THE UNIVERSITY OF CHICAGO

TRACING INTERACTIONS BETWEEN PICOPHYTOPLANKTON AND PHAGE THROUGH DISSOLVED ORGANIC MATTER

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF THE GEOPHYSICAL SCIENCES

BY
XIUFENG MA

CHICAGO, ILLINOIS
MARCH 2018

Dedicated to my family for their love and support.

Thou canst not stir a flower/ Without troubling of a star. - The Mistress of Vision (1913)

by *Francis Thompson*

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

It's a challenging and rewarding journey under the guidance of my advisors - Jacob Waldbauer and Maureen Coleman. I am really grateful to have worked in their labs and have met a handful of lovely co-workers and colleagues along the way. I learned a lot about biogeochemistry and microbiology, how to think and how to be a better person. I am also indebted to my two other committee members - Albert Colman and David Archer who had offered insights to my research and life as a graduate student. Albert was the very first person who introduced me to Jake and Maureen's lab and who gave me a lovely impression of Hinds - our Geophysical Sciences department. Moreover, I am thankful to nice people I met and friends I made in the department. Love was brewed during the happy hour and coffee-tea time. I owe many thanks for Gerry who has maintained the liveliness of our third floor and helped with the vertical mixing of laughters between floors.

Together, we are grateful to Michael Rust and Eugene Leypunskiy for helping with mechanical lysate preparation, to Lichun Zhang for operation and maintenance of the LC-MS, and to Aric Mine and Kate Campbell for collaboration on cell cultivation and viral infection. We also thank Gerald Olack, Mark Anderson, Sean Gibbons, Albert Colman and Sara Paver for insightful discussions. We also thank Mark Anderson and Sean Gibbons for helping with R packages and Hao Xiong for reviewing Python scripts. This work was supported by the Gordon and Betty Moore Foundation Marine Microbiology Initiative (Award #3305).

ABSTRACT

This thesis explores for the first time how the mechanism of dissolved organic matter (DOM) release from the biomass of abundant marine phytoplanktons - exudation, mechanical lysis or viral lysis - affects the composition of the released DOM. Experiments were carried out both in lab conditions and in a field-based incubation. For lab culture experiments, two model systems were explored separately: *Synechococcus* WH7803 and phage S-SM1, and *Prochlorococcus* MED4 and phage PHM2. We performed these molecular composition measurements using a broad, untargeted high-resolution mass spectrometry approach, and analyzed the large resulting dataset with a novel computational pipeline that revealed the molecular features that make each of the three types of DOM compositionally distinct. For the *Synechococcus* experiment in particular, we find evidence that the picocyanobacterium *Synechococcus* WH7803, releases a set of unsaturated, oxygen-rich and possibly novel biomolecules into the DOM. When lysed by the virus S-SM1, abundant peptides derived from specific proteolysis of the major light-harvesting protein phycoerythrin are released, implicating phage infection of these abundant cyanobacteria as a potentially major source of dissolved organic nitrogen compounds in the oligotrophic surface ocean. For the other model marine phytoplankton - *Prochlorococcus* MED4, peptides related to photosynthesis were found in its exudate as compared to its mechanical lysate and viral lysate. For the 48-hour field incubation experiment, phosphorus concentration and different seawater types were controlled to explore the dynamics of nutrient-virus-phytoplankton interactions. Decrease of viral content in the seawater and phosphorus addition had the most effect on the microbial community structure as indicated by 16S RNA while only viral concentration had a pronounced impact on the DOM composition. Together, these evidence indicates that DOM released from phytoplankton via three modes - exudation, mechanical lysis and viral lysis is compositionally distinct. Controlled cyanobacteria-virus interactions reveal potentially overlooked compound groups in the natural environment and such compounds could serve as future biomarkers.

CHAPTER 1

INTRODUCTION

Oceanic dissolved organic matter (DOM) is one of the largest pools of reduced carbon on Earth, comparable to the amount of carbon to the atmospheric CO2 reservoir. DOM plays a central ecological role in marine microbial communities as the chemical medium through with carbon and nutrients are exchanged between diverse auto- and heterotrophic organisms. Despite its importance to the operation of marine ecosystems, the processes by which the oceanic DOM pool is generated remain little understood at the molecular level. This is due in part to the chemical nature of marine DOM: the standing pool of DOM is dominated by the more biologically recalcitrant compounds that turn over only very slowly (in some cases, on timescales of thousands of years), while the more labile compounds that fuel the highly active biogeochemistry of the surface ocean (and turn over on timescales of hours) are rapidly taken up and metabolized and represent only a tiny portion of the in situ pool. Hence we know relatively little about exactly what kinds of molecules are contributed to the DOM pool by photosynthetic primary producers – the ultimate source of organic carbon in the open ocean – nor how different mechanisms of DOM release influence its composition. Labile DOM is thus also a link to microbial metabolisms and in this sense can be considered as metabolites. Therefore, DOM connects metabolomics study at the micro level to biogeochemistry at the global scale, making researches of DOM on the aquatic environment highly interdisciplinary.

In the oligotrophic low-latitude surface ocean, particularly the vast subtropical gyres that are the largest sunlit biomes on earth, picocyanobacteria of the genera *Synechococcus* and *Prochlorococcus* are the dominant primary producers [49, 83] . The large population sizes of these picocyanobacteria (~10E8 cells per liter of seawater down to ~200m depth) are maintained roughly steady, despite growth rates of 0.5-1 division per day, by equally efficient cell death processes. Chief among these is lytic viral infection, which is estimated to kill 20-40% of marine prokaryotes every day [82].

The understanding of such interplays has been confounded by intrinsic features of DOM

and microbial ecosystems. On the one hand, rapid biological utilization of labile DOM [44] makes it difficult to characterize its compositions [70]. On the other hand, small amount of labile DOM or microbial metabolites (< 1%) of the ocean DOC are overwhelmed by the much larger fraction of non-labile DOM [32], making direct measurement and identification of oceanic labile DOM very challenging. Further, previous lack of computational and analytical tools in high-resolution mass spectrometry and well-annotated databases for metabolomics also hindered a fuller appreciation of these connections [56].

1.1 Dissolved Organic Matter, Picophytoplankton, Phage, and Nutrients in the Ocean

Almost 50% of global primary production occurs in the ocean [21], mainly accomplished by marine picophytoplankton. Particularly, picocyanobacteria contribute to 90% of primary production in the oligotrophic open oceans [49, 60, 83]. DOM is known to be released from the biomass of these unicellular cyanobacteria by at least three principal mechanisms: natural exudation of photosynthate by growing cells (incl. membrane vesicles), lytic viral infection [5], and 'sloppy feeding' (cell breakage) by protistan grazers [12]. These three types of DOM channels have a particular impact on the labile dissolved organic matter pool which is connected with various microbial activities albeit being a small proportion of the total DOC pool (usually < 2%) [31]. The relative importance of these different DOM release channels likely varies in space and time across the oceans, and is intimately interwoven with interacting ecological and biogeochemical factors such as allochthonous nutrient delivery, seasonality, plankton bloom dynamics and higher-order trophic interactions. An outstanding first-order question regarding DOM inputs from planktonic primary producers is to what extent the mechanism of DOM release influences the molecular composition of the released material, and therefore what sorts of compounds might be useful 'markers' for the occurrence of different DOM release mechanisms in natural environments.

There is reason to expect that the mechanism of release – exudation, viral lysis, or sloppy feeding – is likely to affect the composition of the DOM produced from the biomass of a given type of phytoplankton. Exudates are generally seen as either 'metabolic overflow' (e.g., photosynthate in excess of the amount that can productively be combined with the available amounts of other nutrients to produce biomass) and/or compounds that are specifically exported by cells for purposes of signaling, nutrient acquisition or defense. The process of cell breakage by grazers during sloppy feeding results in mechanical disruption of otherwise healthy cells, releasing some intracellular contents and membrane fragments into the dissolved phase.

Lysis of cells by viruses, by contrast, occurs after an extended infection process that results in extensive biochemical remodeling of the host cell, as its metabolism is directed away from cellular replication and towards production of progeny virions. The potential of cyanophages (i.e., viruses infecting cyanobacteria) ability to affect the local food web and global biogeochemical cycles has been more and more supported by the discovery of phage auxiliary metabolic genes (AMGs), their expression and function during the infection (See Section 1.2). DOM released from oceanic diatoms due to viral lysis or 'sloppy feeding' was observed to partition among distinct microbial populations based on diffusivity, and that consumption is skewed toward very few cells due to chemotaxis [80]. At the community level, specific DOM molecular formulae groups have been found to associate with particular bacterial communities [59].

On the other hand, the availability of nutrients to microbes can influence the production and consumption of DOM in the microbial community. Phytoplankton productivity have been found to be influenced by inorganic nutrient sources such as ammonium, nitrate [18] and phosphate [70]. A recent study done by Goldberg et al. [29] has shown that enrichments of both two nitrogen sources (ammonium and nitrate) stimulated bulk phytoplankton, bacterial and DOM production and and enriched *Synechococcus* and *Flavobacteriaceae*.

1.2 Influence of Viruses on DOM and Biogeochemistry

Viruses infecting cyanobacteria (i.e., cyanophages) are increasingly found to play an important role in global biogeochemical cycling [76]. The interaction among cyanobacteria, viruses and labile DOM with fast turnover rate on the scale of days to weeks play a crucial role in the microbial loop. Viral infection leads to a surprising 20-50% of microbial biomass turnover each day in the ocean [82] releasing viral lysate into the seawater. These phages carry auxiliary metabolic genes (AMGs) that can rewire various metabolic pathways in the infected host cell. A few dominant cyanophages' genome contain, for example, photosynthetic genes [48], phosphate uptake genes [81], and genes encoding fatty acid desaturase [68], and so on. Entangled with other microbial processes such as protistan grazing on cyanobacteria and natural exudation, the effect of viral-cyanobacteria needs to be explored in order to better understand DOM cycling in the microbial loop [37].

Most previous studies focused on examining DOM produced in single cultures of marine phytoplankton [8, 62, 22, 70]. A recent study have looked into the interactions between phage and heterotrophic bacteria [5]. Yet little is explored about DOM generated via phage-host interaction or viral lysing of dominant primary producers such as cyanobacteria, which is one of the crucial channel through which labile DOM is infused into the microbial loop.

1.3 DOM Research in the Future

The study of labile DOM or metabolites belongs to the realm of Natural Products (NPs) research. NPs refer to, for example, crude extracts of various origin (e.g., plants, marine organisms, and microorganisms) [4]. Tools and methods developed in general NPs research can be shared and applied in environmental metabolomic studies.

Emerging research fields such as environmental metabolomics [46] will facilitate the understanding of DOM as a link of complex ecosystem network. Untargeted DOM or metabolomics study will continue to benefit from a deeper and more mature pure culture

or model biological systems researches. On the technology side, measurement platforms such as mass spectrometry and nuclear magnetic resonance have enabled high precision and high sensitivity measurement of DOM. Such stat-of-art instruments will gradually populate through labs at research institutions. Meanwhile, adoption of computational techniques and improvement on metabolomic database curation will largely benefit DOM studies.

Lack of well-annotated metabolomic database has become a bottleneck for identification of metabolomic compounds or DOM and cross-lab comparisons. Current popular databases, such as PubChem and KEGG, have limited compound candidates from biological samples. Particularly, well-curated MS2 (i.e., fragmentation spectra from parent ions) databases are needed to facilitate untargeted mass spectrometry based metabolomic discovery. This type of database, together with new MS2 computation tools, such as MetFrag [69], will push metabolomic and DOM research to a new level.

1.3.1 Data Science in Metabolomics

Labile DOM or metabolomics researches are advancing hand in hand with cutting-edge computational tools in the 'omics studies. Compared to proteomics and genomics, physically small metabolites lack of a repetitive pattern and therefore often demands more information for identification and analysis. Metabolomic-specific tools have evolved rapidly over the past decade.

Identification of metabolites : Identifying metabolites are of crucial importance in metabolomic research. There have been a range of techniques developed, such as elemental formula calculation [61, 84], and molecular networking based on matching fragmentation patterns. Many of these techniques have conveniently been deployed as new cloud-computing services as compared to conventional stand-alone software application on the client side.

Elemental formula calculation has been helpful for untargeted metabolomics and screening of broad patterns for large groups of molecular clusters, as shown in this thesis study (For example, Figure 2.2b). We adopted Chemcalc [61] which is an easy-to-query RESTful (Rep-

resentational state transfer) web service and developed client side Python pipelines using its API (Application Programming Interface). Chemcalc provides fast brutal-force elemental formula calculations based on a variant of FIND-ALL algorithm that has been used to solve the Money Changing Problem [10]. We were able to feed the query with molecular mass, charge state and a predefined elemental-atom-limitation and get possible formula candidate results within a certain mass error.

Various novel bioinformatics approaches such as molecular networking (MN) and in-silico fragmentation tools have emerged recently and provide new perspective for early metabolite identification in natural products (NPs) research [3]. Global Natural Products Social Molecular Networking (GNPS) [84] has becoming increasingly popular for metabolomic researches and was a critical step for our own metabolomic analysis. It scales well for our 30-50GB dataset each run. GNPS compares mass-to-charge ratio (m/z) and MS2 fragmentation pattern of parent ions. MS2 spectra that have parent m/z values within a defined tolerance ($\leqslant 0.03$ Dalton) and sufficiently similar fragmentation patterns as measured by cosine scores ($\geqslant 0.7$) are considered members of the same compound ion.

Yet, there are challenging issues ahead. Metabolome annotation and the corresponding library construction are of critical importance in the lengthy process of de novo identification of the diverse metabolites [3, 43]. While high-throughput and high-resolution spectroscopic data on NPs has been achieved particularly via high-resolution mass spectrometry, the primary bottleneck still exists in annotating the high-quality spectral data and thus the accurate identification of compounds of interest. Besides simple molecular formula determination (e.g., with brutal force calculator such as [61], orthogonal types of information (e.g., chemotaxonomy, retention time, fragmentation patterns [84, 69, 34], etc.) are increasingly been used to generate scoring schemes for the most possible and robust identification of natural compounds. In recent years, several big metabolome databases focusing on human [85], pharmaceuticals [79] and life science [33] have been developed, but development of metabolomic databases for environmental microbial communities are particularly needed.

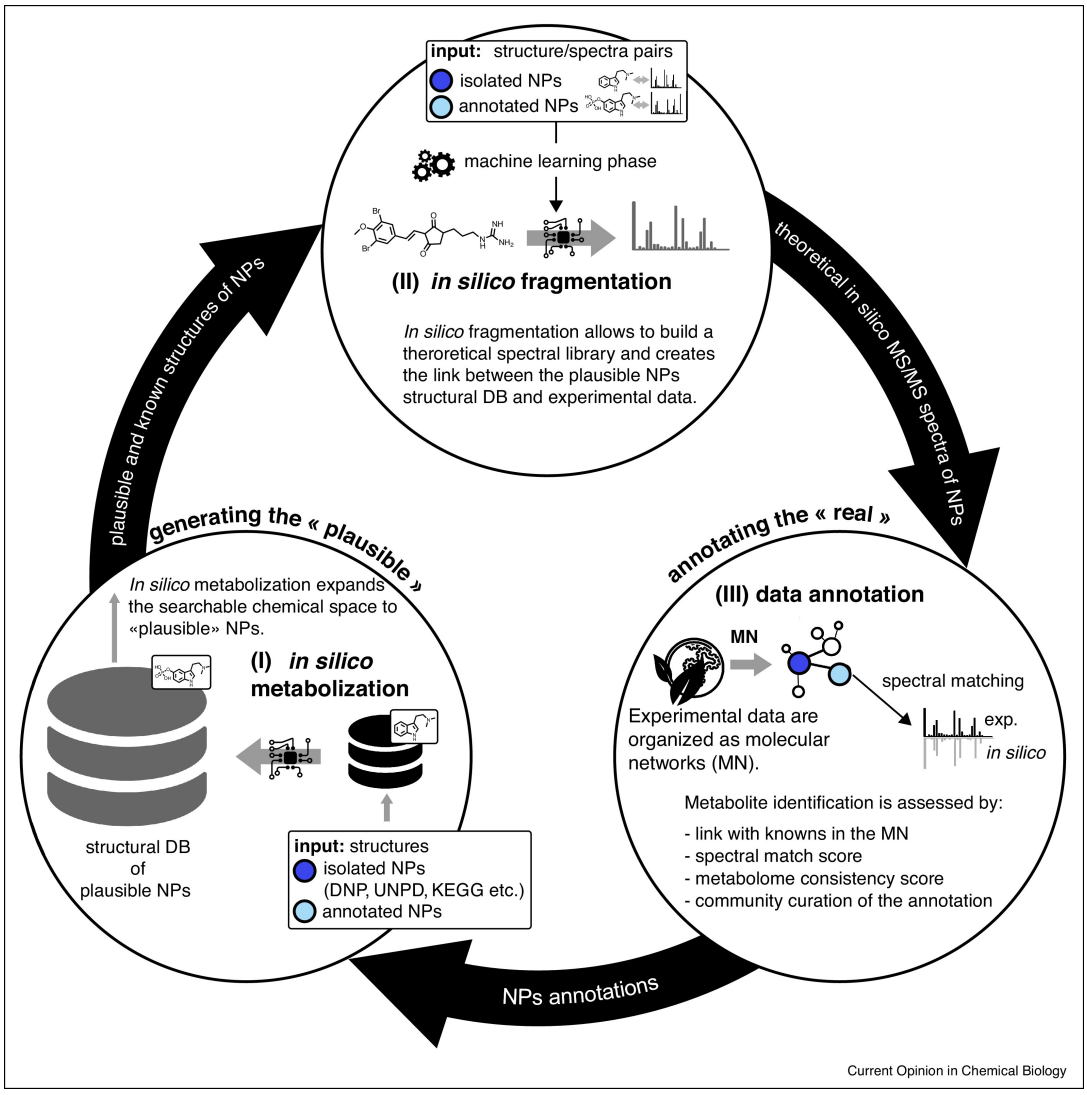


Figure 1.1: Conceptual virtuous cycle of metabolite identification. From Allard et al [3]

Analysis of metabolomic data : In order to understand metabolomic data, a crucial first step is to understand the MS/MS and MSn (i.e., the fragmentation data) since MS1 data provides limited information about ions or molecules detected in the LC-MS. For MS2 data, fragmentation data can be organized based on the concept of Molecular Networks (MN). Global Natural Products Social Molecular Networking (GNPS) [84] platform has populated this concept in recent years and now widely used for metabolomic research. For MSn data, the idea of fragmentation trees (molecular networks at different levels) is adopted [3]. Based on molecular networks or fragmentation trees, spectra can be grouped into spectral clusters for identification or analysis.

Although metabolomic data requires special pipelines (e.g., acquisition, treatment, organization, annotation) as compared with techniques used in genetics and proteomics, some general statistical analysis can be transplanted for metabolomic analysis [66]. There exist two different approaches for analyzing metabolomic data: chemometric and quantitative analysis. Chemometric analysis applies statistical methods directly on spectral patterns or counts and signal intensity data while only looking into the identity of metabolites if needed. Quantitative metabolomics tries to identify all metabolites first and then perform the analysis [66]. Due to the lack of annotations for environmental metabolomics, chemometric analysis is mostly used in DOM research. Table 1.1 is a summary of popular metabolomics data analysis techniques adapted from Ren et al [66].

Table 1.1: Summary of commonly used metabolomics data analysis methods

Type of analysis	Basic methods		Goal	Application	Input	Output	Software
Basic statistical testing	Fold Change		Biomarker discovery; feature selection for supervised learning	Fold change of metabolite concentrations between two groups	Data tables with class membership; each row represents one subject and each column represents concentration of a metabolite/MS and NMR peak list or spectral bin	Lists of selected metabolites with p-values, volcano plot (for two groups)	MetaboAnalyst, R and MATLAB
	Statistical testing	Two sample tests (e.g., t test)		Identify significantly different expressed metabolites between two groups			
		ANOVA with post hoc analysis		Identify significantly different expressed metabolites for multiple groups			
Unsupervised learning	Principal component analysis (PCA)		Data grouping and visualization	Reduce dimensionality and check clusters visually	Same as above, but no need for class memberships	Scores and loadings plots for visualization	MetaboAnalyst, R and MATLAB
	Clustering	K-means, fuzzy c-means, K-means in reduced space		Group the subjects into k different clusters		Cluster labels for each subject	
		Hierarchical		Show how subjects form different clusters		Heatmap, dendrogram	
	Self-organizing map (SOM)		Reduce dimensionality and check clusters visually		SOM		

Continued on following page

Table 1.1, continued

Type of analysis	Basic methods	Goal	Application	Input	Output	Software
Supervised learning-classification	Support vector machine (SVM) Partial least square-discriminant analysis(PLS-DA), OPLS-DA and SPLS-DA	Predict class memberships (disease diagnostic), biomarker discovery	All of them are classification methods suitable for metabolomics research; need to compare their performance in real data analysis	Same as above, but class memberships are needed	A prediction model with selected metabolites as predictors	MetaboAnalyst, R and MATLAB
Supervised learning-regression	Support vector regression (SVR) Partial least square (PLS), OPLS and SPLS	Predict continuous variables, calibration, biomarker discovery	All of them are regression methods suitable for metabolomics research; need to compare their performance in real data analysis	The response variables should be continuous vector or matrix; for calibration problem, the response variables are concentrations and covariates are spectral intensities	A prediction model with selected variables as predictors	
Pathway analysis	Over-representation analysis (ORA) Functional class scoring (FCS)/enrichment analysis	Find biologically meaningful metabolite sets or pathways that are associated with certain diseases or characters	Find related pathways or metabolite sets using statistical testing	Significant metabolites list and reference pathways Metabolites list with concentrations and reference pathways or metabolite sets	A list of selected pathways with their p-values or FDRs	MetaboAnalyst, bioconductor
	Elementary modes/ Extreme pathways	Simulation and pathway reconstruction	Analyze or reconstruct metabolic pathways	Pathway, stoichiometric matrix	Several different possible flux distributions	MATLAB

Continued on following page

Table 1.1, continued

Type of analysis	Basic methods		Goal	Application	Input	Output	Software
	Flux Balance Analysis		Find the flux distribution by maximizing/ minimizing a reaction based on some constraints	Pathway, stoichiometric matrix, constraints, and an objective function	One unique flux distribution		
	Kinetic reaction network model			Simulate the dynamic behavior of metabolites reaction networks	Pathway, stoichiometric matrix, Kinetic parameter and rate equation	Simulation results of network kinetics	
Time course data analysis	Analysis of variance (ANOVA)	Fixed effects ANOVA Repeated measures	Test time dependent effects; compare time profiles for metabolites from different subjects	Two or more given factors ANOVA method when we have multiple within subjects measurements (measured at different time points)	Data tables with labels (class membership indicators): each row represents the observation value of one subject at one time point and columns represent metabolites concentrations/MS and NMR peak lists or spectral bins	Testing results of time dependent effects for each metabolite, time profile plots	MetaboAnalyst, R
	ANOVA-simultaneous component analysis (ASCA)			ANOVA for multivariate response		Scores and loadings plots of sub models, time profiles plots	
	Functional based methods (e.g. smoothing splines mixed effects model)			Estimate sets of curves (metabolic profiles) and quantify their differences		Testing results of the differences in time profiles between groups for each metabolite	

Continued on following page

Table 1.1, continued

Type of analysis	Basic methods	Goal	Application	Input	Output	Software
	Time series models (e.g. ARMA model)	Model the dynamic properties (e.g. biorhythm) of metabolites data	Long time series data modeling		A model that best describes the dynamic properties of underlying biological process	

CHAPTER 2

DOM COMPOSITIONS OF VIRAL LYSATE, INTRACELLULAR METABOLITES AND EXUDATE FROM *SYNECHOCOCCUS*: LAB CULTURE EXPERIMENT

2.1 Overview

Oceanic dissolved organic matter (DOM) is one of the largest pools of reduced carbon on Earth, comparable in size to the CO2 content of the atmosphere [20]. DOM plays a central ecological role in marine microbial communities as the chemical medium through with carbon and nutrients are exchanged between diverse auto- and heterotrophic organisms. Despite its importance to the operation of marine ecosystems, the processes that produce oceanic DOM remain poorly understood at the molecular level. This is due in part to the chemical nature of marine DOM: the standing pool of DOM is dominated by the more biologically recalcitrant compounds that turn over only very slowly (in some cases, on timescales of thousands of years), while the more labile compounds that fuel the highly active biogeochemistry of the surface ocean (and turn over on timescales of hours) are rapidly taken up and metabolized and represent only a tiny portion of the in situ pool [56]. Hence we know relatively little about exactly what kinds of molecules are contributed to the DOM pool by photosynthetic primary producers – the predominant source of organic carbon in the open ocean – nor how different mechanisms of DOM release influence its composition.

In the oligotrophic low-latitude surface ocean, particularly the vast subtropical gyres that constitute the largest sunlit biomes on earth, picocyanobacteria of the genera *Synechococcus* and *Prochlorococcus* are the numerically dominant primary producers [72, 24]. The large population sizes of these picocyanobacteria (~108 cells per liter of seawater down to ~200m depth) are held roughly constant over time, despite growth rates of 0.5-1 cell division per day, by equally efficient cell death processes. Prominent among these is lytic viral infection,

which is estimated to kill 20-40% of marine prokaryotes every day [82]. As a result of the lysis process, contents of the infected cells are released to the dissolved phase on bursting. The overall magnitude of viral lysis as a source of marine DOM is poorly constrained, but it is clearly a principal DOM production mechanism [56].

In addition to lytic viral infection, DOM is known to be released from the biomass of unicellular cyanobacteria by at least two other mechanisms: exudation of photosynthate by growing cells [7, 22] and 'sloppy feeding' (cell breakage) by protistan grazers [75, 12]. The relative importance of these different DOM release modes likely varies in space and time across the oceans, and is intimately interwoven with interacting ecological and biogeochemical factors such as allochthonous nutrient delivery, seasonality, plankton bloom dynamics and higher-order trophic interactions. An outstanding first-order question regarding DOM inputs from planktonic primary producers is to what extent the mechanism of DOM release influences the molecular composition of the released material, and therefore what sorts of compounds might be useful 'markers' for the occurrence and prevalence of various DOM release mechanisms in natural environments.

There is reason to expect that the mechanism of release – exudation, viral lysis, or sloppy feeding – is likely to affect the composition of the DOM produced from the biomass of a given type of phytoplankton. Exudates are generally seen as either 'metabolic overflow' (e.g., photosynthate in excess of the amount that can productively be combined with the available amounts of other nutrients to produce biomass) and/or compounds that are specifically exported by cells for purposes of signaling, nutrient acquisition or defense. The process of cell breakage by grazers during sloppy feeding results in mechanical disruption of otherwise generally healthy cells, releasing some intracellular contents and membrane fragments into the dissolved phase. Lysis of cells by viruses, by contrast, occurs after an extended infection process that results in extensive biochemical remodeling of the host cell, as its metabolism is directed away from cellular replication and towards production of progeny virions. In comparing the composition of DOM released from phage-lysed *Sulfitobacter* to that from

uninfected cells, Ankrah et al. [5] found substantial enrichment of cell wall constituents and coenzyme A metabolites, and depletion of several amino acids, in the lysates. Hence we expect that, to some degree, these three modes of DOM release contribute biochemically distinct components of the phytoplankton cell to the dissolved pool.

