View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by University of Birmingham Research Portal

UNIVERSITY OF BIRMINGHAM

Research at Birmingham

Using community metabolomics as a new approach to discriminate marine microbial particulate organic matter in the western English Channel

Llewellyn, Carole A.; Sommer, Ulf; Dupont, Chris L.; Allen, Andrew E.; Viant, Mark

DOI:
10.1016/j.pocean.2015.04.022

License:
Creative Commons: Attribution-NonCommercial-NoDerivs (CC BY-NC-ND)

Document Version
Peer reviewed version

Citation for published version (Harvard):
Llewellyn, CA, Sommer, U, Dupont, CL, Allen, AE & Viant, MR 2015, 'Using community metabolomics as a new approach to discriminate marine microbial particulate organic matter in the western English Channel', Progress in Oceanography, vol. 137, pp. 421-433. https://doi.org/10.1016/j.pocean.2015.04.022

Link to publication on Research at Birmingham portal

Publisher Rights Statement:
NOTICE: this is the author's version of a work that was accepted for publication in Progress in Oceanography. Changes resulting from the publishing process, such as peer review, editing, corrections, structural formatting, and other quality control mechanisms may not be reflected in this document. Changes may have been made to this work since it was submitted for publication. A definitive version was subsequently published in Progress in Oceanography, early online May 2015, DOI: 10.1016/j.pocean.2015.04.022.

After embargo period this copy of the work is subject to a Creative Commons Attribution Non-Commercial No-Derivatives license.

Checked June 2015

General rights
Unless a licence is specified above, all rights (including copyright and moral rights) in this document are retained by the authors and/or the copyright holders. The express permission of the copyright holder must be obtained for any use of this material other than for purposes permitted by law.

• Users may freely distribute the URL that is used to identify this publication.
• Users may download and/or print one copy of the publication from the University of Birmingham research portal for the purpose of private study or non-commercial research.
• User may use extracts from the document in line with the concept of 'fair dealing' under the Copyright, Designs and Patents Act 1988 (?)
• Users may not further distribute the material nor use it for the purposes of commercial gain.

Where a licence is displayed above, please note the terms and conditions of the licence govern your use of this document.

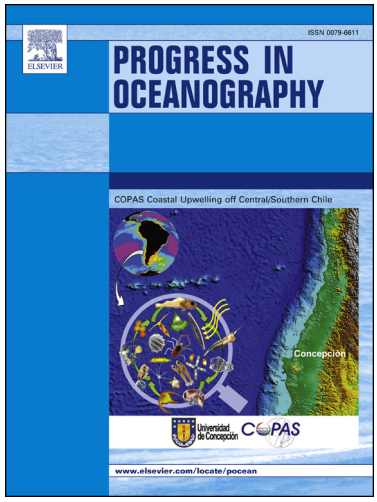
When citing, please reference the published version.

Take down policy
While the University of Birmingham exercises care and attention in making items available there are rare occasions when an item has been uploaded in error or has been deemed to be commercially or otherwise sensitive.

If you believe that this is the case for this document, please contact UBIRA@lists.bham.ac.uk providing details and we will remove access to the work immediately and investigate.

Download date: 01. Feb. 2019

Accepted Manuscript



Using community metabolomics as a new approach to discriminate marine microbial particulate organic matter in the western English Channel

Carole A. Llewellyn, Ulf Sommer, Chris L. Dupont, Andrew E. Allen, Mark R. Viant

PII: S0079-6611(15)00089-0
DOI: http://dx.doi.org/10.1016/j.pocean.2015.04.022
Reference: PROOCE 1576

To appear in: *Progress in Oceanography*

Please cite this article as: Llewellyn, C.A., Sommer, U., Dupont, C.L., Allen, A.E., Viant, M.R., Using community metabolomics as a new approach to discriminate marine microbial particulate organic matter in the western English Channel, *Progress in Oceanography* (2015), doi: http://dx.doi.org/10.1016/j.pocean.2015.04.022

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

ACCEPTED MANUSCRIPT

Using community metabolomics as a new approach to discriminate marine microbial particulate organic matter in the western English Channel

Carole A. Llewellyn*[a,b], Ulf Sommer *[c], Chris L. Dupont[d], Andrew E. Allen[d], Mark R. Viant[c]

*CAL and US joint first authors.

[a]. Plymouth Marine Laboratory, Prospect Place, The Hoe, Plymouth, PL1 3DH, UK.
[b]. current address: Centre for Sustainable Aquatic Research, Swansea University, Swansea, SA2 8PP, UK.
[c].NERC Biomolecular Analysis Facility – Metabolomics Node (NBAF-B), School of Biosciences, University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK.
[d]. Microbial and Environmental Genomics Group, J. Craig Venter Institute, San Diego, CA 92104

ABSTRACT

Metabolomics provides an unbiased assessment of a wide range of metabolites and is an emerging 'omics technique in the marine sciences. We use 'non-targeted' community metabolomics to determine patterns in metabolite profiles associated with particulate organic matter (POM) at four locations from two long-term monitoring stations (L4 and E1) in the western English Channel. The polar metabolite fractions were measured using ultra-high performance liquid chromatography Fourier transform ion cyclotron resonance mass spectrometry (UHPLC-FT-ICR-MS), and the lipid fractions by direct infusion Fourier transform ion cyclotron resonance mass spectrometry (DI-FT-ICR-MS); these were then analysed to statistically compare the metabolite distributions. Results show significantly different profiles of metabolites across the four locations with the largest differences for both the polar and lipid fractions found between the two stations relative to the smaller differences associated with depth. We putatively annotate the most discriminant metabolites revealing a range of amino-acid derivatives, diacylglyceryltrimethylhomoserine (DGTS) lipids, oxidised fatty acids (oxylipins), glycosylated compounds, oligohexoses, phospholipids, triacylglycerides (TAGs) and oxidised TAGs. The majority of the polar metabolites were most abundant in the surface waters at L4 and least abundant in the deep waters at E1 (E1-70m). In contrast, the oxidised TAGs were more abundant at E1 and most abundant at E1-70m. The differentiated metabolites are discussed in relation to the health of the phytoplankton as indicated by nutrients, carbon and chlorophyll, and to the dominance (determined from metatranscript data) of the picoeukaryote *Ostreococcus*. Our results show proof of concept for community metabolomics in discriminating and characterising polar and lipid metabolite patterns associated with marine POM.

Highlights

- First application of community metabolomics to discriminate marine POM.
- Significantly different metabolite profiles across the four English Channel locations.
- Polar metabolites most abundant in the surface waters at L4.
- Oxidised TAGs most abundant at depth at E1.

Keywords

Meta-metabolomics; polar metabolites; lipidomics; direct infusion mass spectrometry; particulate organic matter; marine microbes; phytoplankton; UK-western English Channel.

ACCEPTED MANUSCRIPT

Abbreviations
ANOVA, analysis of variance; Chl-a, chlorophyll-a; CID, collision-induced dissociation; DAG, diacylglyceride; DGTS, diacylglyceryltrimethylhomoserine; DI, direct infusion; FT-ICR, Fourier transform ion cyclotron resonance; GC-MS, gas chromatography – mass spectrometry; IRMPD, infrared multiphoton dissociation; JCVI, J. Craig Venter Institute; LV, latent variable; MS, mass spectrometry or mass spectrometric; OVOCs, oxygenated volatile organic compounds; PCA, principal component analysis; PEG, polyethylene glycol; PLS-DA, partial least squares discriminant analysis; POM, particulate organic matter; PUFA, polyunsaturated fatty acids; QC, quality control (sample); RP, reversed-phase; RSLC®, rapid separation liquid chromatography; UHPLC, ultra-high performance liquid chromatography; TAG, triacylglyceride; WEC, western English Channel.

1. Introduction

Particulate organic matter (POM) in the ocean plays a crucial role in global carbon cycling in terms of the turnover of organic metabolites, driving the biological pump and the generation of climatically active gases. The composition of marine POM is largely determined by microbes, principally the carbon fixing phytoplankton. Fixed phytoplankton carbon and other elements are incorporated into a wide range of organic compounds or metabolites which are then acted on by biotic factors including interactions between bacteria, viruses and zooplankton, resulting in recycling and remineralisation of POM. In addition to biotic factors, a diverse range of abiotic interactions such as light, temperature and salinity also affect POM composition.

Lipids, carbohydrates and amino acids are the primary groups of metabolites that make up the fundamental building blocks of microbes in the oceans. These primary metabolites and other groups of secondary metabolites, especially pigments, have often been used as organic biomarkers to investigate the source, composition and degradation of marine POM especially its alteration down through the water column and into the sediment (e.g. Handa and Tominaga 1969; Wakeham and Lee 1989; Lee et al. 2004; Rontani et al. 2011). As a sub-set of the lipids, the fatty-acids are key nutrients affecting physiological performance, and have been used as organic biomarkers to assess trophic transfer and food quality (e.g. Kainz et al. 2004). Pigments, central to light harvesting in photosynthesis, have been used widely to provide chemotaxonomic characterisation of phytoplankton in a wide range of contrasting oceans (see review by Jeffrey et al, 1997). Pigments together with pigment degradation products and particulate carbon have also been used to track the fate of POM down the water column (Bidigare et al. 1986; Llewellyn and Mantoura 1996). Overall though, a lack of biochemical techniques has hindered the full chemical identification of POM and a significant proportion remains uncharacterised (Lee et al. 2004). Recent advancements in analytical and computational tools are now enabling a revolution in the investigation of microbial communities and their interactions with the environment (Larsen et al. 2012).

Advancements in mass spectrometry (MS), hyphenated technologies and associated software have enabled the development of the newest of the ‘omic techniques, metabolomics. Metabolomics involves the non-targeted unbiased analysis of large suites of low molecular weight organic molecules or metabolites (typically 50-1500 Da) and combined

ACCEPTED MANUSCRIPT

with statistical analysis enables the discovery of relationships between metabolites, organism physiology and the environment. Metabolomics complements genomics, transcriptomics and proteomics and represents an important addition to the 'omics toolkit especially because it provides the closest molecular link to phenotype (Vemuri & Aristidou 2005). This unbiased analysis of organic matter contrasts to the more traditional targeted analysis of predefined compound groups, the latter remaining important for the testing of specific hypotheses. As the polarity of molecules within organic material is highly diverse, the extraction and analysis of all metabolites using one method cannot be achieved. Therefore extraction and analysis in metabolomics is generally divided into that required for the polar or hydrophilic metabolite fraction and that required for non-polar or lipophilic metabolite fraction, often termed lipidomics.

Metabolomics has already demonstrated its important role in several research fields, including bioenergy, environmental interactions, functional genomics and gene discovery, secondary metabolism, genome-wide association mapping, and metabolic modelling in higher organisms and microbial systems (Tang 2011). It has also been used to study environmental stress responses in plants (reviewed in Arbona et al. 2013). Metabolomics has also been applied in studies of individual strains of microalgae, e.g., on the model algae *Chlamydomonas* (Lee & Fiehn 2008; May et al. 2008), the cyanobactera *Synechococcus* and *Synechocystis* (Baran et al. 2011; Schwarz et al. 2013) and on the diatom *Skeletonema marinoi* (Vidoudez & Pohnert 2011). Notably non-targeted metabolomics has revealed a number of unexpected metabolites in *Synechococcus* sp. PCC 7002, such as histidine betaine (hercynine), its derivatives and several unusual oligosaccharides including a range of oligohexoses (Baran et al. 2011). The potential of combining metabolomics and genomics for the identification of novel biosynthetic genes was recently highlighted in a study on a diverse range of cyanobacteria (Baran et al. 2013). Metabolomics has also revealed that shifts from high to low CO_2 levels induce a coordinated change in the central C/N-metabolism in *Synechocystis* 68034 (Schwarz et al. 2011).

Metabolomics, when applied to whole systems or communities direct from the environment, is termed community or meta-metabolomics, akin to metagenomics. An example of where community metabolomics is being used widely is in determining the effects of gut microflora on human health (Nicholson et al. 2012; Turnbaugh & Gordon 2008). It was also used recently in a soil ecology study to assess the entire microbial community of a soil sample to determine how it responds to factors such as pollution and climate change (Jones et al. 2014). There have been few community metabolomics studies in aquatic or terrestrial environments to date and it has not yet been used to study natural populations of marine microbes.

The temperate marine ecosystem of the western English Channel (WEC) provides an excellent platform to assess the metabolite compositions of the POM in an un-biased manner and to provide proof of concept for marine community metabolomics. Monitoring in the WEC has been occurring for over forty years making it one of the best studied marine regions in the world. The two main monitoring stations, L4 and E1, are seasonally stratified from late April until September and both have a spring and autumn phytoplankton bloom. Long-term monitoring of phytoplankton using microscopy counts at L4 over a period of 15 years has revealed a consistent pattern of bloom formation with diatoms reaching maximum abundance in mid-April followed by peaks in abundance of *Phaeocystis* and

ACCEPTED MANUSCRIPT

coccolithorphorids (Widdicombe et al. 2010). Phyto-flagellates numerically dominate throughout the year gradually increasing in spring with maximum abundance towards late May (Widdicombe et al. 2010). Overall the biological community in the WEC is variable, shifting over the annual cycle in response to abiotic factors such as seasonal fluctuations in light and nutrients, turbulence, temperature and other meteorology factors such as wind and cloud (Widdicombe et al. 2010, Smyth et al. 2010).

