UNIVERSITY OF CALIFORNIA, SANTA CRUZ

MAMIELLOPHYCEAE: PHYLOGENETIC AND BIOGEOGRAPHIC INSIGHTS

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

DOCTOR OF PHILOSOPHY

in

OCEAN SCIENCE

by

Melinda P. Simmons

June 2014

The Dissertation of Melinda Simmons is approved by:

_____
Associate Adjunct Professor Alexandra Z. Worden, Chair

_____
Professor Jonathan P. Zehr

_____
Associate Adjunct Professor Steven H. D. Haddock

_____
Professor Raphael Kudela

_____
Tyrus Miller Vice Provost and Dean of Graduate Studies

TABLE OF CONTENTS

Title Page…….……...…….…. i

Table of Contents……..……... iii

List of Tables and Figures……..…… ……..……..……. vi

Abstract……..…….…….……..... viii

Acknowledgments……..…….…….…..... x

1 Chapter 1: Introduction..... 1

1.1 The Importance of Marine Phytoplankton within the Class Mamiellophyceae 1

1.2 Diversity within the Mamiellophyceae 2

1.3 Mamiellophyceae Abundance 5

1.4 Biogeography of Mamiellophyceae 8

1.5 Aims and Objectives 10

2 Chapter 2: Micromonas Genome Architecture..... 16

2.1 Abstract 16

2.2 Introduction 17

2.3 Methods..... 20

2.3.1 Randomized analysis of gene structure 20

2.3.2 Introner element analysis..... 21

2.4 Results 21

2.5 Discussion 26

2.5.1 COP Formation 26
2.5.2 COP Function 27
2.5.3 Introner Elements 31
2.6 Conclusion 32
3 Chapter 3: Prolific but distinct repetitive introns in picoeukaryote species highlight Antarctic Micromonas 47
3.1 Abstract 47
3.2 Introduction 48
3.3 Methods 52
3.3.1 Culturing and nucleic acid extraction 52
3.3.2 PCR, cloning and sequencing 54
3.3.3 Clustering and phylogenetics 55
3.3.4 Metagenome searches 57
3.4 Results 58
3.5 Discussion 68
3.6 Conclusions 75
4 Chapter 4: Biogeography of photosynthetic picoeukaryotes in the North Pacific Ocean 90
4.1 Abstract 90
4.2 Introduction 91
4.3 Materials and Methods 93

4.3.1 Field sampling 93
4.3.2 Flow cytometry 94
4.3.3 QPCR 95
4.3.4 Metatranscriptome library construction and sequencing 96
4.3.5 Metatranscriptome analyses 98
4.3.6 Determining genetic divergence among putative Bathycoccus and Ostreococcus ecotypes 100
4.3.7 Determining Ecomarkers 101
4.3.8 Identifying Ecotype Ecomarkers in Meta-omic Data 102
4.3.9 Analysis of SOD and nitrogen transporter gene families 102
4.4 Results & Discussion 103
4.4.1 Environmental conditions 103
4.4.2 Phytoplankton abundance 104
4.4.3 Establishment of Bathycoccus ecotype genetic distances 109
4.4.4 Diversity assessment based on metatranscriptomic data 110
4.4.5 Exploratory gene expression analyses 113
4.5 Conclusion 118
5 Chapter 5: Conclusions and Perspectives 137

List of Tables and Figures

Table 1.1 12

Table 2.1 34

Table 2.2 34

Table 2.3 35

Table 2.4 35

Table 2.5 36

Figure 2.1 37

Figure 2.2 38

Figure 2.3 39

Figure 2.4 39

Figure 2.5 40

Figure 2.6 41

Table 3.1. 77

Table 3.2. 78

Table 3.3. 79

Figure 3.1 80

Figure 3.2 81

Figure 3.3 82

Table 4.1 119

Table 4.2 120

Table 4.3 122

Figure 4.1 123

Figure 4.2 125

Figure 4.3 127

Figure 4.4 128

MAMIELLOPHYCEAE: PHYLOGENETIC AND BIOGEOGRAPHIC INSIGHTS

By Melinda P. Simmons

Abstract

Marine phytoplankton perform approximately half of global CO_2 uptake. The phytoplankton communities that perform this photosynthesis are diverse and belong to different size classes. In many marine settings picophytoplankton (size) are abundant and among eukaryotic picoplankton the prasinophyte class Mamiellophyceae is particularly well-represented. Three Mamiellophyceae genera found in many marine settings are *Micromonas*, *Ostreococcus* and *Bathycoccus*. This thesis provides analyses of features shared by the Mamiellophyceae, features unique to an individual genus, and overall environmental distributions of the above genera. Specifically, Mamiellophyceae genome features were characterized quantitatively in *Micromonas* (Chapter 2, 3) and led to the detection of repetitive elements that can be used as tools for identifying *Micromonas* clades (Chapter 3). Application of these 'markers' to metagenomic data led to the discovery of *Micromonas* in the Southern Ocean and at salinities and depths it was previously not known to inhabit. In *Ostreococcus* and *Bathycoccus* different types of molecular markers were used to define distances between putative ecotypes within these genera (Chapter 4). This analysis showed two coastal *Bathycoccus* targeted metagenomes belong to the same clade as a cultured species, *Bathycoccus prasinos*, while sequences from a Tropical Atlantic *Bathycoccus* belong to a separate clade. Using this comprehension of

Bathycoccus and *Ostreococcus* diversity, ecomarker genes were selected and used to analyze ecotypes of these genera in metatranscriptome and metagenome data. These analyses were combined with qPCR, flow cytometry and nutrient measurements (Chapter 4). The results provided in this thesis improve our knowledge of the diversity and biogeography of members of the Mamiellophyceae class. Collectively, these studies provide a baseline for future comparisons, as speciation and environmental adaptations are ongoing and may increase with global climate change.

Acknowledgments

I would first like to thank my PhD advisor Alexandra Z. Worden for her support and guidance during my time in her laboratory with additional thanks to my dissertation reading committee- Steven H. D. Haddock, Raphael Kudela and Jonathan P. Zehr. I would also like to express my deep gratitude to the Monterey Bay Aquarium Research Institute staff that made even the toughest days brighter. Much thanks goes to all my friends and lab-mates for their immense encouragement and help through the years. Finally, special thanks to my husband, Dr. Christopher Perle, our daughter, Gloria Perle, and Gloria's grandparents, Ms. Virginia Perle, Mr. Robert Perle and Drs. Richard and Sylvia Simmons, for providing unyielding support and love, without which I could not have completed this work.

This research was financially supported by grants from the Gordon and Betty Moore Foundation and the National Science Foundation. Several collaborators contributed to the work included in chapters three and four of this thesis and will be coauthors on the manuscripts. For chapter three, collaborators include Charles Bachy, Sebastian Sudek, Marijke J. van Baren, Manuel Ares, Jr. and Alexandra Z. Worden, while Sebastian Sudek, Adam Monier, Christopher Perle, Alexander J. Limardo, Valeria Jimenez, J. Timothy Pennington, Marijke J. van Baren, Francisco Chavez and Alexandra Z. Worden contributed to chapter four.

1 Chapter 1: Introduction

1.1 The Importance of Marine Phytoplankton within the Class Mamiellophyceae

Prasinophytes are marine algae composed of seven distinct clades. Clade II consists of the Mamiellophyceae, a diverse class that contains the picoeukaryotic (<2 or 3μm diameter, depending on the research group) genera, as well as genera with larger cell sizes (Marin and Melkonian, 2010). The Mamiellophyceae are widespread in marine systems and many have been cultured (Moreau et al., 2012). Picoeukaryotes as a whole can dominate carbon biomass in oligotrophic waters, despite relatively low cell abundances compared to the cyanobacteria *Prochlorococcus* and *Synechococcus* (Li, 1994; Cuvelier et al., 2010). In the Arabian Sea and equatorial Pacific, picoeukaryotes have made up well over 30% of depth integrated phytoplankton biomass and picophytoplankton biomass, respectively (Shalapyonok et al., 2001; Mackey et al., 2002). In an Eastern Pacific Ocean study, picoeukaryotes were found to be responsible for 76% of the net picoplankton production (Worden et al., 2004). Another study, extending from 49°N and 46°S in the Atlantic Ocean, found that picoeukaryotes often contributed 25 to 60% of total C biomass (Tarran et al., 2006). Even in certain regions of the arctic a large portion of the picophytoplankton population consists of Mamiellophyceae taxa (Not et al., 2005; Lovejoy et al., 2007).

In addition to environmental importance, the Mamiellophyceae (and prasinophytes as a whole) are valuable for evolutionary studies because they help reveal features in the last common ancestor of land plants (streptophytes) and

unicellular green algae (Lewis and McCourt, 2004). Thus, because of both their evolutionary and ecological importance, five Mamiellophyceae taxa have had their genomes sequenced: *Ostreococcus tauri* (Derelle et al., 2006), *Ostreococcus lucimarinus* (Palenik et al., 2007), *Micromonas pusilla* strain CCMP1545 and *Micromonas* sp. RCC299 (Worden et al., 2009) and *Bathycoccus prasinos* (Moreau et al., 2012). Additionally, a genome for *Ostreococcus* isolate RCC809 is publically available and three targeted *Bathycoccus* genomes have been published (Monier et al., 2011; Vaulot et al., 2012). This thesis explores Mamiellophyceae genome characteristics, and investigates a subset of these to develop a better understanding of Mamiellophyceae diversity and biogeography.

1.2 Diversity within the Mamiellophyceae

Genera within the Mamiellophyceae are similar in size and pigmentation, but have structural differences. *Micromonas*, *Ostreococcus* and *Bathycoccus* have simple cellular organizations consisting of a single nucleus, mitochondrion and chloroplast inside a <2 μm diameter cell (Marin and Melkonian, 2010). However, these genera have some large morphological differences. *Micromonas* has a flagella used for rapid swimming, while *Ostreococcus* and *Bathycoccus* are non-motile. *Bathycoccus* is covered in mineralized scales (as are the majority of other known prasinophytes), which are absent from *Micromonas* and *Ostreococcus*.

It is more difficult to detect Mamiellophyceae intra-genus differentiation than inter-genus differentiation. *Micromonas*, *Ostreococcus* and *Bathycoccus* harbor cryptic species, with cells within each genus appearing morphologically identical,

while actually being evolutionarily distinct. In some cases the 18S ribosomal RNA (rRNA) gene can be used to differentiate clades. For example, several distinct clades within the *Micromonas* and *Ostreococcus* genera have been identified and are thought to correspond to species level differences (Guillou *et al*. 2004, Slapeta et al., 2006, Worden 2006, Viprey *et al.* 2008), while *Bathycoccus* is a single clade based on the 18S rRNA gene.

18S rRNA gene phylogenetic analyses of *Micromonas* isolates and environmental sequences typically show five clades (Worden, 2006; Viprey et al., 2008; Worden et al., 2009) (Table 1.1). An early study described only three (Guillou et al., 2004), however, another study amplified and sequenced *Micromonas* genes from the nuclear genome, mitochondrial genome and plastid genome, and found evidence for five clades (Slapeta et al., 2006). Notably, there are variations in the five clade divisions, as Worden (2006) incorporated a group of environmental *Micromonas* sequences, some of which formed an uncultured clade. However, this study could not distinguish two of the cultured clades proposed by Slapeta and colleagues (Table 1.1). In this thesis, a combination of the Slapeta et al., (2006) *Micromonas* clade designations (A-E) and the Worden (2006) clade designations (I-V) will be used, as will additional differentiations resulting from analyses herein.

Studies using the 18S rRNA gene (Guillou et al., 2004; Worden, 2006; Viprey et al., 2008; Worden and Not, 2008) and Internal Transcribed Spacer (ITS) sequences (Rodriguez et al., 2005) support four *Ostreococcus* clades A, B, C and D, of which A

and B have representatives in culture from multiple marine regions. Additional 18S rRNA phylogenies suggest *Ostreococcus* can be categorized as clades OI and OII (Worden and Not, 2008; Worden et al., 2009; Demir-Hilton et al., 2011). In these groupings, clades A and C belong to Clade OI (due in part to grouping driven by a single Clade C sequence) and Clade OII consists of Clade B. Clade D was excluded, due to a lack of sufficient published sequences available during analysis.

In contrast to *Micromonas* and *Ostreococcus*, *Bathycoccus* appeared to have relatively homogeneous genetic diversity (Marin and Melkonian, 2010). However, recent metagenomic studies revealed inserts in the pre-mRNA processing factor 8 (*PRP*8) gene from an oceanic *Bathycoccus* targeted metagenome, which branched separately from coastal *Bathycoccus* and *B. prasinos* Bban7 sequences, despite having 99% 18S rRNA nucleotide identity with *B. prasinos* Bban7 (Monier et al., 2011; Monier et al., 2013). Although the number of geographical sites sampled was limited, the results suggested insert-bearing *Bathycoccus PRP8* sequences were specific to oligotrophic, open ocean waters, while insert-less sequences were from mesotrophic waters. Thus, *Bathycoccus* may have two or more ecotypes (Monier et al., 2013), raising questions about the concept of a single *Bathycoccus* type thriving in a wide range of marine environments (Zhu et al., 2005; Lovejoy et al., 2007; Treusch et al., 2012). .

18S rRNA genes can be too conserved to assess protistan diversity at ecologically and evolutionarily meaningful levels (Piganeau et al. 2011). When Slapeta and colleagues (2006) amplified and sequenced genes from the nucleus,

mitochondria and chloroplast of 17 *Micromonas* strains, they found greater variation than with the 18S rRNA gene. For example, while small-subunit rRNA genes only diverged by 1.5%, *coxI* sequences differed by 16.9%. In another study, two *Micromonas* genomes with 97% 18S rRNA gene similarity were found to share only 90% of their protein encoding genes, based on best reciprocal BLASTP results and additional TBLASTN to ensure "unique" protein encoding genes were truly unique to that strain (Worden et al., 2009). It has been hypothesized the remaining 10% (not detected when used as TBLASTN queries against the other strain) relate to niche differentiation (Worden et al., 2009; McDonald et al., 2010). Analyses of genetic markers and features beyond the 18S rRNA gene are valuable for understanding the Mamiellophyceae class and genera therein. Determining the distributions of these genera (and clades within each) will aid understanding of niche partitioning, speciation and where these taxa contribute most to primary production. To date, few studies have addressed the distributions of *Micromonas*, *Bathycoccus* and *Ostreococcus* using quantitative methods.

1.3 Mamiellophyceae Abundance

Abundance data is limited for Mamiellophyceae taxa in nature. One study used quantitative polymerase chain reaction (qPCR) primers and fluorescent *in situ* hybridization (FISH) to separately quantify *Micromonas*, *Bathycoccus* and *Ostreococcus*, at a coastal Mediterranean Sea site (Zhu et al., 2005). Through this work an annual cycle was discovered where *Micromonas* bloomed in February and *Bathycoccus* bloomed in March. QPCR results, using 18S rRNA gene primers,

showed that each genera made up 12 to 20% of the total picoeukaryotic population while blooming (depending on the year). Using genera-specific FISH probes these estimates increased for *Micromonas* (35 to 65% of the picoeukaryote population), but remained consistent for *Bathycoccus* (12.5 to 27%). QPCR results for both genera were extremely low between June and November, while *Micromonas* was quantifiable by FISH during this same period. Both FISH and qPCR data showed a small increase in *Ostreococcus* abundance during October (0.5% and 3% of picoeukaryotes based on qPCR and FISH, respectively). By either method, *Ostreococcus* counts were extremely low from March through July. Another Mediterranean Sea study using the same qPCR primers and methods as Zhu and colleagues, but on oligotrophic summer samples, reported that *Micromonas*, *Bathycoccus* and *Ostreococcus* maxima were extremely low, at 0.6, 5.5 and 5.3 18S rRNA gene copies ml^{-1}, respectively (Marie et al., 2006). Two copies of the 18S rRNA gene are present in most of the sequenced Mamiellophyceae genomes (Palenik et al., 2007; Demir-Hilton et al., 2011; Moreau et al., 2012).

Published *Micromonas* and *Bathycoccus* qPCR primers and FISH probes do not discriminate between known clades. However, qPCR primers have been developed to discriminate *Ostreococcus* clades OI and OII (Demir-Hilton et al., 2011). Despite conclusions in early culture studies that *Ostreococcus* ecotypes arose through adaptations to varying irradiance levels (Rodríguez et al., 2005), Demir-Hilton *et al*. found distributions indicative of more complex ecophysiological factors impacting the abundance and distribution of *Ostreococcus* clades (Demir-Hilton et al., 2011).

Ostreococcus Clade OI was found to be most abundant (19,555 ± 37 18S rRNA gene copies ml^{-1}) in cool (14 ± 3°C) waters along temperate coasts, while *Ostreococcus* Clade OII was most prevalent (891 ± 107 18S rRNA copies ml^{-1}) in warmer waters (22 ± 3°C) of the open ocean.

Clade specific *Ostreococcus* primers were used with general *Bathycoccus* and *Micromonas* qPCR primers on a limited number of samples from the Bermuda Atlantic Time Series station (BATS) (Treusch et al., 2012). *Ostreococcus* Clade OI was not detected in these samples. During the summer, when a deep chlorophyll maximum formed between 80 and 120 m, *Micromonas* and *Ostreococcus* Clade OII were also undetected, but *Bathycoccus* was present (18S rRNA gene copies reached 111 copies ml^{-1} at 120 m). When the waters at this same station were deeply mixed in winter, the chlorophyll *a* maximum was at 5 m and nutrient and chlorophyll *a* concentrations were generally higher. At this time *Ostreococcus* and *Micromonas* were present at >1,000 18S copies ml^{-1}, while *Bathycoccus* was present at <100 18S copies ml^{-1}. In the spring, following the winter deep mixing, 18S rRNA peak concentrations were located deeper in the water column at 80 m and *Ostreococcus* dominated with values higher than the winter sample, while *Micromonas* and *Bathycoccus* showed <400 18S copies ml^{-1}, (again *Ostreococcus* Clade OI was undetected). These previous quantitative analyses were preliminary steps that helped define distributions of Mamiellophyceae taxa.

1.4 Biogeography of Mamiellophyceae

The Mamiellophyceae as a whole are geographically widespread, but taxa specific studies have shown spatial differentiation within this class (Lovejoy et al., 2007; Foulon et al., 2008; Demir-Hilton et al., 2011). In the aforementioned study by Demir-Hilton et al., *Ostreococcus* clades OI and OII appeared to thrive in different marine environments. For further resolution, the isolation locations of strains within the other *Ostreococcus* classification system of clades A, B, C and D were analyzed revealing the majority of *O*. Clade C strains (representative *O*. *tauri*) were lagoonal, while the majority of *O*. Clade D strains (representative *O*. strain RCC501) were isolated from brackish waters. Therefore, when used on oceanic samples, *Ostreococcus* Clade OI specific qPCR primers mainly target *O*. Clade A (representative *O*. *lucimarinus*), while *O*. Clade OII specific qPCR primers mainly target *O*. Clade B (representative *O*. strain RCC809). The *Ostreococcus* clades each appear to thrive in a different aquatic setting with Clade A in temperate coastal waters, Clade B in warm open ocean waters, Clade C in lagoons and Clade D in brackish waters.

Another Mamiellophyceae biogeography study focused on *Micromonas.* In this study, based on the three-clade classification (Guillou et al., 2004), the *Micromonas* genus was located in all studied regions, including temperate Atlantic, Mediterranean, Indian Ocean and Arctic Ocean waters (Foulon et al., 2008). Clades A, B and C (*sensu* (Guillou et al., 2004)) co-existed year round in locations such as the English Channel (see Table 1.1, for how these designations correspond to clade naming in

other studies). While Clade A was detected at all sampling locations, clades B and C were not, although sampling time of year could not be ruled out as the causative factor in these cases. Clade B was common in warmer more stratified waters, but was oddly detected in Arctic waters as well. This was attributed to the fact that clade B can be split into two sub-clades, based on the five clade classification system. The abundance of clade C was found to be inversely proportional to total *Micromonas* abundance, increasing with distance from shore and with depth. The largest contributions, from Clade C to total *Micromonas* assemblages, were detected in open ocean waters of the Indian Ocean at depth and it was therefore predicted to be a low light-adapted clade. No ecological preferences could be determined for Clade A, as it was the most abundant and ubiquitous. However, similar to the case with Clade B, this clade can be subdivided into different clades, depending on the classification system (Slapeta et al., 2006; Worden et al., 2009), and therefore biogeographies of these potentially distinct clades have not been determined. Foulon and colleagues could not find significant correlations between temperature or chlorophyll *a* and clade contribution(s) to total *Micromonas* abundance. The reasons suggested for this included involvement of other potential environmental parameters, which were not analyzed, and/or the need for finer scale genetic differentiation in their analysis.

Although *Bathycoccus* has been considered a single species genus, a few recent studies have suggested potential ecotypic differentiation for this Mamiellophyceae genus as well (Vaulot et al., 2012; Monier et al., 2013). Monier et al., (2011), recovered *Bathycoccus* intein/intron containing prp8 gene sequences from

oligotrophic, open-ocean waters, while insert-less types were located in more nutrient rich sites or times of year including the Gulfs of Lion, France, Naples, Italy, BATS in spring, mesotrophic Eastern North Pacific waters and North Atlantic Slope waters. *Bathycoccus* targeted metagenomes from Chilean coastal waters (Vaulot et al., 2012) were also found to have insert-less prp8 genes.

The California Current System (CCS) is one region where all three Mamiellophyceae genera and the majority of clades therein have been retrieved (Worden, 2006; Demir-Hilton et al., 2011). Due to the range of environmental conditions encompassed (temperature, nutrients, salinity, productivity) and frequent ship access, this eastern boundary of the North Pacific subtropical gyre is an ideal region in which to research picoplankton. The dynamism resulting from upwelling and oligotrophic ocean influences (Collins et al., 2003) makes the CCS useful for studying evolutionary adaptations to a variety of niches.

1.5 Aims and Objectives

Few studies have investigated distributions or abundances of specific clades within the Mamiellophyceae genera *Micromonas*, *Ostreococcus* and *Bathycoccus*. Here, differentiation between these clades is explored and new measures for identifying different clades or ecotypes are developed. These measures are applied to explore Mamiellophyceae biogeography. The overarching goals of this thesis are to increase the understanding of Mamiellophyceae diversity and *Micromonas*, *Ostreococcus* and *Bathycoccus* distributions in the field. Specific goals are as follows:

1. Highlight differentiation in *Micromonas* clades through comparative genome analyses investigating shared properties and elements (Chapter 2).
2. Reveal presence/absence polymorphisms of *Micromonas* introns in cultured Mamiellophyceae and environmental samples (Chapter 3).
3. Use novel conserved polymorphic intron motifs, known as introner elements (IEs), to characterize the biogeographies of different *Micromonas* clades based on metagenomic data (Chapter 3).
4. Explore picoeukaryotes and distributions of Mamiellophyceae taxa across the CCS using abundance and transcriptomic data (Chapter 4).

Table 1.1 *Micromonas* clade designations as proposed by: (Guillou et al., 2004; Slapeta et al., 2006; Worden, 2006; Viprey et al., 2008)

Micromonas	Clade Designations			
Culture/ Environmental Clone	Guillou *et al*. 2004	Slapeta *et al*. 2006	Worden *et al*. 2006	Viprey *et al*. 2008
CCMP1195	A	C	I	A.BC.1+A.A.2
RCC472	A	B	I	A.BC.1+A.A.2
NEPCC29	A	C	I	A.BC.1+A.A.2
CS222	A	C	I	A.BC.1+A.A.2
RCC451	A	A	II	A.BC.1+A.A.2
RCC299			II	
CCMP1646	B	E	III	B.E.3
CCMP2099	B	E	III	B.E.3
UEPACOp3	B		IV	B. .4
CCMP490	C	D	V	C.D.5
CCMP1545	C	D	V	C.D.5

References

Collins, C.a., Pennington, J.T., Castro, C.G., Rago, T.a., and Chavez, F.P. (2003) The California Current system off Monterey, California: physical and biological coupling. *Deep Sea Research Part II: Topical Studies in Oceanography* **50**: 2389-2404.

Cuvelier, M.L., Allen, A.E., Monier, A., McCrow, J.P., Messié, M., Tringe, S.G. et al. (2010) Targeted metagenomics and ecology of globally important uncultured eukaryotic phytoplankton. *Proceedings of the National Academy of Sciences U S A* **107**: 14679-14684.

Demir-Hilton, E., Sudek, S., Cuvelier, M.L., Gentemann, C., Zehr, J.P., and Worden, A.Z. (2011) Global distribution patterns of distinct clades of the photosynthetic picoeukaryote *Ostreococcus*. *The ISME Journal* **5**: 1095-1107.

Derelle, E., Ferraz, C., Rombauts, S., Rouze, P., Worden, A.Z., Robbens, S. et al. (2006) From the Cover: Genome analysis of the smallest free-living eukaryote *Ostreococcus tauri* unveils many unique features. *Proc Natl Acad Sci U S A* **103**: 11647-11652.

Foulon, E., Not, F., Jalabert, F., Cariou, T., Massana, R., and Simon, N. (2008) Ecological niche partitioning in the picoplanktonic green alga Micromonas pusilla: evidence from environmental surveys using phylogenetic probes. *Environ Microbiol* **10**: 2433-2443.

Guillou, L., Eikrem, W., Chretiennot-Dinet, M., Le Gall, F., Massana, R., Romari, K. et al. (2004) Diversity of picoplanktonic prasinophytes assessed by direct nuclear SSU rDNA sequencing of environmental samples and novel isolates retrieved from oceanic and coastal marine ecosystems. *Protist* **155**: 193-214.

Lewis, L.A., and McCourt, R.M. (2004) Green algae and the origin of land plants. *Am. J. Bot.* **91**: 1535-1556.

Li, W.K.W. (1994) Primary production of prochlorophytes, cyanobacteria, and eukaryotic ultraplankton: Measurements from flow cytometric sorting. *Limnology and Oceanography* **39**: 169-175.

Lovejoy, C., Vincent, W.F., Bonilla, S., Roy, S., Martineau, M.J., Terrado, R. et al. (2007) Distribution, phylogeny, and growth of cold-adapted picoprasinophytes in arctic seas. *Journal of Phycology* **43**: 78-89.

Mackey, D., Blanchot, J., Higgins, H., and Neveux, J. (2002) Phytoplankton abundances and community structure in the equatorial Pacific. *Deep Sea Res Part I* **49**: 2561-2582.

Marie, D., Zhu, F., Balague, V., Ras, J., and Vaulot, D. (2006) Eukaryotic picoplankton communities of the Mediterranean Sea in summer assessed by molecular approaches (DGGE, TTGE, QPCR). *FEMS Microbiol Ecol* **55**: 403-415.

Marin, B., and 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**: 304-336.

McDonald, S.M., Plant, J.N., and Worden, A.Z. (2010) The mixed lineage nature of nitrogen transport and assimilation in marine eukaryotic phytoplankton: a case study of *Micromonas*. *Mol Biol Evol* **27** 2268-2283.

Monier, A., Sudek, S., Fast, N.M., and Worden, A.Z. (2013) Gene invasion in distant eukaryotic lineages: discovery of mutually exclusive genetic elements reveals marine biodiversity. *The ISME journal* **7**: 1764-1774.

Monier, A., Welsh, R.M., Gentemann, C., Weinstock, G., Sodergren, E., Armbrust, E.V. et al. (2011) Phosphate transporters in marine phytoplankton and their viruses: cross-domain commonalities in viral-host gene exchanges. *Environmental Microbiology* **14**: 162-176.

Moreau, H., Verhelst, B., Couloux, A., Derelle, E., Rombauts, S., Grimsley, N. et al. (2012) Gene functionalities and genome structure in Bathycoccus prasinos reflect cellular specializations at the base of the green lineage. *Genome Biol* **13**: R74.

Not, F., Massana, R., Latasa, M., Marie, D., Colson, C., Eikrem, W. et al. (2005) Late summer community composition and abundance of photosynthetic picoeukaryotes in Norwegian and Barents Seas. *Limnology and Oceanography* **50**: 1677-1686.

Palenik, B., Grimwood, J., Aerts, A., Rouze, P., Salamov, A., Putnam, N. et al. (2007) The tiny eukaryote *Ostreococcus* provides genomic insights into the paradox of plankton speciation. *Proc Natl Acad Sci U S A* **104**: 7705-7710.

Piganeau, G., Eyre-Walker, A., Grimsley, N., and Moreau, H. (2011) How and Why DNA Barcodes Underestimate the Diversity of Microbial Eukaryotes. *PLoS ONE* **6**.

Rodriguez, F., Derelle, E., Guillou, L., Le Gall, F., Vaulot, D., and Moreau, H. (2005) Ecotype diversity in the marine picoeukaryote *Ostreococcus* (Chlorophyta, Prasinophyceae). *Environmental Microbiology* **7**: 853-859.

Shalapyonok, A., Olson, R.J., and Shalapyonok, L.S. (2001) Arabian Sea phytoplankton during South West and Northeast Monsoons 1995: composition, size structure and biomass from individual cell properties measured by flow cytometry. *Deep-Sea Research Part II* **48**: 1231-1261.

Slapeta, J., Lopez-Garcia, P., and Moreira, D. (2006) Global dispersal and ancient cryptic species in the smallest marine eukaryotes. *Molecular Biology and Evolution* **23**: 23-29.

Tarran, G.A., Heywood, J.L., and Zubkov, M.V. (2006) Latitudinal changes in the standing stocks of nano- and picoeukaryotic phytoplankton in the Atlantic Ocean. *Deep Sea Research Part II: Topical Studies in Oceanography* **53**: 1516-1529.

Treusch, A.H., Demir-Hilton, E., Vergin, K.L., Worden, A.Z., Carlson, C.a., Donatz, M.G. et al. (2012) Phytoplankton distribution patterns in the northwestern Sargasso Sea revealed by small subunit rRNA genes from plastids. *The ISME journal* **6**: 481-492.

Vaulot, D., Lepere, C., Toulza, E., De la Iglesia, R., Poulain, J., Gaboyer, F. et al. (2012) Metagenomes of the Picoalga Bathycoccus from the Chile Coastal Upwelling. *PLoS ONE* **7**.

Viprey, M., Guillou, L., Ferreol, M., and Vaulot, D. (2008) Wide genetic diversity of picoplanktonic green algae (Chloroplastida) in the Mediterranean Sea uncovered by a phylum-biased PCR approach. *Environmental Microbiology* **10**: 1804-1822.

Worden, A. (2006) Picoeukaryote diversity in coastal waters of the Pacific Ocean. *Aquatic Microbial Ecology* **43**: 165-175.

Worden, A.Z., and Not, F. (2008) Ecology and diversity of picoeukaryotes. In *Microbial Ecology of the Oceans*. Kirchman, D.L. (ed). Hoboken: Wiley, p. 594.

Worden, A.Z., Nolan, J.K., and Palenik, B. (2004) Assessing the dynamics and ecology of marine picophytoplankton: The importance of the eukaryotic component. *Limnology and Oceanography* **49**: 168-179.

Worden, A.Z., Lee, J.H., Mock, T., Rouze, P., Simmons, M.P., Aerts, A.L. et al. (2009) Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes *Micromonas*. *Science* **324**: 268-272.

Zhu, F., Massana, R., Not, F., Marie, D., and Vaulot, D. (2005) Mapping of picoeucaryotes in marine ecosystems with quantitative PCR of the 18S rRNA gene. *FEMS Microbiol Ecol* **52**: 79-92.

2 Chapter 2: *Micromonas* Genome Architecture

2.1 Abstract

The cultivability, small genome size and molecular diversity of the Mamiellophyceae make these marine prasinophytes useful for exploring links between genome architecture and ecology and evolution. Of the five published Mamiellophyceae genomes sequenced to date, all possess a genomic region, atypical in other eukaryotic genomes. This region is characterized by low GC (LGC) content relative to the genome average and less gene synteny than the rest of the genome. In a prior study, published *Ostreococcus* genomes were found to have more transposable elements and convergent overlapping gene pairs (COPs) in this LGC region. In the work herein, a finished *Micromonas* genome sequence (*Micromonas* RCC299) was found to have four times as many COPs in the LGC region compared to the rest of the normal GC (NGC) genome. The LGC region was also found to have lower gene density than its NGC counterpart, which is interesting because previous studies suggested that COPs represent a form of genome compaction. Additionally, LGC COP genes typically have more introns (as do the majority of genes in the LGC region), longer UTRs and higher abundances of expressed sequence tags (ESTs), than non-COP genes. These characteristics appear to also be present in *Micromonas* CCMP1545, although a comprehensive analysis was not performed, due to a greater number of incomplete gene models in this genome. A major issue for gene modeling in *M.* CCMP1545 stemmed from a significant number of repetitive elements, termed "introner elements" (IEs) located in introns of this genome, which do not necessarily have canonical splice sites. IE numbers are almost absent from the *M.* CCMP1545

LGC region and completely absent from the *M.* RCC299 genome. The existence of genomic regions with lower gene density, but higher COP frequency in *M.* RCC299 (and other likely Mamiellophyceae taxa), as well as the prevalence of repetitive non-coding sequences throughout the *M.* CCMP1545 genome are examples of shared (but unusual) and unique aspects of their genomes. Collectively, these shared and strain specific traits demonstrate aspects of the unique genomic landscape of the Mamiellophyceae.

