nature microbiology

ARTICLES

https://doi.org/10.1038/s41564-020-00859-8

Check for updates

Microbial production and consumption of hydrocarbons in the global ocean

Connor R. Love[1,4], Eleanor C. Arrington[1,4], Kelsey M. Gosselin[1], Christopher M. Reddy[2], Benjamin A. S. Van Mooy[2], Robert K. Nelson[2] and David L. Valentine[3]✉

Seeps, spills and other oil pollution introduce hydrocarbons into the ocean. Marine cyanobacteria also produce hydrocarbons from fatty acids, but little is known about the size and turnover of this cyanobacterial hydrocarbon cycle. We report that cyanobacteria in an oligotrophic gyre mainly produce *n*-pentadecane and that microbial hydrocarbon production exhibits stratification and diel cycling in the sunlit surface ocean. Using chemical and isotopic tracing we find that pentadecane production mainly occurs in the lower euphotic zone. Using a multifaceted approach, we estimate that the global flux of cyanobacteria-produced pentadecane exceeds total oil input in the ocean by 100- to 500-fold. We show that rapid pentadecane consumption sustains a population of pentadecane-degrading bacteria, and possibly archaea. Our findings characterize a microbial hydrocarbon cycle in the open ocean that dwarfs oil input. We hypothesize that cyanobacterial hydrocarbon production selectively primes the ocean's microbiome with long-chain alkanes whereas degradation of other petroleum hydrocarbons is controlled by factors including proximity to petroleum seepage.

Hydrocarbons are released into the ocean via natural oil seeps and industrial spills associated with extraction, transportation and consumption of oil, totalling ~1.3 Tg per year[1]. Photosynthetic production also contributes hydrocarbons (C_{15}–C_{19} alkanes and alkenes) to the ocean[2–5], with a hypothetical contribution that exceeds petroleum by two orders of magnitude, based on scaling of a laboratory cultivation study[6]. Nonetheless, the biogeochemical cycle of oceanic hydrocarbons has not been directly observed or closed and the ecological ramifications of this input are scarcely considered beyond an untested hypothesis that biohydrocarbons prime the oceans for consumption of petroleum[6].

Our efforts focus on the North Atlantic subtropical oligotrophic gyre for which productivity is dominated by hydrocarbon-producing cyanobacteria *Prochlorococcus* and *Synechococcus*[7], genera estimated to account for ~25% of the global ocean's net primary production[8,9]. Subtropical oligotrophic gyres comprise ~40% of the planet's surface[10,11], tend to host predominantly cyanobacterial productivity[9] (Extended Data Fig. 10) and are far from the continents and associated petroleum sources that could mask the signal of cyanobacterial hydrocarbons. Here we target the primary production of hydrocarbons by cyanobacteria in oligotrophic settings and the associated consumption by hydrocarbon-oxidizing microbes to establish the spatial context, flux and controls on the cycle. We also explore the ocean's capacity to consume petroleum-derived hydrocarbons, incorporating biogeography to assess degradation capacity across a gradient from open ocean to active oil seepage.

Pentadecane is abundant and vertically structured in the oligotrophic ocean

To investigate the abundance pattern of cyanobacterial alkanes we quantified their depth distribution at seven locations in the western North Atlantic, five of which represent oligotrophic conditions and two that were more nutrient replete (Stations 3 and 9, Fig. 1). In total, we quantified alkane concentration in 441 particulate samples (≥0.2 µm), mainly in triplicate (Methods, Supplementary Table 4 and Extended Data Fig. 1). Pentadecane (nC_{15}) was the most abundant hydrocarbon in each sample from the five stations located in oligotrophic waters (Fig. 1). Concentrations of pentadecane ranged from 2–65 ng l^{-1} in the subtropical gyre, with maximum values of ~80 ng l^{-1} for the Gulf Stream (station 3) and ~130 ng l^{-1} for a *Synechococcus* bloom (station 9). Heptadecane (nC_{17}) was found at concentrations up to 12 ng l^{-1} but was often near our detection limit of ~2 ng l^{-1}; additionally, heptadecane was always lower in abundance than pentadecane in waters off of the continental shelf. No other hydrocarbons of measurable concentration were found in these samples.

Depth profiles of pentadecane concentration in oligotrophic waters reveal a distinctive subsurface maximum that coincides with both fluorescence and cyanobacteria cell counts (Fig. 1), aligning with the deep chlorophyll maximum (DCM). Concentrations above the DCM at the surface are lower but detectable (10–15 ng l^{-1} in *Prochlorococcus*-dominated waters), while they become undetectable below the DCM (Fig. 1) near the base of the euphotic zone (150–200 m). The observed coupling of pentadecane concentration with cell abundance is consistent with pentadecane occurrence primarily within cyanobacterial cells[12] (>98%), a finding further supported by observations of diel cycling (Fig. 2c,d and Extended Data Fig. 4) and cultivation work (Methods). Heptadecane shows no coherent spatial patterns or relationships with other variables likely due to the inability of our analytical procedure to measure concentrations <2 ng l^{-1} with suitable precision.

The geographic and vertical distribution of pentadecane is consistent with the ecology of *Prochlorococcus* and *Synechococcus*. The subsurface pentadecane maximum exhibits a decrease in magnitude and a deepening from ~50 m in the Gulf Stream to ~100 m at the most southerly station in the North Atlantic subtropical gyre, which is reflective of *Prochlorococcus* and *Synechococcus* cell abundance

[1]Interdepartmental Graduate Program for Marine Science, University of California—Santa Barbara, Santa Barbara, CA, USA. [2]Department of Marine Chemistry and Geochemistry, Woods Hole Oceanographic Institution, Woods Hole, MA, USA. [3]Department of Earth Science and Marine Science Institute, University of California—Santa Barbara, Santa Barbara, CA, USA. [4]These authors contributed equally: Connor R. Love, Eleanor C. Arrington. ✉e-mail: valentine@ucsb.edu

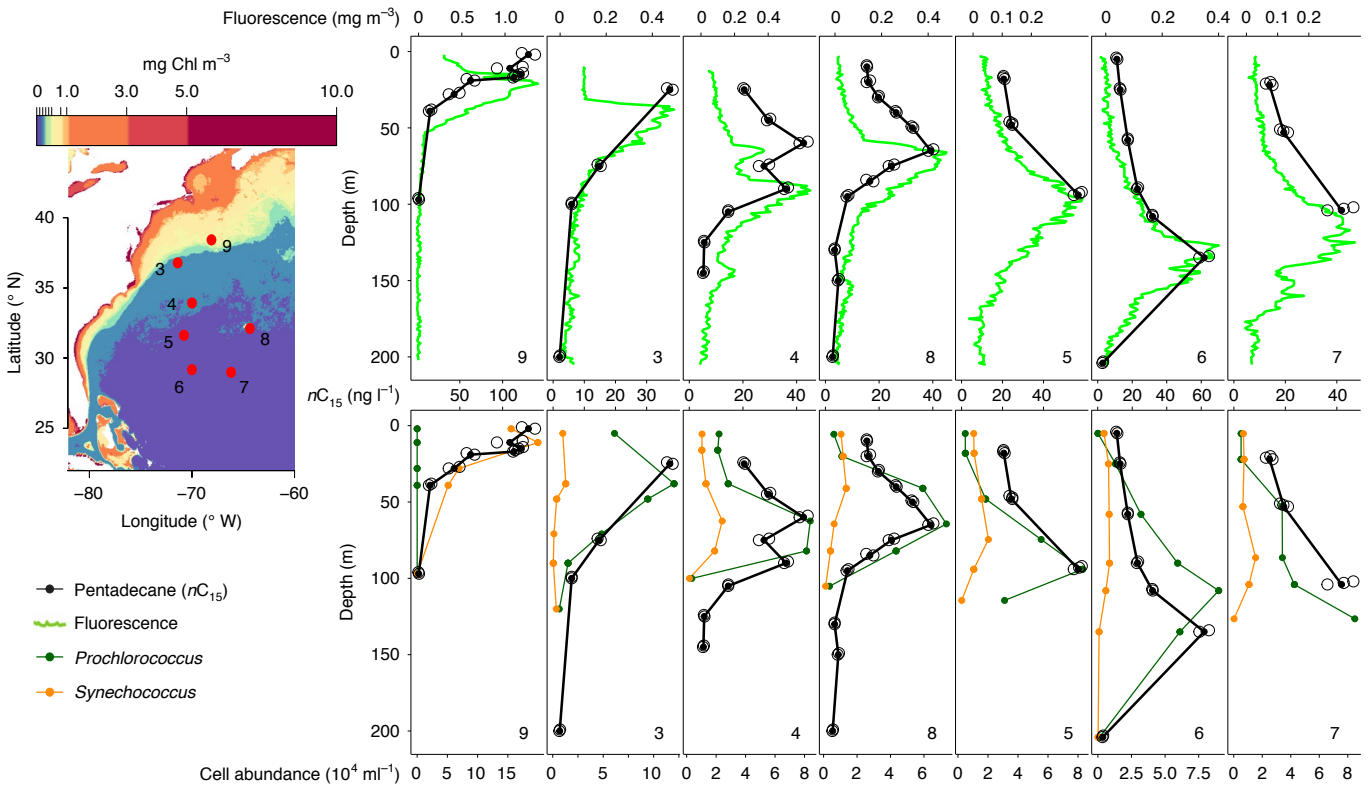


Fig. 1 | Pentadecane maps onto trends in ocean fluorescence and cyanobacteria abundance. Study area (at left) shows station coordinates mapped onto 4-km resolution MODIS-Aqua satellite chlorophyll (Chl) concentration for 2017. Station 3 was located in the Gulf Stream and station 9 targeted a *Synechococcus* bloom; all other stations captured more 'typical' *Prochlorococcus*-dominated oligotrophic water. Pentadecane depth distributions for each station are displayed with fluorescence (top row) and cyanobacterial abundance (bottom row). Depth distributions are organized by descending latitude with pentadecane distribution and station number duplicated for ease of comparison. Open black circles show biologically independent pentadecane measurements; each data point represents the contents of one distinct sample bottle (Methods). Replicates are sequentially moved 1 m below the other for visualization (water was taken from same depth, depth of top replicate); solid black circles indicate mean of $n=2$ at stations 9, 4, 8 and 6 and represent mean of $n=3$ for stations 3, 5 and 7 (Supplementary Table 4).

distributions[13] (Fig. 1). Pentadecane was slightly decoupled from cyanobacteria cell abundance at stations 6 and 7 (Fig. 1), possibly due to differential cell-specific hydrocarbon content for *Prochlorococcus* ecotypes at different parts of the photic zone[6,14].

Rapid pentadecane production in the lower euphotic zone

To quantify production patterns of cyanobacterial alkanes, we amended shipboard incubations with ^{13}C-enriched dissolved inorganic carbon (DIC) to 480‰ and quantified changes in hydrocarbon concentration (Extended Data Figs. 2 and 3) and ^{13}C enrichment of pentadecane. Incubations were conducted shipboard at ambient temperature and light level (Methods). In total, we quantified alkane production in 31 samples, from five stations, mainly in triplicate (Supplementary Table 4). Pentadecane production varies between ~3 and 30 ng nC_{15} l^{-1} d^{-1} within oligotrophic gyre waters (Fig. 2a), and has a higher maximum (~50 ng nC_{15} l^{-1} d^{-1}) in the Gulf Stream at the DCM (Supplementary Note). For each of the (four) oligotrophic stations tested (stations 4, 5, 7 and 8), volumetric pentadecane production is greatest near the DCM, where approximately 1% of photosynthetically active radiation (PAR) penetrates (Fig. 2a). Three of these stations (stations 4, 5 and 8) exhibit pentadecane production of 5–8 ng nC_{15} l^{-1} d^{-1} at 30% PAR depths, increasing with depth to ~30 ng nC_{15} l^{-1} d^{-1} at 1% PAR. (Fig. 2a). Diel variability in pentadecane concentration is also greatest at the DCM and 1% PAR, further consistent with hotspot production there (Fig. 2c and Extended Data Fig. 4).

By normalizing volumetric pentadecane production to cyanobacteria abundance (*Pro.* + *Syn.*), we find that 1% PAR has a higher average cellular production rate of pentadecane (0.37 ± 0.13 fg per cell per day) compared with 30% and 10% PAR (0.26 ± 0.05 and 0.13 ± 0.05 fg per cell per day, respectively) (Fig. 2b), indicating that cyanobacteria at or near the DCM produce more pentadecane per cell per unit of time. Furthermore, steady-state pentadecane replenishment time (production rate divided by concentration), calculated from ^{13}C incorporation and pentadecane concentration, is approximately twice as rapid at 1% PAR compared with 10% and 30% PAR (Fig. 2f). It is notable that we consistently observed greater production of pentadecane in the lower photic zone (1% PAR, near the DCM) than the upper photic zone (30% PAR) because depth profiles of primary production in oligotrophic gyres typically have greater production closer to the surface[15,16]. The reason underlying this productivity inversion is unclear, but is potentially related to a role for hydrocarbons in low-light and cold adaption of cyanobacteria[12,17].

Our findings of increased cell-specific pentadecane production and variability in the lower euphotic zone for the North Atlantic subtropical gyre are informed by differences in per-cell pentadecane content (nC_{15}/(*Pro.* + *Syn.*)) and dissolved nitrite concentrations. Relative importance analysis for physicochemical parameters (Fig. 2 and source data) ammonium, nitrite, depth, light and cyanobacterial pentadecane content (stations 4, 5, 7 and 8) in determining cell-specific production rate of pentadecane revealed that per-cell