As part of the monitoring at these stations an extensive database has been compiled providing information on the phytoplankton and zooplankton community populations. Additional routine measurements at these stations include irradiance, salinity, temperature, chlorophyll, nutrients, carbon and nitrogen, phytoplankton and zooplankton counts, and photosynthetic pigments (www.weco.uk). In terms of metabolite analysis, targeted analysis of pigments using HPLC has been undertaken in the WEC for over ten years although correlating pigments with phytoplankton carbon and particulate carbon remains a challenge (Llewellyn et al. 2005). Short term, targeted metabolite studies at L4 have focussed on fatty acids to determine zooplankton fecundity (Pond et al. 1996). Additionally a group of UV sunscreen metabolites, mycosporine-like amino acids, have been studied at L4 showing temporal variation according to phytoplankton composition and solar irradiance (Llewellyn and Harbour 2003). Recently, preliminary metagenome and metatranscriptome analyses have been used to characterise the microbial populations at L4 revealing a robust seasonal structure for the bacterial community (Gilbert et al. 2010a; Gilbert et al. 2010b).

Here we build on our long term understanding of the western English Channel describing the first preliminary community metabolomics study to chemically characterise the POM in the WEC. Our study is focused on the > 0.7 µm to < 200 µm fraction of POM primarily composed of phytoplankton. Our investigation was enhanced by collecting samples in collaboration with JCVI (J. Craig Venter Institute) in May 2009, whose aim was to molecularly and genetically characterize the microbes in the WEC. There were four main aims to our study; 1. to evaluate community metabolomics as a new state-of-the- art approach to statistically discriminate different microbial populations in the WEC; 2. to putatively annotate abundant lipid and polar metabolites to determine trends across the sampling locations; 3. to compare metabolite profiles with the physico-chemical, carbon and chlorophyll measurements across the sampling locations and 4. to compare metabolite profiles with phytoplankton community transcriptional activity across the sampling locations.

2. Methods

2.1. Sample collection

Samples were collected from the WEC at the coastal station L4 (50° 15′N, 4° 13′W) on 21st May 2009 at a surface depth of 2m and below the thermocline at a depth of 17m (L4-2m, L4-17m) and at the open shelf station E1 (50° 02′N, 4° 22′W) on 28th May 2009 at a surface depth of 1m and below the thermocline at a depth of 70m (E1-1m, E1-70m; Table 1). At each sampling location, 1L of < 200 µm mesh pre-filtered seawater (n=12) was filtered under vacuum on-board ship onto a 25mm glass fibre GF/F filter paper (Whatman; nominal cut-off at 0.7 µm).

ACCEPTED MANUSCRIPT

Table 1
Physico-chemical properties of the water at time of sampling the two stations.

	L4-2m	L4-17m	E1-1m	E1-70m
Date	21st May 2009	21st May 2009	28th May 2009	28th May 2009
Number of samples	12	12	12	12
Time	12:00pm	12:00pm	10:30am	10:30am
Latitute	50.25	50.25	50.03	50.03
Longitude	-4.22	-4.22	-4.34	-4.34
Total Water Column (m)	55	55	73.2	73.2
Thermocline (m)	13	13	20	20
Sample Depth (m)	2	17	1	70
Temperature (°C)	12	11	12.44	10.77
Salinity (PSU)	35.00	35.00	35.18	35.28
Oxygen (µmol/kg)	6.10	6.10	5.98	6.20
pH (log of[H^+])	8.4	8.4	8.4	8.3

2.2. Metabolite extraction

Samples were lysed and extracted from the filters with 1 mL methanol for 20 min at 4 °C and the supernatant was removed with a glass pipette. The extraction was repeated with 1 mL of methanol: water (1:1), the extracts combined, and dried *in vacuo* (Thermo Savant, Holbrook, NY) for ca. 3 h. The dried extracts were dissolved in water: methanol: chloroform (300 µL: 270 µL: 300 µL), vortexed for 30 s, and then centrifuged for 10 min at 1800 rcf and 4 °C (Wu et al. 2008). The polar extract (upper phase) was dried *in vacuo* while the non-polar (lipid) extract (lower phase) was dried under a stream of nitrogen to minimise oxidation. Samples were stored at -80 °C until analysis.

2.3. Mass spectrometry based metabolomics and lipidomics

Direct infusion Fourier transform ion cyclotron resonance mass spectrometry (DI-FT-ICR-MS) based lipidomics was performed on a LTQ-FT Ultra (Thermo Fisher Scientific, Bremen, Germany) with a chip-based Triversa direct infusion nanoelectrospray source (Advion Biosciences, Ithaca, NY). Non-polar (lipid) samples were taken up in the original volume of methanol:chloroform (3:1) containing 5% ammonium acetate. Samples were centrifuged (10 min, 4 °C) to remove any particular matter. They were then analysed in positive ion mode in a controlled-randomized sequence different from the extraction sequence, with each sample analysed as three technical replicates. A quality control (QC) sample was pooled from all samples and analysed repeatedly at the start, end, and equidistantly throughout the sequence. Data was acquired at a nominal resolution of 100,000 (at *m/z* 400*)* in eight increasing SIM (selected ion monitoring) windows of 200 Da width, from *m/z* 120 to 1440 (Weber et al. 2011).

Reversed-phase ultra-high performance liquid chromatography Fourier transform ion cyclotron resonance mass spectrometry (RP UHPLC-FT-ICR-MS) based metabolomics of the polar samples was carried out on a Thermo Scientific Dionex Ultimate RSLC 3000 system on the same FT-ICR mass spectrometer. Samples were each taken up in 40 µL methanol and 360 µL water and centrifuged for 10 min at 4 °C and 22000 rcf. Five µL of each sample was injected onto a Hypersil Gold column (Thermo Scientific, 2.1 x 100 mm, 1.9 µm particles) and separated at 40 °C with a flow rate of 400 µL/min and a gradient from 0.1%

ACCEPTED MANUSCRIPT

formic acid in water (solvent A) to 0.1% formic acid in methanol (solvent B). The flow was held at A for 1 min, followed by a 3 min gradient to B, held there for 4 min before reverting over 1 min back to A and re-equilibrating for another 3 min before the next injection. For the first 0.5 min, flow was diverted to waste. One UHPLC-FT-ICR-MS analysis was performed per sample. Data was acquired in positive ion mode from *m/z* 100-1000 at a nominal resolution of 50,000 in centroid mode. A QC sample was pooled from all biological samples and analysed repeatedly at the start, end, and equidistantly throughout the sequence. After statistical analysis (see below), peaks of interest were subjected to further MS analysis using the same instrumentation, using wide SIM windows and spiked polyethylene glycol (PEG) standards (Sigma-Aldrich, UK) for additional internal calibration, narrow SIM windows for the determination of isotope patterns, and MS^2 / MS^n fragmentation using collision-induced dissociation (CID) and infrared multiphoton dissociation (IRMPD).

2.4. Data processing and peak annotation

DI-FT-ICR-MS lipidomics data were processed using the SIM-stitching algorithm (Southam et al. 2007; Payne et al. 2009; Weber et al. 2011), using an in-house Matlab script (SIMStitch_2_10, freely available upon request) and a series of internal mass calibrants derived from known lipid identities. High quality reproducible data was achieved by implementing a series of peak filtering algorithms (Payne et al. 2009): peaks were picked with a signal-to-noise ratio of greater than 3.5:1, a 'replicate filter' was applied such that only peaks in two (or more) of the three analytical replicates (per sample) were retained, then a 'sample filter' was applied to retain only those peaks in >30% of all samples. At the same time a 'blank filter' was applied to discard peaks that occurred in an extraction blank sample (i.e. a sample prepared as indicated above but with no biological material present) with peaks retained if they exceed a minimum sample-to-blank intensity ratio of 2, creating a peaklist and an intensity matrix. Missing values were imputed using a KNN algorithm (Hrydziuszko & Viant 2012) in an in-house R script, and the intensity matrix was normalized using the PQN algorithm (Dieterle et al. 2006). This matrix was subject to univariate statistical analysis. The same matrix was transformed using the generalised logarithm (Parsons et al. 2007) to stabilise the technical variance across the measured peaks prior to analysis using multivariate statistics. This DI-FT-ICR-MS processing algorithm has been described in detail elsewhere (Kirwan et al. 2014).

UHPLC-FT-ICR-MS metabolomics data were initially converted into netcdf (.cdf) format using Xcalibur 2.1 and processed using XCMS online (https://xcmsonline.scripps.edu/; Tautenhahn et al. 2012) to generate an intensity matrix, list of peak retention times and metabolite annotation from Metlin. The intensity matrix was imported into our SIMStitch pipeline immediately after the replicate filter, and hence included sample filtering, blank filtering, PQN normalization, KNN missing value imputation and the generalised logarithm transformation, as for DI-FT-ICR-MS processing above. The sample filter was set to 75% as no technical replicate filtering could be applied.For both the polar and lipid datasets, peaks were annotated and putative empirical formulae calculated using the MI-Pack software (Weber & Viant 2010), and by searching the KEGG and LipidMaps databases (http://www.genome.jp/kegg/; http://www.lipidmaps.org/). Polar data was also annotated with retention times and the identification output from Metlin. Those peaks that were found to differ significantly between the four sampling locations (see below) were reviewed in the original spectra in Xcalibur 2.1 (Thermo Scientific) taking into account isotopic information,

ACCEPTED MANUSCRIPT

and databases such as Chemspider and the Dictionary of Natural Compounds were used to infer compositions, especially in cases where only one molecular formula was predicted but no annotation was available. The identification of selected metabolites, using MS fragmentation, was performed as described above. The list was subject to manual filtering to remove implausible results, e.g. ^{41}K adducted peaks were removed if the corresponding, higher abundance ^{39}K adduct was not detected. Also annotated manually were chlorine isotope clusters, inorganic ions, and an oligoglycan series, which were not recognized by the automated searches.

2.5. Statistical analyses of metabolomics and lipidomics measurements

Initially, principal components analysis (PCA) was used to assess the overall metabolic similarities and differences between the four sampling locations in an unbiased manner, using the PLS_Toolbox (version 6.5, Eigenvector Research, Manson, WA, USA) within Matlab (version 7.8; The MathsWorks, Natick, MA, USA). All resulting PC scores data were tested using ANOVA and a Tukey-test to determine whether there were significant differences in the metabolic and lipid profiles between the sampling stations and depths, using an in-house Matlab script. Supervised multivariate analyses were performed using partial least squares discriminant analysis (PLS-DA), again using the PLS_Toolbox, with internal cross-validation and permutation testing to determine the quality of the models (Venetian blinds, 1000 permutations each) using in-house Matlab scripts (Westerhuis et al. 2008). Univariate statistical analyses were used to investigate whether individual MS peaks differed significantly between sampling station and depth. Specifically, ANOVAs were conducted using an in-house Matlab script (with a false discovery rate (FDR) of 5% to correct for multiple hypothesis testing; Benjamini & Hochberg 1995).

2.6. Supporting biological, physical and chemical measurements

The protocols used for physical, chemical and biological measurements including zooplankton counts are as described on the WEC website at www.westernchannelobservatory.org.uk and by Smyth et al. 2010. As the purpose of this study was to assess community metabolomics, the metatranscript data was used here only to provide taxonomic characterisation.

For metatranscriptomic sampling, 200 µm filtered seawater was passed through a 0.2 µm sterivex filter for 30 minutes (typically 1.5-2L), after which the sterivex was capped, flash frozen in liquid nitrogen, and frozen at -80ºC. RNA was purified from filters using the Trizol reagent (Life Technologies; Carlsbad, CA) and, treated with DNase (Qiagen, Valencia, CA, USA) and cleaned with the RNeasy MinElute Kit (Qiagen, Valencia, CA, USA). For polyA primed cDNA, 200 ng of DNase treated total community RNA was amplified using the MessageAmpII aRNA Amplification kit (Life Technologies, Carlsbad, CA, USA) with two rounds of *in vitro* transcription at 37 ºC for 14 hours with T7 Oligo(dT) priming. Amplified RNA was then converted to double stranded cDNA using the SuperScript III First-Strand Synthesis System (Life Technologies, Carlsbad, CA, USA) with random hexamers for first strand synthesis, and the SuperScript Double-Stranded cDNA synthesis kit (Life Technologies, Carlsbad, CA, USA) for second-strand synthesis. cDNA in the 0.3-3.0 kb size range was purified from agarose gels using QIAquick Gel Extraction Kit reagents and

ACCEPTED MANUSCRIPT

protocols (Qiagen, Valencia, CA, USA), further purified with Ampure XP beads (Beckman Coulter, Brea, CA, USA) and used directly for pyrosequencing.

For sequence annotation, all metatranscriptomic sequence libraries were filtered to remove near identical reads using CD-hit-454 (Niu et al 2010). Metatranscriptomic sequences were compared against SILVA to remove rRNA (Pruesse et al 2007). We also aligned reads against an in-house database of rRNA sequences and whole rRNA operons, including ITS sequences. All hits with E-value < 10e-10 were considered to be ribosomal RNA and were removed from further analysis. The remaining reads were compared to PhyloDB 1.02 in two separate BLAST searches to establish phylogenetic annotation. PhyloDB is a combination of many public protein sequence databases including KEGG (Kanehisa et al. 2011), IMG (Markowitz et al 2010), GenBank (Benson et al. 2011), Ensembl (Flicek et al. 2011), several in-house assemblies of algal uniculture transcriptomic sequences, the metagenomic assemblies of SAR86 (Dupont et al 2012) and HNLC *Prochlorococcus* (Rusch et al 2010), and the single cell genomes of SAR324 (Chitsaz et al 2011) and SAR86 (Dupont et al 2012). PhyloDB protein sequences (n=14 million) come from a wide array of sources, but only proteins directly annotated in KEGG serve as the source of annotations such as EC and KO. All phylogenetic annotations are generated from the best hit to any protein in PhyloDB. The cutoff used for BLASTing phylodb was 1e-5.

