods

DNA extraction

A culture of *Pyramimonas parkeae* Norris and Pearson (NIES254) was acquired from National Institute of Environmental Studies, Japan (NIES). The culture was propagated in Alga-Gro® seawater medium (Carolina Biological Supply Company, Burlington, NC, USA), and was maintained in a walk-in growth room with 16:8 daily light/dark cycle at 20˚C. Cells were harvested during the exponential phase. Total DNA was prepared by using FastDNA® SPIN Kit for Soil (MP Biomedicals, Solon, OH) and sequenced by Illumina Miseq technologies at the University of Wisconsin-Madison Biotechnology Center.

Data pre-processing and genome construction

The raw paired-end Illumina data consisted of 13,232,998 reads with average read length of 251 bp. The data were trimmed by Trimmomatic v 0.33 (Bolger et al. 2014) in order to obtain the quality score of at least 28 on the phred 64 scale. The chloroplast genome was initially constructed using *de novo* sequence assembly, which proved challenging because repeat regions were longer than individual reads. Therefore, we employed a baiting and iterative mapping method described in Satjarak et al. (2016) using MIRA v 4.0.2 and MITObim v 1.8 (Hahn et al. 2013). Protein coding sequences of *P. parkeae* CCMP726 chloroplast genome (Turmel et al. 2009) available in GenBank (accession number: FJ493499.1) were used as baits.

Sequence analyses

To determine the chloroplast genome coverage, we aligned the trimmed reads against the newly constructed NIES254 chloroplast genome by using BWA non-model species alignment v 0.7.4 (Li and Durbin 2009) and calculated the coverage of every position in the plastid genome by using Bedtools Genome Coverage BAM v 2.19.1 (Quinlan 2014) implemented in iPlant Collaborative (Goff et al. 2016). The functions of the open reading frames (ORFs) with length of at least 100 bp were predicted by using BLAST search against the NCBI non-redundant protein databases accessed in February 2016 (http://blast.ncbi.nlm.nih.gov/Blast.cgi). tRNAs and rRNAs were predicted using tRNAscan-SE v 1.21 (Schattner et al. 2005) and RNAmmer v 1.2 (Lagesen et al. 2007). Base frequencies, amino acid frequencies, and codon usage were calculated using statistics option in Geneious v 9.0.4 (Kearse et al. 2012). The circular genome was drawn using OGDraw v 1.2 (Lohse et al. 2013). The resulting annotated sequence has been deposited at the GenBank under accession number KX013546.

Relationship between Pyramimonas parkeae NIES254 and CCMP726

We used a phylogenetic approach to assess the relationship between *P. parkeae* NIES254 and CCMP726. The 18S rDNA of *P. parkeae* NIES254 was constructed from pair-end Mi-Seq reads sequenced from whole genomic DNA of *P. parkeae* NIES254 using methods described in Satjarak et al. (2016). The gene was assembled using 18S rDNA of *P. parkeae* Hachijo (accession number AB017124, Nakayama et al. 1998) as a bait. The average coverage of the sequence was estimated using BWA non-model species alignment v 0.7.4 (Li and Durbin 2009) and Bedtools Genome Coverage BAM v 2.19.1 (Quinlan 2014) implemented in iPlant

Collaborative (Goff et al. 2016). 18S rDNA was predicted using RNAmmer v 1.2 (Lagesen et al. 2007). The final 18S rDNA construct was a linear molecule of 1,802 bp. The average coverage of every position of the gene was 144 fold. The resulting annotated sequence has been deposited at the GenBank under accession number KX611141.

To assess the relationship between the two strains, we performed phylogenetic analyses of 3 genes: 18S rDNA, 16S rDNA, and *rbcL*. *Pyramimonas* 18S rDNA, 16S rDNA, and *rbcL* sequences publicly available in GenBank (accessed in July 2016) were used in the analyses. *Cymbomonas tetramitiformis* DNA sequences of corresponding genes were used as the outgroups.

The accession numbers of 18S rDNA sequences used in the phylogenetic analyses were FN562438, AB017126, AB052289, AJ404886, FN562440, AB017121, HQ111511, HQ111509, HQ111510, KF422615, FN562442, AB017122, KF615765, FN562443, KT860881, AB017124, AB017123, FN562441, AB999994, AB853999, AB854000, AB854001, AB854002, KF899837, AB854003, AB854004, AB854006, AB854005, AB854007, AB854008, AB854009, AB854010, AB854011, AB854012, AB854013, AB854014, AB854016, AB854015, AB854017, AB854018, AB854021, AB854020, AB854019, AB854022, AB854023, AB854024, AB854025, JN934670, JF794047, JF794048, JN934689, KT860923, AB854026, AB854027, AB854028, AB854029, AB854030, AB854031, AB854032, AB854033, AB854034, AB854035, AB854036, AB854037, AB854038, AB854039 and KX611141 (Nakayama et al. 1998, Moro et al. 2002, Duanmu et al. 2004, Suda 2004, Marin and Melkonian 2010, Balzano et al. 2012, Suda et al. 2013, Bhuiyan et al. 2015). The accession numbers of 16S rDNA sequences used in the phylogenetic analyses were AF393608, L34687, LK391817, LK391818, K391819, LK391820, LN735316, LN735321, LN735377, LN735378, LN735435, KX013545.1, FJ493499.1, and KX013546 (Daugbjerg et al.

1994, Turmel et al. 2002, Turmel et al. 2009, Decelle et al. 2015, Satjarak et al. 2016). The accession numbers of *rbcL* sequences used in the phylogenetic analyses were AB052290, L34776, L34814, L34819, L34779, L34810, L34812, L34811, L34817, L34815, L34816, L34813, L34777, L34833, L34778, LC015748, LC015747, L34834, KP096399, L34818, KX013545.1, FJ493499.1, and KX013546 (Daugbjerg et al. 1994, Suda 2004, Bhuiyan et al. 2015, Satjarak et al. 2016).

We aligned the sequences using Geneious; setting free end gaps and identity to (1.0/0.0) resulted in 1,922 bp unambiguously aligned sequences of 18S rDNA, 706 bp of 16S rDNA, and 1,089 bp of *rbcL*. For each gene, the nucleotide substitution model was computed using jModelTest2 (Darriba et al. 2012). Maximum-Likelihood (ML) analysis was performed using RAxML (v 8.2.8) (Stamatakis 2014) on the CIPRES XSEDE Portal (Miller et al. 2010) using a GTR + I + F substitution model, employing the rapid bootstrapping method with 1000 replications for bootstrap analyses. Baysian analyses were performed with MrBayes v 3.2.6 (Ronquist and Huelsenbeck 2003) using a GTR + I + F substitution model. Four independent chains were run for 1,100,000 cycles and the consensus topologies were calculated after the burn-in of 100,000 cycles.

Comparative analysis of Pyramimonas parkeae chloroplast genomes

The analysis of syntenic conservation between *P. parkeae* NIES254 and CCMP726 was performed using progressiveMauve alignment v 2.4.0 (Darling et al. 2010). The two genomes were also aligned using LAST (Kiełbasa et al. 2011), with the following parameters: maximum score, max multiplicity for initial matches = 10, minimum length for initial matches = 1, step-size along reference sequences = 1, step-size along query sequences = 1, query letters per random alignment = 1e6. SNPs within the whole genome and within the protein coding regions were identified using Geneious alignments v 9.0.4 (Kearse et al. 2012). Synonymous (Ks) and nonsynonymous (Ka) substitution sites as well as the Ka/Ks ratio were calculated using MEGA6 v 6.06 (Tamura et al. 2013). To compare variability at the intraspecific level, we also calculated Ka/Ks ratios of protein coding sequences of *Ostreococcus tauri*.

Results

Chloroplast genome of P. *parkeae NIES254*

We sequenced the chloroplast genome of *Pyramimonas parkeae* NIES254 for comparison to *P. parkeae* CCMP726 to investigate intra-specific genetic diversity. The newly sequenced *P. parkeae* NIES254 chloroplast genome was observed to have quadripartite structure of a 104,809 bp-long mapping circular molecule. The genome featured two copies of the inverted repeat (IR; 11,971 bp encompassing 22.84% of the genome), which separated the large single copy region (LSC; 65,441 bp) from the small single copy region (SSC; 15,426 bp) (Fig.1). The coverage of every position of the chloroplast genome ranged from 286 to 939 fold. GC content was 34.2%. The coding capacity of NIES254 was the same as that of CCMP726. This NIES254 chloroplast genome encoded 112 conserved genes including 2 rRNAs, 25 tRNAs, and 85 protein coding genes (Fig. 1). The genome of NIES254 was longer than that of CCMP726 by 3,204 bp. NIES254 LSC was 288 bp longer, the SSC was 5,088 bp-longer, and the IR was 1,086 bp shorter than that of CCMP726.

Relationship between Pyramimonas parkeae NIES254 and CCMP726

ML and Baysian assessments of 18S and 16S rDNA and *rbcL* sequences to infer the relationship between *P. parkeae* NIES254 and CCMP726 resulted in the same tree topology. All of these gene trees resolved a monophyletic clade of *P. parkeae* strains (Fig. 2-4).

Comparative analyses of Pyramimonas parkeae chloroplast genomes

Mauve alignment analysis of synteny of the two chloroplast genomes —*P. parkeae* strains NIES254 and CCMP726 showed that these genomes exhibited a collinear relationship, as only one syntenic block from each strain was present (Fig. 5). Although the genomes were collinear, the Mauve alignment showed four large hotspot regions where similarity values were almost zero. Such regions could be classified into 3 categories: 1) the 6 kb intergenic region between *psbA-trnS* and *ndhB* in LSC, 2) 2 kb intron of *atpB*, and 3) boundaries of IR and SSC, 5.7 kb at IRB-SSC and 6.8 kb at SSC-IRA (Fig. 2).

The size of the first large hotspot region (located between *psbA-trnS* and *ndhB* in LSC) was 6,848 bp in NIES254 and 6,240 bp in CCMP726. This intergenic region, which represented the largest intergenic region, contained different ORFs. However, we did not find orthologous protein products of the ORFs by similarity searches between the two genomes and against the non-redundant protein databases. A second hotspot region was the intron of *atpB* gene. The region was 2,144 bp long in NIES254 and 2,757 bp long in CCMP726. Both were group II introns with conserved region of reverse transcriptases of group II intron origin. The third and the fourth large hotspots occurred at the border between IRB-SSC and SSC-IRA. These hotspot regions resulted from boundary movement. The shift observed at the IRB-LSC boundary was minor, but the shift at the IRA-SSC boundary was greater. These movements and nucleotide variation at the boundaries resulted in difference in length and presence of ORFs and genes within IRB-SSC-IRA region of the two algae. Only one of the hypothetical ORFs (orf 454, 1,365 bp) in the CCMP726 IRs was present in those of NIES254 but in the latter it was fragmented into three separated ORFs having lengths of 126, 402 and 186 bp. Also, the boundary movement caused re-positioning and change in copy number of three genes: *ycf20, psaC,* and *ndhE*. In CCMP726, these three genes were present on both copies of the IR region, whereas in NIES254 they were present on one end of the SSC region.

Variability between the two chloroplast genomes was present in all of three informative regions: 1) protein coding regions, 2) intronic regions, and 3) intergenic regions. Alignment of NIES254 and CCMP726 chloroplast using LAST resulted in 93 similar regions due to the high variability present in the intergenic regions. This high variability made it challenging to identify the variable positions throughout the whole genome. Therefore, we were only able to perform comparative analyses of protein coding regions.

The total number of polymorphic sites in protein coding genes and ORF (*orf91*) was 3,111 positions including 2,684 SNPs and 44 indels. *ftsH* exhibited the highest number of SNPs and indels: 246 SNPs (7 positions per 100 bpbase pair) and 6 indels. The number of substitutions per 100 nucleotides of protein coding genes and *orf91* showed that mutations were randomly distributed across the chloroplast genomes (Fig. 6). Among plastid coding sequences, *psbT* possessed the highest Ks (57 positions per 100 nucleotides), *ftsH* possessed the highest Ka (16 positions per 100 nucleotides), and *petA* exhibited the highest Ka/Ks ratio (1.00) (Fig. 6).

Discussion

The availability of a sequenced chloroplast genome for *P. parkeae* NIES254 provided the opportunity for comparative analysis of chloroplast genome structure between *P. parkeae* strains NIES254 and CCMP726. These *P. parkeae* chloroplast genomes were similar in gene content. Of three ORFs (*orf91*, *orf454*, *orf608*) present in CCMP726, two were also present in NIES254, though o*rf454* present as a single unit in CCMP726 was fragmented into three separate pieces in NIES254, and o*rf608* present in the *atpB* introns of both genomes differed in nucleotide sequence. These differences in non-coding sequences may reflect lower constraint than experienced by coding regions. In the prasinophyte species *Ostreococcus tauri* a group II intron similarly evolved rapidly, resulting in sequence loss in some strains (Blanc-Mathieu et al. 2013).

Comparison of these two *Pyramimonas parkeae* plastid genomes also indicated IRB-SSC and SSC-IRA boundary movement. The expansion and contraction of the IR regions at the inter-specific level is not uncommon (Goulding et al. 1996). Comparative plastome studies in embryophyte families showed that boundaries between the IR and single copy regions are not static, but rather have been subjected to dynamic and random processes that allow the conservative expansion and contraction of IR regions. Movement of IR boundaries is likely to be unique for each species, and hypothesized to reflect relationship among embryophyte families (Zhu et al. 2015, Wang et al. 2016) and contribute to the expansion of the genome (Dugas et al. 2015, Zhu et al. 2016).

Most observations of boundary movements have involved boundary shifts at IRA-LSC and LSC-IRB, which have been hypothesized to be lineage-specific and tend to be minor; across the embryophytes, most such shifts resulted in loss and gain of a few nucleotides or partial genes, which often gave rise to a pseudogene at one end of the borders. While movements of IR-LSC boundaries have been known to be evolutionary markers, IR-SSC boundaries of closely related embryophyte species tend to be static, with shifts involving only a few nucleotides (Zhu et al. 2015). Extreme cases included 1) the medicinal plant *Eucommia ulmoides*, where the IR was expanded by 5 kb in comparison to other angiosperms, and 2) the legumes *Acacia* and *Inga*, where the IR was expanded by 13 kb. The boundary shifts in *Eucommia ulmoides* were hypothesized to be the result of genome rearrangement, whereas the shifts in the legumes were accompanied by the presence of an increased number of tandem repeats in the genome (Dugas et al. 2015, Wang et al. 2016).

The chloroplast genome IR regions of *P. parkeae* genomes are similar to those of other green algae and embryophytes in clustering *rrl* and *rrs.* However, the IR boundaries of prasinophytes seem less stable (e.g. Turmel et al. 2009) than in of most land plants (Zhu et al. 2015, Zhu et al. 2016). Our study provides an example of such instability in the form of evidence for boundary movement that has affected both length and copy number of some genes.

A more complete understanding of the mechanism underlying intra-specific plastid genome contraction/expansion will require analysis of additional *Pyramimonas* strains. However, we can speculate about processes that may have been involved in their origin. Contraction of the IRs might be as simple as DNA deletion in one IR copy. This deletion would leave one copy of the IR nucleotides on either LSC or SSC. A more complicated scenario would be IR expansion, which might arise from repair after a double-strand DNA break (Goulding et al. 1996).

Synonymous (Ks) and nonsynonymous (Ka) substitution sites as well as the Ka/Ks ratio calculated from protein coding sequence and a common ORF from *P. parkeae* NIES254 and CCMP726 suggested that mutation in these chloroplast genomes occurred in a random fashion. This contrasts with results of some other studies (Ogihara and Tsunewaki 1988, Birky and Walsh 1992, Zhu et al. 2016), where the observed substitution rates in IR regions were lower than in single copy regions. However, the NIES254 and CCMP726 *P. parkeae* IRs contain *rrl*, *rrs*, and tRNAs clusters that are highly conserved. Therefore, if we include rRNAs and tRNAs in the analyses, the mutation rate will be relatively lower in the inverted repeat regions. This depressed substitution rate in the IR regions is hypothesized to provide copy-dependent repair mechanism during the D-loop replication of the chloroplast genome (Zhu et al. 2015, Zhu et al. 2016).

These nucletotide substitutions may alter nucleotide sequences, resulting in change in GC content that if occurring in coding regions, may alter amino acid frequencies and codon usage. Given that no RNA editing processes have as yet been found in the green algae (Stern et al. 2010), we deduced the frequency of amino acid and codon usage based on protein coding sequences and tRNAs. Our results showed that the amino acid frequencies and codon usage differed slightly between the two strains, but the GC content remained the same (data not shown).

Nucleotide substitution rate varies within genes, among genes, and across lineages (Wolfe et al. 1987). Knowing the extent of this variation aids understanding the mode of evolution of protein coding plastid genes in prasinophytes. The observed disproportional increases in Ka/Ks suggest a history of relaxed purifying selection and/or increase in positive selection acting on a subset of plastid genes. The ratio differences also suggest that changes in selection pressure may be associated with specific biochemical pathways or functions rather than across the entire genome (Magee et al. 2010).

One explanation for observed high variability of prasinophyte chloroplast genomes at the intra-specific level may be long divergence time. It is known that divergence time is correlated with the number of substitutions, because nucleotide substitutions accumulate over time in independent populations. When compared to the prasinophyte *Ostreococcus tauri* (Blanc-Mathieu et al. 2013), at the intra-specific level, *P. parkeae* chloroplast genomes contained fewer variable positions overall (37,873 positions in *O. tauri* and ~16,700 positions in *P. parkeae* estimated using whole genome Geneious alignment). However, the variability within protein coding sequences of *P. parkeae* was much higher. 3,111 variable positions (2,684 SNPs and 44 indels) were present in protein coding genes of *P. parkeae* while only 153 SNPs were present in that of *O. tauri* (Supporting Information Table S1).

It is also possible that *P. parkeae* chloroplast genomes contain a trait that allows the chloroplast genomes to evolve at a higher rate when compared to those of other organisms in the green lineage. This hypothesis is supported by presence of high intra-specific variability of some euglenoid chloroplast genomes (Bennett and Triemer 2015), which were inherited from a *Pyramimonas*-like chloroplast donor (Palmer 1987, Turmel et al. 2009). Similar to *P. parkeae* chloroplast genomes, those euglenoid chloroplast genomes exhibit intra-specific variability, however, with a higher mutation rate (Bennett and Triemer 2015). It might be possible that a *Pyramimonas*-like chloroplast genome progenitor had a trait that favors mutation and was passed on to descendants.

Another potential explanation for observed high variability of prasinophyte chloroplast genomes at the intra-specific level is recombination of bi-parentally inherited chloroplast genomes. Evidence for chloroplast DNA recombination has been reported for the prasinophyte *O. tauri* (Blanc-Mathieu et al. 2013). These observations indicate that earliest diverging green algae may display bi-parental chloroplast genome inheritance. If so, uni-parental chloroplast inheritance may have evolved independently in chlorophyte and streptophyte lineages.

Last, but not least, our observation of greater than expected variability between the chloroplast genomes might indicate that NIES254 and CCMP726 are actually different species of *Pyramimonas*. However, phylogenetic analysis of publically-available *Pyramimonas* 18S rDNA, 16S rDNA, and *rbcL* sequences were consistent with previous studies (Balzano et al. 2012, Suda et al. 2013) in resolving all *P. parkeae* strains known to date as a monophyletic clade. Additional *Pyramimonas* strains and molecular data may clarify diversification patterns for this ecologically and evolutionarily important genus.

Summary

The availability of a newly sequenced chloroplast genome for *Pyramimonas parkeae* NIES254 made it possible to examine intra-specific variation of chloroplast genomes in early diverging green algae. Although plastid genomes of CCMP726 and NIES254 have identical gene content, these genomes exhibited some of the highest variability known to occur at the intra-specific level in the green lineage: 1) the NIES254 chloroplast genome is longer than that of CCMP726 by 3,024 bp; 2) there are four large hotspot regions where the similarity value between the two studied strains is close to zero; 3) inverted repeat boundaries have shifted; and

4) boundaries of the inverted repeat at the IR-SSC junction have undergone contraction or expansion for not just a few nucleotides, but for about 2.5 kb, resulting in differences in copy number for the three protein coding genes *ycf20*, *psaC*, and *ndhE*.

Acknowledgements

This study was supported in part by NSF grant DEB-1119944. We thank the Laboratory of Genetics for access to computational infrastructure and support.

References

Allender, C.J., Allainguillaume, J., Lynn, J. & King, G.J. 2007. Simple sequence repeats reveal uneven distribution of genetic diversity in chloroplast genomes of *Brassica oleracea* L. and (n=9) wild relatives. *Theor. Appl. Genet.* 114(4):609-618.

Balzano, S., Gourvil, P., Siano, R., Chanoine, M., Marie, D., Lessard, S., Sarno, D. & Vaulot, D. 2012. Diversity of cultured photosynthetic flagellates in the northeast Pacific and Arctic Oceans in summer. *Biogeosciences* 9(11):4553-4571.

Bennett, M.S. & Triemer, R.E. 2015. Chloroplast genome evolution in the Euglenaceae. *J. Eukaryot. Microbiol.* 62(6):773-785.

Blanc-Mathieu, R., Sanchez-Ferandin, S., Eyre-Walker, A. & Piganeau, G. 2013. Organellar inheritance in the green Lineage: insights from *Ostreococcus tauri*. *Genome Biol. Evol.* 5(8):1503-1511.

Birky, C.W. & Walsh, J.B. 1992. Biased gene conversion, copy number, and apparent mutation rate differences within chloroplast and bacterial genomes. *Genetics* 130(3):677-683.

Bolger, A. M., Lohse, M. & and Usadel, B. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* 30:2114-2120.

Brouard, J.S., Otis, C., Lemieux, C. & Turmel, M. 2010. The exceptionally large chloroplast genome of the green alga *Floydiella terrestris* illuminates the evolutionary history of the Chlorophyceae. *Genome Biol. Evol.* 2:240-256.

Bhuiyan, M.A.H., Faria, D.G., Horiguchi, T., Sym, S.D. & Suda, S. 2015. Taxonomy and phylogeny of *Pyramimonas vacuolata* sp. nov. (Pyramimonadales, Chlorophyta). *Phycologia* 54(4):323-332.

Darling, A.E., Mau, B. & Perna, N.T. 2010. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PloS ONE* 5(6):e11147.

Daugbjerg, N., Moestrup, Ø. & Arctander, P. 1994. Phylogeny of the genus *Pyramimonas* (Prasinophyceae, Chlorophyta) inferred from the *rbcL* gene. *J. Phycol.* 30(6):991-999.

De Las Rivas, J., Lozano, J.J. & Ortiz, A.R. 2002. Comparative analysis of chloroplast genomes: functional annotation, genome-based phylogeny, and deduced evolutionary patterns. *Genome Res.* 12(4):567-583.

Decelle, J., Romac, S., Stern, R.F., Bendif, E.M., Zingone, A., Audic, S., Guiry, M.D., Guillou, L., Tessier, D., Le Gall, F. & Gourvil, P. 2015. PhytoREF: a reference database of the plastidial 16S *rRNA* gene of photosynthetic eukaryotes with curated taxonomy. *Mol. Ecol. Resour.* 15(6):1435-1445.

Duanmu, D., Bachy, C., Sudek, S., Wong, C.H., Jiménez, V., Rockwell, N.C., Martin, S.S., Ngan, C.Y., Reistetter, E.N., van Baren, M.J. & Price, D.C. 2014. Marine algae and land plants share conserved phytochrome signaling systems. *Proc. Natl. Acad. Sci. U.S.A.* 111(44):15827-15832.

Dugas, D.V., Hernandez, D., Koenen, E.J., Schwarz, E., Straub, S., Hughes, C. E., Jansen, R.K., Nageswara-Rao, M., Staats, M., Trujillo, J.T. & Hajrah, N.H. 2015. Mimosoid legume plastome evolution: IR expansion, tandem repeat expansions, and accelerated rate of evolution in *clpP*. *Sci. Rep.* 5:16958.

Goff, S.A., Vaughn, M., McKay, S., Lyons, E., Stapleton, A.E. Gessler, D., Matasci, N., Wang, L., Hanlon, M., Lenards, A. & Muir, A. 2011. The iPlant collaborative: cyberinfrastructure for plant biology. *Front. Plant Sci.* 2.

Goulding, S.E., Wolfe, K.H., Olmstead, R.G. & Morden, C.W. 1996. Ebb and flow of the chloroplast inverted repeat. *Mol. Gen. Genet.* 252(1-2):195-206.

Hahn, C., Bachmann, L. & Chevreux, B. 2013. Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—a baiting and iterative mapping approach. *Nucleic Acids Res.* 41(13):e129-e129.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C. & Thierer, T. 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28(12):1647-1649.

Kiełbasa, S.M., Wan, R., Sato, K., Horton, P. & Frith, M.C. 2011. Adaptive seeds tame genomic sequence comparison. *Genome Res*. 21(3):487-493.

Lagesen, K., Hallin, P., Rødland, E.A., Stærfeldt, H.H., Rognes, T. & Ussery, D.W. 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* 35(9):3100-3108.

Leliaert, F., Tronholm, A., Lemieux, C., Turmel, M., DePriest, M.S., Bhattacharya, D., Karol, K.G., Fredericq, S., Zechman, F.W. & Lopez-Bautista, J.M. 2016. Chloroplast phylogenomic analyses reveal the deepest-branching lineage of the Chlorophyta, Palmophyllophyceae class. nov. *Sci. Rep.* 6:25367.

Lemieux, C., Otis, C. & Turmel, M. 2014. Six newly sequenced chloroplast genomes from prasinophyte green algae provide insights into the relationships among prasinophyte lineages and the diversity of streamlined genome architecture in picoplanktonic species. *BMC Genomics* 15(1):1.

Lemieux, C., Otis, C. & Turmel, M. 2016. Comparative chloroplast genome analyses of streptophyte green algae uncover major structural alterations in the Klebsormidiophyceae, Coleochaetophyceae and Zygnematophyceae. *Front. Plant Sci.* 7:6971.

Li, H. & Durbin, R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 25(14):1754-1760.

Lohse, M., Drechsel, O., Kahlau, S. & Bock, R. 2013. OrganellarGenomeDRAW—a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. *Nucleic Acids Res.* 41(W1):W575-581.

Magee, A.M., Aspinall, S., Rice, D.W., Cusack, B.P., Semon, M., Perry, A.S., Stefanović, S., Milbourne, D., Barth, S., Palmer, J.D. & Gray, J.C. 2010. Localized hypermutation and associated gene losses in legume chloroplast genomes. *Genome Res.* 20(12):1700-1710.

Marin, B. & Melkonian, M. 2010. Molecular phylogeny and classification of the Mamiellophyceae class. nov. (Chlorophyta) based on sequence comparisons of the nuclear-and plastid-encoded rRNA operons. *Protist* 161(2):304-336.

Moro, I., La Rocca, N., Valle, L.D., Moschin, E., Negrisolo, E. & Andreoli, C. 2002. *Pyramimonas australis* sp. nov. (Prasinophyceae, Chlorophyta) from Antarctica: fine structure and molecular phylogeny. *Eur. J. Phycol.* 37(1):103-114.

Nakayama, T., Marin, B., Kranz, H.D., Surek, B., Huss, V.A., Inouye, I. & Melkonian, M. 1998. The basal position of scaly green flagellates among the green algae (Chlorophyta) is revealed by analyses of nuclear-encoded SSU rRNA sequences. *Protist* 149(4):367-380.

Ogihara, Y. & Tsunewaki, K. 1988. Diversity and evolution of chloroplast DNA in *Triticum* and *Aegilops* as revealed by restriction fragment analysis. *Theor. Appl. Genet.* 76(3):321-332.

Palmer, J.D. 1987. Chloroplast DNA evolution and biosystematic uses of chloroplast DNA variation. *Am. Nat.* 130:S6-S29.

Quinlan, A. R. 2014. BEDTools: the Swiss-army tool for genome feature analysis. *Curr. Protoc. Bioinformatics* 11-12.

Robbens, S., Derelle, E., Ferraz, C., Wuyts, J., Moreau, H. & Van de Peer, Y. 2007. The complete chloroplast and mitochondrial DNA sequences of *Ostreococcus tauri*: organelle genomes of the smallest eukaryote are examples of compaction. *Mol. Biol. Evol.* 24(4):956-968.

Ronquist, F. & Huelsenbeck, J.P. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19(12):1572-1574.

Satjarak, A., Paasch, A.E., Graham, L.E. & Kim, E. 2016. Complete chloroplast genome sequence of phagomixotrophic green alga *Cymbomonas tetramitiformis*. *Genome Announc*. 4(3):e00551-16.

Schattner, P., Brooks, A. N. & Lowe, T. M. 2005. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res.* 33(suppl 2):W686-W689.

Stern, D. B., Goldschmidt-Clermont, M. & Hanson, M. R. 2010. Chloroplast RNA metabolism. *Annu. Rev. Plant Biol.* 61:125-155.

Suda, S. 2004. Taxonomic characterization of *Pyramimonas aurea* sp. nov. (Prasinophyceae, Chlorophyta). *Phycologia* 43(6):682-692.

Suda, S., Bhuiyan, M.A.H. & Faria, D.G. 2013. Genetic diversity of *Pyramimonas* from Ryukyu Archipelago, Japan (Chlorophyceae, Pyramimonadales). *J. Mar. Sci. Technol.* 21:285-296.

Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. 2013. MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* 30(12):2725-2729.

Turmel, M., Otis, C. & Lemieux, C. 1999. The complete chloroplast DNA sequence of the green alga *Nephroselmis olivacea*: insights into the architecture of ancestral chloroplast genomes. *Proc. Natl. Acad. Sci. U.S.A.* 96(18):10248-10253.

Turmel, M., Ehara, M., Otis, C. & Lemieux, C. 2002. Phylogenetic relationships among streptophytes as inferred from chloroplast small and large subunit *rRNA* gene sequences. *J. Phycol.* 38(2):364-375.

Turmel, M., Gagnon, M. C., O'Kelly, C. J., Otis, C. & Lemieux, C. 2009. The chloroplast genomes of the green algae *Pyramimonas*, *Monomastix,* and *Pycnococcus* shed new light on the evolutionary history of prasinophytes and the origin of the secondary chloroplasts of euglenids. *Mol. Biol. Evol.* 26(3):631-648.

Turmel, M., Otis, C. & Lemieux, C. 2015. Dynamic evolution of the chloroplast genome in the green algal classes Pedinophyceae and Trebouxiophyceae. *Genome Biol. Evol.* 7:2062-2082.

Turmel, M., de Cambiaire, J.C., Otis, C. & Lemieux, C. 2016. Distinctive architecture of the chloroplast genome in the chlorodendrophycean green algae *Scherffelia dubia* and *Tetraselmis* sp. CCMP 881. *PloS ONE* 11(2):e0148934.

Wang, L., Wuyun, T.N., Du, H., Wang, D. & Cao, D. 2016. Complete chloroplast genome sequences of *Eucommia ulmoides*: genome structure and evolution. *Tree Genet. Genomes* 12(1):1-15.

Wolfe, K.H., Li, W.H. & Sharp, P.M. 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proc. Natl. Acad. Sci. U.S.A.* 84(24):9054-9058.

Worden, A.Z., Lee, J.H., Mock, T., Rouzé, P., Simmons, M.P., Aerts, A.L., Allen, A.E., Cuvelier, M.L., Derelle, E., Everett, M.V. & Foulon, E. 2009. Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes *Micromonas*. *Science* 324(5924):268-272.

Zhu, A., Guo, W., Gupta, S., Fan, W. & Mower, J.P. 2015. Evolutionary dynamics of the plastid inverted repeat: the effects of expansion, contraction, and loss on substitution rates. *New Phytol.* 209(4):1747-1756.

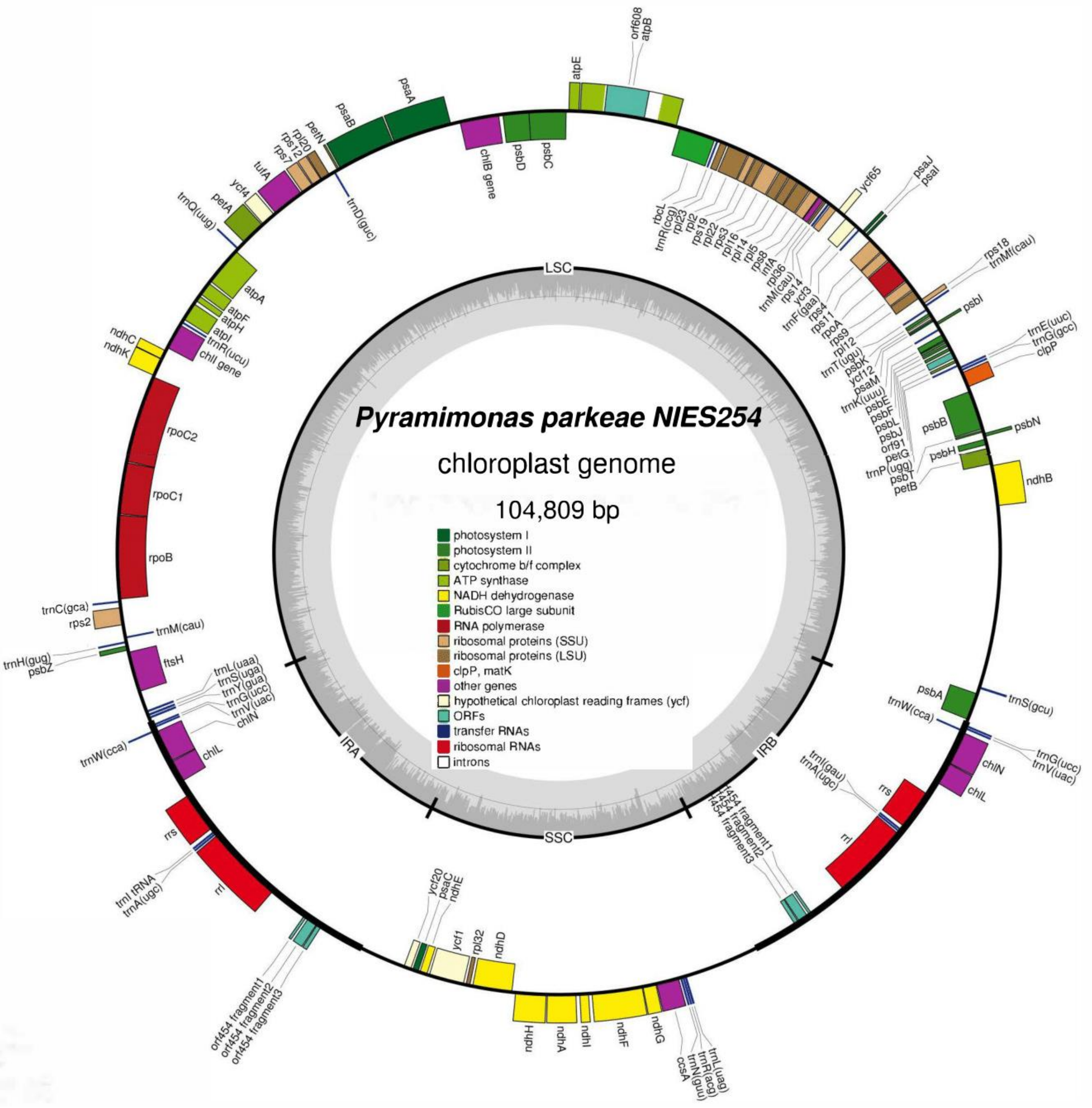


Figure 1. Map of *Pyramimonas parkeae* NIES254 chloroplast genome. The thick lines indicate the extent of the inverted repeat regions (IRA and IRB), which separate the genome into LSC and SSC regions. Genes outside the map are transcribed counterclockwise and those inside the map are transcribed clockwise.

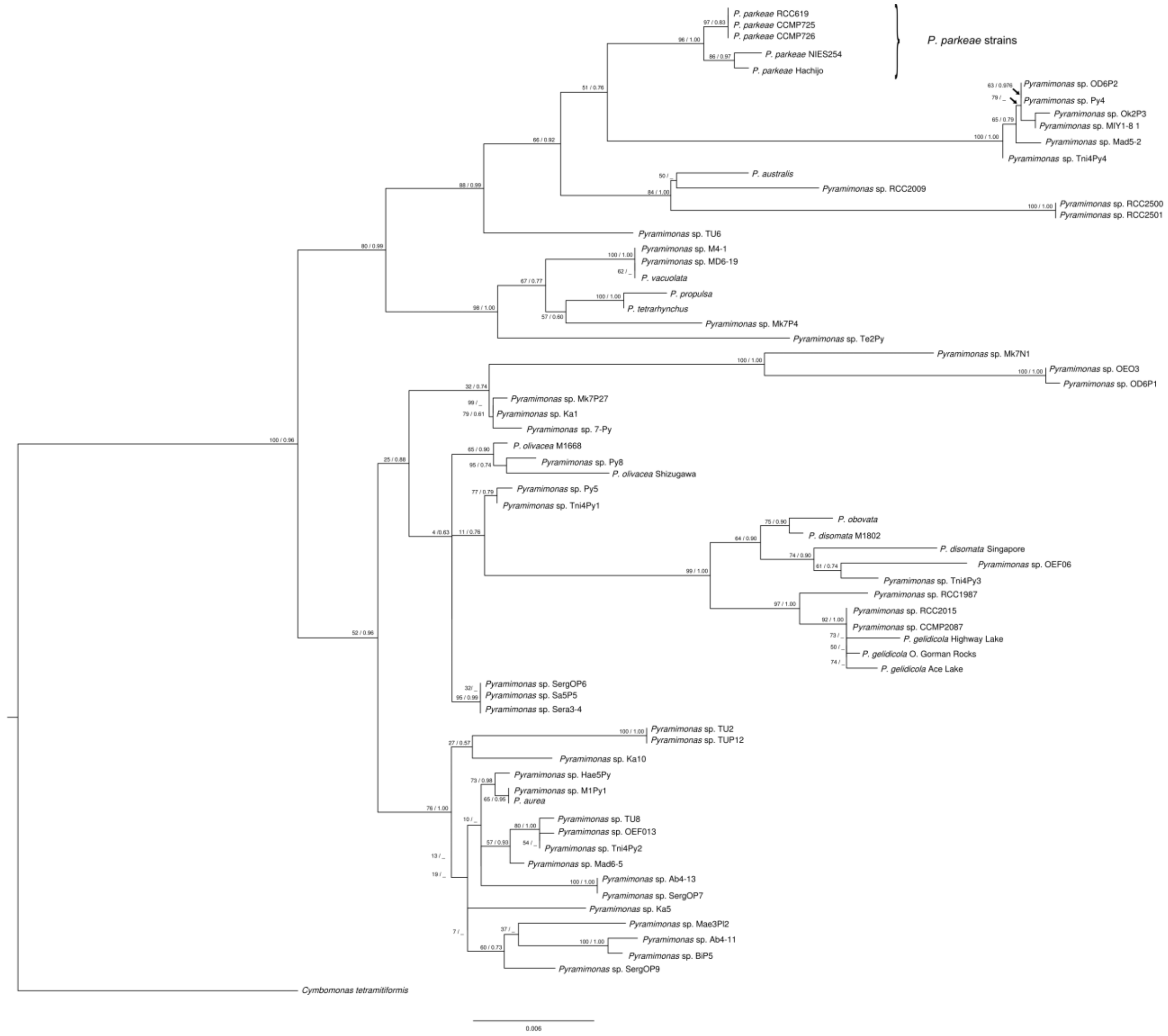


Figure 2. Maximum-Likelihood tree inferred from 18S rDNA sequences of 65 *Pyramimonas* spp. using a GTR+I+F model. The bootstrap and posterior probability values are reported at the respective nodes. The scale bar represents the estimated number of nucleotide substitutions per site. The box indicates the monophyletic relationship of *P. parkeae* strains. *Cymbomonas tetramitiformis* was used as an outgroup.

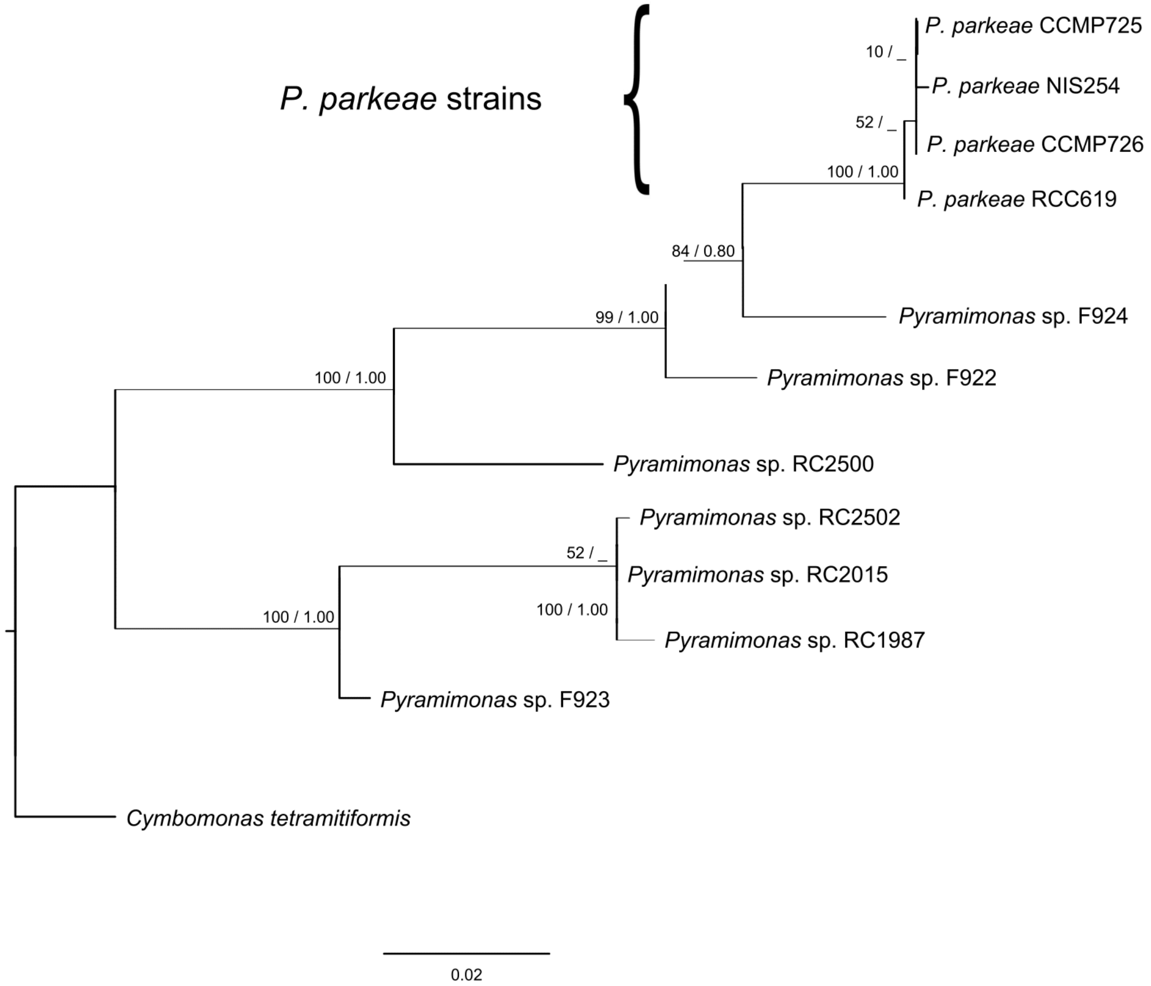


Figure 3. Maximum-Likelihood tree inferred from 16S rDNA sequences of 11 *Pyramimonas* spp. using a GTR+I+F model. The bootstrap and posterior probability values are reported at the respective nodes. The scale bar represents the estimated number of nucleotide substitutions per site. The box indicates the monophyletic relationship of *P. parkeae* strains. *Cymbomonas tetramitiformis* was used as an outgroup.

