

Diel Oscillations of Particulate Metabolites Reflect Synchronized Microbial Activity in the North Pacific Subtropical Gyre

Angela K. Boysen¹, Laura T. Carlson¹, Bryndan P. Durham², Ryan D. Groussman¹, Frank O.
Aylward³, François Ribalet¹, Katherine R. Heal¹, Edward F. DeLong⁴, E. Virginia Armbrust¹,
Anitra E. Ingalls¹

Affiliations:

¹School of Oceanography, University of Washington, Seattle, WA, USA. ² Department of
Biology, Genetics Institute, University of Florida, Gainesville, FL, USA. ³ Department of
Biological Sciences, Virginia Tech, Blacksburg, VA, USA. ⁴ Daniel K. Inouye Center for
Microbial Oceanography: Research and Education (C-MORE), University of Hawaii, Honolulu,
HI, USA

Supplemental Methods

Metabolite sample extraction

Metabolites were extracted as in as in Boysen et al.(1). Briefly, frozen filters were cut
into small pieces and put into bead beating tubes with silica beads, heavy isotope-labeled internal
standards (Table S1), and cold aqueous (50:50 methanol:water) and organic solvents
(dichloromethane). The samples were shaken on a FastPrep-24 Homogenizer for 30 seconds and
chilled in a -20 °C freezer, repeated for three cycles. The organic and aqueous layers were
separated by spinning samples in a microcentrifuge at 5,000 rpm for 90 seconds at 4 °C. The
aqueous layer was removed to a new combusted glass centrifuge tube. The remaining organic
fraction was rinsed three times with 50:50 methanol:water. All aqueous rinses were combined

with the original aqueous extract and dried down under N₂ gas. The remaining organic layer was transferred into a clean glass centrifuge tube and the bead beating tube was rinsed two more times with cold organic solvent. The combined organic rinses were centrifuged, transferred to a new tube, and dried under N₂ gas. Dried aqueous and organic fractions were re-dissolved in 400 μ L of water and 400 μ L of 1:1 water:acetonitrile, respectively. Isotope-labeled injection standards were added to both fractions (Table S1).

Metabolite data acquisition

LC-MS parameters were as in Boysen et al.(1). Data for most compounds were collected on a Waters Xevo TQ-S triple quadrupole (TQS), but for compounds outside the linear range or which were originally detected in sub-optimal polarity, a Thermo QExactive HF (QE) mass spectrometer with ESI was used. Chromatography and mass spectrometry methods follow those reported in Boysen et al.(1), and details are provided in the supplemental methods. The organic fraction was analyzed using reversed phase chromatography (Waters Acquity UPLC HSS Cyano column, 1.8 μ m particle size, 2.1 mm x 5 mm) and the aqueous fraction was analyzed with both reversed phase and hydrophilic liquid interaction chromatography (HILIC - SeQuant ZIC-PHILIC column, 5 mm particle size, 2.1 mm x 150 mm, from Millipore). We monitored 101 compounds with reversed phase and 110 compounds with HILIC.

Data collected on a Thermo QExactive HF (QE) with ESI was used for nucleic acids and nucleosides (adenine, guanine, thymine, cytosine, adenosine, guanosine, thymidine, cytidine) and compounds which were overloaded on the triple quadrupole (glycine betaine, homarine, phenylalanine, DMSP) but which were within the linear range on the QE. On the QE, for HILIC, a full scan method with polarity switching was used with 60,000 resolution. For RP, positive

ionization mode was used with a resolution of 120,000. For the QE data, proteowizard was used to convert .raw files to .mzxml(2).

Peak integrations were performed using Skyline for small molecules(3). Isoleucine and leucine did not always chromatographically separate and were treated as a single metabolite, (iso)leucine. Data were subjected to in-house quality control (QC) that removed misidentified compounds, removed compounds with a low signal to noise ratio ($S/N < 4$), and flagged compounds that were detected in the blanks (signal intensity in the sample must be greater than 3 times that in the blank). Compounds which were below the level of the blank in greater than 30% of samples (40 out of 132) were discarded. Compounds which did not pass the QC for 6 or more samples were discarded. For compounds which had fewer than 6 samples fail the QC, if the compound was detected in a sample but comparable to the blank, that data was retained. If the signal in a sample was less than the signal in the blank, a value equal to half the average blank value was filled back in to reflect the limit of detection.