3. Results

3.1. Distinguishing sampling locations using metabolomics

UHPLC-FT-ICR-MS of the polar metabolite extracts, coupled with rigorous data processing, yielded a final data matrix of 47 biological and 9 QC samples and 173 unique *m/z* values (one E1-1m sample was lost before extraction); the QC samples were later removed from the dataset. This relatively low number of peaks detected reflects the relatively low concentrations of metabolites in these filtered seawater samples compared to biofluids and tissue extracts that are more routinely investigated in a metabolomics study. PCA was used initially to visualize the metabolic differences between the four sampling locations, with the scores plots revealing that the largest metabolic differences (along PC1 axis) occurred spatially, between L4 and E1, relative to smaller metabolic differences (along PC4 axis) between the near-surface and deeper samples (Supplementary Information, Fig. SI1).

The clustering of the QC samples within the PCA scores plot indicates high technical quality of this polar metabolomics dataset. The apparent metabolic differences between sampling locations were evaluated statistically by testing the significance of the separation between the groups along the PC axes (Supplementary Information, Table SI1): The effect of both stations and depth on the metabolic profiles was significant, with each of the four locations being significantly different from all other locations along PC1 ($p(PC1)<1E-8$), by both depth and site.

Having confirmed that the metabolic profiles differ between stations and as a function of depth, we re-analysed the UHPLC-FT-ICR-MS metabolomics dataset using supervised multivariate analyses (PLS-DA), a more powerful approach for discovering which peaks in the mass spectra are primarily responsible for these differences. The optimal PLS model

ACCEPTED MANUSCRIPT

comprised of 4 latent variables (LVs) and 36 forward-selected variables (*m/z* values), derived by minimising the group classification errors obtained through internal cross-validation. The resulting PLS-DA scores plots confirm that the largest metabolic differences occur between stations E1 and L4 (Fig. 1A), relative to the smaller metabolic differences associated with sampling depths (Figs. 1B & C).

Specifically, the LV1 axis describes mostly sampling station differences in the metabolic profiles, LV2 describes depth differences at E1, and LV3 describes depth differences at L4; low classification error rates were obtained for all four groups, and permutation testing was used to confirm the statistical significance of these results (Table SI2; Westerhuis et al. 2008).

The *m/z* values that are responsible for the separation of these groups (derived from the LV1, LV2 and LV3 weightings) were then putatively annotated and are presented in Table SI3a for L4 vs. E1, in Table SI3b for E1-1m vs. E1-70m, and Table SI3c for L4-2m vs. L4-17m, all with extensive metadata.

A summary of these findings showing which putatively annotated polar metabolites differ the most between stations (L4 and E1) and depths is shown in Table 2. As an additional assessment for the robustness of these findings, ANOVA was conducted on each of the 173 peaks in the polar metabolite dataset, revealing that 35 of the 36 forward-selected *m/z* values were significantly different between groups (at FDR<5%; Tables SI3).

Several groups of annotated polar metabolites were found to differ between the sampling locations; these included aromatic amino acids and derivatives, glycosylated compounds, oligohexoses and a range of fatty acids and oxylipins (Table 2 and SI3). Overall polar metabolites were considerably more abundant at L4-2m, L4-17m and E1-1m than at E1-70m. Notably glycosylated compounds and oligoglycans were a lot more abundant at L4-2m. For the amino-acids and derivatives, the related metabolites phenylalanine and tyrosine showed similar distributions with slightly higher abundance at L4-2m. In contrast, another aromatic amino acid, mycosporine-glycine with UV sunscreen and antioxidant properties, was most abundant at E1-1m. Its MS^2 & MS^3 spectra correspond well to those of an authentic compound. Pyloricidin C, an antibiotic, showed a different distribution again with much higher abundance at L4-17m. Another unusual metabolite that was differentially discriminated was annotated as a betaine lipid matching ulvaline, constituted of the headgroup plus glycerol of homoserine betaine with a diacylglyceryltrimethylhomoserine (DGTS) backbone, (Dictionary of Natural Compounds; Abe & Kaneda, 1975). IRMPD fragmentation shows the expected headgroup ion at *m/z* 144.10171 (calc. 144.10191). Ulvaline was found to be most abundant in the surface samples, and in particular at L4-2m.

ACCEPTED MANUSCRIPT

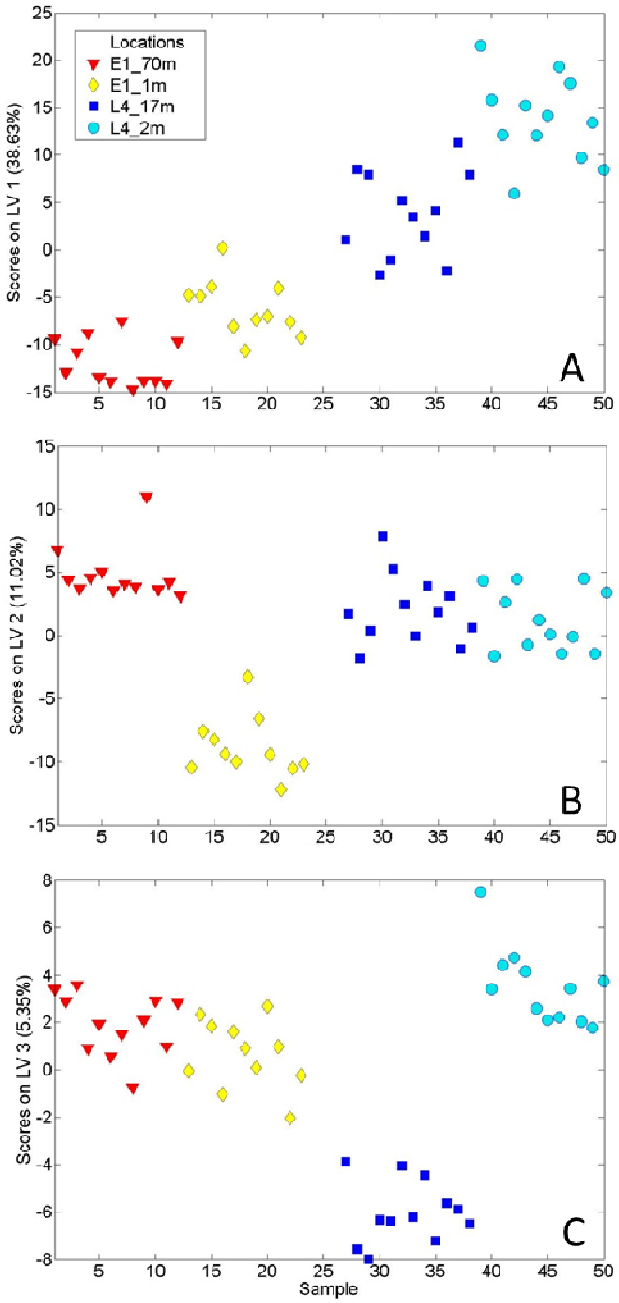


Figure 1: PLS-DA scores plots from analysis of the polar metabolite extracts of marine POM showing the effects of site location and sampling depth. **A** sample number sorted according to sampling location plotted against latent variable (LV)1, highlighting the differences between stations E1 and L4; **B** sample number plotted against LV2, highlighting the influence of the 1m vs. 70m sampling depth at station E1; **C** sample number plotted against LV3, highlighting the influence of the 2m vs. 17m sampling depths at station L4. The classification error rates and significance of the differences in the metabolic profiles are listed in Table SI2.

ACCEPTED MANUSCRIPT

3.2. Distinguishing sampling locations using lipidomics

DI-FT-ICR-MS of the lipid extracts coupled with rigorous data processing resulted in lipid profiles comprising of 1896 peaks and a final data matrix without QC samples comprising of 47 biological samples. PCA was used initially to visualize the similarities or differences between the lipid profiles from the four sampling locations. One L4-2m sample, an outlier in the PCA, was excluded from subsequent modeling. Consistent with the polar metabolite measurements, the PCA scores plot revealed that the largest lipid differences occurred spatially, between L4 and E1, relative to smaller or no apparent differences between the near-surface and deeper samples (Fig. SI2); these observations are supported by statistical analyses of the group separations along the PC axes (Table SI4). Specifically, the effect of sampling station on the lipid profiles was significant, with the near-surface locations L4-2m vs. E1-1m samplings differing significantly, and the deeper L4-17m vs. E1-70m also differing significantly ($p=1.67x10^{-15}$). While the effect of depth at L4 was significant ($p=3.98x10^{-3}$), the lipid profiles were not significantly different between E1-1m and E1-70m.

Using the same strategy as for the polar metabolites, the lipidomics dataset was re-analysed using PLS-DA to discover which peaks were primarily responsible for the differences between the sampling locations. The optimal PLS model comprised of 4 LVs and 134 forward-selected variables (*m/z* values), and the resulting scores plots confirm that the largest lipid differences occur between stations (Fig. 2A) relative to the more subtle differences between sampling depths (Figs. 2B & C). Specifically, the LV1 axis again describes sampling station differences, while LV2 describes the depth differences at L4, and LV3 the depth differences at E1. Relatively low classification error rates were obtained for all four groups, and permutation testing was used to confirm the statistical significance of these results (Table SI5).

The *m/z* values of the lipids that are responsible for these group separations, along with their putative annotation and associated metadata, are listed in Table SI6A for L4 vs. E1, in Table SI6B for L4-2m vs. L4-17m, and Table SI6C for E1-1m vs. E1-70m. A summary of these findings, showing which putatively annotated lipid metabolites differed the most between stations and depths, is shown in Table 2. As an additional assessment for the robustness of these findings, ANOVA was conducted on each of the 1896 peaks in the lipidomics dataset, revealing that 76 of the 134 forward-selected *m/z* values were significantly different between groups (at FDR<5%; Table SI6).

The lipid fraction contained suites of fatty acids, TAGs and DAGs as well as their oxidised products. The membrane lipids, DGTS and some phospholipids were also prominent (Tables 2 and SI6). In relation to abundance patterns there were some distinct differences across classes (Table 2 and SI7). Overall most of the lipid classes, as with the polar metabolites, were least abundant in the E1-70m. The exception to this was for a range of oxidised TAGs, docosanedioic acid and TAG 48:4 which were in contrast the most abundant at E1-70m. The oxidised TAGs were on average five times higher in abundance in the E1-70m compared to at both the L4 sample locations. The DGTS lipids were typically 3 to 4 times higher at L4-2m, L4-17m and at E1-1m compared to at E1-70m. Phospholipids were more evenly distributed with on average highest abundance in the L4-17m and lowest abundance in the L4-2m sample (Table 2 and SI6).

ACCEPTED MANUSCRIPT

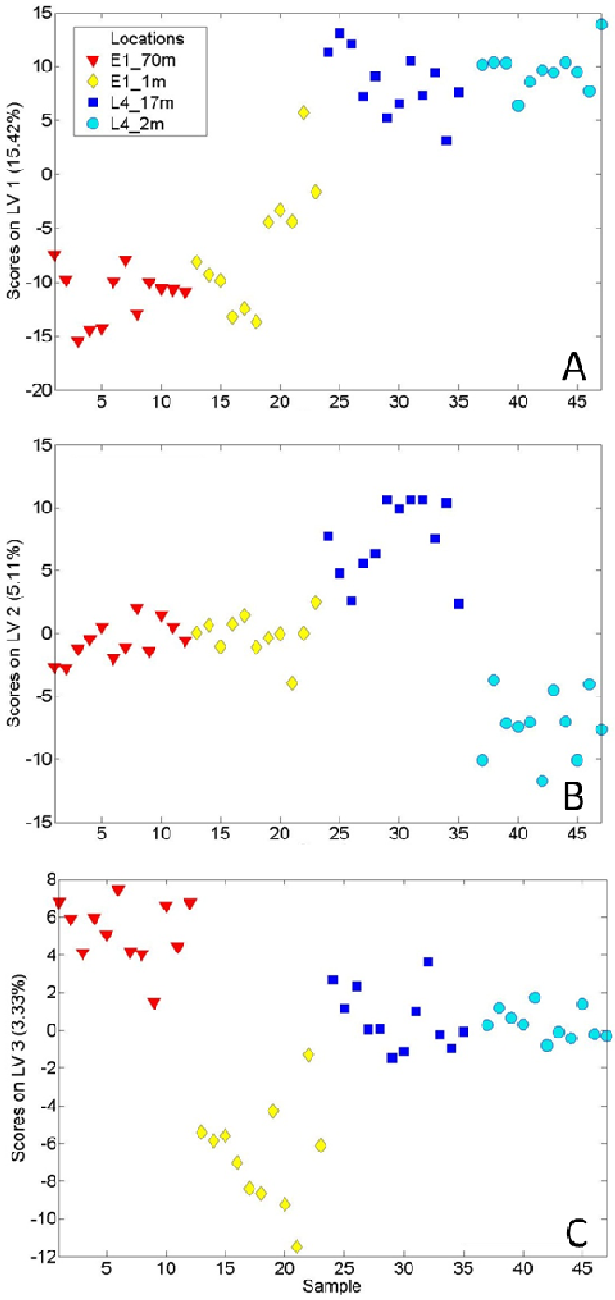


Figure 2: PLS-DA scores plots from analysis of the lipid metabolite extracts of marine POM showing the effects of site location and sampling depth. A sample number sorted according to sampling location plotted against latent variable (LV)1, highlighting the differences between stations E1 and L4; B sample number plotted against LV2, highlighting the influence of the 2m vs. 17m sampling depth at station L4; C sample number plotted against LV3, highlighting the influence of the 1m vs. 70m sampling depths at station E1. The classification error rates and significance of the differences in the metabolic profiles are listed in Table SI5.