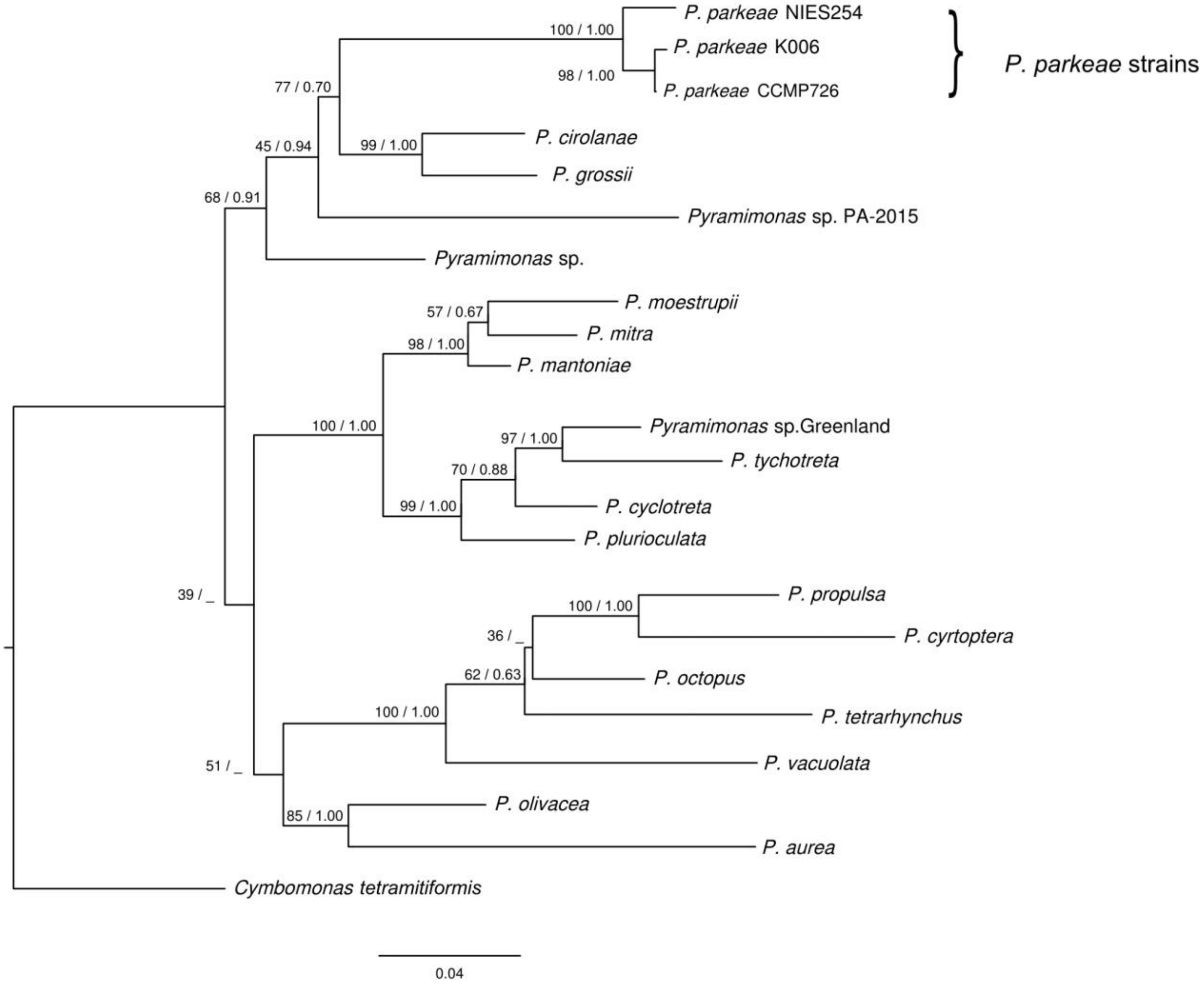


Figure 4. Maximum-Likelihood tree inferred from *rbcL* sequences of 21 *Pyramimonas* spp. using a GTR+I+F model. The bootstrap and posterior probability values are reported at the respective nodes. The scale bar represents the estimated number of nucleotide substitutions per site. The box indicates the monophyletic relationship of *P. parkeae* strains. *Cymbomonas tetramitiformis* was used as an outgroup.

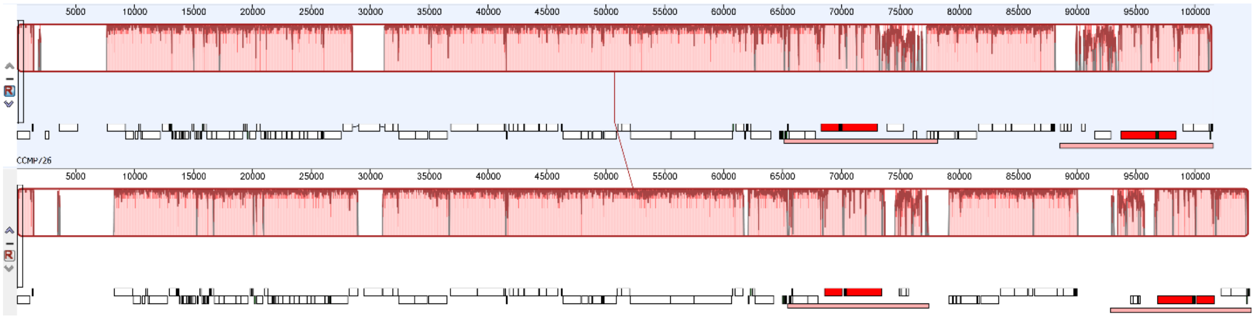


Figure 5. Mauve alignment of *Pyramimonas parkeae* NIES254 and CCMP726 chloroplast genomes showing shared synteny. The vertical line connecting the two syntenic regions between NIES254 and CCMP726 represents the collinear synteny of the two chloroplast genomes. The histogram inside each block represents pairwise nucleotide sequence identity. The large four areas where the heights of the histograms are almost equal to zero represent the four large hotspot regions: 1) the 6 kb intergenic region between *psbA*-*trnS* and *ndhB* in LSC, 2) 2 kb intron of *atpB*, and 3) 5.7 kb and 6.8 kb located at the boundaries of IR and SSC.

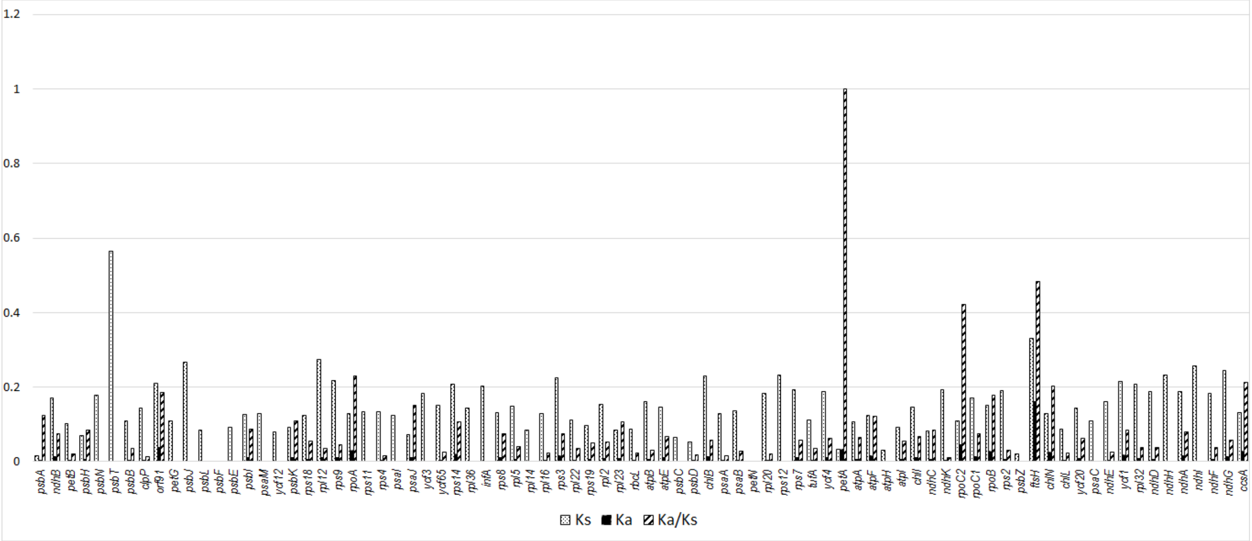


Figure 6. Comparison of NIES254 and CCMP726 shows that substitution in *P. parkeae* chloroplast genomes is unevenly distributed. The x-axis shows protein coding genes present in the chloroplast genomes, in genomic order. The y-axis is the value for the substitution rate (per 100 bp) and the value for Ka/Ks ratio. The dotted bar indicates synonymous substitution (Ks), the solid bar indicates nonsynonymous substitution (Ka), and the diagonal bar indicates the ratio of nonsynonymous substitution to synonymous substitution (Ka/Ks).

Supporting Information Table S1. Variable positions present within the protein coding regions of *Pyramimonas parkeae* and *Ostreococcus tauri* identified using Geneious alignments v 9.0.4 (Kearse et al. 2012). Synonymous (Ks) and nonsynonymous (Ka) substitution sites as well as the Ka/Ks ratio using MEGA6 v 6.06 (Tamura et al. 2013).

Genes	*Pyramimonas parkeae*				*Ostreococcus tauri*			
	Alignment length	# of variable position		Ka/Ks	Alignment length	# of variable position		Ka/Ks
		SNPs	# of Indels (bp)			SNPs	# of Indels (bp)	
psbA	246	2	0	0.125	1062	3	0	0
ndhB	1,539	71	0	0.076	Gene(s) not present in the chloroplast genomes			
petB	648	16	0	0.020	648	2	0	0
psbH	237	5	0	0.085	237	6	0	2.500
psbN	135	5	0	0	135	0	0	
psbT	96	9	0	0	96	0	0	NA
psbB	1,527	37	0	0.037	1332	9	0	0.143
clpP	597	20	0	0.014	603	0	0	NA
orf91	276	19	0	0.186	Gene(s) not present in the chloroplast genomes			
petG	114	3	0	0	114	13	0	0
psbJ	129	8	0	0	126	0	0	NA
psbL	117	2	0	0	117	0	0	NA
psbF	126	0	0	0	117	0	0	NA
psbE	246	5	0	0	249	0	0	NA
psbI	115	4	0	0.087	117	0	0	NA
psaM	102	3	0	0	96	0	0	NA
ycf12	102	2	0	0	102	0	0	NA
psbK	138	4	0	0.110	138	0	0	NA
rps18	207	7	1 (12)	0.056	213	0	0	NA
rpl12	393	24	0	0.036	Gene(s) not present in the chloroplast genomes			
rps9	402	22	0	0.046	384	2	0	0
rpoA	1,056	49	4 (14, 19, 15, 8)	0.231	1071	5	0	0
rps11	393	12	0	0	393	2	0	0
rps4	687	19	2 (71, 6)	0.015	594	2	0	0
psaI	111	3	0	0	111	1	0	0
psaJ	126	3	0	0.151	129	0	0	NA
ycf3	510	19	0	0	501	1	0	0
ycf65	294	9	0	0.027	Gene(s) not present in the chloroplast genomes			
rps14	303	17	0	0.106	303	4	0	0
rpl36	114	3	0	0	114	0	0	NA

Genes	*Pyramimonas parkeae*				*Ostreococcus tauri*			
	Alignment length	# of variable position		Ka/Ks	Alignment length	# of variable position		Ka/Ks
		SNPs	# of Indels (bp)			SNPs	# of Indels (bp)	
infA	228	9	0	0	237	2	0	0
rps8	387	14	0	0.076	369	1	0	0
rpl5	546	19	0	0.040	555	0	0	NA
rpl14	369	7	0	0	360	0	0	NA
rpl16	411	13	1 (6)	0.023	426	1	0	0
rps3	639	36	1 (13)	0.075	636	0	0	NA
rpl22	393	24	0	0.036	Gene(s) not present in the chloroplast genomes			
rps19	279	7	0	0.051	279	0	0	NA
rpl2	828	34	0	0.052	834	0	0	NA
rpl23	279	7	0	0.106	264	0	0	NA
rbcL	1,428	31	0	0.023	1428	12	0	0.125
atpB	1,551	72	2 (76, 9)	0.031	1452	12	0	0.250
atpE	410	15	1 (8)	0.068	429	1	0	0
psbC	1,422	21	0	0	1422	4	0	0
psbD	1,059	14	0	0.019	1059	2	0	0
chlB	1,530	80	5 (6, 2, 4, 6, 6)	0.057	Gene(s) not present inthe chloroplast genomes			
psaA	2,256	64	0	0.016	2256	12	0	0
psaB	2,232	72	0	0.029	2202	7	0	0
petN	93	0	0	0	Gene(s) not present in the chloroplast genomes			
rpl20	357	15	0	0.022	336	1	0	0
rps12	375	19	0	0	375	0	0	NA
rps7	471	23	0	0.057	471	0	0	NA
tufA	1,230	34	0	0.036	1230	1	0	0
ycf4	561	26	0	0.063	Gene(s) not present in the chloroplast genomes			
petA	971	32	3 (2, 2, 2)	1.000	924	5	0	0
atpA	1,521	46	0	0.065	1515	3	0	0
atpF	540	20	0	0.121	507	1	0	0
atpH	249	2	0	0	306	0	0	NA
atpI	729	19	0	0.055	711	1	0	0
chlI	1,017	39	0	0.068	Gene(s) not present in the chloroplast genomes			
ndhC	363	9	0	0.084				
ndhK	697	35	0	0.010				
rpoC2	3,455	215	6 (12, 2, 2, 2, 2, 6)	0.422	3036	20	0	0.250
rpoC1	2,006	89	3 (9, 2, 2)	0.076	2223	8	0	0.5
rpoB	3,197	159	2 (2, 5)	0.179	3267	6	0	0
rps2	678	27	1 (9)	0.032	681	3	0	0
psbZ	189	1	0	0	Gene(s) not present in the chloroplast genomes			
ftsH	1,686	277	3 (6, 12, 98)	0.483				

Genes	*Pyramimonas parkeae*				*Ostreococcus tauri*			
	Alignment length	# of variable position		Ka/Ks	Alignment length	# of variable position		Ka/Ks
		SNPs	# of Indels (bp)			SNPs	# of Indels (bp)	
chlN	1,320	62	2 (30, 15)	0.203				
chlL	888	18	1 (17)	0.023				
ycf20	297	10	0	0.062				
psaC	246	6	0	0	246	0	0	
ndhE	306	12	0	0.025	Gene(s) not present in the chloroplast genomes			
ycf1	1,356	75	3 (9, 2, 8)	0.084				
rpl32	174	8	0	0.038	219	0	0	
ndhD	1,530	69	0	0.037	Gene(s) not present in the chloroplast genomes			
ndhH	1,176	54	0	0				
ndhA	1,104	56	0	0.080				
ndhI	354	17	0	0				
ndhF	1,929	84	0	0.038				
ndhG	519	32	0	0.057				
ccsA	817	40	4 (1, 1, 5, 4)	0.212				
Total		2571				153	0	

References

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C. & Thierer, T. 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28(12):1647-1649.

Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. 2013. MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol*. 30(12):2725-2729.

CHAPTER 3: COMPLETE MITOCHONDRIAL GENOMES OF PRASINOPHYTE ALGAE *PYRAMIMONAS PARKEAE* AND *CYMBOMONAS TETRAMITIFORMIS*[1]

Research article

Anchittha Satjarak[2]

Department of Botany, University of Wisconsin-Madison, 430 Lincoln drive, Madison, Wisconsin, USA

[2]corresponding author: e-mail: satjarak@wisc.edu, phone: +16082620657, fax: +16082627509

John A. Burns

Division of Invertebrate Zoology and Sackler Institute for Comparative Genomics, American Museum of Natural History, Central Park West at 79th Street, New York, New York, USA

Eunsoo Kim

Division of Invertebrate Zoology and Sackler Institute for Comparative Genomics, American Museum of Natural History, Central Park West at 79th Street, New York, New York, USA

Linda E. Graham

Department of Botany, University of Wisconsin-Madison, 430 Lincoln drive, Madison, Wisconsin, USA

Key word: mitochondrial genome; prasinophyte; *Pyramimonas parkeae*; *Cymbomonas tetramitiformis*

Running header: Mitochondrial genomes of *Pyramimonas parkeae* and *Cymbomonas tetramitiformis*

Abbreviations: ORF, open reading frame; ML, Maximum-likelihood; MLGO, Maximum Likelihood for Gene Order Analysis; LSC, large single-copy region; IRA, inverted repeat region A; IRB, inverted repeat region B; *atp*1, 4, 6, 8, 9, genes for ATP synthase subunit1, 4, 6, 8, 9; *tatC,* gene for sec-independent protein translocase; *nad*1, 2, 3, 4, 4L, 5, 6, 7, 9, genes for NADH dehydrogenase subunit 1,2, 3, 4, 4L, 5, 6, 7, 9; *cob*, gene for apocytochrome b; *cox*1, 2, 3, gene for cytochrome oxidase subunit 1, 2, 3; *rps*2, 3, 4, 7, 8, 10, 11, 12, 13, 14, 19, genes for ribosomal protein S2, 3, 4, 7, 8, 10, 11, 12, 13, 14, 19; *rpl*5, 6, 14, 16, genes for ribosomal protein L5, 6, 14, 16; *rrn*L, gene for 23S ribosomal RNA; *rrn*S, gene for 16S ribosomal RNA.

Abstract

Mitochondria are archetypal eukaryotic organelles that were acquired by endosymbiosis of an ancient species of alpha-proteobacteria by the last eukaryotic common ancestor. The genetic information contained within the mitochondrial genome has been an important source of information for resolving relationships among eukaryotic taxa. In this study, we utilized mitochondrial and chloroplast genomes to explore relationships among prasinophytes. Prasinophytes are represented by diverse early-diverging green algae whose physical structures and genomes have the potential to elucidate the traits of the last common ancestor of the Viridiplantae (or Chloroplastida). We constructed *de novo* mitochondrial genomes for two prasinophyte algal species, *Pyramimonas parkeae* and *Cymbomonas tetramitiformis*, representing a prasinophyte clade (clade I) for which mitochondrial genomes were not previously available in public databases. Comparisons of genome structure and gene order between these species and to those of other prasinophytes revealed that the mitochondrial genomes of *P. parkeae* and *C. tetramitiformis* are more similar to each other than to other prasinophytes, consistent with other molecular inferences of the close relationship between these two species. Phylogenetic analyses using the inferred amino acid sequences of mitochondrial and chloroplast protein-coding genes resolved a clade consisting of *P. parkeae* and *C. tetramitiformis*; and this group (representing the prasinophyte clade I) branched with the clade II, consistent with previous studies based on the use of nuclear gene markers.

Introduction

Mitochondria, found in most eukaryotic cells, perform essential cellular functions, including energy transduction, homeostasis, intermediary metabolism, and apoptosis (Gray et al. 1999, Burger et al. 2003; Osellame et al. 2012). Mitochondria are hypothesized to have had a single origin, derived from the endosymbiosis of an alpha-proteobacterial endosymbiont in the last eukaryotic common ancestor. In consequence, the mitochondrial genome has been widely used to help infer evolutionary history across diverse eukaryotic lineages (Gray et al. 1999).

Mitochondrial genomic features have been particularly useful for inferring the evolutionary history of the green algae and land plants, together known as Viridiplantae (or Chloroplastida), because mitochondrial genomes are more structurally diverse than are chloroplast genomes in this group. For example, Viridiplantae mitochondrial genomes seem to have higher rates of gene breakage and fusion and consequently, higher variation in genomic organization, size, and coding capacity than do Viridiplantae chloroplast genomes (Gray 1999, Lang et al. 1999, Nosek and Tomáška 2003).

Prasinophyte algae, occurring primarily as flagellate or non-flagellate unicells, are particularly important as the earliest diverging representatives of Viridiplantae (Graham et al. 2016). Their morphological and molecular features have the potential to reflect characters of the last common ancestor of the green lineage. Most previous phylogenetic analyses of prasinophytes have been based on protein sequences from chloroplast genomes (Lemieux et al. 2014, Leliaert et al. 2016). However, information from chloroplast genomes has not adequately resolved prasinophyte relationships thus far.

Since early diverging green algal mitochondrial genomes evolve more rapidly than chloroplast genomes (Smith and Keeling 2015), and can yield phylogenetic tree topologies that differ from those based upon chloroplast genome features (Blanc-Mathieu et al. 2013), mitochondrial genome comparisons offer the potential to improve the resolution of prasinophyte phylogeny. A few previous studies have reported paraphyletic assemblages of prasinophytes inferred from mitochondrial genomes. However, the number of taxa used in these studies was limited (Turmel et al. 1999, Turmel et al. 2010), and the results were judged to be highly susceptible to systemic errors of phylogeny reconstruction, e.g., presence of long-branch attraction artifacts (Turmel et al. 2010).

To date, mitochondrial genomes are publicly available for eight prasinophyte taxa representing only five of approximately seven clades as described in Leliaert et al. (2016): *Pyramimonas parkeae* SCCAP K-0007 (clade I); *Micromonas sp.* RCC299, *Bathycoccus prasinos*, *Ostreococcus tauri*, *Monomastix sp.* OKE-1 (clade II); *Nephroselmis olivacea* (clade III); *Pycnococcus provasolii* (clade V); and *Prasinoderma coloniale* (clade VI) (Turmel et al. 1999, Robbens et al. 2007, Worden et al. 2009, Turmel et al. 2010, Vaulot et al. 2012, Pombert et al. 2013, Turmel et al. 2013, Hrdá et al. 2016). To improve representation of mitochondrial genomes, we performed *de novo* reconstructions of the mitochondrial genomes of *Pyramimonas parkeae* NIES254 and *Cymbomonas tetramitiformis* PLY262, representing prasinophyte clade I based on previous chloroplast 16S rRNA gene phylogenetic analyses (Leliaert et al. 2012, Leliaert et al. 2016). These new mitochondrial genomes offer the opportunity to revisit phylogenetic relationships of prasinophytes. The new data also foster an assessment of mitochondrial genome variability for comparison to chloroplast genome variability in this group of green algae, as some previous studies measured variability in prasinophyte chloroplast nucleotide sequences, genome structures, and their gene contents (Turmel et al. 2009, Satjarak et al. 2016). Comparisons of the new mitochondrial genomes to each other, and to those of other prasinophytes, revealed that the two new mitochondrial genomes exhibit lower variation in gene order and DNA sequence than occurs more widely among prasinophyte species. Also, our phylogenetic analyses of 32 concatenated proteins derived from mitochondrial protein-coding sequences supported a clade consisting of *P. parkeae* and *C. tetramitiformis*, as well as monophyly of the known prasinophyte clades.

Materials and Methods

Algal strains, genome sequencing, and assembly

A culture of *P. parkeae* Norris and Pearson (NIES254) was acquired from the National Institute of Environmental Studies, Japan. The culture was propagated in Alga-Gro® seawater medium (Carolina Biological Supply Company, Burlington, NC, USA), and was maintained in a walk-in growth room with 16:8 daily light/dark cycle at 20˚C. Cells were harvested during the exponential phase. Total DNA was prepared by using FastDNA® SPIN Kit for Soil (MP Biomedicals, Solon, OH) and sequenced using Illumina Miseq at the University of Wisconsin-Madison Biotechnology Center.

P. parkeae raw paired-end Illumina data consisted of 13,232,998 sequences with an average read length of 251 bp. The data were trimmed by Trimmomatic v 0.33 (Bolger et al. 2014) in order to obtain a quality score of at least 28 on the phred 64 scale. The *P. parkeae* mitochondrial genome was assembled by using MIRA v 4.0.2 and MITObim v 1.8 (Hahn et al. 2013) using *P. parkeae* NIES254 mitochondrial *cox*1 mRNA for cytochrome c oxidase subunit

1, partial coding DNA sequence available in GenBank (accession number: AB491639.1) as a reference sequence.

A culture of *C. tetramitiformis* Schiller (PLY262) was acquired from the Plymouth Culture Collection of Marine Microalgae; and its total DNA was prepared for sequencing on the Illumina MiSeq platform as described in Satjarak et al. (2016). The mitochondrial genome of *C. tetramitiformis* was initially assembled in MIRA v 4.0.2 and MITObim v 1.8 (Hahn et al. 2013) using *P. parkeae* NIES254 mitochondrial protein-coding sequences (obtained in this study) as reference sequences. However, the variability present between mitochondrial sequences of *P. parkeae* and *C. tetramitiformis* caused the mitochondrial assembly of *C. tetramitiformis* to be fragmented. Therefore, we used the genes sequences presented in the initial assembly of *C. tetramitiformis* as reference sequences. To obtain the reference sequences, we predicted and annotated ORFs present in the initial assembly using the "Find ORFs" function implemented in Geneious v 9.0.4 (Kearse et al. 2012) and BLAST search against the NCBI non-redundant protein database accessed in March 2016. Then the eukaryotic mitochondrial ORFs with known functions related to electron transport, ATP synthesis, or translation were used as reference sequences for *C. tetramitiformis* secondary assembly using MIRA v 4.0.2 and MITObim v 1.8 (Hahn et al. 2013).

Mitochondrial genome sequence analyses

We calculated the coverage of every position of the two mitochondrial genomes by aligning trimmed reads against the newly constructed mitochondrial genomes using BWA non-model species alignment v 0.7.4 (Li and Durbin 2009). Then, coverage of every position in the mitochondrial genome was calculated by using Bedtools Genome Coverage BAM v 2.19.1 (Quinlan 2014) implemented in iPlant Collaborative (Goff et al. 2011). Open reading frames (ORFs) were predicted using the “Find ORFs” function implemented in Geneious v 9.0.4 (Kearse et al. 2012). Then, the ORFs having a length of at least 100 bp were annotated by homology search using BLAST search against the NCBI non-redundant protein databases accessed in March 2016 (http://blast.ncbi.nlm.nih.gov/Blast.cgi). Intron boundaries were determined by comparing intron-containing genes with intron-less homologs. tRNAs and rRNAs were predicted using tRNAscan-SE v 1.21 (Schattner et al. 2005) and RNAmmer v 1.2 (Lagesen et al. 2007). Base frequencies, amino acid frequencies, and codon usage were calculated using the statistics option in Geneious v 9.0.4 (Kearse et al. 2012). The circular mapping genomes were drawn using OGDraw v1.2 (Lohse et al. 2013). The resulting annotated mitochondrial genome sequences of *P. parkeae* and *C. tetramitiformis* have been deposited at GenBank under accession numbers KX013547.1 and KX013548.1, respectively.

Comparative analysis of Pyramimonas parkeae mitochondrial genomes

The analysis of syntenic conservation between *P. parkeae* NIES254 and SCCAP K-0007 mitochondrial genomes was performed using progressive Mauve alignment v 2.4.0 (Darling et al. 2010) and LAST (Kiełbasa et al. 2011), with the following parameters: maximum score, max multiplicity for initial matches = 10, minimum length for initial matches = 1, step-size along reference sequences = 1, step-size along query sequences = 1, query letters per random alignment = 1e6. Single nucleotide polymorphisms (SNPs) within the whole genome and within the protein coding regions were identified using Geneious alignments, v 9.0.4 (Kearse et al.

2012). Synonymous (Ks) and nonsynonymous (Ka) substitution sites as well as the Ka/Ks ratio were calculated using MEGA6 v 7.0.14 (Kumar et al. 2016).

Comparative analysis of mitochondrial genomes of prasinophytes clade I and II

Syntenic conservation between the mitochondrial genomes of *P. parkeae* NIES254 and *C. tetramitiformis* PLY262 and among prasinophytes clade II (Table 1) was assessed by using progressive Mauve alignment v 2.4.0 (Darling et al. 2010). The phylogeny reconstruction from gene-order data was performed using Maximum Likelihood for Gene Order Analysis (MLGO) (Hu et al. 2014). Gene content and the density of coding regions were compared across prasinophyte clades I and II.

Phylogenetic analyses

The dataset for phylogenetic analyses included an alignment of 32 concatenated unambiguously aligned mitochondrial protein sequences inferred from protein-coding sequences of complete and partial mitochondrial genomes publicly available in Genbank (accessed in May 2016) as listed in Table 1. This set of proteins was selected—starting from the prasinophyte mitochondrial genomes—based on the following criteria: 1) it is present in at least 50% of the taxa analyzed and 2) it is conserved enough to be aligned well across distinct eukaryotic groups. The deduced amino acid sequences were aligned using MAFFT v 7.205 (Katoh and Standley 2013) and were trimmed using trimAl v 1.2 (Capella-Gutiérrez et al. 2009). The alignment, including 7,239 amino acid positions, has been submitted to TreeBASE under submission number 20167. The amino acid substitution models for each gene and the concatenated protein sequences were computed using ProtTest v 3.4.2 (Darriba et al. 2011), and Maximum-likelihood (ML) analysis was performed using RAxML v 8.2.8 (Stamatakis 2014) on the CIPRES XSEDE Portal (Miller et al. 2010) using an LG + I + G + F substitution model, rapid bootstrapping method with 1,000 replications for bootstrap analyses. Baysian analysis was performed using MrBayes v 3.2.6 (Ronquist and Huelsenbeck 2003) using the substitution model LG + I + G + F. Four independent chains were run for 1,100,000 cycles and the consensus topologies were calculated after the burn-in of 100,000 cycles.

To resolve the relationship of the deeper branching events, we added proteins inferred from chloroplast protein-coding sequences of the corresponding algal species (Table 1). The set of chloroplast proteins was selected by the same criteria used for selecting the mitochondrial proteins. The added proteins included the derived protein sequences of *acc*D, *atp*A, *atp*B, *atp*E, *atp*F, *atp*H, *atp*I, *ccs*A, *cem*A, *chl*B, *chl*I, *chl*L, *chl*N, *clp*P, *ftsH*, *inf*A, *pet*A, *pet*B, *pet*D, *pet*G, *pet*L, *psa*A, *psa*B, *psa*C, *psa*I, *psa*J, *psa*M, *psb*A, *psb*B, *psb*C, *psb*D, *psb*E, *psb*F, *psb*H, *psb*I, *psb*L, *psb*K, *psb*L, *psb*M, *psb*N, *psb*T, *psb*Z, *rbc*L, *rpl*2, *rpl*5, *rpl*12, *rpl*14, *rpl*16, *rpl*19, *rpl*20, *rpl*22, *rpl*23, *rpl*32, *rpl*36, *rpo*A, *rpo*B, *rpo*C1, *rpo*C2, *rps*2, *rps*3, *rps*4, *rps*7, *rps*8, *rps*9, *rps*11, *rps*14, *rps*18, *rps*19, *tufA*, *ycf1*, *ycf3*, *ycf4*, and *ycf12.* The concatenated alignment of combined mitochondrial and chloroplast data resulted in 25,059 amino acid positions. This alignment has been submitted to TreeBASE under submission number 20167. ML analysis was performed using RAxML (v 8.2.8) (Stamatakis 2014) on the CIPRES XSEDE Portal (Miller et al. 2010) using a CpREV+I+G+F substitution model, rapid bootstrapping method with 1,000 replications for bootstrap analyses. Baysian analysis was performed using MrBayes v 3.2.6 (Ronquist and Huelsenbeck 2003) using a CpREV+I+G+F substitution model. Four independent chains were run for 1,100,000 cycles and the consensus topologies were calculated after the burn-in of 100,000 cycles.

Results

General features and gene content

The mitochondrial genome of *P. parkeae* NIES254 assembled to a 53,406 bp long circular-mapping molecule. The mitochondrial genome exhibited quadripartite structure consisting of 31,222 bp of large single copy region (LSC), 3,228 bp of small single copy region (SSC), and two copies of a 9,478 bp inverted repeat (IR) encompassing 34 % of the genome. The IR regions encoded six protein coding genes (*nad*1, *nad*2, *nad*6, *rps*2, *rps*4, and *rps*12) and 10 tRNA genes (*trnE*, *F*, *G*, *I*, *M*, *N*, *P*, *Q*, *R*, and *W*) (Fig. 1). The average coverage of all positions of the mitochondrial genome was 310 fold. The GC content of the genome was 30.50%. This mitochondrial genome encoded 58 unique conserved genes: 2 rRNAs, 22 tRNAs, 19 genes encoding respiratory proteins, 14 genes encoding ribosomal subunits, and 1 Sec-independent protein translocase protein (Table 2). The genic regions accounted for 68.35 (36,502 bp), and the intergenic regions accounted for 31.65% (16,904 bp) of the entire genome (Table 3). There were 4 regions of overlapping genes: 1) 19 bp at the 5′end of *rps*19 and the 3′end of *rps*10, 2) 17 bp at the 5′end of *rps*3 and the 3′end of *rps*19, 3) 4 bp at the 5′end of *rpl*14 and the 3′end of *rpl*16, 4) 29 bp at the 5′end of *rps*13 and the 3′end of *rps*11, and 5) 32 bp at the 5′end of *rns* and the 3′end of *tatC*.

The mitochondrial genome of *C. tetramitiformis* assembled to a 73,520 bp long circular-mapping molecule. The mitochondrial genome exhibited quadripartite structure consisting of

29,132 bp of LSC, 24,953 bp of SSC, and two copies of a 9,718 bp IR encompassing 13.22 % of the genome. The IR regions encoded three tRNA species (*trnE*, *F*, and *M*) and *rns* (Fig. 2). The average coverage of each position of the mitochondrial genome was 419 fold. GC content of the genome was 37.8%. This mitochondrial genome encoded 56 unique conserved genes: 2 rRNAs, 23 tRNAs, 16 protein-coding genes encoding respiratory proteins, and 15 protein-coding genes encoding ribosomal subunits (Table 2). The genic regions account for 55.67 % (40,930 bp) and the intergenic regions accounted for 44.33 % (32,590 bp) of the entire genome (Table 3). There were 6 regions of overlapping genes: 1) 2 bp at the 5′end of *rps*13 and the 3′end of *rpl*6, 2) 11 bp at the 5′end of *rps*14 and the 3′end of *rpl*5, 3) 10 bp at the 5′end of *rpl*5 and the 3′end of *rpl*14, 4) 4 bp at the 5′end of *rpl*14 and the 3′end of *rpl*16, 5) 1 bp at the 5′end of *rps*19 and the 3′end of *rps*10, and 6) 14 bp at the 5′end of *rps*12 and the 3′end of *rps*2.

Comparative analyses of *Pyramimonas parkeae* mitochondrial genomes

Comparative analyses between *P. parkeae* NIES254 and SCCAP K-0007 showed that their mitochondrial genomes sizes were different. NIES254 was 53,406 bp while SCCAP K-0007 was 43,294 bp. The LSC, SSC, and IR regions for NIES254 were 4,184 bp, 2,807 bp, and 314 bp longer than those of SCCAP K-0007.

The mitochondrial genomes of two *P. parkeae* strains had similar gene content and gene arrangements. The lengths of most genes from the two strains were similar. The same (or at least a variant of the) trans-spliced *cox*1 intron present in *P. parkeae* SCCAP K-0007 appeared to be present in *cox*1 of *P. parkeae* NIES254. However, a tRNA (*trnT*) present in SCCAP K-0007 was absent in NIES254.The density of the coding regions in NIES254 (33,784 bp, 63.26%) was lower than that of SCCAP K-0007 (35,391 bp, 81.75%).

Mauve alignment analysis of synteny of the two *P. parkeae* mitochondrial genomes showed that these genomes exhibited a collinear relationship, as only one syntenic block from each strain was present (Fig. 3a). However, the alignment showed two large hotspot regions where similarity values were almost zero. The first hotspot region was located between *cox*1 and *cob* in LSC. The region was about 2.9 kb in NIES254 while it was about 150 bp in SCCAP K-0007. The second hotspot region was located between *rps*4-*trnF* and *trnI*-*trnP*-*nad*2 in IRs. The region was 2.4 kb in NIES254 while it was about 100 bp in SCCAP K-0007. The gene content of the *P. parkeae* mitochondrial genomes was similar.

LAST whole genome alignment resulted in 131 fragments, suggesting high variability in the intergenic regions. Due to the difficulty in identifying orthologous positions in those intergenic regions, we only identified SNPs and calculated the number of synonymous (Ks) and nonsynonymous (Ka) nucleotide substitutions along with Ka/Ks ratio of the protein coding regions. Ks values ranged from 12.8 (*nad*1) to 77.2 (*atp*6) nucleotide substitutions per 100 bp. Ka values ranged from 0 (*atp*9*, nad*4L*, and nad*10) to 17.2 (*rps*10) nucleotide substitutions per 100 bp. The Ka/Ks ratios ranged from 0 (*atp*9*, nad*4L*, and nad*10) to 0.9 (*rps*4) (Fig.4).

Comparative analyses of mitochondrial genomes of prasinophyte clade I and II

The organizational structures of the *P. parkeae* and the *C. tetramitiformis* mitochondrial genomes were similar to those of most Viridiplantae in having quadripartite structure. The mitochondrial genome of *P. parkeae* was smaller (53,406 bp) than that of *C. tetramitiformis*

(73,520 bp) but it had an equivalent number of coding genes, resulting in a higher density of coding regions (Table 3). The sizes of the inverted repeats of the two genomes were also different. *P. parkeae* had somewhat smaller inverted repeat regions (9,478 bp, containing *rps*2, *rps*4, *rps*12, *nad*1, *nad*2, *nad*6, and 10 tRNAs) than that of *C. tetramitiformis* (9,718 bp, containing *rns* and 3 tRNAs). However, the *P. parkeae* inverted repeats had a much higher density of coding regions (46.61%) than that of *C. tetramitiformis* (16.64%).

The gene content, density of coding regions, and gene order and syntenic regions of the mitochondrial genomes were also compared within and across prasinophyte clades I and II. Gene content of the algal genomes was similar. However, they exhibited a few differences: *atp*1 was absent in *C. tetramitiformis* and *B. prasinos*; *atp*4 was absent in *O. tauri* and *Micromonas* sp; *atp*9 was absent in *C. tetramitiformis*; *cob* was absent in *O. tauri*; *nad*10 was absent in *C. tetramitiformis* and *Micromonas* sp.; *rps*8 was absent in *P. parkeae*. Some genes were present in only a few species: *tatC* in *P. parkeae, mtt*2 in *Micromonas* sp and *Monomastix* sp., *ymf16*, *ymf39*, and *cyt*B in *O. tauri* (Robbens et al. 2007, Worden et al. 2009, Vaulot et al. 2012, Turmel et al. 2013, Hrdá et al. 2016).

The mitochondrial genomes of the prasinophytes from the same clade displayed more similarity in gene order and syntenic regions (Fig. 3b-c). The size of the genomes ranged from 43,294 bp (*P. parkeae* SCCAP K-0007) to 73,520 bp (*C. tetramitiformis*). The density of the protein coding regions ranged from 45.91% (33,753 bp, *C. tetramitiformis*) to 89.97 % (39,801 bp, *O. tauri*). Introns were present in *cox*1 of *P. parkeae* SCCAP K-0007, *cox*1 of *C. tetramitiformis,* and *cox*1 and *rnl* of *Monomastix* sp. (Fig. 5) (Robbens et al. 2007, Worden et al. 2009, Vaulot et al. 2012, Turmel et al. 2013, Hrdá et al. 2016). The *cox*1 introns in both *P. parkeae* SCCAP K-0007 and *C. tetramitiformis* were group II introns that were homologous to reverse transcriptase (cd01651), while the introns present in *Monomastix* sp. were homologous to LAGLIDADG homing endonuclease (Turmel et al. 2013).

Phylogenetic analyses

We used concatenated sequences of 32 mitochondrial protein-coding genes (Table 1) to construct a phylogenetic tree under ML and Baysian frameworks. As our main purpose was to examine relationships among members of prasinophytes, the taxonomic sampling of this study was focused on green algal and land plant diversity (62 taxa); for the out-groups, one red alga, two glaucophytes, two cryptophytes, and two jakobids were used. Our phylogenetic tree (Fig. 6) supported the close relationship between *P. parkeae* and *C. tetramitiformis* (representing prasinophyte clade I), with 100% bootstrap support and a posterior probability value of 1.0, to the exclusion of other green algae examined in this study, and indicated monophyly of prasinophyte clade II, Ulvophyceae, Trebouxiophyceae, Chlorophyceae, and streptophytes (streptophyte algae and embryophytes). However, the deeper level relationships among prasinophytes were not resolved. Most of the prasinophyte clades were resolved as paraphyletic assemblages.

To better resolve the deeper branching events, we added a data set of 74 chloroplast protein-coding sequences to the mitochondrial data. The final data set included an alignment of 25,059 amino acid positions of 51 taxa (a subset of the taxon list from Table 1). Similarly, to the mitochondrial analysis, the tree constructed from the combined data set robustly supported the close relationship between *P. parkeae* and *C. tetramitiformis* (representing prasinophyte clade I), and indicated the monophyly of prasinophyte clade II, Ulvophyceae, Trebouxiophyceae, Chlorophyceae, and streptophytes (streptophyte algae and embryophytes) (Fig. 7).

Discussion

The assembled mitochondrial genomes of *P*. *parkeae* NIES254 and *C. tetramitiformis* represent mitochondrial genomes for the prasinophyte clade I, providing the opportunity to revisit prasinophyte phylogeny using mitochondrial genome characters and to investigate mtDNA variability present within and among prasinophyte clades.

Comparative analyses of Pyramimonas parkeae NIES254 and SCCAP K-0007 *mitochondrial genomes*

Pyramimonas parkeae NIES254 and SCCAP K-0007 mitochondrial genomes are different in size and density of protein-coding regions. This difference was the result of changes in non-coding regions between the two strains (Fig. 5). There were two hotspot regions caused by insertions in *P*. *parkeae* NIES254 intergenic regions at 1) between *cox*1 and *cob* in LSC and 2) between *rps*4-*trnF* and *trnI*-*trnP*-*nad*2 in IRs. In the first hotspot region, we found an ORF (*orf4*, 2,742 bp) that is homologous to reverse transcriptase (cd01651). Similarly, in the second hotspot region, we found *orf77* (210 bp) and *orf78* (177 bp), which are homologous to partial sequences of putative integrase/recombinase protein (Fig. 3a). Ka/Ks ratio analyses suggest that the mitochondrial genes were subjected to purifying selection, indicated by Ka/Ks ratios ranging from 0 (*atp*9, *nad*4L, and *nad*10) to 0.9 (*rps*4) (Fig.4).

The presence of this intra-specific variability in the *P. parkeae* mitochondrial genomes might be a result of long divergence time as a similar degree of intra-specific polymorphism was also observed in another prasinophyte *O. tauri* (Blanc-Mathieu et al. 2013). While such differences might suggest the possibility that the two investigated *P. parkeae* strains are different species, a phylogeny based on 18S rDNA, *rbc*L, and chloroplast 16S rDNA sequences indicated that the two strains together form a monophyletic group separate from other studied *Pyramimonas* strains (Satjarak and Graham 2017). Additional molecular data from other *P. parkeae* strains and *Pyramimonas* species would be useful for future examination of intra-specific variability in mitochondrial genomes and its relevance to evolutionary diversification patterns.

Comparative analyses of mitochondrial genomes of prasinophyte clade I

Gene content and introns

The gene content of *P. parkeae* and *C. tetramitiformis* mitochondrial genomes was similar to that of other prasinophyte algae. All contain genes coding for protein subunits that are involved in electron transport, ATP synthesis, and translation (Table 2). However, there are slight differences between the gene content of *P. parkeae* and *C. tetramitiformis* mitochondrial genomes. The *P. parkeae* mitochondrial genome does not contain *rps8 – a* gene~~s~~ coding for small subunits of ribosomal proteins that is present in the mitochondrial genome of *C. tetramitiformis*. Perhaps the *rps8* gene has been transferred to the nuclear genome; however, the complete nuclear genome of *P. parkeae* is needed to investigate this possibility. This gene is also absent from the mitochondrial genomes of *Pycnococcus provasolii* (prasinophyte clade V) and

Pedinomonas minor, additional early-diverging green algal species (Turmel et al. 1999, Turmel et al. 2010), suggesting one or more independent loss events or transfer to the nuclear genomes.

The *P. parkeae* mitochondrial genome contains *atp*1, a gene coding for ATP synthase subunit 1, as do the mitochondrial genomes of *Ostreococcus tauri*, *Monomastix* sp., and *Nephroselmis olivacea*, but the majority of other prasinophyte mitochondrial genomes do not contain this gene (Turmel et al. 1999, Robbens et al. 2007, Turmel et al. 2013). Interestingly, while the currently available data suggest the general presence of *atp9* in prasinophyte mitochondrial genomes, we did not find this gene in the mitochondrial genome of *C. tetramitiformis*. *atp*1 and and *apt*9 are also absent in the mitochondrial genomes of *Pedinomonas* and *Chalmydomonas* spp. (Turmel et al. 1999, and the references therein).

