Dissolved organic matter produced by *Thalassiosira pseudonana*

Krista Longnecker, Melissa C. Kido Soule, and Elizabeth B. Kujawinski*.

Woods Hole Oceanographic Institution, Marine Chemistry and Geochemistry, Woods Hole, MA 02543, U.S.A.

For submission to: *Marine Chemistry*

Date submitted: March 11, 2014 ; 1st revised version submitted September 10, 2014, 2nd revision submitted October 28, 2014

Running title: Phytoplankton metabolomics

*Corresponding author. Mailing address: WHOI MS#4, Woods Hole, MA 02543. Phone: (508) 289-3493. Fax: (508) 457-2164. E-mail: ekujawinski@whoi.edu

Keywords: metabolomics, marine phytoplankton, dissolved organic matter

Abstract

Phytoplankton are significant producers of dissolved organic matter (DOM) in marine ecosystems but the identity and dynamics of this DOM remain poorly constrained. Knowledge on the identity and dynamics of DOM are crucial for understanding the molecular-level reactions at the base of the global carbon cycle. Here we apply emerging analytical and computational tools from metabolomics to investigate the composition of DOM produced by the centric diatom *Thalassiosira pseudonana*. We assessed both intracellular metabolites within *T. pseudonana* (the endo-metabolome) and extracellular metabolites released by *T. pseudonana* (the exo-metabolome). The intracellular metabolites had a more variable composition than the extracellular metabolites. We putatively identified novel compounds not previously associated with *T. pseudonana* as well as compounds that have previously been identified within *T. pseudonana*'s metabolic capacity (e.g. dimethylsulfoniopropionate and degradation products of chitin). The resulting information will provide the basis for future experiments to assess the impact of *T. pseudonana* on the composition of dissolved organic matter in marine environments.

1 Introduction

Autotrophic microbes play a central role in the global carbon cycle because they fix inorganic carbon into organic compounds. A fraction of this organic material is released into the surrounding environment as dissolved organic matter (DOM), where it supports microbial growth or is respired to carbon dioxide (del Giorgio and Cole, 1998; Kirchman, 2008). The rates of utilization or remineralization of individual compounds are determined by their structure, concentration, and the metabolic properties of ambient microorganisms (Azam and Worden, 2004). Thus, the molecular-level composition of DOM is an important factor in our understanding of the global carbon cycle. Despite the significance of photosynthesis in the production of organic matter, we know little about the molecular-level composition of photosynthetically-derived DOM and the environmental factors that govern its production (reviewed in: Carlson, 2002; Kujawinski, 2011).

Centric and pennate diatoms bloom in both coastal and open ocean settings where up to 40% of carbon fixation in marine ecosystems is attributed to these organisms (Nelson et al., 1995; Tréguer et al., 1995). The centric diatom *Thalassiosira pseudonana* has received significant attention as a laboratory model organism (Bowler et al., 2010). It was the first diatom with a completed genome, although function could only be determined for half of the genes (Armbrust et al., 2004). More recently, there have been genomic, transcriptomic, and proteomic investigations of *T. pseudonana* which have revealed dynamic responses to growth state, light, and nutrients (Dyhrman et al., 2012; Montsant et al., 2007; Norden-Krichmar et al., 2011; Nunn et al., 2009; Shi et al., 2013). Yet our knowledge of *T. pseudonana*'s impact on the composition of organic matter in marine environments has not been well-explored.

Metabolomics is an emerging analytical approach that seeks to characterize metabolites produced by an organism during growth or released following cell death. In targeted metabolomics, a limited set of known metabolites is quantified as a function of the process under study. In contrast, untargeted metabolomics investigations (e.g., Böttcher et al., 2008; Long et al., 2011) have no pre-defined list of metabolites and use qualitative, or semi-quantitative, mass spectrometry to examine all possible features. Untargeted metabolomics datasets are immense with thousands of resolved features, and informatics and statistical tools are employed to identify the subset of biologically relevant compounds (Patti et al., 2012). Although complete characterization is not feasible with available analytical methodologies, practitioners have used electrospray ionization (ESI) mass spectrometry (MS) and nuclear magnetic resonance spectrometry (NMR) to resolve and identify important molecules within plant systems (Iijima et al., 2008; Quanbeck et al., 2012) and within model microorganisms such as *Escherichia coli* (Rabinowitz and Kimball, 2007). These projects have provided valuable information on method development and computational tools which have allowed detailed examinations of the chemical interactions between biological entities and their habitats.

In the marine ecosystem, metabolic assessments of microorganisms have focused on phytoplankton such as cyanobacteria (Baran et al., 2010; Bennette et al., 2011) and diatoms (Paul et al., 2009). Experiments with these microbes have revealed that variability in phytoplankton-derived metabolites can be linked to growth stage (Barofsky et al., 2009; Vidoudez and Pohnert, 2012), nutrient limitation (Bromke et al., 2013), and are affected by the presence of co-cultured phytoplankton (Paul et al., 2009). Although recovery of targeted compounds has been used to optimize metabolite extraction and analysis methods (Bennette et al., 2011), structural characterization and identification of most metabolites remains challenging (Baran et al., 2010).

The goal of this project was an exploration of the molecular-level composition of metabolites produced by an autotrophic microorganism in order to characterize the metabolites released into the marine environment as a result of photoautotrophic processes. We extracted intracellular and extracellular metabolites from a laboratory culture of *T. pseudonana* and examined their composition over time with liquid chromatography coupled to ultrahigh resolution mass spectrometry (LC/FT-ICR-MS). Our analysis confirmed the presence of metabolites previously identified as part of *T. pseudonana*'s metabolic capacity as well as specific metabolites not previously known to occur in *T. pseudonana.*

2 Experimental section

2.1 Culturing Thalassiosira pseudonana

The diatom *Thalassiosira pseudonana* (CCMP culture #1335) was cultured axenically in a modified version of L1 media made with an artificial salt solution (Turks Island Salts) with extra silicate (212 µmol L^{-1}) and 10 µmol L^{-1} selenous acid. The cultures were initiated by adding 30 ml of *T. pseudonana* in exponential growth to twelve flasks with an additional six flasks serving as cell-free controls; each flask initially contained 300 ml of media. The cultures were incubated at 12ºC under a 12h:12h light:dark cycle. Samples were collected six hours into the light cycle on days 0, 1, 3, 7, 8, and 10. Three flasks were destructively sampled at each time: two replicates with *T. pseudonana* and one cell-free control. In order to characterize the temporal variability in DOM and include cell-free controls, we could not accommodate more than two replicates with *T. pseudonana* and the cell-free control for each time point.

2.2 Ancillary samples

At each time point, sample aliquots were removed for total organic carbon and nutrient analyses, and for cell counts. Unfiltered water samples for total organic carbon were acidified to

pH = 2 with 12 M hydrochloric acid, and stored at 4ºC until analysis on a Shimadzu TOC-V_{CSH} total organic carbon analyzer. The coefficient of variability between replicate injections was <1%. Comparisons to standards provided by Prof. D. Hansell (University of Miami) were made daily to verify that the measured concentrations of the standard fell within the consensus values for total organic carbon. The unfiltered water samples were also used to obtain concentrations of nitrate + nitrite, ammonium, silicate, and phosphate using a Lachat Instruments QuickChem 8000 continuous flow injection system. For cell counts, samples were fixed with 10% formaldehyde (final concentration) and stored at -80ºC until cells were counted using a Reichert hemocytometer. The formaldehyde-fixed cells were also stained with DAPI and viewed with an epifluorescence microscope to check for potential contamination by heterotrophic microorganisms. Contamination was not observed at any time point during the experiment.

2.3 Extraction of metabolites

Our initial experiments testing different extraction and mass spectrometry methods showed that extraction protocols appropriate for freshwater microorganisms such as a methanol/chloroform extraction (Winder et al., 2008) cannot be readily applied to marine organisms. The salt in seawater and the growth media is problematic for ESI mass spectrometry because salt suppresses the ionization of the organic molecules. For this reason, we were not able to analyze intracellular metabolites by direct infusion into the mass spectrometer. Rather, for both intracellular and extracellular metabolites, we opted for a reversed-phase LC/FT-MS method, in which salt co-elutes with the solvent front and is removed from compounds that are retained on the chromatography column. We adapted existing extraction and analysis methods to distinguish the organic compounds produced by *T. pseudonana* from the organic compounds in

the growth media. The methods for extracting intracellular and extracellular metabolites proceeded identically for the flasks with *T. pseudonana* and the cell-free controls.

The intracellular metabolites were extracted using a method developed by Rabinowitz and Kimball (2007). Briefly, 1.5 ml samples were centrifuged at 16,000 x *g* at 4ºC for 30 minutes and the supernatant discarded. The resulting cell pellet was extracted three times with ice-cold extraction solvent (acetonitrile:methanol:water with 0.1 M formic acid, 40:40:20). The combined extracts were neutralized with 0.1 M ammonium hydroxide, dried in a vacufuge, and then re-dissolved in 1 mL of 90:10 (v/v) water:acetonitrile for analysis on the mass spectrometer.

Prior to sampling the extracellular metabolites, the cells were removed by gentle vacuum filtration through 0.2 µm Omnipore filters (hydrophilic PTFE membranes, Millipore). Barofsky et al. (2009) have observed filtration may release intracellular metabolites into the exometabolome, and this potential bias must be considered in the discussion of our results. The acidified filtrate was extracted using solid phase extraction with PPL cartridges (Varian Bond Elut PPL cartridges) as previously described (Dittmar et al., 2008). After eluting with methanol, the extracts were dried in a vacufuge, and then re-dissolved in 1 mL 90:10 water:acetonitrile prior to analysis.

We quantified the extraction efficiency of the solid phase extraction resin in the following manner. The water:acetonitrile solution was dried completely using a vacufuge and re-dissolved in MilliQ water. The extract was then added to MilliQ water that had been acidified with hydrochloric acid and analyzed on a Shimadzu TOC-V_{CSH} total organic carbon analyzer as described in section 2.2. The carbon concentration from the extract was compared to the concentration of dissolved organic carbon in the filtrate to calculate the percent of organic carbon retained by the PPL cartridges.

2.4 Analysis of metabolites

All metabolomics analyses were conducted using liquid chromatography (LC) coupled by electrospray ionization to a hybrid linear ion trap - Fourier-transform ion cyclotron resonance (FT-ICR) mass spectrometer (7T LTQ FT Ultra, Thermo Scientific). Samples were stored at -20ºC until mass spectrometric analysis. Analysis of the extracts was conducted within 48 hours of sample processing, except for the extracts from day three which were analyzed six days after sample processing. LC separation was performed on a Synergi Fusion reversed-phase column using a binary gradient with solvent A being water with 0.1% formic acid and solvent B being acetonitrile with 0.1% formic acid. Samples were eluted at 250 μl min^{-1} with the following gradient: hold at 5% B for 0-2 min, ramp from 5 to 65% B between 2 and 20 min, ramp from 65 to 100% B between 20 and 25 min, hold at 100% B from 25-32 min, and then ramp back to 5% B between 32 and 32.5 min for re-equilibration (32.5-40 min). Both full MS and MS/MS data were collected. The MS scan was performed in the FT- ICR cell from *m/z* 100-1000 at 100,000 resolving power (defined at 400 *m/z*). In parallel to the FT acquisition, MS/MS scans were collected at nominal mass resolution in the ion trap from the two features with the highest peak intensities in each scan. Separate autosampler injections were made for analysis in positive and negative ion modes.

Electrospray and mass spectrometry conditions were initially optimized by infusing a mixture of metabolite standards in positive and negative ion modes. The list of compounds within this solution and sample spectra are given in Fig.S1. The majority of these standards preferentially ionize in either positive or negative ion mode. The LTQ FT Ultra was also externally calibrated weekly using a standard mixture of caffeine (Sigma Aldrich), L-methionyl-arginyl-phenylalanyl-alanine acetate (MRFA) (Sigma Aldrich), Ultramark 1621 (Alfa Aesar),

acetic acid (Sigma Aldrich), sodium dodecyl sulfate (Sigma Aldrich), and sodium taurocholate (Sigma Aldrich). The instrument has a mass accuracy of < 2 ppm after external calibration.