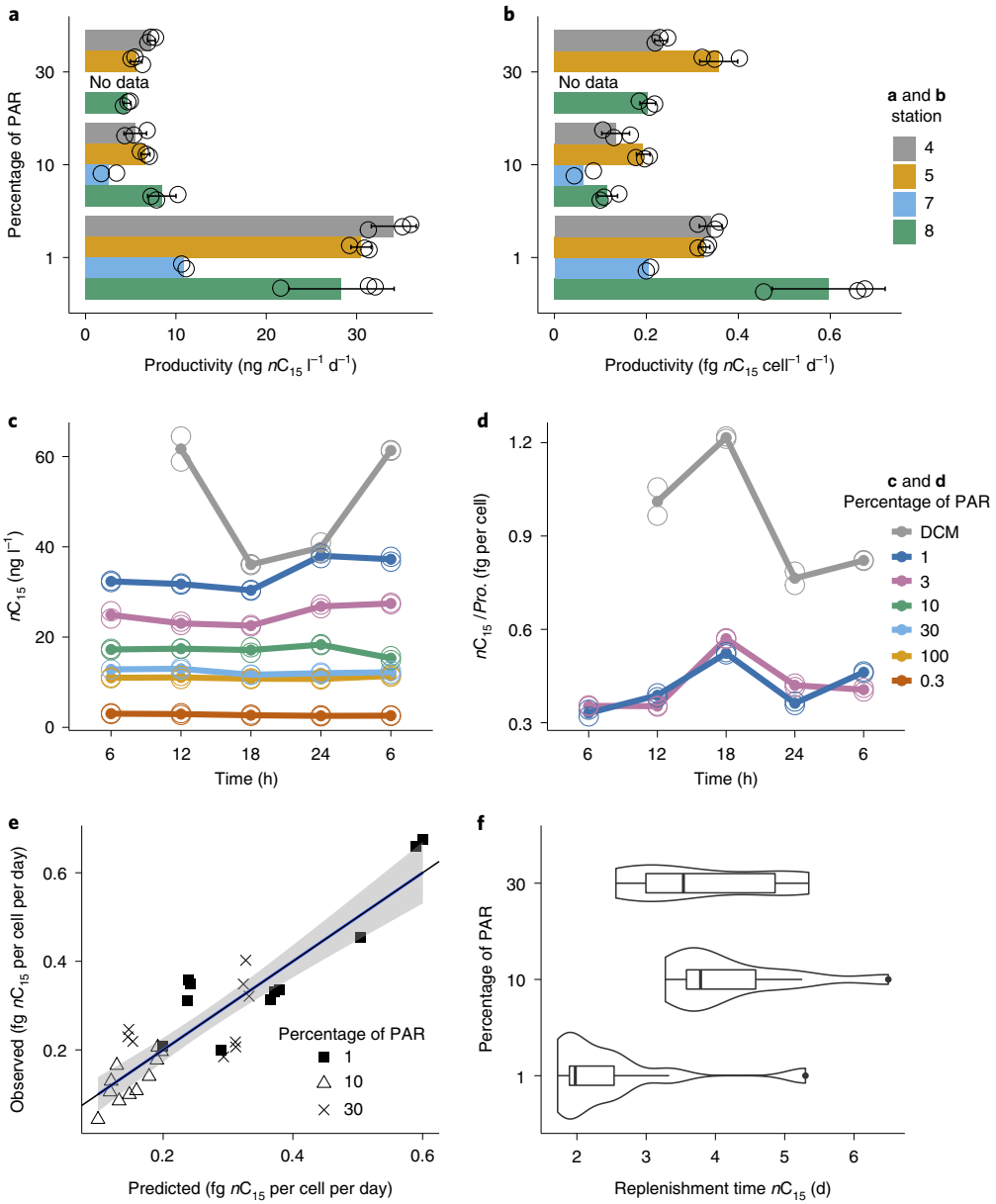


Fig. 2 | Most pentadecane production in lower euphotic zone. Pentadecane production and diel dynamics from ^{13}C-DIC enrichments and diel sampling grouped by light penetration depth. **a**, Volumetric pentadecane production separated by light penetration depth and station, calculated using pentadecane concentration and ^{13}C enrichment from incubation experiments (Methods). **b**, Cellular pentadecane production (cell = *Pro.* + *Syn.*) separated by light penetration depth and station, calculated by dividing volumetric production by cyanobacteria abundance. Both volumetric and cellular pentadecane production data are displayed as open black circles with bars representing mean production rate; error bars show standard deviation for $n = 3$. **c**, Diel changes in pentadecane concentration separated by light penetration depth and feature. DCM, deep chlorophyll maximum. **d**, Diel changes in pentadecane per *Prochlorococcus* cell separated by light penetration depth and feature. Diel plots show the lower euphotic zone and particularly the DCM is most dynamic (Extended Data Figs. 4 and 5); data are plotted as open circles, and means of replicates are plotted as solid circles ($n = 2$). **e**, Results of a multiple linear regression ($n = 31$) using nitrite and per-cell pentadecane content (nC_{15}/(*Pro.* + *Syn.*)) to predict cell-specific production (blue line); grey shadings indicate 95% confidence intervals; black line is 1:1. **f**, A density plot overlaid on a box and whisker plot of pentadecane replenishment time, grouped by light depth (centre line, median; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range; points, outliers); replication by light depth is as follows: 30 PAR ($n = 9$), 10 PAR ($n = 11$), 1 PAR ($n = 11$). For all panels, '*n*' describes the number of biologically independent pentadecane measurements.

pentadecane content and nitrite are the most powerful and only significant predictors at 33% and 34%, respectively (nC_{15}/(*Pro.* + *Syn.*): $P < 0.001$; nitrite: $P < 0.001$). In addition, ammonium, depth and light have 6%, 5% and 4% predictive power, respectively, for a total predictive power of 80% ($R^2 = 0.80$). A similar predictive capacity was found when the number of predictor variables was reduced to only per-cell pentadecane content and nitrite concentrations (Fig. 2e). Given a constant cell growth rate, the cell-specific production rate of pentadecane would be dependent on cell-specific pentadecane content, logically explaining its predictive power. The reason underlying nitrite's predictive power is less clear, but it is possible that low-light *Prochlorococcus* ecotypes can utilize nitrite more effectively than high-light ecotypes[18], driving production of pentadecane at the DCM via shoaling of the nitricline.

Global geochemical budget of pentadecane

Based on our measures of productivity and concentration, we sought to quantify key terms in the geochemical budget of cyanobacterial pentadecane—namely global standing stock (that is, reservoir magnitude) and global production of pentadecane produced by *Prochlorococcus* and *Synechococcus* (that is, turnover rate or input). Importantly, we assume consumption balances production (that is, steady state, with consumption discussed more in the following section) at the regional and global scale. We focus on pentadecane production by *Prochlorococcus* and *Synechococcus* because we found them to be the main drivers of the biological hydrocarbon cycle in the oligotrophic ocean (Extended Data Fig. 10). Two distinct approaches are applied for each budget term, low-end values based on pentadecane stock and production rates encountered in the study area (Figs. 1 and 2a) scaled by oligotrophic ocean area (method 1, representative of global oligotrophic ocean contribution), and higher values based on scaling of observed cellular properties (pentadecane content per cell) using a previous model[9] (method 2, representative of global cyanobacterial contribution, both outlined in the Methods). The water column integrated approach (method 1) is representative of the pentadecane stock in the oligotrophic gyres insomuch as the locations (North Atlantic subtropical gyre) are scalable; considering population estimates[9] and time-series data (Extended Data Fig. 10), we note that the Atlantic tends to have relatively low cyanobacterial abundance causing a potential low bias to method 1. We estimate the global standing stock of pentadecane to be 0.70 ± 0.17 Tg by method 1 and 1.78 ± 1.24 Tg by method 2, the latter of which is similar to an estimate based on laboratory cultivation[6]. We further estimate the global production rate of pentadecane to be 131 ± 13 Tg of pentadecane per year by method 1 and 274–649 Tg of pentadecane per year by method 2 (Supplementary Table 1). By comparison, the total quantity of petroleum estimated to reach the ocean annually from all sources is 1.3 Tg (ref. [1]), indicating that biohydrocarbon input to the ocean exceeds petroleum input by a factor of ~100–500. Interestingly, the global production rate of pentadecane by cyanobacteria is similar in magnitude to the atmospheric release for two other important hydrocarbons: methane[19] and isoprene[20,21].

To assess the reasonableness of our measurements and global scaling, we further check the replenishment time of pentadecane relative to known population turnover for wild *Prochlorococcus* and *Synechococcus*. Replenishment time of pentadecane was calculated from independent measures of water column integrated stock and production at three oligotrophic stations (Methods), yielding a value of 1.9 ± 0.5 d. This value is taken to represent the turnover time of cellular pentadecane and is within the range of cellular turnover time observed for environmental *Prochlorococcus* (1–2 d)—weighted slightly towards the slower environmental turnover of *Synechococcus* (1–6 d)[8,22–25]. Furthermore, since water column integrated turnover aligns with 1% PAR replenishment time (Fig. 2f), this further bolsters our finding that the low-light euphotic zone is driving most pentadecane flux, where elevated pentadecane concentrations and rapid turnover coincide.

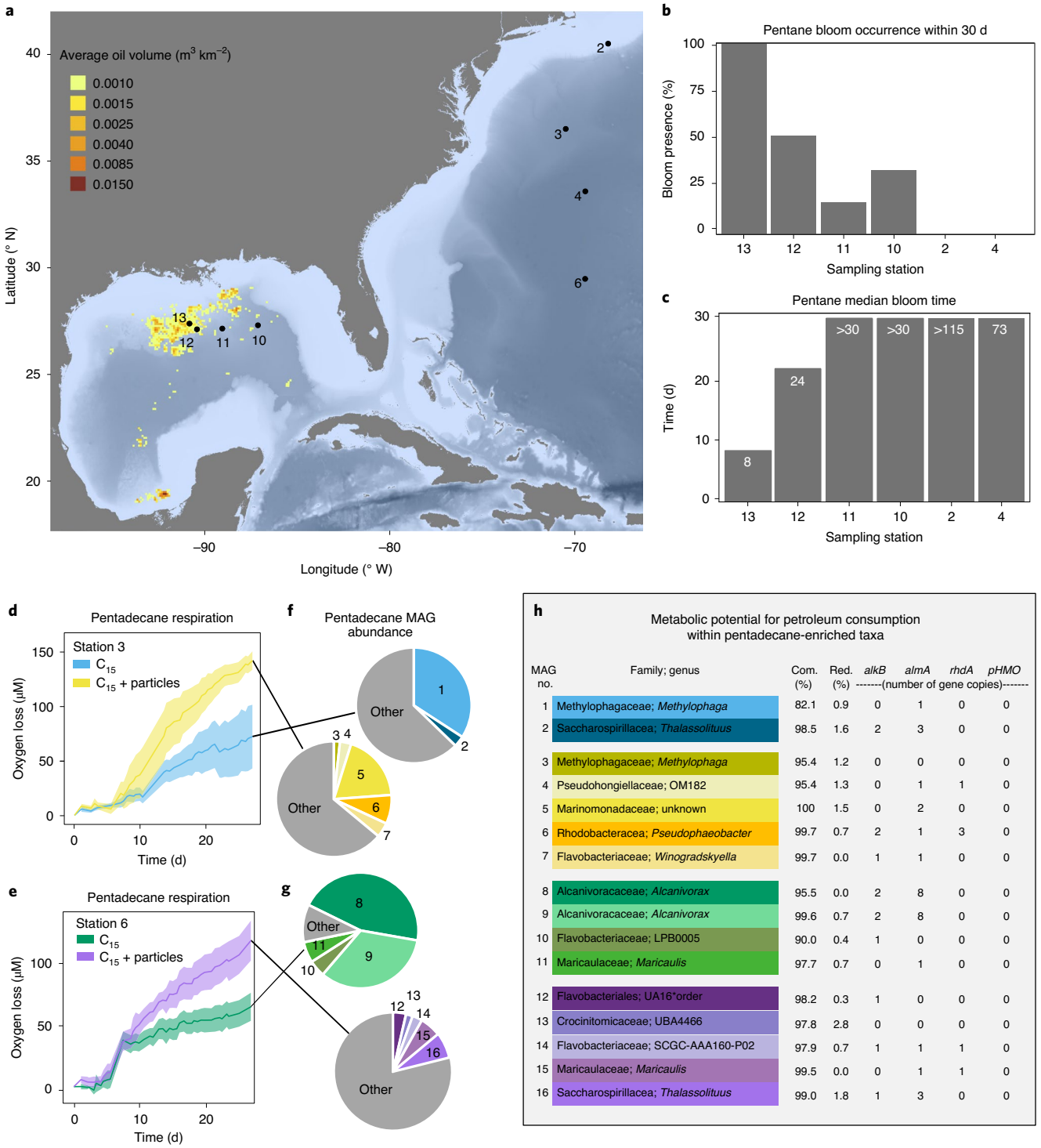
Biohydrocarbon consumption decoupled from petroleum

Our findings indicate a biological pentadecane cycle at steady state based on rapid production, consistent concentrations and the tight coupling to cyanobacterial physiology—spanning ≥40% of the Earth. Pentadecane and other long-chain *n*-alkanes can also be major components of spilled oil[26–28], and thus a priming effect has been proposed by which populations of petroleum degraders are sustained in a latent hydrocarbon cycle, enabling a rapid response to oil spills[6]. However, petroleum contains thousands of compounds in addition to *n*-alkanes[29–31], leading us to question the extent to which steady-state biohydrocarbon consumption primes the ocean with a microbial community capable of rapidly consuming this myriad of other compounds. Biodegradation was thus investigated to differentiate factors driving consumption of biological versus petroleum hydrocarbons.

Ocean-going experimentation revealed that waters in the mesopelagic underlying the North Atlantic subtropical gyre photic zone hosted *n*-alkane-degrading bacteria that bloomed rapidly when fed pentadecane, exhibiting exponential oxygen decline within ~1 week (Fig. 3). Parallel experiments performed with sinking particles collected in situ from beneath the DCM—representing an export flux of particulate-phase pentadecane and its bacterial consumers from the euphotic zone—exhibited similarly rapid bloom timing with pentadecane, but with greater oxygen decline. Despite similar timing of respiratory blooms on pentadecane with and without particles, each station and treatment displayed a distinctive bacterial response by a limited number of taxa (Fig. 3f–h and Extended Data Fig. 6). Blooms on pentadecane were dominated by *Alcanivorax* at station 6 and *Methylophaga* at station 3, whereas the addition of particles favoured *Thalassolituus* and an uncharacterized genus belonging to the family *Marinomonadaceae* at these respective stations (Fig. 3f–h).

Using metagenomics, we compiled genomes for the dominant organisms that bloomed on pentadecane, finding sequences for *alkB* (alkane-1-monooxygenase) and *almA* (flavin-binding monooxygenase) (Fig. 3h)—genes known to encode proteins that act on medium- (C_5–C_{11}) and long-chain *n*-alkanes (C_{12}–C_{30})—to be common among these taxa, with up to ten copies (*almA* + *alkB*) per genome (Extended Data Fig. 9 and Supplementary Table 2)[32–34]. Each recovered genome also encodes beta-oxidation functionality, essential for shunting alkane-derived carboxylic acids into central carbon metabolism. The genomes containing the greatest number of *almA* + *alkB* copies belong to the genera *Alcanivorax* and *Thalassolituus*, neither of which contain key genes for catabolism of aromatic (*rhdA*—ring-hydroxylating dioxygenases) or short-chain (*pHMO*—particulate hydrocarbon monooxygenase subunits A, B, C) alkanes, and both of which bloomed at the North Atlantic subtropical gyre stations. We interpret these genomes to indicate a specialization in long-chain *n*-alkanes (that is, biohydrocarbons) with an undefined upper limit on the carbon chain length. Testing for potential crossover catabolism to aromatic hydrocarbons using the approach of González-Gaya et al.[35], we also analysed the genomes for *rhdA*, finding copies in *Pseudophaobacter*, *Flavobacteria, Maricaulis* and *Pseudohongiellaceae* genomes. However, further analysis of protein hits for *rhdA* reveals that hits within our pentadecane-enriched taxa could originate from amino acid metabolism rather than aromatic hydrocarbon catabolism and points to a need to analyse *rhdA* hits in close detail before assuming hydrocarbon oxidation functionality (Supplementary Table 3). Despite ambiguity with the function of *rhdA*, observed blooms of *n*-alkane specialists underlying the oligotrophic ocean point to a decoupling between biohydrocarbons and dissimilar petroleum hydrocarbons such as aromatics.