We conducted a set of experiments using *Synechococcus* WH7803 and its lytic phage S-SM1 as a model virus-host system to investigate the composition of DOM released from primary producer biomass by three major mechanisms: viral lysis, exudation, and mechanical cell breakage. The composition of DOM released by each mechanism was analyzed by high-resolution mass spectrometry. We developed an integrated data pipeline using the GNPS mass spectrometry data platform [84] and statistical tools adapted from gene expression analysis [51] to find differentially abundant chemical compounds among the three modes of DOM production.

2.2 Results

This study is designed to determine compositional distinctions between dissolved organic matter produced from picocyanobacterial biomass via three modes of release: viral lysis, exudation, and mechanical lysis (Figure 2.1). Using a broad, untargeted analytical approach, between 24,000-65,000 MS2 spectra were collected from biological triplicate samples of *Synechococcus* WH7803 mechanical lysate, exudate and S-SM1 viral lysate DOM (Table 2.1). We used mechanical lysis to simulate DOM release via grazer sloppy feeding, rather than co-culture with a protistan predator, in order to avoid having another cell type in the culture that would be consuming and reworking the primary released DOM, confounding comparisons with exudate and viral lysate. Our chosen DOM extraction method (solid-phase extraction with BondElut PPL) is widely used for studies of marine DOM [15], though it is somewhat biased towards more hydrophobic and higher-molecular weight compounds, and is unlikely to reveal information about small, highly polar molecules that are accessible in more targeted approaches (e.g., Ankrah et al. 2014 [5], Fiore et al. 2015 [23]). Spectra from each

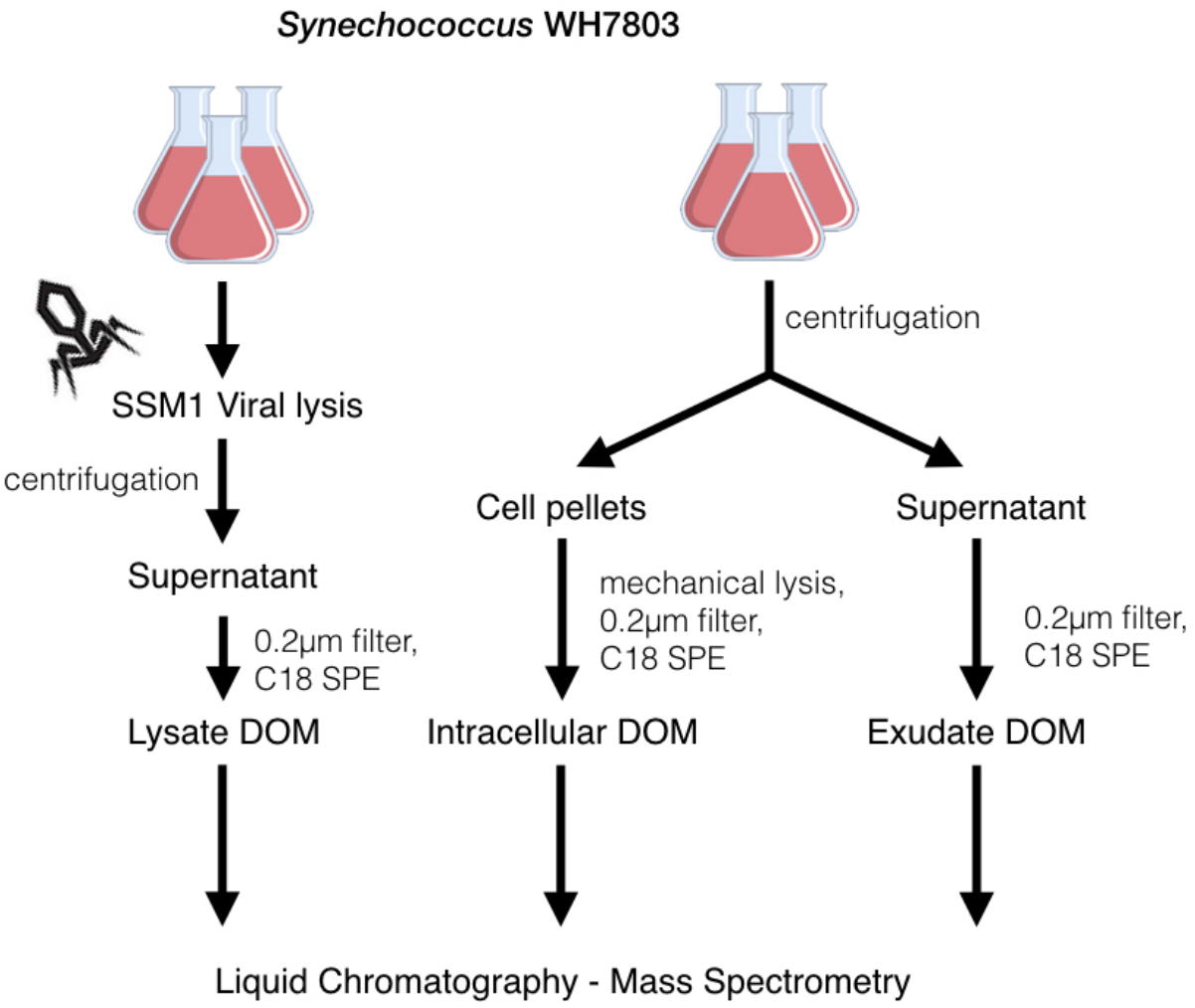


Figure 2.1: Experimental design for preparation of S-SM1 viral lysate, exudate and mechanical lysate DOM from *Synechococcus* WH7803 cultures.

sample were grouped into spectral clusters by the GNPS MS2 algorithm [84]. A spectral cluster can be viewed as representing a chemical species with a certain mass-to-charge ratio (m/z) and MS2 fragmentation pattern. MS2 spectra that have parent m/z values within a defined tolerance ($\leqslant$ 0.03 Da) and sufficiently similar fragmentation patterns as measured by cosine scores ($\geqslant$ 0.7) are considered members of the same spectral cluster. MS2 spectral clusters can thus serve as a basis for comparisons within and between samples.

The entire 9-sample dataset comprises 507,385 MS2 spectra, which were organized by GNPS into 55,913 clusters. We removed 5799 clusters with parent masses that matched

those also observed in the artificial seawater media blank, and restricted further analyses to clusters that were observed at least in all three replicates of one condition; there were 6781 such clusters (12.1% of the total; Table 2.1), which constituted 284,158 spectra (56.0% of the total). The clusters excluded from downstream analysis due to inconsistent occurrence in the full dataset were predominantly of low spectral count (<5 total spectra per cluster; Fig 2.6). Of the 6781 consistently-observed clusters, 87 (comprising 28,151 spectra) were present in all 9 samples, while between 1352 and 3369 clusters were unique to each of the DOM types, and generally smaller numbers were seen in 2 of 3 conditions (Fig 2.2a). The viral lysate DOM contained the highest number of compound clusters and the highest proportion of clusters found in only one DOM type (Fig 2.2a, Table 2.1), indicative of viral lysis as a source of particularly complex and compositionally distinct organic matter.

Best-match elemental formula assignments to each cluster based on exact mass and biomolecular composition rules indicate a broad range of biomolecular classes in all three types of DOM (Fig 2.2b). The viral and mechanical lysates contain more high-MW compounds than does the exudate DOM (Fig 2.2c), suggesting the importance of cellulolytic processes as inputs of larger molecules to the DOM pool. A substantial number of clusters are predicted to have low H/C ratios (~0.3-1.2) and high O/C ratios (~0.6-1.0) (Fig 2.2b). This unsaturated, oxygen-rich organic matter appears particularly characteristic of viral lysate, and also somewhat so of the mechanical lysate, though the clusters' relatively low spectral counts mean they do not meet the statistical criteria used in this study (see below) for differentially abundant compounds.

2.2.1 Compounds Distinct to Each Mode of DOM Release

The number of MS2 spectra belonging to a given spectral cluster (the spectral count) can be used as a measure of relative abundance of the compound represented by that cluster, at least when analytical conditions are held close to constant between samples. Spectral clusters that are differentially abundant between the three modes of DOM release are identified based on

statistically significant differences (as determined by DESeq2) in spectral counts between the three experimental treatments. Those with p value $\leqslant 0.01$ (Fig 2.7a) against the null hypothesis of equal abundance between each pair of experimental conditions are considered statistically significant.

A total of 18 clusters (0.3% of the 6781 consistently-observed clusters) were found to be differentially abundant between at least one pair of conditions by these criteria (Table 2.1, Table 2.2). These 18 clusters comprise a total of 18,193 MS2 spectra, or 6.4% of the dataset. We identified four groups of differentially abundant clusters (Fig 2.3). One group (Group I) is enriched in both exudate and viral lysate DOM relative to mechanical lysate. The clusters of Group II were enriched in the viral lysate DOM relative to the other two conditions, and Group III was similarly enriched only in the exudate. Group IV, which comprises only a single cluster, was enriched in the mechanical lysate.

2.2.2 Differentially Abundant Low-MW DOM Components

Fourteen of the differentially abundant clusters, including members of all four Groups, are of relatively low molecular weight (412 Da or less). We sought to identify these compounds by a combination of elemental formula assignments to their precise intact masses and matching of fragmentation patterns to known structures via MetFrag [69], querying against the PubChem and KEGG databases. While candidate elemental formulas could be calculated for each cluster?s intact mass (Table 2.2), none of the fragmentation patterns produced a clear MetFrag match that would indicate a particular molecular structure. Similarly low identification rates using spectral library matching were observed and some potential causes discussed by Petras et al. (2017) [63]. In the absence of specific structural identification, we used elemental formula assignments (allowing for multiple assignments within parent mass uncertainty) in order to suggest at least the broad compound classes that these clusters might represent (Fig 2.4). Notably, most of these clusters' monoisotopic parent masses have a negative mass defect (i.e., a mass slightly less than an integer value), which suggests a

relatively low hydrogen content (since 1H has a positive mass defect of 7.8 mDa) and requires the presence of nuclides with negative mass defects, such as 16O, 32S or 35Cl, and/or adduction of 23Na.

Consistent with this mass defect, the candidate elemental formulas predicted from the observed cluster parent masses tend to be both oxygenated and hydrogen-poor compared to common classes of biochemicals (Fig 2.4). Such relatively unsaturated, oxygen-rich metabolites are known from other phototrophic organisms, notably the flavonoids and phenylpropanoids that are abundant and diverse secondary metabolites [28, 77] and the UV-absorbing scytonemin and microsporines [26]. Cyanobacteria are known to both synthesize [36] and respond to [35, 88] flavonoids. Many of the candidate formulae predicted from the differentially-abundant spectral clusters in this study also contain sulfur and/or chlorine (Fig 2.4, Table 2.2). Sulfurylation could increase the solubility of condensed, aromatised structures, helping to explain their presence in the DOM. The single cluster from Group IV (enriched in mechanical lysate) has a formula match to C17H31SO3, putatively a sulfonated, polyunsaturated C17 lipid; analogous sulfonolipids are known from a variety of aquatic bacteria [14, 2]. Halogenation of polyphenolics is known in a variety of cyanobacteria [19, 30, 39]. Marine *Synechococcus* also produce a variety of halomethanes [38], and WH7803 may well possess as-yet unelucidated pathways for halogenation of secondary metabolites. These results suggest that some of the compounds most characteristic of DOM release from marine *Synechococcus* cells by exudation (Group III), viral lysis (clusters 16453 and 1256, Group II) or both (Group I) could constitute a new (sub)class of unsaturated (likely aromatic), oxygenated and sulfurylated and/or halogenated natural products. Further work will be required to isolate these compounds in sufficient quantity and purity to enable clear structural elucidation.

DOM Type	Sample	MS2 Spectra	Spectral Clusters	Clusters Observed in All 3 Replicates	Differentially Abundant Spectral Clusters (out of Total) per Pairwise DOM Type Comparison		
					Exudate	Mechanical Lysate	Viral Lysate
Exudate	Exudate A	28646	4368	1703		7 / 3412	0 / 5255
	Exudate B	34315	6798				
	Exudate C	27263	4206				
Mechanical Lysate	Mechanical Lysate A	71201	13160	1887			16 / 5429
	Mechanical Lysate B	65552	12345				
	Mechanical Lysate C	62602	9033				
Viral Lysate	Viral Lysate A	75529	16239	3812	Total 18 / 6781		
	Viral Lysate B	73178	15015				
	Viral Lysate C	69099	12848				

Table 2.1: Summary of numbers of MS2 spectra collected from each DOM sample, numbers of clusters those spectra were grouped into, and the numbers of clusters observed in all 3 replicates of each DOM type. At right is a matrix showing how many clusters were determined to be significantly differentially abundant ($p \leqslant 0.01$) in each pairwise comparison of DOM types, with the total number of clusters in each comparison after the slash; note that some clusters were significant in more than one comparison.

2.2.3 High-MW DOM Components Enriched in Viral Lysate

Four of the spectral clusters of Group II (Nos. 89916, 69841, 81454, and 86941) that are enriched in the viral lysate DOM have precursor masses > 1000 Da and are in +2 or +3 charge states; under our electrospray ionization conditions, peptides are a likely source of such high-MW, multiply-charged ions. The peptide sequences represented by each cluster were identified with high confidence by SEQUEST search against the proteomes of WH7803 and S-SM1 (Fig 2.5a, Table 2.4). Remarkably, all four peptides were found to derive from protein components of phycoerythrin. Phycoerythrin is a pigment-protein complex that forms a major part of the light-absorbing antenna of the phycobilisome, which is the principal light-harvesting apparatus for *Synechococcus*. Two classes of phycoerythrin genes (I and II, encoded by the genes cpeAB and mpeAB, respectively) are present in WH7803 [78]. Phycoerythrin consists of heterodimers of two subunits, the α- and β-chains, which are further organized into trimeric ring structures, which stack on the outside of the phycobilisome to form the light-harvesting antenna.

Among the compounds enriched in the viral lysate DOM, one peptide was found from each of the α- and β-chains of each of the two classes of phycoerythrins (I and II) present

in WH7803 (Fig 2.5a). When mapped back to their parent full-length protein sequences, it is clear that these four phycoerythrin peptides are not random fragments, but rather derive from specific portions of the α- and β-chains that are homologous between the two phycoerythrin classes. Furthermore, the eight termini of the peptides are all adjacent to basic amino acid residues (either lysine or arginine), either N-terminal (6/8) or C-terminal (2/8) to the basic site, which suggests they might be produced by the action of an endoprotease with specificity for cleavage at such positions.

The specificity of excision of these peptides from the phycoerythrin complex is reinforced by their proximity in the three-dimensional protein structure (Fig 2.5b). The peptides from the α- and β-chains derive from adjacent portions of a region of contact between the two chains, and which is located on the exterior of the trimeric $(\alpha\beta)3$ phycoerythrin ring structure. One explanation for this striking pattern could be that these peptides are produced by a protease that specifically targets those portions of the phycoerythrin structure in order to initiate phycobilisome degradation or disassembly. Taken together, this evidence suggests that specific proteolytic degradation of phycobilisomes could occur during phage infection of *Synechococcus* WH7803. Whether this degradation is mediated by a host- or phage-encoded protease is not clear at present. Phycobilisome degradation is a well-known and highly coordinated stress response in cyanobacteria, particularly to N starvation [57]. Alternatively, phage may carry their own (as yet unrecognized) protease to facilitate recycling of phycobilisome components into progeny phage proteins. It is also possible that this degradation pattern represents a host defense mechanism in response to phage-mediated increases in phycoerythrin, which have been observed during infection of WH7803 by another phage, P-SM2 [73].

2.3 Discussion

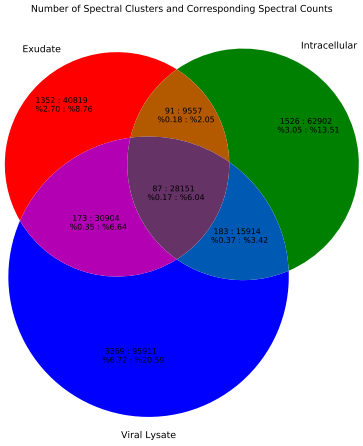
We have shown that the DOM released from biomass of *Synechococcus* WH7803 by each of three different mechanisms – exudation, cell breakage, and viral lysis – is compositionally

distinct. In particular, *Synechococcus* WH7803/S-SM1 viral lysate is enriched in peptide fragments of phycoerythrin, possibly generated during infection as either part of phage-directed remodeling of host photosynthesis, or as a host defense mechanism against phage infection. Other compounds found enriched in exudate and/or viral lysate (as compared to DOM released by mechanical cell breakage) could not yet be clearly identified, but indicated unsaturated, oxygen-rich stoichiometries, some of which were sulfurylated or halogenated and may comprise a new (sub)group of natural products. These compounds that are specifically associated with a particular mode of DOM release from primary producer biomass could in the future serve as markers for the relative importance of DOM generation mechanisms in different parts of the ocean; establishing an index of such markers will require further such studies with a wider array of phytoplankton. The use of molecular networking via GNPS for analyses of natural marine DOM has recently been demonstrated [63]; future developments may see convergence between directed experimental investigations of DOM sources and sinks with characterizations of the in situ oceanic DOM pool employing similar methodologies, thereby enabling a more integrated understanding of the production and processing of this central ocean carbon reservoir.

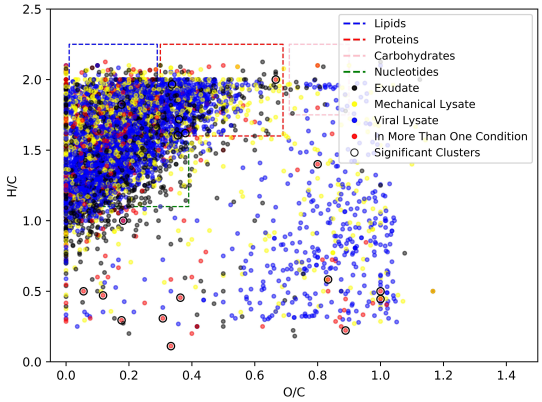
Our finding of the (possibly proteolytic) release of phycobilisome components during infection of *Synechococcus* WH7803 by phage S-SM1 adds a new element to the growing body of evidence for viral influence on the photosynthetic machinery of cyanobacterial hosts. Besides the occurrence of the photosystem II reaction center genes in numerous cyanophage genomes, many cyanophages carry genes related to biosynthesis of bilin chromophores [65]. One of these, CpeT, was recently characterized from a *Prochlorococcus* cyanophage, and may play a role as a bilin lyase and/or a transcription factor during infection [27]. Shan et al. [73] reported that infection of *Synechococcus* WH7803 with phage S-PM2 induced heightened levels of phycoerythrin synthesis; whether this induction is related to the release of phycoerythin components to the dissolved phase during viral lysis remains to be determined. *Synechococcus* WH7803 lacks homologs of the Nbl genes, which are involved in proteolytic

phycobilisome degradation in some freshwater cyanobacteria [16, 6, 58] and have also been found in phages [25, 86], and the molecular mechanisms involved in phycobilisome processing in marine *Synechococcus* remain unclear.

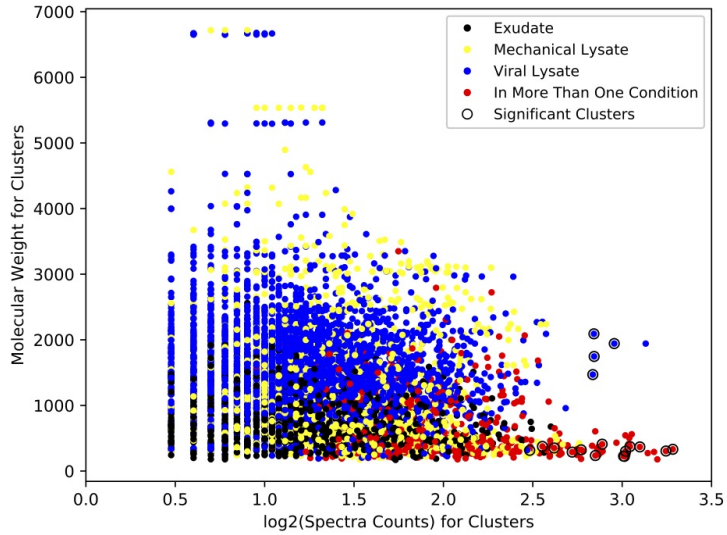
The involvement of phycobilisome processing in phage infection could also have significant consequences for nitrogen biogeochemistry in the surface ocean. Phycoerythrin is among the most abundant proteins in marine *Synechococcus* cells, constituting up to 25% of cellular nitrogen under light-limiting conditions [40]. The specific release of phycobilisome degradation products during viral lysis may represent a substantial input flux of dissolved organic nitrogen (DON) to surface waters. In oligotrophic subtropical gyre regions, DON to dissolved inorganic nitrogen ratios are generally high [45], and DON is key to fueling microbial productivity in surface waters via nutrient recycling. This infection-associated phycobilisome remodeling, then, could be an important mechanism by which lytic aquatic viruses mediate dissolved nutrient regeneration and feed the microbial loop.



(a) Venn diagram of occurrence of spectral clusters (and their member spectra) in DOM produced by the three release mechanisms.



(b) van Krevelen (atomic H:C ratio vs atomic O:C ratio) diagram for best-fit molecular formulas (lowest mass error from observed cluster molecular weight) calculated for clusters in the DOM dataset, colored by their presence in one or multiple DOM types. The 18 clusters that show significant ($p<0.01$) differential abundance between the DOM types are highlighted (see also Fig 2.4). Approximate composition regions for some major classes of biochemicals are indicated.



(c) Cross-plot of cluster molecular weight (i.e., precursor mass) vs log of spectral count for the DOM dataset. The 18 differentially-abundant clusters are highlighted; note that all are of high spectral count, since higher counts yield greater statistical power to detect abundance differences.

Figure 2.2: Distribution, classification and features of compound clusters and their spectra

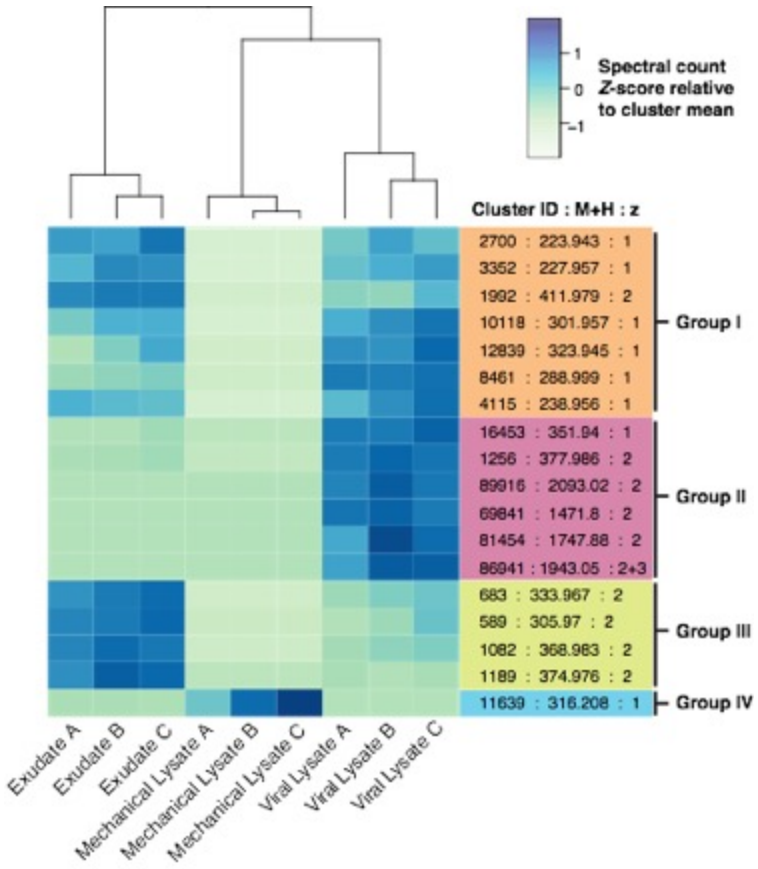


Figure 2.3: Heatmap of abundance of spectral clusters that are significantly ($p<0.01$) differentially abundant between DOM released by the three mechanisms. Heatmap is colored based on spectral counts in a given sample relative to the cluster mean across samples. On the right side, compound mass in expressed as monoisotopic M+H mass and its observed charge state(s) is listed for each spectral cluster. Column dendrogram shows that biological replicates within each DOM type group together. Four groups (I-IV) of differentially abundant spectral clusters are found between experimental treatments.

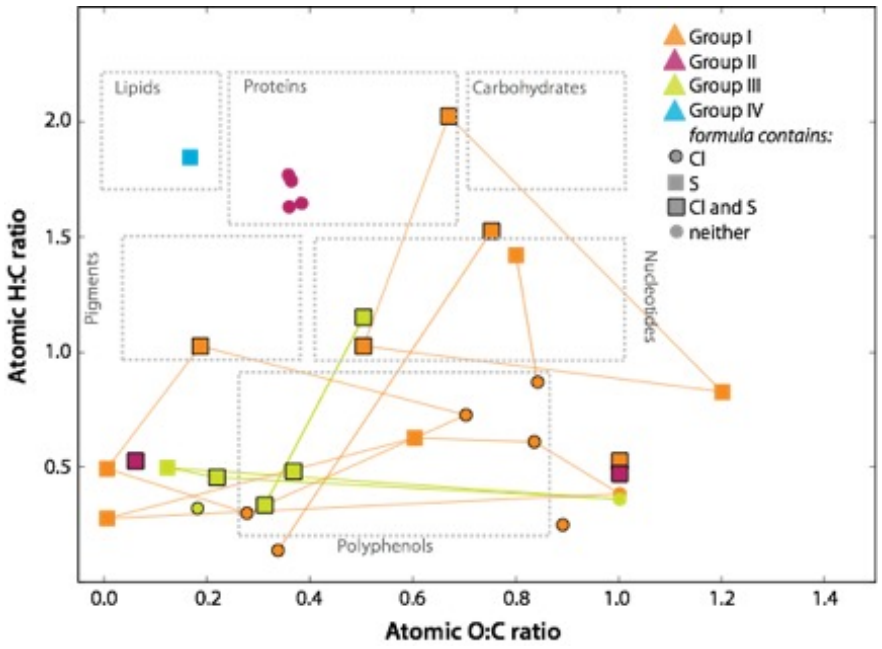
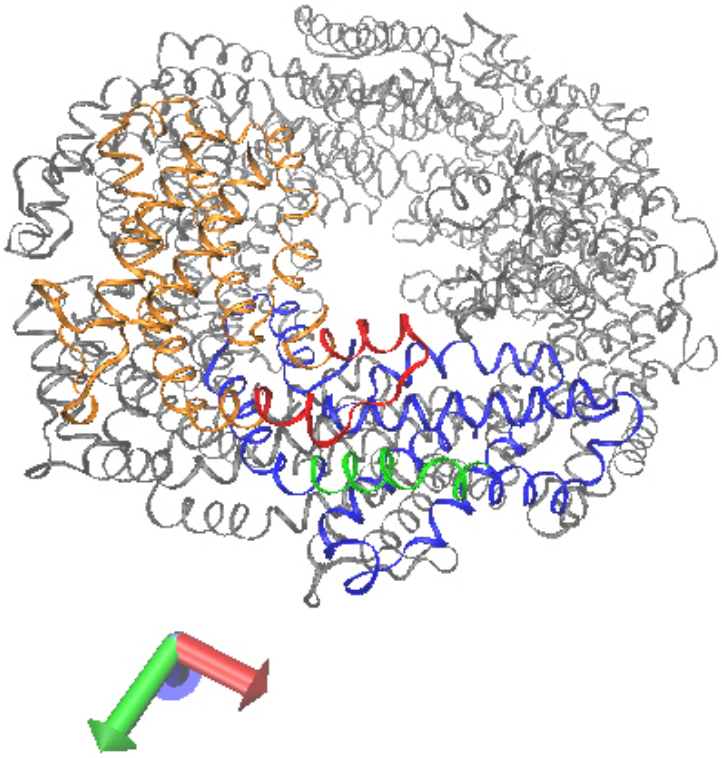


Figure 2.4: van Krevelen (atomic H:C ratio vs atomic O:C ratio) diagram for molecular formulas calculated for low-MW (<412 Da) differentially abundant clusters. For each cluster, up to 4 molecular formulas with mass errors <0.03 Da were calculated, and the resultant atomic ratios plotted as vertices of a polygon. Formulae that contain sulfur and/or chlorine are indicated by symbols. Points are colored by grouping with regard to differential abundance between DOM release modes (see Fig 2.3).