Metabolite data curation

The normalization procedure for metabolomics data follows the Best-Matched Internal Standard method described in Boysen et al.(1) Pooled samples at full and half strength (diluted 1:1 with water, for the aqueous fraction, or solvent, for the organic fraction) were run after the samples were run in order to train the normalization algorithm. An improvement of 10% in the relative standard deviation (RSD) of the pooled sample was required in order to apply normalization. Compounds with a raw RSD of less than 0.01 were not normalized. The low cutoff was used to encourage normalization because the data were collected over a period of five weeks of instrument time; normalizing to an internal standard is an improvement over the instrument variability that we expect to influence mass spectral measurements over this time

period. All QE Orbitrap data were either used as raw data (no normalization) or, in the case of phenylalanine, normalized to its isotopologue internal standard, as the high resolution instrumentation experiences less ion suppression compared to that on the TQ-S instrument(1).

Prior to normalization, the peak areas of the internal standards in each sample were assessed to detect run quality and extraction efficiency. This assessment showed that a number of samples, mostly those collected as part of the second sampling period, clearly had half or twice the appropriate standard concentration added. (samples with internal standards adjusted in the 1st sampling period: 5R1, 6R1, second sampling period: 26R2, 26R3, 27R1, 27R2, 27R3, 28R1, 28R3, 29R2, 30R1, 31R1, 32R3, 34R1, 34R3, 35R1, 35R2, 37R1, 39R1). Peak areas were adjusted according to these observations as well as laboratory notebook records for the samples which had incorrect concentrations. The periodicity analysis was done both with and without these replicates and the results are robust to this change. Additionally, samples collected between July 29th 14:00 and July 30th 02:00 did not have some of the standards added during sample processing. Data from the effected compounds were removed for those samples as normalization was not possible (as indicated in Table S2).

Euclidean distance of samples based on z-score standardized metabolite profiles showed that single replicates from two different samples (31R2, 41R1) in the second sampling period were outliers (> 3 standard deviations away from the mean average distance), so these samples were removed prior to further analysis.

Quantification of select metabolites

Metabolites with isotope-labeled authentic standards (see list in Table 3) were quantified using the following formula:

$$[M]_{SW} = \frac{PkA_M}{PkA_{IS}} * [IS]_{vial} * \frac{V_{vial}}{V_{filtered}}$$

where M indicates metabolite, V indicates volume, IS indicates internal standard, SW indicates seawater, and PkA is the integrated LC-MS peak area.

Isotope-labeled standards for trehalose and sucrose were purchased after the full sample set had been processed. A subset of samples ($n = 19$) was spiked with the internal standards and re-analyzed on the LC-TQS-MS, concentrations were determined as above, and a linear regression relating peak area to trehalose and sucrose concentration was fitted according to the formula:

$$[M]_{vial} = A \times (NormalizedPkA)$$

where A is a constant. This regression was used to calculate the concentrations of trehalose and sucrose in the rest of the samples that were not reanalyzed.

An authentic standard for 2,3-dihydroxypropane-1-sulfonate (DHPS) was obtained after the full sample set had been processed and the quantification of this compound using standard additions is described in Durham et al(4).

Estimated Concentrations of Metabolites

Approximate concentrations correcting for ionization efficiency and ion suppression were calculated using the equation:

$$Concentration = \frac{Area}{IE} * \frac{V_{vial}}{V_{filtered}} * \frac{1}{RF_{ratio}}$$

where IE is the ionization efficiency, calculated by taking the average peak area of a standard injected in water divided by the concentration of the standard. RF_{ratio} is the response factor ratio published in Boysen and Heal et al. (2018), calculated by taking the ratio of the peak area of a standard injected in environmental matrix (less the ambient matrix signal, if applicable) to the peak area of the standard injected in water. Published values were used with the exceptions of taurine, chitobiose, and for nucleosides and nucleotides where we used values calculated from a different sample set also from near Station ALOHA, so with similar matrix effects.

We did not make estimates of concentration for five compounds: methylthioadenosine (MTA) in the aqueous fraction because we didn't have appropriate standards at the time of sample analysis, EPA, DHA, ergosterol, and trans retinal because insufficient isotope labeled internal standards were used for the lipid fraction.

Particulate carbon and particulate nitrogen discrete sample analysis

Samples for particulate nitrogen and particulate carbon were collected from the ship's underway flow through seawater, which sits at ~7 m water depth. Samples were collected onto combusted GFF filters. Filters were folded and stored in combusted aluminum foil at -80 °C until analysis, when they were thawed and dried overnight at 60 °C, balled into Ag and Sn boats, and analyzed using high-temperature combustion (1020 °C) on a ThermoQuest NC 2500 elemental analyzer.