2.5 Processing of mass spectral data and feature identification

Data were collected as XCalibur RAW files which were converted to mzXML files using the msConvert tool within ProteoWizard (Chambers et al., 2012). Features were extracted from the LC-MS data using XCMS (Smith et al., 2006), where a feature is defined as a unique combination of a mass-to-charge (*m/z*) ratio and a retention time. Peak finding was performed with the centWave algorithm (Tautenhahn et al., 2008), and only peaks that fit a Gaussian shape were retained. Features were aligned across samples based on retention time and *m/z* value using the group.nearest function in XCMS; fillPeaks was used to reconsider features missed in the initial peak finding steps. CAMERA was used (1) to find compounds differing by adduct ion and stable isotope composition (Kuhl et al., 2012) and (2) to extract the intensities and *m/z* values for the associated MS/MS spectra. Finally, the list of features with their retention time, *m/z* value, and intensity from the extracted ion chromatographs (EIC peak heights) were exported to MATLAB for further processing. Positive and negative ion mode data were processed as separate datasets in XCMS and MATLAB.

In order to compare the data from the LC-based analysis with analyses generally done for direct infusion ESI FT-ICR MS data, we calculated the elemental formulas for the *m/z* values from the mzXML files processed by XCMS. We used the Compound Identification Algorithm developed by Kujawinski and colleagues (Kujawinski and Behn, 2006; Kujawinski et al., 2009) with a formula error of 1 ppm, and a relationship error of 20 ppm. The mass limit above which elemental formulas were assigned only by functional group relationships was 500 Da. Elements considered are C, H, O, N, S, and P. These elemental formulas were then divided into compound

187 classes that have been defined based on elemental ratios as approximated from data within
188 Hedges and Kim et al. (Hedges, 1990; Kim et al., 2003).

189 Several databases and *in silico* tools were consulted in order to make putative
190 identifications of select features from the untargeted metabolomics data. The databases included
191 the Madison Metabolomics Consortium Database (MMCD, Cui et al., 2008), METLIN (Smith et
192 al., 2005), and to a lesser extent MassBank (Horai et al., 2010), PubChem, and KEGG. The
193 database searches described in the present project allowed a 2 ppm window between the
194 measured and the database (calculated) *m/z* values.

195 ### *2.6 Statistical analysis*

196 Non-metric multidimensional scaling (NMS) (Kruskal, 1964; Mather, 1976) was used to
197 analyze variability in metabolite composition. For this analysis, only the compounds that were
198 not found in the controls were considered; the list was not pruned to remove isotopologues or
199 adducts. Differences between individual samples were calculated based on the presence or
200 absence of features with the Bray-Curtis distance measure using the Fathom toolbox (D.L. Jones,
201 pers. comm.). The statistics toolbox in MATLAB was used to run the NMS analyses. The
202 dimensionality of the data set was assessed by comparing 40 runs with real data to 50 runs with
203 randomized data. Additional axes were considered if the addition of the axis resulted in a
204 significant improvement over the randomized data (at $p \leq 0.05$) and the reduction in stress was
205 greater than 0.05. Stress is a metric of goodness of fit in NMS data, and thus large reductions in
206 stress indicate that the additional axis significantly improved the presentation of the data. The
207 proportion of variation represented by each axis was assessed by using a Mantel test to calculate
208 the coefficient of determination (r^2) between distance in the ordination space and distance in the
209 original space.

Model I regressions were used to quantify changes in EIC peak heights during the experiment. The non-parametric Spearman's rank correlation implemented in MATLAB was used to test changes in compound classes, cell abundance, TOC concentration, and inorganic nitrogen concentrations during the experiment.

3 Results and Discussion

3.1 Using metabolomics to assess the impact of an organism on its chemical environment

Metabolomics seeks to describe and quantify metabolites produced by organisms in response to their chemical microenvironment (Patti et al., 2012). To address this goal for the centric diatom, *T. pseudonana*, we analyzed our untargeted metabolomics data in two ways. First, we consider the pattern of shared metabolites in order to examine the similarities (or differences) between samples or along a time series. These comparisons do not require identification of unknown compounds; rather all detected intracellular and extracellular metabolites can be compared over time. Second, the list of *m*/*z* values was mined to obtain putative compound identifications. While there is still need for increased coverage of metabolomics databases (Kind et al., 2009), obtaining the identity of a compound greatly expands our ability to understand the chemical impact of microbes such as diatoms. For example, with a compound identification we can consider the environmental conditions that affect the concentration of a metabolite, describe the biochemical pathways in which it occurs, and examine the sources and sinks of this compound in the environment. This remains a major challenge in environmental metabolomics, and is one that cannot be achieved solely within the context of our work with *T. pseudonana*. In the following sections, we discuss the pattern of

compounds produced by *T. pseudonana* and then the compounds that we were able to putatively identify.

3.2 Temporal patterns in T. pseudonana *metabolites*

Over the course of the ten-day experiment, there were significant increases in the abundance of *T. pseudonana* and the concentration of total organic carbon concurrent with significant decreases in the inorganic nutrient concentrations (Spearman rank correlations, p-values all <<0.0001, Fig. S2). The inoculum into the flasks with *T. pseudonana* resulted in the transfer of organic compounds into the flasks at the beginning of this experiment as is apparent in the higher total organic carbon concentrations observed at the first sampling point in the flasks with *T. pseudonana* compared to the cell-free controls (Fig. S2). The cell-free controls did not show statistically significant changes over time in cell abundance, TOC, or inorganic nutrient concentrations (Spearman rank correlations, p-values > 0.05).

We use two measures to compare our extracellular extracts with previous research on dissolved organic matter. First, we consider the fraction of dissolved organic carbon that was recovered with solid phase extraction. We recovered between 14 and 41% of the organic compounds with the PPL cartridges, with increased extraction efficiencies at the later growth stages of the experiment (Fig. 1). By comparison, Becker et al. (2014) recovered between 2 and 24% of dissolved organic carbon from their phytoplankton cultures, with variability in the extraction efficiency correlated to phytoplankton phylogeny. Both studies observe efficiencies below the 40-60% extraction efficiency measured by Dittmar et al. (2008) for water samples from marine and estuarine sites. Second, *m/z* values from ultrahigh resolution mass spectrometry data have previously been sorted into compound classes based on their elemental ratios (e.g., Bhatia et al., 2010; Minor et al., 2012). Using this approach, the DOM from *T. pseudonana*

contains primarily protein-like, condensed hydrocarbon-like, and lipid-like compounds (Fig. 2); with lignin-like and carbohydrate-like compounds comprising less than 1% of the assigned elemental formulas. However, only a small fraction of the DOM could be sorted into compound classes based on elemental formulas. The proteins, hydrocarbons, and lipids from the positive ion mode data showed statistically significant increases with time; in negative ion mode, only the correlation between proteins and sampling time was statistically significant (Spearman's rho, p-values <0.05). While these compound classes do not distinguish between structural isomers, they provide a means to compare the composition of different samples. Here, we show that increases in *T. pseudonana* abundance are linked to higher numbers of protein-like, condensed hydrocarbon-like, and lipid-like compounds. These broad classifications, however, also highlight the limitations of analyses based on elemental formulas, which cannot distinguish between structural isomers. As we will note in section 3.4, we observed instances of the same *m/z* value at different retention times, confirming the presence of structural isomers in this dataset.

The number of features detected within the sample extracts analyzed with our untargeted metabolomics method showed slight variability during the course of the experiment (Table 1, Fig. S3). The data summarized in Table 1 includes all unique combinations of an *m/z* value and a retention time; specific compounds will appear in this list multiple times if they were observed with different adducts (e.g., $M\text{-}H^+$ or $M\text{-}Na^+$) or with ^{13}C substitutions (i.e., isotopologues). Almost 75% of all features were observed only in the treatments with *T. pseudonana* and were absent in the cell-free controls. The features found both in the treatments with *T. pseudonana* and in the cell-free controls were not analyzed further.

Throughout the experiment, we detected more features in the exometabolome compared to the endometabolome in both positive and negative ion modes (Fig. S3). There are several

278 factors that contribute to this observation. First, a smaller volume of sample was processed to
279 obtain the intracellular metabolites. Thus, metabolites present at lower concentrations may not
280 have sufficient signal strength to be detected by the mass spectrometer. Second, methodological
281 constraints required us to use a different extraction method to assess the endometabolome
282 compared to the exometabolome. This might have affected the number and type of features
283 retained and likely impacted the patterns of features observed within the intracellular and
284 extracellular metabolite pools. The primary goal of extraction methods for both intra- and
285 extracellular metabolites is the capture of the broadest suite of compounds at sufficient
286 concentration and with minimal salt interference. Due to the different sample matrices, this
287 requires two different extraction methods. In the intracellular methods, simple cell lysis liberates
288 a diverse pool of compounds and salt removal occurs during the LC step. For extracellular
289 methods, compounds must be concentrated from the saltwater media, requiring the use of solid-
290 phase extraction resins. De-salting occurs at the same time as extraction in this method. The
291 method for the intracellular metabolites is optimal for low molecular weight compounds that are
292 more polar compared to the slightly less polar, moderate molecular weight compounds which are
293 captured by the PPL solid-phase extraction cartridges used for the extracellular metabolites.
294 Nevertheless the boundaries of polarity and molecular weight are not exclusive to each method,
295 and we expected overlap in the compounds observed in the intracellular and extracellular
296 metabolites extracted from *T. pseudonana*. Yet, only a small number of compounds (9 in
297 negative ion mode and 35 in positive ion mode) from *T. pseudonana* were observed in both the
298 intracellular and extracellular metabolites (Table 1). Whether or not this was due to changes in
299 the metabolites after they were exuded from the cells cannot be determined based on our current
300 data. Previous research has noted that filtration of diatom cells may cause intracellular

metabolites to leak from cells which would bias characterization of extracellular metabolites (Barofsky et al., 2009). Of the limited research that has been done into metabolites of marine microorganisms, only two studies that we are aware of have attempted to examine both intracellular and extracellular metabolites (Baran et al., 2010; Rosselló-Mora et al., 2008), and neither of these studies assessed the overlap between the intracellular and extracellular metabolites in their organisms.

Changes in the extracted-ion-current (EIC) peak heights can be a semi-quantitative measure of the amount of a feature within a sample. EIC peak heights may vary because of (1) variability in the mass spectrometer, (2) ionization efficiency of the different compounds, and (3) the concentration of a compound within a sample. By definition, the present project took a finite amount of time and we opted to analyze the samples a constant length of time after extraction in order to minimize changes to the extract. This precludes analysis of the samples in a randomized fashion, which would reduce the impact of instrument variability. An alternative option is to group the sample extractions required for one project. This allows a pooled sample to be created which can constrain differences in EIC peak heights that are due to analytical variability (Dunn et al., 2011). Almost 9,000 features were observed in positive and negative ion mode and absent from the cell-free controls (Table 1, Fig. S3). A fraction of these features showed significant increases or decreases in EIC peak heights over the experiment (Table S1). More of the extracellular metabolites showed increases in EIC peak heights over time compared to the intracellular metabolites in both positive and negative ion modes (Table S1). Furthermore, a higher percentage of features decreased over time in the extracellular compared to the intracellular metabolites (Table S1). Such temporal variability over different growth stages in extracellular metabolites released by *T. pseudonana* has also been noted by Barofsky et al.