Two group II introns were observed in the *cox*1 gene of *C. tetramitiformis*. A BLAST search suggested that these introns contained a reverse transcriptase (cd01651 domain), a protein commonly present in bacterial and mitochondrial genomes that originated by insertion of transposable elements (Zimmerly et al. 1995). According to a BLAST homology search, this protein domain was also present in *cox*1 introns of the Viridiplantae species *Oltmannsiellopsis viridis*, *Chlorokybus atmophyticus, Chara vulgaris*, *Nitella hyalina*, and *Physcomitrella patens* (Turmel et al. 2003, Pombert et al. 2006, Terasawa et al. 2007, Turmel et al. 2007, Turmel et al. 2013). However, the insertion sites differed, suggesting that these introns likely arose independently.

Gene arrangement and gene overlapping regions

A whole genome alignment between the two newly constructed mitochondrial genomes and prasinophyte clade II mitochondrial genomes showed less variability between *P. parkeae* NIES254 and *C. tetramitiformis* (Fig. 3b) than across the prasinophyte clades I and II (Fig. 84). The *P. parkeae* NIES254 and *C. tetramitiformis* mitochondrial genomes contain the same 5 gene clusters: 1) *atp*4 – *atp*8, 2) *cox*2 – *cox*3 – *atp*6, 3) *nad*4 – *rps*7 – *nad*5 – *rps*12, 4) *rps*11 – *rps*13 – *rpl*6, and 5) *rps*14 – *rpl*5 – *rpl*14 – *rpl*16 – *rps*3 – *rps*19 – *rps*10. The first and the last two clusters are also present in the mitochondrial genomes of *Andalucia godoyi* and *Reclinomonas americana*, jakobids remarkable for having the most bacterial-like, gene-rich mitochondrial genomes (Lang et al. 1997, Burger et al. 2013). Commonality of gene clusters among *P. parkeae*, *C. tetramitiformis* and the jakobids suggests that these clusters might represent relics of ancestral operons.

Both *P. parkeae* NIES254 and *C. tetramitiformis* mitochondrial genomes contained regions where two genes overlapped. Among the overlapping regions, two were commonly present in both algal species – overlapping regions at the 5′ end of *rpl*14 and the 3′ end of *rpl*16 and at the 5′ end of *rps*19 and the 3′ end of *rps*10. The overlap of *rpl*14 and *rpl*16 was also observed in mitochondrial genomes of prasinophyte *Bathycoccus prasinos* (Vaulot et al. 2012) and jakobid *Andalucia godoyi* (Burger et al. 2013).

Other overlapping regions present in *P. parkeae* NIES254 and *C. tetramiformis* were also observed in other mitochondrial genomes in several lineages. In *P. parkeae* NIES254 the overlapping region at the 5′ end of *rps*19 and the 3′ end of *rps*10 was also present in glaucophyte *Glaucocystis nostochinearum* (Price et al. 2012), and streptophyte algae *Chaetosphaeridium globusum* (Turmel et al. 2002), *Chara vulgaris* (Turmel et al. 2003), and *Nitella hyalina* (Turmel et al. 2013). The overlapping region at the 5′ end of *rps*11 and the 3′ end of *rps*13 was also present in cryptophyte *Hemiselmis andersenii* (Kim et al. 2008) and streptophyte alga *Microspora stagnorum* (Turmel et al. 2013). Similarly, for *C. tetramitiformis*, the overlapping region at the 5′ end of *rps*13 and the 3′ end of *rpl*6 was also present in streptophytes *Mesostigma viride* (Turmel et al. 2002), *Entransia fimbriata*, and *Roya obtusa* (Turmel et al. 2013). The overlapping region at the 5′ end of *rps*14 and the 3′ end of *rpl*5 was also present in jakobid *Andalucia godoyi* (Burger et al. 2013), cryptophyte *Rhodomomas salina* (Hauth et al. 2005), prasinophytes *Micromonas pusilla* (Worden et al. 2009) and *Ostreococcus tauri* (Robbens et al. 2007), streptophyte algae *Microspora stagnorum* and *Entransia fimbriata* (Turmel et al. 2013). The presence of these overlapping regions in various lineages may suggest the ancestral gene arrangement/ancestral operon or represent the trace of the mitochondrial genome re-arrangements that occurred during the evolutionary processes. On the other hand, that the overlapping regions occurred in different locations in either *P. parkeae* NIES254 or *C. tetramitiformis* suggests that they may have originated from independent events, where the evolution of the mechanisms related to translational coupling was preferred. This hypothesis of translational coupling due to overlapping regions in genes was suggested by Burger et al. (1995) after observing the protein products of the overlapping mitochondrial genes in stoichiometric amounts.

Comparative analyses of mitochondrial genomes of prasinophyte clade I and II

Previous phylogenetic analyses using 16S rDNA, 28S rDNA, complete chloroplast genomes (Lemieux et al. 2014, Leliaert et al. 2016), together with our phylogenetic pattern inferred from complete mitochondrial genomes suggested a close relationship between prasinophytes clade I and II. To gain more information between these closely related clades, we performed comparative analyses across prasinophyte clade I and II.

Our comparison showed that the gene content and the size of the coding regions were similar for the prasinophytes used in the comparison. The difference in genome size was related to different sizes of non-coding regions (Fig. 5). The presence of group II introns in *cox*1 of *P. parkeae* SCCAP K-0007 and *C. tetramitiformis*, along with the presence of *orf4*—an ORF that is located next to *cox*1 in *P. parkeae* NIES254, suggests that these introns and ORF might have originated from a common event, as they are present in a similar region and are similarly homologous to group II introns (reverse transcriptase, cd01651). Alternatively, the introns present in *Mononastix* sp. might have arisen from a different event, as a similar sequence of LAGLIDADG homing endonuclease is not found in the introns of mitochondrial genomes of prasinophyte clades I and II.

Our analyses of gene arrangement using Mauve alignment and MLGO showed that syntenic regions and gene arrangement of the mitochondrial genomes from prasinophytes of the same clade showed higher similarity than that of different clades (Fig. 3b-c). However, the results from the phylogenetic analyses using gene order were not congruent with those inferred from protein coding sequences (Fig. 8). This incongruence suggested different evolutionary patterns of gene re-arrangement and nucleotide substitution that have accumulated in the algal mitochondrial genomes. However, this comparison was made from molecular information for only a few algal species. More inclusive taxon sampling is needed to increase the resolution of relationships among the prasinophytes of clades I and II.

Phylogenetic analyses

Our newly sequenced mitochondrial genomes of *P. parkeae* NIES254 and *C. tetramitiformis* allowed us to explore the relationship of these two species among prasinophytes. A phylogenetic analysis based on concatenated mitochondrial protein-encoded sequences (Fig. 6) revealed a monophyletic Viridiplantae, consistent with other molecular analyses (Rodríguez-Ezpeleta et al. 2007, Turmel et al. 2007, Leliaert et al. 2016). The concatenated mitochondrial data also indicated monophyly of a clade containing *P. parkeae* and *C. tetramitiformis*, as well as monophyly of several other eukaryotic clades: jacobids, cryptophytes, rhodophytes, prasinophyte clades II, Ulvophyceae, Trebouxiophyceae, Chlorophyceae, and streptophytes, and overall tree topology was consistent with other molecular analyses (Kim et al. 2008, Pombert et al. 2010, Burger et al. 2013, Smith et al. 2013, Turmel et al. 2013, Jackson and Reyes-Prieto 2014, Leliaert et al. 2016, Zhou et al. 2016). The occurrence of a clade consisting of *P. parkeae* and *C. tetramitiformis* was also consistent with our whole mitochondrial genome alignment results (Figs. 3 & 8) that showed more similarity in genome organization between these two species than among other prasinophytes.

To resolve the deep branching events, we combined the chloroplast protein-encoded sequences to the mitochondrial analyses. The tree generated from a combined data set (Fig. 7) was similar to that from mitochondrial data (Fig. 6) in resolving a monophyly of a clade containing *P. parkeae* and *C. tetramitiformis*, as well as monophyly of several other eukaryotic clades: prasinophyte clade II, Ulvophyceae, Trebouxiophyceae, Chlorophyceae, and streptophytes. The bootstrap support and posterior probability values for some of the deep-branching nodes were increased in the combined mitochondrial and chloroplast tree.

Both the mitochondrial tree and the combined mitochondrial and chloroplast tree suggest the hypothesized clade consisting of *P. provasolii* (prasinophyte clade V) and *P. coloniales* (VI). However, this clade is not congruent with results from other studies based on chloroplast or nuclear gene markers (e.g. Leliaert et al. 2012, Turmel et al. 2013, Lemieux et al. 2014 Leliaert et al. 2016). Also, we did not observe this clade in any single gene trees generated from both mitochondrial and plastid data (not shown) and the representation of these two clades is limited. Therefore, we cannot rule out the possibility that the union of the two taxa in our concatenated tree stemmed from an artefact of long-branch attraction or from limited taxon sampling.

The topologies of the trees inferred from mitochondrial genomes (Fig. 6) and from the combined data of mitochondrial and chloroplast genomes (Fig. 7) were slightly different. One difference was the varying position of prasinophyte clade III (*Nephroselmis olivaceae*). The clade was sister to streptophytes in the tree estimated from mitochondrial data, whereas it was sister to a clade consisting of prasinophyte clade I and II in the tree estimated from the combined data. Another difference in topology was the position of *Pycnococcus provasolii* (clade V) plus *Prasinoderma colonial* (clade VI). In the mitochondrial tree, they were sister to a clade consisting of Pedinophyceae, Ulvophyceae, and Chlorophyceae whereas, in the combined tree, they were sister to a clade consisting of Pedinophyceae, Chlorophyceae, Ulvophyceae, and Trebouxiophyceae. Last, in the mitochondrial tree, Charales was sister to embryophytes whereas, in the combined tree, Zygnematales was sister to the embryophytes.

The difference in topology of *Nephroselmis olivaceae* (prasinophyte clade III) and *Prasinoderma coloniale* (prasinophyte clade VI) might have resulted from limited molecular data and limited taxon sampling, as other studies have shown that different sources of molecular data and different numbers of taxa used in phylogenetic analyses can result in different topologies (Robbens et al. 2007, Turmel et al. 2013, Lemieux et al. 2014, Leliaert et al. 2016).

The presence of prasinophyte clade V (*Pycnococcus provasolii*) as a sister group to core Chlorophyta (Ulvophyceae, Trebouxiophyceae, and Chlorophyceae) and the presence of Zygnematales as a sister group to embryophytes in the combined mitochondrial and chloroplast tree is congruent with the results of previous studies from complete chloroplast genomes (Leliaert et al. 2012, Lemieux et al. 2014, Leliaert et al. 2016). Differences of these topologies in the mitochondrial tree vs the combined mitochondrial and chloroplast tree likely arose from the addition of the chloroplast data, which contributed more information in the phylogenetic analysis (17,648 aligned amino acid positions of a total of 25,059 aligned amino acid positions).

Summary

Understanding the early diversification of Viridiplantae is of interest, because it can reveal fundamental traits of the green plants on which humans depend and of ecologically important green algae. However, most phylogenetic analyses focused on the early diversification of Viridiplantae, particularly diversification of prasinophytes, have been based on chloroplast genomes, which, though valuable, have not yet yielded adequate resolution of the group. As an alternative approach, we employed newly assembled mitochondrial genomes for an important prasinophyte clade and analyzed them with available mitochondrial genomic data from other groups to infer prasinophyte relationships. To improve phylogenetic resolution, we constructed the complete mitochondrial genomes of prasinophytes from clade I, *P. parkeae* and *C. tetramitiformis*. Comparisons among mitochondrial genomes in prasinophytes showed that the newly constructed mitochondrial genomes of *P. parkeae* and *C. tetramitiformis* were more similar to each other than to other prasinophytes. Our phylogenetic analyses of concatenated amino acid sequences derived from mitochondrial genomes and from both mitochondrial and chloroplast genomes of selected taxa resolved a clade consisting of *P. parkeae* and *C. tetramitiformis* representing prasinophyte clade I, and its sister relationship to the clade II. However, the analyses indicated that existing mitochondrial and chloroplast genomes were not sufficient to firmly resolve deep branching events, e.g., relationships among prasinophyte clades. An improved taxon sampling in combination with more data from particularly from the nuclear compartment are still needed for resolving the earliest events in Viridiplantae diversification.

Acknowledgements

This study was supported in part by NSF grants DEB-1119944 awarded to L.G. and CAREER-1453639 awarded to E.K. We thank the Laboratory of Genetics for access to computational infrastructure and support.

References

Blanc-Mathieu, R., Sanchez-Ferandin, S., Eyre-Walker, A. & Piganeau, G. 2013. Organellar inheritance in the Green Lineage: insights from *Ostreococcus tauri*. *Genome Biol. Evol.* 5(8):1503-1511.

Bolger, A.M., Lohse, M. & Usadel, B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* btu170.

Burger, G., Plante, I., Lonergan, K.M. & Gray, M.W. 1995. The mitochondrial DNA of the amoeboid protozoon, *Acanthamoeba castellanii*: complete sequence, gene content and genome organization. *J. Mol. Biol.* 245(5):522-537.

Burger, G., Gray, M.W. & Lang, B.F. 2003. Mitochondrial genomes: anything goes. *Trends Genet.* 19(12):709-716.

Capella-Gutiérrez, S., Silla-Martínez, J.M. & Gabaldón, T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25(15):1972-1973.

Darling, A.E., Mau, B. & Perna, N.T. 2010. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PloS ONE* 5(6):e11147.

Darriba, D., Taboada, G.L., Doallo, R. & Posada, D. 2011. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* 27(8):1164-1165.

Goff, S.A., Vaughn, M., McKay, S., Lyons, E., Stapleton, A.E., Gessler, D., Matasci, N., Wang, L., Hanlon, M., Lenards, A. & Muir, A. 2011. The iPlant collaborative: cyberinfrastructure for plant biology. *Front. Plant Sci.* 2.

Graham, L.E., Graham, J.M., Wilcox, L.W., Cook, M.E. 2016. Algae. 3rd ed.
LJLM Press, Madison, WI, 595 pp.

Gray, M.W., Burger, G. & Lang, B.F. 1999. Mitochondrial evolution. *Science* 283(5407):1476-1481.

Hahn, C., Bachmann, L. & Chevreux, B. 2013. Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—a baiting and iterative mapping approach. *Nucleic Acids Res.* 41(13):e129-e129.

Hrdá, Š., Hroudová, M., Vlček, Č. & Hampl, V. 2016. Mitochondrial Genome of Prasinophyte Alga *Pyramimonas parkeae*. *J. Euk. Microbiol.* 0:1-10.

Hu, F., Lin, Y. & Tang, J. 2014. MLGO: phylogeny reconstruction and ancestral inference from gene-order data. *BMC bioinformatics* 15:354.

Jackson, C.J. & Reyes-Prieto, A. 2014. The mitochondrial genomes of the glaucophytes *Gloeochaete wittrockiana* and *Cyanoptyche gloeocystis*: multilocus phylogenetics suggests a monophyletic Archaeplastida. *Genome Biol. Evol.* 6(10):2774-2785.

Katoh, K. & Standley, D.M. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30(4):772-780.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C. & Thierer, T. 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28(12):1647-1649.

Kiełbasa, S.M., Wan R., Sato K., Horton P. & Frith M.C. 2011. Adaptive seeds tame genomic sequence comparison. *Genome Res.* 21(3):487-93.

Kim, E., Lane, C.E., Curtis, B.A., Kozera, C., Bowman, S. & Archibald, J.M. 2008. Complete sequence and analysis of the mitochondrial genome of *Hemiselmis andersenii* CCMP644 (Cryptophyceae). *BMC genomics* 9(1):215.

Kumar, S., Stecher, G. and Tamura, K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* 33(7):1870–1874.

Lagesen, K., Hallin, P., Rødland, E.A., Stærfeldt, H.H., Rognes, T. & Ussery, D.W. 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* 35(9):3100-3108.

Lang, B.F., Gray, M.W. & Burger, G. 1999. Mitochondrial genome evolution and the origin of eukaryotes. *Annu. Rev. Genet.* 33(1):351-397.

Lemieux, C., Otis, C. & Turmel, M. 2014. Six newly sequenced chloroplast genomes from prasinophyte green algae provide insights into the relationships among prasinophyte lineages and the diversity of streamlined genome architecture in picoplanktonic species. *BMC genomics* 15(1):1.

Li, H. & Durbin, R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics 25*(14):1754-1760.

Leliaert, F., Smith, D.R., Moreau, H., Herron, M.D., Verbruggen, H., Delwiche, C.F. & De Clerck, O. 2012. Phylogeny and molecular evolution of the green algae. *Crit. Rev.Plant Sci.* 31(1):1-46.

Leliaert, F., Tronholm, A., Lemieux, C., Turmel, M., DePriest, M.S., Bhattacharya, D., Karol, K.G., Fredericq, S., Zechman, F.W. and Lopez-Bautista, J.M. 2016. Chloroplast phylogenomic analyses reveal the deepest-branching lineage of the Chlorophyta, Palmophyllophyceae class. nov. *Sci.Rep*. 6:25367

Lohse, M., Drechsel, O., Kahlau, S. & Bock, R. 2013. OrganellarGenomeDRAW—a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. *Nucleic Acids Res.* gkt289.

Miller, M.A., Pfeiffer, W. & Schwartz, T. 2010. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. In Gateway Computing Environments Workshop (GCE), *IEEE*. 1-8.

Nosek, J. & Tomáška, Ľ. 2003. Mitochondrial genome diversity: evolution of the molecular architecture and replication strategy. *Curr. Genet.* 44(2):73-84.)

Osellame, L.D., Blacker, T.S. & Duchen, M.R. 2012. Cellular and molecular mechanisms of mitochondrial function. *Best Pract. Res. Clin. Endocrinol. Metab*. 26(6):711-723.

Pombert, J.F. & Keeling, P.J. 2010. The mitochondrial genome of the entomoparasitic green alga *Helicosporidium*. *PLoS ONE* 5(1):e8954.

Pombert, J.F., Otis, C., Turmel, M. & Lemieux, C. 2013. The mitochondrial genome of the prasinophyte *Prasinoderma coloniale* reveals two trans-spliced group I introns in the large subunit rRNA gene. *PloS ONE* 8(12):e84325.

Quinlan, A.R. 2014. BEDTools: the Swiss-army tool for genome feature analysis. *Curr. Protoc. Bioinformatics* 11-12.

Robbens, S., Derelle, E., Ferraz, C., Wuyts, J., Moreau, H. & Van de Peer, Y. 2007. The complete chloroplast and mitochondrial DNA sequence of *Ostreococcus tauri*: organelle genomes of the smallest eukaryote are examples of compaction. *Mol. Biol. Evol.* 24(4):956-968.

Rodríguez-Ezpeleta, N., Brinkmann, H., Burger, G., Roger, A.J., Gray, M.W., Philippe, H. & Lang, B.F. 2007. Toward resolving the eukaryotic tree: the phylogenetic positions of jakobids and cercozoans. *Cur. Biol*. 17(16):1420-1425.

Ronquist, F. & Huelsenbeck, J.P. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19(12):1572-1574.

Satjarak, A., Paasch, A.E., Graham, L.E., Kim, E. 2016. Complete chloroplast genome sequence of phagomixotrophic green alga *Cymbomonas tetramitiformis. Genome Announc.* 4(3):e00551-16.

Satjarak. A. & Graham, L.E. 2017. Comparative DNA sequences analyses of *Pyramimonas parkeae* (Prasinophyceae) chloroplast genomes. *J. Phycol.* In press.

Schattner, P., Brooks, A.N. & Lowe, T.M. 2005. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res.* 33(suppl 2):W686-W689.

Smith, D.R., Hua, J., Archibald, J.M. & Lee, R.W. 2013. Palindromic genes in the linear mitochondrial genome of the nonphotosynthetic green alga *Polytomella magna*. *Genome Biol. Evol.* 5(9):1661-1667.

Smith, D.R. & Keeling, P.J. 2015. Mitochondrial and plastid genome architecture: reoccurring themes, but significant differences at the extremes. *Proc. Natl. Acad. Sci. U.S.A.* 112(33):10177-10184.

Stamatakis, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* btu033.

Turmel, M., Lemieux, C., Burger, G., Lang, B.F., Otis, C., Plante, I. & Gray, M.W. 1999. The complete mitochondrial DNA sequences of *Nephroselmis olivacea* and *Pedinomonas minor*: two radically different evolutionary patterns within green algae. *Plant Cell* 11(9):1717-1729.

Turmel, M., Otis, C. & Lemieux, C. 2010. A deviant genetic code in the reduced mitochondrial genome of the picoplanktonic green alga *Pycnococcus provasolii*. *J. Mol. Evol.* 70(2):203-214.

Turmel, M., Otis, C. & Lemieux, C. 2013. Tracing the evolution of streptophyte algae and their mitochondrial genome. *Genome Biol. Evol.* 5(10):1817-1835.

Vaulot, D., Lepere, C., Toulza, E., De la Iglesia, R., Poulain, J., Gaboyer, F., Moreau, H., Vandepoele, K., Ulloa, O., Gavory, F. & Piganeau, G. 2012. Metagenomes of the picoalga *Bathycoccus* from the Chile coastal upwelling. *PLoS ONE* 7(6):e39648.

Worden, A.Z., Lee, J.H., Mock, T., Rouzé, P., Simmons, M.P., Aerts, A.L., Allen, A.E., Cuvelier, M.L., Derelle, E., Everett, M.V. & Foulon, E. 2009. Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes *Micromonas*. *Science* 324(5924):268-272.

Zhou, L., Wang, L., Zhang, J., Cai, C. & He, P. 2016. Complete mitochondrial genome of *Ulva linza*, one of the causal species of green macroalgal blooms in Yellow Sea, China. *Mitochondrial DNA Part B* 1(1):31-33.

Table 1. Taxa used in phylogenetic analyses. The asterisk (*) indicates the taxon used only in phylogenetic estimation using concatenated amino acids sequences from mitochondrial genomes. The colored box indicates the presence of mitochondrial gene and the empty box indicates the absence of the gene.

Organisms	mitochondrial genome accession #	Mitochondrial genes																																chloroplast genome accession #
		atp 1	*atp 4*	*atp 6*	*atp 8*	*atp 9*	*cob*	*cox 1*	*cox 2*	*cox 3*	*nad 1*	*nad 2*	*nad 3*	*nad 4*	*nad 4L*	*nad 5*	*nad 6*	*nad 7*	*nad 9*	*rpl 5*	*rpl 6*	*rpl 14*	*rpl 16*	*rps 2*	*rps 3*	*rps 4*	*rps 7*	*rps 10*	*rps 11*	*rps 12*	*rps 13*	*rps 14*	*rps 19*	
Prasinophytes																																		
Clade I																																		
Pyramimonas parkeae NIES254	KX013547.1	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	KX013546.1
Pyramimonas parkeae SCCAP K-007*	KX756655	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	
Cymbomonas tetramitiformis	KX013548.1		■	■	■		■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	KX013545.1
Clade II																																		
Bathycoccus prasinos	FO082258.2		■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	NC_024811
Micromonas sp. RCC299	FJ859351.1		■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	FJ858267
Micromonas pusilla (partial)*	FJ858268.1		■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■		■	■	■	■				■	■	■	
Ostreococcus tauri	CR954200.2	■		■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	NC_008289
Monomastix sp. OKE	KF060939.1	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	FJ493497
Clade III																																		
Nephroselmis olivacea	AF110138.1	■		■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	AF137379
Clade V																																		
Pycnococcus provasolii	GQ497137.1		■	■	■	■	■	■	■	■	■	■	■	■	■	■	■								■	■				■				FJ493498
Clade VI																																		
Prasinoderma coloniale	KF387569.1			■	■	■	■	■	■	■	■	■	■	■	■	■	■		■	■	■	■	■		■	■			■	■	■	■	■	KJ746598
Pedinophyte																																		
Pedinomonas minor	AF116775.1			■	■		■	■			■	■	■	■	■	■	■																	FJ968740
Ulvophyceae																																		
Oltmannsiellopsis viridis	DQ365900.1	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■				■	■	■				■	■	■	■	■	DQ291132
Ulva fasciata	KT364296.1	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■		■		■	■	■	■	■		■	■	■	■	■	■	KT882614
Ulva linza	KU189740.1	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■		■		■	■	■		■		■	■	■	■	■	■	KX058323
*Ulva prolifera**		■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■		■		■	■	■	■	■		■	■	■	■	■	■	
Pseudendoclonium akinetum	AY359242.1	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■		■		■	■	■	■	■		■	■	■	■	■	■	AY835431
Trebouxiophyceae																																		
Chlorella sorokiniana	KM241869.1	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■		■	■	■	■	■	■	■	■	■	■	■	KJ742376
Chlorella sp. ArM0029B	KF554428.1	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■		■	■	■	■	■	■	■	■	■	■	■	KF554427
Chlorella variabilis	KM252919.1	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■		■	■	■	■	■	■	■	■	■	■	■	KP271969
Auxenochlorella protothecoides	KC843974.1	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■		■	■	■	■	■	■	■	■	■	■	■	KC843975
Prototheca wickerhamii	U02970.1	■		■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■		■	■	■	■	■	■	■	■	■	■	■	KJ001761
Helicosporidium sp.	GQ339576.1	■	■	■	■	■	■		■	■	■	■	■	■	■	■	■	■	■	■	■		■	■	■	■	■	■	■	■	■	■	■	DQ398104
Botryococcus braunii	KR057902.1	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■			■	■	■	■	■	■	■	■	■	■	■	KM462884
Lobosphaera incisa	KP902678.1	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■		■	■	■	■	■	■	■	■	■	■	■	KM462871
Trebouxiophyceae sp. MX-AZ01	JX315601.1	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■			■	■	■	■	■	■		■	■	■	■	JX402620
Chlorophyceae																																		
Coccomyxa sp. C-169	HQ874522.1	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■			■	■	■	■	■	■		■	■	■	■	HQ693844
*Chlamydomonas leiostraca**	KP696389.1						■	■			■	■		■		■	■																	
*Chlamydomonas moewusii**	AF008237.1						■	■			■	■		■		■	■																	
Chlamydomonas reinhardtii	U03843.1						■	■			■	■		■		■	■																	FJ423446
Gonium pectorale	AP012493.1						■	■			■	■		■		■	■																	AP012494
Pleodorina starrii	JX977845.1						■	■			■	■		■		■	■																	JX977846
*Ourococcus multisporus**	KJ806272.1			■		■	■	■	■	■	■	■	■	■	■	■	■																	
*Polytomella capuana**	EF645804.1						■	■			■	■		■		■	■																	
Dunaliella salina	GQ250045.1						■	■			■	■		■		■	■																	GQ250046
*Dunaliella viridis**	KP691602.1						■	■			■	■		■		■	■																	
*Polytomella magna**	KC733827.1						■	■			■	■		■		■	■																	
*Polytoma uvella**	KP696388.1						■	■			■	■		■		■	■																	
Bracteacoccus aerius strain UTEX 1250	KJ806265.1			■		■	■	■	■	■	■	■	■	■	■	■	■																	KT199254
Bracteacoccus minor	KJ806266.1			■		■	■	■	■	■	■	■	■	■	■	■	■																	KT199253
Chlorotetraedron incus	KJ806267.1			■		■	■	■	■	■	■	■	■	■	■	■	■																	KT199252
Chromochloris zofingiensis	KJ806268.1			■		■	■	■	■	■	■	■	■	■	■	■	■																	KT199251
Mychonastes homosphaera	KJ806270.1			■		■	■	■	■	■	■	■	■	■	■	■	■																	KT199249
Pseudomuriella schumacherensis	KJ806273.1			■		■	■	■	■	■	■	■	■	■	■	■	■																	KT199256
streptophyte algae																																		
Mesostigma viride	AF353999.1	■		■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	AF166114
Chlorokybus atmophyticus	EF463011.1	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	DQ422812
Klebsormidium flaccidum (partial)	DF238763.1	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■			■	■		■	■	■		■	■		■		KJ461680
*Microspora stagnorum**	KF060942.1	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■		■	■	■	■	■	■		■	■	■	■		
Entransia fimbriata	KF060941.1	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■		■	■	■	■		KU646490
Chara vulgaris	AY267353.1	■		■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■		■	■	DQ229107
*Nitella hyalina**	JF810595.1	■	■	■	■	■	■		■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■		■	■	
Chaetosphaeridium globosum	AF494279.1	■		■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■		■	■	■	■	■	■	■	■	■	■	■	AF494278
Roya obtusa	KF060943.1	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■		■	■	■	■	■	■	■	■	■	■	■	KU646496
Closterium baillyanum	KF060940.1	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■		■	■	■	■	■	■	■	■	■	■	■	NC_030314
Embryophytes																																		
Marchantia polymorpha	NC001660.1	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■		■	■	■		■	■	■	■	■	■	■	■	■	■	■	NC_001319.1
Physcomitrella patens	AB251495.1	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■		■	■	■	■	■		■	■	■	■	■	AP005672
*Anomodon rugelii**	JF973314.1	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■		■	■	■	■	■		■	■	■	■	■	
Cycas taitungensis	AP009381.1	■	■	■	■	■	■	■	■	■	■	■	■	■	■		■	■	■	■			■	■	■	■	■	■		■	■	■	■	NC_009618
Oryza sativa	JF281153.1	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■			■	■	■	■	■		■	■	■		■	JN861110
Beta vulgaris	FP885845.1	■		■	■	■	■	■	■	■	■	■	■	■	■		■	■	■	■					■	■	■			■	■			KR230391
Nicotiana tabacum	KR780036.1	■	■	■		■	■	■	■	■	■	■	■	■	■	■	■	■	■	■			■		■	■		■		■	■	■	■	Z00044
Arabidopsis thaliana	Y08501.2	■		■		■	■	■	■	■	■	■	■	■	■		■	■	■	■			■		■	■	■			■				KX551970
Jakobids																																		
*Andalucia godoyi**	KC353352.1	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	
*Reclinomonas americana**	AF007261.1	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	
Cryptophytes																																		
Rhodomonas salina	NC002572.1	■		■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■		■	■	■	■	■	EF508371
*Hemiselmis andersenii**	EU651892.1	■		■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■		■	■	■	■	■	
Rhodophyte																																		
Cyanidioschyzon merolae	D89861.1			■	■	■	■	■	■	■	■	■	■	■	■	■	■			■	■	■	■		■	■			■	■		■		AB002583
Glaucophytes																																		
*Glaucocystis nostochinearum**	NC015117			■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■		■		■		■	■	■	■	■		■	
Cyanophora paradoxa	HQ849544.1		■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■		■	■	■	■	■	■	■		■	U30821

Table 2. Functional classification of the genes present in the mitochondrial genomes of *P. parkeae* and *C. tetramitiformis.*

Biological Process	Genes
Translation	
LSU ribosomal protein	*rp*l5,6,14,16
SSU ribosomal protein	*rps*2,3,4,7,8[b],10,11,12,13,14,19
Ribosomal RNAs	*rnl*, *rns*
Transfer RNAs	*trnA*,*C*,*D*,*E*,*F*,*G*,*H*,*I*,*K*,*L*,*M*,*N*,*P*,*Q*,*R*,*S*,*T*, *W*,*Y*
Electron transport and ATP synthesis	
NADH dehydrogenase (complex I) subunits	*nad*1,2,3,4,4L,5,6,7,8,9
Cytochrome c oxidase (complex IV) subunits	*cox*1[c],2,3
Cytochrome bc1 complex (complex II) subunits	*cob*
ATP synthase (complex V) subunits	*atp*1[a]4,6,8,9[a]
Translocase protein	*tatC*[a]

[a] Genes present in *P. parkeae* but not in *C. tetramitiformis* mitochondrial genomes

[b] Genes present in *C. tetramitiformis* but not in *P. parkeae* genomes

[c] Genes with introns in *C. tetramitiformis*

Table 3. Comparison of *P. parkeae* and *C. tetramitiformis* mitochondrial genome features

	Pyramimonas parkeae	*Cymbomonas tetramitiformis*
Genome size (bp)	53,406	73,520
Percent coding DNA (including introns)	68.35	56.70%
%GC	30.5	37.8
Size of inverted repeat regions (percent of total genome length)	9,478 bp (17.75%)	9,718 (13.22%)
Group II introns	Not present	2 group II introns are present in *cox*1
Number of genes	58 genes (22 tRNAs)	56 genes (23 tRNAs)

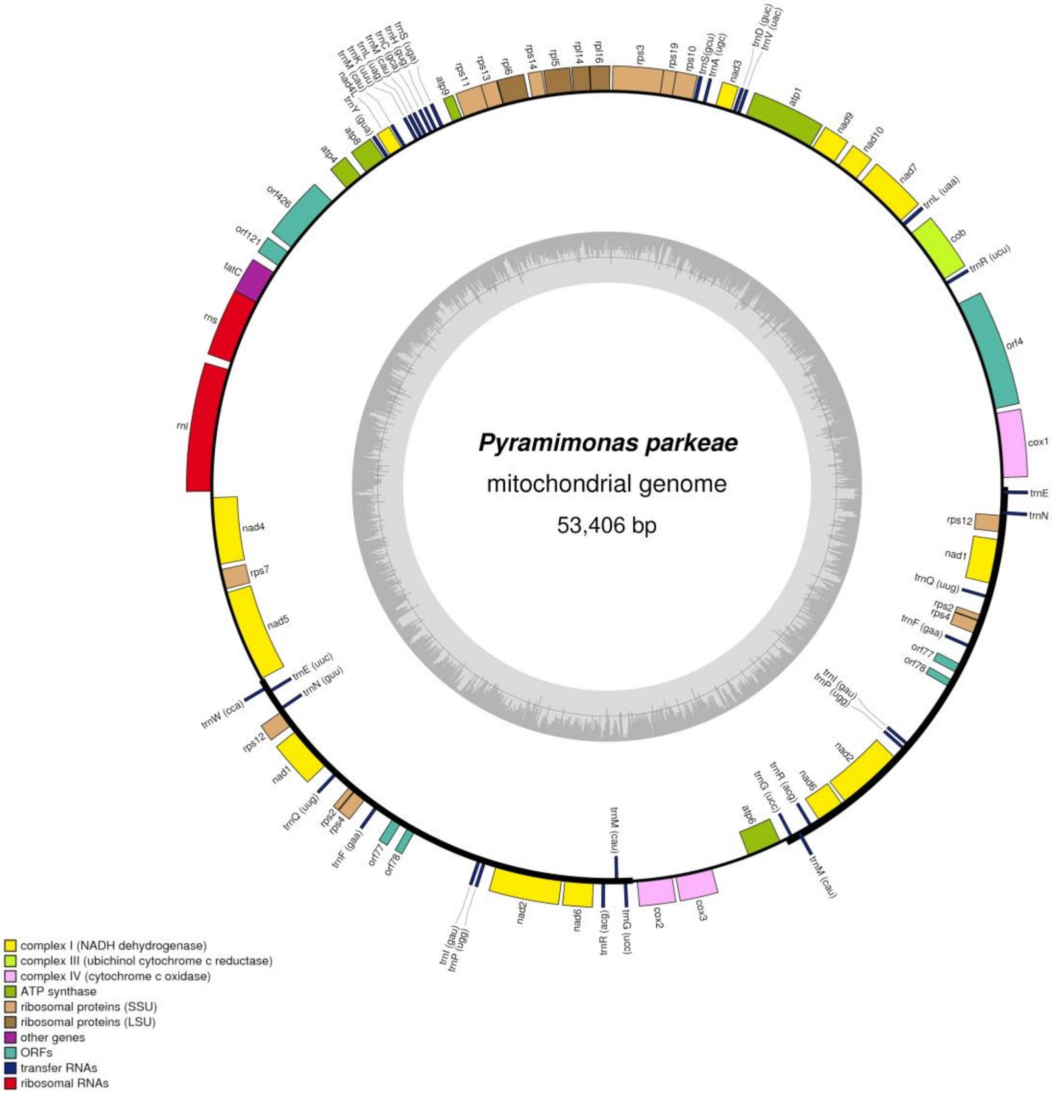


Figure 1. Map of *Pyramimonas parkeae* NIES254 mitochondrial genome. Genes positioned on the outside of the map are transcribed counter clockwise and those inside the map are transcribed clockwise. The thick lines indicate the extent of the inverted repeat regions.

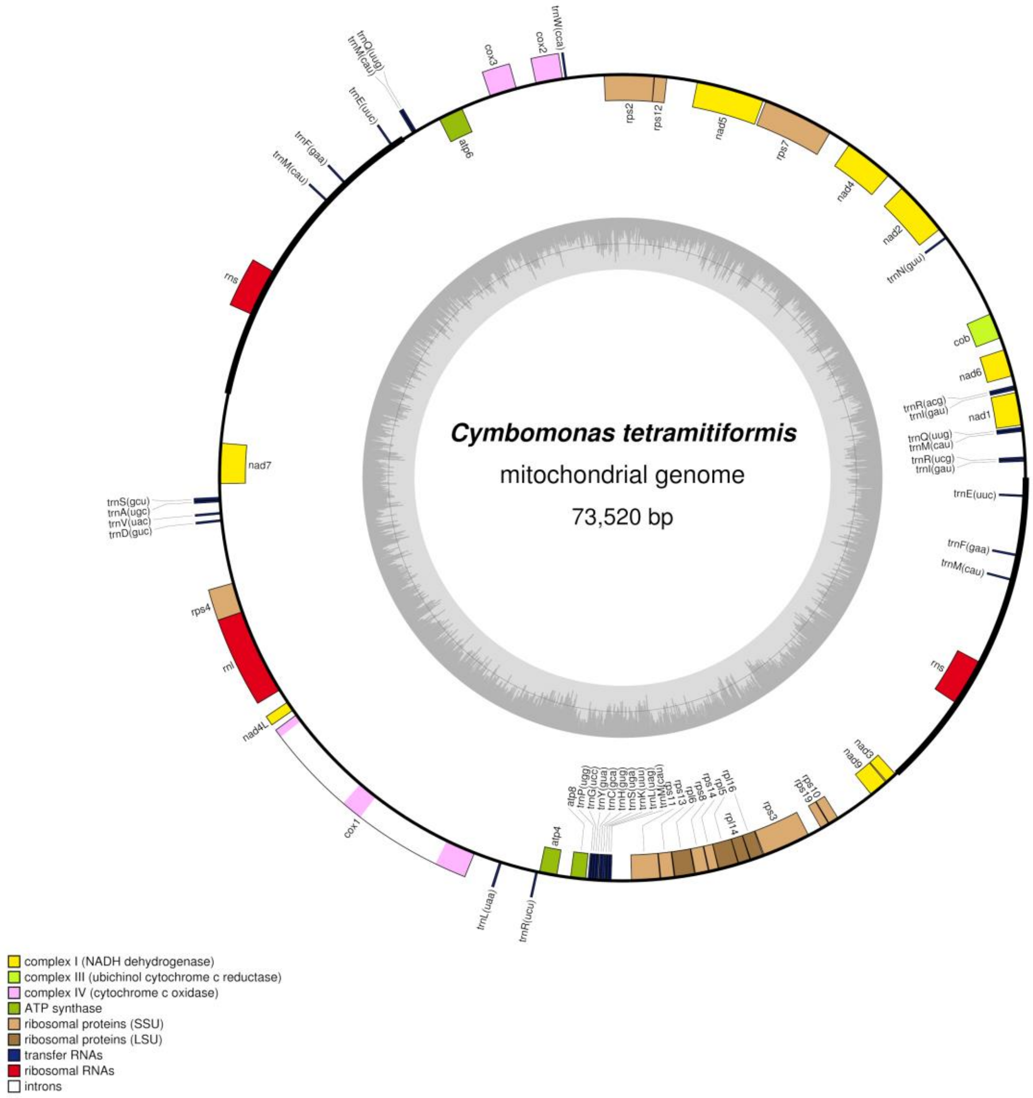


Figure 2. Map of *Cymbomonas tetramitiformis* PLY262 mitochondrial genome. Genes positioned on the outside of the map are transcribed counter clockwise and those inside the map are transcribed clockwise. The thick lines indicate the extent of the inverted repeat regions.

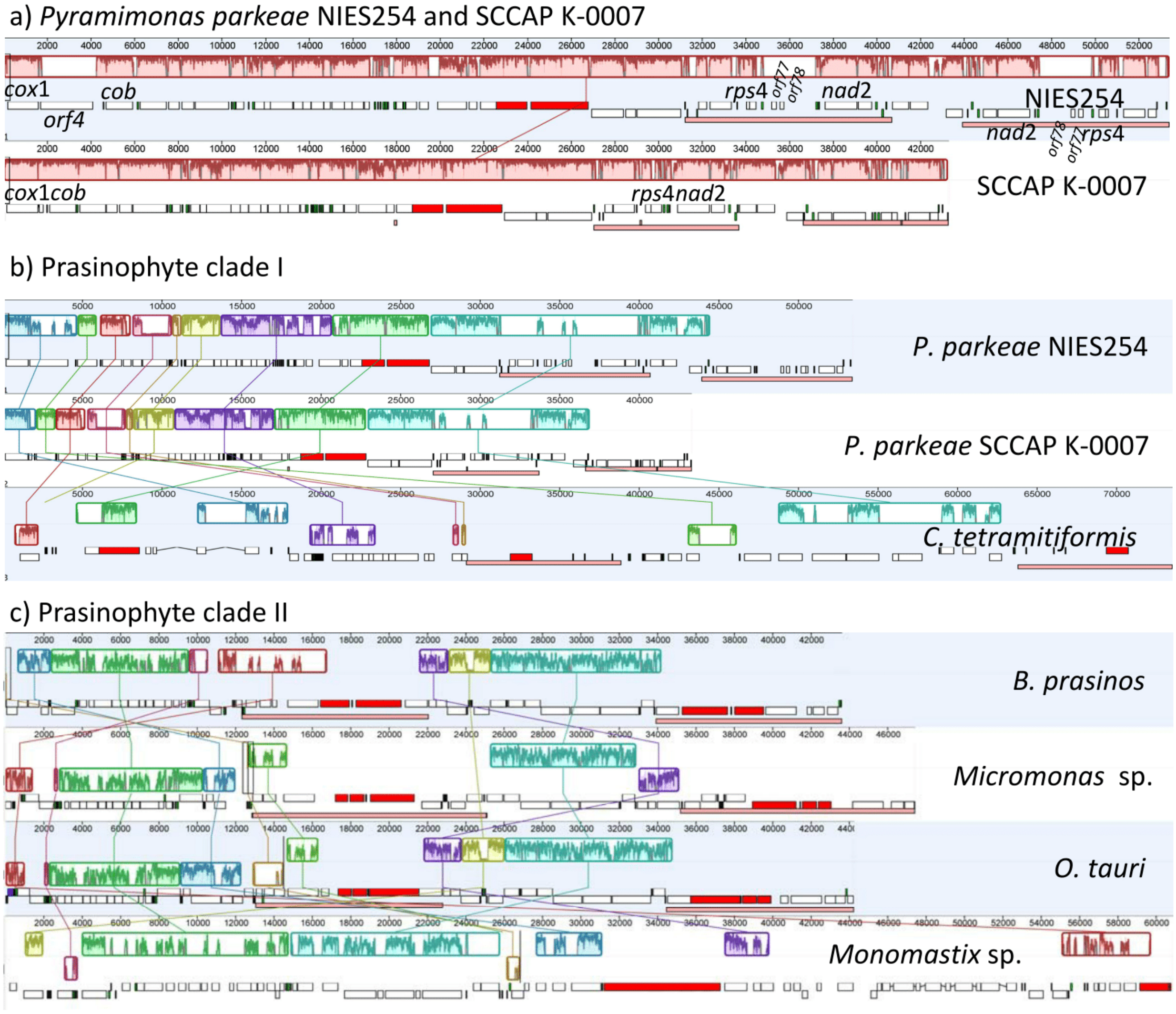


Figure. 3 Mauve alignment of mitochondrial genomes of a) *Pyramimonas parkeae* NIES254 and *P. parkeae* SCCAP K-007, b) prasinophyte clade I including *P. parkeae* NIES254 and SCCAP K-007 and *Cymbomonas tetramitiformis* PLY262, and c) prasinophyte clade II including *Bathycoccus prasinos*, *Micromonas* sp. RCC299, *Ostreococus tauri*, and *Monomastix* sp. OKE. The algal mitochondrial genomes with quadripartite structure were arranged in the same direction (LSC-IRB-SSC-IRA). Regions of homology between prasinophyte species are represented by syntenic blocks of the same color and are connected by vertical bars. The histogram inside each block represents pairwise nucleotide sequence identity. The three large

areas where the heights of the histograms are almost equal to zero (a) represent the large hotspot regions in *P. parkeae* NIES254 and SCCAP K-0007 at 1) between *cox*1 and *cob* in LSC and 2) between *rps*4-*trnF* and *trnI*-*trnP*-*nad*2 *in* IRs.