(a) Clusters enriched in viral lysate identified as phycoerythrin peptides. The detected peptides are highlighted within the aligned sequences of the α- and β-chains of the class I and II phycoerythrin genes of *Synechococcus* WH7803.



(b) Location of detected lysate-enriched peptides in the phycoerythrin hexameric ring structure. The detected peptides (highlighted with same coloration as (A)) derive from the exterior of the ring, at the contact interface between the α- and β-chains.

Figure 2.5: Peptide sequences and ring structure of phycoerythrin

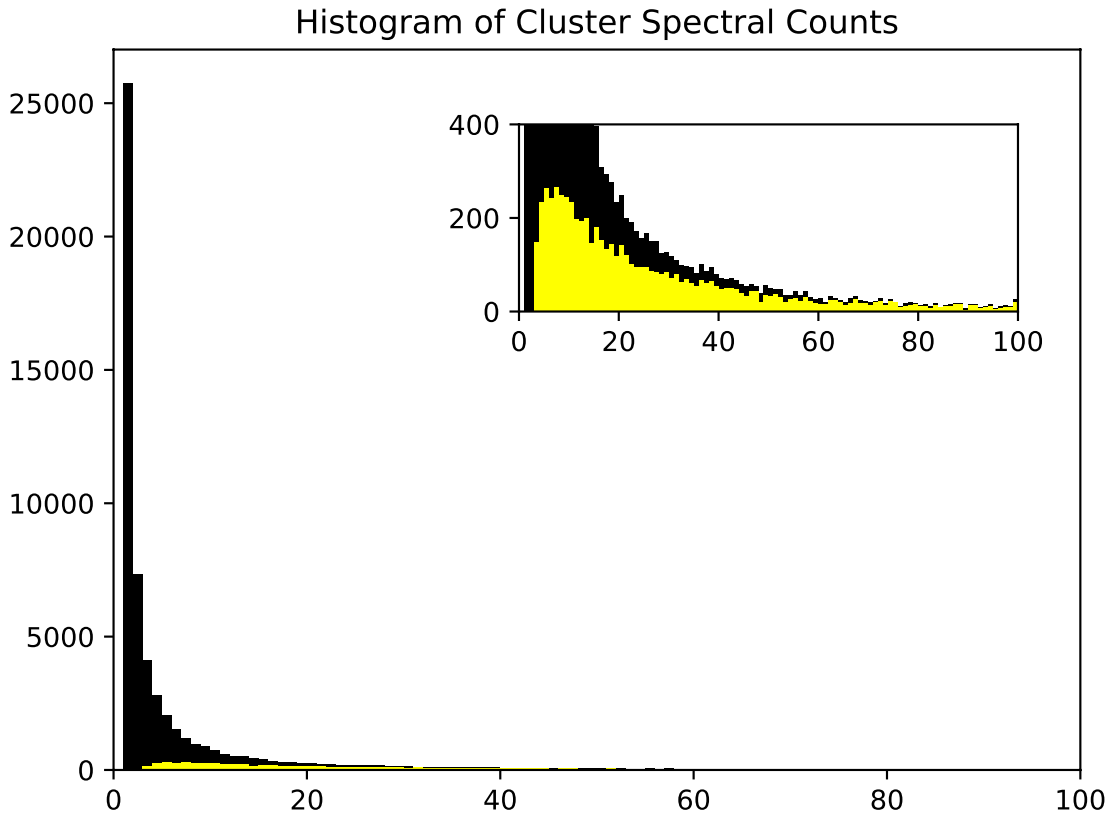
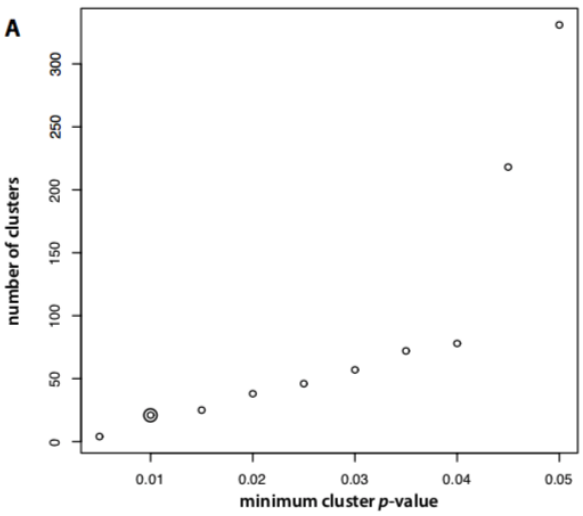
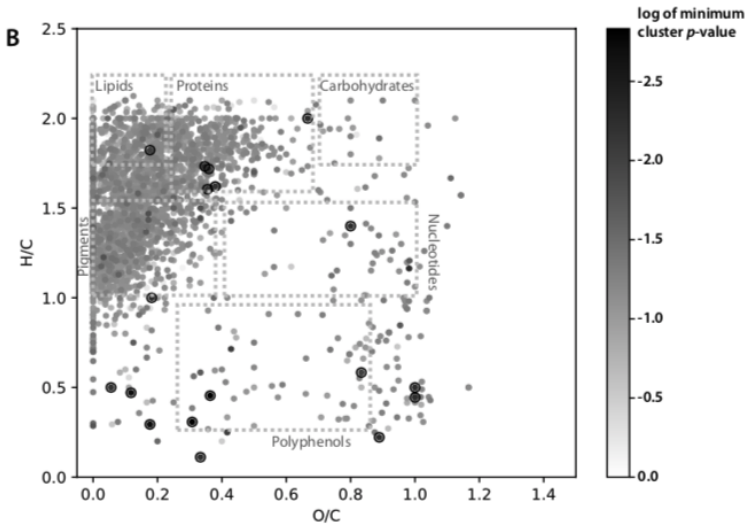


Figure 2.6: Histogram of spectral counts for all 50,114 spectral clusters in the 9-sample dataset (black) after removal of clusters also found in the media blank and of the 6781 clusters observed in all 3 replicates of at least one DOM type (yellow). Inset: enlargement of the 0-100 spectral count region to highlight the spectral count distribution of consistently-observed clusters.



(a) Numbers of clusters found with p-values below a given cuto against the null hypothesis of equal spectral count in at least one pairwise comparison between DOM types. The chosen cutoff ($p \leqslant 0.01$) is highlighted.



(b) van Krevelen (atomic H:C ratio vs atomic O:C ratio) diagram for molecular formulas calculated the 6781-cluster dataset, colored by lowest cluster p-value. Differentially abundant clusters with $p \leqslant 0.01$ are highlighted.

Figure 2.7: Supporting information: p-value plot and van Krevelen diagram

						Sample Spectral Counts																		
Cluster Index	Group	Cluster Total Spectral Count	Precursor M+H+	Precursor Charge	Precursor *m/z*	Exudate A	Exudate B	Exudate C	Mechanical Lysate A	Mechanical Lysate B	Mechanical Lysate C	Viral Lysate A	Viral Lysate B	Viral Lysate C	pVal Ex-ML	pVal Ex-VL	pVal VL-ML	CandidateFormula1	Candidate Formula1 Mass Error	CandidateFormula2	Candidate Formula2 Mass Error	CandidateFormula 3	Candidate Formula3 Mass Error	CandidateFormula4	Candidate Formula4 Mass Error
2700	Group_I	1029	223.943	1	223.943	184	177	233	0	0	0	121	176	138	**0.009**	0.777	**0.005**	C9H1O3P1Cl1	0.012	C4H6O3S1P1Na1Cl1	0.014				
1992	Group_I	773	411.979	2	206.493	162	179	177	0	0	0	72	67	116	**0.009**	0.554	0.011	C20H5N5S3	-0.001	C15H9N1O9S2	0.038	C12H7N3O10Na1Cl1	0.042	C14H5N1O14	0.057
3352	Group_I	1018	227.957	1	227.957	149	201	196	0	0	0	131	161	180	**0.009**	0.855	**0.005**	C5H7O4S3	0.001	C6H5O5Cl2	0.002				
10118	Group_I	1045	301.957	1	301.957	117	162	165	0	0	1	165	201	234	0.030	1.000	**0.005**	C9H2N2O8Cl1	0.033						
12839	Group_I	590	323.945	1	323.945	28	62	104	0	0	0	122	119	155	0.118	0.972	**0.005**	C8H4N2O8S1Cl1	0.033						
8461	Group_I	525	288.999	1	288.999	38	46	56	0	0	0	127	124	134	0.196	0.806	**0.005**	C15H7N1S2Na1	0.001	C11H11N1O2S2Cl1	0.010	C11H3N5O3Cl1	0.013	C10H7N1O7Cl1	0.029
4115	Group_I	708	238.956	1	238.956	110	101	95	0	0	0	101	137	164	0.040	1.000	**0.005**	C3H6N3O2S2Na1Cl1	0.010	C6H6N1O3S1P1Cl1	0.014	C5H4N1O6S2	0.025		
16453	Group_II	416	351.940	1	351.940	9	8	21	0	0	0	119	120	139	0.809	0.217	**0.005**	C9H4N2O9S1Cl1	0.037						
1256	Group_II	358	377.986	2	189.493	15	17	21	0	0	0	96	109	100	0.592	0.430	**0.006**	C18H9N1O1S2Na1Cl1	-0.002						
89916	Group_II	695	2093.020	2	1047.010	0	0	0	0	0	0	206	265	224	1.000	0.034	**0.006**	C87H141N27O33	-0.005						
69841	Group_II	685	1471.800	2	736.400	0	0	0	0	0	0	223	249	213	1.000	0.034	**0.006**	C64H110N16O23	0.000						
81454	Group_II	697	1747.880	2	874.442	0	0	0	0	0	0	151	296	250	1.000	0.034	**0.006**	C76H122N20O27	-0.002						
86941	Group_II	1354	1943.050	2 + 3	972.027	5	0	0	0	0	0	565	893	794	1.000	0.028	**0.005**	C82H143N25O29	-0.006						
683	Group_III	1920	333.967	2	167.487	531	632	667	0	0	0	157	230	263	**0.001**	0.485	**0.005**	C11H5N3O4S1Na1Cl1	0.016						
589	Group_III	1758	305.970	2	153.485	408	442	491	0	0	0	69	116	232	**0.001**	0.343	**0.006**	C13H4N1O4S1Cl1	0.010	C8H9N1O4S2Na1Cl1	0.012				
1082	Group_III	1256	368.983	2	184.995	287	334	315	0	0	0	77	109	134	**0.002**	0.376	**0.008**	C17H8N2O2S3	0.008	C14H6N4O3S1Na1Cl1	0.012	C12H4N2O12	0.050		
1189	Group_III	1105	374.976	2	187.988	352	429	426	0	0	0	136	103	153	**0.002**	0.116	0.032	C17H5N2O3P1Na1Cl1	0.006						
11639	Group_IV	304	316.208	1	316.208	6	5	4	42	108	139	0	0	0	1.000	0.100	**0.009**	C17H31O3S	-0.001						

Table 2.2: Properties of the 18 spectral clusters determined to be significantly differentially abundant between the three DOM types. p-values considered significant ($\leqslant 0.01$) for each pairwise comparison of DOM types are indicated in bold. For each cluster, up to 4 candidate formulas that passed the filtering rules (Table 2.4) were calculated.

Formula property	Acceptable values
C atoms	0-200
H atoms	0-400
N atoms	0-100
O atoms	0-200
S atoms	0-3
P atoms	0-3
Na atoms	0-1
Cl atoms	0-4
H/C ratio	0.33-2.25
O/C ratio	≤1.2
N/C ratio	≤0.5
S/C ratio	≤0.2
P/C ratio	≤0.1
Degree of Unsaturation	Integers
O/P ratio	≥3

Table 2.3: Rules applied to elemental formula calculations based on accurate intact mass measurements. Degree of unsaturation was calculated after Badertscher et al. (2001).

Cluster ID	Confidence	Sequence	Qvality PEP	Qvality q-value	# PSMs in Lysate samples	Master Protein Accessions	Theo. MH+ [Da]	Max XCorr Sequest HT	Percolator q· Value	Percolator PEP	Formula	H:C	O:C
69841	High	LEAAEKLAAGLDAVT	2.05E-04	0.00E+00	304	VIMSS3341049	1471.800	4.56	0.00E+00	2.71E-05	C64H110N16O23	1.72	0.36
81454	High	LEAAEKIAQNYDAIAQ	7.73E-05	0.00E+00	275	VIMSS3341055	1747.886	5.14	0.00E+00	5.91E-05	C76H122N20O27	1.61	0.36
86941	High	RSVVSADAKTAAVGAGDIAAL	3.62E-06	0.00E+00	739	VIMSS3341048	1943.056	6.50	0.00E+00	5.08E-07	C82H143N25O29	1.74	0.35
89916	High	RQAVSADSSGSFIGGAQLNDL	6.49E-08	0.00E+00	617	VIMSS3341056	2093.026	7.56	0.00E+00	1.08E-09	C87H141N27O33	1.62	0.38

Table 2.4: Details of SEQUEST peptide identi cations of 4 high-MW clusters enriched in viral lysate.

2.4 Experimental Procedures

2.4.1 Synechococcus *WH7803 Culture*

Triplicate *Synechococcus* WH7803 cultures in artificial seawater medium (salt mix with 428mM NaCl, 9.8mM MgCl26H2O, 6.7mM KCl, 14.2mM MgSO47H2O, 3.4mM CaCl22H2O, 9.1mM Tris Base, and trace metals of 46.278uM H3BO3, 9.15uM MnCl24H2O, 0.772uM ZnSO47H2O, 0.032uM CuSO45H2O, 0.025uM CoCl26H2O, 1.616uM Na2MoO42H2O, as well as macro nutrients of 8.8mM NaNO3, 2.0mM NH3Cl, 0.13mM NaH3PO4H2O, 5.9mM NaHCO3) were maintained in a Percival incubator on a 24-hour light/dark cycle (14-hour light and 10-hour dark with an average light intensity of around 25 µmEinstein/m2/s). A 200ml media blank sample was extracted as described below for subtraction of media- and sample preparation-associated compound clusters from the experimental dataset.

2.4.2 Exudate Collection and Mechanical Cell Lysis

Cultures were harvested in late-log phase. Each 200ml biological replicate was centrifuged at 4700 × g for 10 minutes to pellet cells. Supernatant was 0.2 µm filtered and the filtrate was then extracted as described below to prepare the exudate DOM samples. Cell pellets were lysed mechanically using an Avestin EmulsiFlex-C3 (1500 psi for 20 cycles) in DOC-free artificial seawater prepared with LC-MS grade water and ashed reagent-grade salts. This mechanically disrupted cell material was centrifuged again to remove unlysed cells and debris, and the supernatant was 0.2 µm filtered then extracted as described below to prepare the mechanical lysate DOM samples.

2.4.3 S-SM1 Viral Lysis

Triplicate WH7803 cultures for viral infection were maintained as above. S-SM1 cyanophage was added at 9.67E+08 per ml in late-log phase. Viral lysates were harvested when the culture was cleared. Each biological replicate was centrifuged at 4700 × g for 10 min. Su-

pernatant was collected, 0.2 µm filtered, and DOM extracted as described below to prepare the viral lysate DOM samples.

2.4.4 DOM Extraction

DOM extraction procedure generally follows [15]. DOM samples were acidified to pH 2 before solid phase extraction. Agilent Bond Elut PPL Cartridges (1gm 6mL) used in the experiment is retentive for non-polar to moderately polar compounds. Cartridges were washed first with 1 column volume of methanol, equilibrated with 1 column volume of ultrapure water, followed by sample loading. After washing with 2 column volumes of 0.01M HCl, DOM compounds attached to the C18 column were eluted with 6 ml methanol. Eluate was then freeze-dried (Centrivap, Labconco) and stored at -80C.

2.4.5 LC-MS Analysis

Samples were re-dissolved with 2% acetonitrile + 0.1% and 6l aliquots were injected onto a trapping column (OptiPak C18, Optimize Technologies) and separated on a capillary C18 column (Thermo Acclaim PepMap 100 A, 2um particles, 50um I.D. × 50cm length) using a water-acetonitrile + 0.1% formic acid gradient (2-50% AcN over 210 min) at 90nl/min using a Dionex Ultimate 3000 LC system. Ionization was by nanoelectrospray (Proxeon Nanospray Flex) in positive mode. Mass spectra were collected on an Orbitrap Elite mass spectrometer (Thermo) operating in a data-dependent acquisition (DDA) mode, with one high-resolution (240,000 m/delta m) MS1 parent ion full scan triggering 15 MS2 CID fragment ion scans of intensity-selected precursors.

2.4.6 Data Processing

Mass spectrometry data in Thermo RAW format were converted to mzXML format with ProteoWizard [41]. mzXML datafiles were uploaded to Global Natural Products Social

Molecular Networking (GNPS) [84]), a platform for clustering and network analysis of mass spectrometry data. Spectral clustering in GNPS based upon similarity cosine scoring of MS2 spectra. MS2 fragmentation patterns that are have sufficiently high cosine scores ($\geqslant 0.7$)), and differences between their parent masses' mass-to-charge ratio is within 100 Th, are grouped into the same spectral cluster. Across different LC-MS runs, the same cluster can be identified and the number of spectra belonging to it counted in multiple samples. We used a Python data pipeline (available at https://github.com/WaldbauerLab/metabolomics) to clean, transform, and store spectral cluster data from GNPS, including merging spectral clusters which have parent masses within mass spectrometer analytical error ($\leqslant 0.005$ Da) and/or which appear to be isotopologues.

2.4.7 Statistical Analysis

To test for differential abundance of spectral clusters between the three types of DOM, we used DESeq2, an R statistical package with Generalized Linear Model at its core [50]), which enables quantification and statistical inference of systematic changes between experimental conditions with a wide variety of discrete data types. Here we applied it for differential analysis of GNPS-clustered mass spectrometric data, including three pair-wise comparisons between our DOM types: Mechanical Lysate vs. Exudate, Mechanical Lysate vs. Viral Lysate, and Viral Lysate vs. Exudate. A strength of DESeq2 for this purpose is its stable estimation of effect sizes and corrections for noisiness coming from small sample size and candidates with large dynamic range (i.e., low counts versus high counts). It enables a more quantitative analysis focused on the strength rather than the mere presence of differential expression [50]. Spectral clusters with p values $\leqslant 0.01$ against the null hypothesis of equal abundance in each pairwise condition comparison are considered significantly differentially abundant. A total of 18 statistically significant spectral clusters differentially expressed between the three modes of DOM production are found (Table 2.1, Table 2.2).

2.4.8 Compound Identification

Compound identification of the 18 differentially abundant spectral clusters was explored using multiple approches. Two methods, brute-force stoichiometric calculation and MS2 fragmentation-pattern matching, were attempted for the identification of spectral clusters with smaller masses and lower charge state (Figure 3). The brute-force stoichiometric calculation based on monoisotopic mass of spectral clusters was done via ChemCalc [61]. Monoisotopic masses of these spectral clusters were queried through ChemCalc web service with criteria constraining formula results to be chemically plausible (Table 2.3). MetFrag [69] draws known compounds from molecular structure databases, e.g., PubChem and KEGG, fragments them in silico, and compares observed MS2 spectra with predicted fragment masses. To identify spectral clusters with peptide-like features (i.e., high molecular weight and charge state; Fig 2.3) the raw MS data was searched against a protein database comprising the *Synechococcus* WH7803 and phage S-SM1 translated genomes using SEQUEST HT implemented in Proteome Discoverer (Thermo Scientific), using unspecific protein cleavage and controlling peptide- and protein-level FDRs to 0.01 using Percolator.

CHAPTER 3

DOM COMPOSITIONS OF VIRAL LYSATE, INTRACELLULAR METABOLITES AND EXUDATE FROM *PROCHLOROCOCCUS*: LAB CULTURE EXPERIMENT

3.1 Overview

Prochlorococcus and marine *Synechococcus* are considered to have diverged from a common ancestor ca. 150 million years ago [17], but the majority of the two groups' members would be regarded as the same species based on their 16S rRNA sequences [9]. While they still share many phenotypic and ecological traits, differences in cell size and photosynthetic pigments yield distinct flow cytometry profiles. Although the phylogenies of many individual gene families cluster low-light (LL)-adapted *Prochlorococcus* more closely with *Synechococcus* than with high-light (HL)-adapted *Prochlorococcus* [67, 87], whole-genome phylogenies clearly separate *Prochlorococcus* from *Synechococcus* [47]. Although some data question the distinction between these two genera, there are several physiological and ecological factors supporting their separation [9].

The geographic distributions of *Prochlorococcus* and *Synechococcus* also suggest the forces that mediate their niche partitioning. *Synechococcus* lives in almost all marine environments, whereas *Prochlorococcus* is confined to warmer, oligotrophic oceans, such as subtropical gyres and the eastern Mediterranean Sea, and is absent from colder, nutrient-rich waters at high latitude as well as in most nutrient-rich coastal waters. A few pieces of evidence could explain these differences. *Synechococcus* can tune its phycobilisome antenna systems to acclimate to changing temperatures, which may contribute to its greater geographical range [52, 64]. It is also less susceptible to copper toxicity than *Prochlorococcus* [53], which might partially explain its dominance in coastal waters. Moreover, *Synechococcus* strains have higher maximum growth rates [55] than *Prochlorococcus* and they are prey for many of the

same predators.

Thus, it is worth investigating *Prochlorococcus* DOM separately. Using a similar approach as for *Synechococcus* WH7803, we investigated DOM compositions from another major strain of cyanobacteria – *Prochlorococcus* MED4.

3.2 Results

This set of experiment is very similar to the one done using *Synechococcus* WH7803 strain (See Experimental Design in Figure 2.1 on Page 16). Using the same broad, untargeted analytical approach, between 46,000-54,000 MS2 spectra were collected from triplicate samples of intracellular, extracellular and lysate DOM (Table 3.1). The data were also organized into three groups: Exudate, Viral Lysate and Mechanical Lysate. Data of three biological replicates for each group look consistent from the Principle Component Analysis (Figure 3.1).

DOM Type	Sample	MS2 Spectra	Spectral Clusters	Clusters Observed in All 3 Replicates	Differentially Abundant Spectral Clusters (out of Total) per Pairwise DOM Type Comparison		
					Exudate	Mechanical Lysate	Viral Lysate
Exudate	Exudate A	49262	4280	2318		4 / 241	1 / 383
	Exudate B	50400	4481				
	Exudate C	56288	5265				
Mechanical Lysate	Mechanical Lysate A	60617	10492	2631	Total 13/3856		9 / 116
	Mechanical Lysate B	60979	8700				
	Mechanical Lysate C	52227	6285				
Viral Lysate	Viral Lysate A	57176	6530	1608			
	Viral Lysate B	56038	6056				
	Viral Lysate C	62011	12164				

Table 3.1: *Prochlorococcus* MED4 Samples - Summary of MS2 spectral and spectral cluster counts from different DOM sources

The entire 9-sample dataset comprises 455,736 MS2 spectra, which were organized by GNPS into 32,733 clusters. We restricted further analyses to clusters that were observed in all three replicates of at least one condition; there were 4739 such clusters (14.5% of the total), which included 306,536 spectra (67.3% of the total). When the 9-sample of MED4 was analyzed together with the Media Blank in GNPS, there were 6053 clusters

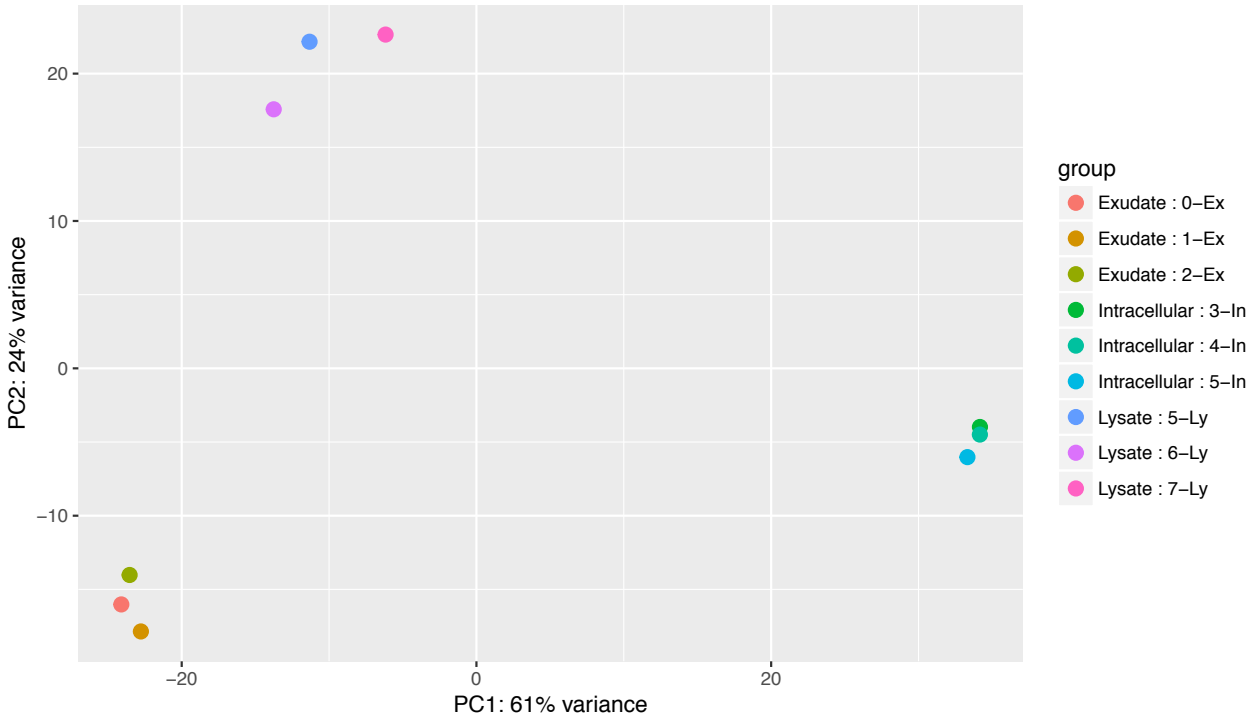


Figure 3.1: Principle Component Analysis of DOM Samples from *Prochlorococcus* MED4

and 41,668 MS2 spectra that appeared in the Media Blank. 886 clusters out of the 4736 subset of MED4 data were found among the 6053 clusters that appeared in the Media Blank. These 886 clusters were subtracted from the MED4 dataset. Therefore, the 9-sample MED4 dataset has 3856 clusters and 221,912 MS2 spectra with Media Blank subtracted. Of the 3856 consistently-observed clusters, 283 (comprising 66,182 spectra) were present in all 9 samples, while between 501 and 1394 clusters were unique to each of the DOM types, and generally smaller numbers were seen in 2 of 3 conditions (Fig 3.2). The venn diagram gives an overview of the distribution of spectral clusters and spectra among samples. A spectral cluster's existence in one experimental conditions is defined to be that it has at least one spectrum found in each of its three replicates. It shows that 1) a small percentage of the spectral clusters (ca.%0.56) contains a significant number of the MS2 spectra (ca.%14.21) (the grey purple region in the center which has the highest spectra percentage to cluster percentage ratio); 2) a significant number of spectra (ca.%14.21) are shared in all three

replicates in all three experimental conditions.