(2009). Yet, when all samples over the six sampling days were considered, the majority of features did not exhibit statistically significant changes in EIC peak heights during the experiment. The metabolites not exhibiting temporal variability could be compounds replenished by *T. pseudonana* at a constant rate or compounds that are not affected by changes in the growth conditions of the present project. While we cannot exclude the possibility that a subset of these metabolites were transferred with the inoculum at the beginning of the experiment, these features were absent from the controls and therefore not present in the media used to grow *T. pseudonana*.

3.3 Statistical analysis of occurrence patterns of metabolites

We used NMS to analyze the pattern of features found in the intracellular and extracellular metabolites. In positive ion mode, the NMS calculation (Fig. 3A and B) resulted in an ordination with a final stress of 0.13 and $r^2 = 0.89$ with more variability on axis one than on axis two (r^2 on axis 1 = 0.66, r^2 on axis 2 = 0.32). In negative ion mode, the NMS calculation (Fig. 3C and D) resulted in an ordination with a final stress of 0.12 and $r^2 = 0.78$ with more variability on axis one than on axis two (r^2 on axis 1 = 0.57, r^2 on axis 2 = 0.20). In both positive and negative ion modes, the NMS revealed a larger variability in the intracellular metabolites compared to the variability in the extracellular metabolites. In negative ion mode, this pattern was predominantly due to differences observed in one replicate on days 1, 8, and 10. This variability between replicates sampled on days 1, 8, and 10 was also evident in positive ion mode. We do not have an explanation for this inter-replicate variability for select days of the experiment, but the fact that it was observed in both positive and negative ion modes suggests either variability in biological activity or sample processing, and not analytical variability. The positive ion mode data exhibited greater differences among samples collected at all of the time

points. By day 7 of the experiment, the NMS revealed small changes in the composition of extracellular metabolites as shown by the tight clustering of symbols for days 7, 8, and 10 in Fig. 3. This indicates that the composition of metabolites produced by *T. pseudonana* at the conclusion of the experiment was less dynamic and fewer new compounds were being produced compared to the more variable composition of metabolites observed during exponential growth.

3.4 Annotating metabolites from T. pseudonana

A major challenge with an untargeted metabolomics assessment is the task of fully identifying the tens of thousands of features detected within a single dataset (Daly et al., 2014; Schymanski and Neumann, 2013). In the ideal case, these identifications are validated using authentic standards and multiple analytical methods coupled to iterative comparisons to different databases (Sumner et al., 2007). This labor-intensive process currently renders identification of the ~18,000 features found in the present project (Table 1) infeasible. Therefore, we culled our dataset to focus attention on those compounds that would have the highest potential for significant interest. To address our scientific goal of identifying compounds produced by *T. pseudonana* and subsequently released into the environment, we focused our attention on features detected in both intracellular and extracellular extracts. As described in the methods section, we required a feature (a) to be absent from the cell-free controls, and additionally required features (b) to be present in both replicates with *T. pseudonana* in order to increase our confidence in the observation of each feature, and (c) to be present at more than one time point in order to avoid considering transient features within the dataset. In the end, nine compounds in negative ion mode and 35 compounds in positive ion mode (Table 1) met these stringent criteria and we attempted to identify them based on exact mass and MS/MS data.

We used an iterative process to annotate and putatively identify the compounds, with comparisons to multiple databases. We used a classification scheme proposed by Sumner et al. (2007) to rate the strength of our putative metabolite identifications (Table 2). The strongest identifications, level 1, are those for which we have an authentic standard and have analyzed it on our mass spectrometer. Level 2 identifications are putatively annotated without chemical reference standards, but are based on spectral similarities with data from public or commercial libraries. Compounds with only a match based on *m/z* value are rated as level 3 classifications within the Sumner et al. (2007) format. Finally, level 4 classifications are unknown compounds. Searches based on comparisons of exact mass and the KEGG database have been previously used to help characterize organic compounds (Longnecker and Kujawinski, 2011; Romano et al., 2014; Suhre and Schmitt-Kopplin, 2008). Here, our first step in identification was comparison of exact mass values with masses of metabolites present in METLIN. In addition, for features with MS/MS fragmentation in our experiment, we compared our MS/MS spectra with METLIN database spectra to assign a putative identification to the feature. However, not all compounds in the METLIN database have associated MS/MS spectra. When no MS/MS data or matches to the METLIN database were available for a selected feature, we consulted the MMCD database. Finally, four of the features in positive ion mode were eliminated from consideration because the features had the same *m/z* value with different retention times, suggesting the presence of structural isomers. Identifying these features would require analysis of authentic standards of possible isomers to establish appropriate retention times and MS/MS data for comparison with our experimental dataset. In the following sections we discuss the implications of the compounds we were able to putatively identify (Table 2, Table S2).

3.5 *Observation of compounds previously associated with* T. pseudonana

Chitin is produced by a variety of marine organisms, and crustacean shells are the largest pool of chitin in the marine environment. After cellulose, chitin is the second most abundant biopolymer on earth yet its low levels in marine sediments imply that it is readily recycled within marine ecosystems despite its poor aqueous solubility (Gooday, 1990; Jeuniaux and Voss-Foucart, 1991). In *T. pseudonana*, the chitin is found in the cell wall (Brunner et al., 2009; Durkin et al., 2009) and flexible chitin fibers extend through the silica channels surrounding the cell (Hildebrand et al., 2009). In the present project, we observed both tri-N-acetylchitotriose (Fig. 4A) and chitobiose, which corroborates previous observations from culture experiments with *T. pseudonana* (Smucker and Dawson, 1986). The extracted ion chromatogram of tri-N-acetylchitotriose is shown in Fig. S4. The putative identification of tri-N-acetylchitotriose was well supported because the METLIN database provided a match for its exact mass and a match for the measured MS/MS spectra (Fig. 4B). Furthermore, the retention time for an authentic standard matched the retention time measured in the culture experiment. The *T. pseudonana* genome contains the synthetic pathway for chitin and 22 putative chitinases, which have a potential for chitin degradation (Armbrust et al., 2004). One hypothesis is that *T. pseudonana* degrades chitin to alter its sinking rate or to change the thickness of its cell wall to modulate the influx of compounds (Armbrust et al., 2004). An alternate hypothesis is that *T. pseudonana* does not express its chitinases for chitin degradation and that tri-N-acetylchitotriose and chitobiose are lost from chitin fibers during cellular growth. This hypothesis is consistent with culture experiments that did not reveal measurable levels of chitinase activity under different growth conditions (Štrojsová and Dyhrman, 2008). However, this would not explain the presence of multiple chitinases within the *T. pseudonana* genome. While we cannot distinguish between

these two hypotheses with the present dataset, the observed increases in tri-N-acetylchitotriose and possibly chitobiose suggest that the metabolic potential for chitin degradation in the genome could be realized during the growth of *T. pseudonana*. This is a good example of a case where genomic data are helpful in describing potential microbial metabolisms, yet metabolomics data are required to quantify the actual metabolic processes active within marine environments.

As the experiment progressed, *Thalassiosira* retained increasing amounts of intracellular dimethylsulfoniopropionate (DMSP; Fig. S5). DMSP was identified by exact *m/z* matches in the MMCD database and through analysis of an authentic standard. The identification of DMSP also indicates that we were able to distinguish chemical compounds that are important in marine environments, even with an untargeted metabolomics approach. DMSP is an organic sulfur compound that can act as an osmolyte for marine phytoplankton (Kirst, 1990). In addition, internal DMSP might scavenge potentially damaging reactive oxygen species or serve as a sink for carbon during periods of unbalanced growth (Stefels et al., 2007). In *T. pseudonana*, the production of DMSP is well-established (Keller et al., 1999) and increased amounts of DMSP are produced when the cells are nitrogen-limited (Bromke et al., 2013; Bucciarelli and Sunda, 2003; Franklin et al., 2012). In the present project, we did not add DSMP to the media and we did not detect DMSP in the cell-free controls. Yet, the amount of DMSP inside the cells increased during the experiment which might reflect *T. pseudonana*'s response to decreasing nitrogen availability, although the concentration of inorganic nitrogen always remained above 600 μM during the experiment (Fig. S2). While DMSP was observed at low levels in the external metabolites, DMSP is not efficiently recovered by PPL cartridges and thus its detection in this pool was likely underestimated significantly (W. Johnson, personal communication). The amount of DMSP produced by different species of phytoplankton varies over several orders of

magnitude, with diatoms having lower DMSP-to-carbon ratios than other phytoplankton (Stefels et al., 2007). However, our observation and the recent observations by Franklin et al. (2012) indicate that *Thalassiosira* may be important in the production of DMSP and therefore plays a role within the marine sulfur cycle.

3.6 *Identification of compounds not previously associated with* T. pseudonana

Compound identification is a major challenge in environmental metabolomics. In the following discussion, we present putatively annotated features, but recognize that there is still uncertainty associated with these identifications. Thus, a definitive identification will require additional verification before we can hypothesize as to the role of these compounds within the metabolism of *T. pseudonana*. Once we have confirmed the identity of these compounds, we can develop an appropriate quantitative assay and conduct laboratory experiments to address hypotheses about the importance of these compounds in the chemical ecology of *T. pseudonana*.

The first compound is bryotoxin A, which could only be identified with exact *m/z* matches in the MMCD database (Figure S6). While we collected an associated MS/MS spectrum, there were no corresponding listings in METLIN or MassBank which could help confirm the structure and identification. This compound is potentially interesting because it has not previously been observed in marine systems and studies on its toxicity are limited to experiments with cattle (McKenzie et al., 1987; McKenzie et al., 1989).

The second compound is a complex organic compound containing both iodine and chlorine. The putative elemental formula for this feature is $C_{22}H_{22}Cl_2I_2N_2O_7$ and we observed the feature primarily in the intracellular metabolites (Fig. 5). We observed the *m/z* values of both ^{35}Cl- and ^{37}Cl-isotopologues at the same retention time (Fig. S7), confirming the presence of chlorine. Macroalgae are the primary source of halogenated organic compounds in marine

ecosystems (Carpenter et al., 2000; Gschwend et al., 1985; Schall et al., 1994) and *T. pseudonana* releases CH_3I, a simple organic halogenated compound (Hughes et al., 2006). Yet, while *T. pseudonana* assimilates both iodide and iodate (de la Cuesta and Manley, 2009), to our knowledge this is the first observation that *T. pseudonana* may produce a complex organic compound containing both iodine and chlorine. In marine systems, between 40 and 90% of all soluble iodine-containing compounds are organic compounds (Gilfedder et al., 2008; Lai et al., 2011; Wong and Cheng, 1998), which play an important role in chemical ecology (Vanelslander et al., 2012) and marine atmospheric chemistry (O'Dowd et al., 2002; Saiz-Lopez et al., 2011).

We also observed over one hundred features that were tentatively identified as peptides (Table S3), mostly tripeptides (e.g., Arg-Tyr-Tyr) and a smaller number of dipeptides (e.g, Ala-Pro). Matches to *m*/*z* values in the METLIN database provide insight into the possible combinations of amino acids that compose the peptides we observed. However, peptide identification is challenging because (1) the amino acid sequence cannot be determined solely based on exact mass and (2) there are structural isomers among the twenty possible amino acids (He et al., 2004). While we will require additional analyses to identify the peptides we observed, the prevalence of *m*/*z* values matching peptides is noteworthy even without a putative identification of the amino acid sequences. The peptides were primarily present in the external metabolites, and were not always simultaneously present in the intracellular metabolite pool. Furthermore, plots of the peptides' EIC peak heights over time revealed statistically significant increases in peak heights (Pearson correlation coefficients with $p < 0.05$) for the majority of the peptides (Table S3). This suggests that most of the peptides were increasing in concentration over the course of the experiment.