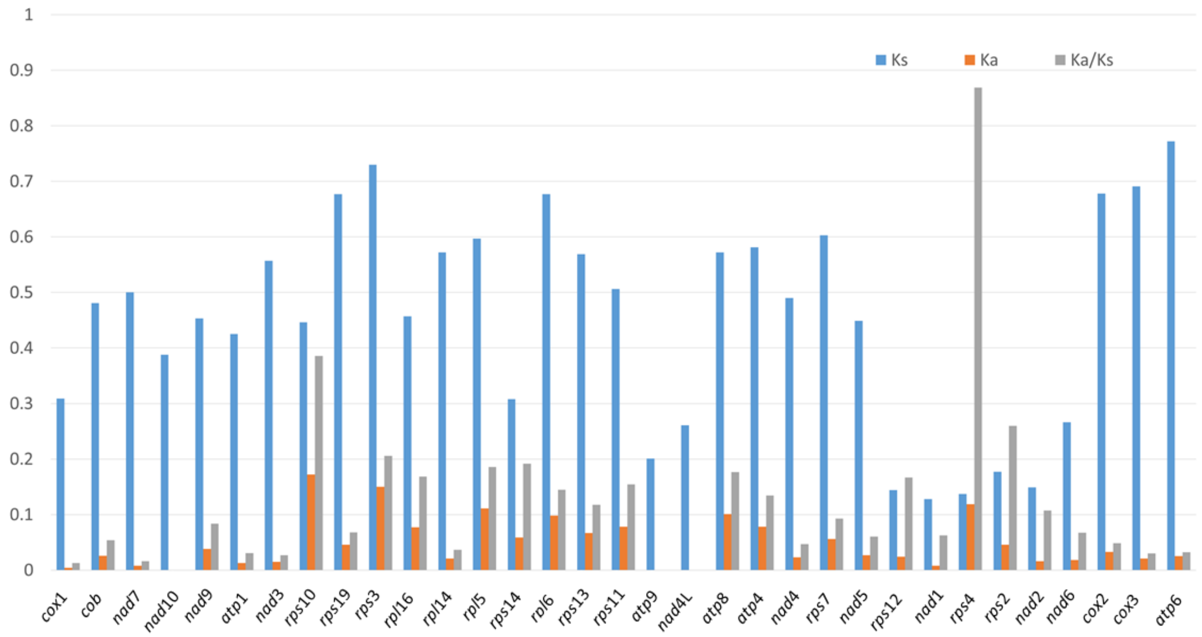


Figure 4. Comparison of NIES254 and SCCAP K-0007 shows that substitution in *P. parkeae* mitochondrial genomes is unevenly distributed. The x-axis shows protein-coding genes present in the mitochondrial genomes, in genomic order. The y-axis is the value for the number of substitution (per 100 bp) and the value for Ka/Ks ratio. The bar with blue color indicates synonymous substitutions (Ks), the bar with orange color indicates non-synonymous substitutions (Ka), and the bar with grey color indicates the ratio of non-synonymous substitutions to synonymous substitutions (Ka/Ks).

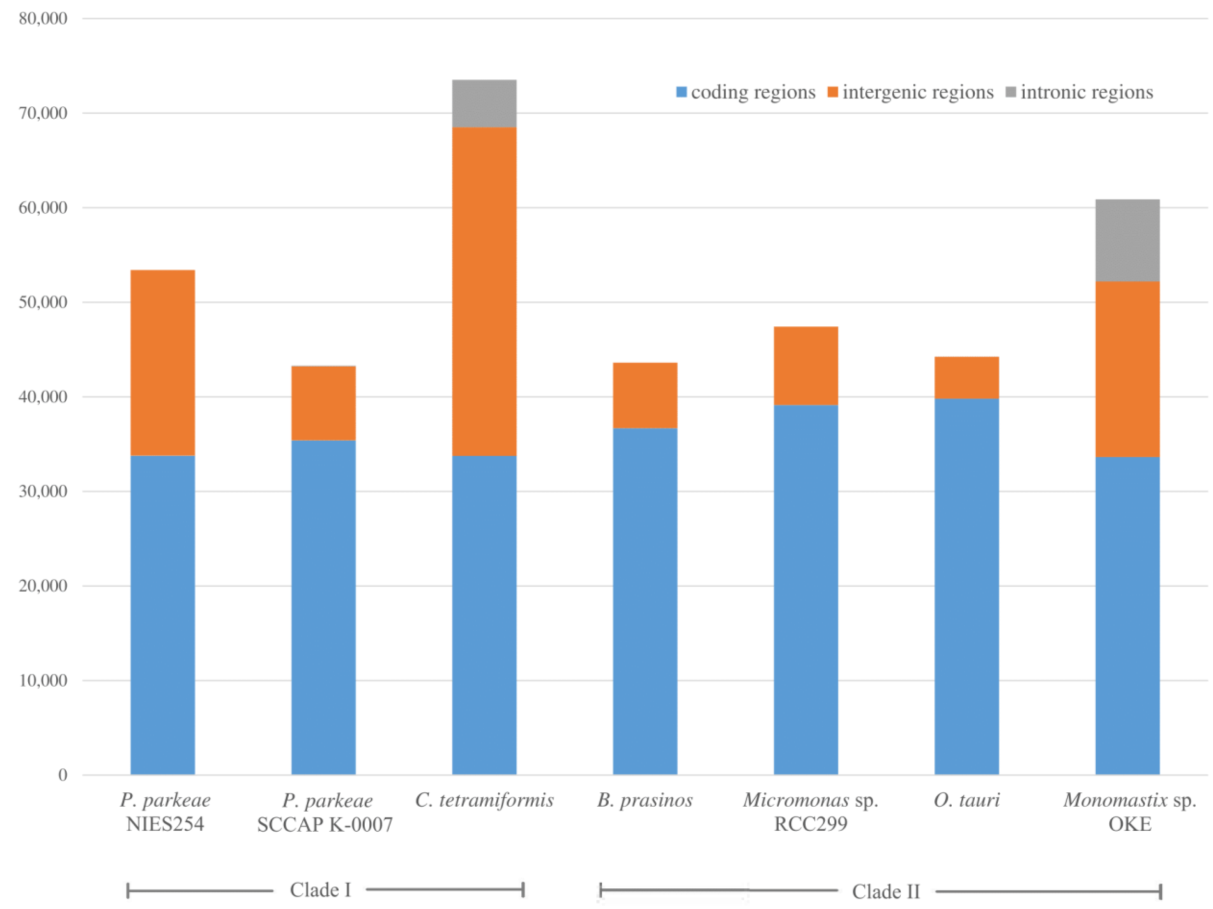


Figure 5. Distribution of coding regions (blue), intergenic regions (orange), and intronic regions (grey) within the mitochondrial genomes of prasinophytes from clades I and II. The x-axis represents length in base pairs. The single intron from *P. parkeae* SCCAP K-0007 is only 71 bp long, too small to be resolved in this plot.

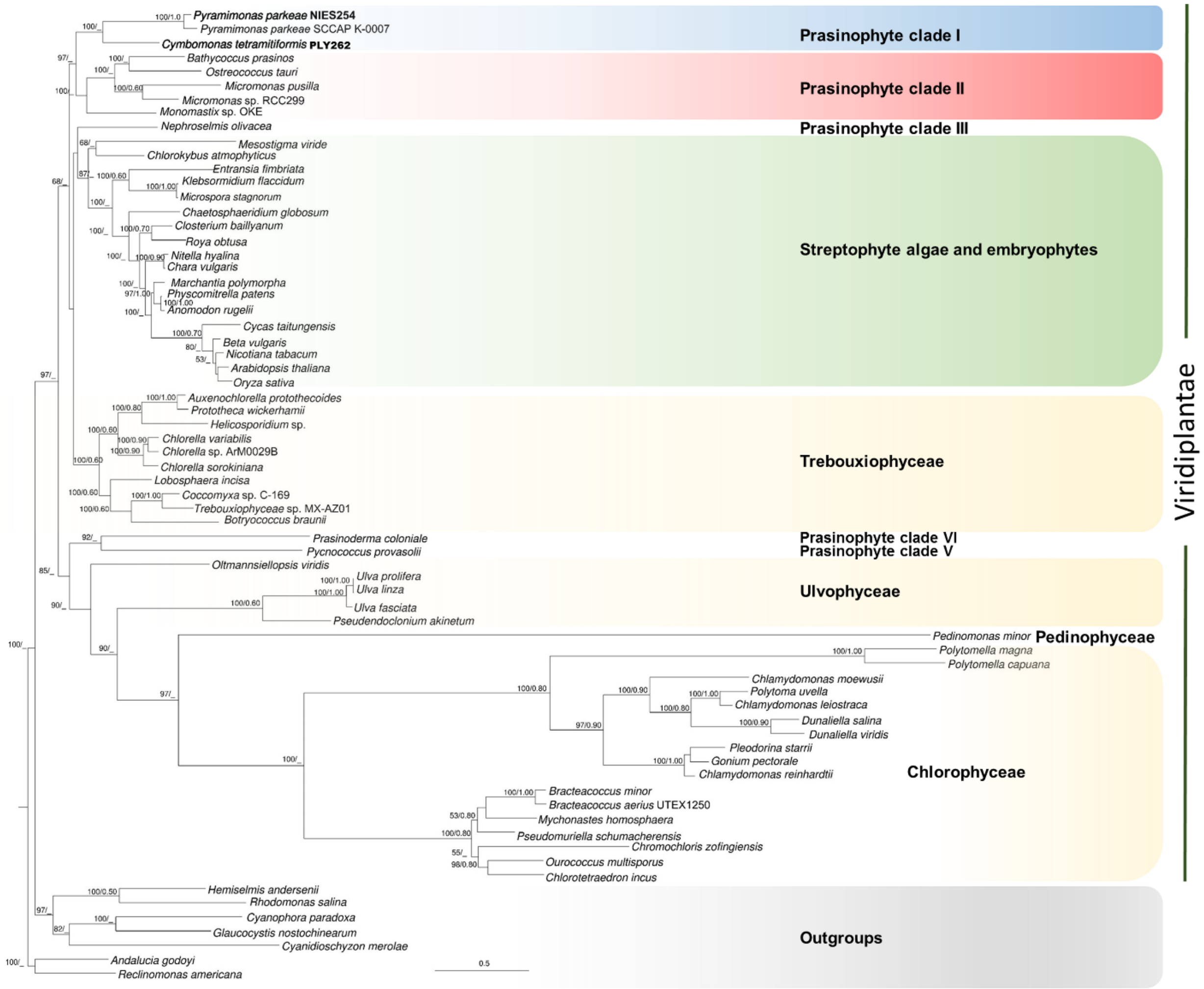


Figure 6. Maximum-Likelihood (ML) tree inferred from concatenated sequences of 32 mitochondrial-encoded proteins from 69 organisms (7,239 amino acid positions) as indicated in Table 1. ML bootstrap values (50 or higher) and posterior probability value (0.5 or higher) from Bayesian inference are shown at the respective nodes. The scale bar represents the estimated number of amino acid substitution per site. The jakobids, cryptophytes, rhodophyte, and glaucophytes were used as outgroup

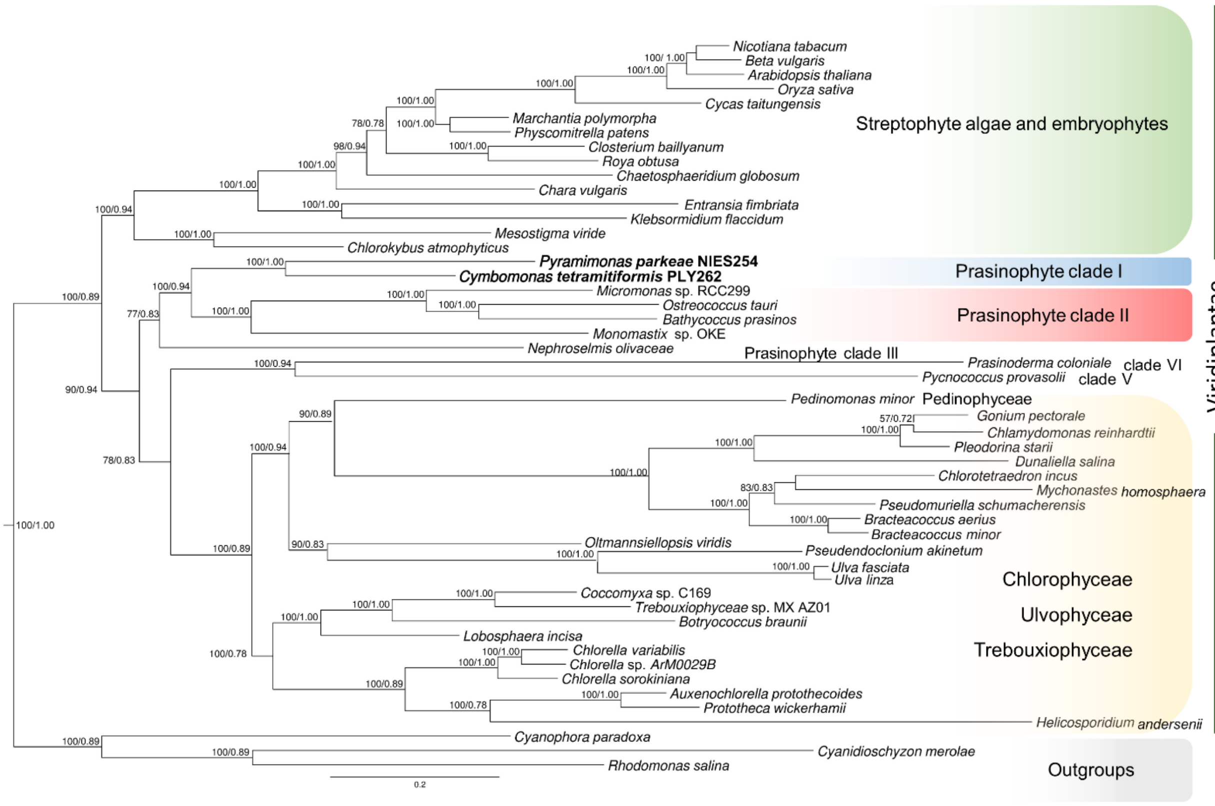


Figure 7. Maximum-Likelihood tree inferred from concatenated sequences of 32 mitochondrial and 74 chloroplast-encoded proteins (25,059 amino acid positions total) of 51 organisms as indicated in Table 1. ML bootstrap values (50 or higher) and Bayesian posterior probability values are shown at the respective nodes. The scale bar represents the estimated number of amino acid substitution per site. The cryptophyte, rhodophyte, and glaucophyte were used as outgroups.

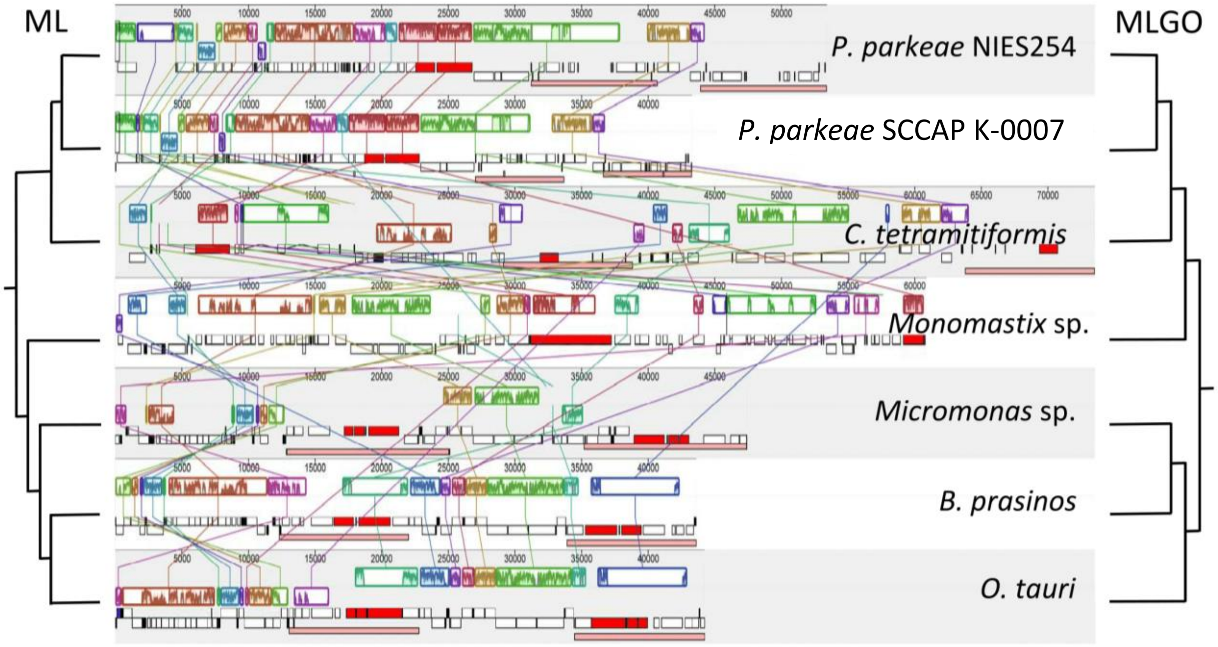


Figure 8. Mauve alignment of mitochondrial genomes of prasinophyte clades I and II including *Pyramimonas parkeae* NIES254, *P. parkeae* SCCAP K-007, *Cymbomonas tetramitiformis*, *Bathycoccus prasinos*, *Micromonas* sp. RCC299, *Ostreococus tauri*, and *Monomastix* sp. OKE. The algal mitochondrial genomes with the quadripartite structure were arranged in the same direction (LSC-IRB-SSC-IRA) Regions of homology between the prasinophyte species are represent by the same color of syntenic blocks and are connected by vertical bars. The histogram inside each block represents pairwise nucleotide sequence identity. At right (not drawn to scale) is a maximum-likelihood tree inferred from concatenated sequences of 32 mitochondrial protein-encoded proteins and 74 plastid protein-encoded proteins of the prasinophyte species. At left (not drawn to scale) is a maximum-likelihood tree inferred from the mitochondrial gene order of the prasinophyte species.

CHAPTER 4: WHOLE GENOME SEQUENCING OF *PYRAMIMONAS PARKEAE* (PRASINOPHYCEAE) REVEALS GENES ENCODING CARBOHYDRATE ACTIVE ENZYMES[1]

Research article

Anchittha Satjarak[2]

Department of Botany, University of Wisconsin-Madison, 430 Lincoln drive, Madison, Wisconsin, USA

[2]corresponding author: e-mail: satjarak@wisc.edu, phone: +16082620657, fax: +16082627509

Linda E. Graham

Department of Botany, University of Wisconsin-Madison, 430 Lincoln drive, Madison, Wisconsin, USA

Key word: *Pyramimonas parkeae,* carbohydrate active enzymes, comparative genomics, comparative transcriptomics, glycosyl transferases, glycoside hydrolases,

Running header: CAZymes of *Pyramimonas parkeae*

Abstract

The wall-less green flagellate *Pyramimonas parkeae* is classified in clade I of the prasinophytes, a paraphyletic assemblage representing the last common ancestor of Viridiplantae, a monophyletic group composed of the green algae and land plants. Consequently, *P. parkeae* and other prasinophytes illuminate early-evolved Viridiplantae traits likely fundamental in the systems biology of green algae and land plants. Cellular structure and organellar genomes of *P. parkeae* are now well understood, and transcriptomic sequence data are also publically available for one strain of this species, but corresponding nuclear genomic sequence data are lacking. For this reason, we obtained shotgun genomic sequence and assembled a draft nuclear genome for *P. parkeae* NIES254 to use along with existing transcriptomic sequence to focus on carbohydrate active enzymes. We found that the *P. parkeae* nuclear genome encodes carbohydrate active protein families similar to those previously observed for other prasinophytes, green algae, and early-diverging embryophytes for which full nuclear genomic sequence is publically available. Sequences homologous to genes related to biosynthesis of starch and cell wall carbohydrates were identified in the *P. parkeae* genome, indicating molecular traits common to Viridiplantae. For example, the *P. parkeae* genome includes sequences clustering with bacterial genes that encode cellulose synthases (Bcs), including regions coding for domains common to bacterial and plant cellulose synthases; these new sequences were incorporated into phylogenies aimed at illuminating the evolutionary history of cellulose production by Viridiplantae. Genomic sequences related to biosynthesis of xyloglucans, pectin, and starch likewise shed light on the origin of key Viridiplantae traits.

Introduction

Viridiplantae (green algae and land plants) are generally characterized by the distinctive combination of starch production in plastids and carbohydrate-rich cell coverings, commonly cellulosic cell walls. Because land plants inherited these economically important features from green algal ancestors, the evolutionary origin of starch and cellulose biosynthetic pathways is of interest, though not yet well understood. Prasinophyte green algae, which diverged prior to the two main Viridiplantae lineages (Chlorophyta and Streptophyta), produce starch in plastids as a main energy storage, but lack cellulosic cell walls. Hence, understanding carbohydrate biosynthetic pathways of prasinophytes may illuminate the evolution of these and other Viridiplantae features.

Starch and the main components of the plant cell wall–cellulose, hemicellulose and matrix carboxylic polysaccharide–are synthesized, modified, and degraded by Carbohydrate-Active enZymes (CAZymes), a group of enzymes having protein domains common to all living organisms. CAZymes also play key roles in a diverse array of additional cellular processes such as signaling, defense, and carbohydrate-related post-translational modifications. CAZymes have been grouped into four classes based on distinctive enzymatic domains: Glycosyl Transferases (GTs), Glycoside Hydrolases (GHs), Polysaccharide Lyases (PLs), Carbohydrate Esterases (CEs), but also include the non-enzymatic carbohydrate binding modules (CBMs). GTs catalyze the formation of glycosyl bonds between a donor sugar substrate and another molecule, typically another sugar. For this reason, GTs are mostly responsible for production of glucans, carbohydrate storage, and signaling processes. GHs hydrolyze the glycosyl bonds between sugars in carbohydrate biopolymers, and thus play an important role in the modification of biopolymers to be introduced into the cell wall, as well as abscission and dehiscence of plant organs. PLs are implicated in non-hydrolytic cleavage of activated glycosidic bonds, such as cleavage of uronic acids from pectins. CEs de-acetylate polysaccharide side-chains and are thought to modify the cross-linking of hemicellulose with lignin. CBMs allow for specific binding to different carbohydrate biopolymers, thereby facilitating precise biopolymer modification before addition to the cell wall (Cantarel et al. 2008).

Many algal CAZyme-encoding gene sequences are known from studies focused on evolution of Viridiplantae cellulose. Biochemical analyses have shown that cellulose is a major component of most chlorophyte and streptophyte cell walls (reviewed by Popper et al. 2011). However, the cellulose synthesizing complexes (CSCs) present in the two main Viridiplantae lineages differ in macromolecular structure; chlorophyte CSCs take rectangular forms related to the linear CSCs of bacteria, while streptophyte CSCs are primarily known to occur as rosettes (reviewed by Tsekos 1999). Chlorophytes possess only gene sequences that encode the bacterial type of CSC, whereas streptophytes have genes encoding both bacterial and rosette CSCs (Kumar and Turner 2015). Phylogenetic analyses have suggested to some investigators that streptophyte CSC-encoding genes might have had a common origin before divergence into separate clades (Mikkelsen et al. 2014). Chlorophytes and streptophytes also differ in other cell-wall carbohydrate polymers, such as hemicelluloses and matrix carboxylic polysaccharides. Although xyloglucans and mannans are present in the cell walls of representatives of both green lineages (though in different ratios), glucuronans and ulvans seem to only occur in chlorophytes, and mixed-linkage glucans and pectins are associated only with streptophytes (reviewed by Popper et al. 2011).

Among Viridiplantae, the main storage polysaccharide is starch that is synthesized in plastids, a unique and defining feature of this clade. Starch biosynthesis is thought to have arisen in Viridiplantae after the endosymbiotic acquisition of the cyanobacterial endosymbiont ancestral to green plastids (Ball et al. 2011). One hypothesis posits the origin of Viridiplantae starch biosynthesis from host glycogen metabolism (Cenci et al. 2014). The early-diverging position of prasinophytes within Viridiplantae suggests that better understanding the genetic foundation of starch metabolism in these green algae might illuminate the process by which Viridiplantae first acquired starch. However, little is known about prasinophyte CAZymes related to starch metabolism.

To better understand early-evolved genes involved in Viridiplantae cell wall and starch metabolism, we investigated selected CAZyme gene families derived from whole genome sequence obtained for the wall-less prasinophyte *Pyramimonas parkeae* NIES254, whose chloroplast and mitochondrial genomes we have previously described (Satjarak and Graham 2017, Satjarak et al. 2017). *Pyramimonas,* classified in prasinophyte clade I thought to have diverged relatively early within Viridiplantae (Leliaert et al. 2016), has also been the source of substantial transcriptomic data (Keeling et al. 2014) that have not previously been used to survey CAZymes. We hypothesized that combining these transcriptomic and genomic data might help to illuminate early evolutionary diversification of CAZymes and the origin of polysaccharide metabolism in Viridiplantae.

Materials and methods

Culture source and DNA extraction

A unialgal but non-axenic culture of *Pyramimonas parkeae* Norris and Pearson (NIES254) was acquired from the National Institute of Environmental Studies, Japan. The culture was propagated in Alga-Gro® seawater medium (Carolina Biological Supply Company, Burlington, NC, USA), and maintained in a walk-in growth room with 16:8 daily light/dark cycle at 20˚C. Cells were harvested during the exponential phase of growth. Total DNA was prepared by using the FastDNA® SPIN Kit for Soil (MP Biomedicals, Solon, OH) and sequenced by Illumina Miseq and Hiseq technologies at the University of Wisconsin-Madison Biotechnology Center.

Data pre-processing and genome construction

Raw paired-end Illumina Miseq data consisted of 13,232,998 reads with average read length of 251 bp; Hiseq data consisted of 267,637,390 reads with average read length of 101 bp. The data were trimmed by Trimmomatic v 0.33 (Bolger et al. 2014) to obtain a quality score of at least 28 on the phred 64 scale. The *Pyramimonas parkeae* genome was then assembled into contigs using Newbler *de novo* sequence assembly (v 2.9) available at http://454.com/products/analysis-software/. Among the several assembler systems investigated, Newbler was judged to work best with our sequence data.

CAZymes prediction and annotation

Genomic DNA

Assembled contigs included sequence from nuclear, chloroplast, and mitochondrial genomes. Therefore, we used BLAST methods to select only contigs relevant to putative CAZymes. The BLAST CAZyme query, a compilation of eukaryotic protein sequences from the

Carbohydrate Active enZYme database (CAZY) available at http://www.cazy.org/ (Lombard et al. 2014), was employed to search against our assembled contigs using a local TBLASTN, with a threshold of expected value of least 1E-10 (Gertz et al. 2006). Queries for eukaryotic CAZymes included 19,112 glycoside hydrolases (GHs), 38,042 glycosyltransferases (GTs), 3,131 carbohydrate esterases (CEs), 4,083 polysaccharide lyases (PLs), and 3,520 carbohydrate binding modules (CBMs).

We then used *P. parkeae* contigs containing putative CAZyme-encoding sequences to search against the NCBI non-redundant protein database, using NCBI-BLASTX with a threshold of expected value equal to 1E-10. Contigs of non-Viridiplantae origin (e.g. those of contaminating bacteria) were removed from the data set. Finally, we annotated the remaining contigs using the MAKER annotation pipeline (Cantarel et al. 2008) to obtain gene models, and used dbCAN HMMs v. 5.0 (Yin et al. 2012) to infer CAZyme families.

The MAKER annotation pipeline employs a combination of *ab initio* and evidence-based gene prediction and annotation. Because our *P. parkeae* assembled genome was fragmented and the results from AUGUSTUS training showed differences between the results from AUGUSTUS *ab initio* prediction and those based on expressed sequence tag (EST) alignment, we only used the evidence-based method implemented in the MAKER annotation pipeline to predict gene models directly from EST alignments built from ESTs obtained for *P. parkeae* CCMP726 as part of the Marine Microbial Eukaryote Transcriptome Sequencing Project (Keeling et al. 2014). The exons of contigs that failed to be predicted and annotated using the MAKER annotation pipeline were annotated using BLASTX against the NCBI non-redundant protein database (accessed in August 2016) with a threshold of expected value equal to 1E-10.

To predict the CAZyme families of putative protein-coding genes present in the assembled contigs, we used a web server and database for Carbohydrate-active enzyme Annotation (dbCAN) available at http://csbl.bmb.uga.edu/dbCAN/ (Yin et al. 2014). Protein sequences derived from the annotated contigs were used to search against dbCAN HMMs using HMMER3 implemented in dbCAN HMMs v. 5.0 (Yin et al. 2014) to obtain CAZyme families. Then, we used Blast2GO (Conesa et al. 2005) and TBLASTN search to obtain the sequence description with a threshold of expected value equal to 1E-10. To obtain *P. parkeae* NIES254 assembled contigs that were potentially CAZymes involved in the metabolism of selected polysaccharides but not identified due to fragmentation, we used *P. parkeae* CCMP726 CAZyme transcripts (see below) to search against the NIES254 assembled genome using local BASTN with a threshold of expected value equal to 1E-10.

Transcriptomic DNA

Transcriptomic contigs and derived proteins for *P. parkeae* CCMP726 were obtained from the Marine Microbial Eukaryote Transcriptome Sequencing Project under accession numbers MMETSP0058 and MMETSP0059 available at http://data.imicrobe.us (Keeling et al. 2014). The data include a total of 23,233 *P. parkeae* CCMP726 transcripts and their derived protein sequences. Similarly, we used the compilation of eukaryotic protein sequences from the CAZY database to search against the *P. parkeae* CCMP726 transcriptomic data using TBLASTN in order to obtain putative CAZyme sequences. Then we removed non-Viridiplantae transcripts and annotated CAZyme families with derived protein sequences using dbCAN

HMMs v. 5.0 (Yin et al. 2014). Finally, we used Blast2GO (Conesa et al. 2005) and TBLASTN search to obtain sequence descriptions having threshold of expected value equal to 1E-10.

Comparative analyses of P. parkeae *CAZyme families and those of other chlorophyte and streptophyte species*

We compared CAZyme families and numbers of family members identified in *P. parkeae* with those of other chlorophyte and streptophyte species. For this comparison, we used all prasinophytes for which complete genomes are available, including *Bathycoccus prasinos* (Moreau et al. 2012), *Micromonas* sp. RCC299, *Micromonas pusilla* (Worden et al. 2009), *Ostreococcus tauri* (Derelle et al. 2006), and *Ostreococcus lucimarinus* (Palenik et al. 2007); all chlorophytes for which complete genomes are available, including *Chlamydomonas reinhardtii* (Merchant et al. 2007), *Chlorella variabilis* (Blanc 2010), *Coccomyxa subellipsoidea* (Blanc et al. 2012), *Monoraphidium neglectum* (Bogen et al. 2013) and *Volvox carteri* (Prochnik et al. 2010); a complete genome of the streptophyte alga *Klebsormidium flaccidum* (Hori et al. 2014); and complete genomes for the early-diverging embryophytes *Physcomitrella patens* (Rensing et al. 2007) and *Selaginella moellendorffii* (Banks et al. 2011).

The CAZymes of some of the selected species had previously been annotated and were available at dbCAN database (http://csbl.bmb.uga.edu/dbCAN/). These included CAZymes identified from *Micromonas* sp. RCC299, *Micromonas pusilla, Ostreococcus tauri*, *Ostreococcus lucimarinus*, *Chlamydomonas reinhardtii*, *Chlorella variabilis*, *Volvox carteri*, *Physcomitrella patens,* and *Selaginella moellendorffii*. To identify CAZymes for species incompletely annotated (*Bathycoccus prasinos, Coccomyxa subellipsoidea*, *Monoraphidium* *neglectum, and Klebsormidium flaccidum*), we obtained protein sequences from GenBank and annotated CAZyme families using dbCAN HMMs v. 5.0 (Yin et al. 2014). Then, we compared the presence of CAZyme families and numbers of family members across species. Although this data set is not ideal, because the data were obtained using different approaches, we considered that the results of this analysis would provide a broad perspective regarding CAZymes in early-diverging Viridiplantae.

Phylogenetic analyses

We further investigated a *P. parkeae* CAZyme classified as GT2 in order to evaluate relationship with cellulose synthases and cellulose synthase-like proteins. For this protein sequence (CAMPEP_0191478436), homologous sequence was identified in NIES254 genomic data, though the latter was more fragmented. Therefore, for the best alignment results, we only used the CCMP726 GT2 inferred protein sequences in subsequent analysis. We aligned the *P. parkeae* GT2 protein sequences with the predicted function “cellulose synthase” available in NCBI Reference Sequence Database NCBI (Table 1) using MAFFT v.7.222 (Katoh and Standley 2013) and the Auto algorithm and BLOSUM62 scoring matrix. The amino acid substitution model was computed using ProtTest v. 3.4.2 (Darriba et al. 2011). Maximum-likelihood (ML) analysis was performed using RAxML v. 8.2.8 (Stamatakis 2014) on the CIPRES XSEDE Portal (Miller et al. 2010) using an LG+I+G+F substitution model, rapid bootstrapping method with 1,000 replications for bootstrap analyses.

Results

CAZyme family identification from new genomic assemblies and database transcriptomic sequence data

105,787 contigs having a length of at least 500 bp were assembled from a total of 106,589,986 Illumina Miseq and Hiseq reads (N50 statistic 3,945 bp) for *P. parkeae* NIES254. From these new genomic sequence data, 118 putative CAZymes-containing contigs were classified: 12 GTs, 84 GHs, 1 CEs, and 21 CMBs (Table 2). In addition, from the *P. parkeae* CCMP726 transcriptomic data, 519 putative CAZymes were classified: 329 GTs, 85 GHs, 33 CEs, 1 PLs, and 54 CMBs (Table 3).

Comparative analyses of CAZyme families of P. parkeae *and those of other chlorophyte and streptophyte species*

We compared the presence of selected CAZyme families and the number of the predicted proteins of these families present in *P. parkeae* to homologs present in other chlorophyte algae and early-diverging streptophyte species. The results showed that most of these species shared common CAZyme families. The numbers of predicted proteins for those common families were similar in prasinophytes, chlorophytes, and streptophyte algae, whereas the number of such sequences were increased dramatically in *Physcomitrella patens* and *Selaginella moellendorffii* (Table 4-6).

P. parkeae *CAZymes involved in metabolism of cell wall components*

In this study, 9 *P. parkeae* CCMP726 transcripts and homologous regions in NIES254 assembled contigs were classified with the GT2 family (Table 7). Of those, one protein sequence, CAMPEP_0191478436, contained QXXRW domains, Ds residues, and the DXD motif, known to be catalytic sites of bacterial and embryophyte cellulose synthases. However, by contrast to other cellulose synthases, this sequence contained only one transmembrane region (Figure 1). By using TBLASTN, though fragmented, we found the genomic data that were homologous to this protein sequence. The putative exons of these genomic contigs covered at least 73% of the protein sequence inferred from the transcriptomic data.

Phylogenetic analyses using cellulose synthase protein sequences from green algae, moss, clubmoss, and cyanobacteria (Table 1) resolved 2 monophyletic clades, both with bootstrap values of 98. One clade consisted only of protein sequences from embryophytes (moss and clubmoss) while the other clade consisted of protein sequences from all selected species used in the analysis (Table 1, Figure 2). Within this second clade, cellulose synthase protein sequences from the same species did not resolve a monophyletic clade.

With respect to xyloglucans, in this study, we did not find evidence in *P. parkeae* genomic or transcriptomic sequence data for α-fucosyl transferase, α-fucosidase, or xyloglucan transglycosylase/hydrolase. However, in the *P. parkeae* transcriptome, we found 2 xylosyltransferases, 2 β-galactosyltransferases, 3 β-glucosidases, 3 α-xylosidases, and 2 β-galactosidases, and we found homologous regions in *P. parkeae* genomic contigs (Table 8). Regarding pectin-related genes, we found genomic and transcriptomic sequences of putative galacturonosyltransferases (GAUT1) associated with the synthesis of homogalacturonan, a pectic polysaccharide. Also, we found sequences related to pectin-degrading enzymes (GH28) and pectin acetyl esterases (CE13) predicted to be responsible for de-acetylation of pectin in embryophytes (Table 9). However, we did not find pectin methyl esterases (CE8) that function in homogalacturonan de-esterification.

Other CAZymes

In this study, we identified sequences putatively encoding enzymes involved in starch metabolism: ADP-glucose pyrophosphorylase, starch phosphorylase, starch synthases (soluble starch synthase and granule-bound starch synthase), alpha-amylase, isoamylase, 1,4-alpha-glucan-branching enzyme, pullulanase, and beta-amylase (Table 10). Also, we found sequence evidence for putative *P. parkeae* trehalose phosphate synthase and 6 sequences of trehalase (Table 11). Additionally, we found a sequence in the *P. parkeae* transcriptomic data, annotated as GT51, for which a homolog was also present in the genomic data, though fragmented (Tables 3, and 12).

Discussion

The availability of the new assembled draft nuclear genome and archived transcriptome for *P. parkeae* provides the opportunity to explore the CAZymes present in this early diverging green alga. Also, the new genomic data can be used with transcriptomic sequence to compare CAZymes present in early diverging green algae with those of selective species from chlorophytes and streptophytes. Such comparisons have the potential to indicate expansion or acquisition of CAZyme families associated with multicellularity or divergence of the two green lineages – chlorophytes and streptophytes.

Comparative analyses of CAZymes in selected species

The sequence evidence presented here for presence of CAZymes in *P. parkeae* was congruent with the CAZyme encoding genes in other Viridiplantae. The gene families and numbers of sequences per family observed for *P. parkeae* were similar to those of other prasinophytes and other unicellular green algae (Tables 4-6). Because prasinophytes diverged early within Viridiplantae, CAZyme families present in *P. parkeae* and other prasinophytes might represent those present in the last common Viridiplantae ancestor. Consequently, comparisons between earlier and later-diverging Viridiplantae should indicate patterns in CAZymes evolutionary diversification.

Our comparative analysis suggests that the numbers of CAZyme sequences in the genomes of the colonial *V. carteri* and filamentous *K. flaccidum* are similar to those of unicellular green algae. This is congruent with results reported by Prochnik et al. (2016) indicating that increased morphological complexity in *V. carteri* is associated with modification of lineage-specific proteins rather than the acquisition of new proteins. Although CAZymes families were not the focus of a genomic study of *K. flaccidum*, the overall number of gene families was reported to be comparable to those of other green algae (Hori et al. 2014).

By contrast, our comparisons revealed that the number of CAZyme family members was considerably higher in the early-diverging streptophytes examined–*Selaginella moellendorffii* and *Physcomitrella patens*–than in the green algal taxa for which whole genome data are publically available (Tables 4-6). This difference is consistent with the conclusion of Banks et al. (2011) that nuclear gene content approximately doubled during the transition from green algae to multicellular land plants. Increases in gene family size are generally attributed to local or whole genome duplication. For example, a whole genome duplication is proposed to explain why the *Populus trichocarpa* genome has 1.6 times more CAZymes genes than does *Arabidopsis thaliana* (Henrissa et al. 2001, Geisler-Lee et al. 2006). Although it is theoretically possible for CAZymes gene numbers known for *S. moellendorffii* and *P. patens* to likewise have resulted from ancient whole genome duplications, *P. patens* genome duplication is thought to occurred too recently (30 -60 million years ago) (Rensing et al. 2007) to explain the observed differences with algae, and no evidence has been reported for genome duplication in *S. moellendorffii*.

CAZymes involved in biosynthesis or degradation of cell wall components

Plant genomes appear to have relatively higher ratios of CAZymes genes to total gene number than do other eukaryotes or bacteria. This higher ratio has been attributed to increased number of GTs and GHs related to biosynthesis and degradation of complex polysaccharides in plant cell walls (Coutinho et al. 2003). If so, a preadaptive increase in GTs and GHs might be expected to have occurred during evolution of the modern green algae, whose ancestors were closely related to those of plants. Our analysis of *P. parkeae* CAZyme genes supports this concept.

Although *Pyramimonas parkeae* cells are enclosed by scale layers rather than a typical cell wall (Pearson and Norris 1975), we found that most of the annotated *P. parkeae* CAZymes are GTs and GHs, and include the CAZyme families GT2, GT8, GT34, GT37, GT43, GT47, and GT77 (Table 4). These CAZyme families include genes known to be involved in the biosynthesis or degradation of cell wall components in embryophytes. Our genomic evidence for presence of these CAZymes in wall-less early-diverging green algae suggests that many of the proteins essential for biosynthesis and degradation of cell wall components existed prior to the acquisition of the cell wall in Viridiplantae. These proteins include homologs of embryophyte proteins known to be involved in biosynthesis or degradation of cellulose, xyloglucan, and pectin.

Cellulose- and cellulase-encoding genes

Cellulose, the most abundant component of non-lignified plant cell walls, is composed of microfibrils constructed from linear molecules of β-1,4-linked glucan held together by intra- and intermolecular hydrogen bonds and van der Waals forces. As in the case of land plants, algal microfibrillar cellulose is synthesized at the plasma membrane by protoplasmic-face complexes known as cellulose synthesizing complexes (CSC) or terminal complexes (TC) (Brown and Montezinos 1976, Tsekos 1999).

Plasma membrane-bound cellulose synthesizing complexes have been demonstrated by microscopic analyses to occur in diverse algal lineages, including Cyanobacteria (Zaar 1979), Glaucophyta (Willison and Brown 1978), Rhodophyta (Tsekos and Reiss 1992, Tsekos et al.1999), Chlorophyta (Brown and Montezinos 1976, Itoh 1990), Phaeophyta (Katsaros et al. 1996) and other photosynthetic stramenopiles (Okuda et al. 2004), at least some alveolate dinoflagellates (Okuda and Sekida 2007) and streptophyte algae (Okuda and Brown 1992, Hotchkiss and Brown 1989), consistent with the production of celluloses by these groups.

However, the shapes of autotroph CSCs differ. Known chlorophyte CSCs occur as linear complexes consisting of rows of subunits; by contrast, streptophyte algal CSCs occur as rosettes that generally consist of six subunits that each have six-fold rotational symmetry, as in the case of embryophytes (Tsekos et al. 1999, Kumar and Turner 2015). CSC shape differences correspond with differing architectures of cellulose microfibrils (Giddings et al. 1980, Herth 1983, Sugiyama et al. 1994, Kim et al. 1996, Tsekos 1999, Kumar and Turner 2015).

Most analyses of the genetic basis for CSC structure and function have focused on streptophyte *CesA*–a nuclear gene that is known to encode CesA protein (GT2) that forms the core of CSCs (Kumar and Turner 2015). CesA-CesA interaction has been proposed to underlie rosette-shaped CSCs (Arioli et al. 1998, Peng et al. 2001, Kurek et al. 2002, Gardiner et. al. 2003). If correct, the absence of rosette-shape CSCs in non-streptophytes might result from absence of one or more domains distinctive for streptophyte CesA: a Zinc-binding domain, a plant-conserved region, and a hypervariable region (Arioli et al. 1998, Peng et al. 2001, Kurek et al. 2002, Gardiner et. al. 2003). Nobles et al. (2001) hypothesized that Viridiplantae *CesA* gene was acquired by horizontal transfer from the cyanobacterial ancestor of plastids. Seedless streptophyte genomes include genes encoding both bacterial (known as Bcs) and rosette-forming CesAs (Harhold et al. 2012, Ulvskov et al. 2013, Mikkelsen et al. 2014), suggesting that earlier-diverging green lineages might also encode both types of cellulose synthases.