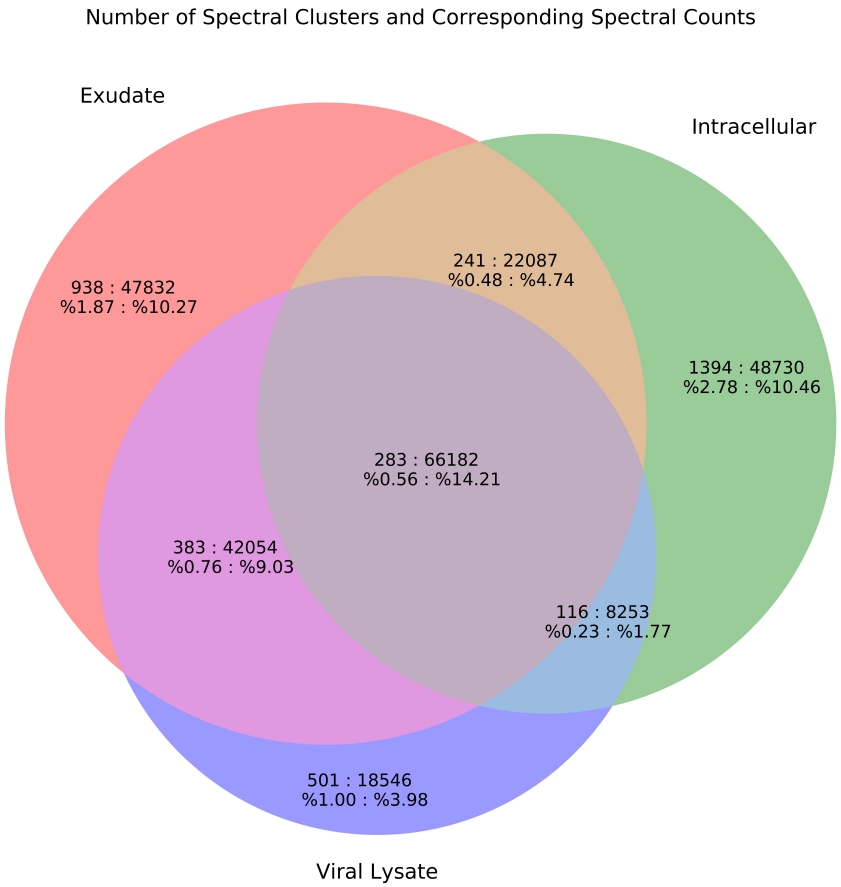


Figure 3.2: Venn Diagram of Spectral Clusters and Spectra Counts. The first number in each region is the number of spectral clusters while the second being their corresponding number of spectra. The two numbers below the counts are their respective percentage of the total number of spectral clusters and total number of spectra.

Principle Component Analysis (PCA) (Fig 3.1) confirms the reproducibility of replicates in each sample type: three biological replicates for each type of DOM, i.e., Exudate, Viral Lysate and Mechanical Lysate, cluster together in the PCA diagram. A histogram of cluster

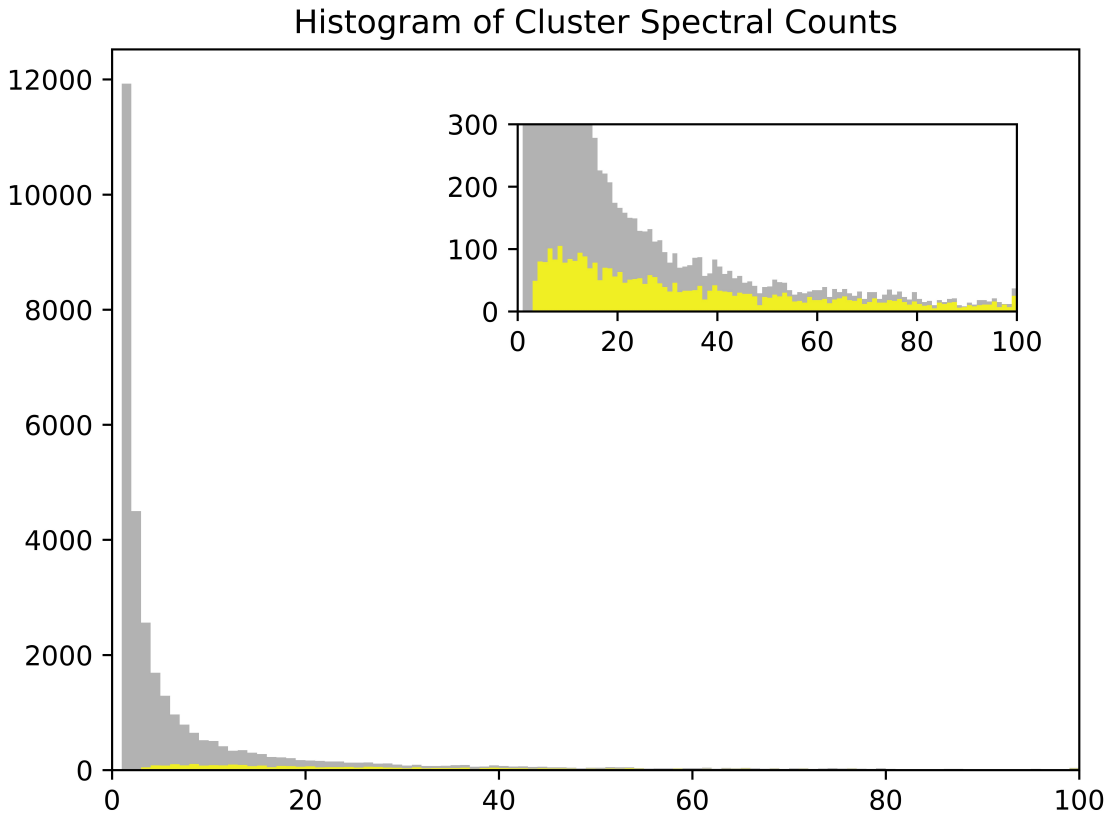
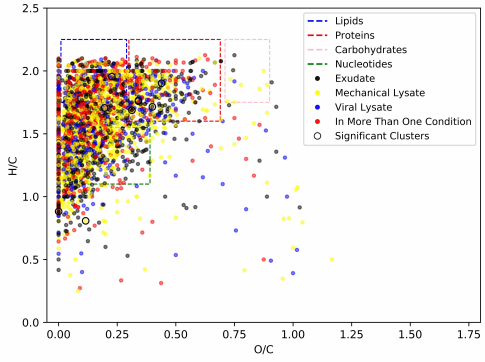


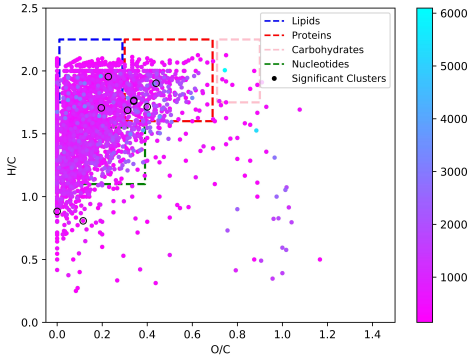
Figure 3.3: Histogram of Spectra Counts. Grey: spectra counts for the entire 9-sample dataset (32,733 clusters and 455,736 MS2 spectra); Yellow: a subset of the data where clusters were observed in all three replicates of at least one condition and Media Blank has been substracted (3856 clusters and 221,912 MS2 spectra)

spectra counts (Fig 3.1) suggests that when there are about more than 90 spectra counts for each cluster, the particular cluster occurred in all the replicates for at least one experimental condition. Elemental formula calculation is attempted for each compound cluster (Figures 3.4a, 3.4b, and 3.4c). They suggest that most compounds from the samples belong to lipids, proteins, and nucleotides groups. Each of the 3 elemental ratio plots is visualized based on other information about the spectral cluster or its corresponding elemental formula. Most candidate formulae (which represent spectral clusters' compound candidates) fall into the upper left corner of the diagram while Mechanical Lysate and Viral Lysate have a few dozen candidate formulae in the lower right region (high O/C ratio and H/C ratio) (Fig 3.4a).

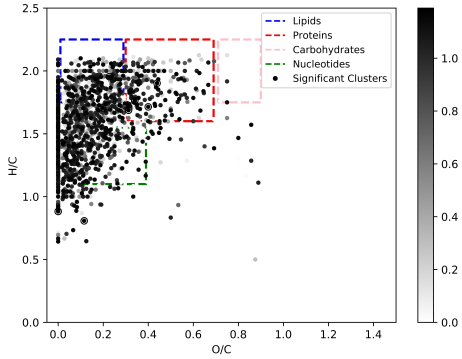
Formulae with higher molecular weights mainly concentrate in the proteins and nucleotides regions (Fig 3.4b), whereas those with high P values from the pair-wise comparisons among three conditions mostly concentrate on the upper left corner of the diagram (Fig 3.4c) – the region where majority of the formulae (or spectral clusters) are. The spectra counts vs molecular weight (Fig 3.4d) plot shows that clusters represented by higher spectra counts have a molecular weight around 500~1000Da.



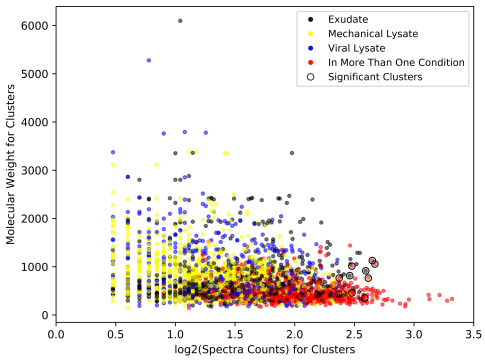
(a) Colored by Experimental Conditions



(b) Colored by Molecular Weight



(c) Colored by abs(log10(p value))



(d) log10(Spectra Count) vs Molecular Weight

Figure 3.4: O/C vs H/C for Spectral Clusters Found in at Least One Experimental Condition in MED4 samples

Elemental ratio plots (van Krevlen diagram). Each elemental formula is calculated using ChemCalc web API [61].

A total of 13 clusters (out of 32,733 total unique clusters in the dataset) were found to be differentially abundant in the pair-wise comparison of Exudate, Viral Lysate and Mechanical Lysate samples. 4 of them (clusters 7228, 8740, 10427, and 43487) were found in the Media

Blank samples but with only cluster 10427 having more than 7 spectra found in the Media Blank. Therefore, only cluster 10427 was removed from the significant clusters. Among the rest of the 12 significant clusters, 2 of them (cluster IDs 7228 and 1411) didn't get a formula through Chemcalc query. These 12 clusters comprise a total of 4413 MS2 spectra, or 0.97% of the total 455,736 spectra in the 9-sample dataset. Three groups of differentially abundant clusters have been identified, based on comparison of the DOM produced by the 3 release mechanisms (Figure 3.5). One group (Group I) is enriched in Exudate DOM relative to Mechanical Lysate and Viral Lysate. The clusters of Group II were enriched in the Mechanical Lysate relative to the other two conditions, and Group III which comprises only a single cluster, was enriched in the Viral Lysate samples.

Column dendrogram shows that biological replicates within each DOM group well together. Differentially abundant spectral clusters are found between DOM groups. Group I: spectral clusters differentially abundant in Exudate generally have higher molecular weight and higher charge states. Group II: spectral clusters enriched in Mechanical Lysate are generally smaller in size and lower in charge state. Group III: only one statistically significant spectral cluster that's more abundant in Viral Lysate and Exudate.

Similar compound identification procedures were taken as we did for *Synechococcus* WH7803 (See Subsection 2.4.8 in Page 36). Elemental ratio plots of small and low-charge compounds show that are differentially abundant (Figure 3.5). The brutal-force stoichiometric calculation based on monoisotopic mass of spectral clusters was done through ChemCalc's web service[61]. See details in Section 2.4.8 on Page 36

3.2.1 Differentially Abundant Low-MW DOM Components

Five of the differentially abundant clusters, including members of Group II and III, are of relatively low molecular weight (431 Da or less). The candidate elemental formulas predicted from the observed cluster parent masses in Group II and III belong to the broad categories of nucleotides and proteins (Fig 3.6). Further work will be required to isolate these compounds

in sufficient quantity and purity to enable clear structural elucidation.

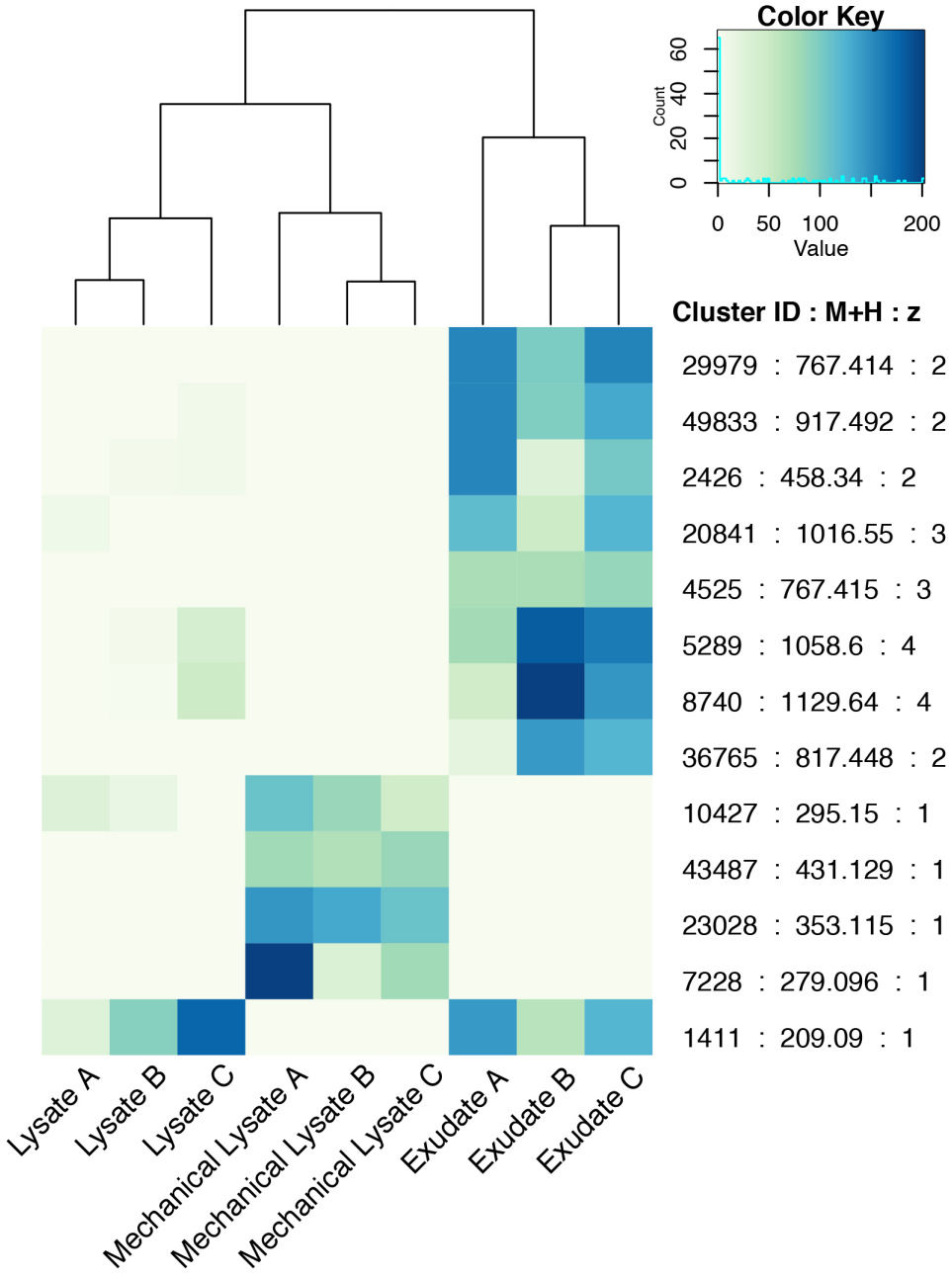


Figure 3.5: *Prochlorococcus* MED4 Samples - Distinct DOM Compositions Between 3 Modes Revealed By Statistical Analysis

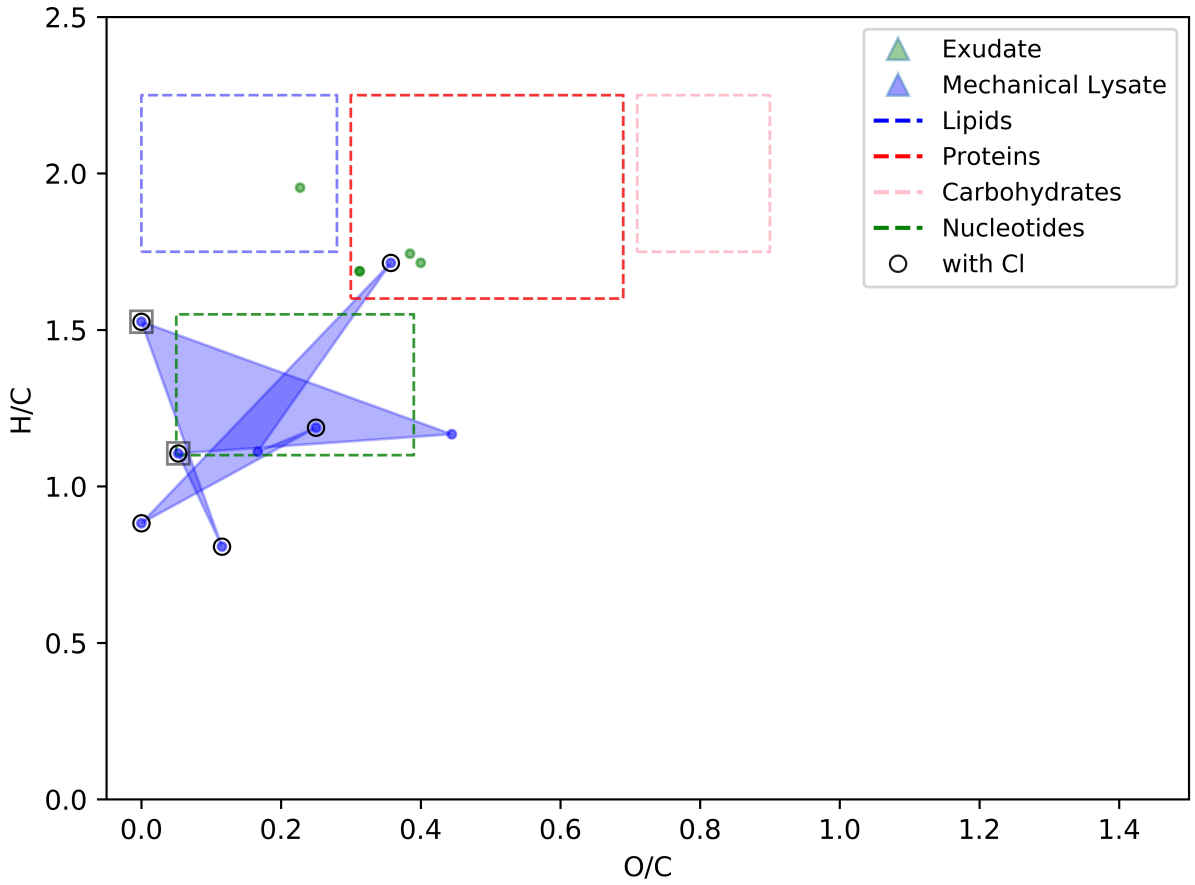


Figure 3.6: O/C vs H/C for Compound Candidate Formulae in MED4 samples

3.2.2 Differentially Abundant Peptides

Clusters 29979 and 4525 were identified by SEQUEST as the peptide GLGKNAAH (M+H+ 767.416, in the 2+ and 3+ charge states, respectively), derived from hypothetical protein PMM1011. This small (81-residue) protein, of whose function nothing is known, appears unique to *Prochlorococcus*.

Cluster 49833 was identified by SEQUEST as the peptide SESKAPVVT (M+H+ 917.494), derived from PsbE, the alpha-subunit of cytochrome b559. Clusters 5289, 8740 and 36765 were identified by FindPept (against exact parent ion mass) and MS-Product (against fragment masses) as the peptide AAEELKEVIK (M+H+ 1129.646) and its subsequences AEELKEVIK (M+H+ 1058.609) and AEELKEV (M+H+ 817.430). These peptides derive

from the thylakoid-associated protease FtsH4.

Intriguingly, both FtsH4 and cytochrome b559 are involved in the assembly-photoprotection-repair cycle of Photosystem II in cyanobacteria [71, 13]. Cytochrome b559 is an internal electron pathway in PSII, allowing for de-excitation of the acceptor-side quinone sites via cyclic electron flow or a water-water cycle when a membrane plastoquinone is not available, thereby preventing photoinhibition [13]. FtsH4 is the least well-characterized of the ATP-dependent FtsH zinc metalloprotease family, of which other homologs (FtsH1-3) are involved in repair of photodamaged D1 protein. A Synechocystis knockout of FtsH4 generated no discerable phenotype [54], though it is clearly localized to the thylakoid periphery [71]. Uniquely among the Fts proteases, FtsH4 appears to interact with the KaiC circadian clock protein and the PstS periplasmic phosphate-binding protein (Yu PhD thesis 2013 https://spiral.imperial.ac.uk/handle/10044/1/28957). In *Prochlorococcus*, FtsH4 expression responds to N- and P-starvation, and to DCMU and DBMIB treatments that affect the redox state of the thylakoid plastoquinone pool (Gomez-Baena et al 2009 Res Micro), as well as co-culture with heterotrophic bacteria [1]. Exactly why particular peptide fragments of these two proteins should be found in MED4 exudates is not clear.

3.3 Discussion

DOM released from biomass of *Prochlorococcus* MED4 and phage PHM2 system by each of three different mechanisms – exudation, cell breakage, and viral lysis – is compositionally distinct. In particular, the distinction is more pronounced between the Exudate and the Mechanical Lysate. The exudate appeared to have preferentially abundant peptides related to the host photo synthesis systems. Other compounds found enriched in exudate and/or viral lysate (as compared to DOM released by mechanical cell breakage) could not yet be clearly identified, but indicated nucleotide-like products. Different from that of *Synechococcus* WH7803 lab culture experiment, *Prochlorococcus* and PHM2 host-viral system did not exhibit strong distinctions between the viral lysate and other DOM types. This may be due

to the fact that the two systems are different in terms of their host and viral genomes.

3.4 Materials and Methods

Besides using a different strain of cyanobacteria – *Prochlorococcus* MED4, the methods are similar to those for the *Synechococcus* WH7803 experiment. See Section 2.4 on page 33 for details.

CHAPTER 4

DOM COMPOSITIONS OF VIRAL LYSATE, INTRACELLULAR METABOLITES AND EXUDATE FROM PICOPHYTOPLANKTON: FIELD INCUBATION EXPERIMENT

4.1 Overview

Our two controlled lab culture experiments on *Synechococcus* WH7803 and phage S-SM1, as well as *Prochlorococcus* MED4 and phage PHM2 have been good at separating factors such as cyanobacteria host and phage which both play a role in the DOM production process. However, the real oceanic environment is more complicated. Not only are there other autotrophs and heterotrophs in the microbial network, but also a complex media exist in microbial food web in the surface ocean. A field incubation experiment is thus designed to understand the nutrient-virus-phytoplankton interactions. One challenge of this microcosm incubation experiment is to separate the effect of viral infection from that of zooplankton grazing (which is simulated by mechanical lysis in the lab culture experiments) on the mortality of phytoplankton, because both these two top-down controls can affect the primary productivity and potentially DOM compositions, as we have shown in the lab culture setting. Therefore, at the core of this larger scale field incubation experiment is a dilution assay aiming to partition phytoplankton mortality into virus- versus grazing-induced fractions [42]. By mixing whole seawater (80um-nitex-mesh filtered) with grazer (cell) - free seawater (< 0.45 um filtrate from pre-filtered whole seawater) or virus-free seawater (<100 kDa filtrate from cell-free seawater) at different proportions, a pressure gradient of viral infection or grazing activity can be established and thus the two effects can be evaluated individually . Moreover, for each dilution series, two nutrient conditions (with or without phosphorus addition) were created. Thus, at the end of incubation (48 hours in this case), the compositions of

DOM produced under each seawater and nutrient condition can be compared. DOM samples from 0 hour – the starting point were also collected for comparison. This field microcosm experiment encompassed a wide set of measurements and analyses, such as flow-cytometry, genomics, proteomics and metabolomics or dissolved organic matter analysis.

4.2 Materials and Methods

4.2.1 Field Experiment and Sampling

The field experiment was carried out in Bermuda during September 12-18th, 2014, when surface oceans presumably reaches a maximum nutrient deplete condition due to stratification. Thus, there is a natural phosphorus-deplete temporary environment. Adding P to different seawater types can easily create a phosphorus-replete condition. 21 4L bottles were setup for 48-hour mesocosm incubation experiment with three seawater conditions and two nutrient environments. DOM composition from mesocosm incubations using seawater as media is compared to show difference between DOM generated by different microbial communities. Three incubations in parallel will be diluted with the following types of seawater:

- Whole seawater (W): undiluted seawater pre-filtered through 80um nitex mesh; it's supposed to contain grazers, picophytoplankton and phages
- Cell (grazer)-free seawater (CF): $<$ 0.45 um filtrate from pre-filtered whole seawater; it's supposed to contain phages but neither grazers nor picophytoplankton
- 100 kDa cut-off virus-free seawater (VF): $<$100 kDa filtrate from cell-free seawater. it's supposed to contain contain only dissolved organic compounds, but no grazers, picophytoplankton or phages

Three seawater conditions are:

- 100% whole seawater;

- 20% whole seawater mixed with 80% cell-free water (CF);
- 20% whole seawater mixed with 80% 100 kDa cut-off virus-free water (VF).

For the 48-hour incubation, each water condition has two nutrient conditions: normal-P and plus-P. Phosphorus is added as NaH2PO4-H2O to achieve a final concentration of 30 nM in the plus-P condition. In total, there are 9 groups with 21 samples (See Table 4.1 on page 52).