The unexpected observation of a large number of different peptides raises the question as to why *T. pseudonana* releases so many different peptides and whether the peptides would be readily assimilated by marine heterotrophic microorganisms. Free amino acids have been quantified inside and outside of the cells in *Skeletonema costatum* cultures and they represented a minor fraction (<5%) of the released dissolved organic carbon (Granum et al., 2002). More recently, a metabolomic investigation of *Synechococcus* revealed the presence of amino acids and dipeptides both inside the cells and in the growth media (Baran et al., 2010). Thus, the peptides we observe during *T. pseudonana* growth may be a product of protein turnover within the cell which is subsequently released into the media. Dissolved peptides are easily hydrolyzed to their constituent amino acids and then are rapidly consumed by heterotrophic microorganisms (Hollibaugh and Azam, 1983; Kirchman and Hodson, 1986). As a result, any peptides released by *T. pseudonana* are likely to have a short residence time in the marine environment. Further work is needed to assess whether the peptides observed here will be consumed at different rates depending on the carbon, nitrogen, and sulfur demands of the *in situ* heterotrophic bacterial community.

Metabolomics lags far behind other 'omics' investigations and, until recently, metabolomics research has focused on the development of laboratory methods, analytical methods, and computational tools. In the present project we applied these emerging tools to characterize the organic compounds that *T. pseudonana* may release into the environment. Through this analysis we putatively identified compounds not previously associated with *T. pseudonana* metabolism. While there have been only limited metabolomic investigations into the impact of marine microorganisms on their chemical environment, the continued development of

computational tools and establishment of new databases will be crucial in facilitating the comparison of metabolites across organisms and ecosystems.

Acknowledgments

We thank the people at www.metabolomics-forum.com for answering questions about the use of XCMS and CAMERA, Matthew Monroe at PNNL for information about data file conversions, Erin Bertrand for discussions about the peptides, Paul Henderson for analyzing the nutrient samples, and Winn Johnson for information about the analysis of DMSP. Comments from the manuscript reviewers were much appreciated and helped clarify the presentation of the project. Instrumentation in the WHOI FT-MS facility was funded by the National Science Foundation MRI program (OCE-0619608) and by the Gordon and Betty T. Moore Foundation. This work was supported by NSF grant OCE-0928424 to EBK.

Figure Legends

Fig. 1. Extraction efficiency for the solid phase extraction of extracellular metabolites during the course of the experiment. The extraction efficiencies for the treatments with *T. pseudonana* is the percentage of the dissolved organic carbon retained by the PPL cartridges as a fraction of the measured dissolved organic carbon in the filtrate.

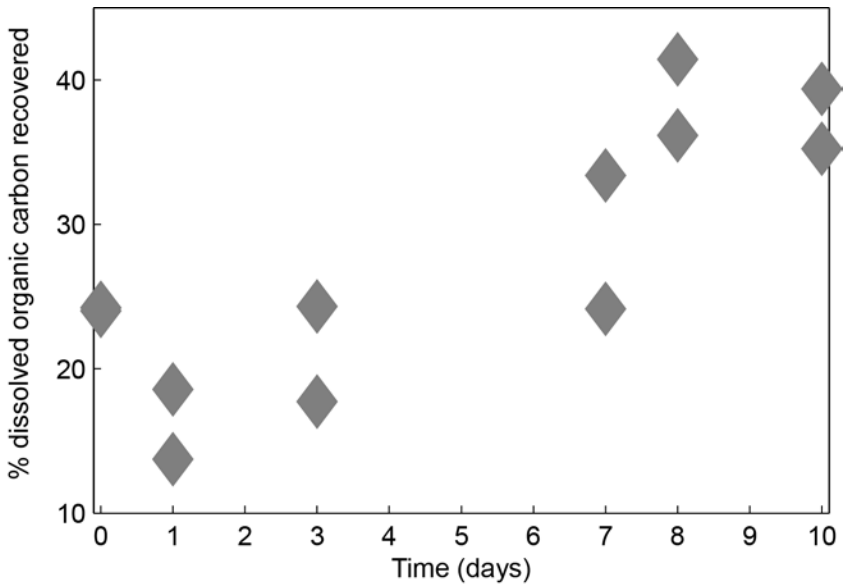
Fig. 2. The number of (A) protein-like, (B) condensed hydrocarbon-like, and (C) lipid-like compounds in the samples with *T. pseudonana* over the course of the experiment. The features within the cell-free controls were removed from the dataset prior to calculating the elemental formulas needed to define features based on their elemental ratios.

Fig. 3. NMS analysis based on presence or absence of features showing the differences in the composition of organic matter analyzed in (A and B) positive ion mode and (C and D) negative ion mode. The panels contain the same data coded differently to highlight (A and C) the differences between intracellular and extracellular metabolites or (B and D) the sampling time for the intracellular and extracellular metabolites.

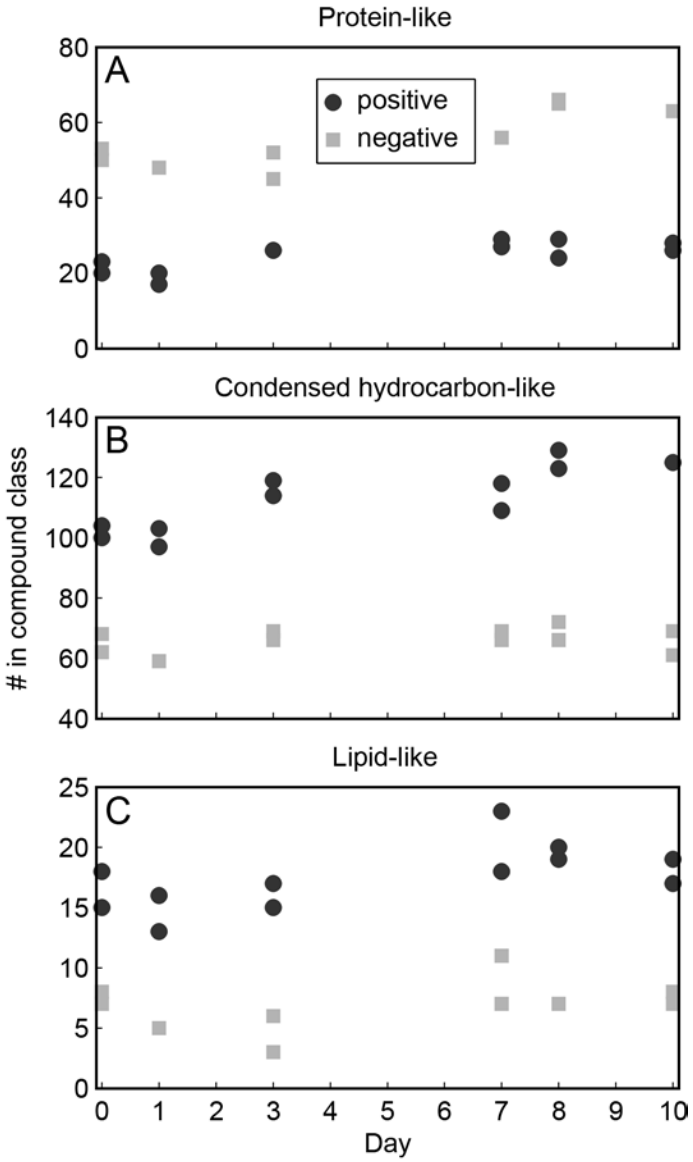
Fig. 4. Changes in the (A) EIC peak height of a feature putatively identified as tri-N-acetylchitotriose. Data from both replicates with *T. pseudonana* are shown in the figure. The identification was based on *m/z* value and (B) comparison of the MS/MS spectrum with data available in METLIN. The structure of the compound is given within panel (A). The MS/MS spectrum shown in (B) is that from our unknown feature and the table lists the *m/z* values and relative intensities given in METLIN.

Fig. 5. The organic compound potentially containing both chlorine and iodine was (A) observed in negative ion mode as $[M - 2H + Na]^-$. The structure of the compound as shown in PubChem is given in (B).

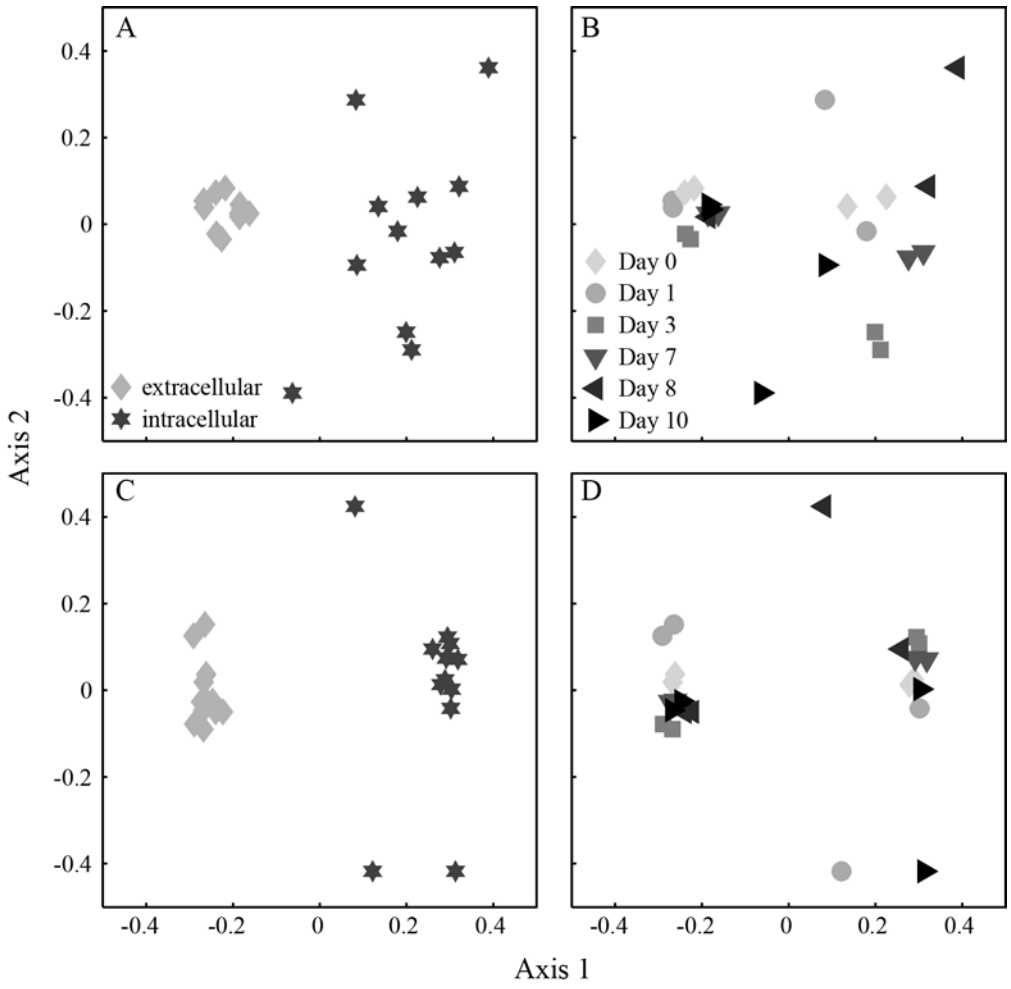
539 Longnecker, Kido Soule, Kujawinski
540 Fig. 1
541