ACCEPTED MANUSCRIPT

Table 2

Summary of polar and lipid metabolites differentiating the four sampling locations: Column titles are as follows: *M/Z*: *m/z* values of the peaks; Extr: polar (P) or lipid (L) extract; rank: ranking of metabolites by latent variables from PLS-DA (E1-L4 differentiates the two stations, L4 distinguishes L4-2m and L4-17m, and E1 differentiates E1-1m and E1-70m); Adj.P: adjusted p-value from t-test with FDR < 5 %; L4-2m, L4-17m, E1-1m: fold change of average intensity relative to E1-70m; Ion form: adduct form detected: Final annotation: selected from KEGG or LipidMap hits, spectral interpretation and literature searches (note that the three compounds annotated as dioic acids (*m/z* 357.20370, 393.18047, and 393.29785) are also isobaric with oxylipins); compound group: by chemical similarity. A more detailed version of this table containing more compounds and descriptions can be found as Table SI7.

Observed		Statistics							Annotation		
		rank	rank	rank		ratio					
M/Z	Extr	E1-L4	L4	E1	Adj.P	L4_2m	L4_17m	E1_1m	Ion form	Final Annotation	Compound group
166.08614	P	39	18	89	4.5E-03	2.99	1.12	1.24	$[M+H]^+$	Phenylalanine	Amino acids and derivatives
182.08107	P	62	12	41	4.7E-02	2.22	1.05	1.44	$[M+H]^+$	Tyrosine	Amino acids and derivatives
246.09711	P	42	35	3	1.0E-06	4.40	4.14	7.98	$[M+H]^+$	Mycosporine-glycine	Amino acids and derivatives
343.14985	P	78	1	10	0.0E+00	3.49	34.12	14.21	$[M+H]^+$	Pyloricidin C	Amino acids and derivatives
236.14915	P	21	6	11	0.0E+00	10.88	2.27	4.57	$[M+H]^+$	Ulvaline	Amino acids and derivatives, DGTS backbone
682.56203	L	58	724	4	1.6E-04	4.65	5.73	4.51	$[M+H]^+$	DGTS 30:1	DGTS lipid
704.54636	L	115	1149	3	1.7E-04	3.30	2.83	3.51	$[M+H]^+$	DGTS 32:4	DGTS lipid
730.56191	L	69	516	5	1.1E-04	4.16	5.62	4.21	$[M+H]^+$	DGTS 34:5	DGTS lipid
732.57806	L	36	174	1	1.3E-04	5.43	8.86	6.30	$[M+H]^+$	DGTS 34:4	DGTS lipid
736.60936	L	1751	1632	37	7.0E-02	1.56	1.31	2.07	$[M+H]^+$	DGTS 34:2	DGTS lipid
758.59355	L	132	285	11	1.7E-04	2.83	4.47	3.51	$[M+H]^+$	DGTS 36:5	DGTS lipid
804.57769	L	119	978	16	2.0E-05	3.73	3.06	2.89	$[M+H]^+$	DGTS 40:10	DGTS lipid
856.60868	L	49	1378	13	1.1E-04	4.80	4.44	3.13	$[M+H]^+$	DGTS 44:12	DGTS lipid
289.17741	P	11	19	13	0.0E+00	6.30	4.56	0.73	$[M+Na]^+$	Hexadecanoid	Fatty acids and oxylipins
291.19309	P	6	4	24	0.0E+00	10.47	1.49	0.78	$[M+Na]^+$	Hexadecanoid	Fatty acids and oxylipins
301.21638	P	36	44	117	2.0E-06	4.71	3.58	2.18	$[M+H]^+$	Eicosahexaenoic acid (20:6)	Fatty acids and oxylipins
313.17772	L	165	24	814	0.0E+00	4.10	1.42	0.93	$[M+Na]^+$	Octadecanoid	Fatty acids and oxylipins
315.19312	P	5	100	14	0.0E+00	9.47	3.78	0.82	$[M+Na]^+$	Octadecanoid	Fatty acids and oxylipins
319.16451	P	44	7	110	0.0E+00	3.63	5.61	2.41	$[M+2Na-H]^+$	Octadecapentaenoic acid	Fatty acids and oxylipins
341.20876	P	30	26	81	0.0E+00	3.84	3.04	1.74	$[M+Na]^+$	Eicosanoid	Fatty acids and oxylipins
357.20370	P	32	86	151	1.0E-06	3.73	2.10	1.37	$[M+Na]^+$	Eicosatetraenedioic acid	Fatty acids and oxylipins
393.18047	P	16	85	19	0.0E+00	5.87	3.24	0.78	$[M+2Na-H]^+$	Tetracosadecaenoic acid	Fatty acids and oxylipins
393.29785	L	11	141	712	1.6E-05	0.21	0.24	0.92	$[M+Na]^+$	Docosanedioic acid	Fatty acids and oxylipins
252.14406	P	28	72	22	0.0E+00	5.52	2.65	2.70	$[M+H]^+$	Gluconamide or hexapyranoside	Glycosylated compound
277.08928	P	7	11	104	0.0E+00	9.34	7.02	2.22	$[M+Na]^+$	Hexosyl-glycerol	Glycosylated compound
329.13419	P	31	36	1	0.0E+00	28.62	10.47	41.10	$[M+H]^+$	Cyanogenic glycoside	Glycosylated compound
347.14476	P	12	2	2	0.0E+00	27.67	2.58	11.61	$[M+H]^+$	poss. glycoside	Glycosylated compound
573.23173	P	146	39	9	0.0E+00	1.71	1.26	3.85	[M+K]+	Glycoside	Glycosylated compound
434.11807	P	17	78	44	1.0E-06	18.99	10.09	5.49	$[M+K+H]^{2+}$	Hex5 (2+)	oligoglycan
527.15814	P	23	38	42	0.0E+00	16.68	10.75	6.00	$[M+Na]^+$	Hex3	oligoglycan
649.21804	P	4	5	12	0.0E+00	21.11	16.31	6.63	$[M+H]^+$	Hex4-H2O	oligoglycan
671.20000	P	15	14	38	0.0E+00	19.45	14.47	7.04	$[M+Na]^+$	Hex4-H2O	oligoglycan
811.27102	P	8	8	17	0.0E+00	17.72	14.02	6.65	$[M+H]^+$	Hex5-H2O	oligoglycan
851.26387	P	13	50	35	1.0E-06	22.38	11.02	5.97	$[M+Na]^+$	Hex5	oligoglycan
457.25669	L	399	23	1204	2.6E-02	1.43	3.91	1.01	$[M+H]^+$	lysoPG 14:0	Phospholipids
730.47734	L	838	35	1520	1.3E-02	0.50	1.34	1.14	$[M+K]^+$	PC 29:0 or PE 32:0	Phospholipids
828.55313	L	1853	31	1542	1.5E-01	0.84	1.95	1.17	$[M+H]^+$	PC 40:9 or PE 43:9	Phospholipids
908.70797	L	14	998	1131	4.7E-04	0.27	0.33	0.87	$[M+Na]^+$	PC 43:1 or PE 46:1	Phospholipids
914.66069	L	50	353	1398	0.0E+00	0.33	0.24	0.82	$[M+Na]^+$	PC 44:5 or PE 47:5	Phospholipids
450.35838	L	122	1711	144	6.4E-04	0.25	0.30	0.45	$[M+H]^+$	TAG 48:4	Tri- or diacylglycerides
739.52620	L	498	22	24	3.5E-01	0.69	2.66	2.63	$[M+Na]^+$	DAG 44:10	Tri- or diacylglycerides
893.69890	L	202	358	191	2.2E-02	2.63	1.50	0.85	$[M+K]^+$	TAG 42:4	Tri- or diacylglycerides
921.72974	L	99	19	157	1.2E-01	4.90	2.18	2.14	$[M+K]^+$	TAG 44:4	Tri- or diacylglycerides
803.54429	L	2	1	606	4.0E-04	0.22	0.04	0.68	$[M+Na]^+$	2x oxidized TAG 45:8	Triacylglyceride - oxidized
907.70356	L	15	672	190	2.0E-06	0.19	0.18	0.59	$[M+H]^+$	oxidized TAG 54:8	Triacylglyceride - oxidized
913.65717	L	28	139	1071	2.0E-06	0.31	0.20	0.75	$[M+Na]^+$	2x oxidized TAG 55:12	Triacylglyceride - oxidized
927.67120	L	6	27	10	3.0E-06	0.27	0.14	0.53	$[M+H]^+$	2x oxidized TAG 56:12	Triacylglyceride - oxidized
929.68055	L	150	228	1450	4.1E-04	0.59	0.43	0.93	$[M+Na]^+$	2x oxidized TAG 54:8	Triacylglyceride - oxidized
983.73598	L	1	338	746	2.0E-05	0.08	0.04	0.62	$[M+H]^+$	2x oxidized TAG 60:12	Triacylglyceride - oxidized

ACCEPTED MANUSCRIPT

3.3. Nutrients, carbon and chlorophyll

There were also distinct differences between the stations and depths for nutrient, total particulate carbon (C) and chlorophyll-a (Chl-a) concentrations. Nutrient concentrations, consistent with stratification in the water column typical for May, were generally higher in the deeper samples at both stations, with a degree of depletion in the surface samples (Fig. 3). Most striking was the depleted level for all nutrients at E1-1m.

Chl-a concentrations were overall higher in deep samples compared to the surface at both stations although the Chl-a concentrations at E1-70m (0.70 µg/L) were only slightly higher than at the E1 surface (0.46 µg/L; Fig. 3). The highest concentrations of Chl-a (1.5 µg/L) were found at L4-17m where they were at least double those in the other samples. Surface Chl-a concentrations were similar for both stations (0.46 and 0.50 µg/L at E1-1m and L4-2m respectively).

Notably, the pattern in Chl-a concentration did not reflect the pattern for C (Fig. 3). The C in both the surface and deep sample at station E1 was over double (>100 µg/L) that compared to both depths at L4. This resulted in variable C:Chl-a ratios at L4 (63 and 28 at L4-2m and L4-17m, respectively) and high C:Chl-a ratios at E1 (225 and 148 at E1-1m and E1-70m, respectively; Fig. 3).

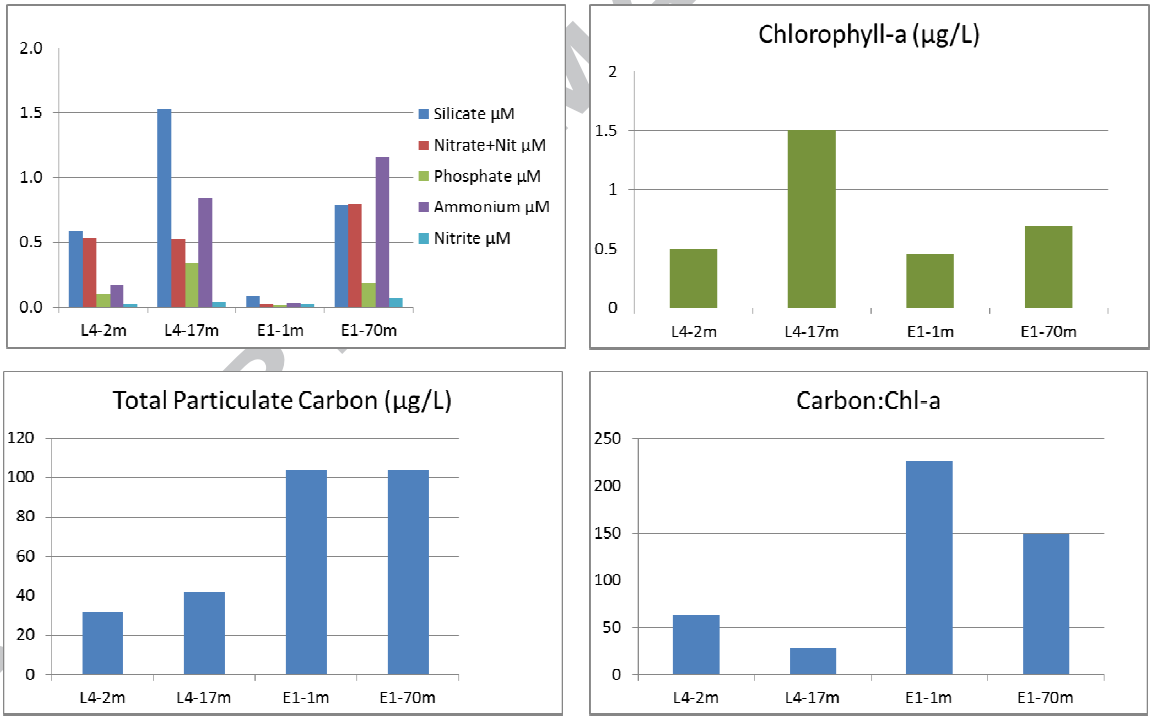


Figure. 3. Nutrients, carbon and chlorophyll-a concentrations and the carbon:chlorophyll-a ratio for the four sample locations in the western English Channel.