Eukaryotic metatranscriptome assembly

Metatranscriptome assembly was conducted using Trinity(5) on the Pittsburgh Supercomputing Center's Bridges Large Memory system. Parameters include using in-silico normalization, a

minimum k-mer coverage of 2, and a minimum contig length of 300. The raw assemblies were quality controlled with Transrate v1.0.3(6). To eliminate redundancy and duplication, the assemblies were merged and clustered at the 99% amino-acid identity threshold level with linclust in the MMseqs2 package(7).

Eukaryotic metatranscriptome reference database

We supplemented the curated MarineRefII reference database (<http://roseobase.org/data/>) with additional representatives of marine animal, fungal, protist and viral reference sequences (Table S5), totaling 641 marine eukaryotes and prokaryotes.

Detecting periodicity and estimating time of peak concentration/abundance

When conducting the RAIN analysis outliers removed during quality control were filled back in by taking an average of the other two replicates collected at that time point.

The time of peak concentration for each oscillating metabolite and peak abundance for each oscillating transcript was calculated by fitting a periodic oscillator according to the function:

$$[M] = A * \cos\left(\frac{2\pi}{24} * HourCollected\right) + B * \sin\left(\frac{2\pi}{24} * HourCollected\right)$$

as in Ottesen et al, 2014(8). The lag time between metabolites and transcripts coding for proteins that use or produce them was estimated simply by taking the difference between the two peak times, in hours.

Multivariate analysis

Metabolite concentrations were standardized to their z-score across the samples, having a mean of zero and standard deviation of one. Samples and metabolites were clustered based on Euclidean distances with the vegdist and metaMDS functions from the vegan R package (version 2.5-6)(9). Average-linkage clustering was calculated using the hclust function from the stats R package (version 3.6.1).

We used nonmetric multidimensional scaling (NMDS) to explore differences in samples based on their metabolite profiles. NMDS is an ordination technique that uses rank-order similarity between samples rather than absolute differences between samples to reduce dimensionality in the data(10). NMDS using a Euclidean distance matrix was chosen rather than principal components analysis (PCA) because NMDS relaxes the assumption that there are fewer variables than samples, which is not the case in metabolomics where data on the abundance of hundreds of compounds is collected for each sample. Additionally, NMDS avoids the assumption of linear relationships among variables.

NMDS was run with two ordination axes and 100 random starts. Significance of the stress value was tested with a Monte Carlo randomization test. Goodness of fit was assessed by correlating NMDS ordination results and Euclidean distances using both a non-metric and linear fit. Analysis of Group Similarities (ANOSIM) was used to test differences between the times of day that samples were collected as well as differences between the days of collection. ANOSIM was conducted on a Euclidean distance matrix with 1000 permutations. Pairwise ANOSIM was conducted to further clarify which times were significantly different.

Phytoplankton culture growth conditions

Cultures of *Crocospaera watsonii* WH8501 were obtained from the Zehr lab at the University of California, Santa Cruz. They were grown with a square 12:12 light:dark cycle with 50 $\mu\text{mol photons m}^{-2} \text{ s}^{-1}$, at 26 °C, and had an average growth rate of 0.17 d⁻¹. Cells were collected in mid-to-late exponential phase by gentle vacuum filtration onto 0.2 μm Omnipore filters using combusted borosilicate filter towers. Cells were enumerated using a Beckman coulter counter.

Mixed layer depth

Mixed layer depth was calculated from CTD profiles and is defined by a 0.03 kg/m³ density offset from 10 db.

Supplemental calculation: Trehalose fueling nitrogen fixation

To convert half a mol N₂ to one mol NH₃ requires 3 mol electrons and 8 mol ATP(11). One mol glucose can provide 24 mol electrons or 30–36 mol ATP, leading to a stoichiometry of between 2.08–2.35 moles CH₂O to produce sufficient ATP and electrons to produce one mole of fixed NH₃(12). From Wilson and Aylward et al. *Crocospaera* is responsible for 7.3 nmol N l⁻¹ d⁻¹ (13), which therefore requires between 15.2 and 17.2 nM of respired carbon. Our data has a range of 1.6 to 4.3 nM C in the form of trehalose drawn down every night, with a mean drawdown of 3.2 nM C in the form of trehalose. Using this range of required carbon and trehalose drawn down, we estimate that trehalose catabolism could fuel between 9% and 28% of N₂ fixation, with a mean value of 20%.