To further probe alkane specialization among oligotrophic microbial populations, we analysed gene abundance from the Tara Oceans dataset[36]. The relative abundance of *alkB*-like genes in the upper oligotrophic ocean (relative to the single-copy gene, *recA*) was substantially greater than for genes involved in the activation of other hydrocarbons including C1-C5 alkanes, phenanthrene, benzene, toluene, naphthalene, xylene, cymene, and biphenyl (Supplementary Data 1), supporting a greater capacity for long-chain *n*-alkane degradation relative to other hydrocarbons. A detailed analysis of the Tara Oceans dataset in the North Atlantic reveals *alkB*-related genes are abundant in the surface ocean and DCM, and are phylogenetically distinct from the related delta-9 fatty acid desaturase (Extended Data Figs. 7 and 8). Notably, a dominant clade of *alkB*-like monooxygenases belongs to the globally abundant Marine Group II (MGII) archaea (with possible occurrence also in Marine Group III archaea) and is consistently present in all surface



h

Metabolic potential for petroleum consumption within pentadecane-enriched taxa

MAG no.	Family; genus	Com. (%)	Red. (%)	*alkB*	*almA*	*rhdA*	*pHMO*
				(number of gene copies)			
1	Methylophagaceae; *Methylophaga*	82.1	0.9	0	1	0	0
2	Saccharospirillacea; *Thalassolituus*	98.5	1.6	2	3	0	0
3	Methylophagaceae; *Methylophaga*	95.4	1.2	0	0	0	0
4	Pseudohongiellaceae; OM182	95.4	1.3	0	1	1	0
5	Marinomonadaceae; unknown	100	1.5	0	2	0	0
6	Rhodobacteracea; *Pseudophaeobacter*	99.7	0.7	2	1	3	0
7	Flavobacteriaceae; *Winogradskyella*	99.7	0.0	1	1	0	0
8	Alcanivoracaceae; *Alcanivorax*	95.5	0.0	2	8	0	0
9	Alcanivoracaceae; *Alcanivorax*	99.6	0.7	2	8	0	0
10	Flavobacteriaceae; LPB0005	90.0	0.4	1	0	0	0
11	Maricaulaceae; *Maricaulis*	97.7	0.7	0	1	0	0
12	Flavobacteriales; UA16*order	98.2	0.3	1	0	0	0
13	Crocinitomicaceae; UBA4466	97.8	2.8	0	0	0	0
14	Flavobacteriaceae; SCGC-AAA160-P02	97.9	0.7	1	1	1	0
15	Maricaulaceae; *Maricaulis*	99.5	0.0	0	1	1	0
16	Saccharospirillacea; *Thalassolituus*	99.0	1.8	1	3	0	0

Fig. 3 | Pentadecane rapidly consumed by specialists in waters underlying the oligotrophic ocean. Microbial respiratory blooms on pentane and pentadecane quantified via contactless optical oxygen sensors, followed by metagenomic analysis. **a**, Oceanographic sampling stations relative to natural petroleum seepage[42] with increasing distance from intense seepage as follows: 13, 12, 11, 10, 6, 4, 3 and 2. **b**, Occurrence of respiratory blooms on pentane (petroleum proxy compound) at 1,000 m with increasing distance from seepage ($n = 6$ at stations 10–13, $n = 8$ at stations 2 and 4). **c**, Timing of respiratory blooms on pentane with experiment duration 30 d at stations 10–13 and 115 d for stations 2 and 4. Numbers listed in white are median bloom times over the entire duration of the experiment, which varied from 30 d to 115 d, assuming all bottles bloom given enough time. **d**, Pentadecane respiration at station 3 (500 m) ± particles (solid line indicates mean; shading indicates ± standard deviation; $n = 6$ at station 3). **e**, Pentadecane respiration at station 6 (500 m) ± particles (solid line indicates mean; shading indicates ± standard deviation; $n = 4$ at station 6). **f**, Relative abundance of metagenomes (MAGs) at final time point of pentadecane incubations (~28 d) from station 3. **g**, Relative abundance of metagenomes at final time point of pentadecane incubations (~28 d) from station 6. **h**, Genome quality and metabolic potential for MAGs. Abbreviations include completion (Com); redundancy (Red); *alkB* (alkane-1-monooxygenase); *almA* (flavin-binding monooxygenase); *rhdA* (ring-hydroxylating dioxygenase subunit a); and *pHMO* (particulate hydrocarbon monooxygenase subunits A, B and C). *MAG 12 is unclassified at the family and genus level, and therefore we have listed the class and order. For panels **b**–**e**, '*n*' describes the number of biologically independent incubations.

and DCM stations (Extended Data Fig. 8 and Supplementary Data 2). This highly successful group of surface-ocean-dwelling archaea is known for a chemoorganoheterotrophic lifestyle targeting lipids, proteins and amino acids[37,38] and can utilize photoheterotrophy, but a key role in biohydrocarbon cycling would be unexpected as MGII archaea are not among the ~300 genera of bacteria and archaea[39,40] previously identified as hydrocarbon degraders.

We ask whether efficient turnover of pentadecane in the surface oligotrophic ocean is consistent with steady-state cycling by chemoorganoheterotrophic MGII archaea that occupy this environment. To do so, we quantify the contribution of pentadecane to export flux ($1.5 \pm 0.8 \times 10^{-4}$% of POC flux) from the euphotic zone at stations 2, 7 and 8, and compare the results with integrated production ($1.76\,mg\,nC_{15}\,m^{-2}\,d^{-1}$) from the same three stations, finding that only $\sim 1 \times 10^{-4}$% of pentadecane production was exported below 150 m. The coincidence of efficient pentadecane recycling, euphotic zone niche specialization for MGII archaea and the high prevalence of MGII archaeal *alkB*-like monooxygenases in oligotrophic surface waters collectively point to a potentially important role for MGII archaea in biohydrocarbon cycling. A capacity for MGII to consume *n*-alkanes, coupled with an inability to outcompete bacteria in a spill scenario, is consistent with existing theories of archaea's ecological specialization[41] and, if correct, provides a further example of decoupling from petroleum, given the lack of reported MGII in oil spill bloom response.

Natural seeps prime petroleum hydrocarbon consumption

Results from our investigation indicate that the biohydrocarbon cycle primes the ocean for consumption of long-chain *n*-alkanes, but that this effect is at least partially decoupled from the consumption of petroleum hydrocarbons by bacterial *n*-alkane specialization and a possible role for cosmopolitan MGII archaea. We therefore explore an alternative hypothesis, that priming for petroleum hydrocarbon degradation occurs by proximity to petroleum sources. The test of this biogeographic hypothesis is a bloom response experiment conducted at a single depth across seven stations representing a gradient of natural oil seep intensity, spanning from the seep-ridden northern Gulf of Mexico to the North Atlantic subtropical gyre[42] (Fig. 3a). Pentane was used for these experiments as it is a model nonbiogenic compound that is unique to and abundant in petroleum, a structural analogue to pentadecane (linear chain with an odd number of carbons), readily bioavailable based on its higher vapour pressure and aqueous solubility, relatively low in toxicity and known to partition to the ocean's interior following release from the seafloor[43,44]. We find the microbial response to *n*-pentane to be structured by proximity to seepage, with ~9× more rapid bloom onset in the Gulf of Mexico versus the North Atlantic subtropical gyre (Fig. 3c). Notably, the bloom onset for pentane underlying the North Atlantic subtropical gyre is ~10× slower than for pentadecane in the same region, albeit with experiments conducted at different depths and stations. Results demonstrate a clear biogeographic dependence on natural seepage for biodegradation of a petroleum hydrocarbon, providing another example of decoupling between petroleum versus biological hydrocarbon consumption, and pointing to source-specific priming by which the capacity for rapid consumption of a petroleum-derived hydrocarbon is defined by proximity to petroleum inputs.

Through our studies in the subtropical North Atlantic Ocean, we have confirmed the existence and magnitude of a cryptic hydrocarbon cycle as proposed by Lea-Smith et al.[6]. We further demonstrate a decoupling between biological alkanes and petroleum-derived hydrocarbons that points to a complex interplay of chemical composition and biogeography that structures the ocean's response to oil spills. Importantly, our findings are most applicable to the oligotrophic regions of the ocean, encompassing ~40% of Earth's surface. Other oceanic regions may harbour abundant Eukaryotic phytoplankton, many of which are capable of producing hydrocarbons[45–47]. Based on our qualitative observations from productive waters on the continental shelf of the Northwest Atlantic Ocean (Supplementary Note), we expect such environments to harbour a dynamic and complex hydrocarbon cycle including biological alkanes and alkenes, structured in part by proximity to continental sources and interaction with the seafloor. Cryptic hydrocarbon cycling and its relationship to biogeographic structuring of microbial populations represents an important factor in understanding the metabolic response capacity of the oceanic microbiome to oil inputs and should be incorporated as a predictive tool in oil spill response planning.

Methods

In situ sampling and quantification of hydrocarbon production. Water was collected with a rosette equipped with 12-l Niskin bottles just after sunrise (~8:00) for all sampling except for the diel experiment. Salinity, density, temperature, fluorescence and % PAR were measured semi-continuously for each hydrocast. For diel sampling, a Lagrangian framework was used by following deployed particle traps set just below the DCM (150 m) and sampled at 6-h intervals through a full 24-h cycle. Sampling targeted six light-penetration levels with depths held constant following initial collection, plus the DCM, which is a depth-variable feature. Water was collected from the Niskin into 2-l polycarbonate bottles via a polyvinyl chloride tube equipped with a 200-µm mesh to filter out large zooplankton. Precautions were taken to avoid contamination from the vessel and validated with controls. For example, all Niskin bottles were cleaned with a brush and MilliQ water before the cast and was moved into a secure bay for sampling. To avoid exhaust and fumes, the vessel was oriented into the wind during sampling and certain activities were disallowed during sampling (that is, smoking and painting). Control samples were collected by pouring clean MilliQ water into the Niskin bottles, waiting for 30 min and then filtering the water using the same procedure as for all samples. No pentadecane of considerable quantity ($>2\,ng\,l^{-1}$) was found in control samples, thus validating efforts to minimize contamination. As a secondary check, we also collected diesel from the vessel and extracted and ran the extract on the gas chromatograph. This diesel had a distinct multi-hydrocarbon fingerprint that we did not observe in any of our chromatograms. For in situ hydrocarbon concentration measurements, water in the 2-l polycarbonate bottles was immediately filtered through a 0.22-µm Teflon filter under gentle vacuum with an oil-less vacuum pump. Captured particles (sediment trap deployed for 24 h at 150 m) were also filtered onto 0.22-µm Teflon filters. For the hydrocarbon production experiment, ^{13}C-bicarbonate tracer solution (with $45\,g\,l^{-1}$ NaCl to sink the tracer to the bottom of the bottle) made from ^{13}C-sodium bicarbonate (Cambridge Isotope Laboratories, ^{13}C 99%) was added to the 2-l polycarbonate bottles to achieve a 480‰ enrichment in seawater DIC. Dark control bottles were covered completely beforehand with aluminium foil before tracer addition and kill control bottles were treated with zinc chloride to 2% $ZnCl_2$ (*m*/*v*) before tracer addition. Then, 2-l bottles were immediately placed into black mesh bags to attenuate light to the value from which it was collected (either 30%, 10% or 1% PAR) and placed into on-board seawater incubators with a continuous flow of surface water; this was marked as the start of incubation. No artificial light was used. Black mesh bags were made by stitching together rolls of commercial-grade neutral-density window screen material[48] and photosynthetically active radiation attenuation by the bags was quantified using a spherical quantum sensor (Licor). Bottles were harvested at 0-h (initial), 5-h, 10-h, 20-h and 30-h (final) time points for the 30% PAR light bags and at $t = 0$ h and $t = 30$ h (final) for the 10% and 1% light levels; care was taken to reduce light exposure in the shipboard laboratory when preparing for incubation by placing bottles into covered tubs. A 2-ml aliquot was taken for ^{13}C-DIC before filtration. Filters were placed into precombusted aluminium foil packets and immediately frozen at −20 °C for later analysis.

A preliminary culture experiment was conducted to assess the percentage of all cyanobacterial hydrocarbons within membranes; that is, what percentage of total cyanobacterial hydrocarbons our extraction protocol was capturing. We compared two types of extractions, the modified Bligh and Dyer used in this study (described below) to extract membrane lipids from cells filtered on a 0.22-µm Teflon filter and an extraction of frozen cell culture that includes cells and the culture medium. A comparison of these results provides the proportion of hydrocarbons found within cell membranes versus total hydrocarbons inclusive of those interior and exterior to cells. We conducted a triplicate measurement of this ratio from a culture of *Synechocystis*. Of the two hydrocarbons that *Synechocystis* makes in abundance (*n*-heptadecane and 8-heptadecene), we found that 98 ± 17% of total *n*-heptadecane and 82 ± 9% of 8-heptadecene were cell associated. We interpret this to mean that the majority of hydrocarbons, particularly saturated *n*-alkanes, reside within the biological membranes of cyanobacteria or adsorb to particulate matter including cellular necromass. This is further supported by work done by Lea-Smith et al.[6,12], and the low solubility of straight-chain hydrocarbons 15–17 carbons in length.

Hydrocarbon extraction and analysis. A modified Bligh–Dyer[49] method was used to extract hydrocarbons from membranes of frozen cells collected on Teflon filters. Dodecahydrotriphenylene (internal standard) and C23 ethyl ester (secondary internal standard and transesterification standard if needed) were added to the dry filter before extraction. Two-thirds of the amount of each solvent was used according to Van Mooy et al.[49] and a 10-min sonication step was added after addition of the first solvents. An additional extraction into 1.0 ml of DCM was conducted after the first lower organic phase was removed to extract any remaining hydrocarbons from the filter; this was added to the first DCM extract for a final extract volume of 3.0 ml of DCM. Once extracted into dichloromethane, sodium sulfate was added for drying, ~40 µl of toluene was added to prevent complete dryness of the extracts and then the solution was rotary evaporated to ~30 µl and placed into a 2-ml gas chromatography (GC) vial with a combusted glass insert. Before analysis, a small volume of C23 methyl ester (external standard) was added. All glassware and solid chemicals were precombusted before use. Concentration analysis was done on a gas chromatograph flame ionization detector (GC-FID), HP-Agilent 6890 GC-FID. Chromatography was performed with a 30 m × 0.25 mm internal diameter (ID), 0.25 µm pore size, fused silica Restek 13323 Rxi-1 MS capillary column with a splitless 2-µl injection. Initial oven temperature was set at 70 °C and held for 2 min, followed by a 3 °C min^{-1} ramp to 120 °C, then a 6 °C min^{-1} ramp to the final temperature of 320 °C. A standard mix of pentadecane, heptadecane, internal standard, external standard and transesterification standard was run to calibrate response factors for every batch of samples (~20 per batch). Blanks were run approximately every six samples and peaks were manually integrated; there were no co-eluting peaks for pentadecane or heptadecane in oligotrophic samples (all stations but station 1 on continental shelf). Comprehensive two-dimensional chromatography, GC × GC-FID and GC × GC-TOF (where TOF represents time of flight), was used on select samples to check for other hydrocarbons, contaminants and quality of blank filters run through the extractive process.

GC × GC-FID and -TOF chromatographic analyses were performed on Leco systems consisting of an Agilent 7890A GC configured with a split/splitless auto-injector (7683B series) and a dual-stage cryogenic modulator (Leco). Samples were injected in splitless mode. The cold jet gas was dry N_2 chilled with liquid N_2. The hot jet temperature offset was 15 °C above the temperature of the main GC oven and the inlet temperature was isothermal at 310 °C. Two capillary GC columns were utilized in this GC × GC experiment. The first-dimension column was a Restek Rxi-1ms (60 m length, 0.25 mm ID, 0.25 µm d_f), and second-dimension separations were performed on a 50% phenyl polysilphenylene-siloxane column (SGE BPX50, 1.2 m length, 0.10 mm ID, 0.1 µm d_f). The temperature programme of the main oven was held isothermal at 50 °C (15 min) and was then ramped from 50 °C to 335 °C at a rate of 1.5 °C min^{-1}. The second-dimension oven was isothermal at 60 °C (15 min) and was then ramped from 60 °C to 345 °C at a rate of 1.5 °C min^{-1}. The hot jet pulse width was 0.75 s, while the modulation period between stages was 7.50 s with a 3.00-s cooling period for the FID method, and 10.00 s with a 4.25-s cooling period for the TOF method. FID data were sampled at an acquisition rate of 100 data points per second, while the TOF data were sampled at an acquisition rate of 50 spectra per second in the mass range of 40–500 atomic mass units. Different modulation periods were used due to differences in the GC × GC instrument; for example, the GC × GC-FID combusts the column effluent at atmospheric pressure, while in the GC × GC-TOF instrument, column effluent has to move through a heated transfer line into the ion source. Since the total distance between detector and secondary oven is different between these two instruments, optimization of the chromatographic plane requires slight modifications to the GC × GC methods.