2.2 Introduction

Mamiellophyceae are part of an ecologically successful class of prasinophytes that, along with chlorophyte algae, form a sister clade to land plants. Mamiellophyceae genome sizes are some of the smallest eukaryotic genomes, ranging from 12.6 to 13.3 Mb for *Ostreococcus* (Derelle et al., 2006; Palenik et al., 2007), 20.9 to 21.9 Mb for *Micromonas* (Worden et al., 2009) and 15.1 Mb for *Bathycoccus* (Moreau et al., 2012). These are similar in size to small yeast genomes, such as *Saccharomyces cerevisiae* (12.1 Mb), but smaller than genomes of marine algae such as diatoms (27.4 to 34.5 Mb (Armbrust et al., 2004; Bowler et al., 2008)) and haptophytes (141.7 Mb (Read et al., 2013)). Thus, based on their small cell and genome sizes, the Mamiellophyceae are thought to have relatively compact genomes that show signs of genome reduction.

Genome streamlining, where genome size is reduced relative to that of the ancestral genome, has been proposed frequently for marine taxa. In organisms with small cell sizes, such as specific bacterial taxa that thrive in oligotrophic

environments, it is thought to offset low availability of nutrients and other constituents (Giovannoni et al., 2005). Additionally, larger population sizes, such as those of the marine heterotrophic bacterial clade SAR11 and cyanobacterium *Prochlorococcus*, should reduce the effect of genetic drift and allow natural selection to hinder excess DNA aggregation (Lynch, 2006). These concepts also apply to large protistan populations (Lynch, 2006). For example, *Ostreococcus* has been found to reach concentrations of 3.2×10^5 cells ml^{-1} off the California coast (Countway and Caron, 2006). The Mamiellophyceae have reduced gene numbers (relative to other green algae), gene fusions, reduced chromatin and selenoproteins, which are more catalytically active than similar enzymes lacking selenium and are thought to reduce the need for genes encoding these less effective enzymes (Palenik et al., 2007; Peers and Niyogi, 2008; Worden et al., 2009). All of these features have been considered indicative of genome streamlining, or compaction, in these taxa.

Genomes from three genera, *Ostreococcus*, *Micromonas* and *Bathycoccus*, have been published to-date. These genomes have two outlier chromosomes, one big (BOC) and one small (SOC), which are apparent synapomorphies (shared with a last common ancestor) (Derelle et al., 2006; Palenik et al., 2007; Worden et al., 2009; Moreau et al., 2012). The entire SOC and the majority of the BOC for each sequenced Mamiellophyceae have been found to have lower GC content than the rest of their genomes. For example, in *M.* RCC299 and *M.* CCMP1545 LGC regions were found to be 7% and 8% of the genomes, respectively, and to have 14% lower GC than the rest of the genomes (Worden et al., 2009). Genes from BOC LGC regions are more

highly expressed than genes from elsewhere in the genomes and while these genes are not collinear within the Mamiellophyceae, they tend to cluster in the LGC despite lack of synteny (Palenik et al., 2007; Moreau et al., 2012). In other green algae (*Chlamydomonas*, *Chlorella*, *Volvox*) orthologs of the Mamiellophyceae LGC genes are scattered throughout their genomes (Moreau et al., 2012). After a comparison of two *Ostreococcus* genomes, *O. tauri* and *O. lucimarinus*, it was suggested that the LGC content of the BOC led to enhanced transposon activity, which in turn resulted in decreased gene collinearity (Palenik et al., 2007). The prevalence of transposable elements (TEs) may also explain the presence of overlapping genes reported from this region of the *Ostreococcus* genomes (Palenik et al., 2007). While this seems plausible for *O. tauri* and *O. lucimarinus*, no known TEs were detected in *Bathycoccus* or *M.* RCC299 and only a few TEs were found in *M.* CCMP1545 (Worden et al., 2009; Moreau et al., 2012). Due to the mutational burden of mobile genetic elements (Lynch, 2006), the presence of transposons in *Ostreococcus* is more surprising than the lack in *Bathycoccus* and *Micromonas*,

When first sequenced, *M.* RCC299 was one of very few eukaryotic genomes assembled without gaps, from telomere to telomere. Although *M.* CCMP1545 assembly and gene prediction proved more difficult, with 21 scaffolds representing 19 complete chromosomes, comparative analyses proved revealing. Despite sharing 97% 18S rRNA gene identity, only 90% of predicted protein-encoding genes were shared between the two strains (Worden et al., 2009). Comparisons of *Micromonas* genome architecture revealed both synapomorphic and apomorphic (specialized)

characteristics that raise questions about the extent of genome streamlining in these species which can have very large population sizes (Countway and Caron, 2006).

2.3 Methods

2.3.1 Randomized analysis of gene structure

Gene structures in the LGC and NGC were investigated by first dividing the *M.* RCC299 genome (20,984,628 bp) into 210 equally sized contiguous fragments (160,188 bp). Fragments truncated by the end of the chromosome were combined with the start of the next chromosome to form a complete fragment (Figure 2.1). Subsequently, a pseudo-random sequence of numbers, created in Matlab, was used to select eight genome fragments for analysis in an unbiased manner. Two non-randomly selected fragments, from the LGC region of chromosome one were analyzed using the same approach. All models with support from directionally cloned ESTs (clones were bi-directionally Sanger sequenced) on the ten fragments were manually verified or remodeled according to transcript data, allowing more precise estimates of gene characteristics (Table 2.2) than the automatically predicted gene catalog. Specifically, intron splice junctions were corrected if incorrectly modeled, 5´ and 3´ UTRs were added or extended when necessary, coding sequences (CDSs) were edited (as supported by data) if models included erroneous stop codons (generally due to incorrect intron predictions). For non-expressed models, we collected data for the best model (e.g., those containing start and stop codons rather than purely homology-based models, which frequently lacked start and stop codons) and only remodeled when BLASTP (Altschul et al., 1997) of orthologs provided support to do so. Our

data provide a conservative estimate of gene overlap with only EST supported UTRs included.

2.3.2 Introner element analysis

Once IEs were observed through manual scrutiny of introns, a kmer analysis described in Worden *et al* (2009) was performed by collaborators. Initially detected IE sequences were used as BLASTN (Altschul et al., 1997) queries against the *M.* CCMP1545 genome and retrieved multiple hits (unexpected for introns). Over 6,000 IEs were revealed in the *M.* CCMP1545 genome and most appeared to be within introns (Worden et al., 2009). IEs located within intergenic spaces or on the non-coding strand were more closely examined. It was apparent, based on EST data and/or the presence of stop codons within the CDS of the predicted gene model, that these cases were an artifact of poor modeling and they were then manually remodeled. IE alignments were made using ClustalW (Thompson et al., 1994).

2.4 Results

M. RCC299 gene architectural features, including introns, exons, untranslated regions (UTRs) and gene overlap, were analyzed in detail. Special attention was paid to how these characteristics differed between the 1.5 Mb LGC region and the NGC genome, which has 64.6% GC (excluding the SOC). The genome was divided into 210 contiguous and equal length (160,188 bp) fragments. From this a subset of fragments were selected, eight at random from throughout the NGC, and two specifically from the BOC LGC region (48% GC). Gene models from these fragments were manually curated using Sanger-bidirectional expressed sequence tag

(EST) reads, as well as reciprocal BLASTPs of orthologs to improve gene model predictions (Figure 2.1, Table 2.1).

Genes from the majority of the *M.* RCC299 genome had longer mean intron and exon lengths than genes from the LGC region, although LGC genes were longer on average (Table 2.2). Comparison of the average nucleotides (nt) per gene, consisting of introns, CDS and UTR, to the average nt per transcript, consisting of CDS and UTR, showed the majority of this size difference was due to an increased number of introns per gene in the LGC region, albeit smaller introns. The comparison of arithmetic means of CDS length and transcript length revealed longer UTRs (for genes in the LGC region) also contributed to this difference. Both 5´ and 3´ UTRs were longer in the LGC region, than in the NGC. There may be potential biases in these results since more than twice as many genes from the LGC region had ESTs than did NGC genes (Table 2.3, leading to higher UTR recognition and therefore potentially more COPs observed), or from potential differences in cDNA recovery or cloning efficiencies related to GC content.

Close examination of ESTs showed COP forming genes (for examples, see Figures 2.2, 2.3, 2.5), and UTRs for both COP forming and non-overlapping genes, were frequently miscalled by gene prediction and model prioritization programs. These programs typically exclude the possibility of overlapping genes. Estimates from this work are improvements over those predicted based on gene modeling algorithms, but are likely still under estimates, as overlap generally involved the 3´

UTR of at least one of the COP forming genes (Figures 2.2, 2.3, 2.5) and UTRs could only be predicted when ESTs were present. The analysis of manually corrected gene models on randomly selected genome fragments (and non-randomly selected LGC fragments) revealed that at least 66% and 33% of LGC and NGC region genes, respectively, formed a COP (Table 2.3).

Twice as many *M.* RCC299 LGC genes formed COPs than did NGC region genes and the differences in overlap extent were notable (Table 2.3). On average, the length of overlapping sequence shared between two COP forming genes was 567±641 nt and 44±57 nt, from the LGC and NGC regions respectively. The most extreme overlap lengths observed were >2,000 nt and >350 nt, respectively, for LGC and NGC regions (Figure 2.2A, B).

COPs were not found to have more CDS per gene (e.g., due to longer exons or more exons per gene) than non-overlapping genes, but rather longer total intragenic noncoding sequences, including introns and 3´ UTRs. These common characteristics of COP genes were more pronounced in the LGC region, where the average 3´ UTR length was 468±564 compared to 105±118 for the NGC region. UTR lengths were also shorter for EST-bearing non-overlapping LGC genes, when compared to COP-forming genes from the same region (Table 2.4). However, when non-overlapping genes from the LGC fragments and the NGC fragments were compared the differences in UTR length were minimal. It could not be determined whether some COPs were missed due to insufficient EST coverage to observe UTR overlap, but in

the LGC there was an apparent association between longer 3´ UTRs and an increased frequency of COPs. In the NGC region both COP-forming and non-overlapping genes had almost identical (within 1 nt) average 3´ UTR lengths (Table 2.4).

In addition to shorter UTRs, genes from NGC fragments had reduced intron numbers compared to genes from the LGC region (Table 2.4). In the LGC non-overlapping genes had fewer introns than COPs. However, the opposite was seen when comparing COP and non-overlapping genes from the NGC region.

A quantitative analysis of gene overlap similar to that performed in *M.* RCC299 was not carried out for *M.* CCMP1545, as gene models were generally weaker for this strain. However, a preliminary manual comparison of *M.* CCMP1545 orthologs was performed, specifically those found to form COPs in *M.* RCC299. This analysis resulted in several unusual findings. First, although gene overlap appeared prevalent in *M.* CCMP1545 as well as *M.* RCC299, genes found to form COPs in *M.* RCC299 did not necessarily form a COP in *M.* CCMP1545. Secondly, COP forming genes in *M.* RCC299 sometimes overlapped with different genes in *M.* CCMP1545 (e.g., from Figure 2.2A, The *M.* RCC299 ABHG gene (Joint Genome Institute protein ID number (PID) 113982) had no BLASTP hits to *M.* CCMP1545, while its convergent overlapping pair member DUF339 (PID113604) hit to a conserved uncharacterized protein encoding gene in *M.* CCMP1545 (PID70267) with an E-value of $8.69e^{-40}$. PID70267 also formed a COP, but with a gene encoding yet another predicted protein (PID122667)).

The most striking difference observed, when the gene structure of orthologs in *M.* RCC299 and *M.* CCMP1545 were compared, was the presence of many more introns in *M.* CCMP1545 gene models. These were especially unusual because they were repetitive in the *M.* CCMP1545 genome and could be recovered using a simple BLASTN, similar to transposons, but they lacked known transposon characteristics (e.g., repetitive-intron-flanking exons lacked clear sequence consensus and the repetitive introns did not include known transposase or reverse transcriptase sequences). A significant fraction of *M.* CCMP1545 gene models were found to harbor these repetitive introns, (or IEs), in CDS (Table 2.5) (Worden et al., 2009). Examples were even found where both genes in a *M.* CCMP1545 COP included IEs (Figure 2.3). IEs, in COP forming genes or otherwise, had also contributed to the gene prediction issues in this genome project.

The majority of IEs appeared to be intronic. Those IEs located outside of intron splice sites were manually examined, using directional ESTs as reliable guides for mapping actual CDs sequences, and were often found to belong in introns that were initially incorrectly modeled . Although not all incidences could be inspected and remodeled, the majority appeared to be completely intronic. A subset of cases where IEs were observed on the strand opposite of the CDs were manually investigated. These represented falsehoods in that directionally cloned ESTs clearly demonstrated the IEs should be modeled as an intron in a gene on the same strand as the IE (intron). New models were then created, or more frequently existing models that had not been

prioritized at that locus were prioritized over the prior "catalog" model (on the opposite strand)..

IEs were so prevalent in *M.* CCMP1545 they made up nearly the entire 1 Mb size difference between this genome and that of *M.* RCC299 (where IEs were not detected) (Worden et al., 2009). Some genes formed of a single exon in *M.* RCC299 had multiple IEs in their *M.* CCMP1545 ortholog (Figure 2.4). However, IEs did not appear to cause gene inactivation, since most gene models indicated complete protein sequences were encoded and that IEs did not generate erroneous stop codons.

An exhaustive search of public nucleic acid databases returned no IE hits to public genomes or sequences from cultured organisms. However since relatively few marine algae had been sequenced at the time, metagenomes were also queried. In these, hits were recovered from Sargasso Sea metagenomic data with flanking sequences from Sargasso Sea IEs showing higher identity to *M.* CCMP1545 than *M.* RCC299 (Venter et al., 2004; Worden et al., 2009).

2.5 Discussion

2.5.1 COP Formation

COPs were more frequent in *M.* RCC299 than in sequenced "model" eukaryotes. Thirty seven percent of expressed genes in *M.* RCC299 formed COPs, while human, mouse, *Drosophila* and *Arabidopsis* have lower reported overlap frequencies (4-9%, 1.7-14%, 15% [of annotated genes], and 7% respectively), with the ranges dependent on sample size analyzed and search criteria used (Boi et al., 2004; Jen et al., 2005). Prior publications have suggested gene overlap evolves

through transposition (Shintani et al., 1999; Makalowska, 2008). This scenario is clearly plausible in the transposon abundant LGC regions of *Ostreococcus* (Palenik et al., 2007). However, only six transposons were identified in *M.* CCMP1545 and none were found in *M.* RCC299 (Worden et al., 2009) making this an unlikely mechanism for formation of *Micromonas* (and presumably all Mamiellophyceae) COPs.

Gene overlap, like that represented by *M.* RCC299 COPs, may arise through a variety of processes including chromosome rearrangement, gene creation (duplication) and gene alteration (e.g., UTR change) (Shintani et al., 1999; Dan et al., 2002; Johnson and Chisholm, 2004; Makalowska, 2008). One study compared the structure of human and mouse orthologs, which overlapped in only one of the species, and found overlapping orthologs are frequently longer and possess more exons than non-overlapping orthologous genes (Solda et al., 2008). In another study of overlapping mammalian genes, alterations of 3' UTRs appeared to play a key role in determining COP formation (Sanna et al., 2008). Both of these gene structural characteristics were commonly observed in *Micromonas* COPs.

2.5.2 COP Function

Despite having more COP genes, the LGC region of *M.* RCC299 was found to be less gene-dense (with longer stretches of intergenic space between genes). This suggests COPs are not directly linked to genome compaction. It is peculiar, based on population-genetic principles to find a less gene dense region of a microbial genome since microbes with large population sizes, both prokaryotic and eukaryotic, are expected to have streamlined genomes (Lynch, 2006). When there are more copies of

an allele, or in this case genome architecture variations, the effect of genetic drift is presumably diminished and the forces of natural selection dominate. Since a larger genome can be equated with a larger mutational burden, one would expect selection for microbial genomes that are as compact as possible (Lynch 2006).

In *M.* RCC299, many LGC region genes were found to have longer UTRs (particularly 3´), higher intron numbers and more frequent overlap than those in the NGC. The noncoding portions of genes, consisting of 5´ and 3´ UTRs and introns, are known to be the dominant regions involved in regulation of gene expression (Barrett et al., 2012). The genes within the LGC region had more EST coverage than those found in the NGC region of the *M.* RCC299 genome. It is plausible that LGC gene organization is adapted for transcriptional regulation efficiency. However, how this would occur is unclear. COPs could perform in a manner similar to prokaryote operons. Genes arranged in operons decrease genetic vulnerability, by reducing the number of required regulatory elements and enabling co-transcription (Lynch, 2006). Although evidence for (or against) the idea of COP involvement in co-regulation is lacking, it seems plausible for genes with related functions For example, in *M.* RCC299 a gene (PID 105064) containing a diacylglycerol (DAG) kinase catalytic domain, which acts as a protein kinase C activator, forms a COP with a eukaryotic protein kinase domain bearing gene (PID 113565, Figure 2.5A). Another example involved three genes with EST support, a mitochondrial import inner membrane translocase (PID 112700), a mitochondrial carrier (MC) family protein encoding gene (PID 112701) and PSP5 (PID 113395), a Prasinophyte-specific protein of unknown

function (Figure 2.5B). Given that the first two genes are functionally related, one might speculate the third gene also has a mitochondrial-related function. Unfortunately because many of the encoded proteins are of unknown function, including many that are present in multiple genomes, it is difficult to assess functional relationships in a more quantitative way. Knowing the functions of at least one gene in a COP may help with future gene predictions for unknown function COP genes.

COP arrangements were not always conserved between the two *Micromonas* genomes. However, this does not rule out the possibility of COP co-regulation. As was the case in the example from Figure 2.2A, if the *M.* CCMP1545 genome only includes a single ortholog sequence from the two *M.* RCC299 COP forming genes, perhaps selection led to the formation of a COP from the *Micromonas* ortholog and a different CCMP1545 gene, which could be co-regulated under the same circumstances. Since 10% of protein encoding genes were unique to each of the *Micromonas* genomes, such gene rearrangements would likely be necessary. Unfortunately, two of the three genes involved in this example were of unknown function.

The higher expression levels observed in the *M.* RCC299 LGC may indicate increased transcript stability for COP genes, while increased LGC intergenic space could facilitate higher transcriptional activity, aiding promoter sharing, proximity and access for transcriptional machinery. There are several theories regarding COP function in gene expression. One resulting from bacterial work is that single stranded

mRNAs of COPs bind along their complementary sequences, blocking ribosome binding sites and translation (Wagner et al., 2002). Following genome analyses of multicellular eukaryotes, Boi et al., (2004) postulated that COPs may result in degraded dsRNA, since a negative association was found between gene expression and the overlapping of 3' UTRs. However, there is also evidence from an Arabidopsis study (Jen et al., 2005) and a mammalian study (Faghihi and Wahlestedt, 2006) contrary to the notion that RNA degradation is the predominant effect of dsRNA formation, as would result from the expression of both genes in a COP.

There is experimental evidence of both co-expression and differential expression of overlapping transcripts in both plants and animals (Vanhee-Brossollet and Vaquero, 1998; Jen et al., 2005). When human overlapping gene data was compared with data from a breast cancer transcriptome, the rate of co-expression among overlapping genes was four times higher than predicted by the random probability of co-expression of any two genes (7.3%) (Solda et al., 2008). Prior studies have reported even higher rates of co-expression in humans (44.9% and 35.1%) (Chen et al., 2005; Galante et al., 2007) and it is theorized that overlapping genes may conceal gene destabilizing elements (AU-rich) resulting in a transcriptional increase (Chen and Shyu, 1995). Whether COPs lead to increased or decreased expression rates, COP arrangements are more conserved than the general organization of genes. Increasing distance between typically COP forming genes is selected against, signaling the functional importance of COPs (Dahary et al., 2005).

COPs may not appear to be a genome space-saving mechanism. However, they could result from natural selection for streamlined genomes, if they enable co-regulation, promoter sharing or increased proximity and access for transcriptional machinery. It is possible the increased intergenic space of the LGC is a function of genome streamlining by encoding unidentified regulatory motifs or noncoding RNAs that enhance regulation. Clearly transcriptome focused studies will be required to resolve the function(s) of COPs, although current high throughput RNA sequencing methods are ill-suited for such future studies.

2.5.3 Introner Elements

As shown here (Figure 2.6A-D) and in Worden et al. (2009), despite high sequence conservation IEs appear to be noncoding and intronic, based on EST supported gene models. This sequence conservation is highly unusual because introns are thought to be neutrally evolving and free to mutate without deleterious effects. IEs could lead to a more streamlined *M.* CCMP1545 genome, if they have deleterious effects (e.g., intron splicing disruption) that allow organism survival, but lead to gene loss. IE prevalence within the *M.* CCMP1545 genome suggests they either recently invaded or are difficult to purge. It is beyond the scope of this study to determine how IEs spread throughout the *M.* CCMP1545 genome. However, since the majority of IEs are in intragenic space, some transcriptional level process seems possible. Whether IEs inserted and spread through the *M.* CCMP1545 genome post-divergence from *M.* RCC299 or *M.* RCC299 has been highly successful at IE elimination is not known. However, the elimination of these noncoding DNA

elements would be consistent with microbial genome streamlining. Future analyses of *Micromonas* strains from other clades and metagenomic data released since the time of this analysis will help clarify this.

2.6 Conclusion

The large size of microbial populations, whether marine or terrestrial, theoretically reduces the likelihood of genetic drift and allows for selection of streamlined genomes (Lynch, 2006). However, this is not directly apparent from analyses of *Micromonas* genomes. In *M*. RCC299 the vast majority of overlapping genes is in the LGC region where gene density is reduced and noncoding regions of genes are more numerous (introns) or longer (UTRs). The LGC regions and their unique gene structures presumably influence the development and biology of this lineage. Moreover the characteristic LGC region has not been observed in other green algae with sequenced genomes (although notably, genomes from other prasinophyte classes are not available). In other organisms, COP gene arrangements are conserved and selection has been found to act against increased distance between COP forming genes (Dahary et al., 2005). That the large size of microbial populations should lead to the elimination of unnecessary genetic material is also challenged when considering *M*. CCMP1545. This genome was found to contain enough non-coding elements to increase the genome size by 1 Mb. IEs do not appear to have a function or a gene functional preference. It is currently unknown why IEs are absent from *M*. RCC299 and the LGC region of *M*. CCMP1545. Future research on how these

elements are gained or lost may provide insights not only into IEs themselves, but the potential functional role of the Mamiellophyceae characteristic LGC region.

Table 2.1 Coordinates and gene counts for eight randomly selected normal GC content (NGC) and two non-randomly selected low GC content (LGC) fragments manually curated and analyzed for *Micromonas* RCC299 transcript data. The coordinates for the low GC region of chromosome 1 are approximately 265,000 - 1,817,500.

Fragment number	Chromosome Number	%GC	Coordinates		Number of Gene Models	
			Left	Right	Total	EST-supported
NGC 1	14	62	473061	633248	79	30
NGC 2	12	66	168664	328851	79	26
NGC 3	1 - 2	67	1922257	29397	80	32
NGC 4	3	67	1486806	1646993	66	28
NGC 5	4	67	1168547	1328734	86	23
NGC 6	10	64	71654	231841	70	32
NGC 7	2	67	670150	830337	83	35
NGC 8	10	67	872594	1032781	78	44
LGC 1	1	48	270000	430188	70	54
LGC 2	1	48	430189	590377	63	55

Table 2.2 Analysis of *Micromonas* RCC299 genome characteristics based on transcript data and manual curation of randomly selected fragments from the normal GC content (NGC) region, as well as non-randomly selected fragments from the low GC content (LGC) region of Chromosome 1. Values in parentheses represent standard deviations. Abbreviations: ME, multi-exon; CDS, coding sequence; UTR, untranslated region; COP, convergent overlapping pair.

	Genome	Normal GC	Low GC
Region size (Mb)[a]	21	19.5	1.5
G+C (%)[a]	64.6	65.9	48
Mean exon size in ME genes (nt)	993[b]	1,024 (1,271)	597 (864)
Mean intron size (nt)	178[b]	185 (122)	94 (56)
Introns per ME gene	2[b]	1.50 (1.09)	2.95 (3.25)
Gene size (nt)	1,804	1,791 (1,381)	1,967 (1,075)
Transcript size (nt)	1,676	1,675 (1,379)	1,684 (1,028)
CDs size (nt)	1,521	1,535 (1,406)	1,344 (944)
5´ UTR size (nt)	110	109 (185)	123 (144)
3´ UTR size (nt)	127	106 (90)	408 (487)

[a]Data from (Worden et al., 2009)
[b]Values extrapolated from the relative proportions of analyzed fragment types

Table 2.3 Gene characteristics comparison of COP and non-overlapping genes derived from randomized sub-sampling and manual analysis of normal GC content (NGC) and low GC content (LGC) genome fragments. Numbers in parenthesis are standard deviations.

	Normal GC	**Low GC**
EST supported models	250	109
Expressed genes (%)	40 (3)	82 (4)
COP forming genes (%)	33 (6)	66 (1)
Overlap (nt) COP genes	44 (57)	567 (641)
Range of overlap (nt) COP genes	1 to 352	1 to 2,407
Transcript size (nt) COP genes	1,563 (1,053)	1,808 (1,149)
Transcript size (nt) non-overlapping genes	1,732 (1,506)	1,490 (787)
Exon per COP genes	1.43 (0.72)	2.92 (2.94)
Exon per non-overlapping genes	1.13 (1.74)	2.85 (3.05)

Table 2.4 Comparison of structural characteristics of COP and non-overlapping genes from normal GC content (NGC) and low GC content (LGC) regions of the *Micromonas* RCC299 genome. Data was collected during manual curation of randomly selected fragments from NGC regions, as well as non-randomly selected fragments from the BOC LGC region, for transcript bearing sequences only. Numbers in parenthesis are standard deviations. Abbreviations: COP, convergent overlapping pair; ME, multi-exon; UTR, untranslated region.

	COPs		Non-Overlapping	
	Normal GC	Low GC	Normal GC	Low GC
Introns per ME gene	1.35 (0.63)	3.1 (3.30)	1.54 (1.20)	2.38 (1.81)
5´ UTR size (nt)	94 (109)	127 (169)	113 (204)	116 (92)
3´ UTR size (nt)	105 (118)	468 (564)	106 (77)	273 (188)

Table 2.5 Genome coordinates and protein IDs (of the sequence containing the particular IE) of example IEs from each of the four IE categories (IE1, IE2, IE3 and IE4) found in *Micromonas* CCMP1545. Coordinates given are relative to the plus strand of the genome, regardless of which strand the IE was located on. EST? indicates whether the gene model and IE had EST support (yes) or not (no) (Worden et al., 2009). Alignments are shown in Figure 2.6A-D.

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

Figure legends

Figure 2.1

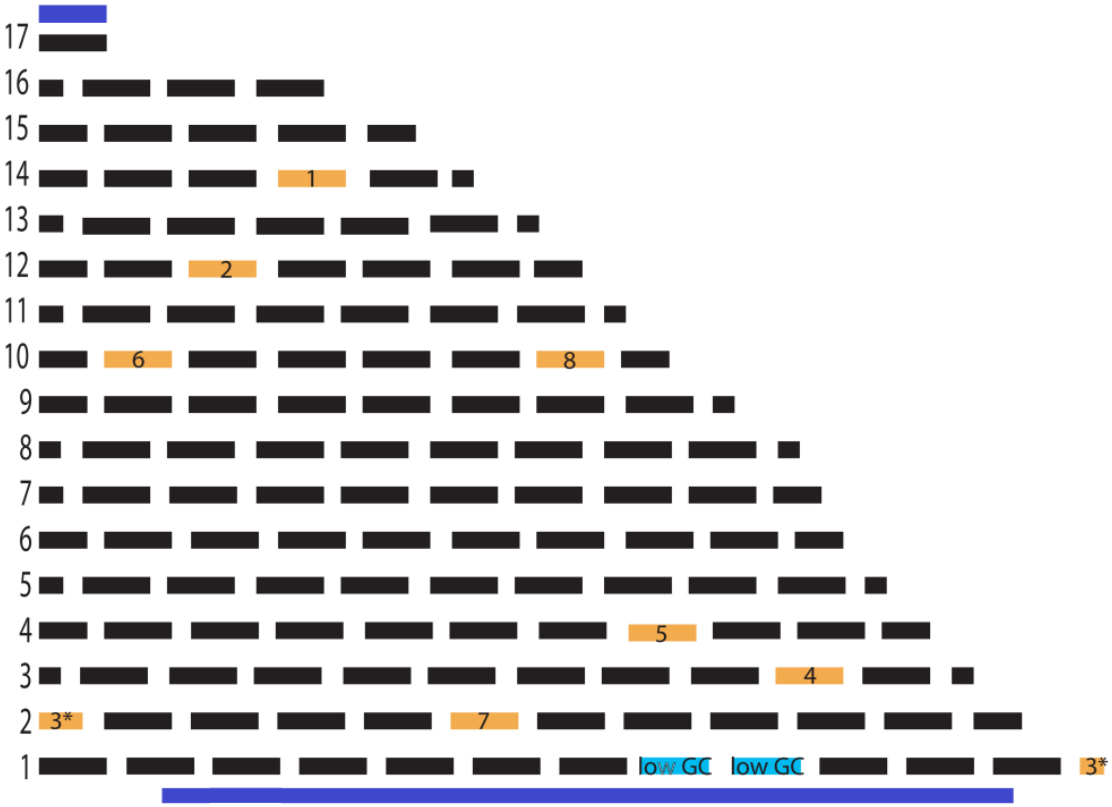

Figure 2.1 Analytical approach to genome architecture analysis in *Micromonas* RCC299. (A) The *M.* RCC299 genome was divided into 210 contiguous 160,188 bp fragments. A pseudo-random number sequence was used to select fragments (orange) for detailed analyses. Additional low GC fragments (turquoise) were analyzed for comparison. Numbers along the left side denote chromosomes and dark blue lines show regions with low GC content (LGC, 14% less than the rest of the genome). For *M.* RCC299 chromosome 17 is the small outlier chromosome (SOC) and chromosome 1 is the big outlier chromosome (BOC). Note that all of the SOC (Chr.17) is LGC, while the majority (Chr1: 265,000-1,817,500bp) of the BOC is LGC. 7% of the *M.* RCC299 genome was found to be LGC (Worden et al., 2009).

Figure 2.2

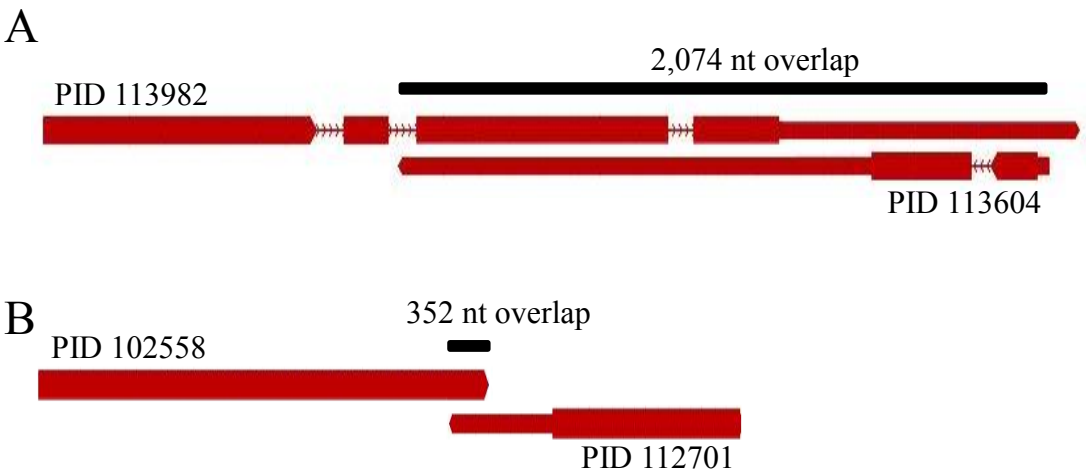

Figure 2.2 Examples of extreme overlap observed during LGC and NGC fragment analyses of COPs from the *Micromonas* RCC299 genome. (A) In the LGC, over half of a green algal specific Aspartyl/Asparaginyl beta-hydroxylase (ABHG) encoding gene (PID 113982), including the 3´ UTR, two exons, one whole intron and part of a second intron, was found to overlap with 2,074 nt (thin black bar) of a green algal specific domain of unknown function (DUF339) containing gene (PID 113604), including the 3´ UTR, two exons, one intron and 5´ UTR. (B) The 3´ CDS of a single exon putative transporter encoding gene (PID 102558) overlaps with 352 nt (thin black bar) of the 3´ UTR of a single exon gene (PID 112701), which encodes an uncharacterized protein with a BSD domain (a domain found in BTF2-like transcription factors, Synapse-associated proteins and DOS2-like proteins). All of these models had EST support (not shown).

