Spring 2017

Phylogenetic-Based Characterization of Microbial Eukaryote Community Structure and Diversity of an Estuary in the Salish Sea

Taylor L. Clement
Western Washington University, tclem26@gmail.com

PHYLOGENETIC-BASED CHARACTERIZATION OF MICROBIAL EUKARYOTE COMMUNITY STRUCTURE AND DIVERSITY OF AN ESTUARY IN THE SALISH SEA

By

Taylor L. Clement

Accepted in Partial Completion
of the Degree requirements for the Degree
Master of Science

Kathleen L. Kitto, Dean of the Graduate School

ADVISORY COMMITTEE

Chair, Dr. Robin Kodner

Dr. Heather Fullerton

Dr. Craig Moyer

Dr. Suzanne Strom

MASTER'S THESIS

In presenting this thesis in partial fulfillment of the requirements for a master's degree at Western Washington University, I grant to Western Washington University the non-exclusive royalty-free right to archive, reproduce, distribute, and display the thesis in any and all forms, including electronic format, via any digital library mechanisms maintained by WWU.

I represent and warrant this is my original work, and does not infringe or violate any rights of others. I warrant that I have obtained written permissions from the owner of any third party copyrighted material included in these files.

I acknowledge that I retain ownership rights to the copyright of this work, including but not limited to the right to use all or part of this work in future works, such as articles or books.

Library users are granted permission for individual, research and non-commercial reproduction of this work for educational purposes only. Any further digital posting of this document requires specific permission from the author.

Any copying or publication of this thesis for commercial purposes, or for financial gain, is not allowed without my written permission.

Taylor Clement

April 27, 2017

PHYLOGENETIC-BASED CHARACTERIZATION OF MICROBIAL EUKARYOTE COMMUNITY STRUCTURE AND DIVERSITY OF AN ESTUARY IN THE SALISH SEA

A Thesis
Presented to
The Faculty of
Western Washington University

In Partial Fulfillment
Of the Requirements for the Degree
Master of Science

By
Taylor Clement
April, 2017

ABSTRACT

Coastal estuaries are highly dynamic environments. Due to a multitude of environmental variables, microbial communities in these systems are generally complex and difficult to predict. A majority of studies in estuaries have focused on relating environmental factors to changes in community response through indirect measurements such as biomass or chlorophyll. Though these studies are important for our understanding of these systems, they treat community members as a "black box" by focusing on the environmental input and biological output of the system on a broad scale. However, community composition and diversity dictates how cohabitants respond to both environmental stimuli and each other, which in turn impacts their community-level response. We utilized high throughput sequencing of the small subunit ribosomal RNA gene to phylogenetically characterize the diversity of microbial eukaryotes in a coastal estuary located in the Northeastern Salish Sea, and explore spatial and temporal patterns in community structure. Sampling was carried out daily to biweekly for four years during the summer seasons of 2013 to 2016. While diatoms and dinoflagellates were found to be the most abundant amplicon sequences in our samples, there was a diverse assemblage of less represented amplicon sequences that showed spatial and temporal patterns. We also found that our assessment of stability in this system varied based on timescale. Our results highlight the importance of repeated sampling to characterize microbial eukaryotes in dynamic environments, and the importance of including small and/or rare taxa in future ecological studies of these systems. We also discuss current challenges and limitations of this methodological approach to evaluate microbial eukaryotes.

ACKNOWLEDGEMENTS

First and foremost, my warmest thanks go out to my advisor, Dr. Robin Kodner, for her patience, support, and guidance, and for the opportunities she provided throughout the course of my thesis research. Additionally, thanks to Dr. Craig Moyer for his expertise in microbial genomics and for graciously allowing me to utilize some of his lab equipment. Thanks to Dr. Heather Fullerton for her invaluable support and wisdom during troubleshooting and lab setbacks, as well as manuscript editing. Also, thanks to Dr. Suzanne Strom, for her help in analyzing the community composition in my dataset. This thesis would not have been accomplished without the collaboration of multiple undergraduate assistants, who completed the cell counts and assisted with data collection, including: Ciara Asomato, Rachel Lekanoff, and Staci Nazario. Thank you to the "carpet scientists" of the Kodner Lab, Eric Hervol and Ryan McLaughlin, for providing endless guidance and support during my data analysis, and Kevin Hager, for helping me work through the lab protocols. Big thanks to the captains during research cruises: Nate Schwarck and Jay Dimond, Marco Hatch, and Andrés Quesada. I would also like to express my gratitude toward my family, fellow graduate students, and the faculty and staff at WWU as well as the Salish Sea Research Center at Northwest Indian College (NWIC). This work would not have been possible without the support of WWU's Office of Research and Sponsored Programs providing a Project Development Award to Dr. Kodner and a Graduate Research Award to me in 2015. I also received generous funding from Rosemary Flora, through the Flora Scholarship and summer support through the Biology Department's Fraser Endowment.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

Marine microbial eukaryotes are incredibly diverse groups of organisms spanning multiple eukaryotic superkingdoms. They fill numerous key ecological roles as primary producers, symbionts, parasites, mixotrophs, and phagotrophs (Sherr *et al.* 2007). Photosynthetic members of this community contribute a significant portion of marine global primary production, with a majority of it occurring in coastal regions (Field *et al.* 1998; Worden *et al.* 2004; Massana 2011). Because they have important roles in nutrient cycling and carbon fixation and form the basal trophic level (Arrigo 2005; Jardillier *et al.* 2010), changes within these communities directly and indirectly impact every part of the food web (Michaels 1988; Frederiksen *et al.* 2006). These community changes can be on short or long timescales and are often catalyzed by shifts in physical or chemical environmental conditions, such as temperature and nutrient availability (Vigil *et al.* 2009; Kim *et al.* 2014).

Members of this group comprise some of the earliest divergence of eukaryotic lineages and have diverse evolutionary histories (Falkowski *et al.* 2004). Within marine environments, distantly related organisms may fill similar roles as primary producers and prey. In turn, many phylogenetically similar organisms fill different ecological roles. An example of similar organisms fulfilling different niches can be observed within the species level of *Ditylum brightwellii*, a common coastal diatom, where two co-occurring strains of different cell-sizes showed differential abundances during subsequent blooms (Rynearson & Armbrust 2005; Rynearson *et al.* 2006). During the first seasonal bloom, a genetically distinct strain with smaller cells comprised a majority of the population, and during a later bloom, a strain with larger cells was more abundant, suggesting different physiological capabilities and interactions with environmental conditions (Rynearson & Armbrust 2004;

Rynearson *et al.* 2006). Dinoflagellates also exhibit divergent ecological niches, where closely related species can be parasites, autotrophs, phagotrophs or mixotrophs (Sherr *et al.* 2007; Worden & Not 2008). In addition, many diverse groups of phytoplankton form harmful algal blooms (HABs), including diatoms, raphidophytes, and dinoflagellates; however, not all members of these lineages produce harmful blooms, and some members rarely form blooms.

In order to understand the roles and relationships between microeukaryote groups in the environment, we need to be able to characterize natural communities. This can be done microscopically, but recently large-scale high throughput sequencing (HTS) of microbial eukaryote DNA has revealed extensive hidden genetic diversity not captured with microscopic methods (Logares *et al.* 2014; Shade *et al.* 2014; Ignatiades & Gotsis-Skretas 2014; de Vargas *et al.* 2015). Specifically, taxa that are low in abundance or small in size, such as picoplankton (0.2-2 μm) and nanoplankton (2-20 μm), were especially difficult to characterize and often escaped microscopic identification. Using HTS methods, these groups have since been found to be highly abundant and diverse in marine systems (Moon-van der Staay *et al.* 2001; Worden *et al.* 2004; Worden & Not 2008; Massana 2011).

Though numerous studies characterizing marine microeukaryote diversity exist, until recently a relatively small proportion of studies have employed a clone-free HTS approach to characterize highly dynamic estuarine systems (Lallias *et al.* 2014; Abad *et al.* 2016; Brannock *et al.* 2016; Marquardt *et al.* 2016; Xu *et al.* 2017). Coastal estuarine systems are complex, dynamic environments, characterized by high rates of primary productivity (Boyle & Silke 2010). Microbial communities in these systems are simultaneously subject to multiple environmental stimuli including: the atmosphere, inflow from offshore waters,

freshwater inputs, the nearby terrestrial environment, and underlying sediments (Cloern 1996; Boyle & Silke 2010; Cloern & Jassby 2010). As a result of their dynamic environment, microbial communities in these regions may operate on smaller timescales than microbial eukaryotes in open-ocean systems (Sverdrup 1953; Cushing 1959; Cloern & Jassby 2010), leading to variable, somewhat unpredictable community dynamics (Boyle & Silke 2010).

A study comparing temporal phytoplankton patterns in 84 estuarine-coastal sites suggested that a universal rule to describe the dominant timescales of phytoplankton dynamics in these systems may not exist (Cloern & Jassby 2010). Though it may be true that coastal and estuarine communities are unpredictable, this particular study used only chlorophyll *a* (chl *a*) to measure temporal changes in phytoplankton. This method quantifies photosynthetic biomass but does not capture taxonomic patterns within photosynthetic groups and does not profile the whole community structure. An inherent variability in the taxonomic composition of microbial eukaryotic communities may impart a competitive advantage for these communities, if cohabitants can respond quickly to ecosystem change. A dynamic response allows for rapid community reassembly but is diversity dependent (Caron & Countway 2009). Since assembly and reassembly of community structure cannot be determined through chl *a* measurements alone, the next step is to use a more targeted approach that provides taxonomic context for community structure (Cloern & Jassby 2010). Microscopy has traditionally been used to characterize community structure, but this method is not comprehensive because it cannot always distinguish between species and cannot identify rare or small taxa. In contrast, HTS is a promising way to capture full community diversity because it characterizes the whole community with a single method. Studies using molecular approaches to explore estuarine microbial diversity have greatly expanded our

characterization of these environments. A number of studies have found sequences identified taxa that were unlikely to be seen using solely light microscopy, including Cryptophyta, Prasinophyta, Haptophyta, Oomycota, Perkinsozoa, Chrysophyta, Apicomplexa, and Xanthophyceae (Romari & Vaulot 2004; Herfort *et al.* 2011; Bazin *et al.* 2013, 2014; Lallias *et al.* 2014; Brannock *et al.* 2016; Marquardt *et al.* 2016; Xu *et al.* 2017). In addition, many sequences from these studies were annotated as "uncultured" and "environmental" eukaryotes, categories of sequences from environmental studies that are not well characterized and have not been cultured in the laboratory.

Characterization of microbial eukaryote community structure and diversity is an important component in understanding estuarine ecology, but it is challenging to capture this structure because this system is highly dynamic. Because HTS can capture most of the microbial community diversity in an individual sample, it holds the potential to accomplish a broad-scale analysis of estuarine community structure. To test this, we surveyed community structure and diversity on a range of timescales from Bellingham Bay, our local highly dynamic estuarine system located in Northwest Washington State.

Bellingham Bay is a shallow, coastal embayment of the Salish Sea (Figure 1), and is subject to both naturally occurring and anthropogenic activity from adjacent terrestrial and offshore environments (Nesbitt *et al.* 2015). It is influenced heavily by snowmelt from the North Cascades via the Nooksack River and circulation from the Strait of Georgia (Wang *et al.* 2010; Nesbitt *et al.* 2015). This semi-urban body of water is critically important to the local economy and tribal community, and experiences seasonal hypoxia and phytoplankton

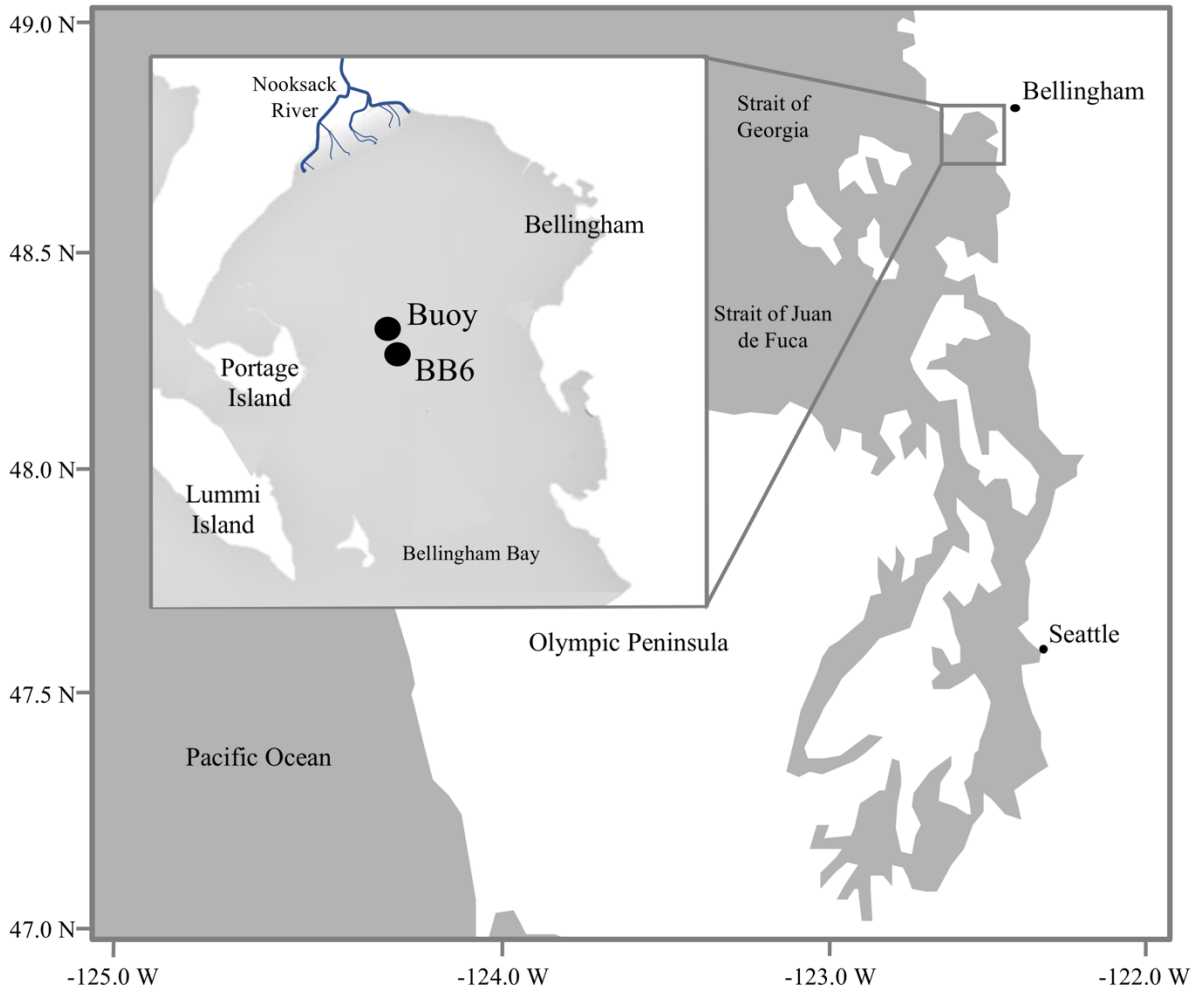


Figure 1. Map of Northwestern Washington State and the Salish Sea, with an insert of Bellingham Bay sampling site locations (BB6: 48.7155 N, -122.561 W and the buoy: 48.7237 N, -122.5765 W) on the coast of Bellingham, Washington.

blooms, including harmful algal blooms (HABs) (Nesbitt *et al.* 2015). Our study was designed to 1) broadly observe the microbial eukaryote community structure in the environment and to resolve any possible spatial or temporal patterns, 2) characterize phylogeny-based diversity changes over time, and 3) compare community variation over yearly, weekly, and daily timescales. To accomplish these goals, we employed a HTS approach to sample the center of the bay over the summer of 4 years. Over the 4 years of the study, we generated 70 samples, with sampling efforts ranging from every 2 weeks to daily.

METHODS

Sample collection

CTD casts and discrete water samples were completed in Bellingham Bay, in northwestern, Washington State from 2013-2016. Summer 2013 samples were collected every 2 weeks (July 11, 2013 through August 15, 2013) from the center of Bellingham Bay (Figure 1, Supplementary Table S1). In 2014, sampling efforts were increased to include 1 spring sample from April 22, 2014 and 1 sample per week from June 24, 2014 through August 8, 2014. In 2015 sampling effort remained the same as in 2014 from June 24, 2015 through August 19, 2015. In 2016 sampling was done on a weekly basis from June 30, 2016 until August 22, 2016, with additional high frequency sampling occurring every day between August 8th and August 12th.

During each sampling event, discrete water samples were collected using the CTD Niskin rosette at two functional depths: chlorophyll maximum and deep. The chlorophyll maximum (CM) was determined by measuring chl *a* in situ fluorescence with a CTD. Deep samples were functionally defined as samples collected 1 m above the bottom. In 2016, the CM was moved to a surface sample (1 m below surface), to align with sensors affixed to a buoy that was deployed at the sampling site. This was justified after a pilot study comparing surface to CM in the bay suggested minimal taxonomic differences in populations at these depths, likely because the chlorophyll maximum is shallow in the bay during the summer (average depth = 5.6 m, S.D. = 3.1) (Supplementary Figure S1, Supplementary Table S1). After water was collected at each depth, Supor-200 0.2 μm Pall filters (47 mm) (Pall Corporation, Port Washington, NY) were used to filter 500 mL of the seawater, which were frozen at -20°C before DNA extraction.

Aliquots of seawater (50 mL) were also taken from CM/Surface and deep samples and fixed with 300 μL of acid Lugols for cell counts. Samples were settled in Utermöehl chambers for 48 h. Quantitative cell counts were completed using a Palmer cell counter at 200X on an inverted light microscope. Each member of chain-forming cells was counted as an individual. Though we collected environmental data at the time of sampling, it is not presented, as this study does not attempt to make any assertions about specific environmental drivers of community patterns.

DNA was extracted from the frozen 0.2 μm Pall filters using a Qiagen DNeasy Plant mini kit (Qiagen, Germantown, MD), following the manufacturer's protocol, with a minor modification in the first steps to increase the recovery of organisms. First, the filters containing the collected cells were flash-frozen in liquid nitrogen, then cut using sterile scissors, and ground using a micropestle to help break up diatom frustules. After denaturing the cells via adding the extraction buffer (AP1) and RNase, the samples were heated at 65°C for 10 min. The liquid was then pipetted off and placed in a centrifuge tube. After completion of these steps, the standard extraction protocol was followed.