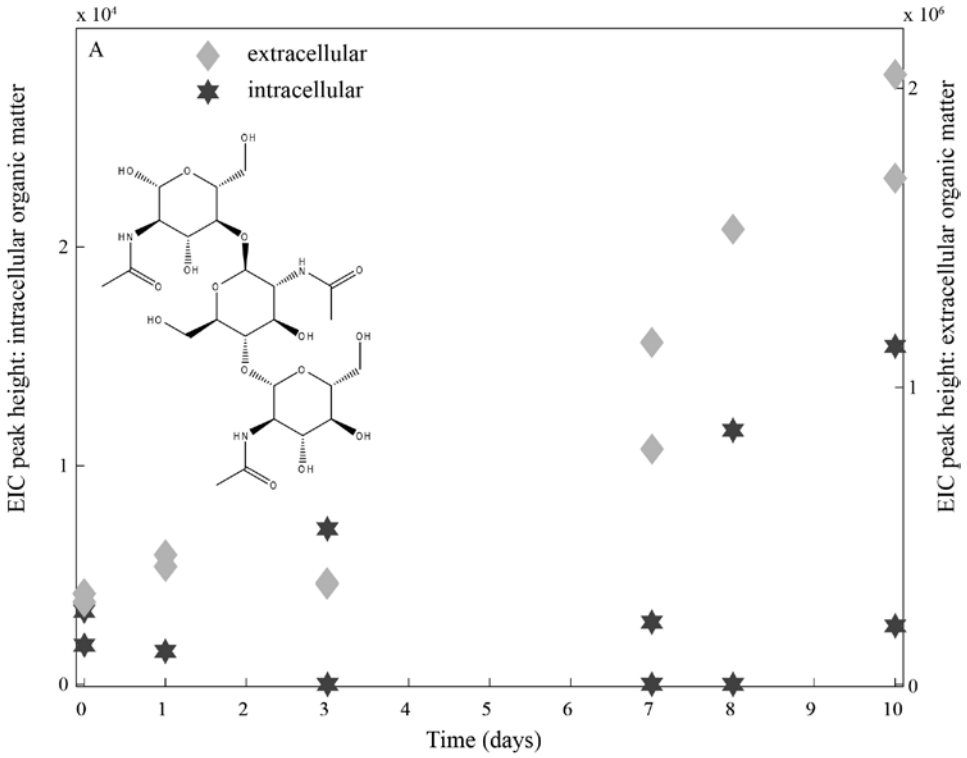
542 Longnecker, Kido Soule, Kujawinski
543 Fig. 2



545 Fig. 3


A
B
C
D
extracellular
intracellular
Day 0
Day 1
Day 3
Day 7
Day 8
Day 10
Axis 2
Axis 1
0.4
0.2
0
-0.2
-0.4

547 Fig. 4


A
extracellular
intracellular
x 10⁴
x 10⁶
EIC peak height: intracellular organic matter
EIC peak height: extracellular organic matter
Time (days)
B
407.08
204.08
222.17
425.17
481.08
MS2 listed in METLIN
204.0890 (84% intensity)
222.0940 (7% intensity)
407.1640 (100% intensity)
425.1790 (10% intensity)
Relative Abundance
m/z

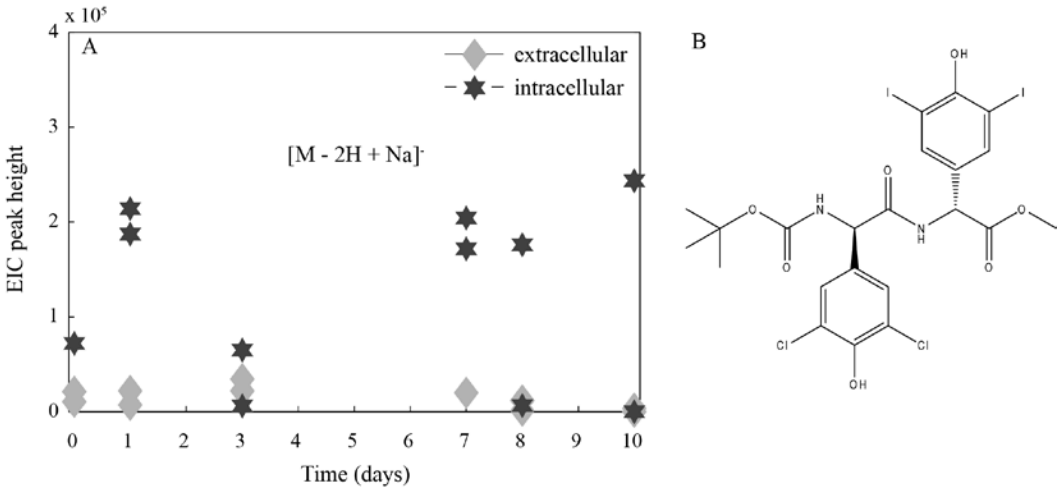
548

549 Longnecker, Kido Soule, Kujawinski

550 Fig. 5

551

552



553 Table 1. Summary of extracellular and intracellular metabolites exuded or retained by *T.*

554 *pseudonana*, as collected by LC/FT-MS in either positive or negative ion mode. Each feature is a

555 unique combination of an *m/z* value and a retention time.

	Negative	Positive
Total # of unique features	5484	12443
Number (and percent) of features remaining after deleting features found in the controls	4015 (73%)	9042 (73%)
# of features in the extracellular metabolites[§]	1630	2203
# of features in the intracellular metabolites[§]	1458	3685
# of features found in both the intracellular and extracellular metabolites[†]	9	35

556

557 [§]Features present in both replicates of either the extracellular or intracellular metabolites.

558 [†]Each feature had to be present in both replicates of the intracellular and extracellular metabolites

559 and present at more than one time point during the experiment.

560

561 Table 2. Details on the compounds putatively identified in the experiments with *T. pseudonana*.
562 The table includes the information on the measured *m/z* value and the ionization mode.
563 Additional details on each compound are presented in Table S2. Identification level is described
564 in the text and follows the convention proposed by Sumner et al. (2007).

Putative annotation	**Measured mass/charge**	**Ionization mode**	**Reference numbers**	**Identification level**
Tri-N-acetylchitotriose	628.255600	Positive	PubChem CID 444514	1
Chitobiose	425.176513	Positive	KEGG C01674	2
Dimethylsulfoniopropionate (DMSP)	135.047394	Positive	KEGG C04022	1
Bryotoxin A	619.275800	Positive	KEGG C08853	3
Organo-iodine compound	770.865479	Negative	PubChem CID 11535056	3

565

566

References

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

Azam, F. and Worden, A.Z., Microbes, molecules, and marine ecosystems. Science 2004;303:1622-4.

Baran, R., Bowen, B.P., Bouskill, N.J., Brodie, E.L., Yannone, S.M. and Northen, T.R., Metabolite identification in *Synechococcus sp. PCC 7002* using untargeted stable isotope assisted metabolite profiling. Anal Chem 2010;82:9034-42.

Barofsky, A., Vidoudez, C. and Pohnert, G., Metabolic profiling reveals growth stage variability in diatom exudates. Limnol Oceanogr Meth 2009;7:382-90.

Becker, J.W., Berube, P.M., Follett, C.L., Waterbury, J.B., Chisholm, S.W., DeLong, E.F. et al., Closely related phytoplankton species produce similar suites of dissolved organic matter. Frontiers in Microbiology 2014;5.

Bennette, N.B., Eng, J.F. and Dismukes, G.C., An LC-MS-based chemical and analytical method for targeted metabolite quantification in the model cyanobacterium *Synechococcus* sp. PCC 7002. Anal Chem 2011;83:3808-16.

Bhatia, M.P., Das, S.B., Longnecker, K., Charette, M.A. and Kujawinski, E.B., Molecular characterization of dissolved organic matter associated with the Greenland ice sheet Geochim Cosmochim Acta 2010;74:3768-84.

Böttcher, C., Edda von, R.-L., Schmidt, J., Schmotz, C., Neumann, S., Scheel, D. et al., Metabolome analysis of biosynthetic mutants reveals a diversity of metabolic changes and allows identification of a large number of new compounds in Arabidopsis. Plant Physiology 2008;147:2107-20.

Bowler, C., Vardi, A. and Allen, A.E., Oceanographic and biogeochemical insights from diatom genomes. Annu Rev Mar Sci 2010;2:333-65.

Bromke, M.A., Giavalisco, P., Willmitzer, L. and Hesse, H., Metabolic analysis of adaptation to short-term changes in culture conditions of the marine diatom *Thalassiosira pseudonana*. PLoS ONE 2013;8.

Brunner, E., Richthammer, P., Ehrlich, H., Paasch, S., Simon, P., Ueberlein, S. et al., Chitin-based organic networks: an integral part of cell wall biosilica in the diatom *Thalassiosira pseudonana*. Angew Chem Int Ed 2009;48:9724-7.

Bucciarelli, E. and Sunda, W.G., Influence of CO_2, nitrate, phosphate, and silicate limitation on intracellular dimethylsulfoniopropionate in batch cultures of the coastal diatom *Thalassiosira pseudonana*. Limnol Oceanogr 2003;48:2256-65.

Carlson, C.A., 2002. Production and removal processes. In: D.A. Hansell and C.A. Carlson (Editors), Biogeochemistry of marine dissolved organic matter. Academic Press, pp. 91-151.

Carpenter, L.J., Malin, G., Liss, P.S. and Kupper, F.C., Novel biogenic iodine-containing trihalomethanes and other short-lived halocarbons in the coastal East Atlantic. Global Biogeochem Cy 2000;14:1191-204.

Chambers, M.C., Maclean, B., Burke, R., Amodei, D., Ruderman, D.L., Neumann, S. et al., A cross-platform toolkit for mass spectrometry and proteomics. Nat Biotech 2012;30:918-20.

Cui, Q., Lewis, I.A., Hegeman, A.D., Anderson, M.E., Li, J., Schulte, C.F. et al., Metabolite identification via the Madison Metabolomics Consortium Database. Nat Biotech 2008;26:162-4.

Daly, R., Rogers, S., Wandy, J., Jankevics, A., Burgess, K.E.V. and Breitling, R., MetAssign: probabilistic annotation of metabolites from LC–MS data using a Bayesian clustering approach. Bioinformatics 2014.

de la Cuesta, J.L. and Manley, S.L., Iodine assimilation by marine diatoms and other phytoplankton in nitrate-replete conditions. Limnol Oceanogr 2009;54:1653-64.

del Giorgio, P.A. and Cole, J.J., Bacterial growth efficiency in natural aquatic systems. Annu Rev Ecol Syst 1998;29:503-41.

Dittmar, T., Koch, B., Hertkorn, N. and Kattner, G., A simple and efficient method for the solid-phase extraction of dissolved organic matter (SPE-DOM) from seawater. Limnol Oceanogr Meth 2008;6:230-5.

Dunn, W.B., Broadhurst, D., Begley, P., Zelena, E., Francis-McIntyre, S., Anderson, N. et al., Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. Nat Protocols 2011;6:1060-83.

Durkin, C.A., Mock, T. and Armbrust, E.V., Chitin in diatoms and its association with the cell wall. Eukaryot Cell 2009:EC.00079-09.

Dyhrman, S.T., Jenkins, B.D., Rynearson, T.A., Saito, M.A., Mercier, M.L., Alexander, H. et al., The transcriptome and proteome of the diatom *Thalassiosira pseudonana* reveal a diverse phosphorus stress response. PLoS ONE 2012;7:e33768.

Franklin, D.J., Airs, R.L., Fernandes, M., Bell, T.G., Bongaerts, R.J., Berges, J.A. et al., Identification of senescence and death in *Emiliania huxleyi* and *Thalassiosira pseudonana*: Cell staining, chlorophyll alterations, and dimethylsulphoniopropionate (DMSP) metabolism. Limnol Oceanogr 2012;57:305-17.

Gilfedder, B.S., Lai, S.C., Petri, M., Biester, H. and Hoffmann, T., Iodine speciation in rain, snow and aerosols. Atmos Chem Phys 2008;8:6069-84.

635 Gooday, G.W., 1990. The ecology of chitin degradation. In: K.C. Marshall, R.M. Atlas, J.G.
636 Jones and B.B. Jorgensen (Editors), Adv Microb Ecol. Plenum, New York, New York, pp. 387-
637 430.

638 Granum, E., Kirkvold, S. and Myklestad, S.M., Cellular and extracellular production of
639 carbohydrates and amino acids by the marine diatom *Skeletonema costatum*: diel variations and
640 effects of N depletion. Mar Ecol Prog Ser 2002;242:83-94.

641 Gschwend, P.M., Macfarlane, J.K. and Newman, K.A., Volatile halogenated organic compounds
642 released to seawater from temperate marine macroalgae. Science 1985;227:1033-5.