In bacteria, genes that encode cellulose synthase occur as an operon (*Bcs* operon). Within the operon, *BcsA* and *BcsB* are considered crucial for the synthesis of cellulose because mutation of either *BcsA* or *BcsB* is reported to result in the absence of cellulose production (Hu et al. 2015). *BcsA* encodes the protein responsible for production of the linear cellulose glucan chain, while *BcsB* encodes an activator of cellulose synthesis (Romling et al. 2002). These two genes occur together in the operon and are sometimes fused (Kimuar et al. 2001). Our finding of genomic and transcriptomic evidence for the presence of a Bcs-like protein in *P. parkeae* is consistent with GenBank evidence that Bcs-encoding sequences generally occur in the genomes of prasinophytes, including two *Ostreococus* species, *Micromonas commoda*, and *Bathycoccus prasinos*. However, the *P. parkeae* Bcs we found, though displaying QXXRW, Ds, and DXD motifs distinctive for CesA, is so far unique in having only a single transmembrane region. By contrast, *O. tauri CesA*-like sequences contained four transmembrane regions, though only one of those contained QXXRW residues. Similarly, all three *CesA*-like sequences from *O. lucimarinus* had four transmembrane regions, but only one contained QXXRW residues. Both such sequences from *M. commoda* and both from *B. prasinos* likewise contained four transmembrane regions but all lacked the QXXRW residues; they had been annotated as Bcs because Ds, DXD, and transmembrane regions were present. We employed all of these putative prasinophyte sequences in our phylogenetic analysis.

Our ML analysis indicated that the *P. parkeae* CesA protein sequence grouped within a strongly-supported monophyletic clade that included other *P. parkeae* GT2 proteins and bacterial-type CesA sequences. However, relationships of protein sequences within this clade were not fully resolved. The *P. parkeae* protein forms a poorly-supported clade with a protein sequence (XP_013894101) from a chlorophyte alga *Monoraphidium neglectum* that has been annotated as cellulose synthase (UDP-forming) (Bogen et al. 2013), and thus may have descended from bacterial *BcsB*. More biochemical and molecular work will be needed in order to investigate possible homologies between the *P. parkeae* CesA-like protein and the bacterial Bcs operon, and to understand the function of CesA-like proteins in prasinophytes.

Cellulose is hydrolyzed by cellulase, an enzyme classified into GH family 9. A study of *Brassica napus* showed that this enzyme could hydrolyze crystalline cellulose, xyloglycan, xylan, (1→3)(1→4)-β-D-glucan and other polysaccharides and oligosaccharides (Mølhøj et al.

2001). Although we did not identify a sequence that classified to GH9, our BLAST search of genomic sequences indicated the presence of partial cellulase-encoding sequences (contig16083 length=4296 and contig66435 length=992). Because non-eukaryotic sequences were informatically removed from genomic data prior to the search process, we hypothesize that these sequences are located within the *P. parkeae* genome. If so, more work would be needed to determine their functions.

Xyloglucans

Xyloglucans, the most abundant hemicelluloses in the plant primary cell wall, cross-link cellulose microfibrils, giving strength and preventing the aggregation of cellulose microfibrils and other wall matrix polymers (Thomson 2005). Xyloglucans consist of a backbone β-1,4 glucan chain that is decorated with xylosyl units, which carry additional glucosyl residues, such as D-galactose, D-xylose, L-arabinopyranose, L-arabinofuranose, D-galacturonic acid, L-fucose, and L-galactose. These glucosyl residues vary among plant groups, tissue types, cell types, developmental stages, and even positions within the cell wall (Scheller and Ulvskov 2010, Schultink et al. 2014). Xyloglucan synthesis occurs in Golgi bodies, whereas the glucan backbone is hypothesized to arise by the action of cellulose synthase-like proteins. Xylosyltransferases (GT34) then transfer xylose to the glucan backbone. This xylosyl group is often additionally glycosylated with fucosyl and galactosyl residues. The enzyme hypothesized to be responsible for fucosylation is α-fucosyl transferase and the enzyme responsible for galactosylation is MUR3 (GH47) that exhibits β-galactosyltransferase activity in xyloglucan synthesis. A study in *Arabidopsis* showed that mutations of the *mur3* gene reduced the level of fucose and galactose in the cell wall (Madson et al. 2003).

We found three *P. parkeae* sequences that classified as xylosyltransferases. However, these proteins and their homologous sequences in prasinophytes do not classify as GT34 as in embryophytes, but instead with GT47. Our phylogenetic analysis suggested that these putative xylosyltransferases are closely related to embryophyte exostosin proteins (Supplementary Figure 1). More work is needed to determine if these putative prasinophyte xylosyltransferases have functions similar to those of embryophyte exostosins.

We did not find α-fucosyl transferase in the *P. parkeae* draft genome or transcriptome. However, we found *P. parkeae* transcriptomic sequences, together with corresponding partial genomic sequences, that were predicted as β-galactosyltransferases (Table 8). The protein sequences inferred from their transcripts belong to GT47, as do MUR3 proteins. Also, our phylogenetic analyses using these inferred protein sequences indicated that these *P. parkeae* proteins are sister to the clade consisting of embryophyte β-galactosyltransferases and MUR3 proteins (Supplementary Figure 1). Though more work is needed, the presence of these proteins suggests that prasinophyte β-galactosyltransferases may represent a preadaptation key to the origin of xyloglucans.

The degradation of xyloglucan is the reverse of the synthesizing steps, and includes the removal of glycosyl groups from the xyloglucan backbone. The enzymes responsible include α-fucosidase, xyloglucan transglycosylase/hydrolase, β-glucosidase, α-xylosidase, and β-galactosidase (reviewed by Del Bem and Vincenz 2010). In this study, we did not find α-fucosidase or xyloglucan transglycosylase/hydrolase. However, we found three transcriptomic sequences representing putative α-xylosidases, two transcriptomic sequences indicating putative β-galactosidase and three indicating β-glucosidases, along with homologous genomic sequences (Table 8).

α-xylosidase is classified as GH31. In embryophytes, this protein functions in cleaving the α-xylosyl residue from the glucose residue on xyloglucan-oligosaccharide (Sampedro et al. 2001). In this study, we identified *P. parkeae* protein sequences inferred from transcriptomic nucleotide data belonging to GH31, along with their partial genomic homologs. Our BLAST results (Table 8) suggest that these *P. parkeae* sequences are similar to α-glucosidase and α-xylosidase present in other prasinophytes and in streptophytes. The presence of these sequences is congruent with the hypothesis that this gene was a feature of the common ancestor of chlorophytes and streptophytes (Del Bem and Vincentz 2010).

β-galactosidase is present across a wide range of taxonomic groups. In embryophytes, these enzymes are classified as GH2 (dbCAN database, accessed in November 2016) and is involved in the hydrolysis of cell-wall galactose-containing polymers. This protein type was not identified from completely sequenced genomes of the chlorophytes *Chlamydomonas reinhardtii* and *Volvox carteri* (Del Bem and Vincentz 2010). However, our BLAST results (Table 8) suggest evidence for β-galactosidase in the form of genomic and transcriptomic sequences of *P. parkeae*, indicating ancient origin in the green lineage. Another gene involved in the degradation of xyloglucan is β-glucosidase. Although we found *P. parkeae* genomic and corresponding transcriptomic sequences indicating β-glucosidase (Table 8), blast results suggested that these sequences likely represented chloroplastic β-glucosidases.

Pectin-related gene sequences

Pectin, synthesized in the Golgi apparatus, provides strength flexibility to the cell wall. This carbohydrate enhances the strength of plant cell walls by cross-linking to cellulose microfibrils, and pectins having specific side chains that foster localization to specific regions of the cell wall can soften during directional cell wall expansion (Pelloux et al. 2007). Pectic polysaccharides identified in embryophytes include homogalacturonan, xylogalacturonan, apiogalacturonan, rhamnogalacturonan I, and rhamnogalacturonan II. Homogalacturonan, consisting of unbranched homopolymer chains of α-1,4-linked D-GalUA (galacturonic acid), makes up about 65% of embryophyte pectin. The enzyme responsible for synthesis of homogalacturonan is galacturonosyltransferase (GAUT1) (Sterling et al. 2006). This GT family 8 enzyme is hypothesized to have originated prior to the divergence of bryophytes, as homologs have been identified in *Selaginela moellendorffii* and *Physcomitrella patens* (Moller et al. 2007). Likewise, we found evidence for galacturonosyltransferase homologs in the *P. parkeae* genome and transcriptome (Table 9), suggesting that this protein originated early in Viridiplantae history.

The CAZymes responsible for pectin degradation and remodeling are classified as GH28, CE8, and CE13. In *Arabidopsis*, the pectin-degrading enzyme GH28 is responsible for hydrolyzing alpha-1,4 glycosidic bonds between galacturonic acid residues (Rao and Paran 2003), while pectin methyl esterase (CE8) and pectin acetyl esterase (CE13) are responsible for homogalacturonan de-esterification and de-acetylation of pectin (Geisler-Lee et al. 2006). Our results (Table 9) suggest that the *P. parkeae* genome and transcriptome contained sequences indicating pectin degrading enzymes (GH28) and pectin acetyl esterase (CE13), but we did not find evidence for presence of pectin methyl esterase (CE8).

CAZymes involved in biosynthesis or degradation of other polysaccharides

Viridiplantae starch biosynthesis and degradation occur exclusively within the plastid compartment. Starch biosynthesis requires four steps: substrate activation, chain elongation, chain branching, and chain de-branching (Preiss et al. 1991, Zeeman et al. 2010). The production of ADP-Glucose is completed by ADP-Glucose pyrophosphorylase. Then starch synthases (SSs, including soluble starch synthases and granule-bound starch synthases) catalyze the elongation of α-1,4-glucan by transferring the glucosyl moiety from sugar nucleotide to the non-reducing end of the growing polyglucan chain. Starch branching enzymes (SBEs) cleave the linear glucan chains and transfer cleaved portions to glucose residues in acceptor chains via α-1,6 linkage to form branches. Isoamylases (ISAs) facilitate granule crystallization by removing wrongly-positioned branches (Zeeman et al. 2010). These starch-degrading enzymes are of mixed origin. GBSS, SS and ISA are thought to have derived from the cyanobacterial endosymbiont ancestral to green plastids, while SBEs are considered to be of host origin (Deschamps et al. 2008, Busi et al. 2014).

Starch degradation begins with phosphorylation at the granule surface, which increases accessibility to β-amylase. Phosphoglucan phosphatase releases the phosphate, allowing complete degradation. α-amylase, ISA, and pullulanase hydrolysis products are branched and linear glucans, and β-amylase hydrolyzes linear glucan to maltose. Then maltose and glucose-1-phosphate, a product of starch phosphorylation, are transported to the cytosol (Zeeman et al. 2010). Our results indicated that all genes known to encode enzymes required for plant starch biosynthesis and degradation are present in the *P. parkeae* genome and transcriptome (Table 10), as in the cases of other algal species for which genomes have been completely sequenced – two chlorophytes (*Chlamydomonas reinhardtii* and *Volvox carterii*), four prasinophytes (*Ostreococcus tauri, O. lucimarinus,* and two *Micromonas pusilla* strains) (Ulvskov et al. 2013), and a streptophyte alga *Klebsormidium flaccidum* (Hori et al. 2014). The ubiquitous presence of these genes suggests that starch biosynthesis and degradation processes are conserved across Viridiplantae.

Other CAZymes

Trehalose

Trehalose is a non-reducing disaccharide hypothesized to stabilize proteins and membranes under stress conditions, especially during desiccation (Crowe et al. 1998, Wingler 2012). Trehalose synthesis is catalyzed by trehalose phosphate synthase (GT20) and trehalase (GH37) hydrolyzes this compound. The presence of putative trehalose phosphate synthases and trehalases in *P. parkeae* indicated by genomic and transcriptomic data (Table 11) is congruent with evidence for presence of these genes in other Viridiplantae and cyanobacteria. These genes are thought to be involved in osmoregulation (Wingler 2012).

GT51

A sequence of GT51 (penicillin-binding protein) identified in our study of the *P. parkeae* genome is congruent with the presence of CAZyme family GT51 in the prasinophyte *Micromonas* sp., the bryophyte *Physcomitrella patens*, and the lycophyte *Selaginella* *moellendorffii* (Table 4; Ulvskov et al. 2013). The *P. parkeae* GT51 is hypothesized to originate from a cyanobacterial endosymbiont by gene transfer during the acquisition of the chloroplast. In cyanobacteria, members of this protein family catalyze the final step of synthesis of peptidoglycan. Peptidoglycan is associated with chloroplasts of *P. patens* (Hirano et al. 2016) and proteins involved in peptidoglycan biosynthesis are essential for chloroplast division in some bryophytes (Machida et al. 2006).

Functional homology of P. parkeae *CAZymes*

In this study, which employed HMMs CAZyme classification and sequence similarities indicated by BLAST and phylogenetic approaches, we found *P. parkeae* sequences homologous to those encoding enzymes having known carbohydrate biosynthesis or degradation functions in Viridiplantae. However, this is not sufficient evidence to conclude that functional homology has been conserved.

One possibility is that sequences in the genomic data might be degenerate or otherwise non-functional. Alternatively, homologous sequences in *P. parkeae* might not have the same function as their homologs in other organisms. This issue of functional homology is even more complicated for *P. parkeae* because we lack evidence of the presence of enzyme products such as cell wall components. *Pyramimonas parkeae* is known to produce starch (Pearson and Norris 1975). However, in order to consider *P. parkeae* starch completely homologous to that of other Viridiplantae, the starch should be structurally identical and synthesized via the same pathway. Therefore, more work is needed to determine the biological function of *P. parkeae* sequences that seem related to starch biosynthesis and degradation, and to understand the regulatory network controlling *P. parkeae* starch metabolism.

In this study, due to limitations in the current CAZyme database, we were not able to determine whether any *P. parkeae* CAZyme sequences originated by horizontal gene transfer. More genomic data for green algae and early-diverging plants would not only clarify phylogenetic relationships, but also help to determine which CAZymes are ancestral, lineage-specific, or inherited through horizontal gene transfer.

How complete is this census?

Our census focused on CAZyme proteins identifiable from our draft genome of *P. parkeae* NIES254 and database transcriptomic data for *P. parkeae* CCMP726. However, our statistics for the assembled scaffolds indicate that the draft genome is still highly fragmented, making it challenging to obtain full length gene sequences. Consequently, CAZyme annotation generated false negative results when coding regions needed to meet annotation criteria were missing. To gain more information regarding CAZYmes present in *P. parkeae*, we included *P. parkeae* transcriptome data in the analyses, which significantly increased the number of CAZymes identified. However, it is possible that some CAZyme-encoding genes were present but not expressed at the time of RNA extraction for transcriptomics, and were therefore not detected.

An additional reason that this census is likely to be incomplete is inadequate database coverage of CAZymes. A CAZyme protein family can be defined only when its members have been fully characterized at the biochemical level. Therefore, gene families that have not been characterized biochemically could not be annotated in this study. Also, some of the *P. parkeae* protein sequences might be too divergent to recognize with methods we employed for comparison to known CAZymes.

Summary

We report CAZymes detected in new *P. parkeae* genomic sequence and publically available transcriptomic data. Our results showed that most of the CAZyme families previously identified in Viridiplantae are present in this early-diverging green alga. We also found that some of the CAZymes present in wall-less *P. parkeae* exhibit homology with embryophyte proteins known to be involved in biosynthesis or degradation of cell wall components–cellulose, xyloglucan, and pectin–as well as starch and trehalose. In particular, sequence evidence for the presence of *P. parkeae* GT2 protein sequence exhibiting catalytic domains characteristic of bacterial, green algal, and streptophyte cellulose synthases may shed light on evolutionary acquisition of *Bcs* and/or *CesA* genes in Viridiplantae. Lastly, although this census of *P. parkeae* CAZymes is likely incomplete, our findings provide foundational information for further study of CAZymes in Viridiplantae.

Acknowledgements

This study was supported in part by NSF grants DEB-1119944 awarded to L.G. We thank the Laboratory of Genetics for access to computational infrastructure and support.

References

Adair, W.S., Steinmetz, S.A., Mattson, D.M., Goodenough, U.W. & Heuser, J.E. 1987. Nucleated assembly of *Chlamydomonas* and *Volvox* cell walls. *J. Cell Biol*. 105(5):2373-2382.

Arioli, T., Peng, L., Betzner, A.S., Burn, J., Wittke, W., Herth, W., Camilleri, C., Höfte, H., Plazinski, J., Birch, R. & Cork, A. 1998. Molecular analysis of cellulose biosynthesis in *Arabidopsis*. *Science 279*(5351):717-720.

Ball, S., Colleoni, C., Cenci, U., Raj, J.N. & Tirtiaux, C. 2011. The evolution of glycogen and starch metabolism in eukaryotes gives molecular clues to understand the establishment of plastid endosymbiosis. *J. Exp. Bot.* 62(6):1775-1801.

Banks, J.A., Nishiyama, T., Hasebe, M., Bowman, J.L., Gribskov, M., Albert, V.A., Aono, N., Aoyama, T., Ambrose, B.A., Ashton, N.W. & Axtell, M.J. 2011. The *Selaginella* genome identifies genetic changes associated with the evolution of vascular plants. *Science* 332(6032):960-963.

Blanc, G., Duncan, G., Agarkova, I., Borodovsky, M., Gurnon, J., Kuo, A., Lindquist, E., Lucas, S., Pangilinan, J., Polle, J. & Salamov, A. 2010. The *Chlorella variabilis* NC64A genome reveals adaptation to photosymbiosis, coevolution with viruses, and cryptic sex. *Plant Cell* 22(9):2943-2955.

Blanc, G., Agarkova, I., Grimwood, J., Kuo, A., Brueggeman, A., Dunigan, D.D., Gurnon, J., Ladunga, I., Lindquist, E., Lucas, S. and Pangilinan, J. 2012. The genome of the polar eukaryotic microalga *Coccomyxa subellipsoidea* reveals traits of cold adaptation. *Genome Biol.* 13(5):R39.

Bogen, C., Al-Dilaimi, A., Albersmeier, A., Wichmann, J., Grundmann, M., Rupp, O., Lauersen, K.J., Blifernez-Klassen, O., Kalinowski, J., Goesmann, A. & Mussgnug, J.H. 2013. Reconstruction of the lipid metabolism for the microalga *Monoraphidium neglectum* from its genome sequence reveals characteristics suitable for biofuel production. *BMC genomics* 14(1):926.

Brown, R.M. & Montezinos, D. 1976. Cellulose microfibrils: visualization of biosynthetic and orienting complexes in association with the plasma membrane. *Proc. Natl. Acad. Sci. U. S. A.* 73(1):143-147.

Busi, M.V., Barchiesi, J., Martín, M. & Gomez-Casati, D.F. 2014. Starch metabolism in green algae. *Starch* 66(1-2):28-40.

Cantarel, B.L., Coutinho, P.M., Rancurel, C., Bernard, T., Lombard, V. & Henrissat, B. 2009. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. *Nucleic Acid Res.* 37(suppl 1):D233-D238.

Cantarel, B.L., Korf, I., Robb, S.M., Parra, G., Ross, E., Moore, B., Holt, C., Alvarado, A.S. & Yandell, M. 2008. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* 18(1):188-196.

Cao, X., Wu, X., Ji, C., Yao, C., Chen, Z., Li, G. & Xue, S. 2014. Comparative transcriptional study on the hydrogen evolution of marine microalga *Tetraselmis subcordiformis*. *Int. J. Hydrogen Energy* 39(32):18235-18246

Cenci, U., Nitschke, F., Steup, M., Minassian, B.A., Colleoni, C. & Ball, S.G. 2014. Transition from glycogen to starch metabolism in Archaeplastida. *Trends. Plant Sci.* 19(1):18-28.

Chatterjee, M., Berbezy, P., Vyas, D., Coates, S. & Barsby, T. 2005. Reduced expression of a protein homologous to glycogenin leads to reduction of starch content in *Arabidopsis* leaves. *Plant Sci.* 168(2):501-509.

Conesa, A., Götz, S., García-Gómez, J.M., Terol, J., Talón, M. & Robles, M. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21(18):3674-3676.

Coutinho, P.M., Stam, M., Blanc, E. & Henrissat, B. 2003. Why are there so many carbohydrate-active enzyme-related genes in plants? *Trends Plant Sci.* 8(12):563-565.

Del Bem, L.E. & Vincentz, M.G. 2010. Evolution of xyloglucan-related genes in green plants. *BMC Evol. Biol.* 10(1)341.

Derelle, E., Ferraz, C., Rombauts, S., Rouzé, P., Worden, A.Z., Robbens, S., Partensky, F., Degroeve, S., Echeynié, S., Cooke, R. & Saeys, Y. 2006. Genome analysis of the smallest free-living eukaryote *Ostreococcus tauri* unveils many unique features. *Proc. Natl. Acad. Sci. U. S. A.* 103(31):11647-11652.

Domozych, D.S., Sørensen, I. & Willats, W.G. 2009. The distribution of cell wall polymers during antheridium development and spermatogenesis in the Charophycean green alga, *Chara corallina*. *Ann. Bot.* 104(6):1045-1056.

Eastmond, P.J., Van Dijken, A.J., Spielman, M., Kerr, A., Tissier, A.F., Dickinson, H.G., Jones, J.D., Smeekens, S.C. & Graham, I.A. 2002. Trehalose-6-phosphate synthase 1, which catalyses the first step in trehalose synthesis, is essential for *Arabidopsis* embryo maturation. *Plant J.* 29(2):225-235.

Darriba, D., Taboada, G.L., Doallo, R. & Posada, D. 2011. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* 27(8):1164-1165.

Deschamps, P., Colleoni, C., Nakamura, Y., Suzuki, E., Putaux, J.L., Buléon, A., Haebel, S., Ritte, G., Steup, M., Falcón, L.I. & Moreira, D. 2008. Metabolic symbiosis and the birth of the plant kingdom. *Mol. Biol. Evol.* 25(3):536-548.

Gardiner, J.C., Taylor, N.G. & Turner, S.R. 2003. Control of cellulose synthase complex localization in developing xylem. *Plant Cell 15*(8):1740-1748.

Graham, L.E., Graham, J.M., Wilcox, L.W., Cook, M.E. 2016. Algae. 3rd ed.
LJLM Press, Madison, WI, 595 pp.

Geisler-Lee, J., Geisler, M., Coutinho, P.M., Segerman, B., Nishikubo, N., Takahashi, J., Aspeborg, H., Djerbi, S., Master, E., Andersson-Gunnerås, S. & Sundberg, B. 2006. Poplar carbohydrate-active enzymes. Gene identification and expression analyses. *Plant Physiol.* 140(3):946-962.

Gertz, E.M., Yu, Y.K., Agarwala, R., Schäffer, A.A. & Altschul, S.F. 2006. Composition-based statistics and translated nucleotide searches: improving the TBLASTN module of BLAST. *BMC Biol.* 4:41.

Giddings, T.H., Brower, D.L. & Staehelin, L.A. 1980. Visualization of particle complexes in the plasma membrane of *Micrasterias denticulata* associated with the formation of cellulose fibrils in primary and secondary cell walls. *J. Cell Biol.* 84(2):327-339.

Harholt, J., Sørensen, I., Fangel, J., Roberts, A., Willats, W.G., Scheller, H.V., Petersen, B.L., Banks, J.A. & Ulvskov, P. 2012. The glycosyltransferase repertoire of the spikemoss *Selaginella moellendorffii* and a comparative study of its cell wall. *PloS ONE* 7(5):e35846.

Henrissat, B., Coutinho, P.M. & Davies, G.J., 2001. A census of carbohydrate-active enzymes in the genome of *Arabidopsis thaliana*. *Plant Mol. Biol.* 47:55-72.

Herth, W. 1983. Arrays of plasma-membrane "rosettes" involved in cellulose microfibril formation of *Spirogyra*. *Planta* 159(4):347-356.

Hirano, T., Tanidokoro, K., Shimizu, Y., Kawarabayasi, Y., Ohshima, T., Sato, M., Tadano, S., Ishikawa, H., Takio, S., Takechi, K. & Takano, H. 2016. Moss chloroplasts are surrounded by a peptidoglycan wall containing D-amino acids. *Plant Cell* 28(7):1521-1532.

Hori, K., Maruyama, F., Fujisawa, T., Togashi, T., Yamamoto, N., Seo, M., Sato, S., Yamada, T., Mori, H., Tajima, N. & Moriyama, T. 2014. *Klebsormidium flaccidum* genome reveals primary factors for plant terrestrial adaptation. *Nat. Commun.* 5.

Hotchkiss, A.T. & Brown, R.M., 1987. The association of rosette and globule terminal complexes with cellulose microfibril assembly in *Nitella translucens* var. axillaris (Charophyceae). *J. Phycol.* 23(s2):229-237.

Hu, L., Grim, C.J., Franco, A.A., Jarvis, K.G., Sathyamoorthy, V., Kothary, M.H., McCardell, B.A. & Tall, B.D. 2015. Analysis of the cellulose synthase operon genes, *bcs*A, *bcs*B, and *bcs*C in *Cronobacter* species: Prevalence among species and their roles in biofilm formation and cell–cell aggregation. *Food Microbiol.* 52:97-105.

Imam, S.H., Buchanan, M.J., Shin, H.C. & Snell, W.J., 1985. The *Chlamydomonas* cell wall: characterization of the wall framework. *J. Cell Biol.* 101(4):1599-1607.

Itoh, T., 1990. Cellulose synthesizing complexes in some giant marine algae. *J. Cell Sci.* 95(2):309-319.

Iwai, H., Masaoka, N., Ishii, T. & Satoh, S. 2002. A pectin glucuronyltransferase gene is essential for intercellular attachment in the plant meristem. *Proc. Natl. Acad. Sci. U. S. A.* 99(25):16319-16324.

Katsaros, C., Reiss, H.D. & Schnepf, E. 1996. Freeze-fracture studies in brown algae: putative cellulose-synthesizing complexes on the plasma membrane. *Eur. J. Phycol.* 31(1):41-48.

Katoh, K. & Standley, D.M. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30(4):772-780.

Keeling, P.J., Burki, F., Wilcox, H.M., Allam, B., Allen, E.E., Amaral-Zettler, L.A., Armbrust, E.V., Archibald, J.M., Bharti, A.K., Bell, C.J. & Beszteri, B. 2014. The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. *PLoS Biol.* 12(6):e1001889.

Keskiaho, K., Hieta, R., Sormunen, R. & Myllyharju, J. 2007. *Chlamydomonas reinhardtii* has multiple prolyl 4-hydroxylases, one of which is essential for proper cell wall assembly. *Plant Cell* 19(1):256-269.

Kim, N.H., Herth, W., Vuong, R. & Chanzy, H. 1996. The cellulose system in the cell wall of *Micrasterias*. *J. Struct. Biol.* 117(3):195-203.

Kim, E. & Maruyama, S. 2014. A contemplation on the secondary origin of green algal and plant plastids. *Acta Soc. Bot. Pol.* 83(4).

Kumar, M. & Turner, S. 2015. Plant cellulose synthesis: CESA proteins crossing kingdoms. *Phytochemistry* 112:91-99.

Kurek, I., Kawagoe, Y., Jacob-Wilk, D., Doblin, M. & Delmer, D. 2002. Dimerization of cotton fiber cellulose synthase catalytic subunits occurs via oxidation of the zinc-binding domains. *Proc. Natl. Acad. Sci. U.S.A.* 99(17):11109-11114.

Lahaye, M., Jegou, D. & Buleon, A. 1994. Chemical characteristics of insoluble glucans from the cell wall of the marine green alga *Ulva lactuca* (L.) Thuret. *Carbohydr. Res.* 262(1):115-125.

Leliaert, F., Tronholm, A., Lemieux, C., Turmel, M., DePriest, M.S., Bhattacharya, D., Karol, K.G., Fredericq, S., Zechman, F.W. and Lopez-Bautista, J.M. 2016. Chloroplast phylogenomic analyses reveal the deepest-branching lineage of the Chlorophyta, Palmophyllophyceae class. nov. *Sci.Rep*. 6:25367

Lombard, V., Ramulu, H.G., Drula, E., Coutinho, P.M. & Henrissat, B. 2014. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.* 42(D1):D490-D495.

Machida, M., Takechi, K., Sato, H., Chung, S.J., Kuroiwa, H., Takio, S., Seki, M., Shinozaki, K., Fujita, T., Hasebe, M. & Takano, H. 2006. Genes for the peptidoglycan synthesis pathway are essential for chloroplast division in moss. *Proc. Natl. Acad. Sci. U.S.A.* 103(17):6753-6758.

Madson, M., Dunand, C., Li, X., Verma, R., Vanzin, G.F., Caplan, J., Shoue, D.A., Carpita, N.C. & Reiter, W.D. 2003. The MUR3 gene of Arabidopsis encodes a xyloglucan galactosyltransferase that is evolutionarily related to animal exostosins. *Plant Cell* 15(7):1662-1670.

Merchant, S.S., Prochnik, S.E., Vallon, O., Harris, E.H., Karpowicz, S.J., Witman, G.B., Terry, A., Salamov, A., Fritz-Laylin, L.K., Maréchal-Drouard, L. & Marshall, W.F. 2007. The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. *Science* 318(5848):245-250.

Michel, G., Tonon, T., Scornet, D., Cock, J.M. & Kloareg, B. 2010. The cell wall polysaccharide metabolism of the brown alga *Ectocarpus siliculosus*. Insights into the evolution of extracellular matrix polysaccharides in Eukaryotes. *New Phytol.* 188(1):82-97.

Mikkelsen, M.D., Harholt, J., Ulvskov, P., Johansen, I.E., Fangel, J.U., Doblin, M.S., Bacic, A. & Willats, W.G. 2014. Evidence for land plant cell wall biosynthetic mechanisms in charophyte green algae. *Ann. Bot.* 114(6):1217-1236.

Miller, M.A., Pfeiffer, W. & Schwartz, T. 2010. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. In Gateway Computing Environments Workshop (GCE), *IEEE*. 1-8.

Minic, Z. & Jouanin, L. 2006. Plant glycoside hydrolases involved in cell wall polysaccharide degradation. *Plant Physiol. Biochem.* 44(7):435-449.

Mølhøj, M., Ulvskov, P. & Dal Degan, F. 2001. Characterization of a functional soluble form of a *Brassica napus* membrane-anchored endo-1, 4-β-glucanase heterologously expressed in *Pichia pastoris*. *Plant Physiol*. 127(2):674-684.

Moller, I., Sørensen, I., Bernal, A.J., Blaukopf, C., Lee, K., Øbro, J., Pettolino, F., Roberts, A., Mikkelsen, J.D., Knox, J.P. & Bacic, A. 2007. High-throughput mapping of cell-wall polymers within and between plants using novel microarrays. *Plant J.* 50(6):1118-1128.

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

Nakamura, Y., Takahashi, J.I., Sakurai, A., Inaba, Y., Suzuki, E., Nihei, S., Fujiwara, S., Tsuzuki, M., Miyashita, H., Ikemoto, H. & Kawachi, M. 2005. Some cyanobacteria synthesize semi-amylopectin type α-polyglucans instead of glycogen. *Plant Cell Physiol.* 46(3):539-545.

Nobles, D.R., Romanovicz, D.K. & Brown, R.M. 2001. Cellulose in cyanobacteria. Origin of vascular plant cellulose synthase?. *Plant Physiol.* 127(2):529-542.

Okuda, K. & Brown Jr, R.M. 1992. A new putative cellulose-synthesizing complex of *Coleochaete scutata*. *Protoplasma* 168(1-2):51-63.

Okuda, K., Sekida, S., Yoshinaga, S. & Suetomo, Y. 2004. Cellulose-synthesizing complexes in some chromophyte algae. *Cellulose* 11(3-4):365-376.

Okuda, K. & Sekida, S. 2007. Cellulose: molecular and structural biology. Springer Netherlands, 379 pp.

Palenik, B., Grimwood, J., Aerts, A., Rouzé, P., Salamov, A., Putnam, N., Dupont, C., Jorgensen, R., Derelle, E., Rombauts, S. & Zhou, K. 2007. The tiny eukaryote *Ostreococcus* provides genomic insights into the paradox of plankton speciation. *Proc. Natl. Acad. Sci. U. S. A.* 104(18):7705-7710.

Pearson, B.R. & Norris, R.E. 1975. Fine structure of cell division in *Pyramimonas parkeae* Norris and Pearson (Chlorophyta, Prasinophyceae). *J. Phycol. 11*(1):113-124.

Peña, M.J., Darvill, A.G., Eberhard, S., York, W.S. & O'Neill, M.A. 2008. Moss and liverwort xyloglucans contain galacturonic acid and are structurally distinct from the xyloglucans synthesized by hornworts and vascular plants. *Glycobiology* 18(11):891-904.

Pelloux, J., Rusterucci, C. & Mellerowicz, E.J. 2007. New insights into pectin methylesterase structure and function. *Trends Plant Sci.* 12(6):267-277.

Peng, L., Xiang, F., Roberts, E., Kawagoe, Y., Greve, L.C., Kreuz, K. & Delmer, D.P. 2001. The experimental herbicide CGA 325′ 615 inhibits synthesis of crystalline cellulose and causes accumulation of non-crystalline β-1, 4-glucan associated with CesA protein. *Plant Physiol.* 126(3):981-992.

Popper, Z.A. & Fry, S.C. 2003. Primary cell wall composition of bryophytes and charophytes. *Ann. Bot.* 91(1):1-12.

Preiss, J., Ball, K., Smith-White, B., Iglesias, A., Kakefuda, G. & Li, L. 1991. Starch biosynthesis and its regulation. Biochem Soc. Trans. 19(3):539-547.

Prochnik, S.E., Umen, J., Nedelcu, A.M., Hallmann, A., Miller, S.M., Nishii, I., Ferris, P., Kuo, A., Mitros, T., Fritz-Laylin, L.K. & Hellsten, U. 2010. Genomic analysis of organismal complexity in the multicellular green alga *Volvox carteri*. *Science* 329(5988):223-226.

Rao, G.U. & Paran, I. 2003. Polygalacturonase: a candidate gene for the soft flesh and deciduous fruit mutation in *Capsicum*. *Plant Mol. Biol.* 51(1):135-141

Rensing, S.A., Ick, J., Fawcett, J.A., Lang, D., Zimmer, A., Van de Peer, Y. & Reski, R. 2007. An ancient genome duplication contributed to the abundance of metabolic genes in the moss *Physcomitrella patens*. *BMC Evol. Biol.* 7(1):130.

Römling, U. 2002. Molecular biology of cellulose production in bacteria. *Res. Microbiol.* 153(4):205-212.

Sampedro, J., Sieiro, C., Revilla, G., González-Villa, T. & Zarra, I. 2001. Cloning and expression pattern of a gene encoding an α-xylosidase active against xyloglucan oligosaccharides from *Arabidopsis*. *Plant Physiol.* 126(2):910-920.

Scheller, H.V. & Ulvskov, P. 2010. Hemicelluloses. *Annu. Rev. Plant Biol.* 61(1):263-289.

Schultink, A., Liu, L., Zhu, L. & Pauly, M. 2014. Structural diversity and function of xyloglucan sidechain substituents. *Plants* 3(4):526-542.

Stamatakis, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* btu033.

Sterling, J.D., Atmodjo, M.A., Inwood, S.E., Kolli, V.K., Quigley, H.F., Hahn, M.G. & Mohnen, D. 2006. Functional identification of an *Arabidopsis* pectin biosynthetic homogalacturonan galacturonosyltransferase. *Proc. Natl. Acad. Sci. U. S. A.* 103(13):5236-5241.

Sugiyama, J., Vuong, R. & Chanzy, H. 1991. Electron diffraction study on the two crystalline phases occurring in native cellulose from an algal cell wall. *Macromolecules* 24(14):4168-4175.

Taylor, J.S. & Raes, J. 2004. Duplication and divergence: the evolution of new genes and old ideas. *Annu. Rev. Genet.* 38:615-643.

Taji, T., Ohsumi, C., Iuchi, S., Seki, M., Kasuga, M., Kobayashi, M., Yamaguchi-Shinozaki, K. & Shinozaki, K. 2002. Important roles of drought-and cold-inducible genes for galactinol synthase in stress tolerance in *Arabidopsis thaliana*. *Plant J*. 29(4):417-426.

Thompson, D.S. 2005. How do cell walls regulate plant growth? *J. Exp. Bot.* 56(419):2275-2285.

Tsekos, I. & Reiss, H.D. 1992. Occurrence of the putative microfibril-synthesizing complexes (linear terminal complexes) in the plasma membrane of the epiphytic marine red alga *Erythrocladia subintegra* Rosenv. *Protoplasma* 169(1-2):57-67.

Tsekos, I. 1999. The sites of cellulose synthesis in algae: diversity and evolution of cellulose-synthesizing enzyme complexes. *J. Phycol.* 35(4):635-655.

Tsekos, I., Orologas, N. & Herth, W. 1999. Cellulose microfibril assembly and orientation in some bangiophyte red algae: relationship between synthesizing terminal complexes and microfibril structure, shape, and dimensions. *Phycologia* 38(3):217-224.

Ulvskov, P., Paiva, D.S., Domozych, D. & Harholt, J. 2013. Classification, naming and evolutionary history of glycosyltransferases from sequenced green and red algal genomes. *PloS ONE* 8(10):e76511.

Velasquez, S.M., Ricardi, M.M., Dorosz, J.G., Fernandez, P.V., Nadra, A.D., Pol-Fachin, L., Egelund, J., Gille, S., Harholt, J., Ciancia, M. & Verli, H. 2011. O-glycosylated cell wall proteins are essential in root hair growth. *Science* 332(6036):1401-1403.

Voigt, J., Woestemeyer, J. & Frank, R. 2007. The chaotrope-soluble glycoprotein GP2 is a precursor of the insoluble glycoprotein framework of the *Chlamydomonas* cell wall. *J.Biol. Chem.* 282(42):30381-30392.

Willison, J.H. & Brown, R.M. 1978. Cell wall structure and deposition in *Glaucocystis*. *J.Cell Biol.* 77(1):103-119.

Wingler, A. 2002. The function of trehalose biosynthesis in plants. *Phytochemistry* 60(5):437-440.

Worden, A.Z., Lee, J.H., Mock, T., Rouzé, P., Simmons, M.P., Aerts, A.L., Allen, A.E., Cuvelier, M.L., Derelle, E., Everett, M.V. & Foulon, E. 2009. Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes *Micromonas*. *Science* 324(5924):268-272.

Wray, G.A. 2007. The evolutionary significance of cis-regulatory mutations. *Nature Rev. Genet.* 8(3):206-216.

Yin, Y., Mao, X., Yang, J., Chen, X., Mao, F. & Xu, Y. 2012. dbCAN: a web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res.* 40(W1):W445-W451.

Zaar, K. 1979. Visualization of pores (export sites) correlated with cellulose production in the envelope of the gram-negative bacterium *Acetobacter xylinum*. *J. Cell Biol.* 80(3):773-777.

Zeeman, S.C., Kossmann, J. & Smith, A.M. 2010. Starch: its metabolism, evolution, and biotechnological modification in plants. Annu. Rev. Plant Biol. 61:209-234.

Zhong, R. & Ye, Z.H. 2003. Unraveling the functions of glycosyltransferase family 47 in plants. *Trends Plant Sci.* 8(12):565-568.

Table 1. Protein sequences used in phylogenetic analyses of GT2 protein family.