ACCEPTED MANUSCRIPT

3.4. Microbial community structure

The phylum Viridiplantae dominated the phytoplankton community population based on transcript abundance at all sampling locations (Fig. 4). This phylum was most active at L4-2m with transcriptional activity almost double that at L4-17m and E1-1m and almost five times that at E1-70m. All other phyla/classes were insignificant in terms of transcription activity compared to the Viridiplantae at L4-2m. In contrast, at the other stations the activity of most other classes increased relative to the Viridiplantae (Fig.4). At E1-70m, in particular, there were increased relative activities of all groups with highest activities for the Dinophyceae and Phaeocystales (Fig.4B). The Bacillariophyta were active in all samples, with slightly more activity overall at E1, with most activity at E1-1m and least activity at L4-17m. The Dinophyceae were also active in all the samples with highest activity in both of the deep sample locations.

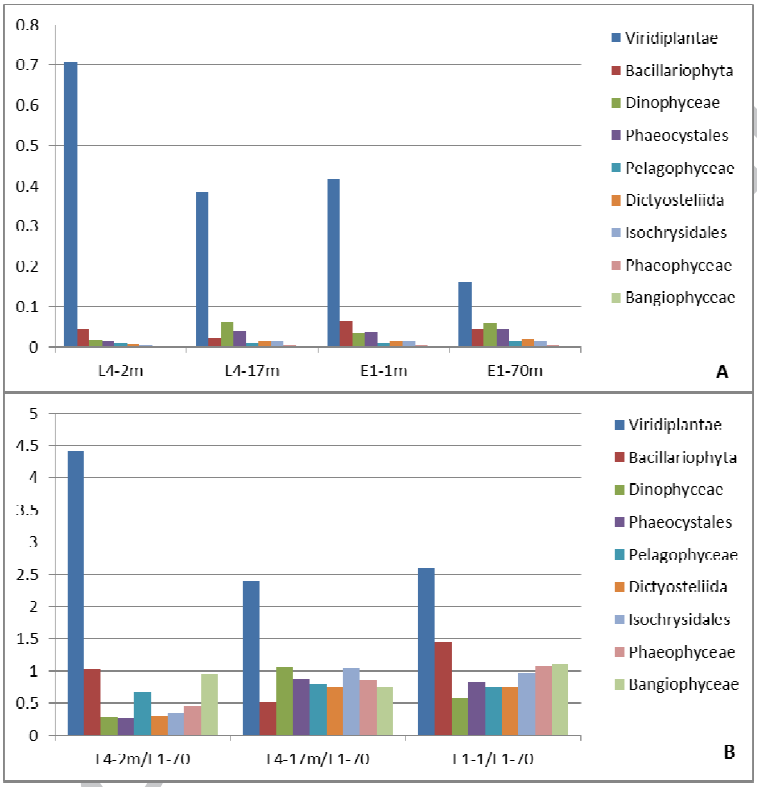


Figure.4. Relative transcriptional activity abundance of the phytoplankton classes. **A**. Relative abundance for each class at each sampling location. **B**. Relative abundance compared to E1-70m.

The Viridiplantae were dominated by the picoeukaryote flagellates of the *Ostreococcus* genus (Division Chlorophyte, Class Prasinophyte: Fig.4). In particular, *Ostreococcus lucimarinus* dominated transcript activity at all stations except at E1-70m (where it was second highest in activity) with transcript activity at L4-2m at least five times higher compared to other species, and at L4-17m and E1-1m approximately three times higher (Fig.5). *Ostreococcus tauri* was the second most active species at L4-2m and third active species at the other sites. An unidentified Viridplantae, Streptophyta species (with a genome match to *Physcomitrella patens*) was also highly active at all four stations and dominated the E1-1m signature.

ACCEPTED MANUSCRIPT

Given the proportion of transcripts attributed to *O. lucimarinus,* they potentially comprise a large portion of the total community, contributing to the observed metabolites (Fig.5). Thus these reads were examined in more detail in similarity to the reference genome. A fragment recruitment analysis of these *O. lucimarinus* reads revealed that greater than 60% had greater than 95% nucleotide identity to the type strain, CCE9901, which was isolated from coastal California waters (Palenik et al. 2011). E1-70m was the exception with less than 10% of the reads showing greater than 95% similarity. Essentially, at E1-70m, not only were *O. lucimarinus* transcripts less abundant, they likely originated from a different strain, while those from the other samples are highly similar to the reference genome. The relative number of transcripts recruited to each chromosome were similar between stations, but not between chromosomes. Specifically, chromosome 18 was consistently underrepresented in each metatranscriptome, while chromosomes 8 and 12, while recruiting more reads than other chromosomes, recruited the least at high identity (Fig. SI4).

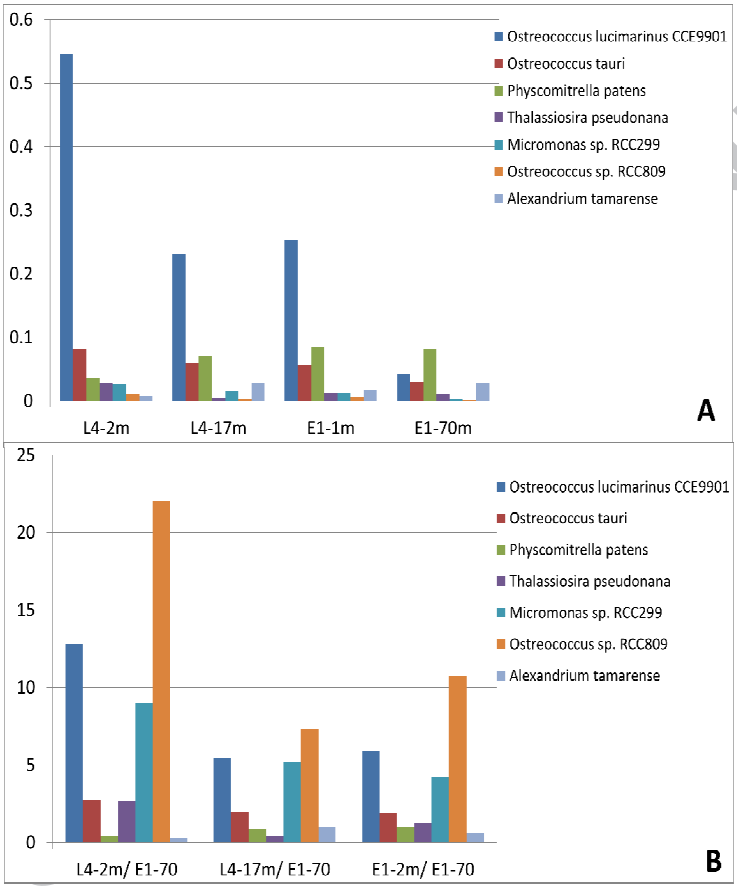


Fig. 5. Relative abundance of dominant phytoplankton species based on metatranscript activity. **A**. Relative abundance of each species compared to total number of species present. **B.** Abundance of each species relative to abundance at E1-70m. For further species abundance refer to Fig SI3.

The most abundantly active diatom closely matched *Thalassiosira pseudonama,* this centric diatom was notably transcriptionally active in the L4-2m sample compared to the other samples (Fig. 5). Another picoeukaryotic prasinophyte *Micromonas* sp., was also prominent at L4-2m, with lesser activity in L4-17m and E1-1m and least activity at E1-70m. *Phaeocystis globosa* dominated the Phaeophyceae.

ACCEPTED MANUSCRIPT

4. Discussion

4.1. Linking metabolite profiles with the environment

The profiles of metabolites in the POM for both the polar and lipid fractions across the four locations in the English Channel were significantly different with the largest differences found between the two stations relative to the smaller differences associated with depth. There were in addition some noteworthy observations, beyond proving the technological proof of concept, on the types of metabolites that were discriminated at the different stations and on how these related to differences in nutrients, C, Chl-a and the phytoplankton community populations.

In our study the majority of the polar metabolites were most abundant at L4-2m. Within the polar fraction, the oligoglycan and glycoside metabolite groups were most notably abundant at L4 compared to at E1 (Table 2). At L4-2m the levels of nutrients and a C:Chl-a ratio of 63 indicated the phytoplankton were in a healthy state (Fig. 3). This coincided with the high levels of transcriptional activity for *Ostreococcus* observed at this station. This suggests that the abundance of polar metabolites at L4-2m were associated with higher levels of nutrients and healthy *Ostreococcus* dominated phytoplankton communities.

Polar metabolites were least abundant at E1-70m. In comparison to L4, the depleted nutrients at E1-1m together with relatively low Chl-a (0.4 µg/L), relatively high carbon (>100 µg/L) and high C:Chl-a (225) are indicative of a phytoplankton post-bloom situation, typical for May (Widdicombe et al. 2010). It is likely that during this post-bloom situation that the POM was composed of phytoplankton that are no longer viable and/or in the process of being grazed or degraded. Consistent with this was the transcriptome data which indicated that there were far fewer active cells at E1-70m (Figs. 4, 5). This suggests that the lower abundance of polar metabolites particularly at E1-70m, was associated with POM likely to have contained compromised phytoplankton cells.

In contrast, within the lipid fraction, the oxidised TAGs were most abundant at E1-70m and generally more abundant at E1 than at L4 (Table 2). Zooplankton abundance, as measured in the surface waters, was considerably higher at E1 than at L4 (Table SI8). Indeed zooplankton grazing activity has been linked to the production of oxidised TAGs (Ianora et al. 2011) indicating a possible link between our observations and zooplankton abundance (also see Section 4.2).

The metatranscriptomes of each site were dominated by *Ostreococcus*, with *Ostreococcus lucimarinus* being particularly prevalent, though all three reference genomes (*O. lucimarinus, tauri,* RCC809) were detected. The WEC *Ostreococcus* population contain a distinct, but highly similar strain of *Ostreococcus lucimarinus* CCE9901*,* based on the high nucleotide identity, and general evenness of the majority of reads mapped to the chromosomes. Chromosome 18 recruited substantially fewer reads, which is consistent with it being the most divergent between species (Palenik et al. 2007). Alternatively, this chromosome is enriched in cell surface modification proteins thought to be involved in predator defence, thus the low expression might indicate a lack in predation pressure at the time of sampling. Chromosome 8 and 12 have the most reads recruited to them, but at the lowest sequence similarity relative to other chromosomes. Interestingly, in *O. lucimarinus*, these

ACCEPTED MANUSCRIPT

chromosomes share small internal duplications, which can act as chromosome recombination sites. Potentially the WEC population contains larger scale duplications, which result in sequence divergence. A metabolomics study of *Ostreococcus* cultures has not been performed to our knowledge, but provides an ideal follow-up experiment.

4.2. Characterisation of the metabolites and their possible roles

The lipid fraction was composed of a wide range of fatty acids and oxidised fatty acids (oxylipins), including compounds ranging from C15 fatty acids to oxidised triacylglycerides with up to C60 total fatty acid content. A variety of oxylipin metabolites were putatively annotated (Table 2). Because of the involvement of free radicals and other reactive oxygen species in the production of oxylipins it is possible that they could be useful markers of oxidative stress in the marine environment. There are, however, other potential implications of oxylipins including mediation of physiological and ecological processes in the plankton. Oxylipins have been found to impact food webs by interfering with the reproductive success of herbivores therefore introducing a new perspective on phytoplankton-zooplankton interactions (Ianora et al. 2011). Such metabolites are suggested to have multiple simultaneous functions: They not only deter herbivore feeding but some also act as allelopathic agents against other phytoplankton cells, thereby affecting the growth of competitors, and signalling population-level cell death and termination of blooms, with possible consequences for food web structure and community composition. Some oxylipins also play a role in driving marine bacterial community diversity, with neutral, positive or negative interactions depending on the species, thereby shaping the structure of bacterial communities during diatom blooms (Ianora and Miralto 2010).

Oxylipins may play another important role acting as precursors to the production of volatile compounds in the ocean. Currently there is poor understanding of the sources of volatiles important in cloud condensation affecting climate (Dixon et al. 2013). Oxylipins could act as important precursors to oxygenated volatile organic carbons (OVOCs). Polyunsaturated fatty acids (PUFAs) form free lipid radicals in the presence of preformed radicals, light or iron ions amongst other things. The highly reactive free lipid radical is oxidised to form lipid peroxyl radicals (LOO•) which in turn react with a new lipid molecule to form lipid hydroperoxides (Laguerre et al. 2007) which can decompose to form OVOCs. The production of OVOCs from the ocean, whilst recognised as being important in climate change, is poorly understood (Dixon et al. 2013). The importance of oxidised fatty acids has recently been highlighted in a paper studying the heterogeneous oxidation of PUFAs at the air-sea interface (Zhou et al. 2014). It is clear that a better understanding on the types and distributions of oxylipins is required to determine the potential important roles that these compounds play in the marine environment.

Polar metabolites were also assigned putative annotations. One of the most striking features in the differentially abundant metabolites was a suite of oligoglycans (Table 2). Oligoglycans or oligosaccharides, like the oxylipin compounds, are components of cell wall membranes. More unusual was the detection of the glycerolipid ulvaline (glycerol homoserine betaine) a betaine lipid with a diacylglyceryl-N-trimethylhomoserine (DGTS) backbone structure. Whilst betaine lipids are known to be widely distributed in cell membranes of photosynthetic

ACCEPTED MANUSCRIPT

bacteria and eukaryotes, less is known about the distribution of betaine lipids in microalgae, and DGTS betaine lipids are more unusual (Kato et al. 1996, Armada et al. 2013). Interestingly, whilst the low abundance of these DGTS related metabolites clearly distinguished the E1-70m location, phospholipids remained relatively evenly distributed in all sample locations (Table 2). DGTS betaine lipids are poorly understood: They are increasingly being recognized as important to the composition and metabolism of marine algae, especially with respect to the relationship between nutrients such as phosphate and phytoplankton (Van Mooy et al. 2009; Armada et al.2013). More specifically, betaine lipids have been shown to substitute phospholipids in phytoplankton where phosphate is scarce (Van Mooy et al. 2009). However, in our study, there was no obvious relationship between the levels of nutrients and the observed differences in these two classes of lipids. Indeed correlating intact polar lipid composition to species abundance is non-trivial (Brandsma et al. 2012) and a wider number of samples and multivariate statistics would be required to study this intriguing relationship in more detail.

