nature research

Corresponding author(s): David Valentine

Last updated by author(s): 14_12_2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	no software was used for data collection
Data analysis	FlowJo version 9.8, RStudio with R base stats package version 1.2.1335, DADA2 R package version 1.9.3, Trimmomatic version 0.36, Sickle version 1.33,metaSPAdes version 3.8.1 , QUAST version 5.02, Bowtie2 version 2.3.4.1, Samtools version 1.7, Centrifuge version 1.0.1, CheckM version 1.0.7, GTDB-Tk version 1.0.2, Prodigal version 2.6.3, HMMER version 3.1b2, Pfam database version 31.0, KofamScan version 1.1.0, Geneous Prime version 2019.2.3, MUSCLE version 3.8.425, TrimAL version 1.2, RAxML version 8.2.9, FigTree version 1.4.3, BLAST Version 2.7.1

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All oceanographic, chemical and cell count data is available at the Biological and Chemical Oceanography Data Management Office website under project code NSF OCE-1635562 (doi:10.26008/1912/bco-dmo.826878.1). Metagenomes are available through NCBI in BioProject PRJNA657625 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA657625). Databases accessed were the Genome Taxonomy Database (https://data.ace.uq.edu.au/public/gtdb/data/releases/release89/89.0/, Version r89), the Pfam database (ftp://ftp.ebi.ac.uk/pub/databases/Pfam/releases/Pfam31.0, Version 31.0) and the Ocean Microbial Reference Gene Catalogue (http://ocean-microbiome.embl.de/companion.html).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	The study involved chemical and biological sampling of waters in the Gulf of Mexico and the North Atlantic Subtropical Gyre. Work conducted in the North Atlantic Subtropical Gyre primarily involved numerical chemical, cell count and nutrient data. Experiments were conducted with n = 3 replication per condition and environmental sampling (depth profiles, diel sampling) were conducted with n = 2 replication. Hydrocarbon consumption incubations were conducted in the Gulf of Mexico and the North Atlantic Subtropical Gyre and involved incubation experiments with n = 6 replication per condition, data are primarily numerical chemical data and sequence data. Data also include sequence data from TARA Oceans dataset.
Research sample	The research sample was primarily chemical compounds and sequences from native microbial populations within ocean water or populations altered by bottle incubation experiments (both hydrocarbon production and consumption). The rationale for this sample choice was to provide accurate identification and quantification of microbial communities that were responsible for the production and consumption of pentadecane and other hydrocarbons.
Sampling strategy	Sampling procedure primarily involved filtering water for chemical and sequence data. No sample-size calculation was performed. Sample size was chosen simply by the maximum amount of work that could be conducted on the respective cruises. These sample sizes are sufficient for experimentation because of tight precision between replicates (n = 3) and for more broad scale analyses (across all stations of cruise), sample sizes exceed 25 distinct samples for statistical tests.
Data collection	Pentadecane quantification and isotope enrichment data from environmental sampling and hydrocarbon production experiments were collected by Connor Love and Kelsey Gosselin (quantification only) by Gas Chromatography Flame Ionization Detection and Gas Chromatography Combustion Isotope Ratio Mass Spectrometer. Data were collected via Gas Chromatography and peak integration of extracted samples. Eleanor Arrington collected oxygen loss data via oxygen optodes and sequence data from hydrocarbon consumption experiment incubations. Benjamin Van Mooy collected data for nutrients and cyanobacteria cell counts (via flow cytometry).
Timing and spatial scale	Samples were taken on a cruise during May 2017 (5/5/17- 5/20/17) in the North Atlantic Subtropical Gyre spanning ~10 degrees of latitude. Sampling for pentadecane production experiments and depth profiles typically switched off each day with an aim to capture the standing stock of pentadecane and production dynamics equally. Pentadecane consumption experiments were conducted twice on this cruise because of the large time and material expenditure of these experiments, with daily monitoring of the bottles after the cruise to check for oxygen loss. This sampling scheme was chosen to maximize sampling opportunity during the cruise. Gulf of Mexico samples were taken during a cruise from 6/15/2015-6/29/2015, samples were taken 6/15/2015-6/17/2020 due to only a few days being permitted for water collection. After collection, respiration experiments ran days-months with daily oxygen monitoring. This sampling scheme was chosen to fit in with the operations of the research cruise that was more geared towards submersible exploration.
Data exclusions	The only data excluded are from the waters overlying the continental shelf of the eastern United States. Exclusion criteria was not pre-established but due to low cyanobacteria abundance, coelutions making chemical quantification inaccurate and the absence of using a mesh when collecting water as was used with all other stations (this allowed visible zooplankton to enter the bottles) we chose to exclude this data. Furthermore, this region did not target our site of study (the oligotrophic ocean) and was out of scope of the project, this is further discussed in the supplemental note.
Reproducibility	Reproduction of experiments at the same station was not possible due to time constraints, space on-board and resources.
Randomization	Sequence of filer extraction for pentadecane quantification was randomized by batch. Oxygen data measurements were randomized by bottle number. DNA extraction was randomized by batch.
Blinding	Blinding was not possible for our study because the people performing data acquisition were also those involved in collection and experimental design.

Did the study involve field work? ☒ Yes ☐ No

Field work, collection and transport

Field conditions	The environmental conditions that were relevant to the study pertain primarily to the conditions ideal for the cyanobacteria Prochlorococcus and Synechococcus, this includes temperature and dissolved nutrients. Temperature in the euphotic zone (0-200 m) ranged from 24-20 degrees Celsius, with some anomalous stations (station 9) going down to ~15 degrees Celsius. Nutrients were extremely low for all stations except for anomalous stations like station 9 where a Synechococcus bloom was encountered.

Location	Northwest Atlantic Ocean (Lat = 25-40 N, Long = 60-75 W), primarily in the euphotic zone (0-200 m) with some sampling at 500 m depth of the Subtropical Gyre/ Sargasso Sea. Gulf of Mexico (Lat = 25-30 N, 85-95 W) primarily at 1000 m depth.
Access & import/export	Open ocean habitats were accessed by research vessel, water was collected with a CTD rosette and no permits were required. Sample import to land was done via frozen filters or bottled ocean water.
Disturbance	No disturbance was caused by this study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☐	☒ Flow cytometry
☒	☐ MRI-based neuroimaging

Flow Cytometry

Plots

Confirm that:

☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☐ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☐ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	One mL of seawater was fixed to 0.5% paraformaldehyde and frozen in liquid nitrogen for long-term storage.
Instrument	Influx cytometer at Bigelow Laboratory for Ocean Sciences
Software	FlowJo version 9.8
Cell population abundance	Prochlorococcus was typically abundant an order of magnitude more than Synechococcus, with Prochlorococcus coming in at ~10^5 cells per mL
Gating strategy	Flow cytometry was conducted at the Bigelow Lab for Ocean Science using a standard approach published previously by Lomas et al and described in the methods sections. Unlike other approaches to flow cytometry, native fouresence was used to differentiate the two major taxa of cyanobacteria, based on the occurrence of phycoerythrin in Syn. Pico-autotrophs were identified as either Synechococcus or Prochlorococcus based upon cell size and the presence or absence of phycoerythrin, respectively. Based upon these gating criteria, the number of cells in each identified population was enumerated and converted to cell abundances by the volume-analyzed method (Sieracki et al., 1993).

☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.