To analyze the eukaryotic community diversity, the 418-420 basepair (bp) V4 region of the SSU rRNA gene was amplified via polymerase chain reaction (PCR) using the primers TAReuk454FWD1 (5′-CCAGCA(G/C)C(C/T)GCGGTAATTCC-3′, *S. cerevisiae* position 565-584) and TAReukREV3 (5′-ACTTTCGTTCTTGAT(C/T)(A/G)A-3′, *S. cerevisiae* position 964-981) (Stoeck *et al.* 2010). To be compatible with Illumina MiSeq (Illumina Inc., San Diego, CA) technology, both the forward and reverse primer were synthesized together with an overhang adaptor sequence suggested by Illumina (5' TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG was attached to the forward primer

and 5' GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG was attached to the reverse primer) (Part # 15044223 Rev. B; Illumina Inc.). Amplicons were generated using a two-step 25 μL reaction PCR using Apex 2x Taq Master Mix (Genessee Scientific, San Diego, CA) with the following conditions: 95°C for 5 min, followed by 10 cycles consisting of 94°C for 30 s, 57°C for 45 s, and 72°C for 1 min, which was followed by 25 further cycles consisting of 94°C for 30 s, 48°C for 45 s, and 72°C for 1 min; with a final 2 min extension at 72°C and hold at 4°C. The resultant amplicon libraries were cleaned using the Qiagen Qiaquick PCR Clean-up kit (Qiagen, Hilden, GER.). These purified libraries were then indexed using the Illumina Nextera XT index kit (Illumina Inc., San Diego, CA) for parallel high throughput sequencing. The 50 μL reaction for the index PCR was performed with 2X KAPA HiFi HotStart ReadyMix (Kapa Biosystems, Wilmington, MA), and the conditions were: 95°C for 3 min, 8 cycles of 95°C for 30 s, 30 s at 55°C, and 30 s at 72°C, with a final elongation at 72°C for 5 min. After, the samples underwent a second PCR cleanup, this time using Agencourt AMPure XP beads (Beckman Coulter, Brea, CA) with the Illumina directed protocol (Illumina Inc.).

DNA Concentrations and size fragments of purified libraries were quantified with a Qubit 2.0 fluorometer using the dsDNA HS kit (Thermo Fisher Scientific, Waltham, MA), validated with an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA), brought to 4 nM, pooled, and then reanalyzed with the Qubit. The pooled library was brought to a final concentration of 6 pM and 20% 6 pM PhiX was spiked in. Sequencing was performed using the 500 cycle V2 kit (Illumina Inc.) on an Illumina MiSeq following the manufacturer's protocol, which generates 2 x 500 bp reads.

Sequence Analysis

After sequencing was completed, output sequences were analyzed using FastQC (Andrews 2010). Forward and reverse reads were combined using Paired-End reAd mergeR (PEAR) (Zhang *et al.* 2014). The FASTX-Toolkit was used to trim and quality filter samples; specifically, fastx_trimmer to trim forward and reverse primers off sequences, fastq_quality_trimmer to trim off basepairs with a Q-score below 28, and fastq_quality_filter to filter out any reads below 75% confidence (http://hannonlab.cshl.edu/fastx_toolkit/). The remaining files were then deduplicated using Vsearch (Rognes *et al.* 2016), and put through a pipeline which uses HMMER hmmsearch to align reads to a pre-computed reference set of microeukaryote SSU rRNA gene sequences (Matsen *et al.* 2010; Mclaughlin 2016) and pplacer to place reads on a pre-computed microeukaryote SSU rRNA gene tree (Matsen *et al.* 2010). The reference package for pplacer was built using PhyloSift (Darling *et al.* 2014) and a custom micro-eukaryote focused reference-database that includes the Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP) (Keeling *et al.* 2014).

Using the output from pplacer, annotation tables were manipulated for taxon-specific analyses using NCBI taxonomy. The tables were reduplicated to reflect counts of each classification. Sequences classified as Opisthokonta were filtered out from the tables, although this did remove choanoflagellates. Rhodophyta and Streptophyta were also filtered out, as they were very sparsely represented in the sequence libraries, and considered to be mainly artifacts of land plants (i.e. pollen, dead plant matter), and seaweed detritus or spores. Lastly, in each sample, classifications which had fewer than 10 representative sequences were also removed from analyses. While this may have caused us to miss some extremely rare community members, studies have shown that these are likely artifacts of sequencing,

and can give false impressions of richness and the appearance of transient taxa (Stoeck *et al.* 2009; Brown *et al.* 2015; Lynch & Neufeld 2015). These extremely rare sequences could not be removed from the diversity metric calculations, but because weighted measures were used, they likely had minimal impact on final calculations.

To avoid missing patterns due to overrepresentation of some sequences in the dataset and to more accurately assess trends in all of the phyla present (not just abundant groups), amplicon sequences were sorted into two groups: less-represented amplicon sequences (LRAS) and highly-represented amplicon sequences (HRAS). LRAS were defined as sequences annotated within phyla that were found in abundances lower than 1% of the total population for at least 3 years of the study. This included every phylum in the sequence libraries except Dinoflagellata (dinoflagellates) and Bacillariophyta (diatoms). Most subsequent analyses were completed twice, first with both the HRAS and LRAS included, and then with only LRAS.

Diversity and statistical analyses

The guppy analysis package, part of the pplacer software package, was used to explore ecological patterns. Edge principle components analysis (ePCA) was used to examine whole community structure within each sample. In ePCA, principle components are derived from a large distance matrix calculated from the distances between reads on the same reference tree across samples (Matsen IV *et al.* 2013). This method was used to compare samples by using phylogenetic relationships between reads in each sample. Each ePCA returns a phylogenetic tree associated with each principle component that describes the

lineages influencing each principal component (PC) (Matsen IV *et al.* 2013). These trees were viewed in Archaeopteryx (Han & Zmasek 2009).

Two diversity measures were also computed using the guppy package. The first was balance-weighted phylogenetic diversity (BWPD), which measures alpha diversity, but is more robust than classical measures as it incorporates the similarity of sequences more efficiently than discrete name-based measures (McCoy & Matsen 2013). The BWPD metric used an un-rooted tree, and weighted tree edges based on the balance of neighboring read fractions (McCoy & Matsen 2013). This approach helped normalize amplicon data, so BWPD more closely represented community diversity. The second diversity measure was the weighted Unifrac distance, the phylogenetic representation of beta diversity (Evans & Matsen 2012). The weighted Unifrac distance computes the distance between placements from each sample on a tree to quantitatively describe weighted sequence distributions (Evans & Matsen 2012). Comparison of both depths each year were made by merging sample files, reducing 72 individual samples to 8 merged samples.

Rarefaction curves were completed on low-read samples using guppy to explore sequencing saturation (Supplementary Figure S2). In addition, the R 3.2.3 package Heatmap.2 was used to make heatmaps and ggplot was used to make a bubble plot for comparing the similarity and/or abundance of taxa within and between samples.

RESULTS

Sequencing Results

HTS produced 19,432,818 total reads, and 84.67% of these had a Phred quality score greater than or equal to 30 (indicating 99.9% base call accuracy), producing a final count of 8,088,642 paired-end reads (contigs). After quality filtering and trimming, 4,545,710 amplicons were analyzed using the pplacer pipeline (Matsen *et al.* 2010). Of these, 3,060,881 were placed on the phylogenetic tree produced by the reference package (Supplementary Table S2). Unplaced sequences did not pass the HMMER filter of a e^{-10} cutoff. Of the successfully placed amplicons, 168,856 (5.5%) were classified as "no match", "unclassified", or "environmental sample" (Supplementary Table S3).

Patterns in Dominant Amplicon Sequences of Bellingham Bay

Using a taxonomy-based approach derived from phylogenetic data, the presence and proportions of successfully placed and classified sequences from Bellingham Bay were characterized using the annotation tables produced from the reduplicated output files of pplacer. At a high taxonomic level of classification, the communities appeared fairly stable year-to-year, and were dominated by Stramenopiles and Alveolates (Figure 2). A majority of the Stramenopile sequences were classified as diatoms, and the vast majority of Alveolate sequences were classified as dinoflagellates. Consequently, the most consistent pattern in the dataset was that diatom and dinoflagellate amplicon sequences were consistently highly represented at both the CM and at depth in each year of the study, while other lineages were variably present, and generally found in much lower proportions (Figure 3). Dinoflagellate sequences were dominant overall, and diatoms were the second-most

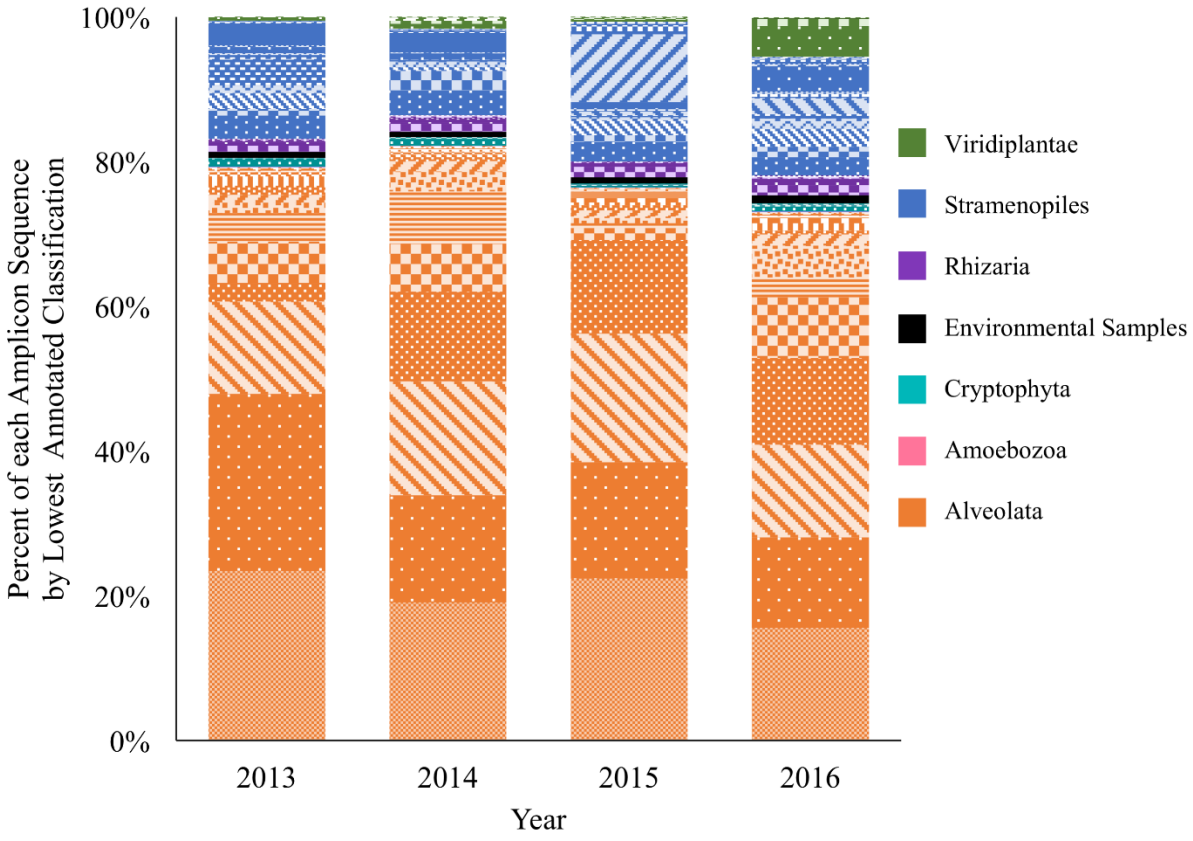


Figure 2. Stacked bar graph showing the relative proportions (as a percent) of broad-level classification (color) and lowest-level classification (pattern) from annotations of microbial eukaryote SSU rRNA gene sequences from each year of sampling (2014-2016) in Bellingham Bay, WA.

abundant. The dominance of diatoms and dinoflagellates in estuarine environments is well characterized, and the dynamics within these groups are well studied. As such, the dominance of diatoms and dinoflagellates in the Bellingham Bay dataset was expected. However, the relative proportions of these groups in Bellingham Bay contradict preliminary cell count data, which found higher relative proportions of diatoms than dinoflagellates (Supplemental Figure S2). The overrepresentation of dinoflagellate sequences was consistent in all of the samples.

Year-to-year, the highly-represented amplicon sequences (HRAS) contained the same diatom taxa based on annotations, and their relative proportions were fairly stable in both CM and deep samples (Figure 3). The diatoms were very consistent over time, with the exception of a noticeable increase in the relative abundance of sequences classified as Coscinodiscophyceae in 2016 (Figure 3, Figure 4), and a high proportion of sequences only classified to the phylum level (Bacillariophyta) in 2013. The dinoflagellates were slightly more variable. In the amplicon data, there were numerous sequences found in the deep samples of 2014 and 2016 that were classified within the dinoflagellate lineage as environmental samples (Figure 4). These represent sequences observed in other HTS studies, but not thoroughly characterized. Additionally, the proportions of Polykrikaceae increased substantially in the CM each year over the four-year study while the Peridiniales decreased in proportion each year (Figure 3, Figure 4). This variation was not observed in the deep samples.

Sequencing also revealed the presence of a parasitic dinoflagellate genus, *Amoebophrya* (Figure 3, Figure 4), which parasitizes numerous dinoflagellates that appeared in Bellingham Bay cell counts (Supplementary Table S4, Supplementary Table S5).

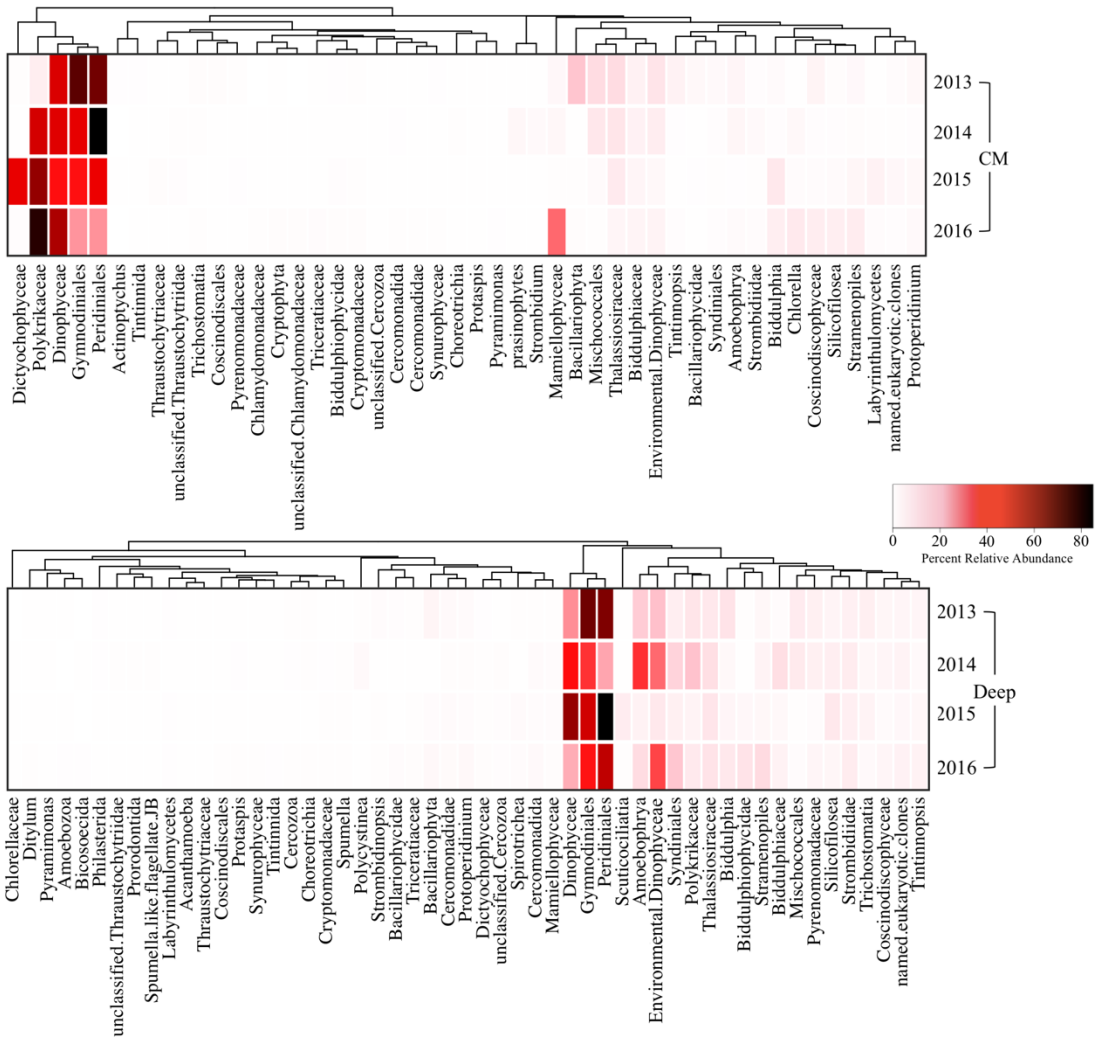


Figure 3. Heatmap showing the relative proportions (as a percentage) of each uniquely placed microbial eukaryote SSU rRNA gene sequence found in chlorophyll maximum (CM) **(top)** and deep (1m above the bottom) samples **(bottom)** during each year of sampling in Bellingham Bay, WA from 2014-2016. Sequence annotations are shown at the lowest classification level from annotations based on alignment to a fixed reference tree. Sequences representing less than 0.1% of the dataset were removed. A dendrogram was utilized to cluster the sequences by similarities in relative abundance patterns. The scale bar indicates the colors associated with relative abundance, which range from black (85% of the total sequences) to red (43.5%) to white (0%).

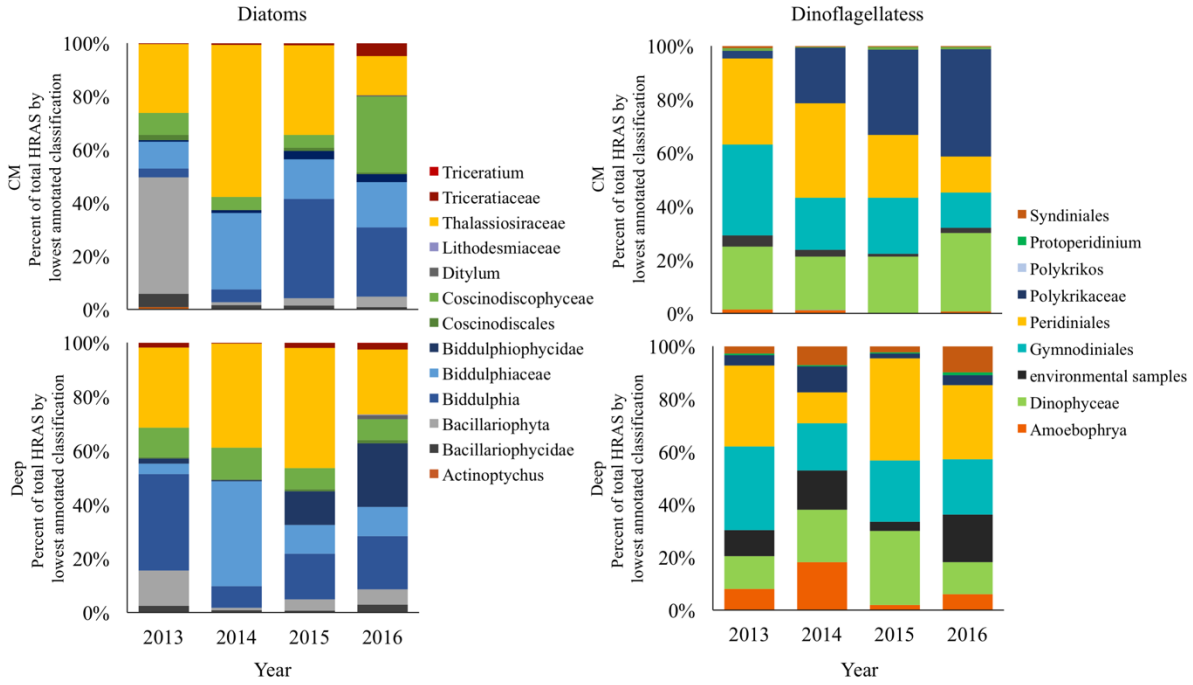


Figure 4. Stacked bar graphs indicating the relative abundance of Diatom SSU rRNA gene sequences in the chlorophyll maximum **(top left)** and deep samples **(bottom left)** of and Dinoflagellate sequences in the chlorophyll maximum **(top right)** and deep samples **(bottom right)** in Bellingham Bay, WA each year from 2014-2016. Deep samples were collected 1m above the bottom of the bay. Similar shades of the same color indicate differing levels of classification within highly similar taxonomic groups.