Figure 2.3

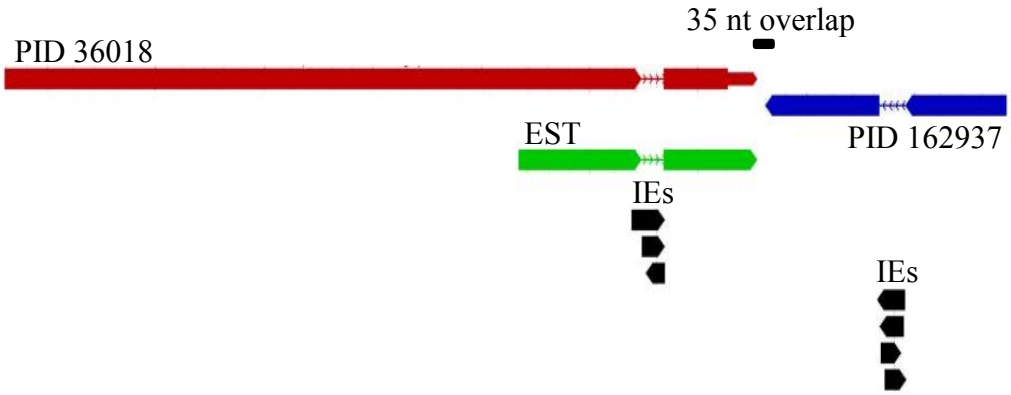

Figure 2.3 Example of *Micromonas* CCMP1545 IE-bearing genes in a convergent overlapping pair (COP). The 3´ UTR of a Ca^{2+} transporting ATPase (PID36018, red) overlaps with 35 nt (thin black bar) of coding sequence from a translation release factor (PID162937, blue). Directional EST support is shown (green bar with 3´ end pointing right) with a gap where an IE (black bar with 3´ end pointing right) is located in the ATPase. An IE (black bar with 3´ end pointing left) was also detected in the release factor, but EST support was lacking. Note that multiple IEs were identified within each gene due to high IE sequence identities.

Figure 2.4

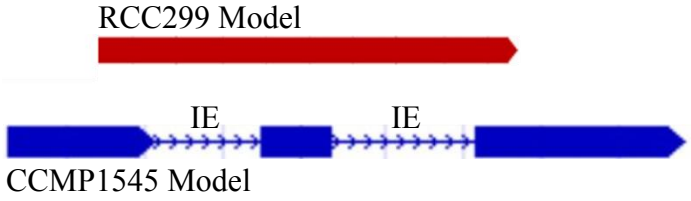

Figure 2.4 *Micromonas* orthologs with and without IEs. A depiction of how a single exon *Micromonas* RCC299 gene (red), can have a multi-exon *Micromonas* CCMP1545 ortholog (blue), due to the presence of IEs.

Figure 2.5

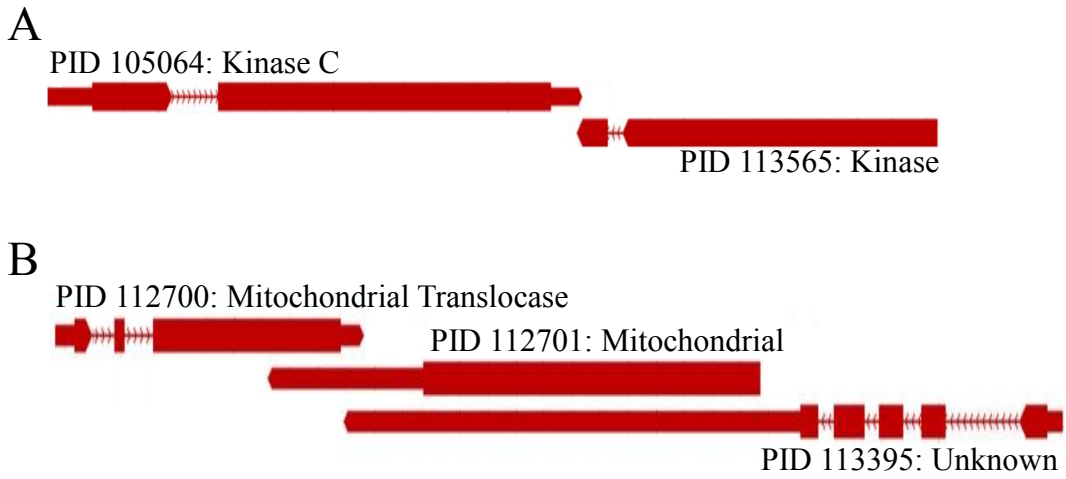

Figure 2.5 Examples of *Micromonas* RCC299 functionally related overlapping genes. (A) Part of the 3´ UTR of a gene (PID 105064) with a diacylglycerol (DAG) kinase catalytic domain, which acts as a protein kinase C activator, overlaps with a region of the 3´ UTR from a gene (PID 113565), which has a eukaryotic protein kinase domain. (B) The entire 3´ UTR and partial CDs of a mitochondrial import inner membrane translocase encoding gene (PID 112700) overlaps with part of the 3´ UTR of a mitochondrial carrier (MC) family protein encoding gene (PID 112701). Part of the translocase 3´ UTR also overlaps with part of the 3´ UTR from the PSP5 gene (PID 113395), a Prasinophyte-specific protein of unknown function. (Note that while not a COP formation, all of the CDs and most of the 3´ UTR from the MC family protein encoding gene overlaps with most of the PSP5 3´ UTR.) All of these models had EST support (not shown).

Figure 2.6

A

```
IE1.4_FN257193      GCGCGTTCTGTCTCACACTGGTCCCCGTACGACCGCGTCGGCGTGTTGAACGCCGATCCT 60
IE1.5_FN257195      GCGCGTTCTAT---ACACTGGTCCCCATACGACCGCGTTCTCGTGGTGAACGCCGATCCT 57
IE1.1_FN257190      GCGCGTTCTATCTCAAACTGGTCCCCATACGACCGCGTCGGCGTGGTGCACGCCGATCCT 60
IE1.2_FN257191      GTGAGTTG-AC---ACACTGGTCCCCATACGACCCCGTCGGCGTGGTGAACGCCGTTTCT 56
IE1.3_FN257192      GTGCGTTCTAT---ACACTGGTCCCCATACGACCCCAATGGCGAGGTGGACGCCGATCCT 57
                    * * ***       * ********** ******* *     ** * ** ****** * **

IE1.4_FN257193      TAAGGACTTTCTCTCCCGGCGTGTCTCTCCGTCCATCACCCCTCGCTTTCAATCCCCCCG 120
IE1.5_FN257195      TAAGGACTTT---TCCCGGCGATTCTCTCCGCCCAT---CCCTCGCTTTCAATCCCC--G 109
IE1.1_FN257190      TAAGGACTTTTCTTCCCGTCGCATCTCTCCGCCTAC---CCCACGGTTTCAATCCCGACA 117
IE1.2_FN257191      TAAGGACTTT---GCCCGTCGTTTCTCTCCGCCCAC---CCCACGGTTTCAATCCCC--G 108
IE1.3_FN257192      TAAGGACTTT---GCCCGTCGTATCTCTCCGCCCAC---ACCTCGCTTTCAATCCCC--G 109
                    **********    **** **  ******** * *     ** ** **********

IE1.4_FN257193      C-CCTCGACGCCTTTCAACTCCA--TCTGACGCCTTTGAACCTCACCCCGATATTCGCT- 176
IE1.5_FN257195      A-CCTCGACGCCTTTCAACTCCA--TCTGACACCTTTCAACTCCACCCCGACAT-CGC-- 163
IE1.1_FN257190      CACCGCGATGCCTTTCAACTCCGCTTCTGACGCCTTTGAACTCCACCCCGACGT-CGCTT 176
IE1.2_FN257191      C-CCGCGACGCCTTTCAACTCCGCTTCTGACGCCTTTGAACTCCACCCCGACGTTCGCT- 166
IE1.3_FN257192      C-CTTCGACGCCTTTCAACTCC-CAACTGACGCCTTTCAACTCCACCCCGACGTTCGCT- 166
                      *  *** *************    ***** ***** ***  ********  * ***

IE1.4_FN257193      CGTACGGACCCTCGACCCTCAG 198
IE1.5_FN257195      -----------------CTCAG 168
IE1.1_FN257190      CGTACGGACCCTCGACCCTCAG 198
IE1.2_FN257191      CGTATGGACCCTCGACCCTCAG 188
IE1.3_FN257192      CGTATAG--------------- 173
```

B

```
IE2.3_FN257197      GTGCGTTCAGG------------------GTG-ACAA---AAAGTTAGTTTTT--CACCC 36
IE2.4_FN257198      GTGCGTTCT-------------------------ATAC---AAA---AGTTTTT--CACCC 28
IE2.5_FN257199      GTGCGTTCTAC------------------GTGTATAC---AAA---CGTTTTT--CACCC 34
IE2.1_FN257196      GCGCGTC--------------------------GCGT----CGCGCCGTCTC--GCGCCC 28
IE2.2_FN257196      GTGCGTCTGACCGCTCCCCATACGACCCCGTTCGCGTTTCGCGCGTCGTTTCTGAAGCCC 60
                    * ****                                          ** *      ***

IE2.3_FN257197      GTATTGCCCGG-------------TTTCATCAACATTTGATCGCGTCCCCTTTCAACAAA 83
IE2.4_FN257198      --ATCGTCCGG-------------TTTCA---ACGTTTGATCGCGTCCCCTTTCAAC--- 67
IE2.5_FN257199      --ACCGCTCGG-------------TTTCA---ACACTTGATCGCGTCCCCTTTCAAC--- 73
IE2.1_FN257196      GCGTCCCTCGGTGG----------TTTCA---ACGTTTGATCGCGTTCCCTTTCAACTGA 75
IE2.2_FN257196      TTTTTCTTCACCCGCGCTTTCCGCTTTCA---ATATTTGATCGCGTCCCCTTTCAACTG- 116
                           *              *****   *   ********** **********

IE2.3_FN257197      TGACCGGTGAACTTTTTTTGTACGGCGGAATGGCCCTCATCATGCAG 130
IE2.4_FN257198      TGACTGGTGAACTATTTTTGTAT---GGAATGGCCCTAA--AAG--- 106
IE2.5_FN257199      TGACCGATGAACATTTTTTGTAT---GGAACGACCCTCATCAG---- 113
IE2.1_FN257196      TGACCGACGCATCGCCCTCCTTC-----TACAG-------------- 103
IE2.2_FN257196      --ACCGATGAACGACCATC------------AG-------------- 135
                      ** *  * *      *
```

C

```
IE3.3_FN257202      GTGAGACTGCTTCCCATACGACCCCGTTCGCGTGGTGAACGCCGTTCCTTAAGGACTTTT 60
IE3.5_FN257204      GTGAGACTGCTTCCCATACGACCCCGTTCGCGTGGCGCGCGCCGTTCCTTAAGGACTTTT 60
IE3.2_FN257201      GTGAGACTGCTTCCCATACGACCCCGTTCGCGTGGTGCACGCCGTTCCTTAAGGACTTTT 60
IE3.1_FN257200      GTGAGACTGCTTCCCATACGACCCCGTTCGCGTGGTGCACGCCATTCCTTAAGGACTTTT 60
IE3.4_FN257203      GTGAGACTGCTTCCCATACGACCCTGTTCGCGTGGTGCACGTCGTTCCTTGAGGACTTTT 60
                    ************************ ********** *  ** * ****** *********

IE3.3_FN257202      CCCGTCGTCACTCTTCACCCGCGTTTCCCTTTCAACGCT----TGACCGGTAAGACGTTC 116
IE3.5_FN257204      CCCGTCGACACTCTTCACCCGCGCTTCCCTTTCAACGTT----TGACCGGTACGACGTTC 116
IE3.2_FN257201      CCCGTCTTCACTCTTCACCCGCGCTCCCCTTTCAACGTTCGTTTGACCGGTAAGACGTTC 120
IE3.1_FN257200      CCCGTCGTCACTCTTCACCCGCGCTTCCCTTTCAACGTT----TGACCGGTAAGACGTTC 116
IE3.4_FN257203      CCCGTCGTCACTCTTCACCCGCGCTTCCCCTCCAACGTT----TGACCGGTAAGACGTTC 116
                    ******  *************** * *** * ***** *    ********* *******

IE3.3_FN257202      GACTGACCGATCGCTTCACCCACGCAG 143
IE3.5_FN257204      GACTGACCGATCGCTTCACCCACGCAG 143
IE3.2_FN257201      GACTGACCGATCGCTTCACCCACGCAG 147
IE3.1_FN257200      GACTGACCGATCGCTTCACCCACGCAG 143
IE3.4_FN257203      GACTGACCGATCGCTTCACCCACGCAG 143
                    ***************************
```

D

```
IE4.1_FN257205      GTGAGACTG-------------------------GTTCCCATACGACCCCGTTCTCGTG 34
IE4.3_FN257207      GTGAGACTG-------------------------CTTCCCGTACGACTCCGTTCGCGTG 34
IE4.4_FN257208      GTGAGACTG-------------------------CTTCCCGTACGACCCCGTTCGCGCG 34
IE4.2_FN257206      GTGAGACTGGAAGATGAGATCACAGACTGCGACTGCTCCCCGTACGACCACGTTCGCGTG 60
                    *********                          * *** ******  ***** ** *

IE4.1_FN257205      TTGATCGTCGTTTCTTAAGGAGTTCTGTGAGACCGCTTCCCGTACGACCTACCCCGTTCG 94
IE4.3_FN257207      TTGATCGTCGTTTCTTAAGGAGTTCT----------TTCCT---------------TTCG 69
IE4.4_FN257208      TTGATCGTCGTTTCTTAAGGAGTTCG------TTTTTTCCT---------------TTCG 73
IE4.2_FN257206      TTGATCGTCGTTTCTTAAGGCGTTCG-----------TTAT---------------TTCG 94
                    ******************** ****           *                 ****

IE4.1_FN257205      TTCGCGCGGTGAACGCCGTTTCTTAAGGCGTACTTTCCTTTCCTTTCGCGCGCTGTGAAC 154
IE4.3_FN257207      --CGCGCTATGA-----------------------------------------TATGAAC 86
IE4.4_FN257208      --CGCGCTATGT-----------------------------------------CATGAAT 90
IE4.2_FN257206      --CGCGC-----------------------------------------------TATGAAC 106
                      *****                                               ****

IE4.1_FN257205      CTCACGTTCGTTTTGACACGTGGGAATGATATCCACCAATCACATGCACGCGCGACTGAC 214
IE4.3_FN257207      CTCACTT----TTTGACAATCGGGAATGATATACACCAATCACATGCACACGTGACTGAC 142
IE4.4_FN257208      CTCACGTTCGTTTTGACACGTGTGAAGTATATCCACCAATAACATCCTCGCGTGACTGAC 150
IE4.2_FN257206      CTCACGT----TTTGACACGTGGGATTTATCTCCACCAATCATATGTGTACGCGACTGAC 162
                    ***** *    *******   * **   ** * ******* * **     ** *******

IE4.1_FN257205      ACGTGTCTTCCCTCGGCC---------TATCAC-----AG 240
IE4.3_FN257207      ACGTGTTTTCCCTCGGCC---------TATCAC-----AG 168
IE4.4_FN257208      ACGTGTCTTCCCTCGGCC---------TATCGCGTCGCAG 181
IE4.2_FN257206      ACGTGTCTTCCCTCGGCCAATGAATGATATCGC-----AG 197
                    ****** ***********         **** *     **
```

Figure 2.6 Alignments of *Micromonas* CCMP1545 IE sequences by category. Examples of IEs aligned by category (A) IE1 examples, (B) IE2 examples, (C) IE3 examples and (D) IE4 examples. IE sequence identifiers and accession numbers are listed along the left sides of the alignments. The number of nucleotides aligned per row is listed on the right side. Conserved positions are noted by stars on the bottom row of each alignment. Also see Table 2.5.

References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389-3402.

Armbrust, E.V., Berges, J.A., Bowler, C., Green, B.R., Martinez, D., Putnam, N.H. et al. (2004) The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. *Science* **306**: 79-86.

Barrett, L., Fletcher, S., and Wilton, S. (2012) Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements. *Cellular and Molecular Life Sciences* **69**: 3613–3634.

Boi, S., Solda, G., and Tenchini, M.L. (2004) Shedding light on the dark side of the genome: overlapping genes in higher eukaryotes. *Curr Genomics* **5**: 509-524.

Bowler, C., Allen, A.E., Badger, J.H., Grimwood, J., Jabbari, K., Kuo, A. et al. (2008) The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes. *Nature* **456**: 239-244.

Chen, C.-Y.A., and Shyu, A.-B. (1995) AU-rich elements: characterization and importance in mRNA degradation*Trends in Biochemical Sciences* **20**: 465–470.

Chen, J.J., Sun, M., Hurst, L.D., Carmichael, G.G., and Rowley, J.D. (2005) Genome-wide analysis of coordinate expression and evolution of human cis-encoded sense-antisense transcripts. *Trends in Genetics* **21**: 326-329.

Countway, P.D. and Caron, D.A. (2006) Abundance and distribution of Ostreococcus sp. in the San Pedro Channel, California, as revealed by quantitative PCR. *Appl. Environ. Microbiol*. **72**: 2496-2506.

Dahary, D., Elroy-Stein, O., and Sorek, R. (2005) Naturally occurring antisense: Transcriptional leakage or real overlap? *Genome Research* **15**: 364-368.

Dan, I., Watanabe, N.M., Kajikawa, E., Ishida, T., Pandey, A., and Kusumi, A. (2002) Overlapping of MINK and CHRNE gene loci in the course of mammalian evolution. *Nucleic Acids Research* **30**: 2906-2910.

Derelle, E., Ferraz, C., Rombauts, S., Rouze, P., Worden, A.Z., Robbens, S. et al. (2006) From the Cover: Genome analysis of the smallest free-living eukaryote *Ostreococcus tauri* unveils many unique features. *Proc Natl Acad Sci U S A* **103**: 11647-11652.

Faghihi, M., and Wahlestedt, C. (2006) RNA interference is not involved in natural antisense mediated regulation of gene expression in mammals. *Genome Biology* **7**: R38.

Galante, P.A.F., Vidal, D.O., de Souza, J.E., Camargo, A.A., and de Souza, S.J. (2007) Sense-antisense pairs in mammals: functional and evolutionary considerations. *Genome Biology* **8**.

Giovannoni, S.J., Tripp, H.J., Givan, S., Podar, M., Vergin, K.L., Baptista, D. et al. (2005) Genome streamlining in a cosmopolitan oceanic bacterium. *Science* **309**: 1242-1245.

Jen, C.H., Michalopoulos, I., Westhead, D.R., and Meyer, P. (2005) Natural antisense transcripts with coding capacity in Arabidopsis may have a regulatory role that is not linked to double-stranded RNA degradation. *Genome Biol* **6**: R51.

Johnson, Z., and Chisholm, S. (2004) Properties of overlapping genes are conserved across microbial genomes. *Genome Research* **14**: 2268-2272.

Lynch, M. (2006) Streamlining and simplification of microbial genome architecture. In *Annual Review of Microbiology*, pp. 327-349.

Makalowska, I. (2008) Comparative analysis of an unusual gene arrengement in the human chromosome 1. . *Gene* **423**: 172-179.

Moreau, H., Verhelst, B., Couloux, A., Derelle, E., Rombauts, S., Grimsley, N. et al. (2012) Gene functionalities and genome structure in Bathycoccus prasinos reflect cellular specializations at the base of the green lineage. *Genome Biol* **13**: R74.

Palenik, B., Grimwood, J., Aerts, A., Rouze, P., Salamov, A., Putnam, N. et al. (2007) The tiny eukaryote *Ostreococcus* provides genomic insights into the paradox of plankton speciation. *Proc Natl Acad Sci U S A* **104**: 7705-7710.

Peers, G., and Niyogi, K.K. (2008) Pond scum genomics: The genomes of Chlamydomonas and Ostreococcus. *Plant Cell* **20**: 502-507.

Read, B.A., Kegel, J., Klute, M.J., Kuo, A., Lefebvre, S.C., Maumus, F. et al. (2013) Pan genome of the phytoplankton Emiliania underpins its global distribution. *Nature* **499**: 209-213.

Sanna, C.R., Li, W.H., and Zhang, L. (2008) Overlapping genes in the human and mouse genomes. *BMC Genomics* **9**: 169.

Shintani, S., O'HUigin, C., Toyosawa, S., Michalova, V., and Klein, J. (1999) Origin of gene overlap: The case of TCP1 and ACAT2. *Genetics* **152**: 743-754.

Solda, G., Suyama, M., Pelucchi, P., Boi, S., Guffanti, A., Rizzi, E. et al. (2008) Non-random retention of protein-coding overlapping genes in Metazoa. *Bmc Genomics* **9**.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**: 4673-4680.

Vanhee-Brossollet, C., and Vaquero, C. (1998) Do natural antisense transcripts make sense in eukaryotes? *Gene* **211**: 1-9.

Venter, J., Remington, K., Heidelberg, J., Halpern, A., Rusch, D., Eisen, J. et al. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**: 66-74.

Wagner, E.G.H., Altuvia, S., and Romby, P. (2002) Antisense RNAs in bacteria and their genetic elements. *Homology Effects* **46**: 361-398.

Worden, A.Z., Lee, J.H., Mock, T., Rouze, P., Simmons, M.P., Aerts, A.L. et al. (2009) Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes *Micromonas*. *Science* **324**: 268-272.

3 Chapter 3: Prolific but distinct repetitive introns in picoeukaryote species highlight Antarctic *Micromonas*

3.1 Abstract

Introner elements (IEs) are introns with high sequence conservation that occur repetitively within genomes, distinguishing them from regular spliceosomal introns. Large numbers of repetitive introns were first identified in the marine picoeukaryote *Micromonas* CCMP1545 and more recently in fungi. To systematically examine the distribution of IEs in all *Micromonas* clades, we designed PCR primers to *Micromonas* and *Ostreococcus* (a related prasinophyte) orthologs known to contain one or more IEs in *M*. CCMP1545. IEs confirmed in *M*. CCMP1545 were identified in a second cultured Clade D strain and recovered from coastal Pacific Ocean clones. Some of the environmental clones had 100% identity to the Clade D North Atlantic isolates with IEs at the same loci, but IEs were also found at non-homologous loci. Although introns with high identity to D-lineage IEs were not found in other *Micromonas* lineages, additional intron presence/absence polymorphisms led to the discovery of new classes of IEs. Introner elements (termed ABC-IEs) were present in representatives from the monophyletic *Micromonas* clades A, B and C, as well as metagenomic data, and had a predicted 3-loop secondary structure. Two other types of novel IEs were discovered in *Micromonas* Clade E2. These each had high identity with sequences in Antarctic metagenomic data, including a meromictic lake, and expanded the known range for isolate *M*. CCMP2099 (previously considered endemic to the Arctic) to the massive Southern Ocean biome. Polar *Micromonas* E2-IEs were not observed in temperate or tropical metagenomic data, or other Clade E cultures.

Collectively, our data demonstrate presence of *Micromonas* in the euphotic zone of all ocean provinces. Furthermore, the previous rarity of polymorphic intron observations appears to be due to biases in taxon sampling.

3.2 Introduction

The evolution of spliceosomal introns is one of the great mysteries of molecular biology. These intervening, non-coding sequences are distinctly eukaryotic features intimately linked with eukaryote evolution. Introns are thought to have 'sped up' eukaryotic evolution by generating new proteins when alternative splicing occurs (Gilbert, 1978; Koonin, 2006). Mutations in splicing that convey an advantage, for example, by allowing formation of new mRNA that encodes an improved protein, can be retained. Introns play influential roles in eukaryotic biology. Some have single nucleotide polymorphisms (SNPs) implicated in disease states (Frayling et al., 2007; Robbens et al., 2008). Others modulate gene expression, via mechanisms such as intron mediated enhancement, and many enable alternative spliceoforms that result in multiple protein products translated from an individual gene (Modrek and Lee, 2002).

Evolutionary processes leading to intron presence or absence are still not well understood. The presence of introns in homologous positions within orthologous genes of divergent taxa (e.g., mammals and plants) suggests the last eukaryotic common ancestor (LECA) had an intron-rich genome (Koonin, 2006; Roy and Gilbert, 2006). This idea has drawn considerable support for the introns-early theory of intron evolution, where noncoding sequences interrupted genes in the earliest evolutionary stages and the typical lack of introns in prokaryotes resulted from

massive streamlining (Roy, 2003; Roy and Gilbert, 2006). Alternatively, while introns likely entered early, some hypothesize they became widespread only with the emergence of eukaryotes (Koonin, 2006) and modern day intron patterns have been dominated by intron loss processes since divergence from ancestral eukaryotes (Roy, 2006; Csuros et al., 2011). Moreover, most introns are neutrally evolving and even orthologously positioned introns (the 25% of introns found at orthologous loci in *Homo sapiens* and *Arabidopsis thaliana* genes), lack sequence homology (Sverdlov et al., 2007; Rogozin et al., 2012).

Recently, several studies have demonstrated that polymorphic introns are more common than initially thought (Li et al., 2009a; Worden et al., 2009; Denoeud et al., 2010; Torriani et al., 2011; van der Burgt et al., 2012; Verhelst et al., 2013). The first report of multiple (24) intron presence/absence polymorphisms (PAPs, used here solely to refer to introns that are not in orthologous positions in related taxa) was detected while comparing genomes from two *Daphnia pulex* isolates (Li et al., 2009a). When the regions containing these PAPs were analyzed, in 84 natural *D. pulex* isolates, most were attributed to recent gains, often at homologous loci. Additionally, short repeats (5-12 nt) flanking the inserted introns appeared to be common, leading to the hypothesis that polymorphic introns were derived from double-strand break (DSB) repair.

The most striking PAPs are repetitive sequences found in introns across a single genome. These were first observed in a single species within the eukaryotic

supergroup Archaeplastida, specifically the green alga *Micromonas pusilla* and termed introner elements (Worden et al., 2009). Less numerous cases occur in two types of Opisthokonta, the larvacean tunicate *Oikopleura dioica* (Denoeud et al., 2010) and several species of fungi (Torriani et al., 2011; van der Burgt et al., 2012). Collectively, these are distinct from previously reported PAPs because they do not represent a single gain event at a particular locus, but rather the spread of a conserved sequence in apparently 'new' intron sites within multiple genes in a given genome. What remains unclear is whether these repetitive/conserved introns reflect oddities in a few unusual organisms or are common in many extant taxa, 'missed' until now due to taxon under-sampling.

In the appendicularian genome of *O. dioica*, four genes were identified that each contained a pair of highly conserved introns referred to as 'nearly identical introns', NII (Denoeud et al., 2010). The proposed means of propagation was reverse splicing at the pre-mRNA stage. Because adjacent introns were observed in many *O. dioica* genes, it was surmised that other newly acquired introns may have propagated this way and subsequently diverged. Highly similar introns, termed introner-like elements (ILEs) in the fungi *Mycosphaerella graminicola* (Torriani et al., 2011), *Cladosporium fulvum*, *Dothistroma septosporum*, *Mycosphaerella fijiensis*, *Hysterium pulicare* and *Stagonospora nodorum*, number around 10 to 108 per cluster with each fungus having 1 to 8 clusters (van der Burgt et al., 2012). ILEs differ from regular spliceosomal introns (RSIs) by being longer, having high sequence conservation and Gibbs free energy (ΔG) predictions connected with potential

secondary structures. Because they are always found on the coding strand, it is hypothesized that reverse splicing followed by reverse transcription led to ILE-spread through the genome. ILEs appear prone to degeneration of both sequence conservation and length, leading to the hypothesis that ILEs are the source of the majority of fungal RSIs (van der Burgt et al., 2012).

The first and most prolific repetitive introns, IEs, were discovered in *Micromonas* (Worden et al., 2009), an ecologically important genius of unicellular marine green alga, found from the tropics to the Arctic (Worden, 2006; Foulon et al., 2008; Not et al., 2008; Li et al., 2009b; Worden et al., 2009). *Micromonas* belongs to the Mamiellophyceae class of the prasinophytes, which together with chlorophyte algae and land plants form the Viridiplantae. Comparative genome analysis of *Micromonas pusilla* (isolate CCMP1545) and *Micromonas* sp. RCC299 (Worden et al., 2009), representing two of six known clades, were found to share at most 90% of their protein encoding genes (Slapeta et al., 2006; Worden et al., 2009). The *M.* CCMP1545 genome is 1 Mb larger than that of *M.* RCC299. This is due almost entirely to the presence of four classes of IEs (IE1-IE4), with an average length (~173 nt) shorter than known transposable element (TE) classes and that occur over 6,000 times in the *M.* CCMP1545 genome. IEs were not found in *M.* RCC299 (Worden et al., 2009). Multiple IEs can be retrieved using a simple BLASTN against the genome and a series of motifs have been constructed to recognize these elements, including more divergent versions (Worden et al., 2009; Verhelst et al., 2013). The extent to which IEs are unique to *M.* CCMP1545, a strain isolated from the North Atlantic in

the 1950s, or are present in other *Micromonas* strains or clades is still unclear. At the time of their discovery, IEs were also detected in Sargasso Sea metagenomic data (Venter et al., 2004) from the North Atlantic Ocean (Worden et al., 2009).

To better understand intron gain and loss in *Micromonas* as a whole, we designed a study to establish presence/absence patterns of *M.* CCMP1545 IEs in another D-lineage taxa and additional *Micromonas* clades with isolates from around the world. We analyzed representative isolates from each of the cultured *Micromonas* clades by designing primers to orthologous genes present in *Micromonas* and related prasinophytes. Then, we investigated PAPs in environmental sequences generated for two of these genes, for which the primers designed for use on pure cultures, would be suitable for use in DNA samples containing diverse taxa and still allow recovery of the target organisms. Finally, we analyzed metagenomic data and used PCR data and transcriptomes, generated herein, to establish the contributing *Micromonas* clade and demonstrate IE splicing. The results reveal complex presence/absence patterns, as well as new types of IEs that correspond to the diversification of different *Micromonas* lineages. Additionally, we demonstrate that *Micromonas* is widespread in the Southern Ocean, as well as an Antarctic meromictic lake, with near perfect identity to an Arctic strain.

3.3 Methods

3.3.1 Culturing and nucleic acid extraction

Ten *Micromonas* isolates were acquired from culture collections, specifically the Culture Collection of Marine Phytoplankton (CCMP), the Roscoff Culture

Collection (RCC), the North East Pacific Culture Collection (NEPCC) and the Australian National Algae Culture Collection (CS), and bear isolate numbers prefaced by the culture collection abbreviation. These were grown at 200 µE m^{-2} s^{-1} on a 14-h/10-h light/dark cycle in standard media and conditions (Table 3.1). Additionally, *M.* RCC434 and *M.* RCC1614 were grown and harvested to improve 18S rRNA gene sequence availability for two clades. Cells were harvested by centrifugation at 6,000 or 8,000 × *g*, the supernatant removed immediately and pelleted cells frozen at -80^{o}C until extraction. Environmental samples were collected from surface water (0.5 m) adjacent to the end of the Scripps Institution for Oceanography pier (32° 53' N, 117° 15' W) in April and October 2001, as part of a previous study (Worden, 2006). DNA was extracted using a QIAGEN DNeasy Plant or Blood and Tissue Kit (Germantown, Maryland, USA) according to the manufacturer's instructions. *M.* CCMP1764 was also grown for genome sequencing. In this case cells were pelleted as above, but in larger volumes. DNA was extracted using a standard CTAB protocol (outlined athttp://www.mbari.org/phyto-genome/Resources.html).