4.2.2 DOM Sample Preparation at the Field

19 bottles from the initial 21 bottles were recovered for sample processing. 3L water sample in each bottle was used for DOM extraction. The key method is Solid Phase Extraction (SPE) - a widely used technique for extracting dissolved organic matter from the seawater (Dittmar, et al 2008; Kujawinski, et al 2009). Thermo Scientific HyperSep C18 Cartridges used in the experiment feature a highly retentive alkyl-bonded phase for nonpolar to moderately polar compounds. The hydrophobic reversed phase material is retentive for most organic analytes from aqueous matrices. This is ideal for capturing a wide range of DOM from seawater.

A general procedure is described below:

1. Filter the incubation water using 0.2 um Milipore;
2. Acidify filtrate to $PH \approx 2$;
3. Rinse C18 SPE cartridge with 6ml methanol;
4. Re-equilibrate SPE cartridge with 6ml MiliQ;
5. Run the 3L water sample through the cartridge;
6. Desalt using 12ml 0.01M HCl;
7. Store cartridges at 4 degree C.

SPE cartridges were transported in 4 degree C cooler back from Bermuda to Jacob Waldbauer's Biogeochemistry Lab at the University of Chicago. To recover DOM retained on the C18 cartridges and further measure on the instrument, please see 2.4 on page 33.

4.3 Results

One of the main goals of the Bermuda field incubation experiment is to to determine compositional distinctions of natural viral-host-nutrient environment as compared to an environment with reduced viral and grazing activity. Unlike the lab culture experiments we did for *Synechococcus* and *Prochlorococcus*, where sources of DOM were designed to be different, the Bermuda field microcosm incubation experiment had more controlled features, such as phosphorus concentration, time duration of incubation, different mixtures of water types (created by the size-fraction-dependent filtration).

Flow-cytometry results showed that 7 of the samples from Condition 1, 6 and 7 (where 20% whole seawater plus 80% cell-free water is involved) have been contaminated. These 7 samples are therefore excluded from further DOM and GNPS analysis. All the rest 12 samples from Conditions 2, 3, 4, 5, 8, and 9, were processed for LC-MS measurement of their DOM compositions. Using a broad, untargeted analytical approach, between 31,000-38,000 MS2 spectra were collected from incubation water DOM samples from different time points and phosphorus concentrations (Table 4.1). The number of MS spectra and that of spectral clusters (~4000-5000) in each sample varies within about 20% (Table 4.2).

Samples are grouped into 9 experimental conditions for analysis.

- Condition 1 (1 biological replicate): Time-0, 20% whole seawater plus 80% cell-free water
- Condition 2 (1 biological replicate): Time-0, 20% whole seawater plus 80% 100 kDa cut-off viral-free water
- Condition 3 (1 biological replicate): Time-0, 100% whole seawater

Treatment	Collection (hrs)	Replicates	Volume(L)
W	0	1	4 ea.
CF80	0	1	
VF80	0	1	
W	48	3 (2)	
W+P	48	3 (2)	
CF80	48	3	
CF80+P	48	3	
VF80	48	3	
VF80+P	48	3	

Table 4.1: Experimental design

- Condition 4 (2 biological replicates): Time-48hr, 100% whole seawater
- Condition 5 (2 biological replicates): Time-48hr, 100% whole seawater with phosphorus addition
- Condition 6 (3 biological replicates): Time-48hr, 20% whole seawater plus 80% cell-free water
- Condition 7 (3 biological replicates): Time-48hr, 20% whole seawater plus 80% cell-free water with phosphorus addition
- Condition 8 (3 biological replicates): Time-48hr, 20% whole seawater plus 80% 100 kDa cut-off viral-free water
- Condition 9 (3 biological replicates): Time-48hr, 20% whole seawater plus 80% 100 kDa cut-off viral-free water with phosphorus addition

Clustering of 16S RNA data suggest general consistency between biological replicates of incubation. At T 48 hours, samples from CF80 and VF80 water types (which have different

size fraction) deviated from each other. For the same size fraction (VF80), sample sets with phosphorus addition are quite different from those without. These indicate that nutrients and phages have an impact on the microbial community structure.

The entire 12-sample dataset comprises 408,885 MS2 spectra, which were organized by GNPS into 12,995 clusters. First, statistical hypothesis testing was done using the Generalized Linear Model from R package DESeq2 in order to find clusters that are differentially abundant between different subsets of experimental conditions (e.g., water types, phosphorus conditions, time period of incubation, etc.). Second, we did compound identification on the clusters. We restricted further identification analyses to clusters that were observed in each replicate of at least one of the W48, W48+P, V48, or V48+P conditions; there were 7475 such clusters (57.5% of the total), which constituted 390,242 spectra (95.44% of the total). The clusters excluded from downstream analysis due to inconsistent occurrence in the full dataset were nearly all of low spectral count (fewer than 15 total spectra per cluster) . Of the 7475 consistently-observed clusters, 1000 (comprising 321,513 spectra) were present in all 12 samples.

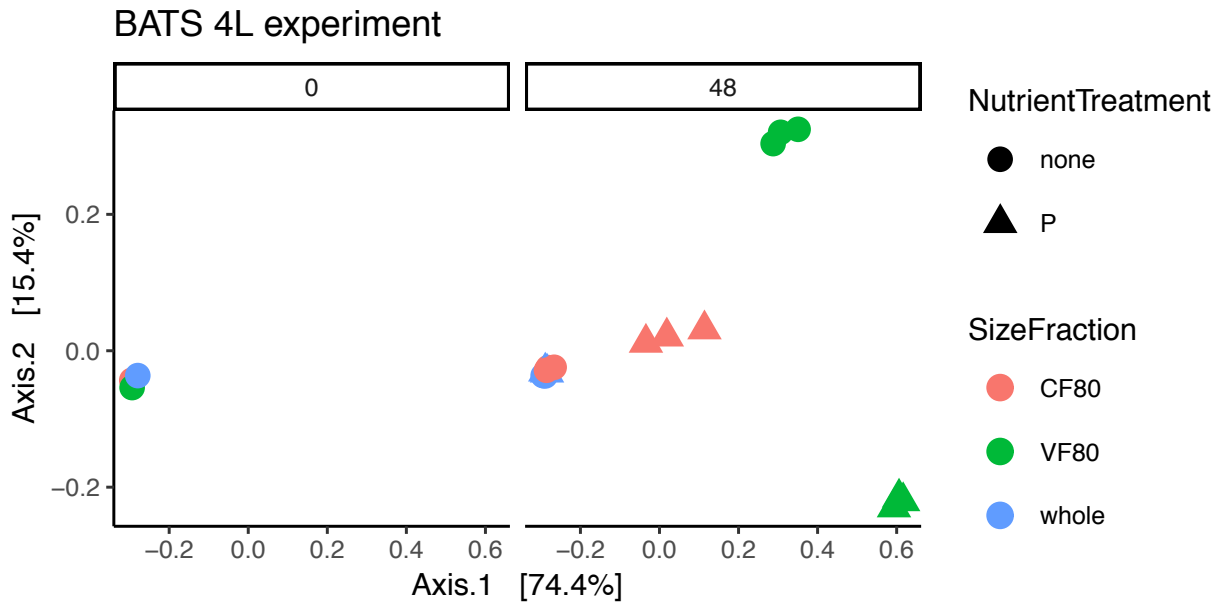


Figure 4.1: BATS Field Incubation Samples - 16S RNA data for water types with different size fraction and nutrient treatment

4.4 Discussion

Sample ID	Sample Description	MS2 Spectra	Spectral Clusters	Significant Spectral Clusters (P<=0.01)
BATS_T0_(72)	Whole water T0	42594	4999	35 (when the red and green groups are compared)
BATS_T0_(74)	80% Viral-free water T0	39324	4778	
BATS_T48_68	Whole water T48	34824	4608	
BATS_T48_69	Whole water T48	38898	3986	
BATS_T48_71	Whole water +P T48	37452	4631	
BATS_T48_72	Whole water +p T48	35142	4446	
BATS_T48_79	80% Viral-free water T48	34015	4558	
BATS_T48_80	80% Viral-free water T48	38131	4849	
BATS_T48_81	80% Viral-free water T48	39419	4789	
BATS_T48_82	80% Viral-free water +P T48	36246	4886	
BATS_T48_83	80% Viral-free water +P T48	36802	4819	
BATS_T48_84	80% Viral-free water +P T48	38632	4924	

Table 4.2: BATS Samples - Summary of MS2 spectral and spectral cluster counts from different DOM sources

Overall, spectra counts data for each of the 12 samples analyzed through GNPS platform indicate that the amount of spectra and spectral clusters are relatively consistent across all the replicates and on the same order of magnitude respectively (Fig 4.2). A histogram of cluster spectra counts suggests that when there are about more than 10 spectra counts for each cluster, the particular cluster occurred in all the replicates for at least one experimental condition. Note that some experimental conditions only have one biological replicates. (Fig 4.2)

Two van Krevlen diagrams (elemental ratio – O/C vs H/C plots) for the subset of spectral clusters which appeared in at least one of the 9 experimental conditions described on Page 51 also showed some broad patterns of compound candidates inferred from the molecular weight of compound clusters detected. Additional features were added on top of the elemental ratio plots: one being molecular weight while the other being the type of hetero-atoms (N, S and/or P). Figure 4.3a and 4.3b suggests that the samples mostly contain compounds that belong to the lipids, proteins, and nucleotides groups. Formulae with higher molecular weights were more likely to occur in the proteins, nucleotides, and lipids region (Fig 4.3a).

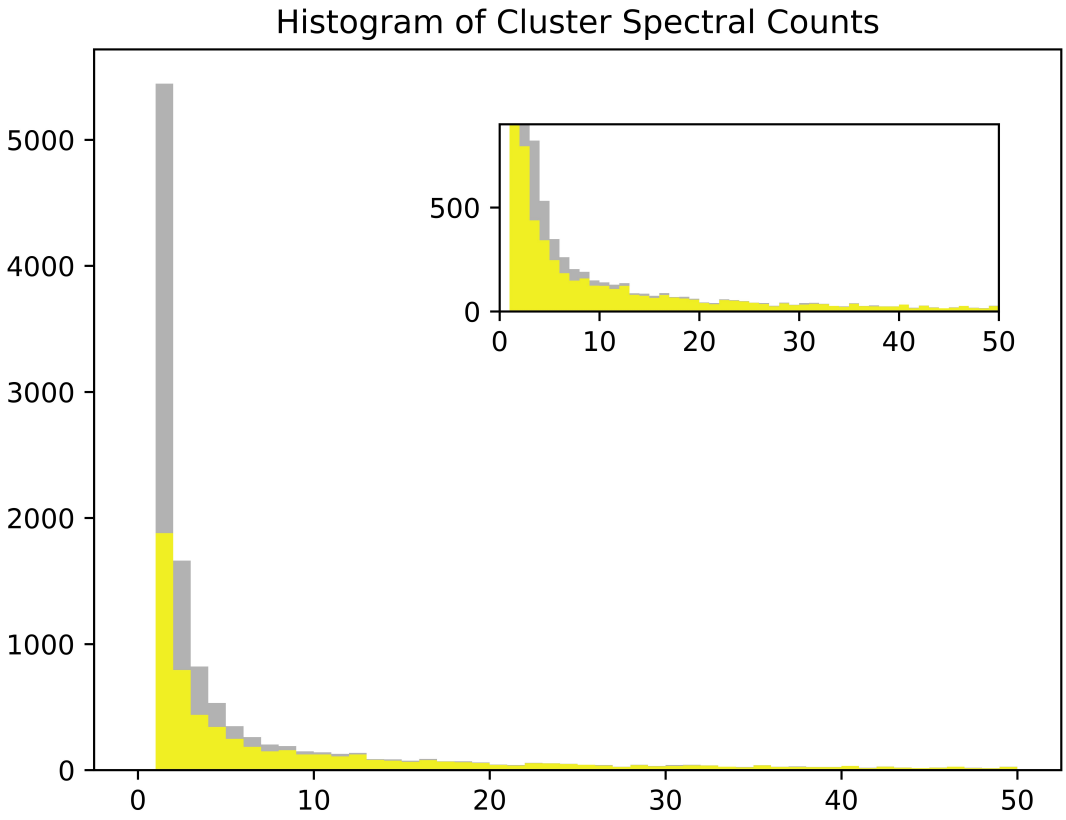


Figure 4.2: Histogram of Spectra Counts for Each Spectral Cluster. Grey: for the entire data set. Yellow: for a subset of the data where each spectral cluster appeared in all the replicates for at least one of the 9 experimental conditions describe above on Page 51.

4.4.1 Differentially Abundant DOM Between Low-viral-content Seawater and Whole Seawater

35 statistically significant spectral clusters were found between the comparison of 80% 100 kDa cut-off viral-free water samples and that of the rest (Fig 4.4). Heatmap was constructed based on their raw spectral counts and normalized over each row. On the right side, three columns represent Cluster Index, Compound Monoisotopic Mass (M+H), and Charge State. Column dendrogram shows that 1) samples from the same time point cluster together; 2) samples from the same water type cluster together, i.e., whole water samples cluster together while 80% VF water samples group well; 3) phosphorus addition does not seem to have a

big effect on the DOM compositions; 4) Differentially abundant spectral clusters are found among three sample groups: Time-0 samples, Time-48 whole water samples, Time-48 80% VF water samples.

Elemental ratio plot was made for the 35 statistically significant spectral clusters shown in the heatmap (Figure 4.4). Molecular formulae is estimated through Chemcalc See Subsection 2.4.8 on Page 36. Formulae candidates from ChemCalc for each spectral cluster were ranked by mass error. Top 4 formulae with smallest mass error were chosen (note that some spectral clusters have less than 4 formulae candidates from ChemCalc) as representatives. Candidates for the same compound cluster are connected by a polygon (e.g., quadrilateral, triangle) or a line (in the case where only two formula candidates are available for the compound cluster). Polygons or lines are colored by their differential enrichment in the different sample types (Figure 4.4). It suggests: 1) most candidates of the 35 spectral clusters fall in the region of lipids, neocleotides, and proteins; 2) compound candidates of 80% VF water incubation (i.e., the condition where much fewer viruses were present) samples seemed to cover a smaller region of the plot than those of the whole water incubation samples, indicating viruses in the whole water incubation samples could lead to a more complex DOM composition.

4.4.2 Differentially Abundant DOM Between Low-viral-content Seawater Incubation at T0 and T48hr

52 statistically significant spectral clusters were found between the comparison of 80% 100 kDa cut-off viral-free water samples incubated for 48 hours and that of collected at the beginning of incubation (T0) (Fig 4.8a and Fig 4.6). Heatmap was constructed based on their raw spectral counts and normalized over each row. On the right side, three columns represent Cluster Index, Compound Monoisotopic Mass (M+H), and Charge State. Column dendrogram shows that 1) samples from the same time point cluster together; 2) samples from the same water type cluster together; 3) phosphorus addition does not seem to have a big effect on the DOM compositions; 4) differentially abundant spectral clusters are found

between the two sample groups.

7 out of the 52 differentially abundant clusters appear to have more prominent signals by the end of the 48-hour incubation, indicating that the decrease of viruses in the water (due to the fact that the viral content is reduced in the 80% 100 kDa cut-off viral-free water) could potentially lead to an accumulation of different compounds. For the same reasoning, reduction of viral pressure also seemed to have decreased the concentration of certain molecules in the environment (the rest 45 out of 53 compound clusters).

There's one caveat: there's only one biological replicate for the T0 80% 100 kDa cut-off viral-free water condition so this may lead to a less robust statistical hypothesis testing when the Generalized Linear Model is applied.

4.4.3 Hypothesis Testing on Other Sub-sample Groups

Other comparisons were also made to test for differentially abundant compound clusters between groups by evaluating the the log2 fold change of their spectral counts (Fig 4.8). Unlike the two comparisons discussed before, a few other sub-sample groups did not exhibit differentially abundant compound clusters at significance. For instance, samples harvested at Time-48hr were divided into two subgroups: one with phosphorus addition while the other without. P values for this set of statistical testing using generalized linear model are above 0.99 (Fig 4.8b. See Subsection 2.4.7 for details). Among 80% 100 kDa cut-off viral-free water samples at Time-48hr, we compared the sub-group which had phosphorus addition with the other sub-group which did not have phosphorus addition but did not find significance either (Fig 4.8c). A similar result was found for the comparison of whole water samples at Time-48hr with or without phosphorus addition (Fig 4.8d).

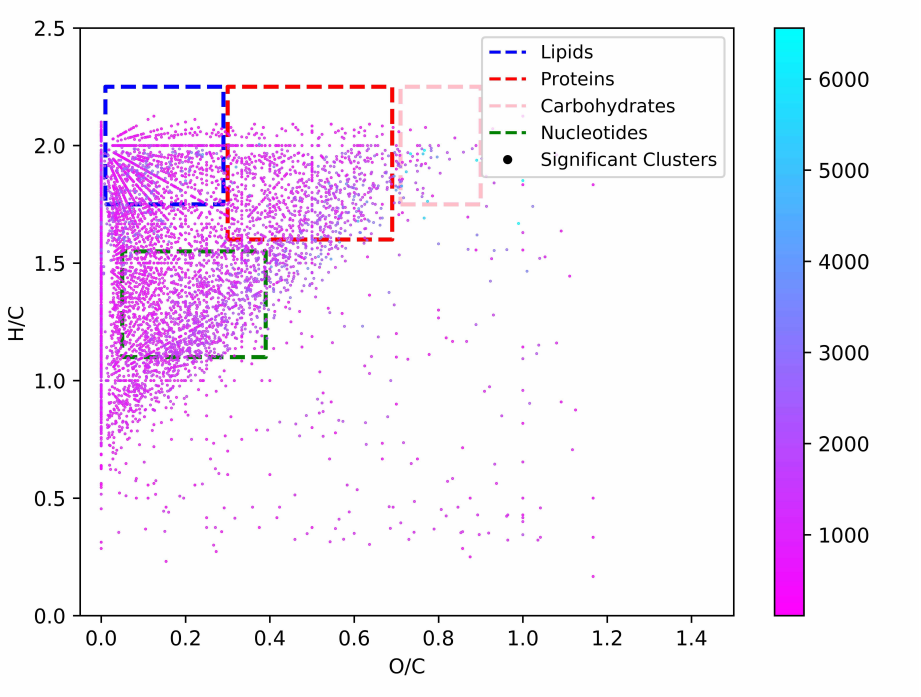
4.4.4 Comparing Results between Field Incubation and Lab Cultures

When the elemental ratio plots from *Synechococcus* lab culture experiment and the BATS field incubation are compared, one intriguing feature appears to be that there are much

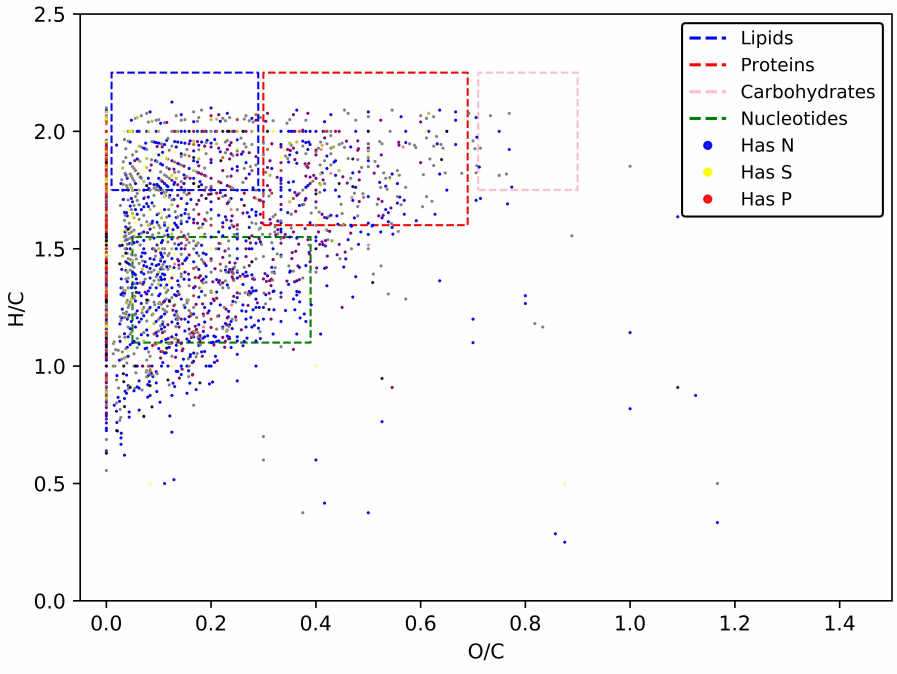
fewer compound clusters plotted in the lower right region (low hydrogen and high oxygen content) in the BATS field incubation samples. There are a couple hypotheses: a) these hydrogen poor but oxygen rich compounds may be consumed in the field as nutrients by other microbes in the microbial loop albeit usually at a slower rate than alkane consumption [11]; b) photo degradation of aromatic compounds (high O/C and low H/C) in the open ocean could reduce the presence of this compound group [74].

4.4.5 Summary

BATS field incubation produced a comprehensive data set. At the genetic level, 16S RNA analysis has shown that the 48-hour incubation led to a differentiation of microbial communities (Fig 4.1). The Time-0hr samples cluster together in the PCA plot while Time-48hr samples dispersed into subgroups based on their specific treatments (e.g., phosphorus addition and seawater types). Phosphorus concentration had a pronounced effect in the microbial community structure where grazers or viruses had been removed (in 80% cell-free water and 80% 100 kDa cut-off viral-free water respectively). 80% 100 kDa cut-off viral-free water seemed to have a bigger impact on differentiating the microbial communities from that of the whole water and the 80% cell-free water. At the molecular level in terms of metabolomics or DOM analysis, we see differentially abundant compound clusters in the comparisons of 80% 100 kDa cut-off viral-free water Time-0hr vs Time-48hr samples (Fig 4.6), as well as 80% 100 kDa cut-off viral-free water Time-48hr vs. whole seawater samples (Fig 4.4). However, we did not see the phosphorus effect as we did in the 16S RNA analysis.



(a) Colored by Molecular Weight



(b) Colored by Hetero-atom Type

Figure 4.3: O/C vs H/C for Spectral Clusters Found in at Least One Experimental Condition in BATS Samples

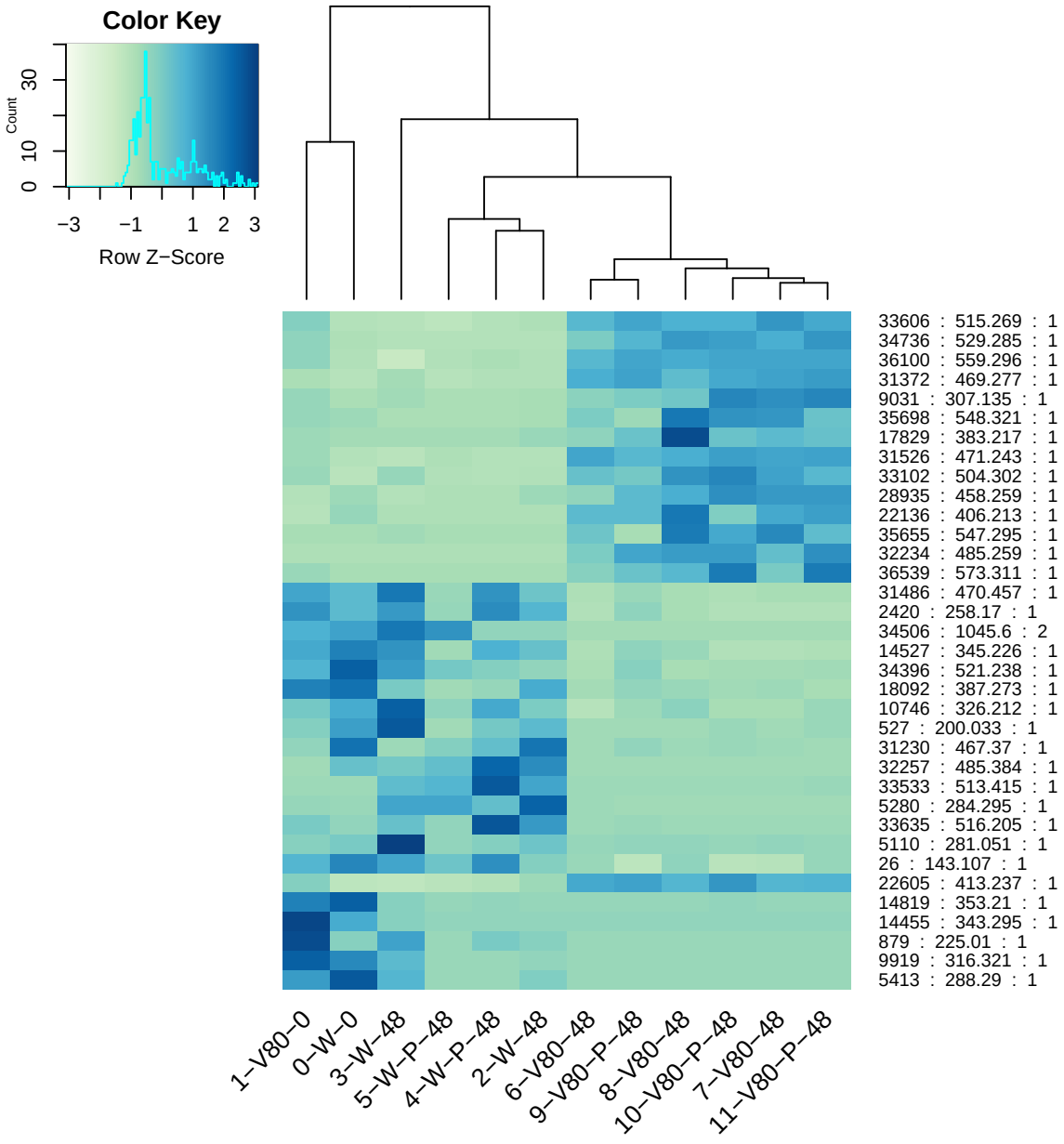


Figure 4.4: Distinct DOM Compositions of Seawater Incubation at BATS: VFT48 vs All Other