This parasitic genus was more abundant in the sequence libraries of 2013 and 2014 than the latter two years' libraries, with a peak in relative abundance in 2014, but there were no inverse relationships found between the sequence abundance of *Amoebophrya* and any other dinoflagellates (Figure 3).

Patterns in the less-represented amplicon sequences

In addition to the highly abundant dinoflagellate and diatom sequences, the samples in Bellingham Bay included numerous amplicon sequences found in particularly low abundance (Figure 3). The relative proportions of LRAS were overall more variable than the HRAS (Figure 5). Some sequences were present in both the CM samples and deep samples each year, while others were only found predominantly at one of the sampled depths (Figure 5). The proportions of LRAS in the CM were fairly dynamic over time, while the proportions of LRAS in the deep samples were comparatively more stable over the four-year sampling period (Figure 5). The deep samples were dominated by sequences classified within: Rhizaria, environmental eukaryotes, Stramenopiles, Cryptophyta, and Ciliophora (Alveolates) (Figure 5), though a few taxa within this groups were also found in similar proportions in the CM. Though there were sequences that remained stable over time in the CM, the samples also contained a few LRAS found in highly fluctuating proportions between years. These variable groups included: Mischococcales and Dictyochophyceae (both Stramenopiles), as well as Prasinophytes, *Chlorella* and Mamiellophyceae (picoplanktonic Viridiplantae) (Figure 5). These specific amplicon sequences were all identified as photosynthetic taxa that either fall into the pico- or nanoeukaryote size range or are generally found in low abundance in Bellingham Bay. Because the reads were found in low

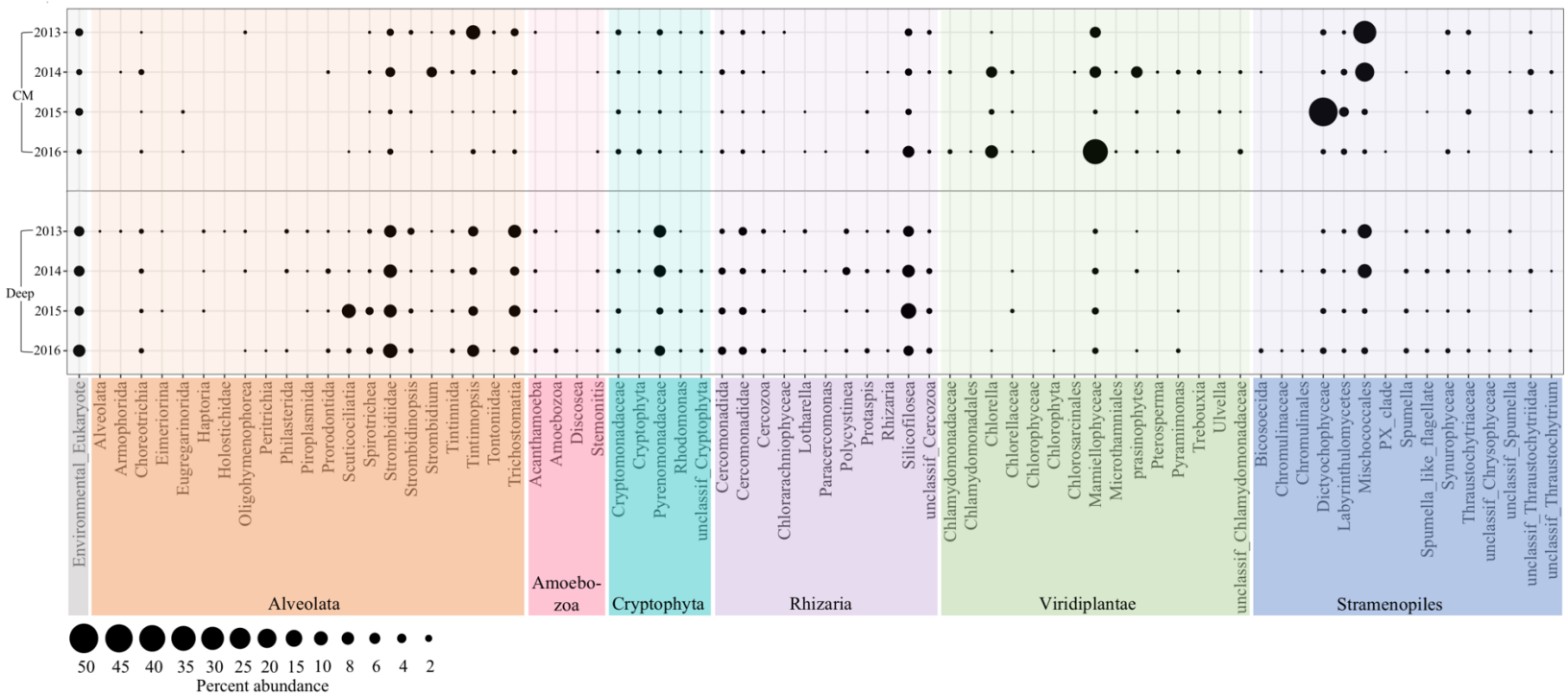


Figure 5. Bubble plot indicating the relative proportions (as a percent) of LRAS (sequences that were successfully classified, but found in low abundances each year (2013-2016) at each of the two depths sampled based on SSU rRNA gene sequences. Samples were collected) from the chlorophyll maximum (CM) and 1m above the bottom (Deep). Diatoms and Dinoflagellate sequences were removed from the analysis, and sequences accounting for less than 0.001% of the total sequences each year were also removed. These included sequences annotated as: *Eimeria*, *Intramacronucleata*, Stemonitidae, *Tubulinida*, Pyrenomonadales, Euglyphidae, and *Prasiola*. Taxonomic classifications are listed at lowest level of annotation, and colored by higher taxonomic ranks.

proportions for a majority of the study, their increased representation in a few samples may have ecological significance. The variation in abundance within specific taxa over time was less likely to be associated with large disparities in copy number, and was more likely to represent small-scale blooms of these primary producers.

To assess the effect of sequencing depth, two samples were compared from 2015 (one CM and one deep) that were sequenced twice at different sequencing efforts. Greater sequencing effort did reveal sequences from taxa that were absent from the results of the smaller sequencing effort runs. Specifically, there were 4 sequences present in the "increased effort" samples from the deep samples that were not found in any of standard sequencing effort used across the study. These included 3 classified as ciliates and an Apicomplexan. The increased-effort sequencing data also contained 2 different sequences in the CM that were not present in our dataset, including one of the same ciliates not observed in the deep samples, and a Chrysophyte. When the taxonomic annotations were pooled by year we missed environmental Ciliophora sequences because they were not present in any of the 2015 sequence libraries, but this lineage was present in other years. However, all the sequences not observed due to shallower sequencing depth were found in extremely sparse quantities in the deeply sequenced samples.

Edge principle components analysis was used to investigate how samples varied by year, week, and depth, within all microeukaryote sequences (HRAS and LRAS combined) and only the LRAS. We found that 56.75% of the variability in the combined microbial eukaryote data was explained by PC1, and 11.7% by PC2 (Figure 6a). The samples did not cluster by year or week, but did cluster loosely by depth (Figure 6a). When the HRAS were

removed, samples still did not cluster by year, clustered slightly by week, and clustered more tightly between depths despite the PCs explaining less of the variability in these sequences, with PC1 accounting for 38.24% of the variability, and PC2 accounting for 23.85% (Figure 6b). Clustering was more distinct in PC1, with surface samples on the left and deep samples on the right. Though there was no strong evidence of clustering by week, none of the samples from the beginning of any sampling seasons were found on the upper portion of the graph. The corresponding phylogenetic tree shows that LRAS classified as Viridiplantae (Trebouxiophytes and Mamiellophyceae) were responsible for clustering samples on the left side of the plot, while sequences identified as Alveolates (Ciliophora) and Rhizaria (Cercozoa) were responsible for samples clustering on the right side of the plot (Figure 7a). Though the deep samples still clustered within PC2, the CM samples encompassed a larger range (Figure 6b). Amplicon sequences classified as Stramenopiles (Labrynthulomycetes and Dictochophyceae) were responsible for clustering samples on the upper region of the plot, while sequences classified as Viridiplantae (Trebouxiophytes and Mamiellophyceae), Alveolates (Ciliophora), and Rhizaria (Cercozoa) were responsible for samples clustering on the lower region of the plot Figure 7b). The loose clustering in the CM samples on PC2 may be occurring because Stramenopiles are more closely related to Rhizaria and Ciliophora than Viridiplantae, or may be due to the decent amount of some ciliates (Strombiidae and Tintinnids) that show up in the CM and deep sequences. This indicates that while there was variation over time, there was considerable overlap in the community structure between years, some slight separation by week, and a more distinct pattern in community structure due to depth.

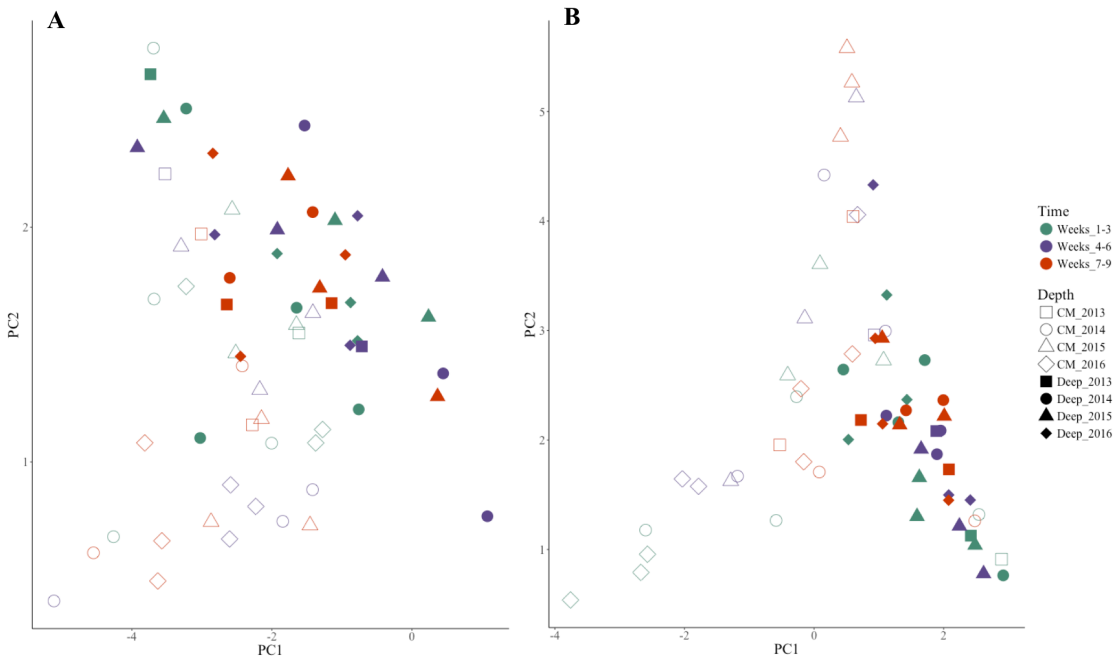


Figure 6. Edge Principle Components of Bellingham Bay samples patterned by different timescales and depths **A**. PC1 and PC2 of microbial eukaryote sequences (LRAS and HRAS) by year, with 56.57% of the variability is explained by PC1, and 11.70% explained by PC2. Different shapes represent different years of sampling. Weeks 1-3 of each year are colored teal, 4-6 are purple, and 7-9 are red. Filled shapes are deep samples (taken 1 m above the bottom of Bellingham Bay) and non-filled shapes are samples taken from the chlorophyll maximum (CM). **B**. PC1 and PC2 of less-represented amplicon sequences (LRAS) using the same color and pattern template, with 38.24% of the variability is explained by PC1, and 23.85% is explained by PC2.

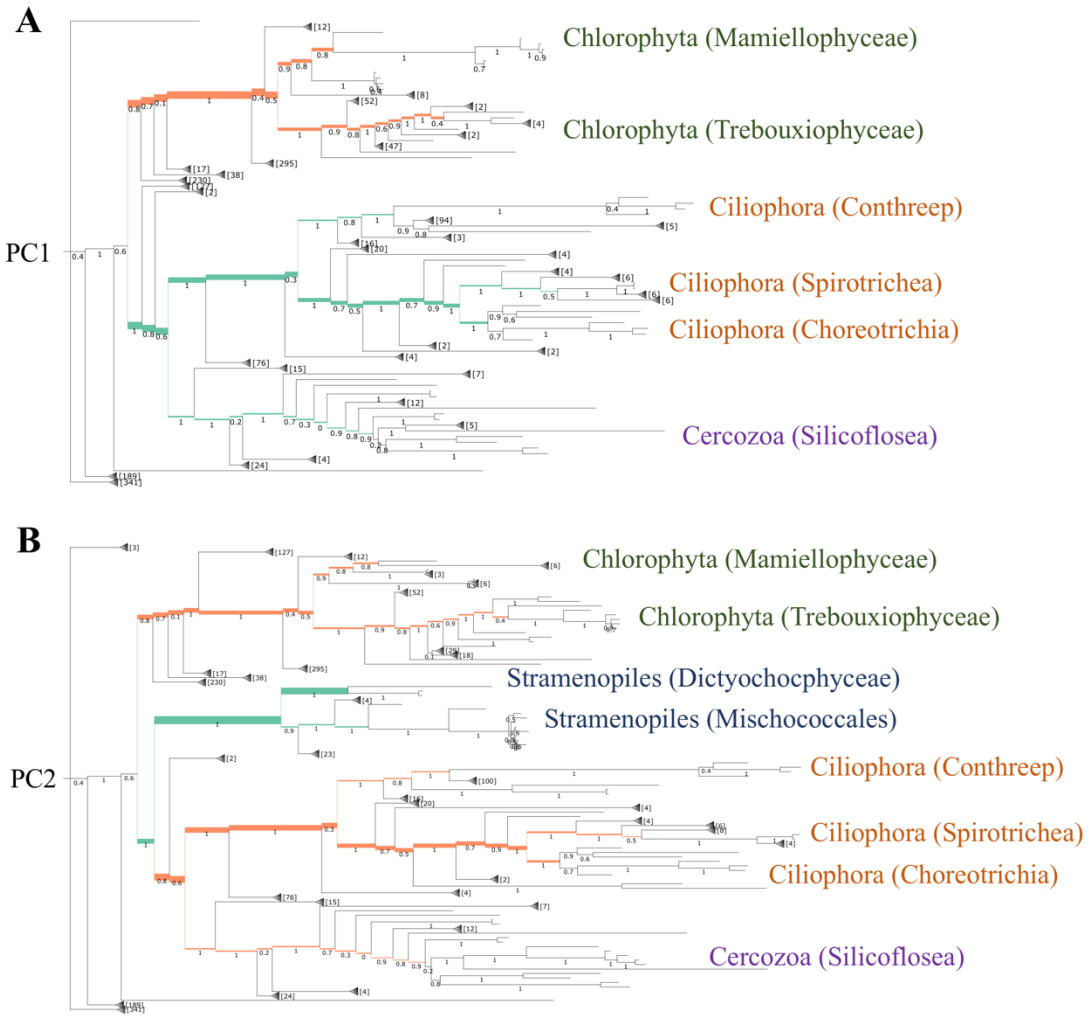


Figure 7. Archaeopteryx Fat Trees showing important lineages involved in explaining the patterns and variation in the dataset of less-represented amplicon sequences through principle components. Taxa influencing principle components of each sample are labelled and colored by lineage: Alveolata are colored orange, Rhizaria are purple, Stramenopiles are blue, and Viridiplantae are green.
A. PC1 explains 38.24% of the variability. Lineages with orange branches are responsible for pushing sequences to the left, and lineages with teal branches are responsible for pushing sequences to the right. **B**. PC2 0.23.85%. Lineages with orange branches are responsible for pushing sequences down, and lineages with teal branches are responsible for pushing sequences to up.

Diversity Patterns of Estuarine Microbial Eukaryotes

We explored patterns in alpha and beta diversity using phylogenetic-based diversity measures to complement our taxonomic characterization of the amplicon sequences. We calculated the balance-weighted phylogenetic diversity (BWPD) in each sample, which is a weighted phylogenetic representation of alpha diversity (McCoy & Matsen 2013). Alpha diversity (BWPD) varied both spatially and temporally in Bellingham Bay. Week-to-week, the BWPD of microbial eukaryotes (HRAS and LRAS combined) followed an oscillating pattern at both sampled depths (Figure 8). The deep samples had slightly higher alpha diversity (BWPD) overall, but this trend was not consistent in every sample. When we removed the highly-represented amplicon sequences from the analysis and focused just on the LRAS, we saw that the BWPD significantly negatively correlated to the BWPD of all microbial eukaryote sequences (p-value = <0.0001, rho = -0.47) (Figure 8). In both the CM and deep samples, when the alpha diversity (BWPD) within the entire sequence library was low (likely due to certain taxa representing a large proportion of the community), the BWPD within just the LRAS was higher (Figure 8). In turn, when total community BWPD was high, the diversity of the LRAS was low. The highly-represented amplicon sequences, specifically the dinoflagellate sequences, appeared to drive the trends we observed in total community alpha diversity (BWPD), and were responsible for reducing the overall BWPD. However, BWPD was also affected by the dynamics of less abundant lineages. In the CM, the diversity of the LRAS was consistently higher and generally less variable than when diatoms and dinoflagellates were included, with the exception of 2016.