For RNA extraction *M.* CCMP2099 and *M.* NEPCC29 cells were grown at 80 and 100 µE m^{-2} s^{-1}, respectively, filtered onto a 47 mm, 0.2 µm pore-size Supor filter (Pall Gelman), flash frozen in liquid N_2 and transferred to -80^{o}C until extraction using the TotallyRNA kit (Life Technologies, Grand Island, NY, USA). Manufacturer's methods were used, except for an added step of 1 min bead beating at the outset using ~200 µl of a 1:1 mixture by volume of 0.1 mm and 0.5 mm diameter autoclaved glass beads (Biospec Products, Bartlesville, OK, USA) in 1 ml kit-supplied lysis buffer.

The extract was treated using the TurboDNA-free kit (Life Technologies) following manufacturer's instructions, integrity was evaluated on a bioanalyzer (Agilent, Santa Clara, CA, USA), and quantity determined using a Qubit fluorometer (Life Technologies).

3.3.2 PCR, cloning and sequencing

PCR primers were designed to conserved regions of four gene homologs found in the genomes of *Ostreococcus tauri*, *Ostreococcus lucimarinus*, *M.* RCC299 and *M.* CCMP1545, spanning IEs in the latter (Table 3.2). In addition, 18S rRNA gene primers (18SEUKF: 5′-ACCTGGTTGATCCTGCCAG-3′; 18SEUKR: 5′-TGATCCTTCYGCAGGTTCAC-3′) were used to verify isolate identity (Worden et al., 2004). 25µl PCR reactions consisted of 9 µl RNAase free water, 12.5 µl Qiagen HotStar Master Mix (HotStar Taq Polymerase, PCR buffer with 3 mM $MgCl_2$, and 400 mM of each dNTP), 1.25 µl each of forward and reverse primers, and 1 µl of DNA. For negative controls, an additional 1 µl of RNAase free water was used in place of DNA. PCR products were purified using either the QIAquick PCR Kit or QIAquick Gel Extraction Kit (Qiagen), if multiple product bands were observed.

Cloning was performed using TOPO TA Cloning (Invitrogen, Carlsbad, CA). For each culture 2-16 clones were purified for subsequent sequencing. For environmental samples preliminary clone libraries were constructed using the same methods, and a small set of clones sequenced, followed by construction of two libraries per gene (cloning of bands extracted based on size, one reflecting intron/element-less sequences) from which 96 clones each were sequenced. Culture

plasmids were purified using the QIAprep Miniprep Kit and products were sequenced bi-directionally on an ABI 3100 using BigDye terminator v.3.1 chemistry (Applied Biosystems Inc., Foster City, California) with M13F: (5′-CTGGCCGTCGTTTTAC-3′) and M13R (5′-CAGGAAACAGCTATGAC-3’) primers. Two additional sequencing primers were used for internal regions of the 18S rRNA gene, 502F (5′-GGAGGGCAAGTCTGGT-3′) and EUK1174R: (5′-CCCGTGTTGAGTCAAA-3′). Plasmids from larger environmental clone libraries were purified (Davis, 1986) and sequenced using a 3730xl DNA Analyzer (Applied Biosystems). Sequences were assembled using DNAStar (Lasergene) and manual curation.

The *M*. CCMP2099 and *M*. NEPCC29 transcriptomes were sequenced using pair-ended Illumina sequencing. Assembly was performed as in Keeling et al., (in press). *M*. CCMP1764 DNA was sequenced using the 454-FLX platform, and the reads have been deposited in CAMERA under project CAM_PROJ_CCMP1764.

3.3.3 Clustering and phylogenetics

Genomic DNA and cDNA sequences were aligned using ClustalW, or manually in DNASTAR or Bioedit. Clustering was performed using BLASTClust (Altschul, 1990) with required coverage specified by a 100% identity threshold over 100% of the sequence length.

For gene homolog phylogenetic analysis, introns and IEs were removed and nucleotide sequences translated. The resulting sequences were compared to published gene models for *M*. RCC299 and *M*. CCMP1545, as well as cDNA evidence for the former two and *M*. CCMP2099. Non-prasinophyte sequences were removed from

environmental clone libraries based on an initial BLASTX and BLASTN (Altschul et al., 1997) evaluation to NCBI's non-redundant database and in-house genome databases. Pairwise intron and IE nucleotide identities were computed using Emboss Water (employing the Smith-Waterman algorithm), unless otherwise specified. Logos were constructed from alignments using WebLogo (Crooks et al., 2004) after manual curation of insertion sequence alignments.

For the 18S rRNA gene phylogeny, sequences were aligned using MAFFT (Katoh et al., 2005). Regions of unambiguous alignment were identified using MUST (Philippe, 1993) and, except ten positions (corresponding to nucleotide 645 to 655 in the *M*. CCMP1545 sequence #AY954994) that help resolve *Micromonas* clade differences, all gap-containing positions were removed. Phylogenetic tree reconstructions were statistically evaluated using Bayesian inference (BI) and maximum likelihood (ML) methods from 1,646 homologous positions. We used GTR+Γ+I as the model of nucleotide substitution for both analyses. Phylogenetic analyses were calculated using MrBayes for BI (Ronquist et al., 2012) and Treefinder for ML (Jobb et al., 2004). Bayesian analyses were performed with two independent runs and 1,000,000 generations per run. After a burn-in of 350,000 trees per run, the remaining trees were used to reconstruct a consensus tree and to get posterior probabilities for node supports. Bootstrap values were calculated using 1,000 replicates with the same substitution model.

For the Calcium ATPase phylogeny sequences from cultures and representative environmental sequences were aligned using MAFFT (Katoh et al., 2005). Regions with unambiguous alignment were identified using MUST (Philippe, 1993) and all gap-containing positions were removed. A ML phylogeny was built with Treefinder from 734 homologous positions using the TVM+G model including relaxing parameters of first, second and third codon positions. The model was selected using Modeltest (Posada and Crandall, 1998) as implemented in Treefinder (Jobb et al., 2004). Bootstrap statistics were performed as described for the 18S rRNA gene tree.

3.3.4 Metagenome searches

Metagenomes were BLASTN queried using IE1.1 (Worden et al., 2009) and the ABC transporter PID68853 IE to represent Clade D elements. The NEPCC29 ABC Transporter element and the element from model *M*. RCC299 Mipur011i11380 (Verhelst et al., 2013) represented ABC-IEs in the metagenome queries, while the *M*. CCMP2099 ABC Transporter elements E2-IEt1 and E2-IEt2 represented E2_IEs in the CAMERA implemented queries (Sun et al., 2011). The complexity filter was off and only hits with E-values $<e^{-5}$ were returned. Those with repetitive elements typically ranged from e^{-7} to e^{-100}. Only metagenomic reads with flanking sequence on either side of the 'hit' alignment region were further investigated. Sequences with E-values $<e^{-7}$ were verified as being repetitive elements via alignment (or, for the recovered sequences, were used as BLASTN queries against the *M*. CCMP2099 and *M*. NEPCC29 transcriptomes). Nucleotide identities for Antarctic sequences and *M*.

CCMP2099 were typically 99% between metagenomic flanking sequence and transcripts. This was also the case for ABC-IEs and *M*. NEPCC29, which had higher identity to each other than to *M*. RCC299.

3.4 Results

Our study was designed to determine whether IEs were present in other *Micromonas* clades, as well as Clade D isolates other than *M*. CCMP1545. The established *Micromonas* clades are thought to reflect species level differences and have been defined using multiple marker genes, including the 18S rRNA gene (Slapeta et al., 2006; Worden, 2006; Worden et al., 2009). Divergence of the D-lineage from other *Micromonas* lineages is estimated at 66 ± 10 My ago, while, as a reference, the split between gymnosperms and angiosperms is estimated at between 290-320 My ago (Slapeta et al., 2006). Because our experimental design relied on robust clade discrimination, we reanalyzed relationships between different *Micromonas* isolates using new sequence data from cultures (Table 3.2). For groups with sparse representation, we included sequences from environmental clone library studies. Our phylogenetic analysis confirmed the presence of six clades, including the purely environmental Clade _.IV, which lacks Slapeta clade lettering because it is not represented in culture collections and hence was not analyzed by Slapeta (Figure 3.1A). Clades A.II and B.I do not retain support based on the 18S rRNA gene and are more unstable based on partial 18S rRNA gene trees (Worden et al., 2009). However, these two clades can be distinguished using proteins sequences (Slapeta et al., 2006). Our phylogenetic reconstruction indicated that Clade E should be reclassified as two

groups (Figure 3.1A). The previous Clade E structure appears to have resulted from taxon-undersampling and potentially by long-branch attraction. Here, the level of divergence between Clade E members warrants division at the clade level, but for simplicity, we term these groups E1 and E2. Notably, E2 is composed of the Arctic strain *M.* CCMP2099 and an Arctic environmental 18S rRNA gene sequence, but not sequences from more temperate or tropical regions.

To ask whether IEs were present in the broader *Micromonas* radiation, we selected cultures representing each clade, including E1 and E2 (Table 3.1, Figure 3.1A). We then designed PCR primers to amplify regions spanning one or more IEs in *M.* CCMP1545 (Table 3.3). Regions were selected from four protein encoding genes that had orthologs in all published *Micromonas* (*M.* CCMP1545, *M.* RCC299 and genomic contigs generated herein for *M.* CCMP1764) and *Ostreococcu*s (*O. lucimarinus* and *O. tauri*) genomes. These encoded a putative calcium ATPase, a putative NADH dehydrogenase [ubiquinone] flavoprotein 1 (referred to hereafter as NADH dehydrogenase), actin and an ABC transporter. Given their presence in multiple members of two Mamiellophyceae genera, these genes were presumed present in other *Micromonas*, specifically the ten isolates investigated here. Our primers generated data that collectively represented each of the five cultured clades (Table 3.1).

IEs were recovered in Clade D representative *M.* CCMP490, isolated in the western North Atlantic, always at the same insertion locus as for *M.* CCMP1545

(Figure 3.1B-E). The *M*. CCMP490 IEs were also at the same phase (i.e., the same codon position) as in *M*. CCMP1545, with phase 0 representing introns located between two codons. The identities for these culture IEs at homologous positions were 100% and 99% for the two in actin and 100%, 98% and 97% for the single amplified IEs in the calcium ATPase, ABC transporter and NADH dehydrogenase, respectively. For both isolates, three of these IEs were phase 0, while the 5′-most IE in the actin gene fragment was phase 1 and the ABC transporter IE was phase 2. All five of these IEs had 58-60% G+C, lower than the overall genome average of 65% G+C content (Supplemental Table 3.1). While the actin 3′ IE, calcium ATPase and NADH dehydrogenase IEs had canonical GT/AG 5′ss/3′ss (ss, splice site, also known as donor and acceptor sites, indicated here as Donor/Acceptor), the actin 5′ IE and ABC transporter had GC/AG ss.

Clade D IEs were not found in homologs from representatives of clades A, B, C and E (Figure 3.1B-E; note that E1 and E2 data was lacking for the calcium ATPase, as was E2 data for actin). Although IEs were not identified, introns of other types were found in the ABC transporter (Figure 3.1E). Specifically, the same PAP-intron was present in all three Clade C strains and two different PAP-introns were found in Clade E2. The Clade C PAP-intron was phase 0 and had complete conservation in motifs associated with splicing, including the branch point (CTGAC). A single nucleotide variation was present in this intron between the three isolates sequenced. Although PCR products from Clade A.II (and B.1) did not contain introns, six hits were recovered in the *M*. RCC299 genome using the Clade C PAP-intron as a query.

Sixty-four hits (E-values e^{-05} to e^{-08}, with nucleotide identities 88-94%) were recovered from Clade B isolate *M*. CCMP1764 genomic DNA generated herein using the 454-platform (Supplementary Figure 3.1A). The two best *M*. RCC299 hits (e^{-7} and e^{-6}) were to chromosome 1 introns, one in a putative calmodulin-binding protein (Prot ID 55550) and the other in ribonuclease H (Prot ID 105055) (Supplementary Figure 3.1B). The *M*. RCC299 introns had 79 and 81% identities to the *M*. NEPCC29 PAP-intron, respectively, higher than expected for RSIs.

RSI similarities can be characterized for the *Micromonas* β-tubulin gene because sequences exist for all cultured clades and this gene has three introns. Two of these introns represent homologous RSIs (termed 5′ and 3′ here) for which RSI-locus comparisons show 73 (5′ RSI) and 51% (3′ RSI) nucleotide identities between *M*. NEPCC29 (Clade C.I) and *M*. RCC299 (Clade A.II). RSIs at the same loci in clades D, E1 and E2 have <50% identity to those in the other clades. Furthermore, BLASTN queries, of the *M*. RCC299 and *M*. CCMP1545 β-tubulin RSIs to their respective genome sequences, attain only self-hits and pairwise alignment of these sequential β-tubulin RSIs renders identities <50% within each strain. Unlike the strikingly high identity between the Clade C ABC transporter PAP and introns in Clade A.II (RCC299), RSIs such as the β-tubulin example above, are much more divergent.

We also tested the Clade C PAP-intron, which shared high identity with intronic sequences in clades A and B, for potential secondary structure, a characteristic of noncoding RNAs with specific functions (Mathews et al., 2010).

This sequence formed a three stem structure (Supplementary Figure 3.1C) also observed for fungal ILEs (van der Burgt et al., 2012), with the branch point located in a loop. Although we lacked genomic data for Clade C the presence of highly similar sequences in multiple clade A and B genes led us to term these repetitive sequences ABC-IEs, with the prefix referring to the monophyletic *Micromonas* A, B and C clades.

The Clade E2 PAP-introns (identified in isolate CCMP2099), classical D-lineage IEs and ABC-IEs were in independent loci within the ABC transporter. The two 5′-most PAP-introns in the *M*. CCMP2099 ABC transporter had 89% identity to one another, higher than expected for RSIs (Figure 3.1E alignment, hereafter referred to as Type 1). The third E2 PAP-intron, located nearer the 3′ end of the ABC transporter PCR product, appeared to be distinct from these and was much longer (hereafter termed Type 2; 185 nt versus 74 and 75 nt). These introns were phase 0, phase 2 and phase 0, moving in the 5′ to 3′ direction, and had canonical splice sites (GT/AG). BLASTN queries against the *M*. RCC299 and *M*. CCMP1545 genome assemblies, as well as the *M*. CCMP1764 cDNA sequences, did not recover significant hits. Thus, our results indicated that conserved elements, akin to the four examples of nearly identical introns in *O. dioica* (Denoeud et al., 2010), were present in *Micromonas* Clade E2. Given the lack of genome data for this strain, it remained unclear whether these might be present in other *M*. CCMP2099 genes as well, similar to IEs in Clade D.

M. CCMP490 and *M*. CCMP1545 are North Atlantic isolates, hence we were interested in whether IEs were present in Pacific Ocean Clade D members and, if present, their level of conservation with Atlantic IEs. We examined samples from the Southern California Bight (eastern North Pacific Ocean) in which both Clade D and the uncultured Clade _.IV had previously been detected, along with other *Micromonas* clades (Worden, 2006). We used two of the PCR primer sets to construct environmental clone libraries from a spring and an autumn sample. Of the total clones sequenced, 123 and 352 came from the correct calcium ATPase and actin homologs, respectively. These were composed of both IE-bearing and non-IE-bearing sequences, the latter coming from *Micromonas* and *Ostreococcus* (and sometimes other more distant taxa). In each sample, the majority of Clade D sequences from both genes contained IEs at the same loci as seen in cultured Clade D strains (Figure 3.2A, B, Supplement Dataset 3.1). Some were identical sequences throughout the gene and IE(s), while others showed a limited number of nucleotide variations (Figure 3.2C, Supplementary Figure 3.2).

Several new PAP-introns were observed, some of which represented new additions to known IEs. In the fall sample alone, six calcium ATPase clones (Cluster D, clustered at the 100% nucleotide identity level) contained an additional 5′ IE. Cluster D coding sequences (CDS) and 3′ IE sequences had nucleotide variations distinct from other IE-bearing environment clones and from *M*. CCMP1545 and *M*. CCMP490 (Figure 3.2C). Phylogenetic analysis of the calcium ATPase itself showed this environmental cluster had a basal position to the entire *Micromonas* Clade D-

lineage with complete support (Supplementary Figure 3.3). The limited phylogenetic distance between CDS from Clade D and environmental Cluster D makes it unlikely that Cluster D sequences are derived from the uncultured *Micromonas* Clade _.IV, which shows considerable divergence based on the 18S rRNA gene (Figure 3.1A). The 5′ and 3′ss were canonical (GT/AG) for the Cluster D 5′ IE and for 3′ IE sequences from 14 Environmental Clusters (including cluster A, which contains culture sequences). (The environmental Cluster G IE had GC/AG ss.)

Like the calcium ATPase clones, the majority of actin-IE-bearing sequences from the Pacific samples also had high identity to CDS (≥99%) and IEs (12 identical clones) from Clade D cultures (Supplementary Figure 3.2). Still, one spring Cluster (S4, composed of a single sequence) contained only the 3′ IE, despite 99% nt identity to CDS from Clade D cultures (Figure 3.2B, Supplementary Figure 3.2). In actin sequences bearing two IEs, the 5′ IE had 97-100% nucleotide identities and the 3′ IEs had 98-100% identities. The environmental Clone S4 IE had lower identity (maximum 93%) to the IE at the homologous loci. In both culture and environmental sequences, 5′ and 3′ss were canonical GT/AG in actin 3′ IEs and GC/AG in 5′ IEs. Two clones, from the spring and fall samples, with identical CDS, contained a PAP-intron located between the actin IE-loci (Figure 3.2B). These clones shared 96% CDS identity to Clade D cultures and much less to other cultured *Micromonas* clades (~90-91%). The clones' introns had a single nucleotide variation between them and non-canonical ss (TG/GG). High amino acid identity for the amplified region (across all *Micromonas* actin sequences) and the overall limited number of positions amplified

precluded phylogenetic analyses. These clones potentially represent uncultured Clade _.IV, or more likely, given their still relatively high identify to *M*. CCMP1545, a group basal to the Clade D lineage. Nevertheless, the actin PAP does not appear to be an IE, based on a BLASTN query of the *M*. CCMP1545 genome.

Metagenomic data was used to further explore the various PAP-introns identified and whether they represented RSIs or IEs. Multiple hits were attained with the two shorter, nearly identical Clade E2 Type 1 introns (Figure 3.3B, D), some with 100% identity, and other hits were attained with the longer Clade E2 Type 2 intron (Figure 3.3D, Supplemental Figure 3.4B). The Type 1 introns appeared to be intervening sequences within genes encoding different proteins, some with recognizable functions, including e.g., an autophagocytosis-associated protein, a putative aminopeptidase, transcriptional repressors and an Early Light Induced Protein (ELIP) according to BLASTX analyses using the NCBI nr database. Samples from which these elements arose were obtained in the Antarctic (Ace Lake, Southern Ocean, Ross Sea), a region where *Micromonas* has never been reported. Notably, metagenomic data are not available from Arctic waters, but the Antarctic intervening sequences were highly similar to the Type 1 introns from Arctic *Micromonas* strain *M*. CCMP2099 (Figure 3.3B). They were therefore termed E2-IEs, specifically E2-IEt1 to differentiate them from the Type 2 intron. We confirmed E2-IEt1s from Antarctic metagenomic reads were intronic by comparison to a *M*. CCMP2099 transcriptome sequenced herein (Figure 3.3B).

The Type 2 Clade E2 ABC transporter element (E2-IEt2) had high similarity to sequences in multiple other protein encoding genes as well (Supplementary Figure 3.4b). These included a putative amidophosphoribosyl transferaseone, a pre-mRNA-processing-splicing factor, cytochrome P450 monooxygenase, eukaryotic translation initiation factor 6 and transcription initiation factor TFIID sub.10. The average length of E2-IEt2s was 186 ± 8 nt, but ranged up to 207 nt and a more degenerate potential Type 2 element was 144 nt long. As with E2-IEt1s, the ABC Transporter E2-IEt2 hit metagenomic reads from on or near Antarctica (Figure 3.3D). While E2-IEt1s were detected at locations without E2-IEt2s, E2-IEt2s were always collocated with E2-IEt1s, suggesting the latter are more abundant.

The ABC transporter ABC-IE also received multiple metagenomic read hits with flanking sequences of recruited reads encoding different proteins, e.g., a putative intraflagellar transport protein (IFT 140) and a putative Superfamily I helicase. ABC-IEs appeared to be less conserved than E2-IEs (Figure 3.3C, Supplementary Figure 3.5). Alignment of metagenomic and culture-derived sequences allowed a more robust evaluation of ABC-IE compositional features, than the single example from Clade C PCR products. These included a conserved 5'ss (GTYNGYTT), a branch point ((G)AcTGACRt), a pyrimidine track (YGTgTTTTgYY) and a 3'ss (YACAG). The Clade C ABC-IE central branch point positions were identical to those in *M.* RCC299 (Verhelst et al., 2013). Upstream of the branch point we also identified the shared motif YRGGCARTYA(G).

Micromonas Clade D IEs were present in metagenomic samples from temperate and tropical sites (Figure 3.3D). Metagenome searches for introns with high identity to the *Micromonas* Env. Clade D-like PAP-introns identified in Pacific Ocean actin PCR products (Figure 3.2B) returned zero hits (E-value cutoff e^{-5}). We conclude these represent a polymorphic RSI, not a new repetitive element example. Clear distribution patterns emerged for taxa containing ABC-IEs and E2-IEs, although little is otherwise known about their global distributions. While, E2-IEs appear restricted to polar environments, IEs and ABC-IEs were commonly detected in the same temperate water samples, supporting previous observations of co-location of taxa from clades A, B, C and D (Worden, 2006; Foulon et al., 2008). However, IEs were also detected in samples from lower latitudes. The lowest latitudes from which ABC-IEs were detected were significantly north and south of the tropics of Cancer and Capricorn, respectively. These results demonstrate the utility of repetitive-intronic elements in mapping *Micromonas* distributions using metagenomic data. Access to appropriate metadata could help refine our understanding of the ecological niches of these different clades.

The identification of E2-IEt1s and E2-IEt2s, as well as ABC-IEs, in metagenomic reads encoding proteins of diverse function confirms these are not neutrally-evolving (RSI) introns. These sequences represent new examples of repetitive, intronic elements similar to IEs in terms of presence across a genome, but unrelated at the sequence level with D-lineage IEs.

3.5 Discussion

Across the *Micromonas* clades, we found that only CCMP490 and environmental sequences contained the classical D-lineage IEs initially reported in Clade D strain *M.* CCMP1545. Remarkably, multiple *Micromonas* PAP-introns were identified, including RSIs and novel repetitive introns, in both cultured species and metagenomic sequences from wild *Micromonas*. Thus, we show that multiple *Micromonas* clades have repetitive introns, previously termed Introner Elements (Worden et al., 2009). However, because the repeat type, or types, is 'novel' in each major lineage and lacks homology to those of distant clades, each is designated according to its host-genome lineage and can be used to map *Micromonas* distributions using metagenomic data.

Approximately 200 repetitive introns were recently reported in *M*. RCC299 genes and termed IEC (Verhelst et al., 2013). The ABC-IE first observed in the Clade C PAP-intron sequences, and then in other *M*. CCMP1764 and *M*. RCC299 genes had high identity to a reported IEC example from yet another *M*. RCC299 gene (73% between the Clade C ABC transporter ABC-IE and reported Mipur01i11380 IEC; Supplementary Figure 3.1A). Verhelst et al., (2013) also retrieved and assembled our *M*. CCMP1764 data and noted similar sequences therein. Because of the pervasiveness of elements identified here, we have not adopted nomenclature from Verhelst and colleagues (Verhelst et al., 2013), which renamed IEs. Instead we have favored lineage-based IE naming and the original nomenclature for *M*. CCMP1545 elements (Worden et al., 2009).

The fact that Clade C repetitive introns were identified in cultured isolates from clades A and B, in addition to metagenomic samples from temperate Atlantic and Pacific sites, to just north of the Southern Ocean boundary (60^{o}S; Figure 3.3D), is not completely surprising. Clades C.I, A.II (containing RCC299) and Clade B.I group together in phylogenetic analyses of the 18S rRNA gene (Figure 3.1A) and other markers (Slapeta et al., 2006). In many 18S rRNA gene phylogenies clades A.II and B.I are not robustly separated and some strains shift between them. The presence of ABC-IEs in all three clades indicates these elements were present at the ancestral node.

Our results suggest heterogeneity in the ABC-IE landscape, although complete genome sequencing of representatives from all clades will be necessary to quantitatively evaluate distributions. The *M*. RCC299 ortholog of the ABC transporter (where we identified Clade C ABC-IEs), is a single exon gene (Figure 3.1E). Likewise, PCR products from other Clade A.II and B.I strains do not have introns. This suggests the Clade C.I element was present in the ancestor of these three clades, but later purged from this locus (in clades A.II and B.I) and that the clades may be undergoing a process of ABC-IE loss. Alternatively, this element could have been present in the A/B/C ancestor, but proliferated after these clades diverged (therefore not necessarily colonizing the same loci).

The geographic origins of E2-IEt1s and E2-IEt2s in metagenomic data led to the discovery of *Micromonas* in many Antarctic samples (Figure 3.3D). Our cultured representative of Clade E2, *M*. CCMP2099, is psychrophilic, known to grow from 0

to 12^{o}C, with maximal growth rates from 2-10^{o}C. Originally isolated from a North Water Polynya between Ellesmere Island and Greenland, it has now been reported frequently in the Canadian Arctic (Lovejoy et al., 2007; Li et al., 2009b). Lovejoy *et al*. (2007) proposed *M*. CCMP2099 was endemic to the Arctic and restricted to this region's semi-enclosed system with geographic barriers and ecological filters. However, protein-encoding portions of the metagenomic reads from the Antarctic were largely identical to transcripts from *M*. CCMP2099, as are the few E2-IEs that could be compared from these Polar Regions (Figure 3.3B).

If originally endemic to the Arctic, then how did Clade E2 *Micromonas* come to be present in the Antarctic as well? Slapeta *et al.* (2006) suggested that *Micromonas* may be circulated around the world in a low metabolic state by deep-sea currents (Slapeta et al., 2006). Although only asexual reproduction has been observed, the genomes of *M*. CCMP1545 and *M*. RCC299 contain meiotic recombination genes, putative sex chromosomes, and numerous genes encoding hydroxyproline-rich glycoproteins (HRGP), cell wall components implicated in diploid spore-forming in *Chlamydomonas reinhardtii* and plants (Worden et al., 2009). Such spores are frequently resistant to extreme conditions and might serve as a morphotype stable over long durations during transport in deep sea currents, such as the North Atlantic bottom waters formed in the Norwegian Sea and upwelled in the Southern Ocean.

In addition to discovering *Micromonas* in the Southern Ocean, we found E2-IEs in Antarctica's Ace Lake. This meromitic lake formed more than 9,200 years ago and

is thought to have undergone little change during the past 4,000 years (Fulford-Smith and Sikes, 1996). A limited input of seawater has caused some vertical mixing, but this input stopped ~ 5,500 year ago and the lake began to return to meromixis. Salinities of the Ace Lake samples, where we detected *Micromonas* E2-IEs, ranged down to 22 with temperatures of 0.42 – 1°C, while other Antarctic samples showed higher salinities, but lower temperatures. For example, in the Southern Ocean salinity was 33 to 34 and temperatures ranged down to -1.9°C. E2-IEs were also present in a 330 m sample, indicating *Micromonas* either sank out of the euphotic zone or can potentially assume a mixotrophic lifestyle, as recently reported for *M.* CCMP2099 (McKie-Krisberg and Sanders, 2014). *M.* CCMP2099 has been shown to reach growth rates of >0.4 d^{-1} at 2°C, with higher rates at the highest of three tested light levels. 18S rRNA gene sequences with 100% identity to this strain have been observed at Arctic sites with salinities from 27 to 34 (Lovejoy et al., 2007), but none as low as 22. Given the stability of Ace Lake, it seems likely that *Micromonas* are growing under these conditions. This suggests that results showing an increase in *Micromonas* in the Canadian Arctic correlated with reduced salinity (declined to ~29; associated with increased ice melt) (Li et al., 2009b) could potentially be amplified with further salinity reductions. Laboratory experiments with *M.* CCMP2099 should help define the salinity tolerance of this polar strain.

The fact that Clade E2 strain *M.* CCMP2099 contains two types of novel IEs in the ABC transporter gene yet none were found in Clade E1 may reflect under-sampling of *M.* CCMP1646. That is, if E2-IEs have a similarly heterogeneous

distribution as IEs and ABC-IEs they may be absent from the orthologs investigated here, but present in other *M*. CCMP1646 genes. However, the fact that E2-IEs were not detected in temperate (akin to *M*. CCMP1646 isolation site) or tropical metagenomic data suggests they are restricted to Polar Clade E2 (Figure 3.3D). This indicates E2-IEs were likely gained after E1 and E2 clades diverged or potentially contributed to their divergence.

IEs, the repetitive elements originally described in *M*. CCMP1545 (Worden et al., 2009), and shown in *M*. CCMP490 here, are thus far the best characterized. This is in part because a complete genome sequence is available for *M*. CCMP1545 and because IEs are so numerous (>6,000 in the *M*. CCMP1545 genome). The *M*. CCMP1545 IE groups range in length, but the most abundant group (6,112 members) is 173 nt on average (Worden et al., 2009; Verhelst et al., 2013), slightly shorter than the 192 nt average for 3,553 RSIs in *M*. CCMP1545. Clade D IEs are present in both the Atlantic and Pacific Oceans, including tropical waters (Figure 3.3D). It is presumed that IEs entered *Micromonas* Clade D post-divergence (~65MYA) from other clades, or contributed to its divergence. Notably, multiple Pacific environmental clones had 100% IE identity to those from the two North Atlantic *Micromonas* Clade D isolates. This was surprising because sequences thought to lack functional roles, such as introns, are considered neutrally evolving and these oceans have been separate for ~3 MY (and relevant populations within them presumably longer, given the circulation patterns in the regions where *M*. CCMP1545 and *M*. CCMP490 were isolated). The retrieval of IE-bearing *Micromonas* homologs differing from cultured

strains in the number of IEs possessed (Figure 3.2A, B) was initially surprising. However, together with polymorphic positions of ABC-IEs in cultured taxa, these findings suggest a heterogeneous landscape in the distribution of these elements throughout the genomes of extant taxa.

Collectively, Clade D IEs do not share sequence similarity with the new elements identified here. Moreover they are longer than ABC-IEs and E2-IEt1s, but shorter than E2-IEt2s (Figure 3.3B, Supplementary Figures 3.1, 3.4).

Although we show repetitive introns are present in non-homologous positions in different strains and environmental sequences (Figure 3.2A, B), we also observed introns at homologous positions. An intron-rich LECA would explain the occurrence of introns with homologous loci in gene orthologs from divergent eukaryotes. We observed a phase 0 intron with canonical splice sites in a deposited actin sequence from the prasinophyte *Pterosperma cristatum* (Figure 3.1A, Supplementary Figure 3.6). This intron was located at the same codon as the phase 2 *Micromonas* Clade D actin 3′ IE. If LECA derived, the ancestral intron would presumably have been replaced by an IE. We also found that the *M*. CCMP1545 homolog of the E2-IE-containing Antarctic ELIP gene, has two phase-2 IEs (Supplementary Figure 3.7). The *M*. CCMP1545 ELIP 3′ IE is at the same codon as the E2-IE, but the latter is at phase 0, rather than phase 2. An alternative hypothesis to LECA intron contributions is that some regions, or types of sequence composition, are predisposed to intron-insertion. Li et al., (2009a) reported parallel gains at homologous loci in independent allelic lineages of *Daphnia duplex*, similar to those seen here for our more divergent

Micromonas clades. The different intron phases observed support the idea of independent gains at these codons.