643 He, F., Emmett, M.R., Håkansson, K., Hendrickson, C.L. and Marshall, A.G., Theoretical and
644 experimental prospects for protein identification based solely on accurate mass measurement. J
645 Proteome Res 2004;3:61-7.

646 Hedges, J.I., 1990. Compositional indicators of organic acid sources and reactions in natural
647 environments. In: E.M. Perdue and E.T. Gjessing (Editors), Organic Acids in Aquatic
648 Ecosystems. John Wiley & Sons Ltd.

649 Hildebrand, M., Kim, S., Shi, D., Scott, K. and Subramaniam, S., 3D imaging of diatoms with
650 ion-abrasion scanning electron microscopy. J Struct Biol 2009;166:316-28.

651 Hollibaugh, J.T. and Azam, F., Microbial degradation of dissolved proteins in seawater. Limnol
652 Oceanogr 1983;28:1104-16.

653 Horai, H., Arita, M., Kanaya, S., Nihei, Y., Ikeda, T., Suwa, K. et al., MassBank: a public
654 repository for sharing mass spectral data for life sciences. J Mass Spectrom 2010;45:703-14.

655 Hughes, C., Malin, G., Nightingale, P.D. and Liss, P.S., The effect of light stress on the release
656 of volatile iodocarbons by three species of marine microalgae. Limnol Oceanogr 2006;51:2849-
657 54.

658 Iijima, Y., Nakamura, Y., Ogata, Y., Tanaka, K., Sakurai, N., Suzuki, T. et al., Metabolite
659 annotations based on the integration of mass spectral information. Plant J 2008;54:949 - 62.

660 Jeuniaux, C. and Voss-Foucart, M.F.o., Chitin biomass and production in the marine
661 environment. Biochem Syst Ecol 1991;19:347-56.

662 Keller, M.D., Kiene, R.P., Matrai, P.A. and Bellows, W.K., Production of glycine betaine and
663 dimethylsulfoniopropionate in marine phytoplankton. II. N-limited chemostat cultures. Mar Biol
664 1999;135:249-57.

665 Kim, S., Kramer, R.W. and Hatcher, P.G., Graphical method for analysis of ultrahigh-resolution
666 broadband mass spectra of natural organic matter, the van krevelen diagram. Anal Chem
667 2003;75:5336-44.

668 Kind, T., Scholz, M. and Fiehn, O., How large is the metabolome? A critical analysis of data
669 exchange practices in chemistry. PLoS ONE 2009;4:e5440.

670 Kirchman, D.L., 2008. Introduction and Overview. In: D.L. Kirchman (Editor), Microbial
671 Ecology of the Oceans. Wiley-Blackwell, pp. 1-44.

672 Kirchman, D.L. and Hodson, R.E., Metabolic regulation of amino acid uptake in marine waters.
673 Limnol Oceanogr 1986;31:339-50.

674 Kirst, G.O., Salinity tolerance of eukaryotic marine algae. Annu Rev Plant Phys 1990;41:21-53.

675 Kruskal, J.B., Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis.
676 Psychometrika 1964;29:1-27.

677 Kuhl, C., Tautenhahn, R., Böttcher, C., Larson, T.R. and Neumann, S., CAMERA: an integrated
678 strategy for compound spectra extraction and annotation of liquid chromatography/mass
679 spectrometry data sets. Anal Chem 2012;84:283-9.

680 Kujawinski, E.B., The impact of microbial metabolism on marine dissolved organic matter.
681 Annu Rev Mar Sci 2011;3:567-99.

682 Kujawinski, E.B. and Behn, M.D., Automated analysis of electrospray ionization Fourier-
683 transform ion cyclotron resonance mass spectra of natural organic matter. Anal Chem
684 2006;78:4363-73.

685 Kujawinski, E.B., Longnecker, K., Blough, N.V., Del Vecchio, R., Finlay, L., Kitner, J.B. et al.,
686 Identification of possible source markers in marine dissolved organic matter using ultrahigh
687 resolution mass spectrometry. Geochim Cosmochim Acta 2009;73:4384-99.

688 Lai, S.C., Williams, J., Arnold, S.R., Atlas, E.L., Gebhardt, S. and Hoffmann, T., Iodine
689 containing species in the remote marine boundary layer: The link to oceanic phytoplankton.
690 Geophys Res Lett 2011;38:L20801.

691 Long, J.Z., Cisar, J.S., Milliken, D., Niessen, S., Wang, C., Trauger, S.A. et al., Metabolomics
692 annotates ABHD3 as a physiologic regulator of medium-chain phospholipids. Nature chemical
693 biology 2011;7:763-5.

694 Longnecker, K. and Kujawinski, E.B., Composition of dissolved organic matter in groundwater.
695 Geochim Cosmochim Acta 2011;75:2752-61.

696 Mather, P.M., 1976. Computational methods of multivariate analysis in physical geography. J.
697 Wiley & Sons, London, 532 pp.

698 McKenzie, R.A., Franke, F.P. and Dunster, P.J., The toxicity to cattle and bufadienolide content
699 of six Bryophyllum species. Aust Vet J 1987;64:298-301.

700 McKenzie, R.A., Franke, F.P. and Dunster, P.J., The toxicity for cattle of bufadienolide cardiac
701 glycosides from *Bryophyllum tubiflorum* flowers. Aust Vet J 1989;66:374-6.

Minor, E.C., Steinbring, C.J., Longnecker, K. and Kujawinski, E.B., Characterization of dissolved organic matter in Lake Superior and its watershed using ultrahigh resolution mass spectrometry. Org Geochem 2012;43:1-11.

Montsant, A., Allen, A.E., Coesel, S., Martino, A.D., Falciatore, A., Mangogna, M. et al., Identification and comparative genomic analysis of signaling and regulatory components in the diatom *Thalassiosira pseudonana*. J Phycol 2007;43:585-604.

Nelson, D.M., Treguer, P., Brzezinski, M.A., Leynaert, A. and Queguiner, B., Production and dissolution of biogenic silica in the ocean - revised global estimates, comparison with regional data and relationship to biogenic sedimentation. Global Biogeochem Cy 1995;9:359-72.

Norden-Krichmar, T.M., Allen, A.E., Gaasterland, T. and Hildebrand, M., Characterization of the small RNA transcriptome of the diatom, *Thalassiosira pseudonana*. PLoS ONE 2011;6:e22870.

Nunn, B.L., Aker, J.R., Shaffer, S.A., Tsai, Y.H., Strzepek, R.F., Boyd, P.W. et al., Deciphering diatom biochemical pathways via whole-cell proteomics. Aquat Microb Ecol 2009;55:241-53.

O'Dowd, C.D., Jimenez, J.L., Bahreini, R., Flagan, R.C., Seinfeld, J.H., Hämeri, K. et al., Marine aerosol formation from biogenic iodine emissions. Nature 2002;417:632-6.

Patti, G.J., Yanes, O. and Siuzdak, G., Metabolomics: the apogee of the omics trilogy. Nat Rev Mol Cell Bio 2012;13:263-9.

Paul, C., Barofsky, A., Vidoudez, C. and Pohnert, G., Diatom exudates influence metabolism and cell growth of co-cultured diatom species. Mar Ecol Prog Ser 2009;389:61-70.

Quanbeck, S.M.M., Brachova, L., Campbell, A.A., Guan, X., Perera, A., He, K. et al., Metabolomics as a hypothesis-generating functional genomics tool for the annotation of *Arabidopsis thaliana* genes of "unknown function". Frontiers in Plant Science 2012;3.

Rabinowitz, J.D. and Kimball, E., Acidic acetonitrile for cellular metabolome extraction from *Escherichia coli*. Anal Chem 2007;79:6167-73.

Romano, S., Dittmar, T., Bondarev, V., Weber, R.J.M., Viant, M.R. and Schulz-Vogt, H.N., Exo-metabolome of *Pseudovibrio* sp. FO-BEG1 analyzed by ultra-high resolution mass spectrometry and the effect of phosphate limitation. PLoS ONE 2014;9:e96038.

Rosselló-Mora, R., Lucio, M., Peña, A., Brito-Echeverría, J., López-López, A., Valens-Vadell, M. et al., Metabolic evidence for biogeographic isolation of the extremophilic bacterium *Salinibacter ruber*. ISME J 2008;2:242-53.

Saiz-Lopez, A., Plane, J.M.C., Baker, A.R., Carpenter, L.J., von Glasow, R., Gómez Martín, J.C. et al., Atmospheric chemistry of iodine. Chem Rev 2011.

Schall, C., Laturnus, F. and Heumann, K.G., Biogenic volatile organoiodine and organobromine compounds released from polar macroalgae. Chemosphere 1994;28:1315-24.

737 Schymanski, E. and Neumann, S., The Critical Assessment of Small Molecule Identification
738 (CASMI): challenges and solutions. Metabolites 2013;3:517-38.

739 Shi, X., Gao, W., Chao, S.-h., Zhang, W. and Meldrum, D.R., Monitoring the single-cell stress
740 response of the diatom *Thalassiosira pseudonana* by quantitative real-time reverse rranscription-
741 PCR. Appl Environ Microbiol 2013;79:1850-8.

742 Smith, C.A., O'Maille, G., Want, E.J., Qin, C., Trauger, S.A., Brandon, T.R. et al., 2005.
743 METLIN: A Metabolite Mass Spectral Database 9th International Congress of Therapeutic Drug
744 Monitoring and Clinical Toxicology, Louisville, Kentucky.

745 Smith, C.A., Want, E.J., O'Maille, G., Abagyan, R. and Siuzdak, G., XCMS: processing mass
746 spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and
747 identification. Anal Chem 2006;78:779 - 87.

748 Smucker, R.A. and Dawson, R., Products of photosynthesis by marine phytoplankton: Chitin in
749 TCA "protein" precipitates. J Exp Mar Biol Ecol 1986;104:143-52.

750 Stefels, J., Steinke, M., Turner, S., Malin, G. and Belviso, S., Environmental constraints on the
751 production and removal of the climatically active gas dimethylsulphide (DMS) and implications
752 for ecosystem modelling. Biogeochemistry 2007;83:245-75.

753 Štrojsová, A. and Dyhrman, S.T., Cell-specific β-N-acetylglucosaminidase activity in cultures
754 and field populations of eukaryotic marine phytoplankton. FEMS Microbiol Ecol 2008;64:351-
755 61.

756 Suhre, K. and Schmitt-Kopplin, P., MassTRIX: mass translator into pathways. Nucleic Acids
757 Res 2008;36:W481-4.

758 Sumner, L., Amberg, A., Barrett, D., Beale, M., Beger, R., Daykin, C. et al., Proposed minimum
759 reporting standards for chemical analysis. Metabolomics 2007;3:211-21.

760 Tautenhahn, R., Bottcher, C. and Neumann, S., Highly sensitive feature detection for high
761 resolution LC/MS. BMC Bioinformatics 2008;9:504.

762 Tréguer, P., Nelson, D.M., Bennekom, A.J.V., DeMaster, D.J., Leynaert, A. and Quéguiner, B.,
763 The silica balance in the world ocean: a reestimate. Science 1995;268:375-9.

764 Vanelslander, B., Paul, C., Grueneberg, J., Prince, E.K., Gillard, J., Sabbe, K. et al., Daily bursts
765 of biogenic cyanogen bromide (BrCN) control biofilm formation around a marine benthic
766 diatom. Proc Natl Acad Sci USA 2012.

767 Vidoudez, C. and Pohnert, G., Comparative metabolomics of the diatom *Skeletonema marinoi* in
768 different growth phases. Metabolomics 2012;8:654-69.

769 Winder, C.L., Dunn, W.B., Schuler, S., Broadhurst, D., Jarvis, R., Stephens, G.M. et al., Global
770 metabolic profiling of *Escherichia coli* cultures: an evaluation of methods for quenching and
771 extraction of intracellular metabolites. Anal Chem 2008;80:2939-48.