Pairwise comparisons of both depths each year were completed by calculating the weighted Unifrac beta diversity (Kantorovich–Rubinstein metric) and supported the

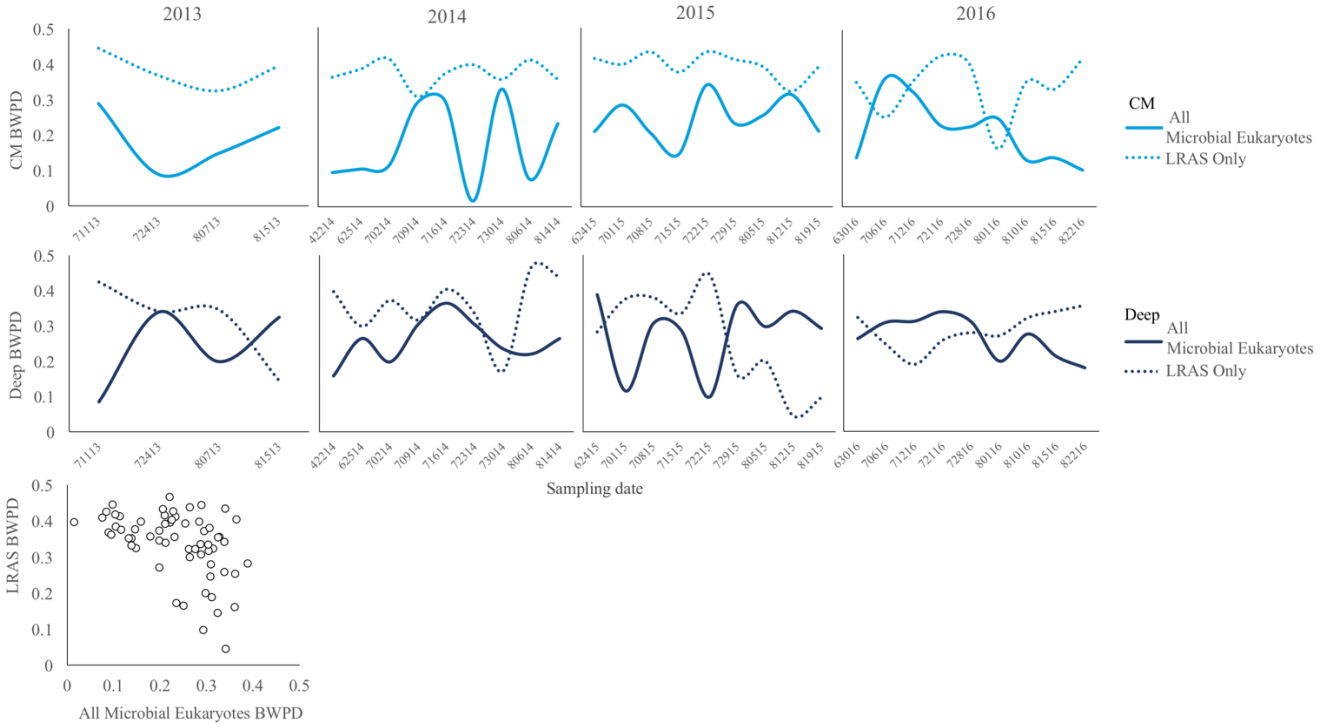


Figure 8. Diversity within samples (alpha diversity) measured in terms of balance-weighted phylogenetic diversity (BWPD), during each date of sampling from 2014-2016. (**Top**) BWPD for each sample (by date) each year in the chlorophyll maximum (CM). All Y axes are the same. Light blue solid lines represent values calculated from all microbial eukaryote sequences in sample. Light blue dotted lines represent values calculated from just the less-represented amplicon sequences (LRAS) (diatoms and dinoflagellates excluded). (**Middle**) BWPD for each sample (by date) from each year collected 1m above bay floor (Deep). All Y axes are the same. Dark blue solid lines represent values calculated from all microbial eukaryote sequences in sample. Dark blue dotted lines represent values calculated from just the LRAS (diatoms and dinoflagellates excluded). (**Bottom**) Significant negatively-correlated relationship between LRAS BWPD and all microbial eukaryotes BPWD. Spearman's rank correlation p-value = 0.000137, rho -0.4703734

taxonomic analysis suggesting samples were similar overall but not identical (Figure 9). There was higher variability across the CM samples, while the deep samples were fairly similar to each other (Figure 9). The 2014 CM samples seemed the most disparate from the other years at both depths. Based on the taxonomy, one of the leading contributors to difference found in 2014 may be due to exceptionally high presence of sequences annotated as Peridiniales in the CM during 2014 (Figure 3). However, even these samples still had 55% or more shared sequences with the other 3 years.

Resolving Patterns in a Dynamic Environment

Because coastal estuarine systems are highly dynamic and microbial eukaryotes are often directly impacted by environmental changes, characterization of these communities may require a more significant sampling effort than a closed, stable system. To test this, we examined how sampling timescales affected our estimate of stability over time in the relative proportions of sequence annotations and diversity metrics, which we used as a proxy to measure stability of the microbial eukaryote communities. We compared the general patterns observed in Bellingham Bay based on three different timescales of sampling: yearly, weekly, and daily.

With samples binned into years, the community looked fairly stable on yearly timescales (Figure 2). These communities were dominated by seven major lineages found in similar relative proportions each year, with the exception of Viridiplantae (Figure 2). Overall, the Alveolata (which contain the dinoflagellate lineage) and Stramenopiles (which contain the diatoms) were the most common (Figure 2). Within each lineage, we observed variation in the relative proportions of uniquely classified sequences, but overall, the yearly sampling

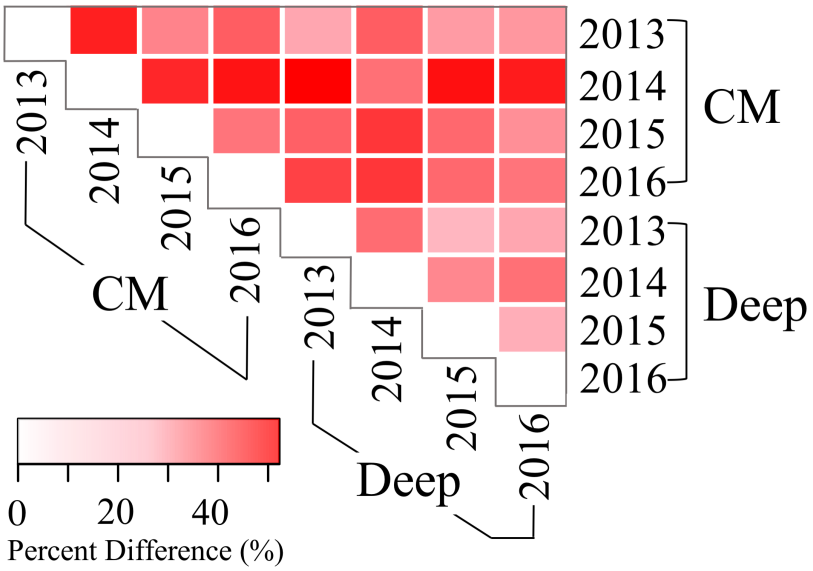


Figure 9. Heatmap showing differences between and within each of the two depths sampled over 4 years (2014-2016) of sampling in Bellingham Bay, WA. The depths compared were chlorophyll maximum (CM) and deep. Differences were measured pairwise using the weighted Unifrac metric, to calculate phylogenetic beta diversity. Lighter colors indicate a lower weighted Unifrac value (more similarity between samples), and darker colors indicate higher weighted Unifrac values (less similarity).

suggested a resistant community at the division level of classification. When looking within each depth on a yearly timescale, the deep samples seemed to follow a bi-annual pattern not observed in the CM samples (Figure 10a-b). The deep samples also had consistently higher alpha diversity (BWPD) than the CM samples. The lower BWPD values calculated from the CM samples seemed dependent on the proportions of dinoflagellate sequences, and had slightly more variability between years (Figure 10a-b).

When we increased the resolution by looking at samples on a weekly scale, our perception of community stability changed. We still saw the dominance of Alveolata and Stramenopiles at both depths, due to the dinoflagellate and diatom (HRA) sequences, but relative to the yearly timescale there was more variation in relative abundance within major lineages (Figure 10a-b). We also saw that sequences classified within Viridiplantae were proportionally more abundant during the first half of the sampling season than the latter, and were consistently sparse in the deep samples (Figure 10a-b). There was also more variation among the calculated alpha diversity (BWPD), especially in the CM, which again seemed closely tied to sequences placed within the Alveolata lineage and followed an oscillatory pattern (Figure 10a-b). When the diatom and dinoflagellate sequences were filtered out of the analysis to better analyze changes occurring week-to-week within the LRAS, we observed even greater variation, and saw changes in relative proportions both within and between lineages (Figure 11).

For five days in 2016, we sampled daily, which was our most intensive sampling timescale. During these five days, BWPD stayed relatively constant (Figure 10a-b), but there was subtle variation in the proportions of taxonomically identified sequences (Figure 10a-b, Figure 11, and Figure 12). The first two days and the latter three days of the intensive-week

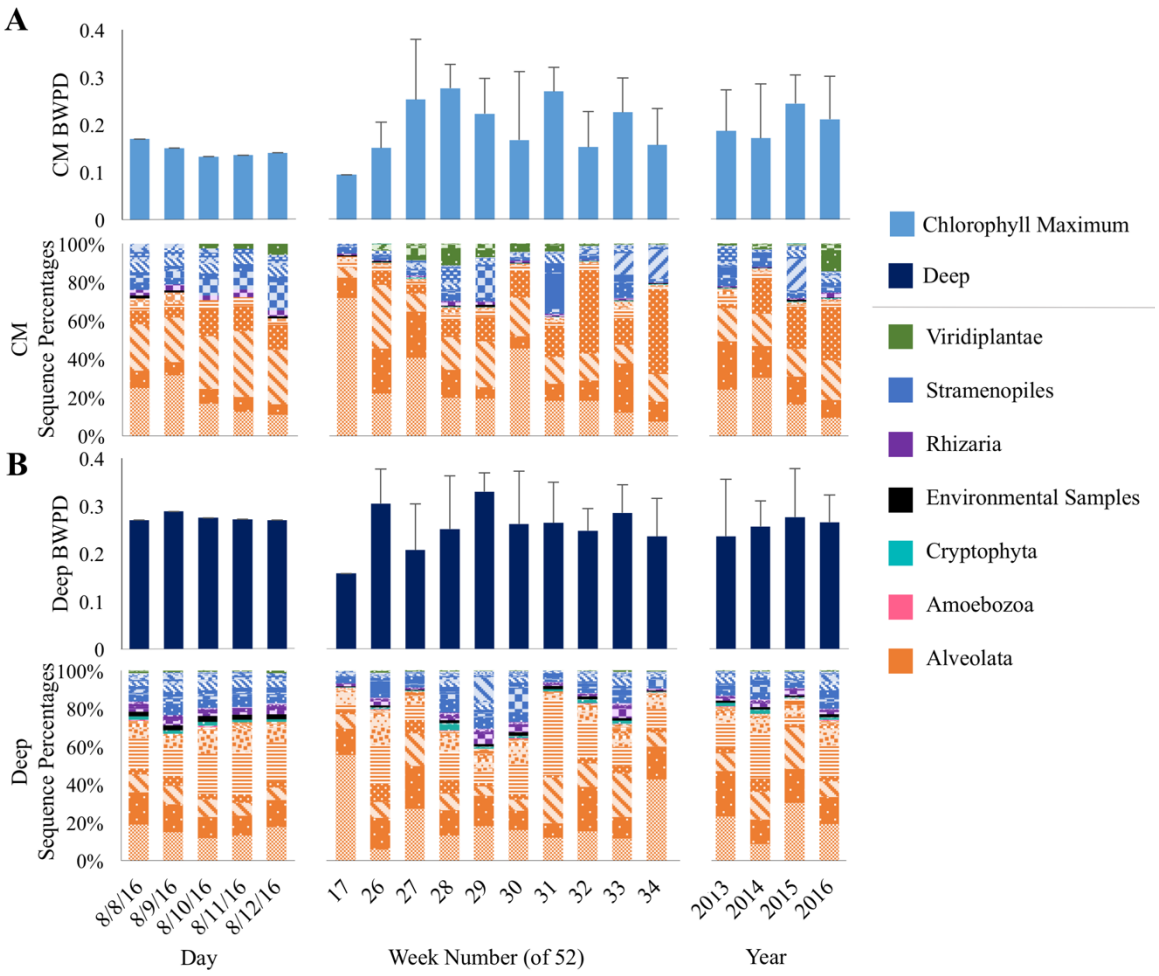


Figure 10. Bar graphs showing proportions (as a percent) of community composition and diversity in Bellingham Bay, WA at three increasing timescales over a four-year period (2014-2016) based on microbial eukaryote SSU rRNA gene sequences. The first timescale (left) is daily, spanning 5 days in 2016, the second (middle) is weekly, with week number (out of 52) and values derived from averaging the matching days across the sampling season from each year. The longest timescale (right) is yearly, with sequences binned into the years in which they were collected. **A.** (**Top**) Bar graph with standard deviation showing the average Balance Weighted Phylogenetic Diversity (BWPD) of successfully placed SSU rRNA sequences from microbial eukaryotes in the chlorophyll maximum (CM) during the three previously mentioned timescales of sampling in Bellingham Bay, WA. BWPD measures the diversity within samples (alpha diversity). (**Bottom**) Stacked bar graph showing the relative proportions (as a percent) of broad-level classification (color) and lowest-level classification (pattern) from annotations of microbial eukaryote SSU rRNA gene sequences from CM samples **B**. (**Top**) Bar graph with standard deviation showing the average BWPD of successfully placed SSU rRNA sequences from microbial eukaryotes 1m above the bottom of the bay (Deep) during the same three timescales previously mentioned. (**Bottom**) Stacked bar graph showing the relative proportions (as a percent) of broad-level classification (color) and lowest-level classification (pattern) from annotations of microbial eukaryote SSU rRNA gene sequences from deep samples.

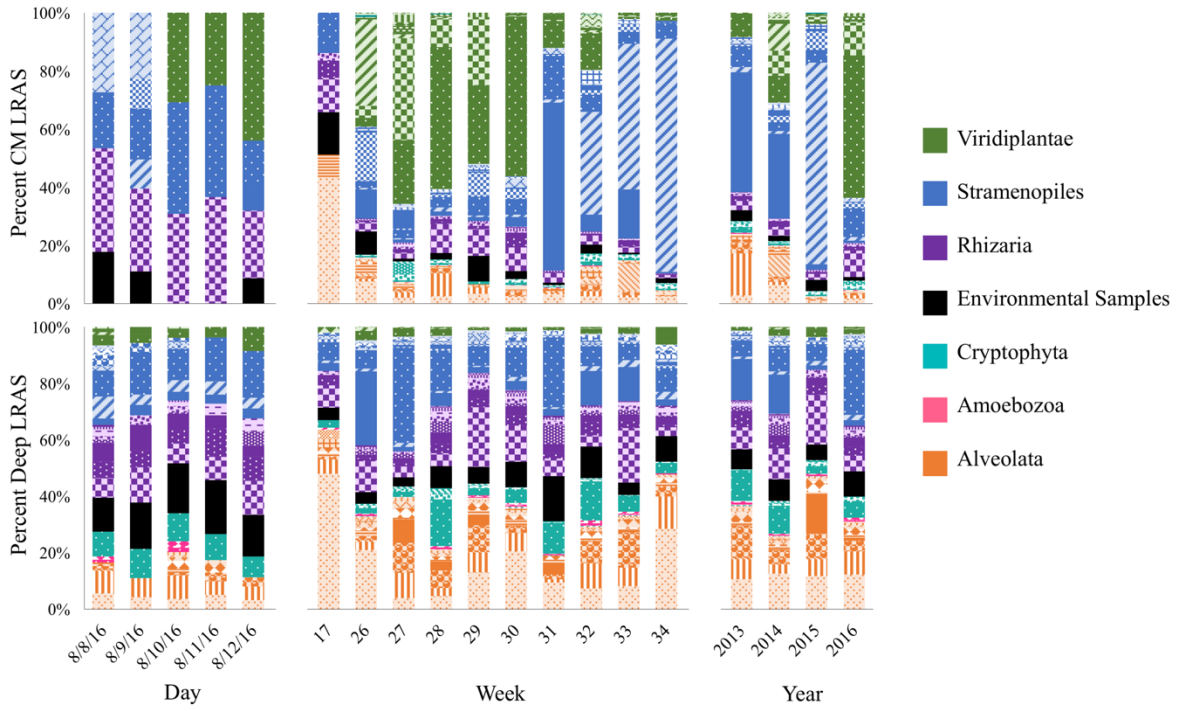


Figure 11. Bar graphs showing relative proportions (as a percent) of broad-level classification (color) and lowest-level classification (pattern) from annotations of less represented amplicon sequences (LRAS) of the SSU rRNA gene in Bellingham Bay, WA at three increasing timescales over a four-year period (2014-2016) based on microbial eukaryote SSU rRNA gene sequences. The first timescale (left) is daily, spanning 5 days in 2016, the second (middle) is weekly, with week number (out of 52) and values derived from averaging the matching days across the sampling season from each year. The longest timescale (right) is yearly, with sequences binned into the years in which they were collected. The top graph shows samples from the chlorophyll maximum (CM) and the bottom graph shows samples from 1m above the bottom of the bay (Deep). Diatoms and dinoflagellates are excluded.