Compound-specific and DIC isotope measurements. Compound-specific isotope analysis was performed after concentration analysis on a gas chromatograph combustion isotope ratio mass spectrometer with a Trace GC (Thermo Finnigan) set up to a GC-C/TC III (Finnigan) interface and a Delta^{plus} XP isotope ratio mass spectrometer (Thermo Finnigan). A J&W Scientific DB-5 Capillary column (30 m, 0.25 mm, 0.25 µm) was used with 2-µl manual injections. A temperature ramp was conducted, starting at 70 °C and held for 2 min, followed by a 3 °C min^{-1} ramp to 120 °C, held for 0 min, then a 6 °C min^{-1} ramp to 185 °C, held for 0 min, then a 120 °C min^{-1} ramp to 290 °C, held for 3 min. Inlet temperature was 260 °C; flow rate was held at 2.2 ml He min^{-1} with a splitless injection held for 0.5 min after injection. Isotope ratio accuracy was calibrated with a C_{14} fatty acid methyl ester Schimmelmann reference material to Vienna PeeDee Belemnite. Precision was accounted for with a standard mix of nC_{15}, nC_{16} and nC_{17} at ~1.2 ng µl $^{-1}$, run between every batch of ~20 samples. Peaks were manually integrated after establishing the baseline; analytical precision was ~0.9‰ $\delta^{13}C$ for pentadecane.

DIC ^{13}C isotope ratio measurements were made on a Gas Bench II (Thermo Finnigan) interfaced to the same Delta^{plus} XP isotope ratio mass spectrometer (Thermo Finnigan) used for the compound-specific analysis. Sample preparation and analysis closely followed the protocol outlines by the University of California, Davis, Stable Isotope Facility (https://stableisotopefacility.ucdavis.edu/dictracegas.html).

Respiration experiment and analysis. Pentadecane respiration incubations were conducted at station 3 (36° 50.93′ N, 71° 23.94′ W) and station 6 (29° 4.79′ N, 69° 44.38′ W) with water collected from 500 m. Pentane respiration incubations were conducted at stations 2 (40° 9.14′ N, 68° 19.889′ W), 4 (33° 58.21′ N, 69° 43.38′ W), 10 (27° 30.41′ N, 87° 12.41′ W), 11 (27° 15.00′ N, 89° 05.05′ W), 12 (27° 11.60′ N, 90° 41.75′ W) and 13 (27° 38.40′ N, 90° 54.98′ W) with water collected from 1,000 m. Water samples from the CTD Niskin bottles were transferred to 250-ml glass serum vials using a small length of Tygon tubing. Vials were filled with at least three volumes of water to overflow. Care was taken to ensure no bubbles were present before sealing with a Teflon-coated rubber stopper and crimp cap. Abiotic controls were amended with 14 g of zinc chloride before sealing. All bottles except for unamended blank controls immediately received 10 µl of pentadecane or pentane using a gas-tight syringe and were maintained in the dark at in situ temperature (15 °C for pentadecane, 4 °C for pentane). Sediment traps at stations 3 and 6 were deployed for 24 h at 150 m. For each particle addition, 10 ml of particle-laden seawater was vortexed lightly for 1 min, then 2 ml of the vortexed seawater was added to the bottom of each serum bottle with a pipet via displacement. Each serum bottle was fixed with a contactless optical oxygen sensor (OXSP5, Pyroscience) on the inner side with silicone glue and oxygen content was monitored approximately every 12 h with a fibre optic oxygen meter (FireStingO2, Pyroscience). Observed changes in oxygen content were normalized to abiotic controls and to unamended seawater to correct for variability due to temperature and background respiration not caused by alkane addition. In the case of the pentadecane particle incubations, oxygen losses from particles and seawater were subtracted from particle plus pentadecane treatments to isolate pentadecane respiration. Bloom onset is operationally defined as three consecutive time points with oxygen loss >0.21 µM h^{-1}. At the end of each respiration experiment incubations were sacrificially harvested and filtered on a 0.22-µm polyethersulfone filter. DNA extraction was performed from ¼ of each filter using the PowerSoil DNA Extraction kit (Qiagen) with modifications to standard protocol (described below).

Cell counts and dissolved nutrient analysis. Sampling for nutrients and cell counts was conducted on the CTD cast immediately before the casts for hydrocarbon sampling (~1-h difference); these casts were all at approximately sunrise. Parallel sampling was conducted with the same cast water for the diel sampling. Flow cytometry analysis was performed by the Bigelow Laboratory for Ocean Sciences using a slightly modified protocol from Lomas et al.[50]. Samples were fixed with paraformaldehyde (0.5% final concentration) and stored at ~4 °C for 1–2 h before long-term storage in liquid nitrogen. An Influx cytometer was used with a 488-nm blue excitation laser and appropriate chlorophyll *a* (692 ± 20 nm) and phycoerythrin (585 ± 15 nm) bandpass filters, and was calibrated daily with 3.46-µm Rainbow Beads (Spherotech). Each sample was run for 4–6 min (~0.2–0.3 ml total volume analysed), with log-amplified chlorophyll *a* and phycoerythrin fluorescence, and forward and right-angle scatter signals were recorded. Data files were analysed from two-dimensional scatter plots based on red or orange fluorescence and characteristic light-scattering properties[51] using FlowJo 9.8 Software (Becton Dickinson). Pico-autotrophs were identified as either *Synechococcus* or *Prochlorococcus*, pico-eukaryotes or nano-eukaryotes, based upon cell size and the presence or absence of phycoerythrin. Nutrients were analysed by the University of Washington Marine Chemistry Laboratory.

Calculations and analyses, statistics and reproducibility. All statistics and points within figures were conducted with distinct samples (not replicated measurements of the same sample). Pentadecane production from compound-specific isotope enrichment measurements was calculated using a published equation[52]. The time duration used in the equation was from complete set up of the incubation to completion of filtering the water through the filter. The value used for ^{13}C-DIC was the average of the whole dataset ($\delta^{13}C = 480‰$) and the value used for unlabelled pentadecane was from a nonenriched sample ($\delta^{13}C = -20‰$) because of variations in the time-zero values from a slight but inevitable enrichment when bottles were filtered in the laboratory (roughly 1 h to filter the whole bottle in a well-illuminated laboratory space).

Statistical analyses were conducted using R within RStudio v.1.2.1335. Statistical analyses of single linear models were done using the R base stats package. Relative importance of regressors in multiple linear models was found using the R package 'relaimpo' and the function 'calc.relimp()'. Reproduction of experiments at the same station was not possible due to time constraints, space on-board and resources.

Quantification of global stock and production for cyanobacterial alkanes: method 1. Method 1 draws from direct observations of water column pentadecane stock and production rates encountered in the North Atlantic subtropical gyre. We integrated the depth profiles of pentadecane concentration for stations 4, 6 and 8 to calculate a mean water column integrated stock of pentadecane with standard deviation and further integrated primary production rates of pentadecane for stations 4, 5 and 8 from our isotope enrichment incubation experiments, to obtain a mean water column production rate with standard deviation. Calculation of pentadecane stock results in an average water column integrated stock of pentadecane of 3.42 ± 0.83 mg m^{-2}, and when scaled by the mean areal extent of the oligotrophic ocean (estimated at 204×10^6 km^2) results in a standing stock of

0.70 ± 0.17 Tg (Supplementary Table 1). Calculation of pentadecane production rate results in 1.76 ± 0.17 mg pentadecane $m^{-2} d^{-1}$, which multiplied by the areal extent of the oligotrophic ocean yields 131 ± 13 Tg of pentadecane per year (Supplementary Table 1).

To integrate pentadecane stock in the water column, we integrated station 4, 6 and 8 depth profiles because of suitable data coverage. Integration was performed by taking a data point to be the centre of a rectangle, with the ends of rectangles meeting halfway between data points on the depth axis. For the data closest to the surface we assume that the stock stays at that value from the depth of collection to the surface. If the deepest data are shallower than 200 m (station 4), we assume that the pentadecane concentration attenuates to 0 ng l^{-1} at 200 m depth, and thus integrated the area from the deepest rectangle to 200 m as a triangle. If the deepest data go to 200 m or deeper (stations 6 and 8), we integrated the height of the deepest rectangle as the value of the data found beyond 200 m, and chose these data to be the deepest endmember of our integration.

To integrate pentadecane production rate throughout the water column, we used 'typical' oligotrophic stations that had production measurements at 30%, 10% and 1% PAR (stations 4, 5 and 8). All three stations had a very similar trend in productivity (Fig. 2a). We integrated by taking the data to be the height (pentadecane productivity) of the rectangle and the width of the rectangle (depth) to be the depth halfway between data points. Integration to the surface was done by assuming that productivity remained the same from the shallowest data point to the surface. For the deep endmember we chose to retain the distance between the middle (10%) and deepest (1%) data points and carry the rectangle this same distance below the 1% PAR data point depth.

Quantification of global stock and production for cyanobacterial alkanes: method 2. Method 2 draws from all samples with co-occurring measured pentadecane concentrations as well as *Prochlorococcus* and *Synechococcus* abundance ($n = 67$) to establish average per-cell quantities of pentadecane across all our stations. We then used previously modelled global populations of *Prochlorococcus* (2.9 ± 0.1 × 1,027) and *Synechococcus* (7.0 ± 0.3 × 1,026)[9] to scale our estimates for a global stock and utilized known doubling rates (1–2 d for *Prochlorococcus*, 1–6 d for *Synechococcus*)[8,22–25] to scale the average per-cell pentadecane content from our data to estimate a global production rate.

To differentiate the pentadecane contributions from each genus in our data, we created a multiple linear model using *Prochlorococcus* and *Synechococcus* cell counts as separate independent variables, yielding values of 0.47 ± 0.42 fg per cell for *Prochlorococcus* and 0.60 ± 0.35 fg per cell for *Synechococcus* ($R^2 = 0.768$). These values are similar to those from pure cultures of three ecotypes of *Prochlorococcus* (0.49 ± 0.23 fg per cell) and are slightly higher than reported for four strains of *Synechococcus* (0.25 ± 0.04 fg per cell), also from culture[6]. From this approach we estimate the global standing stock of pentadecane from *Prochlorococcus* to be 1.4 ± 1.2 Tg and from *Synechococcus* to be 0.42 ± 0.25 Tg, for a total of 1.78 ± 1.24 Tg. See Supplementary Table 1 for estimates and comparisons with Lea-Smith et al.[6].

DNA extraction. The PowerSoil DNA Extraction (Qiagen) was used according to manufacturer recommendations with the following modifications: 200 µl of bead beating solution was removed in the initial step and replaced with phenol chloroform isoamyl alcohol, the C4 bead binding solution was supplemented with 600 µl of 100% ethanol and we added an additional column-washing step with 650 µl of 100% ethanol. Extracts were purified and concentrated with ethanol precipitation.

16S rRNA gene amplification and analysis. We amplified and barcoded the V4 region of the 16S rRNA gene using the method described previously[53] with small modifications to the 16Sf and 16Sr primers[54]. Amplicon PCR reactions contained 1 µl of template DNA, 2 µl of forward primer, 2 µl of reverse primer and 17 µl of AccuPrime Pfx SuperMix. Thermocycling conditions consisted of 95°C for 2 min; 30 cycles of 95 °C for 20 s, 55 °C for 15 s and 72 °C for 5 min; and a final elongation at 72 °C for 10 min. Sample DNA concentrations were normalized using the SequelPrep Normalization Kit, cleaned using the DNA Clean and Concentrator kit, visualized on an Agilent Tapestation and quantified using a Qubit fluorometer. Samples were sequenced and demultiplexed at the University of California Davis Genome Center on the Illumina MiSeq platform with 250-nucleotide paired-end reads. A PCR-grade water sample was included in extraction, amplification and sequencing as a negative control to assess for DNA contamination.

Trimmed fastq files were quality-filtered using the fastqPairedFilter command within the dada2 R package, v.1.9.3 (ref. [55]), with the following parameters: truncLen = c(190,190), maxN = 0, maxEE = c(2,2), truncQ = 2, rm.phix = TRUE, compress = TRUE, multithread = TRUE. Quality-filtered reads were dereplicated using derepFastq command. Paired dereplicated fastq files were joined using the mergePairs function with the default parameters. A single-nucleotide variant table was constructed with the makeSequenceTable command and potential chimeras were removed de novo using removeBimeraDenovo. Taxonomic assignment of the sequences was done with the assignTaxonomy command using the Silva taxonomic training dataset formatted for DADA2 v.132. If sequences were not assigned, they were left as not applicble (NA).

Metagenome assembly, binning and relative abundance calculation. Metagenomic library preparation and shotgun sequencing were conducted at the University of California Davis DNA Technologies Core. DNA was sequenced on the Illumina HiSeq4000 platform, producing 150-base pair (bp) paired-end reads with a targeted insert size of 400 bp. Quality control and adaptor removal were performed with Trimmomatic[56] (v.0.36; parameters: leading 10, trailing 10, sliding window of 4, quality score of 25, minimum length 151 bp) and Sickle[57] (v.1.33 with paired-end and Sanger parameters). Concatenation of high-quality reads for replicate samples (for coassembly) was conducted before assembly (see Supplementary Table 2 for more details on coassembly). The trimmed high-quality reads were assembled using metaSPAdes[58] (v.3.8.1; parameters $k = 21, 33, 55, 77, 88, 127$). The quality of assemblies was determined using QUAST[59] (v.5.0.2 with default parameters). Sequencing coverage (and differential coverage for coassemblies) was determined for each assembled scaffold by mapping high-quality reads to the assembly using Bowtie2 (ref. [60]) (v.2.3.4.1; default parameters) with Samtools[61] (v.1.7). Contigs greater than 2,500 bp were manually binned using Anvi'o with Centrifuge[62,63] (v.1.0.1) based on coverage uniformity (v.5). Quality metrics for metagenome-assembled genomes (MAGs) were determined using CheckM[64] (v.1.0.7; default parameters). The taxonomic identity of each MAG was determined using GTDB-Tk[65] (v.1.0.2) against The Genome Taxonomy Database[66] (https://data.ace.uq.edu.au/public/gtdb/data/releases/release89/89.0/, v.r89). The length-normalized relative abundance of MAGs was determined for each sample as by Tully et al.[67].

Metagenomic annotation of hydrocarbon degradation genes. Open reading frames were predicted for MAGs using Prodigal[68] (v.2.6.3; default parameters). Functional annotation was determined using HMMER3 (ref. [69]) (v.3.1b2) against the Pfam database[70] (v.31.0) with an expected value (e-value) cutoff of 1×10^{-7} and KofamScan (v.1.1.0)[71] against the hidden Markov model (HMM) profiles for Kyoto Encyclopedia of Genes and Genomes and Kegg Orthology (KEGG/KO) with a score cutoff of 1×10^{-7}. For Fig. 3h, to find the number of hits for *almA* we used Pfam (PF00743), for *rhdA* we used Pfam (PF00848) and for *pHMO* we summed Pfam hits (subunit a: PF02461; subunit b: PF04744; subunit c: PF04896).