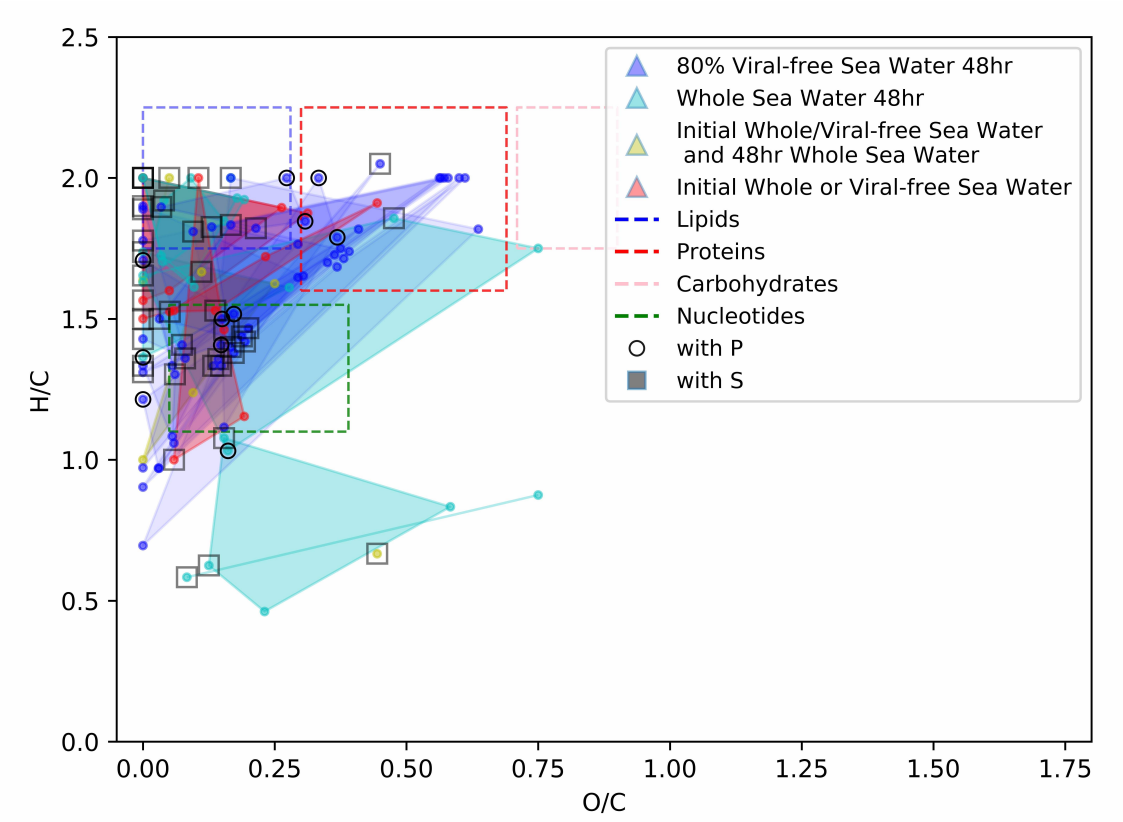


Figure 4.5: Elemental Ratio Plot - O/C vs H/C for Differentially Abundant Compounds in BATS Samples

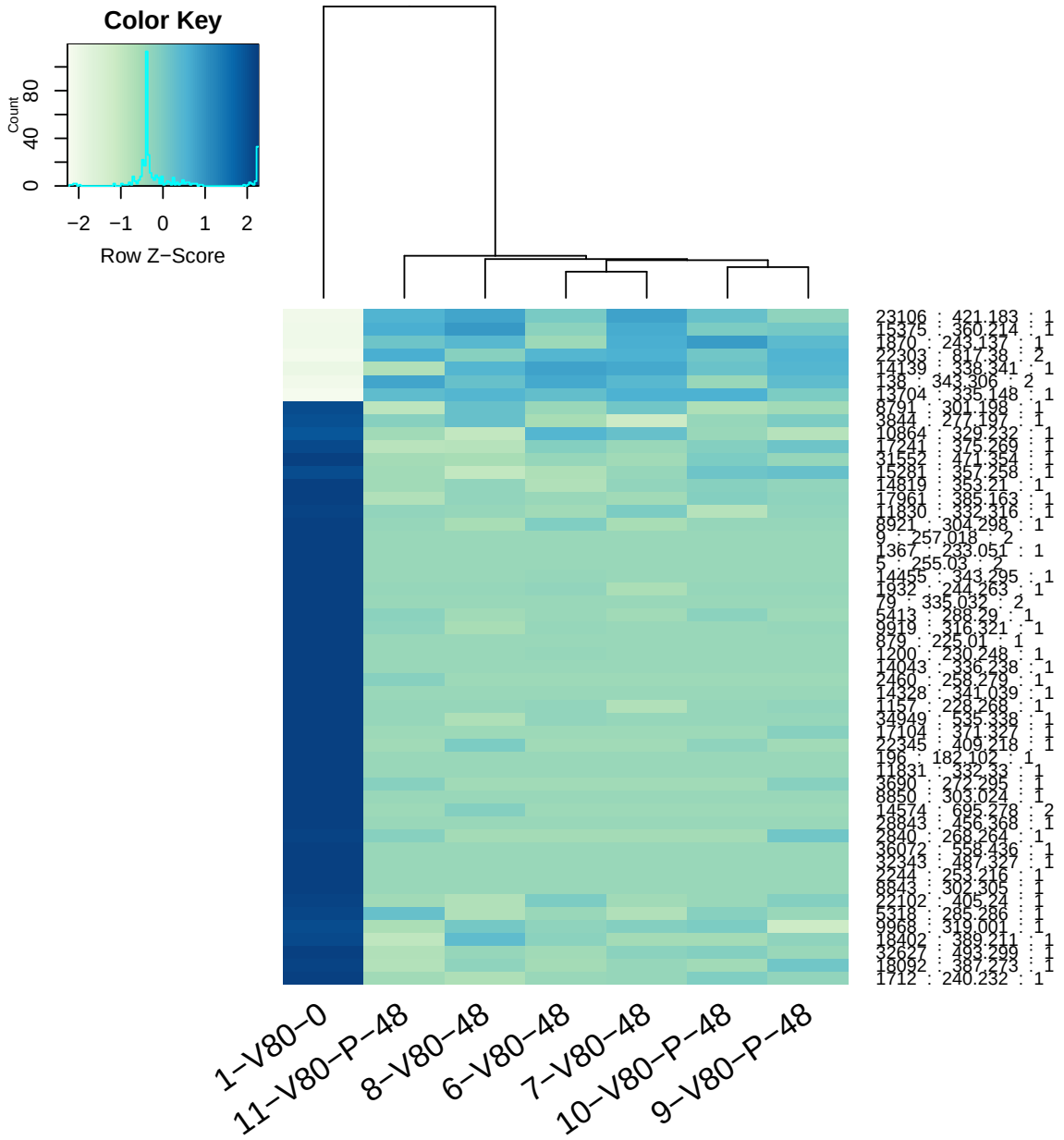


Figure 4.6: Distinct DOM Compositions of Seawater Incubation at BATS: VFT48 vs VFT0

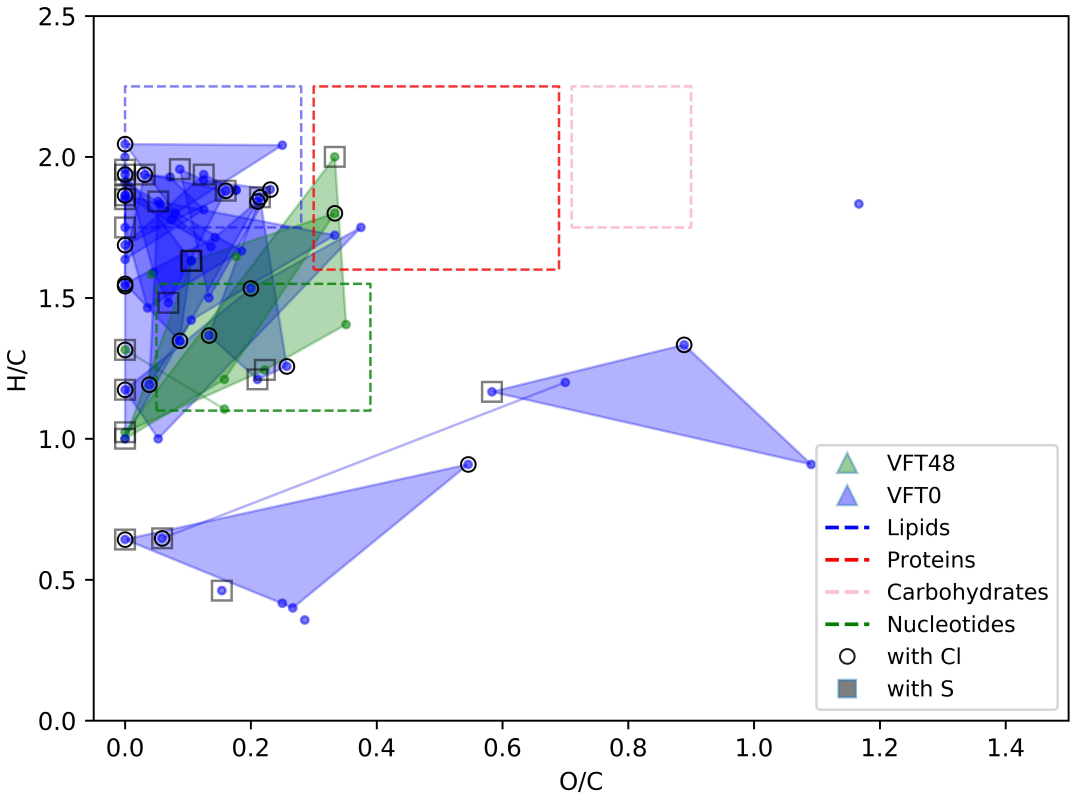
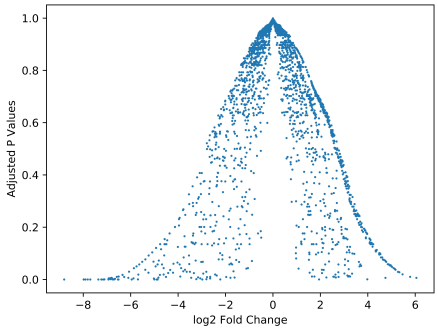
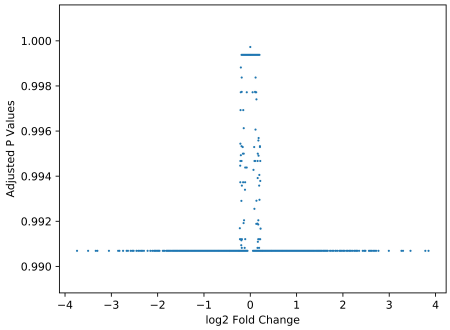


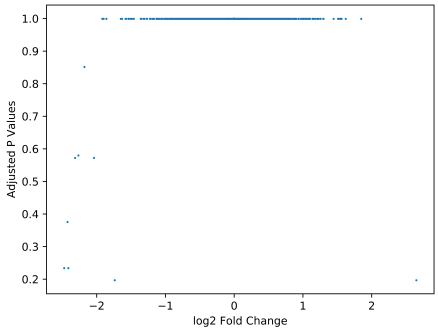
Figure 4.7: Elemental Ratio Plot - O/C vs H/C for Differentially Abundant Compounds in BATS Samples



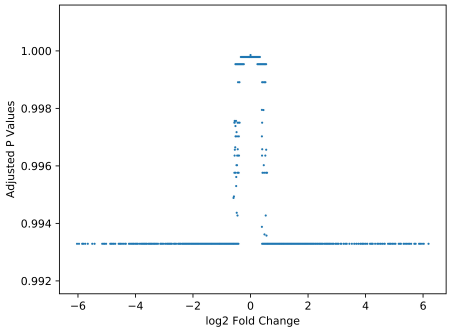
(a) 80% 100 kDa cut-off viral-free water samples: Time-48hr vs. Time-0hr



(b) All samples at time-48hr: with phosphorus addition vs. without phosphorus addition



(c) 80% 100 kDa cut-off viral-free water samples at time-48hr: with phosphorus addition vs. without phosphorus addition



(d) Whole water samples at time-48hr: with phosphorus addition vs. without phosphorus addition

Figure 4.8: Hypothesis Testing for Cluster Spectral Counts between Sub-sample Groups.

APPENDIX A

TRACING PHAGE NUTRIENT SOURCES DURING INFECTION OF CYANOBACTERIA

A.1 Overview

During viral infection, host gene expression is usually shut down and thus phage particles must be assembled from building blocks present in the host cell or by taking additional nutrients from extracellular environment. To test the hypothesis whether phages pick up external nutrients and to further quantify the usage of such extracellular nutrients, isotope labeling method is well suited for this purpose. Particularly, stable nitrogen isotope (15N) is used to trace nitrogen incorporation during phage infection. This Appendix A summarizes a method developed to measure nitrogen isotopes of phages particles using Isotope Ratio Mass Spectrometry (IRMS)

A.2 Results and Discussion

To trace the nitrogen source of phages during infection of cyanobacteria, two sets of experiments using different forms of nitrogen compounds (nitrate and ammonium) in the media were conducted. In the one using nitrate media, phages showed about 16% incorporation of 15N after infection (Figure A.1 and A.2). It indicates that phages use a majority (ca. 84%) of nitrogen from host cell components to replicate during infection while drawing around 16% of their nitrogen from extra-cellular nitrate media.

A.3 Materials and Methods

A.3.1 Experimental Design

Goal: quantify nutrient incorporation into phage particles during infection through isotopes.

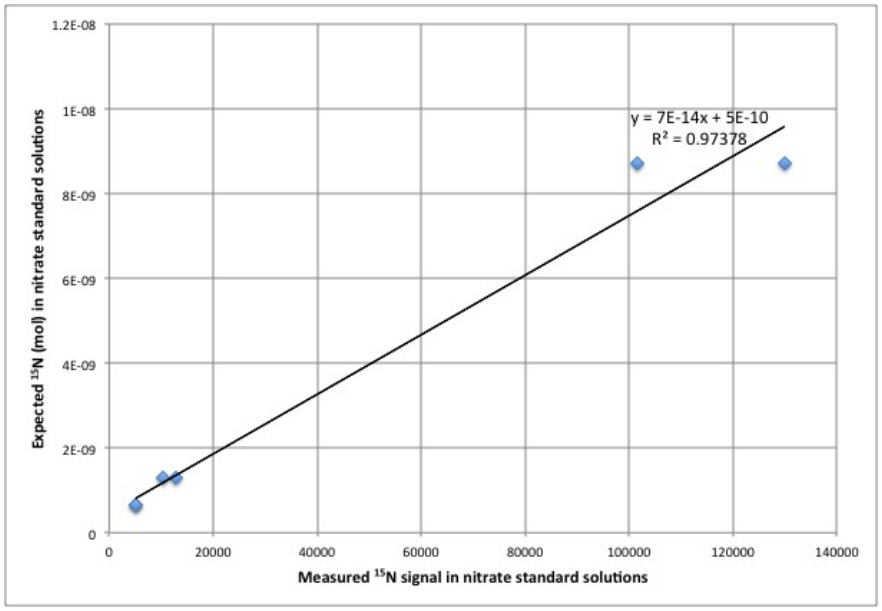


Figure A.1: Measure delta 15N signal versus expected 15N (mol) in nitrate standard solutions analyzed with Na15NO3 labeling experiment. Nitrate standard solutions analyzed with Na15NO3 labeling experiment. It serves as a delta 15N calibration curve which can be used to calculate measured 15N in labeled phage particles.

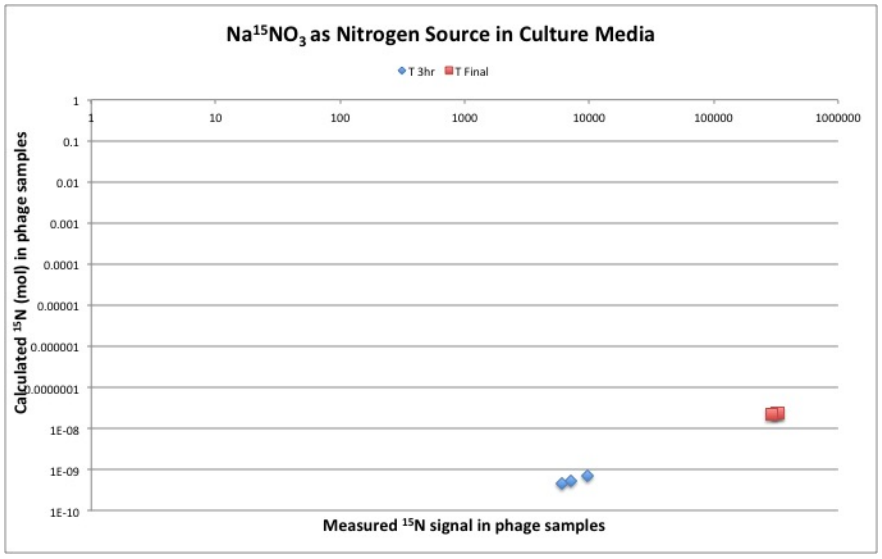


Figure A.2: Measure delta 15N signal versus calculated 15N (mol) in phage samples of Na15NO3 labeling experiment. Nitrate standard solutions analyzed with Na15NO3 labeling experiment. It serves as a delta 15N calibration curve which can be used to calculate measured 15N in labeled phage particles.

two experiments targeting two different sources of nitrogen: 15N labeled nitrate; 15N labeled ammonium.

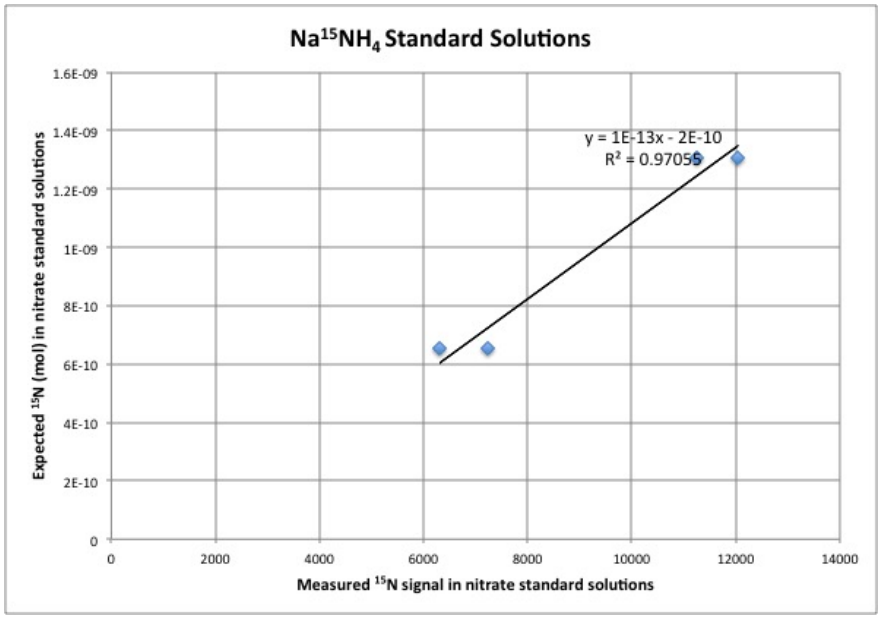


Figure A.3: Measure delta 15N signal versus expected 15N (mol) in nitrate standard solutions analyzed with Na15NH4 labeling experiment. Nitrate standard solutions analyzed with Na15NH4 labeling experiment. It serves as a delta 15N calibration curve which can be used to calculate measured 15N in labeled phage particles.

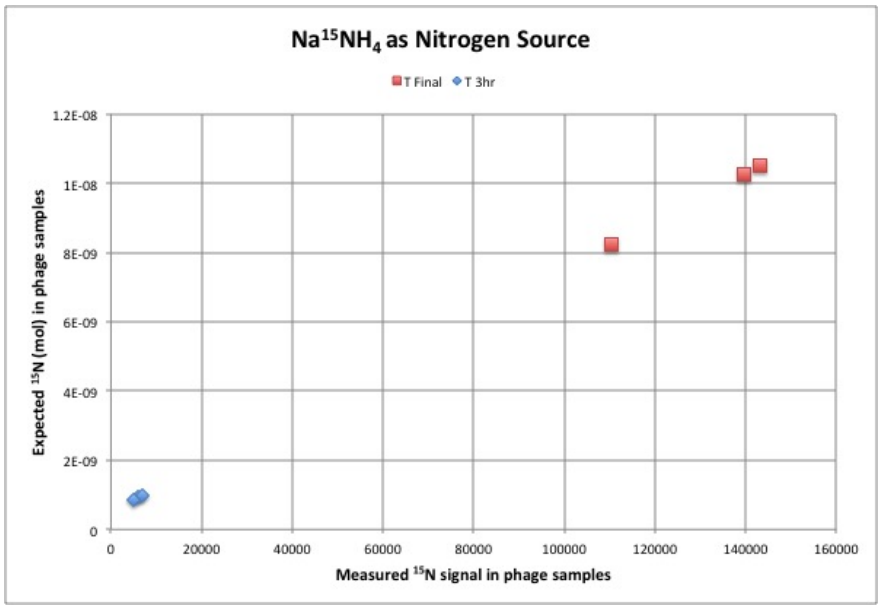


Figure A.4: Measure delta 15N signal versus expected 15N (mol) in nitrate standard solutions analyzed with Na15NH4 labeling experiment. Nitrate standard solutions analyzed with Na15NH4 labeling experiment. It serves as a delta 15N calibration curve which can be used to calculate measured 15N in labeled phage particles.

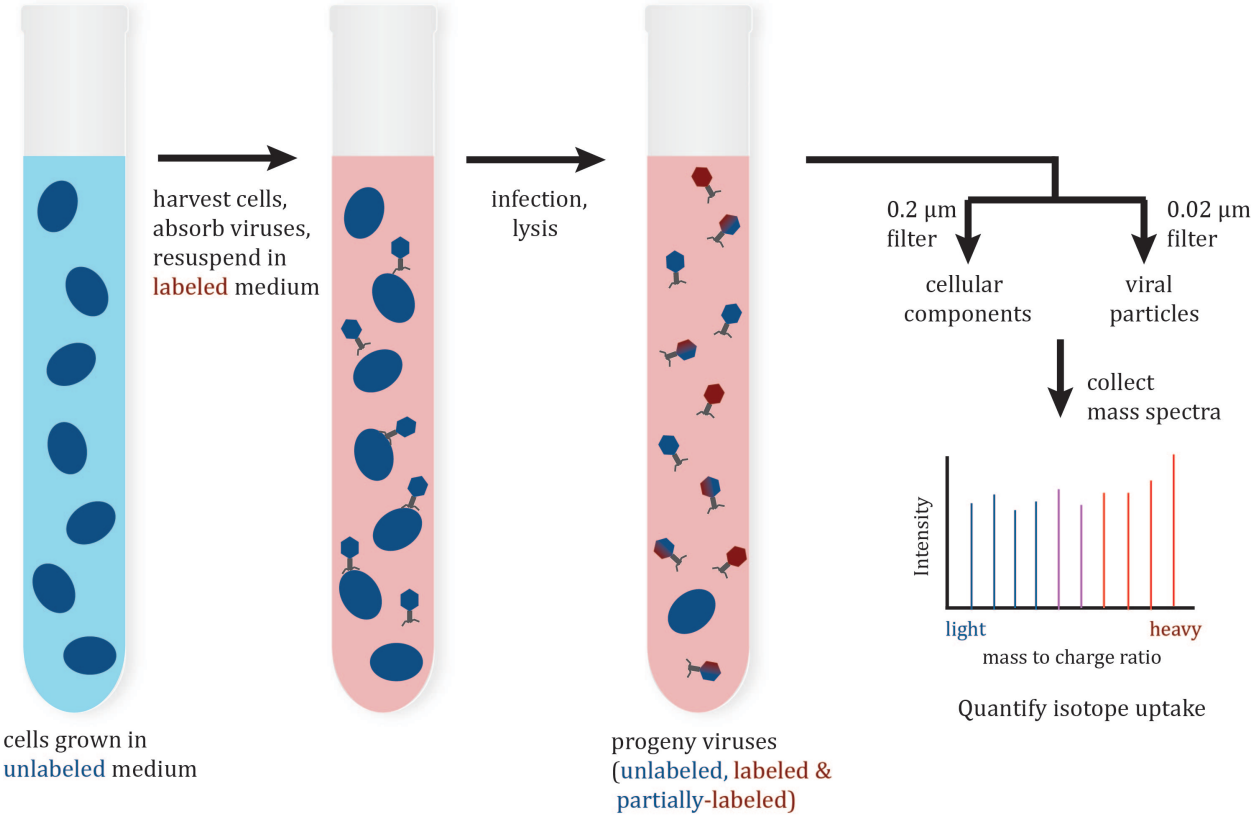


Figure A.5: Quantify nutrient uptake during viral infection and its contribution to viral progeny through isotope labeling. Adapted from Coleman, et al. Expected results: for both conditions (Group A and B), viral progeny and/or host cell components are expected to have heavy isotopes (labeled nutrients) derived from the labeled inorganic substrates. This means that nutrient uptake can happen during viral infection and that such uptake directly fuels viral production. The ratio of labeled and unlabeled nutrients in the viral particles could indicate proportions of nutrients drawn from the host and from the outside environment. For example, if there are more labeled nutrients than unlabeled ones in the viral particles, it means that viruses mainly use nutrients from extracellular medium, indicating intracellular nutrients are far from enough or not in readily accessible forms for viruses to use. Compared with Group B, Group A is expected to have a smaller ratio of heavy-isotope-labeled nutrients over unlabeled nutrients, meaning that nutrient uptake from extracellular environment is proportionally enhanced in nutrient-limited conditions.

Phage-specific Considerations

1. Elemental composition of phage Since m2 is estimated based on phage counting (e.g., SYBR count), we need to estimate the nitrogen content of one phage particle. An Excel file named ElementalCompositionPhage shows the process to estimate N content of a single

phage based on other literatures.

2. Phage numbers before and after infection Before infection, unlabeled phages (with natural 14N enriched nitrogen composition) are added to host cells. Among those phages, only a certain proportion (ca. 5%) of them are considered infective. The number of infective phages divided by the number of hosts is defined as Multiplicity of Infection (MOI). It is estimated that each infective phage can lead to 30 more phage progeny at cell burst. Only these progeny, cellular components that grow after infection, and uninfected cells which still grow after infection, could possible have 15N labeled material.

Therefore, the absolute number of unlabeled phages versus labeled phages matters a lot in order to get a good measurement on the Mass Spec. The measurement of phage isotopic composition is complicated by cell debris from (mostly) uninfected cells, the amount of which is largely affected by MOI.

On the one hand, a high MOI potentially leads to less complication from cell debris. On the other hand, a high MOI also means adding more unlabeled phages before infection and this would overwhelm the isotopic signal from labeled phage progeny, thus leading to a poor measurement.

An Excel file named PhageNumbersPrep.xlsx is included to help prepare the experiment through calculations concerning the number of phages and host cells before and after infection.

A.3.2 Testing Standards on EA-IRMS

The major goal is to establish a calibration curve for 15N measurement on the EA-IRMS in Albert Colman 's lab at Department of the Geophysical Sciences, University of Chicago. This process has been tailored and optimized towards measuring 15N-labeled phage particles using this calibration curve. Delta 15N ranges in 50 - 7000 per mil, which is favored by the mass spectrometer settings, has been tested (Figure A.6, A.7, A.8, and A.9). Delta 15N values have been brought to this window through controlling the amount of 15N labeled

phage particles. See 15Ncalculation.xlsx for details.

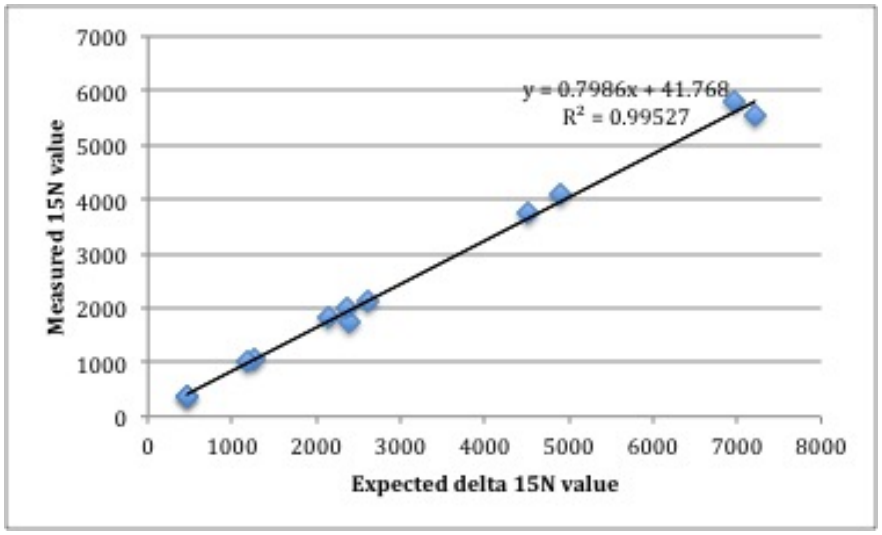


Figure A.6: Measure delta 15N versus expected delta 15N (absolute value) in a mixture of 15N labeled nitrate solution and in-lab standard cocoa. To establish a delta 15N calibration curve. Mix different amount of cocoa and Na15NO3 solution.