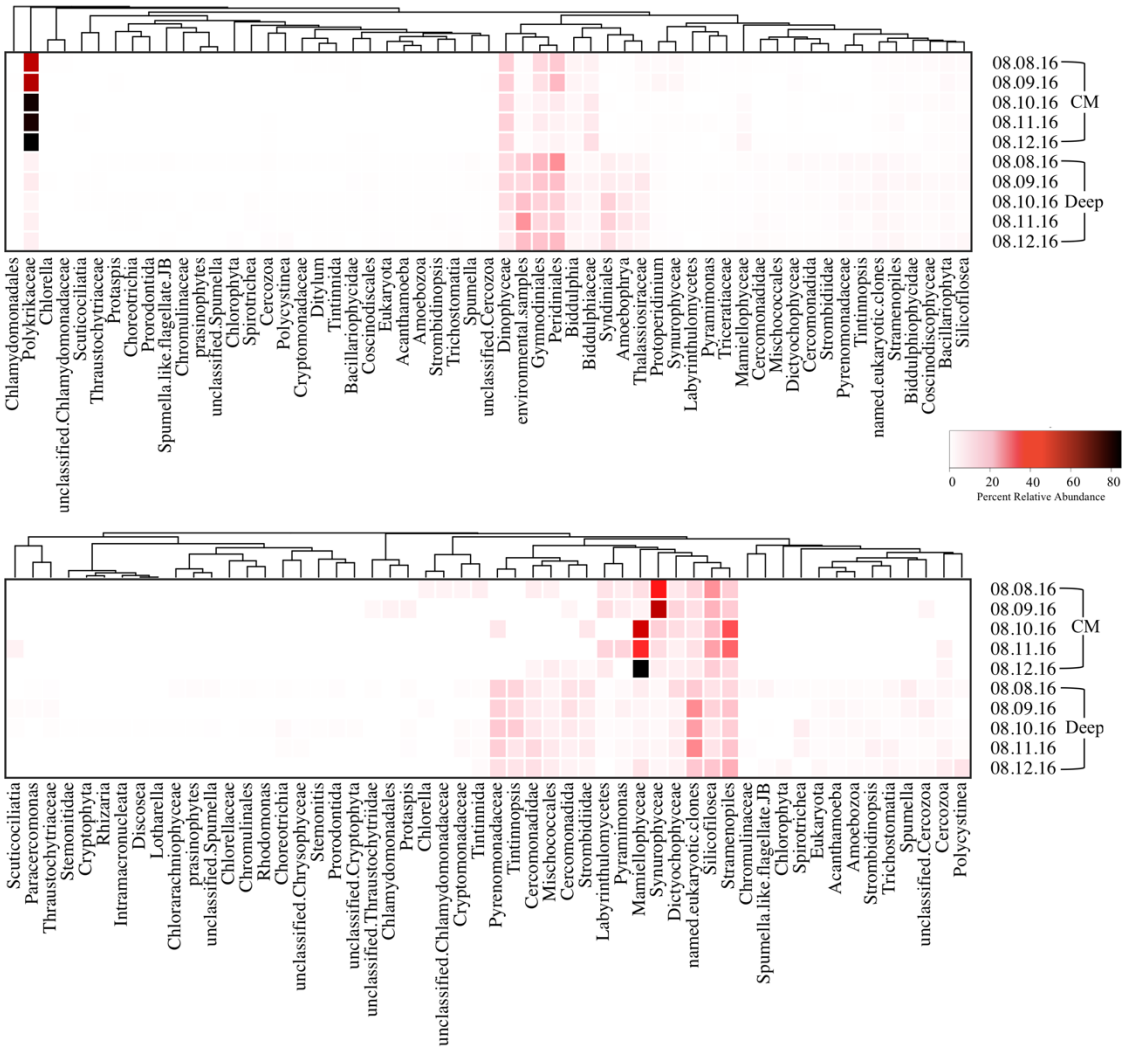


Figure 12. Heatmaps showing (**top**) the percent of different microbial eukaryote sequences found in Bellingham Bay samples taken between August 8-12, 2016 from the chlorophyll maximum (CM) and 1m above the floor of the bay (Deep). The sequences found in percentages lower than 0.1% were removed. (**Bottom**) Less-represented amplicon sequences (LRAS) (diatoms and dinoflagellates omitted) in Bellingham Bay samples between August 8-12, 2016 in the CM and Deep. Sequences found in percentages lower than 0.001% of the community were removed. Dendrograms were used in both heatmaps to cluster the sequences by similarities in abundance patterns. Scale bar shows color associated with relative abundance of sequences, with black indicating a maximum of 85% abundance in the total sequences for the associated time period and depth, and white accounting for 0% of the sequences.

samples had different dominant sequences in the CM samples (Figure 12). In addition, the sequences classified as organisms other than diatoms and dinoflagellates seemed to be fairly variable in the surface samples (Figure 12). Though the change in relative proportions was subtle day-to-day, it accumulated over the course of a week. Also, though the differences were minimal overall, the samples were not identical, and because the taxonomic classifications never reached the species or strain level, we can hypothesize that there was an additional level of variation within some groups that we did not capture, as most samples had multiple placements within internal nodes. Overall, our interpretation of Bellingham Bay's stability seemed to vary based on timescale. The Bay looked most variable on weekly timescales, and more stable on yearly and daily timescales, although never completely static (Figure 10a-b, Figure 11).

DISCUSSION

This study shows that distribution of microbial eukaryotes in a dynamic estuarine bay is not homogenous. There are highly-represented and more sparsely-represented amplicon sequences, and they show different spatial patterns of alpha diversity and temporal changes in community structure. The HRAS were all classified as taxa within two phyla (diatoms and dinoflagellates), while the sequences that were found in low abundance contained a diverse assemblage of lineages. Compared to dinoflagellates and diatoms, the less dominant and/or smaller microbial eukaryotes of estuarine environments have been studied to a lesser extent (Romari & Vaulot 2004; Bazin *et al.* 2013, 2014; Brannock *et al.* 2016; Marquardt *et al.* 2016), but these less-represented amplicon sequences (LRAS) included many important microbial eukaryote groups (Caron & Countway 2009; Logares *et al.* 2014; Ignatiades & Gotsis-Skretas 2014; Jousset *et al.* 2017). Both the HRAS and less common sequences showed patterns that tied to nutritional modes, with autotrophs found in higher abundance during the beginning of the season and in the CM, and heterotrophs common throughout the season, but at higher proportions in deep samples. We also observed undulating diversity over time, and found a negative correlation between diversity of the LRAS and HRAS. At times, sequences that were generally sparse were capable of substantially increasing in abundance, which have been previously observed (Caron & Countway 2009; Sjöstedt *et al.* 2012; Shade *et al.* 2014). As such, the heterogeneity seen in patterns between and within the HRAS and LRAS may indicate that the organisms represented by LRAS have a specific role in dynamic estuarine environments, making them critical to understanding community assembly and function (Allan *et al.* 2011; Jousset *et al.* 2017).

Though HTS of amplicons has proven to be a powerful technique for surveying microbial community diversity, tying patterns in the sequencing data back to the ecology of an estuarine system is challenging. This issue manifests itself when determining patterns between rare and abundant taxa. Though the use of HTS has recently become a popular method for assessing diversity and accessing the rare biosphere, disentanglement of the truly rare biosphere from smaller cells with fewer numbers of copies of their SSU rRNA gene that may appear rare in comparison with big cells with many copy numbers is not possible through sequencing amplicons alone. Also, our cell count data suggest that diatoms are more common than dinoflagellates in Bellingham Bay, a finding that is not reflected in our sequencing data, and relative abundances of some additional taxa contradict previous observation. For example, from other studies of this region, we know some ciliate taxa represented in our sequence libraries are found in much closer levels of abundance to Gymnodiniales (a class of dinoflagellates) than our sequencing data suggest (Paul 2010; Brown 2013). To make assertions about the ecology of a system, it is important to consider the biases introduced through HTS methods.

As HTS approaches to assess microbial eukaryote communities continue increasing in popularity, a number of biases inherent in the methodology are emerging. With an amplicon approach to sequencing (in lieu of shotgun sequencing), the use of polymerase chain reactions (PCR) are employed, and one region of the genomic DNA is amplified. We use this as a representation of diversity within each sample. However, it has been demonstrated that diversity measured from amplicon data is dependent on the primer (Dawson & Hagen 2009; Engelbrektson *et al.* 2010; Stoeck *et al.* 2010; Fredriksson *et al.* 2013; Tanabe *et al.* 2016). Not surprisingly, there are conflicting views on which region "best" captures the

diversity. Although the SSU rRNA gene remains the popular choice for amplification in diversity studies of both bacteria and eukaryotes (Campbell *et al.* 2011; Bik *et al.* 2012; Logares *et al.* 2013; de Vargas *et al.* 2015), the preferred regions for amplification within the small subunit vary, especially in eukaryotes (Stoeck *et al.* 2010; Hugerth *et al.* 2014; Tanabe *et al.* 2016). The most common regions used in HTS studies of eukaryotes have been the V4 and V9 regions (Pernice *et al.* 2013; de Vargas *et al.* 2015; Piredda *et al.* 2016; Tanabe *et al.* 2016). Previously, the V4 region was limited to use with 454 pyrosequencing because Illumina did not support the longer read lengths. Consequently, many published microeukaryote diversity studies using the Illumina platform, including the TARA project, have focused on the smaller (<200 bp) V9 region (Dawson & Hagen 2009; Amaral-Zettler *et al.* 2009; Brown *et al.* 2009; Stoeck *et al.* 2010; de Vargas *et al.* 2015). However, with the advent of Illumina kits able to generate longer reads, sequencing the V4 region in now possible.

Multiple comparative analyses have shown the V4 region is a strong choice for capturing genetic diversity (Stoeck *et al.* 2010; Nolte *et al.* 2010; Hugerth *et al.* 2014; Piredda *et al.* 2016; Tanabe *et al.* 2016). Recently, a paper comparing biases from the V4 and V8-V9 regions on a mock dataset also showed that the V4 region is better at representing sequences found in low abundance, which is critical in studies characterizing composition (Bradley *et al.* 2016). However, they also found that the two representative Haptophytes in their mock community, *Isochrysis galbana* and *Prymnesium parvum*, were significantly underrepresented in the V4 sequences. This was attributed to the high CG content common in Haptophytes in the V4 region of the SSU rRNA gene, which increased mismatches during PCR and resulted in shortened reads during sequencing (<100 bp). Using these primers, we

saw very low Haptophyte representation (6 sequences) in our dataset, though Haptophyta are known to inhabit nearby waters (Hopkinson *et al.* 2011). We also completed shotgun sequencing of a marine metagenome from Bellingham Bay using the same extraction technique as our amplicon sequences and identified Haptophyte sequences, validating their presence (data not shown). However, the primers used in our study and by Bradley et al. (2016) had previously been vetted and successfully amplified this lineage, with better coverage than the V9 (Stoeck *et al.* 2010). Despite these contradictory assertions, we believe problems arising during the PCR significantly reduced amplification of Haptophyte DNA and/or caused truncated reads during sequencing, which were filtered out during quality control steps (Bradley *et al.* 2016). Because this is an ecologically important group in marine systems, the new V8-V9 primers introduced in Bradley et al (2016) may be a better choice going forward, as these include the tail end of the V8 region to increase sequence length. Another option is to sequence multiple regions in the SSU rRNA gene to address the previously mentioned shortcomings in the more widely-used primers.

Another challenge with HTS studies is the bias created by sequencing effort. As is typical for amplicon studies, we multiplexed our samples, which increases sample quantity, but reduces the sequencing depth of each sample. We also had an issue with under-clustering, which reduced the number of total reads, though we were still able to achieve deeper sequencing than possible using 454 pyrosequencing (Smith & Peay 2014). While these factors probably do not affect our ability to resolve general trends in the community, they may affect our coverage of the rare taxa. A study of microbial eukaryotes in a French coastal estuary (Bay of Biscay) found rarefaction curves plateaued after 3000 reads (Abad *et al.* 2016). We averaged 43,727 reads in our sequence libraries, and rarefaction curves of the

smallest libraries (<4000 successfully placed reads) suggest we captured almost of the diversity, with the exception of one sample. However, the effect of sequencing depth on coverage and the possibility of missing taxa must be taken into account when making assertions about the Bellingham Bay, specifically within the rare community members (Supplementary Figure S2). In addition to the previously mentioned missing Haptophytes, there may be additional unresolved diversity.

The concept of transient taxa in dynamic environments has been introduced and identified in other studies characterizing communities with HTS (Nolte *et al.* 2010; Gobet *et al.* 2011; Shade *et al.* 2014). However, it has been suggested that deeper sampling effort reduces or diminishes the likelihood of finding sequences that are inconsistently present in surveys within a given system (Dolan & Stoeck 2011). Our comparison of sequencing depth highlights the importance of assuming not every "transient" member of our dataset is actually transient. Using a limited sampling depth may have caused minor overrepresentation of transient taxa in the Bellingham Bay dataset, but pooling samples for each year somewhat mitigated this issue by increasing the likelihood that the absent sequences truly represented absent taxa from the environment as a whole.

Methods to optimize analysis of eukaryotic communities, specifically, are still evolving, because the reference databases for eukaryotes are significantly smaller than those for bacteria (Dawson & Hagen 2009). Phylogenetic annotations for environmental sequences make the most of limited reference sequence data by allowing environmental sequences to be placed at internal nodes. Using this approach, sequences not found in the database can still be given context, as they are placed intermediate to or near known sequences. However, this approach may limit the number of low-level classifications within the dataset because the

reference tree to which we align our environmental sequences is a highly-reduced representation of the total microbial eukaryote database. Samples in our dataset frequently were placed on internal nodes within a general SSU ribosomal reference tree, giving many high taxonomic levels of annotation. Interpreting these placements as taxonomic names, instead of a place within a tree, homogenizes the sequence diversity and often generalizes placements at high taxonomic levels. However, there is sequence variability within lineages in environmental samples which represent strain and population level variability that can be highlighted with phylogenetic tree-based diversity measures such as BWPD, weighted-Unifrac, and ePCA . This highlights the importance of phylogeny-based and name-independent diversity measures, which offer valuable insights into otherwise-hidden variability, since they show relative relationships of the organisms and utilize computationally robust methods for microbial communities (Leinster & Cobbold 2012; Evans & Matsen 2012; McCoy & Matsen 2013).

The small subunit of ribosomal DNA has established itself as the preferred region of choice on the genome for microbial diversity exploration, however this region creates some inherent complications with eukaryotes, regardless of whether a taxonomy- or phylogeny-based analysis approach is used. Counts of particular taxa may be over-represented because there is high interspecies variability in the number of copies of the rRNA gene in eukaryotes based on cell size (Prokopowich *et al.* 2003; Godhe *et al.* 2008; Medinger *et al.* 2010). As mentioned in our results, we saw a higher proportion of dinoflagellate sequences in comparison to diatom sequences, a finding that contradicted preliminary cell count data (Supplementary Figure S2). The high representation of diatoms and dinoflagellates in our amplicon sequences is likely due their high abundance in the bay, but is exaggerated because

taxa in these groups generally have large cells, and there is a positive relationship between biomass and SSU rRNA gene copy number (Godhe *et al.* 2008). In addition to often having large cells, dinoflagellates have large, complex genomes with disproportionately high numbers of SSU rRNA gene copies per individual cell, which likely explains their dominance in the dataset (Prokopowich *et al.* 2003; Godhe *et al.* 2008; Abad *et al.* 2016). However, the contradictory cell count ratio may also be somewhat distorted in favor of diatoms. Because we counted cells in chains as individuals, each chain found during cell counts dramatically increased their total abundance. For example, chains of a specific genus may only show up sparingly within a sample, but because the proximity of the cells in the chain allows them to dominate the field-of-view, that genus will be increase its total count number at a faster rate than solitary cells, and falsely appear to dominate the total proportions for that sample. Disparities between the total counts of chain-forming taxa and solitary taxa often reached multiple orders of magnitude. In the cell counts, 3 genera of chain-forming diatoms accounted for 84% of the total cells counted, while the remaining taxa were often found in more similar relative abundances (Supplementary Table S4, Supplementary Table S5).

Though bias in the diatom counts may be exaggerating disparities between diatom:dinoflagellate ratios in the datasets, overrepresentation of the dinoflagellates due to genomic copy number of the ribosomal gene still needs to be addressed, as some dinoflagellates may have hundreds of thousands of copies per cell (Prokopowich *et al.* 2003; Bik *et al.* 2012; Grossmann *et al.* 2016). Recently, similar HTS studies have found discrepancies involving inflated dinoflagellate ratios (Medinger *et al.* 2010; Grossmann *et al.* 2016).

A predictive phylogenetic-based correction exists for prokaryotic ribosomal sequences (Kembel *et al.* 2012), and there have been suggestions on how to approach the copy number issue in eukaryotes, including: normalization of sequence ratios based on the known SSU rRNA gene copy number for each taxon found in the sequencing data or adjustment of sequence counts based on the biomass of each type of cell identified. However, we do not know the exact copy number of SSU rRNA genes for many microbes, and the ratio of biomass to the SSU rRNA gene copy number is not consistent within or between all lineages (Godhe *et al.* 2008). Until this issue is resolved, HTS of amplicons will remain a semi-quantitative method (Amend *et al.* 2010; Abad *et al.* 2016). However, because our methodology was consistent while preparing the molecular samples, and because we used weighted diversity measures, analyzing changes across our samples was still appropriate despite current limitations.

Interpreting Abundantly-Represented Amplicon Sequences

Though various biases may be convoluting patterns between diatoms and dinoflagellates, we were able to observe patterns within these lineages. The diatom sequences included many cosmopolitan lineages present in similar proportions in the CM and deep samples. This is likely due to cells mixing by tidal flow or sinking as these cells are non-flagellated (Worden *et al.* 2015). Consequently, not all of the diatoms in the deep sequences may have been metabolically active (Campbell *et al.* 2011; Koid *et al.* 2012). Although not all of the sequences of these taxa may represent active community members, the sinking of inactive cells to depth is still important for biogeochemical cycling (Cloern 1996; Calbet & Landry 2004; Worden *et al.* 2015). The dinoflagellates, which also have

some members with large cells and are important in biogeochemical transportation (Cloern 1996; Calbet & Landry 2004; Worden *et al.* 2015), contained the same members at each depth. However, they showed slightly more vertical stratification in their relative abundance in the sequencing data, most notably in the uncharacterized "environmental" sequences, which were found in high proportions in the deep samples. Because dinoflagellates are flagellated, and thus more motile than diatoms, their variability between depths may be more related to niche partitioning (Mouritsen & Richardson 2003), and more of the sequences in the deep samples may represent active cells. While diatoms are predominantly photoautotrophic, dinoflagellates also contain many mixotrophic species and non-photosynthetic heterotrophs, and thus face different competition for resources and less dependence on staying in the photic zone (Sherr *et al.* 2007; Worden *et al.* 2015).

We observed a surprisingly large abundance of Polykrikaceae sequences in the dataset. These taxa were absent from preliminary cell counts, likely due to the smaller volume of water surveyed using microscopy-based identification. This lineage includes heterotrophs that graze on other dinoflagellates, and often found in low abundances (Reñé *et al.* 2015). Because they are large in size, their copy number is likely inflating their relative proportions in the bay.

Amoebophrya, a dinoflagellate found in variable proportions throughout the 4-year study, parasitizes a range of host dinoflagellates, including many responsible for harmful algal blooms (HABs) (Park *et al.* 2002) such as *Akashiwo sanguinea*, *Alexandrium fundyense*, and *Dinophysis norvegica*. All of these taxa have caused HABs in the Salish Sea (Trainer *et al.* 2013; Moore *et al.* 2015; Ikeda *et al.* 2016). The populations of the host and parasite are thought to be negatively correlated (Mazzillo *et al.* 2011). Although we did not

capture any strong inverse relationship between the relative abundance of lineages containing these harmful algae and *Amoebophrya* sequences, further investigation into the dynamics of these parasites within the bay could be valuable for the local bloom-monitoring efforts, since studies have shown these can directly control HAB-forming dinoflagellates in other bodies of water (Park *et al.* 2002; Chambouvet *et al.* 2008).

Interpreting Sparse Sequences

The low representation of numerous amplicon sequences in our dataset may be attributed to a multitude of causes, and ecologically derived or due to experimental design. Most likely, these sequences represent taxa that are rare in the ecosystem and/or taxa that are small in size and thus have fewer SSU rRNA gene copies per cell than microplankton. Regardless of their size or abundance, these taxa are still important to include when attempting to comprehensively describe the microbial eukaryote community as they may impact Bellingham Bay's ecology.