Another abundant polar metabolite in the L4-2m sample was putatively annotated as a terpenoid, 3'-hydroxy-geranylhydroquinone (Table SI3). 3'-hydroxy-geranylhydroquinone has been identified as a precursor to shikonin found in the Chinese herbal plant *Lithospermum.* and is known to have potent cancer efficacy (Duan et al. 2014). Definitive annotation is therefore required on this unusual terpenoid. Similar geranyl compounds or isoprenoids are precursors to both the phytol side chain of chlorophyll and to the backbone of carotenoids so further in depth investigation would be required to confirm the role of 3'-hydroxy-geranylhydroquinone in biosynthetic or biodegradation pathways.

A number of amino acid and related metabolites were differentially abundant including the putatively annotated aromatic amino acids phenylalanine and tyrosine (Table 2). Such amino-acids been shown to provide an alternate and sole source of nitrogen to diatoms and haptophytes especially when deprived of nitrate (Landymore and Antia 1977). The identity of the mycosporine–like amino acid (MAA), mycosporine-glycine, which has UV sunscreen and antioxidant properties, was also confirmed using MS^2, yielding a fragmentation pattern consistent with that reported in MS targeted MAA analysis (Llewellyn and Airs 2010). The abundance of mycosporine-glycine in the WEC is consistent with a previous study on MAAs where high levels (up to 8 µg/L) of mycosporine-glycine were found in springtime and corresponded to increases in *Phaeocystis pouchetti* (Llewellyn and Harbour 2003). The detection of this known metabolite using a MS based community metabolomics approach provides a degree of validation of the workflow presented here. Although it is well understood that UV irradiation results in the up-regulation of pathways leading to the production of MAAs, little is known of the role aromatic amino acids such as phenylalanine and tryosine play in shunting nitrogen under environmental stress.

Another amino-acid metabolite identified was pyloricidin-C, a natural novel antibiotic known to possess potent and highly selective activity against *Helicobacter pylori* (Hasuoka et al 2002). Further work would be required to determine if such a metabolite is used in microbial population control. Further unambiguous identification of the large number of metabolites detected here would require further extensive chemical characterisation.

ACCEPTED MANUSCRIPT

4.3. Methods and limitations

Whilst there were clear differences in both the polar metabolite and lipid profiles at the four locations, more detailed sampling and metabolomics profiling under different environmental scenarios would be required to confirm correlation between metabolite profiles and the physico-chemical and biological environment. Our samples were taken from stations that have been well studied in terms of the physico-chemical environment and community characterisation; however, the logistics of our sampling specifically the low number of samples being collected a week apart and small number of sampling locations studied, limited a truly spatial comparison. Given our samples were set within the context of a well-studied site within the WEC, we were able to generate substantial hypotheses regarding the discriminatory patterns of the metabolite features and in terms of the possible role of annotated suites of metabolites in contributing to the cycling of organic carbon and nitrogen.

There were also limitations associated with the sampling protocol used: only metabolites associated with particulates retained on GF/F filters were investigated. A substantial number of metabolites associated with smaller heterotrophic bacteria and perhaps more importantly large suites of metabolites associated with the dissolved organic matter (DOM) will have been discarded in this process. Recently a study has shown that the production of DOM in a range of cultured phytoplankton is important is the main source of organic substrates for heterotrophic bacteria and acts as a link between autotrophic and heterotrophic microbial community structure (Becker et al. 2014).

The use of the relatively undeveloped community metabolomics on environmental samples provides significant analytical challenges, in particular, absolute identification is inherently difficult without further in depth MS fragmentation and (if standards are not available) NMR analysis (Viant & Sommer 2013). While there are compound databases for lipids and polar metabolites, these are not focussed on microbial organisms sampled from the environment. Indeed, one of the biggest challenges facing metabolomics is that of standardisation and metabolic identification (Tang 2011). Level 1 metabolite identification (as opposed to level 2 putative metabolite annotation), as defined by the Metabolomics Standards Initiative, requires two orthogonal measurements of a compound in the biological sample as well as the authentic chemical standard (e.g., measurement of exact mass and MS^2 characterisation) (Sumner et al. 2007). In this regard metabolomics lags behind other 'omics techniques. Regardless of these limitations, this study highlights the power of metabolomics to discriminate marine particulate organic matter based on profiling rather than compound specific analysis.

4.4. Broader key future challenges

A key challenge in understanding microbial communities is to use a systems biology approach combining metagenome, metatranscriptome, metaproteome and metabolome results linking the genotype with the phenotype to give a more complete picture. With respect to the metagenomic and metatranscript data, there are many challenges associated with the interpretation of microbial gene expression patterns at the community level. These arise again in part from the remarkable diversity and complexity of microbial communities in the ocean environment, and the lack of comprehensive representation in metagenomic databases (Frias-Lopez et al. 2008). In addition, correlation of results may not be

ACCEPTED MANUSCRIPT

straightforward, since a direct link between genes and metabolites often does not exist, for example, microorganisms have fewer metabolites than genes (Tang 2011). At this stage, considering the pitfalls associated with each 'omics method, and especially as outlined above with those associated with metabolomics, it is currently premature to directly compare community metabolomics data with metagenomic and metatranscript data. However, looking forward, a community metabolomics approach united with metagenomics, metatranscriptomics and metaproteomics should provide a powerful approach to reconstruct microbial ecosystems and understand their parts and network connectivity. Notably metabolomics should enable better annotation of hypothetical proteins by their association with known metabolites.

Another broad challenge is understanding the role that metabolomics could play in contributing functional trait information and providing a mechanistic foundation to better predict the function of communities. Trait based approaches have been used widely in terrestrial plant communities and have more recently been applied to provide a mechanistic foundation for understanding the structure and dynamics of phytoplankton in the English Channel and in U.S. lakes (Edwards et al. 2013a; Edwards et al. 2013b). Combining such a functional trait approach with community metabolomics could be powerful in revealing the mechanisms underlying community structure and in shaping marine ecosystem processes.

5. Conclusions

We have provided proof of concept in terms of using community metabolomics as an approach to discriminate metabolite patterns associated with marine POM and marine microbial communities. Using community metabolomics we could discriminate and characterise both the polar and lipid metabolite patterns. Our study highlights the power of metabolomics to discriminate marine POM without being restricted to focus on specific compound classes. Specifically, we were able to statistically distinguish different metabolite distributions in the four sampling locations revealing larger differences between the multiple samples taken from different sampling stations (and/or time points) compared to the more subtle differences associated with depth. Furthermore, using a 'non-targeted' metabolomics approach revealed differences in several individual and classes of metabolites present at these sites, reflecting and shaping the microbial community structure. Such a non-targeted approach has the advantage of highlighting compounds that have not yet been recognised to play an important role in microbial interactions and/or biogeochemical cycles. The majority of the metabolites that we putatively annotated were associated with oxylipins, oxidised TAGs and oligoglycans (simple carbohydrates). The preponderance of oxylipins could be particularly important in informing on the health of the community with a possible intriguing link to the formation of oxygenated volatile organic compounds that are important in atmospheric chemistry and in influencing climate. Within this manuscript we have highlighted some of the metabolites that showed differences in abundance at the sampled locations, many further metabolites are reported in the Supplementary Tables and this data will be further used in subsequent work linking with metagenome and metatranscriptome data.

This preliminary study shows that community metabolomics has the potential to be a powerful technique contributing to more comprehensive and unbiased characterisation of marine microbial populations. Combining metabolomics data with the massive and recent acceleration in genome sequencing capacities and increased resource of genetic information

ACCEPTED MANUSCRIPT

for marine phytoplankton will in the future greatly enhance our understanding of the metabolic processes by which microbes interact with their environment, as well as the evolution of their underlying metabolic pathways. Although such approaches are clearly still in their infancy, they should ultimately allow a systems biology approach to better understand how the microbes in the marine environment function and interact to control and drive the production and the biogeochemical cycles of our planet.

Acknowledgements

We thank Jack Gilbert for his role in coordinating the JCVI Expedition in the WEC and for support of this study, Dennis Cummings for supplying supporting physicochemical data, Rachel Harmer for supplying zooplankton abundance data and Ben Temperton for providing guidance on interpreting transcriptome data and for critical evaluation of the manuscript. This work was supported by the UK Natural Environmental Research Council's (NERC) Biomolecular Analysis Facility at the University of Birmingham (R8-H10-61) through an NBAF award (458: METMAP) and PML RP funding to CAL. The mass spectrometer was obtained through the Birmingham Science City Translational Medicine: Experimental Medicine Network of Excellence project, with support from Advantage West Midlands. Financial support was provided by the Beyster Family Fund of the San Diego Foundation and the Life Technologies Foundation to JCVI. Finally, we thank the anonymous reviewers who helped us improve the manuscript.

References

Abe, S., & Kaneda, T. (1975). Studies on effect of marine products on cholesterol metabolism in rat-XI, isolation of a new betaine, ulvaline, from a green laver *Monostroma nitidum* and its depressing effect on plasma cholesterol levels. Bulletin of the Japanese Society for the Science of Fish 41,567-57. doi: 10.2331/suisan.41.567

Arbona, V., Manzi, M., Ollas, C. De, & Gómez-Cadenas, A. (2013). Metabolomics as a tool to investigate abiotic stress tolerance in plants. International Journal of Molecular Sciences, 14(3), 4885-911. doi:10.3390/ijms14034885

Armada, I., Hachero-Cruzado, I., Mazuelos, N., Ríos, J. L., Manchado, M., & Cañavate, J. P. (2013). Differences in betaine lipids and fatty acids between *Pseudoisochrysis paradoxa* VLP and *Diacronema vlkianum* VLP isolates (Haptophyta). Phytochemistry, 95, 224-33. doi:10.1016/j.phytochem.2013.07.024

Baran, R., Bowen, B.P., Bouskill, N.J.,Brodie, E.L., Yannone, S.M., & Northen, T.R. (2010). Metabolite identification in Synechococcus sp. PCC 7002 using untargeted stable isotope assisted metabolite profiling. Analyical Chemistry. 82, 9034–9042, doi:10.1021/ac1020112.

Baran, R., Ivanova, N. N., Jose, N., Garcia-Pichel, F., Kyrpides, N. C., Gugger, M., & Northen, T. R. (2013). Functional genomics of novel secondary metabolites from diverse cyanobacteria using untargeted metabolomics. Marine Drugs, 11(10), 3617–31. doi:10.3390/md11103617

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate—a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological), 57,289-300. http://www.jstor.org/stable/2346101

ACCEPTED MANUSCRIPT

Becker, J. W., Berube, P. M., Follett, C. L., Waterbury, J. B., Chisholm, S. W., Delong, E. F., & Repeta, D. J. (2014). Closely related phytoplankton species produce similar suites of dissolved organic matter. Frontiers in Microbiology, 5, 111. doi:10.3389/fmicb.2014.00111

Benson, D.A., Karsch-Mizrachi, I., Clark, K., Lipman, D.J., Ostell, J., & Sayers, E.W. (2010) GenBank. Nucleic Acids Research. 38:D46-D51. doi: 10.1093/nar/gkp1024

Bidigare R.R., Frank T, Zastrow C & Brooks J.M. (1986). The distribution of algal chlorophylls and their degradation products in the Southern Ocean. Deep-Sea Research, 33 (7). 923-937. doi: 10.1016/0198-0149(86)90007-5

Brandsma, J., Hopmans, E.C., Brussaard, C.P.D., Witte, H.J., Schouten, S., & Damsté, J.S.S. (2012). Spatial distribution of intact polar lipids in North Sea surface waters: Relationship with environmental conditions and microbial community composition. Limnology and Oceanography, 57(4), 959-973. doi:10.4319/lo.2012.57.4.0959

Chitsaz, H., Yee-Greenbaum, J., Tesler, G., Lombardo, M.-J., Dupont, C.L., Badger, J.H., Novotny, M., Rusch, D.B., Fraser, L.J., Gormley, N.A., Schulz-Trieglaff, O., Smith, G.P., Evers, D.J., Pevzner, P.A., & Lasken, R.S. (2011) Efficient de novo assembly of single-cell bacterial genomes from short-read data sets. Nature Biotechnology, 29, 915-921. doi:10.1038/nbt.1966

Dieterle, F., Ross, A., Schlotterbeck, G., & Senn, H. (2006). Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in H-1 NMR metabonomics. Analytical Chemistry 78, 4281-90. doi: 10.1021/ac051632c

Dixon, J. L., Beale, R., & Nightingale, P. D. (2013). Production of methanol , acetaldehyde , and acetone in the Atlantic Ocean, Geophysical Research Letters, 40, 4700-4705. doi:10.1002/grl.50922

Duan, D., Zhang, B., Yao, J., Liu, Y., & Fang, J. (2014). Shikonin targets cytosolic thioredoxin reductase to induce ROS-mediated apoptosis in human promyelocytic leukemia HL-60 cells. Free radical biology & medicine, 70, 182–93. doi:10.1016/j.freeradbiomed.2014.02.016