For alkane-1-monooxygenase (*alkB*) we used both HMMER3 against the Pfam database (PF00487) and KofamScan against the KEGG/KO HMM profiles (K00496). Each hit was manually curated using Geneious Prime v.2019.2.3 (https://www.geneious.com) to search for the eight-histidine residues considered catalytically essential for function[72]. The base seed alignments for both PF00487 and K00496 include the ancestrally related protein, fatty acid desaturase; therefore, we found it necessary to phylogenetically analyse each hit to determine relation to *alkB* or fatty acid desaturase. Through this method we learned that the HMMER3 method with Pfam ID PF00487 identifies more hits within each MAG for *alkB* than KofamScan with K00496; however, those additional hits were generally more closely related to fatty acid desaturases than *alkB*. Furthermore, the phylogenies produced are necessary to determine similarity to the related xylene monooxygenase protein which acts on the methyl groups of xylene. We excluded any hits that formed a well-supported monophyletic clade with xylene monooxygenase from our final number of copies of *alkB*. In total, we used KofamScan with K00496 to search for *alkB*, manually curated the results to ensure presence of eight-histidine residues essential for function and phylogenetically analysed each hit for relation to *alkB* compared with fatty acid desaturase and xylene monooxygenase.

Phylogenetic analyses. Each putative *alkB* hit was aligned using MUSCLE[73] (v.3.8.425). For Extended Data Fig. 7, all manually curated hits for *alkB* in the Tara Oceans dataset were sequentially clustered by 90%, 80% and 60% identity using H-CD-HIT[74]. All columns with >95% gaps were removed using TrimAL[75] (v.1.2). Phylogenetic analysis of concatenated *alkB* was inferred by RAxML[76] (v.8.2.9; parameters: raxmlHPC -T 4 -s input -N autoMRE -n result -f a -p 12345 -x 12345 -m PROTCATLG). Resulting trees were visualized using FigTree[77] (v.1.4.3).

Extracting data from Tara Oceans dataset. To quantify the abundance of genes involved with hydrocarbon degradation we queried the Ocean Microbial Reference Gene Catalogue (OM-RGC) dataset (see http://ocean-microbiome.embl.de/companion.html) from the Tara Oceans expedition[36] for KEGG identifiers of interest. These included genes for the activation of alkanes such as alkane-1-monooxygenase (K00496), flavin-binding monooxygenase (K10215) and particulate hydrocarbon monooxygenase (K10944, K10945, K10946), as well as aromatic hydrocarbons such as toluene dioxygenase (K03268), naphthalene 1,2-dioxygenase (K14579, K14580, K14578, K14581), toluene methyl-monooxygenase (K15757 and K15758), p-cymene methyl-monooxygenase (K10616), benzene/toluene/chlorobenzene dioxygenase (K18089) and biphenyl 2,3-dioxygenase (K08689, K15750). We extracted the abundance of each gene from the Tara Oceans OM-RGC profiles dataset which was calculated from read counts mapped to each reference gene normalized by the gene length[36]. The abundance of select genes involved in hydrocarbon oxidation was analysed from the Tara Oceans dataset. The total abundance of OM-RGC sequences matching the reference gene identifier was normalized to the total abundance of the single-copy gene *recA* (KEGG identifier: K03553), as performed in previous studies, to

calculate abundance on a per-genome level[78,79]. The resulting data are included in Supplementary Data 1.

For select Tara Oceans stations we conducted further analysis of alkane-1-monooxygenase to assess the diversity and abundance of the gene for oceanographic settings underlying the North Atlantic subtropical gyre (Extended Data Figs. 7 and 8). First, we took the assembled Tara Oceans[36] data (see http://ocean-microbiome.embl.de/companion.html) and used Prodigal[68] (v.2.6.3; default parameters) to identify open reading frames. The resulting protein sequences were scanned for *alkB* using the above method (KofamScan for K00496, manual curation for eight-histidine residues and phylogenetically analysed). Each curated hit was assigned a taxonomic classification through homology search using BLAST[80] (v.2.7.1) against the nr database (v.38 accessed December 2019) (see resulting data in Supplementary Data 2). Read mapping of high-quality reads from each respective station using Bowtie2 (ref. [60]) (v.2.3.4.1) was used to determine the abundance of each unique *alkB*-like protein at each station.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

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

Received: 30 July 2020; Accepted: 17 December 2020;
Published online: 1 February 2021

References

1. *Oil in the Sea III* (National Research Council, 2003).
2. Han, J., McCarthy, E. D., Hoeven, W. V., Calvin, M. & Bradley, W. H. Organic geochemical studies II. A preliminary report on the distribution of aliphatic hydrocarbons in algae, in bacteria, and in recent lake sediment. *Proc. Natl Acad. Sci. USA* **59**, 29–33 (1968).
3. Li, X., del Cardayre, S. B., Popova, E., Schirmer, A. & Rude, M. A. Microbial biosynthesis of alkanes. *Science* **329**, 559–562 (2010).
4. Coates, R. C., Podell, S., Korobeynikov, A., Lapidus, A. & Pevzner, P. Characterization of cyanobacterial hydrocarbon composition and distribution of biosynthetic pathways. *PLoS ONE* **9**, 85140 (2014).
5. White, H. K. et al. Examining inputs of biogenic and oil-derived hydrocarbons in surface waters following the Deepwater Horizon oil spill. *ACS Earth Space Chem.* **3**, 1329–1337 (2019).
6. Lea-Smith, D. J. et al. Contribution of cyanobacterial alkane production to the ocean hydrocarbon cycle. *Proc. Natl Acad. Sci. USA* **112**, 13591–13596 (2015).
7. Chisholm, S. W. et al. A novel free-living prochlorophyte abundant in the oceanic euphotic zone. *Nature* **52**, 169–173 (1988).
8. Field, C. B., Behrenfeld, M. J., Randerson, J. T. & Falkowski, P. Primary production of the biosphere: integrating terrestrial and oceanic components. *Science* **281**, 237–240 (1998).
9. Flombaum, P. et al. Present and future global distributions of the marine Cyanobacteria *Prochlorococcus* and *Synechococcus*. *Proc. Natl Acad. Sci. USA* **110**, 9824–9829 (2013).
10. Karl, D. M. & Church, M. J. Microbial oceanography and the Hawaii Ocean Time-series programme. *Nat. Rev. Microbiol.* **12**, 699–713 (2014).
11. Polovina, J. J., Howell, E. A. & Abecassis, M. Ocean's least productive waters are expanding. *Geophys. Res. Lett.* **35**, 2–6 (2008).
12. Lea-Smith, D. J. et al. Hydrocarbons are essential for optimal cell size, division, and growth of Cyanobacteria. *Plant Physiol.* **172**, 1928–1940 (2016).
13. Cavender-Bares, K. K., Karl, D. M. & Chisholm, S. W. Nutrient gradients in the western North Atlantic Ocean: relationship to microbial community structure and comparison to patterns in the Pacific Ocean. *Deep Sea Res. I Oceanogr. Res. Pap.* **48**, 2373–2395 (2001).
14. Johnson, Z. I. et al. Niche partitioning among *Prochlorococcus* ecotypes along ocean-scale environmental gradients. *Science* **311**, 1737–1740 (2006).
15. Grande, K. D. et al. Primary production in the North Pacific gyre: a comparison of rates determined by the ^{14}C, O_2 concentration and ^{18}O methods. *Deep Sea Res. A Oceanogr. Res. Pap.* **36**, 1621–1634 (1989).
16. Karl, D. M. & Church, M. J. Ecosystem structure and dynamics in the North Pacific subtropical gyre: new views of an old ocean. *Ecosystems* **20**, 433–457 (2017).
17. Knoot, C. J. & Pakrasi, H. B. Diverse hydrocarbon biosynthetic enzymes can substitute for olefin synthase in the cyanobacterium *Synechococcus* sp. PCC 7002. *Sci. Rep.* **9**, 1360 (2019).
18. Martiny, A. C., Kathuria, S. & Berube, P. M. Widespread metabolic potential for nitrite and nitrate assimilation among *Prochlorococcus* ecotypes. *Proc. Natl Acad. Sci. USA* **106**, 10787–10792 (2009).
19. Saunois, M. et al. The global methane budget 2000–2017. *Earth Syst. Sci. Data* **12**, 1561–1623 (2020).
20. Guenther, A. B. et al. The model of emissions of gases and aerosols from nature version 2.1 (MEGAN2.1): an extended and updated framework for modeling biogenic emissions. *Geosci. Model Dev.* **5**, 1471–1492 (2012).
21. McGenity, T. J., Crombie, A. T. & Murrell, J. C. Microbial cycling of isoprene, the most abundantly produced biological volatile organic compound on Earth. *ISME J.* **12**, 931–941 (2018).
22. Vaulot, D., Marie, D., Olson, R. J. & Chisholm, S. W. Growth of *Prochlorococcus*, a photosynthetic prokaryote, in the equatorial Pacific. *Science* **268**, 1480–1482 (1995).
23. Mann, E. L. & Chisholm, S. W. Iron limits the cell division rate of *Prochlorococcus* in the eastern equatorial Pacific. *Limnol. Oceanogr.* **45**, 1067–1076 (2000).
24. Zubkov, M. V. Faster growth of the major prokaryotic versus eukaryotic CO_2 fixers in the oligotrophic ocean. *Nat. Commun.* **5**, 3776 (2014).
25. Liu, H. B., Campbell, L. & Landry, M. R. Growth and mortality rates of *Prochlorococcus* and *Synechococcus* measured with a selective inhibitor technique. *Mar. Ecol. Prog. Ser.* **116**, 277–288 (1995).
26. Head, I. M., Jones, D. M. & Larter, S. R. Biological activity in the deep subsurface and the origin of heavy oil. *Nature* **426**, 344–352 (2003).
27. Reddy, C. M. et al. Composition and fate of gas and oil released to the water column during the Deepwater Horizon oil spill. *Proc. Natl Acad. Sci. USA* **109**, 20229–20234 (2012).
28. Head, I. M., Jones, D. M. & Röling, W. F. M. Marine microorganisms make a meal of oil. *Nat. Rev. Microbiol.* **4**, 173–182 (2006).
29. Frysinger, G. S., Gaines, R. B., Xu, L. & Reddy, C. M. Resolving the unresolved complex mixture in petroleum-contaminated sediments. *Environ. Sci. Technol.* **37**, 1653–1662 (2003).
30. McKenna, A. M. et al. Unprecedented ultrahigh resolution FT-ICR mass spectrometry and parts-per-billion mass accuracy enable direct characterization of nickel and vanadyl porphyrins in petroleum from natural seeps. *Energy Fuels* **28**, 2454–2464 (2014).
31. Wardlaw, G. D. et al. Disentangling oil weathering at a marine seep using GCxGC: broad metabolic specificity accompanies subsurface petroleum biodegradation. *Environ. Sci. Technol.* **42**, 7166–7173 (2008).
32. Wang, W. & Shao, Z. Diversity of flavin-binding monooxygenase genes (*almA*) in marine bacteria capable of degradation long-chain alkanes. *FEMS Microbiol. Ecol.* **80**, 523–533 (2012).
33. van Beilen, J. B., Li, Z., Duetz, W. A., Smits, T. H. M. & Witholt, B. Diversity of alkane hydroxylase systems in the environment. *Oil Gas Sci. Technol.* **58**, 427–440 (2003).
34. Smits, T. H. M., Balada, S. B., Witholt, B. & Van Beilen, J. B. Functional analysis of alkane hydroxylases from Gram-negative and Gram-positive bacteria. *J. Bacteriol.* **184**, 1733–1742 (2002).
35. González-Gaya, B. et al. Biodegradation as an important sink of aromatic hydrocarbons in the oceans. *Nat. Geosci.* **12**, 119–125 (2019).
36. Sunagawa, S. et al. Structure and function of the global ocean microbiome. *Science* **348**, 6237 (2015).
37. Rinke, C. et al. A phylogenomic and ecological analysis of the globally abundant Marine Group II archaea (*Ca*. Poseidoniales ord. nov.). *ISME J.* **13**, 663–675 (2019).
38. Tully, B. J. Metabolic diversity within the globally abundant Marine Group II Euryarchaea offers insight into ecological patterns. *Nat. Commun.* **10**, 271 (2019).
39. Hazen, T. C., Prince, R. C. & Mahmoudi, N. Marine oil biodegradation. *Environ. Sci. Technol.* **50**, 2121–2129 (2016).
40. Prince, R. C., Amande, T. J. & McGenity, T. J. in *Taxonomy, Genomics and Ecophysiology of Hydrocarbon-Degrading Microbes* (ed. McGenity, T. J.) 1–39 (Springer, 2019).
41. Valentine, D. L. Adaptations to energy stress dictate the ecology and evolution of the Archaea. *Nat. Rev. Microbiol.* **5**, 316–323 (2007).
42. MacDonald, I. R. et al. Natural and unnatural oil slicks in the Gulf of Mexico. *J. Geophys. Res. Oceans* **120**, 8364–8380 (2015).
43. Ryerson, T. B. et al. Atmospheric emissions from the Deepwater Horizon spill constrain air-water partitioning, hydrocarbon fate, and leak rate. *Geophys. Res. Lett.* **38**, L07803 (2011).
44. Ryerson, T. B. et al. Chemical data quantify Deepwater Horizon hydrocarbon flow rate and environmental distribution. *Proc. Natl Acad. Sci. USA* **109**, 20246–20253 (2012).
45. Sorigué, D. et al. Microalgae synthesize hydrocarbons from long-chain fatty acids via a light-dependent pathway. *Plant Physiol.* **171**, 2393–2405 (2016).
46. Sorigué, D. et al. An algal photoenzyme converts fatty acids to hydrocarbons. *Science* **357**, 903–907 (2017).

47. Aleksenko, V. A. et al. Phylogeny and structure of fatty acid photodecarboxylases and glucose-methanol-choline oxidoreductases. *Catalysts* **10**, 1072 (2020).
48. Reshkin, S. J. & Knauer, G. A. Light stimulation of phosphate uptake in natural assemblages of phytoplankton. *Limnol. Oceanogr.* **24**, 1121–1124 (1979).
49. Van Mooy, B. A. S., Moutin, T., Duhamel, S., Rimmelin, P. & Van Wambeke, F. Phospholipid synthesis rates in the eastern subtropical South Pacific Ocean. *Biogeosciences* **5**, 133–139 (2008).
50. Lomas, M. W. et al. Increased ocean carbon export in the Sargasso Sea linked to climate variability is countered by its enhanced mesopelagic attenuation. *Biogeosciences* **7**, 57–70 (2010).
51. Durand, M. D. & Olson, R. J. Contributions of phytoplankton light scattering and cell concentration changes to diel variations in beam attenuation in the equatorial Pacific from flow cytometric measurements of pico-, ultra and nanoplankton. *Deep Sea Res. II Top. Stud. Oceanogr.* **43**, 891–906 (1996).
52. López-Sandoval, D. C., Delgado-Huertas, A. & Agustí, S. The ^{13}C method as a robust alternative to ^{14}C-based measurements of primary productivity in the Mediterranean Sea. *J. Plankton Res.* **40**, 544–554 (2018).
53. Kozich, J. J., Westcott, S. L., Baxter, N. T., Highlander, S. K. & Schloss, P. D. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Appl. Environ. Microbiol.* **79**, 5112–5120 (2013).
54. Caporaso, J. G. et al. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J.* **6**, 1621–1624 (2012).
55. Callahan, B. J. et al. DADA2: high-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**, 581–583 (2016).
56. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
57. Joshi, N. & Fass, J. Sickle: a sliding-window, adaptive, quality-based trimming tool for FastQ files v.1.33 (2011); https://github.com/najoshi/sickle
58. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. MetaSPAdes: a new versatile metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).
59. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).
60. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
61. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
62. Kim, D., Song, L., Breitwieser, F. P. & Salzberg, S. L. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res.* **26**, 1721–1729 (2016).
63. Eren, A. M. et al. Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* **2015**, e1319 (2015).
64. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
65. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925–1927 (2019).
66. Parks, D. H. et al. A complete domain-to-species taxonomy for Bacteria and Archaea. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-020-0501-8 (2020).
67. Tully, B. J., Wheat, C. G., Glazer, B. T. & Huber, J. A. A dynamic microbial community with high functional redundancy inhabits the cold, oxic subseafloor aquifer. *ISME J.* **12**, 1–16 (2018).
68. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
69. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, 1002195 (2011).
70. El-Gebali, S. et al. The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**, D427–D432 (2019).
71. Aramaki, T. et al. KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* **36**, 2251–2252 (2020).
72. Shanklin, J., Whittle, E. & Fox, B. G. Eight histidine residues are catalytically essential in a membrane-associated iron enzyme, stearoyl-CoA desaturase, and are conserved in alkane hydroxylase and xylene monooxygenase. *Biochemistry* **33**, 12787–12794 (1994).
73. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
74. Huang, Y., Niu, B., Gao, Y., Fu, L. & Li, W. CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* **26**, 680–682 (2010).
75. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
76. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
77. Rambaut, A. FigTree v.1.4.3 (2012); http://tree.bio.ed.ac.uk/software/figtree/
78. Sosa, O. A., Repeta, D. J., DeLong, E. F., Ashkezari, M. D. & Karl, D. M. Phosphate-limited ocean regions select for bacterial populations enriched in the carbon–phosphorus lyase pathway for phosphonate degradation. *Environ. Microbiol.* **21**, 2402–2414 (2019).
79. Martinez, A., Tyson, G. W. & Delong, E. F. Widespread known and novel phosphonate utilization pathways in marine bacteria revealed by functional screening and metagenomic analyses. *Environ. Microbiol.* **12**, 222–238 (2010).
80. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).