Similar to bacteria, microbial eukaryote community contribution in ecosystems may not necessarily be dependent on abundance (Lynch & Neufeld 2015). Rare taxa can be keystone species in microbial communities and have large effects on the ecosystem (Giovannoni & Stingl 2005). Additionally, rare taxa may aid in resilience of the ecosystem because they increase diversity and functional redundancy (Yachi & Loreau 1999; Sogin *et al.* 2006; Caron & Countway 2009; Sjöstedt *et al.* 2012; Lynch & Neufeld 2015). This may be especially true with rare microbial eukaryotes in dynamic environments, as they can have high turnover rates, and many if not all sexually reproduce, further increasing genetic

variation and diversity, although sexual reproduction is dependent on the mating types encountering each other (Dunthorn *et al.* 2014).

Ample diversity is critical after a disturbance event, which, in an estuary, may be manifested by the introduction of contaminants, a severe weather event, or changes in the environmental conditions that do not follow seasonal patterns (Yachi & Loreau 1999; Lynch & Neufeld 2015). For example, a local disturbance event occurred in 2015 when the west coast of North America experienced a warm water anomaly (WWA) (nicknamed "the blob") that increased sea surface temperatures (SST), resulting in increased vertical stratification, reduced nutrient flux, and overall reduced phytoplankton biomass (Cavole *et al.* 2016). It subsequently coincided with a shift in the microbial community, resulting in record-breaking and sustained HABs dominated by *Pseudo-nitzschia* (Cavole *et al.* 2016). Though "the blob" did not reach Bellingham Bay, events such as this may occur again as global temperatures rise, increasing SST. The LRAS sequenced in Bellingham Bay spread across numerous, deeply divergent lineages, indicating that sequences representing rare taxa may already be contributing to community stability in this dynamic environment, and could contribute substantially to the resiliency of this ecosystem if future environmental or human-induced disturbances occur (Caron & Countway 2009; Dawson & Hagen 2009).

Many of the LRAS in Bellingham Bay represent taxa that are likely offering similar, but not identical, ecosystem services as the diatoms and dinoflagellates in regards to food web dynamics, nutrient cycling, and physiological constraints (Dunthorn *et al.* 2014), which enables differentiation in temporal patterns. A similar study that examined microbial eukaryote diversity in an Austrian lake also found that highly-represented taxa and rarer taxa followed different temporal patterns (Nolte *et al.* 2010). The study revealed a stable,

endemic group of major taxa and rare taxa that were proportionally more variable and/or inconsistently present (Nolte *et al.* 2010). Similarly, in Bellingham Bay, the composition and relative abundances of LRAS were more variable over time than the HRAS, and more closely followed patterns based on their roles in the food web, with heterotrophs found in higher abundances in deep samples, and photosynthesizers found in higher and more variable abundances in the CM samples.

A number of studies suggest diverse, rare microbial prokaryotes may act as seed banks throughout ecosystems, and this is now hypothesized as a mechanism operating in microbial eukaryote ecosystems as well (Nolte *et al.* 2010; Caron *et al.* 2012; Dunthorn *et al.* 2014). The community data we present here included a few sequences that may indicate the presence of transient taxa, similar to other studies of rare microbes (Nolte *et al.* 2010; Shade *et al.* 2014; Alonso-Sáez *et al.* 2015). These transient taxa may highlight dispersal potential within this open system and consequently contribute to the diversity of the rare biosphere (Nolte *et al.* 2010; Caron *et al.* 2012; Dunthorn *et al.* 2014; Logares *et al.* 2014). A few of the extremely rare and/or inconsistently present sequences belong to taxa common in coastal ocean or freshwater environments. For example, *Pyramimonas* (a green picoplankton) and *Rhodamonas* (a Cryptophyte) were sparse in our samples, but abundant in a study of small photosynthetic eukaryotes in coastal waters of the western North Pacific Ocean (Kataoka *et al.* 2016). We also observed a low, inconsistent presence of *Spumella* and *Spumella*-like flagellate sequences, which are both Chrysophytes (within Stramenopiles) found mainly in freshwater environments, including rivers (Nolte *et al.* 2010; Grossmann *et al.* 2015). These have been observed in other estuarine environments as well (Bazin *et al.* 2014). The examples highlighted here and other rare and/or inconsistent sequences in our dataset may

offer evidence of water mixing from the Strait of Georgia, as well as the mouth of the Nooksack River at the north end of the Bay. However, deciphering whether the appearance of transient taxa is due to environmental conditions and dispersal events or to experimental protocol remains a challenge (Caporaso *et al.* 2012). As such, we cannot unequivocally confirm that all of the inconsistently present taxa in our samples are truly transient, though some likely are.

As previously mentioned, the LRAS also include sequences that may not actually be rare, but have low representation due to small cell size, and consequently fewer copy numbers of the SSU rRNA gene than other taxa in our samples. The LRAS in the CM were predominantly identified as lineages dominated by pico- and nano- autotrophs. These cells may be important in surface food webs, as high surface:volume ratio in combination with a lack of mineralized tissues increases buoyancy, and many are also flagellated, and thus motile (Worden *et al.* 2015; Kataoka *et al.* 2016). The LRAS showed higher temporal variation in the CM than in the deep samples. We observed a higher abundance of Viridiplantae (green algae) in the first half of each season, and observed a small-scale bloom of Dictyochophyceae in 2015, which was otherwise only found in low proportions. Blooms of these taxa may increase competition for nutrients with the diatoms and dinoflagellates. The deep community seemed comparatively more stable than the CM over time based on taxonomy, but encompassed greater alpha diversity (BWPD) within samples. Deep samples were dominated by mixotrophs and heterotrophs, including many Ciliophora (within Alveolates), which were important taxa in determining the patterns in our ePCA of the LRAS. Heterotrophic taxa all contribute to nitrogen recycling as well as food web dynamics,

competing with consumers across multiple trophic levels such as bacteria and zooplankton (Massana *et al.* 2002; Vigil *et al.* 2009; Grossmann *et al.* 2015).

We cannot unequivocally differentiate between small and rare taxa in this study, because we did not size-fractionate cells. Even so, whether the sequences represent small, rare, or small and rare taxa, they are all likely contributing substantially to the alpha diversity (BWPD) in each sample. Recently the TARA project, which constitutes one of the most extensive microbial eukaryote surveys to date using HTS, measured diversity in different size fractions of cells and found that ribosomal diversity increased with decreasing cell size, and as stated earlier, rare taxa substantially increase diversity as well (Debroas *et al.* 2015). The higher diversity measured in our samples once HRAS were removed indicates that this was true in our study as well. The BWPD values increased more in the CM than the deep when dinoflagellate and diatom sequences were removed, suggesting the HRAS were more similar or more abundant in the CM. The LRAS in the CM also likely determined the higher weighted-Unifrac values as they were more variable between weeks, unlike the deep LRAS. This higher variability may also explain the larger spread of the CM samples in our ePCA of depth. Dictyochophyceae sequences, which increased in abundance later in the sampling season were one of the important drivers of samples to higher PC2 values. Notably, there were no samples from the early sampling weeks in this region of PC2, meaning seasonal succession was a component in the spread.

The patterns in temporal community composition also suggest an interaction between the taxa represented in the LRAS and HRAS. When we examined the alpha diversity (BWPD) of all the microbial eukaryotes in our samples and the BWPD of just the LRAS, we saw a significant negative correlation. The BWPD measurements follow an inverse

oscillating pattern. Though this clearly indicates that the dinoflagellate and diatom sequences heavily influence the diversity, it is difficult to definitively tie this inverse relationship to the ecology of the system. However, a relationship between diatoms and dinoflagellates (which we know are abundant in Bellingham Bay) and the rest of the taxa in the system is likely, since they are in close proximity and subject to the same environmental conditions (Roy & Chattopadhyay 2007). The data may indicate that homogeneity in the dinoflagellates and/or diatoms coincides with diversification of taxa represented by the LRAS, that can take advantage of unused resources (Caron & Countway 2009), or that conditions favoring diatoms and dinoflagellates are less favorable for small flagellates. BWPD measurements in the CM decreased substantially when diatoms and dinoflagellates were introduced in the measurement. This may be indicative of blooms within these abundant lineages, which we saw intermittently appear in the cell count data. When blooms form and subsequently sink, the cells provide food for bacteria and promote growth in these populations, which are then eaten by ciliates and other heterotrophic eukaryotes (Amin *et al.* 2012). However, growth-periods should cause a lagged correlation, which we did not capture on weekly timescales. To explore this interaction further, more frequent sampling may be necessary, as well as the addition of functional gene analysis, so diversity patterns in these groups can be tied back to function.

Evaluation of Patterns Based on Timescale

Assessing marine microbial eukaryote diversity and community structure is a challenge, in part because the communities change very quickly in comparison with terrestrial communities. This is especially evident in open and dynamic ecosystems, such as

estuaries, as these communities are sensitive to constantly changing local environmental factors such as tides, river discharge, and oceanic circulation, and thus have high turnover rates (Boyle & Silke 2010; Cloern & Jassby 2010). When elucidating diversity in dynamic environments, single snapshots may cause vast underestimation of the community diversity, and bias our understanding of the community structure (Vigil *et al.* 2009; Nolte *et al.* 2010). In our dataset, the diversity varied weekly, as community composition shifted throughout sampling seasons.

Repeated sampling is a necessary step forward, but knowing at what intervals to sample is another challenge (Dornelas *et al.* 2012). Multiple studies have attempted to elucidate temporal patterns using molecular approaches in microbial eukaryotes at various timescales with mixed, and sometimes contrasting, results. A study of the western North Pacific took samples at 4 different locations at 5 3-month intervals, and found samples were similar between sites but significantly different by season (Kataoka *et al.* 2016). The previously mentioned study of an Austrian lake, a closed system, collected 10 samples from the same location at 3-week intervals and found fluctuating relative proportions of a stable set of taxa among samples, as well as taxa that would disappear and reappear throughout the sampling season (Nolte *et al.* 2010). This shows that even in closed, dynamic systems, high variability has been observed (Nolte *et al.* 2010). Weekly sampling of a fjord in Norway revealed fairly stable temporal community structure, with more variability seen during seasonal shifts (Marquardt *et al.* 2016). A study that used terminal restriction fragment length polymorphism (T-RFLP) to sample a mid-Atlantic estuarine environment at 1-3 week intervals for 2 consecutive summers found that dominant taxa changed frequently and significantly between sampling, but that samples taken 1 year apart at the same location

showed high similarity (Vigil *et al.* 2009). In contrast, a study implementing cloning to complete seasonal sampling in an estuary within the English Channel for 1 year, followed by monthly sampling the next year found some seasonal patterns, but marked differences between samples taken 1 year apart (Romari & Vaulot 2004). Though we did not find any studies measuring daily variability using molecular methods, cell counts of phytoplankton in a Lebanese harbor showed substantial variation day-to-day (Saab 1992). The variable patterns of these past studies in estuaries and other systems led us to explore three different time intervals during our sampling period.

In order to determine how sampling timescale and effort affected pattern resolution and our perception of community stability in dynamic environments, we designed our study to include daily, weekly, and yearly intervals. The most consistent observation was that Bellingham Bay shows annually resetting cycles, similar to the mid-Atlantic estuary study by Vigil et al. (2009). We saw this both in the ePCAs, which lacked any clustering by year, and our taxonomy-based analyses, which showed high similarity when samples were pooled by year. As such, we found it appropriate and beneficial to pool data by week, as this timescale captured more taxonomic diversity and reduced the risk of single-sample variability. This seemed especially important when assessing the presence of transient taxa. Samples pooled into weeks showed strong seasonal succession in Bellingham Bay, but this was limited to the CM. The high similarity in the daily samples suggest more stability than previous microscopy-based analyses have suggested (Saab 1992), likely because we can characterize higher volumes of water using HTS. Because daily samples are highly similar, but weekly samples seem quite variable, sampling twice a week may be the best way to effectively capture transitions in the community, while keeping sampling efforts and costs reasonable.

However, specific questions will dictate sampling needs. Repeated sampling at small timescales allows for analytical flexibility.

Study Insights

There are numerous opportunities for more directed research questions in Bellingham Bay and other estuarine systems. This study generated much-needed initial survey data on microbial communities in Bellingham Bay, which can be utilized to design improved, highly-directed studies. Our ability to make specific predictions about Bellingham Bay in this study was limited by the often broad-scale classifications, a product of the state of bioinformatic limitations at the time of this study. However, we can improve on the resolution of the data presented here, by using the preliminary results to make additional taxon-specific reference packages to increase resolution of specific groups. This will allow a more detailed taxon-specific analysis of existing data.

Another next step could build off the observation of the taxon-specific temporal patterns to do population-level studies. An appropriate focus would be the green algae, which showed temporal patterns indicative of possible blooms, and/or the parasitic dinoflagellates that may be involved in bloom control of blooming autotrophic dinoflagellates. Another fruitful next step, given we now know that major and rare taxa within estuarine environments show spatial and temporal patterns in their community structure and distribution, is to tie these changes with environmental data. We found no strong relationships between environmental variables and phytoplankton cell counts or amplicons in this study (data not shown), so we know that we did not capture the tie between physical and biological factors in

our study design. Additional sampling designs are necessary, specifically utilizing the new buoy in Bellingham Bay, *Se'lhaem*, which takes continuous environmental data.

Understanding and ultimately predicting nearshore microbial eukaryote community dynamics has important ecological and economic implications and may help govern decisions in environmental policy (Hallegraeff 2010; Lallias *et al.* 2014; Piredda *et al.* 2016). These nearshore communities often contain bloom-forming species, some of which create harmful toxins. Historically, it has been difficult to predict exactly when blooms will occur, and more so which species will bloom, and this may continue to be the case as climate change continues to affect the chemistry and physics of our oceans (Hallegraeff 2010). However, Better resolution may also increase our ability to model community structure over time and predict HABs (Dornelas *et al.* 2012; Giovannoni & Vergin 2012; Piredda *et al.* 2016).

CONCLUSION

Microbial eukaryotic communities in estuarine systems are dynamic and variable, and Bellingham Bay is no exception. Although the community was moderately stable on a yearly timescale, there was weekly variability, even on a class or higher level. We now have a multi-year record of spatial and temporal community dynamics within the microbial eukaryotes of Bellingham Bay, a representative coastal estuarine environment. Through employing a HTS approach, we found that Bellingham Bay, which is dominated by diatoms and dinoflagellates, also includes a diverse assemblage of other taxa, which may have an important function in this ecosystem. Changes in phytoplankton communities can be difficult to predict. However, by adding a more comprehensive technique into our analyses of these communities, we may start to better understand nearshore microbial community dynamics. In this study specifically, we were able to gain data about the entire community structure and how it shifted throughout the summer seasons, using an all-encompassing direct approach. As we overcome the limitations of HTS, both in sample preparation and sequencing analysis, and as costs continue to lower, we hope this method will continue to gain popularity for estuarine microbial eukaryote studies, specifically in the context of time-series analysis and exploration within the rare and small taxa. With a directed approach, future studies would likely benefit from applying HTS of amplicons as a way to monitor microbial communities' responses to disturbance events, natural environmental fluctuations, and the changing climate.

LITERATURE CITED

Abad D, Albaina A, Aguirre M *et al.* (2016) Is metabarcoding suitable for estuarine plankton monitoring? A comparative study with microscopy. *Marine Biology*, **163**, 1–13.

Allan E, Weisser W, Weigelt A *et al.* (2011) More diverse plant communities have higher functioning over time due to turnover in complementary dominant species. *Proceedings of the National Academy of Sciences*, **108**, 17034–17039.

Alonso-Sáez L, Díaz-Pérez L, Morán XAG (2015) The hidden seasonality of the rare biosphere in coastal marine bacterioplankton. *Environmental microbiology*, **17**, 3766–3780.

Amaral-Zettler LA, McCliment EA, Ducklow HW, Huse SM (2009) A method for studying protistan diversity using massively parallel sequencing of V9 hypervariable regions of small-subunit ribosomal RNA genes. *PloS one*, **4**, e6372.

Amend AS, Seifert KA, Bruns TD (2010) Quantifying microbial communities with 454 pyrosequencing : does read abundance count ? *Molecular Ecology*, **19**, 5555–5565.

Amin SA, Parker MS, Armbrust EV (2012) Interactions between diatoms and bacteria. *Microbiology and molecular biology reviews : MMBR*, **76**, 667–84.

Andrews S (2010) FastQC A Quality Control tool for High Throughput Sequence Data. *http://www.bioinformatics.babraham.ac.uk/projects/fastqc/*.

Arrigo KR (2005) Marine microorganisms and global nutrient cycles. *Nature*, **437**, 349–355.

Bazin P, Jouenne F, Deton-Cabanillas A-F, Pérez-Ruzafa Á, Véron B (2013) Complex patterns in phytoplankton and microeukaryote diversity along the estuarine continuum. *Hydrobiologia*, **726**, 155–178.

Bazin P, Jouenne F, Friedl T *et al.* (2014) Phytoplankton diversity and community

composition along the estuarine gradient of a temperate macrotidal ecosystem: combined morphological and molecular approaches. *PloS one*, **9**, e94110.

Bik HM, Porazinska DL, Creer S *et al.* (2012) Sequencing our way towards understanding global eukaryotic biodiversity. *Trends in Ecology & Evolution*, **27**, 233–243.

Boyle SO, Silke J (2010) A review of phytoplankton ecology in estuarine and coastal waters around Ireland. *Journal of Plankton Research*, **32**, 99–118.

Bradley IM, Pinto AJ, Guest JS (2016) Gene-Specific Primers for Improved Characterization of Mixed Phototrophic Communities. *Applied and Environmantal Microbiology*, **82**, 5878–5891.

Brannock PM, Ortmann AC, Moss AG, Halanych KM (2016) Metabarcoding reveals environmental factors influencing spatio-temporal variation in pelagic micro-eukaryotes. *Molecular Ecology*, **25**, 3593–3604.

Brown KL (2013) Synechococcus distribution and abundance in the San Juan Archipelago , Salish Sea By. Western Washington University.

Brown M V, Philip GK, Bunge JA *et al.* (2009) Microbial community structure in the North Pacific ocean. *The ISME journal*, **3**, 1374–86.

Brown SP, Veach AM, Rigdon-Huss AR *et al.* (2015) Scraping the bottom of the barrel: Are rare high throughput sequences artifacts? *Fungal Ecology*, **13**, 221–225.

Calbet A, Landry MR (2004) Calbet, Albert, and Michael R. Landry. Phytoplankton growth, microzooplankton grazing, and carbon cycling in marine systems. Limnol. Oceanogr., 49(1), 2004, 51–57. *Limnol. Oceanogr*, **49**, 51–57.

Campbell BJ, Yu L, Heidelberg JF, Kirchman DL (2011) Activity of abundant and rare bacteria in a coastal ocean. *Proceedings of the National Academy of Sciences of the*

United States of America, **108**, 12776–12781.

Caporaso JG, Paszkiewicz K, Field D, Knight R, Gilbert JA (2012) The Western English Channel contains a persistent microbial seed bank. *The ISME journal*, **6**, 1089–93.