Organism	Accession number	Description
Prasinophyte algae:		
Pyramimonas parkeae CCMP726	CAMPEP_0191463744	glycosyl transferase family 2 protein
Pyramimonas parkeae CCMP726	CAMPEP_0191465826	glycosyl transferase family 2 protein
Pyramimonas parkeae CCMP726	CAMPEP_0191472438	glycosyl transferase family 2 protein
Pyramimonas parkeae CCMP726	CAMPEP_0191478436	glycosyl transferase family 2 protein
Pyramimonas parkeae CCMP726	CAMPEP_0191479288	glycosyl transferase family 2 protein
Pyramimonas parkeae CCMP726	CAMPEP_0191486766	glycosyl transferase family 2 protein
Pyramimonas parkeae CCMP726	CAMPEP_0191494686	glycosyl transferase family 2 protein
Pyramimonas parkeae CCMP726	CAMPEP_0191507404	glycosyl transferase family 2 protein
Pyramimonas parkeae CCMP726	CAMPEP_0191507432	glycosyl transferase family 2 protein
Ostreococcus lucimarinus CCE9901	XP_001415821	predicted protein
Ostreococcus lucimarinus CCE9901	XP_001417592	predicted protein
Ostreococcus lucimarinus CCE9901	XP_001421850	predicted protein
Micromonas commoda	XP_002507213	predicted protein
Micromonas commoda	XP_002509326	glycosyltransferase family 21 protein
Ostreococcus tauri	XP_003074299	unnamed protein product
Ostreococcus tauri	XP_003083844	unnamed protein product
Bathycoccus prasinos	XP_007515131	predicted protein
Bathycoccus prasinos	XP_007515371	IPT/TIG domain-containing protein
Chlorophyte algae:		
Chlamydomonas reinhardtii	XP_001698563	predicted protein, partial
Volvox carteri	XP_002950756	hypothetical protein VOLCADRAFT_91271
Volvox carteri	XP_002955514	hypothetical protein VOLCADRAFT_96455
Volvox carteri	XP_002958451	hypothetical protein VOLCADRAFT_119967
Volvox carteri	XP_002960031	hypothetical protein VOLCADRAFT_101543, partial
Monoraphidium neglectum	XP_013892410	hypothetical protein MNEG_14572

Organism	Accession number	Description
Monoraphidium neglectum	XP_013895267	hypothetical protein MNEG_11716
Monoraphidium neglectum	XP_013896539	cellulose synthase (UDP-forming)
Monoraphidium neglectum	XP_013901581	cellulose synthase (UDP-forming)
Monoraphidium neglectum	XP_013904643	hypothetical protein MNEG_2335
Monoraphidium neglectum	XP_013905209	hypothetical protein MNEG_1768
Monoraphidium neglectum	XP_013905932	cellulose synthase/ transferase, transferring glycosyl group
Monoraphidium neglectum	XP_013906044	beta-mannan synthase, partial
Coccomyxa subellipsoidea	XP_005643139	Six-hairpin glycosidase
Coccomyxa subellipsoidea	XP_005647231	hypothetical protein COCSUDRAFT_63825
Coccomyxa subellipsoidea	XP_005649406	hypothetical protein COCSUDRAFT_41138
Coccomyxa subellipsoidea	XP_005651400	hypothetical protein COCSUDRAFT_64690
Chlorella variabilis	XP_005845657	hypothetical protein CHLNCDRAFT_58525
Auxenochlorella protothecoides	XP_011396950	Endoglucanase B
Auxenochlorella protothecoides	XP_011399938	hypothetical protein F751_2068
Embryophytes:		
Physcomitrella patens	XP_001753310	putative cellulose synthase 3, glycosyltransferase family 2
Physcomitrella patens	XP_001755866	predicted protein
Physcomitrella patens	XP_001757832	predicted protein
Physcomitrella patens	XP_001757887	cellulose synthase 5, glycosyltransferase family 2
Physcomitrella patens	XP_001759209	cellulose synthase-like A3, glycosyltransferase family 2 protein
Physcomitrella patens	XP_001759264	cellulose synthase-like A2, glycosyltransferase family 2 protein
Physcomitrella patens	XP_001762809	cellulose synthase-like D2, glycosyltransferase family 2
Physcomitrella patens	XP_001764061	cellulose synthase-like A1, glycosyltransferase family 2 protein
Physcomitrella patens	XP_001764498	cellulose synthase-like D8, glycosyltransferase family 2 protein
Physcomitrella patens	XP_001764905	predicted protein
Physcomitrella patens	XP_001767133	cellulose synthase 4, glycosyltransferase family 2
Physcomitrella patens	XP_001769140	cellulose synthase-like D3, glycosyltransferase family 2 protein
Physcomitrella patens	XP_001769175	cellulose synthase-like D4, glycosyltransferase family 2 protein

Organism	Accession number	Description
Physcomitrella patens	XP_001769255	cellulose synthase 8, glycosyltransferase family 2
Physcomitrella patens	XP_001774233	predicted protein
Physcomitrella patens	XP_001775315	predicted protein
Physcomitrella patens	XP_001775317	predicted protein
Physcomitrella patens	XP_001775646	predicted protein
Physcomitrella patens	XP_001776974	cellulose synthase 10, glycosyltransferase family 2
Physcomitrella patens	XP_001778677	cellulose synthase-like D1, glycosyltransferase family 2 protein
Physcomitrella patens	XP_001778678	cellulose synthase-like D7, glycosyltransferase family 2
Physcomitrella patens	XP_001779999	cellulose synthase-like D5, glycosyltransferase family 2
Physcomitrella patens	XP_001781718	cellulose synthase-like D6, glycosyltransferase family 2 protein
Selaginella moellendorffii	XP_002960291	family 2 glycosyltransferase
Selaginella moellendorffii	XP_002960719	hypothetical protein SELMODRAFT_73698
Selaginella moellendorffii	XP_002960761	cellulose synthase 4-1
Selaginella moellendorffii	XP_002960802	ceramide beta-glucosyltransferase
Selaginella moellendorffii	XP_002962367	hypothetical protein SELMODRAFT_404096
Selaginella moellendorffii	XP_002962487	hypothetical protein SELMODRAFT_404311
Selaginella moellendorffii	XP_002963103	hypothetical protein SELMODRAFT_78846
Selaginella moellendorffii	XP_002963530	glycosyltransferase family 2 protein
Selaginella moellendorffii	XP_002963550	family 2 glycosyltransferase
Selaginella moellendorffii	XP_002964575	cellulose synthase-like D1-2, glycosyltransferase family 2 protein
Selaginella moellendorffii	XP_002965783	cellulose synthase-like D2-1, glycosyltransferase family 2 protein
Selaginella moellendorffii	XP_002966771	cellulose synthase-like D3-1, glycosyltransferase family 2 protein
Selaginella moellendorffii	XP_002967423	hypothetical protein SELMODRAFT_86720
Selaginella moellendorffii	XP_002968635	family 2 glycosyltransferase
Selaginella moellendorffii	XP_002968763	hypothetical protein SELMODRAFT_90812
Selaginella moellendorffii	XP_002971505	glycosyltransferase, CAZy family GT2
Selaginella moellendorffii	XP_002971746	hypothetical protein SELMODRAFT_412321

Organism	Accession number	Description
Selaginella moellendorffii	XP_002977978	cellulose synthase-like D3-2, glycosyltransferase family 2 protein
Selaginella moellendorffii	XP_002980221	hypothetical protein SELMODRAFT_112511
Selaginella moellendorffii	XP_002981528	family 2 glycosyltransferase
Selaginella moellendorffii	XP_002981551	family 2 glycosyltransferase
Selaginella moellendorffii	XP_002983910	family 2 glycosyltransferase
Selaginella moellendorffii	XP_002983952	family 2 glycosyltransferase
Selaginella moellendorffii	XP_002984435	cellulose synthase-like D2-2, glycosyltransferase family 2 protein
Selaginella moellendorffii	XP_002988822	cellulose synthase-like D1-2, glycosyltransferase family 2 protein
Selaginella moellendorffii	XP_002989886	glycosyltransferase family 2 protein
Selaginella moellendorffii	XP_002990324	hypothetical protein SELMODRAFT_428828
Selaginella moellendorffii	XP_002990646	hypothetical protein SELMODRAFT_132027
Selaginella moellendorffii	XP_002991857	glycosyltransferase, CAZy family GT2
Selaginella moellendorffii	XP_002993014	glycosyltransferase, CAZy family GT2
Selaginella moellendorffii	XP_002993016	glycosyltransferase, CAZy family GT2
Selaginella moellendorffii	XP_002994516	glycosyltransferase, CAZy family GT2
Cyanobacteria:		
Calothrix sp. PCC 7103	WP_019491820	cellulose synthase catalytic subunit (UDP-forming)
Calothrix sp. PCC 7507	WP_015129306	cellulose synthase catalytic subunit (UDP-forming)
Coleofasciculus chthonoplastes	WP_006102136	cellulose synthase
Cyanothece sp. PCC 8801	WP_012596409	cellulose synthase
Dolichospermum circinale	WP_028082292	cellulose synthase
Leptolyngbya sp. PCC 7375	WP_006515141	cellulose synthase catalytic subunit (UDP-forming)
Nostoc sp. PCC 7120	WP_044521920	cellulose synthase
Pseudanabaena sp. PCC 7367	WP_015165179	cellulose synthase catalytic subunit
Scytonema hofmannii	WP_017748626	cellulose synthase catalytic subunit (UDP-forming)
Synechococcus sp. PCC 7002	WP_012307721	cellulose synthase

Organism	Accession number	Description
Calothrix sp. PCC 6303	AFZ01982	Cellulose synthase BcsB
Calothrix sp. PCC 7507	AFY33497	Cellulose synthase BcsB
Gloeocapsa sp. PCC 7428	AFZ32009	Cellulose synthase BcsB
Leptolyngbya boryana dg5	BAS60480	Cellulose synthase BcsB
Leptolyngbya boryana IAM M-101	BAS54132	Cellulose synthase BcsB
Leptolyngbya sp. NIES-2104	WP_059000847	Cellulose synthase BcsB

Table 2. Listing of CAZymes obtained from assembled genomic data of *Pyramimonas parkeae* NIES254.

CAZymes families	Sequence name	Sequence description	Length (bp)	min E-value	Identity
GT5	contig04676	starch synthase	7619	3.00E-111	52.00%
GT5	contig05032	starch synthase	7409	6.90E-30	82.60%
GT5	contig08440	starch synthase partial	6056	4.40E-33	87.95%
GT5	contig11430	soluble starch synthase ii- partial	5239	8.00E-48	77.05%
GT5	contig17390	granule-bound starch synthase I partial	4076	8.20E-114	73.55%
GT5	contig03984	starch synthase	8121	9.40E-26	77.95%
GT5	contig21417	soluble starch synthase iii	3508	1.00E-20	80.95%
GT5	contig37096	soluble starch synthase iii-1	2103	1.80E-13	67.00%
GT5	contig67273	starch synthase chloroplastic amyloplastic	976	6.70E-42	83.00%
GT20	contig22451	alpha,alpha-trehalose-phosphate synthase [UDP-forming] 1-like	3380	6.80E-48	87.55%
GT20	contig19238	alpha,alpha-trehalose-phosphate synthase [UDP-forming] 9	3807	1.80E-35	86.80%
GT20	contig39533	trehalose-6-phosphate synthase	1959	6.10E-35	76.75%
GH1	contig15531	beta-glucosidase-like chloroplastic	4399	8.00E-10	67.75%
GH1	contig15764	beta-glucosidase-like chloroplastic	4351	1.40E-09	74.55%
GH2	contig18865	beta-d-galactosidase	3869	6.90E-14	74.30%
GH2	contig03292	beta-d-galactosidase	8738	8.00E-16	75.20%
GH2	contig32014	glycoside hydrolase family 2 protein	2455	5.10E-20	75.60%
GH3	contig25234	probable beta-d-xylosidase 5	3065	5.40E-14	75.45%
GH3	contig35436	beta-d-xylosidase 3-like	2208	1.90E-32	74.80%
GH3	contig42275	glycoside hydrolase family 3 protein	1809	4.60E-12	71.30%
GH5	contig14376	mannan endo- -beta-mannosidase 6	4614	7.20E-15	60.95%
GH5	contig01083	mannan endo-1,4-beta-mannosidase	12472	1.10E-11	56.75%
GH5	contig17623	mannan endo- -beta-mannosidase 6	4041	9.10E-14	65.35%
GH5	contig65664	mannan endo- -beta-mannosidase 8-like	1011	1.90E-16	64.55%
GH5	contig83413	Soluble Starch synthase	711	3.50E-11	72.80%
GH13	contig14566	alpha-1,6-glucosidase, pullulanase-type	4580	2.30E-14	70.10%
GH13	contig09312	isoamylase chloroplastic	5803	3.60E-30	87.75%
GH13	contig12693	alpha amylase	4959	4.10E-25	71.55%
GH13	contig27155	alpha-amylase chloroplastic	2876	2.30E-22	75.85%
GH13	contig27652	alpha-amylase chloroplastic	2828	4.10E-10	73.00%
GH13	contig22050	Alpha-amylase [Auxenochlorella protothecoides]	3433	5.60E-66	72.35%
GH13	contig27553	glycoside hydrolase family 13 protein	2839	3.80E-19	66.60%
GH13	contig04624	starch branching enzyme	7655	7.70E-31	84.30%
GH13	contig09068	starch branching enzyme partial	5874	7.90E-27	72.35%
GH13	contig04569	starch branching enzyme partial	7691	4.10E-45	86.00%

CAZymes families	Sequence name	Sequence description	Length (bp)	min E-value	Identity
GH13	contig02858	alpha-amylase	9188	1.20E-20	65.05%
GH13	contig06515	alpha-amylase	6715	1.00E-09	59.92%
GH13	contig27219	alpha-amylase	2869	1.80E-20	60.55%
GH13	contig15065	alpha amylase	4483	1.40E-10	67.55%
GH13	contig15272	pullulanase chloroplastic isoform x1	4444	1.70E-08	89.15%
GH13	contig53879	probable alpha-amylase 2	1336	4.00E-48	79.25%
GH13	contig78622	probable alpha-amylase 2	774	2.90E-17	83.60%
GH13	contig00885	isoamylase-type starch debranching enzyme or isoamylase III	13134	6.30E-12	80.60%
GH13	contig14681	pullulanase type debranching partial	4556	1.90E-08	61.45%
GH13	contig36560	alpha-glucan-branching enzyme	2136	5.00E-33	74.75%
GH13	contig39384	glycoside hydrolase family 13 protein	1968	1.90E-14	67.80%
GH13	contig06824	isoamylase chloroplastic	6601	1.70E-10	69.75%
GH13	contig39045	alpha amylase	1987	7.40E-24	81.65%
GH13	contig33122	alpha-amylase	2372	7.70E-22	65.80%
GH14	contig06265	beta-amylase	6807	1.80E-22	68.40%
GH18	contig04002	glycoside hydrolase family 18 protein	8106	1.30E-11	64.50%
GH27	contig00040	alpha-galactosidase	23774	-	-
GH27	contig01204	alpha-galactosidase	12127	4.20E-25	76.75%
GH30	contig05282	glucosylceramidase	7267	5.00E-15	52.80%
GH30	contig07782	glucosylceramidase	6268	1.50E-20	63.05%
GH31	contig11730	glycoside hydrolase family 31 protein	5170	2.50E-15	71.55%
GH31	contig04107	alpha-glucosidase	8031	1.60E-38	66.80%
GH31	contig11469	glycoside hydrolase family 31 protein	5229	2.90E-30	78.10%
GH31	contig02068	glycosyl family 31	10242	1.70E-34	68.50%
GH31	contig04204	alpha-glucosidase 2	7944	5.70E-33	69.05%
GH31	contig58236	glycosyl hydrolase family 31	1199	3.80E-10	61.75%
GH31	contig13755	alpha-xylosidase 2	4742	6.90E-09	65.75%
GH31	contig17374	glycoside hydrolase family 31 protein	4079	3.50E-13	59.50%
GH33	contig01618	sialidase	11101	7.90E-16	57.63%
GH33	contig03041	glycosyl hydrolase	8976	1.40E-15	65.75%
GH33	contig04027	sialidase	8089	5.60E-15	63.67%
GH33	contig00061	sialidase	22407	-	-
GH33	contig13452	putative exo-alpha-sialidase	4805	2.00E-13	61.07%
GH33	contig20375	probable sialidase	3669	4.60E-13	57.85%
GH33	contig02185	probable sialidase	10044	1.40E-12	59.33%
GH33	contig14116	sialidase	4664	1.70E-13	60.33%
GH33	contig17377	putative exo-alpha-sialidase	4078	1.40E-21	60.27%
GH33	contig39574	sialidase	1957	7.20E-11	60.75%
GH33	contig15607	probable sialidase	4382	1.10E-10	67.00%
GH33	contig17767	probable sialidase	4017	9.30E-07	69.00%

CAZymes families	Sequence name	Sequence description	Length (bp)	min E-value	Identity
GH37	contig07730	probable -trehalose-phosphate synthase	6282	9.30E-28	64.50%
GH37	contig17973	probable trehalase	3988	9.00E-11	75.45%
GH37	contig35069	probable trehalase	2233	2.70E-09	78.05%
GH38	contig42973	lysosomal alpha-mannosidase	1773	3.00E-27	72.35%
GH43	contig08211	glycosyl hydrolase family 43 protein	6131	8.70E-29	72.90%
GH47	contig33102	probable alpha-mannosidase i mns4	2373	3.50E-20	72.35%
GH47	contig02096	mannosyl-oligosaccharide-alpha-mannosidase	10197	2.00E-20	68.30%
GH47	contig01408	probable alpha-mannosidase i mns5	11569	1.70E-14	78.00%
GH47	contig18648	probable alpha-mannosidase i mns4	3886	9.30E-13	74.50%
GH47	contig70711	putative alpha-mannosidase I MNS4 [Triticum urartu]	907	4.80E-19	80.60%
GH47	contig13947	mannosyl-oligosaccharide-alpha-mannosidase	4701	5.20E-10	77.50%
GH47	contig20141	mannosylglycoprotein endo-beta-mannosidase-like isoform x2	3679	1.90E-16	71.45%
GH47	contig78462	endoplasmic reticulum mannosyl-oligosaccharide -alpha-mannosidase-like	776	1.40E-13	89.75%
GH47	contig40603	mannosyl-oligosaccharide -alpha-mannosidase mns1-like	1898	1.40E-14	71.60%
GH47	contig106924	mannosyl-oligosaccharide -alpha-mannosidase mns1-like	492	7.50E-26	74.70%
GH47	contig09490	mannosyl-oligosaccharide -alpha-mannosidase mns1-like isoform x1	5757	1.90E-10	74.50%
GH47	contig22551	mannosyl-oligosaccharide -alpha-mannosidase mns1	3369	9.90E-23	64.70%
GH77	contig18612	glycoside hydrolase family 77 protein, 4-alpha-glucanotransferase [Auxenochlorella protothecoides]	3892	1.10E-17	73.60%
GH77	contig49443	glycoside hydrolase family 77 protein	1493	4.60E-12	79.90%
GH77	contig68542	4-alpha-glucanotransferase	950	4.10E-12	64.35%
GH99	contig03961	glycoprotein endo-alpha-mannosidase	8147	1.40E-13	72.95%
CE13	contig24607	pectin acetylesterase 9	3132	4.90E-14	61.70%
CBM1	contig41735	cellulose-binding fungal	1837	1.20E-06	62.50%
CBM20	contig01449	glycoside hydrolase family 13 protein	11464	3.90E-41	64.05%
CBM20	contig07217	glycoside hydrolase family 13 protein	6461	1.30E-13	62.30%
CBM20	contig09842	carbohydrate-binding module family 20 protein, glucan 1,4-alpha-glucosidase	5652	4.60E-16	66.17%
CBM20	contig23722	phosphoglucan, water dikinase, chloroplastic	3230	1.70E-41	79.70%
CBM20	contig13722	phosphoglucan, water dikinase, chloroplastic	4749	2.20E-33	81.15%
CBM20	contig00908	phosphoglucan, water dikinase, chloroplastic	13058	1.10E-08	59.70%
CBM20	contig06195	phosphoglucan water dikinase, alpha-glucan water dikinase	6839	1.50E-36	75.55%
CBM20	contig23721	phosphoglucan, water dikinase, chloroplastic	3231	2.90E-22	78.20%
CBM20	contig57670	alpha-glucan water dikinae	1216	1.80E-24	73.55%
CBM20	contig00396	phosphoglucan water dikinase, alpha-glucan water dikinase	16217	7.00E-21	69.20%
CBM20	contig27539	phosphoglucan, water dikinase, chloroplastic	2839	1.50E-29	79.75%
CBM20	contig20903	phosphoglucan, water dikinase, chloroplastic	3574	4.70E-15	72.75%

CAZymes families	Sequence name	Sequence description	Length (bp)	min E-value	Identity
CBM20, GH13	contig00845	alpha-amylase	13305	4.00E-10	71.65%
CBM20	contig59551	carbohydrate-binding module family 20 protein	1163	8.40E-11	52.90%
CBM48	contig06276	starch branching enzyme	6807	1.60E-14	83.10%
CBM48	contig18195	carbohydrate-binding module family 48 protein	3953	4.50E-17	77.30%
CBM48	contig22473	carbohydrate-binding module family 48 protein	3379	1.60E-15	72.40%
CBM53	contig06109	starch synthase	6875	3.30E-30	66.55%
CBM53, GT5	contig00712	Soluble Starch synthase	13996	3.00E-18	56.85%
CBM53, GT5	contig06202	soluble starch synthase iii	6836	5.50E-39	68.15%

Table 3. Listing of CAZyme protein sequences inferred from assembled transcriptomic data of *Pyramimonas parkeae* CCMP726.

CAZymes families	Sequence name	Sequence description	Length (aa)	min E-value	Identity
GT1	CAMPEP_0191473538	UDP-glycosyltransferase	350	5.50E-24	48.95%
GT2	CAMPEP_0191478558	glycosyl transferase	436	2.00E-96	65.75%
GT2	CAMPEP_0191507432	glycosyl transferase family 1	844	0.00E+00	54.75%
GT2	CAMPEP_0191463840	glycosyl transferase family 2	812	2.30E-173	53.80%
GT2	CAMPEP_0191494686	glycosyl transferase family 1	1018	0.00E+00	53.80%
GT2	CAMPEP_0191465826	glycosyl transferase	212	1.30E-61	64.85%
GT2	CAMPEP_0191463744	glycosyl transferase family 1	439	2.80E-57	58.50%
GT2	CAMPEP_0191503650	udp- c:betagal beta- -n-acetylglucosaminyltransferase-like protein 1 isoform x2	322	1.00E-89	61.70%
GT2	CAMPEP_0191486766	family 2 glycosyl transferase	403	5.50E-121	71.15%
GT2	CAMPEP_0191472438	glycosyl transferase family 1	230	6.40E-57	61.10%
GT2	CAMPEP_0191507404	glycosyl transferase family 2	729	1.10E-54	46.80%
GT2	CAMPEP_0191478436	glycosyl transferase family 2	781	0.00E+00	53.10%
GT2	CAMPEP_0191481704	dolichol-phosphate mannosyltransferase subunit 1-like	238	3.30E-116	81.70%
GT2	CAMPEP_0191484158	dolichyl-phosphate beta-glucosyltransferase	333	1.80E-94	64.75%
GT2	CAMPEP_0191479288	glycosyl transferase family 2	420	2.50E-52	43.73%
GT4	CAMPEP_0191501924	glycosyltransferase family 4 protein	839	5.70E-165	68.15%
GT4	CAMPEP_0191481616	digalactosyldiacylglycerol synthase chloroplastic-like	552	3.60E-142	66.10%
GT4	CAMPEP_0191496276	glycosyl transferase family 1	218	6.30E-27	48.85%
GT4	CAMPEP_0191468554	glycosyl transferase family 2	777	3.00E-39	44.35%
GT4	CAMPEP_0191468964	glycosyl transferase family 2	818	1.10E-52	48.90%
GT4	CAMPEP_0191463596	cazy family gt4	510	1.20E-45	48.20%
GT4	CAMPEP_0191472944	cazy family gt4	471	2.30E-145	64.25%
GT4	CAMPEP_0191462542	group 1 family glycosyltransferase	135	5.40E-36	65.20%
GT4	CAMPEP_0191483674	glycosyl transferase family 1	204	2.00E-46	61.40%
GT4	CAMPEP_0191484842	glycosyl transferase group 1	314	1.70E-79	49.50%
GT4	CAMPEP_0191472690	glycosyltransferase family 4 protein	503	8.40E-140	65.60%
GT4	CAMPEP_0191472730	alpha- -mannosyltransferase alg2	413	2.60E-133	64.40%
GT4	CAMPEP_0191471204	group 1 family protein	437	7.40E-56	47.30%
GT4	CAMPEP_0191485464	glycosyl group 1	386	1.90E-141	53.40%
GT4	CAMPEP_0191484126	gdp-man:man c -pp-dol alpha- -mannosyltransferase	491	2.10E-142	63.50%
GT4	CAMPEP_0191465134	glycosyl family 1	137	3.90E-24	59.94%
GT5	CAMPEP_0191492284	soluble starch synthase	647	0.00E+00	67.35%
GT5	CAMPEP_0191482370	soluble starch synthase chloroplastic amyloplastic	652	0.00E+00	67.45%
GT5	CAMPEP_0191479448	granule-bound starch synthase chloroplastic amyloplastic	583	1.20E-168	65.30%
GT5	CAMPEP_0191486440	glycosyltransferase family 5 protein	399	3.10E-170	72.85%
GT5	CAMPEP_0191482422	soluble starch synthase i	739	0.00E+00	71.80%
GT5	CAMPEP_0191479490	granule-bound starch synthase chloroplastic amyloplastic-like	614	0.00E+00	71.20%
GT7	CAMPEP_0191484224	glycosyltransferase family 7 protein	419	7.00E-44	58.20%
GT8	CAMPEP_0191507902	probable galacturonosyltransferase 13 isoform x1	522	1.30E-16	47.25%
GT8	CAMPEP_0191471398	glucuronosyltransferase pgsip8	455	2.70E-57	49.90%

CAZymes families	Sequence name	Sequence description	Length (aa)	min E-value	Identity
GT8	CAMPEP_0191493730	probable galacturonosyltransferase-like 10	327	6.90E-84	61.70%
GT8	CAMPEP_0191477494	glycosyltransferase 8 domain-containing protein 1 isoform x2	359	1.90E-31	50.40%
GT8	CAMPEP_0191479436	glycogenin-1	425	1.80E-133	45.80%
GT8	CAMPEP_0191494888	glycosyl transferase	687	3.00E-17	45.75%
GT8	CAMPEP_0191464514	probable galacturonosyltransferase-like 9	342	1.20E-17	43.90%
GT8	CAMPEP_0191506248	protein	831	3.60E-53	46.70%
GT8	CAMPEP_0191485104	protein	370	1.40E-55	46.70%
GT8	CAMPEP_0191466058	protein	891	1.30E-51	46.10%
GT8	CAMPEP_0191477662	probable galacturonosyltransferase-like 4	353	3.50E-82	62.10%
GT8	CAMPEP_0191479232	Glycogenin	716	1.60E-31	42.20%
GT8	CAMPEP_0191476798	glycosyltransferase 8 domain-containing protein 2	395	1.20E-90	51.75%
GT8	CAMPEP_0191507894	NA	407	-	-
GT8	CAMPEP_0191463788	protein cdi-like	291	4.30E-152	71.40%
GT8	CAMPEP_0191497872	protein cdi-like	735	4.30E-97	67.00%
GT8	CAMPEP_0191499140	protein cdi-like	735	2.50E-97	66.95%
GT8	CAMPEP_0191466230	unknown protein	413	2.50E-68	44.95%
GT10	CAMPEP_0191464916	glycoprotein 3-alpha-l-fucosyltransferase	950	5.90E-14	42.10%
GT10	CAMPEP_0191507100	alpha-( )-fucosyltransferase-like	888	9.80E-13	52.60%
GT10	CAMPEP_0191465730	glycoprotein 3-alpha-l-fucosyltransferase	687	2.30E-28	47.10%
GT10	CAMPEP_0191508450	hypothetical protein RFI_13014	918	2.40E-23	42.33%
GT13	CAMPEP_0191462936	alpha- -mannosyl-glycoprotein 2-beta-n-acetylglucosaminyltransferase	618	2.00E-158	55.45%
GT13	CAMPEP_0191495048	alpha- -mannosyl-glycoprotein 2-beta-n-acetylglucosaminyltransferase isoform x1	516	1.00E-121	57.95%
GT13	CAMPEP_0191466322	protein o-linked-mannose beta- -n-acetylglucosaminyltransferase 1-like	634	6.20E-29	43.85%
GT14	CAMPEP_0191505294	PREDICTED: uncharacterized protein LOC101763968	405	3.70E-37	48.60%
GT15	CAMPEP_0191481456	mannosyltransferase ktr2	603	2.20E-27	43.90%
GT15	CAMPEP_0191480912	glycosyltransferase family 15 protein	421	2.80E-42	50.55%
GT17	CAMPEP_0191464216	glycosyltransferase family 17 protein	513	2.50E-09	49.13%
GT18	CAMPEP_0191499838	alpha- -mannosylglycoprotein 6-beta-n-acetylglucosaminyltransferase a isoform x1	497	3.20E-29	45.30%
GT19	CAMPEP_0191463488	lipid-a-disaccharide synthase	540	6.70E-107	59.70%
GT20	CAMPEP_0191501708	trehalose-phosphate synthase	851	0.00E+00	66.55%
GT20	CAMPEP_0191486514	trehalose-phosphate synthase	1043	0.00E+00	75.90%
GT21	CAMPEP_0191497790	glucosylceramide synthase (udp-glucose-dependent)	499	6.10E-110	51.35%
GT22	CAMPEP_0191494504	dol-p-man:man c -pp-dol alpha- -mannosyltransferase isoform x1	489	2.20E-130	60.15%
GT22	CAMPEP_0191498252	dol-p-man:man c -pp-dol alpha- -mannosyltransferase	545	3.30E-104	55.65%
GT22	CAMPEP_0191469544	gpi mannosyltransferase 3	570	1.10E-150	62.05%
GT23	CAMPEP_0191506532	alpha-( )-fucosyltransferase-like	264	1.30E-38	46.13%
GT23	CAMPEP_0191471810	exostosin-like glycosyltransferase	1029	8.30E-117	48.25%
GT23	CAMPEP_0191486714	glycosyltransferase family 37 protein	514	1.90E-74	42.30%
GT23	CAMPEP_0191465268	alpha-( )-fucosyltransferase	204	1.90E-51	56.80%
GT23	CAMPEP_0191496390	alpha-( )-fucosyltransferase	363	7.30E-86	46.00%
GT23	CAMPEP_0191469296	protein	502	8.50E-36	43.88%
GT23	CAMPEP_0191487928	protein	481	2.00E-34	46.17%
GT23	CAMPEP_0191472094	hypothetical protein GUITHDRAFT_120848	431	7.00E-13	45.80%

CAZymes families	Sequence name	Sequence description	Length (aa)	min E-value	Identity
GT23	CAMPEP_0191484812	alpha-(1,6)-fucosyltransferase	361	-	-
GT23	CAMPEP_0191473040	glycosyltransferase family 23 protein	629	-	-
GT24	CAMPEP_0191480530	udp-glucose:glycoprotein glucosyltransferase isoform x1	1678	0.00E+00	56.50%
GT28	CAMPEP_0191463570	udp-n-acetylglucosamine--n-acetylmuramyl-pyrophosphoryl-undecaprenol n-acetylglucosamine transferase	425	2.60E-94	57.00%
GT28	CAMPEP_0191508050	monogalactosyldiacylglycerol synthase	376	8.30E-119	61.45%
GT29	CAMPEP_0191488638	cmp-n-acetylneuraminate-beta-galactosamide-alpha- -sialyltransferase 1	381	3.40E-41	47.25%
GT29	CAMPEP_0191467772	cmp-n-acetylneuraminate-beta-galactosamide-alpha- -sialyltransferase 1	619	2.00E-32	46.55%
GT29	CAMPEP_0191462938	cmp-n-acetylneuraminate-beta-galactosamide-alpha- -sialyltransferase 1	731	2.20E-51	47.05%
GT29	CAMPEP_0191482780	cmp-n-acetylneuraminate-beta-galactosamide-alpha- -sialyltransferase 1-like	364	5.30E-29	46.85%
GT29	CAMPEP_0191505836	cmp-n-acetylneuraminate-beta-galactosamide-alpha- -sialyltransferase 2	498	1.80E-37	47.60%
GT29	CAMPEP_0191481044	cmp-n-acetylneuraminate-beta-galactosamide-alpha- -sialyltransferase 2-like	670	1.70E-132	48.60%
GT29	CAMPEP_0191462540	cmp-n-acetylneuraminate-beta-galactosamide-alpha- -sialyltransferase 2-like	598	4.30E-18	48.05%
GT29	CAMPEP_0191497626	tenascin-x isoform x1	458	1.30E-52	63.40%
GT29	CAMPEP_0191463640	tenascin XB [Bathycoccus prasinos]	556	4.30E-68	52.75%
GT29	CAMPEP_0191497976	alpha- -sialyltransferase 8f	371	1.00E-95	46.45%
GT29	CAMPEP_0191492814	beta-galactoside alpha- -sialyltransferase 1	431	1.30E-38	48.10%
GT29	CAMPEP_0191473564	cmp-n-acetylneuraminate-beta-galactosamide-alpha- -sialyltransferase 1	510	2.00E-45	47.70%
GT29	CAMPEP_0191475084	cmp-n-acetylneuraminate-beta-galactosamide-alpha- -sialyltransferase 1	216	2.10E-76	54.25%
GT29	CAMPEP_0191471284	cmp-n-acetylneuraminate-beta-galactosamide-alpha- -sialyltransferase 1	788	2.00E-18	51.25%
GT29	CAMPEP_0191479862	beta-galactoside alpha- -sialyltransferase 1	514	2.80E-30	45.25%
GT29	CAMPEP_0191476494	beta-galactoside alpha- -sialyltransferase 2	796	1.50E-45	50.15%
GT29	CAMPEP_0191495590	beta-galactoside alpha- -sialyltransferase 2-like	426	1.30E-37	51.20%
GT29	CAMPEP_0191508568	cmp-n-acetylneuraminate-beta-galactosamide-alpha- -sialyltransferase 1	583	2.70E-44	49.35%
GT29	CAMPEP_0191472478	cmp-n-acetylneuraminate-beta-galactosamide-alpha- -sialyltransferase 1	595	5.60E-48	47.45%
GT29	CAMPEP_0191502578	cmp-n-acetylneuraminate-beta-galactosamide-alpha- -sialyltransferase 1	479	1.70E-57	47.15%
GT29	CAMPEP_0191502696	cmp-n-acetylneuraminate-beta-galactosamide-alpha- -sialyltransferase 1	591	2.60E-47	49.35%
GT29	CAMPEP_0191473938	beta-galactoside alpha- -sialyltransferase 1	376	3.10E-127	47.15%
GT29	CAMPEP_0191473366	sialyltransferase-like protein 4	202	1.20E-31	60.10%
GT29	CAMPEP_0191496140	cmp-n-acetylneuraminate-beta-galactosamide-alpha- 3-sialyltransferase partial	448	1.20E-29	47.95%
GT29	CAMPEP_0191476620	cmp-n-acetylneuraminate-beta-galactosamide-alpha- -sialyltransferase 1	532	4.60E-33	48.90%
GT29	CAMPEP_0191477380	cmp-n-acetylneuraminate-beta-galactosamide-alpha- -sialyltransferase 1	308	4.10E-46	47.95%

CAZymes families	Sequence name	Sequence description	Length (aa)	min E-value	Identity
GT29	CAMPEP_0191500154	cmp-n-acetylneuraminate-beta-galactosamide-alpha- -sialyltransferase 1	311	2.80E-44	48.45%
GT29	CAMPEP_0191493970	protein	480	5.90E-41	45.25%
GT29	CAMPEP_0191463516	cmp-n-acetylneuraminate-beta-galactosamide-alpha- -sialyltransferase 1	844	1.70E-74	50.00%
GT29	CAMPEP_0191494076	cmp-n-acetylneuraminate-poly-alpha- 8-sialyltran sferase	651	5.80E-47	57.40%
GT29	CAMPEP_0191474792	beta-galactoside alpha- -sialyltransferase 1-like	452	4.40E-18	55.35%
GT29	CAMPEP_0191505888	cmp-n-acetylneuraminate-beta-galactosamide-alpha- -sialyltransferase 2-like isoform x1	561	4.30E-38	43.85%
GT29	CAMPEP_0191471854	beta-galactoside alpha- -sialyltransferase 2	394	1.60E-60	50.25%
GT29	CAMPEP_0191466678	cmp-n-acetylneuraminate-beta-galactosamide-alpha- -sialyltransferase	432	4.20E-26	45.75%
GT29	CAMPEP_0191465110	cmp-n-acetylneuraminate-beta-galactosamide-alpha- -sialyltransferase 2 isoform x1	903	2.10E-17	50.44%
GT29	CAMPEP_0191477642	cmp-n-acetylneuraminate-beta-galactosamide-alpha- -sialyltransferase 1	417	1.70E-94	47.05%
GT29	CAMPEP_0191504412	beta-galactoside alpha- -sialyltransferase 2	935	8.90E-75	49.85%
GT29	CAMPEP_0191503738	beta-galactoside alpha- -sialyltransferase 2	390	2.10E-87	47.40%
GT29	CAMPEP_0191500624	beta-galactoside alpha- -sialyltransferase 2	544	1.00E-85	51.90%
GT29	CAMPEP_0191474906	cmp-n-acetylneuraminate-beta-galactosamide-alpha- 3-sialyltransferase partial	453	1.00E-134	52.60%
GT29	CAMPEP_0191472112	cmp-n-acetylneuraminate-beta-galactosamide-alpha- -sialyltransferase 1	306	3.10E-83	48.65%
GT29	CAMPEP_0191463750	cmp-n-acetylneuraminate-beta-galactosamide-alpha- -sialyltransferase 1	185	7.60E-15	58.54%
GT29	CAMPEP_0191464126	sialyltransferase	248	3.90E-42	65.05%
GT29	CAMPEP_0191473892	cmp-n-acetylneuraminate-beta-galactosamide-alpha- -sialyltransferase 1	376	1.20E-32	45.35%
GT29	CAMPEP_0191477158	cmp-n-acetylneuraminate-beta-galactosamide-alpha- -sialyltransferase 1	279	4.60E-16	46.10%
GT29	CAMPEP_0191468270	beta-galactoside alpha- -sialyltransferase 1	424	1.30E-50	46.55%
GT29	CAMPEP_0191484834	beta-galactoside alpha- -sialyltransferase 2	357	9.80E-56	47.55%
GT29	CAMPEP_0191466782	cmp-n-acetylneuraminate-poly-alpha- 8-sialyltran sferase	461	1.80E-40	43.35%
GT29	CAMPEP_0191486994	beta-galactoside alpha- -sialyltransferase 2	526	3.60E-29	50.25%
GT29	CAMPEP_0191481596	cmp-n-acetylneuraminate-beta-galactosamide-alpha- -sialyltransferase 1	447	3.20E-55	47.95%
GT29	CAMPEP_0191465100	beta-galactoside alpha- -sialyltransferase 1-like isoform x1	641	1.20E-91	48.30%
GT29	CAMPEP_0191485130	cmp-n-acetylneuraminate-beta-galactosamide-alpha- -sialyltransferase 1	436	3.40E-31	49.55%
GT29	CAMPEP_0191503690	cmp-n-acetylneuraminate-beta-galactosamide-alpha- -sialyltransferase 1	499	1.10E-55	48.50%
GT29	CAMPEP_0191476730	cmp-n-acetylneuraminate-beta-galactosamide-alpha- -sialyltransferase 1	275	4.40E-36	45.20%
GT29	CAMPEP_0191506910	alpha- -sialyltransferase 8b	469	2.80E-36	48.05%
GT29	CAMPEP_0191489574	alpha- -sialyltransferase st3gal i-r2	231	5.50E-72	57.00%
GT29	CAMPEP_0191466008	alpha-n-acetylgalactosaminide alpha- -sialyltransferase 2	535	5.80E-36	47.55%
GT29	CAMPEP_0191481458	beta-galactoside alpha- -sialyltransferase 1	805	1.60E-24	46.20%
GT29	CAMPEP_0191465328	beta-galactoside alpha- -sialyltransferase 2	469	3.30E-128	45.75%

CAZymes families	Sequence name	Sequence description	Length (aa)	min E-value	Identity
GT29	CAMPEP_0191465510	beta-galactoside alpha- -sialyltransferase 2	896	4.90E-135	48.65%
GT29	CAMPEP_0191462906	protein	701	9.00E-170	47.05%
GT29	CAMPEP_0191488628	st3 beta-galactoside alpha- -sialyltransferase 2-like	284	7.60E-53	58.75%
GT29	CAMPEP_0191480270	st3 beta-galactoside alpha- -sialyltransferase 2-like	240	3.50E-73	53.00%
GT29	CAMPEP_0191508608	beta-galactoside alpha- -sialyltransferase 2	516	9.10E-108	50.80%
GT29	CAMPEP_0191472038	beta-galactoside alpha- -sialyltransferase 1	422	1.10E-51	48.70%
GT29	CAMPEP_0191471498	cmp-n-acetylneuraminate-poly-alpha- 8-sialyltran sferase	290	1.00E-47	44.30%
GT29	CAMPEP_0191506406	cmp-n-acetylneuraminate-beta-galactosamide-alpha- -sialyltransferase 1	658	1.20E-39	46.80%
GT29	CAMPEP_0191483520	beta-galactoside alpha- -sialyltransferase 1	400	1.40E-97	47.50%
GT29	CAMPEP_0191475670	cmp-n-acetylneuraminate-beta-galactosamide-alpha- -sialyltransferase 1	689	3.20E-42	46.95%
GT29	CAMPEP_0191506900	beta-galactoside alpha- -sialyltransferase 2-like isoform x1	483	3.00E-52	45.90%
GT29	CAMPEP_0191469014	cmp-n-acetylneuraminate-beta-galactosamide-alpha- -sialyltransferase 1	576	1.30E-38	49.20%
GT29	CAMPEP_0191494410	beta-galactoside alpha- -sialyltransferase 2	511	9.00E-115	46.85%
GT29	CAMPEP_0191465212	cmp-n-acetylneuraminate-beta-galactosamide-alpha- 3-sialyltransferase partial	416	8.80E-26	48.50%
GT29	CAMPEP_0191465882	cmp-n-acetylneuraminate-beta-galactosamide-alpha- -sialyltransferase 1	342	4.10E-19	44.25%
GT29	CAMPEP_0191486838	cmp-n-acetylneuraminate-beta-galactosamide-alpha- -sialyltransferase 1	343	2.60E-55	50.30%
GT29	CAMPEP_0191479916	cmp-n-acetylneuraminate-beta-galactosamide-alpha- -sialyltransferase 1	1006	2.50E-63	46.85%
GT29	CAMPEP_0191462816	cmp-n-acetylneuraminate-beta-galactosamide-alpha- -sialyltransferase 1	513	1.00E-20	45.05%
GT29	CAMPEP_0191505692	cmp-n-acetylneuraminate-beta-galactosamide-alpha- -sialyltransferase 1	690	2.00E-56	45.50%
GT29	CAMPEP_0191486822	cmp-n-acetylneuraminate-beta-galactosamide-alpha- -sialyltransferase 1	316	1.40E-35	50.20%
GT29	CAMPEP_0191470746	cmp-n-acetylneuraminate-beta-galactosamide-alpha- -sialyltransferase 1	451	5.60E-51	49.15%
GT29	CAMPEP_0191472030	cmp-n-acetylneuraminate-beta-galactosamide-alpha- -sialyltransferase 1	705	3.30E-47	50.15%
GT29	CAMPEP_0191477010	cmp-n-acetylneuraminate-beta-galactosamide-alpha- -sialyltransferase 1	619	2.50E-32	46.35%
GT29	CAMPEP_0191477576	cmp-n-acetylneuraminate-beta-galactosamide-alpha- -sialyltransferase 1	516	5.90E-49	48.50%
GT29	CAMPEP_0191503476	cmp-n-acetylneuraminate-beta-galactosamide-alpha- -sialyltransferase 1	585	1.80E-35	44.00%
GT29	CAMPEP_0191462460	cmp-n-acetylneuraminate-poly-alpha- 8-sialyltran sferase	468	8.20E-43	46.10%
GT29	CAMPEP_0191492696	cmp-n-acetylneuraminate-poly-alpha- 8-sialyltran sferase	179	5.70E-26	50.70%
GT29	CAMPEP_0191496004	cmp-n-acetylneuraminate-poly-alpha- 8-sialyltran sferase	465	1.20E-40	47.40%
GT29	CAMPEP_0191495106	cmp-n-acetylneuraminate-poly-alpha- 8-sialyltran sferase	142	1.90E-31	51.55%

CAZymes families	Sequence name	Sequence description	Length (aa)	min E-value	Identity
GT29	CAMPEP_0191495488	cmp-n-acetylneuraminate-poly-alpha- 8-sialyltransferase	154	1.50E-18	48.38%
GT29	CAMPEP_0191495898	sialyltransferase-like protein	395	6.50E-68	45.65%
GT29	CAMPEP_0191478430	alpha- -sialyltransferase st3gal i-r2	569	5.60E-126	51.00%
GT29	CAMPEP_0191471910	beta-galactoside alpha- -sialyltransferase 1	341	4.50E-73	48.86%
GT29	CAMPEP_0191472278	beta-galactoside alpha- -sialyltransferase 1	256	4.40E-27	50.67%
GT29	CAMPEP_0191495610	beta-galactoside alpha- -sialyltransferase 1	392	6.90E-27	44.50%
GT29	CAMPEP_0191473792	beta-galactoside alpha- -sialyltransferase 1	736	4.20E-27	45.25%
GT29	CAMPEP_0191487808	cmp-n-acetylneuraminate-beta-galactosamide-alpha- -sialyltransferase 1	478	8.90E-23	46.05%
GT29	CAMPEP_0191506066	cmp-n-acetylneuraminate-beta-galactosamide-alpha- -sialyltransferase 1	373	1.80E-71	46.71%
GT29	CAMPEP_0191476522	cmp-n-acetylneuraminate-beta-galactosamide-alpha- -sialyltransferase 1	212	1.20E-18	49.67%
GT29	CAMPEP_0191483276	cmp-n-acetylneuraminate-beta-galactosamide-alpha- -sialyltransferase 1 isoform x1	430	3.40E-42	47.30%
GT29	CAMPEP_0191483084	protein	573	1.70E-82	48.25%
GT29	CAMPEP_0191466490	sialyltransferase	210	5.80E-11	54.00%
GT29	CAMPEP_0191487264	sialyltransferase	383	1.40E-17	55.00%
GT29	CAMPEP_0191477490	sialyltransferase	162	4.00E-28	51.50%
GT29	CAMPEP_0191506582	sialyltransferase-like protein	574	1.90E-12	51.00%
GT29	CAMPEP_0191477824	st3 beta-galactoside alpha- -sialyltransferase 2-like	337	6.80E-38	47.00%
GT29	CAMPEP_0191477124	type 2 lactosamine alpha- -sialyltransferase	236	9.70E-68	52.33%
GT29	CAMPEP_0191506186	beta-galactoside alpha- -sialyltransferase 1	816	1.30E-30	45.60%
GT29	CAMPEP_0191484606	beta-galactoside alpha- -sialyltransferase 2	331	8.20E-34	50.65%
GT29	CAMPEP_0191499084	cmp-n-acetylneuraminate-beta- -galactoside alpha- -sialyltransferase-like isoform x1	261	2.10E-19	42.80%
GT29	CAMPEP_0191478920	cmp-n-acetylneuraminate-beta-galactosamide-alpha- -sialyltransferase 1-like	304	3.00E-17	45.85%
GT29	CAMPEP_0191483344	cmp-n-acetylneuraminate-beta-galactosamide-alpha- -sialyltransferase 1	97	2.00E-35	69.10%
GT29	CAMPEP_0191470606	cmp-n-acetylneuraminate-beta-galactosamide-alpha- -sialyltransferase 1	569	2.40E-14	59.45%
GT29	CAMPEP_0191463230	beta-galactoside alpha- -sialyltransferase 2	373	5.90E-19	52.65%
GT29	CAMPEP_0191508666	cmp-n-acetylneuraminate-beta-galactosamide-alpha- -sialyltransferase 2-like isoform x1	757	2.70E-72	50.20%
GT29	CAMPEP_0191486400	beta-galactoside alpha- -sialyltransferase 2	560	2.90E-106	51.90%
GT29	CAMPEP_0191486732	beta-galactoside alpha- -sialyltransferase 1-like isoform x2	599	3.40E-22	59.90%
GT29	CAMPEP_0191508286	cmp-n-acetylneuraminate-beta-galactosamide-alpha- -sialyltransferase 2-like	354	7.50E-89	47.90%
GT29	CAMPEP_0191481002	tenascin-x isoform x1	475	9.20E-53	63.40%
GT29	CAMPEP_0191479124	cmp-n-acetylneuraminate-beta-galactosamide-alpha- -sialyltransferase 1	339	6.60E-37	47.45%
GT29	CAMPEP_0191465804	cmp-n-acetylneuraminate-beta-galactosamide-alpha- -sialyltransferase 1	765	1.40E-11	57.15%
GT29	CAMPEP_0191466594	beta-galactoside alpha- -sialyltransferase 2	245	4.40E-85	47.65%
GT29	CAMPEP_0191467962	cmp-n-acetylneuraminate-beta-galactosamide-alpha- -sialyltransferase 1	266	1.20E-73	51.40%
GT29	CAMPEP_0191505308	beta-galactoside alpha- -sialyltransferase 2	547	8.20E-78	51.35%
GT29	CAMPEP_0191471000	beta-galactoside alpha- -sialyltransferase 2	366	2.80E-71	47.15%