Acknowledgements

We thank G. Paradis for analytical support; the R/V *Neil Armstrong* captain and crew for support at sea; the Hawai'ian Oceanographic Time Series programme and the crew of the R/V *Kilo Moana* for enabling a preliminary study; J. Hayes and P. Chisholm for their advice on the study; J. Ossolinski, H. Fredricks, B. Jenkins and R. Swarthout for assistance on the R/V *Armstrong* cruise; T. McKinnon for assistance in dissolved inorganic carbon isotope measurements; A. Ebling for total particulate phosphate measurements; N. Poulton for flow cytometry analysis; and K. Krogslund for nutrient analysis. For bioinformatic analysis, this work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation (NSF) grant no. ACI-1548562. Specifically, it used the Bridges system, which is supported by NSF award no. ACI-1445606, at the Pittsburgh Supercomputing Center. This project was supported by NSF grant nos. OCE-1635562, OCE-1536346, OCE-1756254 and OCE-1634478.

Author contributions

C.R.L. carried out the organization of field experiments and measurements of pentadecane production and concentration, and corresponding data analysis. E.C.A. carried out the biodegradation experiments and bioinformatic work. B.A.S.V.M. contributed nutrient data, cell count data and sediment trap particles for experimentation. K.M.G. carried out the pentadecane concentration measurements. C.M.R. built the methodology for pentadecane quantification. R.K.N. carried out the two-dimensional gas chromatography quality checks. D.L.V. contributed towards experimental design and data interpretation.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41564-020-00859-8.

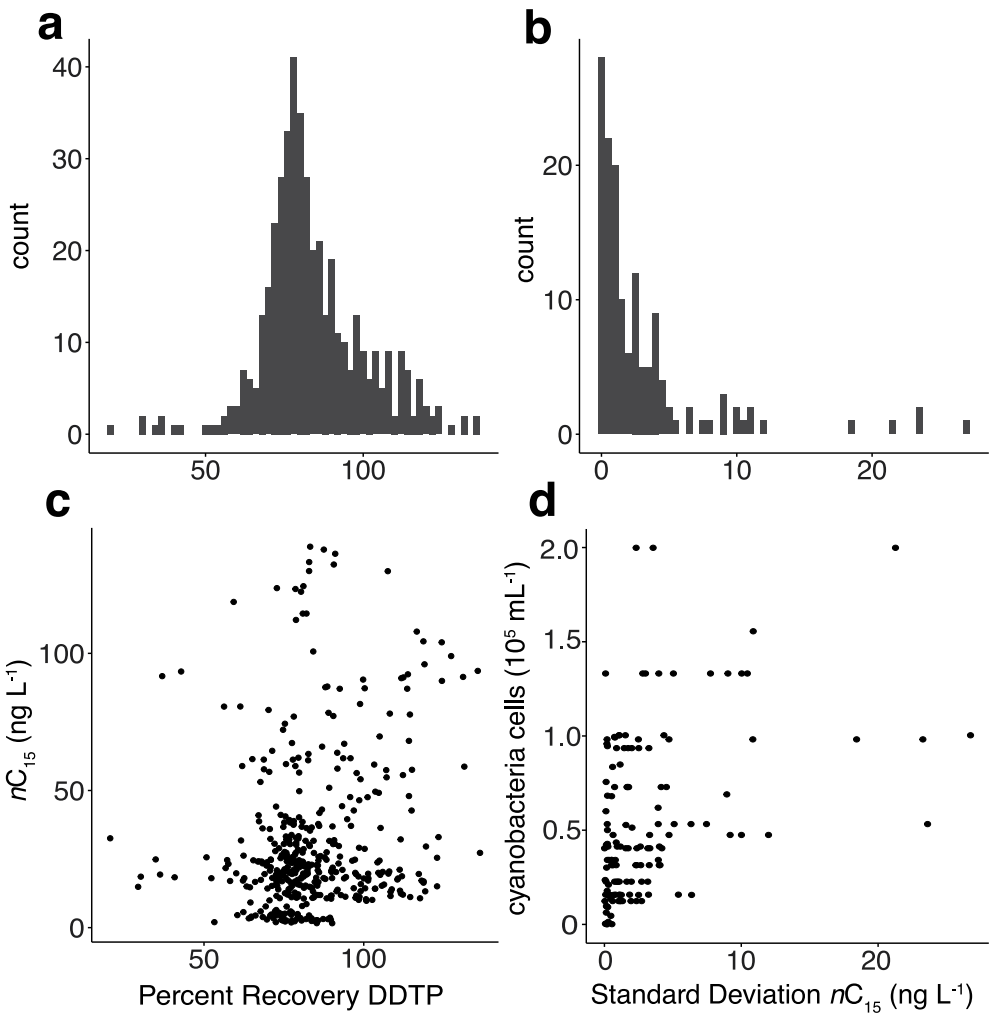
Supplementary information is available for this paper at https://doi.org/10.1038/s41564-020-00859-8.

Correspondence and requests for materials should be addressed to D.L.V.

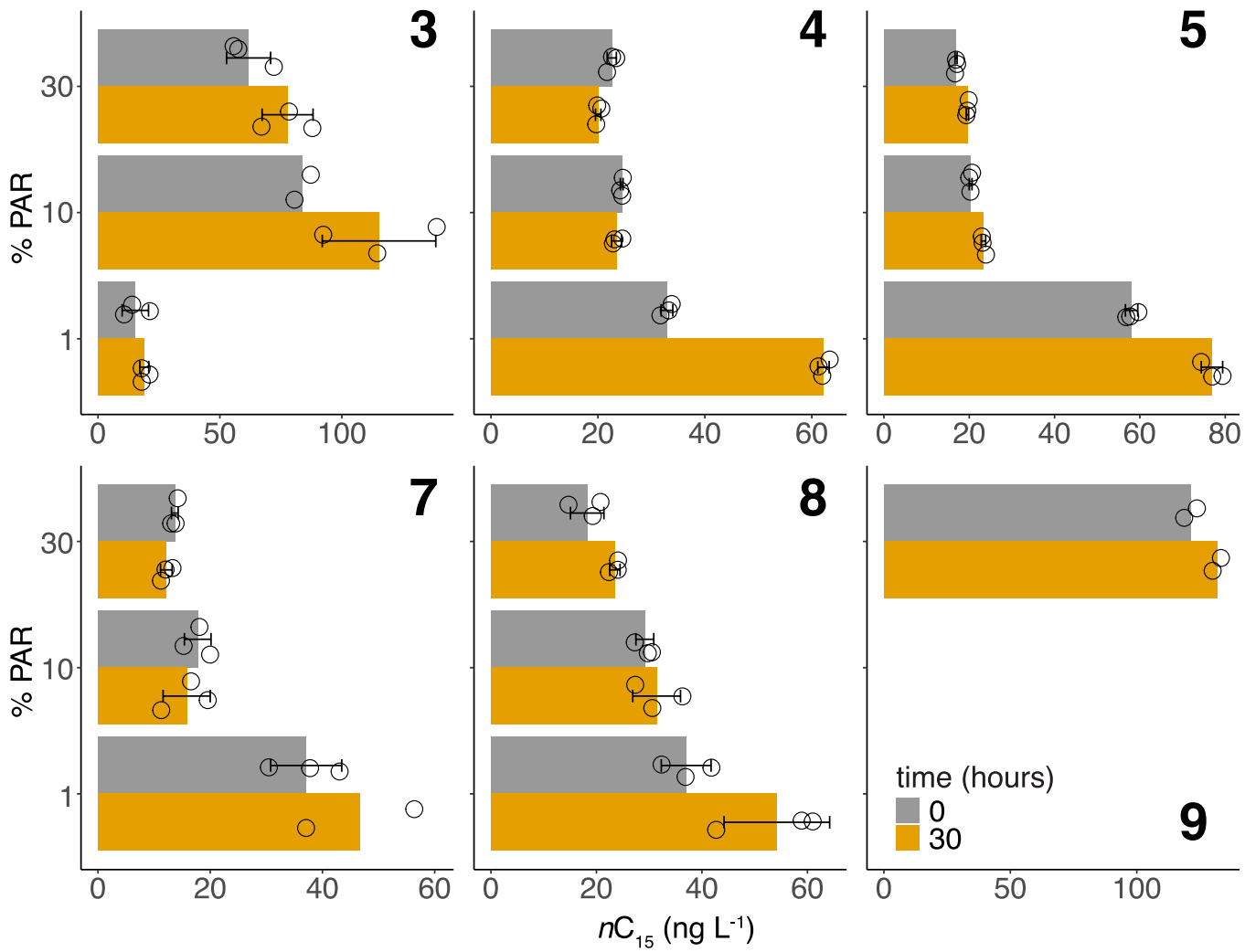
Peer review information *Nature Microbiology* thanks Alexandra Turchyn and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at www.nature.com/reprints.

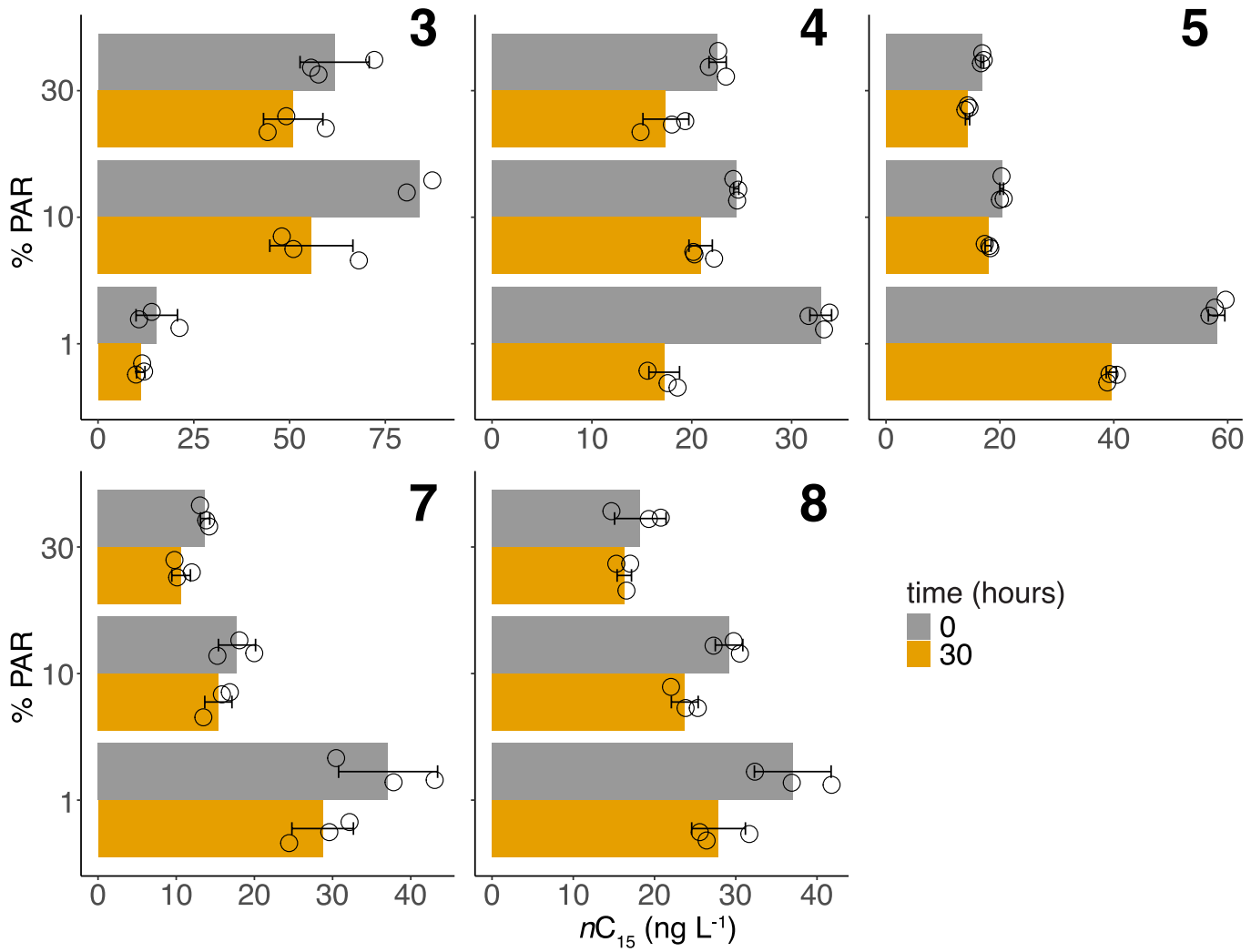
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