Caron DA, Countway PD (2009) Hypotheses on the role of the protistan rare biosphere in a changing world. *Aquatic Microbial Ecology*, **57**, 227–238.

Caron DA, Countway PD, Jones AC, Kim DY, Schnetzer A (2012) Marine Protistan Diversity. *Annu. Rev. Mar. Sci*, **4**, 467–93.

Cavole LM, Demko AM, Diner RE *et al.* (2016) Biological impacts of the 2013–2015 warm-water anomaly in the Northeast Pacifc: Winners, losers, and the future. *Oceanography*, **29**, 273–285.

Chambouvet A, Morin P, Marie D, Guillou L (2008) Control of Toxic Marine Dinoflagellate Blooms by Serial Parasitic Killers. *Science*, **322**.

Cloern JE (1996) Phytoplankton bloom dynamics in coastal ecosystems: A review with some general lessons from sustained investigation of San Francisco Bay, California. *Reviews of Geophysics*, **34**, 127–168.

Cloern JE, Jassby AD (2010) Patterns and scales of phytoplankton variability in estuarine-coastal ecosystems. *Estuaries and Coasts*, **33**, 230–241.

Cushing DH (1959) The Seasonal Variation in Oceanic Production as a Problem in Population Dynamics. *Journal du Conseil International pour l'Exploration de la Mer*, **45**, 455–464.

Darling AE, Jospin G, Lowe E *et al.* (2014) PhyloSift: phylogenetic analysis of genomes and metagenomes. *PeerJ*, **2**, e243.

Dawson SC, Hagen KD (2009) Mapping the protistan "rare biosphere". *Journal of biology*,

8, 105.

Debroas D, Hugoni M, Domaizon I (2015) Evidence for an active rare biosphere within freshwater protists community. *Molecular Ecology*, **24**, 1236–1247.

Dolan JR, Stoeck T (2011) Repeated sampling reveals differential variability in measures of species richness and community composition in planktonic protists. *Environmental Microbiology Reports*, **3**, 661–666.

Dornelas M, Magurran AE, Buckland ST *et al.* (2012) Quantifying temporal change in biodiversity: challenges and opportunities. *Proceedings. Biological sciences / The Royal Society*, **280**, 20121931.

Dunthorn M, Stoeck T, Clamp J, Warren A, Mahé F (2014) Ciliates and the rare biosphere: A review. *Journal of Eukaryotic Microbiology*, **61**, 404–409.

Engelbrektson A, Kunin V, Wrighton KC *et al.* (2010) Experimental factors affecting PCR-based estimates of microbial species richness and evenness. *The ISME journal*, **4**, 642–647.

Evans SN, Matsen FA (2012) The phylogenetic Kantorovich-Rubinstein metric for environmental sequence samples. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **74**, 569–592.

Falkowski PG, Katz ME, Knoll AH *et al.* (2004) The Evolution of Modern Eukaryotic Phytoplankton Published by : American Association for the Advancement of Science Stable URL : http://www.jstor.org/stable/3837504 . *Science*, **305**, 354–360.

Field CB, Behrenfeld MJ, Randerson JT, Falkowski P (1998) Primary Production of the Biosphere : Integrating Terrestrial and Oceanic Components. *Science*, **281**.

Frederiksen M, Edwards M, Anthony J, Halliday NC, Wanless S (2006) From plankton to

top predators : bottom-up control of a marine food web across four trophic levels. *Journal of Animal Ecology*, **75**, 1259–1268.

Fredriksson NJ, Hermansson M, Wilén B-M (2013) The choice of PCR primers has great impact on assessments of bacterial community diversity and dynamics in a wastewater treatment plant. *PloS one*, **8**, e76431.

Giovannoni SJ, Stingl U (2005) Molecular diversity and ecology of microbial plankton. *Nature*, **437**, 343–348.

Giovannoni SJ, Vergin KL (2012) Seasonality in Ocean Microbial Communities. *Science*, **335**, 671–676.

Gobet A, Böer SI, Huse SM *et al.* (2011) Diversity and dynamics of rare and of resident bacterial populations in coastal sands. *The ISME Journal*, **6**, 542–553.

Godhe A, Asplund ME, Härnström K *et al.* (2008) Quantification of diatom and dinoflagellate biomasses in coastal marine seawater samples by real-time PCR. *Applied and Environmental Microbiology*, **74**, 7174–7182.

Grossmann L, Bock C, Schweikert M, Boenigk J (2015) Small but Manifold - Hidden Diversity in "Spumella-like Flagellates." *The Journal of Eukaryotic Microbiology*, **63**, 419–39.

Grossmann L, Jensen M, Heider D *et al.* (2016) Protistan community analysis: key findings of a large-scale molecular sampling. *The ISME Journal*, 1–11.

Hallegraeff GM (2010) Ocean climate change, phytoplankton community responses, and harmful algal blooms: A formidable predictive challenge. *Journal of Phycology*, **46**, 220–235.

Han M V, Zmasek CM (2009) phyloXML : XML for evolutionary biology and comparative

genomics. *BMC Bioinformatics*, **6**, 1–6.

Herfort L, Peterson TD, Mccue LA, Zuber P (2011) Protist 18S rRNA gene sequence analysis reveals multiple sources of organic matter contributing to turbidity maxima of the Columbia River estuary. *Marine Ecology Progress Series*, **438**, 19–31.

Hopkinson BM, Xu Y, Shi D, Mcginn PJ, Morel MM (2011) The effect of CO 2 on the photosynthetic physiology of phytoplankton in the Gulf of Alaska. *Limnology and Oceanography*, **55**, 2011–2024.

Hugerth LW, Muller EEL, Hu YOO *et al.* (2014) Systematic Design of 18S rRNA Gene Primers for Determining Eukaryotic Diversity in Microbial Consortia. *PLoS ONE*, **9**, 1–11.

Ignatiades L, Gotsis-Skretas O (2014) The contribution of rare species to coastal phytoplankton assemblages. *Marine Ecology*, **35**, 132–145.

Ikeda CE, Cochlan WP, Bronicheski CM, Trainer VL, Trick CG (2016) The effects of salinity on the cellular permeability and cytotoxicity of *Heterosigma akashiwo* (J Raven, Ed,). *Journal of Phycology*, **52**, 745–760.

Jardillier L, Zubkov M V, Pearman J, Scanlan DJ (2010) Significant CO 2 fixation by small prymnesiophytes in the subtropical and tropical northeast Atlantic Ocean. *The ISME Journal*, **4**, 1180–1192.

Jousset A, Bienhold C, Chatzinotas A *et al.* (2017) Where less may be more: how the rare biosphere pulls ecosystems strings. *The ISME Journal*, 1–10.

Kataoka T, Yamaguchi H, Sato M *et al.* (2016) Seasonal and geographical distribution of near-surface small photosynthetic-eukaryotes in the western North Pacific determined by pyrosequencing of 18S rDNA. *FEMS Microbiology Ecology*, 1–16.

Keeling PJ, Burki F, Wilcox HM *et al.* (2014) The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing. *PLoS Biology*, **12**.

Kembel SW, Wu M, Eisen JA, Green JL (2012) Incorporating 16S Gene Copy Number Information Improves Estimates of Microbial Diversity and Abundance. *PLoS ONE*, **8**, 16–18.

Kim DY, Countway PD, Jones AC *et al.* (2014) Monthly to interannual variability of microbial eukaryote assemblages at four depths in the eastern North Pacific. *The ISME Journal*, **8**, 515–530.

Koid A, Nelson WC, Mraz A, Heidelberg KB (2012) Comparative analysis of eukaryotic marine microbial assemblages from 18S rRNA gene and gene transcript clone libraries by using different methods of extraction. *Applied and environmental microbiology*, **78**, 3958–65.

Lallias D, Hiddink JG, Fonseca VG *et al.* (2014) Environmental metabarcoding reveals heterogeneous drivers of microbial eukaryote diversity in contrasting estuarine ecosystems. *The ISME journal*.

Leinster T, Cobbold CA (2012) Measuring diversity: the importance of species similarity. *Ecology*, **93**, 477–489.

Logares R, Audic S, Bass D *et al.* (2014) Patterns of Rare and Abundant Marine Microbial Eukaryotes. *Current Biology*, **24**, 813–821.

Logares R, Sunagawa S, Salazar G *et al.* (2013) Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities. *Environmental microbiology*, **16**, 2659–71.

Lynch MDJ, Neufeld JD (2015) Ecology and exploration of the rare biosphere. *Nature reviews Microbiology*, **13**, 217–229.

Marquardt M, Vader A, Stübner EI, Reigstad M, Gabrielsen TM (2016) Strong seasonality of marine microbial eukaryotes in a high-Arctic fjord (Isfjorden, in West Spitsbergen, Norway). *Applied and Environmental Microbiology*, **82**, 1868–1880.

Massana R (2011) Eukaryotic picoplankton in surface oceans. *Annual Review of Microbiology*, **65**, 91–110.

Massana R, Guillou L, Díez B, Pedrós-Alió C (2002) Unveiling the Organisms behind Novel Eukaryotic Ribosomal DNA Sequences from the Ocean Unveiling the Organisms behind Novel Eukaryotic Ribosomal DNA Sequences from the Ocean. *Applied and Environmantal Microbiology*, **68**, 4554–4558.

Matsen IV FA, Evans SN, Jaccard P *et al.* (2013) Edge Principal Components and Squash Clustering: Using the Special Structure of Phylogenetic Placement Data for Sample Comparison (A Moustafa, Ed,). *PLoS ONE*, **8**, e56859.

Matsen FA, Kodner RB, Armbrust EV *et al.* (2010) pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics*, **11**, 538.

Mazzillo FFM, Ryan JP, Silver MW (2011) Parasitism as a biological control agent of dinoflagellate blooms in the California Current System. *Harmful Algae*, **10**, 763–773.

McCoy CO, Matsen FA (2013) Abundance-weighted phylogenetic diversity measures distinguish microbial community states and are robust to sampling depth. *PeerJ*, **1**, e157.

Mclaughlin RJ (2016) Microbial Diversity Across an Oxygen Gradient Using Large-scale

Phylogenetic-based Analysis of Marine Metagenomes. Weestern Washington University.

Medinger R, Nolte V, Pandey RV *et al.* (2010) Diversity in a hidden world: potential and limitation of next-generation sequencing for surveys of molecular diversity of eukaryotic microorganisms. *Molecular Ecology*, **19**, 32–40.

Michaels AFAMWS (1988) Primary production, sinking fluxes and the microbial food web. *Deep-Sea Research*, **35**, 473–490.

Moon-van der Staay SY, De Wachter R, Vaulot D (2001) Oceanic 18S rDNA sequences from picoplankton reveal unsuspected eukaryotic diversity. *Nature*, **409**, 607–610.

Moore SK, Johnstone JA, Banas NS, Salathé EP (2015) Present-day and future climate pathways affecting Alexandrium blooms in Puget Sound, WA, USA. *Harmful Algae*, **48**, 1–11.

Mouritsen LT, Richardson K (2003) Vertical microscale patchiness in nano- and microplankton distributions in a stratified estuary. *Journal of Plankton Research*, **25**, 783–797l.

Nesbitt EA, Martin RA, Martin DE, Apple J (2015) Rapid deterioration of sediment surface habitats in Bellingham Bay, Washington State, as indicated by benthic foraminifera. *Marine Pollution Bulletin*, **97**, 273–284.

Nolte V, Pandey RV, Jost S *et al.* (2010) Contrasting seasonal niche separation between rare and abundant taxa conceals the extent of protist diversity. *Molecular Ecology*, **19**, 2908–2915.

Park M, Cooney S, Yih W, Coats D (2002) Effects of two strains of the parasitic dinoflagellate Amoebophrya on growth, photosynthesis, light absorption, and quantum

yield of bloom-forming dinoflagellates. *Marine Ecology Progress Series*, **227**, 281–292.

Paul BM (2010) Polyunsaturated aldehyde production by a temporally varying field assemblage of diatoms in the San Juan Island Archipelago : can diatom metabolites affect microzooplankton grazing ? Western Washington University.

Pernice MC, Logares R, Guillou L, Massana R (2013) General patterns of diversity in major marine microeukaryote lineages. *PloS one*, **8**, e57170.

Piredda R, Tomasino MP, D'Erchia AM *et al.* (2016) Diversity and temporal patterns of planktonic protist assemblages at a Mediterranean LTER site. *FEMS Microbiology Ecology*, **in press**, fiw200.

Prokopowich CD, Gregory TR, Crease TJ (2003) The correlation between rDNA copy number and genome size in eukaryotes. *Genome / National Research Council Canada = Genome / Conseil national de recherches Canada*, **46**, 48–50.

Reñé A, Camp J, Garcés E (2015) Diversity and Phylogeny of Gymnodiniales (Dinophyceae) from the NW Mediterranean Sea Revealed by a Morphological and Molecular Approach. *Annals of Anatomy*, **166**.

Rognes T, Flouri T, Nichols B, Quince C, Mahé F (2016) VSEARCH : a versatile open source tool for metagenomics. *PeerJ*, 1–22.

Romari K, Vaulot D (2004) Composition and temporal variability of picoeukaryote communities at a coastal site of the English Channel from 18S rDNA sequences. Limnol. Oceanogr., 49(3), 2004, 784–798. *Limnol. Oceanogr*, **49**, 784–798.

Roy S, Chattopadhyay J (2007) Towards a resolution of " the paradox of the plankton ": A brief overview of the proposed mechanisms. *Ecological Complexity*, **4**, 26–33.

Rynearson TA, Armbrust EV (2004) Genetic Differentiation Among Populations of the

Planktonic Marine Diatom Ditylum brightwellii (BACILLARIOPHYCEAE). *Journal of Phycology*, **40**, 34–43.

Rynearson T, Armbrust EV (2005) Maintenance of clonal diversity during a spring bloom of the centric diatom Ditylum brightwellii. *Molecular Ecology*, **14**, 1631–1640.

Rynearson TA, Newton JA, Armbrust E V. (2006) Spring bloom development, genetic variation, and population succession in the planktonic diatom Ditylum brightwellii. *Limnology and Oceanography*, **51**, 1249–1261.

Saab MA abi (1992) Day-to-day variation in phytoplankton assemblages during spring blooming in a fixed station along the Lebanese coastline. *Journal of Plankton Research*, **14**, 1099–1115.

Shade A, Jones SE, Caporaso JG *et al.* (2014) Conditionally Rare Taxa Disproportionately Contribute to Temporal Changes in Microbial Diversity. *American Society for Microbiology*, **5**, 1–9.

Sherr B, Sherr E, Caron D, Vaulot D, Worden A (2007) Oceanic Protists. *Oceanography*, **20**, 130–134.

Sjöstedt J, Koch-Schmidt P, Pontarp M *et al.* (2012) Recruitment of members from the rare biosphere of marine bacterioplankton communities after an environmental disturbance. *Applied and Environmental Microbiology*, **78**, 1361–1369.

Smith DP, Peay KG (2014) Sequence depth, not PCR replication, improves ecological inference from next generation DNA sequencing. *PLoS ONE*, **9**.

Sogin ML, Morrison HG, Huber JA *et al.* (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere" *Proceedings of the National Academy of Sciences*, **103**, 12115–12120.

Stoeck T, Bass D, Nebel M *et al.* (2010) Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. *Molecular Ecology*, **19**, 21–31.

Stoeck T, Behnke A, Christen R *et al.* (2009) Massively parallel tag sequencing reveals the complexity of anaerobic marine protistan communities. *BMC Biol*, **7**, 72.

Sverdrup HU (1953) On Conditions for the Vernal Blooming of Phytoplankton. *Journal du Conseil International pour l'Exploration de le Mer*, **18**, 287–295.

Tanabe AS, Nagai S, Hida K *et al.* (2016) Comparative study of the validity of three regions of the 18S-rRNA gene for massively parallel sequencing-based monitoring of the planktonic eukaryote community. *Molecular ecology resources*, **16**, 402–414.

Trainer V, Moore L, Bill B *et al.* (2013) Diarrhetic Shellfish Toxins and Other Lipophilic Toxins of Human Health Concern in Washington State. *Marine Drugs*, **11**, 1815–1835.

de Vargas C, Stephane A, Nicolas H *et al.* (2015) Eukaryotic plankton diversity in the sunlit ocean. *Science*, **348**, 1–12.

Vigil P, Countway PD, Rose J *et al.* (2009) Rapid shifts in dominant taxa among microbial eukaryotes in estuarine ecosystems. *Aquatic Microbial Ecology*, **54**, 83–100.

Wang T, Yang Z, Khangaonkar T (2010) Development of a Hydrodynamic and Transport model of Bellingham Bay in Support of Nearshore Habitat Restoration. , 52.

Worden AZ, Follows MJ, Giovannoni SJ *et al.* (2015) Rethinking the marine carbon cycle: Factoring in the multifarious lifestyles of microbes. *Science*, **347**.

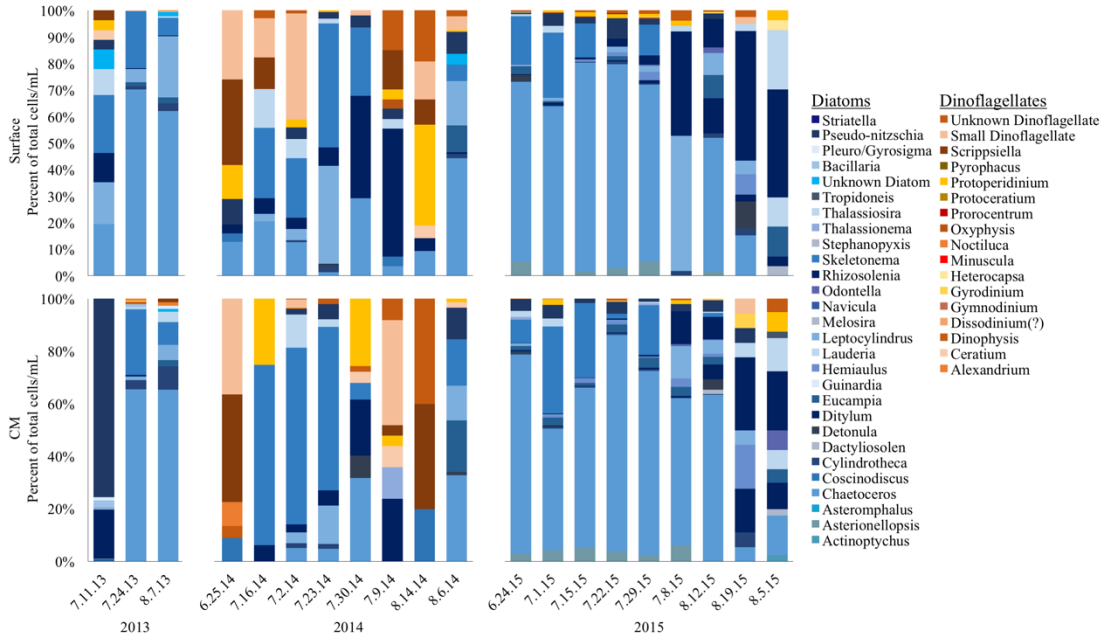
Worden AZ, Nolan JK, Palenik B (2004) Assessing the dynamics and ecology of marine picophytoplankton: The importance of the eukaryotic component. *Limnology and Oceanography*, **49**, 168–179.