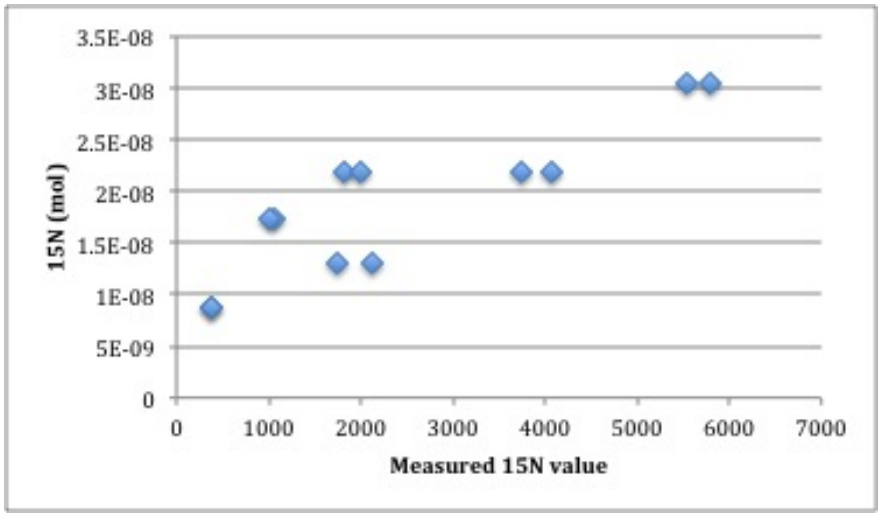


Figure A.7: Measure delta 15N versus expected delta 15N (mol) in a mixture of 15N labeled nitrate solution and in-lab standard cocoa. To establish a delta 15N calibration curve. Mix different amount of cocoa and Na15NO3 solution.

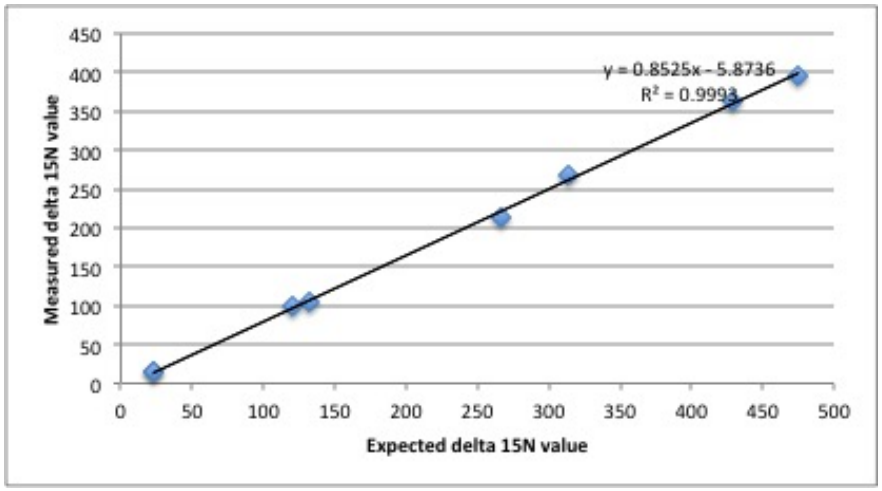


Figure A.8: Measure delta 15N versus expected delta 15N (absolute value) in a mixture of 15N labeled nitrate solution, in-lab standard cocoa and Anodisc filter. To establish a delta 15N calibration curve. Mix different amount of cocoa, Na15NO3 solution and Anodisc filter.

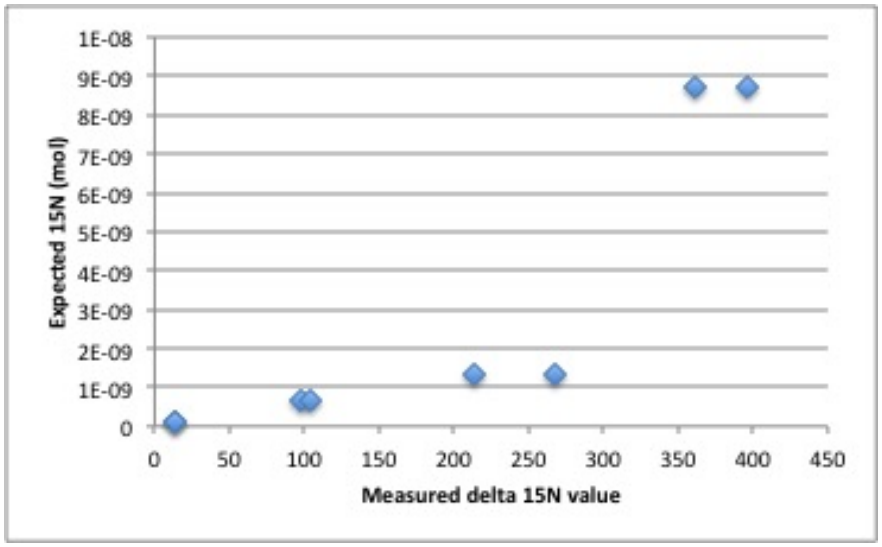


Figure A.9: Measure delta 15N versus expected delta 15N (mol) in a mixture of 15N labeled nitrate solution, in-lab standard cocoa and Anodisc filter. To establish a delta 15N calibration curve. Mix different amount of cocoa, Na15NO3 solution and Anodisc filter.

A.3.3 Phage Sample Packets Preparation for EA-Mass Spectrometry

Each sample is eventually loaded onto the EA-IRMS sample holder as a small tin foil packet or ball which usually fits into a 96-well plate's hole. Phage particles need to be packed in this tin foil. Before this, phages are separated from liquid media and collected and concentrated

on a 0.02 um filter (See Lysate Preparation for details). As technical challenges prevent isolating phage particles from it, the filter - Anodisc will also be packed along into the tin foil. Steps are presented below.

Breaking Down the Anodisc Filter

The filter is ground into small pieces on a aluminum foil supported by the petri dish. Two sets of forceps of similar style are used to tear apart the anodisc and then further press the pieces to smaller ones repetitively until they break down to an average diameter of ca. 2-4mm. Aluminum foil is folded up during grinding with a small opening at the top for a set of forceps to get through and press the Anodisc into smaller pieces, due to the fact that the filter is brittle and can scatter around.

Packaging Ground Anodisc into Tin Capsule

Small Anodisc pieces are transferred into a tin capsule (5 by 9mm) that's regularly been used to carry samples for EA-IRMS measurement. The tin capsule will then be folded and slightly compressed with great care to avoid puncturing by the brittle Anodisc pieces. Packaged tin capsules are then ready for EA-IRMS.

Preparing Blanks or References

Tin capsule blanks are pure or empty tin capsules, folded and packaged for direct analysis. Reference standards are tin capsules filled with tiny droplets of 15N labeled nitrate or ammonium solutions, dried under room temperature and then packaged for analysis. During early standard testing stage, cocoa powder was also be used as reference.

APPENDIX B

DOM COMPOSITIONS OF VIRAL LYSATE FROM OSTREOCOCCUS

B.1 *Ostreococcus* DOM analysis

Using a broad, untargeted analytical approach, between 30,000-40,000 MS2 spectra were collected from triplicate samples of Control, OlVX and OlV7 lysate DOM (Table B.1 Cluster Statistics Summary). Spectra from each sample are grouped into spectral clusters by the GNPS MS2 spectra matching algorithm[84]. A spectral cluster can be viewed as a chemical species with a certain mass-to-charge ratio (m/z) and MS2 fragmentation pattern. MS2 spectra that have parent m/z values within a defined tolerance (< 0.03 Da) and sufficiently similar fragmentation patterns as measured by cosine scores (< 0.7) are considered members of the same spectral cluster. MS2 spectral clusters can thus serve as a basis for comparisons within and between samples; here the data were organized into three groups: OlVX Lysate, OlV7 Lysate and Control.

The number of MS2 spectra belonging to a given spectral cluster (the spectral count) is a measure of abundance of the compound represented by that cluster, at least when analytical conditions are held close to constant. Spectral cluster count data for each sample are analyzed together with statistical clustering algorithm (See Method Section). Spectral clusters of interest are identified based on statistically significant differential abundance between experimental conditions. Those with P value < 0.01 against the null hypothesis of equal abundance are considered statistically significant and chosen for further analysis. Differentially abundant spectral clusters (in OlVX vs OlV7, OlV7 vs Control, and OlVX vs Control) are represented in a heatmap based on their spectral cluster counts (Figure B.4). A total of 29 clusters (out of 43553 total in the dataset) were found to be differentially abundant between the conditions by these criteria.

Spectral clusters have molecular weight from 321 Da to about 1453 with most clusters

being in the 500-600 Da range. We sought to identify these compounds by a combination of elemental formula assignments to their precise intact masses and matching of fragmentation patterns to known structures via MetFrag [69]. While several candidate elemental formulas could be calculated for each cluster's intact mass, none of the fragmentation patterns produced a clear MetFrag match that would suggest a particular molecular structure, despite searching against the PubChem and KEGG databases. In the absence of specific structural identification, we used the elemental formula assignments to suggest at least the broad compound classes that these clusters might represent (Figure B.5).

Based on 112 candidate elemental formulas assigned to the 29 clusters that were enriched in either or both viral lysate(s), many of the lysate-specific DOM compounds appear to be roughly nucleotide-like (Figure B.5) in elemental composition. Notably, 85% (95/112) of the candidate formulas contain nitrogen, suggesting that viral lysis of *Ostreococcus* could contribute preferentially to dissolved organic N release. And while most of the putative elemental compositions contain N, only 23% (26/112) contain P, implying that de- or unphosphorylated nucleoside-like compounds may represent a distinct component of this virally-released DOM. The fact that the observed fragmentation patterns of these clusters could not be confidently matched to known nucleotide structures could be due to either chemical modification (e.g., oxidation) due to either biological activity or during sample preparation and analysis, or due to unrecognized fragmentation pathways or adduct formation under our mass spec conditions.

					Significant Spectral Clusters		
Sample	Replicates	MS2 Spectra	Spectral Clusters	Significant MS2 Spectra	Control	Olvx	Olv7
Control	+P_Con_A	34749	4591	106	Total 29	28	6
	+P_Con_B	40527	4520	103			
	+P_Con_C	39805	5097	60			
Olvx	+P_Olvx_A	31976	4498	1367			0
	+P_Olvx_B	32451	5088	807			
	+P_Olvx_C	40138	4785	1583			
Olv7	+P_Olv7_A	32232	4765	498			
	+P_Olv7_B	40469	5100	575			
	+P_Olv7_C	37528	5109	584			

Table B.1: A summary of spectral and compound cluster counts. "MS2 spectra" column shows the raw MS2 spectra number for each sample. They are grouped/clustered based on MS2 pattern into spectral clusters that can be considered as compounds. Then hypothesis testing using DESeq2 is applied to Olvx vs. Control, Olv7 vs. Control and Olv7 vs Olvx (3 pair-wise comparisons). "Significant MS2 Spectra" column shows the number of MS2 spectra that are statistically significant after the DESeq2 analysis. "Significant Spectral Clusters" section shows the number of spectral clusters that "Significant MS2 Spectra" represent.

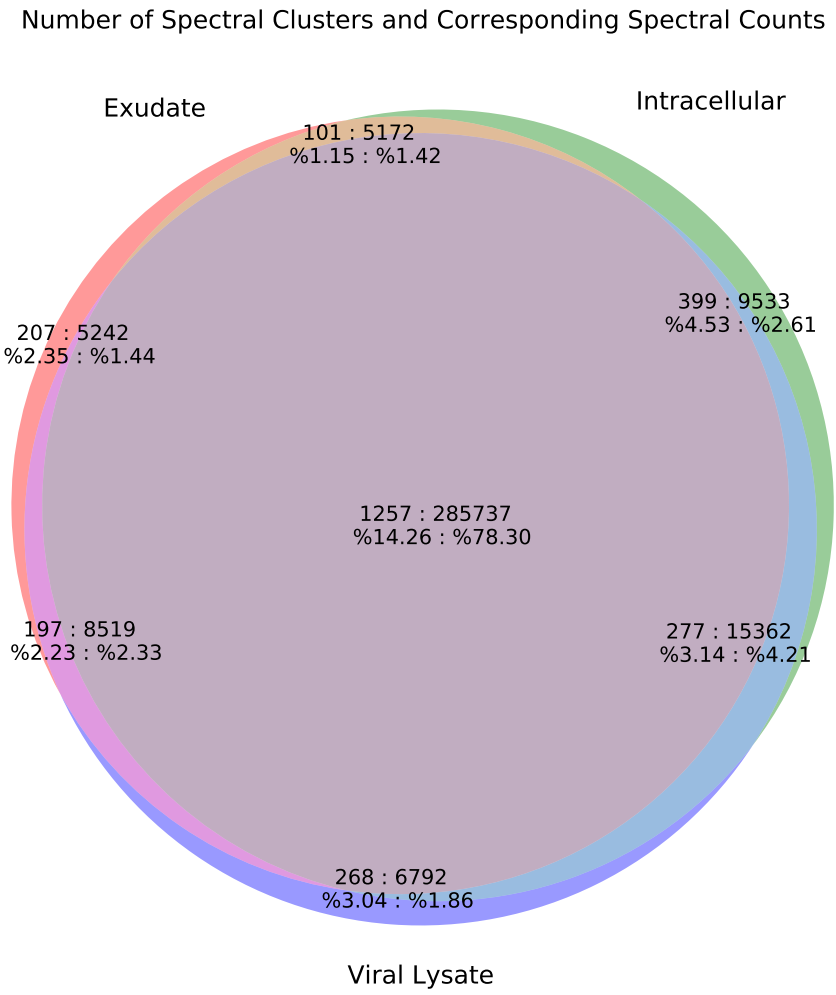


Figure B.1: *Ostreo* - Venn Diagram of Spectral Clusters and Spectra Counts

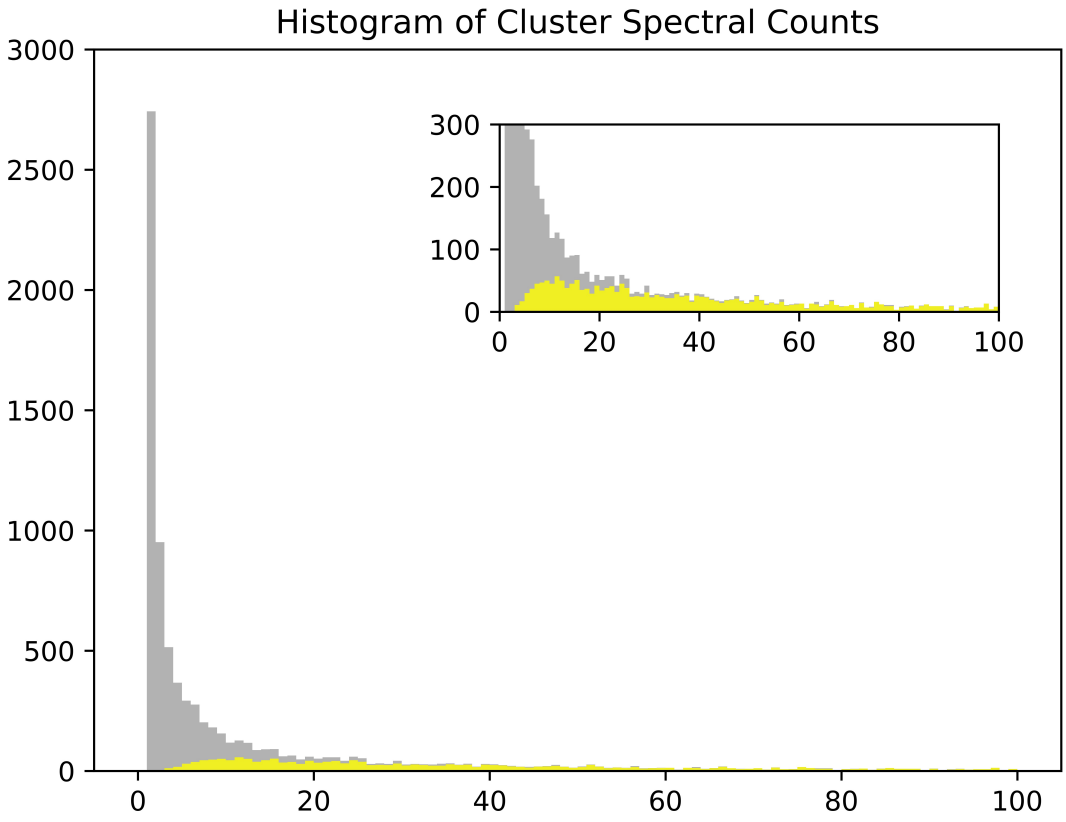
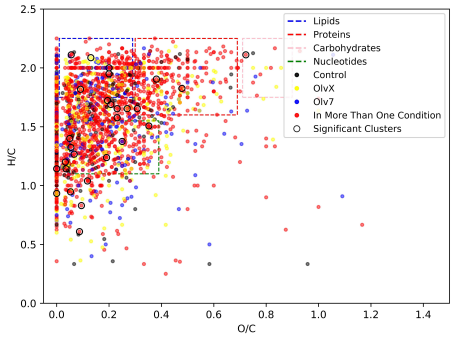
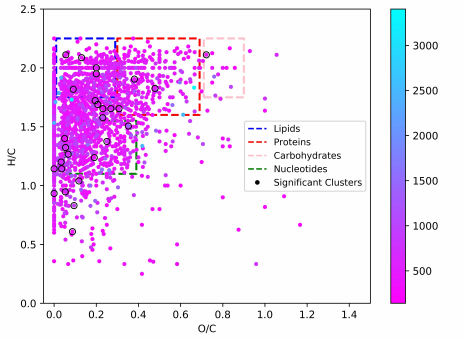


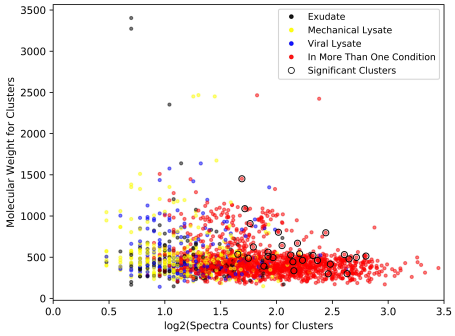
Figure B.2: *Ostreo* - Histogram of Spectra Counts



(a) Colored by Experimental Conditions



(b) Colored by Molecular Weight



(c) log10(Spectra Count) vs Molecular Weight

Figure B.3: O/C vs H/C for Spectral Clusters Found in at Least One Experimental Condition in Ostreo samples

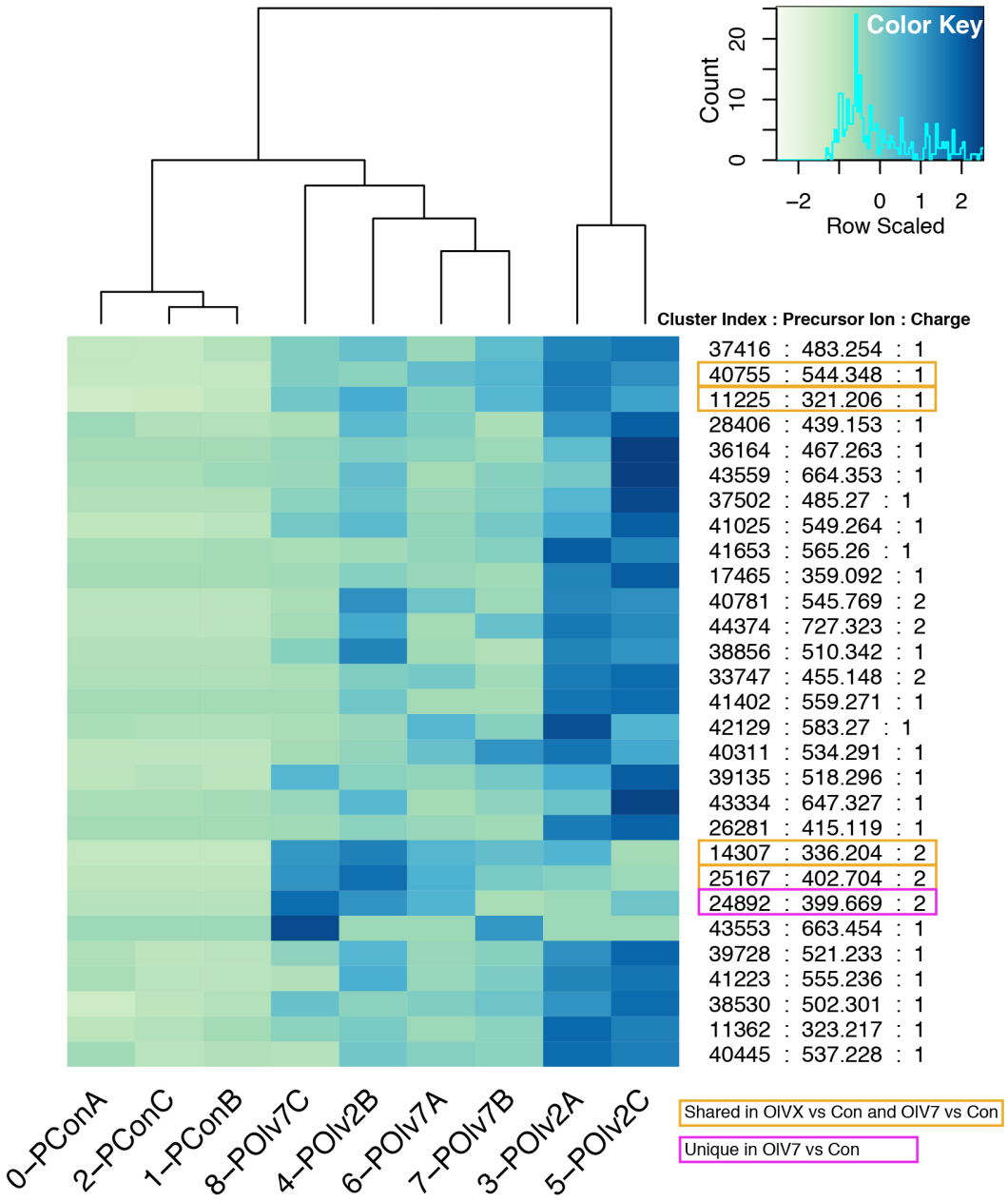


Figure B.4: Distinct DOM Compositions Between 3 Modes Revealed By Statistical Analysis. A heatmap showing cluster analysis on statistically significant spectral clusters. Spectral clusters appeared in both Olvx vs Control and Olv7 vs Control are labeled yellow. The one only appeared in both Olv7 vs Control is labeled purple. Heatmap is constructed based on raw spectral counts and normalized over each row. On the right side, compound mass in expressed as monoisotopic precursor ion mass and its charge state is listed for each spectral cluster. Column dendrogram shows that biological replicates within each DOM group well together.

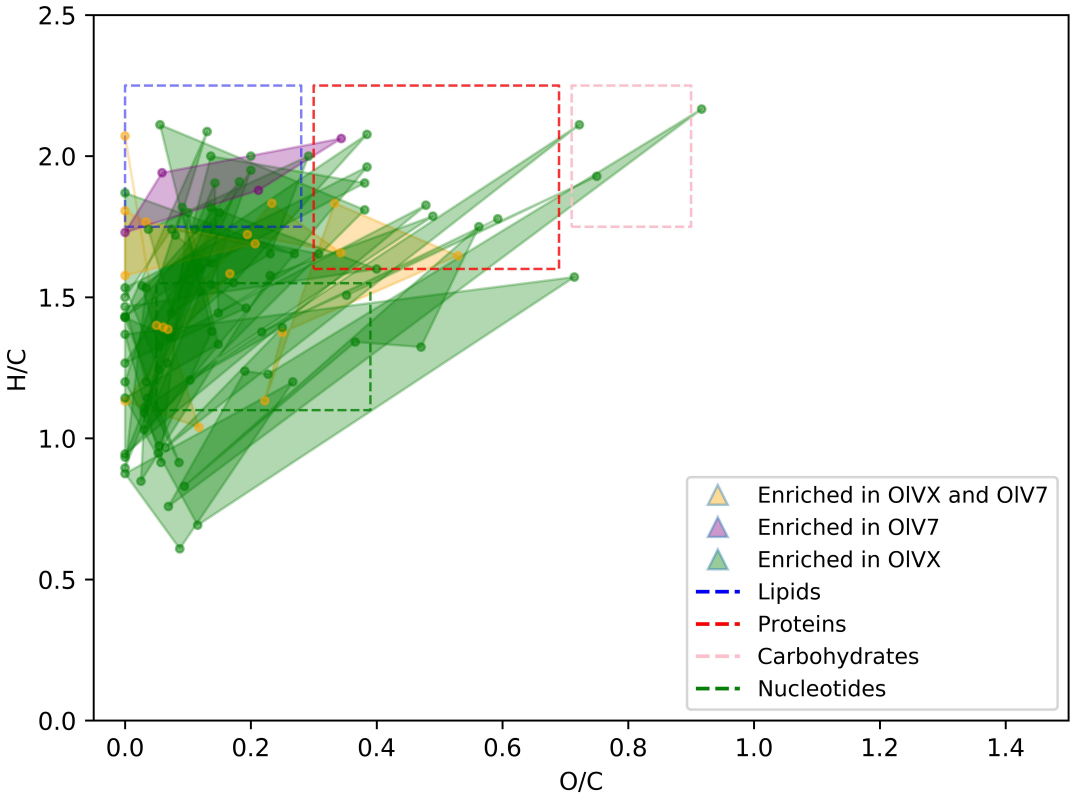


Figure B.5: O/C vs H/C for Compound Candidate Formulae. Elemental ratio plots of statistically significant clusters. Brutal-force stoichiometric calculation based on monoisotopic mass of spectral clusters was done through ChemCalc's web service[61]. Monoisotopic masses of these spectral clusters were queried with criteria constraining formula results to be chemically plausible. Formulae candidates from ChemCalc for each spectral cluster were ranked by mass error. Top 4 formulae with smallest mass error were chosen (note that some spectral clusters have less than 4 formulae candidates from ChemCalc) as representatives. Candidates for the same spectral cluster are connected by a polygon (e.g., quadrilateral, triangle), a line (in the case where only two formula candidates are available for the compound cluster), or simply exists as a dot (in case where only one formula candidate is available.

REFERENCES

[1] Dikla Aharonovich and Daniel Sher. Transcriptional response of prochlorococcus to co-culture with a marine alteromonas: differences between strains and the involvement of putative infochemicals. *The ISME journal*, 10(12):2892–2906, 2016.