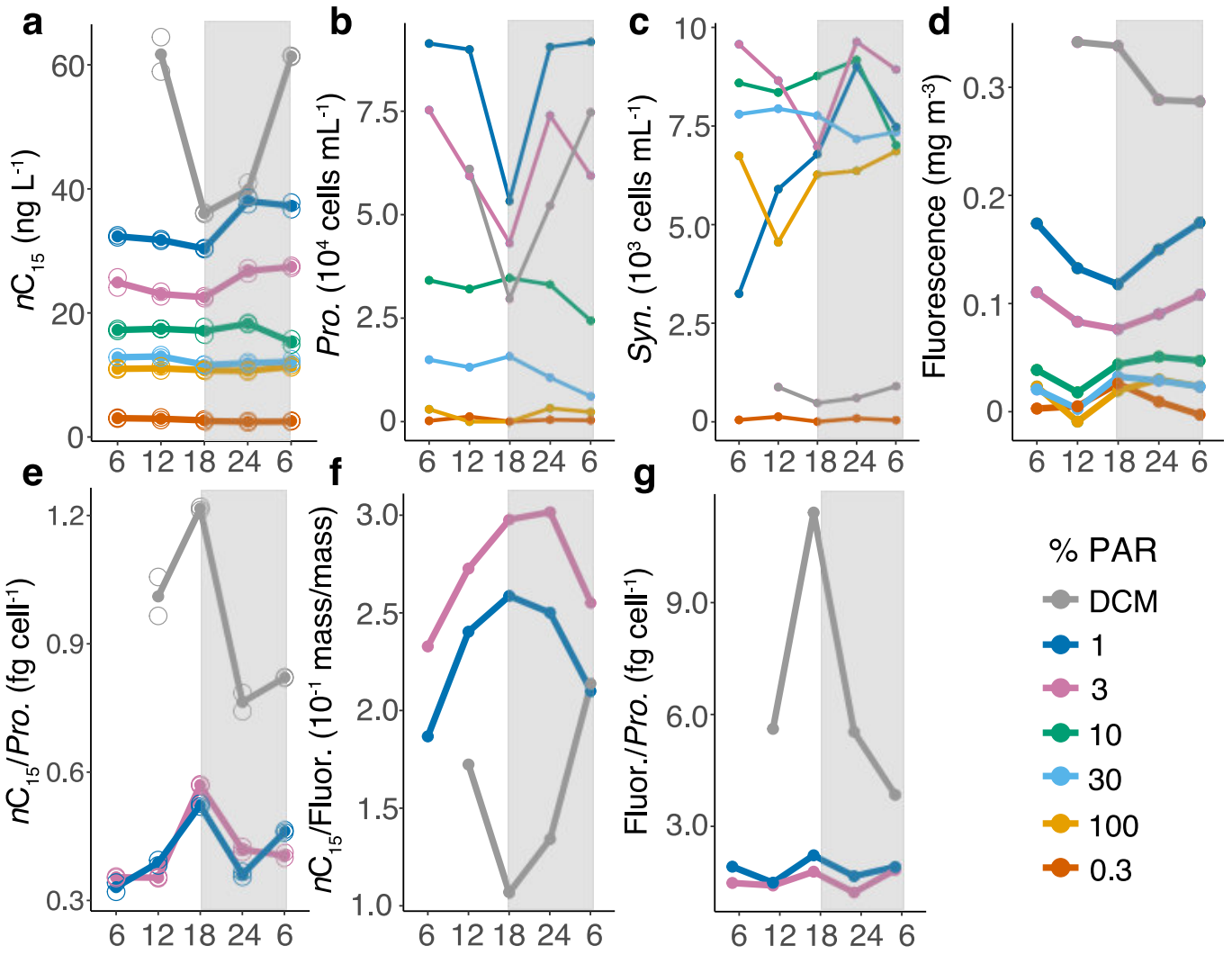
Extended Data Fig. 1 | Pentadecane extraction data quality check. a, Histogram of percent recovery of the internal standard (DDTP). **b**, Histogram of standard deviation of replicates of pentadecane concentration measurements; only a few replicates have a standard deviation > 15 ng L^{-1}. **c**, Pentadecane concentration data vs. percent recovery of DDTP; there is no coherent trend of greater recovery with higher concentration. **d**, Cyanobacterial cell abundance (*Pro.* + *Syn.*) vs. standard deviation of pentadecane concentration between replicates; points with high standard deviation and low cyanobacterial cell abundance were further investigated (see Supplementary Note).



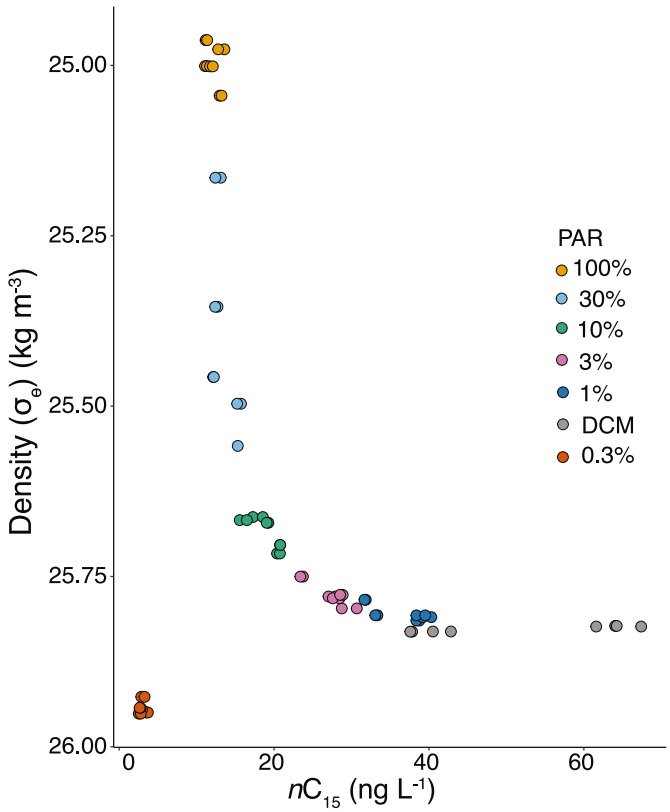
Extended Data Fig. 2 | Pentadecane concentration changes during light incubations. Concentration of pentadecane at beginning and end of 30-hour light incubations (time = 0 and 30 hours) at three light penetration depths for stations 3, 4, 5, 7, 8, 9 (indicated by number at right of each panel). Water was incubated at the light level from which it was collected (see Methods). Data are plotted as black open circles and represent biologically independent measurements; bar indicates mean of replicates at that light depth, error bars indicate standard deviation of n = 3 replication.



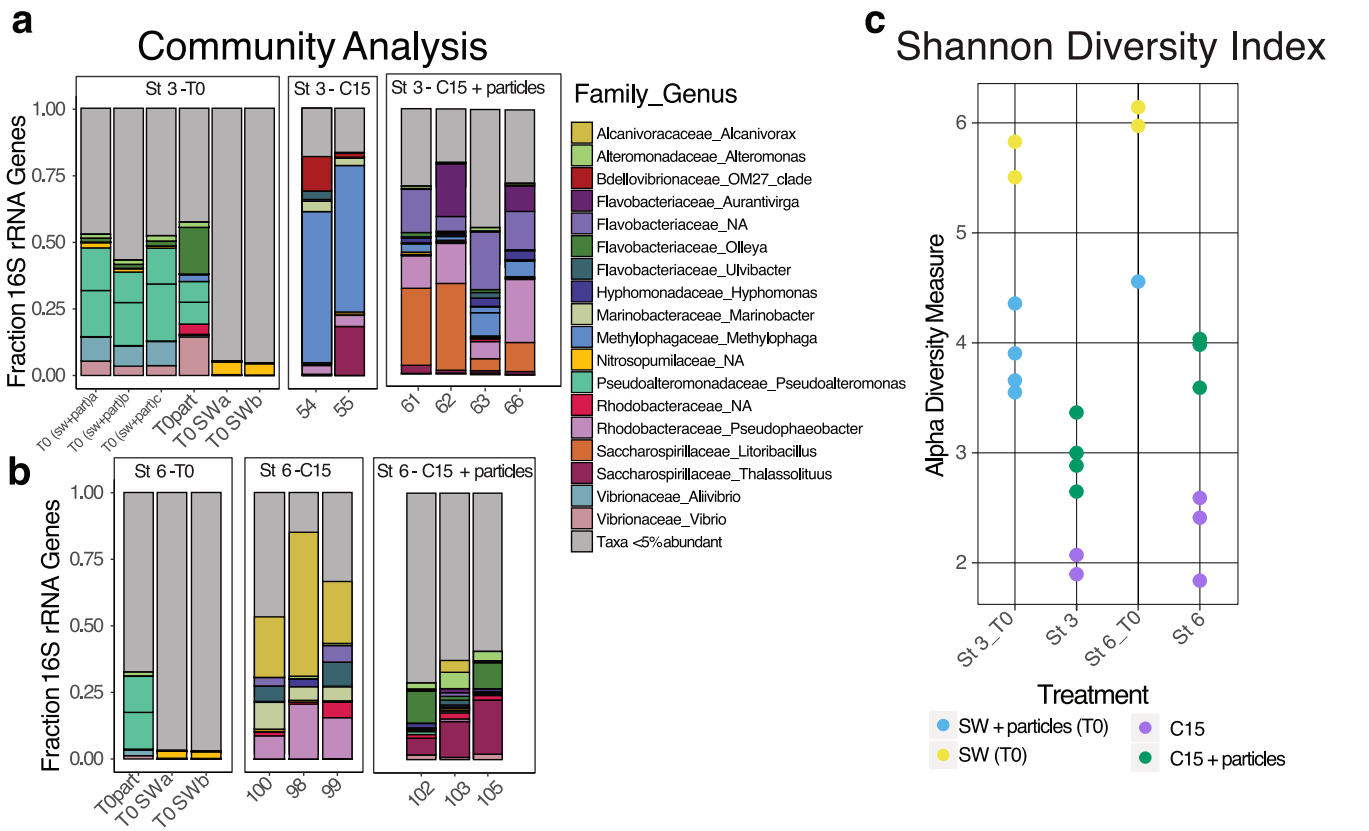
Extended Data Fig. 3 | Pentadecane concentration changes during dark control incubations. Concentration of pentadecane at beginning and end of 30-hour dark control incubations (time = 0 and 30 hours) at three light penetration depths for stations 3, 4, 5, 7, 8 (indicated by number at right of each panel). Data are plotted as black open circles and represent biologically independent measurements; bar indicates mean of replicates at that light depth, error bars indicate standard deviation of n = 3 replication.



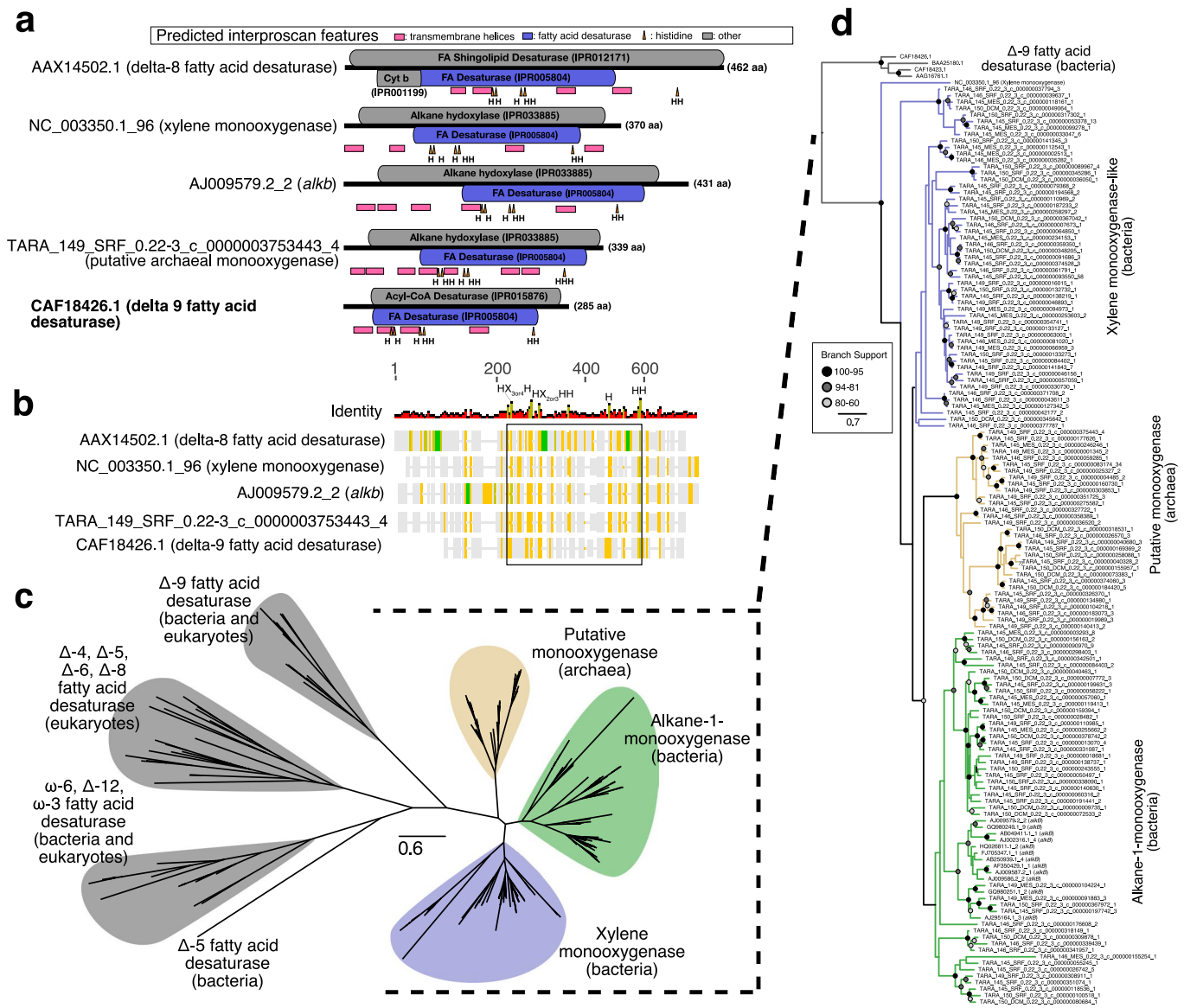
Extended Data Fig. 4 | Station 6 diel patterns across different light depths. Light depths kept constant through Lagrangian sampling framework whereas the DCM is a depth variable feature throughout the diel cycle (see Methods). The x-axis represents time of day in hours, with gray shading representing night. Diel patterns of **a** pentadecane, **b** *Prochlorococcus*, **c** *Synechococcus*, **d** fluorescence (averaged with 1-meter resolution data with 2 data points above and 2 data points below to smooth signal, n = 5) and **e-g** selected ratios (see Supplementary Note). **a**, **e**, Data are plotted as open circles with n = 2 biologically independent pentadecane measurements, solid circles indicate mean.



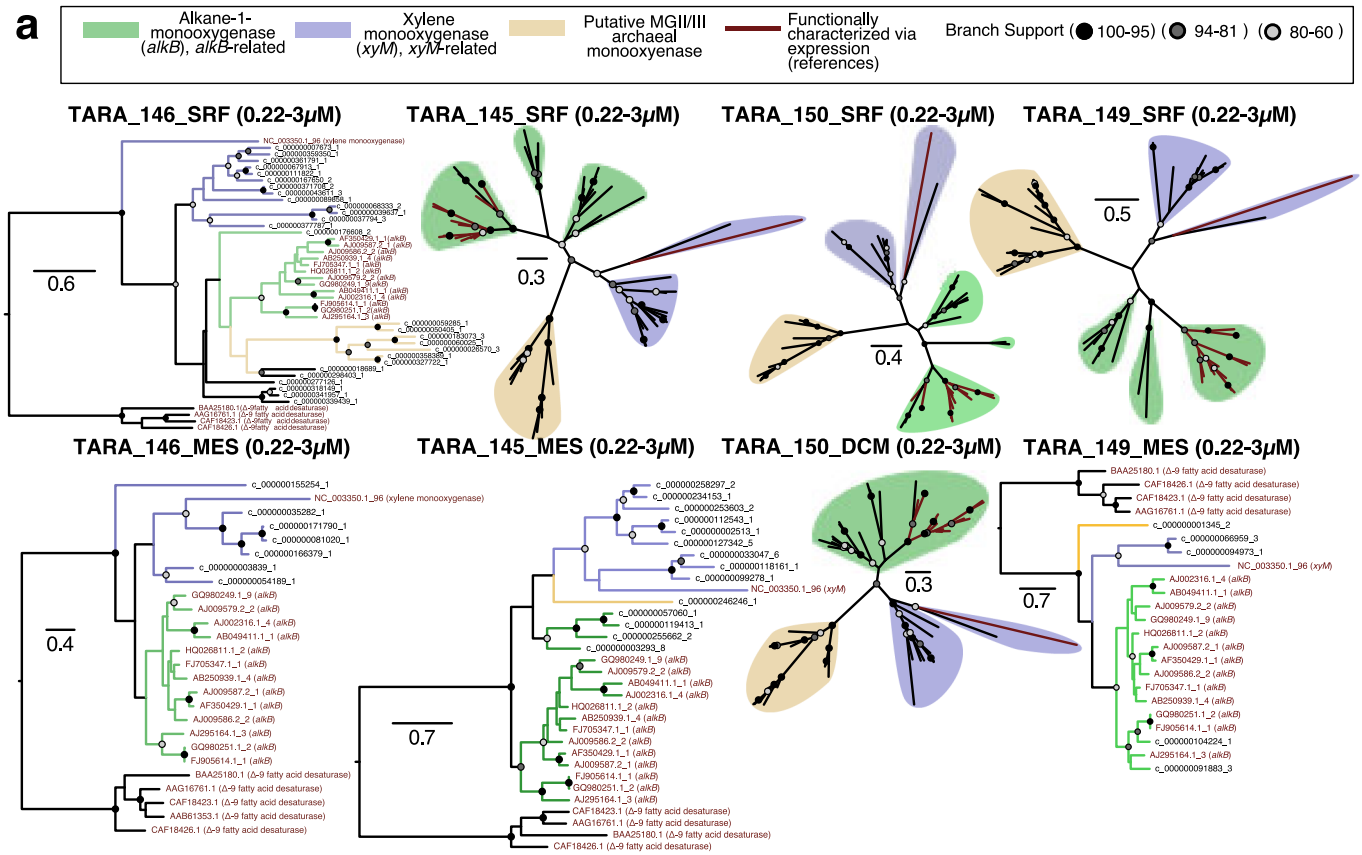
Extended Data Fig. 5 | Water mass proxies show consistent diel sampling for lower euphotic zone. Seawater density plotted against pentadecane concentration colored by light penetration depth and feature (DCM). In this plot, seawater density acts as a proxy for water mass identity in diel sampling. The closer the vertical spread of points of the same color means that samples are more likely to have originated from the same water mass, whereas the further spread means that samples may have originated from different water masses. The horizontal spread of points of the same color represents different concentrations of pentadecane found in the diel cycle. 3% PAR, 1% PAR and particularly the DCM, have pronounced changes in pentadecane over the diel cycle with minimal shifts in seawater density. We conclude this to mean that pentadecane patterns at these depths can be attributed to biological origin, rather than sampling of different water masses. Further information on sampling and data in Methods and Supplementary Note.