High intra-genomic conservation is suggestive of selective pressure and functional roles. Short, conserved, non-coding DNA sequences, known as CNSs (conserved non-coding sequences) are motifs that serve regulatory functions in plants and animals (Reineke et al., 2011). *Micromonas*-repetitive elements are longer than average plant (20-30 bp) and mammal (15-60 bp) CNSs, although some CNSs extend beyond 100 bp (Kaplinsky et al PNAS 2002, Freeling & Shabarinath Plant Bio 2009). Moreover, *Micromonas* repetitive elements are in genes with diverse functions, while CNSs are typically found in genes with upstream regulatory functions (Inada et al., 2003). CNSs also have a phylogenetic footprint, preserved between species (Lockton and Gaut, 2005), unlike repetitive introns from different *Micromonas* clades. *Micromonas* repetitive elements also lack the target site duplications and terminal inverted repeats of short interspersed elements (SINEs). If IEs were Class I retrotransposons there should be evidence of reverse transcriptases in the IE sequences and if they were Class II – DNA TEs the required enzymes should be encoded and repeated flanking sequences should be obvious. No transposable elements (TEs) were found in *M.* RCC299 and those located in *M*. CCMP1545 were degenerate or relic with little similarity to known TEs. Class II TE structural elements were completely absent from the genomes (Worden et al., 2009).

Availability of genomes from representatives of *Micromonas* clades will facilitate future determination of repetitive intron proliferation extent. This should

also enable motif development for recognizing these repetitive elements. In the future new nomenclature and groupings should be possible, once *Micromonas* repetitive elements can be defined according to characteristics other than sequence homology.

3.6 Conclusions

Intron gains were once thought to be rare events. The increase in available genome sequences and broader taxonomic sampling has revealed major exceptions to this. The *Micromonas* radiation provides an extraordinary case due to the number, variety and repetitive nature of polymorphic introns, which appear absent from other sequenced Mamiellophyceae. After analyzing strains representing different *Micromonas* lineages, environmental clone libraries from samples known to harbor the uncultured *Micromonas* Clade _.IV and metagenomes, we have laid a foundation for future research on the heterogeneity and functional implications of the *Micromonas* intron landscape. We hypothesize that invasion of distinct types of repetitive elements facilitated the divergence of extant *Micromonas* lineages from their last common ancestor. Differential loss of gene orthologs, due to the deleterious impact of invasive elements that disrupted splicing, could have reduced the core gene number, while exon shuffling, facilitated by phase 0 insertions, could have promoted establishment of new proteins contributing to the accessory genome, (at least those not shared with taxa outside the *Micromonas* lineage). We expect that as genome sequences become available for more branches of the eukaryotic tree of life, other cases of rampant invasion by repetitive intronic elements will be discovered. Together with analyses on potential functional roles of conserved elements, such studies will

facilitate a more comprehensive view of intron gain and its influence on eukaryotic diversity.

Table 3.1 *Micromonas* isolates grown and number of assembled sequences attained from clones for each gene ortholog investigated.

Culture	**Slapeta (Slapeta et al., 2006)**	**Worden (Worden, 2006)**	**Actin 48012**	**Calcium ATPase 36018**	**ABC Trans-porter 68853**	**NADH De-hydrogenase 26689**
RCC299		II	2	2	1*	2
CCMP492	A	II	2	2	0*	2
CCMP1764	B	I	2	2	2	2
NEPCC29	C	I	2	0†	2	2
CS222	C	I	2	2	2	2
CCMP1195	C	I	2	2	2	2
CCMP490	D	V	2	2	2	2
CCMP1545	D	V	2	2	2	2
CCMP1646	E	III	2	0‡	2	1
CCMP2099	E	III	0‡	0‡	2	2

*The primers consistently produced sequences from a different predicted ABC transporter in Clade A strain *M*. CCMP492 and in *M*. RCC299; the correct *M*. RCC299 gene homolog was obtained from the sequenced genome (but *M*. CCMP492 data was not included in analyses of this gene).

†Repeated attempts did not produce sequence from the correct calcium ATPase homolog in *M*. NEPCC29, but were successful for other Clade C strains.

‡Comparison to transcript sequences, attained later for Clade E isolates, revealed extensive primer mismatches for actin and the calcium ATPase in *M*. CCMP2099 and the calcium ATPase from *M*. CCMP1646, thus no DNA product was attained.

Table 3.2 Growth conditions for strains cultured and sequenced for this study. Additional details for media preparation can be found at: http://www.mbari.org/phyto-genome/Resources.html and recipes for K, L1 and F/2 are available in (Anderson, 2005). H_2SeO_3 was amended at 0.01 µM (final concentration). Abbreviations: ASW, artificial seawater; SS, Sargasso Sea water; MB Monterey Bay seawater; CCMP, Culture Collection of Marine Phytoplankton; RCC, Roscoff Culture Collection; NEPCC, North East Pacific Culture Collection; CS, CSIRO (Australian National Algae Culture Collection).

	Clade designations from various studies							
Culture	Guillou et al. 2004	Slapeta et al. 2006	Viprey et al. 2008	Worden et al. 2006	T (°C)	Media	Sea water	Amendments
CCMP1195	A	C	A.BC.1+A.A.2	I	16	F/2	MB	H_2SeO_3
CCMP1764	A	B	A.BC.1+A.A.2	I	21	L1	SS	
NEPCC29	A	C	A.BC.1+A.A.2	I	16	F/2	MB	H_2SeO_3
CS222	A	C	A.BC.1+A.A.2	I	21	F/2	MB	H_2SeO_3
CCMP492 (RCC451)	A	A	A.BC.1+A.A.2	II	21	L1	SS	
CCMP2706 (RCC299)*	n.a.	n.a.	n.a.	II	21	K	AS	
CCMP1646	B	E	B.E.3	III	21	K	AS	
CCMP2099	B	E	B.E.3	III	6	K	SS	
CCMP490	C	D	C.D.5	V	21	K	SS	
CCMP1545*	C	D	C.D.5	V	21	L1	SS	

*This isolate has been genome sequenced.

Table 3.3 PCR Primers designed for amplification of four gene homologs in *Micromonas* (also present in *Ostreococcus*) genomes. Protein IDs are for *M.* CCMP1545 and regions selected to span IEs were based on examination of the predicted gene model from the *M.* CCMP1545 genome. Accession numbers for these *M.* CCMP1545 models are XM_003062703.1, XM_003061058.1, XM_003060502.1, XM_003058664.1. Degenerate positions: B: C, G, T; M: A, C; N: A, C, G, T; R: A, G; S: C, G; Y: C.

	Primers (5′ to 3′)		**Product Length (bp)**	
Protein	**Forward**	**Reverse**	***M.* CCMP1545**	***M.* RCC299**
36028	GGTBCTCGCMGACGACAA	TCCAGCGGSACGATG	867	765
48012	TGGGACGACATGGAGAAGATC	ACGTACGCGAGCTTCTCCTT	791	422
68853	TGACGTGGCTCGARGAGTT	CGCTCGTCGTGCGANAC	1119	922
26689	CCSGGSACGTGCAARGA	CACTGBCCGCASGAYTCG	925	782

Figure 3.1

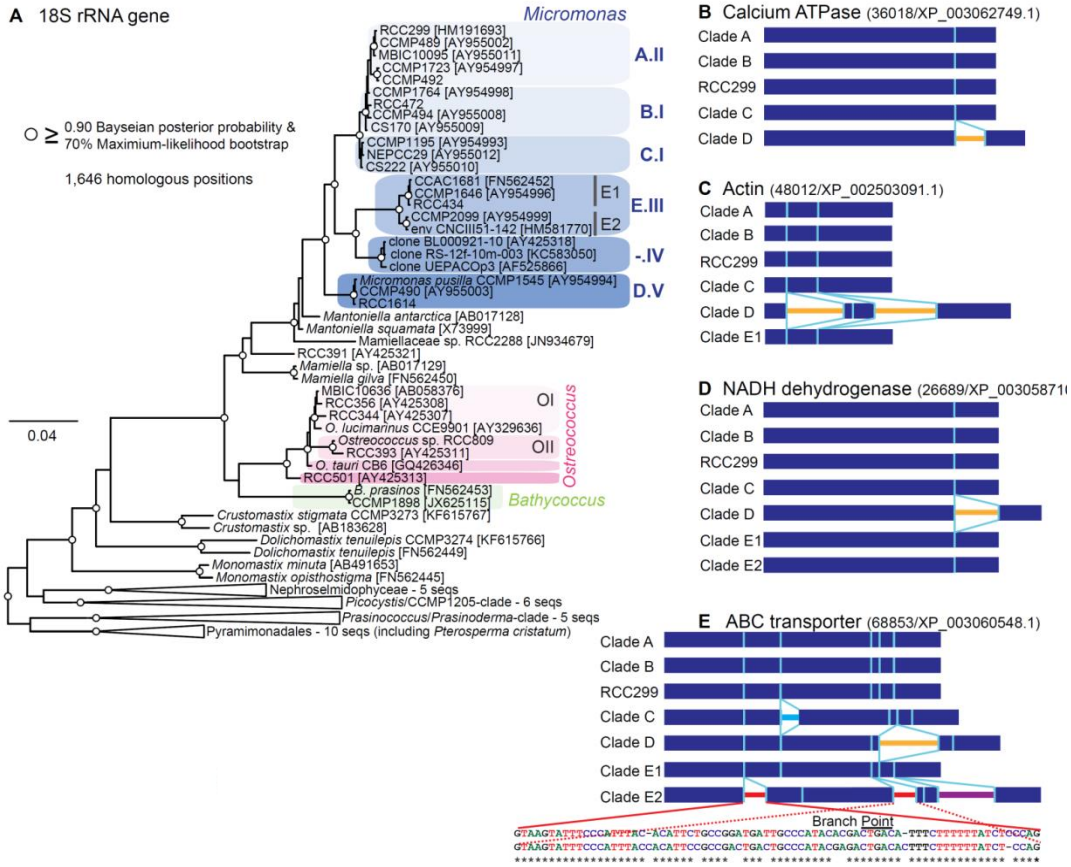

Figure 3.1 Molecular phylogeny of *Micromonas* lineages and insertion sequences in gene homologs from different clades. (A) This Bayesian reconstruction used nearly full length 18S rRNA gene sequences (1,646 positions) from Mamiellophyceae and other prasinophytes and the GTR+Γ+I model of substitution. Six *Micromonas* clades (blue), previously reported divergent at the species level (Slapeta et al., 2006), are highlighted. Clade names are designated with letters, as in (Slapeta et al., 2006) and roman numerals, as in (Worden et al., 2009), with an additional differentiation between clades E1 and E2, resolved here using new sequence data. Clade _.IV contains environmental sequences only. Other widespread Mamiellophyceae genera shown, *Ostreococcus* (pink) and *Bathycoccus* (green), also have genome sequenced representatives. The tree is rooted by the *Pycnococcus*-clade for display purposes. (B-E) Four protein-encoding genes were investigated using PCR in multiple representatives from cultured *Micromonas* clades (Table 3.1). Thick bars (blue) represent exons, vertical turquoise lines denote loci where introns were present (accompanied by thin horizontal intron lines) or absent (vertical line only). Thin horizontal lines represent Clade D IEs (yellow) and newly identified introns in Clade C (blue) and Clade E2 (red, purple) orthologs of the putative ABC transporter. High sequence identity (alignment in panel E) was observed between the first two Clade E2 introns (red).

Figure 3.2

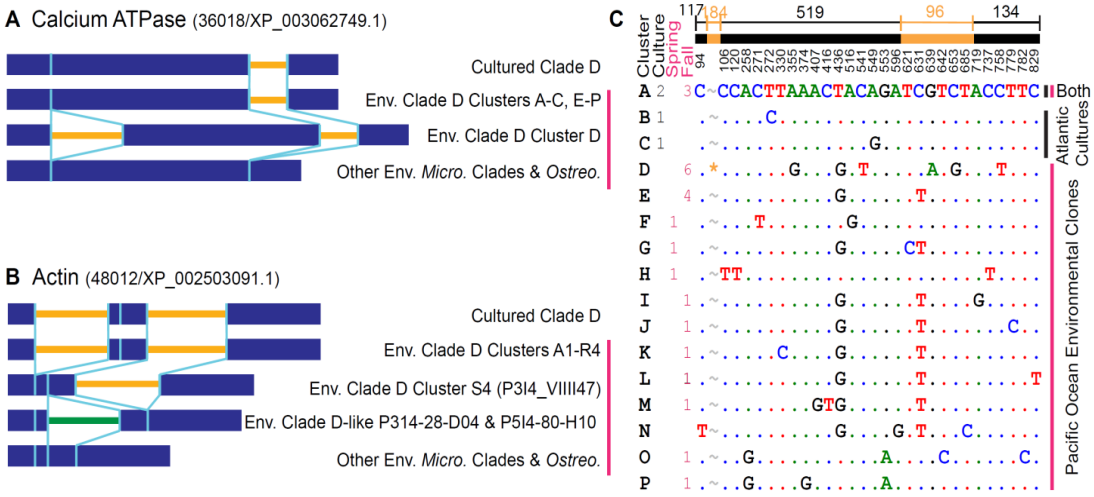

Figure 3.2 Intron presence/absence patterns in environmental clones. Architecture for partial regions of the genes encoding the (A) putative calcium ATPase and (B) actin are shown. Thick bars (blue) represent exons, vertical turquoise lines denote loci where introns were present (accompanied by thin horizontal intron lines) or absent (vertical line only). Thin horizontal lines represent Clade D IEs (yellow) and a newly identified PAP (green) in environmental clones belonging to Clade D. Calcium ATPase Cluster D consists of six environmental sequences, while actin Cluster S4 and the RSI-bearing Clade D-like type consist of one and two sequences, respectively. (C) Nucleotide variations in the amplified region of IE-bearing calcium ATPase homologs. Coding regions and IEs are indicated by black and orange bars, respectively, with the lengths shown above and positions with nucleotide variations below. Each cluster of identical sequences was assigned a letter, listed in the first column. Adjacent columns show the number of sequences per cluster derived from either cultures or environmental clones from spring or fall Eastern Pacific Ocean samples. Dots indicate identical nucleotides to those of the first sequence and letters denote nucleotide variations. Only sites with variants are shown and numbering below the bar corresponds to position in the overall sequence. The asterisk (orange) represents a 5′ IE (184 nt) in Cluster D sequences, lacking in all other calcium ATPase sequences (variant nucleotide numbering does not include the Cluster D 5′ IE). Cluster D 5′ and 3′ IEs had 83% identity to each other. Cluster A consisted of three environmental clones and two *Micromonas* Clade D culture sequences (from *M.* CCMP1545 and *M.* CCMP490). The second calcium ATPase sequences from these strains made up Clusters B and C, respectively, and may represent PCR artifacts rather than actual nucleotide variations. Results for actin are in Supplementary Figure 3.2.

Figure 3.3

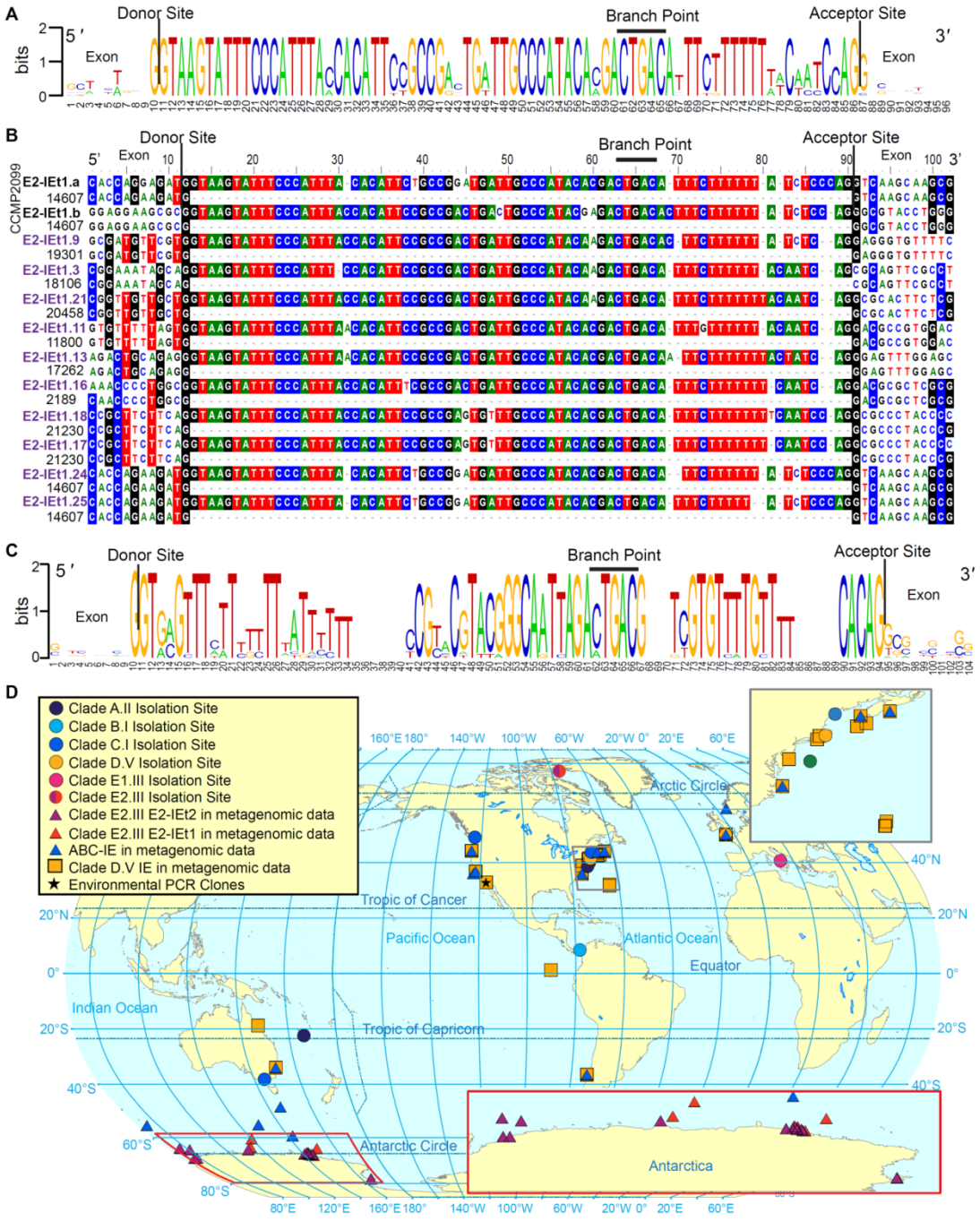
A
Donor Site
Branch Point
Acceptor Site
bits
5′
3′
Exon
Exon
B
Donor Site
Branch Point
Acceptor Site
5′
Exon
Exon
3′
CCMP2099
E2-IEt1.a
14607
E2-IEt1.b
14607
E2-IEt1.9
19301
E2-IEt1.3
18106
E2-IEt1.21
20458
E2-IEt1.11
11800
E2-IEt1.13
17262
E2-IEt1.16
2189
E2-IEt1.18
21230
E2-IEt1.17
21230
E2-IEt1.24
14607
E2-IEt1.25
14607
C
Donor Site
Branch Point
Acceptor Site
bits
5′
3′
Exon
Exon
D
Clade A.II Isolation Site
Clade B.I Isolation Site
Clade C.I Isolation Site
Clade D.V Isolation Site
Clade E1.III Isolation Site
Clade E2.III Isolation Site
Clade E2.III E2-IEt2 in metagenomic data
Clade E2.III E2-IEt1 in metagenomic data
ABC-IE in metagenomic data
Clade D.V IE in metagenomic data
Environmental PCR Clones
Arctic Circle
Tropic of Cancer
Pacific Ocean
Atlantic Ocean
Equator
Indian Ocean
Tropic of Capricorn
Antarctic Circle
Antarctica

Figure 3.3 The global distribution of *Micromonas* introns in metagenomic data and the discovery of conserved elements in Polar *Micromonas*. (A) Consensus sequence for E2-IEt1. The consensus represents intron data from Antarctic metagenomic reads encoding eight different proteins. (B) Aligned Antarctic reads, as well as the two E2-IEt1 sequences from the calcium ATPase in *M.* CCMP2099, and metagenomic reads encoding the same protein, but coming from different samples (2 examples; excluded from A to avoid over-representing element conservation levels). Aligned *M.* CCMP2099 transcripts demonstrate splicing. In the case of the ABC transporter, regions that flanked the Arctic *M.* CCMP2099 E2-IEt1.a and Antarctic metagenomic E2-IEt1.24 and E2-IEt1.25 (from the same ABC transporter but from different Antarctic samples) were identical. Arctic E2-IEt1.a is identical to E2-IEt1.24 and E2-IEt1.25, with the exception of nucleotide variations (a single "t" each, but at different positions in E2-IEt1.24 and E2-IEt1.25, potentially representing homopolymer accuracy issues in the 454-sequencing platform). Consensus sequences and alignments for E2-IEt2 are show in Supplemental Figure 3.4. (C) Consensus sequence for *Micromonas* Clade C elements, termed ABC-IEs after identification in other genes from clade A and B *Micromonas* strains (Supplementary Figure 3.1). (D) Locations of isolation sites for cultured *Micromonas* strains (circles) and sample site for environmental clone libraries (star). Sites where multiple BLASTN hits were recovered in publically available metagenomic data (CAMERA "all metagenomic 454" dataset as of 1 March 2014 using elements from cultures as queries) are color-coded as indicated on legend. Boarders to insets are color-coded to show expanded regions on map. Note that for Clade E2, beneath every purple triangle (representing E2-IEt2 sequences) is a red triangle (E2-IEt1).

Chapter 3 Supplementary Figure Legends

Supplementary Figure 3.1 (A) Alignment of the repetitive intron in the Clade C *M.* NEPCC29 ABC transporter (ABC-IE) to genomic data from *M.* CCMP1764 (Clade B). For the latter 15 flanking nucleotides that encode up and downstream exons (of different proteins) are also shown. *M.* CCMP1764 reads were selected to reflect hits with lowest E-values (e^{-05} to highest, e^{-08}; influenced but the length of the gDNA read). (B) Alignment of ABC-IE from the *M.* NEPCC29 ABC-transporter (and present in other Clade C strains) to the two best hits (according to E-value) in *M.* RCC299, specifically JGI Prot ID 55550 and 105055, as well as an example IEC sequence "Mipur" provided in (Verhelst et al., 2013) and a representative metagenomic sequence (CAM READ) identified here. (C) 2D structure prediction for the ABC-IE identified in the *M.* NEPCC29 ABC transporter.

Supplementary Figure 3.2 Nucleotide variations in the amplified region of IE-bearing actin homologs. Coding regions and IEs are indicated by black and orange bars, respectively, with the lengths shown above and variant-positions below. Each cluster of identical sequences was assigned a letter, listed in the first column. Cluster A1 consists of a *M.* CCMP1545 sequence and 12 environmental sequences. Dots indicate identical nucleotides to those of the first sequence, and letters denote nucleotide variations; only sites with variants are shown and numbering below bar corresponds to position in the overall sequence.

Supplementary Figure 3.3 Maximum-likelihood phylogenetic reconstruction of the calcium ATPase with representative environmental sequences. 734 positions from Mamiellophyceae and other prasinophytes were used with the TVM+G model. Sequence identifiers starting with P3 are from spring, and P5 are from fall; however, they represent clusters that may contain sequences from both.

Supplementary Figure 3.4 Alignment of *M.* CCMP2099 E2-IEt2 retrieved from metagenomic data. (A) The consensus sequence is derived from metagenomic reads encoding seven different proteins (sequences 1, 3, 4, 5, 6, and 8). Letters are shown above for nucleotides in which at least four consecutive nucleotides were identical in all sequences. (B) The alignment contains ten DNA sequences (bold) and matching transcript data (denoted MMETSP…) that demonstrate effective splicing of the E2-IEt2. The first and last 15 positions represent CDS, not intronic sequence. In addition to sequences included in (A) is the *M.* CCMP2099 ABC transporter E2-IEt2 and two more divergent E2-IEt2s found in metagenomic reads (bottom two). Sequences 1 and 2 came from the same gene, but were collected from different sites (and have nucleotide variations in the E2-IEt2 sequence). All metagenomic reads recovered came from Antarctica (Southern Ocean, Ross Sea, Ace Lake).

Supplementary Figure 3.5 Alignment of ABC-IEs recovered in metagenomic data. The first sequence shown is the *M.* NEPCC29 ABC transporter ABC-IE sequence (with flanking nucleotides). DNA-derived sequences (bold names) and matching transcripts from *M.* NEPCC29 provide evidence of splicing. The first and last ten

positions represent CDS, not the intronic sequence. Transcript contig numbers are provided from MMETSP1386-20130603 (harvested at 5:30am, two hours before lights on) or, where marked with an asterisk, MMETSP1082-20130531 (harvested at 9:30am, two hours after lights on).

Supplementary Figure 3.6 Comparison of intron presence absence patterns in actin homologs from *Micromonas* Clade D, environmental clones and *P. cristatum*, all of which have an element at a homologous locus within this gene. Thick blue bars represent exons, while thin yellow lines represent Clade D IEs. The blue line in the *P. cristatum* actin model represents a regular spliceosomal intron (RSI). Vertical turquoise lines denote loci where elements were detected in environmental clones. Numbers of environmental sequences in each gene structure type are provided in Supplementary Dataset 3.1. Without a sequenced genome it is unclear whether the *P. cristatum* intron represents an RSI or a new type of repetitive intron.

Supplementary Figure 3.7 Insertion sites for repetitive elements in an Early Light Induced Protein (ELIP). *M.* RCC299 and *M.* CCMP1764 do not contain introns in the region analyzed. The *M.* CCMP2099 transcript and Cam_Read are most similar to each other (the transcript is identical to the exonic sequence and is therefore not shown in the second portion of the alignment). Insertion positions for the IE (*M.* CCMP1545) and E2-IEt1 (from the Antarctic metagenomic read) sequences are shown.

References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389-3402.

Anderson, R. (2005) *Algal culturing techniques*. San Francisco: Elsevier Academic Press.

Crooks, G.E., Hon, G., Chandonia, J.M., and Brenner, S.E. (2004) WebLogo: a sequence logo generator. *Genome Res* **14**: 1188-1190.

Csuros, M., Rogozin, I.B., and Koonin, E.V. (2011) A detailed history of intron-rich eukaryotic ancestors inferred from a global survey of 100 complete genomes. *PLoS Comput Biol* **7**: e1002150.

Davis, L.G. (1986) Plasmid "Mini-Prep" Method. In *Basic Methods in Molecular Biology*. Davis, L.G., Dibner, M.D., and Battey, J.F. (eds): Elsevier Science Publishing Co, Inc, pp. 102-104.

Denoeud, F., Henriet, S., Mungpakdee, S., Aury, J.M., Da Silva, C., Brinkmann, H. et al. (2010) Plasticity of animal genome architecture unmasked by rapid evolution of a pelagic tunicate. *Science* **330**: 1381-1385.

Foulon, E., Not, F., Jalabert, F., Cariou, T., Massana, R., and Simon, N. (2008) Ecological niche partitioning in the picoplanktonic green alga Micromonas pusilla: evidence from environmental surveys using phylogenetic probes. *Environ Microbiol* **10**: 2433-2443.

Frayling, T.M., Timpson, N.J., Weedon, M.N., Zeggini, E., Freathy, R.M., Lindgren, C.M. et al. (2007) A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* **316**: 889-894.

Fulford-Smith, S.P., and Sikes, E.L. (1996) The evolution of Ace Lake, Antarctica, determined from sedimentary diatom assemblages. *Palaeogeography, Palaeoclimatology, Palaeoecology* **124**: 73-86.

Gilbert, W. (1978) Why genes in pieces? *Nature* **271**: 501.

Inada, D.C., Bashir, A., Lee, C., Thomas, B.C., Ko, C., Goff, S.A., Freeling, M. (2003) Conserved noncoding sequences in the grasses. *Genome Res.* **13**: 2030-2041.

Jobb, G., von Haeseler, A., and Strimmer, K. (2004) TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. *BMC Evol Biol* **4**: 18.

Katoh, K., Kuma, K., Toh, H., and Miyata, T. (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res* **33**: 511-518.

Keeling, J.P. *et al*. Community page. (2014) The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. *PLoS Biology* in press.

Koonin, E.V. (2006) The origin of introns and their role in eukaryogenesis: a compromise solution to the introns-early versus introns-late debate? *Biol Direct* **1**: 22.

Li, W., Tucker, A.E., Sung, W., Thomas, W.K., and Lynch, M. (2009a) Extensive, recent intron gains in Daphnia populations. *Science* **326**: 1260-1262.

Li, W.K.W., McLaughlin, F.A., Lovejoy, C., and Carmack, E.C. (2009b) Smallest algae thrive as the Arctic Ocean freshens. *Science* **326**: 539.

Lockton, S., and Gaut, B.S. (2005) Plant conserved non-coding sequences and paralogue evolution. *Trends Genet* **21**: 60-65.

Lovejoy, C., Vincent, W.F., Bonilla, S., Roy, S., Martineau, M.J., Terrado, R. et al. (2007) Distribution, phylogeny, and growth of cold-adapted picoprasinophytes in arctic seas. *Journal of Phycology* **43**: 78-89.

Mathews, D.H., Moss, W.N., and Turner, D.H. (2010) Folding and finding RNA secondary structure. *Cold Spring Harb Perspect Biol* **2**: a003665.

McKie-Krisberg, Z.M., and Sanders, R.W. (2014) Phagotrophy by the picoeukaryotic green alga *Micromonas*: implications for Arctic Oceans. *The ISME Journal* **advance online publication 20 February 2014**.

Modrek, B., and Lee, C. (2002) A genomic view of alternative splicing. *Nat Genet* **30**: 13-19.

Not, F., Latasa, M., Scharek, R., Viprey, M., Karleskind, P., Balague, V. et al. (2008) Protistan assemblages across the Indian Ocean, with a specific emphasis on the picoeukaryotes. *Deep-Sea Research Part I-Oceanographic Research Papers* **55**: 1456-1473.

Philippe, H. (1993) MUST, a computer package of Management Utilities for Sequences and Trees. *Nucleic Acids Res* **21**: 5264-5272.

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

Reineke, A.R., Bornberg-Bauer, E., and Gu, J. (2011) Evolutionary divergence and limits of conserved non-coding sequence detection in plant genomes. *Nucleic Acids Res* **39**: 6029-6043.

Robbens, S., Rouze, P., Cock, J.M., Spring, J., Worden, A.Z., and Van de Peer, Y. (2008) The FTO gene, implicated in human obesity, is found only in vertebrates and marine algae. *J Mol Evol* **66**: 80-84.

Rogozin, I.B., Carmel, L., Csuros, M., and Koonin, E.V. (2012) Origin and evolution of spliceosomal introns. *Biol Direct* **7**: 11.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D.L., Darling, A., Hohna, S. et al. (2012) MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol* **61**: 539-542.

Roy, S.W. (2003) Recent evidence for the exon theory of genes. *Genetica* **118**: 251-266.

Roy, S.W. (2006) Intron-rich ancestors. *Trends Genet* **22**: 468-471.

Roy, S.W., and Gilbert, W. (2006) The evolution of spliceosomal introns: patterns, puzzles and progress. *Nat Rev Genet* **7**: 211-221.

Slapeta, J., Lopez-Garcia, P., and Moreira, D. (2006) Global dispersal and ancient cryptic species in the smallest marine eukaryotes. *Molecular Biology and Evolution* **23**: 23-29.

Sun, S., Chen, J., Li, W., Altintas, I., Lin, A., Peltier, S. et al. (2011) Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis: the CAMERA resource. *Nucleic Acids Res* **39**: D546-551.

Sverdlov, S.V., Rogozin, I.B., Babenko, v.N., and Koonin, E.V. (2007) Conservation versus parallel gains in intron evolution. *Nucleic Acids Res* **33**: 1741-1748.

Torriani, S.F., Stukenbrock, E.H., Brunner, P.C., McDonald, B.A., and Croll, D. (2011) Evidence for extensive recent intron transposition in closely related fungi. *Curr Biol* **21**: 2017-2022.

van der Burgt, A., Severing, E., de Wit, P.J., and Collemare, J. (2012) Birth of new spliceosomal introns in fungi by multiplication of introner-like elements. *Curr Biol* **22**: 1260-1265.

Venter, J., Remington, K., Heidelberg, J., Halpern, A., Rusch, D., Eisen, J. et al. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**: 66-74.

Verhelst, B., Van de Peer, Y., and Rouze, P. (2013) The complex intron landscape and massive intron invasion in a picoeukaryote provides insights into intron evolution. *Genome Biol Evol* **5**: 2393-2401.

Worden, A.Z. (2006) Picoeukaryote diversity in coastal waters of the Pacific Ocean. *Aquatic Microbial Ecology* **43**: 165-175.

Worden, A.Z., Nolan, J.K., and Palenik, B. (2004) Assessing the dynamics and ecology of marine picophytoplankton: The importance of the eukaryotic component. *Limnology and Oceanography* **49**: 168-179.