CAZymes families	Sequence name	Sequence description	Length (aa)	min E-value	Identity
GT29	CAMPEP_0191476450	beta-galactoside alpha- -sialyltransferase 2	773	5.20E-91	46.15%
GT29	CAMPEP_0191495540	beta-galactoside alpha- -sialyltransferase 2	522	7.60E-78	48.30%
GT29	CAMPEP_0191476156	beta-galactoside alpha- -sialyltransferase 2	580	2.70E-85	51.95%
GT29	CAMPEP_0191470034	cmp-n-acetylneuraminate-beta-galactosamide-alpha- -sialyltransferase 1	394	2.30E-134	49.85%
GT29	CAMPEP_0191472396	cmp-n-acetylneuraminate-beta-galactosamide-alpha- -sialyltransferase 1	341	5.70E-29	48.74%
GT29	CAMPEP_0191476670	cmp-n-acetylneuraminate-beta-galactosamide-alpha- -sialyltransferase 1	476	2.90E-51	51.40%
GT30	CAMPEP_0191466260	probable 3-deoxy-d-manno-octulosonic acid mitochondrial isoform x1	460	5.20E-77	52.40%
GT31	CAMPEP_0191480260	glycoprotein-n-acetylgalactosamine 3-beta-galactosyltransferase 1-like	450	2.10E-35	42.35%
GT31	CAMPEP_0191479844	predicted protein	285	2.30E-14	46.00%
GT31	CAMPEP_0191479908	beta- -galactosyltransferase 7-like	199	8.90E-58	66.65%
GT32	CAMPEP_0191469992	glycosyltransferase family 32 protein	1002	5.70E-167	51.90%
GT32	CAMPEP_0191464796	lactosylceramide 4-alpha-galactosyltransferase-like	498	3.30E-16	52.75%
GT32	CAMPEP_0191480984	alpha- -galactosyltransferase	374	1.50E-170	59.05%
GT33	CAMPEP_0191462972	chitobiosyldiphosphodolichol beta-mannosyltransferase-like	463	3.50E-113	58.00%
GT34	CAMPEP_0191495664	galactosyl transferase gma12 mnn10 domain protein	472	1.30E-47	49.75%
GT34	CAMPEP_0191483136	galactosyl transferase	335	1.40E-17	43.06%
GT34	CAMPEP_0191484676	conserved unknown protein	407	4.80E-18	44.33%
GT34	CAMPEP_0191474686	glycosyltransferase family 34 protein	427	6.20E-10	41.33%
GT34	CAMPEP_0191462640	galactosyltransferase	397	-	-
GT35	CAMPEP_0191501634	starch phosphorylase	895	0.00E+00	68.90%
GT35	CAMPEP_0191480620	starch phosphorylase	1026	0.00E+00	75.85%
GT41	CAMPEP_0191478520	probable udp-n-acetylglucosamine--peptide n-acetylglucosaminyltransferase sec	979	0.00E+00	60.20%
GT41	CAMPEP_0191467944	probable udp-n-acetylglucosamine--peptide n-acetylglucosaminyltransferase sec	717	1.50E-71	49.80%
GT41	CAMPEP_0191474014	probable udp-n-acetylglucosamine--peptide n-acetylglucosaminyltransferase sec	917	3.00E-106	48.65%
GT41	CAMPEP_0191499074	probable udp-n-acetylglucosamine--peptide n-acetylglucosaminyltransferase spindly isoform x1	356	6.60E-119	72.90%
GT47	CAMPEP_0191506784	exostosin family protein	595	4.20E-33	41.80%
GT47	CAMPEP_0191503426	exostosin-like glycosyltransferase	488	8.80E-46	45.25%
GT47	CAMPEP_0191466934	exostosin family protein	1232	1.60E-51	40.80%
GT47	CAMPEP_0191484264	exostosin family protein, acetylglucosaminyltransferase	416	6.40E-46	46.90%
GT47	CAMPEP_0191464984	exostosin-like glycosyltransferase	497	2.40E-52	43.75%
GT47	CAMPEP_0191471540	exostosin-like glycosyltransferase	747	2.20E-41	41.50%
GT47	CAMPEP_0191482834	exostosin-like glycosyltransferase	449	4.20E-46	45.65%
GT47	CAMPEP_0191508210	exostosin-like glycosyltransferase	672	3.70E-88	44.15%
GT47	CAMPEP_0191470578	exostosin-like glycosyltransferase	848	1.20E-44	50.55%
GT47	CAMPEP_0191472808	exostosin-like glycosyltransferase	537	3.50E-49	42.20%
GT47	CAMPEP_0191473562	exostosin-like glycosyltransferase	762	1.20E-65	41.75%
GT47	CAMPEP_0191504294	exostosin-like glycosyltransferase	590	3.90E-25	43.40%
GT47	CAMPEP_0191494648	glucuronoxylan glucuronosyltransferase	456	6.80E-90	55.35%
GT47	CAMPEP_0191466848	glycosyltransferase family 47 protein	411	1.40E-24	44.80%

CAZymes families	Sequence name	Sequence description	Length (aa)	min E-value	Identity
GT47	CAMPEP_0191481546	exostosin-like glycosyltransferase	293	5.30E-27	48.15%
GT47	CAMPEP_0191479484	exostosin-like glycosyltransferase	501	3.50E-39	43.65%
GT47	CAMPEP_0191462684	exostosin-like glycosyltransferase	395	4.00E-32	57.80%
GT47	CAMPEP_0191475362	exostosin-like glycosyltransferase	497	5.60E-56	43.40%
GT47	CAMPEP_0191500752	exostosin-like glycosyltransferase	304	4.50E-29	46.80%
GT47	CAMPEP_0191467648	exostosin-1-like isoform x1	490	9.80E-37	40.60%
GT47	CAMPEP_0191470976	probable glucuronoxylan glucuronosyltransferase f8h isoform x2	310	2.20E-53	55.20%
GT47	CAMPEP_0191489314	probable glucuronoxylan glucuronosyltransferase irx7	141	1.00E-17	46.40%
GT47	CAMPEP_0191501274	galactosyltransferase-like protein	551	3.80E-16	40.45%
GT47	CAMPEP_0191508824	exostosin family protein isoform 1	569	1.90E-106	54.75%
GT47	CAMPEP_0191506398	exostosin-like glycosyltransferase	525	3.40E-35	42.90%
GT47	CAMPEP_0191474612	exostosin family protein	463	3.30E-37	44.65%
GT47	CAMPEP_0191507502	xyloglucan galactosyltransferase katamari1 homolog	395	1.60E-34	44.95%
GT47	CAMPEP_0191476310	probable beta- -xylosyltransferase irx10	353	4.70E-57	51.10%
GT47	CAMPEP_0191508120	probable beta- -xylosyltransferase irx10	546	1.30E-156	51.70%
GT47	CAMPEP_0191495078	probable glucuronosyltransferase gut1	571	4.50E-118	63.05%
GT47	CAMPEP_0191478050	probable beta- -xylosyltransferase irx10	603	2.00E-154	55.05%
GT47	CAMPEP_0191504172	probable beta- -xylosyltransferase irx10	423	1.90E-150	55.30%
GT49	CAMPEP_0191494174	glycosyltransferase-like protein large2	1825	5.40E-71	50.70%
GT50	CAMPEP_0191471784	gpi mannosyltransferase 1	427	6.00E-130	61.45%
GT50	CAMPEP_0191494004	gpi mannosyltransferase 1	307	5.00E-95	61.30%
GT51	CAMPEP_0191464422	penicillin-binding protein	488	1.10E-179	58.75%
GT54	CAMPEP_0191491604	alpha- -mannosyl-glycoprotein 4-beta-n-acetylglucosaminyltransferase c	360	1.90E-66	54.25%
GT58	CAMPEP_0191496814	dol-p-man:man c -pp-dol alpha- -mannosyltransferase	484	5.80E-123	68.55%
GT60	CAMPEP_0191463200	glycosyltransferase family 60 protein	623	8.70E-116	47.70%
GT60	CAMPEP_0191502412	Glycosyltransferase (GlcNAc)	581	4.80E-74	51.85%
GT60	CAMPEP_0191476576	N-acetylglucosaminyltransferase family protein	698	1.20E-120	52.00%
GT61	CAMPEP_0191507488	beta-( )-xylosyltransferase-like	675	3.40E-43	46.65%
GT64	CAMPEP_0191506908	exostosin-2	471	3.90E-22	48.50%
GT64	CAMPEP_0191493882	glycosyltransferase family 64 protein c4-like	274	1.90E-67	60.40%
GT65	CAMPEP_0191498552	fucosyltransferase	546	2.10E-19	38.60%
GT65	CAMPEP_0191472904	protein	648	1.90E-29	49.75%
GT66	CAMPEP_0191502564	oligosaccharyl transferase-like protein	769	0.00E+00	60.65%
GT66	CAMPEP_0191472868	dolichyl-diphosphooligosaccharide--protein glycosyltransferase subunit stt3b	714	0.00E+00	74.95%
GT68	CAMPEP_0191466776	fucosyltransferase partial	435	-	-
GT68	CAMPEP_0191485632	fucosyltransferase	453	-	-
GT68	CAMPEP_0191473528	fucosyltransferase partial	294	-	-
GT68	CAMPEP_0191505114	unknown protein	927	2.90E-65	66.00%
GT71	CAMPEP_0191480316	glycosyltransferase family 71 protein	441	3.60E-45	48.70%
GT71	CAMPEP_0191464288	glycosyltransferase family 71 protein	334	5.70E-52	54.40%
GT75	CAMPEP_0191469706	PREDICTED: uncharacterized protein LOC107338883	504	1.50E-97	59.05%
GT76	CAMPEP_0191464406	gpi mannosyltransferase 2-like	236	8.50E-44	53.20%
GT77	CAMPEP_0191507582	glycosyltransferase family 77 protein	853	7.80E-104	47.45%
GT77	CAMPEP_0191462762	glycosyltransferase family 77 protein	205	2.10E-18	53.25%
GT77	CAMPEP_0191505962	glycosyltransferase family 77 protein	696	1.10E-69	47.00%

CAZymes families	Sequence name	Sequence description	Length (aa)	min E-value	Identity
GT77	CAMPEP_0191485946	glycosyltransferase family 77 protein	527	6.40E-72	49.40%
GT77	CAMPEP_0191468940	glycosyltransferase family 77 protein	662	3.20E-100	49.75%
GT77	CAMPEP_0191479504	glycosyltransferase cazy family gt77-like protein	510	2.50E-131	62.40%
GT77	CAMPEP_0191484364	glycosyltransferase family 77 protein	785	0.00E+00	58.85%
GT77	CAMPEP_0191466612	glycosyltransferase family 77 protein	701	1.00E-68	42.85%
GT77	CAMPEP_0191483312	glycosyltransferase family 77 protein	643	1.40E-120	56.05%
GT77	CAMPEP_0191473826	glycosyltransferase family 77 protein	535	3.70E-148	57.40%
GT77	CAMPEP_0191467086	glycosyltransferase family 77 protein	401	6.60E-118	58.05%
GT77	CAMPEP_0191478736	arabinosyltransferase xeg113-like	787	0.00E+00	63.35%
GT77	CAMPEP_0191504410	protein	904	1.70E-28	45.30%
GT77	CAMPEP_0191469294	glycosyltransferase family 77 protein	694	2.60E-54	46.00%
GT77	CAMPEP_0191506294	nucleotide-diphospho-sugar transferase	525	1.20E-11	51.60%
GT77	CAMPEP_0191472762	predicted protein	768	3.00E-18	52.25%
GT77	CAMPEP_0191466862	protein	509	1.30E-34	43.00%
GT77	CAMPEP_0191464458	NA	897	-	-
GT81	CAMPEP_0191472990	glycosyl transferase	256	1.50E-62	63.35%
GT90	CAMPEP_0191463696	o-glucosyltransferase rumi homolog	532	8.00E-63	47.65%
GT90	CAMPEP_0191489310	o-glucosyltransferase rumi homolog	506	1.00E-101	45.80%
GT90	CAMPEP_0191471772	hypothetical protein AURANDRAFT_61289	507	2.30E-10	43.00%
GT90	CAMPEP_0191465446	O-glucosyltransferase partial	184	1.80E-17	53.83%
GT90	CAMPEP_0191486202	o-glucosyltransferase rumi-like protein	562	6.00E-137	51.60%
GT90	CAMPEP_0191485316	lipopolysaccharide-modifying enzyme	209	2.30E-27	58.20%
GT90	CAMPEP_0191507108	hypothetical protein	429	4.90E-69	50.85%
GT92	CAMPEP_0191505846	protein	1311	4.00E-43	46.15%
GT95	CAMPEP_0191463778	PREDICTED: uncharacterized protein LOC104826730	521	1.70E-110	68.85%
GT95	CAMPEP_0191471512	protein	515	4.90E-144	68.75%
GT95	CAMPEP_0191471578	hypothetical protein Ctob_000494	702	7.90E-139	56.90%
GT95	CAMPEP_0191493676	hypothetical protein EMIHUDRAFT_434753	558	9.80E-143	56.65%
GT95	CAMPEP_0191470942	protein	337	4.60E-88	58.00%
GT95	CAMPEP_0191472468	protein	497	2.40E-80	55.50%
GT96	CAMPEP_0191484714	protein	624	2.00E-141	57.70%
GT96	CAMPEP_0191482934	protein	630	3.00E-83	59.25%
GT96	CAMPEP_0191493606	protein	695	1.20E-72	44.80%
GH1	CAMPEP_0191505234	beta-glucosidase-like chloroplastic	625	1.50E-157	63.90%
GH2	CAMPEP_0191499494	beta-galactosidase	1153	0.00E+00	54.30%
GH2	CAMPEP_0191502620	beta-galactosidase isoform x1	566	3.40E-104	50.80%
GH3	CAMPEP_0191481108	beta glucosidase	914	0.00E+00	59.30%
GH3	CAMPEP_0191485454	glycoside hydrolase family 3	461	6.20E-103	62.10%
GH13	CAMPEP_0191482168	alpha-amylase isozyme 3c	589	1.90E-155	67.45%
GH13	CAMPEP_0191492276	isoamylase chloroplastic	757	0.00E+00	70.45%
GH13	CAMPEP_0191492516	alpha-amylase	788	0.00E+00	66.30%
GH13	CAMPEP_0191465998	alpha-amylase isozyme 3c	618	2.50E-155	67.45%
GH13	CAMPEP_0191463886	alpha-amylase	601	6.30E-137	57.55%
GH13	CAMPEP_0191503872	alpha-amylase	577	8.90E-75	53.30%
GH13	CAMPEP_0191508254	1,4-alpha-glucan-branching enzyme	461	4.90E-138	62.90%
GH13	CAMPEP_0191495696	alpha-amylase partial	251	1.00E-57	53.20%
GH13	CAMPEP_0191484072	pullulanase chloroplastic-like isoform x1	531	0.00E+00	73.50%
GH14	CAMPEP_0191505954	beta-amylase chloroplastic-like	627	0.00E+00	65.25%
GH18	CAMPEP_0191481372	chitinase domain-containing protein 1	436	1.00E-90	55.80%
GH18	CAMPEP_0191482906	chitinase	291	9.00E-76	47.90%

CAZymes families	Sequence name	Sequence description	Length (aa)	min E-value	Identity
GH20	CAMPEP_0191476350	glycoside hydrolase family 20 protein, N-acetyl-beta-D-glucosaminidase	753	4.30E-88	46.60%
GH23	CAMPEP_0191503196	sgnh hydrolase- partial	161	5.90E-46	68.40%
GH27	CAMPEP_0191466088	alpha-galactosidase	421	8.90E-123	75.60%
GH27	CAMPEP_0191501180	alpha-galactosidase	443	9.80E-123	75.40%
GH28	CAMPEP_0191480306	polygalacturonase pg1	651	1.10E-67	47.55%
GH28	CAMPEP_0191476068	pectin lyase-like superfamily partial	493	1.60E-75	48.40%
GH31	CAMPEP_0191501208	probable glucan -alpha-glucosidase	918	0.00E+00	65.70%
GH31	CAMPEP_0191503670	glycosyl family 31	840	1.20E-169	50.35%
GH31	CAMPEP_0191503980	alpha-glucosidase	792	1.30E-166	54.80%
GH31	CAMPEP_0191476758	alpha-glucosidase	951	0.00E+00	61.55%
GH31	CAMPEP_0191475472	glycoside hydrolase family 31 protein	874	0.00E+00	58.90%
GH32	CAMPEP_0191469190	glycosyl five-bladed beta-propellor domain-containing protein	389	2.00E-86	57.30%
GH33	CAMPEP_0191508014	probable sialidase	518	8.20E-43	48.50%
GH33	CAMPEP_0191508028	glycosyl hydrolase	395	2.40E-20	50.60%
GH33	CAMPEP_0191494858	probable sialidase	635	1.60E-79	43.10%
GH33	CAMPEP_0191462506	bnr repeat-containing glycosyl hydrolase	173	2.20E-26	58.65%
GH33	CAMPEP_0191492844	bnr repeat-containing glycosyl hydrolase	491	2.90E-70	54.20%
GH33	CAMPEP_0191500810	bnr repeat-containing glycosyl hydrolase	512	9.40E-33	55.10%
GH33	CAMPEP_0191472470	glycosyl hydrolase	446	6.10E-74	54.55%
GH33	CAMPEP_0191478036	glycosyl hydrolase	300	6.50E-18	49.65%
GH33	CAMPEP_0191499918	probable sialidase	327	4.90E-85	47.65%
GH33	CAMPEP_0191480588	probable sialidase	489	3.60E-82	46.65%
GH33	CAMPEP_0191506866	probable sialidase	627	1.30E-64	44.80%
GH33	CAMPEP_0191465104	probable sialidase	649	2.20E-126	45.45%
GH33	CAMPEP_0191465430	probable sialidase	479	3.60E-131	45.55%
GH33	CAMPEP_0191463294	probable sialidase	344	1.70E-60	50.00%
GH33	CAMPEP_0191468838	probable sialidase	427	2.30E-75	48.05%
GH33	CAMPEP_0191505830	probable sialidase	802	4.90E-52	45.85%
GH33	CAMPEP_0191481526	probable sialidase	678	2.30E-72	47.95%
GH33	CAMPEP_0191472310	probable sialidase	520	1.20E-83	44.90%
GH33	CAMPEP_0191464106	probable sialidase	315	7.30E-70	48.35%
GH33	CAMPEP_0191464066	probable sialidase	197	1.00E-32	52.60%
GH33	CAMPEP_0191475446	probable sialidase	309	6.60E-84	47.30%
GH33	CAMPEP_0191478548	probable sialidase	433	1.20E-47	48.80%
GH33	CAMPEP_0191502930	sortilin	888	1.10E-142	45.70%
GH33	CAMPEP_0191483358	probable sialidase	371	1.60E-76	50.95%
GH33	CAMPEP_0191470994	probable sialidase	477	7.40E-71	47.45%
GH33	CAMPEP_0191495292	probable sialidase	322	2.40E-44	46.95%
GH33	CAMPEP_0191472932	probable sialidase	561	1.70E-41	41.50%
GH33	CAMPEP_0191477112	probable sialidase	527	2.70E-55	47.30%
GH33	CAMPEP_0191503686	probable sialidase	364	9.90E-47	49.00%
GH33	CAMPEP_0191478162	probable sialidase	428	4.50E-58	44.90%
GH33	CAMPEP_0191475498	probable sialidase	469	2.80E-53	47.30%
GH33	CAMPEP_0191493976	exported exo-alpha-sialidase	471	1.90E-71	50.50%
GH33	CAMPEP_0191468856	neuraminidase	544	6.70E-153	54.80%
GH33	CAMPEP_0191500092	probable sialidase	243	1.80E-59	51.75%
GH33	CAMPEP_0191478714	alpha-rhamnosidase-like protein	421	6.30E-113	61.70%
GH33	CAMPEP_0191486130	exported exo-alpha-sialidase	815	9.60E-74	56.15%
GH33	CAMPEP_0191473540	hypothetical protein POPTR_0018s10610g	381	9.50E-110	67.90%

CAZymes families	Sequence name	Sequence description	Length (aa)	min E-value	Identity
GH36	CAMPEP_0191468216	alpha-galactosidase	658	1.00E-131	52.25%
GH38	CAMPEP_0191481190	glycosyl hydrolase family 38 protein	1176	8.20E-162	68.45%
GH47	CAMPEP_0191477294	mannosyl-oligosaccharide -alpha-mannosidase mns1-like	809	2.40E-121	62.70%
GH47	CAMPEP_0191475492	probable alpha-mannosidase i mns5 isoform x2	331	3.30E-120	72.40%
GH47	CAMPEP_0191483182	mannosyl-oligosaccharide -alpha-mannosidase mns3	509	2.70E-130	61.35%
GH47	CAMPEP_0191480982	mannosyl-oligosaccharide -alpha-mannosidase mns1-like	600	9.10E-180	67.65%
GH51	CAMPEP_0191487994	alpha-l-arabinofuranosidase	835	9.30E-53	41.00%
GH51	CAMPEP_0191491676	alpha-l-arabinofuranosidase	860	1.40E-52	41.00%
GH74	CAMPEP_0191504274	60s ribosomal protein l44	574	1.50E-46	88.05%
GH77	CAMPEP_0191478640	glycoside hydrolase family 77 protein	614	0.00E+00	66.10%
GH79	CAMPEP_0191503944	heparanase-like protein 3	755	1.50E-86	50.40%
GH99	CAMPEP_0191484074	glycoprotein endo-alpha- -mannosidase	239	2.30E-78	64.50%
GH103	CAMPEP_0191489810	lytic transglycosylase	446	6.70E-66	55.95%
GH103	CAMPEP_0191474586	lytic transglycosylase	440	7.20E-66	55.95%
GH103	CAMPEP_0191478210	lytic transglycosylase	112	6.30E-23	72.20%
GH109	CAMPEP_0191492160	oxidoreductase and D-galacturonic acid reductase	419	0.00E+00	69.00%
GH116	CAMPEP_0191505756	non-lysosomal glucosylceramidase-like isoform x1	920	5.70E-66	62.80%
GH127	CAMPEP_0191462882	hypothetical protein	738	1.80E-147	50.40%
GH127	CAMPEP_0191463354	uncharacterized protein	1011	4.10E-179	54.35%
CE1	CAMPEP_0191472458	chlorophyllase- chloroplastic	323	9.10E-44	51.85%
CE1	CAMPEP_0191479654	alpha beta hydrolase	338	8.50E-35	48.10%
CE1	CAMPEP_0191492480	hydrolase	300	2.10E-44	51.05%
CE1	CAMPEP_0191481276	2-hydroxy-6-oxononadienedioate 2-hydroxy-6-oxononatrienedioate hydrolase isoform x2	490	4.60E-64	46.90%
CE1	CAMPEP_0191480626	acyl-protein thioesterase 1	221	2.70E-54	58.10%
CE1	CAMPEP_0191463688	phospholipase carboxylesterase	466	3.60E-56	48.85%
CE1	CAMPEP_0191485492	prolyl oligopeptidase	707	0.00E+00	66.45%
CE1	CAMPEP_0191476470	probable glutamyl chloroplastic isoform x2	503	0.00E+00	72.95%
CE1	CAMPEP_0191494326	protein phosphatase methylesterase 1	353	1.30E-88	61.00%
CE1	CAMPEP_0191464556	s-formylglutathione hydrolase	354	1.60E-118	72.15%
CE1	CAMPEP_0191481642	acyltransferase-like protein chloroplastic	715	8.30E-180	55.90%
CE3	CAMPEP_0191463010	protein	302	1.50E-29	49.15%
CE3	CAMPEP_0191471802	lipolytic protein g-d-s-l family	339	6.60E-17	45.65%
CE3	CAMPEP_0191495476	sgnh hydrolase	356	6.60E-37	49.70%
CE3	CAMPEP_0191477560	platelet-activating factor acetylhydrolase ib subunit gamma	492	4.10E-18	43.55%
CE3	CAMPEP_0191503550	o-antigen related protein	348	9.20E-39	51.25%
CE3	CAMPEP_0191487086	sgnh hydrolase	165	1.60E-14	48.33%
CE7	CAMPEP_0191503516	alpha beta- partial	449	7.90E-88	64.15%
CE7	CAMPEP_0191497594	alpha beta-hydrolases superfamily protein	533	6.80E-61	52.50%
CE7	CAMPEP_0191474108	PREDICTED: uncharacterized protein LOC106391074 isoform X2	757	2.80E-48	59.80%
CE8	CAMPEP_0191504784	f-box only protein 11-like	331	7.50E-39	52.10%
CE10	CAMPEP_0191470146	carboxylesterase	636	1.50E-54	45.40%
CE10	CAMPEP_0191469032	carboxylesterase	635	1.90E-45	44.60%
CE10	CAMPEP_0191462712	peptidase	684	0.00E+00	63.95%
CE10	CAMPEP_0191474166	para-nitrobenzyl esterase	655	7.90E-38	46.90%
CE10	CAMPEP_0191472178	acylamino-acid-releasing enzyme-like isoform x1	220	1.70E-76	75.20%

CAZymes families	Sequence name	Sequence description	Length (aa)	min E-value	Identity
CE11	CAMPEP_0191466984	UDP-3-O-acyl N-acetylglucosamine deacetylase,C-terminal [Ostreococcus tauri]	465	2.80E-85	61.25%
CE12	CAMPEP_0191468582	gdsl esterase lipase at5g45920-like	367	1.20E-57	57.40%
CE12	CAMPEP_0191472526	gdsl esterase lipase at5g62930	255	1.00E-60	57.85%
CE13	CAMPEP_0191506802	palmitoleoyl-protein carboxylesterase notum	348	1.10E-61	45.90%
CE13	CAMPEP_0191486696	pectin acetylesterase 5-like	508	2.60E-67	51.60%
CE13	CAMPEP_0191481718	pectin acetylesterase 5-like	503	1.70E-49	45.40%
CE14	CAMPEP_0191475414	probable n-acetylglucosaminyl-phosphatidylinositol de-n-acetylase	271	6.80E-63	59.50%
PL9	CAMPEP_0191468676	protein	343	1.70E-39	51.30%
CBM20	CAMPEP_0191495466	alpha-glucan water dikinase	1152	6.80E-106	49.30%
CBM20	CAMPEP_0191480820	carbohydrate-binding module family 20 protein	468	1.10E-60	55.95%
CBM20	CAMPEP_0191479434	carbohydrate-binding module family 20 protein	347	1.60E-19	51.65%
CBM20	CAMPEP_0191481524	carbohydrate-binding module family 20 protein	366	2.80E-25	54.15%
CBM20	CAMPEP_0191502054	glycoside hydrolase family 13 protein	786	0.00E+00	57.40%
CBM20	CAMPEP_0191505532	alpha-glucan water dikinase	1353	2.00E-161	61.35%
CBM20	CAMPEP_0191468148	alpha-amylase	497	2.10E-156	58.30%
CBM20	CAMPEP_0191469022	starch binding domain-containing protein	372	1.30E-09	54.95%
CBM20	CAMPEP_0191491012	starch-binding domain-like protein	290	1.30E-09	55.60%
CBM20	CAMPEP_0191474744	kynurenine 3-monooxygenase and related flavoprotein monooxygenases	366	4.70E-13	43.90%
CBM20	CAMPEP_0191503532	alpha-amylase	697	4.70E-13	44.40%
CBM20	CAMPEP_0191497726	starch-binding domain protein	630	5.70E-11	54.60%
CBM20	CAMPEP_0191497884	starch-binding domain-like protein	281	1.40E-09	52.90%
CBM23	CAMPEP_0191478802	-like family protein	659	3.60E-145	54.30%
CBM25	CAMPEP_0191469866	sucrose phosphatase	378	6.40E-97	55.45%
CBM32	CAMPEP_0191473856	anaphase-promoting complex subunit 10	179	1.30E-82	75.75%
CBM32	CAMPEP_0191475388	peptide-n -(n-acetyl-beta-glucosaminyl)asparagine amidase	740	2.50E-112	59.55%
CBM32	CAMPEP_0191465994	peptide-n -(n-acetyl-beta-glucosaminyl)asparagine amidase	587	1.60E-75	63.25%
CBM32	CAMPEP_0191477794	capsular associated protein	462	3.50E-18	58.45%
CBM32	CAMPEP_0191469798	intraflagellar transport protein 25 homolog	221	1.30E-42	67.35%
CBM32	CAMPEP_0191463180	btb poz domain-containing protein at2g30600	658	9.40E-52	60.55%
CBM45	CAMPEP_0191480240	alpha-glucan water dikinase 2	1136	0.00E+00	54.00%
CBM45	CAMPEP_0191486510	alpha-amylase chloroplastic	1157	0.00E+00	77.35%
CBM45	CAMPEP_0191504288	carbohydrate-binding module family 45 protein	1340	0.00E+00	59.40%
CBM45	CAMPEP_0191482936	alpha-glucan water chloroplastic isoform x2	308	6.30E-19	45.60%
CBM45	CAMPEP_0191469916	alpha-glucan water chloroplastic-like	1459	0.00E+00	61.10%
CBM47	CAMPEP_0191469266	protein	1043	1.60E-32	57.25%
CBM48	CAMPEP_0191491654	carbohydrate-binding module family 48 protein	274	1.10E-81	76.00%
CBM48	CAMPEP_0191497806	isoamylase chloroplastic	811	0.00E+00	66.00%
CBM48	CAMPEP_0191498746	isoamylase chloroplastic	1016	0.00E+00	71.25%
CBM48	CAMPEP_0191499298	isoamylase chloroplastic	1016	0.00E+00	71.25%
CBM48	CAMPEP_0191492416	phosphoglucan phosphatase chloroplastic-like isoform x2	412	1.60E-50	54.75%
CBM48	CAMPEP_0191480358	alpha-glucan-branching enzyme chloroplastic amyloplastic-like isoform x1	815	0.00E+00	70.75%
CBM48	CAMPEP_0191484066	isoamylase chloroplastic isoform x2	890	7.00E-98	46.50%
CBM48	CAMPEP_0191466762	starch branching enzyme 4	825	0.00E+00	74.20%
CBM48	CAMPEP_0191501000	alpha catalytic domain-containing protein	711	1.60E-147	56.05%
CBM48	CAMPEP_0191471384	protein	813	1.00E-96	51.80%

CAZymes families	Sequence name	Sequence description	Length (aa)	min E-value	Identity
CBM48	CAMPEP_0191481024	isoamylase chloroplastic	772	0.00E+00	70.30%
CBM50	CAMPEP_0191483960	peptidoglycan-binding protein	444	4.60E-09	55.79%
CBM50	CAMPEP_0191465794	peptidoglycan-binding protein	335	5.80E-27	53.25%
CBM50	CAMPEP_0191508626	beta-lactamase family protein	593	1.40E-29	47.55%
CBM50	CAMPEP_0191508190	peptidoglycan-binding protein	926	1.90E-06	48.00%
CBM50	CAMPEP_0191501478	protein	375	1.10E-25	56.35%
CBM50	CAMPEP_0191480760	peptidoglycan-binding protein	270	4.40E-33	54.95%
CBM50	CAMPEP_0191493836	peptidase m23	343	5.50E-50	55.65%
CBM50	CAMPEP_0191481482	serine threonine protein phosphatase	932	4.60E-50	53.05%
CBM53	CAMPEP_0191495926	alpha-dextrin endo-1	755	2.00E-112	50.45%
CBM53	CAMPEP_0191502748	alpha beta superfamily hydrolase	1178	3.40E-116	52.25%
CBM53	CAMPEP_0191483898	starch synthase chloroplastic amyloplastic-like	335	1.00E-30	47.85%
CBM53	CAMPEP_0191488556	glycosyltransferase family 5 protein	1011	0.00E+00	59.00%
CBM53	CAMPEP_0191504260	soluble starch synthase iii-1	1641	0.00E+00	59.90%
CBM53	CAMPEP_0191469058	starch synthase chloroplastic amyloplastic	831	0.00E+00	63.90%
CBM53	CAMPEP_0191492668	chloroplast post-illumination chlorophyll fluorescence increase protein	176	2.10E-42	60.20%
CBM53	CAMPEP_0191503982	hypothetical protein EUTSA_v10022443mg	275	3.10E-60	59.75%

Table 4. Glycosyltransferase families present in selected prasinophyte, chlorophyte, and streptophytes species. Blue boxes represent presence of the protein family indicate the number of identified sequences.

CAZymes Family	*Pyramimonas parkeae*	*Ostreococcus lucimarinus*	*Ostreococcus tauri*	*Bathycoccus prasinos*	*Micromonas pusilla*	*Micromonas sp.*	*Chlamydomonas reinhardtii*	*Volvox carteri f. nagariensis*	*Chlorella variabilis*	*Monoraphidium neglectum*	*Coccomyxa subellipsoidea*	*Klebsormidium flaccidum*	*Physcomitrella patens*	*Selaginella moellendorffii*	CAZymes Family	*Pyramimonas parkeae*	*Ostreococcus lucimarinus*	*Ostreococcus tauri*	*Bathycoccus prasinos*	*Micromonas pusilla*	*Micromonas sp.*	*Chlamydomonas reinhardtii*	*Volvox carteri f. nagariensis*	*Chlorella variabilis*	*Monoraphidium neglectum*	*Coccomyxa subellipsoidea*	*Klebsormidium flaccidum*	*Physcomitrella patens*	*Selaginella moellendorffii*
GT1	1	3	1	■	1		4	1	8	10	8	9	22	225	GT37							1					4	16	23
GT2*	14	20	15	13	15	17	15	14	16	8	12	31	68	62	GT41	4	83	82	1	96	98	105	115	85	1	2	2	152	211
GT4	16	19	22	13	17	18	18	20	18	15	13	24	41	26	GT43				1	1							2	5	6
GT5	15	12	21	8	18	20	27	21	17	11	8	8	27	10	GT45				2										
GT7	1	1	1		4	2			1						GT47	32	3	3	10	1	3	36	31	20	6	10	23	56	57
GT8	18	2	1	4	1	2	3	3	9	4	3	16	41	33	GT48		2	2	2		3	9	5	3			2	12	16
GT9												1	2	2	GT49	1	2	1	1			9	19	6	3	2			
GT10	4	2	3	6		3	1	1	1	3	4	8	5	3	GT50	2	1	1	1	1	1	1	3	1	1		1	1	1
GT11						1					2	2			GT51	1				1	1						1	2	1
GT12				1											GT54	1				1	2					1	1		
GT13	3				2	4		3	2	1	1	1	6	3	GT57					3	3	3	2	2	2	2	2	2	3
GT14	1				1				8		8	5	26	12	GT58	1	1	1	1	1	1					1	1	1	1
GT15	2						1	3	3		2				GT59							■	1	2	1	1	1	1	1
GT16												2	2	1	GT60	3	1	1	1	2	2	6	3	3	1	3	3		
GT17	1	1			1				1		3	4	2	3	GT61	1						1	1	1		1	18	5	5
GT18	1				1		1		1				1	2	GT62										1			1	
GT19	1									1		3	5	1	GT64	2	1	2	2	1	1	1	2	4			2	5	7
GT20	5	3	2	2	2	2	6	2	2	2	2	2	8	7	GT65	2			1			3	2	5	1	1	2	5	3
GT21	1				2	2	2	2	2	2	1	1	2	3	GT66	2	1	2	1	2	2	2	3	1	2	1	2	3	2
GT22	3	5	6	4	4	4			2		2	3	3	6	GT68	4			2			4	3	1			11	14	15
GT23	10		4	8		3	3	1		1	11	16	1	21	GT69		1					6	12		4	5	6		
GT24	1	2	2	1	2	2		3	2		1	1		2	GT71	2	1	3	1	2	1	6	1	3	4	3	9		
GT25		8	12	6	4	2	2		1		1	8			GT75	1		1				3	4	1	2	3	2	7	9
GT26										1		1			GT76	1	1	1	1	1	1	1	2	1			1	1	2
GT27								1					3		GT77	18	15	15	16	13	20	22	22	26	10	11	12	7	8
GT28	2	2	2	2	4	4	4	4	2	2	3	4	8	5	GT78												1	2	2
GT29	130			71		1						6	6	11	GT81	1	1	2		1	1						1		
GT30	1									1	1	2	1	1	GT83		1	1				1			6		1	2	1
GT31	3		3		1	2	6	3	7	1	8	14	28	41	GT90	7	6	5	10	3	4	36	23	9	2	3	21		22
GT32	3	6	3	1	1	2	6	3	7	6	7	4	6	2	GT92	1		2				6	2	7	1	6	7	17	14
GT33	1	2	2	1	2	1	2	2	1	1	1	1	2	2	GT94												2		
GT34	5	2	7	5	5	2	2	2	1		6	9	24	19	GT95	6			6						1	3	2		
GT35	2	3	6	2	3	3	3	5	5		2	5	5	3	GT96	3			2						5	6	2		

Table 5. Glycoside hydrolase families present in prasinophyte, chlorophyte, and streptophytes species for which complete genomic sequence has been reported. Blue boxes represent the presence of the protein family and number of the identified sequences.

CAZymes Family	*Pyramimonas parkeae*	*Ostreococcus lucimarinus*	*Ostreococcus tauri*	*Bathycoccus prasinos*	*Micromonas pusilla*	*Micromonas sp.*	*Chlamydomonas reinhardtii*	*Volvox carteri f. nagariensis*	*Chlorella variabilis*	*Monoraphidium neglectum*	*Coccomyxa subellipsoidea*	*Klebsormidium flaccidum*	*Physcomitrella patens*	*Selaginella moellendorffii*
GH1	3	1	1				4	9	5	10	8	4	21	71
GH2	5	1	2	1	1	2	3	6	3	5	8	4	4	2
GH3	5				2	1	1		4	1		3	9	14
GH4		1	1	1	1	1			1			1	3	
GH5	5	5	5	4	5	3	9	8	12	18	26	7	20	16
GH8											2	2		
GH9							3	3	6	7	9	11	20	26
GH10								2		2		11	1	4
GH13	31	13	16	8	16	18	14	11	18	7	11	17	20	30
GH14	2	2	5	3	2	2	3	2	5	6	3	5	7	6
GH16		1	1	1		1	6	4	11	3	1	14	39	40
GH17										1		10	32	70
GH18	3	1	1	1	1	1	6	5	6	4	2	2	7	32
GH19									1	1		1	11	34
GH20	1			1				2	3		2	2	2	16
GH23	1											1	2	5
GH24										1	2	2		
GH25												1	1	4
GH27	4						4	4	16		4	5	7	73
GH28	2			1			1	3	1		1	2	15	41
GH29												2	2	27
GH30	2									1	1		1	2
GH31	13	7	7	3	7	6	2	2	8	2	7	9	20	8
GH32	1						7	10	7	6	3	4	12	11
GH33	49			25					6		1	1	4	2
GH35									2		7	2	6	10
GH36	1	1			1	1	1	1	4				5	2
GH37	3	2	2	1	1	2	3	4	3	1	1	1	15	17
GH38	2					3	1	2	1		1	3	5	10
GH39								1						
GH42							3	3	7					
GH43	1								4	1	3	4	10	7
GH44														
GH45									1				1	
GH46									5					
GH47	16	2	3	3	3	3	1	2	6	1	6	4	4	11
GH50													1	
GH51	2			1				1	3		3	2	2	4
GH53													4	
GH55							1		35				4	6

CAZymes Family	*Pyramimonas parkeae*	*Ostreococcus lucimarinus*	*Ostreococcus tauri*	*Bathycoccus prasinos*	*Micromonas pusilla*	*Micromonas sp.*	*Chlamydomonas reinhardtii*	*Volvox carteri f. nagariensis*	*Chlorella variabilis*	*Monoraphidium neglectum*	*Coccomyxa subellipsoidea*	*Klebsormidium flaccidum*	*Physcomitrella patens*	*Selaginella moellendorffii*
GH59													1	
GH63							2	2	1	1		4	7	7
GH71														4
GH72			3											
GH74	1	1	3	2	2	1	2	1	4	1		2	2	2
GH75										1				
GH76								1					1	
GH77	4	2	4	2		2	3	2	3	2	2	2	4	6
GH78							1						6	2
GH79	1									13		2	4	10
GH81							3	2				1	1	4
GH82														1
GH85							1	1	1	1	1	1	1	2
GH87								1						
GH88									2	2			1	
GH89						4				1	1	1	1	2
GH91									4					
GH93														1
GH95													3	2
GH99	2		1				2	1	2		1	1		
GH100												2	8	8
GH101														
GH102														1
GH103	3				1	1		2				1	1	2
GH105										2			7	
GH106														2
GH109	1			1	1	1	2	1	1	1	2	3	4	7
GH110											5			
GH113													2	2
GH114							1	1	1	21				
GH116	1				1	1						1	4	8
GH117			1	1									10	
GH119							1							
GH123													1	2
GH125							1	1	6	2	1		1	
GH127	2										1	1		
GH128											2			
GH130											1	1		
GH131										1				
GH135										5		1		

Table 6. Carbohydrate Esterases (CEs), Polysaccharide Lyases (PLs), and carbohydrate binding modules (CBMs) families present in prasinophyte, chlorophyte, and streptophytes species for which complete genomic sequence has been reported. Blue boxes represent presence of the protein family and number of the identified sequences.