Extended Data Fig. 6 | Community composition and diversity in pentadecane incubations. a-b, Microbial community composition within pentadecane incubations informed via the V4 region of the 16 S rRNA gene for initial samples and those harvested at 27 days (station 3) and 29 days (station 6). Labels on x axis are sample IDs of biologically independent DNA samples with the following abbreviations (T0: time point 0, T0part: initial sediment trap particle community, T0SW (**a** and **b**): initial seawater community, T0sw + part (**a**, **b**, and **c**): initial seawater community immediately after particles added, #: pentadecane enrichment). Nucleotide variants are grouped by genus and are listed under associated family and genus; if genus is unclassified than it is listed as NA. All taxa less than 5% are aggregated and shaded gray. **c**, Shannon diversity index (see Supplementary Note) for each biologically independent DNA sample. Shannon indices for pentadecane (n = 2 at station 3, n = 3 at station 6) and pentadecane + particles (n = 4 at station 3, n = 3 at station 6).



Extended Data Fig. 7 | *AlkB* diversity underlying the North Atlantic subtropical gyre. Phylogenetic analysis of genes closely related to *alkB* from Tara Ocean dataset reveal bacterial and archaeal clades distinct from xylene monooxygenase and fatty acid desaturases. **a**, Protein domain architecture across select representatives of xylene monooxygenase, alkane-1 monooxygenase, fatty acid desaturase, and related proteins from the Tara Oceans dataset which share a core fatty acid desaturase-like region (blue) expanded on in panel **b**. **b**, Abbreviated protein alignment for phylogenetic analyses (for details see Methods). Each column of alignment figure represents a sliding window of 5 bp with the following identity to consensus sequence coloration: green (100%), mustard (80-99% similar), yellow (60-79% similar), gray (<60% similar). The black box represents the region containing the eight histidine residues considered catalytically essential which were used for phylogenetic analyses in panels **c**-**d**. **c**, Maximum-likelihood phylogenetic tree with scale bar of substitutions per site. For clarity, bootstrap values are not shown for the full tree. Δ-X indicates activity X carbons from the carboxylic end of the fatty acid and ω-X indicates activity X carbons from the methyl end of the fatty acid. **d**, Expanded subtree of membrane monooxygenases and delta-9 fatty acid desaturases (outgroup). Clade coloration in panel **d** is according to position in panel **c**. NCBI accession codes are given for functional representatives in the subtree (accession_ORF#).

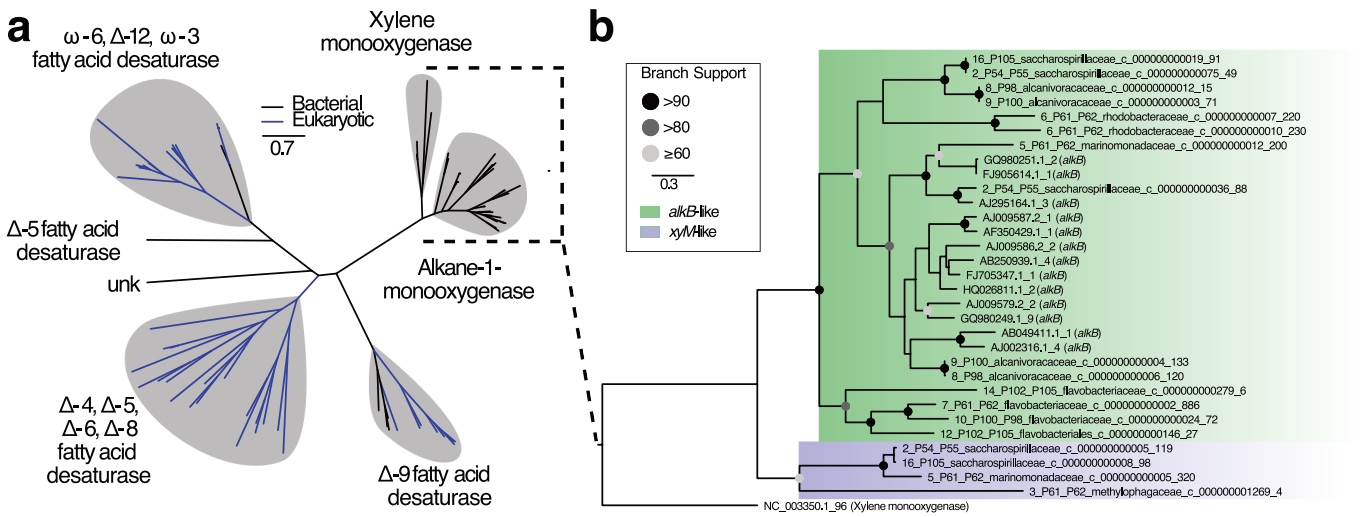


b

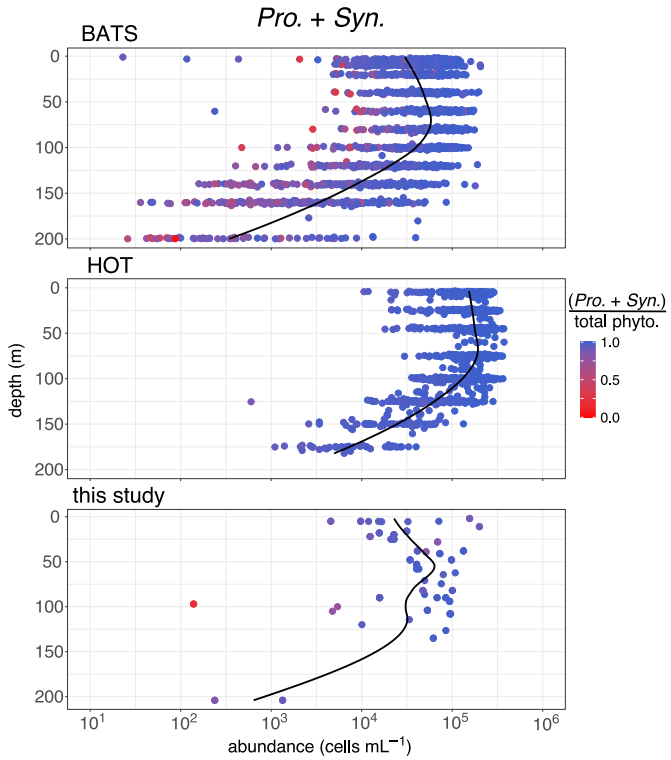
Sample ID	Latitude (degrees North)	Longitude (degrees East)	Sampling depth (m)	Abundance (% coverage)	Diversity (# unique)
TARA_146_MES_0.22-3	34.6663	-71.2907	640	0.0	0
				100.0	7
				0.0	0
TARA_146_SRF_0.22-3	34.6712	-71.3093	5	27.6	7
				53.7	13
				37.4	7
TARA_149_MES_0.22-3	34.0771	-49.8233	740	34.2	1
				36.4	2
				29.5	3
TARA_149_SRF_0.22-3	34.1132	-49.9181	5	39.0	15
				44.8	12
				16.2	8
TARA_145_MES_0.22-3	39.2392	-70.0343	590	8.0	1
				65.8	9
				26.2	4
TARA_145_SRF_0.22-3	39.2305	-70.0377	5	34.5	9
				24.8	14
				40.7	15
TARA_150_DCM_0.22-3	35.8427	-37.1526	40	40.7	14
				28.7	10
				30.6	13
TARA_150_SRF_0.22-3	35.9346	-37.3032	5	35.5	9
				30.0	9
				38.5	11

Clade Type
Archaea
Xylene
AlkB

Extended Data Fig. 8 | Putative archaeal *alkB* consistently abundant at surface and DCM stations. a, Maximum-likelihood phylogenetic analysis for each station with scale bar of substitutions per site. Clade designations as follows: green (alkane-1-monoxygenase representatives and related Tara hits), blue (xylene monooxygenase representative and related Tara hits), yellow (putative Marine Group II/III archaeal monooxygenase). See Supplementary Data 2 for homology search results for putative MG II/III monooxygenase hits. Trees <27 unknown Tara sequences are out-grouped with delta-9 fatty acid desaturases, whereas trees with >27 unknown Tara sequences are left unrooted. **b**, Meta-data for each Tara station and abundance of unique hits derived from read-mapping. % Coverage indicates the fraction of reads that map to genes within each clade (xylene monooxygenase, *alkB*, or archaeal monooxygenase) over the total reads mapped to all *alkB*-like, xylene-like, and archaeal monooxygenases found at each station.



Extended Data Fig. 9 | Phylogenetic confirmation of AlkB presence in MAGs from pentadecane incubations. a, Maximum likelihood tree of *alkB* hits within metagenomes compared to fatty acid desaturase, xylene monooxygenase, and *alkB* functionally expressed/characterized representatives (See Supplementary Note for identification details). Δ-X indicates activity X carbons from the carboxylic end of the fatty acid and ω-X indicates activity X carbons from the methyl end of the fatty acid. **b**, Expanded view of the alkane-1-monooxygenase, xylene monooxygenase, and related hits from metagenomes. Coloration in panel **b** is according to position in panel **a**. Gene copies for *alkB* in MAGS (in green) used in Fig. 3h.



Extended Data Fig. 10 | Proportional prokaryote contribution to phytoplankton community compared across time-series stations. Depth profiles of ~20 years of data from the Bermuda Atlantic Time-series (BATS, at top, data obtained from Bermuda Atlantic Time-series Study http://bats.bios.edu/bats-data/), the Hawaii Ocean Time-series (HOT, in middle, data obtained from Hawaii Ocean Time-series HOT-DOGS application; University of Hawai'i at Mānoa, National Science Foundation Award #1756517), and this study (at bottom). Data points are colored on a gradient by the proportional contribution to the phytoplankton community by *Prochlorococcus* and *Synechococcus* (total phytoplankton community is calculated as *Pro.* + *Syn.* + pico- + nano-Eukaryotes for BATS and this study, and *Pro.* + *Syn.* + pico-Eukaryotes for HOT). BATS and HOT data are each from a single station measured nearly monthly for ~20 years whereas measurements from this study incorporate spatial variability (see Fig. 1) with minimal temporal variability (all measurements taken in May 2017). The proportional contribution of *Prochlorococcus* and *Synechococcus* is >90 % of the phytoplankton community at BATS 84% of the time. At HOT, *Pro.* + *Syn.* is > 90 % of phytoplankton community ~100% of the time. For this study, *Pro.* + *Syn.* is >90% of the phytoplankton community (*Pro.* + *Syn.* + pico + nano-Eukaryotes) for 80% of the measurements, with most cases of lower proportional prokaryote abundance due to an anomalous nutrient pulse observed at station 9 (a *Synechococcus* bloom) or at low absolute abundance of *Pro.* + *Syn.*

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.

Terms and Conditions

Springer Nature journal content, brought to you courtesy of Springer Nature Customer Service Center GmbH (“Springer Nature”).
Springer Nature supports a reasonable amount of sharing of research papers by authors, subscribers and authorised users (“Users”), for small-scale personal, non-commercial use provided that all copyright, trade and service marks and other proprietary notices are maintained. By accessing, sharing, receiving or otherwise using the Springer Nature journal content you agree to these terms of use (“Terms”). For these purposes, Springer Nature considers academic use (by researchers and students) to be non-commercial.
These Terms are supplementary and will apply in addition to any applicable website terms and conditions, a relevant site licence or a personal subscription. These Terms will prevail over any conflict or ambiguity with regards to the relevant terms, a site licence or a personal subscription (to the extent of the conflict or ambiguity only). For Creative Commons-licensed articles, the terms of the Creative Commons license used will apply.
We collect and use personal data to provide access to the Springer Nature journal content. We may also use these personal data internally within ResearchGate and Springer Nature and as agreed share it, in an anonymised way, for purposes of tracking, analysis and reporting. We will not otherwise disclose your personal data outside the ResearchGate or the Springer Nature group of companies unless we have your permission as detailed in the Privacy Policy.
While Users may use the Springer Nature journal content for small scale, personal non-commercial use, it is important to note that Users may not:

1. use such content for the purpose of providing other users with access on a regular or large scale basis or as a means to circumvent access control;
2. use such content where to do so would be considered a criminal or statutory offence in any jurisdiction, or gives rise to civil liability, or is otherwise unlawful;
3. falsely or misleadingly imply or suggest endorsement, approval , sponsorship, or association unless explicitly agreed to by Springer Nature in writing;
4. use bots or other automated methods to access the content or redirect messages
5. override any security feature or exclusionary protocol; or
6. share the content in order to create substitute for Springer Nature products or services or a systematic database of Springer Nature journal content.

In line with the restriction against commercial use, Springer Nature does not permit the creation of a product or service that creates revenue, royalties, rent or income from our content or its inclusion as part of a paid for service or for other commercial gain. Springer Nature journal content cannot be used for inter-library loans and librarians may not upload Springer Nature journal content on a large scale into their, or any other, institutional repository.
These terms of use are reviewed regularly and may be amended at any time. Springer Nature is not obligated to publish any information or content on this website and may remove it or features or functionality at our sole discretion, at any time with or without notice. Springer Nature may revoke this licence to you at any time and remove access to any copies of the Springer Nature journal content which have been saved.
To the fullest extent permitted by law, Springer Nature makes no warranties, representations or guarantees to Users, either express or implied with respect to the Springer nature journal content and all parties disclaim and waive any implied warranties or warranties imposed by law, including merchantability or fitness for any particular purpose.
Please note that these rights do not automatically extend to content, data or other material published by Springer Nature that may be licensed from third parties.
If you would like to use or distribute our Springer Nature journal content to a wider audience or on a regular basis or in any other manner not expressly permitted by these Terms, please contact Springer Nature at

onlineservice@springernature.com