Worden, A.Z., Lee, J.H., Mock, T., Rouze, P., Simmons, M.P., Aerts, A.L. et al. (2009) Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes *Micromonas*. *Science* **324**: 268-272.

4 Chapter 4: Biogeography of photosynthetic picoeukaryotes in the North Pacific Ocean

4.1 Abstract

Eukaryotic algae within the picoplankton size class (<2 μm diameter) are important marine primary producers. In 2009-2010, field work was conducted in the North Pacific along the historic California Cooperative Fisheries Investigations (CalCOFI) and Long Term Ecological Research (LTER) transect Line-67 focusing on both a broad survey of picophytoplankton populations and a semi-Lagrangian experiment. Small eukaryotes (pico and ultraplankton) and *Synechococcus* reached maximum abundances of 1.44×10^5 and 3.37×10^5 cells ml^{-1}, respectively, in mesotrophic waters, while *Prochlorococcus* reached 1.95×10^5 cells ml^{-1} at 75 m in the open ocean (OO). Quantitative PCR showed the picoprasinophyte *Bathycoccus* was present at all stations investigated, reaching 21,368 ± 327 18S rRNA gene copies ml^{-1}. *Micromonas* and *Ostreococcus* Clade OI were present in colder more nutrient rich and coastal waters, while *Ostreococcus* Clade OII was present at low levels in the OO. *Micromonas* and *Ostreococcus* Clade OI always co-occurred in our samples. To resolve putative *Bathycoccus* ecotypes we established genetic distances for 1,104 marker genes. The results indicated the existence of two ecotypes, named here BI (represented by coastal isolate *Bathycoccus prasinos*) and BII (an uncultured, oceanic type). Exploratory analyses of relative gene expression in *Bathycoccus* (and *Ostreococcu*s OI) showed differential use of NH_4 and NO_3 transporters. We also identified a nickel super oxide dismutase encoding gene is present in all of the genome sequenced picoprasinophytes, the *Bathycoccus* version of which was

expressed most highly in the OO. Finally, unlike *Ostreococcus* Clades OI and OII which rarely co-occur, marker gene analysis of the metatranscriptomes indicated that *Bathycoccus* ecotypes co-occur in mesotrophic and oligotrophic environments.

4.2 Introduction

Marine phytoplankton are responsible for roughly half of global net primary production (Field et al., 1998) and picoeukaryotes play important roles in this production (Li, 1994; Worden et al., 2004; Grob et al., 2007; Cuvelier et al., 2010; Jardillier et al., 2010). However, picoeukaryotes are diverse with distinguishing morphological features that are limited and difficult to visualize (Lopez-Garcia et al., 2001; Worden et al., 2006; Viprey et al., 2008). Thus, efficiently resolving different picoeukaryotic taxa in nature remains a challenge hampering studies of the complex biological, physical and chemical forces influencing picoeukaryote-dynamics.

One widespread class of picoeukaryotes is the Mamiellophyceae, green algae that belong to the prasinophytes. The Mamiellophyceae genera *Micromonas* and *Ostreococcus* have a number of clades or ecotypes represented in culture, which are thought to represent species level divisions (Slapeta et al., 2006; Worden, 2006; Palenik et al., 2007; Worden et al., 2009). Genome analyses have demonstrated extensive diversity within *Micromonas*, and lesser, but marked, diversity between *Ostreococcus* species. For a third Mamiellophyceae genera, *Bathycoccus*, only one cultured strain (Bban7, also known as RCC1105) has been genome sequenced (Moreau et al., 2012), in part due to the notion that this genus consists of a single species, *Bathycoccus prasinos*. Cultured *Bathycoccus* strains and environmental

clones have 100% 18S rRNA gene identity, unlike the various clades defined within the genera *Micromonas* and *Ostreococcus* (Guillou et al., 2004; Worden, 2006; Monier et al., 2012).

In addition to culture based approaches, molecular and metagenomic methods have been essential for gaining insight into Mamiellophyceae taxa and distributions. Following the analysis of targeted metagenomic data from the tropical Atlantic it was proposed that different ecotypes may exist in the *Bathycoccus* genus as well (Monier et al., 2012; Monier et al., 2013). Monier et al., (2013) recovered sequences from a tropical Atlantic wild *Bathycoccus* population with insert-bearing *PRP8* genes that had either an intron or intein at the same insertion position, and nearly identical exonic sequence. However, these environmental insert-bearing *PRP8* genes had extensive third codon variation from *PRP8* genes of cultured *Bathycoccus* strains and the latter were insert-less. Insert-bearing *Bathycoccus PRP8* sequences were recovered from oligotrophic, open ocean waters, while insert-less sequences were primarily from mesotrophic waters or coastal isolates. Additionally, the Internal Transcribed Spacer (ITS) from the tropical Atlantic wild *Bathycoccus* metagenome branched separately from cultured strains (Monier et al., 2013), while ITS from cultured strains and coastal Chilean *Bathycoccus* metagenomes were very similar (Vaulot et al., 2012). Ultimately, results from these targeted metagenomic studies led to hypotheses that two (Monier et al., 2013) or more (Vaulot et al., 2012) *Bathycoccus* ecotypes exist. However, to date little is known about overall genetic distances and whether they correspond to ecotype-level differences. More generally,

relatively little data are available on the natural abundances of Mamiellophyceae taxa (Zhu et al., 2005; Marie et al., 2006; Demir-Hilton et al., 2011; Treusch et al., 2012).

Here, we investigated Mamiellophyceae genera in several environmental regimes. First, we established the abundance of the three major picophytoplankton groups, picoeukaryotes and the picocyanobacteria *Prochlorococcus* and *Synechococcu*s, from two cruises in the eastern North Pacific Ocean. We then used qPCR to enumerate Mamiellophyceae taxa in coastal, mesotrophic and open ocean sites along the cruise route and sequenced metatranscriptomes from each of these regions. To further ecotype resolution we established genetic divergence levels for both *Ostreococcus* and *Bathycoccus* ecotypes. This information was then applied to the exploration of ecotype distributions within the context of the eastern North Pacific Ocean phytoplankton community. Our study provides an exploration of Mamiellophyceae-gene expression and new insights into the natural distributions of several Mamiellophyceae ecotypes.

4.3 Materials and Methods

4.3.1 Field sampling

The Eulerian survey was carried out as part of the Activities and Dynamics cruise in 2009 (WFAD09) along the historic California Cooperative Fisheries Investigations and Long Term Ecological Research transect Line-67, while the semi-Lagrangian survey was part of the Controlled, Agile and Novel Observing Network expedition in 2010 (CANON10), which followed a drifting Environmental Sample Processor (ESP) (Ottesen et al., 2013) and used ship-based and remote efforts to

repeatedly sample the same water mass starting near Line-67. ESP collection depths ranged from 23 to 25 m. Our analysis of σT and temperature show that, of the 12 samples reanalyzed herein, three came from within the mixed layer and the others were from 10 to 20 m below the mixed layer (Supplementary Figure 4.1). Both expeditions were on the RV Western Flyer, originating from the central Californian Coast near Monterey (USA), at the same time of year (boreal autumn). Methods and results for NO_3 and NH_4 were published as part of a previous study (Santoro et al., 2010). Samples for nutrient analyses, flow cytometry and qPCR were collected using Niskin bottles mounted on a rosette with a CTD and fluorometer. For WFAD09 metatranscriptomics, 50 l samples were collected at sites 25 km, 172 km and 785 km from shore (Table 4.1), gravity filtered through 20 μm Nitex mesh and subsequently onto a 0.8 μm pore size 142 mm diameter Supor filter (Pall Scientific, Port Washington NY, USA), using a peristaltic pump. Bisected filters were transferred to 50 ml tubes and frozen at -80°C approximately 1 hour after CTD retrieval.

4.3.2 Flow cytometry

Flow cytometry samples were preserved in 3 ml aliquots with 30 μl of EM grade 25% gluteraldehyde and aliquoted into three 1.2 ml cryovials. Samples were fixed at room temperature in the dark for 20 min and then frozen in liquid N_2, where they remained until analysis. Prior to analysis, samples were thawed in the dark and Yellow Green fluorescent polystyrene beads (0.75 μm diameter, Polysciences) were added as standards. Samples were analyzed at approximately 25 μl min^{-1} using PBS sheath fluid on a Becton Dickinson (Franklin Lakes, NJ, USA) Influx flow cytometer

as described previously (Cuvelier et al., 2010). Winlist 6.0 and 7.1 (Verity Software House) were used to analyze listmodes and characterize *Prochlorococcus*, *Synechococcus* and small eukaryote (pico and ultraplankton size classes) populations, defined based on Forward Angle Light Scatter and autofluorescence from photosynthetic pigments (Chisholm et al., 1986; Olson et al., 1991).

4.3.3 QPCR

Bathycoccus, *Micromonas* and *Ostreococcus* Clades OI and OII were enumerated using qPCR assays (Demir-Hilton et al., 2011). Triplicate reactions, along with inhibition tests and no-template control reactions, were performed in 25 µl volumes. A master mix was used, consisting of 12.5 µl Taqman Universal Master Mix (Applied Biosystems Foster City, CA, USA), 2.5 µl of both forward and reverse primers (900 nM final concentration), 2.5 µl of probe (250 nM final concentration) and 3 µl of water. For inhibition tests, 2 µl of DNA template was added per reaction, as was an additional 2 µl of 18S plasmid (0.5×10^4 or 0.5×10^5 copies ul^{-1}) and water was reduced to 1 µl. Based on inhibition test results, environmental template solutions were diluted between 1:4 and 1:40 to prevent inhibition. Thermal cycling conditions consisted of 10 min at 95°C (initial denaturation), followed by 45 cycles of 15 sec at 95°C and 1 min at 60°C, using an AB7500 (Applied Biosystems); data were collected during the annealing phase. Ten-fold plasmid serial dilutions were used to generate standard curves. Threshold and baseline values were calculated using AB7500 software (Foster City, CA, USA). Copy numbers per reaction were calculated using the linear regression of Ct values against log scale copy numbers of standards and

converted to 18S rRNA gene copies ml^{-1} based on the volume of seawater extracted (usually 1,000 ml) and amount of template used. 100% extraction efficiency was assumed. *B. prasinos* RCC1105, *O. lucimarinus*, *O.* RCC809 and *M.* CCMP1545 each have two 18S rRNA gene copies per genome, while *O. tauri* appears to have one and *M.* RCC299 has three (Derelle et al., 2006; Palenik et al., 2007; Demir-Hilton et al., 2011; Moreau et al., 2012).

4.3.4 Metatranscriptome library construction and sequencing

Because metagenomes for the picoplankton size fraction can be dominated by prokaryotic sequences, for the Eulerian samples we sequenced polyA selected RNA to avoid low eukaryotic read count issues (Worden and Allen, 2010; Worden et al., 2011). These metatranscriptomes were generated from 10 m at the coastal and TZ sites, and from 15 m (OOsurf) and 105 m depths (OODCM) at the OO site. The semi-Lagrangian samples were generated as described in (Ottesen et al., 2013).

RNA extraction supplies were obtained from Life Technologies (Grand Island, NY, unless otherwise noted). Buffers were prepared with nuclease-free H_2O and filter-sterilized prior to use. Frozen 0.8 μm filters were transferred to sterile petri dishes, covered with ~2 ml of lysis buffer (see below) and sliced into ~1 cm^2 pieces. The filter pieces and buffer were transferred to polypropylene tubes and lysis buffer (5 ml of RNAlater, 25% sucrose, 2.5 mg ml^{-1} lysozyme, 5 mM tris(hydroxymethyl) aminomethane pH 8 and 27.5 $mmol \cdot L^{-1}$ each of ethylenediaminetetraacetic acid and ethylene glycol tetraacetic acid; Sigma, St. Louis MO, USA) was added to a volume of 20 ml, and incubated for 1 hour at 37°C. Four mg of proteinase K (Qiagen,

Valencia, CA, USA) was added and samples were frozen (liquid N_2) and thawed (55°C) three times. Subsequently, 4 mg Proteinase K and sodium dodecyl sulfate (Sigma, St. Louis MO, USA) was added to 1% (volume/volume). Samples were incubated at 55°C for 2 hours under agitation, and centrifuged for 2 min at 4500 x *g*. Supernatants were transferred to fresh tubes and filter pieces stored at -80°C for later re-extraction. Extractions used pH 8 phenol and nucleic acids were recovered by isopropanol precipitation following standard protocols (Sambrook, 2001). Pellets were re-suspended in 1 ml RNA extraction buffer (4 M guanidine thiocyanate, 25 mM sodium citrate, 0.5 % sarkosyl; Sigma, St. Louis MO, USA) and acidified with 50 µl 2 $mol \cdot L^{-1}$ sodium acetate, pH 4 (Sigma). The resulting solutions were extracted with acidic pH 4 phenol, chloroform and isoamyl alcohol (ratios of 125:24:1 respectively; Sigma) and RNA was precipitated with isopropanol (Sambrook, 2001). Pellets were washed twice with 75% ethanol and re-suspended in water. RNA was purified using the RNeasy kit and the RNase-free DNase kit (Qiagen, Valencia CA, USA). The entire procedure was repeated using the remaining filter pieces and the resulting two RNA aliquots per sample were combined, yielding 1.8 to 12.3 µg of total RNA from ~25 l of seawater. 500 ng aliquots of RNA were amplified in two rounds of *in-vitro* transcription (yielding 237 µg of antisense RNA) using the Message AMP II aRNA Amplification kit following the manufacturer's protocol except for using the primer T7-BpmI $_{16}$VN (5′-GCCAGTGAATTGTAATACGACTCACTATAGGGGCGACTGGAGTTTTTTTTTTTTTTTTTVN-3′ instead of the supplied T7 Oligo (Stewart et al., 2012). 60 µg aRNA

was converted to single stranded cDNA using the Superscript III First Strand Synthesis System (Invitrogen, Carlsbad CA, USA) with random hexamer primers. Double stranded cDNA was produced by incubating DNA for 2 hours at 16°C with 40U of DNA polymerase I, 10U of DNA ligase and 200 $\mu mol \cdot L^{-1}$ dNTPs in 5x second-strand buffer. RNA was digested with 2U RNase H. cDNAs were blunt-ended by adding 5U T4 DNA polymerase and incubating at 16°C for 5 min. Purification used Agencourt AMPure XP magnetic beads (Beckman Coulter, Indianapolis IN, USA). The resulting material was size-selected (300-3000 bp) on a low-melting point, 1% agarose gel and extracted using the QIAquick Gel Extraction kit (Qiagen, Valencia CA, USA). Poly(A) tails were removed by digestion with BpmI (NEB, Ipswich MA, USA), according to manufacturer's protocol, and cDNA purified again with AMPure XP beads. cDNA quantity (1.2 to 1.5 μg) and quality was assessed with Qubit and Bioanalyzer (Agilent, Santa Clara CA, USA). Sequencing was performed on a 454 FLX+ instrument using Titanium chemistry (Roche/454, Branford CT, USA).

4.3.5 Metatranscriptome analyses

Twelve of 13 published semi-Lagrangian metatranscriptomes (Ottesen et al., 2013) from 4 hour intervals starting at 14:00 on 16 September 2010 at 36°2.712, -123°1.302 and ending 44 hours later at 35°47.412, -122°42.54 were reanalyzed here. Reanalysis was necessary because the initial publication (Ottesen et al., 2013) did not include reference genomes from *Bathycoccus* and several other relevant phytoplankton. Semi-Lagrangian reads were generated using methods for prokaryotic

transcript recovery and fewer reads were generated per sample than in the Eulerian survey.

For these and the Eulerian metatranscriptomes redundant sequences, considered to be 454-artifacts (Gomez-Alvarez et al., 2009), were removed using CD-Hit v4.6 (Li and Godzik, 2006) and rRNAs were removed using Ribopicker (Schmieder et al., 2012), both with default settings. 454-reads were assigned to taxonomic groups using BLASTX (Altschul et al., 1997) against predicted proteomes from sequenced genomes and culture-based transcriptomes from the following taxa: *A. anophagefferens*, *B. prasinos*, *B. natans*, *C. reinhardtii*, *E. siliculosus*, *E. huxleyi*, *G. theta*, *M.* RCC299, *M.* CCMP1545, *M.* CCMP2099, *O. tauri*, *O. lucimarinus*, *O.* RCC809, *P. parkeae*, *P. tricornutum*, *T. pseudonana*), algal viruses and vascular plants. Predicted proteins from genome projects were retrieved from GenBank or the Joint Genome Institute genome portal (Grigoriev et al., 2011). The results were filtered with cutoffs of >60% identity, bit score >50 and E-value $\leq e^{-20}$ with the reference protein sequence. Sequences that passed this filter were used as BLASTX queries against the GenBank nr database (Sayers, 2011), using CAMERA (Sun et al., 2011), and removed from the assigned set if they matched a different taxon ID (e.g., bacteria) with a better overall score. Relatively few reads were removed at this step, most of which appeared to be of prokaryotic origin. To count reads per reference protein, assigned reads were divided into species-specific sets and the proteins of the corresponding species were used as TBLASTN queries against each of the parsed read sets. Hits were only counted if they matched the protein uniquely or if the

protein was present in duplicate. Results from these samples were summed and expressed as percentages of the total reads mapped to that taxon per analyzed sample (Supplementary Datasets 4.2, 4.3, Figures 4.4, 4.5, 4.6). Each semi-Lagrangian metatranscriptome was treated separately during taxon parsing. However, because fewer reads were generated per semi-Lagrangian sample, the data were later merged.

4.3.6 Determining genetic divergence among putative *Bathycoccus* and *Ostreococcus* ecotypes

To identify genes for exploring *Bathycoccus* and *Ostreococcus* genomic diversity and to use as ecomarkers in environmental samples, groups of homologous sequences were identified in the predicted proteomes from *Bathycoccus* and *Ostreococcus* genome projects (taxa listed above). *Bathycoccus* data were also retrieved from two metagenomic sequencing projects, a targeted tropical Atlantic Ocean metagenome (Monier et al., 2012) and two coastal Chilean targeted metagenomes (Vaulot et al., 2012), from which open reading frames (ORFs, i.e., sequence between stop codons with a minimal length of 60 amino acid residues) were identified. Diversity and ecomarker analysis was not performed for *Micromonas* because of the six established clades only two have representative complete genome sequences. Clusters of homologs were computed using OrthoMCL (Li et al., 2003), based on all versus all BLASTP (Altschul et al., 1997) searches of *Bathycoccus* and *Ostreococcus* proteomes, using a e^{-100} E-value cutoff . Clusters containing one sequence from each input source were kept, to ensure analysis of exclusively single-copy orthologs (i.e., clusters composed of sequences for four and three *Bathycoccus* and *Ostreococcus* predicted proteomes, respectively). Clusters producing alignments

shorter than 100 amino acid residues were discarded. Final datasets were composed of 1,104 *Bathycoccus* clusters, and 3,212 *Ostreococcus* clusters. Protein sequences from each cluster were aligned with T-Coffee (Notredame et al., 2000) and back-translated to codon-alignments to ensure better nucleotide gene alignments. Columns with gaps were removed from the nucleotide alignments and for *Bathycoccus* and *Ostreococcus* sets, identities of each codon-alignment were used to calculate an average percent identity at the genus level.

4.3.7 Determining Ecomarkers

To identify ecotype/species ecomarkers, the prior all versus all BLASTP results were rerun in OrthoMCL, with *Bathycoccus* and *Ostreococcus* data run together. Only groups composed of single-copy genes and populated by all seven sequence sources (i.e., homolog groups containing seven ORFs, one from each distinct *Bathycoccus* and *Ostreococcus* sequence source) were retained. To avoid potential ORF length biases in sequence recruitment (mainly due to large indels or to Bban7 having longer ORFs likely related to the gene modeling algorithms employed), the sequences within a cluster were 'corrected' by discarding indels; sequences from the same cluster were aligned using T-Coffee and positions with more than three gaps were removed. Only groups of 'corrected' sequences were retained, for which the longest sequence was ≤30% longer than the smallest. Sequences with mean identities <50% or >95% were discarded to recruit an informative set of 112 homolog groups used as ecomarkers for classifying meta-omic reads among *Bathycoccus* and *Ostreococcus* types (Supplementary Dataset 4.2).

4.3.8 Identifying Ecotype Ecomarkers in Meta-omic Data

To identify *Bathycoccus*/*Ostreococcus* ecomarkers in North Pacific Ocean WFAD09 and CANON2010 metatranscriptomes, the 112 *Bathycoccus*/*Ostreococcus* ecomarkers were used as TBLASTN (Altschul et al., 1997) queries against environmental reads. TBLASTN hits (bit-score >50, E-value $<e^{-5}$) were used as BLASTX queries against nr and only reads with best BLASTX hits to a *Bathycoccus*/*Ostreococcus* ecomarker (and bit-score >50, E-value $<e^{-5}$), were retained for downstream classification. These candidate reads were then used as BLASTN queries against the four *Bathycoccus* and three *Ostreococcus* ORF datasets. ORF datasets from *Micromonas* CCMP1545 and RCC299 (Worden et al., 2009) were used to remove false positives (i.e., reads with best BLASTN hits to a *Micromonas* sequence were removed). Reads were classified as environmental *Bathycoccus* or *Ostreococcus* sequences if they produced a BLASTN hit with a bit-score ≥250, a nucleotide identity ≥95%. The classified reads were then binned to one of the *Bathycoccus* or *Ostreococcus* types based on their best BLASTN hit. Those with the same statistics for two or more distinct *Bathycoccus* or *Ostreococcus* sequences were discarded due to lack of resolution (7.6% of the WFAD09 candidate reads, 4.7% for CANON2010). We also performed this analysis on historical metagenomic data from the same stations along Line-67 (Monier et al., 2012). From these 4.9% were discarded due to lack of resolution.

4.3.9 Analysis of SOD and nitrogen transporter gene families

Reciprocal BLASTP (Altschul et al., 1997) queries against nr, the Joint Genome Institute (JGI) browser and the Bioinformatics Ghent browser were used to

retrieve putative super oxide dismutase and nitrogen transporter orthologs from publicly available Mamiellophyceae genomes. For the AMT phylogeny, alignments from McDonald *et al*. (2010) were amended with *Bathycoccus prasinos* Bban7 and *Ostreococcus* RCC809 sequences by realigning sequences using ClustalW (Thompson et al., 1994). These results were manually adjusted in BioEdit and ambiguously aligned or non-homologous positions were subsequently masked. Phylip was used to construct NJ distance trees (Felsenstien, 2005) and maximum likelihood (ML) methods were performed in PhyML (Guindon and Gascuel, 2003) with 100 bootstrap replicates.

4.4 Results & Discussion

4.4.1 Environmental conditions

Two fundamentally different research cruises were conducted in the North Pacific Ocean to investigate picophytoplankton populations (Figure 4.1A) with an emphasis on Mamiellophyceae taxa. Eulerian sampling focused on euphotic zone waters in three oceanographic zones, referred to here as coastal (a near-shore upwelling zone with high phytoplankton abundances and nutrient concentrations (Moisen et al., 1996)), mesotrophic transition (TZ, a non-homogenous, hydrodynamically unstable zone, where nutrient rich coastal waters are advected hundreds of km offshore ((Brink and Cowels, 1991; Moisen et al., 1996) and open ocean (OO, a nutrient poor zone beyond the California Current) (Figure 4.1B). For in-depth biological characterization, analyses were performed on samples collected from at least two stations per region. PO_4 was always detectable, while both NO_3 and NH_4

dropped below detection limits (0.02 and 0.01 μmol·L^{-1}, respectively) at the OO stations (Figure 4.2A, Supplementary Figure 4.2A). The coastal region had the highest nutrient concentrations and Chlorophyll *a* values observed during the Eulerian transect (Table 4.1, Supplementary Dataset 4.1). Surface-water nutrient concentrations decreased with distance from shore and the chlorophyll maximum deepened (Figure 4.1B). Thus, at the most offshore stations the deep chlorophyll maximum (DCM) ranged from 85 to 105 m. NO_3 and NH_4 concentrations were the lowest observed, below detection in surface waters and ranging from 8.06×10^{-4} - 1.97×10^{-2} μmol·L^{-1} and 5.54×10^{-4} μmol·L^{-1}, respectively at the DCM (Supplementary Figure 4.2A). Within the eight days between outbound and inbound TZ sampling, the water column changed from having a more stratified structure, with a slight chlorophyll maximum around 30 to 40 m (Casts C8 and C10), to having more homogenous chlorophyll concentrations from 0 to 35 m (inward bound Cast, C42) (Supplementary Figure 4.2A). Nutrient concentrations during the 2010 semi-Lagrangian study were most similar to those of the coastal region Cast C2 (2009), although clear differences and some similarities to the TZ were observed, in terms of water column structure and nutrient depth gradients (Figure 4.2A, Supplementary Figure 4.3A, Supplementary Dataset 4.1).

4.4.2 Phytoplankton abundance

Picophytoplankton groups varied dramatically between the zones investigated. At the coast, small eukaryotes and *Synechococcus* were present at the same order of magnitude and declined in abundance below 10 m, while *Prochlorococcus* was

undetectable (Figure 4.2B, Supplementary Figure 4.2B, Supplementary Dataset 4.1). Eukaryotes and *Synechococcus* were present at the same order of magnitude in the semi-Lagrangian study, but extended deeper in the water column and were more abundant, and *Prochlorococcus* was detected (Figure 4.2B, Supplementary Figure 4.3B, Supplementary Dataset 4.1). In the TZ, all three groups were present, but they were more abundant during inward than outward-bound sampling (Supplementary Figure 4.2B, Supplementary Dataset 4.1). The offset was less between *Prochlorococcus* and *Synechococcus* abundances during inward-bound Station 67-70 TZ sampling (with increased *Prochlorococcus* cell numbers), and small eukaryotes ranged from ~1.8×10^3 to 2.8×10^4. Inward-bound NH_4 concentrations were also higher in the TZ, while NO_3 was lower than in outward-bound samplings (Supplementary Figure 4.2A). In the OO, eukaryote and *Prochlorococcus* abundances increased with depth within the euphotic zone, while *Synechococcus* abundances declined (Figure 4.2B, Supplementary Figure 4.2B, Supplementary Dataset 4.1). The DCM was just above the nutricline (Figure 4.2A) and the eukaryotic maximum (3.67×10^3 cells ml^{-1}, 105 m, Cast C21) occurred here, while the *Prochlorococcus* peak (1.95×10^5 cells ml^{-1}) was ~30 m above (Supplementary Figure 4.2B).

Since Mamiellophyceae have been reported previously in the California Current System, we used qPCR to discriminate some of the taxa within this class (Demir-Hilton et al., 2011). General primer-probe sets were used to quantify *Bathycoccus* and *Micromonas* (i.e., primer-probes that should amplify 18S rRNA genes from all known clades within these genera), while clade specific primers were

used for *Ostreococcus* (Demir-Hilton et al., 2011). Therefore, Mamiellophyceae 18S copies ml^{-1} values can be considered equivalent to twice the number of cells ml^{-1} (depending on the cell cycle stage) (Derelle et al., 2006; Palenik et al., 2007; Demir-Hilton et al., 2011; Moreau et al., 2012). *Ostreococcus* Clade OII was only detected in the OO (at very low abundance, maximum: ≤ 72 18S copies ml^{-1}), while *Ostreococcus* Clade OI was detected in relatively cool (13 ± 2°C), more nutrient rich coastal and TZ waters, as well as semi-Lagrangian samples, but not in the OO. This clade reached nearly 40,000 18S copies ml^{-1} at the surface in the semi-Lagrangian study (Figure 4.2C, Supplementary Figure 4.3C, Supplementary Dataset 4.1), but in the TZ Clade OI was higher at depth (and overall) during outward versus inward sampling. Previous laboratory studies indicate isolates composing clades OI and OII correspond to 'high-light' and 'low-light' adapted *Ostreococcus* ecotypes, respectively (Rodríguez et al., 2005; Cardol et al., 2008; Six et al., 2008). Our data support previous field-based results (Demir-Hilton et al., 2011), which indicated more complex ecophysiological factors than irradiance, dictate *Ostreococcus* ecotype distributions.

Like *Ostreococcus* Clade OI, *Micromonas* was undetectable in the OO region. Maximum TZ *Micromonas* 18S copies ml^{-1} (949 ± 170) were considerably lower than in coastal samples (1.46×10^4 ± 196) and the semi-Lagrangian study (8.08×10^3 ± 641). In contrast to *Ostreococcus* and *Micromonas*, *Bathycoccus* was relatively abundant at all stations investigated in both studies. In the Eulerian survey, *Bathycoccus* 18S copies ml^{-1} (2.14×10^4 ± 327) were the highest observed for the

Mamiellophyceae and dominated most profiles. During the semi-Lagrangian study they were also relatively high (e.g., 1.08×10^4 18S copies $ml^{-1} \pm 436$), although *Ostreococcus* Clade OI was dominant (Supplementary Figure 4.3C). Peak *Bathycoccus* abundances occurred in surface coastal and semi-Lagrangian waters, but at other sites the maximum typically occurred deeper, e.g., 95m at the OODCM (e.g., $7.25 \times 10^3 \pm 289$ 18S copies ml^{-1}, Station 67-155) and 40 m during outward-bound TZ sampling (Supplementary Dataset 4.1, Supplementary Figure 4.2C).

The most apparent patterns were the absences of *Ostreococcus* Clade OI and *Micromonas* in OO samples and *Ostreococcus* Clade OII in samples from inshore of the OO. In the coastal region, where NH_4, NO_3 and PO_4 concentrations are high in surface waters (0.17 $\mu mol \cdot L^{-1}$, 6.80-7.21 $\mu mol \cdot L^{-1}$ and 0.80 $\mu mol \cdot L^{-1}$, respectively) *Bathycoccus*, *Ostreococcus* Clade OI and *Micromonas* had maximum 18S copies ml^{-1} at the surface. When the TZ was stratified (Cast C8 outward-bound sampling of Station 67-70), the *Ostreococcus* Clade OI peak occurred at 30 m, as NH_4, NO_3 and PO_4 increased with depth, while *Bathycoccus* and *Micromonas* peaked at 40 m, coincident with greater increases in NO_3 and NH_4 (Figure 4.2, Supplementary Figure 4.2, Supplementary Dataset 4.1). At TZ Station 67-75 (C10) *Bathycoccus* 18S copies ml^{-1} were higher than those from either TZ Station 67-70 cast (C8 or C42), increasing at depths where relatively high (0.48 to 1.17 $\mu mol \cdot L^{-1}$) NH_4 was detected, with peak abundance co-occurring with those of *Ostreococcus* Clade OI and *Micromonas* at 40 m depth. In contrast, when the TZ mixed layer extended down to 35 m, with higher surface nitrogen concentrations than outward-bound sampling, the maximum 18S

copies ml^{-1} for *Bathycoccus*, *Ostreococcus* Clade OI and (very low) *Micromonas* were again near the surface, as seen in coastal and semi-Lagrangian samples.

Our results increase *Bathycoccus* abundance data considerably. Of the few other quantitative studies, Zhu *et al*. (2005) showed *Bathycoccus* peaked during winter in the coastal Mediterranean Sea (based on 18S rRNA gene qPCR), contributing 12 to 20% to the total picoeukaryotic population (copies per ml were not provided), but was undetectable in fall. Only minimal *Ostreococcus* 18S copies ml^{-1} were recovered at either time. *Bathycoccus*, *Ostreococcus* (only *Ostreococcus* Clade OII was detected) and *Micromonas* have been enumerated at the Bermuda Atlantic Time Series station (BATS) where they reached 96, 2,273, and 1,344 18S copies ml^{-1}, respectively, at 40 m during spring deep mixing (Treusch et al., 2012). *Micromonas* and *Ostreococcus* were not detected in the well-developed summertime DCM (80-120 m), but *Bathycoccus* was still present at >100 18S copies ml^{-1}. A major difference between the North Pacific OO region and BATS is PO_4 availability, which is below detection (<10 nM) at BATS in summer, but in the 0.25 - 0.84 $\mu mol \cdot L^{-1}$ range in the OO.

Only *Bathycoccus* was present at all stations investigated. Therefore, we wondered if this might reflect mixed signals coming from the proposed *Bathycoccus* ecotypes (both of which would be amplified by our qPCR primer-probe set). From TZ Cast C42 samples, two abundance peaks were observed, one at the surface and a second smaller peak at 40 m (Supplementary Dataset 4.1). This would be expected if the putative *Bathycoccus* ecotypes were adapted for different depth-associated

parameters, such as light or nutrients (or both). Thus, we sought to test the hypothesis that *Bathycoccus* omni-presence in our study reflected success of a single cosmopolitan species, *B. prasinos*, or differential contributions from different putative ecotypes. Unfortunately, no study was available that defined ecotype level divergence for Mamiellophyceae genera, or with a robust analysis of genetic distances between the putative *Bathycoccus* ecotypes, which have 100% 18S rRNA gene identity.