[2] Rosanna A Alegado, Laura W Brown, Shugeng Cao, Renee K Dermenjian, Richard Zuzow, Stephen R Fairclough, Jon Clardy, and Nicole King. A bacterial sulfonolipid triggers multicellular development in the closest living relatives of animals. *elife*, 1:e00013, 2012.

[3] Pierre-Marie Allard, Grégory Genta-Jouve, and Jean-Luc Wolfender. Deep metabolome annotation in natural products research: towards a virtuous cycle in metabolite identification. *Current opinion in chemical biology*, 36:40–49, 2017.

[4] Pierre-Marie Allard, Tiphaine Péresse, Jonathan Bisson, Katia Gindro, Laurence Marcourt, Van Cuong Pham, Fanny Roussi, Marc Litaudon, and Jean-Luc Wolfender. Integration of molecular networking and in-silico ms/ms fragmentation for natural products dereplication. *Analytical chemistry*, 88(6):3317–3323, 2016.

[5] Nana Yaw D Ankrah, Amanda L May, Jesse L Middleton, Daniel R Jones, Mary K Hadden, Jessica R Gooding, Gary R LeCleir, Steven W Wilhelm, Shawn R Campagna, and Alison Buchan. Phage infection of an environmentally relevant marine bacterium alters host metabolism and lysate composition. *ISME J*, 8(5):1089–1100, May 2014.

[6] Antje Baier, Wiebke Winkler, Thomas Korte, Wolfgang Lockau, and Anne Karradt. Degradation of phycobilisomes in synechocystis sp. pcc6803 evidence for essential formation of an nbla1/nbla2 heterodimer and its codegradation by a clp protease complex. *Journal of Biological Chemistry*, 289(17):11755–11766, 2014.

[7] Jamie W. Becker, Paul M. Berube, Christopher L. Follett, John B. Waterbury, Sallie W. Chisholm, Edward F. Delong, and Daniel J. Repeta. Closely related phytoplankton species produce similar suites of dissolved organic matter. *Frontiers in Microbiology*, 5:111, 2014.

[8] S Bertlisson, Olof Berglund, Michael J Pullin, and Sallie W Chisholm. Release of dissolved organic matter by prochlorococcus. *Vie et Milieu*, 55(3-4):225–232, 2005.

[9] Steven J Biller, Paul M Berube, Debbie Lindell, and Sallie W Chisholm. Prochlorococcus: the structure and function of collective diversity. *Nature Reviews Microbiology*, 13(1):13–27, 2015.

[10] Sebastian Bocker and Zsuzsanna Lipták. A fast and simple algorithm for the money changing problem. *Algorithmica*, 48(4):413–432, 2007.

[11] Rob JW Brooijmans, Margreet I Pastink, and Roland J Siezen. Hydrocarbon-degrading bacteria: the oil-spill clean-up crew. *Microbial biotechnology*, 2(6):587–594, 2009.

[12] Francis Chan, Roxanne L. Marino, Robert W. Howarth, and Michael L. Pace. Ecological constraints on planktonic nitrogen fixation in saline estuaries. ii. grazing controls on cyanobacterial population dynamics. *Marine Ecology Progress Series*, 309:41–53, 2006.

[13] Hsiu-An Chu and Yi-Fang Chiu. The roles of cytochrome b559 in assembly and photoprotection of photosystem ii revealed by site-directed mutagenesis studies. *Frontiers in plant science*, 6, 2015.

[14] Angela Corcelli, Veronica MT Lattanzio, Giuseppe Mascolo, Francesco Babudri, Aharon Oren, and Morris Kates. Novel sulfonolipid in the extremely halophilic bacterium salinibacter ruber. *Applied and environmental microbiology*, 70(11):6678–6685, 2004.

[15] Thorsten Dittmar, Boris Koch, Norbert Hertkorn, and Gerhard Kattner. A simple and efficient method for the solid-phase extraction of dissolved organic matter (spe-dom) from seawater. *Limnol. Oceanogr. Methods*, 6(6):230–235, 2008.

[16] Nadia Dolganov and Arthur R. Grossman. A Polypeptide with Similarity to Phycocyanin -Subunit Phycocyanobilin Lyase Involved in Degradation of Phycobilisomes. *Journal of Bacteriology*, 181(2):610–617, January 1999.

[17] Alexis Dufresne, Laurence Garczarek, and Frédéric Partensky. Accelerated evolution associated with genome reduction in a free-living prokaryote. *Genome biology*, 6(2):R14, 2005.

[18] Richard W Eppley, Jane N Rogers, and James J McCarthy. Half-saturation constants for uptake of nitrate and ammonium by marine phytoplankton. *Limnology and oceanography*, 14(6):912–920, 1969.

[19] Beatrix S Falch, Gabriele M Koenig, Anthony D Wright, Otto Sticher, Heinz Ruegger, and Gerald Bernardinelli. Ambigol A and B: new biologically active polychlorinated aromatic compounds from the terrestrial blue-green alga Fischerella ambigua. *The Journal of Organic Chemistry*, 58(24):6570–6575, 1993.

[20] Paul Falkowski, RJ Scholes, EEA Boyle, Josep Canadell, D Canfield, J Elser, Nicolas Gruber, Kathy Hibbard, Peter Högberg, S Linder, et al. The global carbon cycle: a test of our knowledge of earth as a system. *science*, 290(5490):291–296, 2000.

[21] Christopher B Field, Michael J Behrenfeld, James T Randerson, and Paul Falkowski. Primary production of the biosphere: integrating terrestrial and oceanic components. *Science*, 281(5374):237–240, 1998.

[22] Cara L Fiore, Krista Longnecker, Melissa C Kido Soule, and Elizabeth B Kujawinski. Release of ecologically relevant metabolites by the cyanobacterium, synechococcus elongatus ccmp 1631. *Environmental Microbiology*, pages n/a–n/a, 2015.

[23] Cara L Fiore, Krista Longnecker, Melissa C Kido Soule, and Elizabeth B Kujawinski. Release of ecologically relevant metabolites by the cyanobacterium synechococcus elongatus ccmp 1631. *Environmental microbiology*, 17(10):3949–3963, 2015.

[24] Pedro Flombaum, José L Gallegos, Rodolfo A Gordillo, José Rincón, Lina L Zabala, Nianzhi Jiao, David M Karl, William KW Li, Michael W Lomas, Daniele Veneziano, et al. Present and future global distributions of the marine cyanobacteria prochlorococcus and synechococcus. *Proceedings of the National Academy of Sciences*, 110(24):9824–9829, 2013.

[25] E.-Bin Gao, Jian-Fang Gui, and Qi-Ya Zhang. A Novel Cyanophage with a Cyanobacterial Nonbleaching Protein A Gene in the Genome. *Journal of Virology*, 86(1):236–245, January 2012.

[26] Ferran Garcia-Pichel, Christopher E. Wingard, and Richard W. Castenholz. Evidence Regarding the UV Sunscreen Role of a Mycosporine-Like Compound in the Cyanobacterium Gloeocapsa sp. *Applied and Environmental Microbiology*, 59(1):170–176, January 1993.

[27] Raphael Gasper, Julia Schwach, Jana Hartmann, Andrea Holtkamp, Jessica Wiethaus, Natascha Riedel, Eckhard Hofmann, and Nicole Frankenberg-Dinkel. Distinct features of cyanophage-encoded t-type phycobiliprotein lyase ϕcpet the role of auxiliary metabolic genes. *Journal of Biological Chemistry*, 292(8):3089–3098, 2017.

[28] Koen Goiris, Koenraad Muylaert, Stefan Voorspoels, Bart Noten, Domien De Paepe, Gino J. E Baart, and Luc De Cooman. Detection of flavonoids in microalgae from different evolutionary lineages. *Journal of Phycology*, 50(3):483–492, June 2014.

[29] SJ Goldberg, CE Nelson, DA Viviani, CN Shulse, and MJ Church. Cascading influence of inorganic nitrogen sources on dom production, composition, lability and microbial community structure in the open ocean. *Environmental Microbiology*, 2017.

[30] Gordon W. Gribble. The diversity of naturally produced organohalogens. *Chemosphere*, 52(2):289–297, July 2003.

[31] Dennis A Hansell and Craig A Carlson. *Biogeochemistry of marine dissolved organic matter*. Academic Press, 2014.

[32] Dennis A Hansell, Craig A Carlson, Daniel J Repeta, and Reiner Schlitzer. Dissolved organic matter in the ocean: A controversy stimulates new insights. *Oceanography*, 22, December 2009. ¡p¿Containing as much carbon as the atmosphere, marine dissolved organic matter is one of Earth's major carbon reservoirs. With invigoration of scientific inquiries into the global carbon cycle, our ignorance of its role in ocean biogeochemistry became untenable. Rapid mobilization of relevant research two decades ago required the community to overcome early false leads, but subsequent progress in examining the global dynamics of this material has been steady. Continuous improvements in analytical skill coupled with global ocean hydrographic survey opportunities resulted in the generation of thousands of measurements throughout the major ocean basins. Here, observations and model results provide new insights into the large-scale variability of dissolved organic carbon, its contribution to the biological pump, and its deep ocean sinks.¡/p¿.

[33] Hisayuki Horai, Masanori Arita, Shigehiko Kanaya, Yoshito Nihei, Tasuku Ikeda, Kazuhiro Suwa, Yuya Ojima, Kenichi Tanaka, Satoshi Tanaka, Ken Aoshima, et al. Massbank: a public repository for sharing mass spectral data for life sciences. *Journal of mass spectrometry*, 45(7):703–714, 2010.

[34] Tao Huan, Chenqu Tang, Ronghong Li, Yi Shi, Guohui Lin, and Liang Li. Mycompoundid ms/ms search: Metabolite identification using a library of predicted fragment-ion-spectra of 383,830 possible human metabolites. *Analytical chemistry*, 87(20):10619–10626, 2015.

[35] Haomin Huang, Xi Xiao, Anas Ghadouani, Jiaping Wu, Zeyu Nie, Cheng Peng, Xinhua Xu, and Jiyan Shi. Effects of natural flavonoids on photosynthetic activity and cell integrity in microcystis aeruginosa. *Toxins*, 7(1):66–80, 2015.

[36] Saadia Ijaz and Shahida Hasnain. Antioxidant potential of indigenous cyanobacterial strains in relation with their phenolic and flavonoid contents. *Natural product research*, 30(11):1297–1300, 2016.

[37] Nianzhi Jiao, Gerhard J Herndl, Dennis A Hansell, Ronald Benner, Gerhard Kattner, Steven W Wilhelm, David L Kirchman, Markus G Weinbauer, Tingwei Luo, Feng Chen, et al. Microbial production of recalcitrant dissolved organic matter: long-term carbon storage in the global ocean. *Nature Reviews Microbiology*, 8(8):593–599, 2010.

[38] Todd L. Johnson, Bianca Brahamsha, Brian Palenik, and Jens Mühle. Halomethane production by vanadium-dependent bromoperoxidase in marine Synechococcus. *Limnology and Oceanography*, 60(5):1823–1835, September 2015.

[39] Adam C Jones, Liangcai Gu, Carla M Sorrels, David H Sherman, and William H Gerwick. New tricks from ancient algae: natural products biosynthesis in marine cyanobacteria. *Current Opinion in Chemical Biology*, 13(2):216–223, April 2009.

[40] Todd M Kana and Patricia M Glibert. Effect of irradiances up to 2000 μe m- 2 s- 1 on marine synechococcus wh7803—i. growth, pigmentation, and cell composition. *Deep Sea Research Part A. Oceanographic Research Papers*, 34(4):479–495, 1987.

[41] Darren Kessner, Matt Chambers, Robert Burke, David Agus, and Parag Mallick. Proteowizard: open source software for rapid proteomics tools development. *Bioinformatics*, 24(21):2534–2536, 2008.

[42] Susan A Kimmance and Corina PD Brussaard. Estimation of viral-induced phytoplankton mortality using the modified dilution method. *Manual of aquatic viral ecology*, pages 65–73, 2010.

[43] Tobias Kind, Hiroshi Tsugawa, Tomas Cajka, Yan Ma, Zijuan Lai, Sajjan S Mehta, Gert Wohlgemuth, Dinesh Kumar Barupal, Megan R Showalter, Masanori Arita, et al. Identification of small molecules using accurate mass ms/ms search. *Mass Spectrometry Reviews*, 2017.

[44] David L Kirchman, Yoshimi Suzuki, Christopher Garside, and Hugh W Ducklow. High turnover rates of dissolved organic carbon during a spring phytoplankton bloom. *Nature*, 352(6336):612–614, 1991.

[45] Angela Landolfi, Andreas Oschlies, and R Sanders. Organic nutrients and excess nitrogen in the north atlantic subtropical gyre. *Biogeosciences Discussions*, 5(1):685–724, 2008.

[46] Brian P Lankadurai, Edward G Nagato, and Myrna J Simpson. Environmental metabolomics: an emerging approach to study organism responses to environmental stressors. *Environmental Reviews*, 21(3):180–205, 2013.

[47] Alla Lapidus, Gregory C Kettler, Adam C Martiny, Katherine Huang, Jeremy Zucker, Maureen L Coleman, Sebastien Rodrigue, Feng Chen, Alla Lapidus, Steven Ferriera, et al. Patterns and implications of gene gain and loss in the evolution of prochlorococcus. *PLoS Genetics*, 3(LBNL-846E), 2007.

[48] Debbie Lindell, Jacob D Jaffe, Zackary I Johnson, George M Church, and Sallie W Chisholm. Photosynthesis genes in marine viruses yield proteins during host infection. *Nature*, 438(7064):86–89, 2005.

[49] Hongbin Liu, Hector A Nolla, and Lisa Campbell. Prochlorococcus growth rate and contribution to primary production in the equatorial and subtropical north pacific ocean. *Aquatic Microbial Ecology*, 12(1):39–47, 1997.

[50] Michael Love, Simon Anders, and Wolfgang Huber. Differential analysis of count data–the deseq2 package. *Genome Biology*, 15:550, 2014.

[51] Michael Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biology*, 15(12):550, 2014.

[52] Katherine RM Mackey, Adina Paytan, Ken Caldeira, Arthur R Grossman, Dawn Moran, Matthew McIlvin, and Mak A Saito. Effect of temperature on photosynthesis and growth in marine synechococcus spp. *Plant physiology*, 163(2):815–829, 2013.

[53] Elizabeth L Mann, Nathan Ahlgren, James W Moffett, and Sallie W Chisholm. Copper toxicity and cyanobacteria ecology in the sargasso sea. *Limnology and Oceanography*, 47(4):976–988, 2002.

[54] Nicholas H Mann, Natalia Novac, Conrad W Mullineaux, Julie Newman, Shaun Bailey, and Colin Robinson. Involvement of an ftsh homologue in the assembly of functional photosystem i in the cyanobacterium synechocystis sp. pcc 6803. *FEBS letters*, 479(1-2):72–77, 2000.

[55] L Moore. Comparative physiology of synechococcus and prochlorococcus: influence of light and temperature on growth, pigments, fluorescence and absorptive properties. *Mar Ecol Prog Ser*, 116:259–275, 1995.

[56] Mary Ann Moran, Elizabeth B Kujawinski, Aron Stubbins, Rob Fatland, Lihini I Aluwihare, Alison Buchan, Byron C Crump, Pieter C Dorrestein, Sonya T Dyhrman, Nancy J Hess, et al. Deciphering ocean carbon in a changing world. *Proceedings of the National Academy of Sciences*, page 201514645, 2016.

[57] Jaclyn Murton, Aparna Nagarajan, Amelia Y Nguyen, Michelle Liberton, Harmony A Hancock, Himadri B Pakrasi, and Jerilyn A Timlin. Population-level coordination of pigment response in individual cyanobacterial cells under altered nitrogen levels. *Photosynthesis research*, 134(2):165–174, 2017.

[58] Amelia Y Nguyen, William P Bricker, Hao Zhang, Daniel A Weisz, Michael L Gross, and Himadri B Pakrasi. The proteolysis adaptor, nbla, binds to the n-terminus of β-phycocyanin: implications for the mechanism of phycobilisome degradation. *Photosynthesis research*, 132(1):95–106, 2017.

[59] Helena Osterholz, Gabriel Singer, Bernd Wemheuer, Rolf Daniel, Meinhard Simon, Jutta Niggemann, and Thorsten Dittmar. Deciphering associations between dissolved organic molecules and bacterial communities in a pelagic marine system. *The ISME Journal*, 2016.

[60] F Partensky, Jean Blanchot, and D Vaulot. Differential distribution and ecology of prochlorococcus and synechococcus in oceanic waters: a review. *BULLETIN-INSTITUT OCEANOGRAPHIQUE MONACO-NUMERO SPECIAL-*, pages 457–476, 1999.

[61] Luc Patiny and Alain Borel. Chemcalc: a building block for tomorrow's chemical infrastructure, 2013.

[62] Byron E. Pedler, Lihini I. Aluwihare, and Farooq Azam. Single bacterial strain capable of significant contribution to carbon cycling in the surface ocean. *Proceedings of the National Academy of Sciences*, 111(20):7202–7207, May 2014.

[63] Daniel Petras, Irina Koester, Ricardo Da Silva, Brandon M Stephens, Andreas Florian Haas, Craig E Nelson, Linda Wegley Kelly, Lihini I Aluwihare, and Pieter C Dorrestein. High-resolution liquid chromatography tandem mass spectrometry enables large scale molecular characterization of dissolved organic matter. *Frontiers in Marine Science*, 4:405, 2017.

[64] Justine Pittera, Florian Humily, Maxine Thorel, Daphné Grulois, Laurence Garczarek, and Christophe Six. Connecting thermal physiology and latitudinal niche partitioning in marine synechococcus. *The ISME journal*, 8(6):1221–1236, 2014.

[65] Richard J. Puxty, Andrew D. Millard, David J. Evans, and David J. Scanlan. Shedding new light on viral photosynthesis. *Photosynthesis Research*, 126(1):71–97, October 2015.

[66] Sheng Ren, Anna A Hinzman, Emily L Kang, Rhonda D Szczesniak, and Long Jason Lu. Computational and statistical analysis of metabolomics data. *Metabolomics*, 11(6):1492–1513, 2015.

[67] Gabrielle Rocap, Daniel L Distel, John B Waterbury, and Sallie W Chisholm. Resolution of prochlorococcus and synechococcus ecotypes by using 16s-23s ribosomal dna internal transcribed spacer sequences. *Applied and Environmental Microbiology*, 68(3):1180–1191, 2002.

[68] Sheila Roitman, Ellen Hornung, Jose Flores-Uribe, Itai Sharon, Ivo Feussner, and Oded Beja. Cyanophage-encoded lipid-desaturases: oceanic distribution, diversity and function. *bioRxiv*, page 109157, 2017.

[69] Christoph Ruttkies, Emma L Schymanski, Sebastian Wolf, Juliane Hollender, and Steffen Neumann. Metfrag relaunched: incorporating strategies beyond in silico fragmentation. *Journal of cheminformatics*, 8(1):3, 2016.

[70] Emily M Saad, Amelia F Longo, Luke R Chambers, Rixiang Huang, Claudia Benitez-Nelson, Sonya T Dyhrman, Julia M Diaz, Yuanzhi Tang, and Ellery D Ingall. Understanding marine dissolved organic matter production: Compositional insights from axenic cultures of thalassiosira pseudonana. *Limnology and Oceanography*, 2016.

[71] Joanna Sacharz, Samantha J Bryan, Jianfeng Yu, Nigel J Burroughs, Edward M Spence, Peter J Nixon, and Conrad W Mullineaux. Sub-cellular location of ftsh proteases in the cyanobacterium synechocystis sp. pcc 6803 suggests localised psii repair zones in the thylakoid membranes. *Molecular microbiology*, 96(3):448–462, 2015.

[72] David J Scanlan, Martin Ostrowski, Sophie Mazard, Alexis Dufresne, Laurence Garczarek, Wolfgang R Hess, Anton F Post, Martin Hagemann, I Paulsen, and Frédéric Partensky. Ecological genomics of marine picocyanobacteria. *Microbiology and Molecular Biology Reviews*, 73(2):249–299, 2009.

[73] Jinyu Shan, Ying Jia, Martha R. J. Clokie, and Nicholas H. Mann. Infection by the 'photosynthetic' phage S-PM2 induces increased synthesis of phycoerythrin in Synechococcus sp. WH7803. *FEMS Microbiology Letters*, 283(2):154–161, June 2008.

[74] Hilla Shemer and Karl G Linden. Aqueous photodegradation and toxicity of the polycyclic aromatic hydrocarbons fluorene, dibenzofuran, and dibenzothiophene. *Water research*, 41(4):853–861, 2007.

[75] E. B. Sherr and B. F. Sherr. Bacterivory and herbivory: Key roles of phagotrophic protists in pelagic food webs. *Microbial Ecology*, 28(2):223–235, September 1994.

[76] Télesphore Sime-Ngando. Environmental bacteriophages: viruses of microbes in aquatic ecosystems. *Roles and mechanisms of parasitism in aquatic microbial communities*, page 7, 2015.

[77] Dhananjaya P Singh, Ratna Prabha, Kamlesh K Meena, Lalan Sharma, and Arun K Sharma. Induced accumulation of polyphenolics and flavonoids in cyanobacteria under salt stress protects organisms through enhanced antioxidant activity. *American Journal of Plant Sciences*, 5(05):726, 2014.

[78] Christophe Six, Jean-Claude Thomas, Laurence Garczarek, Martin Ostrowski, Alexis Dufresne, Nicolas Blot, David J. Scanlan, and Frédéric Partensky. Diversity and evolution of phycobilisomes in marine Synechococcusspp.: a comparative genomics study. *Genome Biology*, 8:R259, December 2007.

[79] Colin A Smith, Grace O'Maille, Elizabeth J Want, Chuan Qin, Sunia A Trauger, Theodore R Brandon, Darlene E Custodio, Ruben Abagyan, and Gary Siuzdak. Metlin: a metabolite mass spectral database. *Therapeutic drug monitoring*, 27(6):747–751, 2005.

[80] Steven Smriga, Vicente I Fernandez, James G Mitchell, and Roman Stocker. Chemotaxis toward phytoplankton drives organic matter partitioning among marine bacteria. *Proceedings of the National Academy of Sciences*, page 201512307, 2016.

[81] Matthew B Sullivan, Maureen L Coleman, Peter Weigele, Forest Rohwer, and Sallie W Chisholm. Three prochlorococcus cyanophage genomes: signature features and ecological interpretations. *PLoS biology*, 3(5):e144, 2005.

[82] Curtis A Suttle. Marine viruses—major players in the global ecosystem. *Nature Reviews Microbiology*, 5(10):801–812, 2007.

[83] Marcel JW Veldhuis, Gijsbert W Kraay, Judith DL Van Bleijswijk, and Martien A Baars. Seasonal and spatial variability in phytoplankton biomass, productivity and growth in the northwestern indian ocean: the southwest and northeast monsoon, 1992–1993. *Deep Sea Research Part I: Oceanographic Research Papers*, 44(3):425–449, 1997.

[84] Mingxun Wang, Jeremy J Carver, Vanessa V Phelan, Laura M Sanchez, Neha Garg, Yao Peng, Don Duy Nguyen, Jeramie Watrous, Clifford A Kapono, Tal Luzzatto-Knaan, Carla Porto, Amina Bouslimani, Alexey V Melnik, Michael J Meehan, Wei-Ting Liu, Max Crusemann, Paul D Boudreau, Eduardo Esquenazi, Mario Sandoval-Calderon, Roland D Kersten, Laura A Pace, Robert A Quinn, Katherine R Duncan, Cheng-Chih Hsu, Dimitrios J Floros, Ronnie G Gavilan, Karin Kleigrewe, Trent Northen, Rachel J Dutton, Delphine Parrot, Erin E Carlson, Bertrand Aigle, Charlotte F Michelsen, Lars Jelsbak, Christian Sohlenkamp, Pavel Pevzner, Anna Edlund, Jeffrey McLean, Jorn Piel, Brian T Murphy, Lena Gerwick, Chih-Chuang Liaw, Yu-Liang Yang, Hans-Ulrich Humpf, Maria Maansson, Robert A Keyzers, Amy C Sims, Andrew R Johnson, Ashley M Sidebottom, Brian E Sedio, Andreas Klitgaard, Charles B Larson, Cristopher A Boya P, Daniel Torres-Mendoza, David J Gonzalez, Denise B Silva, Lucas M Marques, Daniel P Demarque, Egle Pociute, Ellis C O'Neill, Enora Briand, Eric J N Helfrich, Eve A Granatosky, Evgenia Glukhov, Florian Ryffel, Hailey Houson, Hosein Mohimani, Jenan J Kharbush, Yi Zeng, Julia A Vorholt, Kenji L Kurita, Pep Charusanti, Kerry L McPhail, Kristian Fog Nielsen, Lisa Vuong, Maryam Elfeki, Matthew F Traxler, Niclas Engene, Nobuhiro Koyama, Oliver B Vining, Ralph Baric, Ricardo R Silva, Samantha J Mascuch, Sophie Tomasi, Stefan Jenkins, Venkat Macherla, Thomas Hoffman, Vinayak Agarwal, Philip G Williams, Jingqui Dai, Ram Neupane, Joshua Gurr, Andres M C Rodriguez, Anne Lamsa, Chen Zhang, Kathleen Dorrestein, Brendan M Duggan, Jehad Almaliti, Pierre-Marie Allard, Prasad Phapale, Louis-Felix Nothias, Theodore Alexandrov, Marc Litaudon, Jean-Luc Wolfender, Jennifer E Kyle, Thomas O Metz, Tyler

Peryea, Dac-Trung Nguyen, Danielle VanLeer, Paul Shinn, Ajit Jadhav, Rolf Muller, Katrina M Waters, Wenyuan Shi, Xueting Liu, Lixin Zhang, Rob Knight, Paul R Jensen, Bernhard O Palsson, Kit Pogliano, Roger G Linington, Marcelino Gutierrez, Norberto P Lopes, William H Gerwick, Bradley S Moore, Pieter C Dorrestein, and Nuno Bandeira. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat Biotech*, 34(8):828–837, August 2016.

[85] David S Wishart, Timothy Jewison, An Chi Guo, Michael Wilson, Craig Knox, Yifeng Liu, Yannick Djoumbou, Rupasri Mandal, Farid Aziat, Edison Dong, et al. Hmdb 3.0—the human metabolome database in 2013. *Nucleic acids research*, 41(D1):D801–D807, 2012.

[86] Yukari Yoshida-Takashima, Mitsuhiro Yoshida, Hiroyuki Ogata, Keizo Nagasaki, Shingo Hiroishi, and Takashi Yoshida. Cyanophage Infection in the Bloom-Forming Cyanobacteria *Microcystis aeruginosa* in Surface Freshwater. *Microbes and Environments*, 27(4):350–355, 2012.

[87] Olga Zhaxybayeva, W Ford Doolittle, R Thane Papke, and J Peter Gogarten. Intertwined evolutionary histories of marine synechococcus and prochlorococcus marinus. *Genome biology and evolution*, 1:325–339, 2009.

[88] Beata Żyszka, Mirosław Anioł, and Jacek Lipok. Modulation of the growth and metabolic response of cyanobacteria by the multifaceted activity of naringenin. *PloS one*, 12(5):e0177631, 2017.