CAZymes Family	*Pyramimonas parkeae*	*Ostreococcus lucimarinus*	*Ostreococcus tauri*	*Bathycoccus prasinos*	*Micromonas pusilla*	*Micromonas sp.*	*Chlamydomonas reinhardtii*	*Volvox carteri f. nagariensis*	*Chlorella variabilis*	*Monoraphidium neglectum*	*Coccomyxa subellipsoidea*	*Klebsormidium flaccidum*	*Physcomitrella patens*	*Selaginella moellendorffii*		CAZymes Family	*Pyramimonas parkeae*	*Ostreococcus lucimarinus*	*Ostreococcus tauri*	*Bathycoccus prasinos*	*Micromonas pusilla*	*Micromonas sp.*	*Chlamydomonas reinhardtii*	*Volvox carteri f. nagariensis*	*Chlorella variabilis*	*Monoraphidium neglectum*	*Coccomyxa subellipsoidea*	*Klebsormidium flaccidum*	*Physcomitrella patens*	*Selaginella moellendorffii*
CE1	22	18	11	13	17	29	21	21	22	16	13	20	42	57		CBM4								1	1	5	1	1	1	2
CE2		1	1								2					CBM6												5	8	
CE3	6	2	2		1	3	1		10	3	10	1				CMB8										1				
CE4									25	1	1	2	1	1		CBM13							1	9		1		2		9
CE5							1	1	1	1	3	2	5	2		CBM14							6	1	3	2		8		
CE6							1	1			1		1	1		CMB16											1	1		
CE7	3	1	1	1	4	3	3	4	2	1	1	4	9	19		CBM18							5	3	2				10	23
CE8	1	1		1					1			2	47	44		CBM20	16	5	9	9	12	16	17	21	20	19	9	6	12	12
CE9		1	2			4	4	6	6				15	17		CBM21							1	1						
CE10	5	7	7	2	17	15	12	17	9	12	7	10	37	84		CBM22									1			17	7	10
CE11	1	1	2	1	1	2	2	2	1	1	1	1	1	2		CBM23	1	2	2				1			2		2		4
CE12	2	2	1	2	1	1	1	1	1	5	1	2		8		CBM25	1						1	1	3	1	1	1	3	
CE13	4	1	2	2							1	1	1	12		CBM32	6		3	2	1	4	7	3	1	2	2	12	5	28
CE14	1	1	1	1		1	1	1	2	1	1	1	1			CMB37													1	
CE15									1					2		CBM40				2	1	1			1			1		5
CE16							2	2			1	11	47	184		CBM41		1	1	1			1					1	1	
PL1												1	39	35		CBM42							1	1				1		3
PL4												1	2	11		CBM43						1						8	23	58
PL6			1													CBM45	5	5	6	6	5	3	3	4	3	5	4	4	4	5
PL7														2		CBM47	1	5	8		4	2	16	41	6	1	1	3		
PL8						1										CBM48	14	8	7	5	7	7	9	10	5	6	7	11	9	18
PL9	1	2	3				3	2				2				CMB49												4	4	6
PL10														2		CBM50	8	4	4	3	4	4	6	17	29	4	4	25	29	30
PL11					1	1		1	1				1	2		CMB52												2		
PL12													3			CBM53	11	8	14	7	8	6	15	11	6	7	5	5	3	3
PL14									1							CMB55										1				
PL18				1												CBM57											10	36	12	23
PL22		1							2				1	5		CBM61												1		
CBM1	1	2	2	3	4	3				1						CBM63										1	4	2		
CBM2									3		2	3				CBM67										4		1		

Table 7. *P. parkeae* protein sequences inferred from *P. parkeae* CCMP726 transcriptome that were classified as GT2 and their homologous sequences in the genome of *P. parkeae* NIES254.

			BLAST best match				
Family	Sequence name	Length	Sequence description	Accession #	Organism	% identity	E-value
GT2	CAMPEP_0191463744	439 aa	glycosyl transferase	XM_007509908.1	*Bathycoccus prasinos*	32%	5E-58
GT2	CAMPEP_0191465826	212 aa	glycosyl transferase	XM_002504376.1	*Micromonas* sp.	47%	5E-54
GT2	CAMPEP_0191472438	230 aa	glycosyl transferase	XM_007510577.1	*Bathycoccus prasinos*	42%	1E-49
GT2	CAMPEP_0191478436	781 aa	glycosyl transferase	FO082277.1	*Bathycoccus prasinos*	48%	0E+00
GT2	CAMPEP_0191479288	420 aa	glycosyl transferase	XM_002503075.1	*Micromonas* sp.	36%	3E-53
GT2	CAMPEP_0191486766	403 aa	glycosyl transferase	XM_001761873.1	*Physcomitrella patens*	52%	3E-113
GT2	CAMPEP_0191494686	1018 aa	glycosyl transferase	XM_007509908.1	*Bathycoccus prasinos*	49%	0E+00
GT2	CAMPEP_0191507404	729 aa	glycosyl transferase	XM_002499677.1	*Micromonas* sp.	32%	4E-54
GT2	CAMPEP_0191507432	844 aa	glycosyl transferase	XM_007510577.1	*Bathycoccus prasinos*	49%	0E+00
	contig11047	5342 bp	GT2 family protein	XP_001420211.1	*Ostreococcus lucimarinus*	56%	5E-13
	contig07376	6405 bp	GT2 family protein	XP_002503121.1	*Micromonas commoda*	63%	4E-08
	contig03577	8482 bp	GT2 family protein	GAQ85736.1	*Klebsormidium flaccidum*	56%	2E-17
	contig03004	9029 bp	GT2 family protein	OAE19742.1	*Marchantia polymorpha*	65%	6E-05
	contig07653	6310 bp	GT2 family protein	XP_005839308.1	*Guillardia theta*	51%	1E-22
	contig15484	4409 bp	GT2 family protein	GAQ86323.1	*Klebsormidium flaccidum*	53%	2E-08
	contig19689	3742 bp	GT2 family protein	GAQ86323.1	*Klebsormidium flaccidum*	69%	6E-49
	contig11178	5309 bp	GT2 family protein	GAQ86323.1	*Klebsormidium flaccidum*	64%	2E-18
	contig106715	494 bp	GT2 family protein	XP_007510496.1	*Bathycoccus prasinos*	39%	7E-33
	contig58334	1197 bp	GT2 family protein	XP_007510496.1	*Bathycoccus prasinos*	35%	2E-34
	contig01377	11633 bp	GT2 family protein	XP_007509970.1	*Bathycoccus prasinos*	45%	3E-20
	contig08516	6037 bp	GT2 family protein	XP_007509970.1	*Bathycoccus prasinos*	45%	7E-04
	contig16920	4151 bp	GT2 family protein	XP_007509970.1	*Bathycoccus prasinos*	65%	7E-22
	contig17009	4133 bp	GT2 family protein	XP_007509970.1	*Bathycoccus prasinos*	60%	3E-06
	contig19512	3766 bp	GT2 family protein	XP_007509970.1	*Bathycoccus prasinos*	42%	1E-12
	contig21416	3509 bp	GT2 family protein	GAQ86323.1	*Klebsormidium flaccidum*	46%	8E-09
	contig25820	3005 bp	GT2 family protein	XP_007509970.1	*Bathycoccus prasinos*	53%	2E-22

Table 8. *P. parkeae* protein sequences inferred from *P. parkeae* CCMP726 transcriptome that are homologous to proteins involved in xyloglucan biosynthesis or degradation and their homologous sequences in the genome of *P. parkeae* NIES254.

			BLAST best match				
Family	Sequence name	Length	Sequence description	Accession #	Organism	% identity	E-value
GH1	CAMPEP_0191505234	625 aa	beta-glucosidase-like chloroplastic	XM_001759305.1	*Physcomitrella patens*	46%	4E-159
GH2	CAMPEP_0191499494	1153 aa	beta-galactosidase	XM_002987015.1	*Selaginella moellendorffii*	98%	0E+00
GH2	CAMPEP_0191502620	566 aa	beta-galactosidase	XM_002987015.1	*Selaginella moellendorffii*	99%	0E+00
GH3	CAMPEP_0191481108	914 aa	beta glucosidase	XM_629425.1	*Dictyostelium discoideum*	41%	2E-175
GH31	CAMPEP_0191503980	792 aa	alpha-glucosidase/xylosidase	XM_002963915.1	*Selaginella moellendorffii*	74%	4E-134
GH31	CAMPEP_0191476758	951 aa	alpha-glucosidase/xylosidase	CP001575.1	*Micromonas sp.*	81%	0E+00
GH31	CAMPEP_0191475472	874 aa	glycoside hydrolase family 31 protein	XM_002504869.1	*Micromonas sp.*	48%	0E+00
GT47	CAMPEP_0191506784	595 aa	probable beta-1,4-xylosyltransferase	XM_002949444.1	*Volvox carteri*	26%	2E-16
GT47	CAMPEP_0191476310	353 aa	probable beta- -xylosyltransferase	XM_007513592.1	*Bathycoccus prasinos*	33%	6E-53
GT47	CAMPEP_0191478050	603 aa	beta-1,4-xylosyltransferase	XM_007515042.1	*Bathycoccus prasinos*	63%	4E-156
GT47	CAMPEP_0191504172	423 aa	probable beta- -xylosyltransferase	XM_007513592.1	*Bathycoccus prasinos*	56%	5E-152
GT47	CAMPEP_0191474612	463 aa	xyloglucan galactosyltransferase	XM_018608252.1	*Raphanus sativus*	25%	1E-32
GT47	CAMPEP_0191507502	395 aa	xyloglucan galactosyltransferase	XM_013845602.1	*Brassica napus*	31%	2E-34
	contig01969	10441 bp	exostosin family protein	XP_007514533.1	*Bathycoccus prasinos*	52%	2E-22
	contig03117	8905 bp	GH31 family protein	XP_003080945.1	*Ostreococcus tauri*	42%	5E-04
	contig03292	8738 bp	beta-galactosidase	XP_003056549.1	*Micromonas pusilla*	66%	1E-15
	contig04107	8031 bp	alpha-glucosidase	XP_002507949.1	*Micromonas commoda*	67%	7E-33
	contig00551	14936 bp	exostosin family protein	GAQ92390.1	*Klebsormidium flaccidum*	33%	2E-05
	contig00710	14009 bp	exostosin family protein	KOO29371.1	*Chrysochromulina* sp.	35%	2E-04
	contig01838	10653 bp	exostosin family protein	GAQ87674.1	*Klebsormidium flaccidum*	51%	3E+00
	contig08094	6166 bp	alpha-glucosidase	GAQ84110.1	*Klebsormidium flaccidum*	43%	3E-10
	contig16515	4218 bp	probable glucuronosyltransferase	XP_017185468.1	*Malus domestica*	47%	4E-26
	contig17374	4079 bp	alpha-glucosidase	XP_003080945.1	*Ostreococcus tauri*	42%	4E-13
	contig09383	5785 bp	exostosin family protein	XP_007515104.1	*Bathycoccus prasinos*	62%	1E-11
	contig09441	5767 bp	exostosin family protein	XP_001755886.1	*Physcomitrella patens*	53%	2E-07
	contig10008	5603 bp	exostosin family protein	GAQ87674.1	*Klebsormidium flaccidum*	59%	7E-12
	contig11469	5229 bp	alpha-glucosidase	XP_003060696.1	*Micromonas pusilla*	84%	3E-30
	contig21029	3558 bp	exostosin family protein	XP_007515711.1	*Bathycoccus prasinos*	55%	3E-24
	contig27612	2833 bp	xyloglucan-specific galacturonosyltransferase	XP_006477521.1	*Citrus sinensis*	31%	5E-02
	contig28126	2784 bp	beta-galactosidase	XP_002987061.1	*Selaginella moellendorffii*	52%	3E-06
	contig32014	2455 bp	beta-galactosidase	GAQ88227.1	*Klebsormidium flaccidum*	57%	9E-19
	contig33888	2316 bp	beta-glucosidase, chloroplastic	XP_006337995.1	*Solanum tuberosum*	47%	4E-05
	contig32132	2449 bp	beta-galactosidase	XP_002987061.1	*Selaginella moellendorffii*	39%	2E-02
	contig13755	4742 bp	alpha-xylosidase-like	XP_012828156.1	*Erythranthe guttata*	52%	9E-09
	contig15531	4933 bp	beta-galactosidase	KYP52600.1	*Cajanus cajan*	40%	1E-09
	contig15764	4351 bp	beta-glucosidase	AAF23823.1	*Arabidopsis thaliana*	54%	4E-09
	contig15771	4350 bp	xyloglucan galactosyltransferase	XP_011654379.1	*Cucumis sativus*	37%	1E-11
	contig18865	3869 bp	putative beta-galactosidase	XP_003080455.1	*Ostreococcus tauri*	63%	9E-13
	contig20245	3664 bp	exostosin family protein	XP_007515104.1	*Bathycoccus prasinos*	26%	2E-23
	contig58236	1199 bp	alpha-glucosidase	XP_002504915.1	*Micromonas commoda*	59%	4E-10
	contig95904	575 bp	exostosin family protein	XP_007515104.1	*Bathycoccus prasinos*	64%	2E-24

Table 9 *P. parkeae* protein sequences inferred from *P. parkeae* CCMP726 transcriptome that are homologous to proteins involved in pectin biosynthesis or degradation and their homologous sequences in the genome of *P. parkeae* NIES254.

			BLAST best match				
Family	Sequence name	Length	Sequence description	Accession #	Organism	% identity	E-value
GT8	CAMPEP_0191477662	353 aa	probable galacturonosyltransferase-like 4	XM_009386711.1	*Musa acuminata*	47%	4E-83
GT8	CAMPEP_0191464514	342 aa	probable galacturonosyltransferase-like 9	AK321169.1	*Solanum lycopersicum*	27%	2E-17
GT8	CAMPEP_0191493730	327 aa	probable galacturonosyltransferase-like 10	BT011750.1	*Arabidopsis thaliana*	44%	1E-83
GT8	CAMPEP_0191507902	522 aa	probable galacturonosyltransferase	XM_010045536.2	*Eucalyptus grandis*	30%	4E-16
GH28	CAMPEP_0191480306	651 aa	polygalacturonase	XM_009036569.1	*Aureococcus anophagefferens*	34%	6E-58
GH28	CAMPEP_0191476068	493 aa	pectin lyase	XM_007515035.1	*Bathycoccus prasinos*	83%	5E-65
CE13	CAMPEP_0191486696	508 aa	pectin acetylesterase	XM_002532286.2	*Ricinus communis*	32%	3E-44
CE13	CAMPEP_0191481718	503 aa	pectin acetylesterase	NM_111775.4	*Arabidopsis thaliana*	28%	3E-36
	contig01314	11803 bp	GT8 protein	XP_001695484.1	*Chlamydomonas reinhardtii*	29%	2E-04
	contig00492	15382 bp	Putative galacturonosyltransferase-like 2	KHN12736.1	*Glycine soja*	55%	7E-07
	contig19106	3823 bp	probable galacturonosyltransferase	XP_001756118.1	*Physcomitrella patens*	39%	1E+00
	contig20021	3696 pb	pectin acetylesterase 7-like	XP_015170082.1	*Solanum tuberosum*	43%	7E+00
	contig24607	3132 bp	pectin acetylesterase 5-like	BAF27156.1	*Oryza sativa*	48%	6E-09
	contig24684	3122 bp	putative polygalacturonase	XP_007515097.1	*Bathycoccus prasinos*	37%	3E-20
	contig26346	2954 bp	pectinacetylesterase precursor-like protein	CAB71866.1	*Arabidopsis thaliana*	47%	6E-03
	contig33306	2357 bp	probable galacturonosyltransferase-like 4	XP_009384986.1	*Musa acuminata*	44%	1E-28
	contig37112	1980 bp	polygalacturonase	XP_008221862.1	*Prunus mume*	38%	6E+00
	contig47417	1570 bp	probable polygalacturonase	XP_008660018.	*Zea mays*	50%	1E-04
	contig70998	902 bp	Pectinacetylesterase family protein	GAQ92641.1	*Klebsormidium flaccidum*	36%	2E-01
	contig81321	736 bp	pectin acetylesterase	XP_003547731.1	*Glycine max*	46%	3E-03

Table 10 *P. parkeae* protein sequences inferred from *P. parkeae* CCMP726 transcriptome that are homologous to proteins involved in starch biosynthesis or degradation and their homologous sequences in the genome of *P. parkeae* NIES254.

			BLAST best match				
Family	Sequence name	Length	Sequence description	Accession #	Organism	% identity	E-value
GH13	CAMPEP_0191482168	589 aa	alpha-amylase	XM_005649399.1	*Coccomyxa subellipsoidea*	54%	2E-153
GH13	CAMPEP_0191492276	757 aa	isoamylase	NM_001288291.1	*Solanum tuberosum*	54%	0E+00
GH13	CAMPEP_0191492516	788 aa	alpha-amylase	XM_011398487.1	*Auxenochlorella protothecoides*	50%	4E-175
GH13	CAMPEP_0191465998	618 aa	alpha-amylase	XM_005649399.1	*Coccomyxa subellipsoidea*	54%	2E-153
GH13	CAMPEP_0191463886	601 aa	alpha-amylase	XM_002506321.1	*Micromonas* sp.	48%	2E-133
GH13	CAMPEP_0191503872	577 aa	alpha-amylase	XM_002980809.1	*Selaginella moellendorffii*	40%	9E-75
GH13	CAMPEP_0191508254	461 aa	1,4-alpha-glucan-branching enzyme	XM_003058527.1	*Micromonas pusilla*	51%	2E-139
GH13	CAMPEP_0191495696	251 aa	alpha-amylase	XM_002502864.1	*Micromonas* sp.	44%	2E-51
GH13	CAMPEP_0191484072	531 aa	pullulanase	XM_017750943.1	*Gossypium arboreum*	60%	0E+00
GH14	CAMPEP_0191505954	627 aa	beta-amylase	XM_001416933.1	*Ostreococcus lucimarinus*	69%	0E+00
GT5	contig04676	7619 bp	soluble starch synthase (partial)	XP_007513366.1	*Bathycoccus prasinos*	55%	5E-12
GT5	contig05032	7409 bp	starch synthase	XP_016434944.1	*Nicotiana tabacum*	64%	6E-30
GT5	contig08440	6056 bp	starch synthase	XP_016726615.1	*Gossypium hirsutum*	81%	8E-31
GT5	contig11430	5239 bp	starch synthase	XP_013894739.1	*Monoraphidium neglectum*	58%	6E-49
GT5	contig17390	4076 bp	granule-bound starch synthase	CEF98412.1	*Ostreococcus tauri*	70%	6E-115
GT5	CAMPEP_0191492284	647 aa	starch synthase	XM_016026277.1	*Sesamum indicum*	48%	0E+00
GT5	CAMPEP_0191482370	652 aa	soluble starch synthase	XM_001784372.1	*Physcomitrella patens*	58%	0E+00
GT5	CAMPEP_0191479448	583 aa	soluble starch synthase	XM_003083680.1	*Ostreococcus tauri*	51%	2E-167
GT5	CAMPEP_0191486440	399 aa	soluble starch synthase	XM_003083680.1	*Ostreococcus tauri*	63%	4E-172
GT5	CAMPEP_0191482422	739 aa	soluble starch synthase	XM_001784372.1	*Physcomitrella patens*	61%	0E+00
GT5	CAMPEP_0191479490	614 aa	granule-bound starch synthase	XM_007513025.1	*Bathycoccus prasinos*	64%	0E+00
GT5	contig03984	8121 bp	starch synthase	ABK80546.1	*Sorghum bicolor*	70%	7E-27
GT5	contig21417	3508 bp	soluble starch synthase	XP_002958598.1	*Chlamydomonas reinhardtii*	70%	1E-20
GT5	contig37096	2103 bp	soluble starch synthase	ACY56214.1	*Oryza sativa*	45%	6E-13
GT5	contig67273	976 bp	Soluble Starch synthase	XP_003083606.1	*Ostreococcus tauri*	76%	1E-42
	CAMPEP_0191471530	508 aa	ADP-glucose pyrophosphorylase	XM_003054939.1	*Micromonas* sp.	67%	0E+00
	CAMPEP_0191464980	513 aa	ADP-glucose pyrophosphorylase	XM_002507503.1	*Micromonas* sp.	73%	0E+00
	contig03318	8705 bp	ADP-glucose pyrophosphorylase	XP_003054985.1	*Micromonas pusilla*	49%	3E-73
	contig101202	352 bp	ADP-glucose pyrophosphorylase	ADQ38248.1	*Zea mays*	57%	3E-68
	contig08573	6018 bp	ADP-glucose pyrophosphorylase	XP_001693447.1	*Chlamydomonas reinhardtii*	75%	2E-34
	CAMPEP_0191480620	1026 aa	starch phosphorylase	XM_003060330.1	*Micromonas pusilla*	65%	0E+00
	CAMPEP_0191501634	895 aa	starch phosphorylase	XM_002505195.1	*Micromonas* sp.	53%	0E+00
	contig00885	13134 bp	isoamylase-type starch debranching enzyme	XP_005644006.1	*Coccomyxa subellipsoidea*	81%	8E-12
	contig01449	11464 bp	alpha-amylase	XP_003060257.1	*Micromonas pusilla*	53%	1E-41
	contig01035	12660 bp	granule-bound starch synthase	XP_010257576.1	*Nelumbo nucifera*	69%	9E-16
	contig02858	9188 bp	alpha-amylase	GAQ88582.1	*Klebsormidium flaccidum*	61%	1E-20
	contig03318	8705 bp	ADP-glucose pyrophosphorylase	XP_005648650.1	*Coccomyxa subellipsoidea*	49%	1E-74
	contig03984	8121 bp	putative starch synthase	ABK80546.1	*Sorghum bicolor*	70%	7E-27
	contig04569	7691 bp	starch branching enzyme	AMP82281.1	*Prunus tomentosa*	71%	3E-46
	contig04624	7655 bp	starch branching enzyme	XP_001416858.1	*Ostreococcus lucimarinus*	82%	9E-31
	contig06265	6807 bp	beta-amylase	XP_003062547.1	*Micromonas pusilla*	51%	2E-22
	contig04908	7479 bp	soluble starch synthase	XP_003083728.1	*Ostreococcus tauri*	48%	6E-17
	contig05032	7409 bp	starch synthase	XP_013894480.1	*Monoraphidium neglectum*	67%	6E-31
	contig06515	6715 bp	alpha-amylase	XP_005649456.1	*Coccomyxa subellipsoidea*	48%	1E-09
	contig06824	6601 bp	isoamylase	BAD89532.1	*Hordeum vulgare*	59%	4E-10
	contig07231	6455 bp	starch phosphorylase	XP_003060376.1	*Micromonas pusilla*	75%	1E-44
	contig07820	6256 bp	ADP-glucose pyrophosphorylase	BAC16096.1	*Oryza sativa*	54%	1E-12
	contig08440	6056 bp	starch synthase	XP_016726615.1	*Gossypium hirsutum*	81%	9E-31
	contig08573	6018 bp	ADP-glucose pyrophosphorylase	XP_005649170.1	*Coccomyxa subellipsoidea*	77%	3E-35
	contig09068	5874 bp	starch branching enzyme	XP_005536101.1	*Cyanidioschyzon merolae*	64%	1E-26
	contig09312	5803 bp	isoamylase	XP_014631549.1	*Glycine max*	71%	9E-31
	contig10017	5601 bp	alpha-amylase	XP_001785820.1	*Physcomitrella patens*	32%	4E-07
	contig10228	5545 bp	soluble starch synthase	XP_002963372.1	*Selaginella moellendorffii*	68%	1E-19
	contig10857	5391 bp	starch synthase	XP_013894739.1	*Monoraphidium neglectum*	51%	5E-22
	contig11430	5239 bp	starch synthase	XP_013894739.1	*Monoraphidium neglectum*	58%	6E-49
	contig12693	4959 bp	alpha-amylase	XP_001418686.1	*Ostreococcus lucimarinus*	61%	2E-25
	contig12980	4902 bp	ADP-glucose pyrophosphorylase	WP_015078891.1	*Anabaena* sp.	45%	5E-26
	contig14566	4580 bp	alpha-1,6-glucosidase, pullulanase-type	XP_013891818.1	*Monoraphidium neglectum*	64%	1E-14
	contig14681	4556 bp	pullulanase type debranching enzyme	ABK63595.1	*Sorghum bicolor*	70%	3E-09
	contig15065	4483 bp	alpha-amylase	GAQ85297.1	*Klebsormidium flaccidum*	51%	2E-10
	contig15272	4444 bp	Pullulanase	XP_001415537.1	*Ostreococcus lucimarinus*	68%	2E-08
	contig17390	4076 bp	granule-bound starch synthase	XP_003079933.1	*Ostreococcus tauri*	70%	9E-115
	contig17490	4057 bp	starch phosphorylase	XP_005651098.1	*Coccomyxa subellipsoidea*	65%	1E-38
	contig21417	3508 bp	soluble starch synthase	XP_001695327.1	*Chlamydomonas reinhardtii*	72%	1E-20
	contig22050	3433 bp	alpha-amylase	XP_001696014.1	*Chlamydomonas reinhardtii*	67%	2E-64
	contig27155	2976 bp	alpha-amylase	XP_016547075.1	*Capsicum annuum*	63%	3E-22
	contig27219	2869 bp	alpha-amylase	XP_002506367.1	*Micromonas commoda*	40%	3E-17
	contig33122	2372 bp	alpha-amylase	XP_001752981.1	*Physcomitrella patens*	44%	2E-18

Table 11 *P. parkeae* protein sequences inferred from *P. parkeae* CCMP726 transcriptome that are homologous to proteins involved in biosynthesis or degradation of trehalose and their homologous sequences in the genome of *P. parkeae* NIES254.

			BLAST best match				
Family	Sequence name	Length	Sequence description	Accession #	Organism	% identity	E-value
GH37	contig17973	3988 bp	probable trehalase	XP_005537255.1	*Cyanidioschyzon merolae*	60%	7E-10
GH37	contig35069	2233 bp	probable trehalase	EMS66963.1	*Triticum urartu*	71%	3E-09
GT20	CAMPEP_0191501708	851 aa	trehalose-phosphate synthase	XM_003056095.1	*Micromonas pusilla*	62%	0E+00
GT20	CAMPEP_0191486514	1043 aa	trehalose-phosphate synthase	AY884150.1	*Ginkgo biloba*	48%	0E+00
GT20	contig22451	3380 bp	trehalose-phosphate synthase	EPS67601.1	*Genlisea aurea*	78%	2E-48
GT20	contig19238	3807 bp	trehalose-phosphate synthase	XP_008353109.1	*Malus domestica*	70%	7E-36
GT20	contig07730	6282 bp	trehalose-phosphate synthase	KXZ45849.1	*Gonium pectorale*	55%	1E-27
GT20	contig39533	1959 bp	trehalose-phosphate synthase	XP_005646483.1	*Coccomyxa subellipsoidea*	84%	7E-35
GT21	contig17688	4029 bp	trehalose-phosphate synthase	XP_013640501.1	*Brassica napus*	41%	4E-17

Table 12 *P. parkeae* protein sequences inferred from *P. parkeae* CCMP726 transcriptome that was classified as GT52 and homologous sequences in the genome of *P. parkeae* NIES254.

			BLAST best match				
Family	Sequence name	Length	Sequence description	Accession #	Organism	% identity	E-value
GT51	CAMPEP_0191464422	488 aa	penicillin-binding protein	XM_003054860.1	*Micromonas pusilla*	55%	6E-170
	contig114366	447 bp	GT51 protein	XP_003054906.1	*Micromonas pusilla*	49%	3E-31
	contig17412	4071 bp	GT51 protein	XP_003054906.1	*Micromonas pusilla*	62%	1E-08
	contig49490	1493 bp	GT51 protein	XP_003054906.1	*Micromonas pusilla*	54%	8E-09
	contig50594	1448 bp	GT51 protein	XP_002966336.1	*Selaginella moellendorffii*	57%	2E-05
	contig63188	1069 bp	GT51 protein	GAQ89555.1	*Klebsormidium flaccidum*	46%	1E-14

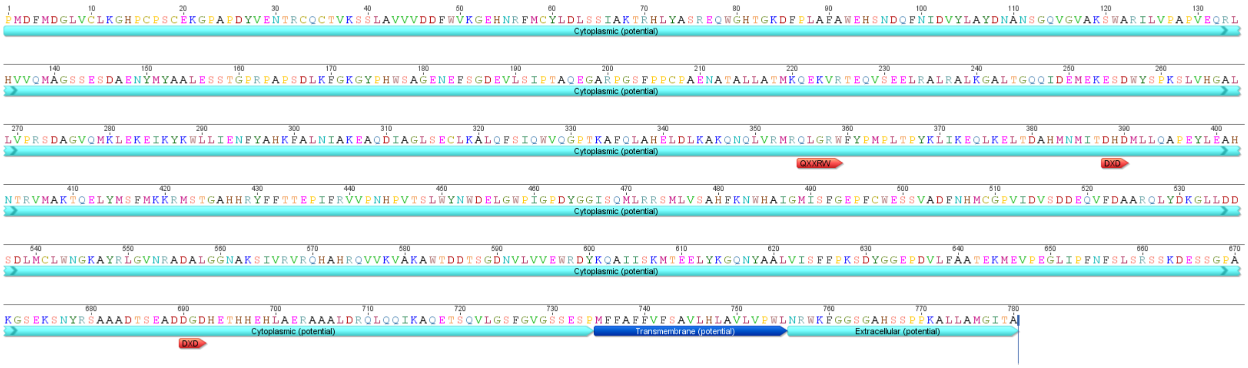

PMDFMDGLVCLKGHPCPSCEKGPAPDYVENTRCQCTVKSSLAVVVDDFWVKGEHNRFMCYLDLSSIAKTRHLYASREQWGHTGKDFPLAFAWEHSNDQFNIDVYLAYDNANSGQVGVAKSWARILVPAPVEQRL
Cytoplasmic (potential)
HVVQMAGSSESDAENYMYAALESSTGPRPAPSDLKFGKGYPHWSAGENEFSGDEVLSIPTAQEGARPGSFPPCPAENATALLATMKQEKVRTEQVSEELRALRALKGALTGQQIDEMEKESDWYSPKSLVHGAL
Cytoplasmic (potential)
LVPRSDAGVQMKLEKEIKYKWLLIENFYAHKFALNIAKEAQDIAGLSECLKALQFSIQWVQGPTKAFQLAHELDLKAKQNQLVRMRQLGRWFYPMPLTPYKLIKEQLKELTDAHMNMITDHDMLLQAPEYLEAH
Cytoplasmic (potential)
QXXRW
DXD
NTRVMAKTQELYMSFMKKRMSTGAHHRYFFTTEPIFRVVPNHPVTSLWYNWDELGWPIGPDYGGISQMLRRSMLVSAHFKNWHAIGMISFGEPFCWESSVADFNHMCGPVIDVSDDEQVFDAARQLYDKGLLDD
Cytoplasmic (potential)
SDLMCLWNGKAYRLGVNRADALGGNAKSIVRVRQHAHRQVVKVAKAWTDDTSGDNVLVVEWRDYKQAIISKMTEELYKGQNYAALVISFFPKSDYGGEPDVLFAATEKMEVPEGLIPFNFSLSRSSKDESSGPA
Cytoplasmic (potential)
KGSEKSNYRSAAADTSEADDGDHETHHEHLAERAAALDRQLQQIKAQETSQVLGSFGVGSSESPMFFAFFVFSAVLHLAVLVPWLNRWKFGGSGAHSSPPKALLAMGITA
Cytoplasmic (potential)
Transmembrane (potential)
Extracellular (potential)
DXD

Figure 1. A *P. parkeae* CCMP726 protein sequence, CAMPEP_0191478436, containing QXXRW domains, Ds residues, and DXD motif known to be catalytic sites of cellulose synthase in bacteria and embryophytes. The red color indicates QXXRW domain and blue color indicated its transmembrane region.

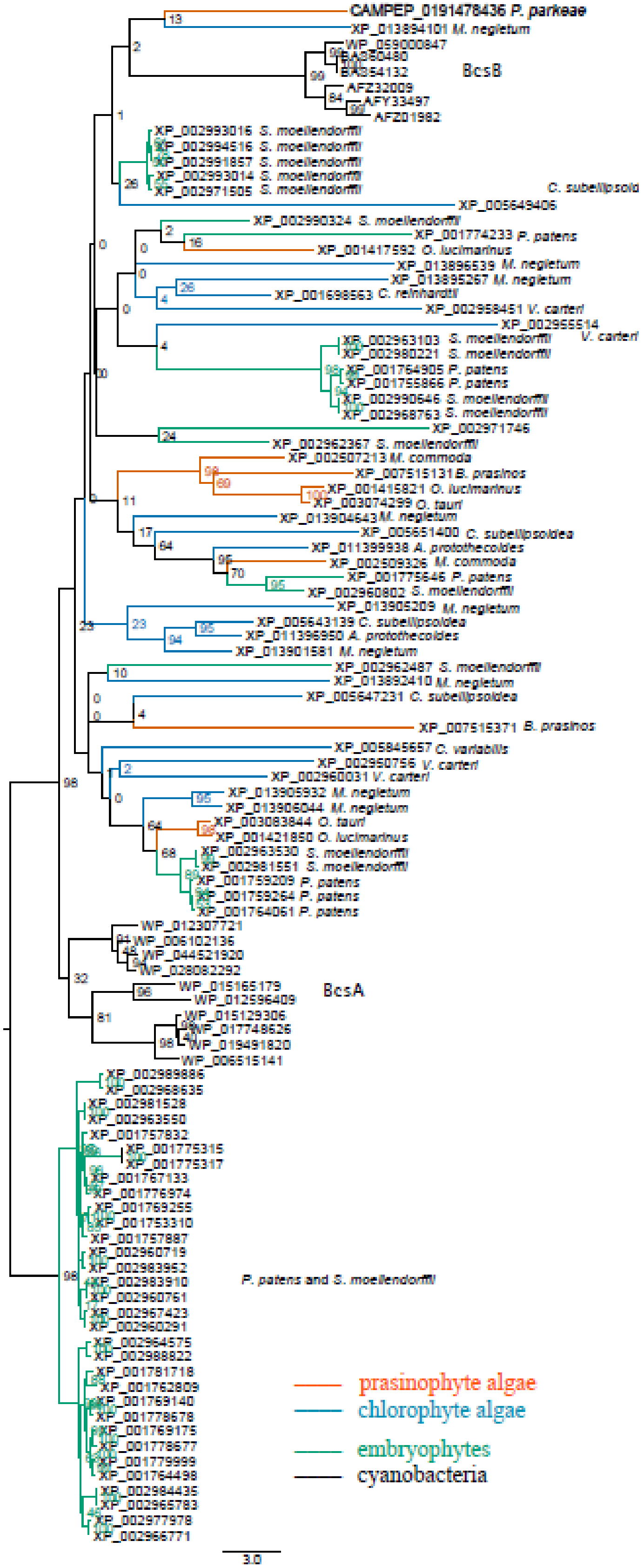
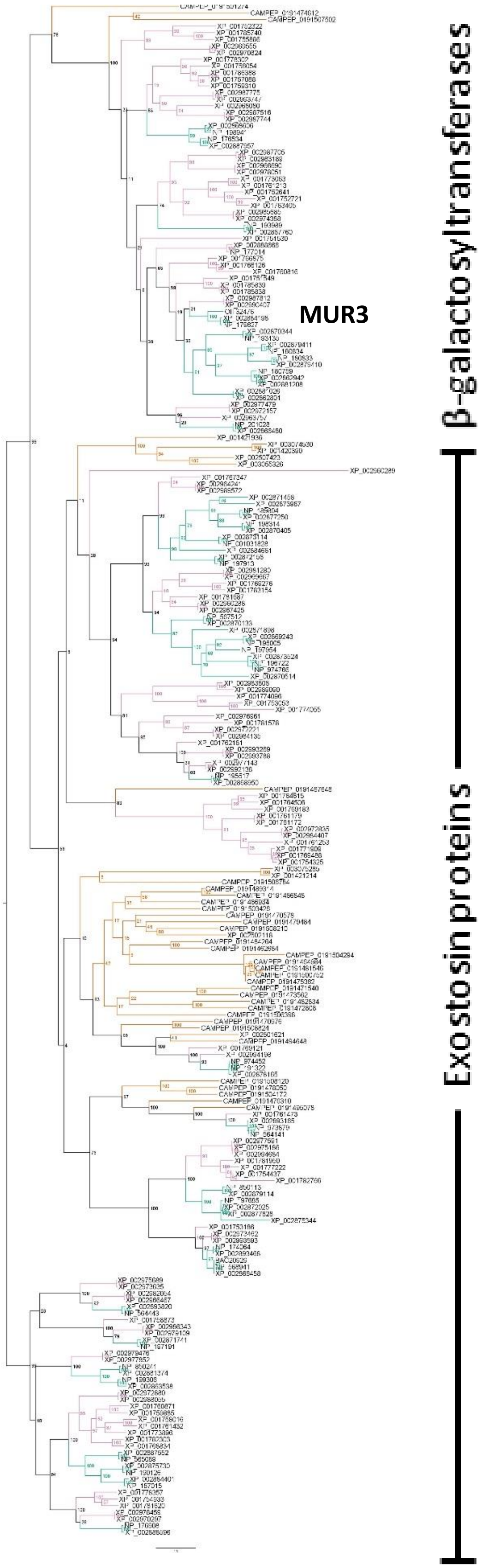

CAMPEP_0191478436 P. parkeae
XP_013894101 M. neglelum
WP_059000847
BAB80480
BAB54132
BcsB
AFZ32009
AFY33497
AFZ01982
XP_002993016 S. moellendorffii
XP_002994516 S. moellendorffii
XP_002991857 S. moellendorffii
XP_002993014 S. moellendorffii
XP_002971505 S. moellendorffii
C. subellipsoid
XP_005649406
XP_002990324 S. moellendorffii
XP_001774233 P. patens
XP_001417592 O. lucimarinus
XP_013896539 M. neglelum
XP_013895267 M. neglelum
XP_001698563 C. reinhardtii
XP_002958451 V. carteri
XP_002955514
V. carteri
XP_002963103 S. moellendorffii
XP_002980221 S. moellendorffii
XP_001764905 P. patens
XP_001755866 P. patens
XP_002990646 S. moellendorffii
XP_002968763 S. moellendorffii
XP_002971746
XP_002962367 S. moellendorffii
XP_002507213 M. commoda
XP_007515131 B. prasinos
XP_001415821 O. lucimarinus
XP_003074299 O. tauri
XP_013904643 M. neglelum
XP_005651400 C. subellipsoidea
XP_011399938 A. protothecoides
XP_002509326 M. commoda
XP_001775646 P. patens
XP_002960802 S. moellendorffii
XP_013905209 M. neglelum
XP_005643139 C. subellipsoidea
XP_011396950 A. protothecoides
XP_013901581 M. neglelum
XP_002962487 S. moellendorffii
XP_013892410 M. neglelum
XP_005647231 C. subellipsoidea
XP_007515371 B. prasinos
XP_005845657 C. variabilis
XP_002950756 V. carteri
XP_002960031 V. carteri
XP_013905932 M. neglelum
XP_013906044 M. neglelum
XP_003083844 O. tauri
XP_001421850 O. lucimarinus
XP_002963530 S. moellendorffii
XP_002981551 S. moellendorffii
XP_001759209 P. patens
XP_001759264 P. patens
XP_001764061 P. patens
WP_012307721
WP_006102136
WP_044521920
WP_028082292
WP_015165179
WP_012596409
BcsA
WP_015129306
WP_017748626
WP_019491820
WP_006515141
XP_002989886
XP_002968635
XP_002981528
XP_002963550
XP_001757832
XP_001775315
XP_001775317
XP_001767133
XP_001776974
XP_001769255
XP_001753310
XP_001757887
XP_002960719
XP_002983962
XP_002983910
P. patens and S. moellendorffii
XP_002960761
XP_002967423
XP_002960291
XP_002964575
XP_002988822
XP_001781718
XP_001762809
XP_001769140
XP_001778678
XP_001769175
XP_001778677
XP_001779999
XP_001764498
XP_002984435
XP_002965783
XP_002977978
XP_002966771
prasinophyte algae
chlorophyte algae
embryophytes
cyanobacteria
3.0

Figure 2. Maximum-Likelihood tree inferred from predicted proteins from completely sequenced prasinophyte algae, chlorophyte algae, embryophytes, and cyanobacteria (Table 1) using an LG+I+G+F amino acid substitution model. The scale bar represents the estimated number of amino acid substitution per site.



Supplementary Figure 1. Maximum-Likelihood tree inferred from GT47 proteins known for the prasinophytes *Pyramimonas parkeae* (CAMPEPs), *Ostreococcus tauri*, *Ostreococcus lucimarinus*, *Micromonas pusilla*, and *Micromonas* sp., the bryophyte *Physcomitrella patens*, the lycophyte *Selaginella moellendorffii*, and the angiosperm *Arabidopsis thaliana*, using a WAG+I+G+F amino acid substitution model. The scale bar represents the estimated number of amino acid substitution per site.

CHAPTER 5: FINAL PERSPECTIVES

The work described in this doctoral thesis focused on organelle and nuclear genomes of *Pyramimonas parkeae*, an early diverging green alga classified into prasinophyte clade I. Recent advances in sequencing technology and computational tools allowed us to perform whole genome sequencing and comparison to genomic data available for other Viridiplantae. Comparative chloroplast genomics are described in Chapter 2, comparative mitochondrial genomics are described in Chapter 3, and comparative genomics related to carbohydrate metabolism are described in Chapter 4. Here, we discuss major findings of this thesis study and limitations that need to be overcome by future work.

Our comparison of two *P. parkeae* chloroplast genomes, one derived from this thesis work and the other previously published for a different strain, showed that the two genomes were identical in protein coding gene content, but varied dramatically in genome size, protein coding sequences, gene copy number, and possible presence of introns. This surprisingly high intra-specific variability raised the question of species delimitation. To assess the relationship between the two studied *P. parkeae* strains, we performed phylogenetic analyses of 3 genes–18S rDNA, 16S rDNA, and *rbcL*–that we obtained from *P. parkeae* strain NIES254 and were available in July 2016 in public databases for other Viridiplantae, which included several additional *P. parkeae* strains. Our results showed that all *P. parkeae* strains used in the analyses resolved a monophyletic clade, but that database limitations did not allow resolution of other relationships, such as prasinophyte species and higher-level diversification patterns, which remain inconclusive. Limited amounts of molecular data also prevented us from resolving such phylogenetic relationships with the use of mitochondrial data. To improve resolution of relationships among *Pyramimonas* species and among prasinophyte clades, more molecular data for these taxa will be needed.

Given observations of high intra-specific variability in the organelle genomes of prasinophytes, comparing mutation rates in *P. parkeae* chloroplast and mitochondrial genomes might have been of interest. However, results reported in this thesis indicated that the mode of evolution of prasinophyte chloroplast and mitochondrial genomes has differed, resulting in differing phylogenetic tree topologies (Chapter 3, figures 6 and 8). In Viridiplantae, chloroplasts and mitochondria are thought to have arisen by endosymbiosis, though obtained at different times and from different endosymbionts. The mitochondrial genome was likely acquired from an ancient alpha-proteobacterial endosymbiont relatively early in the diversification of eukaryotes, while the chloroplast genome was acquired later, and may well have been affected by secondary endosymbiosis (Kim and Maruyama 2014). Therefore, comparative analyses of mutation rate, e.g., the synonymous and non-synonymous substitution rate of nucleotide sequences, might be more complicated to perform and interpret than widely appreciated.

Whole genome sequencing of *P. parkeae* NIES254 using llumina Miseq and Hiseq sequencing methods and comparing the presence of inferred carbohydrate active enzymes to those indicated by other fully sequenced genomes in Viridiplantae revealed that the *P. parkeae* nuclear genome encodes carbohydrate active protein families similar to those previously observed for other prasinophytes, green algae, and early-diverging embryophytes. In particular, *P. parkeae* nuclear-encoded sequences were observed to include those homologous to many other Viridiplantae genes related to biosynthesis and deconstruction of starch and several types of cell wall carbohydrates, indicating molecular traits that evolved early within Viridiplantae.

One complication of this work was fragmentation of the assembled nuclear genome derived from relatively short-read Illumina sequencing. This fragmentation directly affected the completeness of protein coding sequences and posed difficulties in gene annotation and gene family classification. To improve the quality of this work, more *P. parkeae* genomic sequences are needed, and longer reads obtained by other technologies, e.g., PacBio SMRT technology would also be helpful. Transcriptomic sequencing for *P. parkeae* NIES254 would also be useful as it is the direct evidence of RNA production from the same strain we employed in this study~~.~~

Our comparative analyses of CAZyme families of fully sequenced green algal species and selected embryophytes revealed that many common gene families were present. However, taxon sampling limitation precluded determining whether these genes were acquired from vertical or horizontal transfer. To perform such analyses, more whole genome sequencing or at least more sequencing of genes/proteins of interest from green algae – prasinophytes, chlorophyte algae, and streptophyte algae – will be needed.

In addition to insufficient amounts of green algal genomic data, the proteins annotated in fully sequenced green algal genomes were poorly characterized. In this study, we found many sequences that were homologous to protein sequences common to other green algae and streptophytes. Unfortunately, the functions of most *P. parkeae* protein sequences could only be inferred from streptophyte data because annotated proteins for green algae lack sufficient functional prediction. For this reason, a tremendous amount of work in the area of green algal protein characterization will be needed in order to adequately infer functional homology of *P. parkeae* proteins.

Reference

Kim, E. & Maruyama, S. 2014. A contemplation on the secondary origin of green algal and plant plastids. Acta Soc. Bot. Pol. 83(4): 331-336.