4.4.3 Establishment of *Bathycoccus* ecotype genetic distances

We established genetic distances between putative *Bathycoccus* ecotypes to characterize them in our samples. *Ostreococcus* clades have genome sequenced representatives with relatively well defined natural distributions. Therefore, this analysis was first performed for *Ostreococcus*, to serve as a benchmark for interpreting the *Bathycoccus* data. Clade OI isolates, *O. lucimarinus* and *O. tauri*, share approximately 90% of their protein encoding genes (Palenik et al., 2007). Using genes other than 18S rRNA, they consistently separate into different clades, although both are captured by the Clade OI qPCR primer probe set (Demir-Hilton et al., 2011). A comprehensive comparative analysis has not been published with Clade OII (*O.* RCC809), but as shown here, and previously, Clade OI and Clade OII have distinct biogeographies suggesting strong ecotypic differentiation. Thus far, co-occurrence has only been observed where Atlantic continental shelf and Gulf Stream waters meet (Demir-Hilton et al., 2011). We identified 3,212 proteins shared by *O. lucimarinus, O. tauri* and *O.* RCC809 and found coding sequences ranged from 72 to 75% identity between each of these species (Figure 4.3A).

For the putative *Bathycoccus* ecotypes, we identified 1,104 proteins shared between the *B. prasinos* Bban7 genome (Moreau et al., 2012) and *Bathycoccus* targeted metagenomes from the tropical Atlantic (Monier et al., 2012) and coastal Chile (denoted T142 and T149) (Vaulot et al., 2012). Nucleotide identities of the genes encoding these proteins from T142 and T149 (sorted from a well-mixed coastal water column, 2 days and 15 km apart, at 5 and 30 m, respectively) were not statistically different from *B. prasinos*, which was isolated in the coastal Mediterranean (Figure 4.3B). The tropical Atlantic *Bathycoccus* metagenome, from warm oligotrophic waters, had lower (82 ± 6%) identity to *B. prasinos* and the coastal Chilean metagenomes (Figure 4.3B). We concluded the high nucleotide identity between the coastal Chilean marker genes and those of *B. prasinos* were likely more reflective of a single ecotype, than multiple species or ecotypes. Hereafter these three are termed *B. prasinos* or the B1 ecotype. Although the tropical Atlantic *Bathycoccus* metagenome also had higher identity to *B. prasinos* than the three *Ostreococcus* species had to each other, the divergence of shared *Bathycoccus* genes was notable. We therefore termed the tropical Atlantic *Bathycoccus* ecotype B2.

4.4.4 Diversity assessment based on metatranscriptomic data

With our new understanding of *Bathycoccus* and *Ostreococcus* genetic distances, we computed the relative proportions of metatranscriptomic reads derived from *Ostreococcus* and *Bathycoccus* ecotypes. Because metagenomes for the picoplankton size fraction can be dominated by prokaryotic sequences, we sequenced polyA-selected RNA to avoid low eukaryotic read count issues (Worden and Allen,

2010; Worden et al., 2011) resulting in an of average 1.4×10^6 454-reads produced from each of the four Eulerian metatranscriptomes sequenced (Table 4.2, Supplementary Table 4.1). We also reanalyzed 12 of the semi-Lagrangian metatranscriptomes.

Using 112 informative homolog groups (referred to hereafter as 'ecomarkers') from *Ostreococcus* and *Bathycoccus*, we found metatranscriptome read assignments corresponded well with qPCR results. *Ostreococcus* results were used as a benchmark for our methods and to discriminate between *O. lucimarinus* and *O. tauri,* since the *Ostreococcus* Clade OI qPCR primer-probe set does not. The composition of Clade OI qPCR counts has been inferred based on environmental clone library results showing *O. lucimarinus* in coastal environments and *O. tauri* more restricted to bays and lagoons (Demir-Hilton et al., 2011). As seen in the qPCR data, reads hitting *Ostreococcus* Clade OI ecomarkers dominated the semi-Lagrangian samples, as well as Eulerian samples from coastal and TZ regions (Figure 4.3C). Likewise, Clade OII was observed exclusively at depth in the OO, where Clade OI was undetected by either method. Interestingly, no metatranscriptomic reads were assigned to *O. tauri* ecomarkers, providing deeper sequencing support for sub-Clade OI biogeography based on qPCR and environmental clone libraries (Demir-Hilton et al., 2011).

Unlike results for *Ostreococcus*, ecomarkers from *B. prasinos* B1 and *Bathycoccus* B2 co-occurred at the coastal, TZ and OODCM sites (Figure 4.3D). The oceanic B2 ecotype dominated the signal at the OODCM, but *B. prasinos* reads were considerable . In the semi-Lagrangian study, only *B. prasinos* ecomarkers were

detected. These results were replicated by applying the ecomarkers to 2007 Line-67 metagenomic data generated from the same time of year (Monier et al., 2012). The *Ostreococcus* Clades followed the same regional distributions observed in the metatranscriptomes (and qPCR) with Clades OI and OII appearing mutually exclusive, while *O. tauri* was never detected. For *Bathycoccus*, the metagenomic and metatranscriptomic results were also in agreement, with the exception that *Bathycoccus* was not detected in the 2007 OO surface sample. Proportions of *Bathycoccus* ecotypes were stable between the two Line-67 cruises (and the two data types), resulting in 76.2 ± 4.6% B2 reads and 23.8 ± 4.5% *B. prasinos* B1 reads in the OO DCM, versus 99.5 ± 0.7% and 99.7 ± 0.5% *B. prasinos* B1 reads in the TZ and coastal regions, respectively. Thus, the hypothesis that high *Bathycoccus* abundances observed in our study are accounted for by a single cosmopolitan type appears false. Rather, the results demonstrate that the two *Bathycoccus* ecotypes have differential distributions, but with more ecological overlap than *Ostreococcus* ecotypes.

While the majority of eukaryotic reads along the Eulerian transect mapped to the *Bathycoccus* genome and *Ostreococcus* Clade OI recruited the most semi-Lagrangian reads, several other taxa were also represented in the metatranscriptomes (Table 4.2). *Micromonas* reads were cumulatively abundant in the semi-Lagrangian study and in the Eulerian coastal region. In the OOsurf metatranscriptome, where few Mamiellophyceae reads were detected, haptophyte and pelagophyte sequences dominated (represented by *Emiliania huxleyi* and *Aureococcus anophagefferens*, respectively). Reads assigned to the latter were present at several sites and are likely

from the pelagophyte *Pelagomonas,* which can be abundant along Line-67 during the same time of year (Worden et al., 2012). Reads assigned to the diatom *Thalassiosira pseudonana* were notable at the coast.

4.4.5 Exploratory gene expression analyses

Exploratory gene expression analyses were also performed for *Bathycoccus* and *Ostreococcus* Clade OI (represented by *O. lucimarinus*). These taxa were selected because ≥2 reads (on average) were mapped per predicted protein in the genome, in two or more samples. TZ and OODCM metatranscriptomes were sampled at roughly the same time of day (1300 and 1400, respectively), while the coastal metatranscriptome was from 2200. Therefore, in addition to a summation of the 12 semi-Lagrangian metatranscriptomes, for balanced representation of day and night, we also summed the two 1400 and two 2200 metatranscriptomes from the semi-Lagrangian study for comparison to our Eulerian data. The OOsurf sample was excluded due to low read representation for these taxa (Table 4.2). In fitting with known aspects of phytoplankton cell cycles and other published metatranscriptomes (Poretsky et al., 2009; Marchetti et al., 2012) numerous photosynthesis-related genes, such as chlorophyll A/B binding proteins, were in the top 100 *Bathycoccus* genes expressed in mid-day metatranscriptomes, while ribosomal proteins were more prominent in evening samples (Supplementary Dataset 4.3). 26% of the top 100 expressed OODCM *Bathycoccus* genes were of unknown function, while that number was lower in coastal (19%), TZ (19%) and summed-semi-Lagrangian metatranscriptomes (11%). Although *Ostreococcus* Clade OI was too poorly

represented for analysis in the TZ and OO, proteins recruiting the highest percentage of reads from these metatranscriptomes were of similar function (Supplementary Dataset 4.4), as those from *Bathycoccus* in coastal and semi-Lagrangian data.

We investigated how conditions at the OODCM, where the B2 ecotype appeared dominant, might influence gene expression relative to more nutrient rich settings dominated by *B. prasinos* B1. Read representations, of the 2,552 protein-encoding genes that recruited ≥0.01% OODCM *Bathycoccus* reads (≥2 reads per gene), were compared across sites. Of the 1,977 of these with putative functions, 283, 616 and 79 were not detected in the coastal, TZ and summed semi-Lagrangian metatranscriptomes, respectively. Of the 575 unknown function genes, 105, 187 and 49 lacked reads in the same data sets. Because a greater number of *Bathycoccus* reads were sequenced in the coastal and summed semi-Lagrangian samples, detection of *Bathycoccus* gene expression should be greater at these sites than at the OODCM (Table 4.2). Therefore, the higher representation of a particular gene in the OODCM metatranscriptome should reflect responses to OODCM-specific factors and/or ecotype specific physiology. Fourteen known- and 14 unknown-function *Bathycoccus* genes had ≥0.01% read levels in the OODCM, but were undetected elsewhere (Supplementary Table 4.2). These 28 genes could serve as targets for future investigations of acclimation to open-ocean or low-light conditions.

4.4.5.1 Highly expressed genes at the Open Ocean Deep Chlorophyll Maximum

One hundred and forty seven *Bathycoccus* genes received ≥0.1% of OODCM read hits, of which 112 had Interpro-assigned functions (Figure 4.4). Chlorophyll A/B

binding proteins were prevalent and four out of the five had equivalent expression levels at other sites.

Interestingly, a Ni superoxide dismutase (SOD) was detected in the highly expressed (≥0.1%) gene set. There are four known SOD isozymes, which bind different metal centers, i.e., Mn, Fe, Ni and Cu/Zn, but catalyze the same dismutation reaction providing protection from reactive oxygen species (ROS) formed during photosynthesis (Wolfe-Simon et al., 2005). We annotated SODs in the *B. prasinos* and *O.* RCC809 genomes, and improved upon previous annotations in *O. lucimarinus*, *O. tauri* and *M. pusilla* CCMP1545 and *M.* RCC299 (Palenik et al., 2007; Worden et al., 2009). All four SOD metalloforms were identified in these Mamiellophyceae (except *O. tauri*), including previously unreported NiSODs. Each had three different CuZnSODs, forming homolog groups CuZnSOD1, CuZnSOD2 and CuZnSOD3, the latter of which contains homologs of the oldest plastid-targeted CuZnSOD in plants (Kanematsu et al., 2010).

The *Bathycoccus* Fe/MnSOD was represented at equivalent levels (≥0.01% category) at all Eulerian sites and semi-Lagrangian *Ostreococcus* levels were similar (Supplementary Figure 4.4). Interestingly, *Bathycoccus* reads assigned to CuZnSOD2 and CuZnSOD3 were not detected in the OODCM and TZ, although both were detected in coastal and semi-Lagrangian samples. Thus, *Bathycoccus* CuZnSOD2 and CuZnSOD3 expression may increase as a function of increased Fe availability, as seen with diatom CuZnSODs (Marchetti et al., 2012). Strikingly, the NiSOD was close to 10-fold higher at the OODCM than at all other sites, and had higher relative

read counts (again ~10-fold) than other SOD metalloforms (Supplementary Figure 4.4). Although these results were surprising, since ROS-related stresses seem unlikely at the depth of the OODCM (105 m), they may indicate that Ni acquisition can alleviate demand for Mn, Fe or Cu under limiting conditions.

4.4.5.2 Nitrogen Acquisition

The Line-67 transect showed a strong gradient in euphotic zone NO_3 and NH_4 concentrations (Table 4.1, Figure 4.2A, Supplementary Figures 4.1, 4.2). For this reason, and because some highly expressed genes in the *Bathycoccus* (e.g., AMT2.2) and *O. lucimarinus* (e.g., AMT1.1 and AMT2.2) transcriptomes were NH_4 transporters (AMTs), nitrogen transporter expression was investigated. We identified and annotated nitrogen transporters in *B. prasinos* (and *O*. RCC809) and compared them to transporters in *M. pusilla*, *M*. RCC299, *O. lucimarinus* and *O. tauri*, studied previously. Affinities have not been tested for these transporters, but were predicted based on homology to characterized transporters from other taxa (McDonald et al., 2010). In this process, a putative low affinity NO_3 permease (NRT1) and a high affinity NO_3 transporter (NRT2) were identified in *Bathycoccus*.

Bathycoccus NRT1 and NRT2.1 expression was observed at all but the coastal site. Relative percentages were generally low, except at the OODCM where NRT2.1 had higher relative read counts. Unlike levels at all other sites, NO_3 was below the detection limit (<0.02 $\mu mol \cdot L^{-1}$) at the OODCM metatranscriptome collection depth, supporting the identification of NRT2.1 as a high affinity transporter. In contrast, *Bathycoccus* NRT2.1 reads were not detected at the coastal site and had very low

relative read counts in the summed semi-Lagrangian data. *Ostreococcus* NRT2.1 showed no expression. From this, we surmised NRT2.1 may be a high affinity NO_3 transporter and serve as a useful indicator of organism low nitrogen thresholds.

Four AMT1 superfamily and two AMT2 superfamily NH_4 transporters identified in *Bathycoccus* were analyzed using phylogenetic methods (Supplementary Figure 4.6). *Bathycoccus* representatives of two known Mamiellophyceae AMT1 types were found and sistered *Ostreococcus* versions. Two additional *Bathycoccus* AMT1 proteins formed a sister group to *M*. RCC299 AMT1.3, for which close homologs were previously unknown. We termed these AMT1.3a (CCO17111.1) and AMT1.3b (CCO19599.1). Expression of AMT1.2, AMT2.1 (plastid targeted) and AMT2.2 was low or undetectable in the OODCM, but AMT1.1, AMT1.3a and AMT1.3b each contributed >0.02% of total *Bathycoccus* reads (Supplementary Figure 4.5). The *Bathycoccus* 1.3 transporters showed relatively consistent read percentages across sites, indicative of constitutive components of the *Bathycoccus* nitrogen acquisition system. AMT2.2 relative read counts were in the ≥0.01% category for *O. lucimarinus* in the semi-Lagrangian transcriptomes and for *Bathycoccus* in the coastal metatranscriptome. These results suggest AMT2.2 is a low-affinity NH_4 transporter, given that it recruited moderate to high read counts in the semi-Lagrangian and coastal metatranscriptomes, but was not detected in the TZ or OO. Depending on the degree to which transcription represents translation and use, AMT expression at sites where NH_4 was below detection suggests *Bathycoccus* may still acquire this nutrient when standing stocks are low.

4.5 Conclusion

Our data reveal complex patterns in picoeukaryote community composition. While picoeukaryotes are known to be abundant in coastal zones, we show they can also be highly abundant in a North Pacific mesotrophic transition zone extending to eastern boundary currents. Overall, *Ostreococcus* Clade OI showed very high 18S rRNA gene counts in upwelling influenced and coastal waters, while ecomarker analysis showed *O. lucimarinus*-like strains dominated the signal with no contribution from *O. tauri*-like strains. Similar to previous studies, the *Ostreococcus* Clade OI and Clade OII ecotypes did not co-occur and showed distributions that corresponded to differences in environmental conditions (e.g., temperature) (Demir-Hilton et al., 2011). In contrast, *Bathycoccus* ecotypes appear to coexist in areas apart from frontal and convergence zones. Additionally, the diversity and ecomarker analyses indicated two *Bathycoccus* ecotypes exist and that other reported sequence variations proposed to represent other ecotypes are SNPs, which were present at a frequency of less than 3 per 100 nt. Unlike *Ostreococcus*, *Bathycoccus* was abundant in samples where NH_4 and NO_3 were below detection, as well as those with relatively high NH_4 and NO_3 concentrations. Additionally, two *Bathycoccus* AMTs persistently expressed in the field were absent from *Ostreococcus*. Our metatranscriptome data support a high affinity transport role for NRT2.1 and a low affinity transport role for AMT2.2, in the Mamiellophyceae. These results provide targets for future field analyses and testable hypotheses for those taxa in culture.

Table 4.1 Environmental parameters for discrete depths at four representative transect stations and ranges (latitude and longitude) or averages (other parameters) for the 12 semi-Lagrangian stations. Data shown correspond to dates and depths sampled for Eulerian metatranscriptomes. Nutrient and chlorophyll *a* measurement came from the same or nearby depths performed on adjacent casts.

	Eulerian				Semi-Lagrangian
Parameter	OODCM	OOsurf	Transition	Coastal	Ottesen *et al*. 2013
Latitude	33.286	33.286	36.126	36.740	36.045 to 35.790
Longitude	-129.429	-129.429	-123.490	-121.020	-123.022 to -122.709
Depth (m)	106	10	10	10	25
Time of Day (PST)	14:08	09:11	13:02	19:37	diel
NO_3^- ($\mu mol \cdot L^{-1}$)	0.10	0.07	0.26	12.91	8.51±1.87
NH_4 ($\mu mol \cdot L^{-1}$)	0.00	0.00	0.05	0.26	
$Si(OH)_4$ ($\mu mol \cdot L^{-1}$)	2.60	2.03	2.92	9.91	7.76±2.00
PO_4 ($\mu mol \cdot L^{-1}$)	0.40	0.48	0.49	1.15	1.49±0.19
Total Chl *a* ($mg \cdot m^{-3}$)	0.22	0.10	0.50	1.98	0.85±0.08
Synechococcus ($cells \cdot mL^{-1}$)	159	2,797	36,519	21,512	111,300±18,357.38
Prochlorococcus ($cells \cdot mL^{-1}$)	82,502	156,236	71,744	0	35,969±1,271.47
Total Eukaryotes ($cells \cdot mL^{-1}$)	3,674	1,086	16,257	10,574	64,824±13,215.83

[a]Flow cytometry counts provided for the semi-Lagrangian study represent the average and S.D. for two casts corresponding to the two 6 am samplings for which qPCR was also performed.

Table 4.2 Numbers of metatranscriptomic reads assigned to genome sequenced eukaryotic species by location and depth. The number of proteins predicted in genome projects for each species is shown under 'Predicted proteome'.

	Taxa	Taxon Genome Sequenced	Eulerian				Semi-Lagrangian
			OODCM	OOsurf	TZ	Coastal	Ottesen *et al*. 2013[b]
Class II Prasinophytes (Mamiellophyceae)	*Bathycoccus prasinos*	7,921	16,958	602	12,079	30,613	75,887
	Ostreococcus lucimarinus	7,605	451	473	1,604	12,437	180,370
	Ostreococcus tauri	7,987	3402	3504	3908	3367	16,392
	Ostreococcus RCC809	7,492	1,095	802	935	1,634	32,956
	Micromonas RCC299	10,109	982	1,223	1,413	11,388	30,984
	Micromonas CCMP2099[a]	19,316	2,648	2,955	1,831	4,495	12,923
	Micromonas CCMP1545	9,702	591	735	702	3,365	7,059
Class I Prasinophytes	*Pyramimonas parkeae*[a]	20,299	3,493	3,871	3,391	2,666	15,939
Chlorophytes	*Chlamydomonas reinhardtii*	17,113	4,654	4,641	3,207	2,909	10,943
Unassigned Green Origin	*Arabidopsis thaliana*	35,386	3,503	4,862	4,126	3,951	25,117
	Aureococcus anophagefferens	11,501	10,283	5,751	12,684	5,913	48,783
Stramenopiles	*Ectocarpus siliculosus*	16,269	3,181	4,110	3,758	3,138	14,123
	Phaeodactylum tricornutum	10,417	1,271	1,851	1,521	6,417	8,220
	Thalassiosira pseudonana	11,673	1,431	2,819	2,984	9,392	16,886
Cryptophytes	*Guillardia theta*	24,840	2,433	2,935	2,260	4,711	11,185
Rhizarians	*Bigelowiella natans*	21,708	2,059	2,661	2,116	2,408	10,961
Haptophytes	*Emiliania huxleyi*	33,340	6,681	6,778	6,425	6,893	22,894

[a]The predicted proteomes for these taxa are not based on complete genome sequences, but on assembled transcriptomes available at CAMERA under project number CAM_P_0001089. Predicted protein counts are likely inflated due to assembly methods.

[b]Ottesen *et al*. 2013 data shown are the summation of values from the first 12 transcriptome samples collected during the study, the 13th sample was excluded because it led to overrepresentation of one diel timepoint.

Table 4.3 Predicted proteins from the *Bathycoccus prasinos* Bban7 genome (Moreau et al., 2012) and annotation breakdown for proteins recruiting metatranscriptomic reads and for which no metatranscriptomic reads were detected.

Number of *Bathycoccus prasinos* proteins	Known Function	Unknown Function	Total
Predicted proteins in genome	5,698	2,221	7,919
Not detected in any sample	303	337	640
Detected in only 1 sample	811	428	1,239
Detected in >1 Sample	4,584	1,456	6,040
Detected in all samples[a]	1,619	440	2,059

[a]Semi-Lagrangian diel data from Ottesen *et al*. 2013 were summed and considered a single sample.

Figure 4.1

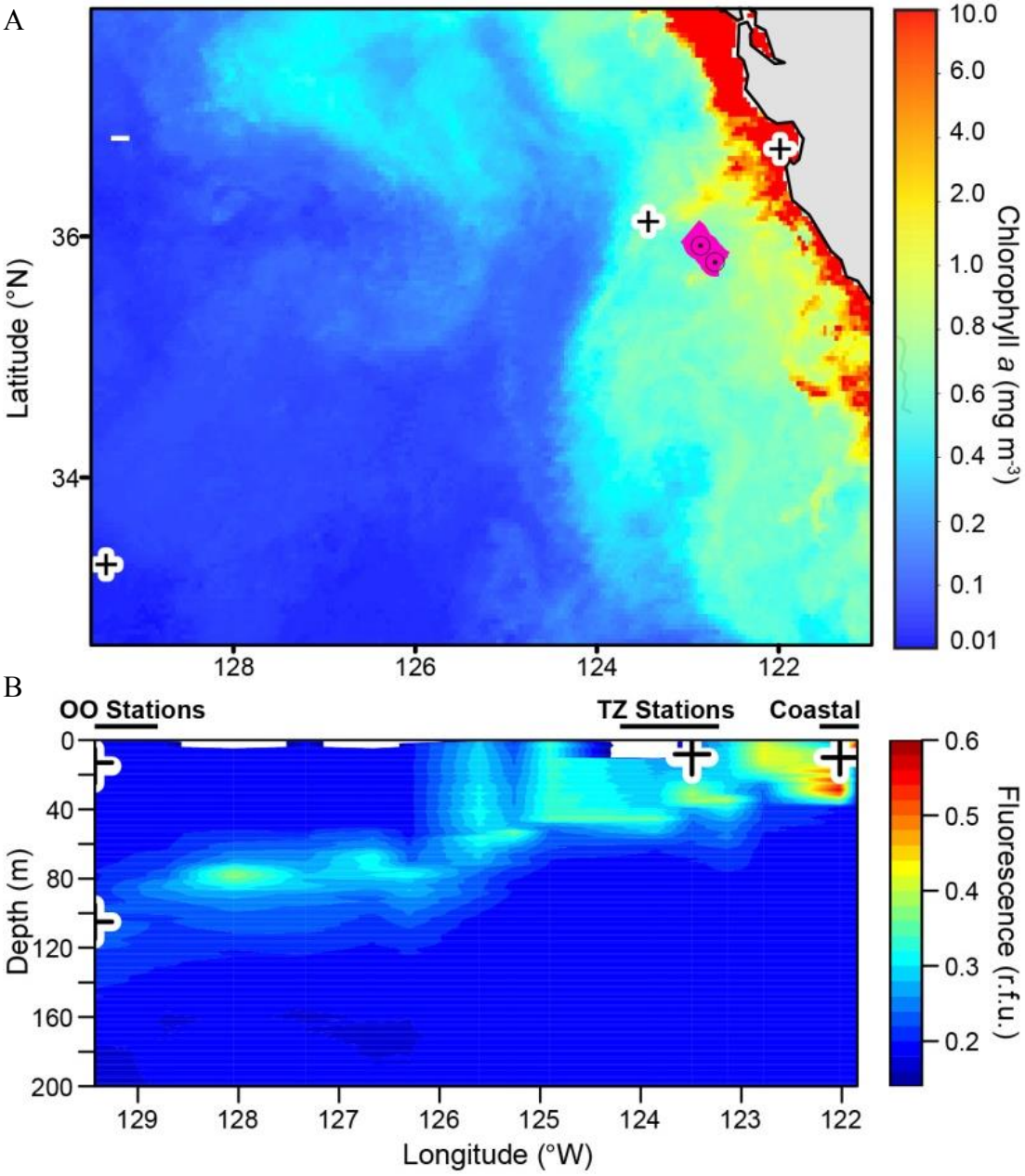
A
Latitude (°N)
36
34
128
126
124
122
10.0
6.0
4.0
2.0
1.0
0.8
0.6
0.4
0.2
0.1
0.01
Chlorophyll a (mg m-3)
B
OO Stations
TZ Stations
Coastal
Depth (m)
0
40
80
120
160
200
129
128
127
126
125
124
123
122
Longitude (°W)
0.6
0.5
0.4
0.3
0.2
Fluorescence (r.f.u.)

Figure 4.1 Geographic locations sampled during North Pacific Cruises. (A) Locations of Eulerian sites where profiles and metatranscriptomes were analyzed as well as the overall trajectory of the ESP drifter (pink) during the semi-Lagrangian survey. Representative profile Casts C44 and C51 in the latter are shown (black dots with circles). The crosses show metatranscriptome sampling sites from the Eulerian study. The map represents remotely-sensed high resolution chlorophyll *a* concentrations from MODIS/Aqua during the month of the semi-Lagrangian study, October 2010. (B) Calibrated *in vivo* fluorescence derived from chlorophyll *a* along the Eulerian transect. The major regions investigated are indicated by horizontal lines (black, above) as are metatranscriptome sampling locations (crosses). The coastal, TZ, and OO metatranscriptome sampling sites were 25 km, 172 km and 785 km from shore, respectively.

Figure 4.2

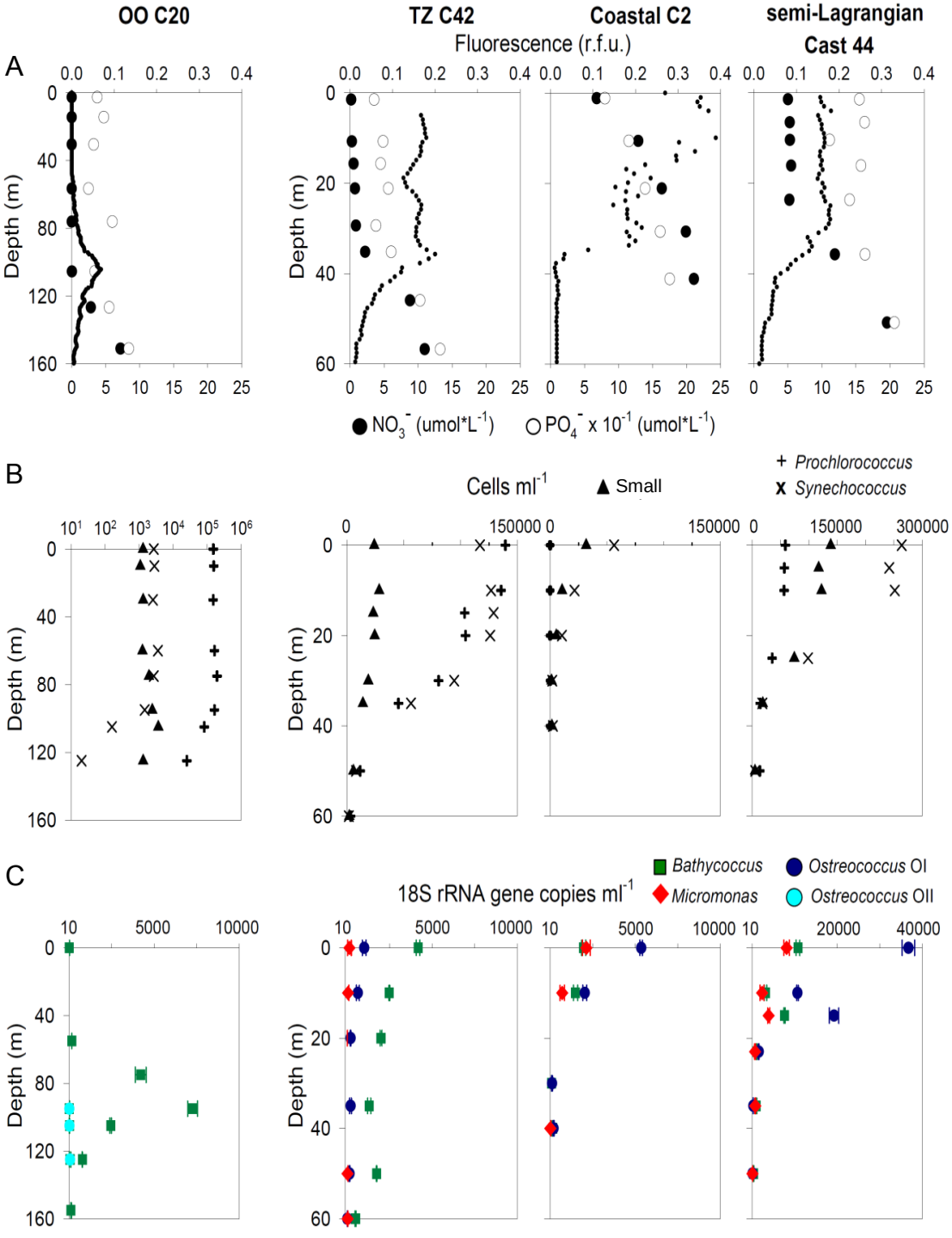
OO C20
TZ C42
Coastal C2
semi-Lagrangian
Cast 44
Fluorescence (r.f.u.)
A
Depth (m)
● NO_3^- (umol*L^{-1})
○ PO_4^- x 10^{-1} (umol*L^{-1})
B
Cells ml^{-1}
▲ Small
+ Prochlorococcus
x Synechococcus
C
18S rRNA gene copies ml^{-1}
Bathycoccus
Micromonas
Ostreococcus OI
Ostreococcus OII

Figure 4.2 Water column characteristics and phytoplankton abundances from representative profiles. Data is shown for three representative Eulerian stations corresponding to sites and dates when metatranscriptomic samples were collected. A representative semi-Lagrangian cast (C44) is also shown. (A) NO_3 and PO_4 concentrations as well as chlorophyll *a* derived *in vivo* fluorescence are represented. Note differences in Y-axis scales for the leftmost panel (OO) from other panels. (B) Flow cytometry counts for major picophytoplankton groups *Prochlorococcus*, *Synechococcus* and small photosynthetic eukaryotes from the same casts. Note differences in Y-axis scales for the leftmost panel (OO) from other panels. (C) 18S rRNA gene copies ml^{-1} for Mamiellophyceae taxa enumerated using primer-probe sets for *Ostreococcus* Clade OI, *Ostreococcus* Clade OII, *Micromonas* and *Bathycoccus* (Demir-Hilton et al., 2011). *Ostreococcus* Clade OII was rarely detected. Error bars represent the standard deviation of technical triplicates. Note differences in Y-axis scales for the leftmost panel (OO) from other panels.