772 Wong, G.T.F. and Cheng, X.-H., Dissolved organic iodine in marine waters: Determination,
773 occurrence and analytical implications. Mar Chem 1998;59:271-81.
774
775

Dissolved organic matter produced by *Thalassiosira pseudonana*

Supporting Information

Krista Longnecker, Melissa C. Kido Soule, and Elizabeth B. Kujawinski*.

Woods Hole Oceanographic Institution, Marine Chemistry and Geochemistry, Woods Hole, MA 02543, U.S.A.

TITLE RUNNING HEAD: Phytoplankton metabolomics

*Corresponding author. Mailing address: WHOI MS#4, Woods Hole, MA 02543. Phone: (508) 289-3493. Fax: (508) 457-2164. E-mail: ekujawinski@whoi.edu

Table S1. Percent of features with changes in EIC peak heights over time for features found in the samples and absent from the cell-free controls. Only statistically-significant (Model I regressions with $p \leq 0.05$) increases or decreases are included in the table. The total # of features is also shown in Table 1 and is the number of features present in both replicates of either the extracellular or intracellular metabolite samples and absent from the cell-free controls.

	Positive		**Negative**	
	Intracellular	**Extracellular**	**Intracellular**	**Extracellular**
% features increased	4%	67%	5%	25%
% features decreased	7%	13%	6%	8%
Total # features	3685	2203	1458	1630

Table S2. Summary of the features of interest which could be putatively identified. The table indicates whether the features were detected in positive (pos) or negative (neg) ion mode, and what charged ion was detected. Error indicates the absolute difference between the observed *m/z* value and the calculated *m/z* value. MS/MS spectra were available for some of the features, and the comments include additional information about each feature. Reference numbers can be used to find additional information about each compound in the indicated database.

Putative annotation	Elemental Formula (exact mass)	Ion mode	Detected as?	Error (ppm)	MS/MS ?	Reference numbers	Comments
Tri-N-acetylchitotriose	$C_{24}H_{41}N_3O_{16}$ (627.248682)	Pos	$[M+H]^+$	0.57	Yes	PubChem CID 444514	1 match at METLIN with MS/MS data corresponding to the observed MS/MS data
Chitobiose	$C_{16}H_{28}N_2O_{11}$ (424.169310)	Pos	$[M+H]^+$	0.17	No	KEGG C01674	Same retention time as tri-N-acetylchitotriose
Dimethylsulfoniopropionate (DMSP)	$C_5H_{10}O_2S$ (134.040150)	Pos	$[M+H]^+$	0.24	Yes	KEGG C04022	1 match at MMCD; no match at METLIN
Bryotoxin A	$C_{32}H_{42}O_{12}$ (618.267627)	Pos	$[M+H]^+$	1.45	No	KEGG C08853	1 match at MMCD; no match at METLIN
Organo-iodine compound*	$C_{22}H_{22}Cl_2I_2N_2O_7$ (749.889361)	Neg	$[M-2H+Na]^-$	1.88	No	PubChem CID 11535056	No matches at MMCD or METLIN; found isotopes for ^{37}Cl

* The full name for the organo-iodine compound is: Methyl (2R)-2-[[(2R)-2-(3,5-dichloro-4-hydroxyphenyl)-2-[(2-methylpropan-2-yl)oxycarbonylamino]acetyl]amino]-2-(4-hydroxy-3,5-diiodophenyl)acetate

Table S3. Summary of peptide data from the intracellular and extracellular metabolites analyzed in positive or negative ion mode. Both dipeptides and tripeptides were observed. Most of the peptides showed increases in EIC peak heights over the course of the experiment.

	Positive		**Negative**	
	Intracellular	**Extracellular**	**Intracellular**	**Extracellular**
Total # of peptide matches	11	88	2	26
# compounds increased during experiment	6	68	2	23
# compounds decreased during experiment	5	20	0	3

Fig. S1. We used a mixture of metabolites to optimize the LC/FT-ICR-MS parameters. This solution consisted of L-methionine, L-proline, L-arginine, L-glutamic acid, L-glutamine, L-threonine, caffeine, n-acetyl D-glucosamine, riboflavin, biotin, thymidine, NAD, succinic acid, malic acid, orotic acid, phosphoenolpyruvate, citric acid, glucose 6-phosphate, fructose 1,6-bisphosphate, and sodium taurocholate. The total ion chromatograph is shown for (A) the analysis of the metabolite mix and a solvent blank (90:10 water:acetonitrile) in negative ion mode, (B) the analysis of the metabolite mixture and a solvent blank in positive ion mode. The text lists the metabolites (and retention time, in seconds) for each ionization mode.

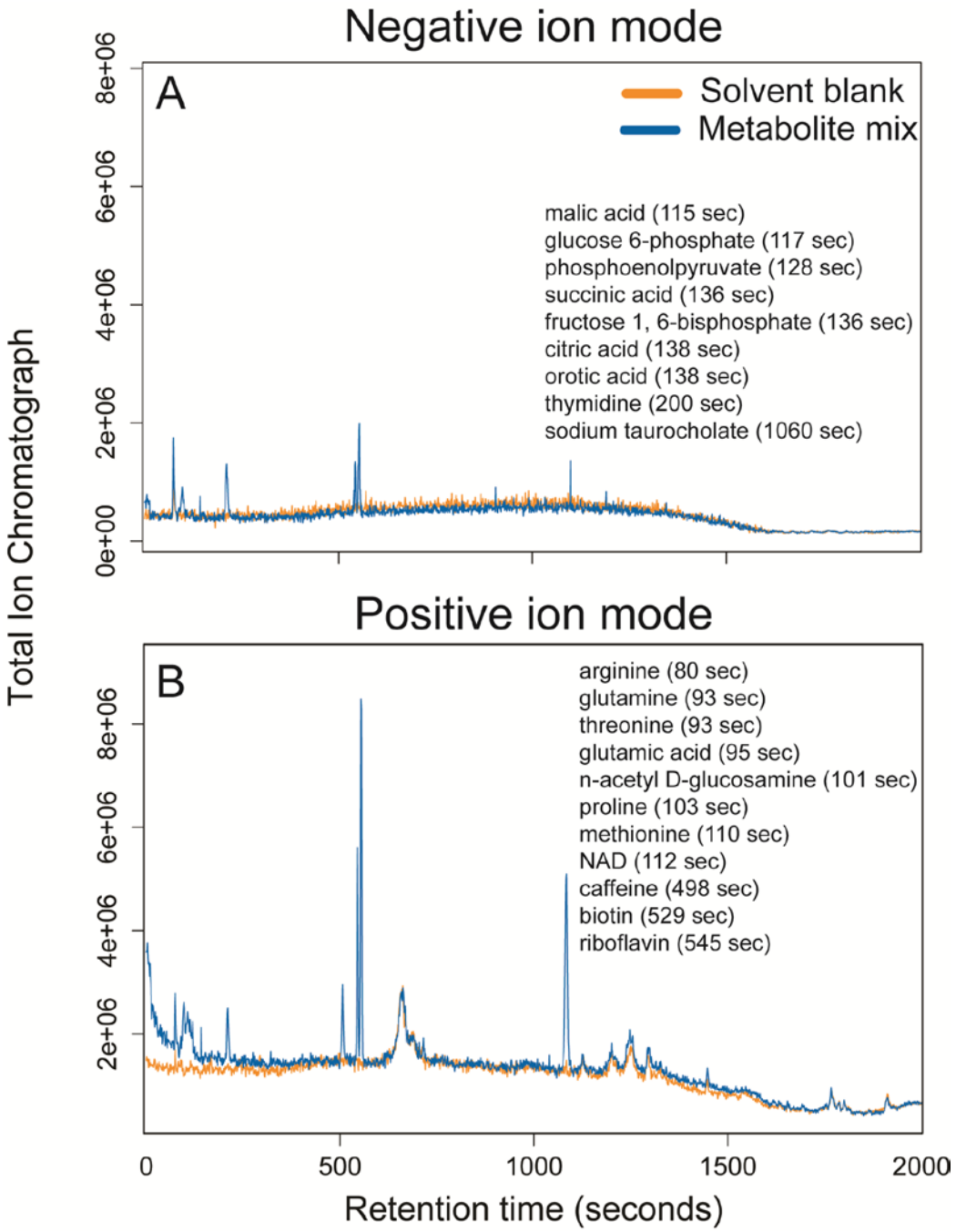


Fig. S2. During the incubation period, there were increases in (A) the abundance of *T. pseudonana* and (B) the concentration of total organic carbon within the flasks. There were no changes in the cell-free controls over the experiment, nor was there evidence of contamination by heterotrophic bacterial cells. The approximately 200 μM of total organic carbon in the cell-free controls is due to the presence of vitamins and EDTA that are required by *T. pseudonana* for cell growth. (C) There was a decrease in the concentration of nitrate+nitrite in the flasks with *T. pseudonana* indicating the consumption of inorganic nutrients concurrent with increases in cellular abundance. Silicate and phosphate showed similar patterns in concentration (data not shown).

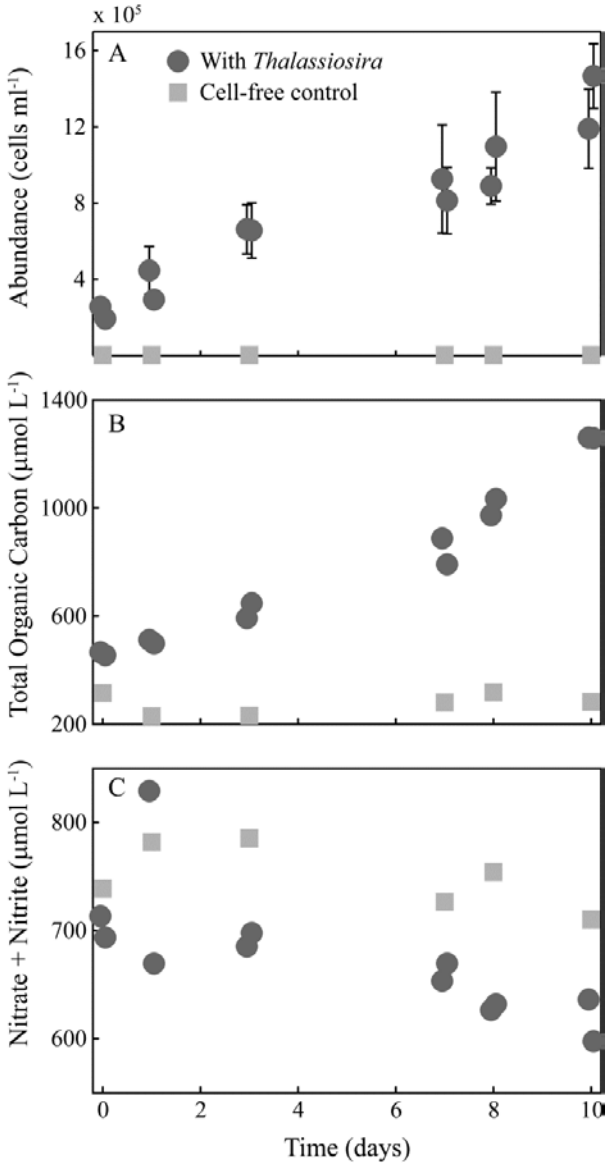


Fig. S3. The number of features (unique combinations of *m/z* value and retention time) for the extracellular and intracellular metabolites during the experiment from (A) positive ion mode and (B) negative ion mode. The points have been jiggered on the time axis to reduce overlap between sample points. A low number of features was observed in one of the extracellular metabolite samples at day 8 within the positive ion mode data. Inspection of the raw data revealed a problem with sample injection and this sample was not considered further.