Supplemental Results and Discussion

196

197 *Multivariate Analysis*

198 NMDS analysis of the samples produced a low stress value of 0.18 which was significant
199 ($p < 0.01$), indicating that the sample scores are robust. The fit between the ordination distance
200 and Euclidean distance had a non-metric R^2 value of 0.97 and a linear fit R^2 value of 0.88. NMDS
201 analysis of the samples from the second sampling period (Figure S2) produced a low stress value
202 of 0.17 which was significant ($p < 0.01$), indicating that the sample scores are robust. The fit
203 between the ordination distance and Euclidean distance had a non-metric R^2 value of 0.97 and a
204 linear fit R^2 value of 0.89. The overall similarity observed at 6:00 during the first collection
205 period is not seen during the second collection period. NMDS and ANOSIM analysis of the
206 samples from the second collection period and full dataset are unable to discriminate between
207 samples collected at different times of day (Supplemental Figure 2), providing additional
208 evidence that community synchrony, as illustrated by overall metabolite composition, weakened
209 as sampling progressed.

210

211 *Metabolites lose diel periodicity in second sampling period*

212 Fewer metabolites had diel periodicity in the second sampling period (Figure S1), with 9
213 and 23 compounds exhibiting diel periodicity when analyzed as molar concentrations (nmol L^{-1})
214 and normalized to POC ($\text{nmol } \mu\text{mol POC}^{-1}$), respectively (Table S2, Figure S1). The reduction of
215 diel oscillation in the second sampling period is seen in multivariate analyses as well, where in
216 the second sampling period no sampling times clustered significantly (Figure S2).

217 Because these samples were processed over several weeks, it is possible that
218 methodological issues arose and masked the diel oscillations during the second half of the

sample set. We deem this unlikely because we still measure strong diel oscillations in some compounds over the course of the whole sample period. Overall there was not a drop in signal intensity from the instruments over the sample processing, and the increase in *Prochlorococcus* concentrations shows, if anything, that there would be more biomass rather than less biomass to produce a robust signal for the latter half of the data (Figure 1).

Supplemental references

1. Boysen AK, Heal KR, Carlson LT, Ingalls AE. Best-Matched Internal Standard Normalization in Liquid Chromatography-Mass Spectrometry Metabolomics Applied to Environmental Samples. *Anal Chem*. 2018;90(2).
2. Chambers MC, MacLean B, Burke R, Amodei D, Ruderman DL, Neumann S, et al. A cross-platform toolkit for mass spectrometry and proteomics. Vol. 30, *Nature Biotechnology*. Nature Publishing Group; 2012. p. 918–20.
3. MacLean B, Tomazela DM, Shulman N, Chambers M, Finney GL, Frewen B, et al. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics*. 2010;26(7):966–8.
4. Durham BP, Boysen AK, Carlson LT, Groussman RD, Heal KR, Cain KR, et al. Sulfonate-based networks between eukaryotic phytoplankton and heterotrophic bacteria in the surface ocean. *Nat Microbiol*. 2019;4(10):1706–15.
5. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol*. 2011;29(7):644–52.
6. Smith-Unna R, Boursnell C, Patro R, Hibberd JM, Kelly S. TransRate: reference-free

242 quality assessment of de novo transcriptome assemblies. *Genome Res.* 2016;26(8):1134–
243 44.

244 7. Steinegger M, Söding J. Clustering huge protein sequence sets in linear time. *Nat*
245 *Commun.* 2018;9(1):2542.

246 8. Ottesen E a., Young CR, Gifford SM, Eppley JM, Marin R, Schuster SC, et al.
247 Multispecies diel transcriptional oscillations in open ocean heterotrophic bacterial
248 assemblages. *Science* (80-). 2014;345(6193):207–12.

249 9. Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O’hara RB, et al. Package
250 ‘vegan.’ *Community ecology package*; 2013. p. 1–295.

251 10. Kruskal JB, Wish M. *Multidimensional scaling*. SAGE Publications; 1978.

252 11. Sohm JA, Webb EA, Capone DG. Emerging patterns of marine nitrogen fixation. *Nat Rev*
253 *Microbiol.* 2011;9(7):499–508.

254 12. Großkopf T, LaRoche J. Direct and indirect costs of dinitrogen fixation in *Crocospaera*
255 *watsonii* WH8501 and possible implications for the nitrogen cycle. *Front Microbiol.*
256 2012;3(JUL):236.

257 13. Wilson ST, Aylward FO, Ribalet F, Barone B, Casey JR, Connell PE, et al. Coordinated
258 regulation of growth, activity and transcription in natural populations of the unicellular
259 nitrogen-fixing cyanobacterium *Crocospaera*. *Nat Microbiol.* 2017;2(July):1–20.

260