Figure 4.3

A

Ostreococcus		*O. tauri*	*O. lucimarinus*
OI	*tauri*	—	73 ± 8
OI	*lucimar*	73 ± 8	—
OII	RCC809	72 ± 8	75 ± 8

B

Bathycoccus	Bban7	T142	T149
Bban7	—	97 ± 4	97 ± 3
T142	97 ± 4	—	97 ± 5
T149	97 ± 3	97 ± 5	—
trop Atlan	82 ± 6	81 ± 7	81 ± 7

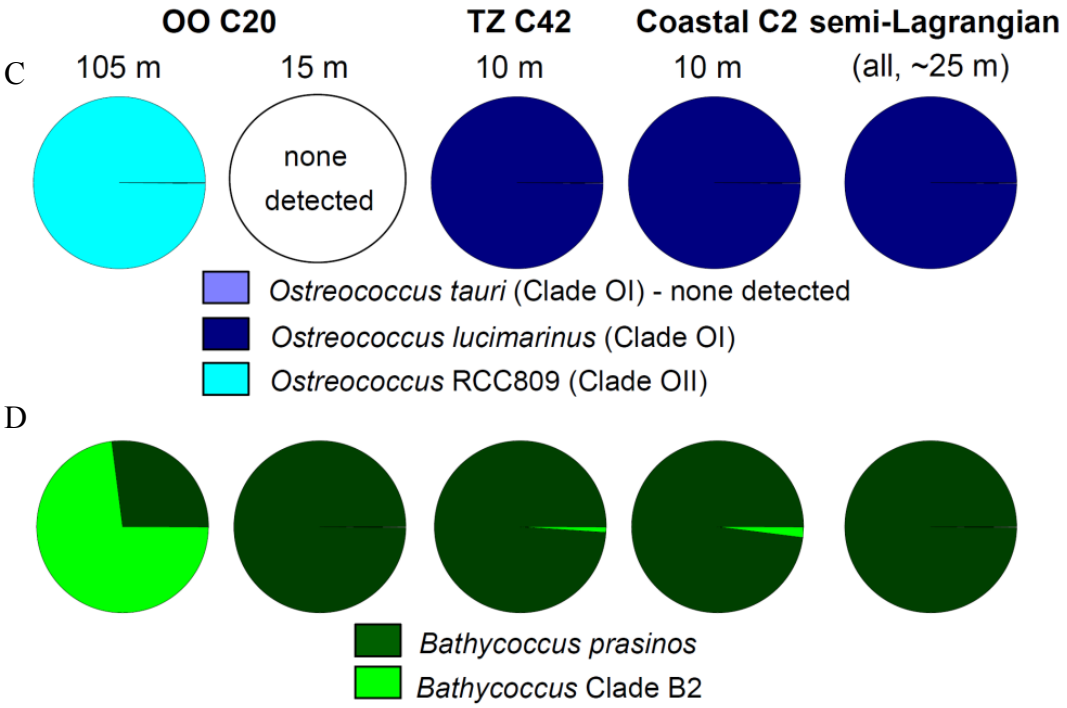

Figure 4.3 *Bathycoccus* and *Ostreococcus* identity matrices and relative abundances of metatranscriptomic ecomarkers. (A) *Ostreococcus* and (B) *Bathycoccus* nucleotide identity matrices based on shared genes identified here. (C) *Ostreococcus* and (D) *Bathycoccus* ecomarker analyses from Eulerian and semi-Lagrangian metatranscriptomic samples.

Figure 4.4


OO
TZ
Coast
S-L
09g00260 Ribulose bisphosphate carboxylase, small chain
08g03010 Band 7 protein
08g01840 Chlorophyll A-B binding protein, plant
11g03740 FAS1 domain
07g02200 Chlorophyll A-B binding protein, plant
07g01020 Chlorophyll A-B binding protein, plant
14g01960 Histone-fold
14g02760 FAD-dependent pyridine nucleotide-disulphide oxidoreductase
14g02170 Chlorophyll A-B binding protein, plant
09g03620 Serine/threonine-protein kinase, active site
09g00740 Chlorophyll A-B binding protein, plant
04g02510 Translation elongation factor EF1A/initiation factor IF2gamma, C-terminal
14g02020 3-deoxy-8-phosphooctulonate synthase
01g05850 N-terminal nucleophile aminohydrolases (Ntn hydrolases)
12g02220 Ribosomal protein S5 domain 2-type fold, subgroup
01g02970 Peptidase, cysteine peptidase active site
07g04750 Chlorophyll A-B binding protein, plant
09g00780 Nucleotide-diphospho-sugar transferase, predicted
11g00460 Chaperonin Cpn60, conserved site
02g03840 Glycosyl transferase, family 8
03g02390 Glucocorticoid receptor-like (DNA-binding domain)
06g01450 S-adenosyl-L-methionine-dependent methyltransferases
14g02590 Photosystem I PsaL, reaction centre subunit XI
06g04240 Xanthine/uracil/vitamin C permease
02g03670 Phosphatidylinositol 3-/4-kinase, catalytic
01g03370 P-loop containing nucleoside triphosphate hydrolases
15g01230 Primosome PriB/single-strand DNA-binding
04g04870 ABC transporter, transmembrane domain, type 1
10g04040 3-oxo-5-alpha-steroid 4-dehydrogenase, C-terminal
10g01660 ATPase, AAA-type, conserved site
08g03700 Uncharacterised protein family UPF0016
13g00240 Peptidase aspartic, active site
03g02160 Cupredoxin
02g00160 Nucleotidyltransferase
07g01960 Ribosomal protein L10e/L16
10g00610 PWWP
11g02970 Aldehyde oxidase/xanthine dehydrogenase, molybdopterin binding
12g00360 Aldolase-type TIM barrel
09g02350 Heat shock protein Hsp20
14g01590 Serine/threonine-protein kinase, active site
06g04900 FAD-linked oxidase, FAD-binding, subdomain 2
04g01510 Ubiquinol-cytochrome c reductase, iron-sulphur subunit
08g01290 Nucleotide-binding, alpha-beta plait
15g01850 Pyridoxal phosphate-dependent enzyme, beta subunit
09g00270 Ribulose bisphosphate carboxylase, small chain
07g02390 Rhodanese-like
09g02680 Glycosyl transferase, family 10
01g00650 Peptidase S8/S53, subtilisin/kexin/sedolisin
01g02940 Twin-arginine translocation pathway, signal sequence
08g01100 P-loop containing nucleoside triphosphate hydrolases
04g04180 P-loop containing nucleoside triphosphate hydrolases
02g02770 N-terminal nucleophile aminohydrolases (Ntn hydrolases)
06g04720 Snf7
07g00670 Concanavalin A-like lectin/glucanase, subgroup
09g01490 Ubiquitin-associated/translation elongation factor EF1B, N-terminal, eukaryote
14g02570 Heat shock protein 70kD (HSP70), peptide-binding domain
04g01150 LytB protein
04g03090 Protein phosphatase 2C, N-terminal
01g02170 EF-Hand 1, calcium-binding site
17g01910 Succinate dehydrogenase/fumarate reductase flavoprotein, catalytic domain
05g01950 P-loop containing nucleoside triphosphate hydrolases
01g02180 Nucleotide-diphospho-sugar transferase, predicted
08g02240 GLUCOSAMINE--FRUCTOSE-6-PHOSPHATE AMINOTRANSFERASE, ISOMERIZING
03g03720 Flagellar calcium-binding protein calflagin
07g01520 Ovarian tumour, otubain
09g01130 14-3-3 protein, conserved site
03g04620 Chlorophyll A-B binding protein, plant
05g03540 Target SNARE coiled-coil domain
15g01150 Alternative oxidase
08g03910 Peptidyl-prolyl cis-trans isomerase, cyclophilin-type
13g01530 DNA-directed DNA polymerase, family B, multifunctional domain
01g06180 Tubulin/FtsZ, 2-layer sandwich domain
15g01440 Peptidyl-prolyl cis-trans isomerase, cyclophilin-type, conserved site
08g02420 Thioredoxin-like fold
01g04000 EF-Hand 1, calcium-binding site
09g00470 Methyltransferase type 11
07g02150 RmlC-like jelly roll fold
10g03670 Fumarate lyase, conserved site
05g04200 PHOSPHORIBOSYLFORMYLGLYCINAMIDINE SYNTHASE
10g03880 Glutathione peroxidase
09g04680 RNA helicase, ATP-dependent, DEAD-box, conserved site
09g04200 Phosphoglycerate/bisphosphoglycerate mutase, active site
01g04200 Uncharacterised protein family UPF0118
15g01340 Phytanoyl-CoA dioxygenase
15g01410 Superoxide dismutase, Nickel-type
10g00130 Pleckstrin homology domain
07g03400 Pheophorbide a oxygenase
13g00910 Sporulation stage V, protein S
03g00920 Pyridine nucleotide-disulphide oxidoreductase, NAD-binding region
05g04640 Proteinase inhibitor I29, cathepsin propeptide
13g00420 WD40/YVTN repeat-like-containing domain
05g04530 Bromodomain
02g01530 Moybdenum cofactor oxidoreductase, dimerisation
10g04240 Thioredoxin fold
03g03000 Glucocorticoid receptor-like (DNA-binding domain)
04g00910 Ribosomal protein S30Ae/sigma 54 modulation protein
07g03030 Proteinase inhibitor I29, cathepsin propeptide
07g03630 Peptidase C19, ubiquitin carboxyl-terminal hydrolase 2, conserved site
01g02200 Lipopolysaccharide-modifying protein
08g03920 P-loop containing nucleoside triphosphate hydrolases
10g03710 PRICHEXTENSN
13g02530 VOLTAGE-GATED CATION CHANNEL (CALCIUM AND SODIUM)
02g04110 Cell division protein FtsZ, conserved site
05g04540 Sec1-like protein
04g00190 Histone-fold
12g02100 Translocon-associated protein (TRAP), alpha subunit
18g01120 SWI/SNF-RELATED MATRIX-ASSOCIATED ACTIN-DEPENDENT REGULATOR OF CHROMATIN
03g01360 Glycyl-tRNA synthetase, class II, heterodimeric
04g00750 Serine/threonine-protein kinase-like domain
06g02210 Glycosyl hydrolase, family 13, subfamily, catalytic domain
08g04420 URICASE AND TRANSTHYRETIN-RELATED
03g01880 Pyrrolo-quinoline quinone beta-propeller repeat

Figure 4.4 Known function *Bathycoccus* proteins highly expressed at the North Pacific OODCM. The heat map shows 112 *Bathycoccus* known function proteins that received ≥ 0.1% of metatranscriptomic reads recovered from the OODCM. Next to the heat map are protein identification numbers (ORCAE) and InterPro Scan assigned functions (run here). Metagenomic dataset abbreviations: OO – open ocean deep chlorophyll maximum, TZ – mesotrophic transition zone, Coast – coastal, S-L – semi-Lagrangian.

Chapter 4 Supplementary Figure, Table and Dataset Legends

Supplementary Figure 4.1 Mixed layer depth during semi-Lagrangian ESP metatranscriptome sampling events, as shown through (A) σT plotted versus depth and (B) temperature plotted versus depth.

Supplementary Figure 4.2 Water column characteristics and phytoplankton abundances for all Eulerian profiles, which correspond directly to sites and specific dates when metatranscriptomic samples were collected. (A) NO_3 and NH_4 concentrations (different than Figure 2A, which shows PO_4 not NH_4) as well as *in vivo* fluorescence are shown. Note differences in Y-axis scales for the two leftmost panels (OO) from other panels. (B) Flow cytometry counts for the major picophytoplankton groups *Prochlorococcus*, *Synechococcus* and small photosynthetic eukaryotes. Note differences in Y-axis scales for the two leftmost panels (OO) from other panels. (C) 18S rRNA gene copies ml^{-1} for Mamiellophyceae taxa enumerated using primer-probe sets for *Ostreococcus* Clade OI, Clade OII, *Micromonas* and *Bathycoccus* (Demir-Hilton et al., 2011). *Ostreococcus* Clade OII was rarely detected. Error bars represent the standard deviation of technical triplicates. Note differences in Y-axis scales for the two leftmost panels (OO) from other panels.

Supplementary Figure 4.3 Water column characteristics and phytoplankton abundances for two representative profiles from the semi-Lagrangian study. Panels are as stated for Supplementary Figure 4.2.

Supplementary Figure 4.4 *Bathycoccus* and *Ostreococcus lucimarinus* SODs as percent of taxon specific total reads at each site. Columns represent samples. Rows are super oxide dismutase isozymes. The three CuSOD homolog groups had amino acid identities of 48% (CuSOD1), 41% (CuSOD2) and 66% (CuSOD3) between *O. lucimarinus* and *B. prasinos*. Fe/MnSOD homologs of *O. lucimarinus* and *B. prasinos* had 60% amino acid identity and NiSOD homologs had 41% amino acid identity. Bathycoccus SOD transcripts were not found in the semi-Lagrangian samples from 1400 (although all were detected, albeit at low levels in the summed set), which corresponds to the TZ and OODCM sampling time points. This suggests they are influenced by cell cycle stage or day:night transitions.

Supplementary Figure 4.5 *Bathycoccus* and *Ostreococcus lucimarinus* nitrogen transporters as percent of taxon specific total reads at each site. Columns represent samples. Rows are nitrogen transporters (NO_3 transporter, NRT; NH_4 transporter, AMT). From a cell cycle perspective, consistent expression of the plastid targeted AMT2.1 might be expected, since transport to the chloroplast is a basic requirement for cell growth (amino acid synthesis, etc.). However, the persistent expression of many of the AMTs suggests NH_4 is an important resource for *Bathycoccus* and *Ostreococcus*.

Supplementary Figure 4.6 Unrooted maximum likelihood phylogenetic tree using protein sequences from AMT1 and AMT2 gene families from a selection of

eukaryotes, bacteria and archaea (some with sequenced genomes). Black circles indicate bootstrap support between 95 and 100 by both ML and NJ methods. White circles represent the same for ML only.

Supplementary Dataset 4.1 Nutrients, cell counts, qPCR data and other parameters for nine profiles.

Supplementary Dataset 4.2 Ecomarkers and relevant data.

Supplementary Dataset 4.3 Top 100 *Bathycoccus prasinos* Bban7 proteins hit by the most metatranscriptomic reads per Eulerian station and summed semi-Lagrangian data.

Supplementary Dataset 4.4 Top 100 *Ostreococcus lucimarinus* proteins hit by the most metatranscriptomic reads per Eulerian station and summed semi-Lagrangian data.

References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389-3402.

Brink, K.H., Beardsley, R. C., Niiler, P. P., Abbott, M., Huyer, A., Ramp, S., Stanton, T., Stuart, D. (1991) Statistical Properties of Near-Surface Flow in the California Coastal Transition Zone. *Journal of Geophysical Research* **96**: 14,693-14,706.

Cardol, P., Bailleul, B., Rappaport, F., Derelle, E., Beal, D., Breyton, C. et al. (2008) An original adaptation of photosynthesis in the marine green alga *Ostreococcus*. *Proc Natl Acad Sci U S A* **105**: 7881-7886.

Chisholm, S.W., Armbrust, E.V., and Olson, R.J. (1986) The individual cell in phytoplankton ecology: cell cycles and flow cytometry. *Can Bull Fish Aquat Sci* **214**: 343-369.

Cuvelier, M.L., Allen, A.E., Monier, A., McCrow, J.P., Messié, M., Tringe, S.G. et al. (2010) Targeted metagenomics and ecology of globally important uncultured eukaryotic phytoplankton. *Proceedings of the National Academy of Sciences of the United States of America* **107**: 14679-14684.

Demir-Hilton, E., Sudek, S., Cuvelier, M.L., Gentemann, C.L., Zehr, J.P., and Worden, A.Z. (2011) Global distribution patterns of distinct clades of the photosynthetic picoeukaryote Ostreococcus. *The ISME journal* **5**: 1095-1107.

Derelle, E., Ferraz, C., Rombauts, S., Rouze, P., Worden, A.Z., Robbens, S. et al. (2006) From the Cover: Genome analysis of the smallest free-living eukaryote *Ostreococcus tauri* unveils many unique features. *Proc Natl Acad Sci U S A* **103**: 11647-11652.

Felsenstien, J. (2005) PHYLIP (Phylogeny Inference Package) version 3.6. In: Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

Field, C.B., Behrenfeld, M.J., Randerson, J.T., and Falkowski, P. (1998) Primary production of the biosphere: Integrating terrestrial and oceanic components. *Science* **281**: 237-240.

Gomez-Alvarez, V., Teal, T.K., and Schmidt, T.M. (2009) Systematic artifacts in metagenomes from complex microbial communities. *Isme Journal* **3**: 1314-1317.

Grigoriev, I.V., Nordberg, H., Shabalov, I., Aerts, A., Cantor, M., Goodstein, D. et al. (2011) The Genome Portal of the Department of Energy Joint Genome Institute. *Nucleic Acids Research* **40**: D26-D32.

Grob, C., Ulloa, O., Claustre, H., Huot, Y., Alarcon, G., and Marie, D. (2007) Contribution of picoplankton to the total particulate organic carbon concentration in the eastern South Pacific. *Biogeosciences* **4**: 837-852.

Guillou, L., Eikrem, W., Chretiennot-Dinet, M., Le Gall, F., Massana, R., Romari, K. et al. (2004) Diversity of picoplanktonic prasinophytes assessed by direct nuclear SSU rDNA sequencing of environmental samples and novel isolates retrieved from oceanic and coastal marine ecosystems. *Protist* **155**: 193-214.

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

Jardillier, L., Zubkov, M.V., Pearman, J., and Scanlan, D.J. (2010) Significant CO_2 fixation by small prymnesiophytes in the subtropical and tropical northeast Atlantic Ocean. *The ISME Journal* **4**: 1180–1192.

Kanematsu, S., Iriguchi, N., and Ienaga, A. (2010) Characterization of CuZn-superoxide dismutase gene from the green alga Spirogyra sp . (Streptophyta): Evolutionary implications for the origin of the chloroplastic and cytosolic isoforms. Bull. Minamikyushu Univ. 40A:65-77.

Li, L., Stoeckert Jr., C.J., and Roos, D.S. (2003) OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. *Brief Bioinformatics* **13**: 2178-2189.

Li, W.K.W. (1994) Primary production of prochlorophytes, cyanobacteria, and eukaryotic ultraplankton: Measurements from flow cytometric sorting. *Limnology and Oceanography* **39**: 169-175.

Li, W., and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**: 1658-1659.

Lopez-Garcia, P., Moreira, D., and Rodriguez-Valera, F. (2001) Diversity of free-living prokaryotes from a deep-sea site at the Antarctic Polar Front. *FEMS Microbiology Ecology* **36**: 193-202.

Marchetti, A., Schruth, D.M., Durkin, C.a., Parker, M.S., Kodner, R.B., Berthiaume, C.T. et al. (2012) Comparative metatranscriptomics identifies molecular bases for the physiological responses of phytoplankton to varying iron availability. *Proceedings of the National Academy of Sciences of the United States of America* **109**: E317-325.

Marie, D., Zhu, F., Balague, V., Ras, J., and Vaulot, D. (2006) Eukaryotic picoplankton communities of the Mediterranean Sea in summer assessed by molecular approaches (DGGE, TTGE, QPCR). *FEMS Microbiol Ecol* **55**: 403-415.

McDonald, S.M., Plant, J.N., and Worden, A.Z. (2010) The mixed lineage nature of nitrogen transport and assimilation in marine eukaryotic phytoplankton: a case study of micromonas. *Molecular biology and evolution* **27**: 2268-2283.

Moisan, J.R., Hoffman, E.E., Haidvoge, D.B. (1996) Modeling nutrient and plankton processes in the California coastal transition zone. *Journal of Geophysical Research* **101**: 22,677-22,691.

Monier, A., Sudek, S., Fast, N.M., and Worden, A.Z. (2013) Gene invasion in distant eukaryotic lineages: discovery of mutually exclusive genetic elements reveals marine biodiversity. *The ISME journal* **7**: 1764-1774.

Monier, A., Welsh, R.M., Gentemann, C., Weinstock, G., Sodergren, E., Armbrust, E.V. et al. (2012) Phosphate transporters in marine phytoplankton and their viruses: cross-domain commonalities in viral-host gene exchanges. *Environmental microbiology* **14**: 162-176.

Moreau, H., Verhelst, B., Couloux, A., Derelle, E., Rombauts, S., Grimsley, N. et al. (2012) Gene functionalities and genome structure in Bathycoccus prasinos reflect cellular specializations at the base of the green lineage. *Genome biology* **13**: R74.

Notredame, C., Higgins, D.G., and Heringa, J. (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* **302**: 205-217.

Olson, R.J., Zettler, E.R., Chisholm, S.W., and Dusenberry, J.A. (1991) Advances in oceanography through flow cytometry. In *Particle Analysis in Oceanography*. Demers, S. (ed). Berlin: Springer-Verlag, pp. 351-399.

Ottesen, E.a., Young, C.R., Eppley, J.M., Ryan, J.P., Chavez, F.P., Scholin, C.a., and DeLong, E.F. (2013) Pattern and synchrony of gene expression among sympatric marine microbial populations. *Proceedings of the National Academy of Sciences of the United States of America* **110**: E488-497.

Palenik, B., Grimwood, J., Aerts, A., Rouzé, P., Salamov, A., Putnam, N. et al. (2007) The tiny eukaryote Ostreococcus provides genomic insights into the paradox of plankton speciation. *Proceedings of the National Academy of Sciences of the United States of America* **104**: 7705-7710.

Poretsky, R.S., Hewson, I., Sun, S., Allen, A.E., Zehr, J.P., and Moran, M.A. (2009) Comparative day/night metatranscriptomic analysis of microbial communities in the North Pacific subtropical gyre. *Environmental microbiology* **11**: 1358-1375.

Rodríguez, F., Derelle, E., Guillou, L., Le Gall, F., Vaulot, D., and Moreau, H. (2005) Ecotype diversity in the marine picoeukaryote Ostreococcus (Chlorophyta, Prasinophyceae). *Environmental microbiology* **7**: 853-859.

Sambrook, J. (2001) *Molecular Cloning: A Laboratoy Manual*: Cold Spring Harbor Laboratory Press.

Santoro, A.E., Casciotti, K.L., and Francis, C.a. (2010) Activity, abundance and diversity of nitrifying archaea and bacteria in the central California Current. *Environmental microbiology* **12**: 1989-2006.

Sayers, E.W., Barret, T., Bensen, D.A., Bolton, E., Bryant, S.H., *et al*. (2011) Databases resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **39**: D38-D51.

Schmieder, R., Lim, Y.W., and Edwards, R. (2012) Identification and removal of ribosomal RNA sequences from metatranscriptomes. *Bioinformatics* **28**: 433-435.

Six, C., Finkel, Z.V., Rodriguez, F., Marie, D., Partensky, F., and Campbell, D.A. (2008) Contrasting photoacclimation costs in ecotypes of the marine eukaryotic picoplankter *Ostreococcus*. *Limnology and Oceanography* **53**: 255-265.

Slapeta, J., López-García, P., and Moreira, D. (2006) Global dispersal and ancient cryptic species in the smallest marine eukaryotes. *Molecular biology and evolution* **23**: 23-29.

Stewart, F.J., Ulloa, O., and DeLong, E.F. (2012) Microbial metatranscriptomics in a permanent marine oxygen minimum zone. *Environmental microbiology* **14**: 23-40.

Sun, S., Chen, J., Li, W., Altintas, I., Lin, A., Peltier, S. et al. (2011) Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis: the CAMERA resource. *Nucleic Acids Res* **39**: D546-551.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**: 4673-4680.

Treusch, A.H., Demir-Hilton, E., Vergin, K.L., Worden, A.Z., Carlson, C.a., Donatz, M.G. et al. (2012) Phytoplankton distribution patterns in the northwestern Sargasso

Sea revealed by small subunit rRNA genes from plastids. *The ISME journal* **6**: 481-492.

Vaulot, D., Lepère, C., Toulza, E., De la Iglesia, R., Poulain, J., Gaboyer, F. et al. (2012) Metagenomes of the picoalga Bathycoccus from the Chile coastal upwelling. *PloS one* **7**: e39648.

Viprey, M., Guillou, L., Ferréol, M., and Vaulot, D. (2008) Wide genetic diversity of picoplanktonic green algae (Chloroplastida) in the Mediterranean Sea uncovered by a phylum-biased PCR approach. *Environmental microbiology* **10**: 1804-1822.

Wolfe-Simon, F., Grzebyk, D., Schofield, O., and Falkowski, P.G. (2005) the Role and Evolution of Superoxide Dismutases in Algae1. *Journal of Phycology* **41**: 453-465.

Worden, A. (2006) Picoeukaryote diversity in coastal waters of the Pacific Ocean. *Aquatic Microbial Ecology* **43**: 165-175.

Worden, A.Z., and Allen, A.E. (2010) The voyage of the microbial eukaryote. *Current Opinion in Microbiology* **13**: 652-660

Worden, A.Z., Nolan, J.K., and Palenik, B. (2004) Assessing the dynamics and ecology of marine picophytoplankton: The importance of the eukaryotic component. *Limnology and Oceanography* **49**: 168-179.

Worden, A.Z., Cuvelier, M.L., and Bartlett, D.H. (2006) In-depth analyses of marine microbial community genomics. *Trends Microbiol* **14**: 331-336.

Worden, A.Z., Dupont, C., and Allen, A.E. (2011) Genomes of uncultured eukaryotes: sorting FACS from fiction. *Genome Biol* **12**: 117.

Worden, A.Z., Janouskovec, J., McRose, D., Engman, A., Welsh, R.M., Malfatti, S. et al. (2012) Global distribution of a wild alga revealed by targeted metagenomics. *Curr Biol* **22**: R675-677.

Worden, A.Z., Lee, J.-H., Mock, T., Rouzé, P., Simmons, M.P., Aerts, A.L. et al. (2009) Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes Micromonas. *Science (New York, NY)* **324**: 268-272.

Zhu, F., Massana, R., Not, F., Marie, D., and Vaulot, D. (2005) Mapping of picoeucaryotes in marine ecosystems with quantitative PCR of the 18S rRNA gene. *FEMS Microbiol Ecol* **52**: 79-92.

5 Chapter 5: Conclusions and Perspectives

Prasinophytes are unicellular eukaryotic phytoplankton found in a wide range of surface ocean environments. They hold an important phylogenetic position, possessing traits of both land plants and other single-celled green algae, yet have small genome sizes. Some cultured strains have relatively short replication times and are amenable to experimental work in the lab and at sea. Thus, these eukaryotic phytoplankton have useful characteristics that facilitate research on genome architecture, microbial diversity and marine ecosystems.

In this thesis, three main genera within the prasinophyte class Mamiellophyceae were studied with a focus on diversity and distributions in the natural environment. The research presented leverages recent advances (including cost-effectiveness) in genome sequencing and high-throughput sequencing, as well as other established methods such as cloning and sequencing and qPCR, to learn about relatively closely related phytoplankton taxa that are well represented in culture.

At some point, each of the main genera, *Bathycoccus*, *Micromonas* and *Ostreococcus*, were thought to consist of a single species. The results presented here build upon studies that indicated 'cryptic' species were present and being overlooked Specifically, using new approaches combined with phylogenetics, the results herein resolve seven distinct clades of *Micromonas* and two types of genetically diverged *Bathycoccus*. This has important ecological implications relating to evolutionary divergence and niche differentiation.

Observations on diversity and distributions of *Micromonas* built upon culture work to better understand IEs. Exploration of IE-prevalence in six cultured clades of *Micromonas* and the environment, revealed the heterogeneity of the *Micromonas* intronome. Several additional repetitive elements were discovered, which had low identity to the original *M.* CCMP1545 IEs. Additionally, polymorphic RSIs and IEs were observed with loci different from to those seen in cultures. The original *M.* CCMP1545 IEs were found in another Clade D.V strain, but not other *Micromonas* Clades, making Clade D-IEs useful for identifying Clade D.V *Micromonas* in metagenomic samples.

The discovery of a novel repetitive element in Clade C.I strains led to the element's detection in Clades A and B as well – and ultimately in environmental samples. These ABC-IEs were likely in the common ancestor of these three clades, inserting post-divergence from Clades E.III, _.IV and D.V. ABC-IEs were recovered from a broader geographic range of metagenomic data than D-IEs. Furthermore, two types of novel elements, E2-IEt1s and E2-IEt2s, were detected from one *Micromonas* Clade E.III strain, *M.* CCMP2099, but not the other, *M.* CCMP1646. This gave additional support to phylogenetic analyses showing Clade E.III should in fact be split into two separate lineages, Clades E1 (representative strain *M.* CCMP1646) and E2 (representative strain *M.* CCMP2099).

Micromonas was detected for the first time in Antarctic waters by searching environmental metagenomic data for E2-IEs. These Antarctic sequences showed high

identity to *M.* CCMP2099 sequences, an isolate believed to be restricted to the Arctic. E2-IEs were detected in the lowest salinity waters yet reported for this genus, suggesting *Micromonas* has a broader salinity tolerance than previously thought. This may be an important finding in light of global climate change freshening polar waters. *Micromonas* was also found in the deepest sample yet reported, suggesting these cells may sink out of the photic zone and could possibly be transported by deep water currents, a potential explanation for the presence of an Arctic strain in the Antarctic.

It is difficult to comprehend microbial biogeography without a meaningful definition of microbial diversity. For another diversity and distribution study *Ostreococcus* clades were used as a diversity benchmark to interpret data from *Bathycoccus*. This was because three sequenced *Ostreococcus* genomes were available and their natural distributions were better defined than those of most picoeukaryotes, including *Bathycoccus*, which had until recently been considered a single species genus. The *Ostreococcus* results were compared to those from a *Bathycoccus* genome (*B. prasinos*) and three targeted metagenomes, suspected to be from different ecotypes (Vaulot et al., 2012; Monier et al., 2013). *B. prasinos* and two metagenomes isolated from coastal Chilean waters were found to be the same lineage, Clade B1, while a third targeted *Bathycoccus* metagenome from oceanic Atlantic waters was found to be a novel lineage, Clade B2.

Using this *Ostreococcus* and *Bathycoccus* diversity information, ecomarker genes were analyzed from North Pacific Ocean metatranscriptomes collected during

an Eulerian transect and a semi-Lagrangian sampling expedition. The results showed ecotype heterogeneity for both *Ostreococcus* and *Bathycoccus* ecotypes, but clearer niche differentiation for the former, and agreed with qPCR data collected during the studies. Metatranscriptomic data gave some very preliminary insights into transcript abundances at different sites. For example, *Bathycoccus* genes that recruited more metatranscriptomic reads at depth in the open ocean (as opposed to other sites) could be targets for future research on acclimation to conditions such as low light levels and nutrient concentrations. These genes may reflect specific characteristics of *Bathycoccus* clade B2 physiology, since this ecotype dominated (73%) the *Bathycoccus* population at this location. Similar ecomarker analyses on future metagenome and metatranscriptome datasets could produce an informative time series on North Pacific Mamiellophyceae populations. Such work will require replicates, not feasible in the context of the presented research, since replication is essential to statistical analyses allowing concrete conclusions.

Additional work should explore the tolerances of various Mamiellophyceae strains and the potential impacts of repetitive elements on diversity and evolution. As climate change continues, ocean salinities and temperatures will also continue to change and it is important to try and predict how marine microbial populations will change and adapt. The conservation of repetitive elements could be used to estimate nucleotide variant accrual. With this information it may be possible to determine when elemental gain and loss occurred and how these events relate to lineage divergence. Along with the potential impacts of repetitive elements on diversity and

evolution, further investigation into the mechanisms of element gain and loss are warranted.

Finally, one important reason for studying Mamiellophyceae is their intriguing genome architectures. Studies herein investigated features of this architecture, including COPs and IEs, and for the latter their relationships to Mamiellophyceae diversity and biogeography. Future studies should focus on the functions of LGC genomic regions and COPs, including exploration of LGC region intergenic space for the presence of regulating motifs. If present, these motifs could explain the lower gene density of a region located within an otherwise streamlined genome and the increased number of expressed sequence tags from the LGC region. Similarly, the overlapping UTR regions of COP sequences should be searched for conserved motifs, such as shared promoters. To test whether COPs result from positive selection for a particular function, the likelihood of non-random COP formation should be evaluated by creating a random prediction model using empirical transcript length data to confirm non-random occurrence of COP genes. Researching patterns of COP conservation between sequenced Mamiellophyceae species may reveal interesting details about gene regulation, as could the comparison of COP gene expression to the expression of non-overlapping gene orthologs.

The work presented in this thesis leveraged recent advances and established methods to reveal complex intron patterns, a new way to detect *Micromonas* clades in metagenomic datasets, additional clade levels of diversity and ecomarkers for

biogeography mapping from metatranscriptomes and metagenomes. By providing new information on Mamiellophyceae diversity and biogeography, hopefully this research provides a foundation for future studies on the environmental factors behind observed patterns.

References

Monier, A., Sudek, S., Fast, N.M., and Worden, A.Z. (2013) Gene invasion in distant eukaryotic lineages: discovery of mutually exclusive genetic elements reveals marine biodiversity. *The ISME journal* **7**: 1764-1774.

Vaulot, D., Lepere, C., Toulza, E., De la Iglesia, R., Poulain, J., Gaboyer, F. et al. (2012) Metagenomes of the Picoalga Bathycoccus from the Chile Coastal Upwelling. *PLoS ONE* **7**.