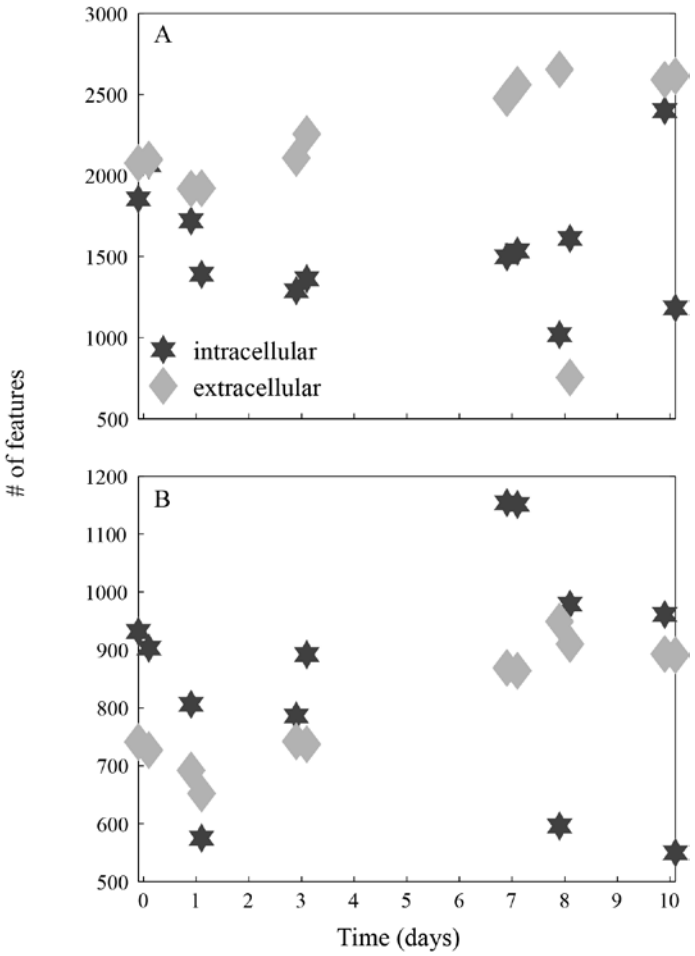


Fig. S4 : Extracted ion chromatogram (EIC) for tri-N-acetylchitotriose which had a measured *m/z* value of 628.52260. XCMS was used to process the data files generated by the LC-FT system. The dark lines are from samples with *T. pseudonana* while the lighter lines are the cell-free controls.

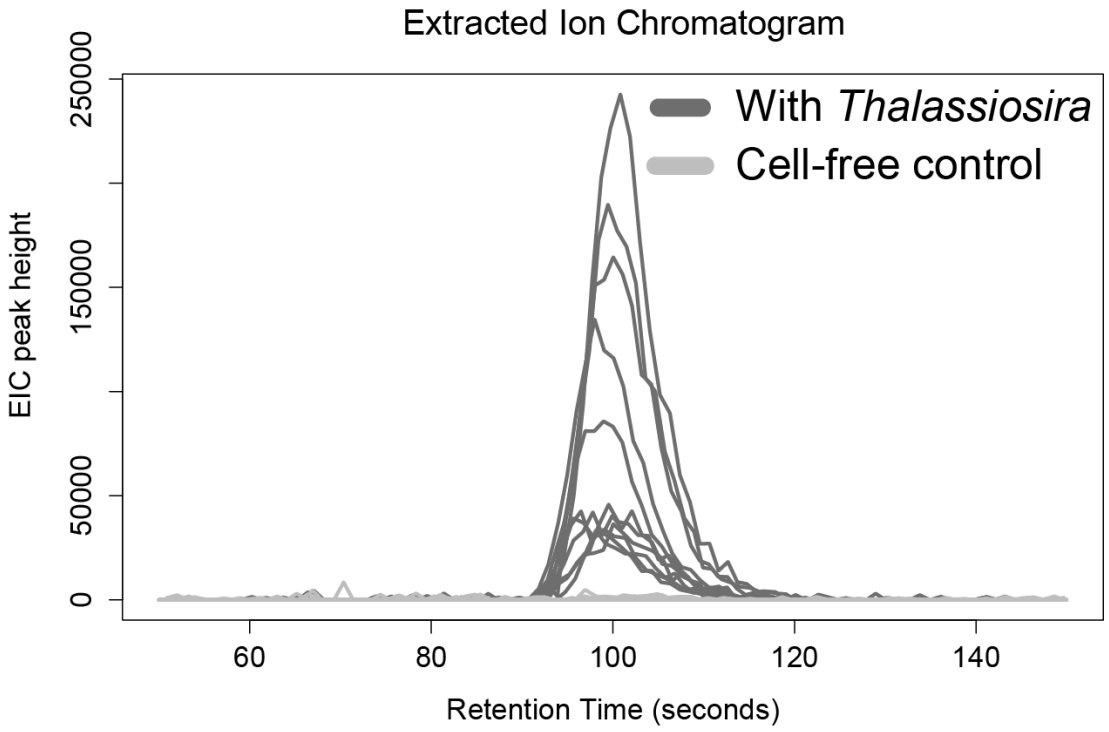


Fig. S5. DMSP was putatively identified in positive ion mode in both the intracellular and extracellular metabolites. The analysis failed for one of the replicates on day 8, and that sample is not plotted on the figure. Note the scale difference in the EIC peak heights between the intracellular and extracellular metabolites. The structure of the compound is also shown.

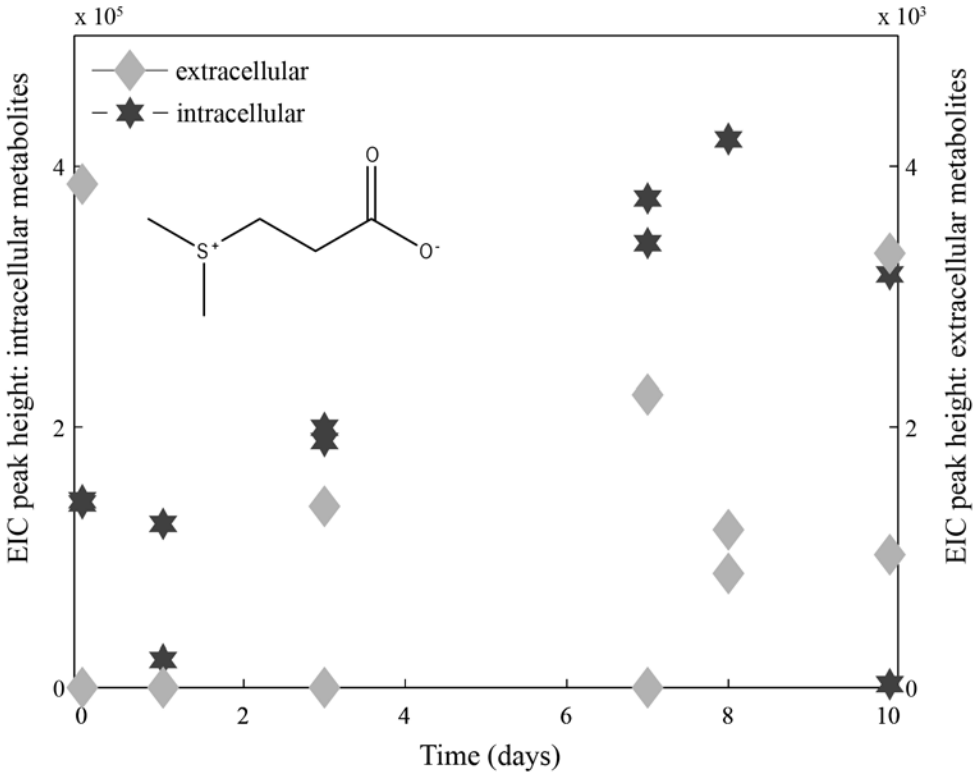


Fig. S6. A feature putatively annotated as bryotoxin A was observed in positive ion mode in both the intracellular and extracellular metabolites. Note the scale difference in the EIC peak heights between the intracellular and extracellular metabolites. The inset shows the structure of bryotoxin A, and the changes in EIC peak heights over the course of the experiment.

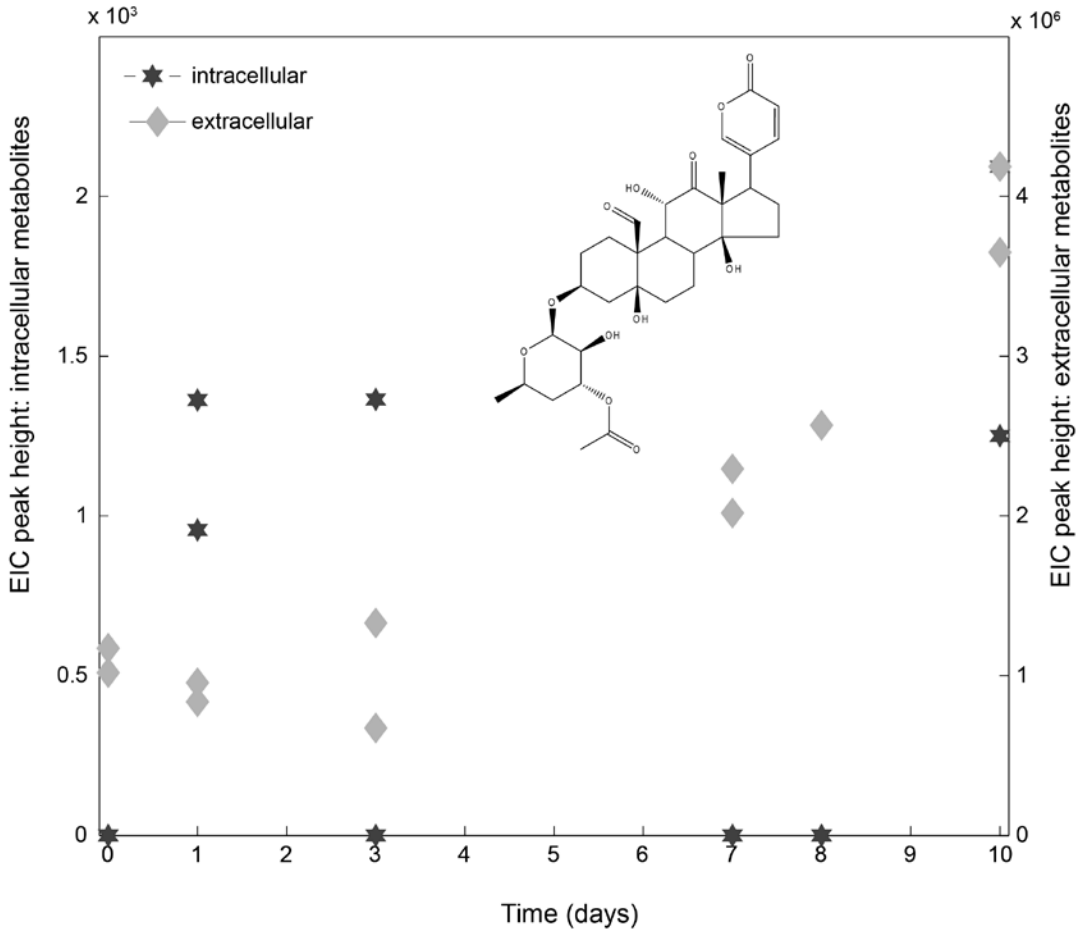


Fig. S7. EIC data for the organo-iodine compound ($C_{22}H_{22}Cl_2I_2N_2O_7$) observed in negative ion mode at (A) day zero and (B) day ten of the experiment. The data in green show the compound with two ^{35}Cl molecules, while the data in orange is the EIC data for a feature that putatively has one ^{35}Cl and one ^{37}Cl. In the environment, chlorine molecules are 75% ^{35}Cl and 25% ^{37}Cl, and our data show the ^{35}Cl organo-iodine compound had higher EIC peak heights supporting our putative identification of a compound containing chlorine.