Dupont, C.L., Rusch, D.B., Yooseph, S., Lombardo, M.-J., Richter, R.A., Valas, R., Novotny, M., Yee-Greenbaum, J., Selengut, J.D., Haft, D.H., Halpern, A.L., Lasken, R.S., Nealson, K., Friedman, R., & Venter, J.C. (2012) Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage. The ISME journal, 6, 1186-1199. doi:10.1038/ismej.2011.189

Edwards, K. F., Litchman, E., & Klausmeier, C. A. (2013a). Functional traits explain phytoplankton community structure and seasonal dynamics in a marine ecosystem. Ecology Letters, 16(1), 56-63. doi:10.1111/ele.12012

Edwards, K. F., Litchman, E., & Klausmeier, C. A. (2013b). Functional traits explain phytoplankton responses to environmental gradients across lakes of the United States. Ecology, 94 (7), 1626-1635. doi: 10.1890/12-1459.1

ACCEPTED MANUSCRIPT

Fernie, A. R., Obata, T., Allen, A. E., Araújo, W. L., & Bowler, C. (2012). Leveraging metabolomics for functional investigations in sequenced marine diatoms. Trends in Plant Science, 17(7), 395-403. doi:10.1016/j.tplants.2012.02.005

Flicek, P., Amode, M.R., Barrell, D., Beal, K., Brent, S., Chen, Y., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gordon, L., Hendrix, M., Hourlier, T., Johnson, N., Kahari, A., Keefe, D., Keenan, S., Kinsella, R., Kokocinski, F., Kulesha, E., Larsson, P…. Proctor, G., Vogel, J., & Searle, S.M. (2011). Ensembl 2012. Nucleic Acids Research, 39, D800-806. doi: 10.1093/nar/gkr991

Frias-Lopez, J., Shi, Y., Tyson, G. W., Coleman, M. L., Schuster, S. C., Chisholm, S. W., & Delong, E. F. (2008). Microbial community gene expression in ocean surface waters. Proceedings of the National Academy of Sciences of the United States of America, 105(10), 3805-10. doi:10.1073/pnas.0708897105

Gilbert, J. A, & Dupont, C. L. (2011). Microbial metagenomics: beyond the genome. Annual Review of Marine Science, 3, 347–71. doi:10.1146/annurev-marine-120709-142811. doi: 10.1146/annurev-marine-120709-142811

Gilbert, J. A, Field, D., Swift, P., Thomas, S., Cummings, D., Temperton, B., Mühling, M. (2010a). The taxonomic and functional diversity of microbes at a temperate coastal site: a "multi-omic" study of seasonal and diel temporal variation. PloS One, 5(11), e15545. doi:10.1371/journal.pone.0015545

Gilbert, J. A, Meyer, F., Schriml, L., Joint, I. R., Mühling, M., & Field, D. (2010b). Metagenomes and metatranscriptomes from the L4 long-term coastal monitoring station in the Western English Channel. Standards in Genomic Sciences, 3(2), 183-93. doi:10.4056/sigs.1202536

Gilbert, J. A, Steele, J. A, Caporaso, J. G., Steinbrück, L., Reeder, J., Temperton, B., Field, D. (2012). Defining seasonal marine microbial community dynamics. The ISME Journal, 6(2), 298–308. doi:10.1038/ismej.2011.107

Giovannoni, S. J., Cameron Thrash, J., & Temperton, B. (2014). Implications of streamlining theory for microbial ecology. The ISME Journal, 1-13. doi:10.1038/ismej.2014.60

Hasuoka, A., Nishikimi, Y., Nakayama, Y., Kamiyama, K., Nakao, M., Miyagawa, K. I., & Fujino, M. (2002). Synthesis and anti-*Helicobacter pylori* Activity of Pyloricidin Derivatives. Part 1. Structure—Activity Relationships on the Terminal Peptidic Moiety. ChemInform, 33(36).

Hrydziuszko, O., & Viant, M.R. (2012). Missing values in mass spectrometry based metabolomics: an undervalued step in the data processing pipeline. Metabolomics 8 (1S),161-74. doi:10.1007/s11306-011-0366-4

Ianora, A., Bentley, M. G., Caldwell, G. S., Casotti, R., Cembella, A. D., Engström-Öst, J., …..Vaiciute, D. (2011). The relevance of marine chemical ecology to plankton and ecosystem function: an emerging field. Marine Drugs, 9(9), 1625-48. doi:10.3390/md9091625

Ianora, A., & Miralto, A. (2010) Toxigenic effects of diatoms on grazers, phytoplankton and other microbes: a review. Ecotoxicolgy, 19(3) 493-511. doi: 10.1007/s10646-009-0434-y

ACCEPTED MANUSCRIPT

Jeffrey, S.W., Llewellyn, C.A., Barlow R.G. & Mantoura R.F.C. (1997). Pigment processes in the sea: a selected biography. pp167-178. In: Phytoplankton pigments in oceanography. Eds Jeffrey, S.W., Mantoura, R.F.C & Wright S.W. SCOR UNESCO Press.

Jones, O. A H., Sdepanian, S., Lofts, S., Svendsen, C., Spurgeon, D. J., Maguire, M. L., & Griffin, J. L. (2014). Metabolomic analysis of soil communities can be used for pollution assessment. Environmental Toxicology and Chemistry / SETAC, 33(1), 61-4. doi:10.1002/etc.2418

Kainz, M., Arts, M. T., & Mazumder, A. (2004). Essential fatty acids in the planktonic food web and their ecological role for higher trophic levels. Limnology and Oceanography. 49(5), 1784–1793.

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., & Tanabe, M. (2011) KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Research. 1-6. doi: 10.1093/nar/gkr988

Kato, M., Sakai, M., Adachi, K, Ikemoto H., Sano H. (1996) Distribution of betaine lipids in marine algae. Phytochemistry, 42 (5). 1341-1345. doi: 10.1016/0031-9422(96)00115-X

Kirwan, J.A., Weber, R.J.M., Broadhurst, D. I., & Viant, M.R. (2014). Direct infusion mass spectrometry metabolomics dataset: a benchmark for data processing and quality control. Sci. Data 1:140012. doi: 10.1038/sdata.2014.12

Laguerre, M., Lecomte, J., & Villeneuve, P. (2007). Evaluation of the ability of antioxidants to counteract lipid oxidation: existing methods, new trends and challenges. Progress in Lipid Research, 46(5), 244-82. doi:10.1016/j.plipres.2007.05.002

Landymore A.F., & Antia N, J. (1977). Growth of a marine diatom and a Haptophycean alga on phenylalanine or tyrosine serving as a sole nitrogen source. Journal of Phycology, 13, 231-238. doi: 10.1111/j.1529-8817.1977.tb02921.x

Lee, C., Wakeham, S., & Arnosti, C. (2004). Particulate organic matter in the sea: the composition conundrum. Ambio, 33(8), 565-75. doi.org/10.1579/0044-7447-33.8.565

Lee, D.Y., & Fiehn, O. (2008). High quality metabolomic data for *Chlamydomonas* reinhardtii. Plant Methods 4(7) doi:10.1186/1746-4811-4-7

Llewellyn, C. A., & Harbour, D. S. (2003). A temporal study of mycosporine-like amino acids in from surface water phytoplankton the English Channel and correlation with solar irradiation, Journal Marine Biology Association of the UK, 83,1-9. doi.org/10.1017/S0025315403006726h

Llewellyn, C. A., Fishwick, J.R, & Blackford, J.C. (2004). Phytoplankton community assemblage in the English Channel: a comparison using chlorophyll a derived from HPLC-CHEMTAX and carbon derived from microscopy cell counts. Journal of Plankton Research, 27(1), 103–119. doi:10.1093/plankt/fbh158

Llewellyn, C. A., & Airs, R. L. (2010). Distribution and abundance of MAAs in 33 species of microalgae across 13 classes. Marine Drugs, 8(4), 1273-91. doi:10.3390/md8041273

ACCEPTED MANUSCRIPT

Llewellyn, C.A., Mantoura, R.F.C. (1996). Pigment biomarkers and particulate carbon in the upper water column compared to the ocean interior of the northeast Atlantic. (1996) Deep-Sea Research Part I 43: (8) 1165-1184. doi: 10.1016/0967-0637(96)00043-X

Markowitz, V.M., Chen, I.M., Palaniappan, K., Chu, K., Szeto, E., Grechkin, Y., Ratner, A., Anderson, I., Lykidis, A., Mavromatis, K., Ivanova, N.N., & Kyrpides, N.C. (2010) The integrated microbial genomes system: an expanding comparative analysis resource. Nucleic Acids Research, 38, D382-390.

May, P., Wienkoop, S., Kempa, S., Usadel, B., Christian, N., Rupprecht, J., Walther, D. (2008). Metabolomics- and proteomics-assisted genome annotation and analysis of the draft metabolic network of *Chlamydomonas reinhardtii*. Genetics, 179(1), 157-66. doi:10.1534/genetics.108.088336

Nicholson, J.K., Holmes, E., Kinross, J., Burcelin, R., Gibson, G., Jia, W., Pettersson, S. (2012). Host-Gut Microbiota Metabolic Interactions. Science, 336(6086):1262-1267. doi: 10.1126/science.1223813

Niu, B.F., Fu, L.M., Sun, S.L., & Li, W.Z. (2010) Artificial and natural duplicates in pyrosequencing reads of metagenomic data. BMC Bioinformatics, 11. 187. doi:10.1186/1471-2105-11-187

Palenik, B., Grimwood, J., Aerts, A., Rouze, P., Salamov, A., Putnam, N., Dupont, C., Jorgensen, R., Derelle, E., Rombauts, S., Zhou, K., Otillar, R., Merchant, S.S., Podell, S., Gaasterland, T., Napoli, C., Gendler, K., Manuell, A., …., Y., Moreau, H., & Grigoriev, I.V. (2007) The tiny eukaryote *Ostreococcus* provides genomic insights into the paradox of plankton speciation. Proceedings of the National Academy of Sciences, 104, 7705-7710. doi: 10.1073/pnas.0611046104

Parsons, H.M., Ludwig, C., Gunther, U.L., & Viant, M.R. (2007). Improved classification accuracy in 1- and 2-dimensional NMR metabolomics data using the variance stabilising generalised logarithm transformation. BMC Bioinformatics, 8, 234. doi:10.1186/1471-2105-8-234

Payne, T.G., Southam, A.D., Arvanitis, T.N., & Viant, M.R. (2009). A signal filtering method for improved quantification and noise discrimination in Fourier transform ion cyclotron resonance mass spectrometry-based metabolomics data. Journal of the American Society for Mass Spectrometry, 20,1087–95. doi: 10.1016/j.jasms.2009.02.001

Pond, J.D., Harris,JR., Head,R. & Harbour. D. (1996) Environmental and nutritional factors determining seasonal variability in the fecundity and egg viability of *Calanus helgolandicus* in coastal waters off Plymouth, UK. Marine Ecology Progress Series 143, 45-63.

Pruesse, E., Quast, C., Knittel, K., Fuchs, B.M., Ludwig, W., Peplies, J., & Glockner, F.O. (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Research, 35, 7188-7196. doi: 10.1093/nar/gkm864

Rusch, D.B., Martiny, A., Dupont, C.L., Halpern, A.L., & Venter, J.C. (2010) Characterization of *Prochlorococcus* clades from iron depleted oceanic regimes. Proceedings of the National Academy of Sciences, 107, 16184-16189. doi: 10.1073/pnas.1009513107

ACCEPTED MANUSCRIPT

Schwarz, D., Nodop, A., Hüge, J., Purfürst, S., Forchhammer, K., Michel, K.-P., … & Hagemann, M. (2011). Metabolic and transcriptomic phenotyping of inorganic carbon acclimation in the Cyanobacterium *Synechococcus elongatus* PCC 7942. Plant Physiology, 155(4), 1640-55. doi:10.1104/pp.110.170225

Schwarz D, Orf I, Kopka J, & Hagemann M. (2013). Recent Applications of Metabolomics Toward Cyanobacteria. Metabolites, 3(1), 72-100. doi:10.3390/metabo3010072

Smyth, T. J., Fishwick, J. R., AL-Moosawi, L., Cummings, D. G., Harris, C., Kitidis, V., … & Woodward, E. M. S. (2010) A broad spatio-temporal view of the Western English Channel observatory. Journal of Plankton Research, 32(5), 585-601. doi:10.1093/plankt/fbp128

Southam A.D., Payne T.G., Cooper H.J., Arvanitis T.N, & Viant M.R. (2007). Dynamic range and mass accuracy of wide-scan direct infusion nanoelectrospray Fourier transform ion cyclotron resonance mass spectrometry based metabolomics increased by the spectral stitching method. Analytical Chemistry, 79, 4595-602. doi: 10.1021/ac062446p

Southam A.D., Lange A., Hines A., Hill E. M., Katsu Y., Iguchi T., Tyler C. R., & Viant M. R. (2011) Metabolomics reveals target and off-target toxicities of a model organophosphate pesticide to roach (*Rutilus rutilus*): Implications for biomonitoring. Environmental Science and Technology, 45, 3759-3767. doi: 10.1021/es103814d

Sumner L.W., Amberg A., Barrett D., Beale M.H., Beger R., Daykin C.A., Fan T.W.M., Fiehn O., Goodacre R., Griffin J.L., Hankemeier T., Hardy N., Harnly J., Higashi R., Kopka J., Lane A.N., Lindon J.C., Marriott P., Nicholls A.W., Reily M.D., Thaden J.J., & Viant M.R. (2007). Proposed minimum reporting standards for chemical analysis, *Metabolomics*, 2007, **3(3)**, 211-221. doi: 10.1007/s11306-007-0082-2