Worden AZ, Not F (2008) Ecology and Diversity of Picoeukaryotes. *Microbial Ecology of the Oceans: Second Edition*, 159–205.

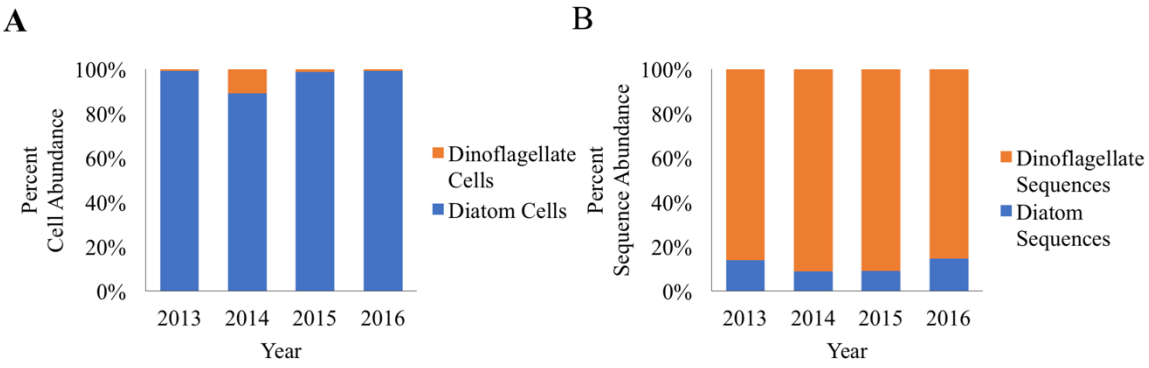
Xu X, Yu Z, Cheng F *et al.* (2017) Molecular diversity and ecological characteristics of the eukaryotic phytoplankton community in the coastal waters of the Bohai Sea , China. *Harmful Algae*, **61**, 13–22.

Yachi S, Loreau M (1999) Biodiversity and ecosystem productivity in a fluctuating environment: the insurance hypothesis. *Proceedings of the National Academy of Sciences*, **96**, 1463–8.

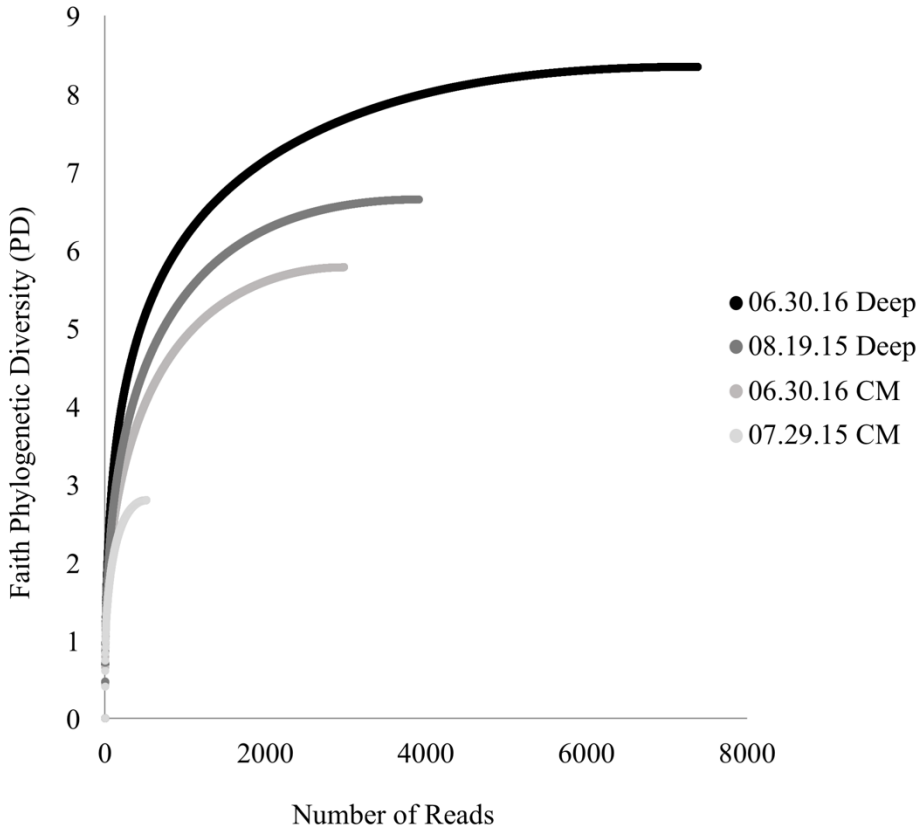
Zhang J, Stamatakis A, Kobert K (2014) Genome analysis PEAR : a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics*, **30**, 614–620.



Supplemental Figure S1. Stacked bar graphs showing the relative proportions of diatoms and dinoflagellate genera in Bellingham Bay, measured in cells/mL. Samples were collected from the surface (1m deep) and chlorophyll maximum (CM) and assessed by cell counts completed during 2014-2015 using light microscopy. Each member of chain-forming cells was counted as an individual. Warm colors indicate dinoflagellates, while cold colors indicated diatoms.



Supplemental Figure S2. Stacked bar graphs showing the relative proportions of diatoms and dinoflagellate in Bellingham Bay using microscopy versus sequencing with the Illumina MiSeq platform. **A**. Diatom:dinoflagellate ratios for cell counts (measured as cells/mL) completed in Bellingham Bay, WA from 2014-2016. **B**. Ratios of SSU rRNA gene sequences annotated as dinoflagellates and diatoms from Bellingham Bay, WA. Samples were collected from the surface (1m deep) and chlorophyll maximum (CM), but are shown averaged together.



Supplemental Figure S3. Rarefaction curves generated by guppy showing the number of reads successfully placed on reference trees versus the unrooted mean of phylogenetic diversity (PD) for microbial eukaryotic SSU rDNA gene of the 4 samples with the lowest number of reads.

Supplemental Table S1. Samples included in the study from cruises in Bellingham Bay, WA and their associated metadata. Surface samples from 2016 (indicated with *) are referred to as chlorophyll maximum (CM) samples throughout study to simplify descriptions.

Name	Date	Week	Latitude	Longitude	Description	Depth (m)
BB071113BCM18S	7/11/13	28	48.7155 N	122.561 W	Chlorophyll Max	No data
BB071113BDp18S	7/11/13	28	48.7155 N	122.561 W	Deep	No data
BB072413BCM18S	7/24/13	30	48.7155 N	122.561 W	Chlorophyll Max	No data
BB072413BDp18S	7/24/13	30	48.7155 N	122.561 W	Deep	No data
BB080713BCM18S	8/7/13	32	48.7155 N	122.561 W	Chlorophyll Max	No data
BB080713BDp18S	8/7/13	32	48.7155 N	122.561 W	Deep	No data
BB081513BCM18S	8/15/13	33	48.7155 N	122.561 W	Chlorophyll Max	No data
BB081513BDp18S	8/15/13	33	48.7155 N	122.561 W	Deep	No data
BB042214BCM18S	4/22/14	17	48.7155 N	122.561 W	Chlorophyll Max	No data
BB042214BDp18S	4/22/14	17	48.7155 N	122.561 W	Deep	No data
BB062514BCM18S	6/25/14	26	48.7155 N	122.561 W	Chlorophyll Max	3.5
BB062514BDp18S	6/25/14	26	48.7155 N	122.561 W	Deep	27.5
BB070214BCM18S	7/2/14	27	48.7155 N	122.561 W	Chlorophyll Max	5.7
BB070214BDp18S	7/2/14	27	48.7155 N	122.561 W	Deep	25
BB070924BCM18S	7/9/14	28	48.7155 N	122.561 W	Chlorophyll Max	2.26
BB070924BDp18S	7/9/14	28	48.7155 N	122.561 W	Deep	27.75
BB071614BCM18S	7/16/14	29	48.7155 N	122.561 W	Chlorophyll Max	4.8
BB071614BDp18S	7/16/14	29	48.7155 N	122.561 W	Deep	28.4
BB072314BCM18S	7/23/14	30	48.7155 N	122.561 W	Chlorophyll Max	2.6
BB072314BDp18S	7/23/14	30	48.7155 N	122.561 W	Deep	26.4
BB073014BCM18S	7/30/14	31	48.7155 N	122.561 W	Chlorophyll Max	2.7
BB073014BDp18S	7/30/14	31	48.7155 N	122.561 W	Deep	26.3
BB080614BCM18S	8/6/14	32	48.7155 N	122.561 W	Chlorophyll Max	No data
BB080614BDp18S	8/6/14	32	48.7155 N	122.561 W	Deep	No data
BB081414BCM18S	8/14/14	33	48.7155 N	122.561 W	Chlorophyll Max	1.6
BB081414BDp18S	8/14/14	33	48.7155 N	122.561 W	Deep	27.4
BB062415BCM18S	6/24/15	26	48.7155 N	122.561 W	Chlorophyll Max	10.8
BB062415BDp18S	6/24/15	26	48.7155 N	122.561 W	Deep	28
BB070115BCM18S	7/1/15	27	48.7155 N	122.561 W	Chlorophyll Max	5.1
BB070115BDp18S	7/1/15	27	48.7155 N	122.561 W	Deep	25
BB070815BCM18S	7/8/15	28	48.7155 N	122.561 W	Chlorophyll Max	7
BB070815BDp18S	7/8/15	28	48.7155 N	122.561 W	Deep	n/a
BB071515BCM18S	7/15/15	29	48.7155 N	122.561 W	Chlorophyll Max	7.9
BB071515BDp18S	7/15/15	29	48.7155 N	122.561 W	Deep	23.2
BB072215BCM18S	7/22/15	30	48.7155 N	122.561 W	Chlorophyll Max	9.81
BB072215BDp18S	7/22/15	30	48.7155 N	122.561 W	Deep	28.1
BB072915BCM18S	7/29/15	31	48.7155 N	122.561 W	Chlorophyll Max	4.6
BB072915BDp18S	7/29/15	31	48.7155 N	122.561 W	Deep	26.2
BB080515BCM18S	8/5/15	32	48.7155 N	122.561 W	Chlorophyll Max	12

BB080515BDp18S	8/5/15	32	48.7155 N	122.561 W	Deep	24
BB081215BCM18S	8/12/15	33	48.7155 N	122.561 W	Chlorophyll Max	4.8
BB081215BDp18S	8/12/15	33	48.7155 N	122.561 W	Deep	26.3
BB081915BCM18S	8/19/15	34	48.7155 N	122.561 W	Chlorophyll Max	4.9
BB081915BDp18S	8/19/15	34	48.7155 N	122.561 W	Deep	24.6
BB063016BSu18S	6/30/16	26	48.7237 N	122.5765 W	Surface*	1
BB063016BDp18S	6/30/16	26	48.7237 N	122.5765 W	Deep	20
BB070616BDp18S	7/6/16	27	48.7237 N	122.5765 W	Deep	0
BB070616BSu18S	7/6/16	27	48.7237 N	122.5765 W	Surface*	1
BB071216BDp18S	7/12/16	28	48.7237 N	122.5765 W	Deep	20
BB071216BSu18S	7/12/16	28	48.7237 N	122.5765 W	Surface*	1
BB072116BDp18S	7/21/16	29	48.7237 N	122.5765 W	Deep	20
BB072116BSu18S	7/21/16	29	48.7237 N	122.5765 W	Surface*	1
BB072816BDp18S	7/28/16	30	48.7237 N	122.5765 W	Deep	20
BB072816BSu18S	7/28/16	30	48.7237 N	122.5765 W	Surface*	1
BB080116BDp18S	8/1/16	31	48.7237 N	122.5765 W	Deep	20
BB080116BSu18S	8/1/16	31	48.7237 N	122.5765 W	Surface*	1
BB080816BDp18S	8/8/16	32	48.7237 N	122.5765 W	Deep	20
BB080816BSu18S	8/8/16	32	48.7237 N	122.5765 W	Surface*	1
BB080916BDp18S	8/9/16	32	48.7237 N	122.5765 W	Deep	20
BB080916BSu18S	8/9/16	32	48.7237 N	122.5765 W	Surface*	1
BB081016BDp18S	8/10/16	32	48.7237 N	122.5765 W	Deep	20
BB081016BSu18S	8/10/16	32	48.7237 N	122.5765 W	Surface*	1
BB081116BDp18S	8/11/16	32	48.7237 N	122.5765 W	Deep	20
BB081116BSu18S	8/11/16	32	48.7237 N	122.5765 W	Surface*	1
BB081216BDp18S	8/12/16	32	48.7237 N	122.5765 W	Deep	20
BB081216BSu18S	8/12/16	32	48.7237 N	122.5765 W	Surface*	1
BB081516BDp18S	8/15/16	33	48.7237 N	122.5765 W	Deep	20
BB081516BSu18S	8/15/16	33	48.7237 N	122.5765 W	Surface*	1
BB082216BDp18S	8/22/16	34	48.7237 N	122.5765 W	Deep	20
BB082216BSu18S	8/22/16	34	48.7237 N	122.5765 W	Surface*	1

Supplemental Table S2. Number of paired-end, quality trimmed and quality filtered sequences put into pplacer, and number of sequences successfully placed on the phylogenetic trees.

Name	Number Paired-End Sequences	Placements
BB071113BCM18S	93651	64374
BB071113BDp18S	35247	24451
BB072413BCM18S	49660	34897
BB072413BDp18S	52250	33695
BB080713BCM18S	61004	40682
BB080713BDp18S	186254	119258
BB081513BCM18S	71395	45236
BB081513BDp18S	96574	60191
BB042214BCM18S	29569	20890
BB042214BDp18S	57593	39181
BB062514BCM18S	135530	92858
BB062514BDp18S	153249	98102
BB070214BCM18S	85970	57643
BB070214BDp18S	30847	20312
BB070924BCM18S	52948	33386
BB070924BDp18S	175437	116227
BB071614BCM18S	65621	46181
BB071614BDp18S	45492	30071
BB072314BCM18S	49475	33984
BB072314BDp18S	48364	30628
BB073014BCM18S	57781	39168
BB073014BDp18S	153535	108677
BB080614BCM18S	161600	119293
BB080614BDp18S	67724	41301
BB081414BCM18S	192159	139813
BB081414BDp18S	78078	47486
BB062415BCM18S	94931	70146
BB062415BDp18S	54883	38918
BB070115BCM18S	41350	28768
BB070115BDp18S	106749	73304
BB070815BCM18S	98855	61519
BB070815BDp18S	54576	34301
BB071515BCM18S	4285	2990
BB071515BDp18S	19477	11490
BB072215BCM18S	11760	8362
BB072215BDp18S	73524	49960
BB072915BCM18S	1014	516
BB072915BDp18S	30603	18377
BB080515BCM18S	57414	37347

BB080515BDp18S	33580	19350
BB081215BCM18S	58545	38690
BB081215BDp18S	18161	10825
BB081915BCM18S	59998	38229
BB081915BDp18S	7797	3909
BB063016BSu18S	5616	2980
BB063016BDp18S	14462	7385
BB070616BDp18S	48770	26738
BB070616BSu18S	36409	21818
BB071216BDp18S	77039	50392
BB071216BSu18S	84819	60964
BB072116BDp18S	78689	54616
BB072116BSu18S	46533	30578
BB072816BDp18S	96602	66159
BB072816BSu18S	87550	63706
BB080116BDp18S	51537	36477
BB080116BSu18S	74302	54235
BB080816BDp18S	93376	63478
BB080816BSu18S	27795	19840
BB080916BDp18S	62350	44756
BB080916BSu18S	26879	19814
BB081016BDp18S	190121	128732
BB081016BSu18S	15200	10633
BB081116BDp18S	53285	35458
BB081116BSu18S	21185	15250
BB081216BDp18S	35693	24463
BB081216BSu18S	28839	20192
BB081516BDp18S	24501	16356
BB081516BSu18S	54920	40784
BB082216BDp18S	66288	43505
BB082216BSu18S	24441	16586

Supplemental Table S3. Unknown taxa counts from Bellingham Bay.

No Match	
Alveolata	131
Amoebozoa	496
Cryptophyta	900
Euglenozoa	9
Eukaryota	642
Glaucocystophyceae	16
Rhizaria	544
Stramenopiles	26,567
Total No Match Reads	**29,305**

Environmental Samples	
Dinophyceae	115,299
Chlorophyta	13
Named eukaryotic clones	16,334
Total Environmental Samples	**131,646**

Unclassified Organisms	
Unclassified Babesia (Apicomplexa)	13
Unclassified Cercozoa	2,911
Unclassified Chlamydomonadaceae	701
Unclassified Chlorophyceae	8
Unclassified Chrysophyceae	265
Unclassified *Spumella* (Chrysophyta)	629
Unclassified Fungi	8
Unclassified Thraustochytriidae (Labrinthulomycete)	2,130
Unclassified Thraustochytrium (Labrinthulomycete)	471
Unclassified Cryptophyta	769
Total Unclassified	**7,905**

Total Unknown Reads	**168,856**
Total Sequence Reads	**3060881**

Supplemental Table S4. Diatom total cell counts (in cells/mL) from Bellingham Bay, WA, categorized by whether they were often found in chains or as solitary cells, and listed in order of decreasing abundance. Samples were collected from the chlorophyll maximum and 1m above the bottom of the bay from 2013-2016.

Diatom Genera	Cells/mL
Chain-Forming	
Chaetoceros	5,584,801
Pseudo-nitzschia	4,981,838
Skeletonema	2,535,527
Leptocylindrus	677,723
Thalassiosira	221,297
Eucampia	146,569
Thalassionema	115,676
Asterionellopsis	84,287
Ditylum	39,233
Hemiaulus	16,842
Detonula	11,913
Dactyliosolen	10,788
Lauderia	5,546
Melosira	4,961
Odontella	3,235
Guinardia	924
Solitary	
Rhizosolenia	479,976
Cylindrotheca	177,926
Unknown Diatom	49,446
Bacillaria	40,371
Pleuro/Gyrosigma	32,297
Coscinodiscus	7,015
Navicula	2,619
Tropidoneis	616
Asteromphalus	308
Actinoptychus	205

Supplemental Table S5. Dinoflagellate total cell counts (in cells/mL) from Bellingham Bay, WA, listed in order of decreasing abundance. Samples were collected from the chlorophyll maximum and 1m above the bottom of the bay from 2013-2016.

Dinoflagellate Genera	Cells/mL
Small Dinoflagellate	103,705
Protoperidinium	40,909
Scrippsiella	26,119
Gyrodinium	20,072
Dinophysis	19,604
Ceratium	15,336
Unknown Dinoflagellate	14,799
Noctiluca	13,405
Prorocentrum	6,388
Oxyphysis	1,510
Gymnodinium	1,420
Heterocapsa	1,322
Protoceratium	462
Dissodinium(?)	154
Minuscula	154