Tang, J. (2011). Microbial metabolomics. Current Genomics, 12(6), 391-403. doi:10.2174/138920211797248619

Tautenhahn, R., Patti, G.J., Rinehart, D., & Siuzdak, G. (2012). XCMS Online: A Web-Based Platform to Process Untargeted Metabolomic Data. Analytical Chemistry, 84 (11), 5035-5039. doi: 10.1021/ac300698c

Turnbaugh, P. J., & Gordon, J. I. (2008). An invitation to the marriage of metagenomics and metabolomics. Cell, 134(5), 708-13. doi:10.1016/j.cell.2008.08.025

Van Mooy, B. A S., Fredricks, H. F., Pedler, B. E., Dyhrman, S. T., Karl, D. M., Koblízek, M., Webb, E. a. (2009). Phytoplankton in the ocean use non-phosphorus lipids in response to phosphorus scarcity. Nature, 458(7234), 69-72. doi:10.1038/nature07659

Vemuri, G. N., Aristidou, A. A., Vemuri, G. N., & Aristidou, A. A. (2005). Metabolic Engineering in the -omics Era : Elucidating and Modulating Regulatory Networks. 69(2). 197-216. doi:10.1128/MMBR.69.2.197

Viant, M.R., & Sommer, U. (2013). Mass spectrometry based environmental metabolomics: a primer and review. Metabolomics 9, S144-58. doi: 10.1007/s11306-012-0412-x

Vidoudez, C., & Pohnert, G. (2011). Comparative metabolomics of the diatom *Skeletonema marinoi* in different growth phases. Metabolomics, 8(4), 654-669. doi:10.1007/s11306-011-0356-6

ACCEPTED MANUSCRIPT

Weber R.J.M., & Viant, M.R. (2010). MI-Pack: increased confidence of metabolite identification in mass spectra by integrating accurate masses and metabolic pathways. Chemom Intell Lab Syst. 104,75-82. doi: 10.1016/j.chemolab.2010.04.010

Weber, RJ.M., Southam, A.D, Sommer, U., & Viant, M.R. (2011). Characterization of isotopic abundance measurements in high resolution FT-ICR and Orbitrap mass spectra for improved confidence of metabolite identification. Analytical Chemistry 83, 3737-43. doi: 10.1021/ac2001803

Westerhuis, J.A., Hoefsloot H.C.J., Smit S., Vis D.J., Smilde A.K., van Velzen E.J.J., van Duijnhoven J.P.M., & van Dorsten F.A. (2008). Assessment of PLSDA cross validation. Metabolomics 4(1), 81-89. doi:10.1007/s11306-007-0099-6

Widdicombe, C. E., Eloire, D., Harbour, D., Harris, R. P., & Somerfield, P. J. (2010). Long-term phytoplankton community dynamics in the Western English Channel. Journal of Plankton Research, 32(5), 643-655. doi:10.1093/plankt/fbp127

Williams, T. J., & Cavicchioli, R. (2014). Marine metaproteomics: deciphering the microbial metabolic food web. Trends in Microbiology, 22(5), 248-260. doi:10.1016/j.tim.2014.03.004

Wu, H.F., Southam, A.D., Hines, A., & Viant, M.R. 2008. High throughput tissue extraction protocol for NMR and MS based metabolomics. Analytical Biochemistry, 372, 204-12. doi: 10.1016/j.ab.2007.10.002

Zhou, S., Gonzalez, L., Leithead, a., Finewax, Z., Thalman, R., Vlasenko, a.,. & Abbatt, J. (2014). Formation of gas-phase carbonyls from heterogeneous oxidation of polyunsaturated fatty acids at the air–water interface and of the sea surface microlayer. Atmospheric Chemistry and Physics, 14(3), 1371-1384. doi:10.5194/acp-14-1371-2014

Supplementary Tables

Table SI1: Results from PCA scores test on model of all 47 biological samples for polar extracts.

Table SI2: PLS-DA for polar extracts.

Table SI3: Statistics and MS signal annotations for polar extracts of WEC samples

Table SI4: Results from PCA scores test on model of all 47 biological samples for lipid extracts.

Table SI5: PLS-DA for lipid extracts.

Table SI6: Statistics and MS signal annotations for lipid extracts of WEC samples

Table SI7: Statistics and MS signal annotations for selected signals (available electronically).

Table SI8: Zooplankton abundance data from the surface waters at L4 and E1.

ACCEPTED MANUSCRIPT

Supplementary Figures

Figure SI1: PCA on all samples (polar extracts) including QC samples.

Figure SI2: PCA on all samples (lipid extracts) including PCA samples.

Figure SI3: Expansion of Figure. 4 providing phytoplankton community transcript activity on a wider range of species at the four sampling locations.

Figure SI4: Metatranscriptome chromosome reads for *Ostreococcus lucimarinus.*

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Table 1. Physico-chemical properties of the water at the time of sampling the two stations

	L4-2m	L4-17m	E1-1m	E1-70m
Date	21st May 2009	21st May 2009	28th May 2009	28th May 2009
Number of samples	12	12	12	12
Time	12:00pm	12:00pm	10:30am	10:30am
Latitute	50.25	50.25	50.03	50.03
Longitude	-4.22	-4.22	-4.34	-4.34
Total Water Column (m)	55	55	73.2	73.2
Thermocline (m)	13	13	20	20
Sample Depth (m)	2	17	1	70
Temperature (°C)	12	11	12.44	10.77
Salinity (PSU)	35.00	35.00	35.18	35.28
Oxygen (µmol/kg)	6.10	6.10	5.98	6.20
pH (log of[H^+])	8.4	8.4	8.4	8.3

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Table 2

Observed		Statistics							Annotation		
		rank	rank	rank		ratio					
M/Z	Extr	E1-L4	L4	E1	Adj.P	L4_2m	L4_17m	E1_1m	Ion form	Final Annotation	Compound group
166.08614	P	39	18	89	4.5E-03	2.99	1.12	1.24	$[M+H]^+$	Phenylalanine	Amino acids and derivatives
182.08107	P	62	12	41	4.7E-02	2.22	1.05	1.44	$[M+H]^+$	Tyrosine	Amino acids and derivatives
246.09711	P	42	35	3	1.0E-06	4.40	4.14	7.98	$[M+H]^+$	Mycosporine-glycine	Amino acids and derivatives
343.14985	P	78	1	10	0.0E+00	3.49	34.12	14.21	$[M+H]^+$	Pyloricidin C	Amino acids and derivatives
236.14915	P	21	6	11	0.0E+00	10.88	2.27	4.57	$[M+H]^+$	Ulvaline	Amino acids and derivatives, DGTS backbone
682.56203	L	58	724	4	1.6E-04	4.65	5.73	4.51	$[M+H]^+$	DGTS 30:1	DGTS lipid
704.54636	L	115	1149	3	1.7E-04	3.30	2.83	3.51	$[M+H]^+$	DGTS 32:4	DGTS lipid
730.56191	L	69	516	5	1.1E-04	4.16	5.62	4.21	$[M+H]^+$	DGTS 34:5	DGTS lipid
732.57806	L	36	174	1	1.3E-04	5.43	8.86	6.30	$[M+H]^+$	DGTS 34:4	DGTS lipid
736.60936	L	1751	1632	37	7.0E-02	1.56	1.31	2.07	$[M+H]^+$	DGTS 34:2	DGTS lipid
758.59355	L	132	285	11	1.7E-04	2.83	4.47	3.51	$[M+H]^+$	DGTS 36:5	DGTS lipid
804.57769	L	119	978	16	2.0E-05	3.73	3.06	2.89	$[M+H]^+$	DGTS 40:10	DGTS lipid
856.60868	L	49	1378	13	1.1E-04	4.80	4.44	3.13	$[M+H]^+$	DGTS 44:12	DGTS lipid
289.17741	P	11	19	13	0.0E+00	6.30	4.56	0.73	$[M+Na]^+$	Hexadecanoid	Fatty acids and oxylipins
291.19309	P	6	4	24	0.0E+00	10.47	1.49	0.78	$[M+Na]^+$	Hexadecanoid	Fatty acids and oxylipins
301.21638	P	36	44	117	2.0E-06	4.71	3.58	2.18	$[M+H]^+$	Eicosahexaenoic acid (20:6)	Fatty acids and oxylipins
313.17772	L	165	24	814	0.0E+00	4.10	1.42	0.93	$[M+Na]^+$	Octadecanoid	Fatty acids and oxylipins
315.19312	P	5	100	14	0.0E+00	9.47	3.78	0.82	$[M+Na]^+$	Octadecanoid	Fatty acids and oxylipins
319.16451	P	44	7	110	0.0E+00	3.63	5.61	2.41	$[M+2Na-H]^+$	Octadecapentaenoic acid	Fatty acids and oxylipins

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

341.20876	P	30	26	81	0.0E+00	3.84	3.04	1.74	$[M+Na]^+$	Eicosanoid	Fatty acids and oxylipins
357.20370	P	32	86	151	1.0E-06	3.73	2.10	1.37	$[M+Na]^+$	Eicosatetraenedioic acid	Fatty acids and oxylipins
393.18047	P	16	85	19	0.0E+00	5.87	3.24	0.78	$[M+2Na-H]^+$	Tetracosadecaenoic acid	Fatty acids and oxylipins
393.29785	L	11	141	712	1.6E-05	0.21	0.24	0.92	$[M+Na]^+$	Docosanedioic acid	Fatty acids and oxylipins
252.14406	P	28	72	22	0.0E+00	5.52	2.65	2.70	$[M+H]^+$	Gluconamide or hexapyranoside	Glycosylated compound
277.08928	P	7	11	104	0.0E+00	9.34	7.02	2.22	$[M+Na]^+$	Hexosyl-glycerol	Glycosylated compound
329.13419	P	31	36	1	0.0E+00	28.62	10.47	41.10	$[M+H]^+$	Cyanogenic glycoside	Glycosylated compound
347.14476	P	12	2	2	0.0E+00	27.67	2.58	11.61	$[M+H]^+$	poss. glycoside	Glycosylated compound
573.23173	P	146	39	9	0.0E+00	1.71	1.26	3.85	[M+K]+	Glycoside	Glycosylated compound
434.11807	P	17	78	44	1.0E-06	18.99	10.09	5.49	$[M+K+H]^{2+}$	Hex5 (2+)	oligoglycan
527.15814	P	23	38	42	0.0E+00	16.68	10.75	6.00	$[M+Na]^+$	Hex3	oligoglycan
649.21804	P	4	5	12	0.0E+00	21.11	16.31	6.63	$[M+H]^+$	Hex4-H2O	oligoglycan
671.20000	P	15	14	38	0.0E+00	19.45	14.47	7.04	$[M+Na]^+$	Hex4-H2O	oligoglycan
811.27102	P	8	8	17	0.0E+00	17.72	14.02	6.65	$[M+H]^+$	Hex5-H2O	oligoglycan
851.26387	P	13	50	35	1.0E-06	22.38	11.02	5.97	$[M+Na]^+$	Hex5	oligoglycan
457.25669	L	399	23	1204	2.6E-02	1.43	3.91	1.01	$[M+H]^+$	lysoPG 14:0	Phospholipids
730.47734	L	838	35	1520	1.3E-02	0.50	1.34	1.14	$[M+K]^+$	PC 29:0 or PE 32:0	Phospholipids
828.55313	L	1853	31	1542	1.5E-01	0.84	1.95	1.17	$[M+H]^+$	PC 40:9 or PE 43:9	Phospholipids
908.70797	L	14	998	1131	4.7E-04	0.27	0.33	0.87	$[M+Na]^+$	PC 43:1 or PE 46:1	Phospholipids
914.66069	L	50	353	1398	0.0E+00	0.33	0.24	0.82	$[M+Na]^+$	PC 44:5 or PE 47:5	Phospholipids
450.35838	L	122	1711	144	6.4E-04	0.25	0.30	0.45	$[M+H]^+$	TAG 48:4	Tri- or diacylglycerides
739.52620	L	498	22	24	3.5E-01	0.69	2.66	2.63	$[M+Na]^+$	DAG 44:10	Tri- or diacylglycerides
893.69890	L	202	358	191	2.2E-02	2.63	1.50	0.85	$[M+K]^+$	TAG 42:4	Tri- or diacylglycerides
921.72974	L	99	19	157	1.2E-01	4.90	2.18	2.14	$[M+K]^+$	TAG 44:4	Tri- or diacylglycerides
803.54429	L	2	1	606	4.0E-04	0.22	0.04	0.68	$[M+Na]^+$	2x oxidized TAG 45:8	Triacylglyceride - oxidized
907.70356	L	15	672	190	2.0E-06	0.19	0.18	0.59	$[M+H]^+$	oxidized TAG 54:8	Triacylglyceride - oxidized
913.65717	L	28	139	1071	2.0E-06	0.31	0.20	0.75	$[M+Na]^+$	2x oxidized TAG 55:12	Triacylglyceride - oxidized
927.67120	L	6	27	10	3.0E-06	0.27	0.14	0.53	$[M+H]^+$	2x oxidized TAG 56:12	Triacylglyceride - oxidized
929.68055	L	150	228	1450	4.1E-04	0.59	0.43	0.93	$[M+Na]^+$	2x oxidized TAG 54:8	Triacylglyceride - oxidized

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

983.73598	L	1	338	746	2.0E-05	0.08	0.04	0.62	$[M+H]^+$	2x oxidized TAG 60:12	Triacylglyceride - oxidized

ACCEPTED MANUSCRIPT